

**EGR361 Analysis of Engineering Data
Fall 2018
Course Handouts
Dr. Tammy VanDeGrift**

Name: _____

If found, call/email: _____

EGR361 Calendar Fall 2018

Calendar subject to change – see Moodle for latest updates. Complete assigned reading prior to lecture.

Week	Monday	Wednesday	Friday
Aug 27 – 31	Course Introduction, Data Reading: pages 1 - 22	Descriptive Statistics Reading: pages 23 – 52 DUE: HW 0	Presentation of Data Graphing Distributions
Sept 3 - 7	Random Variables Probability Reading: pages 57 - 65	Probability DUE: HW 1	Probability Practice
Sept 10 - 14	Discrete Random Variables PMF, CDF Reading: pages 97 - 102	Binomial Distribution Reading: pages 102 - 107	Poisson Distribution DUE: HW 2 Reading: pages 109 - 113
Sept 17 - 21	Continuous Random Variables PDF, CDF Reading: pages 66 - 73	Exponential Distribution Reading: pages 113 - 116	Normal Distribution Reading: pages 74 - 83
Sept 24 - 28	Review DUE: HW 3	EXAM 1: Chapters 1 – 3.9	Random samples Central Limit Theorem Reading: pages 136 - 140
Oct 1 - 5	Approximation of binomial and poisson to normal Reading: pages 119 - 122	Significant Figures	Error Propagation Reading: 129 – 135 DUE: Project Proposal
Oct 8 - 12	Statistical Inference Reading: 148 – 168	Inference on mean of one population, large sample z-test Reading: 169 – 184 DUE: HW 4	Inference on mean of one population, small sample t- test Reading: pages 186 – 196
Oct 15 – 19	FALL BREAK: NO CLASS (ENJOY!)		
Oct 22 - 26	Inference on variance of one population, chi-square Reading: pages 199 – 204	Inference on proportion of one population, z-test Reading: pages 205 - 214	Single Population Inference Practice DUE: Data Collection Plans
Oct 29 – Nov 2	Review DUE: HW 5	EXAM 2: Chapters 3.10 – 4	Inference on means of two populations, z-test and t-test Reading: pages 230 - 252
Nov 5 - 9	Inference on means of two populations, paired t-test Reading: pages 252 - 259	Inference on proportion of two populations, z-test Reading: 265 - 271	Two Population Inference Practice
Nov 12 - 16	One-way ANOVA, Excel Data Analysis Toolkit BRING LAPTOP Reading: pages 272 - 288	Two-way ANOVA, Scatterplots BRING LAPTOP Reading: pages 46 – 52	Review DUE: HW 6
Nov 19 - 23	EXAM 3: Chapter 5	Project Work Time DUE: Project Data Collected	Thanksgiving Break: No class
Nov 26 – 30	Regression BRING LAPTOP Reading: pages 298 – 341 DUE: HW 7	Regression BRING LAPTOP	Project Work Time
Dec 3 - 7	Project Presentations DUE: Project	Course Evaluations Project Presentations	Review for Final Exam DUE: HW 8
Mon, Dec 10, 1:30 – 3:30 pm	Final Exam: Chapters 1 – 6 Time Set By Registrar: Please make travel plans after this date and time		

EGR361: Analysis of Engineering Data

Fall 2018

Course Information

Instructor: Dr. Tammy VanDeGrift

Email: vandegri@up.edu

Office Phone: 503-943-7256

Office: Shiley 223

Website: Course information on Moodle, learning.up.edu

Meetings: MWF 9:15 – 10:10 am

Classroom: Shiley 124

Office Hours: Tentative: M 2:30 – 4:30pm, T 9:30 – 11:30 am; W 8:30 – 9:00am; F 12:30-1:30pm

Bulletin Description: Basic probability and statistical procedures used in the analysis of engineering data and an understanding in measurement. Methods for displaying data, commonly used probability distributions for discrete and continuous random variables, and statistical tools such as simple linear regression are presented. Students are introduced to concepts of statistical experimental design and error mitigation. (Prerequisite: MTH 202)

Student Outcomes

At the end of the course, students should be able to:

- Organize, summarize, present, and analyze data graphically (e.g., Box Plots, and Histograms), and calculate mean, variance, standard deviation, median, and outliers. Ability to perform these using Excel.
- Calculate likelihood of events using probability theory.
- Analyze discrete probabilistic processes using discrete random variables, PMFs and CDFs including Binomial and Poisson distributions.
- Analyze continuous probabilistic processes using continuous random variables, PDFs and CDFs including the Normal and Exponential distributions.
- Explain random samples and the Central Limit Theorem.
- Describe measurement issues (e.g., systematic vs. random error) and calculate error propagation.
- Perform statistical inference on the mean, proportion, and variance of one population using Hypothesis Testing and Confidence Intervals (including z-test, t-test, chi-square test, and p-value).
- Perform statistical inference on the means and proportions of two populations using Hypothesis Testing and Confidence Intervals (including z-test and the Paired t-test).
- Present paired “x-y” data in Scatter Plots and perform Simple Linear Regression.
- Propose a research question, collect appropriate data, and analyze the data to answer the research question.

These goals will be accomplished by:

- Completing homework assignments, preparing and taking exams, and completing a project
- Participating in class discussion and group activities through regular class attendance and in-class collaborative learning
- Seeking help of professor and classmates when necessary
- Communicating ideas orally, graphically, and in writing
- Providing help rather than giving answers to classmates seeking help

Course Philosophy

General: This course is designed to introduce concepts related to probability theory and statistical inference (data science). Note that the course assumes no background in probability and statistics; if you have already taken such a course in high school, you may wish to talk to the Associate Dean about challenging this course. ***Because this course covers topics that build on one another, it is critical that you keep up with the material by completing assignments on***

time and preparing for each class session by reading and viewing assigned material. It's okay to struggle with the concepts. I expect students to be challenged, but it is your responsibility to seek help when you are confused. You will **not** succeed in this course if you start homework assignments the night before they are due.

Code of Academic Integrity: Academic integrity is openness and honesty in all scholarly endeavors. The University of Portland is a scholarly community dedicated to the discovery, investigation, and dissemination of truth, and to the development of the whole person. Membership in this community is a privilege, requiring each person to practice academic integrity at its highest level, while expecting and promoting the same in others. Breaches of academic integrity will not be tolerated and will be addressed by the community with all due gravity. See University Bulletin for policy. (from *UP Bulletin*)

Seeking Help: I expect you to have questions as you learn the course material. You may receive help from classmates (see below about Collaborative Learning) and seek help from the instructor. I encourage you to ask questions during lecture meetings and attend office hours.

Collaborative Learning: Your classmates are a huge resource available to you. Because we understand material in different ways, I encourage you to discuss concepts from the course with your classmates, but ***any work that you turn in must be your own. Unacknowledged copying or using parts of someone else's work, even if it has been modified by you, is plagiarism and is not acceptable.*** When you work with others on homework and projects, *please acknowledge problems in which you received help and by whom.* An acceptable way to collaborate is to discuss problems and potential solutions and then writing the solutions **on your own.** When giving help to classmates, do not give them the answer and do not show them your answer. Instead, ask questions to learn of their understanding and give conceptual explanations - this practice will help you master the material yourself. Remember: you must turn in work that is your own and conceived from your own brain, you must acknowledge the people who helped you, and you are encouraged to seek help when you are confused. Some class sessions will be used to work on problems similar to homework; some work sessions will be collaborative.

Instructor's and Students' Responsibility: In this course, the instructor's job is to guide you in learning about probability and statistics. In addition to traditional lecturing, I will have regular discussions and activities during lectures. I expect your full participation and readiness to learn at all class meetings. Every student learns in a different way; therefore, the instructor will include a variety of activities in the course.

Classroom Conduct: The Shiley School of Engineering is committed to developing and actively protecting a classroom environment in which respect must be shown to everyone in order to facilitate and encourage the expression, testing, understanding and creation of a variety of ideas and opinions. Failure to meet these standards will result in removal from the class session.

In order to maintain a positive learning environment, students should avoid disruptive behaviors such as: receiving cell phone calls during class, leaving class early or coming to class habitually late, talking out of turn, doing assignments for other classes, reading the newspaper, sleeping, and engaging in other activities that detract from the classroom learning experience.

The Learning Commons: The Learning Commons, located in Buckley Center 163, offers a variety of peer tutoring programs that facilitate your active learning and mastery of skills and knowledge. For questions about the Learning Commons, please send all correspondence to Jeffrey White, Administrator, at white@up.edu. The Learning Commons is a program of the Shepard Academic Resource Center.

Math Resource Center: Monday through Thursday, 6:00 p.m. through 9:00 p.m. during the first week of classes. Regular shifts begin the Sunday after the first week. For course-specific schedule; visit www.up.edu/learningcommons, or the reception desk in BC 163.

Writing Assistance: Start brainstorming ideas for your paper with a Writing Assistant. Visit www.up.edu/learningcommons to access our Writing Center schedule.

The Language Studio: Language assistance hotlines to schedule a time to meet throughout the semester at chinesetutor@up.edu, frenchtutor@up.edu, germantutor@up.edu, or spanishtutor@up.edu.

Natural Sciences Center: Please send a request to meet to biotutor@up.edu, chemtutor@up.edu, or physicstutor@up.edu.

Speech & Presentation Lab: Improve your presentations by requesting an appointment at speech@up.edu.

Group Work Lab: Make an appointment for your group project at groupwork@up.edu.

Nursing Tutoring: Our peer tutors for pathophysiology will begin providing peer support in BC 163 during the first week of classes to help you start the semester on the right path. Tutoring is available on a walk-in or appointment basis. Up-to-date schedule information is at www.up.edu/learningcommons/nursing.

Economics and Business Tutoring: For support in economics, OTM, finance, accounting, and business law courses, send requests for appointments to your discipline's respective tutor email hotline: econtutor@up.edu, otmtutor@up.edu, financetutor@up.edu, accountingtutor@up.edu, or bizlaw@up.edu.

Learning Assistance Counselor: Learning assistance counseling is also available in BC 163. The counselor teaches learning strategies and skills that enable students to become more successful in their studies and future professions. The counselor provides strategies to assist students with reading and comprehension, note-taking and study, time management, test-taking, and learning and remembering. Appointments can be made in the on-line scheduler available to all students in Moodle or during posted drop-in hours.

Assessment of Learning

I will assess your learning and mastery based on your submitted work, including homework assignments, exams, a project, a presentation, and in-class activities. Generally, the assignments and in-class activities are your chance to learn, while the exams and project are the main way I know you have learned the material. I expect you to submit your homework by the due date and time. If circumstances arise (e.g. you are ill for an extended period of time, you are out of town for a university-related activity) that prevent you from submitting your work on time, please discuss the reason with me *before* the due date.

Grading Scheme

Course grades will be calculated as follows, although the exact weights may change somewhat, if appropriate. Individual homework assignments could be weighted differently depending on level of difficulty, number of problems, and other factors.

- 20% - Homework Assignments
- 10% - Midterm #1
- 10% - Midterm #2
- 10% - Midterm #3
- 30% - Final Exam
- 10% - Project + Presentation
- 10% - Professionalism + Participation

Homework Assignments: There will be approximately 8 weekly or bi-weekly homework assignments, which will have short answer questions and/or ask you to use Excel. These assignments are intended to help you learn the material.

Homework assignments should be done individually unless otherwise specified. Assignments should be done neatly using a word processor or completed neatly by hand. All homework is submitted electronically (either scan your work or use a word processor to typeset your work).

Exams: Exams are intended to serve as learning tools in addition to helping me evaluate your mastery of concepts. There are 3 midterm exams and one final exam. The midterms will be given in class during the regular lecture hour. The final exam is scheduled by the registrar and cannot be moved to a different time.

Project: Complete a research project that includes collecting or acquiring data and performing appropriate analysis. Projects should be completed in pairs, due to safety and rigor in data collection. Each project will be presented at the end of the semester.

Professionalism and Participation: We will have regular discussions and activities during lecture. Some of these may include items you turn in for credit. Attendance will be taken at all class sessions. You are expected to treat class times as professional work meetings: be prompt, be respectful of your peers, learn with and from your peers, be prepared. If you already know the content of the course, be patient with other students. If you are ill and need to miss class, you should email Tammy before the class session to excuse the absence and find a classmate who can supply notes to you. Readings and videos should be completed prior to class sessions.

Final Grades: Course grades will be assigned based on the total points you earn during the semester divided by the total possible points. The minimum cutoffs for grades will not change. I do reserve the right to raise your grade, but the following minimum percentages are guaranteed. (For example, if you earn 90% of the points, you will get an A-. If you earn 89% of the points, you earn a B+ but I reserve the right to raise your grade to an A-.)

- | | |
|------------------|------------------|
| • $\geq 93\%$ A | • $\geq 73\%$ C |
| • $\geq 90\%$ A- | • $\geq 70\%$ C- |
| • $\geq 87\%$ B+ | • $\geq 67\%$ D+ |
| • $\geq 83\%$ B | • $\geq 63\%$ D |
| • $\geq 80\%$ B- | • $\geq 60\%$ D- |
| • $\geq 77\%$ C+ | • $< 60\%$ F |

Late Assignments: You are granted **two late days** to use at your discretion for submitting late homework assignments without penalty. For example, you may choose to turn in two different assignments 24 hours after their due dates and times. You could choose to use the two days (48 hours) on a single assignment. Weekends count as regular days. If you use a late day, indicate the use of late day(s) with a note on the assignment when you submit your work. If you need to submit an assignment late (due to illness, death in the family, out of town for a university event) and you have already used your two late days, please contact Tammy before the due date and time and we can discuss your options. Late days may **not** be used for exams, the project, and the project presentation.

Logistics

Textbook: The required textbook for the course is *Engineering Statistics* by Montgomery, Runger, and Hubele (Wiley 5th edition). [ISBN: 978-0470631478] The text is available for purchase at the campus bookstore. I ask that you read certain chapters or sections *before* attending the accompanying class session (see the online course calendar for the latest updates to the readings).

Course Calendar and Website: The latest version of the calendar and course materials are posted to Moodle (learning.up.edu). The course calendar lists the lecture topics, assigned readings, exams, and due dates for assignments. The calendar is subject to change as the semester progresses, so check the on-line version frequently.

Accessibility Statement: The University of Portland endeavors to make its courses and services fully accessible to all students. Students are encouraged to discuss with their instructors what might be most helpful in enabling them to meet the learning goals of the course. Students who experience a disability are also encouraged to use the services of the Office for Accessible Education Services [AES], located in the Shepherd Academic Resource Center (503-943-8985). If you have an AES Accommodation Plan, you should make an appointment to meet with your faculty member to discuss how to implement your plan in this class. Requests for alternate location for exams and/or extended exam time should, where possible, be made two weeks in advance of an exam, and must be made at least one week in advance of an exam. Also, you should meet with your faculty member to discuss emergency medical information or how best to ensure your safe evacuation from the building in case of fire or other emergency.

Non-Violence Statement: The University of Portland is committed to fostering a community free from all forms of violence in which all members feel safe and respected. Violence of any kind, and in particular acts of power-based personal violence, are inconsistent with our mission. Together, we take a stand against violence. Join us in learning more about campus and community resources and reporting options, along with our prevention strategy - Green Dot. To get involved or request more information, please contact Tiger Simpson at simpson@up.edu.

Assessment Disclosure Statement: Student work products for this course may be used by the University for educational quality assurance purposes.

Academic Regulation Statement: Policies governing your coursework at the University of Portland can be found in the University Bulletin.

Mental Health Statement: As a college student, you may sometimes experience problems with our mental health that interferes with academic experiences and negatively impact daily life. If you or someone you know experiences mental health challenges at UP, please contact the University of Portland Health and Counseling Center in Orrico Hall at www.up.edu/healthcenter/ or 503-943-7134. Their services are free and confidential, and if necessary they can provide same day appointments. Also know that the University of Portland Public Safety Department (503-943-4444) has personnel trained to respond sensitively to mental health emergencies at all hours. Remember that getting help is a smart and courageous thing to do – for yourself, for those you care about, and for those who care about you.

Improving the Course: I welcome your feedback about the course at any time. I may ask for your feedback periodically and you will have the opportunity to evaluate the course at the end of the semester.

HANDOUTS

Getting Excel Functions for Statistics (get these working at beginning of semester)

1. Open excel.
2. Go to File->Options.
3. In the dialog window, click Add-ins.
4. Under Inactive Application Add-ins, choose Analysis ToolPak. (If you have already added it, it show appear in the Active Application Add-ins list).
5. Click Go...
6. Check the box for Analysis ToolPak.
7. Click OK.

You should now have a Data Analysis button under Data in excel.

Here are the functions that will be useful for you:

- Anova: Single Factor
- Anova: Two-Factor With Replication
- Anova: Two-Factor Without Replication
- Correlation
- F-Test Two-Sample for Variances (we did not cover explicitly in course, but this is the test for two variances from two populations)
- Regression
- Z-test: Two-Sample for Means
- T-Test: Paired
- T-Test: Two-sample assuming equal variances
- T-Test: Two-sample assuming unequal variances

Introduction to Data

What is data?

How is data used in engineering?

Bar chart and histogram activity with majors and birth months.

Terms to think about:

Sample

Population

Random Sample

Histogram

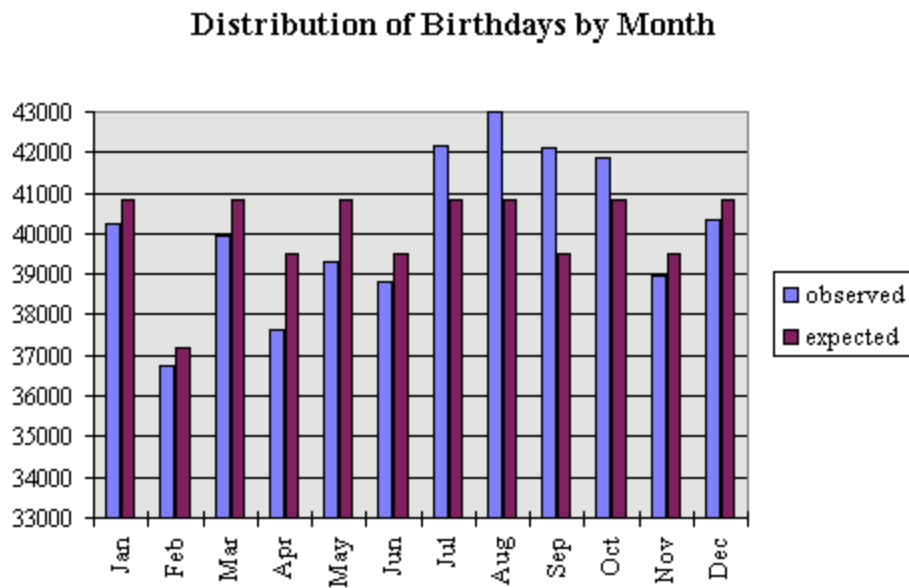


Figure 1: Expected and Actual Number of People by Birth Month; Data: 480,040 insurance policies made between 1981 through 1994. By Ron Murphy.

Does anything surprise you?

What would a uniform expected distribution look like?

What conclusions can you draw from this distribution?

Types of Studies

Enumerative study: where the sample is drawn from the population from which conclusions are drawn.

Analytic study: where the sample is drawn from a population that will be used to predict something about the future population.

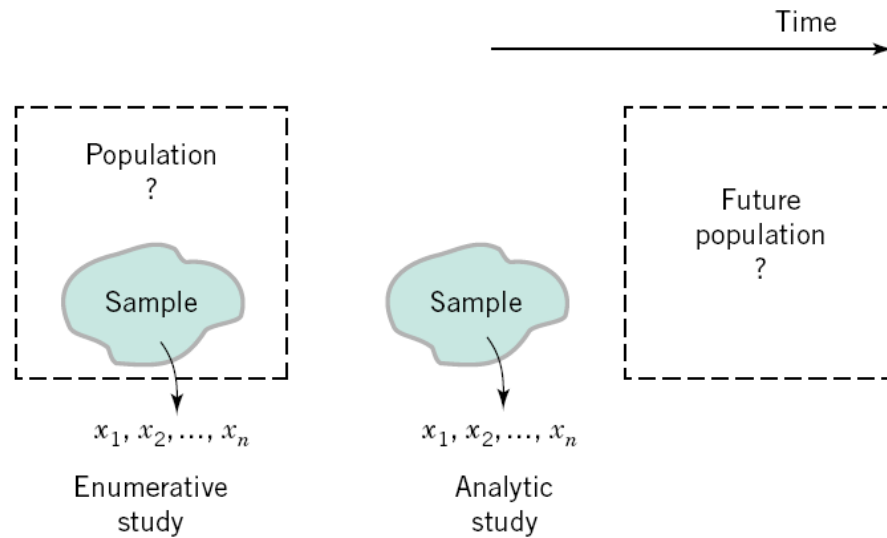


Figure 1-10 Enumerative versus analytic study.

Suppose the population is Shiley students. Below are studies' research questions. Determine if the study is enumerative or analytic.

Study question 1: What is the distribution of home states of current Shiley students?

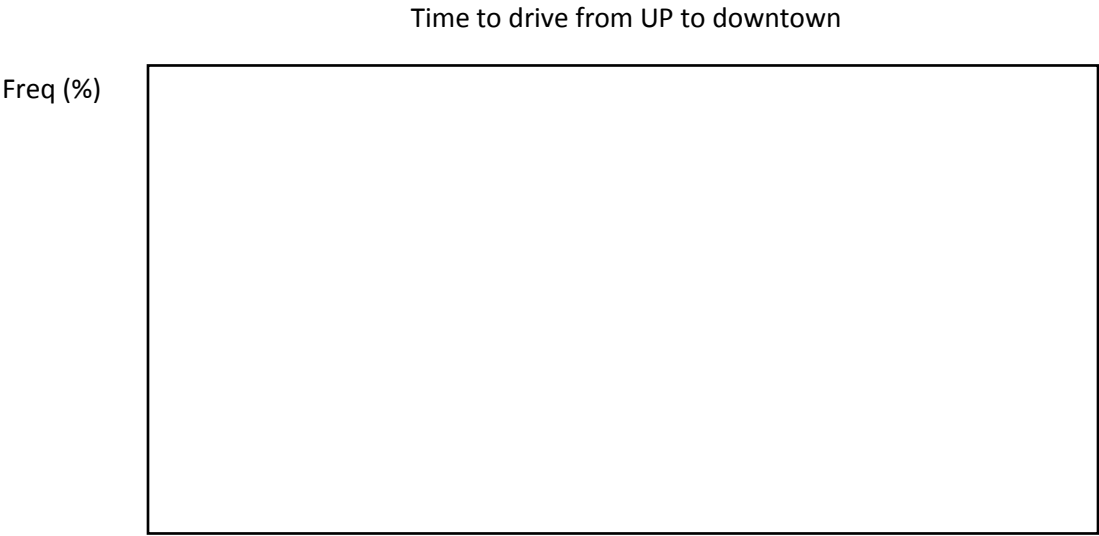
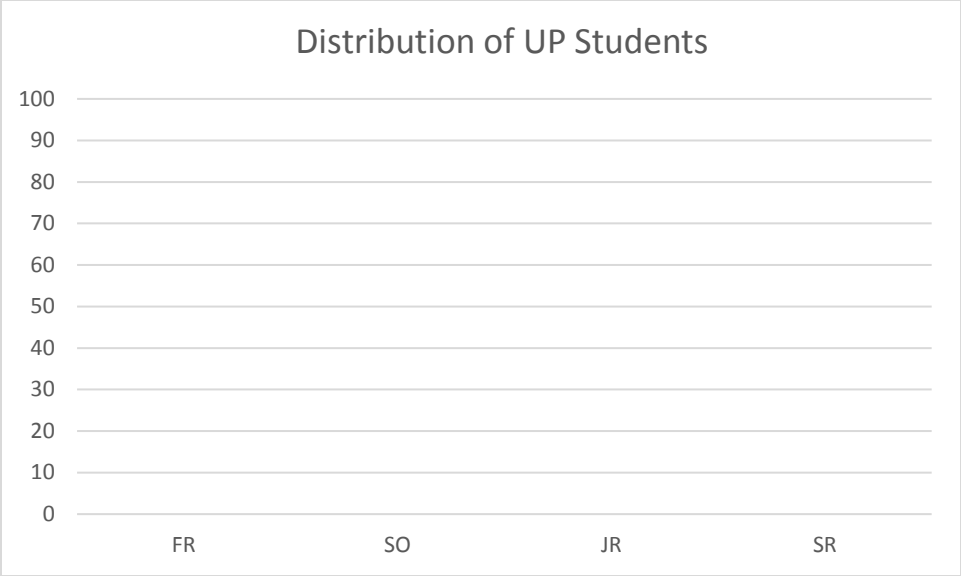
Study question 2: Suppose the Shiley School wants to grow by 20% in enrollment in the next five years. What will be the distribution of home states of Shiley students in five years?

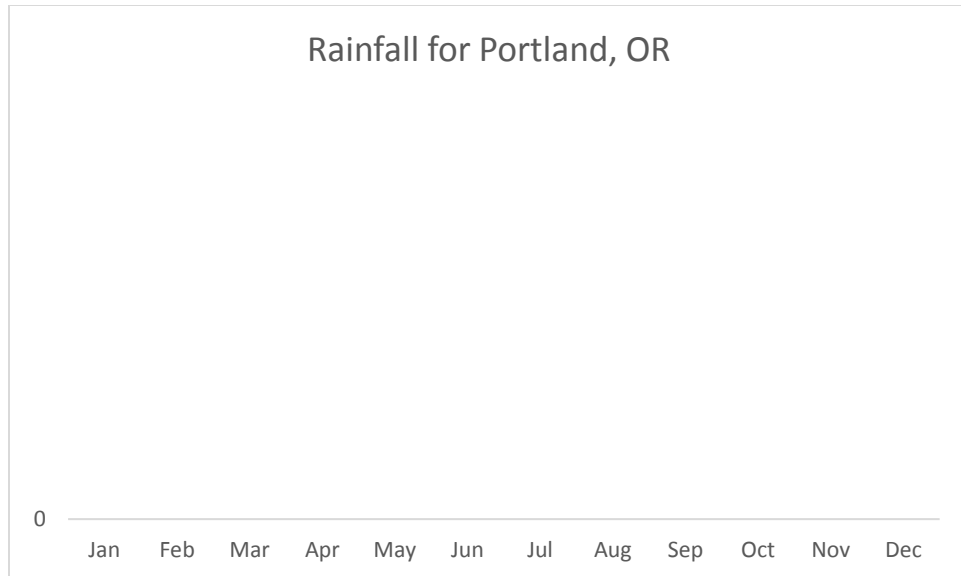
Study question 3: Suppose 25% of Shiley students are wearing jeans today. What percentage of Shiley students will wear jeans tomorrow?

Variance

What is variance or variability?

Now let's think of histograms and the distribution (variability of the population) of some populations. Make a guess at these distributions.





Bar Chart Versus Histogram: Which of the above are bar charts? Which of the above are histograms (ordered data)?

Models

The models that are familiar with from science and theory and are called mechanistic models.

Examples: Current = Voltage / Resistance

Force = Mass * Acceleration

However, some phenomena cannot be modeled mechanistically, so we need an empirical model that is built from data.

Example of an Empirical Model

Suppose we are interested in the number average molecular weight (M_n) of a polymer. Now we know that M_n is related to the viscosity of the material (V), and it also depends on the amount of catalyst (C) and the temperature (T) in the polymerization reactor when the material is manufactured. The relationship between M_n and these variables is

$$M_n = f(V, C, T)$$

say, where the *form* of the function f is unknown.

where the β 's are unknown parameters.

$$M_n = \beta_0 + \beta_1 V + \beta_2 C + \beta_3 T + \epsilon$$

Descriptive Statistics

1. Suppose you set up an experiment to measure the weight of cereal boxes coming off a production line. You make 100 observations at random times during a 24-hour period. What information do you want to know about these samples?
2. In the cereal example, would you want higher variance or lower variance in the observations? Why?

These descriptors are probably familiar to you: mean, standard deviation, variance, and median.

Table 1: Summary of descriptors (N = size of population, n = size of sample)

Descriptor	Population	Population Calculation	Sample	Sample Calculation
Mean	μ	$\frac{\sum_{i=1}^N x_i}{N}$	\bar{x}	$\frac{\sum_{i=1}^n x_i}{n}$
Variance	σ^2	$\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$	s^2	$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$
Standard Deviation	σ	$\sqrt{\sigma^2}$	s	$\sqrt{s^2}$

With some algebra, we can re-write the sample variance calculation to reduce the number of subtraction operations that need to be done. The shortcut method for the sample variance is:

$$s^2 = \frac{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}}{n - 1}$$

Note that the denominator for the sample variance is (n-1) since there are n-1 degrees of freedom. The sum of all the sample differences to the mean is 0, so once (n-1) differences are known, the nth difference is known.

Median

- If number of observations is odd, the median is the middle number of the observations in sorted order
- If number of observations is even, the median is halfway between the middle two observations in sorted order

Minimum = Smallest observation; also called Q_0

Maximum = Largest observation; also called Q_4

Quartiles: Divide data into four equal-sized parts

- Q_1 = value where ~25% of observations are below
- Q_2 = value where ~50% of observations are below (also equal to median)
- Q_3 = value where ~75% of observations are below
- Interquartile Range (IQR) = $Q_3 - Q_1$

Height of Students Activity and Quartiles

Get in sorted order by height (entire class). Then, find Q_0 , Q_1 , Q_2 , Q_3 , and Q_4 of the observations.

There are several methods for determining the quartiles and software packages could vary in terms of which one they implement. Later, we'll see three methods for determining the quartiles.

5-Number Summary (presented at a box of 5 numbers)

Median or Q_2	
Q_1	Q_3
Minimum or Q_0	Maximum or Q_4

Practice with mean, variance, standard deviation, minimum, median, maximum

Suppose the sample of observations of cereal boxes (in ounces) includes:

19.2 19.1 20.2 20.0 19.8 21.1 20.3 21.0 19.9 19.7

What is the sample mean? _____

What is the sample variance using the direct computation (two sig digits after decimal)? _____

What is the sample variance using the shortcut method? _____

What is the sample standard deviation? _____

What is the sample minimum, Q0? _____

What is the sample maximum, Q4? _____

What is the sample median, Q2? _____

(If you finish early) Create a new set of sample values and calculate the statistics, so you can get more practice.

Calculating Quartiles

Now, what about the quartiles? There are many methods to calculate these values. Three such methods are described below.

Method 1:

1. Use the median to divide the ordered data into two halves.
 - a. If the number of observations in original data is odd, do not include median in either half.
 - b. If the number of observations in original data is even, split data exactly in half.
2. $Q1$ = median of lower half of data
3. $Q3$ = median of upper half of data

Method 2:

1. Use the median to divide the ordered data into two halves.
 - a. If the number of observations in original data is odd, include median in both halves.
 - b. If the number of observations in original data is even, split data exactly in half.
2. $Q1$ = median of lower half of data
3. $Q3$ = median of upper half of data

Method 3:

1. If the number of observations is even, use method 1 or 2 (these are equivalent when # of observations is even).
2. If there are $4n+1$ data points:
 - a. $Q1 = .25 * \text{nth data point} + .75 * (n+1)\text{th data point}$
 - b. $Q3 = .75 * (3n+1)\text{th data point} + .25 * (3n+2)\text{th data point}$
3. If there are $4n+3$ data points:
 - a. $Q1 = .75 * (n+1)\text{th data point} + .25 * (n+2)\text{th data point}$
 - b. $Q3 = .25 * (3n+2)\text{th data point} + .75 * (3n+3)\text{th data point}$

Let's figure out the quartiles for the cereal data:

19.2 19.1 20.2 20.0 19.8 21.1 20.3 21.0 19.9 19.7

Sorted data:

19.1 19.2 19.7 19.8 19.9 20.0 20.2 20.3 21.0 21.1

Since the number of observations is even, we can use any of the methods and the results will be the same.

So, we split the data into two halves:

19.1 19.2 19.7 19.8 19.9
20.0 20.2 20.3 21.0 21.1

The median of the first half is 19.7 and the median of the second half is 20.3. So, we have our quartile values:

$Q0 = 19.1$

$Q1 = 19.7$

$Q2 = 19.95$

$Q3 = 20.3$

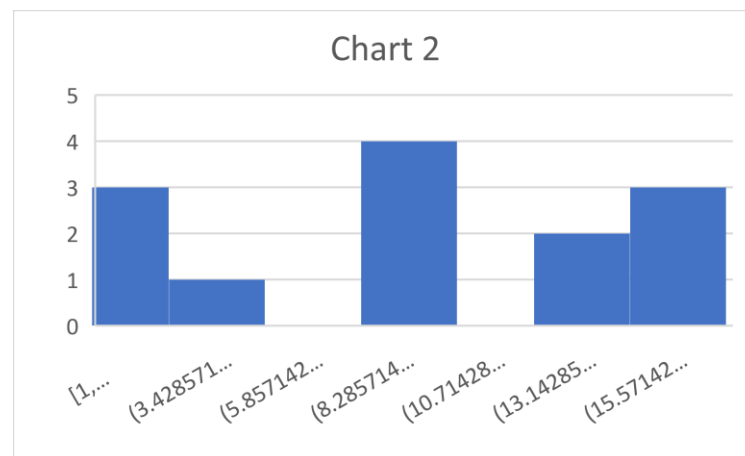
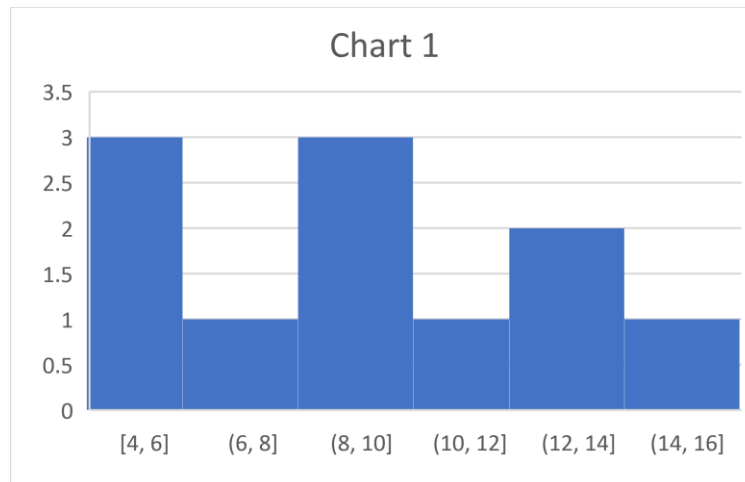
$Q4 = 21.1$

(if time) Produce a dataset of 9 values. Use method 3 to determine the 5-number summary.

Presentation of Data

Summaries of data, such as the 5-number summary, are useful for recording information about a dataset. However, most people like to see a visual representation of the data to better understand the distribution of the data.

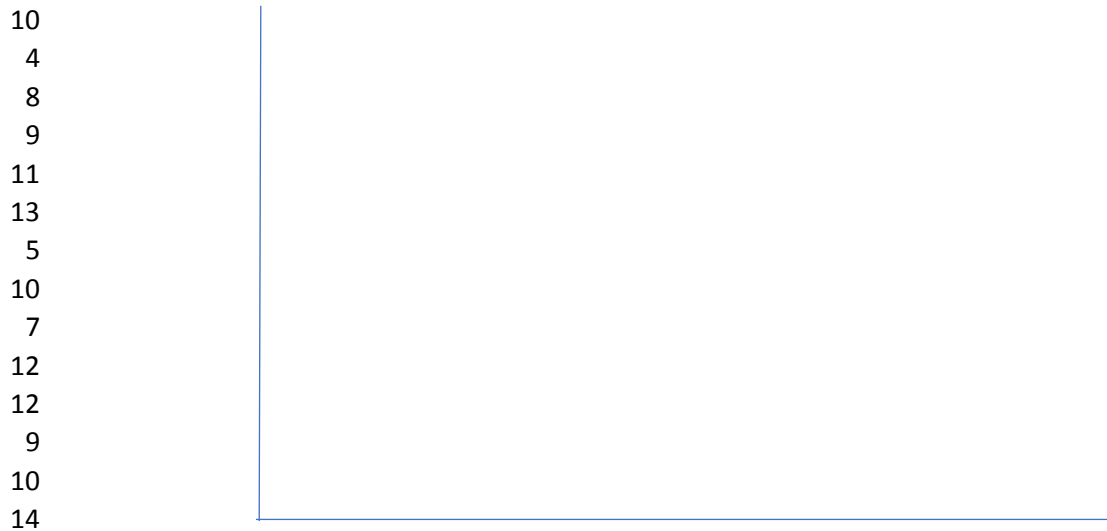
1. Here are two histograms. Which one shows a higher variance of the data?



2. What do you estimate for the mean for the samples shown in Chart 1?
3. What do you estimate for the mean for the samples shown in Chart 2?
4. Look at the y-axis. Does this show counts or relative (percentage) frequency? Histograms can either show frequencies (counts of observed values in that bin) OR show relative frequencies (count / # observations).

7. Where have you seen data presented as a histogram?

8. Below is a dataset. Create a histogram with five bins to display the data. First, figure out what your bin ranges should be, given the values in the dataset. Show the y-axis as counts.



(Read about stem-and-leaf plots in the textbook; no time given during lecture.)

Another way to present data is through a box-plot. A box-plot shows the distribution of observations using boxes (first to second quartile and second to third quartile) and whiskers (smallest data point less than Q1 and largest data point bigger than Q3 that fall into 1.5 IQR) and dots (outliers that are smaller or larger than 1.5 IQR from the boxes). The dots are sometimes open and sometimes full, where open dots are outliers within 3 IQR and full dots are outliers 3 IQR or more away from the boxes. Side-by-side box plots are an effective way to show two or more distributions in a comparative way. For example, let's say a cereal production line was outfitted with new filling sensors. Box-plots of the distribution of sample weights with the old sensor and new sensor could quickly show the engineering team differences between the distributions.

Let's create a box-plot by hand now.

For ease, here are observations of weights from a sample of size 10 (already in sorted order).

4.0 9.2 9.7 9.9 10.1 10.7 11.5 12.1 14.2 19.5

$Q2 = 10.4$ (median)

$Q1 = 9.7$ (using method 1)

$Q3 = 12.1$ (using method 1)

$IQR = 12.1 - 9.7 = 2.4$

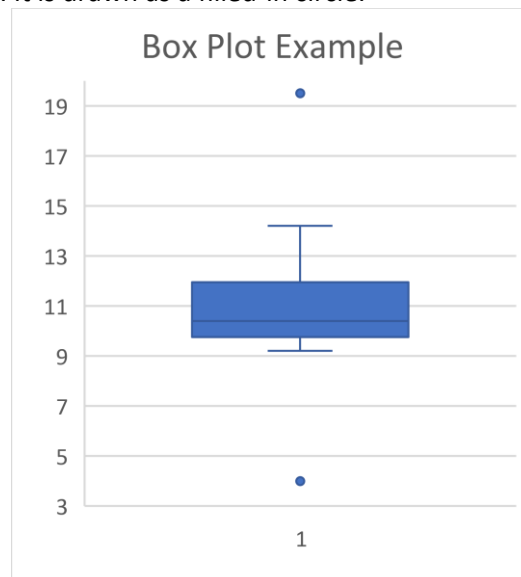
Let's figure out the IQR ranges for 1.5x and 3x.

$1.5 \times 2.4 = 3.6$

$3.0 \times 2.4 = 7.2$

Now, we have the calculation to start drawing.

- First, let's draw the scale from 3 to 20.
- Next, draw the box (Q1 to Q3).
- Next, draw the line for the median (Q2).
- Find the smallest data point in the sample set that is within $1.5 \times \text{IQR}$ of Q1. In this case, $Q1 - 1.5 \times \text{IQR} = 9.7 - 3.6 = 6.1$. The smallest value ≥ 6.1 in the dataset is 9.2. Draw a whisker from the left-edge of the box to 9.2.
- Find the largest data point in the sample set that is within $1.5 \times \text{IQR}$ of Q3. In this case, $Q3 + 1.5 \times \text{IQR} = 12.1 + 3.6 = 15.7$. The largest sample ≤ 15.7 is 14.2. Draw a whisker from the right-edge of the box to 14.2.
- Now, find the outliers. A suspected outlier is within $3 \times \text{IQR}$ of the left or right box edges. $Q1 - 3 \times \text{IQR} = 9.7 - 7.2 = 2.5$. Therefore, the sample value of 4.0 is a suspected outlier and is drawn with an outlined circle (in excel, all outliers are filled in). Now, let's look for an outlier on the high side. $Q3 + 3 \times \text{IQR} = 12.1 + 7.2 = 19.3$. The sample value of 19.5 is bigger than this, so this is considered an outlier. It is drawn as a filled-in circle.



Excel

Now, let's look at how to make histograms and box plots with Microsoft Excel. You can get help within Excel for making charts. You can also find plenty of tutorials with appropriate on-line searches.

Excel Skills:

- Produce a histogram with a specific number or bins (cells).
- Produce a pareto chart.
- Produce a box plot.
- Calculate the mean, variance, and standard deviation of a set of observations. Note that the population and sample calculations for variance and standard deviation are different in excel. Population variance: VARPA, Population Stddev: STDEVPA, Sample variance: VARA, Sample Stddev: STDEVA.
- Calculate the median, Q1, Q3, minimum, and maximum values of a set of observations. Note that excel has 3 functions for quartile. Use Quartile.inc unless otherwise specified in a problem statement.

Probability Activity

Find a partner. Roll the die 10 times. What values did you roll?

- 1.
- 2.
- 3.
- 4.
- 5.
- 6.
- 7.
- 8.
- 9.
- 10.

A. Suppose you roll a 6-sided fair die (with faces having numbers between 1 and 6, inclusive).

1. What is the chance that you roll a 6? _____
2. What is the chance that you roll a 3? _____
3. What is the chance that you roll an odd number? _____

B. Now, suppose you roll two fair 6-sided dice. Think of the outcome as the sum of the two face-up sides after both dice are rolled.

1. What is the outcome set? _____
2. What is the chance that the sum is 12? _____
3. What is the chance that the sum is 1? _____
4. What is the chance that you roll doubles (both dice show the same value)? _____
5. What is the chance that the sum is 7? (Hint: think of all the combinations where the two dice sum to 7.) _____
6. Now roll two dice 20 times.
 - a. How many times did you get doubles? _____
 - b. How many times did you get a sum of 7? _____
 - c. Are these close to the values you found for #4 and #5?

These **chances** are likelihoods, or probabilities. Probability theory comes up in many domains, such as games (like poker or Yahtzee), gambling (like roulette), modeling, simulations, robot navigation, speech recognition, determining significant differences in two or more sets, and this course! We need to understand probability theory, so we can understand distributions and statistical tests.

Probability is the chance of an event happening. The probability of rolling a 3 with a single die is $1/6$. The result of a single trial or observation is called an **outcome**. Rolling a 3 is an outcome. The **sample space** is the set of all possible outcomes. For example, the sample space of a single die roll is {1, 2, 3, 4, 5, 6}, since any of the six sides can end up face up. An **event** consists of a set of outcomes of an experiment. A simple event is an event with one outcome, such as rolling a 3 with one die. A compound event consists of two or more outcomes. A compound event is rolling an odd number with one die. Usually, we just refer to a compound event as an “event”.

Probabilities can be thought about as frequencies and can be determined theoretically (six sides to a fair die), empirically (make several observations and determine the frequencies of the outcomes), or as a predictor (for example, there is a 30% chance that you will need surgery).

$$P(E) = \frac{n(E)}{n(S)} = \text{number of outcomes in } E / \text{Total number of outcomes in sample space}$$

Example: Suppose you draw one card from a standard 52-card playing deck. What is the probability of drawing an Ace?

$$\frac{4}{52} = \frac{1}{13} = 0.077$$

Probability Practice

Q1: Suppose you draw one card from a standard 52-card playing deck. What is the probability of drawing a heart?

Q2: Suppose a family has three children. Find the probability that all the children are boys. Assume the probability of having a girl is 0.5 and the probability of having a boy is 0.5.

Q3: Suppose you draw one card from a standard 52-card playing deck. What is the probability of drawing a 3 **or** a diamond? (Think about the 3 of diamonds and do not over-count).

Explanation

The probability of an event E is a number between 0 and 1, inclusive.

If an event cannot occur, its probability is 0.

If an event is certain, its probability is 1.

The sum of the probabilities of outcomes in the sample space is 1.

The sum of the probability of an event E and the probability of the complement of event E is 1.

When two events A and B are mutually exclusive, the probability that A or B will occur is the sum of $P(A) + P(B)$.

When two events A and B are not mutually exclusive, the probability that A or B will occur is the sum of $P(A) + P(B) - P(A \text{ and } B)$

When two events A and B are independent (the fact that A occurs does not affect the probability of B occurring), the probability of *both* occurring is $P(A) * P(B)$.

When two events A and B are dependent, the probability of both occurring is $P(A \text{ and } B) = P(A) * P(B|A)$.

Because A and B can be thought of symmetrically, we can also write $P(A \text{ and } B)$ as conditional of A on B.

The conditional probability of event B occurs given event A occurs can be written by rearranging terms of the previous statement.

A permutation is an arrangement of n objects in a specific order. The number of permutations of n items is n!.

The number of permutations of n objects taking r objects at a time is the number of permutations of n objects divided by the number of permutations of (n-r) objects that are not selected.

The number of combinations of r objects selected from n objects is the number of permutations (nPr) with r! in the denominators which divides out the duplicate permutations.

Mathematical Statement

$$0 \leq P(E) \leq 1$$

$$P(E) = 0, \text{ if } E \text{ cannot occur}$$

$$P(E) = 1, \text{ if } E \text{ is certain}$$

$$\sum_{E \text{ in Outcome}} P(E) = 1$$

$$P(E) + P(\bar{E}) = 1$$

$$P(A \text{ or } B) = P(A) + P(B),$$

if A and B are mutually exclusive

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B),$$

if A and B are not mutually exclusive

$$P(A \text{ and } B) = P(A) * P(B),$$

when A and B are independent events

$$P(A \text{ and } B) = P(A) * P(B|A),$$

when A and B are dependent events

$$P(A \text{ and } B) = P(B) * P(A|B),$$

when A and B are dependent events

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

$$\begin{aligned} \text{Num permutation of } n \text{ objects} \\ = n * (n - 1) * (n - 2) * \dots 1 \end{aligned}$$

$$nPr = \frac{n!}{(n - r)!}$$

$$nC_r = \frac{n!}{(n - r)! r!}$$

Complements

For each of the following events, what is the complement of the event?

- a) Rolling a single die and getting an even number
- b) Selecting a letter of the alphabet and getting a vowel
- c) Selecting a day of the week and getting a weekday (M – F)

Mutual Exclusion

If two events cannot occur at the same time (no outcomes in common), they are mutually exclusive.

For each of the following outcomes, circle those that are mutually exclusive?

- a) Rolling a single die and getting an odd number and getting an even number
- b) Getting a 5 and getting an odd number
- c) Getting a number less than 4 and getting an odd number
- d) Getting a number less than 4 and getting a number greater than 4
- e) Drawing one card from a standard deck and getting a club and a jack
- f) Drawing one card from a standard deck and getting a heart and a spade

Practice: What is the probability of drawing a single card from a 52-card deck and it is a club **OR** a jack?

What is $P(X = \text{club})$?

What is $P(X = \text{jack})$?

What is $P(X = \text{jack of clubs})$?

What is $P(X = \text{club or jack})$?

Venn Diagrams

\cup = *union* (x is a member of A or B, x is in $A \cup B$)

\cap = *intersection* (x is a member of A and B, x is in $A \cap B$)

If two events A and B are mutually exclusive, then the intersection $P(A) \cap P(B)$ is empty.

Example: Assume there are three possible events, A, B, and C with the following probabilities:

$$P(X \in A) = 0.5$$

$$P(X \in B) = 0.2$$

$$P(X \in C) = 0.4$$

Can events A, B, and C be mutually exclusive? No. The sum of the probabilities is greater than 1, so the intersection of at least two of these events must be non-null.

Now, let's assume the following, continuing with the same example:

$$P(X \in A \cup B) = .65$$

$$P(X \in B \cup C) = .6$$

Are A and B mutually exclusive? No, since the probability of either is .65 and the probability $P(A) + P(B)$ is .70.

Are B and C mutually exclusive? Yes, since the probability of either is .6 and $P(B) + P(C) = .6$.

What does the Venn Diagram look like?

Independence

When the outcome of one event does not impact the outcome of another separate event, the two events are said to be independent. An example is rolling a single die twice. Getting a 5 on the first roll does not impact the value of the second roll. Or rolling two dice at once – the value of one die does not impact the value of another die. Note that it can be confusing with mutual exclusion and independence. Mutual exclusion has to do with events of a SINGLE experiment. Independence has to do with events from two or more experiments.

Example: What is the probability of rolling two 4's with two dice? Since the events are independent, the probability is $(1/6) * (1/6) = (1/36)$.

Practice: What is the probability of getting 3 "heads" on 3 successive coin flips (assume a fair coin)?

Practice: Assume one die is rolled and one coin is flipped. What is the probability of getting an even number on the dice roll and "heads" with the coin flip?

Dependence

When the outcome of one event affects the outcome of another event in such a way that the probability is changed, the events are dependent. Here are some examples of dependent events:

- Drawing a card from the deck, not replacing it, and drawing a second card. Why is this dependent?
- Parking in a no parking zone and getting a parking ticket
- Working as a lifeguard and getting a sunburn.

Conditional Probability

The conditional probability of an event B with relationship to event A is the probability that event B occurs after A has already occurred. It is denoted **$P(B|A)$** . For example, the probability that a second card drawn is a 2 given the first card drawn is an ace is $4/51$ (only 51 cards left for second draw).

$$P(A \text{ and } B) = P(A) * P(B|A) \quad // \text{ when A and B are dependent events}$$

This can be extended to more than 2 events.

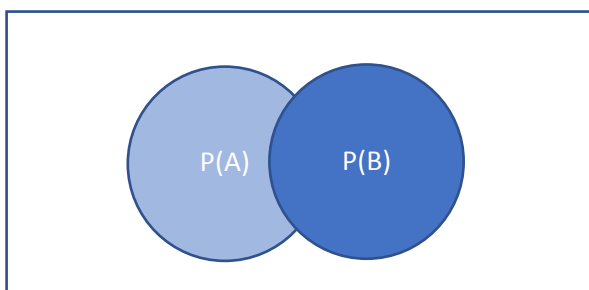
Example: What is the probability of drawing a 2, a 3, and then a 4 (cards are not replaced before each draw)?

$$\begin{array}{ll} P(X=2) = 4/52 & //52 \text{ cards in deck at beginning} \\ P(Y=3) = 4/51 & //51 \text{ cards in deck during second draw} \\ P(Z=4) = 4/50 & //50 \text{ cards in deck during third draw} \end{array}$$

Multiply these together to get $64/132600 = 8/16575$

Practice: What is the probability of getting 5 clubs in a row (cards are not replaced before each draw)?

The Venn diagram of conditional probability:



$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} \quad \begin{array}{l} // \text{intersection} \\ // \text{left circle} \end{array}$$

Becomes the proportion of B that falls in A, where you can think of A as the new sample space.

Permutations and Combinations

Suppose 10 bolts consist of a random sample, taken for testing off a manufacturing line. Suppose each is measured for length. In how many different orders could the bolts be measured?

This is a permutation.

If we label the bolts as B1, B2, B3, ... B10, then one such ordering is B1, B2, B3, ... B10. Another ordering is B10, B1, B2, B3, B4, B5, B6, B7, B8, B9. Let's see how many different orderings there are.

If we think of this as a selection problem, we can consider the following:

There are 10 choices for the first bolt. After that choice is made, there are 9 remaining. There are 9 choices for the second bolt. Now, 8 remain, so there are 8 choices for the third bolt. We can continue with this reasoning and get the following.

$$\# \text{ orders} = 10 * 9 * 8 * 7 * 6 * 5 * 4 * 3 * 2 * 1 = 10!$$

Now, let's modify this example a bit. Suppose we instead choose just 5 of the 10 to measure. How many different orderings of 5 bolts could we have?

Now, there are $10 * 9 * 8 * 7 * 6$ possible orderings.

One way to summarize permutations where r objects are selected from n objects and order matters is:

$${}_nP_r = \frac{n!}{(n-r)!}$$

Let's look at the ordering of 5 of 10 bolts again with this rule in mind.

$${}_{10}P_5 = 10! / 5!, \text{ which is the same as } 10 * 9 * 8 * 7 * 6$$

A combination is simply a set where order DOES NOT matter. For example, a combination is the selection of a presidential advisory committee of 10 faculty members from a total of 200 faculty members. The order in which the committee is selected does not matter – the make-up (set) of the committee is what matters.

Now, we have a function for combinations (choose r items from n items):

$${}_nC_r = \frac{n!}{(n-r)! r!}$$

This is similar to ${}_nP_r$, except there is a factor of $r!$ in the denominator. So, you can think of dividing the number of permutations by $r!$ to eliminate the same combinations among the permutations.

Example: Suppose you want to select a random sample of 10 solar eclipse glasses from 50 glasses. How many different combinations (possible selections of random samples of size 10) are there?

$$50! / (40! * 10!) = 50 * 49 * 48 * 47 * 46 * 45 * 44 * 43 * 42 * 41 / 10 * 9 * 8 * 7 * 6 * 5 * 4 * 3 * 2 * 1$$

Practice with combos

Practice Q1: How many 5-digit zip codes are possible if digits can be repeated?

Practice Q2: How many 5-digit zip codes are possible if digits cannot be repeated?

Practice Q3: Suppose there are 5 quality assurance engineers and 2 product designers on a team. A new project requires a team of two, consisting of one quality assurance engineer and 1 product designer. How many pairings are possible?

Practice Q4: Now, suppose the pairing can be any 2 drawn from the 7 employees. How many pairings are possible?

Random Variables

A random variable is a variable whose values are determined by chance. Often, it is written as a capital letter, such as X . For example, in a single roll of one die, the random variable X is the value of the die.

$$P(X = 1) = 1/6, P(X = 2) = 1/6, P(X = 3) = 1/6, P(X = 4) = 1/6, P(X = 5) = 1/6, P(X = 6) = 1/6$$

In the case of the coin flip, the random variable X is the result of the flip (head or tails). In a fair coin, the chance of getting heads is .5.

$$P(X = \text{heads}) = 0.5, P(X = \text{tails}) = 0.5$$

Example: Let X denote the life of a semiconductor laser (in hours) with the following probabilities:

$$P(X \leq 5000) = 0.05$$

$$P(X > 7000) = 0.45$$

- a) What is the probability that the life span is less than or equal to 7000 hours? _____
- b) What is the probability that the life is greater than 5000 hours? _____
- c) What is $P(5000 < X \leq 7000)$? _____

There are two types of random variables:

- Discrete (finite set of outcomes or countably infinite set of outcomes)
 - Examples: dice roll, megaball in the lottery, coin flip, number of bits transmitted in error, proportion of defective parts among a sample of 1000 parts
- Continuous (interval of real numbers for its range)
 - Examples: electrical current, length, density, weight, temperature

What is another example of a discrete random variable?

What about a continuous random variable?

Probability Activity

Directions: Get into a team of 4 people. Your team will work through problems related to probability. Work together as a team – do not divide and conquer the problems. Each team member is assigned to one of the following roles:

- Notetaker: prepares the master copy to be submitted
- Manager: ensures that group stays on task, manages time, and ensures that everyone has a chance to speak
- Spokesperson: when called upon, speaks on behalf of the team
- Reflector: analyzes team dynamics and makes suggestions for improving team process – is everyone contributing? Does everyone understand how to get to the solution? Should we check the solution?

Names: Notetaker: _____ Manager: _____
Spokesperson: _____ Reflector: _____

1. Suppose a technology company logs data about how many unique visitors (based on IP address) visit any page on the company's website per day. The data suggest the following probabilities, where X is the number of unique customers.

$$P(0 \leq X \leq 10000) = 0.4$$

$$P(X \leq 20000) = 0.7$$

$$P(X \leq 30000) = 0.8$$

$$P(X \leq 40000) = 0.9$$

$$P(X \leq 1000000) = 1.0$$

- a) What is $P(X > 1000000)$? _____
 - b) $P(10000 < X \leq 20000)$? _____
 - c) $P(X > 20000)$? _____
 - d) $P(X > 40000)$? _____
2. A. Could the probability of an event be -3 ? Yes or No?

B. Could the probability of an event be 1.27 ? Yes or No?
 3. Suppose a 5-card hand is dealt to you from a standard 52-card deck. What is the probability that your hand has no face cards and no aces? (Face cards are jacks, queens, and kings)

META-QUESTION: Is everyone contributing to the team? If not, how could the team improve its process?

4. Suppose the Dean's Office tracks where UP graduates are six months after graduation. Of the 120 graduates in 2015, here is the data (this may not be accurate): 65 are employed, 10 are in

graduate school full-time, 5 are employed and doing graduate school part-time, 10 are in the military, 10 are still looking for jobs, and 20 are doing volunteer service.

- a. What is the probability that a student is employed?
 - b. What is the probability that a student is in graduate school?
5. Suppose you roll two dice and the outcome is the product of the two values rolled. For example, if the roll is {2, 5}, then the outcome is 10.
- a. What is the sample space for this experiment? (Which values can you get?)
 - b. What is the probability that the product is less than 10?

Checkpoint 1: Stop for class discussion. If your team has reached this checkpoint and the discussion has not happened, you may move on to the rest of the questions. Be ready to share your team answers with the class for the discussion.

6. Suppose the probability that a student owns a car is 0.65. The probability that a student owns a bicycle is 0.82. If the probability that a student owns a car AND a bicycle is 0.55, draw the Venn Diagram for this situation.
7. Continuing from the prior question, what is the probability that a student owns a car, but not a bicycle?
8. Suppose there are 100 women engineering students. Of these, 35 are members of the crew team, 80 are members of the Society of Women Engineers, and 25 are members of both crew and SWE. What is the probability that a randomly selected female engineering student is neither on crew nor a member of SWE? (You may want to draw a Venn Diagram to help you.)

9. Suppose there are 8 computer chips in a sample, 5 of which have defects and 3 are good. If three are selected at random from the set of 8, what is the probability that all 3 (selected without replacement) will have defects?
10. Suppose a 3-letter string from English (26 letters) is constructed at random with replacement (all 26 letters could be chosen for each position). What is the probability that the word is "cat"?
11. A coin is tossed five times. Find the probability of getting at least one tail. (This problem is probability easier when thinking about the complement... probability of getting no tails and subtracting from 1.)
12. Discuss among your team decisions you have made where you used probability theory. Jot some of them down on the notetaker's master copy.
13. Does your team have further questions about probability theory?

META-QUESTION: Does everyone understand the process used to solve these problems? If not, try to explain how to solve the problem in a different way. In fact, there are several ways to solve some of these problems.

Checkpoint 2: Stop for class discussion. If your team has reached this checkpoint, you may move on to the optional questions. Be ready to share your team answers with the class.

14. (if time) Come up with your own data and question related to probability theory. Try to solve it.
15. (if time) Come up with a theoretical probability (with dice, coins, cards, etc.) and design an experiment to see if the observed frequencies match the theoretical probability.

Discrete Random Variables

Recall that a discrete random variable is one with a finite (or countably infinite) set of real numbers for its range. For example, the roll of a die is a discrete random variable. The number of phone lines in use at a call center is another example.

Let's look at the number of phones simultaneously in use.

$$P(X = 0) = .30$$

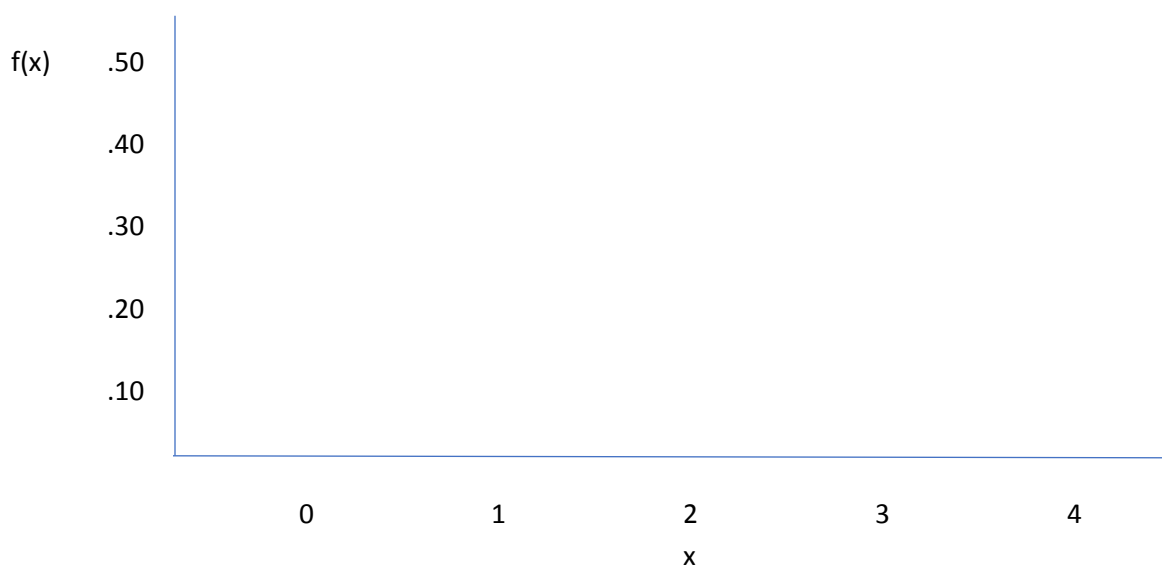
$$P(X = 3) = .10$$

$$P(X = 1) = .40$$

$$P(X = 4) = .05$$

$$P(X = 2) = .15$$

Here, these values show the distribution. What does the **probability distribution** look like for this function?



The function above is a **probability mass function, denoted $P(x)$** .

$$f(x_i) = P(X = x_i)$$

The sum of the probabilities for all outcomes is 1. Verify that the sum of the probabilities does indeed equal 1.

a. What is $P(X = 0)$? _____

d. What is $P(X > 4)$? _____

b. What is $P(X < 3)$? _____

e. What is $P(1 \leq X < 4)$? _____

c. What is $P(X \geq 1)$? _____

Another representation is called the **cumulative distribution function (denoted $F(x)$)** for a random variable. It reflects its name in that it shows the combined (cumulative) probabilities for values of $X \leq x$.

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$$

Using the example above:

$$F(0) = 0.30$$

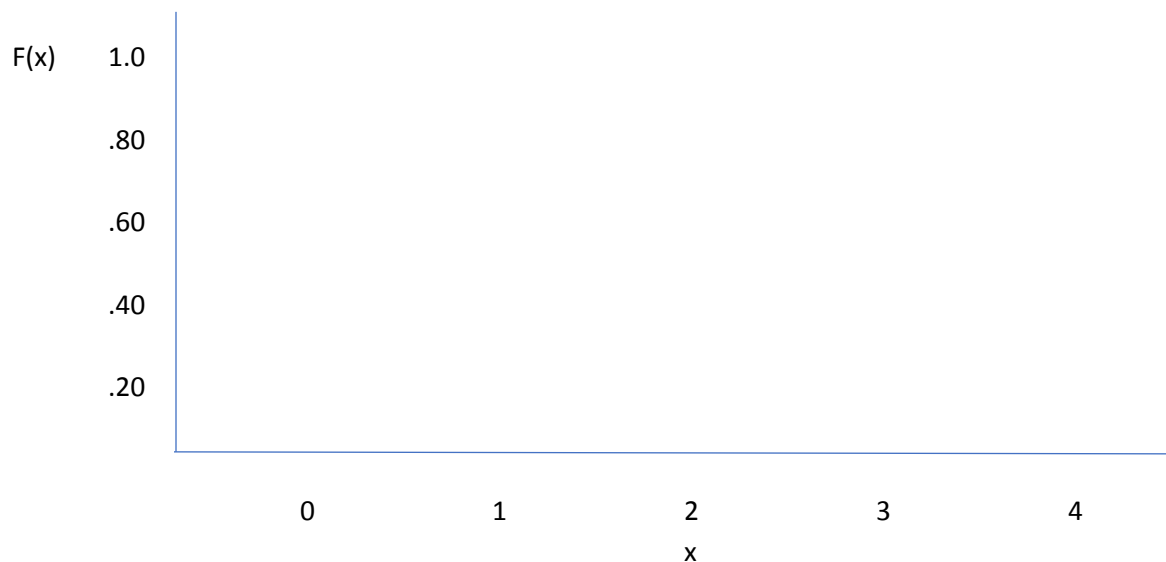
$$F(1) = 0.70$$

$$F(2) = 0.85$$

$$F(3) = 0.95$$

$$F(4) = 1$$

Graph the CDF $F(x)$:



Note that the CDF is defined at non-integer values. $F(0.5) = P(X \leq 0.5) = P(X \leq 0) = 0.30$. There are jumps at integer values, so the function is piecewise continuous. The jump distance at x shows the original probability value of x . For example, the jump at $x = 1$ is a distance of 0.4.

Discrete Random Variables Statistics

The **mean** or **expected value** of X is denoted as μ or $E(X)$.

$$\mu = E(X) = \sum_{i=1}^n x_i f(x_i)$$

In the example above, what is the mean number of simultaneous in-use phone lines?

Variance of a discrete random variable

$$\sigma^2 = V(X) = E(X - \mu)^2 = \sum_{i=1}^n (x_i - \mu)^2 f(x_i) = \sum_{i=1}^n x_i^2 f(x_i) - \mu^2$$

In the example above, what is the variance? _____

The standard deviation of a discrete random variable is the square root of the variance.

In the example above, what is the standard deviation? _____

Coin Experiment

Take a coin. Flip it 10 times and record the number of times that heads “wins” the flip.

- What is the number? _____
- What do you think is the most likely # of heads that are flipped out of 10 trials? _____
- What do you think is the least likely # of heads that are flipped out of 10 trials? _____
- Plot your *estimate* of the probability function for the discrete random variable $X = \#$ of heads flipped of 10 coin flips.

0 1 2 3 4 5 6 7 8 9 10

- Now, we'll plot the values from the entire class.

0 1 2 3 4 5 6 7 8 9 10

The Binomial Distribution

This coin experiment creates data that follows a binomial distribution pattern. A *binomial distribution* arises when an experiment can be thought of as a series of repeated random trials with two outcomes: success or failure. The random variable is the *count* of the number of trials that meet a specific criterion (such as number of times the coin lands heads). It's BINOMIAL because we think of each trial as being successful (meeting the criterion) or being a failure (not meeting the criterion).

A trial with two outcomes (success or failure) is called a *Bernoulli trial* and we make the following assumptions:

- Successive trials are independent. For example, a “heads” flip does not impact the success/failure probability on the next coin flip.
- The probability of success remains constant over the set of trials. For example, the coin's probability of landing heads is 0.5 across all trials.

Let's think about examples where the random variable follows a binomial distribution.

Examples:

- Results show that 40% of patients improve blood pressure when using a certain drug. In the next 50 patients given the drug, let X = the number of patients whose blood pressure improves.
- A 4-answer per question multiple choice exam. X = number of questions answered correctly on an exam of 20 questions, where the student guesses on every question. What is the probability of success in this case?

What is another example? _____

Let's analyze the coin-flip scenario to derive the binomial distribution. Assume you do the coin-flip experiment using **4** trials instead of 10 trials.

How many different outcomes are there? ____ For each, how many heads are flipped? Complete the chart below.

Outcome	NumHeads	Outcome	NumHeads
HHHH	4	HHHT	3
HHTH	3	HTHH	3

- a. How many outcomes have 0 heads? _____
- b. How many outcomes have 1 head? _____
- c. How many outcomes have 2 heads? _____
- d. How many outcomes have 3 heads? _____
- e. How many outcomes have 4 heads? _____

Calculate the frequencies by dividing each of these values by the total number of outcomes. Graph the PMF below.



Because the coin flip is 50/50 chance of landing heads, the probability of success equals the probability of failure. The distribution is symmetric. Let's see what happens when the probability of success does not equal the probability of failure.

Assume the coin is no longer fair. It lands heads 40% of the time. Is the probability of getting {HHHH} smaller or bigger than with the fair coin? _____

$$P(X = \text{HHHH}) = (0.4)^4$$

$$P(X = \text{HHHT}) = (0.4)^3 * (0.6)$$

$$P(X = \text{HHTT}) = (0.4)^2 * (0.6)^2$$

Etc.

So, now the number of outcomes that results in x coin flips is:

$$P(X = x) = (\text{number of outcomes that result in } x \text{ heads}) * (0.4)^x * (0.6)^{4-x}$$

We need to figure out how many of the outcomes results in x heads. Well, we have seen that before when studying combinations. The number of ways to select x items from n items is “n choose x”:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

In our non-fair coin example, the probability that the number of heads is x is:

$$P(X = x) = \binom{4}{x} (0.4)^x (0.6)^{4-x}$$

Let’s look at a specific case of x. What is the probability that the number of heads flipped is 1?

$$P(X = 1) = 4 * (0.4) * (0.6)^3 = 0.346$$

Complete the rest:

$$P(X = 0) \underline{\hspace{2cm}}$$

Hint: 4 choose 0 is 1

$$P(X = 1) \underline{\hspace{2cm}}$$

Hint: 4 choose 1 is 4

$$P(X = 2) \underline{\hspace{2cm}}$$

Hint: 4 choose 2 is 6

$$P(X = 3) \underline{\hspace{2cm}}$$

Hint: 4 choose 3 is 4

$$P(X = 4) \underline{\hspace{2cm}}$$

Hint: 4 choose 4 is 1

Now, graph this PMF:



How does it compare to when the coin was fair?

Now, we can write the general binomial distribution function.

Assumptions:

- n repeated trials; trials are independent
- each trial has one of two outcomes (success and failure)
- probability of success (denoted p) remains constant across trials

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, 2, 3 \dots n$$

$$\mu = E(X) = np$$
$$\sigma^2 = V(X) = np(1-p)$$

Example problem:

The probability of landing a plane with a flight simulator is .80. Nine randomly and independently chosen pilots-in-training are asked to fly the plane with the simulator.

1. What is the probability that ALL pilots-in-training land the plane using the simulator?

First, we need to determine what X is. X = number of students who successfully land the plane with the simulator.

Second, we need to determine p, the probability of success: In this case, p is .80.

Third, we need to determine the n, the number of trials: In this case, n is 9 (9 pilots-in-training).

Now, we have the information we need to solve this problem using the binomial distribution.

$$P(X = 9) = \binom{9}{9} (.8)^9 (.2)^0 = 1 * (.8)^9 * 1 = .1342$$

2. What is the probability that none of the pilots-in-training land the plane? (Note: success is actually failure in this case)

$$P(X = 0) = \binom{9}{0} (.8)^0 (.2)^9 = 1 * 1 * (.2)^9 = .000000512$$

3. What is the probability that <=7 students land the plane?

We can either calculate $P(X = 0) + P(X = 1) + \dots + P(X = 7)$ OR by using set complement, we can determine $1 - P(X=8) - P(X=9)$. When you get these problems, think about what would be easiest or you can use excel in cumulative mode to do these for you.

We already calculated $P(X=9)$ above, so we just need to do $P(X=8)$:

$$P(X = 8) = \binom{9}{8} (.8)^8 (.2)^1 = 9 * (.8)^8 * (.2) = .302$$

$$\text{So } P(X \geq 7) = 1 - .1342 - .302 = .5638$$

Excel

In excel, you can use the BINOM.DIST function. It takes arguments: x, n, p, and TRUE if you want $P(X \geq x)$ or FALSE if you want $P(X = x)$.

```
BINOM.DIST(7, 9, 0.8, TRUE)      // for  $P(X \geq 7)$ 
BINOM.DIST(7, 9, 0.8, FALSE)    // for  $P(X = 7)$ 
```

Fun fact: Plinko from the Price is Right game show is a mechanical version that produces the binomial distribution. You can find simulators online.

Binomial Distribution Practice

Directions: Get into a team of 4 people. Your team will work through problems related to the binomial distribution. Work together as a team – do not divide and conquer the problems. Each team member is assigned to one of the following roles:

- Notetaker: prepares the master copy to be submitted
- Manager: ensures that group stays on task, manages time, and ensures everyone gets to speak
- Spokesperson: when called upon, speaks on behalf of the team
- Reflector: analyzes team dynamics and makes suggestions for improving team process – is everyone contributing? Does everyone understand how to get to the solution?

Names: Notetaker: _____

Manager: _____

Spokesperson: _____

Reflector: _____

Here's the binomial distribution function:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, 2, 3 \dots n$$

$$\mu = E(X) = np$$
$$\sigma^2 = V(X) = np(1-p)$$

Now, let's practice using this distribution. Work with your group to solve these problems

1a) A hip joint replacement part is being stress-tested in a laboratory. The probability of successfully completing the test is .80. Seven randomly and independently chosen parts are tested. What is the probability that exactly two of the seven parts successfully complete the test?

What is the discrete random variable? _____

What is p? _____

What is n? _____

What is x? _____

P(X = 2) = _____

b) Now, what is the probability that at least two of the seven parts successfully complete the test?

What is P(X = 0)? _____

What is P(X = 1)? _____

Then, $P(X \geq 2) = 1 - P(X=0) - P(X=1)$:

c) What is the *expected number* of parts that successfully complete the test?

2) Batches consist of 50 coil springs from a manufacturing line are checked against customer requirements. The mean number of non-conforming coil springs in a batch is 5. Assume that the number of non-conforming springs, denoted as X , is a binomial random variable.

- a) What is n ? _____
- b) What is p ? _____
- c) What is $P(X \leq 2)$? _____

- d) What is $P(X \geq 49)$? _____

META-QUESTION: Does everyone understand the notation? Does everyone understand how to use the binomial distribution? If not, how could the team improve its process, so that all people learn?

Checkpoint 1: Stop for class discussion. If your team has reached this checkpoint, you may move on to the next question. Be ready to share your team answers with the class.

3) Suppose a manufacturing process has 100 customer orders to fill. Each order requires one component that is purchased from a supplier. However, 2% of the components are identified as defective and the components can be assumed to be independent.

- a) If the manufacturer stocks 100 components, what is the probability that the 100 orders can be filled without reordering components?

- b) If the manufacturer stocks 102 components, what is the probability that the 100 orders can be filled without reordering?

c) If the manufacturer stocks 105 components, what is the probability that the 100 orders can be filled without reordering?

Checkpoint 2: Stop for class discussion. If your team has reached this checkpoint, try to come up with your own questions/statements in which the binomial distribution function can be applied. Be ready to share your team answers with the class.

Poisson Process and Distribution

Let's think about the binomial distribution again.

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, 2, 3 \dots n$$

When n goes to infinity and p goes to 0, we can think of the value $np = \lambda$, remaining constant. The limit as n goes to infinity of $P(X = x)$ = binomial distribution is:

$$\frac{e^{-\lambda} \lambda^x}{x!}$$

The mathematics showcasing this limit is in example 3-29 in your textbook or see the video on-line. This limit of n going to infinity means many trials and the probability of success (p) gets smaller and smaller. It turns out that this case represents the Poisson process:

Consider T as an interval of real numbers partitioned into subintervals of small length D and assume D tends to 0:

- The probability of more than one event in a subinterval tends to zero.
- The probability of one event in a subinterval tends to $\lambda D/T$.
- The event in each subinterval is independent of other subintervals.

The subintervals can be thought of as independent Bernoulli trials with success $p = \lambda D/T$ and the number of trials equal to $n = T/D$. So, $pn = \lambda$.

Poisson Distribution

The probability mass function of X (**Poisson random variable**) where random variable X equals the number of events in a Poisson process with parameter $\lambda > 0$ is:

$$f(x) = P(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

$$E(X) = \lambda$$

$$V(X) = \lambda$$

The number e is the natural number, defined as approximately equal to 2.7183. You probably have e as a constant on your calculator. Recall some mathematical properties of e :

$$\text{definition of } e = \sum_{n=0}^{\infty} \frac{1}{n!} \sim 2.7183$$

$$\frac{d}{dx} e^x = e^x$$

$$\int e^x dx = e^x + C$$

Example 1: Customer Arrival

Suppose the number of customers (X) that visit a post office in a 30-minute period follows the Poisson process. Suppose $P(X = 0) = 0.02$.

a) What is the mean of X?

$E(X) = \mu = \lambda$ for the Poisson process. So, we need to determine the value of λ .

$$\begin{aligned} P(X = 0) = 0.02 &= \frac{e^{-\lambda} \lambda^0}{0!} = e^{-\lambda} \\ \ln(e^{-\lambda}) &= \ln(0.02) \\ -\lambda &= -3.912 \\ \lambda &= 3.912 \end{aligned}$$

So, we expect 3.912 customers to visit the post office in a 30-minute period.

b) What is the standard deviation of X?

Since this is a Poisson process, the variance is equal to the mean, which is 3.912. So, the standard deviation is the square root of the variance, which is 1.978.

c) What is the probability that 2 customers arrive in a 30-minute period?

$$P(X = 2, \lambda = 3.912) = \frac{e^{-3.912} 3.912^2}{2!} = .02 * 15.3037 \div 2 = .153$$

The probability of 2 customers arriving in a 30-minute period is .153.

Example 2: Typos

a) Suppose there are 200 typographical errors in a 500-page dissertation. Find the probability that a given page contains exactly three errors. (You may assume the number of errors follows a Poisson process.)

First, we need to find the mean number of errors/page:

$$\lambda = \frac{200}{500} = 0.4 \text{ errors per page}$$

We want to find the probability that $X = 3$ using the Poisson distribution:

$$P(X = 3, \lambda = 0.4) = \frac{e^{-0.4} 0.4^3}{3!} = 0.0072$$

So, there's a less than 1% chance that there are exactly 3 errors on the page. Does this make sense given there are 200 errors in 500 pages?

b) Now, what is the probability that there are no errors on a page?

$$P(X = 0, \lambda = 0.4) = \frac{e^{-0.4} 0.4^0}{0!} = 0.67$$

Does this probability make sense?

REMINDER:

When using the Poisson distribution, it is important to remember to use CONSISTENT units when determining λ .

For example, suppose the mean number flaws is 5.2 flaws per centimeter.

What is the mean number of flaws per millimeter? _____

What is the mean number of flaws per meter? _____

Always check for consistent use of units!!!

Example 3: Road Repair

Suppose the number of potholes in a section of a highway that need repair follows a Poisson distribution with a mean of 2 potholes per mile.

a) What is the probability that there are no potholes that require repair in 5 miles of highway?

We need to first figure out λ in terms of our question. Since our question involves 5 miles of highway, we need to figure out the mean number of potholes in 5 miles. Since there are 2 potholes/mile, there are 10 potholes per 5 miles of highway.

$$P(X = 0, \lambda = 10) = \frac{e^{-10} 10^0}{0!} = 0.0000454$$

Does this probability make sense?

b) What is the probability that *at least* one pothole requires repair in ½-mile of highway?

Again, we need to figure out λ in terms of the consistent units. Since we have 2 potholes per mile, we have 1 pothole per half-mile. We need to figure out $P(X \geq 1) = P(X=1) + P(X=2) + P(X=3) + \dots$. So, we use the complement and can determine this with: $1 - P(X=0)$:

$$1 - P(X = 0, \lambda = 1) = 1 - \frac{e^{-1} 1^0}{0!} = .632$$

Does this probability make sense?

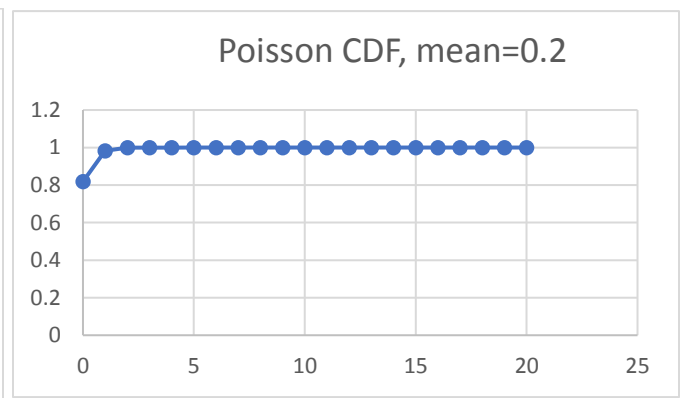
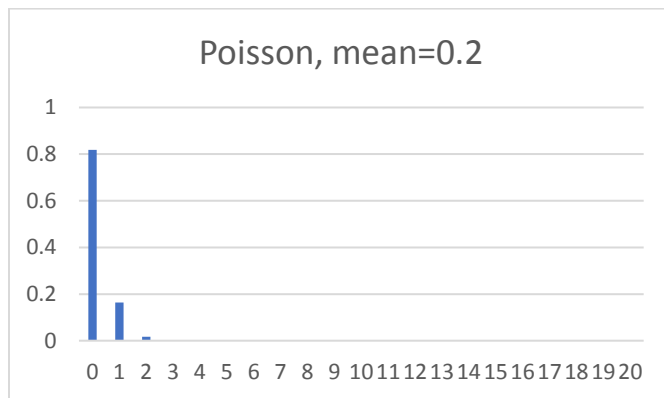
c) If the number of potholes is related to vehicle load on the highway and some sections of the highway have a heavy load of vehicles and some have a light load of vehicles, how do you feel about the assumption of a Poisson distribution for the number of potholes that need repair for all sections?

Excel

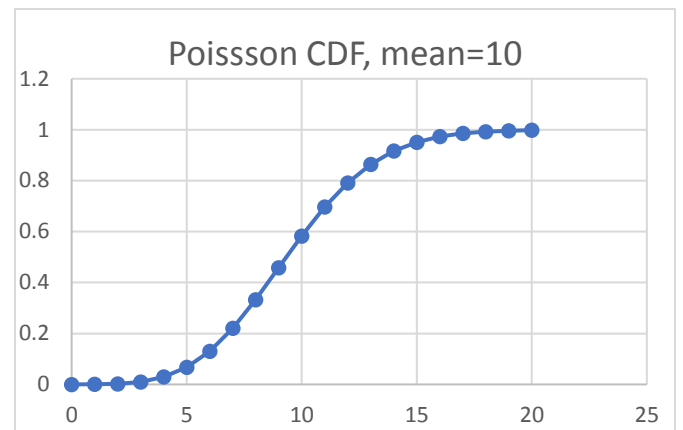
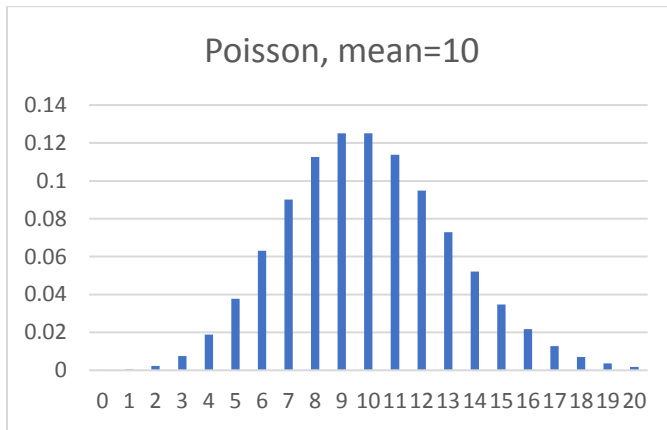
Use POISSON.DIST(X, mean, FALSE) when $P(x=X)$.

Use POISSON.DIST(X, mean, TRUE) when $P(x \leq X)$ [cumulative].

Mean=0.2



Mean=10



FYI: Other common discrete random variable distributions

Multinomial Distribution

When there are more than two outcomes and a fixed number of trials, the probabilities for each outcome can be modeled with a multinomial distribution:

$$P(X) = \frac{n!}{X_1! * X_2! * \dots * X_k!} * p_1^{x_1} p_2^{x_2} p_3^{x_3} \dots p_k^{x_k}$$

Where $X_1 + X_2 + \dots + X_k = n$ and $p_1 + p_2 + p_3 + \dots + p_k = 1$

Example type of problem: Suppose 50% of people choose a movie, 30% choose dinner, and 20% choose shopping as their favorite leisure activity. If a sample of 5 people is chosen, what is the probability that three are going to a movie, one is going to dinner, and one is going shopping?

Here, $n = 5$. $X_1 = 3$. $X_2 = 1$. $X_3 = 1$. $p_1 = .5$. $p_2 = .3$. $p_3 = .2$.

Hypergeometric Distribution

When there are two outcomes and sampling is done WITHOUT replacement, you can use the hypergeometric distribution. The binomial distribution is less accurate when the population is smaller. Also, the binomial distribution assumes replacement since each trial is independent.

This is used when there are two types of objects, such that there are a of one kind and b of the other and a+b is the total population. The probability of selecting a sample of n items with X items of type a and n-X items of type b is:

$$P(X) = \frac{\binom{a}{X} \binom{b}{n-X}}{\binom{a+b}{n}}$$

Here, the $\binom{a}{X}$ means a choose X (the number of combinations of X items from a set of a items).

Example type of problem: Suppose a study finds that 4 of 9 houses are underinsured. If five houses are selected from 9 houses, find the probability that exactly 2 are underinsured.

$a=4$, $b=5$, $n=5$, $X=2$, $n-X=3$

$P(X=2) = 60/126 = 10/21 = .4761$

Poisson Practice

Directions: Get into a team of 4 people. Your team will work through problems related to the binomial, poisson, and exponential distributions. Work together as a team – do not divide and conquer the problems. Each team member is assigned to one of the following roles:

- Notetaker: prepares the master copy to be submitted
- Manager: ensures that group stays on task, manages time, and ensures everyone gets to speak
- Spokesperson: when called upon, speaks on behalf of the team
- Reflector: analyzes team dynamics and makes suggestions for improving team process – is everyone contributing? Does everyone understand how to get to the solution?

Names: Notetaker: _____

Manager: _____

Spokesperson: _____

Reflector: _____

Here's the Poisson distribution function, mean, and variance:

$$f(x) = P(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$
$$E(X) = \lambda$$
$$V(X) = \lambda$$

1) Messages arrive to a computer server according to a Poisson distribution with a mean rate of 10 messages per hour.

a) What is the probability that three messages will arrive in an hour?

b) What is the probability that six messages arrive in 30 minutes?

Hint: remember to update the mean for this unit of time.

c) Do these probabilities make sense?

2) The number of infections at a hospital is modeled with a Poisson distribution with mean 3.5 per month.

a) What is the probability of exactly 3 infections in a month?

b) What is the probability of no infections in a month?

c) What is the expected number of infections per year?

META-QUESTION: Does everyone understand the notation? Does everyone understand how to use the poisson distribution? If not, how could the team improve its process, so that all people learn?

Checkpoint 1: Stop for class discussion. If your team has reached this checkpoint, you may move on to the next question. Be ready to share your team answers with the class.

Continuous Random Variables

Recall that there are types of random variables: discrete random variables and continuous random variables. Examples of continuous random variables:

- Heights of eight-year-old boys
- Concentration of chlorine in water

We no longer think about the probability that $X =$ a definite value for a continuous random variable. It does not make sense for us to say, what is the probability that an eight-year-old's height is 1.589 meters? It makes more sense for us to say, what is the probability that an eight-year-old's height is between 1.5 and 1.6 meters?

With continuous random variables, the probability distribution is a description of the set of probabilities associated with possible values of X . The probability density function (PDF) $f(x)$ can be used to describe the probability distribution of a continuous random variable.

Probability Density Function

PDF $f(x)$ of a continuous random variable X is:

$$P(a < X < b) = \int_a^b f(x)dx$$

Properties of pdf:

- $f(x) \geq 0$ for all x // range is non-negative
- $\int_{-\infty}^{\infty} f(x)dx = 1$ // area under the curve is 1

Because every point on the x -axis has zero-width, we have the following statement:

$$P(X = x) = 0$$

The probability of X at *exactly point* x is 0 for any point x . Instead, we think of probabilities of X being in a range, such as $(-\infty, a)$, (a, b) , or (a, ∞) .

We have the following property:

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$$

PDF Exercise:

Graph each function below. Are the following valid probability density functions? If not, why not?

a) $f(x) = .05$ for $0 \leq x \leq 20$

b) $f(x) = x$ for $0 \leq x \leq 2$

c) $f(x) = 3x$ for $0 \leq x \leq 1$

d) $f(x) = -2x$ for $0 \leq x \leq 2$

e) $f(x) = x$ for $0 \leq x \leq 0.2$
 $f(x) = .1$ for $0.2 \leq x \leq 1.0$
 $f(x) = -x$ for $1.0 \leq x \leq 1.2$

Steps to determine the probability of a random variable:

1. Determine the random variable.
2. Determine the distribution of the random variable.
3. Write probability statement in terms of the random variable.
4. Compute the probability using the statement and distribution.

Example 1:

Suppose the probability density function of X is $f(x) = 0.1$ for $0 \leq x \leq 10$.

a) First, is this a valid pdf (Is the area under the curve equal to 1? Are all values of $f(x) \geq 0$ for all x)?

b) Graph this pdf. You may assume that $f(x) = 0$ if $x < 0$ or $x > 10$.

c) Suppose $f(x)$ is the current measured in copper wire in milliamps. What is the probability that a current measurement is less than 6 milliamps?

1. Determine the random variable:

X = the current measured in milliamps

2. Determine the distribution:

This is given as $f(x)$ and graphed above.

3. Write probability statement:

$$P(X < 6) = \int_0^6 0.1 dx$$

4. Compute the probability:

$$\int_0^6 0.1 dx = 0.1x \Big|_0^6 = .6 - 0 = 0.6$$

$$P(X < 6) = 0.6$$

Cumulative Distribution Function

The CDF of a continuous random variable X with probability density function $f(x)$ is:

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u)du \quad \text{for } -\infty < x < \infty$$

Note that the CDF is *cumulative*, so we're determining the probability of X being less than or equal to the value x . Also, $F(x)$ is always non-negative, since $F(x)$ is cumulative probabilities. Also, $F(x)$ is non-decreasing as x increases. As x goes to infinity, $F(x)$ approaches 1. We use $f(u)du$ as a change in variable name since x is used as the value of the upper bound of the integral.

Example 2:

Suppose the pdf for distance to flaws in a copper wire is:

$$f(x) = \frac{1}{10}e^{-x/10}$$

$$x \geq 0$$

What is the cdf?

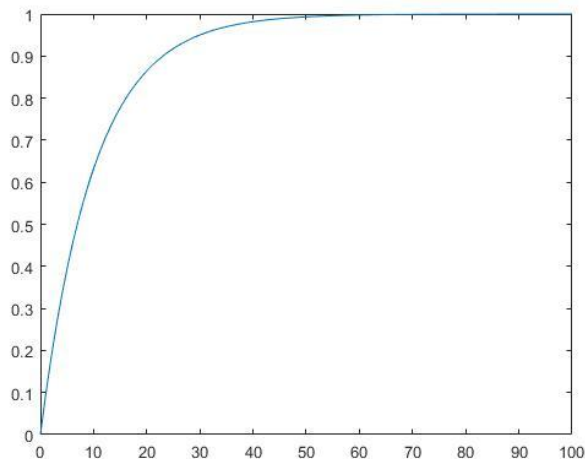
$$F(x) = \int_{-\infty}^x f(u)du$$

Because $f(x)$ is defined only for $x \geq 0$, the lower bound on the integral can be 0:

$$F(x) = \int_0^x \frac{1}{10}e^{-u/10}du$$

$$F(x) = -e^{-x/10} - \left(-e^{0/10}\right) = -e^{-x/10} + 1$$

Let's look at the graph of $F(x)$:



See how it is non-decreasing and approaches 1 as x goes to infinity.

Mean

Assume X is a continuous random variable with pdf $f(x)$. We can calculate the mean as the integral of x multiplied by $f(x)$ over the interval from negative infinity to positive infinity.

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

Variance

The variance of X given a pdf $f(x)$ is:

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = E(X^2) - \mu^2$$

Standard Deviation

The standard deviation of X is:

$$\sigma = \sqrt{V(X)}$$

Example 3:

Suppose the probability density function of X is $f(x) = 0.1$ for $0 \leq x \leq 10$, as in Example 1.

What is the mean of X?

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

$$\mu = E(X) = \int_0^{10} 0.1 x dx$$

$$\mu = E(X) = 0.05x^2 \Big|_0^{10} = 0.05 * 100 - 0 = 5 - 0 = 5$$

What is the variance of X?

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = E(X^2) - \mu^2$$

$$\sigma^2 = V(X) = \int_0^{10} (x - 5)^2 * 0.1 dx = \frac{0.1(x - 5)^3}{3} \Big|_0^{10} = 125 * \frac{0.1}{3} - (-125) * \frac{0.1}{3} = 8.333$$

Practice with PDFs

1. Suppose that contamination particle size (in micrometers) can be modeled as $f(x) = 2x^{-3}$ for $1 < x$ and $f(x) = 0$ for $x \leq 1$.

a. Confirm that $f(x)$ is a pdf. (Is the area under the curve equal to one? Is $f(x) \geq 0$ for all x ?) Note that $f(x)$ is 0 for $x \leq 1$, so the area under the curve for values of $x < 1$ is 0.

$$\int_1^{\infty} 2x^{-3} dx =$$

b. Determine $F(x)$, the cumulative distribution function.

$$\int_1^x 2u^{-3} du =$$

c. Determine the mean.

$$E(X) = \int_1^{\infty} x * 2x^{-3} dx = \int_1^{\infty} 2x^{-2} dx =$$

d. What is the probability that the size of a random particle will be less than 5 micrometers?

$$P(X < 5) =$$

e. Suppose an instrument can detect particles > 7 micrometers in size. What percentage of particles can be detected?

$$P(X > 7) =$$

Exponential Distribution

One distribution over a continuous random variable is the Exponential Distribution. Recall that in the Poisson Distribution, we talk about the number of potholes in a stretch of highway or the number of flaws in a copper wire or the number of customers that arrive in a time window. Here, the unit is always a specific, discrete number ≥ 0 . The Poisson distribution is over a discrete random variable. But, the continuous random variable exponential distribution is related to the Poisson process, too.

What if, instead, we want to know the spacing between the events? For example, what is the distance between the flaws in a copper wire? What is the distance between potholes in a highway? What is the time between customer arrivals to a post office? Note that distance is measured with a continuous valued number.

The number of potholes in 10 miles certainly must be related to the distance between the potholes.

Intuition check: If the number of potholes along a 10-mile stretch of highway goes up, what happens to the distance between the potholes?

We can derive the exponential distribution from the Poisson distribution. We can think of N = number of flaws in x millimeters of copper wire, for example. Then, the distance X between flaws has the following property: $P(X > x) = P(N=0)$ since there are no flaws within x millimeters of copper wire. Assume the mean number of flaws is λ per millimeter, so N has a Poisson distribution with mean λx .

$$P(X > x) = P(N = 0) = \frac{e^{-\lambda x} (\lambda x)^0}{0!} = e^{-\lambda x}$$

Using the complement, we know then that $P(X \leq x)$ is 1 minus $P(X > x)$:

$$P(X \leq x) = 1 - e^{-\lambda x}$$

We know this represents the cumulative distribution function:

$$F(x) = P(X \leq x) = 1 - e^{-\lambda x} = \int_{-\infty}^x f(t) dt$$

We know from the properties of e :

$$\frac{d}{dx} e^x = e^x$$

$$f(x) = \frac{d}{dx} (1 - e^{-\lambda x}) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0$$

This is an exponential distribution. Therefore, the random variable X that equals the *distance* between successive events of a Poisson process with mean $\lambda > 0$ has an exponential distribution with parameter λ . The pdf (probability density function) of X is:

$$f(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0$$

$$E(X) = \frac{1}{\lambda}$$

$$V(X) = \frac{1}{\lambda^2}$$

Note that this is a continuous distribution, since x is any real number greater than or equal to 0. With the exponential distribution, we **integrate** the function from (a, b) if we want to find $P(a < X < b)$. We integrate from (a, ∞) for $P(X > a)$. We integrate from $(-\infty, a)$ to find $P(X < a)$. Since it is a pdf, $P(X > a) = P(X \geq a)$. Likewise, $P(a < X < b) = P(a \leq X \leq b)$.

Again, just like with the Poisson distribution, it is important to use consistent units.

Example 1: Call Times

The call times to a medical clinic is exponentially distributed with a mean time between calls of 12 minutes.

a) What is the probability that there are no calls within a 30-minute interval?

Let X denote the time until the first call. X is exponential and $\lambda = 1/E(X)$. So $\lambda = 1/12$ calls/minute. We need to determine the probability that X is greater than 30 (no calls within 30-minute interval):

$$P(X > 30) = \int_{30}^{\infty} \frac{1}{12} \left(e^{-\frac{x}{12}} \right) dx = -e^{-\frac{x}{12}} \Big|_{30}^{\infty} = 0 - (-e^{-2.5}) = e^{-2.5} = 0.0821$$

Does this probability make sense?

b) What is the probability that at least one call arrives within a 10-minute interval?

Here, we can use the complement. We can find the probability of *no calls* in a 10-minute interval:

$$P(X \leq 10) = 1 - P(X > 10)$$

$$P(X > 10) = \int_{10}^{\infty} \frac{1}{12} \left(e^{-\frac{x}{12}} \right) dx = -e^{-\frac{x}{12}} \Big|_{10}^{\infty} = 0 - (-e^{-5/6}) = 0.4346$$

$$P(X \leq 10) = 1 - 0.4346 = .5654$$

Does this probability make sense?

c) What is the probability that the first call arrives within 5 and 10 minutes?

Now, we find the probability $P(5 < x < 10)$:

$$P(5 < X < 10) = \int_5^{10} \frac{1}{12} \left(e^{-\frac{x}{12}} \right) dx = -e^{-\frac{x}{12}} \Big|_5^{10} = \left(-e^{-\frac{10}{12}} \right) - \left(-e^{-5/12} \right) = 0.2246$$

d) Determine the *length of time* such that the probability of at least one call arriving in the interval is 0.90.

Now, we are trying to figure out x instead of figuring out the probability.

$$\begin{aligned} P(X < x) = .90 &= \int_0^x \frac{1}{12} \left(e^{-\frac{t}{12}} \right) dt = \left(-e^{-\frac{x}{12}} \right) - \left(-e^0 \right) = \left(-e^{-\frac{x}{12}} \right) + 1 \\ .90 &= \left(-e^{-\frac{x}{12}} \right) + 1 \\ -.10 &= \left(-e^{-\frac{x}{12}} \right) \\ \ln(.10) &= -x/12 \\ -12 * \ln(.10) &= x \\ x &= 27.63 \text{ minutes} \end{aligned}$$

Of note: With the Poisson process, the probability does not change after a successful event. So, the starting point of observations does not matter. In the pothole example, a pothole at milepost 1 does not affect the probability of potholes between milepost 1 and 2. It would be the same as the probability of potholes between milepost 2 and 3 (or any other same length interval). So, the Poisson process may not be a good estimate for some models. Take, for example, the phone calls to a medical clinic. The time of day may, in fact, matter (it would be reasonable to have more calls in the morning soon after the office opens). The Poisson process, therefore, has the **lack of memory property** – knowledge of prior events does not impact future subintervals.

The exponential distribution is often used in reliability studies – how long until a device fails. The lifetime of a cell phone camera might be modeled as an exponential random variable with mean 1000 days. So, this model does not take into consideration “wearing out” or “use”. The probability of failure in the next 100 days does not depend on how “old” the camera is. Such reliability studies where “wear” is important may be better modeled by a distribution where there is “memory”.

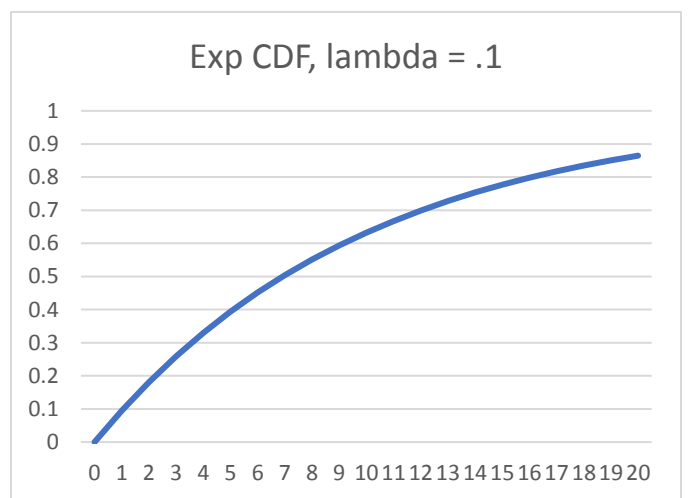
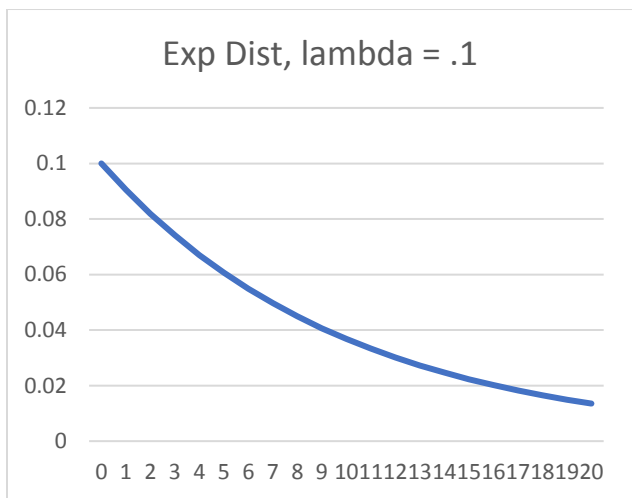
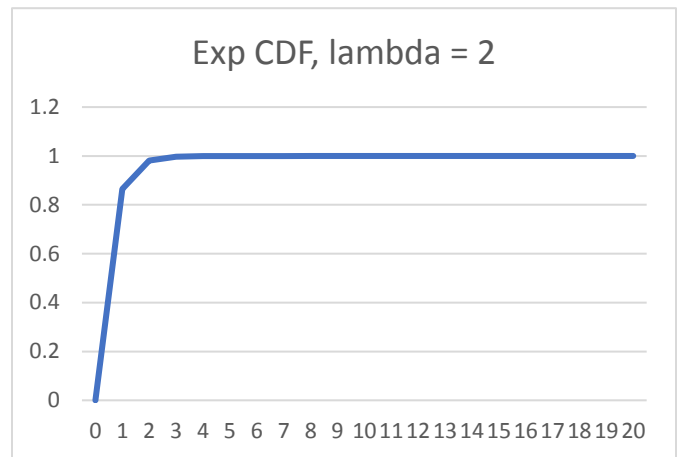
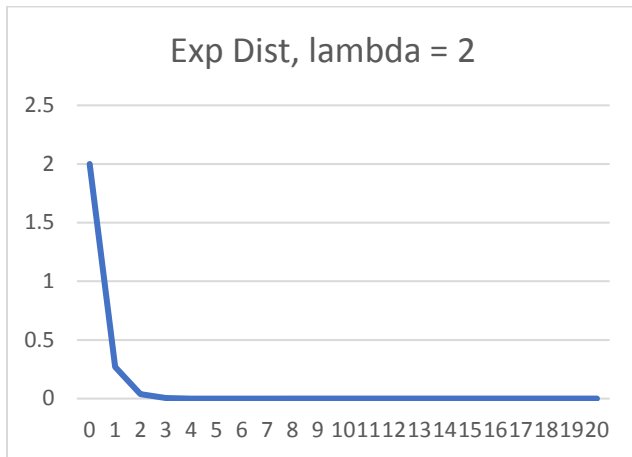
Excel

EXPON.DIST(x, lambda, FALSE)

// for $P(X=x)$

EXPON.DIST(x, lambda, TRUE)

// for $P(X < x)$ [cumulative]



Distribution Practice

Directions: Get into a team of 4 people. Your team will work through problems related to the binomial, poisson, and exponential distributions. Work together as a team – do not divide and conquer the problems. Each team member is assigned to one of the following roles:

- Notetaker: prepares the master copy to be submitted
- Manager: ensures that group stays on task, manages time, and ensures everyone gets to speak
- Spokesperson: when called upon, speaks on behalf of the team
- Reflector: analyzes team dynamics and makes suggestions for improving team process – is everyone contributing? Does everyone understand how to get to the solution?

Names: Notetaker: _____

Manager: _____

Spokesperson: _____

Reflector: _____

Here's the Poisson distribution function:

$$f(x) = P(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$
$$E(X) = \lambda$$
$$V(X) = \lambda$$

Here's the exponential distribution function:

$$f(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0$$

$$E(X) = \frac{1}{\lambda}$$
$$V(X) = \frac{1}{\lambda^2}$$

1. Suppose the counts recorded by a Geiger counter follow a Poisson process with an average of 3 counts per minute.

a) What is the probability that there are no counts in a 30-second interval? (First, use the Poisson distribution for this)

$$P(X = 0, \lambda = 1.5) =$$

(Now, this could also be modeled with the exponential distribution since we could see what the probability is for no counts in a 30-second interval)

$P(X > 0.5) =$ (use exponential distribution with lambda equal to 3) =

Did you get the same value?

b) What is the probability that the first count occurs in less than 10 seconds?

$P(X < 1/6) =$

c) What is the mean time between counts? _____ Does this make sense?

d) What is the length of time x such that the probability that at least one count occurs before time x hours is 0.95?

$P(X < x) = 0.95$

Set this equal to the exponential distribution from 0 to x .

Checkpoint 1: Stop for class discussion. If your team has reached this checkpoint, you can move on. Be ready to share your team answers with the class.

2. For each of these statements, would you use the Binomial Distribution, Poisson Distribution, or Exponential Distribution?

a) An electronic product contains 40 integrated circuits. The probability that any IC is defective is 0.01 and the integrated circuits are independent. The product operates only if there are no defective ICs. What is the probability that product operates?

b) The time between the arrival of email messages is exponentially distributed with a mean of 2 hours. What is the probability that you do not receive an email message during a 2-hour period?

c) Suppose the number of defects in a copper wire follows a Poisson process with mean 5 defects per meter. What is the probability that the distance from the beginning of the wire to the first defect is 2 meters?

d) Suppose the number of defects in a copper wire follows a Poisson process with mean 5 defects per meter. What is the probability that there are 3 defects in the first 2 meters?

e) Suppose the number of earthquake tremors in a 12-month period appears to be distributed as a Poisson random variable with mean of 6. Assume the tremors are independent from each 12-month period to the next. What is the probability that there are more than 5 tremors in a 6-month period?

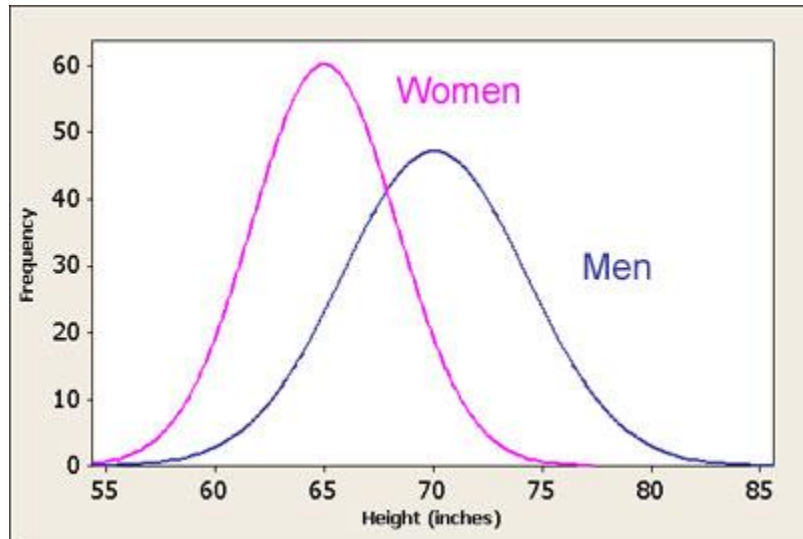
f) Suppose the number of earthquake tremors in a 12-month period appears to be distributed as a Poisson random variable with mean of 6. Assume the tremors are independent from each 12-month period to the next. What is the probability that the time between two tremors is 24 months?

Checkpoint 2: Stop for class discussion. If your team has reached this checkpoint, try to come up with your own questions/statements in which the Poisson distribution or exponential distribution can be applied. Be ready to share your team answers with the class.

The Normal Distribution (also known as Gaussian Distribution)

What is normal?

Let's consider the heights of women and men: (Image courtesy of Matlab tutorial)



The **mean** for women is about 65 inches and the mean for men is about 70 inches. The mean tells us the peak.

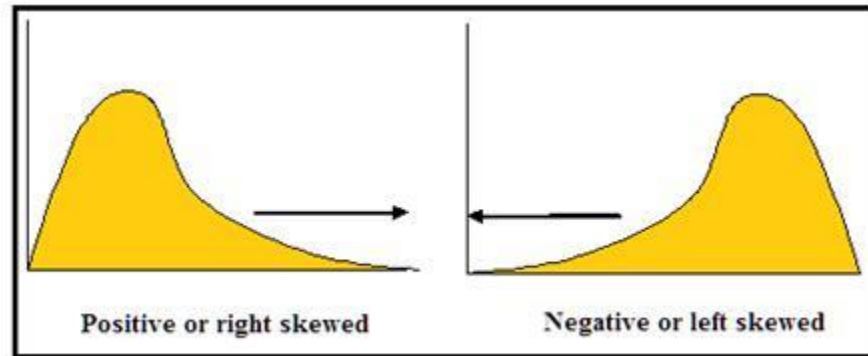
Does the height of women or the height of men have more variance? (Hint: What are the standard deviations for the curves. Which one is wider? The standard deviation tells us the width of the curve.)

The frequencies are percentages, or probabilities. Note that if the frequencies are probabilities, the area under the curve would be 1.

The normal distribution is *symmetric* about the mean. Draw a dotted line from the top of the curve to the x-axis in the above graph.

Aside: Skewness

- If the majority of the data falls to the right of the mean, it is *left-skewed*.
- If the majority of the data falls to the left of the mean, it is *right-skewed*.
- How will I remember this: look for the tail – the tail is the direction of skewness.

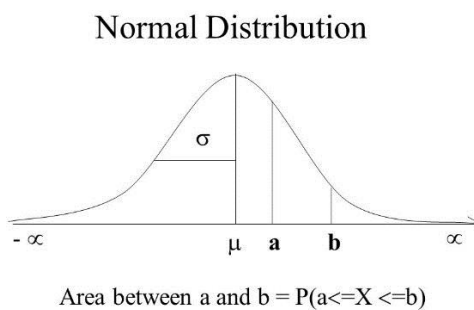


A **normal random variable X** has a normal distribution with parameters μ (population mean) and $\sigma > 0$ (population standard deviation) with the following pdf:

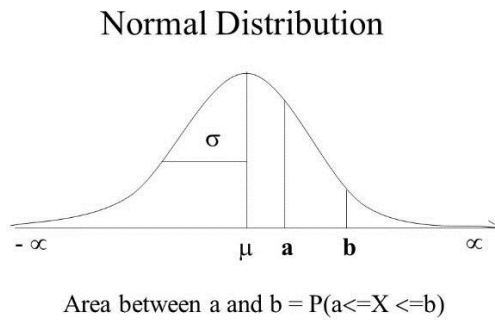
$$f(x) = \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

Where $e \sim 2.7183$ and $\pi \sim 3.1416$. However, this is not easy to integrate to find area under the curve. Instead, we will convert to the standard normal distribution and use a table of values that closely estimate the area under the curve.

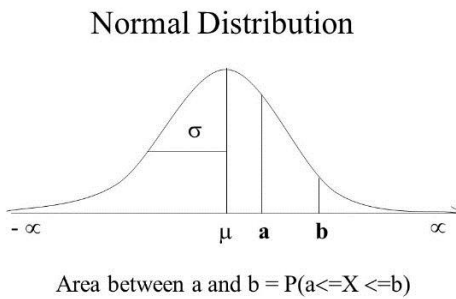
Practice: For each of the distributions below, shade the area for **$P(a \leq X \leq b)$ or $P(a < X < b)$** :



Shade the area for $P(X < b)$:

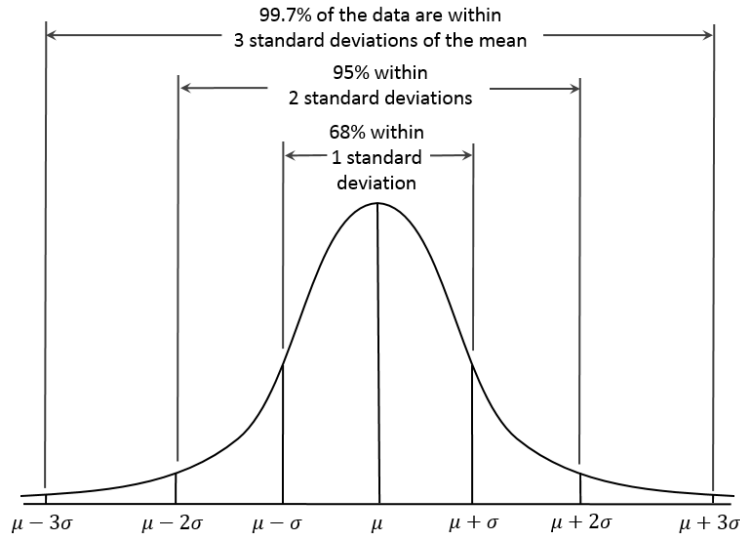


Shade the area for $P(X \geq a)$:



Properties of a normal distribution:

- **mean = median**
- It is **symmetric**.
- There are **no gaps** in the curve.
- It **never touches the x-axis**; gets closer and closer to 0 as x approaches negative infinity and positive infinity.
- The **area under the curve is 1**.
- 99.7% of the area under of the curve is within 3 standard deviations.
- $N(\mu, \sigma)$ denotes a normal distribution with mean μ and standard deviation σ .



Z – standard normal random variable

What's nice about the normal distribution is that we can always map any normal distribution to have mean 0 and standard deviation of 1 by shifting the curve right or left and by squishing or spreading the curve. We can adjust any normal distribution to this *standard normal* random variable Z.

$$\Phi(z) = P(Z \leq z)$$

Is the cdf of the standard normal random variable Z. We cannot determine phi (Φ) directly, so we use a table or computer software to calculate the cdf values.

Look in book on pages 486 and 487 for Table 1. Let's practice using the table:

Find the following values:

$$P(Z < 0)$$

$$P(Z < 2.6)$$

$$P(Z < 1.25)$$

$$P(0 < Z < 2.6)$$

//Hint: This is equal to $P(Z < 2.6) - P(Z < 0)$

//Draw the curve to convince yourself

$$P(Z > 2.6)$$

//Hint: This is equal to $1 - P(Z < 2.6)$

$$P(Z > -1.2)$$

$$P(-1.2 < Z < 2.6)$$

We can map any normal distribution to the **standard normal** by (top is shifting center, bottom is squishing or stretching the curve):

$$Z = \frac{X - \mu}{\sigma}$$

So, standardizing X (normal random variable with mean μ and variance σ^2) is:

$$P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P(Z \leq z) \quad \text{where } z \text{ is the } z\text{-value by standardizing } x$$

Example 1:

Suppose the mean number of hours a person is on their cell phone per day is 3.1 hours with standard deviation of 0.5 hours and the number of hours is distributed normally. What percent of people spend less than 3.5 hours per day on their cell phone?

First, let's determine our random variable X:

X = mean number of hours per day a person is on their cell phone

We want to determine $P(X < 3.5)$.

Second, we need to standardize to z:

$$z = \frac{x - \mu}{\sigma} = \frac{3.5 - 3.1}{0.5} = 0.8$$

We want $P(Z < 0.8)$. Look in the table:

$P(Z < 0.8) = .788145$

So, 78.8% of people are on their phones less than 3.5 hours per day.

Example 2:

The average household generates 28 lbs of garbage per week with a standard deviation of 2 lbs and the weight is distributed normally. If a household is selected at random, what is the probability of it generating:

- a) Between 27 and 31 lbs of garbage?
- b) More than 30.2 lbs of garbage?
- c) What weight do 90% of homes exceed?

First, let's define X:

X = lbs of garbage generated per week

Part a:

We need to standardize to Z:

$$z_1 = \frac{x - \mu}{\sigma} = \frac{27 - 28}{2} = -0.5$$

$$z_2 = \frac{x - \mu}{\sigma} = \frac{31 - 28}{2} = 1.5$$

We want:

$$P(-0.5 < Z < 1.5) = P(Z < 1.5) - P(Z < -0.5) = .933193 - .308538 = .624655$$

Part b:

We need to standardize to Z:

$$z = \frac{x - \mu}{\sigma} = \frac{30.2 - 28}{2} = 1.1$$

$$P(Z > 1.1) = 1 - P(Z < 1.1) = 1 - .864334 = .135666$$

Part c:

We need to figure out the value of x such that $P(X > x) = .90$.

In standard normal, we need to find the value of z such that $P(Z > z) = .90$. Looking at the table, we can find the value of z such that $P(Z < z) = .10$ and that happens at -1.29 (note that z = -1.28 is just a little over .1).

Now, we need to convert to our units for x:

$$-1.28 = \frac{x - 28}{2}$$

$$x = 25.44 \text{ lbs}$$

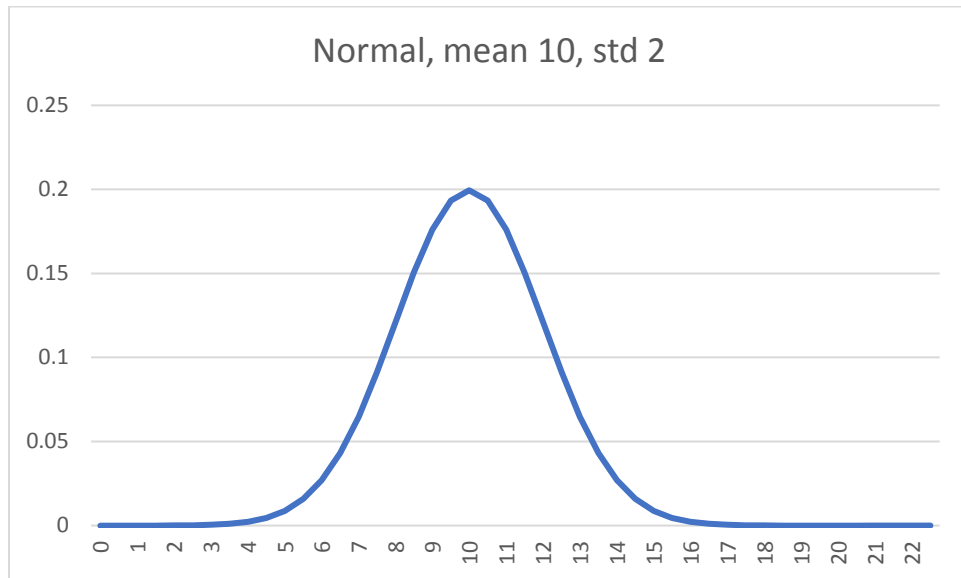
How to check for normality:

1. Plot histogram
2. Use Pearson's Index of Skewness ($PI = 3(\bar{x} - \text{median}) / s$). If $PI \geq 1$ or $PI \leq -1$, the data is skewed and NOT normal
3. Check for outliers ($< Q1 - 1.5 * IQR$ or $> Q3 + 1.5 * IQR$). If two or more, reject normality.
4. Use normal probability plot and special graph paper to plot the data (sorted). The observations are plotted against the cumulative frequency. If the data are normal, the points will fall on a straight line. (See pages 92 – 93 in textbook for this method.)

Excel

=NORM.DIST(x, mean, stddev, FALSE) // PDF

=NORM.DIST(x, mean, stddev, TRUE) // CDF



Normal Distribution Practice

Directions: Get into a team of 4 people. Your team will work through problems related to distributions. Work together as a team – do not divide and conquer the problems. Each team member is assigned to one of the following roles:

- Notetaker: prepares the master copy to be submitted
- Manager: ensures that group stays on task, manages time, and ensures every person gets to speak
- Spokesperson: when called upon, speaks on behalf of the team
- Reflector: analyzes team dynamics and makes suggestions for improving team process – is everyone contributing? Does everyone understand how to get to the solution?

Names: Notetaker: _____

Manager: _____

Spokesperson: _____

Reflector: _____

Review

Recall that the pdf $f(x)$ of a continuous random variable X is:

$$P(a < X < b) = \int_a^b f(x)dx$$

Recall that the cdf $F(x)$ of a continuous random variable X with pdf $f(x)$ is:

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u)du \quad \text{for } -\infty < x < \infty$$

Mean:

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)dx$$

Variance:

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx = E(X^2) - \mu^2$$

Standard deviation is the square root of variance.

Converting to standard normal:

$$z = \frac{x - \mu}{\sigma}$$

Exercises

Work with your group to solve these problems. State with clarity the values you are computing. For example, if the random variable is X and you want to know the probability that $X < 10.5$, use the notation $P(X < 10.5)$.

1. Let's practice using Table I in Appendix A of the book.

Find the values of z such that the following is true:

a. $P(Z < z) = 0.5$

b. $P(Z < z) = 0.001001$

c. $P(Z < z) = 0.881000$

d. $P(Z > z) = 0.866500$ // note that this is $Z > z$

e. $P(-1.3 < Z < z) = .863140$

2. Assume Z has a standard normal distribution. Use Table I to determine the probabilities for Z :

a. $P(Z < 1)$

b. $P(Z < 3)$

c. $P(-1.5 < Z < 1.5)$

d. $P(0 < Z < 2)$

e. Do these probabilities make sense?

META-QUESTION: Does everyone understand the notation? Does everyone understand how to use Table I? If not, how could the team improve its process, so that all people learn?

Checkpoint 1: Stop for class discussion. If your team has reached this checkpoint, you may move on to the next question. Be ready to share your team answers with the class.

3. The compressive strength of samples of cement can be modeled by a normal distribution with mean 6000 kg/sq cm and a standard deviation of 100 kg/sq cm.

a. What is the probability that a sample's strength is less than 6250 kg/cm²?

First, X = the compressive strength of a cement sample.

We want to find $P(X < 6250)$.

Second, convert to the standard normal distribution and find the z-value:

Use Table I to determine $P(Z < z)$:

b. What is the probability that a sample's strength is between 5800 and 5900 kg/cm²?

c. What strength is exceeded by 95% of the samples? ($P(X > x) = .95$; Find x)

4. The fill volume of an automated filling machine used for filling cans of soda is normally distributed with mean 12.4 fluid ounces and a standard deviation of 0.1 fluid ounce.

a. What is the probability that a fill volume is less than 12 fluid ounces?

b. If the manufacturer deems that cans less than 12.1 or greater than 12.6 fluid ounces need to be scrapped, what proportion of cans need to be scrapped?

c. What is the range (symmetric about mean) such that 99% of the cans are included?

5. Does your team have questions about PDFs, CDFs, or the normal distribution?

Checkpoint 2: Stop for class discussion. If your team has reached this checkpoint, you can move on. Be ready to share your team answers with the class.

(If time) Come up with your own problems regarding the normal distribution. Specify a mean, a standard deviation, and what you want to calculate.

EGR361 Exam 1 Review Sheet

Content: Exam 1 will cover chapters 1 to 3.9 of the textbook. Material will be drawn from homework assignments, lectures, in-class activities, and the textbook.

Procedure: Please arrive to class on time. You may use **one** sheet of 8/5" x 11" paper (**both sides**) during the exam. You may use a scientific calculator or regular calculator (**NO cell phone, NO computer**) during the exam. You may NOT use ear buds and other electronic devices during the exam.

Table I (standard normal distribution) will be provided to you during the exam.

Topics: This study guide is not a contract – in other words, the exam may not cover every topic listed below and there may be topics that we covered in class that are not explicitly listed.

Chapter 1:

- Sampling
 - Random samples, simple random sampling
 - Populations
- Types of studies
 - Enumerative vs analytic
 - Retrospective, Observational, Designed
- Models
 - Mechanistic vs empirical

Chapter 2:

- Calculating descriptive statistics
 - Sample mean (\bar{x})
 - Population mean (μ)
 - Sample variance (s^2)
 - Population variance (σ^2)
 - Sample standard deviation (s)
 - Population standard deviation (σ)
 - 5-number summary (Q0 to Q4)
 - 3 Methods for calculating Q1 and Q3 (in class)
- Plots
 - Stem-and-leaf (see book)
 - Histograms (in class and in book)
 - Box Plots (in class and in book)
 - Time series plots (see book)

Chapter 3:

- Random variables (discrete vs continuous) (usually denoted X)
- Probability (likelihood, chance)
 - Venn Diagrams, Sets, Complement, Union, Intersection
 - Outcome: Result of a single trial

- Sample space: Set of all outcomes (sum of probabilities of all outcomes is 1)
- Mutually Exclusive (Intersection is empty between two events)
- Independence (Outcome of event 1 does not impact outcome of event 2)
- Conditional Probability
- Counting
 - Permutations
 - Combinations
- Discrete Random Variables and Distributions
 - PMF (probability mass function) – how to determine if $f(x)$ is a pmf
 - CDF (cumulative distribution function) – piecewise continuous function
 - Calculate mean (μ) or $E(X)$
 - Calculate variance (σ^2) or $V(X)$
 - Calculate standard deviation (σ)
- Binomial Distribution
 - When to use it? How to use it?
 - Example: the probability of 3 parts having flaws of 100 parts, where the probability for a flawed part is 0.1
 - $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, 2, 3 \dots n$
 - $\mu = E(X) = np$
 - $\sigma^2 = V(X) = np(1-p)$
- Poisson Distribution
 - When to use it (intervals are independent, usually referring to # events over an interval, $np = \lambda$), How to use it?
 - Example: the probability of 10 customers arriving to a store in an hour, where the mean is 20 customers per hour and customer arrival follows a Poisson process
 - Remember: λ needs to be in the correct units; must match units in question
 - $f(x) = P(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$, $x = 0, 1, 2, \dots$
 - $\mu = E(X) = \lambda$
 - $\sigma^2 = V(X) = \lambda$
- Continuous Random Variables
 - PDF (probability density function)
 - Area under curve sums to 1
 - $f(x) \geq 0$ for all x
 - $P(a < X < b) = \int_a^b f(x) dx$
 - CDF
 - $F(x) = P(X \leq x) = \int_{-\infty}^x f(u) du$ for $-\infty < x < \infty$
 - $\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$
 - $\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = E(X^2) - \mu^2$
- Exponential Distribution
 - Related to Poisson: The spacing between events in a Poisson distribution
 - Example: the probability that the distance between flaws in a copper wire is between 0.5 and 1 cm where the mean number of flaws is 2 per cm and the number of flaws follows a Poisson process.

- Can also stand on own as a distribution
 - Example: the probability that there are no calls in a 30-minute window and the call times are exponentially distributed with mean time between calls of 10 minutes
- How to use it? Need to integrate $f(x)$ over (a,b) if looking for $P(a < X < b)$.
- Remember: mean needs to be in correct units; must match units in question
- $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$
- $\mu = E(X) = \frac{1}{\lambda}$
- $\sigma^2 = V(X) = \frac{1}{\lambda^2}$
- Normal Distribution
 - Symmetric about mean; mean = median
 - Mean is at peak
 - Standard deviation is related to width
 - $f(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$
 - Z – standard normal random variable
 - $Z = \frac{X-\mu}{\sigma}$
 - Integrating is no fun, so there is a table of probabilities for the standard normal
 - Table I in your book: values are the CDF of the standard normal distribution
 - Checking for normality

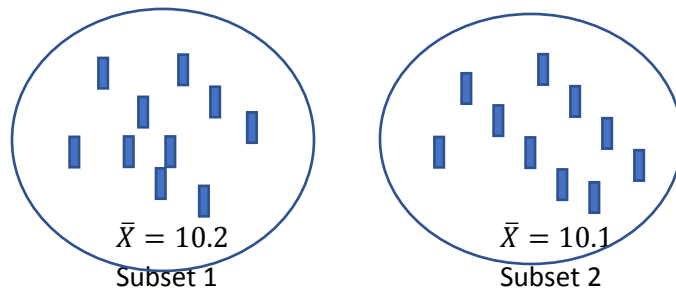
Excel Skills:

- Produce a histogram
- Produce a box plot
- Calculate sample mean, variance, standard deviation of observations
- Calculate population mean, variance, standard deviation of population
- Calculate the 5-number summary
- Produce distributions: Binomial, Poisson, Exponential, Normal

Random samples, statistics, central limit theorem

Suppose we want to study the population of 5000 bolts. We only have time and resources to test 10 bolts. As we learned earlier in the course, we can choose a simple random sample of 10 bolts from 5000 bolts.

As you can see, we have lots of choices for a set of 10 bolts – in fact, we know we have $\binom{5000}{10}$ possible subsets of bolts from our earlier work with combinatorics.

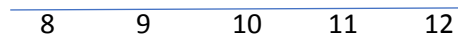


Subset 1: We measure the weight of the bolt for all 50 bolts. We calculate the mean and find it to be **10.2 ounces**.

Subset 2: We measure the weight of the bolt for all 50 bolts. We calculate the mean and find it to be **10.1 ounces**.

Suppose we do this activity for 20 different subsets, with replacement. We have 20 *sample means* now. We can plot the distribution of these means.

What do you expect this distribution of sample means to look like?



This is an example of the sampling distribution of sample means. It is a distribution obtained by using the means computed from random samples of a specific size taken from a population.

The sample means could be different than the population mean μ . These differences are caused by sampling error. If we actually took all possible subsets of size n of N items in our population and calculated all the sample means, then the mean of the sample means will be the same as the population mean.

Do you expect the *standard deviation of the sample means* to be smaller or bigger than the standard deviation of the original distribution?

Smaller

Bigger

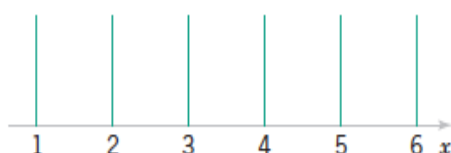
Experiment

Suppose we do a dice experiment. Because this one would take a long time to run in class, we'll work through it mentally. You could do this on your own time if you wish.

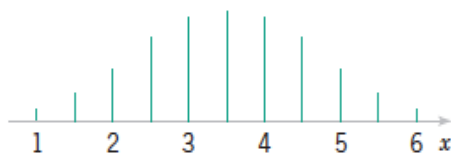
Here's the experiment: Roll a set of N dice. Calculate the mean of the numbers rolled.

What do you expect the mean to be when 2 dice are rolled? _____

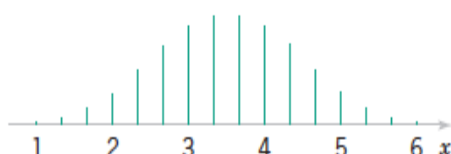
What do you expect the mean to be when 3 dice are rolled? _____



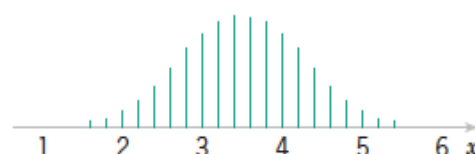
(a) One die



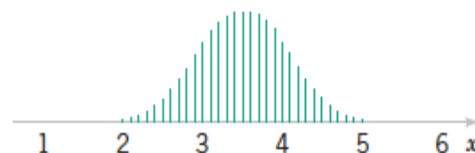
(b) Two dice



(c) Three dice



(d) Five dice



(e) Ten dice

(Figures from textbook, page 139)

What is happening to the distribution as N gets large?

What is the mean converging to?

This leads us to the **Central Limit Theorem**:

As the sample size n increases without limit, the shape of the distribution of the sample means $\{X_1, X_2, X_3, \dots, X_n\}$ taken with replacement from a population with mean μ and standard deviation σ will approach a normal distribution with mean μ and standard deviation σ/\sqrt{n} . This works even when the shape of the probability distribution of the population is unknown.

Because the mean of the sample means is approximately normal, we can standardize to Z to calculate probabilities as we did with the regular normal distribution. REMEMBER: this is the mean of the sample means, so the standard deviation is σ/\sqrt{n} . \bar{X} is the mean of the sample means. μ is the population mean. n is the sample size.

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

We also call the standard deviation of the sample means the **standard error of the mean**:

$$\frac{\sigma}{\sqrt{n}}$$

Note: The central limit theorem applies to the mean of the sample means.

Note: Remember your Z statistic. For a regular normal distribution (calculating probabilities for an individual data value for a normal distribution), you use $Z = \frac{x - \mu}{\sigma}$. When using the central limit theorem about a sample mean, you use

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Note: In general, when n is ≥ 30 , you may use the central limit theorem. With values $n \geq 4$, you may use the central limit theorem if the original population data are normally distributed.

Example 1

The average number of pounds of meat that a person consumes a year is 218.4 pounds. Assume the standard deviation is 25 pounds and the distribution is approximately normal.

a. What is the probability that a person selected at random consumes less than 224 pounds of meat per year?

First, we need to determine if we should use the normal distribution Z calculation or the central limit theorem Z calculation.

Is the problem asking about a specific person? Yes

Is the problem asking about a random sample? No

We use the Z calculation: $Z = \frac{x - \mu}{\sigma}$

$$Z = \frac{224 - 218.4}{25} = 0.22$$

$$P(Z < 0.22) = .587064$$

b. If a sample size of 40 individuals is selected, find the probability that the mean of the sample will be less than 224 pounds per year.

First, we need to determine if we should use the normal distribution Z calculation or the central limit theorem Z calculation.

Is the problem asking about a specific person? No

Is the problem asking about a random sample? Yes

We use the Z calculation:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$
$$z = \frac{224 - 218.4}{25/\sqrt{40}} = 1.42$$

$$P(Z < 1.42) = .922196$$

Does this make sense? The probability of a single person consuming less than 224 pounds of mean is .59 while the probability of the average consumption of a group of 40 people being less than 224 pounds is .92.

Example 2

The average age of a vehicle registered in the US is 8 years (96 months) old. Assume the standard deviation is 16 months. If a random sample of 36 vehicles is selected, find the probability that the mean of their age is between 90 and 100 months.

Which Z should we use? _____

$$z1 = \frac{90 - 96}{16/\sqrt{36}} = -2.25$$
$$z2 = \frac{100 - 96}{16/\sqrt{36}} = 1.5$$

$$P(Z < 1.5) = .933193$$

$$P(Z < -2.25) = .012224$$

$$\text{So, } P(-2.25 < Z < 1.5) = .933193 - .012224 = .920969$$

Example 3

Suppose X is normally distributed with mean 100 and standard deviation 9. Compute the following for n = 16.

a. Mean of \bar{X}

100 (using the central limit theorem, the mean of the sample means of a normally distributed population is the population mean)

b. Variance of \bar{X}

The standard deviation is:

$$\frac{\sigma}{\sqrt{n}}$$

So, the variance is the square of this.

$$\text{Variance} = (9/4)^2 = 81/16$$

c. $P(\bar{X} < 103)$

$$z = \frac{103 - 100}{9/4} = 1.33$$

$$= .908241$$

Normal Approximation to Binomial

The binomial distribution can be modeled with a normal distribution when n gets large with the following parameters:

$$\begin{aligned}\mu &= np \\ \sigma^2 &= np(1 - p)\end{aligned}$$

Note that the binomial distribution is over discrete values and the normal distribution is over continuous values. When the value of p is close to 0 or close to 1, the binomial distribution is quite skewed and the symmetric normal distribution may not be a good approximation.

So, when n is large and p is neither close to 0 nor close to 1 and we are solving a problem that has a binomial distribution, we can approximate with a standard normal:

$$Z = \frac{X - np}{\sqrt{np(1 - p)}}$$

Generally, if $np > 5$ and if $n(1-p) > 5$, we can use the normal approximation.

Normal Approximation to Poisson

When n is large, we can also approximate the Poisson distribution with a normal distribution with the following parameters:

$$\begin{aligned}\mu &= \lambda \\ \sigma^2 &= \lambda\end{aligned}$$

Converting to standard normal:

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

This approximation is good when $\lambda > 5$.

Why is this helpful?

Consider a problem such as the following:

Suppose the number of particles in a liquid sample follows a Poisson distribution with mean of 1000 per fluid ounce. What is the probability that $X \leq 950$ per fluid ounce?

Ugh, we would have to sum all these together if doing this by hand:

$$P(X=0) + P(X=1) + P(X=2) + \dots + P(X=950).$$

Excel does this nicely for us with the cumulative distribution. But we could also use the approximation to the normal distribution to find the z -value that corresponds to 950.5 from $N(1000, \text{sqrt}(1000))$.

When doing the approximation, add 0.5 or subtract 0.5 from the x -value depending on if you are looking for $P(X > x)$ or $P(X < x)$. For example, if we want the probability that $X < 91$, with the approximation, we would use $P(X < 90.5)$. This is due to nature that the normal distribution is *continuous* and the Poisson distribution is *discrete*. If you want the probability that $X > 30$, you would use $P(X > 30.5)$.

Example 4

Suppose you have 362,000 customer accounts for water service. The accounts are metered and billed monthly. The probability that an account has an error in a month is 0.001. Accounts are assumed independent.

- What is the mean number of account errors every month?

$$E(X) = np = 362000 * 0.001 = 362$$

- What is the standard deviation?

This problem would be solved with a binomial distribution, so we can use the approximate normal distribution for this:

$$V(X) = np(1-p) = 362 * (.999) = 35.838$$

The standard deviation is the square root: 19.017

- c. Approximate the probability that fewer than 350 errors happen in a month.

We'll use the normal distribution approximation. Since we are using a continuous distribution to approximate a discrete distribution, we need to add 0.5 or subtract 0.5 from the target. We are looking for less than 350 errors, so we use the target value of x as 349.5.

$$Z = \frac{349.5 - 362}{19.017} = -.657$$

Using Table I, we see that $P(Z \leq -.66)$ is .254627.

- d. Approximate a value x such that the probability of the number of errors being less than x is 0.05.

$$-1.65 = \frac{(x - 0.5) - 362}{19.017}$$

x = 330.7 errors

FYI: Many more continuous distributions

There are other distributions that you may use when modeling engineering data.

Lognormal Distribution

- The natural log of X is normally distributed.
- $X > 0$
- Example use: lifetime of a product degrades over time, such as a semiconductor laser.
- Let W have a normal distribution with mean θ and variance ω^2 ; then $X = \exp(W)$ is a lognormal random variable with pdf:

$$f(x) = \frac{1}{x\omega\sqrt{2\pi}} \exp\left[-\frac{(\ln(x) - \theta)^2}{2\omega^2}\right] \quad 0 < x < \infty$$

$$E(X) = e^{\theta + \omega^2/2}$$
$$V(X) = e^{2\theta + \omega^2}(e^{\omega^2} - 1)$$

See text for:

- Gamma/Erlang Distribution (useful for random experiments)
- Chi-Squared Distribution (special case of gamma)
- Weibull Distribution (useful for time until failure)
- Beta Distribution (flexible but bounded over finite range)

FYI: Correction Factor for Central Limit for Finite Populations

The central limit theorem assumes samples are drawn with replacement OR drawn without replacement from a very large or infinite population. Sampling with replacement may be unrealistic for the problem. There is a correction factor for drawing without replacement from a finite population. The correction factor is:

$$\sqrt{\frac{N - n}{N - 1}}$$

Where N = the size of the population and n = size of the sample.

This factor is multiplied to the standard error of the mean for large samples taken from a small population:

$$\frac{\sigma}{\sqrt{n}} * \sqrt{\frac{N - n}{N - 1}}$$

Let's see what happens when N is large and n is small. The correction factor approaches 1.

Then, the Z value becomes:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n} * \sqrt{\frac{N-n}{N-1}}}$$

When the population is large, we can ignore this correction factor.

Central Limit and Normal Approximations Practice

Directions: Get into a team of 4 people. Your team will work through problems related to distributions. Work together as a team – do not divide and conquer the problems. Each team member is assigned to one of the following roles:

- Notetaker: prepares the master copy to be submitted
- Manager: ensures that group stays on task, manages time, and ensures everyone gets to speak
- Spokesperson: when called upon, speaks on behalf of the team
- Reflector: analyzes team dynamics and makes suggestions for improving team process – is everyone contributing? Does everyone understand how to get to the solution?

Names: Notetaker: _____
Spokesperson: _____

Manager: _____
Reflector: _____

Central Limit Theorem

As the sample size n increases without limit, the shape of the distribution of the sample means $\{X_1, X_2, X_3, \dots, X_n\}$ taken with replacement from a population with mean μ and standard deviation σ will approach a normal distribution with mean μ and standard deviation σ/\sqrt{n} . This works even when the shape of the probability distribution of the population is unknown.

Central limit Z:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Standard error of mean (standard deviation of sample means):

$$\frac{\sigma}{\sqrt{n}}$$

Normal approximation to binomial:

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

Normal approximation to Poisson:

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

Exercises

Work with your group to solve these problems. State with clarity the values you are computing. For example, if the random variable is X and you want to know the probability that $X < 10.5$, use the notation $P(X < 10.5)$.

1. The mean score on a dexterity test for 12-year-olds is 30. The standard deviation is 5 and the scores are normally distributed. If a biomedical engineer administers the test to a group of 22 students, find the probability that the mean of the group will be between 27 and 31.

First, is this asking about a single observation or a group of observations?

What Z value should you use?

$$P(27 < \bar{X} < 31) =$$

2. Use the same data as in problem 1. What is the probability that a randomly selected 12-year-old has a dexterity test score larger than 32?

First, is this asking about a single observation or a group of observations?

What Z value should you use?

$$P(X > 32) =$$

META-QUESTION: Does everyone understand when to use the central limit theorem and when to use the regular standard normal distribution? If not, how could the team improve its process, so that all people learn?

Checkpoint 1: Stop for class discussion. If your team has reached this checkpoint, you may move on to the next question. Be ready to share your team answers with the class.

3. Assume a random sample of 40 observations is drawn from a population with mean 20 and variance

2. Compute the following:

a. Mean of \bar{X}

b. Variance of \bar{X}

c. $P(\bar{X} < 19) =$

d. $P(19 \leq \bar{X} \leq 21.5) =$

4. The viscosity of a fluid can be measured by dropping a small ball into a tube of the liquid and measuring how long it takes to drop one foot. Assume X is the time it takes to drop and that X is normally distributed with mean 20 seconds and standard deviation of 0.5 seconds for a certain liquid.

a. What is the standard deviation (aka standard error of the mean) of the average time of 40 experiments?

b. What is the probability that the average time of 40 experiments will exceed 20.1 seconds?

c. Suppose now, we only have time to conduct 20 experiments. What is the probability that the average time of the 20 experiments will exceed 20.1 seconds?

d. Do the probabilities in (b) and (c) make sense?

5. A semiconductor manufacturer produces chips where 2% are defective. Assume the chips are independent and that a lot contains 1000 chips.

a. Approximate the probability that more than 25 chips are defective. Which approximation are you using? (Remember to use the ± 0.5 correction)

b. Approximate the probability that between 20 and 30 chips are defective.

6. Assume the number of spam emails received in one day follows a Poisson distribution with mean 50. Approximate the following probabilities. Which approximation are you using?

a. More than 40 and less than 60 spam emails a day. (Remember to use the ± 0.5 trick)

b. Less than 40 spam emails a day.

c. More than 340 spam emails in a week.

Checkpoint 2: Stop for class discussion. If your team has reached this checkpoint, you can move on. Be ready to share your team answers with the class.

(if time) Come up with your own problems regarding the central limit and/or normal approximations to the binomial or Poisson distributions.

Measurement: Significant Figures

You want to build a cabinet door to replace a broken door, so you measure the height and width of the existing cabinet door. You get out your tape measure. How *precise* can you measure length with your tape measure?

Now, let's think about a caliper that can measure to 0.01 mm (photo from Home Depot):



A. Which tool is more precise?

Tape Measure

0.01 mm Caliper

B. Would it make sense to record a measurement of 31.236892 feet when using a tape measure?

Yes No

C. Would it make sense to record a measurement of 31.25 inches when using a tape measure?

Yes No

When taking measurements/readings, it is important to record the measurement with the appropriate number of significant figures. Measurements have finite precision (how close it is to the actual value).

D. Suppose two different engineers measure the height of Shiley Hall. One engineer reports 50 feet. Another engineer reports 50.0 feet. Are these the same?

Well, in math, these numbers are equivalent. In measurement/engineering, those numbers are NOT equivalent. Reporting 50 feet just shows that the measurement precision is to the tens place. The 50.0 feet measurement shows the precision is to the tenths place. Always record a number with the appropriate significant figures to show the precision according to the measurement.

If the measurement can be precise to the ones place, you can write it as:

50. feet

Or

5.0×10^1 feet

If you write it as 50, you can only assume precision to the tens place instead of the ones place.

Practice

Let's see if you remember significant figures from prior courses. For each number below, how many significant figures are included?

	# Significant Figures
• 0.00700	_____
• 3.140	_____
• 3500	_____
• 2000.	_____
• -35.200001	_____
• 60.0	_____
• 45	_____
• 3.0020	_____

Let's look at 0.00700

Why does this have 3 significant figures? Well, if the measurement was done in kilometers, we can convert this to 7.00 meters. Maybe the engineer actually measured it with a meter-stick and decided to report the distance in kilometers. The precision level is to the nearest what unit?

Rules for significant figures:

1. Any non-zero digit is significant
2. Zeros in between non-zeros are significant
3. Leading zeros are not significant
4. If there is a decimal point, trailing zeros are significant
5. If there is no decimal point, trailing zeros are ambiguous [assume they are not significant]

This is why you should also put a decimal point, even if the number is 2400 (use 2400.).

If you are making a measurement to the precision of the 100's, and you measure 35000, how do you show that the level of precision is 3 sig figs?

$$3.50 \times 10^4$$

// this is clear to what precision level you made the measurement

Arithmetic: PEMDAS

P: Parentheses or Brackets

E: Exponents

M: Multiplication

D: Division

A: Addition

S: Subtraction

Practice

What is the result of the following expression?

$$1 - \frac{12}{4} + 3^2 \times 5 + 6 =$$

What is the result of the following expression?

$$1 - \frac{12}{4} + 3^2 \times (5 + 6) =$$

Addition/subtraction with significant figures

Suppose you want to cut fabric to cover two adjoining tables. One table is measured to be 1.26 meters long and the other is 2.4 meters long. If we add the lengths of the tables together, in math this is 3.66 meters. But, is this misleading?

It would mean that the precision of measurements of both tables is to the nearest hundredths. So, what matters when adding and subtracting measurements is PRECISION.

$$\begin{aligned} 1.26 \text{ meters} + 2.4 \text{ meters} \\ = 3.66 \text{ (in math)} \\ = 3.7 \text{ meters (in engineering, round to least precise digit of all measurements)} \end{aligned}$$

$$\begin{aligned} 1.26 + 102.3 &= 103.56 \text{ (in math)} \\ &= 103.6 \text{ (in engineering, round to least precise digit)} \end{aligned}$$

Example: Suppose you want the total height of two stacked boxes.

Measurement of first box: 2.09 meters

Measurement of second box: 1.901 meters (used a more precise measuring tool)

$$2.09 + 1.901 = 3.991 \text{ meters (in math)}$$

= 3.99 meters (in engineering, round to least precise measurement)

Example: Suppose you want the total height of a building plus its radio tower.

Height of building: 3.5×10^2 feet

Height of radio tower: 8 feet

A person might say the total height is 358 feet. But, remember the sig figs!

An engineer would say the total height is 3.6×10^2 feet.

+/- Rule: Least precise value

Practice: Combine the following measurements and report with correct # of sig figs

- $35 + 26.2$
- $.002 + 1.1$
- $36.0 - 12$
- $0.91 + 1.2 + 8.4 + 12$
- $1.07 - 0.8826$

Multiplication/division with significant figures

Suppose you want to know the floor area of a room. You can make measurements to the nearest cm:

1.69 meters

2.09 meters

What is the area of the room?

$$1.69 \times 2.09 \text{ meters} = 3.5321 \text{ square meters}$$

What's the problem here? What is the precision of the measurement?

We could only measure to the nearest cm, so the final area should be reported to the nearest cm: 3.53 square meters. We round down since $2 < 5$ and 2 appears in the millimeter position.

\times/\div Rule: Least number of significant figures

Example: Suppose you measure a room and it is 10.1 feet by 12.07 feet (you had two different measurement devices with different levels of precision). A single floor tile is 1.07 square feet. How many tiles fit in the floor?

$$\text{Area} = 10.1 \times 12.07 = 121.907 \text{ square feet (in math)}$$

If this is the final answer, we would then round to 3 sig figs. But we aren't done yet, so DON'T round sig figs until you are **completely finished** with the calculation. Rounding in the middle of a calculation will introduce more error into the final result.

$$\text{Num tiles} = 121.907 / 1.07 = 113.931775701..... \text{ (keeps going) tiles}$$

OK, how many sig figs should we use?

Had 3, 4, 3 sig figs in the original measurement, so we keep 3 sig figs (the least # of sig figs)
114 tiles fit in the floor

Questions about sig figs in multiplication or division?

Example 1 combining $+/-/x/\div$

$$\frac{82.5}{0.18} + 114.25$$

$$459.44 + 114.25 = 573.694$$

Because the minimum of the sig figs for the division is 2, we are uncertain in the tens place of 459.44. The 114.25 means we are uncertain in the hundredths place. The highest uncertainty or left-most uncertain number is in the tens place. So our answer is:
570 or 5.7×10^2 .

Example 2 combining $+/-/x/\div$

$$\frac{0.91 + 1.2 + 8.4}{3.700}$$

$$\frac{10.51}{3.700} = 2.8405$$

Since the precision for the numerator is at best in the tenths, we have 3 sig figs on top and we have 4 sig figs on the bottom, so the final result is 2.84.

Significant Figures Practice

Directions: Get into a team of 4 people. Your team will work through problems. Work together as a team – do not divide and conquer the problems. Each team member is assigned to one of the following roles:

- Notetaker: prepares the master copy to be submitted
- Manager: ensures that group stays on task, manages time, and ensures everyone gets to speak
- Spokesperson: when called upon, speaks on behalf of the team
- Reflector: analyzes team dynamics and makes suggestions for improving team process – is everyone contributing? Does everyone understand how to get to the solution?

Names: Notetaker: _____

Manager: _____

Spokesperson: _____

Reflector: _____

Perform the following calculations and report the answers with the correct number of significant figures. Remember: complete the calculation with full precision first and then determine the sig figs, round the number, and report the final result.

1. $2.8 \times 4.532 + 12.690$

2. $(15.403 - 4.76) / 8.3$

3. $21.3 - 12.39 + 2.432 \times 93.75 / 15$

4. $82.7 / 0.18 + 114.25$

5. $12.62 + 3.4 + 4.11$

Propagation of Error

Let's think about taking measurements. Are they exact? May be in between two marks on a tape measure.

You might find the weight of a bolt to be 3.078 +/- .001 ounces. Measurements are not exact.

Suppose you want to determine the **total height** of Tammy's children. You measure Luke to be 56.25 +/- 0.5 inches and you measure Joel to be 48.50 +/- 0.3 inches. The +/- is the error of the measurement. Now, suppose you want to get the total height of Luke and Joel.

Well, this would be 104.75 inches. But what about that error term?

- It does not make sense to add the error (0.8 inches).
- It does not make sense to take the difference of the error terms (0.2 inches).

It turns out that error values add in **quadrature** (add the squares of the errors and then take the square root, just like the Pythagorean Theorem).

We already have a term that has to do with error in this course – that term is *variance*. So, we'll use our trusty sigma squared (σ^2) notation to represent the square of the error.

Propagation of Error Formula

Suppose $X_1, X_2, X_3, \dots, X_n$ are random variables where X_1 has mean μ_1 and variance σ_1^2 , X_2 has mean μ_2 and variance σ_2^2 , ... X_n has mean μ_n and σ_n^2 .

Suppose $Y = h(X_1, X_2, X_3, \dots, X_n)$. (So, Y is the result of some function h over the random variables.)

Then, the mean and variance of Y can be approximated by the following:

$$E(Y) = \mu_Y \cong h(\mu_1, \mu_2, \dots, \mu_n)$$

// plug in the means of the X_i 's into function h to get the mean for Y

$$V(Y) = \sigma_Y^2 \cong \sum_{i=1}^n \left(\frac{\partial h}{\partial X_i} \right)^2 * \sigma_i^2$$

where the partial derivatives are evaluated at the μ_i values

// calculate partial derivatives of each random variable, evaluate the partial derivative at the
// mean values for that random variable, square the result and multiply by the variance of that
// random variable; then sum all those values to get the total variance of Y

Remember: **Error adds in quadrature.**

Example 1: Total Resistance

Suppose two resistors are connected in parallel and follow the model for total resistance R, as follows:

$$R = \frac{R_1 * R_2}{R_1 + R_2}$$

$$\mu_{R_1} = 20.0 \text{ ohms, standard deviation } \sigma_{R_1} = 0.500 \text{ ohms}$$

$$\mu_{R_2} = 50.0 \text{ ohms, standard deviation } \sigma_{R_2} = 1.000 \text{ ohms}$$

a. What is E(R)?

$$E(R) = \frac{20.0 * 50.0}{20.0 + 50.0} = \frac{1000}{70.0} = 14.3 \text{ ohms}$$

b. What is V(R)?

We need to find and evaluate the partial derivatives for R with respect to R1 and R2. Product rule and chain rule used below. Partial derivative – take derivative with respect to single variable while holding all other variables constant.

$$R = R_1 * R_2 (R_1 + R_2)^{-1}$$

$$\begin{aligned} \frac{\partial R}{\partial R_1} &= R_2 (R_1 + R_2)^{-1} + (-R_1 * R_2 (R_1 + R_2)^{-2}) \\ &= \frac{R_2}{R_1 + R_2} - \frac{R_1 * R_2}{(R_1 + R_2)^2} \\ &= \frac{R_2 (R_1 + R_2) - R_1 * R_2}{(R_1 + R_2)^2} \end{aligned}$$

$$\frac{\partial R}{\partial R_1} = \frac{R_2^2}{(R_1 + R_2)^2}$$

The same procedure will get us the partial with respect to R2:

$$\frac{\partial R}{\partial R_2} = \frac{R_1^2}{(R_1 + R_2)^2}$$

So, now we can calculate the variance of R:

$$\frac{\partial R}{\partial R_1} = \frac{R_2^2}{(R_1 + R_2)^2} = \frac{50.0 * 50.0}{(20.0 + 50.0)(20.0 + 50.0)} = 0.512$$

$$\frac{\partial R}{\partial R_2} = \frac{R_1^2}{(R_1 + R_2)^2} = \frac{20.0 * 20.0}{(20.0 + 50.0)(20.0 + 50.0)} = 0.0816$$

$$\sigma_R^2 = (0.512)(0.512) * (0.5) * (0.5) + (0.0816)(0.0816) * (1.0) * (1.0) = .0722 \text{ ohms squared}$$

Example 2: Area of Rectangle

Suppose we measure the height and width of a rectangle and want to know the area.

H = 4.26 meters +/- 0.05 meters

W = 8.12 meters +/- 0.03 meters

$$A = H * W$$

Again, the error terms can be thought of as uncertainty or as variance. We can employ the propagation of error formula.

- a. What is the expected value (mean of A) for the area?

$$E(A) = 4.26 * 8.12 = 34.6 \text{ sq meters}$$

- b. What is the variance of A?

We need to take the partial derivatives of A with respect to H and with respect to W.

$$\frac{\partial A}{\partial H} = W$$

$$\frac{\partial A}{\partial W} = H$$

When we evaluate these at the means, we get:

$$\sigma_A^2 = (8.12)(8.12) * (0.05)(0.05) + (4.26)(4.26) * (0.03)(0.03) = 0.181 \text{ meters to the fourth}$$

Example 3: Perimeter of Rectangle

Suppose we have two measurements of a rectangle and we want to get the total perimeter, T.

$$T = 2A + 2B$$

A = 6.18 +/- .06 feet

B = 2.09 +/- .04 feet

- a. What is the expected value (mean) of T?

$$E(T) = 2 * 6.18 + 2 * 2.09 = 16.54 \text{ feet}$$

b. What is the variance of T?

We need to take the partials with respect to A and B:

$$\frac{\partial T}{\partial A} = 2$$

$$\frac{\partial T}{\partial B} = 2$$

$$V(T) = (2 * 2) * (.06 * .06) + (2 * 2) * (.04 * .04) = .0208$$

So, the error of the perimeter is .14 feet (take sqrt of V(T)). Does this seem reasonable?

FYI: The propagation of error formula assumes the random variables X_1, X_2, \dots, X_n are independent!! Usually this is the case in engineering. See textbook for how to handle dependent variables (need to calculate covariances) in the case of linear functions.

Special Case: Linear Combination of Independent Random Variables

Note: The propagation of error formula and technique above always works for independent random variables (linear functions, non-linear functions, any model shape). This special case can easily be derived in a similar way to the perimeter example above.

If Y has a random variable that is a linear combination of random variables X_1, X_2, \dots, X_n , then the mean is the linear combination of the means of the random variables and the variance is the sum of the squared coefficients multiplied by the variances of the random variables. Furthermore, if all the random variables are normally distributed, $E(Y)$ is normally distributed with mean $E(Y)$ and variance $V(Y)$.

$$Y = C_0 + C_1X_1 + C_2X_2 + \dots + C_nX_n$$

$$E(Y) = C_0 + C_1\mu_1 + C_2\mu_2 + \dots + C_n\mu_n$$

$$V(Y) = C_1^2\sigma_1^2 + C_2^2\sigma_2^2 + \dots + C_n^2\sigma_n^2$$

Example 1: Practice

Suppose we have the following relationship:

$$P = 8S - 12R$$

$$S = 12.3 \pm 0.11 \text{ inches}$$

$$R = 3.8 \pm 0.05 \text{ inches}$$

What is $E(P)$?

What is $V(P)$?

Example 2: Practice

Suppose we have a pendulum. We find the mean length to be 50 feet and the variance of the length to be 0.1 feet. The pendulum model for the period T is below. π and g are constants.

$$T = 2\pi * g^{0.5} * L^{0.5}$$

- a. What is the expected value (mean) of T ?

$E(T) =$

- b. What is the variance of T ?

Need to calculate the partial derivative with respect to L .

Error Propagation Practice

Directions: Get into your team of 4 people. Your team will work through problems. Work together as a team – do not divide and conquer the problems. Each team member is assigned to one of the following roles:

- Notetaker: prepares the master copy to be submitted
- Manager: ensures that group stays on task, manages time, and ensures everyone gets to speak
- Spokesperson: when called upon, speaks on behalf of the team
- Reflector: analyzes team dynamics and makes suggestions for improving team process – is everyone contributing? Does everyone understand how to get to the solution?

Names: Notetaker: _____

Manager: _____

Spokesperson: _____

Reflector: _____

1. Suppose we have the following relationship for Y and the X_1 and X_2 are normally distributed random variables.

$$Y = 3X_1 + 5X_2$$

The data for the random variables: $E(X_1) = 2$. $V(X_1) = 2$. $E(X_2) = 5$. $V(X_2) = 10$.

a. What is $E(Y)$?

b. What is $V(Y)$?

c. Use the normal distribution to determine the probability that $Y \leq 50$, given the calculations that you just completed for parts a and b.

2. A plastic casing for a magnetic disk is composed of two halves. The thickness of each half is normally distributed with mean 1.5 millimeters and a standard deviation of 0.1 millimeter. The two halves are independent.

a. What is the mean of the total thickness of the two halves?

b. What is the standard deviation of the total thickness of the two halves (remember to convert to variance and then back to standard deviation)?

3. Suppose we find in the lab that the value for Y depends on X as follows:

$$Y = 2X^2$$

Assume X is a random variable with mean 20.0 cm and variance 9.0 cm².

a. What is the mean E(Y)?

b. What is the variance V(Y)?

Checkpoint 1: Stop for class discussion. If your team has reached this checkpoint, you may move on to the next question. Be ready to share your team answers with the class.

4. (if time) The acceleration due to gravity can be measured via distance and time. Assume distance d, here, is a constant value.

$$G = 2d/T^2$$

Suppose E(T) = 5.2 seconds and V(T) = 0.0004 square seconds. Compute the mean and variance of G.

Intro to Statistics

Suppose you want to know the average age of UP students. We could ask a random sample of 100 students for their birth date. Let's say we find that the average age of the set of students is 20.3 years old. From the sample mean, we could *infer* that the average age of UP students is 20.3 years old. This is a *point estimate*. Inferences are made about populations, using the data from a sample.

Statistics involves doing the following things:

- Parameter estimation
 - Examples: What is the mean? What is the variance? What is the proportion? What is the difference between means?
- Hypothesis testing
 - Examples: Is the mean less than 20.3 ounces? Is the average height of men different than the average height of women?

Sample measures give us estimators. For example, the sample mean can give us an estimate of the population mean.

When might the sample mean *not* give us a good estimate of the population mean?

-
-
-

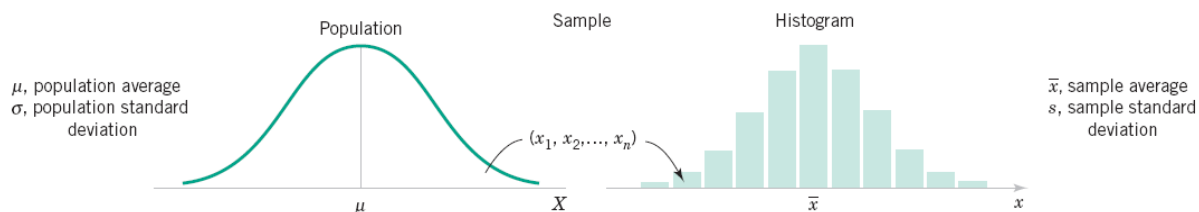


Figure 4-1 Relationship between a population and a sample.

Figures are from textbook

This lecture has a lot of vocabulary that will be used in the remainder of the semester:

Parameter	Alpha: type I error
Statistic	Beta: type II error
Point Estimate	Power
Standard Error (we saw this earlier)	P-value
Null Hypothesis	One-sided
Alternative Hypothesis	Two-sided
Critical Region	

What kinds of statistics would we want to estimate?

- Mean, variance, proportion, difference of means, difference of proportions
- A point estimate of population parameter θ is a single numerical value of a statistic.

Unknown Parameter θ	Statistic $\hat{\Theta}$	Point Estimate $\hat{\theta}$
μ	$\bar{X} = \frac{\sum X_i}{n}$	\bar{x}
σ^2	$S^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1}$	s^2
p	$\hat{P} = \frac{X}{n}$	\hat{p}
$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2 = \frac{\sum X_{1i}}{n_1} - \frac{\sum X_{2i}}{n_2}$	$\bar{x}_1 - \bar{x}_2$
$p_1 - p_2$	$\hat{P}_1 - \hat{P}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$	$\hat{p}_1 - \hat{p}_2$

We want the estimators to be unbiased, which means that the expected value of the statistic is equal to the population parameter. For example, sample mean is an unbiased estimator of μ . However, median is a biased estimator of μ .

The **standard error** of a statistic is the standard deviation of its sampling distribution. If the standard error involves unknown parameters whose values can be estimated, substitution of these estimates into the standard error results in an **estimated standard error**.

Statistical Hypothesis

Statement about parameters of one or more populations.

Think about a comparative experiments. Examples:

- Is the mean weight of bolts manufactured at plant A equal to the mean weight of bolts manufactured at plant B?
- Is the average height of 8-year-old children 50 inches?
- Is the variance of bolts manufactured at plant A equal to 0.05 square grams?
- Is the average weight of 10-year-old boys greater than 75 lbs?

- Do galvanized nails begin to rust later than non-galvanized nails when exposed to the same weather conditions?

We can make inferences from a single population or multiple populations. Which of the questions above refer to single populations?

Example: Heights of 8-year-olds

Suppose you are interesting in studying the height of children. You get the heights of 10 random children at their 8-year-old well child medical visits. You are interested in deciding whether or not the mean height is 50 inches.

Since we are interested in whether the mean is 50 inches, we have a two-sided alternative hypothesis.

H_0 : The mean height is 50 inches.	// null hypothesis
H_1 : The mean height is <u>not</u> equal to 50 inches.	// alternative hypothesis

Remember – hypotheses are about the mean of the population and not the sample!!!. Clearly, we are always definite about the mean of the sample, since we have access to that data.

So, the question is – **when do we decide the sample mean is far enough from 50 inches to warrant a rejection of H_0 .** This is where we need to define a critical region.

Think about it

Suppose the sample mean height of 10 children is 48 inches. Do you have enough evidence to reject H_0 ?

If you need more information, what information would you need?

We can make mistakes in statistical inference.

- Case 1: The true population mean is 50 inches and the sample mean is 50.05 inches. We fail to reject the null hypothesis. **[No error]**
- Case 2: The true population mean is 50 inches and the sample has really tall children. The sample mean is 52 inches. We reject the null hypothesis, but the null hypothesis is really true about the population. **[Type I error]**
- Case 3: The true population mean is different than 50 inches and the sample mean is 50.05 inches. We fail to reject the null hypothesis, but the null hypothesis is false. **[Type II error]**
- Case 4: The true population mean is different than 50 inches and the sample mean is 52 inches. We reject the null hypothesis. **[No error]**

Note that we NEVER get access to the true population mean, unless the sample is the entire population (why would we need to do inference then?).

Table 4-1 Decisions in Hypothesis Testing

Decision	H_0 Is True	H_0 Is False
Fail to reject H_0	No error	Type II error
Reject H_0	Type I error	No error

α : P(Type I error) = P(reject H_0 when H_0 is true)

We sometimes call alpha the *significance level* or type I error or the *size of the test*.

What is alpha? Well, it is the area under the normal curve at the two tails for alternative hypothesis in the form of NOT equal.

- If the critical region is known, we can compute alpha. (using CLT and standard normal)
- If alpha is known, we can compute the critical region values. (using CLT and standard normal)

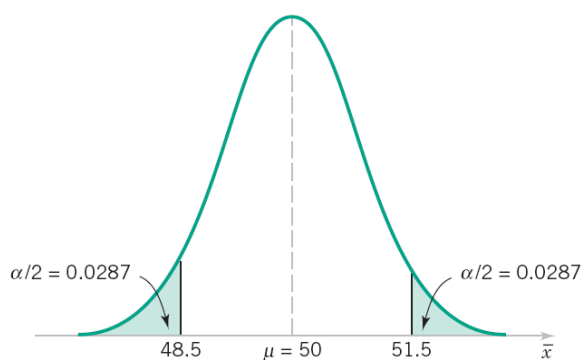


Figure 4-4 The critical region for $H_0: \mu = 50$ versus $H_1: \mu \neq 50$ and $n = 10$.

Suppose the critical region is < 48.5 and > 51.5 for rejecting the null hypothesis. We can compute alpha:

$$\alpha = P(\bar{X} < 48.5) + P(\bar{X} > 51.5)$$

Using the central limit theorem, assume the sample standard deviation of heights in children is 2.5 inches. Then, converting to Z we get:

$$Z_1 = \frac{48.5 - 50}{2.5 \div \sqrt{10}} = -1.90$$

$$Z_2 = \frac{41.5 - 50}{2.5 \div \sqrt{10}} = 1.90$$

Using the z-table, we see that $P(Z < -1.90) = 0.0287$ and $P(Z > 1.90) = 0.0287$. So, $\alpha = .0574$.

Suppose, instead, we start with a particular alpha (risk level for making a type I error) we want to use for the two-sided test. We can use that to determine the critical region values. If the sample error falls outside the critical region, then we reject the null hypothesis.

Assume we set alpha to be 0.04. Since this is a two-sided test, each shaded area under the curve will be 0.02 of the overall distribution. Looking in the z-table, we find Z to be -2.88 and 2.88 for the standard normal for this alpha.

We then convert this back to the original normal distribution and solve for the critical region:

$$2.88 = \frac{X_2 - 50}{2.5 \div \sqrt{10}}$$

$$X_2 = 52.28$$

$$-2.88 = \frac{X_1 - 50}{2.5 \div \sqrt{10}}$$

$$X_1 = 47.72$$

Do these critical region values make sense, given alpha? Now, if the sample height is < 47.72 or > 52.28 , we would reject the null hypothesis.

Think about it

Suppose we gather 50 bolts from a box for testing. The population is all bolts produced from the assembly line. We want to set the significance level alpha for the test at 0.05 that the bolt length is not equal to 2 inches. Assume C1 and C2 are calculated as the critical region endpoints where $C_1 = 1.95$ inches and $C_2 = 2.05$ inches. Now assume that we want the alpha level to be 0.01 instead. Will C1 increase or decrease? Will C2 increase or decrease?

$\beta = P(\text{Type II Error}) = P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false})$

What is beta?

It's the area under the curve where the alternative hypothesis distribution overlaps with the null hypothesis distribution. For the beta calculations, we need a specific value for the alternative hypothesis. Suppose the true population mean of heights of 8-year-olds is really 52 inches and we use the same critical region as before. So, our beta is:

$\beta = P(48.5 \leq \bar{X} \leq 51.5 \text{ when the population mean is } 52) \quad // \text{ use CLT and standard normal}$
 $= 0.2643$

What does this look like? The shaded area shows the probability of getting a sample mean within 48.5 and 51.5 inches when the population mean is really 52 inches. In this case, we use

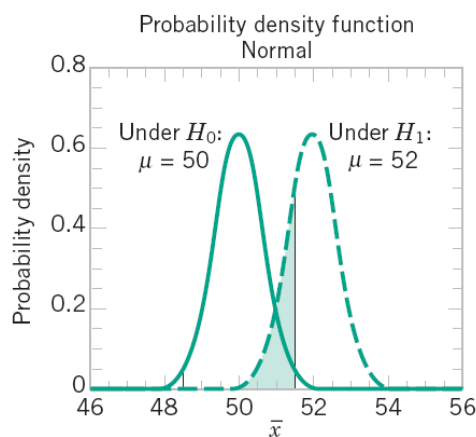


Figure 4-5 The probability of type II error when $\mu = 52$ and $n = 10$.

Note: the area under the curve gets bigger as the alternative hypothesis gets close the null hypothesis. See below when the alternative has mean 50.5. A lot of the curve is now between the critical region values.

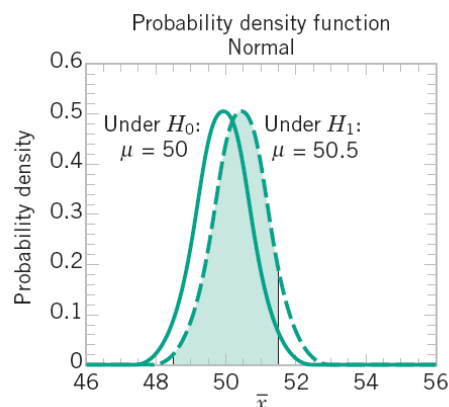


Figure 4-7 The probability of type II error when $\mu = 52$ and $n = 16$.

Relationship among alpha, beta, and sample size:

1. Moving critical region values outward from center reduces alpha.
2. If the sample size is increased and the critical region stays the same, alpha and beta decrease.
[Think about the CLT – more samples means a smaller variance and a skinnier, taller normal distribution.]

Power of a Test

Power = P(rejecting null hypothesis when alternative hypothesis is true) // correctly rejecting a false null hypothesis – this what we want to do, in general (reject null hypothesis when alternative is true)

$$Power = 1 - \beta$$

Want a high power. If the power is .7357, then power gives us a meaning of sensitivity. The test sensitivity for detecting the difference between a mean height of 50 and 52 inches is .7357. If the true mean is really 52 inches, the test will reject the null hypothesis 73.57% of the time.

Justice System

Here's a good time to make an analogy to the US justice system. Assume you are a juror for a trial where person A is accused of stealing money from a store.

As a juror, what is your stance prior to the start of the trial? // innocent until proven guilty
Your null hypothesis is that person A is not guilty

The burden of proof is on the prosecution. There must be enough evidence to have a guilty verdict.
The alternative hypothesis is that person A is guilty

We have to be very confident that there is enough evidence to have a guilty verdict, must like we must have enough evidence that the sample mean is far enough from the hypothesis of 50 to have a verdict that the population mean is not equal to 50.

P-values

You have probably heard of the term p-value. "The test is significant for a p-value of 0.05.". Well, this p-value is like alpha (the significance level). It is the area under the curve at the tail(s) of the distribution. It is the smallest level of significance that would lead to rejection of the null hypothesis.

The p-value is **NOT** the probability that null hypothesis is false. A hypothesis is either TRUE or FALSE (no gray area). The p-value is the risk of wrongly rejecting the null hypothesis.

Example: Propellant burning rate

Suppose the p-value is 0.0226. What does the sample mean need to be to reject the null hypothesis that the population mean is 50? It would need to be greater than 51.8 or less than 48.2. See figure below.

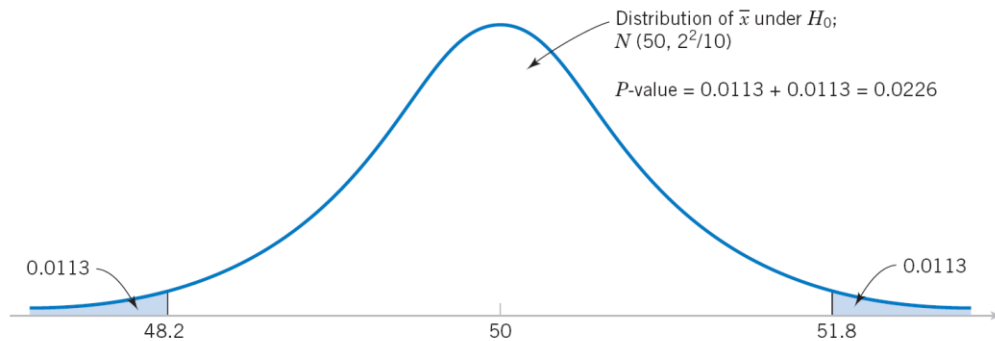


Figure 4-8 Calculating the P -value for the propellant burning rate problem.

One-sided vs Two-sided tests

Note: in the above lecture material about statistics, the entire set of examples assume a two-sided test (alt hypothesis states the mean is NOT equal to some value).

One-sided tests are questions where the inference (alternative hypothesis) has to do with the sample estimate being less than a value or greater than a value. Examples:

- The average weight of bolts made at plant A is *greater than* 8.2 ounces.
- The mean height of babies born in the USA is *less than* 20.26 inches.

$$\begin{array}{ccc} H_0: \mu = \mu_0 & \text{or} & H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 & & H_1: \mu < \mu_0 \end{array}$$

Two-sided tests are questions where the inference has to do with the sample statistic being unequal to a specific value. Examples:

- The average height of male adults in the USA is not 69 inches.
- The average time to assemble a part is not 5.4 hours.

$$\begin{array}{c} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{array}$$

The sided-ness comes from the area under the curve. If we just care about a sample estimate being less than a certain number, our shaded region is just the left tail.

If we care about a sample estimate being greater than a certain number, our shaded region is just the right tail.

If we care about a sample estimate being unequal to a certain number (either smaller or bigger), our shaded region is both tails.

Practice:

For each question below, would you use a one-sided or two-sided test?

1. Is the average number of car accidents on I-5 between the Fremont and Marquam bridges greater than 100 per year?
2. Is the mean weight of 50-year men in the United States 185 pounds?
3. Is the percentage of defective chips produced from Plant A in 2017 less than 2%?
4. Is the average time to failure of an LED less than 1000 days?

Steps in Hypothesis Testing

1. **Parameter of interest:** From the problem context, identify the parameter of interest.
2. **Null hypothesis, H_0 :** State the null hypothesis, H_0 .
3. **Alternative hypothesis, H_1 :** Specify an appropriate alternative hypothesis, H_1 .
4. **Test statistic:** State an appropriate test statistic.
5. **Reject H_0 if:** Define the criteria that will lead to rejection of H_0 .
6. **Computations:** Compute any necessary sample quantities, substitute these into the equation for the test statistic, and compute that value.
7. **Conclusions:** Decide whether or not H_0 should be rejected and report that in the problem context. This could involve computing a P -value or comparing the test statistic to a set of critical values.

Steps 1 – 4 should be done prior to examining the sample data.

Inference on mean of normal population with known variance (z-test)

Case 1: Hypotheses (Two-sided)

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

Case 2: Hypotheses (One-Sided)

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$

OR

$$H_0: \mu = \mu_0$$

$$H_1: \mu < \mu_0$$

When to use z-test:

- Have a random sample of size n : X_1, X_2, \dots, X_n .
- The population is normally distributed or the central limit applies.
- The variance of the population is known.

Test statistic (standard normal with standard deviation accd to sample size):

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad (4-13)$$

Picture of distribution (shaded region shows areas where we reject the null hypothesis):

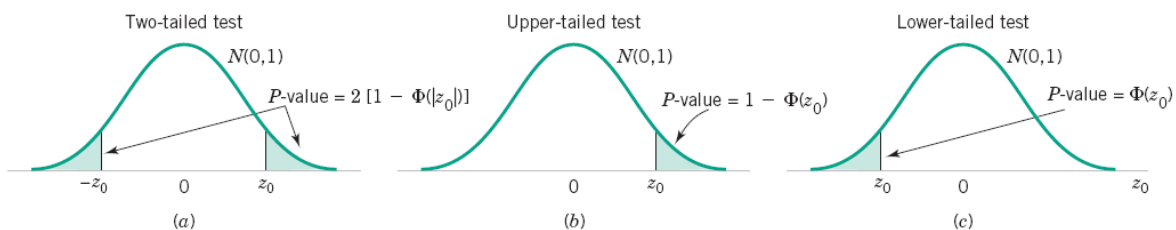


Figure 4-9 The P -value for a z -test. (a) The two-sided alternative $H_1: \mu \neq \mu_0$. (b) The one-sided alternative $H_1: \mu > \mu_0$. (c) The one-sided alternative $H_1: \mu < \mu_0$.

Two-sided test: Reject if sample mean falls into either tail.

One-sided test: Reject if sample mean falls into tail on the side of the alternative hypothesis.

Use our trusty z-table numbers to determine the critical region bounds for z_0 . The capital Phi letter represents the cumulative distribution function for the standard normal distribution.

Summary:

Testing Hypotheses on the Mean, Variance Known (z-Test)

Null hypothesis: $H_0: \mu = \mu_0$

Test statistic: $Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

Alternative Hypotheses	P-Value	Rejection Criterion for Fixed-Level Tests
$H_1: \mu \neq \mu_0$	Probability above $ z_0 $ and probability below $- z_0 $, $P = 2[1 - \Phi(z_0)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1: \mu > \mu_0$	Probability above z_0 , $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1: \mu < \mu_0$	Probability below z_0 , $P = \Phi(z_0)$	$z_0 < -z_\alpha$

The P -values and critical regions for these situations are shown in Figs. 4-9 and 4-10.

Beta (probability of type II error) for two-sided test:

Probability of a Type II Error for the Two-Sided Alternative Hypothesis on the Mean, Variance Known

$$\beta = \Phi\left(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) - \Phi\left(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) \quad (4-24)$$

This means that z falls in between $-z_{\alpha/2}$ and $z_{\alpha/2}$ when alternative hypothesis is true. Here, $\delta + \mu_0$ is the true value of the mean.

Beta for one-sided tests:

$$\beta = \Phi\left(z_\alpha - \frac{\delta\sqrt{n}}{\sigma}\right)$$

How to calculate minimum sample size for two-sided test (with known alpha-level, known beta-level, variance of population, and hypothesized mean μ).

Sample Size for Two-Sided Alternative Hypothesis on the Mean, Variance Known

For the two-sided alternative hypothesis on the mean with variance known and significance level α , the sample size required to detect a difference between the true and hypothesized mean of δ with power at least $1 - \beta$ is

$$n \simeq \frac{(z_{\alpha/2} + z_{\beta})^2 \sigma^2}{\delta^2} \quad (4-26)$$

where

$$\delta = \mu - \mu_0$$

If n is not an integer, the convention is to always round the sample size up to the next integer.

How to calculate minimum sample size for one-sided test (note: z -alpha instead of z -alpha/2 is used):

Sample Size for One-Sided Alternative Hypothesis on the Mean, Variance Known

For the one-sided alternative hypothesis on the mean with variance known and significance level α , the sample size required to detect a difference between the true and hypothesized mean of δ with power at least $1 - \beta$ is

$$n = \frac{(z_{\alpha} + z_{\beta})^2 \sigma^2}{\delta^2} \quad (4-27)$$

where

$$\delta = \mu - \mu_0$$

If n is not an integer, the convention is to round the sample size up to the next integer.

Procedure for hypothesis testing:

1. Specify hypothesis (two-sided, one-sided: upper-tailed or lower-tailed).
2. Specific test statistic (in this case, it is z_0).
3. Specify criteria for rejection (alpha or the P-value).

Warning:

- When sample size is really large, any small departure from the hypothesized mean will probably be detected, even when the difference is of little or no practical significance. [Think central limit theorem...distribution gets really thin when n is large.]

Practice

1. The yield of a chemical process is being studied. From previous experience, the standard deviation of yield is known to be 3%. The past 5 days of plant operation have resulted in the following yields: 91.60%, 88.75%, 90.80%, 89.95%, and 91.30%. Use $\alpha = 0.05$ for the test level.

a. Is there evidence that the mean is not 90%? (Use p-value approach)

Step 1: What is the parameter of interest? _____

Step 2: What is H_0 ? $\mu = 90$

Step 3: What is H_1 ? $\mu \neq 90$

Step 4: Use the z-test statistic.

Step 5: Reject if z_0 is less than $-z_{\alpha/2}$ or greater than $z_{\alpha/2}$. Here, $\alpha/2$ is 0.025 since this is a 2-sided test.

Calculate $z_{0.025}$ (look in table): _____

Step 6: Calculate sample mean: _____

Calculate z-statistic: _____

Step 7: Calculate p-value: $2(1 - \Phi(z\text{-value})) =$ _____

Is this p-value < 0.05 ? _____

What is your conclusion? _____

b. What sample size is necessary to detect a true mean yield of 85% with probability 0.95?

c. What is the type II error probability if the true mean yield is 92%?

Hint: this is beta for 2-sided test

Another method: Confidence Intervals

Assume we take random samples with replacement from our population of interest. Below, you see 20 intervals that showcase the ranges of the sample values. The dot represents the sample mean for each sample. See how the dots may be higher or lower than the population mean?

We have a new idea called a confidence interval. A 95% confidence interval of the mean would be the interval such that 95% of random sample means are contained. Below, a 95% confidence interval would need to contain the dots of 19 of the 20 samples. Which is the most extreme? Probably sample 8. So, we can find the interval around μ such that the other 19 dots are contained and this would be 95% confidence interval of the mean.

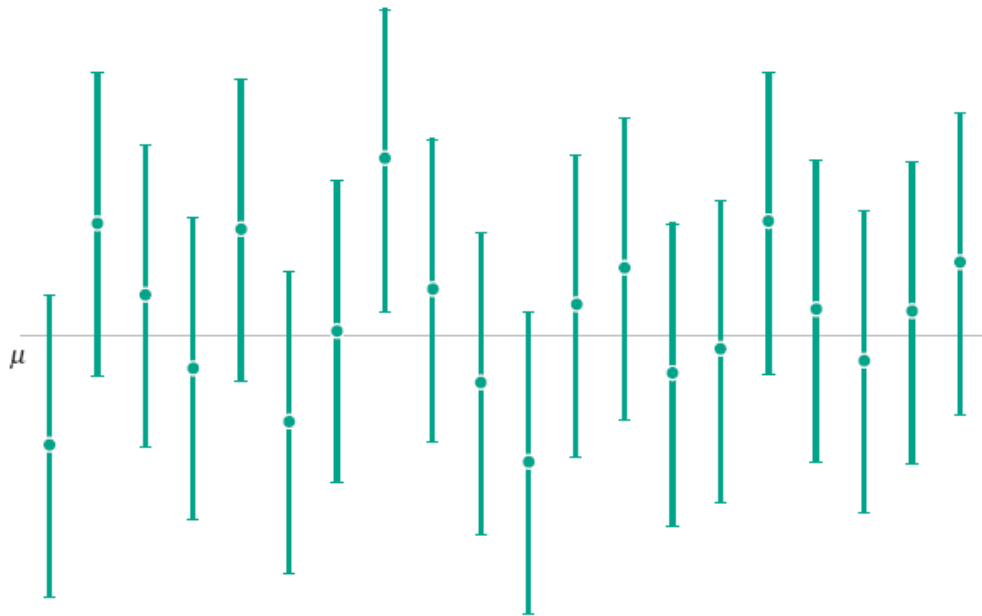


Figure 4-12 Repeated construction of a confidence interval for μ .

Well, this is actually a different view of the same phenomenon we saw with the z-test. We would expect the sample mean to be within the critical region bounds with probability 95% when alpha is 5%. So, we have the same type of calculation to generate the critical region (the confidence interval):

$$P\{-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\} = 1 - \alpha$$

Here $1 - \alpha$ is the confidence interval level. So, if α is 0.05, then the confidence interval level is 0.95. 95% of the area of the under falls between these critical values for z.

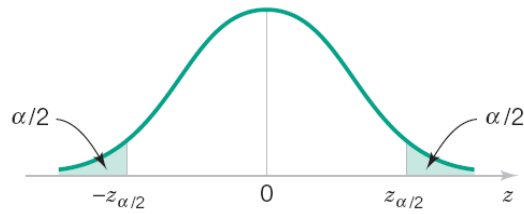


Figure 4-13 The distribution of Z .

Confidence Interval on the Mean, Variance Known

If \bar{x} is the sample mean of a random sample of size n from a population with known variance σ^2 , a $100(1 - \alpha)\%$ **confidence interval on μ** is given by

$$\bar{x} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \quad (4-35)$$

where $z_{\alpha/2}$ is the upper $100\alpha/2$ percentage point and $-z_{\alpha/2}$ is the lower $100\alpha/2$ percentage point of the standard normal distribution in Appendix A Table I.

What about sample size? If we want a $100(1-\alpha)\%$ confidence level that the error (distance between \bar{x} and μ) will not exceed a certain amount E , then the sample size is:

$$n = \left(\frac{z_{\alpha/2}\sigma}{E} \right)^2$$

Let's think about this:

- If α is fixed, as E decreases, n _____.
- If E is fixed, if confidence level increases, n _____.

If we are just looking at a one-sided confidence interval of the mean, we have (note that z just has α as subscript, since this is one-sided):

One-Sided Confidence Bounds on the Mean, Variance Known

The $100(1 - \alpha)\%$ **upper-confidence bound** for μ is

$$\mu \leq u = \bar{x} + z_{\alpha}\sigma/\sqrt{n} \quad (4-37)$$

and the $100(1 - \alpha)\%$ **lower-confidence bound** for μ is

$$\bar{x} - z_{\alpha}\sigma/\sqrt{n} = l \leq \mu \quad (4-38)$$

Examples

1. Suppose a random sample has been taken from a population and two confidence intervals are produced from the same data:

(38.02, 69.98)

(39.95, 68.05)

a. What is the value of the sample mean?

b. One of these intervals is a 90% CI and the other is 95% CI. Identify which one is which.

2. Medical researchers have developed a new artificial heart. The battery pack needs to be recharged about every 4 hours. A random sample of 50 battery packs is selected for a life test. The average life of the batteries is 4.05 hours. Assume that battery life is normally distributed with standard deviation of 0.20 hours.

a. Is there evidence to support the claim that mean battery life exceeds 4 hours? Use $\alpha = 0.05$ for the significance level.

Step 1: Identify parameter: battery life

Step 2: Null hypothesis: $\mu = 4$

Step 3: Alternative hypothesis: $\mu > 4$ (one-sided)

Step 4: Test statistic is z-test. (Normally distributed data, variance known)

Step 5: Reject null hypothesis if z-value is greater than 1.65 (look at table for z-value corresponding to .95)

Step 6: Calculate test statistic:

$$z_0 = \frac{4.05 - 4}{0.2 * \sqrt{50}} = 1.77$$

Step 7: Check statistic. Because 1.77 is bigger than 1.65, we reject the null hypothesis and conclude that battery life exceeds 4 hours at significance level $\alpha = 0.05$.

b. What is the p-value for the result in part (a).

We look in the z-table for 1.77. We see .961636, so the area to the right of this is about 0.04. The p-value is 0.04.

c. Assume the true mean battery life is 4.5 hours. What is the power of the test?

We need to find β . Since we are looking at just a one-sided test, we need to find the difference in the standard z values for the 0.95 level and the z value of the true mean away from the hypothesized mean in standard normal.

$$\beta = \Phi\left(z_{0.05} - \frac{(4.5 - 4)\sqrt{50}}{0.2}\right) = \Phi(1.645 - 17.68) = \Phi(-16.035) = 0$$

Power = $1 - \beta$, so $1 - 0 = 1$. This power is high and makes sense since the true mean is so far from the hypothesized mean with respect to the standard deviation of 0.2.

d. What sample size would be required to detect a true mean battery life of 4.5 hours if we want the power of the test to be at least 0.9? (Remember to use 1-sided alpha).

$$n = \frac{(z_{\alpha} + z_{\beta})^2 \sigma^2}{\delta^2} = \frac{(z_{0.05} + z_{0.1})^2 \sigma^2}{(4.5 - 4)^2} = \frac{(1.645 + 1.29)^2 (0.2)^2}{(0.5)^2} = 1.38,$$

$$n \cong 2$$

e. Now, instead construct a 1-sided 95% confidence interval on the mean life. We need to find the lower bound of the confidence interval (since we are looking to the right of the mean in the distribution).

$$\bar{x} - z_{0.05} \left(\frac{\sigma}{\sqrt{n}} \right) \leq \mu$$

$$4.05 - 1.65 \left(\frac{0.2}{\sqrt{50}} \right) \leq \mu$$

$$4.003 \leq \mu$$

Because the lower limit of the 95% CI is greater than 4, we conclude the average life is greater than 4 hours at $\alpha = 0.05$. The confidence interval is (4.003, infinity)

Z-test practice

Directions: Get into your team of 4 people. Your team will work through problems. Work together as a team – do not divide and conquer the problems. Each team member is assigned to one of the following roles:

- Notetaker: prepares the master copy to be submitted
- Manager: ensures that group stays on task, manages time, and ensures everyone gets to speak
- Spokesperson: when called upon, speaks on behalf of the team
- Reflector: analyzes team dynamics and makes suggestions for improving team process – is everyone contributing? Does everyone understand how to get to the solution?

Names: Notetaker: _____
Spokesperson: _____

Manager: _____
Reflector: _____

1. The yield of a chemical process is being studied. From previous experience, the standard deviation of yield is known to be 3%. The past 5 days of plant operation have resulted in the following yields: 91.60%, 88.75%, 90.80%, 89.95%, and 91.30%. Use $\alpha = 0.05$.

a. Find the 95% two-sided CI on the true mean yield.

b. Use the answer in part (a) to test the original hypothesis. Is hypothesized mean within this confidence interval? If so, we fail to reject the null hypothesis.

2. A civil engineer is analyzing the compressive strength of concrete. Compressive strength is approximately normally distributed with variance 1000.0 psi squared. A random sample of 12 specimens has compressive strength of $\bar{x} = 3255.42$ psi.

a. Test the hypothesis that mean compressive strength is 3500.0 psi. Used a fixed-level test with $\alpha = 0.01$.

Follow the steps: (Hint – this is 2-sided)

b. Construct a 2-sided 95% confidence interval on mean compressive strength.

c. Construct a 2-sided 99% confidence interval on mean compressive strength.

d. Do these confidence intervals make sense?

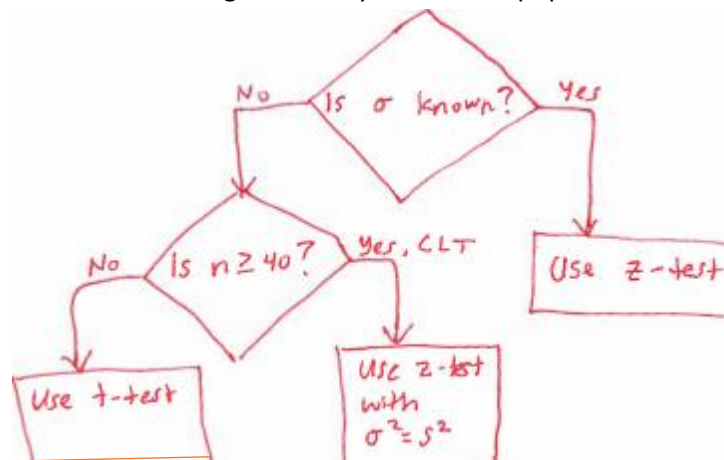
3. Do you have any questions about using the z-test, calculating beta, calculating sample size, or calculating the confidence intervals?

Checkpoint 1: Stop for class discussion. If your team has reached this checkpoint, you may try to formulate your own problem that can be solved via the z-test. Be ready to share your team answers with the class.

Inference on mean of normal population with unknown variance and small sample size (t-test)

See the flowchart below for when to use the z-test and when to use the t-test to test hypotheses about the mean of a population.

Decision Tree: Single normally-distributed population for $H_0: \mu = \mu_0$



We use the t-test when $n < 40$ and when the variance of the population is unknown. (Note: some other references will say to use the t-test when $n < 30$, but we'll stick to 40 to be consistent with your textbook.)

We use the t-distribution for the small sample, variance unknown situation:

Let X_1, X_2, \dots, X_n be a random sample for a normal distribution with unknown mean μ and unknown variance σ^2 . The quantity

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a t distribution with $n - 1$ degrees of freedom.

Notice that this is the same set-up as the z-test except you use S (sample standard deviation) instead of σ .

What does the t-distribution look like?

It depends on k = degrees of freedom. Higher the k – taller it gets and as k goes to infinity, it approaches the normal distribution. In general, the tails on a t -distribution are further above the x -axis than with the standard normal distribution.

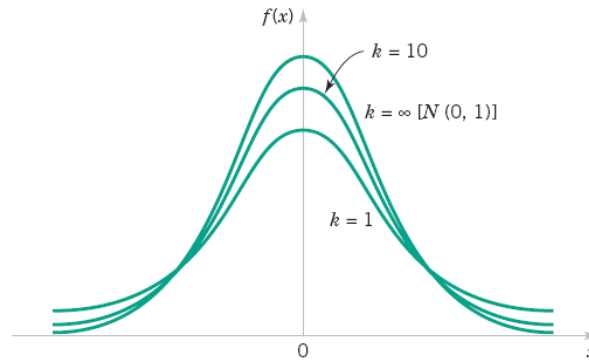


Figure 4-15 Probability density functions of several t distributions.

The statistical procedure is the SAME as with the z -test:

1. Determine parameter of interest
2. Determine null hypothesis (mean = X)
3. Determine alternative hypothesis (mean does not equal X OR mean $> X$ OR mean $< X$)
4. Use t -test statistic
5. Determine critical region or alpha for the cut-off.
6. Compute t -statistic
7. If t -stat is in critical region, reject null hypothesis; otherwise, fail to reject null hypothesis.

Since the t -distribution is no fun to calculate (it is based on the gamma distribution and can be found in the textbook), we use a table for look-up, just as we do with z -scores. Once you have the t -score, use Table II in Appendix A for the lookup.

A note about the table:

- Alpha is set at various values as can be seen on the top row.
- v is the degrees of freedom (or one less than the number of items in sample) and goes down the column.
- The numbers in the table represent the t -score to the right of 0 that corresponds to alpha and v .

Let's practice with the table:

1. Assume the test is 1-sided and alpha is 0.05. The number of samples in the experiment is 10. What t-score corresponds to these values?
2. Assume the test is 2-sided and alpha is 0.05. The number of samples in the experiment is 16. What t-score corresponds to these values?

Testing Hypotheses on the Mean of a Normal Distribution, Variance Unknown		
Null hypothesis:	$H_0: \mu = \mu_0$	
Test statistic:	$T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$	
Alternative Hypotheses	P-Value	Rejection Criterion for Fixed-Level Tests
$H_1: \mu \neq \mu_0$	Sum of the probability above $ t_0 $ and the probability below $- t_0 $,	$t_0 > t_{\alpha/2, n-1}$ or $t_0 < -t_{\alpha/2, n-1}$
$H_1: \mu > \mu_0$	Probability above t_0	or $P = 2P(T_{n-1} > t_0)$ $t_0 > t_{\alpha, n-1}$
$H_1: \mu < \mu_0$	Probability below t_0	$t_0 < -t_{\alpha, n-1}$
The locations of the critical regions for these situations are shown in Fig. 4-19a, b, and c, respectively.		

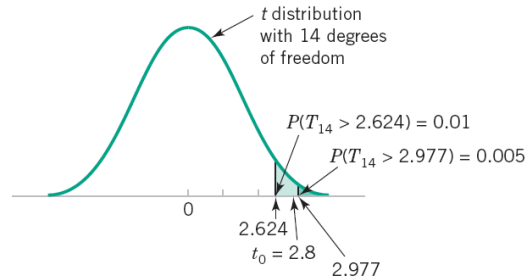
Note: These is the same process as the z-test, but with the t-score instead.

Example: Calculate critical region value

Assume 15 ball bearings are taken from an assembly line and the variance of the population of all ball bearings is unknown. The hypothesized mean of the population is 10.0 ounces and the sample standard deviation is 0.02 ounces. You want to test if the mean is greater than 10.0 ounces. Suppose the alpha level is 0.01. What is the value that which the sample mean needs to exceed to reject the null hypothesis?

We need to find the t-score that corresponds to alpha = 0.01 and v = 14. In the table, we see that this value is 2.624.

Figure 4-18 P -value for $t_0 = 2.8$ and an upper-tailed test is shown to be between 0.005 and 0.01.



You can see in the figure that the lighter shaded area shows the critical region to reject the null hypothesis.

Now, we need to convert back to the weight of the ball bearings using the t-test.

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

$$2.624 = \frac{\bar{X} - 10.0}{.02 / \sqrt{15}}$$

$$2.624 = \frac{\bar{X} - 10.0}{.02 / \sqrt{15}}$$

$$\bar{X} = 10.014$$

Now suppose the sample average is 10.03 ounces. If the test level $\alpha = 0.01$, do you have enough evidence to reject the null hypothesis?

Beta: Type 2 Error for T-test

This is for a two-sided test:

$$\begin{aligned} \beta &= P\{-t_{\alpha/2, n-1} \leq T_0 \leq t_{\alpha/2, n-1} \text{ when } \delta \neq 0\} \\ &= P\{-t_{\alpha/2, n-1} \leq T'_0 \leq t_{\alpha/2, n-1}\} \end{aligned}$$

Again, the t-distribution is approximated with a table where integration is done numerically.

We use charts to get the beta value for one-sided and two-sided tests. Before we can look at the charts, we need to do a simple calculation for d:

$$d = \frac{|\mu - \mu_0|}{\sigma} = \frac{|\delta|}{\sigma}$$

Note that d depends on σ , which we don't know for a t-test. We can estimate σ with the sample standard deviation or we can set d directly. When d is 1, we are detecting a small difference to the mean. If we are detecting large differences to the mean, we can select d = 2.

Let's practice using these charts.

1. Assume alpha = 0.05 and it is a two-sided test. Assume we have calculated d to be 1.5. Assume n = 7.

Estimate the value of beta: _____

2. Assume alpha = 0.05 and it is a one-sided test. Assume we have calculated d to be 0.81 and n = 15.

Estimate the value of beta: _____

What is the power of the test? _____

Confidence Intervals / Bounds for t-test

Just like with the z-test, we can determine confidence intervals for the mean for a t-test.

Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

If \bar{x} and s are the mean and standard deviation of a random sample from a normal distribution with unknown variance σ^2 , a $100(1 - \alpha)\%$ CI on μ is given by

$$\bar{x} - t_{\alpha/2, n-1}s/\sqrt{n} \leq \mu \leq \bar{x} + t_{\alpha/2, n-1}s/\sqrt{n} \quad (4-50)$$

where $t_{\alpha/2, n-1}$ is the upper $100\alpha/2$ percentage point of the t distribution with $n - 1$ degrees of freedom.

If we are calculating a *one-sided confidence bound*, just use the equation 4-50 and use t with α instead of $\alpha/2$. If doing a lower bound, use the left side of μ for the calculation and the right-side becomes infinity. If doing an upper bound, use the right side of μ for the calculation and the left-side becomes negative infinity.

Example Using t-test:

A researcher for a tire manufacturer is studying tire life for a new rubber compound. She has built 10 of these tires and road-tested them. The sample mean life is 61,492 km and the sample standard deviation is 3035.0 km.

a. The researcher would like to demonstrate that the mean life of this tire with new rubber compound exceeds 60,000 km. Formulate the appropriate hypotheses and draw conclusions using the p-value approach.

Step 1: What is the parameter of interest? _____

Step 2: What is H_0 ? _____

Step 3: What is H_1 ? _____

Step 4: What test-statistic should we use? _____

Step 5: What is the critical region boundary? (aka what is alpha? What is the t-value corresponding to this alpha and degrees of freedom) _____

Step 6: Calculate the statistic with the data from the experiment:

Step 7: Conclusion: We reject / fail to reject (*circle answer*) the null hypothesis.

b. Suppose the true mean tire life is at least 61,000 km and the researcher would like to detect this difference with probability at least .90. You may use the sample standard deviation as an estimate of the population standard deviation. Was the sample size of $n = 10$ adequate for the researcher?

1. We need to calculate d:

$$\frac{\delta}{\sigma} = \frac{\mu - \mu_0}{\sigma} = \frac{61000 - 60000}{3035} = 0.3295$$

2. Now, use the OC curve to get the value of beta (one-sided): _____

3. Now, calculate power: _____

4. Does the power exceed .90? If not, then the sample size was not large enough.

5. What sample size should we use (estimate it with the curves) for a power of .90? _____

c. Find a 95% one-sided lower confidence bound on the mean tire life.

d. Use the bound found in part c to test the hypothesis that the mean tire life is $> 60,000$ km.

t-test practice

Directions: Get into your team of 4 people. Your team will work through problems. Work together as a team – do not divide and conquer the problems. Each team member is assigned to one of the following roles:

- Notetaker: prepares the master copy to be submitted
- Manager: ensures that group stays on task, manages time, and ensures everyone gets to speak
- Spokesperson: when called upon, speaks on behalf of the team
- Reflector: analyzes team dynamics and makes suggestions for improving team process – is everyone contributing? Does everyone understand how to get to the solution?

Names: Notetaker: _____

Manager: _____

Spokesperson: _____

Reflector: _____

1. You are working on a new biomedical device and studying the life in hours. The average life is normally distributed and a random sample of 15 devices is selected with an average life of 5625.1 hours. The sample standard deviation is 226.1 hours.

a. Test the hypothesis that the true mean life is *greater than 5500 hours* using the P-value approach.
[Show all 7 steps]

Step 1: (parameter) _____

Step 2: (null hypothesis) _____

Step 3: (alternative hypothesis) _____

Step 4: (test statistic) _____

Step 5: (select alpha and identify critical region boundary)

Step 6: (calculate statistic)

Step 7: (make conclusion)

b. Construct a 95% lower confidence interval on the mean.

c. Compute the power of the test assuming the true mean is 5700.0 hours. You can use sample standard deviation to estimate the population standard deviation. $\alpha = 0.05$.

Checkpoint 1: Stop for class discussion. If your team has reached this checkpoint, you may try to formulate your own problem that can be solved via the t-test. Be ready to share your team answers with the class.

(if time) A physician claims that joggers' maximal volume oxygen uptake is greater than the average of all adults. A sample of 15 joggers has a mean of 40.6 milliliters / kg and a standard deviation of 6.0 ml / kg. If the average of all adults is 36.7 ml / kg, is there enough evidence to support the physician's claim at $\alpha = 0.05$?

Inference on variance of normal population (chi-square-test)

Situation:

We want to infer the *variance* of a *normal* population (not the mean here... but the variance)

Same procedure as before, except now we have a different test statistic.

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_1: \sigma^2 \neq \sigma_0^2$$

OR

$$H_1: \sigma^2 > \sigma_0^2$$

OR

$$H_1: \sigma^2 < \sigma_0^2$$

We use a chi-square statistic (note: it is squared since variance is a square)

$$\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

This statistic has a chi-square distribution with **n-1** degrees of freedom. The probability density function uses the gamma distribution and can be found in the textbook. The mean of this distribution is $k = (n-1)$ = degrees of freedom. The variance of this distribution is $2k$.

First, would this distribution be defined over negative values?

Here's what the distribution looks like:

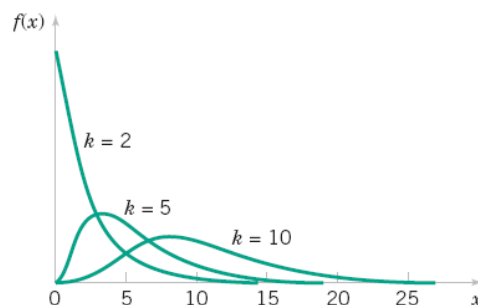


Figure 4-21 Probability density functions of several χ^2 distributions.

$$\mu = k \text{ and } \sigma^2 = 2k$$

We have the same procedure as before:

Testing Hypotheses on the Variance of a Normal Distribution

Null hypothesis: $H_0: \sigma^2 = \sigma_0^2$

Test statistic: $\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$

Alternative Hypotheses

$$H_1: \sigma^2 \neq \sigma_0^2$$

$$H_1: \sigma^2 > \sigma_0^2$$

$$H_1: \sigma^2 < \sigma_0^2$$

Rejection Criterion

$$\chi_0^2 > \chi_{\alpha/2, n-1}^2 \text{ or } \chi_0^2 < \chi_{1-\alpha/2, n-1}^2$$

$$\chi_0^2 > \chi_{\alpha, n-1}^2$$

$$\chi_0^2 < \chi_{1-\alpha, n-1}^2$$

The locations of the critical region are shown in Fig. 4-23.

And the same idea of the critical regions to reject the null hypothesis:

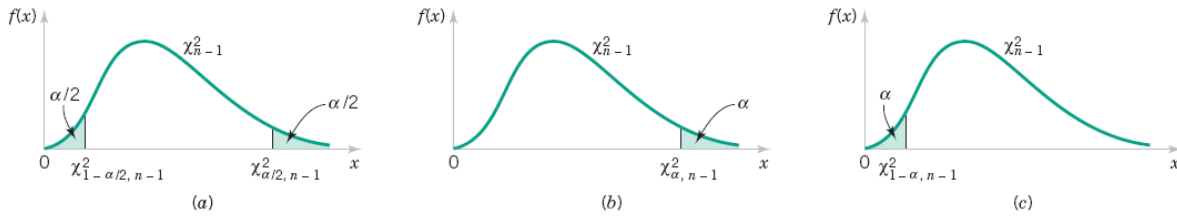


Figure 4-23 Distribution of the test statistic for $H_0: \sigma^2 = \sigma_0^2$ with critical region values for (a) $H_1: \sigma^2 \neq \sigma_0^2$, (b) $H_1: \sigma^2 > \sigma_0^2$, and (c) $H_1: \sigma^2 < \sigma_0^2$.

Note: Figure (a) is for a two-sided hypothesis and Figures (b) and (c) are for the one-sided hypotheses. The confidence interval is also the same procedure:

Confidence Interval on the Variance of a Normal Distribution

If s^2 is the sample variance from a random sample of n observations from a normal distribution with unknown variance σ^2 , a $100(1 - \alpha)\%$ CI on σ^2 is

$$\frac{(n-1)s^2}{\chi_{\alpha/2, n-1}^2} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{1-\alpha/2, n-1}^2} \quad (4-62)$$

where $\chi_{\alpha/2, n-1}^2$ and $\chi_{1-\alpha/2, n-1}^2$ are the upper and lower $100\alpha/2$ percentage points of the chi-square distribution with $n - 1$ degrees of freedom, respectively. To find a CI on the standard deviation σ , simply take the square root throughout in equation 4-62.

For one-sided bounds, replace the $\alpha/2$ with α (just as before).

We're not going to worry about beta/power calculations for this test.

Let's practice with the table:

Just like with the Z-scores and t-scores, we have a table for the chi-square distribution. It is found in Table III of Appendix A in your textbook. Fortunately, the table is laid out in the same manner as the t-test distribution.

1. Suppose the sample size is 20 and it is one-sided test (greater than) with $\alpha = 0.05$. What is the χ^2 value?

2. Suppose the sample size is 15 and it is a two-sided test (not equal) with $\alpha = 0.05$. What is the χ^2 value?

3. Suppose the sample size is 10 and it is a one-sided test (less than) with $\alpha = 0.1$. What is the χ^2 value? (Hint: be sure to look on the left part of the curve instead of the right part of the curve)

Example (two-sided):

The sugar content of syrup in canned peaches is normally distributed and the variance is thought to be $\sigma^2 = 18 \text{ (mg)}^2$.

a. Test the hypothesis that the variance is NOT 18 mg^2 if a random sample of $n = 10$ cans yields a sample standard deviation of $s = 4 \text{ mg}$. Use a fixed-level test of $\alpha = 0.05$. State any necessary assumptions about the underlying distribution of the data.

Step 1: The parameter of interest is the true variance, σ^2 .

Step 2: $H_0: \sigma^2 = 18$.

Step 3: $H_1: \sigma^2 \neq 18$.

Step 4: The test statistic is chi-square: $\chi_0^2 = \frac{(n-1)s^2}{\sigma^2}$

Step 5: We will reject H_0 if $\chi_0^2 < \chi_{1-\alpha/2, n-1}^2$ or $\chi_0^2 > \chi_{\alpha/2, n-1}^2$

We look in the table for the value for these values. Alpha is 0.05, so $\alpha/2 = 0.025$. For 9 degrees of freedom, the first $\chi_0^2 = 2.70$. The second $\chi_0^2 = 19.02$.

Step 6: We calculate the chi-square statistic for our sample values:

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{9(16)}{18} = 8$$

Step 7: The value is 8. Since 8 is not less than 2.70 and 8 is not greater than 19.02, we fail to reject the null hypothesis. There is not sufficient evidence to conclude the true variance of sugar content differs from 18 mg² at the 0.05 test-level.

What assumptions do we need to make? The population needs to be normally distributed.

b. Estimate a p-value for this test.

We have a value of 8 for 9 degrees of freedom. Looking in the chart the $P(X^2 > 8)$ is close to 0.5. For a 2-sided test, the P-value includes both tails, so $0.5 + 0.5 = 1$.

c. Find a 95% confidence interval for the standard deviation.

We first find the CI for the variance:

For $\alpha = 0.05$ and $n = 10$, $\chi_{\alpha/2, n-1}^2 = \chi_{0.025, 9}^2 = 19.02$ and $\chi_{1-\alpha/2, n-1}^2 = \chi_{0.975, 9}^2 = 2.70$

$$\frac{9(16)}{19.02} \leq \sigma^2 \leq \frac{9(16)}{2.70}$$

$$7.57 \leq \sigma^2 \leq 53.33$$

So, now we take the square root of these values to find the CI for the standard deviation:

$$2.75 \leq \sigma \leq 7.30$$

With 95% confidence, the true standard deviation of sugar content is between 2.75 and 7.30 mg.

Example (one-sided):

A researcher for a tire manufacturer is investigating a new rubber compound. She makes 10 tires and road tests them. The sample mean life is 61,492 km and the sample standard deviation is 3035 km. Can you conclude at the $\alpha = 0.05$ level that the standard deviation of the tire life exceeds 3000 km?

Step 1: parameter of interest: _____

Step 2: H_0 :

Step 3: H_1 :

Step 4: Test statistic:

Step 5: Critical region:

Step 6: Calculate test statistic:

Step 7: Conclusion:

Find a bound on the p-value for this test:

95% lower confidence interval for variance:

Inference on proportion (z-test)

What is a proportion?

If X of n samples exhibit some category, then X/n is the *proportion* of the population that is in the category. For example, if the sample consists of 100 people and 25 are EE majors, then the proportion of the population of EE majors is $25/100 = \frac{1}{4}$.

Think back to the binomial distribution (success or failure). We will be using this for the z-test.

Just like before, we set up the data analysis task with a null hypothesis and alternative hypothesis:

$$H_0: p = p_0$$

$$H_1: p \neq p_0 \quad \text{OR} \quad H_1: p > p_0 \quad \text{OR} \quad H_1: p < p_0$$

The test statistic is based on the *binomial distribution*:

$$Z_0 = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}}$$

Where X is the number of observations in a random sample of size n that belongs to the class associated with p .

Good news! We already know how to look up z-scores in the table.

Testing Hypotheses on a Binomial Proportion		
Null hypotheses:	$H_0: p = p_0$	
Test statistic:	$Z_0 = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}}$	
Alternative Hypotheses	P-Value	Rejection Criterion for Fixed-Level Tests
$H_1: p \neq p_0$	Probability above $ z_0 $ and probability below $- z_0 $, $P = 2[1 - \Phi(z_0)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1: p > p_0$	Probability above z_0 , $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1: p < p_0$	Probability below z_0 , $P = \Phi(z_0)$	$z_0 < -z_\alpha$

This is the same procedure as before with the new z-score calculation.

Beta: Type II error calculation:

The approximate β -error for the two-sided alternative $H_1: p \neq p_0$ is

$$\beta = \Phi \left[\frac{p_0 - p + z_{\alpha/2} \sqrt{p_0(1 - p_0)/n}}{\sqrt{p(1 - p)/n}} \right] - \Phi \left[\frac{p_0 - p - z_{\alpha/2} \sqrt{p_0(1 - p_0)/n}}{\sqrt{p(1 - p)/n}} \right] \quad (4-66)$$

If the alternative is $H_1: p < p_0$,

$$\beta = 1 - \Phi \left[\frac{p_0 - p - z_{\alpha} \sqrt{p_0(1 - p_0)/n}}{\sqrt{p(1 - p)/n}} \right] \quad (4-67)$$

whereas if the alternative is $H_1: p > p_0$,

$$\beta = \Phi \left[\frac{p_0 - p + z_{\alpha} \sqrt{p_0(1 - p_0)/n}}{\sqrt{p(1 - p)/n}} \right] \quad (4-68)$$

Sample size calculation (two-sided):

Sample Size for a Two-Sided Hypothesis Test on a Binomial Proportion

$$n = \left[\frac{z_{\alpha/2} \sqrt{p_0(1 - p_0)} + z_{\beta} \sqrt{p(1 - p)}}{p - p_0} \right]^2 \quad (4-69)$$

If n is not an integer, round the sample size up to the next larger integer.

For one-sided calculations, replace $z_{\alpha/2}$ with z_{α} .

Confidence Intervals

Just like before, another way to do statistical inference is through confidence intervals. The process is the same, but just with this updated z-score calculation.

Confidence Interval on a Binomial Proportion

If \hat{p} is the proportion of observations in a random sample of size n that belong to a class of interest, an approximate $100(1 - \alpha)\%$ CI on the proportion p of the population that belongs to this class is

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad (4-73)$$

where $z_{\alpha/2}$ is the upper $100 \alpha/2$ percentage point of the standard normal distribution.

If it is a one-sided bound, just use the left or right side and use z_{α} instead of $z_{\alpha/2}$.

Sample size for specified error:

Sample Size for a Specified Error E on a Binomial Proportion

If \hat{P} is used as an estimate of p , we can be $100(1 - \alpha)\%$ confident that the error $|\hat{P} - p|$ will not exceed a specified amount E when the sample size is

$$n = \left(\frac{z_{\alpha/2}}{E} \right)^2 p(1 - p) \quad (4-74)$$

Example:

Large passenger vans are thought to have a high propensity of rollover accidents when fully loaded. Thirty accidents of these vans were examined and 11 vans had rolled over.

a. Test the claim that the proportion of rollovers exceeds 0.25 with $\alpha = 0.10$.

Step 1: parameter of interest is proportion p of true rollovers

Step 2: $H_0: p = 0.25$

Step 3: $H_1: p > 0.25$ // one-sided

Step 4: Use z-test with binomial distribution approximation

Step 5: Reject H_0 if $z_0 > 1.28$ (1.28 is the z-score for 90% of distribution being to the left)

Step 6: Calculate score

$$x = 11, n = 30, \hat{p} = \frac{11}{30} = 0.367$$

$$z_0 = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}} = \frac{11 - (30 \cdot 0.25)}{\sqrt{30 \cdot 0.25 \cdot 0.75}} = 1.48$$

Step 7: Since $1.48 > 1.28$, we reject the null hypothesis and conclude the percentage of rollovers exceeds 25%.

b. Suppose the true proportion is 0.35 and $\alpha = 0.10$. What is the beta error for this test?

$$\beta = \Phi \left[\frac{p_0 - p + z_{\alpha} \sqrt{p_0(1 - p_0) / n}}{\sqrt{p(1 - p) / n}} \right] = \Phi \left[\frac{0.25 - 0.35 + 1.28 \sqrt{0.25(1 - 0.25) / 30}}{\sqrt{0.35(1 - 0.35) / 30}} \right] = \Phi[0.0137]$$

$$\beta \cong 0.5055$$

c. Suppose the true proportion is 0.35 and alpha = 0.10. How large a sample would we need if we want beta to be 0.10?

$$n = \left[\frac{z_{\alpha} \sqrt{p_0(1-p_0)} + z_{\beta} \sqrt{p(1-p)}}{p - p_0} \right]^2 = \left[\frac{1.28 \sqrt{0.25(1-0.25)} + 1.28 \sqrt{0.35(1-0.35)}}{0.35 - 0.25} \right]^2$$

$$n = 135.67, n \cong 136$$

d. Find a 90% lower confidence bound on the rollover rate of the vans.

$$0.367 - 1.28 \sqrt{\frac{0.367(0.633)}{30}} \leq p$$

$$0.254 \leq p$$

The confidence interval on proportion is (.254, 1]. Or we can say the lower confidence bound is 0.254.

e. Use the confidence bound to test the hypothesis.

The hypothesized mean of 0.25 does not fall in this interval, so we reject the null hypothesis.

Single sample (population) inference practice

Directions: Get into your team of 4 people. Your team will work through problems. Work together as a team – do not divide and conquer the problems. Each team member is assigned to one of the following roles:

- Notetaker: prepares the master copy to be submitted
- Manager: ensures that group stays on task, manages time, and ensures everyone gets to speak
- Spokesperson: when called upon, speaks on behalf of the team
- Reflector: analyzes team dynamics and makes suggestions for improving team process – is everyone contributing? Does everyone understand how to get to the solution?

Names: Notetaker: _____
Spokesperson: _____

Manager: _____
Reflector: _____

1. Formulate the appropriate null and alternative hypotheses to test the following claims. The first one is done for you as an example.

Situation	Null hypothesis, H_0	Alternative hypothesis, H_1
A plastics production engineer claims that at least 99.95% of the plastic tube manufactured by her company meets the engineering specs.	$p = 0.9995$	$p > 0.9995$
A chemical and process engineering team claims that the mean temperature of a resin bath is greater than 45 degrees Celsius.		
The proportion of start-up software companies that successfully market their product within 3 years of company formation is less than 0.05.		
A chocolate bar company claims that, at the time of purchase by a consumer, the mean life of its product is less than 90 days.		
The designer of a computer laboratory at a major university claims that the standard deviation of time of a student on the network is less than 10 minutes.		
A manufacturer of traffic signals advertises that its signals have a mean operating life longer than 2160 hours.		

Recall the 7-step process for statistical inference:

- Step 1: What is the parameter of interest? _____
- Step 2: What is H_0 ? _____
- Step 3: What is H_1 ? _____
- Step 4: What test-statistic should we use? _____
- Step 5: What is the critical region boundary? (based on alpha) _____
- Step 6: Calculate the statistic with the data from the experiment
- Step 7: Conclusion: We reject / fail to reject (*circle answer*) _____ the null hypothesis.

Confidence intervals for statistical inference:

Another way to look at statistical inference is by calculating confidence intervals. If the parameter of interest falls outside the interval, we reject the null hypothesis. If the parameter of interest falls inside the interval, we fail to reject the null hypothesis.

The nice thing about statistical inference is that we have procedures (either 7-step or calculating confidence intervals) that we consistently follow.

2. The life in hours of a furnace heating element is known to be approximately normally distributed. A random sample of 15 heating elements is selected and found to have an average life of 598.14 hours with a *sample* standard deviation of 16.93 hours. Use $\alpha = 0.05$ for the test level. Is the mean life > 550 hours?

Follow the 7 steps:

- Step 1: What is the parameter of interest? _____
- Step 2: What is H_0 ? _____
- Step 3: What is H_1 ? _____

- Step 4: What test-statistic should we use? _____

- Step 5: What is the critical region boundary? (based on alpha) _____

- Step 6: Calculate the statistic with the data from the experiment

- Step 7: Conclusion: We reject / fail to reject (*circle answer*) _____ the null hypothesis.

3. Use the same data in problem 2. Construct a 95% lower confidence bound on the mean life. Do you reject the null hypothesis or fail to reject the null hypothesis?

4. Use the same data in problem 2. Construct a two-sided 95% confidence interval on the population variance.

5a. A manufacturer of cell phones claims that less than 1% of its production output is defective. A random sample of 1200 cell phones contains 8 defective units. Run a statistical test to support or reject this claim at $\alpha = 0.01$ level.

Step 1: What is the parameter of interest? _____

Step 2: What is H_0 ? _____

Step 3: What is H_1 ? _____

Step 4: What test-statistic should we use? _____

Step 5: What is the critical region boundary? (based on α) _____

Step 6: Calculate the statistic with the data from the experiment

Step 7: Conclusion: We reject / fail to reject (*circle answer*) the null hypothesis.

5b. What is the proportion of cell phones that are defective in this sample?

5c. Why do we fail to reject the null hypothesis even though the sample proportion is less than 1%

6a. A standardized test for graduating high school seniors is designed to be completed by 75% of students within 40 minutes. A random sample of 100 graduates showed that 64 completed the test within 40 minutes. Find a 90% two-sided traditional confidence interval on the proportion completing the test in 40 minutes.

6b. Use the data in 6a. Find a 95% two-sided traditional confidence interval on the proportion completing the test in 40 minutes.

6c. At the alpha-level of 0.05, is the proportion significantly different from 0.75?

7. In the production of airbags, a company wants to ensure that the mean distance of the foil to the edge of the inflator is at least 2.00 cm. Measurements on 20 inflators yielded an average value of 2.02 cm. The standard deviation for the measurements in the population is known to be 0.05 cm. Use a significance test level of 0.01.

Step 1: What is the parameter of interest? _____

Step 2: What is H_0 ? _____

Step 3: What is H_1 ? _____

Step 4: What test-statistic should we use? _____

Step 5: What is the critical region boundary? (based on alpha) _____

Step 6: Calculate the statistic with the data from the experiment

Step 7: Conclusion: We reject / fail to reject (*circle answer*) the null hypothesis.

8. Does your team have any questions about statistical inference on the mean, variance, or proportion?

Checkpoint 1: Stop for class discussion. If your team has reached this checkpoint, you may try to formulate your own problem that can be solved via any of the tests. Be ready to share your team answers with the class.

EGR361 Review Sheet – Midterm Exam #2

Content: Exam 2 will cover sections 3.10 to 3.13 and chapter 4 of the textbook. Note that material about significant figures is not in your textbook, but was covered in this section of the course and may be on the exam. Material will be drawn from homework assignments, lectures, in-class activities, and the textbook.

Procedure: Please arrive to class on time. You may use **one** sheet of 8/5" x 11" paper (**both sides**) during the exam. You may use a scientific calculator or regular calculator (NO cell phone, NO computer) during the exam. You may NOT use ear buds and electronic devices during the exam.

You will be given copies of the z-table, t-table, and chi-square table for use during the exam.

Topics: This study guide is not a contract – in other words, the exam may not cover every topic listed below and there may be topics that we covered in class that are not explicitly listed.

- Significant figures
 - Addition rule: least precise
 - Multiplication / Division rule: least number of sig figs

Chapter 3:

- Approximations to normal distribution
 - Poisson

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

- Binomial

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

- Propagation of error
 - $E(Y) = \mu_Y \cong h(\mu_1, \mu_2, \dots, \mu_n)$
 - $V(Y) = \sigma_Y^2 \cong \sum_{i=1}^n \left(\frac{\partial h}{\partial X_i} \right)^2 * \sigma_i^2$
 - Shortcut: linear combination
 - For this course, we will always assume independent random variables
- Central Limit Theorem
 - Random sample (random variables with the same distribution)
 - If we are sampling from a population with an unknown probability distribution, the sampling distribution of the sample means is still approximately normal with mean μ and variance σ^2/n if the sample size n is large.
 - As n goes to infinity:
 - $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$
 - When to use: $n \geq 30$ for any shape of population

- When to use: $n \geq 4$ if distribution is not severely non-normal

Chapter 4: Statistical Inference for One-Sample (One-population)

- Point estimates
- One-sided vs two-sided tests (one-tail versus two-tailed tests)
- Null and alternative hypotheses
- Hypothesis testing: 7 steps
 - Parameter of interest
 - Null hypothesis
 - Alternative hypothesis
 - Test statistic choice
 - Critical region
 - Calculate test statistic
 - Reject / Fail to Reject null hypothesis
 - (Optional) Calculate p-value for the test-statistic
- Type I error = alpha
 - Probability of rejecting the null hypothesis when null hypothesis is true
- Type II error = beta
 - Probability of failing to reject null hypothesis when null hypothesis is false
- P-value
 - If null hypothesis is true, the probability of getting a random sample whose mean (or other stat) is at least as far from the sample mean
 - Area in tail (1-tailed test) or the sum of the area in both tails (two-sided)
- Inference on mean of a single population
 - Z-test (population variance known and population is normally distributed or central limit applies)
 - T-test (population variance unknown, central limit does not apply)
 - Confidence Intervals / Bounds on mean
- Inference on variance of a single population
 - Chi-square test
 - Confidence Interval / Bounds on variance
 - Note: use this test for standard deviation, too (just do the hypothesis testing with variance)
- Inference on proportion of a single population
 - Z-test (binomial distribution)
 - Confidence Interval / Bounds for proportion
- Sample size calculations
- Beta calculations

Excel Skills:

- Z.TEST(array, x, sigma) // data, value, pop standard deviation
 - One-sample test
- T.TEST(array1, array2, 1 or 2-tailed, type)
 - Two-sample test

Inference on mean and proportion of two samples (two populations)

Consider these questions:

- Do more men participate in college sports than women?
- Do smokers have higher blood pressure than non-smokers?
- Are 10-year-old girls taller than 10-year-old boys?
- Is software A faster than software B?
- Does the pass rate for OR 5th graders on a math test exceed the pass rate for CA 5th graders?
- Are the net weight of cereal boxes the same from Plant A and Plant B?

1. What do they have in common?

2. Take the smoker question. What evidence would you need for you to believe that smokers have higher blood pressure than non-smokers?

3. Let's think about this as two distributions.

a. Draw two distributions (smoker and non-smoker) where you would NOT be convinced that the two populations are different in terms of blood pressure (top number of blood pressure reading)?

b. Draw two distributions (smoker and non-smoker) where you would be convinced that the two populations are different in terms of blood pressure?

4. Why are two-sample (or two-population) tests important to engineers and computer scientists?

CASES FOR STATISTICAL INFERENCE ON MEANS OF TWO POPULATIONS (NORMAL or CENTRAL LIMIT THEOREM APPLIES)

Test	Situation
<p>Z-test</p> $Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	<ul style="list-style-type: none"> Variances of both populations are known Delta is often 0 (to compare difference in means)
<p>T-test (with pooled estimator S_p)</p> $S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ $T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ <p>has a t distribution with $n_1 + n_2 - 2$ degrees of freedom.</p>	<ul style="list-style-type: none"> Variances of both populations are unknown Variances of both populations are equal
<p>T-test (with degrees of freedom)</p> $T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (5-10)$ <p>is distributed approximately as t with degrees of freedom given by</p> $v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}} \quad (5-11)$ <p>if the null hypothesis $H_0: \mu_1 - \mu_2 = \Delta_0$ is true. If v is not an integer, round down to the nearest integer.</p>	<ul style="list-style-type: none"> Variances of both populations are unknown Variances are unequal for both populations Delta is often 0

<p>T-test (differences in pairs)</p> $T_0 = \frac{\bar{D} - \Delta_0}{S_D / \sqrt{n}}$	<ul style="list-style-type: none"> • Data collected in pairs (X_{1i}, X_{2i}). Before/After a treatment is often a paired test. Take same item and put it under 2 conditions – this is also data pair. • D-bar is the average of the differences in the pairs • The d's are normally distributed with mean $\mu_1 - \mu_2$. • S_D is the sample standard deviation of the differences in the pairs
--	--

Confidence Interval / Bounds:

See chart on inside of back cover of textbook or see sections 5-2, 5-3, and 5-4 in textbook.

Sample Sizes:

Sample size for Z-test given in textbook (pages 234 – 235, 237)

Sample size for T-test for equal, unknown variances (use operating characteristic charts, see page 246)

Sample size for T-test for unequal, unknown variances (no OC charts available, cannot compute)

When to Pair Samples:

The paired T-test will lead to a smaller value of the variance of the difference in samples means. It does have a disadvantage – losing $n-1$ degrees of freedom compared to the two-sample T-test. Remember, increasing the degrees of freedom increases the power of a test. There are no “hard and fast” rules for determining when to pair samples when collecting data, but here are some guidelines:

- If units are relatively homogeneous (small variance) and correlation within pairs is small, the gain in precision for pairing will be offset by degrees of freedom, so use the two-sample approach (non-paired).
- If units are heterogeneous (large variance) and there is large positive correlation within pairs, use the paired T-test experiment. Generally, this occurs when the experimental units are the same for both treatments such as putting a product in a chemical bath versus not putting a product in a chemical bath.
- If the sample size is large (50 in each sample), then losing half the degrees of freedom by pairing may not be serious. If the sample sizes are small, losing half the degrees of freedom by pairing is potentially serious.

Practice with Decision-Making

Let's try to make the decision about which test to apply.

1. Suppose two types of plastic could be used for an electronics component. The breaking strength of plastic is important. It is known that the population variance of the breaking strength of both plastics is 1.0 psi squared and the population is normally distributed. Suppose plastic 1 has a mean breaking strength of 162.7 and plastic 2 has a mean breaking strength of 155.4. The sample size for plastic 1 is 10 and for plastic 2 is 12. You want to know if the breaking strengths are different. What test-statistic do you use?

Z-test

T-test (with pooled estimator)

T-test (with degrees of freedom)

T-test (paired)

2. The thickness of a plastic film on a substrate material is thought to be influenced by the temperature at which the coating is applied. Eleven substrates are coated at 125 degrees F and 13 are coated at 150 degrees F. The sample at 125 degrees have a mean thickness of 101.28 and sample standard deviation of 5.08 mm. The sample at 150 degrees have a mean thickness of 101.70 and sample standard deviation of 20.15 mm. Are the thicknesses different? What test statistic do you use?

Z-test

T-test (with pooled estimator)

T-test (with degrees of freedom)

T-test (paired)

3. The diameter of steel rods manufactured on two extrusion machines is being investigated. Two samples of size 15 and 17 are selected, with sample mean 1 to be 8.73 and sample mean 2 to be 8.68. The variances of the samples are: sample 1 variance = 0.35, sample 2 variance = 0.40. Assume that the population variance of the two samples is equal and the data come from a normal distribution. Is there evidence to support that the two machines produce rods with different mean diameters? Which test statistic do you use?

Z-test

T-test (with pooled estimator)

T-test (with degrees of freedom)

T-test (paired)

4. An article describes a new equivalent plate analysis method formulation that is capable of modeling aircraft structures. Natural vibration frequencies (in cycles/second) are calculated using both methods with the data showed here:

	FEA	Plate
1	14.58	14.76
2	48.52	49.10
3	97.22	99.99
4	113.99	117.53
5	174.73	181.22
6	212.72	220.14
7	277.38	294.80

Are the mean frequencies for both methods the same? Which test statistic do you use?

Z-test

T-test (with pooled estimator)

T-test (with degrees of freedom)

T-test (paired)

Inference on proportion differences

A related inference procedure is the difference between two proportions from two populations. We use the approximation of the binomial distribution to the normal distribution, just like we did in single-sample testing on proportion. We assume the proportions of the two populations have approximately normal distributions.

The test-statistic calculates z-values:

The quantity

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}} \quad (5-23)$$

has approximately a standard normal distribution, $N(0, 1)$.

Here, \hat{p}_1 and \hat{p}_2 with the little accents on top (hat) refer to the sample observed proportions. p_1 and p_2 are the hypothesized proportions. If you are testing if p_1 is different than p_2 , then $p_1 - p_2$ is equal to 0 (null hypothesis is $p_1 = p_2$, alternative hypothesis is $p_1 \neq p_2$).

Sometimes, you'll see Z defined as:

Testing Hypotheses on the Equality of Two Binomial Proportions

Null hypothesis: $H_0: p_1 = p_2$

Test statistic:
$$Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{P}(1 - \hat{P})(\frac{1}{n_1} + \frac{1}{n_2})}} \quad (5-24)$$

<u>Alternative Hypotheses</u>	<u>P-Value</u>	<u>Rejection Criterion for Fixed-Level Tests</u>
$H_1: p_1 \neq p_2$	Probability above $ z_0 $ and probability below $- z_0 $ $P = 2[1 - \Phi(z_0)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1: p_1 > p_2$	Probability above z_0 , $P = 1 - \Phi(z_0)$	$z_0 > z_{\alpha}$
$H_1: p_1 < p_2$	Probability below z_0 , $P = \Phi(z_0)$	$z_0 < -z_{\alpha}$

Here, $\hat{P} = (X_1 + X_2)/(n_1 + n_2)$

// it is the overall proportion for the two populations

Example: Here is an example problem type in which we would apply the Z-test for proportions.

Two types of polishing solutions are being evaluated for possible use in a tumble-polish operation for manufacturing contact lenses. In the first sample, 253 of 300 had no polishing-induced defects. Another 300 lenses were tumble-polished using a second polishing solution. 196 of those 300 had no polishing-induced defects. Is there any reason to believe the two polishing solutions differ? (Note: assuming normality of proportions – if many more tests were conducted, the proportion p for each sample would be normally distributed.)

FYI – can study ratio of variances of two populations (F test).

See section 5-5. If variances are the same, then we have a ratio of 1. The F distribution looks a bit like the chi-square distribution. The F-tables can be found in Appendix A of the textbook. The procedure is the same, but we are not going to take time to go over this explicitly in lecture.

You may want to consider adding testing of the ratio of variances of two populations to your project if you are studying two populations. Together with establishing if the mean is different, these two tests (mean and variance) can give you important information about your inferences on the population.

Example Calculation

1. Suppose two types of plastic could be used for an electronics component. The breaking strength of plastic is important. It is known that the population variance of the breaking strength of both plastics is 1.0 psi squared. Suppose plastic 1 has a mean breaking strength of 162.7 and plastic 2 has a mean breaking strength of 155.4. The sample size for plastic 1 is 10 and for plastic 2 is 12. The company will only adopt plastic 1 if its mean breaking strength exceeds that of plastic 2 by 10 psi since plastic 1 costs a lot more than plastic 2. What test-statistic do you use?

What do we know about the situation? Known variances for both populations.

Which test? Z-test

7 Steps:

Step 1: Parameter: The parameter of interest is the difference in breaking strengths, $\mu_1 - \mu_2$ and $\Delta_0 = 10$

Step 2: $H_0: \mu_1 - \mu_2 = 10$

Step 3: $H_1: \mu_1 - \mu_2 > 10$

Step 4: Test statistic is Z-test for sample means

$$z_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Step 5: Calculate critical region: Reject H_0 if $z_0 > z_\alpha = 1.65$ (one-sided with $\alpha = 0.05$)

Step 6: Calculate test statistic:

$$\bar{x}_1 = 162.7 \quad \bar{x}_2 = 155.4 \quad \delta = 10$$

$$\sigma_1 = 1.0 \quad \sigma_2 = 1.0$$

$$n_1 = 10 \quad n_2 = 12$$

$$z_0 = \frac{(162.7 - 155.4) - 10}{\sqrt{\frac{(1.0)^2}{10} + \frac{(1.0)^2}{12}}} = -6.31$$

Step 7: Since -6.31 is < 1.65, we fail to reject the null hypothesis. We cannot conclude that sample 1 has a mean breaking strength that is 10 more psi than sample 2.

Addendum

Now, suppose the cost of the plastic 1 goes way down and the company is willing to use plastic 1 if the mean breaking strength is 2 more psi for plastic 1 than for plastic 2.

Run the same steps, except now the delta is 2.

$$z_0 = \frac{(162.7 - 155.4) - 2}{\sqrt{\frac{(1.0)^2}{10} + \frac{(1.0)^2}{12}}} = 12.4$$

Since 12.4 is greater than 1.65, we reject the null hypothesis. Plastic 1's breaking strength exceeds plastic 2's breaking strength by 2 psi.

Two sample (two-population) inference practice

Directions: Get into your team of 4 people. Your team will work through at least one problem. Work together as a team – do not divide and conquer the problems. Each team member is assigned to one of the following roles:

- Notetaker: prepares the master copy to be submitted
- Manager: ensures that group stays on task, manages time, and ensures everyone gets to speak
- Spokesperson: when called upon, speaks on behalf of the team
- Reflector: analyzes team dynamics and makes suggestions for improving team process – is everyone contributing? Does everyone understand how to get to the solution?

Names: Notetaker: _____
Spokesperson: _____

Manager: _____
Reflector: _____

Tammy will assign your team a problem below. If you finish your problem, use the time to try another problem. These are the same problems from lecture – here you will conduct the calculations for the test. You may want to refer to the lecture notes on two-sample inference. We already did problem 1 in class as the sample calculation.

2. The thickness of a plastic film on a substrate material is thought to be influenced by the temperature at which the coating is applied. Eleven substrates are coated at 125 degrees F and 13 are coated at 150 degrees F. The sample at 125 degrees have a mean thickness of 101.28 and sample standard deviation of 5.08 mm. The sample at 150 degrees have a mean thickness of 101.70 and sample standard deviation of 20.15 mm. Are the thicknesses different? What test statistic do you use?

Z-test

T-test (with pooled estimator)

T-test (with degrees of freedom)

T-test (paired)

3. The diameter of steel rods manufactured on two extrusion machines is being investigated. Two samples of size 15 and 17 are selected, with sample mean 1 to be 8.73 and sample mean 2 to be 8.68. The variances of the samples are: sample 1 variance = 0.35, sample 2 variance = 0.40. Assume that the population variance of the two samples is equal and the data come from a normal distribution. Is there evidence to support that the two machines produce rods with different mean diameters? Which test statistic do you use?

Z-test

T-test (with pooled estimator)

T-test (with degrees of freedom)

T-test (paired)

4. An article describes a new equivalent plate analysis method formulation that is capable of modeling aircraft structures. Natural vibration frequencies (in cycles/second) are calculated using both methods with the data showed here:

	FEA	Plate
1	14.58	14.76
2	48.52	49.10
3	97.22	99.99
4	113.99	117.53
5	174.73	181.22
6	212.72	220.14
7	277.38	294.80

Z-test

T-test (with pooled estimator)

T-test (with degrees of freedom)

T-test (paired)

Use this space to work:

What problem are you doing? _____

Step 1:

Step 2:

Step 3:

Step 4:

Step 5:

Step 6:

Step 7:

If you finished your problem, try one involving determining if two proportions are different:

The rollover rate for SUVs is a transportation safety issue. Safety advocates claim that manufacturer A's vehicle has a higher rollover rate than that of manufacturer B. One hundred crashes were studied from each manufacturer. Manufacturer A had 35 rollovers. Manufacturer B had 25 rollovers. You can assume the proportions come from a normal distribution.

Does Manufacturer A's vehicles have a higher rollover rate than Manufacturer B? Use $\alpha = 0.05$.

Follow the 7-step procedure:

Step 1:

Step 2:

Step 3:

Step 4:

Step 5:

Step 6:

Step 7:

If you have more time, choose one of the problems on the first page to complete.

Two sample (two-population) confidence interval practice

Directions: Get into your team of 4 people. Your team will work through at least one problem. Work together as a team – do not divide and conquer the problems. Each team member is assigned to one of the following roles:

- Notetaker: prepares the master copy to be submitted
- Manager: ensures that group stays on task, manages time, and ensures everyone gets to speak
- Spokesperson: when called upon, speaks on behalf of the team
- Reflector: analyzes team dynamics and makes suggestions for improving team process – is everyone contributing? Does everyone understand how to get to the solution?

Names: Notetaker: _____
Spokesperson: _____

Manager: _____
Reflector: _____

These are the same problems from lecture and statistical inference practice (#2 through 4) – here you will conduct the confidence intervals/ bounds. You may want to refer to the chart on calculating confidence intervals.

2. The thickness of a plastic film on a substrate material is thought to be influenced by the temperature at which the coating is applied. Eleven substrates are coated at 125 degrees F and 13 are coated at 150 degrees F. The sample at 125 degrees have a mean thickness of 101.28 and sample standard deviation of 5.08 mm. The sample at 150 degrees have a mean thickness of 101.70 and sample standard deviation of 20.15 mm. Calculate the 95% confidence interval of the difference in mean thicknesses.

Z-test

T-test (with pooled estimator)

T-test (with degrees of freedom)

T-test (paired)

3. The diameter of steel rods manufactured on two extrusion machines is being investigated. Two samples of size 15 and 17 are selected, with sample mean 1 to be 8.73 and sample mean 2 to be 8.68. The variances of the samples are: sample 1 variance = 0.35, sample 2 variance = 0.40. Assume that the population variance of the two samples is equal and the data come from a normal distribution. Calculate the 95% confidence interval on the difference in the sample mean diameters.

Z-test

T-test (with pooled estimator)

T-test (with degrees of freedom)

T-test (paired)

4. An article describes a new equivalent plate analysis method formulation that is capable of modeling aircraft structures. Natural vibration frequencies (in cycles/second) are calculated using both methods with the data showed here. Calculate the 95% confidence interval on the mean difference between the two methods, FEA - Plate.

	FEA	Plate
1	14.58	14.76
2	48.52	49.10
3	97.22	99.99
4	113.99	117.53
5	174.73	181.22
6	212.72	220.14
7	277.38	294.80

Z-test

T-test (with pooled estimator)

T-test (with degrees of freedom)

T-test (paired)

If you finished your problem, try one involving determining if two proportions are different:

The rollover rate for SUVs is a transportation safety issue. Safety advocates claim that manufacturer A's vehicle has a higher rollover rate than that of manufacturer B. One hundred crashes were studied from each manufacturer. Manufacturer A had 35 rollovers. Manufacturer B had 25 rollovers. You can assume the proportions come from a normal distribution.

Calculate the 95% lower confidence bound on $P_A - P_B$.

If you have more time, choose one of the problems on the first page to complete.

Two Sample Experiment

M&Ms

Do not eat your candy yet.

What question(s) do you want to investigate?

Examples:

Is mean number of M&Ms different in plain versus peanut bags?

Is proportion of green M&Ms different in plain versus peanut bags?

What is the question?

What data do we need to collect?

What statistical test should we run?

What test level should we use?

ANOVA stands for Analysis of Variance

When to use: more than two groups / categories of interest and you want to know if the means differ

One-way ANOVA assumptions:

1. One dependent variable
2. Three or more groups (otherwise, can use t-test); also called factors
3. Independence of observations (no relationship between samples between groups or within groups); This means that one person or item should not be in more than one group.
4. No significant outliers
5. Dependent variable should be approximately normal for each group/category
6. Variances for each group are similar (can use Welch ANOVA instead of one-way)

Example: You want to know if a drink impacts the resting heart rate. You give water to Group A. You give orange juice to Group B. You give coffee to Group C. You measure resting heart rate of the individuals in each group.

What is the dependent variable?

What are the groups/factors?

What data is collected?

What if there are two factors? We can use a two-way ANOVA

Example: Is the effect of gender (male/female) on test anxiety influenced by educational level (undergrad/grad)?

What is the dependent variable?

What are the groups/factors?

What data is collected?

(Here gender and educational level are the factors). The data collected is test anxiety. We would have a 4 (2 x 2) sample sets.

Example: Does stress vary according to age and gender?

Here age and gender are the factors. The groups / categories would be “less than 40”, “40 to 50”, “over 50” for age or some other bracketing and “male” and “female” for gender.

The ANOVA test tells you if you can *make the call* that the means are different (more variability between groups than within groups).

You won't be asked to do the ANOVA by hand. However, you should know when to apply the test and what the results mean.

See excel sheets for example calculations.

EGR361 Review Sheet – Midterm Exam #3

Content: Exam 3 will cover chapter 5 of the textbook. Note that material about significant figures is not in your textbook, but was covered in this section of the course and may be on the exam. Material will be drawn from homework assignments, lectures, in-class activities, and the textbook.

Procedure: Please arrive to class on time. You may use **one** sheet of 8/5" x 11" paper (**both sides**) during the exam. You may use a scientific calculator or regular calculator (NO cell phone, NO computer) during the exam. You may NOT use ear buds and other electronic devices during the exam.

You will be given copies of the z-table, t-table, and chi-square table for use during the exam.

Topics: This study guide is not a contract – in other words, the exam may not cover every topic listed below and there may be topics that we covered in class that are not explicitly listed.

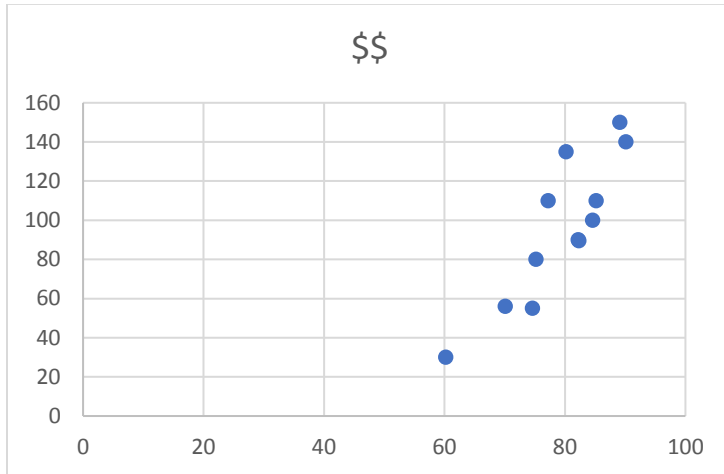
Chapter 5: Statistical Inference for Two-Samples (Two-populations)

- Inference on difference in means
 - Z-test (both populations normal or central limit applies, population variances are known for both populations)
 - T-test, pooled estimator (population variances unknown but equal)
 - T-test, degrees of freedom (population variances unknown and not necessarily equal)
 - Confidence interval / bounds on difference in means
 - Sample sizes
 - Beta calculations
- Inference on difference in paired data
 - Paired T-test (data collected in pairs, calculate difference in pair and this becomes data item)
 - Confidence interval / bounds on mean of pair difference
- Inference on difference in proportions
 - Z-test
 - Sample sizes
 - Beta calculations
 - Confidence interval / bounds on proportion
- ANOVA (Analysis of Variance): Difference in at least two means from multiple populations
 - One-way (single factor, multiple treatments)
 - Two-way (two factor, multiple treatments)
 - F-statistic (excel) and p-values (excel) – how to interpret them

Correlation, Scatterplots, Linear Regression (Simple, one variable), Multiple Regression

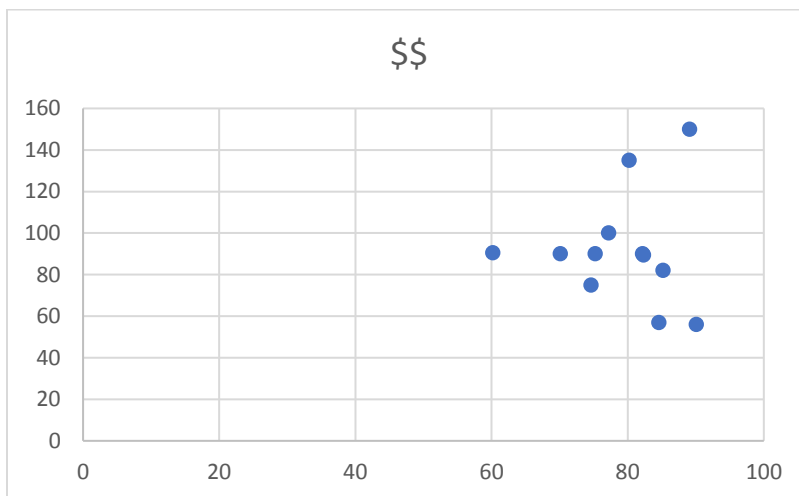
Suppose X = temperature and Y = \$ made selling lemonade

We plot the temperature on the x-axis and \$ on the y-axis. Suppose that plot looks like this. This is a scatterplot and can be built using Excel's Insert->Insert Chart-> X Y (Scatter).



Would you say \$\$ relates to temperature? Why or why not?

Suppose the plot looks like this.



Would you say \$\$ relates to temperature? Why or why not?

Correlation

What is correlation? The **sample correlation coefficient**, r , is defined as:

$$r = \frac{\sum[(x - \bar{x})(y - \bar{y})]}{\sqrt{[\sum(x - \bar{x})^2][\sum(y - \bar{y})^2]}}$$

Here, x and y are values for a single occurrence. The sums are over all data items in the sample.

Let's consider r .

When is r close to 1?

When is r close to 0?

When is r close to -1?

When y positively relates to x

When y does not relate to x at all

When y negatively relates to x

Example correlation calculation:

Blood pressure readings are collected from 6 people. The age and high number from the blood pressure are recorded. Is age correlated with blood pressure?

Subject	Age (x)	Pressure (y)
A	43	128
B	48	120
C	56	135
D	61	143
E	67	141
F	70	152
means	57.5	136.5

1. Calculate $(x - \bar{x})$ and $(y - \bar{y})$ for each sample:

$x - \bar{x}_{\text{bar}}$	$y - \bar{y}_{\text{bar}}$
-14.5	-8.5
-9.5	-16.5
-1.5	-1.5
3.5	6.5
9.5	4.5
12.5	15.5

2. Calculate the squares of each of these and the sum for each column:

$(x - \bar{x}_{\text{bar}})^2$	$(y - \bar{y}_{\text{bar}})^2$
210.25	72.25

	90.25	272.25
	2.25	2.25
	12.25	42.25
	90.25	20.25
	156.25	240.25
sum	561.5	649.5

3. Calculate $(x - \bar{x}) * (y - \bar{y})$ and the sum:

(x-x_bar)(y-y_bar)
123.25
156.75
2.25
22.75
42.75
193.75
541.5

4. Sample correlation coefficient $r = 541.5 / \sqrt{(561.5 * 649.5)} = 0.896673$.

In excel: use the Correlation function in the Data Analysis Toolkit.

With excel:

	<i>Age (x)</i>	<i>Pressure (y)</i>
Age (x)	1	
Pressure (y)	0.896673	1

The top-left box gives you the correlation of age with age. The bottom-right box gives you the correlation of pressure with pressure. The bottom-left box gives you the correlation of age with pressure. Because $r = .896673$, there is a strong positive relationship between age and blood pressure.

NOTE: CORRELATION **DOES NOT** IMPLY CAUSATION. THIS IS ONE OF THE MOST MISUNDERSTOOD CONCEPTS OF DATA ANALYSIS.

Covariance

Covariance is how much two variables vary together. If two variables are independent, then the covariance is 0. When two variables are dependent, the covariance is not 0.

$$cov(X, Y) = \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Example calculation with the data from above:

1. Calculate $(x - \bar{x}) * (y - \bar{y})$ and the sum:

(x-x_bar)(y-y_bar)

123.25

156.75

2.25

22.75

42.75

193.75

541.5

2. Take the sum of 541.5 and divide by 6 (6 samples):

$$Cov(age, pressure) = \frac{541.5}{6} = 90.25$$

In excel: use the Covariance function in the Data Analysis Toolkit.

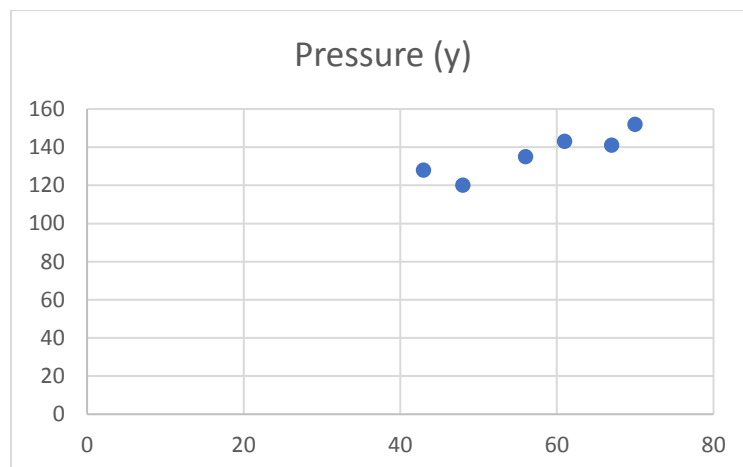
With excel

Covariance:

	Age (x)	Pressure (y)
Age (x)	93.58333	
Pressure (y)	90.25	108.25

Scatterplots

A graph of dots showing the (x,y) values of the data. It is a visual representation of how two variables are related. Here is the scatterplot of (age, pressure) from the data above:

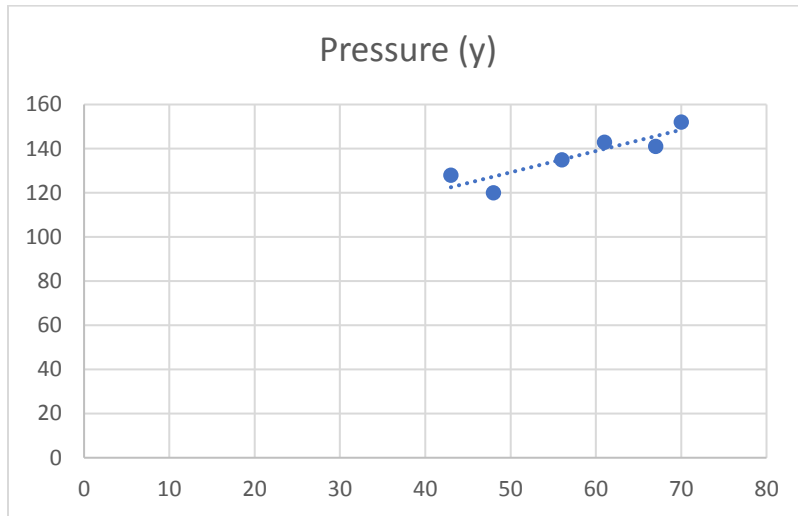


Draw a line that best “fits” these points.

What would such a line look like?

What values would we want to minimize?

Here is the same scatterplot with the regression line (trendline):



In excel: Left-click on the scatterplot. Click on the + box. Select Trendline. The trendline should appear on the chart.

Linear Regression (Simple)

Creating the trendline is linear regression. Why is this trendline useful?

This gives us an empirical model. Remember, not all relationships in engineering come from theoretical mechanical equations, such as $F = ma$. We may have to collect lots of data and then build empirical models. These models can then be used for prediction. For example, suppose a new patient comes to the clinic and that patient is 50 years old. What would you expect the blood pressure to be?

Simple linear regression builds a linear model:

Usually we say the independent variable is the *regressor variable* x

Usually we say the dependent variable is the *response variable* y

In our example, the regressor variable is age and the response variable is blood pressure.

How do we determine the equation for the line?

$$Y = \beta_0 + \beta_1 X + \text{error}$$

We want to minimize the error, so the predicted Y is as close to the observed y value as possible. How do we do this?

We minimize the sum of squares from the predicted Y (trendline estimate) and the observed y for all the data points. This is called *least squares estimation*.

Fortunately, we have equations to estimate the beta values that do least squares estimation.

The errors (observed_y – estimated_y) are called *residuals*.

Here are the equations to compute the beta values:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}$$

β_1 is really $\text{cov}(x,y)$ divided by $\text{cov}(x,x)$.

Example Calculation from the Data Above:

1. Calculate numerator. We did this above and the result is 541.5.
2. Calculate the denominator. We did this above and the result is 561.5.
3. $\beta_1 = 0.964381$
4. Calculate mean of observed y and mean of observed x. We did this above. The mean of y is 136.5. The mean of x is 57.5.
5. $\beta_0 = 81.04809$

In excel: Use Regression in Data Analysis Toolkit:

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.896673
R Square	0.804022
Adjusted R Square	0.755028
Standard Error	5.641091
Observations	6

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	522.2124	522.2124	16.41047	0.015463
Residual	4	127.2876	31.82191		
Total	5	649.5			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	81.04809	13.88088	5.838829	0.004289	42.50858	119.5876	42.50858	119.5876
Age (x)	0.964381	0.238061	4.050984	0.015463	0.303418	1.625344	0.303418	1.625344

RESIDUAL OUTPUT

<i>Observation</i>	<i>Predicted Pressure (y)</i>	<i>Residuals</i>
1	122.5165	5.483526

2	127.3384	-7.33838
3	135.0534	-0.05343
4	139.8753	3.124666
5	145.6616	-4.66162
6	148.5548	3.445236

Let's dissect the output from excel:

1. Multiple R = r (sample correlation coefficient that we calculated earlier).
2. R square = r^2 (proportion of variability explained by linear regression model).
3. Coefficient Intercept = β_0 .
4. Coefficient Age (x) = β_1 .
5. Residual Output shows the residuals for each observation.

What is the ANOVA for?

1. It shows how much the regression model predicts the residuals
2. Total SS (SST) = SS_R + SS_E.
3. See equation 6-21 on page 310 of textbook.
4. $R^2 = 1 - (SS_E / SS_T)$. In this case, it is 0.804022.
5. Often gives redundant info as the regression stats, so you can usually ignore the ANOVA output.

What is the next section with the coefficients?

1. Coefficients are the intercept and slope of the regression line.
2. t Stat is the result of running a t-test on each of these estimates.
 - a. $H_0: B_1 = .964381$; $H_1: B_1$ does not equal .964381
 - b. $H_0: B_0 = 81.04809$; $H_1: B_0$ does not equal 81.04809
3. If p-values for t-test are low, then the regression model is good.
4. The lower 95% and upper 95% give the confidence interval on these predictors. (See section 6-2.3 on page 315 for more information)

What is the residual output?

1. This gives the residuals for each observation (observed value – predicted value)

Multiple Regression

What if you have more than one variable when building an empirical model?

If you want to build a linear model of multiple variables, you may use multiple regression. For example, you might be trying to predict the number of spring babies in a herd of antelope based on the current population, how much precipitation happened over the winter, and the severity of the winter.

Spring Fawn Count (/100)	Antelope Pop (/100)	Precipitation	Winter severity
2.900000095	9.199999809	13.19999981	2
2.400000095	8.699999809	11.5	3
2	7.199999809	10.80000019	4
2.299999952	8.5	12.30000019	2
3.200000048	9.6	12.60000038	3
1.899999976	6.800000191	10.60000038	5
3.400000095	9.699999809	14.10000038	1
2.099999905	7.900000095	11.19999981	3

You want to build a model:

$$FawnCount = \beta_0 + \beta_1 * pop + \beta_2 * precipitation + \beta_3 * winterSeverity$$

Notice that the model is linear and we are trying to minimize the error of the fawn count predictions when estimating the values for the betas. For more information about the equations, see pages 327-328 in your textbook. However, in EGR361, you will not need to do multiple regression by hand. Instead, you can use excel.

In excel: Data -> Data Analysis -> Regression

Choose for Input Y range the predicted variable (in this case, Spring Fawn Count).

Choose the other three columns for the X range.

If you selected with the labels (keep labels if possible), check the box.

Keep the confidence level at 95%.

Select output range as a cell in the current sheet.

Click on Residuals to get the calculations of the residuals.

You can optionally select the other boxes for residuals if you want to see more information.

Look at the excel output.

R = .987, so the regression model is a good predictor.

What is the model? (look in the second box under ANOVA to get the coefficients)

$$FawnCount = -5.922 + 0.338 * pop + 0.402 * precipitation + 0.263 * winterSeverity$$

So, we can use this equation to make predictions. Suppose the population is 8.26, the precipitation is 13.5, and the winter severity is 2. How many spring fawns would you expect?

2.817

Where might multiple regression be used in engineering and computer science?

Summary

How has your perspective about data changed after taking this course?

How has your perspective about reading articles/studies that use statistics changed?

Statistical inference is a scientist's or engineer's method to "prove" differences in populations or "prove" a result from a single population. The tools we have used in this course are parametric tests, because we know the underlying distribution or the data or can make the assumption that the underlying distribution is normal due to the central limit theorem. However, there is an entire class of statistical tests called non-parametric tests that scientists use when the underlying distribution of the data is not known, the assumptions for using a parametric test are not met, or if the parameter of study is the median. Examples of these non-parametric tests include Wilcoxon, Mann-Whitney, Kruskal-Wallis.

Hopefully, this course has given you a foundation on which to organize new statistical testing knowledge that you will gather throughout your career. You can always look up the details of a test or simply use a software tool to run the test. However, you need to know how to choose a test, how much data to collect, when to use a one-sided or two-sided test, and how to make appropriate conclusions. I hope this course has provided practice for these skills.

EGR361 Review Sheet – Final Exam

Content: The final exam covers chapters 1 – 6 of the textbook. Note that material about significant figures is not in your textbook, but is fair game for the exam. Material will be drawn from homework assignments, lectures, in-class activities, and the textbook.

Procedure: Please arrive to the exam on time. You may use **two** sheet of 8/5" x 11" paper (**both sides**) during the exam. You may use a scientific calculator or regular calculator (NO cell phone, NO computer) during the exam. You may NOT use ear buds and other electronic devices during the exam.

You will be given copies of the z-table, t-table, and chi-square-table for use during the exam.

Topics: This study guide is not a contract – in other words, the exam may not cover every topic listed below and there may be topics that we covered in class that are not explicitly listed.

Chapter 1:

- Sampling
 - Random samples, simple random sampling
 - Populations (conceptual versus physical)
- Types of studies
 - Enumerative vs analytic
 - Retrospective, Observational, Designed
- Models
 - Mechanistic vs empirical

Chapter 2:

- Calculating descriptive statistics
 - Sample mean (\bar{x})
 - Population mean (μ)
 - Sample variance (s^2)
 - Population variance (σ^2)
 - Sample standard deviation (s)
 - Population standard deviation (σ)
 - 5-number summary (Q0 to Q4)
 - 3 Methods for calculating Q1 and Q3 (in class)
- Plots
 - Stem-and-leaf (see book)
 - Histograms (in class and in book)
 - Box Plots (in class and in book)
 - Time series plots (see book)
 - Scatterplots (from last part of course)

Chapter 3:

- Random variables (discrete vs continuous) (usually denoted X)
- Probability (likelihood, chance)
 - Venn Diagrams, Sets, Complement, Union, Intersection
 - Outcome: Result of a single trial
 - Sample space: Set of all outcomes (sum of probabilities of all outcomes is 1)
 - Mutually Exclusive (Intersection is empty between two events)
 - Independence (Outcome of event 1 does not impact outcome of event 2)

- Conditional Probability
- Counting
 - Permutations
 - Combinations
- Discrete Random Variables and Distributions
 - PMF (probability mass function) – how to determine if $f(x)$ is a pmf
 - CDF (cumulative distribution function) – piecewise continuous function
 - Calculate mean (μ) or $E(X)$
 - Calculate variance (σ^2) or $V(X)$
 - Calculate standard deviation (σ)
- Binomial Distribution
 - When to use it? How to use it?
 - Example: the probability of 3 parts having flaws of 100 parts, where the probability for a flawed part is 0.1
 - $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, 2, 3 \dots n$
 - $\mu = E(X) = np$
 - $\sigma^2 = V(X) = np(1-p)$
- Poisson Distribution
 - When to use it (intervals are independent, usually referring to # events over an interval, $np = \lambda$), How to use it?
 - Example: the probability of 10 customers arriving to a store in an hour, where the mean is 20 customers per hour and customer arrival follows a Poisson process
 - Remember: λ needs to be in the correct units; must match units in question
 - $f(x) = P(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$, $x = 0, 1, 2, \dots$
 - $\mu = E(X) = \lambda$
 - $\sigma^2 = V(X) = \lambda$
- Continuous Random Variables
 - PDF (probability density function)
 - Area under curve sums to 1
 - $f(x) \geq 0$ for all x
 - $P(a < X < b) = \int_a^b f(x) dx$
 - CDF
 - $F(x) = P(X \leq x) = \int_{-\infty}^x f(u) du$ for $-\infty < x < \infty$
 - $\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$
 - $\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = E(X^2) - \mu^2$
- Exponential Distribution
 - Related to Poisson: The spacing between events in a Poisson distribution
 - Example: the probability that the distance between flaws in a copper wire is between 0.5 and 1 cm where the mean number of flaws is 2 per cm and the number of flaws follows a Poisson process.
 - Can also stand on own as a distribution
 - Example: the probability that there are no calls in a 30-minute window and the call times are exponentially distributed with mean time between calls of 10 minutes
 - How to use it? Need to integrate $f(x)$ over (a, b) if looking for $P(a < X < b)$.
 - Remember: mean needs to be in correct units; must match units in question
 - $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$
 - $\mu = E(X) = \frac{1}{\lambda}$

- $\sigma^2 = V(X) = \frac{1}{\lambda^2}$
- Normal Distribution
 - Symmetric about mean
 - Mean is at peak
 - Standard deviation is related to width
 - $f(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$
 - Z – standard normal random variable
 - $Z = \frac{X-\mu}{\sigma}$
 - Integrating is no fun, so there is a table of probabilities for the standard normal
 - Table I in your book: values are the CDF of the standard normal distribution
 - Checking for normality
- Central Limit Theorem
 - Random sample (random variables with the same distribution)
 - If we are sampling from a population with an unknown probability distribution, the sampling distribution of the sample means is still approximately normal with mean μ and variance σ^2/n if the sample size n is large.
 - As n goes to infinity:
 - $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$
 - When to use: $n \geq 30$ for any shape of population
 - When to use: $n \geq 4$ if distribution is not severely non-normal
- Approximations to normal distribution
 - Poisson

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$
 - Binomial

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$
- Propagation of error
 - $E(Y) = \mu_Y \cong h(\mu_1, \mu_2, \dots, \mu_n)$
 - $V(Y) = \sigma_Y^2 \cong \sum_{i=1}^n \left(\frac{\partial h}{\partial x_i}\right)^2 * \sigma_i^2$
 - Shortcut: linear combination
 - For this course, we will always assume independent random variables
- Significant figures (not in book)
 - Addition rule: least precise
 - Multiplication / Division rule: least number of sig figs

Chapter 4: Statistical Inference for One-Sample (One-population)

- Point estimates
- One-sided vs two-sided tests (one-tail versus two-tailed tests)
- Null and alternative hypotheses
- Hypothesis testing: 7 steps
 - Parameter of interest
 - Null hypothesis

- Alternative hypothesis
- Test statistic choice
- Critical region
- Calculate test statistic
- Reject / Fail to Reject null hypothesis
- (Optional) Calculate p-value for the test-statistic
- Type I error = alpha
 - Probability of rejecting the null hypothesis when null hypothesis is true
- Type II error = beta
 - Probability of failing to reject null hypothesis when null hypothesis is false
- P-value
 - If null hypothesis is true, the probability of getting a random sample whose mean (or other stat) is at least as far from the sample mean
 - Area in one tail (1-sided test) or the sum of the area in both tails (two-sided)
- Inference on mean of a single population
 - Z-test (population variance known and population is normally distributed or central limit applies)
 - T-test (population variance unknown, central limit does not apply)
 - Confidence Intervals / Bounds on mean
- Inference on variance of a single population
 - Chi-square test
 - Confidence Interval / Bounds on variance
 - Note: use this test for standard deviation, too (just do the hypothesis testing with variance)
- Inference on proportion of a single population
 - Z-test (binomial distribution)
 - Confidence Interval / Bounds for proportion
- Sample size calculations
- Beta calculations

Chapter 5: Statistical Inference for Two-Samples (Two-populations)

- Inference on difference in means
 - Z-test (both populations normal or central limit applies, population variances are known for both populations)
 - T-test, pooled estimator (population variances unknown but equal)
 - T-test, degrees of freedom (population variances unknown and not necessarily equal)
 - Confidence interval / bounds on difference in means
 - Sample sizes
 - Beta calculations
- Inference on difference in paired data
 - Paired T-test (data collected in pairs, calculate difference in pair and this becomes data item)
 - Confidence interval / bounds on mean of pair difference
- Note: we skipped inference on the ratio of variances of two normal populations (F stat)
- Inference on difference in proportions
 - Z-test
 - Sample sizes
 - Beta calculations
 - Confidence interval / bounds on proportion

- ANOVA (Analysis of Variance): Difference in at least two means from multiple populations
 - One-way (single factor, multiple treatments)
 - Two-way (two factor, multiple treatments)
 - F-statistic (excel) and p-values (excel) – how to interpret them

Chapter 6:

- Scatterplots
- Correlation, sample correlation coefficient r
- Covariance
- Empirical models
 - Linear regression (simple) – single regressor variable
 - $\beta_0 = \bar{y} - \beta_1 \bar{x}$
 - $\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$
 - Linear regression (multiple) – more than one regressor variable
 - Residuals
 - Confidence intervals on beta values (from excel)

Excel Skills:

- Produce graphs: histogram, box plot, scatterplot
- Calculate sample statistics: sample mean, sample variance, sample standard deviation
- Calculate population statistics: population mean, population variance, population standard deviation
- Calculate the 5-number summary
- Produce distributions: Binomial, Poisson, Exponential, Normal
- Z.TEST(array, x, sigma) // data, value, pop standard deviation
 - One-sample test
- T.TEST(array1, array2, 1 or 2-tailed, type)
 - Two-sample test
- Data Analysis Tools
 - ANOVA: Single Factor
 - ANOVA: Two-Factor With Replication
 - Correlation
 - Covariance
 - Histogram
 - Regression
 - T-Test: Paired Two Sample for Means (same as T.TEST – can use excel function directly)
 - T-Test: Two-Sample Assuming Equal Variances
 - T-Test: Two-Sample Assuming Unequal Variances
 - Z-Test: Two Sample for Means (same as Z.TEST – can use excel function directly)