Econ 378 Lecture Notes

Joseph McMurray

L0 Introduction

1. Opening Prayer
2. About me
    a. Raised in Salt Lake City, mission in Seoul Korea, Economics major at BYU, met my wife at BYU, PhD at University of Rochester, research in political economics (also teach Econ 477), 4 kids, sincerely believe in the Gospel of Jesus Christ and the mission of BYU.
    b. I enjoy teaching Econ 378 because the material is so useful for students, which is rewarding. It is also hard, so making it interesting and easy is a fun challenge.
3. Data analysis in Economics
    a. Scientific method: observe patterns, theorize, test theories, policy implications, policy calibration
    b. Theory: Econ 110, 380-382, 400+
    c. Evidence: Econ 378, 388, 400+, Research, internships, jobs (big data industrial transition)
    d. Economics is *both* (recommend Econ 210 for exploring careers in Economics, also MATH 213/215 linear algebra)
4. Probability and statistics
    a. Often care about population but observe only sample.
    Can't know what's true, but can know what's *probably* true
    b. Probability is the language of uncertainty
    c. Probability theory also useful in theoretical models of risk/uncertainty (e.g. insurance, investments, search, asymmetric information)
5. Syllabus
    a. Materials, participation, homework, exams, project
    b. How to choose a research topic
    c. Finish part 1 (data collection) on time! After the midterm, homework will include data analysis from your own projects.

L1 Math Preview

1. Spiritual thought: like Joseph in Egypt, your time at BYU is 7 years of plenty (spiritual abundance). Likely less so when you go to graduate school or workforce. Store up all you can now (e.g. devotionals, religion classes, student ward, ministering), like wise virgins with oil lamps, to sustain you as you "go forth to serve"
2. In a similar (but temporal) way, this lecture and HW 1 seek to fill your "math lamps" in preparation for the rest of the semester.
3. Factorials
    a. $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$
    b. $0! = 1$
4. Exponents
    a. $e \approx 2.7$ denotes growth
        i. \$1 invested at 100% interest, compound annually, equals \$2 a year later
        ii. \$1 invested at 100% interest, compound continuously, equals \$2.72 a year later

Expression	Simplified / Rewritten
x^{-1}	$1/x$
$x^{1/2}$	$\sqrt{x}$
x^0	1
x^2x^3	x^5
$(x^2)^3$	x^6
$e^x e^y$	e^{x+y}
e^x/e^y	e^{x-y}
e^{x+y}	$e^x e^y$
e^{x-y}	e^x/e^y

5. Logarithms
    a. $\text{Log}_{10}\, 100 = 2$ (How many powers of 10 give you 100?)
        i. $\log(.01) = -2$
    b. $\ln(100) \approx 4.6$ (How many powers of $e \approx 2.7$ give you 100?)

i. $\ln(.01) = -4.6$

c. Logs makes huge numbers smaller, miniscule numbers (e.g. probabilities) bigger

d. Inverse of exponents

i. $\ln(e^5) = 5$ (How many powers of e does it take to reach e^5?)

ii. $e^{\ln(5)} = 5$ (It takes $\ln(5)$ powers of e to make 5; what if we take e to that many powers?)

Expression	Simplified / Rewritten
$\ln(xy)$	$\ln(x) + \ln(y)$
$\ln\left(\frac{x}{y}\right)$	$\ln(x) - \ln(y)$
$\ln(2x)$	$\neq 2\ln(x)$ $\neq \ln(2)\ln(x)$ $= \ln(2) + \ln(x)$
$\ln(x^2)$	$2\ln(x)$
$\ln(x + y)$	Can't simplify
$\ln(x) + \ln(y)$	$\ln(xy)$
$\ln(x) - \ln(y)$	$\ln\left(\frac{x}{y}\right)$
$2\ln(x)$	$\ln(x^2)$

6. Summation

a. $\sum_{k=1}^{5} k^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 55$

b. Column of $n = 500$ observations can be denoted by x_i, with $i = 1, \dots, n$

c. $\frac{1}{n}\sum_{i=1}^{n} x_i$ denotes the average

d. $\sum_{i=1}^{n} 3x_i = 3\sum_{i=1}^{n} x_i$

e. $\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$

f. $\sum_{i=1}^{n}(x_i + 3) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} 3 = \sum_{i=1}^{n} x_i + 3n$

g. Does $\sum_{i=1}^{n}(x_i y_i) = \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$? No!

i. e.g. $2 \cdot 3 + 5 \cdot 4 \neq (2 + 5)(3 + 4)$

7. Derivatives

a. Intuition: limit of slope

b. Finding maximum/minimum

   i. First-order condition: $f'(x) = 0$

   ii. Second-order condition: $f''(x)$ negative for max (slope is decreasing, function makes a frown), positive for min (slope is increasing, function makes a smile)

c. Simple derivatives

$f(x)$	$f'(x)$
x^3	$3x^2$
$4x$	4
4	0
$\frac{1}{x}$	$-\frac{1}{x^2}$
$\sqrt{x}$	$\frac{1}{2}x^{-\frac{1}{2}} = \frac{1}{2\sqrt{x}}$
e^x	e^x
$\ln(x)$	$\frac{1}{x}$

d. Product rule: $\frac{d}{dx}[f(x)g(x)] = f'(x)g(x) + f(x)g'(x)$

   i. $\frac{d}{dx}x^2\ln(x) = 2x\ln(x) + x^2\left(\frac{1}{x}\right)$

   ii. Same pattern for products of 100 terms

e. Chain rule: $\frac{d}{dx}f\big(g(x)\big) = f'\big(g(x)\big)g'(x) = \frac{df}{dg}\frac{dg}{dx}$

   i. Example: $\frac{d}{dx}\ln(x^2 - 3x + 1) = \frac{1}{x^2-3x+1}\cdot(2x-3)$

   ii. Example: $\frac{d}{dx}e^{-3x^2} = e^{-3x^2}(-6x)$

   iii. Same pattern for longer chains

f. [The Quotient rule is useful as well, but I won't require it here.]

8. Integrals

g. Intuition

   i. "sum"/area under f (negative if $f < 0$)

   ii. Anti-derivative: add up $\int_a^b f'(x)dx$ to get $f(b) - f(a)$

$f(x)$	Anti-derivative of $f(x)$
x^2	$\frac{1}{3}x^3$
4	$4x$
$\frac{1}{x^2}$	$-\frac{1}{x}$
$\sqrt{x}$	$\frac{2}{3}x^{\frac{3}{2}}$
e^x	e^x
$x(x-1)$	$\frac{1}{3}x^3 - \frac{1}{2}x^2$

h. Definite integral $\int_4^7 x^2 dx = \left[\frac{1}{3}x^3\right]_{x=4}^{7}$

$= \frac{1}{3}(7)^3 - \frac{1}{3}(4)^3 = \frac{343}{3} - \frac{64}{3} = 93$

   i. $\int_7^4 x^2 dx = \frac{64}{3} - \frac{343}{3} = -93$

i. Useful techniques that I won't cover (or expect you to know)

   i. u-substitution (chain rule in reverse)

   ii. Integration by Parts (product rule in reverse)

j. Double Integrals

   i. Simple: inside integral then outside integral

$$\int_{y=1}^{3}\int_{x=0}^{2} x^2 y dx\, dy = \int_{y=1}^{3}\left[\frac{y}{3}x^3\right]_{x=0}^{x=2} dy = \int_{y=1}^{3}\frac{8}{3}y dy = \cdots = \frac{32}{3}$$

   ii. Note: for rectangular bounds (i.e. bounds of x don't depend on y, and vice versa), can integrate in reverse order

$$\int_{x=0}^{2}\int_{y=1}^{3} x^2 y dy\, dx = \int_{x=0}^{2}\left[\frac{1}{2}x^2y^2\right]_{y=1}^{y=3} dx = \int_{x=0}^{2} 4x^2\, dx = \cdots = \frac{32}{3}$$

   iii. Practice $\int_{y=1}^{3}\int_{x=0}^{2}\frac{x}{y}dxdy$

L2 Statistics preview, Excel

Introduction

1. We recently revised the Econ 378 curriculum. Formerly, we started with basic theory and the basic tools based on it, introduced complex theory with complex tools, then more complex theory and more complex tools. This seemed reasonable, but I realized: "Still to this day, I've never learned to build a car, but even without knowing how to build one, I managed to learn to drive one."
2. Now: learn basic, complex, and more complex tools upfront. Then go learn the underlying theory.

Spreadsheets

1. Unit of observation
2. Quantitative variables
3. Binary variables
    a. Categorical variables as binary (or "dummy") variables

Excel basics

1. Calculations
    a. Arithmetic
    b. Average, etc.
2. Formulas
    a. Example: convert GDP to per capita GDP
    b. Example: convert per capita GDP to per capita GDP change
    c. Example: convert per capita GDP change to per capita GDP % change
3. Help files
4. Transpose
5. Sort
6. Filter
7. Boolean

Data Visualization

1. Single variables
    a. Binary variables: Pie charts
    b. Quantitative variables: Histograms
2. Interactions
    a. Two binary variables: Double pie charts
    b. Binary and quantitative: bar chart
    c. Two quantitative: scatter chart
        i. Quantitative & time: line graph

    d. Three variables
        i. Two binary & quantitative: clustered bar chart
        ii. Two quantitative & binary: color-coded scatter chart
        iii. Three quantitative: bubble chart

Summary statistics

1. Proportions
2. Mean
    a. From histogram, eyeball center of gravity
3. Median/percentiles
4. Mode
5. Standard deviation
    a. Rule of thumb: two standard deviations
    b. Chebyshev's inequality: % of population outside k standard deviations can't exceed $\frac{1}{k^2}$
6. Correlation coefficient
7. Regressions
    a. Slope & intercept
        i. Predict y for any x
        ii. Predict future!
        iii. Counterfactual "experiments" (way less costly than real experiments)
    b. R^2 (coefficient of determination)
    c. Error terms / detrended data
    d. Multiple regression

Estimation

1. Population / samples
    a. Importance of representative sample
2. Point estimates, interval estimates / margin of error
3. Hypothesis test

Learning Statistics

1. We just finished semester (overview). You can now do everything by computer. Rest of semester, we'll go back and do them by hand.
2. Why work by hand? Important to know what computer is doing. (GIGO)
    a. Pushing a button works great unless a situation arises when the standard button is the wrong one to use. We need to know the limitations of the standard techniques and how to modify them appropriately.
    b. Car analogy: why insist on building cars when we could just drive them? Analogy incomplete: I can objectively verify that I've correctly driven a car; I can't objectively verify that I've correctly used statistics. In the real world of research projects,

internships, and jobs, there is no answer key! We only know our answers are correct if we know we've done them correctly, and that's only possible if we understand deeply what theoretical basis underlies the tools we're using.

3. Simple things (e.g. margin of error) mask extremely complex background. Understanding entire background is essential for confidence that we're using statistical formulas appropriately. (Sometimes, tweaks are necessary.)
4. Spiritual analogy: the "why" of the gospel. If atheist friend is kind and righteous already, why need doctrine? Even more happiness. Example: doctrine of eternal marriage informs decisions to resolve conflicts, versus divorce.

L3 Research Design

Research questions

1. There are two important ingredients to a good research study: a good question, and a good methodology for finding an answer
2. Question selection
    a. What ideas do you already have for data analysis projects?
    b. What (topic) are you excited / passionate about?
    c. If you had a crystal ball, what would you ask?
    d. What if you had a crystal ball that could answer anything but that? What would you need to ask so that you can figure out your own answer to the main question?
    e. Continue until so narrow you can collect your own data (the more specific, the better)
    f. Given (time and money) budget constraints, your project may need to settle for similar data
        i. Similar variables
        ii. A few observations
        iii. "Pilot study": this is often what is done in real world
        iv. Proof of concept (consulting sales pitch): can even use fake data
3. Data mining
    a. Given data (e.g. from a private business, a consulting client, etc.), ask, "what can we learn?" and "who is interested?"
    b. Example: private business data
    c. Typically needs to be paired with research question process above
    d. Example: what would CEO want to know?

Correlation may not mean causation!

1. Three possibilities
2. Causation: $X \Rightarrow Y$

   Theory: cell phones ⇒ distraction ⇒ accidents

   Policy implication: banning cell phones will reduce car accidents
3. Reverse causation: $X \Leftarrow Y$

   Not likely in this case (car accidents cause cell phone use?)
4. Lurking variable: $X \Leftarrow Z \Rightarrow Y$

   Example: careless (teenage?) drivers are prone independently both to use cell phones and (regardless of cell phone use) to get in accidents

   Policy implication: banning cell phones will not reduce car accidents
5. Historic instances of conflating correlation with causation
   a. The "Phillips curve" documented a negative correlation between inflation and unemployment, suggesting to policy makers that monetary policy could only avoid one problem by embracing the other. They printed more money in the 1970s, hoping to lower unemployment, but discovered "stagflation": the coincidence of high unemployment and high inflation.
   b. Documenting a positive correlation between on-the-job computer use and income, Krueger (QJE 1993) concluded that computers increase productivity, and recommended policies to increase computer use. Using similar data, however, DiNardo and Pischke (QJE 1997) showed that income is also correlated with pencil use on the job and argued (tongue-in-cheek) that subsidizing pencils would be a much more cost-effective intervention.
   c. Can also conflate lack of correlation with lack of causation: in yesterday's covid example, we derived that $P(cv|vax) = 5.3 * 10^{-5}$ and $P(cv|no\ vax) = \frac{214}{1{,}302{,}912} = 16.4 * 10^{-5}$, so vaccine is 68% effective. Correlation of cv and vax is negative, but weak. If further condition on age (<50 vs. >50):
      i. $P(cv|vax < 50) = \frac{11}{3{,}501{,}118} = .3 * 10^{-5}$

         $P(cv|no\ vax < 50) = \frac{43}{1{,}116{,}834} = 3.9 * 10^{-5}$

         Vaccine 92% effective for this group.

ii. $P(cv|vax > 50) = \frac{290}{2{,}133{,}516} = 13.6 * 10^{-5}$

$P(cv|no\ vax > 50) = \frac{171}{186{,}078} = 91.9 * 10^{-5}$

Vaccine 85% effective for this group.

iii. If condition further, vaccine efficacy by age:

Age	Vaccine efficacy	Age	Vaccine efficacy
12-15	100%	50-59	93%
16-19	100%	60-69	89%
20-29	100%	70-79	90%
30-39	97%	80-89	81%
40-49	94%	90+	92%

iv. Biggest determinant of covid is age (overall, 90+ over 1000 times more likely than 12-15 to be hospitalized with covid), not vaccine. Since people of all ages got vaccinated, corr(vax,cv) gets weaker when not conditioning in age than when conditioning on age. But even for oldest groups (where most "breakthrough" cases are occurring), vaccinated do way better than unvaccinated.

d. These examples underscore importance of careful data work, understanding statistics! Good intentions can easily be led astray by statistical subtleties.

Establishing causation (this is most of the work in economics)

1. Random experiment
    a. Best method
    b. Example: force group A to drive with cell phone, group B to not
    c. Often impractical, ethically and/or logistically (e.g. only one national economy; no control group) or even impossible (e.g. race/gender)
    d. Natural experiment: wait for nature to run experiments
        i. These are rare, might wait a long time
        ii. Government policy randomly allocates permits for some drivers to use cell phones.
        iii. Angrist and Evans (1998): Does having more children affect mother's income? Lurking variables and reverse causation both likely. But parents whose second

child was (randomly) same gender as first child were more likely to have third child, (temporarily) reduced (poor) mother's income

iv. Angrist (1990): What impact (positive or negative) did military service have on men with (randomly) high Vietnam draft numbers had 15% lower incomes years later.

v. Clever researchers keep eyes open for natural randomness, ask "what can we learn?"

vi. Sources of randomness: lottery numbers (e.g. gambling, school choice, scarce social program), random executive decisions (e.g. dorm rooms, judge assignment, advertising), weather, earthquakes, accidents, terrorist attacks

2. Second best: quasi-experiment
    a. Example: cell phones legal in one state, illegal in another
    b. Problem: other reasons for differential accidents (e.g. speed limits, enforcement, roads, recklessness?)
    c. Refinement: large number of states; before/after cell phone law change
    d. Pope (1989, BYU): Geneva Steel closed then reopened six months later, concomitant decrease then increase in local hospitalizations for pneumonia, pleurisy, bronchitis, asthma. (Landmark study in air pollution.)
    e. Sargeant et al. (2004): Restaurant smoking ban in Montana, repealed after six months. Heart attacks dropped 40%, then went back up.
    f. Lee et al. (2004): How does politician (Democrat/Republican) affect policy outcomes? Random election? No. But in close elections (e.g. 48-52% votes), winning or losing was plausibly random.
    g. Possible sources of quasi-randomness: cutoffs (e.g. grades, income thresholds, performance thresholds, birth date), bureaucratic decisions that are not literally random but seem arbitrary (e.g. regulatory decisions, tax levels, regularly/tax hike timing, pre-/post-construction project, mission assignment)
3. Controls
    a. Compare sub-populations to "control" for lurking variables
    b. Most common method (since others infeasible)
    c. Example: compare cell phone use and accidents among teenagers/adults

vii. Other proxies for recklessness: grades? debt? Often imperfect

d. Econ 378: restrict sample (Econ 388: regressions)

Prediction

1. If correlation does not reflect causation, X cannot be used to control Y, but still can be used to predict Y
    a. Example: reduced car insurance premiums for good grades, females, good driving history, yellow cars
    b. Ethics of "statistical discrimination" (e.g. traffic stops for blacks, airport scrutiny for Arabs)
    c. Role of theory is to posit reasons for correlation; essential if anything changes (e.g. cell phones get cheaper).

Research Design for Causal Inference

1. Many of the topics we're interested in seek to establish cause/effect relationships.
    a. What examples did you come up with? (e.g. Do masks reduce covid spread?)
    b. Were there any topics you came up with that do not involve cause/effect relationships? (Probably not.)
2. What is a cause/effect relationship you would like to discover?
3. Which variables might have a simple correlation that suggests the relationship above?
4. Are there any competing forces that might produce the opposite correlation? If the correlation turns out to be consistent with a hypothesized cause/effect relationship, the hypothesized relationship might outweigh any competing forces.
5. But are there other mechanisms that could produce the same correlation? If so, finding a correlation where you expected it does not guarantee that the hypothesized cause/effect relationship is valid.
6. This raises a new question: where could we look for evidence of the hypothesized cause/effect relationship that would not pick up correlations for these other reasons?
7. This is the key question of research design. Note that you can (and should!) think through and plan out your response to these issues before you ever look at the data.

Research design in the quest for spiritual truth

1. A friend, skeptical of spiritual things, recommended the following experiment: go to a hospital, pray for people in every other room. See if they recover more quickly/fully than the others. (His prediction: no.) Is this a valid statistical test of the validity of prayer? Why or why not?
2. Research design is important in answering spiritual questions, too:
    a. What do the scriptures say about experiments to uncover spiritual truth?
    b. "If any of you lack wisdom, let him ask of God, who giveth to all men liberally, and upbraideth not; and it shall be given him. *But let him ask in faith*, nothing wavering." (James 1:5-6, emphasis added)
    c. "And when ye shall receive these things, I would exhort you that ye would ask God, the Eternal Father, in the name of Christ, if these things are not true; and *if ye shall ask with a sincere heart, with real intent, having faith in Christ*, he will manifest the truth of it unto you, by the power of the Holy Ghost." (Moroni 10:4, emphasis added)
    d. "Now, we will compare the word unto a seed. Now, *if ye give place, that a seed may be planted in your heart*, behold, if it be a true seed, or a good seed, *if ye do not cast it out by your unbelief, that ye will resist the Spirit of the Lord*, behold, it will begin to swell within your breasts; and when you feel these swelling motions, ye will begin to say within yourselves—It must needs be that this is a good seed, or that the word is good, for it beginneth to enlarge my soul; yea, it beginneth to enlighten my understanding, yea, it beginneth to be delicious to me." (Alma 32:28, emphasis added)
    e. To me, asking in faith means an honest willingness to follow the promptings received. If I don't honestly intend to follow impressions that are given, the experiment is void.
    f. "If any man will *do his will*, he shall know of the doctrine, whether it be of God, or whether I speak of myself." (John 7:17, emphasis added)
    g. Mission friend: finally prayed about the Book of Mormon but "nothing happened". Zone leader: real intent might mean praying more than once. After continued prayer, he received confirming witness.

L4 Stata: Basics

Stata

1. Introduction

a. Stata especially well-suited to economics (regression analysis).
b. But expensive.
c. Other stats packages are available (e.g. R, SAS, SPSS), can program in Python, C, Matlab, Java). But learning new stats package or programming language is like using a Casio calculator when you're familiar with TI—all the buttons do all the same things, they're just located in different places. So learning one language helps you pick up others quickly, as needed. (Stata has specific value for Economics research assistants, future PhDs.)

2. Basic user interface
    a. Open Stata. Very different GUI (graphical user interface) than Excel, but same idea. Feels more like computer programming software, for good reason.
    b. Click on Data Editor (Edit), enter data by hand: numbers 1-10 in column 1, make up ten values in column 2.
        i. Row entries are called observations (numbered automatically on left side)
        ii. Column entries are called variables, default named "var1" and "var2"
    c. Close data editor. Notice in main screen, a running list of individual commands (and their results), based on what we just did manually.
        i. With our first action, we actually implemented three commands: set "obs" from 0 to 1, generate variable "var1", and set the value of var1 for observation 1.
        ii. As we added more data, we set "obs" to 2, 3, …, 10, replaced var1 for observations 2, 3, …, 10, and then generated a second variable, "var2", and replaced its values for observations 1-10. (Note, we no longer needed to expand "obs" at that point.)
    d. Command line
        i. We can add additional commands using the command line. Type: gen<u></u>erate newvar
        ii. Note: in Stata commands, shortcuts are indicated by underlining the minimal number of letters to convey the same meaning: "gen" or "gene" or "gener" or "genera" or "generat" or "generate" are all equivalent, but "ge" is ambiguous, and will therefore not work.
        iii. Error: when we generate a new variable, we need to define its values. Let's try again: generate varthree=3

iv. Now open Data Editor again to see the result. We have defined a new variable "varthree" to equal 3, not just for a particular observation, but for every observation. This is the power of the programming approach to data: you can do lots of things at once! (In Excel, there are shortcuts to copy and fill, but nothing this quick.) Most of the time, we won't interact with the spreadsheet directly; we'll just be programming.

v. To see this again, close Data Editor and type: replace varthree=2*var2 if var1>7. We see that three real changes were made, and if we open the Data Editor again, we can see what that looks like.

vi. What if we meant to type varthree=2*var2 if var1>5? We could type over again, or we can push "PgUp" to repeat the previous command (and edit it before hitting return). Pushing this multiple times allows us to repeat earlier commands.

vii. We can also use the command line as a calculator: type display sqrt(8*2)

e. Left and right panels

i. On top right-hand side is a list of the variables we have so far: var1, var2, and newvar.

ii. The bottom right-hand side summarizes our data: we have three variables and 10 observations.

iii. (If you ever need to, you can resize these panels by dragging their borders.)

iv. Notice on left-hand side is list of commands we've used so far (red for the ones that returned error codes). Let's highlight the first 26 lines (click on the first, then shift-click on line 26) and push Ctrl+C to copy these. (This creates var1 and var2 and replaces all 10 values of var1 but only the first 5 values of var2.)

3. Do Files

a. Click on "New Do File Editor" at the top of the screen (looks like a Word document, with a pencil). This opens a new window, with a text editor.

b. Type Ctrl+V to paste the command lines that we copied earlier. This becomes a short computer program that, once we execute, will create two variables and replace their values. (In computer programming, a "command" tells the computer to do something. A "script" is a list of commands, to be executed in order. Scripts in Stata are called Do files.)

c. Note: some programming languages require a semicolon or other punctuation to denote the end of one command and the beginning of the next. "Return" plays that role in Stata, so writing on separate commands on separate lines is sufficient.

d. Always use a script, not the command line!

    i. Often, after many steps of manipulating data, you realize that you should have done step 3 differently. In Excel, you might be able to hit "undo" repeatedly (if you haven't saved and closed the program yet), but once you correct step 3, you'll then have to redo all subsequent steps. With a script, you can edit step 3 and recompile in moments.

    ii. A script is also useful for collaborators and replicators, as well as yourself when you come back to a project after a long pause. I learned this the hard way: I downloaded data, manipulated it repeatedly using the command line, found results that were really interesting, and copied and pasted them into my research paper. A more urgent project came up, so it was months before I came back to this. When I came back, I did (what I thought were) the same manipulations but got totally different results. I couldn't figure out how to reproduce the results that I had recorded earlier. Maybe I had been making a mistake? Maybe I'm making a mistake now, but was previously correct? If I had a paper trail of all my data manipulations at the time, I could compare my work now and then to see how they differ, and which is more reliable. …But I don't, and may never publish that paper.

    iii. For your data project, you will need to submit your Do file, not just your results.

    iv. Related programming tip: Never overwrite your original data set. Use a script to open your original data set, make whatever edits need to made, and then save the revised data as a new data set with a different file name. That way, every time you run your script, it pulls the same original data.

4. .dta files

    a. Computer programs are designed to recognize and interpret information, but only in specific formats, specified by file type (e.g. .pdf, .docx, .mp3).

    b. For example, Spreadsheets often saved as .csv ("comma separated values")

        i. year,growth;2021,2%;2020,1%;2019,1.5% can be understood to be a 3x2 table

    c. Excel recognizes .csv or .xlsx, which adds formatting information.

d. Stata stores data using .dta and stores scripts using .do

Topics:

- Help
- Lookfor
- Resize
- Break
- Pwd/cd
- Display
- Ctl+Shft+Right click -> “copy as path”

File types

Variable types

Scripts

Saving data

L5 Stata: Summarizing Data

Topics:

Script tips

Describe

Tabulate (+twoway)

Summarize

Destring/Tostring

Comments

Generate

Replace

If

L6 Stata: Data Visualization

Bysort

Histogram

Help

Scatter

Geodist

Keep/Drop

Collapse

Line

Import

Net

Twoway line

Graph options

Regress

Predict

L7 Probability, Combinatorics (WMS 2.1-6)

1. How many have already collected data? How many know what data to collect?
2. Spiritual thought: many of our greatest priorities are not time sensitive, so it's easy to procrastinate till "later", but often this postpones blessings. Let your daily activities match your eternal priorities! (e.g. Don't delay repentance, marriage/children, promptings, family history, making life plans, for movies/hobbies or even school/work)
3. Probability
    a. Language of uncertainty
    b. Economic applications: risk/insurance, investments, shopping/learning
    c. Data analysis (huge): infer population characteristics from sample
    d. Fundamentally, probability is merely ratio
    e. Sample space / Universe $S = \{1,2,3,4,5,6\}$

f. Subset / Event
    i. $B = \{4,5,6\}$
    ii. $E = \{2,4,6\}$
g. Probability function $P(E) = \frac{n_E}{n_S} = \frac{3}{6}$
    i. A function is like a machine; in this case, output is number but input is set
    ii. $P(S) = 1$
h. Categorical descriptions: 50/500 unemployed = 10% unemployment rate / probability

4. Set Notation
    a. Element
        i. $5 \in B, 5 \notin E$
    b. Subset
        i. $B \subseteq S$
    c. Empty set ∅
    d. Complement
        i. $\bar{B} = \{1,2,3\}$
        ii. $\bar{E} = \{1,3,5\}$
        iii. $P(\bar{E}) = 1 - P(\bar{E})$
        iv. Note: this is a simple but useful tool. Sometimes it's easier to derive $P(\bar{E})$ than to derive $P(E)$. For example: the probability that two or more students in Econ 378 share birthdays is very difficult to find directly, but the complementary event (i.e. that no two students share a birthday) is not as bad.
    e. Intersection ("and")
        i. $B \cap E = \{4,6\}$
        ii. $\bar{B} \cap \bar{E} = \{1,3\}$
    f. Union ("or", "at least one")
        i. $B \cup E = \{2,4,5,6\}$
        ii. $\bar{B} \cup \bar{E} = \{1,2,3,5\}$
    g. Mutually exclusive
        i. $A \cap B = \emptyset$
        ii. $P(A \cap B) = 0$
    h. Collectively exhaustive
        i. $A \cup B = S$

ii. $P(A \cup B) = 1$

5. Total probability: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
    a. Including both $P(A)$ and $P(B)$ "double counts" $P(A \cap B)$
    b. Example: among set S of workers in particular industry, unemployment rate $P(U) = .10$, women $P(W) = .25$, intersection $P(U \cap W) = .05$; $P(U \cup W) = .1 + .25 - .05 = .3$
6. Independence: $P(A \cap B) = P(A)P(B)$
    a. Example: $P(U)P(W) = (.10)(.25) = .025 \neq .05 = P(U \cap W)$
    b. Not the same as mutually exclusive! (Mutually exclusive events are highly negatively correlated)
7. Combinatorics
    a. "mn rule"
        i. 6 pants (2 brown), 10 shirts (3 green); probability that random outfit consists of brown pants and green shirt is $P(B \cap G) = \frac{\#\{B \cap G\ outfits\}}{\#\{outfits\}} = \frac{2 \cdot 3}{6 \cdot 10} = .1$
        ii. Equivalently, since independent, $P(B \cap G) = P(B)P(G) = \frac{2}{6} \cdot \frac{3}{10} = .1$
    b. Number of ways to permute (i.e. order) 10 students: $10! \approx 3.6\ million$
    c. Number of ways to choose 3 out of 10 students: $C_3^{10} = \binom{10}{3} = \frac{10!}{3!7!} = 120$
        i. Numerator: there are $10!$ ways to order the 10 students
        ii. Denominator: but this double-counts ($3!\ 7!$ times) orderings which shuffle the first three and last seven, but don't change the identity of the chosen three
    d. Number of ways to assign 10 students into bins of 3, 5, 2: $\frac{10!}{3!5!2!} = 2{,}520$
8. Applications: discrimination lawsuit after 9 workers (3 immigrants, 6 natives) assigned to 4 dangerous jobs + 5 safe jobs
    a. All 3 immigrants assigned to dangerous jobs
        i. $P(E) = \frac{n_E}{n_S} = \frac{C_3^3 \times C_1^6}{C_4^9} = \frac{\left(\frac{3!}{3!0!}\right)\left(\frac{6!}{1!5!}\right)}{\frac{9!}{4!5!}} = \frac{6}{\left(\frac{9 \cdot 8 \cdot 7 \cdot 6}{4 \cdot 3 \cdot 2 \cdot 1}\right)} = \frac{1}{21} \approx 0.05$
        ii. Alternatively, can think sequentially: $\frac{4}{9} \cdot \frac{3}{8} \cdot \frac{2}{7} = \frac{1}{21}$
        iii. Alternatively, can assign workers to safe jobs: $P(E) = \frac{n_E}{n_S} = \frac{C_0^3 \times C_5^6}{C_5^9} \approx 0.05$
        iv. Alternatively, can assign jobs to workers: $P(E) = \frac{C_3^4 C_0^5}{C_3^9} \approx 0.05$
    b. 2 out of 3 immigrants assigned to dangerous jobs

i. $P(2) = \frac{n_E}{n_S} = \frac{(C_2^3 \times C_2^6)}{C_4^9} = \frac{\left(\frac{3!}{2!1!}\right)\left(\frac{6!}{2!4!}\right)}{\frac{9!}{4!5!}} = \frac{3\left(\frac{6\cdot5}{2\cdot1}\right)}{\left(\frac{9\cdot8\cdot7\cdot6}{4\cdot3\cdot2\cdot1}\right)} = \frac{5}{14} \approx 0.36$

ii. $P(E) = P(2) + P(3) \approx 0.36 + 0.05 = 0.41$

c. 24 workers assigned to 10 safe and 14 dangerous jobs, lawsuit because 6 immigrants all assigned dangerous job

i. $P(E_1) = \frac{n_E}{n_S} = \frac{C_8^{18} C_6^6}{C_{14}^{24}} = \frac{\frac{18!\ \ 6!}{8!10!6!0!}}{\frac{24!}{14!10!}} = \frac{18!14!}{8!24!} \approx .022$

ii. $\frac{C_6^{14} C_0^{10}}{C_6^{24}} = \frac{\frac{14!\ \ 10!}{6!8!0!10!}}{\frac{24!}{6!18!}} = \frac{14!18!}{8!24!} \approx .022$

iii. If 5 out of 6 assigned dangerous job: $P(E_2) = \frac{C_9^{18} C_5^6}{C_{14}^{24}} = \frac{\frac{18!\ \ 6!}{9!9!5!1!}}{\frac{24!}{14!10!}} = \frac{18!14!10\cdot6}{24!9!} \approx .149$,

$P(E) \approx .022 + .149 = .171$

L8 Conditional Probability (WMS 2.7-10)

1. If possible, be prepared next lecture with idea for research project
2. Typically, don't count to determine $\Pr(E)$; estimate from sample

Conditional probability

1. Definition: $\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$
2. This is how online stores (e.g. Ebay, Amazon, Google) figure out what to advertise: given that you purchased a textbook, how likely are you to want a Lego set or motorcycle helmet?
3. Story problem keywords: "given", "conditional on", "among", or "out of"
4. Example 1: Among set S of workers in particular industry, unemployment rate $P(U) = .10$, women $P(W) = .25$, intersection $P(U \cap W) = .05$

d. Rectangular Venn diagram

e. Unemployment rate among women $P(U|W) = \frac{.05}{.25} = .20$

f. Fraction of unemployed who are women $P(W|U) = \frac{.05}{.10} = .50$

g. Practice:

i. Unemployment rate among men $P(U|\bar{W}) = \frac{.05}{.75} = \frac{1}{15} \approx .07$

ii. Fraction of unemployed who are men $P(\bar{W}|U) = \frac{.05}{.10} = .50 = 1 - P(W|U)$

Independence

1. Definition: $P(A|B) = P(A)$, $P(B|A) = P(B)$ (equivalent to $P(A \cap B) = P(A)P(B)$)
2. What is the probability of a person being unemployed? $P(A) = .10$; what if it's raining outside? Then the probability of being unemployed is $P(A|B) = .10$.
3. Surgeon joke (failing to account for independence): the bad news is that this type of surgery is successful only 25% of the time. The good news is that the last three patients all died.

Event decomposition:

1. If $E_1, \dots, E_k$ are mutually exclusive and collectively exhaustive then $P(A) = \sum_{k=1}^{n} P(A \cap E_k)$
2. Example 1: 30% of web traffic comes from a Google add (G), 30% from online newspaper (N), and 40% from a product reviewer's blog (R). 40% of Google traffic, 20% of newspaper traffic, and 30% of reviewer traffic end in a sale (S). What fraction of overall traffic ends in a purchase?
    a. Step 1: draw event tree (first web source, then purchase decision)
    b. Step 2: translate question into notation. $P(G) = .3$, $P(N) = .3$, $P(T) = .4$, $P(S|G) = .4$, $P(S|N) = .2$, $P(S|R) = .3$, wish to find $P(S)$
    c. $P(S) = P(S \cap G) + P(S \cap N) + P(S \cap R)$
    $= P(G)P(S|G) + P(N)P(S|N) + P(R)P(S|R) = .3 \times .4 + .3 \times .2 + .4 \times .3 = .12 + .06 + .12 = .3$
    d. S and R are independent, since $P(S|R) = P(S) = .3$. Is S independent of G? Of N?
3. Bayes' Rule
    a. $P(A \cap B) = \begin{cases} P(A|B)P(B) \\ P(B|A)P(A) \end{cases}$
    b. Therefore, can derive $P(A|B)$ from $P(B|A)$, or vice versa.
    c. $P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A)+P\left(B\middle|\bar{A}\right)P(\bar{A})}$
    d. Practice: find and interpret $P(G|S)$, $P(R|S)$, $P(N|S) = \frac{P(N \cap S)}{P(S)} = \frac{.06}{.3} = .2$ (mere coincidence that $P(N|S) = P(S|N)$)
4. Warning: think carefully about difference between $P(A|B)$, $P(A)$, and $P(B|A)$. Be sure you know which you really want.
5. Note: It's possible for composite probabilities and conditional probabilities to tell rather opposite stories

a. Charig et al. (1986): Kidney stone treatment B looked more effective, but A was more actually effective more effective both with small stones and large stones (but stone size matters, and treatments A and B had been used disproportionately on large and small stones, respectively)

Kidney stone size	Treatment A	Treatment B
Small	81/87=**93%**	234/270=87%
Large	192/263=**73%**	55/80=69%
Both	273/350=78%	289/350=**83%**

b. MLB batting averages: David Justice was better in 1995 and 1996 but Derek Jeter was better in 1995-96. Who is better batter?

Batter	1995	1996	1995-96
Derek Jeter	12/48=.250	183/582=.314	195/630=**.310**
David Justice	104/411=**.253**	45/140=**.321**	149/551=.270

Either could be. Likely depends on which is more predictive of 1997 (depends on other assumptions)

c. Israel covid data: August 2021 (https://www.covid-datascience.com/post/israeli-data-how-can-efficacy-vs-severe-disease-be-strong-when-60-of-hospitalized-are-vaccinated)
   i. When covid Delta variant hit, Israeli hospitals filled up with covid cases: 214 that were unvaccinated and 301 that were vaccinated. Since 60% were vaccinated, superficial conclusion is that vaccines make covid worse, not better!
   ii. But 60%=P(vax|cv). We really want to know P(cv|vax) (actually, want to compare P(cv|vax) and P(cv|no vax))
   iii. $P(cv|vax) = 301/5{,}634{,}634 = 5.3 * 10^{-5}$

   $P(cv|no\ vax) = \frac{214}{1{,}302{,}912} = 16.4 * 10^{-5}$

   Vaccinated only catch covid $\frac{5.3}{16.4} = 32\%$ as often (i.e. vaccine 68% effective)
   iv. Nearly 80% of Israelis over age 12 were vaccinated against covid, so if it were unrelated random draw, 80% of covid patients should have been vaccinated; lower rate than 80% supports hypothesis that treatment helped.

v. Put differently, so many Israelis were vaccinated that even though those vaccinated only got covid 68% as often, there were more vaccinated covid cases than unvaccinated covid cases.

L9 Probability Distributions (WMS 3.1-3)

1. Events are useful for describing binary/categorical information. Next, we'll consider random variables, which are useful for describing quantitative information.
2. Random variables
   a. Distribution Function: Number of cars X owned by a family $P(x) = P(X = x) =$

$$\begin{cases} .10 & if\ x = 0 \\ .30 & if\ x = 1 \\ .40 & if\ x = 2 \\ .20 & if\ x = 3 \\ 0 & else \end{cases}$$

   b. Mean (i.e. average) "mu" μ
      i. We don't know total population size. If we knew there were 100 families, $\mu = \frac{1}{100}(0 \cdot 10 + 1 \cdot 30 + 2 \cdot 40 + 3 \cdot 20) = 1.7$. If population size n is unknown then $\mu = \frac{1}{n}[0(.10n) + 1(.30n) + 2(.40n) + 3(.20n)]$ but this reduces to …
      ii. Expected value (or "expectation") $\mu = E(X) = \sum_x xP(x) = 0(.10) + 1(.30) + 2(.40) + 3(.20) = 1.7$
         1. Note: if all realizations of X are equally likely then $P(x) = \frac{1}{n}$ for all x so $E(X) = \sum_x x\frac{1}{n} = \frac{1}{n}\sum_x x$ reduces to familiar formula
   c. Expected value of functions of X
      i. Example: expected utility $E[u(X)] = E(\sqrt{X})$
      ii. Formula: $E[f(x)] = \sum_x f(x)P(x)$
      iii. Example: $E(X^2) = 0^2(.1) + 1^2(.3) + 2^2(.4) + 3^2(.2) = 3.7$
      iv. Algebra shortcuts
         1. Linear functions, e.g. car maintenance cost $C = 200 + 100X$; average maintenance cost $E(C) = 200(.1) + 300(.3) + 400(.4) + 500(.2) = 370$
         2. Shortcuts: $E(200 + 100X) = E(200) + E(100X)$

$$= 200 + 100E(X) = 200 + 100(1.7) = 370$$

a. Summations: $E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)$
b. Pull out coefficients
c. For constant c, $E(c) = c$

d. Variance, standard deviation
   i. Variance $\sigma^2 = V(X) = E[(X-\mu)^2] = [(0-1.7)^2](.1) + [(1-1.7)^2](.3) + [(2-1.7)^2](.4) + [(3-1.7)^2](.2) = .81$
   ii. Standard deviation $\sigma = \sqrt{\sigma^2} = \sqrt{.81} = .9$
      1. Interpretation: by rule of thumb, "most" families own between -.1 and 3.5 cars
      2. Note: variance, by itself, is difficult to interpret (e.g. units is "cars squared"), but is easier to work with algebraically, because it's technically and average of something, whereas standard deviation is the square root of an average of something.
   iii. Simpler formula: $V(X) = E(X^2) - \mu^2 = 3.7 - (1.7)^2 = .81$
      1. Show equivalent, as homework problem
      2. Remember this formula, as we'll use it repeatedly
   iv. Algebra shortcuts
      1. $E(C^2) = 200^2(.1) + 300^2(.3) + 400^2(.4) + 500^2(.2) = 145{,}000$ doesn't have any easy algebra shortcut; $V(C) = E(C^2) - 370^2 = 8{,}100$
      2. Shortcut: $V(C) = V(200 + 100X) = V(100X) = 100^2 V(X) = 8{,}100$
         a. 200 gets added and subtracted: $V(C) = E\{[(200 + 100X) - E(200 + 100X)]^2\}$
         b. For constant c, $V(c) = 0$
         c. Pull out coefficients, ... but square them! (because Variance is a quadratic function of a random variable)

3. Practice example: number Y of car accidents $P(Y = y) = \begin{cases} .7 & if\ y = 0 \\ .2 & if\ y = 1 \\ .1 & if\ y = 2 \\ 0 & else \end{cases}$
   a. $\mu = 0(.7) + 1(.2) + 2(.1) = .4$
   b. $E(Y^2) = 0^2(.7) + 1^2(.2) + 2^2(.1) = .6$
   c. $V(Y) = E(Y^2) - \mu^2 = .6 - (.4)^2 = .44$

d. $\sigma = \sqrt{.44} \approx .663$

e. Insurance profit $\Pi = \$1200 - \$2000 \cdot Y$

i. $\mathrm{E}(\Pi) = \mathrm{E}(1200 - 2000 \cdot \mathrm{Y}) = 1200 - 2000E(Y) = 1200 - 2000(.4) = \$400$

ii. $V(\Pi) = \mathrm{V}(1200 - 2000\mathrm{Y}) = 0 + (-2000)^2 V(Y) = 4{,}000{,}000(.44) = 1{,}760{,}000$

iii. $\sigma_\Pi = \sqrt{1{,}760{,}000} = \$1{,}326$

f. Review this one more time before attempting your homework!

L10 Correlation (WMS 3.1-8)

[Long lecture; use time efficiently]

1. Correlation coefficient $\rho \in [-1,1]$
   a. Positive correlation means variables X and Y tend to move in same direction (e.g. temperature X and ice cream sales Y)
   b. Negative correlation means variables X and Y tend to move in opposite direction (e.g. frequency of exercise X and body mass index Y)
   c. Magnitude indicates strength of relationship
   d. Independence implies $\rho = 0$
2. Joint distribution function
   a. Employee hours per week X and hourly wage Y

	$Y = 10$	$Y = 15$
$X = 20$	.2	.1
$X = 30$	.1	.2
$X = 40$	.1	.3

Illustrate: $P(x,y) = P(X = x \cap Y = y) = \begin{cases} .2 & if\ (x,y) = (20,10) \\ .1 & if\ (x,y) = (20,15) \\ .1 & if\ (x,y) = (30,10) \\ .2 & if\ (x,y) = (30,15) \\ .1 & if\ (x,y) = (40,10) \\ .3 & if\ (x,y) = (40,15) \end{cases}$

3. Marginal distributions

a. Sum rows or columns: $P(x) = \begin{cases} .3 & if\ x = 20 \\ .3 & if\ x = 30 \\ .4 & if\ x = 40 \end{cases}$ and $P(y) = \begin{cases} .4 & if\ y = 10 \\ .6 & if\ y = 15 \end{cases}$

b. Summary statistics (quick recap)

i. $\mu_x = 20(.3) + 30(.3) + 40(.4) = 31$

ii. $\sigma_x \approx 8.3$

1. $E(X^2) = 20^2(.3) + 30^2(.3) + 40^2(.4) = 1{,}030$
2. $\sigma_x^2 = 1{,}030 - 31^2 = 69$
3. $\sigma_x = \sqrt{69} \approx 8.3$

iii. $\mu_y = 10(.4) + 15(.6) = 13$

iv. $\sigma_y \approx 2.4$

1. $E(Y^2) = 10^2(.4) + 15^2(.6) = 175$
2. $\sigma_y^2 = 175 - 13^2 = 6$
3. $\sigma_y = \sqrt{6} \approx 2.4$

c. Independence

i. Definition: $P(x, y) = P_x(x)P_y(y)$ for every (x, y) pair

ii. Not independent here, since $P(20{,}10) = .20 \neq P(20)P(10) = .3 \times .4 = .12$

4. Expectations of functions of X and Y

a. Average weekly payment $E(XY) = (20 \cdot 10)(.20) + (20 \cdot 15)(.10) + (30 \cdot 10)(.10) + (30 \cdot 15)(.20) + (40 \cdot 10)(.10) + (40 \cdot 15)(.30) = 40 + 30 + 30 + 90 + 40 + 180 = 410$

b. Can do any function $E[f(X, Y)] = \sum_{(x,y)} f(x, y)P(x, y)$

5. Correlation

a. Covariance

$$\sigma_{xy} = E\big[(X - \mu_x)(Y - \mu_y)\big]$$
$$= [(20 - 31)(10 - 13)](.20)$$
$$+[(20 - 31)(15 - 13)](.10)$$
$$+[(30 - 31)(10 - 13)](.10)$$
$$+[(30 - 31)(10 - 15)](.20)$$
$$+[(40 - 31)(10 - 13)](.10)$$
$$+[(40 - 31)(15 - 13)](.30) = 7$$

b. Simpler formula: $\sigma_{xy} = E(XY) - \mu_x\mu_y = 410 - (31)(13) = 7$

c. Correlation $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{7}{(8.3)(2.4)} \approx 0.35$

i. Positive, but fairly weak (again, not independent)

ii. Later: $\rho^2 \approx 0.12$ measures % of variation in Y that covaries with X

6. Algebra shortcuts

a. Covariance of a sum

$Cov(X, 1200 - 2000Y) = Cov(X, 1200) + Cov(X, -2000Y) = 0 + (-2000)\sigma_{xy}$

b. Correlation of a sum

$Corr(X, 1200 - 2000Y) = -\rho$

c. Variance of a sum

$V(X + Y) = V(X) + V(Y) + 2Cov(X, Y)$

d. Variance of a larger sum

$V(X + Y + Z) = V(X) + V(Y) + V(Z) + 2Cov(X, Y) + 2Cov(X, Z) + 2Cov(Y, Z)$

(with 100 variables, $C_2^{100} \approx 5000\ Cov$ terms)

i. Importance of independent samples

7. Application: financial diversification

a. Assume two stocks have same average return $\mu_x = \mu_y = \mu$ and same standard deviation $\sigma_x = \sigma_y = \sigma$.

b. You could buy two shares of X, or one share of X and one share of Y. Since you are indifferent between X and Y, it might seem that you should be indifferent between these two stock portfolios.

c. However, on your homework, you will prove that $E(2X) = E(X + Y)$ but $V(2X) < V(X + Y)$, as long as X and Y are not perfectly correlated (i.e. $\rho < 1$). In fact, if they are perfectly *negatively* correlated then $V(X + Y) = 0$!

d. Thus, the common financial advice to "Diversify your portfolio".

8. Practice [if time]: Cell phone use X and number Y of car accidents

	$Y = 0$	$Y = 1$	$Y = 2$
$X = 0$	.60	.08	.02
$X = 1$	.10	.12	.08

a. Note: numerical values can be assigned to binary categorical variables, so that notion of correlation is still meaningful.

b. Marginal distribution $P(X=x)=\begin{cases}.70 & if\ x=0\\ .30 & if\ x=1\end{cases}$, mean $E(X)=.3$, variance $\sigma_x^2=E(X^2)-\mu_x^2=.3-.3^2=.21$, standard deviation $\sigma_x=\sqrt{.21}\approx 0.458$

c. Marginal distribution $P(Y=y)=\begin{cases}.70 & if\ y=0\\ .20 & if\ y=1\\ .10 & if\ y=2\end{cases}$, mean $E(Y)=.4$, variance $\sigma_y^2=.44$, standard deviation $\sigma_y=\sqrt{.44}\approx 0.663$

d. Not independent since $P(0,0)=.6\neq P_x(0)P_y(0)=(.7)(.7)=.49$

e. $E(XY)=(0)(0)(.60)+(0)(1)(.08)+(0)(2)(.02)+(1)(0)(.10)+(1)(1)(.12)+(1)(2)(.08)=.28$

f. Covariance $\sigma_{xy}=E(XY)-\mu_x\mu_y=.28-(.3)(.4)=.16$

g. Correlation $\rho=\frac{\sigma_{xy}}{\sigma_x\sigma_y}=\frac{.16}{(.458)(.663)}\approx 0.527$ positive and moderately strong

L11 Continuous Distributions (WMS 4.1-3)

1. Continuous random variables
    a. Infinite domain, e.g. sleep hours $x\in[6,9]$
    b. Philosophical view: continuous functions conveniently approximate discrete world, or world is truly infinite but measurement is imprecise
2. Probability density function (pdf) $f(x)$
    a. Measures relative likelihood of individual x values
    b. Individual x values occur with zero probability (and $f(x)>1$ is possible); to find probabilities, must take definite integral $P(7<X<8)=\int_7^8 f(x)dx$
    c. Density must be non-negative and integrate to one over domain (just like probabilities sum to one)
    d. Example $f(x)=k(-x^2+16x-60);\ 6\leq x\leq 9$
        i. Not directly from (finite) data; maybe from calibrated theory
        ii. Find k
            1. $1=\int_6^9 f(x)dx=k\left[-\frac{1}{3}x^3+8x^2-60x\right]_6^9=k[(-243+648-540)-(-72+288-360)]=9k$ requires that $k=\frac{1}{9}$
            2. That is, $f(x)=-\frac{1}{9}x^2+\frac{16}{9}x-\frac{60}{9};\ 6\leq x\leq 9$

iii. Mode solves $f'(x) = -\frac{2}{9}kx + \frac{16}{9}k = 0$; solution at $x = 8$

1. Note: if $f'(x)$ everywhere positive/negative then maximum is at highest/lowest x in range
2. Note: second-order condition $f''(x) = -\frac{2}{9}k \leq 0$ satisfied as long as $k \geq 0$

iv. Probabilities: $P(7 \leq x \leq 8) = \int_7^8 \frac{1}{9}(-x^2 + 16x - 60)dx = \cdots = \frac{11}{27} \approx 0.4$

3. Cumulative distribution function (cdf) $F(x)$

a. $F(x) = P(X \leq x) = \int_6^x \frac{1}{9}(-\tilde{x}^2 + 16\tilde{x} - 60)d\tilde{x}$

$= \left[-\frac{1}{27}\tilde{x}^3 + \frac{8}{9}\tilde{x}^2 - \frac{20}{3}\tilde{x}\right]_{\tilde{x}=6}^{\tilde{x}=x} = -\frac{1}{27}x^3 + \frac{8}{9}x^2 - \frac{20}{3}x + 16$

(This assumes $6 \leq x \leq 9$; if $x < 6$ then $F(x) = 0$ and if $x > 9$ then $F(x) = 1$)

b. Percentiles

Median $F(x) = -\frac{1}{27}x^3 + \frac{8}{9}x^2 - \frac{20}{3}x + 16 = \frac{1}{2}$; solving by computer, $x \approx 7.8$

75th percentile $F(x) = -\frac{1}{27}x^3 + \frac{8}{9}x^2 - \frac{20}{3}x + 16 = .75 \Rightarrow x \approx 8.4$

90th percentile $F(x) = -\frac{1}{27}x^3 + \frac{8}{9}x^2 - \frac{20}{3}x + 16 = .90 \Rightarrow x \approx 8.7$

c. Easier probabilities $P(7 \leq X \leq 8) = F(8) - F(7)$

$= \left(-\frac{1}{27}8^3 + \frac{8}{9}8^2 - \frac{20}{3}8 + 16\right) - \left(-\frac{1}{27}7^3 + \frac{8}{9}7^2 - \frac{20}{3}7 + 16\right) = \frac{11}{27} \approx 0.4$

d. From cdf, get pdf $f(x) = F'(x) = -\frac{1}{9}x^2 + \frac{16}{9}x - \frac{60}{9}$; $6 \leq x \leq 9$, else $f(x) = 0$

4. Moments

a. Mean

$\mu = E(X) = \int xf(x)dx$ (just like $E(X) = \sum xP(x)$)

$= \int_6^9 x\frac{1}{9}(-x^2 + 16x - 60)dx$

$= \int_6^9 \frac{1}{9}(-x^3 + 16x^2 - 60x)dx = \cdots = \frac{31}{4} \approx 7.75$

b. Standard deviation

i. $E(X^2) = \int_6^9 x^2 f(x)dx$

$= \int_6^9 x^2\frac{1}{9}(-x^2 + 16x - 60)dx = \int_6^9 \frac{1}{9}(-x^4 + 16x^3 - 60x^2)dx = \cdots = \frac{303}{5}$

ii. $V(X) = E(X^2) - \mu^2 = \frac{303}{5} - \left(\frac{31}{4}\right)^2 = \frac{43}{80}$

iii. $\sigma_X = \sqrt{\frac{43}{80}} \approx 0.73$

c. Note: algebra tricks still work (e.g. lost wages while sleeping)

i. $E(\$20X) = \$20E(X) = \$20 \cdot 7.75 = \$155$

ii. $V(20X) = 20^2 V(X)$

5. Practice describing steps to classmate: Warehouse stock (as fraction of capacity) $f(x) = -2x^2 + kx + \frac{1}{6}; 0 \le x \le 1$

a. Find $k = 3$

b. mode $= \frac{3}{4}$

c. Draw and interpret pdf (upside-down parabola; warehouse full more often than empty)

d. Find cdf $F(x) = -\frac{2}{3}x^3 + \frac{3}{2}x^2 + \frac{1}{6}x; 0 \le x \le 1$

e. Find $f(x)$ from $F(x)$

f. $P\left(\frac{1}{2} \le X \le \frac{3}{4}\right) = \frac{5}{16}$

g. median $\approx .6$, 75th percentile $\approx .8$

h. mean $\mu \approx 0.58$

i. standard deviation $\sigma \approx 0.26$

j. Insurance payout $\pi = \$1{,}000{,}000X + \$100{,}000$

i. $E(\pi) = \$1{,}000{,}000\mu + \$100{,}000 = \$680{,}000$ $\sigma_\pi = \sqrt{V(\$1{,}000{,}000X + \$100{,}000)} = \$1{,}000{,}000\sigma_x = \$260{,}000$

L12 Continuous Joint Distributions (WMS 5.1-8)

1. Joint Density

a. Compare discrete/continuous pdf and joint pdf

b. Warehouse stocks up to two pallets of cereal X and one pallet of cereal Y, with density $f(x,y) = c(x + 2y); x \in [0,2], y \in [0,1]$.

c. Height of joint pdf represents likelihood of particular (x,y) pairs. Must integrate to one (double integral).

i. $1 = \int_{x=0}^{2}\int_{y=0}^{1} c(x+2y)dy\,dx = \int_{x=0}^{2}(cx + c)dx = 4c$ requires $c = \frac{1}{4}$, or $f(x,y) = \frac{1}{4}x + \frac{1}{2}y; x \in [0,2], y \in [0,1]$.

d. Mode: since upward sloping in both dimensions, mode at $(x,y) = (2,1)$

2. Marginal densities

a. Analogous to margins of table in discrete joint distribution: total probability of particular realization of x is the sum of all joint probabilities of (x, y) pairs, with that particular x value, but any y value in domain.

b. $f_x(x) = \int_{y=0}^{1} \frac{1}{4}(x + 2y)dy = \frac{1}{4}x + \frac{1}{4}; x \in [0,2]$

c. $f_y(y) = \int_{x=0}^{2} \frac{1}{4}(x + 2y)dx = \frac{1}{2} + y; y \in [0,1]$

d. Subscript simply distinguishes $f_x(.5)$ from $f_y(.5)$

e. Moments: means, standard deviations

i. $\mu_x = E(X) = \int_{x=0}^{2} x f_x(x)dx$

$= \int_{x=0}^{2} x\left(\frac{1}{4}x + \frac{1}{4}\right)dx = \frac{2}{3} + \frac{1}{2} = \frac{7}{6}$

ii. $E(X^2) = \int_{x=0}^{2} x^2 f_x(x)dx$

$= \int_{x=0}^{2} x^2\left(\frac{1}{4}x + \frac{1}{4}\right)dx = 1 + \frac{2}{3} = \frac{5}{3}$

iii. $V(X) = E(X^2) - \mu_x^2 = \frac{5}{3} - \left(\frac{7}{6}\right)^2 = \frac{11}{36}$

iv. $\sigma_x = \sqrt{V(X)} = \sqrt{\frac{11}{36}} \approx .55$

v. $\mu_y = E(Y) = \int_{y=0}^{1} y f_y(y)dy$

$= \int_{y=0}^{1} y\left(\frac{1}{2} + y\right)dy = \frac{1}{4} + \frac{1}{3} = \frac{7}{12}$

vi. $E(Y^2) = \int_{y=0}^{1} y^2 f_y(y)dy$

$= \int_{y=0}^{1} y^2\left(\frac{1}{2} + y\right)dy = \frac{1}{6} + \frac{1}{4} = \frac{5}{12}$

vii. $V(Y) = E(Y^2) - \mu_y^2 = \frac{5}{12} - \left(\frac{7}{12}\right)^2 = \frac{11}{144}$

viii. $\sigma_y = \sqrt{V(Y)} = \sqrt{\frac{11}{144}} \approx 0.28$

ix. Could also derive mode, median, cdf, percentiles of X or Y

f. Independence requires $f(x, y) = f_x(x)f_y(y)$ for all (x, y) pairs.

i. X and Y not independent here, since $f(x, y) = \frac{1}{4}(x + 2y) \neq \left(\frac{1}{4}x + \frac{1}{4}\right)\left(\frac{1}{2} + y\right)$ (e.g. when $(x, y) = (0,0)$)

3. Correlation

a. $E(XY) = \int_{x=0}^{2} \int_{y=0}^{1} xyf(x, y)dy\,dx$

$= \int_{x=0}^{2} \int_{y=0}^{1} xy\left[\frac{1}{4}(x + 2y)\right]dy\,dx = \int_{x=0}^{2} \left(\frac{1}{8}x^2 + \frac{1}{6}x\right)dx = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$

b. $\sigma_{xy} = Cov(X,Y) = E(XY) - \mu_x\mu_y$

$= \frac{2}{3} - \left(\frac{7}{6}\right)\left(\frac{7}{12}\right) = -\frac{1}{72}$

c. $\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y} = \frac{-\frac{1}{72}}{(.55)(.28)} \approx -.09$

4. Practice example: $f(x,y) = c(1-xy)$ for $x, y \in [0,1]$

a. Find c: $\int_{x=0}^{1}\int_{y=0}^{1} c(1-xy)dy\,dx = \frac{3}{4}c$ implies $c = \frac{4}{3}$

b. Find marginal densities f_x, f_y: $f_x(x) = \int_{y=0}^{1}\frac{4}{3}(1-xy)dy = \cdots = \frac{4}{3} - \frac{2}{3}x$ for $x \in [0,1]$;

symmetrically, $f_y(y) = \frac{4}{3} - \frac{2}{3}y$ for $y \in [0,1]$

c. Find means μ_x and μ_y and standard deviations σ_x and σ_y:

$\mu_x = E(X) = \int_{x=0}^{1} x\left(\frac{4}{3} - \frac{2}{3}x\right)dx = \cdots = \frac{4}{9}$

$E(X^2) = \int_{x=0}^{1} x^2\left(\frac{4}{3} - \frac{2}{3}x\right)dx = \cdots = \frac{5}{18}$

$\sigma_x^2 = \frac{5}{18} - \left(\frac{4}{9}\right)^2 = \frac{13}{162}$

$\sigma_x = \sqrt{\frac{13}{162}} \approx .283$

Symmetrically, $\mu_y = \frac{4}{9}$, $\sigma_y \approx .283$

d. Correlation ρ:

$E(XY) = \int_{x=0}^{1}\int_{y=0}^{1} xy\frac{4}{3}(1-xy)dy\,dx = \frac{5}{27}$

$\sigma_{xy} = E(XY) - \mu_x\mu_y \approx \frac{5}{27} - \left(\frac{4}{9}\right)^2 = -\frac{1}{81} \approx -.012$

$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y} = \frac{-\frac{1}{81}}{\sqrt{\frac{13}{162}}\sqrt{\frac{13}{162}}} = -\frac{2}{13} \approx -0.154$

L13 Conditional Distributions (WMS 5.3, 11)

1. Recall distribution of cell phone use and car accidents:

	$Y = 0$	$Y = 1$	$Y = 2$
$X = 0$	.60	.08	.02
$X = 1$	.10	.12	.08

a. Recall $\mu_x = .3, \sigma_x \approx .458, \mu_y = .4, \sigma_y \approx .663, \rho \approx .527$

2. Conditional probability

a. Cell phone use among those with two accidents $P(X=1|Y=2)=\frac{.08}{.10}=.80$ versus those with no accidents $P(X=1|Y=0)=\frac{.10}{.70}\approx 0.14$

b. Practice: find $P_y(0|X=0)=\frac{.60}{.7}\approx .86$, $P_y(1|X=0)=\frac{.08}{.7}=.11$, $P_y(2|X=0)=\frac{.02}{.7}\approx .03$, $P_y(0|X=1)=\frac{.10}{.3}\approx .33$, $P_y(1|X=1)=\frac{.12}{.3}=.40$, $P_y(2|X=1)=\frac{.08}{.3}\approx .27$

3. Conditional distribution

a. $P(y|X=0)$ and $P(y|X=1)$ are legitimate distribution functions (numerators sum to denominator)

b. Plot, and compare with $P(y)$

c. Conditional mean

i. $E(Y|X=1)=\sum yP(Y=y|X=1)\approx 0(.33)+1(.40)+2(.27)=.94$

ii. Practice $E(Y|X=0)\approx 0(.86)+1(.11)+2(.03)=.17$

d. Average number of car accidents is higher for cell phone users than for non-users. This still doesn't imply causation!

e. Conditional standard deviation

i. Just like $V(Y)=E(Y^2)-[E(Y)]^2$,
$V(Y|X=x)=E(Y^2|X=x)-[E(Y|X=x)]^2$

ii. (If time) Example : $E(Y^2|X=1)\approx 0^2(.33)+1^2(.40)+2^2(.27)=1.21$, so
In this case, $V(Y|X=1)\approx 1.21-.94^2\approx 0.326$, and $\sigma_{y|X=1}\approx\sqrt{.326}\approx 0.57$
By a similar derivation, $\sigma_{y|X=0}\approx 0.41$; cell phone use increases variance.

4. In an effort to establish causation, could find $P(x,y|Z=z)=\frac{P(x,y,z)}{P_z(z)}$ or $f(x,y|Z=z)=\frac{f(x,y,z)}{f_z(z)}$, and then $\rho_{xy|z}=Corr(X,Y|Z=z)$. For example, find correlation between cell phone use and car accidents *among teenagers*.

5. Continuous densities

a. Recall joint density of cereal inventory, $f(x,y)=\frac{1}{4}x+\frac{1}{2}y; x\in[0,2], y\in[0,1]$

b. Recall marginal densities $f_x(x)=\frac{1}{4}x+\frac{1}{4}; x\in[0,2]$ and $f_y(y)=\frac{1}{2}+y; y\in[0,1]$, means $\mu_x=\frac{7}{6}$, $\mu_y=\frac{7}{12}$, standard deviations $\sigma_x\approx .55$, $\sigma_y\approx 0.28$

c. Conditional density $f_x(x|Y=y)=\frac{f(x,y)}{f_y(y)}=\frac{\frac{1}{4}x+\frac{1}{2}y}{\frac{1}{2}+y}=\frac{x+2y}{2+4y}; x\in[0,2]$. For example, $f_x(x|Y=0)=\frac{x}{2}; x\in[0,2]$

d. Conditional mean and standard deviation

$E(X|Y=0) = \int_0^2 x\left(\frac{x}{2}\right)dx = \left[\frac{1}{6}x^3\right]_0^2 = \frac{8}{6} = \frac{4}{3}$

$E(X^2|Y=0) = \int x^2 f_x(x|0)dx = \int_0^2 x^2\left(\frac{x}{2}\right)dx = \left[\frac{1}{8}x^4\right]_0^2 = 2$

$V(X|Y=0) = E(X^2|Y=0) - [E(X|Y=0)]^2 = 2 - \left(\frac{4}{3}\right)^2 = \frac{2}{9}.$

$\sigma_{x|Y=0} = \sqrt{\frac{2}{9}}$

Thus, when $Y=0$: density of X is steeper, mean of X is higher, variance is lower.

e. More generically, for arbitrary y value,

$E(X|Y=y) = \int x f_x(x|y)dx = \int_0^2 x\left(\frac{x+2y}{2+4y}\right)dx$

$= \frac{1}{2+4y}\int_0^2 (x^2+2xy)dx = \frac{1}{2+4y}\left[\frac{1}{3}x^3 + x^2y\right]_0^2 = \frac{1}{2+4y}\left(\frac{8}{3}+4y\right) = \frac{4+6y}{3+6y}$

For example, $E(X|y=0) = \frac{4}{3}$ as before

Practice: $E(X|y=1) = \frac{10}{9}$

$E(X^2|Y=y) = \int x^2 f_x(x|y)dx$

$= \int_0^2 x^2\left(\frac{x+2y}{2+4y}\right)dx = \cdots = \frac{6+8y}{3+6y}$

$V(X|Y=y) = E(X^2|Y=y) - [E(X|Y=y)]^2$

$= \left(\frac{6+8y}{3+6y}\right) - \left(\frac{4+6y}{3+6y}\right)^2$

$\sigma_{x|Y=y} = \sqrt{\left(\frac{6+8y}{3+6y}\right) - \left(\frac{4+6y}{3+3y}\right)^2}$. For example, $\sigma_{x|Y=0} = \sqrt{\frac{6}{3} - \left(\frac{4}{3}\right)^2} = \sqrt{\frac{2}{9}}$ as before,

$\sigma_{x|Y=1} = \sqrt{\left(\frac{14}{9}\right) - \left(\frac{10}{9}\right)^2} = \sqrt{\frac{26}{81}} \approx \frac{5}{9}$

Thus, when $Y=1$, density of X is less steep, mean is lower, variance is higher.

6. With three variables, can derive joint distribution of X and Y conditional on Z

   a. Israel covid data

      i. When covid Delta variant hit, Israeli hospitals filled up with covid patients. 60% of these patients had already been vaccinated.

      ii. Natural (but erroneous) conclusion: vaccines make things worse, not better!

iii. Nearly 80% of Israelis over age 12 were vaccinated against covid, so if "random draw," 80% of hospital patients should have been vaccinated; lower rate means treatment helped.

iv.

7.

L14 Regressions (WMS 5.3, 11)

1. Regressions
    a. Sir Francis Galton 1886 (cousin of Darwin)
    b. Use data to determine (average) linear relationship $Y = \beta_0 + \beta_1 X$. Once relationship is known, we can predict Y for any X value (even out of sample), as if through a crystal ball!
    c. Examples:
        i. Income Y as function of education X
        ii. Unemployment Y next year as function of (e.g. fiscal or monetary) policy X
        iii. Stock price tomorrow Y as function of today's earnings/price X
        iv. Consultant's "secret formula" predicting sales, etc.
    d. Puts units on correlation ("education and income are strongly correlated" vs. "each year of education is associated with an additional $4k of income")
    e. Working example: education $\mu_x = 15$ years; $\sigma_x = 3$ years; income $\mu_y = \$70{,}000$; $\sigma_y = \$20{,}000$; correlation $\rho = .6$
    f. Any β_0 and β_1 determine line $Y = \beta_0 + \beta_1 X$, implying an error term $\varepsilon = Y - \beta_0 - \beta_1 X$ such that the data satisfies $Y = \beta_0 + \beta_1 X + \varepsilon$. We can choose β_0 and β_1 so that the resulting line is as useful as possible.
    g. "Least squares" regression: choose β_0 and β_1 to minimize $E(\varepsilon^2)$
        i. Equivalently, choose β_0 so that $E(\varepsilon) = 0$ and β_1 to minimize $V(\varepsilon)$
        ii. Can use other criteria (e.g. least absolute deviation $E(|\varepsilon|)$), but less common
2. Intercept
    a. If β_0 is high, most ε_i will be negative; if β_0 is low, most ε_i will be positive
    b. $E(\varepsilon) = \mu_y - \beta_0 - \beta_1 \mu_x = 0$ implies that $\beta_0 = \mu_y - \beta_1 \mu_x$.
    Easier: $\mu_y = \beta_0 + \beta_1 \mu_x$, so regression line passes through (μ_x, μ_y)

c. Example: $\beta_0 = \$70{,}000 - \$4{,}000 \cdot 15 = \$10{,}000$

3. Slope
   a. $V(\varepsilon) = V(Y) + V(-\beta_1 X) + 2Cov(Y, -\beta_1 X) = \sigma_y^2 + \beta_1^2 \sigma_x^2 - 2\beta_1 \sigma_{xy}$
   b. To minimize, $0 = \frac{dV(\varepsilon)}{d\beta_1} = 2\beta_1 \sigma_x^2 - 2\sigma_{xy}$
   c. Solution $\beta_1 = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \frac{\sigma_y}{\sigma_x} = \rho \frac{\sigma_y}{\sigma_x}$
   d. Slope is simply (normalized) correlation coefficient
   e. Example: $\beta_1 = .6 \frac{\$20{,}000}{3yr.} = \$4{,}000$/yr. (e.g. four-year degree provides extra $16,000/yr)
   f. Equivalently, can derive same value by choosing β_1 such that $Cov(X, \varepsilon) = 0$ (HW)
4. Predictions
   a. High school grad ($X^* = 12$) expects $Y^* = \$10k + \$4k(12) = \$58k$
   b. College grad ($X^* = 16$) expects $Y^* = \$10k + \$4k(16) = \$74k$
   c. Three PhDs ($X^* = 31$) expects $Y^* = \$10k + \$4k(31) = \$134k$
      i. This assumes linear trend holds up, constant returns to scale (which may not be reasonable); in econometrics (Econ 388), learn nonlinear regressions
   d. Standardized
      i. $\frac{Y^* - \mu_y}{\sigma_y} = \rho \frac{X^* - \mu_x}{\sigma_x}$ (since $\beta_1 = \rho \frac{\sigma_y}{\sigma_x}$, $\mu_y = \beta_0 + \beta_1 \mu_x$, and $Y^* = \beta_0 + \beta_1 X^*$).
      ii. If X^* is 1 or 2 or k standard deviation above μ_x then Y^* is ρ or 2ρ or $k\rho$ standard deviations above μ_y
   e. Stay in school to get rich?
      i. Maybe. Correlation might reflect causation; if so, staying in school boosts income.
      ii. Regressions still just express correlation, not causation (now in meaningful units)
      iii. Maybe not. Maybe spurious (rich kids enjoy school, but would be rich either way) or maybe helps those who already attend (e.g. engineers) but those who choose not to attend (e.g. mechanics, artists) wouldn't benefit.
      iv. Either way, predict higher incomes for those who do stay in school
      v. But maybe not for those who choose not, but compelled/advised to go to school
   f. Reverse predictions
      i. What if worker makes $100k income and asks for you to guess education?

ii. Could read regression backward, but it was designed to minimize errors in income not errors in education

iii. Better to start over, with income as X and education as Y

5. Errors

   a. $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$

   b. De-trend data (e.g. "skill" or "luck", above and beyond education)

   c. Example: who is more genius (or luckier): $(x, y) = (12, \$80k)$ or $(x, y) = (16, \$100k)$?

      i. $\$80k - (10 + 4 \cdot 12) = \$22k$

      ii. $\$100k - (10 + 4 \cdot 16) = \$26k$

   d. Variance σ_ε^2 of error distribution tells us how far people are off the regression line

   $\sigma_\varepsilon^2 = V(Y - \beta_0 - \beta_1 X) = \sigma_y^2 + \beta_1^2 \sigma_x^2 - 2\beta_1 cov(X, Y)$

   $= 20k^2 + 4k^2 3^2 - 2(4k)(.6 \times 20k \times 3) = (\$16k)^2$

6. Explanatory power

   a. Partition $V(Y) = \beta_1^2 V(X) + V(\varepsilon) = 144 + 256$

      i. Note: $2\beta_1 Cov(X, \varepsilon) = 0$ (see homework) because optimal slope extracts all correlation

      ii. This decomposes $V(Y)$ into "explained" and "unexplained" (e.g. talent, luck, or some other mystery). More accurately, variation that is "related to education" and variation that is "unrelated to education"

   b. "Explained" portion is ρ^2 fraction of $V(Y)$

      i. $\beta_1^2 \sigma_x^2 = \left(\frac{\sigma_y}{\sigma_x}\rho\right)^2 \sigma_x^2 = \rho^2 \sigma_y^2$

      ii. In this case, $.6^2 = 36\%$ ("coefficient of determination")

   c. "Unexplained" portion is $1 - \rho^2$

      i. In this case, $1 - .6^2 = 64\%$

      ii. Shortcut $\sigma_\varepsilon^2 = .64(400) = 256 = (\$16k)^2$

   d. Implicit linearity of ρ

      i. Fundamentally, what does ρ measure?

      ii. X^2 is perfectly predictable from X, but linear regression produces $\rho^2 < 1$

      iii. Thus, $corr(X, X^2) \neq 1$

      iv. ρ fundamentally measures *linear* relationship (see homework)

Exam 1 Review

1. Spiritual thought: prayer through life's trials, faith without works is dead, obedience gives confidence
2. Exam info
    a. Any calculator
    b. No time limit, predict 2-3 hours
    c. Handout provided
    d. Hard: typically C average
3. Study tips
    a. Take it seriously: equal weight with final exam
    b. Start with study guide
    c. Practice exams (first without solutions, then with)
    d. Extra homework problems (or repeat homework problems)
    e. Time saver: talk through steps, don't work out algebra
4. Exam strategies
    a. Don't forget to pray!
    b. Extend familiar material to unfamiliar settings (good practice for real world)
    c. Difficulty is similar to homework, but no TAs or books, so fewer A's than homework
    d. Average score is typically C, which averaged with A- homework gives B- overall.
    e. Show work and list what you know for partial credit (e.g. $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$, even if you can't figure out σ_{xy})
5. Key formulas
    a. Binary events
        i. $P(E) = \frac{\#E}{\#S}$
        ii. $C_k^n = \frac{n!}{k!(n-k)!}$
        iii. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
        iv. Independent events: $P(A \cap B) = P(A)P(B)$ (or $P(A|B) = P(A)$)
        v. $P(A|B) = \frac{P(A \cap B)}{P(B)}$
        vi. $P(A \cap B) = P(B|A)P(A)$
    b. Random variables

i. Legitimate distribution? $\sum P(x) = \int f(x)dx = 1$

ii. Mode maximizes $P(x)$ or $f(x)$ (i.e. $f'(x) = 0$ and $f''(x) < 0$)

iii. $\mu = E(X) = \sum xP(x) = \int xf(x)dx$

iv. $E(X^3) = \sum x^3 P(x) = \int x^3 f(x)dx$

v. $\sigma^2 = V(X) = E[(X-\mu)^2] = E(X^2) - \mu^2; \sigma = \sqrt{V(X)}$

vi. $F(x) = \int_{-\infty}^{x} f(\tilde{x})d\tilde{x}, f(x) = F'(x)$

vii. $P(a < X < b) = F(b) - F(a)$

viii. Percentile: solve $F(\phi_{.5}) = .5$

ix. $f(x) = F'(x)$

c. Joint distributions

i. Legitimate joint distribution? $\sum\sum P(x,y) = \iint f(x,y)dxdy = 1$

ii. Marginal distribution

$P_x(x) = \sum_y P(x,y)$

$f_x(x) = \int f(x,y)dy$

iii. Independent variables

$P(x,y) = P_x(x)P_y(y)$

$f(x,y) = f_x(x)f_y(y)$

iv. $E\left(\frac{X}{Y}\right) = \sum\sum\left(\frac{x}{y}\right)P(x,y) = \iint\left(\frac{x}{y}\right)f(x,y)dxdy$

v. $Cov(X,Y) = E\left[(X-\mu_x)(Y-\mu_y)\right] = E(XY) - \mu_x\mu_y$

vi. $\rho = \frac{Cov(X,Y)}{\sigma_x\sigma_y}$

vii. Conditional distribution

$P(X = x|Y = 3) = \frac{P(x,3)}{P_y(3)}$

$f_x(x|Y = 3) = \frac{f(x,3)}{f_y(3)}$

viii. $E(X|Y = 3) = \sum xP(x|Y = 3) = \int xf(x|Y = 3)dx$

ix. $V(X|Y = 3) = E(X^2|Y = 3) - [E(X|Y = 3)]^2$

d. Regressions

i. $\beta_1 = \frac{\sigma_{xy}}{\sigma_x^2} = \rho\frac{\sigma_y}{\sigma_x}$

ii. $\beta_0 = \mu_y - b\mu_x$

iii. $\frac{V(a+bX)}{V(Y)} = \rho^2$

iv. $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$

6. Algebra tricks
    a. $E(\$100 - \$5X) = \$100 - \$5E(X)$
    b. $V(\$100 - \$5X + \$3Y) = V(\$100) + V(-\$5X) + V(\$3Y) + 2Cov(-\$5X, \$3Y) = 0 + (\$5)^2 V(X) + (\$3)^2 V(Y) - \$30Cov(X,Y)$
    c. $Cov(\$100 - \$5X, Y) = Cov(\$100, Y) + Cov(-\$5, Y) = 0 - \$5Cov(X,Y)$
    d. $Corr(\$100 - \$5X, Y) = Corr(-X, Y) = -Corr(X,Y)$
7. Distributional relationships
    a. If $X \sim N$ then $3X \sim N$ and $X + 7 \sim N$
    b. If $X_1, X_2 \sim N$ then $X_1 + X_2 \sim N$
    c. If $Z \sim N(0,1)$ then $Z^2 \sim \chi^2(1)$
    d. If $W_1 \sim \chi^2(3), W_2 \sim \chi^2(5)$ independent then $W_1 + W_2 \sim \chi^2(8)$ and $\frac{W_1/3}{W2/5} \sim F(3,5)$
    e. If $Z \sim N(0,1)$ and $W \sim \chi^2(\nu)$ independent then $\frac{Z}{\sqrt{\frac{W}{\nu}}} \sim t(\nu)$
8. Rejoice in how much we've learned!

L15 Bernoulli, Uniform, Standard Normal (WMS 4.4-4.5)

Spiritual thought

1. Dealing with disappointment
    a. In grad school, we took two years of courses, then two qualifying exams. If pass, four years of research; if fail, retake or exit with Masters degree. I prepped hard, but on day of exam, got hung up on one really hard question, lost track of time, didn't finish, and failed!
    b. I benefitted from a friend's experience, who had previously been preparing for graduation (robes, parents in town, etc.), when checked grades: E! Couldn't graduate.
        i. First reaction: denial. Must be a mistake!
        ii. Second reaction: blame. Grading is unfair!
        iii. Third reaction: dejection. I'm a failure.
        iv. Fourth reaction: hope. *I'm* not a failure, I just failed at this thing. I can move forward productively to the next step. Retook class, found a summer internship that turned out to be career altering.

c. Scriptures

i. Joseph Smith in Liberty jail: “My son, peace be unto they soul; thine adversity and thine afflictions shall be but a small moment; and then, if thou endure it well, God shall exalt thee on high; thou shalt triumph over all thy foes” (D&C 121:7-8).

ii. “Search diligently, pray always, and be believing, and all things shall work together for your good, if ye walk uprightly and remember the covenant wherewith ye have covenanted one with another” (D&C 90:24).

iii. “…All things work together for good for them that love God…” (Romans 8:28).

d. Midterm exam: If you performed less well than you hoped, press forward with a perfect brightness of hope! Help the Lord make it work toward your good.

i. Learn what went wrong (like spelling bee mistakes, may always remember). Final exam not cumulative per se, but does repeat concepts.

ii. Reassess study habits (e.g. understand every step of every question; don’t just trust TA or study group).

2. $X \sim$ Bernoulli(p) (after Swiss mathematician Jacob Bernoulli, 1713)

a. Recall cell phone use $P(X = x) = \begin{cases} .7 \; if \; x = 0 \\ .3 \; if \; x = 1 \end{cases}$

b. Mean $E(X) = 0(.7) + 1(.3) = .3$

c. Variance $V(X) = E(X^2) - \mu^2 = [0^2(.7) + 1^2(.3)] - .3^2 = .21 = (.3)(.7)$

d. Pattern: $E(X) = p, V(X) = p(1-p)$ for “success” parameter p

3. $X \sim$ Uniform(a, b)

a. $f(x) = k; a \le x \le b$

b. $F(x) = \int_a^x k d\tilde{x} = \cdots = \frac{x-a}{b-a}; a \le x \le b$

c. $\mu = \int_a^b x f(x) dx = \cdots = \frac{a+b}{2}$

d. $\sigma^2 = \int_a^b x^2 f(x) dx - \mu^2 = \cdots = \frac{(b-a)^2}{12}$

e. Example: 90 second stop light

i. Average wait time $E(X) = 45$

ii. Standard deviation $\sigma = \sqrt{\frac{(90)^2}{12}} \approx 26$

iii. Wait less than 30 seconds with probability $F(30) = \frac{30-0}{90-0} = \frac{1}{3}$

1. Standard normal $N(0,1)$

a. $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$ (integrate using polar coordinates or trig substitutions)

b. $E(Z) = \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \cdots = 0$ (u substitution)

c. $V(Z) = \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz - 0^2 = \cdots = 1$ (integration by parts)

d. Practice reading Table A

   i. Excel: NORM.S.DIST(x, cdf?) or NORM.S.INV(percentile)

   ii. $P(-1 < X < 1) \approx .68$

   iii. $P(-2 < X < 2) \approx .95$

   iv. $P(-3 < X < 3) \approx .997$

e. Symmetric: $P(X < -3) = P(X > 3)$

L16 Normal, Chi Square, t Distributions (WMS 4.5-4.6)

1. Standardization (for later reference)

   a. If $E(X) = \mu$ and $V(X) = \sigma^2$ then you can always change units to create a new random variable $Z = \frac{X-\mu}{\sigma}$ such that $E(Z) = 0$ and $V(Z) = 1$

      i. $E(Z) = E\left[\frac{1}{\sigma}(X - \mu)\right] = \frac{1}{\sigma}[E(X) - \mu] = 0$

      ii. $V(Z) = V\left[\frac{1}{\sigma}X - \frac{1}{\sigma}\mu\right] = V\left(\frac{1}{\sigma}X\right) = \frac{1}{\sigma^2}V(X) = 1$

2. Normal (or Gaussian, after German mathematician Carl Friedrich Gauss, 1809) $N(\mu, \sigma^2)$

   a. $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ (integrate using polar coordinates or trig substitutions)

   b. $E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \cdots = \mu$ (u substitution)

   c. $V(X) = \int_{-\infty}^{\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx - \mu^2 = \cdots = \sigma^2$ (integration by parts)

   d. No analytical cdf; instead, approximate numerically

      i. Excel: NORM.DIST(x, mu, sd, cdf?)

      ii. Percentiles: NORM.INV(percentile, mu, sd)

   e. Special Properties

      i. $N + 7 \sim N$

         In other words, adding a constant changes the precise distribution of X but keeps it in the normal family

1. Note: this is true of some other families of random variables (e.g. uniform) but not all (e.g. Bernoulli, binomial, exponential)

ii. $3N \sim N$

In other words, multiplying by a constant keeps X in the normal family

1. Note: this is true of some other families of random variables (e.g. uniform, exponential) but not all (e.g. Bernoulli, binomial)

iii. $N + N \sim N$

That is, if $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$

then $X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy})$

In other words, the sum of two normally distributed random variables is a normally distributed random variable

1. Note: this is true of some other families of random variables (e.g. independent binomials), but not all (e.g. Bernoulli, correlated binomials, uniform, exponential)

3. Standard normal $N(0,1)$

a. Practice reading Table A

i. Excel: NORM.S.DIST(x, cdf?) or NORM.S.INV(percentile)

ii. $P(-1 < X < 1) \approx .68$

iii. $P(-2 < X < 2) \approx .95$

iv. $P(-3 < X < 3) \approx .997$

b. Symmetric: $P(X < -3) = P(X > 3)$

c. Standardized normal $Z = \frac{X-\mu}{\sigma}$ is standard normal $\sim N(0,1)$ (because of special properties of normal X)

d. Example 1: $X \sim N(75,25)$ to find $P(X > 80) = P\left(Z > \frac{80-75}{\sqrt{25}}\right)$

$= P(Z > 1) = 1 - P(Z \leq 1)$

$\approx 1 - .8413 = .1587$

e. Example 2: costs $C \sim N(120,100)$

i. Budget b so that $P(C < b) = .9$

ii. $.90 = P(C < b) = P\left(Z < \frac{b-120}{10}\right) \approx P(Z < 1.28)$ (from Table A)

iii. If $\frac{b-120}{10} \approx 1.28$ then $b \approx 132.8$

f. Example 3: costs $C \sim N(120,100)$) and revenue $R \sim N(150,400)$ are independent; how often are profits $Y = R - C$ positive?

i. $\mathrm{Y} \sim N$

ii. $\mathrm{E(Y)} = \mathrm{E(R)} - \mathrm{E(C)} = 150 - 120 = 30$

iii. $V(\mathrm{Y}) = V(R - C) = V(R) + (-1)^2 V(C) + 2Cov(R, C) = 400 + 100 = 500$

iv. So $Y \sim N(30,500)$

v. $P(\mathrm{Y} > 0) = P\left(Z > \frac{0-30}{\sqrt{500}}\right) \approx P(Z > -1.34) = P(Z < 1.34) \approx .9099$

4. $W \sim \chi^2(\nu)$ (German statistican Friedrich Robert Helmert, 1875)

a. Domain is $[0, \infty)$, roughly bell-shaped (but asymmetric, unlike Normal distribution)

b. ν is often called "degrees of freedom", because in the most common application, it corresponds to how many

c. $E(W) = \nu$ and $V(W) = 2\nu$

d. $f(w) = ugly$ (I won't expect you to know or use)

e. CDF $F(w)$ approximated on Table 6

i. χ^2_α represents a realization of the random variable, where α is the probability to the right of that value (i.e., $1 - F(w)$)

ii. Example: suppose sales follow Chi-square distribution, with average of 30 units

1. Degrees of freedom $\nu = 30$

2. 10th percentile is $\chi^2_{.90} \approx 20.6$, 90th percentile is $\chi^2_{.10} \approx 40.3$

3. Putting these together, $P(20.6 < W < 40.3) \approx .8$

iii. Note: Table 6 only gives 10 points on the cdf. With a computer, you can get the rest. Excel: CHISQ.DIST(x,df, cdf?), CHISQ.INV(percentile, df)

f. Facts

i. If $Z \sim N(0,1)$ then $Z^2 \sim \chi^2(1)$

ii. If $W_1 \sim \chi^2(4)$ and $W_2 \sim \chi^2(7)$ independent then $W_1 + W_2 \sim \chi^2(11)$

iii. Variance is a quadratic function of a random variable, so when we estimate the variance of a random variable that has a normal distribution (in lecture L19), our estimates will follow a χ^2 distribution.

5. t distribution (Friedrich Robert Helmert 1876, Karl Pearson 1900)

a. $T \sim t(\nu)$; as in Chi-square distribution, ν is called "degrees of freedom"

b. Similar to standard normal, but with higher variance (i.e. thicker tails)

c. Approaches $N(0,1)$ as $\nu \to \infty$

d. $f(t) = ugly$ (I won't expect you to know or use)

e. $E(T) = 0, V(T) = \frac{\nu}{\nu-2} \to 1$

f. CDF $F(t)$ approximated on Table C

i. Table is oriented so that probability C lies between $-t^*$ and t^*.

ii. Example: if $T \sim t(20)$ find 90th percentile

1. Following $C = 80\%$ (fifth column) for $df = 20$ leads to $t^* = 1.325$.
2. In other words, 10% of the distribution is left of -1.325, 80% is between -1.325 and 1.325, and 10% is above 1.325.
3. Since $10\% + 80\% = 90\%$ of the distribution is below 1.325 and 10% is above, 1.325 is the 90th percentile of the distribution.
4. Alternatively, can come up from a one-sided p-value of $.10$ or a two-sided p-value of $.20$ (bottom of the table) to reach the same conclusion.

iii. For degrees of freedom greater than 1000, can read z^* row of the table, which corresponds to a standard normal distribution (i.e., ∞ degrees of freedom).

iv. Note: Table C only gives 12 points on CDF. With a computer, you can get the rest. Excel: T.DIST(x, df, cdf?) and T.INV(percentile, df)

g. Fact

i. If $Z \sim N(0,1)$ and $W \sim \chi^2(\nu)$ independent then $\frac{Z}{\sqrt{\frac{W}{\nu}}} \sim t(\nu)$

ii. If we knew the population variance, then estimates of the mean would follow a normal distribution. Since we have to estimate the population variance, and estimates follow a χ^2 distribution, our estimates of the mean follow a t distribution

6. Other distributions

a. The distributions we've gone over are some of the most common; there are many others, with various shapes, properties, and uses.

b. Illustrated: https://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm

c. Discrete

i. Uniform

ii. Binomial

iii. Geometric

iv. Poisson

v. Hypergeometric

d. Continuous

i. Exponential

ii. F

iii. Beta

iv. Gamma

v. Log-normal

vi. Pareto

vii. Weibull

L17 Confidence Intervals (WMS 7.2-3,8.5-9)

Project note

1. If you didn't turn the project in on time, get it in ASAP! Items from part 2 of the project will show up on homework; if you do them with your homework, your project will be finished by the end of the semester.
2. From now on, must start on project as part of homework
3. Keep results, to submit as project
4. Note: If you have a population instead of a sample from a population (e.g. all 50 states), just pretend this is a sample from a larger population (i.e. 50 draws from a population of thousands of U.S. states).

Samples

1. Population vs. sample
    a. So far, our discussion of distributions has presumed an entire population. Often, information is only available from a sample.
        i. Surveys are costly, populations are often huge
        ii. Some of you might have whole populations (e.g. all 50 states, all teams, every week of a company's sales data); for projects, pretend sample even if you actually have population. But be careful:
            1. Sometimes population of interest includes future generations (e.g. NBA rookies, stock returns).

2. Similarly, population of things that actually happened can in some cases be viewed as a sample from the larger set of things that potentially could have occurred instead.

b. Unless entire population is observed, can't know what is true, only what is *probably* true

2. Random sample
   a. i.i.d. (Independently and Identically Distributed): survey answers are independent from each other, and identical to population of interest
   b. Convenience survey (e.g. urban survey of wages): expand sample or limit scope of inference
   c. Selection bias (e.g. survey participation, program participation): administrative records, measurements before participation decided, interpret results narrowly (e.g. benefit of college for those who chose to attend)
   d. Time trends (e.g. daily/weekly sales) – rare "spot check" observations, econometric corrections
3. Estimation
   a. Example: Suppose we wish to estimate the average family size μ of BYU students, along with the standard deviation σ and the correlation ρ between family size and GPA. What pieces of data should be used, and how should they be combined?
   b. Population parameter θ (i.e. generic proxy for $\mu, \sigma, a, b, \rho, \beta, p$, etc.), seek "estimator" function $\hat{\theta}(x_1, x_2, \dots, x_n)$ (commonly denoted by "hat" variable)
      i. Evaluating this "estimator" function with our data provides *point estimates*; next two lectures we'll talk about interval estimates, or margin of error
   c. An estimator is a tool for producing estimates. We'll spend most of the semester talking about a variety of such tools (i.e. estimators for different parameters) but first we need some tool-building tools (i.e. techniques for developing estimators in new settings of interest).

Estimators vs. estimates

1. Example: suppose distribution of income among last year's $8{,}500$ BYU graduates has mean $E(X_i) = \mu = \$48k$ and standard deviation $\sqrt{V(X_i)} = \sigma = \$13k$

But we can't observe this, so we survey $n = 25$ graduates and estimate $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ and $\widehat{\sigma^2} = \frac{1}{n}\int_{i=1}^{n}(x_i - \bar{x})^2$

2. Sampling distribution
    a. Every survey of 25 students yields different estimates $(\hat{\mu}, \widehat{\sigma^2})$. Sampling with replacement, there are $8{,}500^{25} \approx 10^{98}$ such samples.
    (Sampling without replacement is more common in practice, and violates i.i.d. but only slightly, as long as population size is large.)
    b. Before we conduct interviews, survey responses $X_1, X_2, \ldots, X_n$ can be viewed as random variables, each drawn from the population of BYU grads
3. Estimates and estimators
    a. Once we conduct survey, $\hat{\theta}(x_1, x_2, \ldots, x_n)$ provides *estimate* of parameter θ. Before we conduct survey, *estimator* $\hat{\theta}(X_1, X_2, \ldots, X_n)$ is random.
    b. To evaluate estimation procedure, we must think about entire *distribution* of estimates (in other words, evaluate estimator), not individual estimate.
    c. Therefore, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ is random variable with mean $\mu_{\hat{\mu}}$ and variance $\sigma^2_{\hat{\mu}}$
    d. Similarly, $\widehat{\sigma^2} = \frac{1}{n}\int_{i=1}^{n}(X_i - \bar{X})^2$ is random variable with mean $\mu_{\widehat{\sigma^2}}$ and variance $\sigma^2_{\widehat{\sigma^2}}$

Margin of error

1. Recall that
    $\mu_{\bar{X}} = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \cdots = \mu$
    $\sigma^2_{\bar{X}} = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \cdots = \frac{\sigma^2}{n}$
2. Previous estimates are *point* estimates; *margin of error* (e.g. $\pm\$20k$) measures precision, gives *interval* estimate
3. Example
    a. Income X_i of 8,500 BYU graduates has unknown mean μ and known standard deviation $\sigma = \$13k$.
    b. If $n = 25$ then $\bar{X}$ has same mean μ, and standard error $\sigma_{\bar{X}} = \sqrt{\frac{(\$13k)^2}{25}} = \$2.6k$
    c. Rule of thumb: X_i typically within $\mu \pm 2\sigma$, $\bar{X}$ typically within $\mu \pm 2\sigma_{\bar{X}} = \mu \pm \$5.2k$; thus, $\$5.2k$ is "margin of error"

d. Dog and leash principle: 3 ft. leash keeps dog within 3 ft. of owner; symmetrically keeps owner within 3 ft. of dog

e. Observe $\bar{x} = \$47.1k$

Maybe μ as low as $\$41.9k$ and we overestimated

Maybe μ as high as $\$52.3k$ and we underestimated.

4. If σ unknown, can estimate margin of error using $\sqrt{\frac{s^2}{100}}$

Confidence Intervals

1. How often is $\bar{X}$ in interval $\mu \pm 2\sigma$? To compute probability, we need the cdf of $\bar{X}$.
2. Normality of $\bar{X}$

a. If population distribution of X_i is normal then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is normal too. Specifically, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

b. Standardizing, $\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$.

Confidence interval for μ

1. Construction

a. We *want* $\Pr(\# < \mu < \#) = .90$ and from Table A we *know* $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N(\mu, \$2.6k^2)$ (still assuming $\sigma = \$13k$ and $n = 25$)

$$.90 = P\left(-1.64 < \frac{\bar{X}-\mu}{\$2.6k} < 1.64\right)$$
$$= P(-1.64 \cdot \$2.6k < \bar{X} - \mu < 1.64 \cdot \$2.6k)$$
$$\approx P(-\bar{X} - \$4.3k < -\mu < -\bar{X} + \$4.3k)$$
$$= P(\bar{X} + \$4.3k > \mu > \bar{X} - \$4.3k)$$

b. (Can also construct one-sided confidence intervals)

c. Example: $\bar{x} = \$47.1k$ (still assume $\sigma = \$13k$; later we'll estimate)

i. 90% confidence interval $\bar{x} \pm 1.64\sigma_{\bar{x}} = \$47.1k \pm \$4.3k$

ii. 95% confidence interval $\bar{x} \pm 1.96\sigma_{\bar{x}} = \$47.1k \pm \$5.1k$

iii. 99% confidence interval $\bar{x} \pm 2.58\sigma_{\bar{x}} = \$47.1k \pm \$6.7k$

2. Distribution of S^2

a. If $X_i \sim N(\mu, \sigma^2)$ and $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ then $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1)$.

b. Intuitively, expectation of $\chi^2(n-1)$ is $n-1$, expectation of $\frac{S^2}{\sigma^2}$ is 1.

3. Confidence interval for μ when σ unknown
   a. If we replace σ^2 with s^2 then $\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim t(n-1)$.
      i. This is because $\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \cdot \frac{1}{\sqrt{\frac{S^2}{\sigma^2}}}$, which is $N(0,1)$ divided by $\frac{\chi^2(n-1)}{n-1}$
      ii. Note: if n large then $t(n-1) \approx N(0,1)$.
   b. Example: average weekly income $n = 25$, $\bar{x} = \$47.1k$, $s = \$13k$, $\hat{\sigma}_{\bar{X}} = \sqrt{\frac{s^2}{n}} = \$2.6k$
      i. 90% confidence interval $\bar{x} \pm 1.726\hat{\sigma}_{\bar{X}} = \$47.1k \pm \$4.5k$
      ii. 95% confidence interval $\bar{x} \pm 2.093\hat{\sigma}_{\bar{X}} = \$47.1k \pm \$5.4k$
      iii. 99% confidence interval $\bar{x} \pm 2.861\hat{\sigma}_{\bar{X}} = \$47.1k \pm \$7.4k$

Central Limit Theorem (de Moivre 1733, Laplace 1812, Lyapunov 1901)

4. $\sum_{i=1}^{n} X_i \to N$ (and therefore $\bar{X} \to N$) no matter what the distribution of X_i
5. Dice example
   a. Distribution of $X_1 + X_2$ is bell-shaped, even though X_i is (discrete) uniform
   b. Intuition: centrist values frequent (e.g. moderate X_1 and X_2, or X_1 low X_2 high, or vice versa), but extreme values rare (e.g. X_1 and X_2 both low)
   c. $P(\bar{X}_{100} = 1) = \left(\frac{1}{6}\right)^{100} \approx 10^{-78}$; tails become *exponentially* less likely (key feature of normal distribution) as n increases
6. Skewed example
   a. Bernoulli unemployment $P(0,1) = (.7,.3)$
   b. Average of two: $P(0,.5,1) \approx (.5,.4,.1)$
   c. Average of four: $P(0,.25,.5,.75,1) \approx (.25,.4,.25,.1,0)$
7. CLT explains why normal distribution is so prevalent in nature: one attribute is sum total of many, smaller, independent attributes

L18 Hypothesis Tests (WMS 10.2-8)

1. Hypothesis test: old profit $X \sim N(\$400, \$100^2)$, new management; keep or fire?
   a. Null hypothesis (benefit of doubt) $H_0: \mu = 400$

b. Alternative hypothesis (burden of proof) $H_a: \mu < 400$

c. Level $\alpha = \Pr(reject\ H_0 | H_0\ true) = .10$

d. Test statistic $\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$

e. Critical value -1.28, rejection region to left

f. Data: $\bar{x} = \$350$ over 8 weeks

g. If H_0 true, $\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{350-400}{\sqrt{100^2/8}} \approx -1.41 \in RR$; reject H_0

h. "Significantly less than 400" (statistical vs. practical significance)

i. Type 1 error: convict innocent (probability α)

j. Type 2 error: acquit guilty (probability β)

k. Repeat for $\alpha = .01$; critical value -2.33, fail to reject

l. P-value = smallest α such that (for this data) reject H_0; 0.0793 in this case

m. Practice: Reject if $\alpha = .05$?

2. Variations

a. $H_a: \mu < 380$ (expect and tolerate adjustment cost \$20 for new); test statistic increases to $-.85$, p-value increases to 0.20. (At $\alpha = .10$ level, \$350 is *significantly* less than \$400, but not significantly less than \$380)

b. $H_a: \mu > 450$; if still $\alpha = .10$ then critical value $+1.28$; test statistic negative, so (really) fail to reject

c. What if σ^2 unknown, and $s^2 = 100^2$ instead? Use t-distribution with 7 degrees of freedom; critical value if $\alpha = .10$ is 1.415; reject null hypothesis. (p-value not on chart, but by computer is 0.1007)

d. $H_a: \mu \neq 400$; critical values at ± 1.645, now fail to reject; p-value $2(.079) = 0.158$

3. Relationship to confidence intervals

a. In two-sided $\alpha = .05$ level hypothesis test, reject if $\left|\frac{\bar{X}-400}{\sigma_{\bar{x}}}\right| > 1.645$. In other words, if $\bar{X}$ more than $1.645\sigma_{\bar{x}}$ units from 400.

b. Two-sided 95% confidence interval consists of $\bar{X} \pm 1.645\sigma_{\bar{x}}$

c. In other words, $.05$ level hypothesis test merely asks whether 400 lies inside the 95% confidence interval.

L19 Bias and Consistency (WMS 7.2-7.4, 9.1-9.3)

[What if you used median to estimate mean, say in income distribution? Biased.]

Properties of estimators

1. Evaluating $\hat{\theta}$ amounts to evaluating distribution of $\hat{\theta}(X_1, X_2, \dots, X_n)$ relative to true unknown value θ
2. Though θ is unknown, we know how $\hat{\theta}$ relates to X_i and how X_i relates to θ, so can know (probabilistically) how $\hat{\theta}$ relates to θ
3. We'll use this to evaluate estimator goodness and to define *margin of error/interval estimates*, and do *hypothesis test*
4. Moments of $\hat{\mu} = \bar{X}$
    a. $\mu_{\hat{\mu}} = E(\hat{\mu}) = E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} nE(X_i) = E(X_i) = \mu$

    Thus, thought we don't know what μ is, we know that average realization of $\bar{X}$ and average realization of X_i are same

    b. $\sigma_{\hat{\mu}}^2 = V(\hat{\mu}) = V(\bar{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2} nV(X_i) = \frac{\sigma^2}{n}$

    Variance of $\bar{X}$ is much smaller than variance of X_i

    c. Standard error (i.e. standard deviation of estimator) $\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

        i. In population ($n = 1$), incomes typically between $\$48k \pm \$26k$ $[\$22k, \$74k]$

        ii. For $n = 25$, sample average $\bar{X}$ typically between $\$48k \pm \$5.2k$ $[\$43k, \$53k]$

        iii. For $n = 100$, sample average $\bar{X}$ typically between $\$48k \pm \$2.6k$ $[\$44k, \$51k]$

        iv. For $n = 10{,}000$, $\bar{X}$ typically between $\$48k \pm \$0.26k$ $[\$47.7, \$48.3k]$

Consistency

1. Best imaginable case: $\hat{\theta}$ degenerate with $E(\hat{\theta}) = \theta$ and $V(\hat{\theta}) = 0$
2. As $n \to \infty$, $\hat{\theta}$ approaches ideal distribution
    a. That is, $E(\hat{\theta}) \to \theta$ and $V(\hat{\theta}) \to 0$

    Put differently, $\hat{\theta}_n \to \theta$ ("in probability")
3. Example: $\hat{\mu} = \bar{X}$ is consistent estimator of μ
    a. $E(\hat{\mu}) = \mu$ for all n
    b. $V(\hat{\mu}) = \frac{\sigma^2}{n} \to 0$

4. Law of large numbers (Jacob Bernoulli, 1713)
    a. Sample means converge to population means
    b. Higher order moments
        i. $E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^3\right) = E\left(X_i^3\right)$
        ii. $V\left(\frac{1}{n}\sum_{i=1}^{n} X_i^3\right) = \frac{V(X_i^3)}{n} \to 0$
        iii. Sample moments converge to population moments (justification for MOM)
5. Fact: continuous functions of consistent estimators are consistent
6. Fact: MLE are always consistent

Bias

1. Bias $B\left(\hat{\theta}\right) = E\left(\hat{\theta} - \theta\right) = E\left(\hat{\theta}\right) - \theta$
2. On average, does $\hat{\theta}$ produces estimates that are higher or lower than θ?
3. Unbiased estimator: $E\left(\hat{\theta}\right) = \theta$
4. $\bar{X}$ is *unbiased* estimator of μ because $E(\bar{X}) = \mu$
5. Example of biased estimation procedure: sample max from uniform distribution
6. When bias can be measured, can sometimes correct (target analogy)

(Relative) Efficiency

5. Given two estimators, the one with lower variance is more efficient.
6. An estimator cannot be efficient, per se, but only more efficient than another estimator. In some cases in Econ 388, however, it is possible to prove categorically that a particular unbiased estimator is more efficient than any other unbiased estimator.
7. Example: consider throwing out one observation, computing sample average of $n - 1$ observations
    a. $E(\tilde{\mu}) = \mu$ still
    b. $V(\tilde{\mu}) = \cdots = \frac{\sigma^2}{n-1}$
    c. Still unbiased, still consistent, but less efficient than using all available data

Sample Variance

1. $\hat{\sigma}_{MOM}^2$ is biased
    a. $\hat{\sigma}_{MOM}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \cdots = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$

b. $E(\hat{\sigma}^2_{MOM}) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i^2\right) - E(\bar{X}^2)$

$= \frac{1}{n}\sum_{i=1}^{n}(\mu^2 + \sigma^2) - \left(\mu^2 + \frac{\sigma^2}{n}\right)$

(since $\sigma^2 = V(X_i) = E\left(X_i^2\right) - \mu^2$ and $\frac{\sigma^2}{n} = V(\bar{X}) = E(\bar{X}^2) - \mu^2$)

$= \mu^2 + \sigma^2 - \mu^2 - \frac{\sigma^2}{n} = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$

c. $B(\hat{\sigma}^2_{MOM}) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$

d. Still consistent: $B(\hat{\sigma}^2_{MOM}) \to 0$ (and can show that $V(\hat{\sigma}^2_{MOM}) \to 0$)

2. "Sample variance" $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

a. To eliminates bias: $E\left(\frac{n}{n-1}\hat{\sigma}^2_{MOM}\right) = \frac{n}{n-1}E(\hat{\sigma}^2_{MOM}) = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2$

b. So if $S^2 = \frac{n}{n-1}\hat{\sigma}^2_{MOM} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$then $B(S^2) = 0$

i. Example: sample of $n = 4$ student wages, $x_i = \$11, \$10, \$14, \$15$, $\bar{x} = \$13.50$,

$\hat{\sigma}_{MOM} = \sqrt{\frac{1}{4}(1.5^2 + 2.5^2 + 3.5^2 + .5^2)} = \sqrt{\frac{21}{4}} \approx \$2.29$

$s = \sqrt{\frac{1}{3}(1.5^2 + 2.5^2 + 3.5^2 + .5^2)} = \sqrt{\frac{21}{3}} \approx \$2.65$

ii. Excel: use VAR.S or STDEV.S, not =VAR.P or =STDEV.P

c. Correcting bias actually sacrifices some efficiency

L18 Difference in Means, Proportions (WMS 8.3-8,10.3)

1. [Long lecture; use time efficiently.]
2. Difference in means

a. Relating quantitative and binary variables: conditional distributions, conditional means $E(Y|X = 0)$, $E(Y|X = 1)$

b. Wages gap between men and women:

$n_w = 40$, $\bar{x}_w = \$32$, $\sigma_w = \$13$, $n_m = 45$, $\bar{x}_m = \$35$, $\sigma_m = \$15$.

c. 95% confidence intervals for men $[\$30.62, \$39.38]$ and women $[\$27.97, \$36.03]$ overlap, making it difficult to tell true size of wage gap (if any)

d. Trick (used a lot in more advanced settings): combine into single parameter $\theta = (\mu_m - \mu_w)$; MOM estimator $\hat{\theta} = (\bar{X}_m - \bar{X}_w)$

i. $E\left(\hat{\theta}\right) = E(\bar{X}_m - \bar{X}_w) = E(\bar{X}_m) - E(\bar{X}_w) = \mu_m - \mu_w = \theta$; unbiased!

ii. $V(\hat{\theta}) = V(\bar{X}_m - \bar{X}_w) = \frac{\sigma_m^2}{n_m} + (-1)^2 \frac{\sigma_w^2}{n_w} \to 0$; consistent (as long as both sample sizes grow large)!

iii. $\bar{X}_m \sim N\left(\mu_m, \frac{\sigma_m^2}{n_m}\right)$ and $\bar{X}_w \sim N\left(\mu_w, \frac{\sigma_w^2}{n_w}\right)$, so…

iv. $\bar{X}_m - \bar{X}_w \sim N\left(\mu_m - \mu_w, \frac{\sigma_m^2}{n_m} + \frac{\sigma_w^2}{n_w}\right)$

Standardizing, $\frac{\hat{\theta} - \mu_{\hat{\theta}}}{\sigma_{\hat{\theta}}} = \frac{(\bar{X}_m - \bar{X}_w) - (\mu_m - \mu_w)}{\sqrt{\frac{\sigma_m^2}{n_m} + \frac{\sigma_w^2}{n_w}}} \sim N(0,1)$

v. Note: if estimate s_A^2 and s_B^2 then $\frac{(\bar{X}_m - \bar{X}_w) - (\mu_m - \mu_w)}{\sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}} \sim t\left(df = \frac{\left(\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}\right)^2}{\frac{\left(\frac{s_m^2}{n_m}\right)^2}{n_m - 1} + \frac{\left(\frac{s_w^2}{n_w}\right)^2}{n_w - 1}}\right)$

1. (e.g. If $s_m = \$12$ and $s_w = \$10$ then $df \approx 83$)
2. For this class, just use $t(df) \approx N(0,1)$, which is appropriate when n_m and n_w are both large
3. (df between minimum and sum of $(n_m - 1)$ and $(n_w - 1)$)

e. Margin of error: $\pm 2\sqrt{\frac{\sigma_m^2}{n_m} + \frac{\sigma_w^2}{n_w}} = 2(\$1.98) = \$3.96$

f. 95% confidence interval for $(\mu_m - \mu_w)$: $(\bar{X}_m - \bar{X}_w) \pm 1.96\sqrt{\frac{\sigma_m^2}{n_m} + \frac{\sigma_w^2}{n_w}} = [\$0.11, \$7.89]$

g. Test $\mu_m - \mu_w > 0$: test statistic $\frac{(\bar{X}_m - \bar{X}_w) - (\mu_m - \mu_w)}{\sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}} = \frac{4-0}{\sqrt{\frac{144}{100} + \frac{100}{40}}} = 2.02$; p-value 0.0217

h. Test $\mu_m - \mu_w \neq 0$: p-value $2 \cdot 0.0217 = 0.0434$

i. Test $\mu_m - \mu_w > \$2$: test statistic $\frac{(\bar{X}_m - \bar{X}_w) - (\mu_m - \mu_w)}{\sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}} = \frac{4-2}{\sqrt{\frac{144}{100} + \frac{100}{40}}} = 1.01$; p-value 0.1562

j. Note: if we estimated $\mu_w - \mu_m$ instead of $\mu_m - \mu_w$, rejection region would be on left instead of right, and test statistics would be negative instead of positive, but produce same p-values

3. Binary data (i.e. Bernoulli(p))

a. Intuitive estimator: proportion $\hat{p} = \frac{Y}{n}$, where Y = # of 1's in data

i. Example: election survey, $n = 100$, $\hat{p} = \frac{52}{100} = .52$

b. MOM estimator: $\hat{p}_{MOM} = \bar{X}$; actually same, since $Y = \sum_{i=1}^{n} X_i$ (for zeros and ones, adding is the same as counting)

c. Since $Y \sim Bin(n, p)$,

$E(\hat{p}) = E\left(\frac{Y}{n}\right) = \frac{1}{n}E(Y) = \frac{np}{n} = p$; unbiased!

$V(\hat{p}) = \frac{1}{n^2}V(Y) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n} \to 0$; consistent!

d. By Central Limit Theorem, $\hat{p} = \bar{x} \to N\left(p, \frac{p(1-p)}{n}\right) \Rightarrow \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \to N(0,1)$

Note: this is not actually different from $\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$; just special case with $\bar{x} = \hat{p}$, $\mu = p$, and $\sigma^2 = p(1-p)$

Note: does not follow t distribution for small n, because X_i not normal

e. Example: election survey, $n = 100$, $\hat{p} = \frac{52}{100} = .52$

i. Margin of error: $2\sqrt{\frac{p(1-p)}{n}} \approx 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 2\sqrt{\frac{.52 \cdot .48}{100}} \approx 2(.05) = 0.1$

ii. 95% Confidence interval $\approx \hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = .52 \pm 1.96(.05) = [.422, .618]$

iii. Test $H_0: p = 5$ against $H_a: p > 5$: test statistic $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{.52-.5}{\sqrt{\frac{.5 \cdot .5}{100}}} = 0.40$; p-value 0.3446

4. Difference in proportions: unemployment in U.S. and France (2% difference?)

a. $n_F = 1000$, $\hat{p} = \frac{109}{1000} = .109$; $n_{US} = 500$, $\hat{p}_{US} = \frac{38}{500} = .076$

b. 95% confidence intervals $[.090, .128][.053, .099]$

c. Estimate $(p_F - p_{US})$ with MOM estimator $(\hat{p}_F - \hat{p}_{US})$

i. $E(\hat{p}_F - \hat{p}_{US}) = E(\hat{p}_F) - E(\hat{p}_{US}) = p_F - p_{US}$; unbiased!

ii. $V(\hat{p}_F - \hat{p}_{US}) = V(\hat{p}_F) + V(\hat{p}_{US}) = \frac{p_F(1-p_F)}{n_F} + \frac{p_{US}(1-p_{US})}{n_{US}} \to 0$; consistent!

iii. $\frac{(\hat{p}_F - \hat{p}_{US}) - (p_F - p_{US})}{\sqrt{\frac{p_F(1-p_F)}{n_F} + \frac{p_{US}(1-p_{US})}{n_{US}}}} \sim N(0,1)$

d. 95% Confidence interval $(\hat{p}_F - \hat{p}_{US}) \pm 1.96\sqrt{\left(\frac{\hat{p}_F(1-\hat{p}_F)}{n_F} + \frac{\hat{p}_{US}(1-\hat{p}_{US})}{n_{US}}\right)} = [.003, .063]$

e. Test $(p_F - p_{US}) > 0$, test statistic $\frac{(\hat{p}_F - \hat{p}_{US}) - 0}{\sqrt{\left(\frac{\hat{p}_F(1-\hat{p}_F)}{n_F} + \frac{\hat{p}_{US}(1-\hat{p}_{US})}{n_{US}}\right)}} \approx 2.14$; p-value .0162

f. Test $(p_F - p_{US}) > .02$, test statistic $\frac{(\hat{p}_F - \hat{p}_{US}) - .02}{\sqrt{\left(\frac{\hat{p}_F(1-\hat{p}_F)}{n_F} + \frac{\hat{p}_{US}(1-\hat{p}_{US})}{n_{US}}\right)}} \approx 0.84$; p-value .2005

L19 Variance Estimation (WMS 8.9,10.9)

Review

1. $\hat{\sigma}^2_{MOM} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
2. $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Variance estimation

1. Applications: inequality/heterogeneity, quality control, estimation error
2. $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1)$
3. Sample variance: $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1)$
    a. Intuition 1: $X_i \sim N(\mu, \sigma^2)$, so $\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(1)$; $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(n)$; we lose one degree of freedom because we're measuring deviations from $\bar{X}$ rather than deviations from μ
    b. Intuition 2: a single observation conveys information about μ but not σ^2, so if $n = 100$ then we have 100 pieces of information about μ but only 99 pieces of information about σ^2
    c. Intuition 3: $E(S^2) = \sigma^2$, so $E\left[(n-1)\frac{S^2}{\sigma^2}\right] = n - 1$
    d. [Skip] Formal derivation: $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{(X_i-\bar{X})+(\bar{X}-\mu)}{\sigma}\right)^2$

$$= \sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 + \sum_{i=1}^{n}\left(\frac{\bar{X}-\mu}{\sigma}\right)^2 + 2\sum_{i=1}^{n}\frac{(X_i-\bar{X})(\bar{X}-\mu)}{\sigma^2}$$

$$= (n-1)\frac{S^2}{\sigma^2} + \frac{n(\bar{X}-\mu)^2}{\sigma^2} + 2\frac{(\bar{X}-\mu)}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})$$

$$= (n-1)\frac{S^2}{\sigma^2} + \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} + 2\frac{(\bar{X}-\mu)}{\sigma^2}(n\bar{X} - n\bar{X})$$

$$\sim \chi^2(n-1) + \chi^2(1) + 0$$

e. Recall that, in estimating μ, using $\frac{\bar{X}-\mu}{\sqrt{\frac{s^2}{n}}}$ instead of $\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$ required the use of a t distribution instead of a normal. This is because it can be shown that $Z = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$ and $W = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1)$ are independent, implying that $\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}\frac{1}{\sqrt{\frac{S^2}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{W}{n-1}}} \sim t(n-1)$.

4. Example: variance among $n = 71$ sales representatives is $s^2 = 5.3^2$
   a. Confidence interval
      i. Seek $.95 = \Pr(\# < \sigma < \#)$ and know $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(70)$, so from Table 6,
      ii. $\Pr\left(48.76 < (n-1)\frac{S^2}{\sigma^2} < 95.02\right) = \Pr\left(\frac{1}{48.76} > \frac{\sigma^2}{(n-1)s^2} > \frac{1}{95.02}\right)$
      $= \Pr\left(\frac{(n-1)s^2}{48.76} > \sigma^2 > \frac{(n-1)s^2}{95.02}\right) = \Pr\left(\sqrt{\frac{(70)5.3^2}{48.76}} > \sigma > \sqrt{\frac{(70)5.3^2}{95.02}}\right)$
      $= \Pr(6.35 > \sigma > 4.55)$
   b. Hypothesis test
      i. $H_a: \sigma^2 > 4^2, \alpha = .05$
      ii. Critical value 90.53
      iii. Test statistic $(n-1)\frac{S^2}{\sigma^2} = 70\left(\frac{5.3^2}{4^2}\right) = 122.9$, reject H_0 (from Excel, p-value is 10^{-5})

L20 Regression Estimation (WMS 11.1-3)

1. Recall from Lecture 9
   a. Relationship between X and Y can be represented by $\rho = corr(X,Y)$ or by regression line $Y = \beta_0 + \beta_1 X + \varepsilon$
   b. $E(\varepsilon) = 0$ can be guaranteed by choosing intercept to solve $\mu_y = \beta_0 + \beta_1 \mu_x$
   c. Crystal ball: can predict $Y^* = \beta_0 + \beta_1 x^*$ for any x^* value (even out of sample)
   d. σ_ε^2 can be minimized by choosing slope coefficient $\beta_1 = \frac{\sigma_{xy}}{\sigma_x^2} = \rho\frac{\sigma_y}{\sigma_x}$
   e. Fraction of variation in Y associated with X is $\frac{V(\beta_0+\beta_1 X)}{V(Y)} = \frac{\beta_1^2\sigma_x^2}{\sigma_y^2} = \frac{\left(\rho^2\frac{\sigma_y^2}{\sigma_x^2}\right)\sigma_x^2}{\sigma_y^2} = \rho^2$
2. Estimation

a. $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

b. $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

c. $r = \frac{s_{xy}}{s_x s_y}$

d. $\widehat{\rho^2} = r^2$

e. $\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

f. $\hat{\beta}_0 = \bar{y} - b_1\bar{x}$

g. Income predictions $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

h. Individual errors $\hat{\varepsilon}_i = y_i - \hat{y}_i$

i. $s_\varepsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$

3. Can also use "sums of squares", rather than variance (i.e. "total" not "average" deviations)

a. Let $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$

b. Let $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

c. With this notation, $\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{S_{xy}}{S_{xx}}$

d. Let $S_{\varepsilon\varepsilon} = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$

4. Example : Regress Income y (in $ thousands) on Education x (in years)

x_i	$x_i - \bar{x}$	y_i	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$\hat{\beta}_0 + \hat{\beta}_1 x_i$	$\hat{\varepsilon}_i$
11	-4	40	-30	120	37.2	2.8
16	1	80	10	10	78.2	1.8
16	1	70	0	0	78.2	-8.2
14	-1	60	-10	10	61.8	-1.8
18	3	100	30	90	94.6	5.4
$\bar{x} = 15$	$S_{xx} = 28$	$\bar{y} = 70$	$S_{yy} = 2000$	$S_{xy} = 230$		$\overline{\hat{\varepsilon}_i} = 0$
	$s_x^2 = 7$		$s_y^2 = 500$	$s_{xy} = 57.5$		$S_{\varepsilon\varepsilon} = 111$
	$s_x \approx 2.6$		$s_y \approx 22.4$	$r \approx 0.97$		$s_\varepsilon^2 = 37$
				$r^2 \approx 0.94$		$s_\varepsilon = 6.1$
				$\hat{\beta}_1 \approx 8.2$		
				$\hat{\beta}_0 \approx -53$		

a. We'll use this example again in subsequent lecture

b. Predictions

   i. High school graduate $\hat{y}^*_{x^*=12} = -53 + 8.2(12) = 45.4$

   ii. College graduate $\hat{y}^*_{x^*=16} = -53 + 8.2(16) = 78.2$

   iii. PhD graduate $\hat{y}^*_{x^*=20} = -53 + 8.2(20) = 111$

c. Estimated errors

   i. Predict income $\hat{y}_i$ for each individual in sample

   ii. $\hat{\varepsilon}_i = y_i - \hat{y}_i$

      1. Individual

   iii. $\mathrm{S}_{\varepsilon\varepsilon} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = 111$

      1. Alternatively, $\sigma_\varepsilon^2 = (1-\rho^2)\sigma_y^2$, so $SSE = (1-r^2)S_{yy} \approx (1-.9446)(2000) = 111$ (useful if only know summary statistics for X and Y).

   iv. $s_\varepsilon^2 = \frac{1}{n-2} SSE = 37$, $s_\varepsilon = \sqrt{37} \approx 6.1$

d. Illustrate with Excel: Data>Data Analysis>Regression, using education & income data above

5. Preliminaries

a. Algebra trick 1: It can be shown that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

$$= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0$$

Similarly, $\sum_{i=1}^{n}(Y_i - \bar{Y}) = 0$

b. Algebra trick 2: It can be shown that

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})x_i$$
$$= \sum_{i=1}^{n}(x_i - \bar{x})\, x_i - \sum_{i=1}^{n}(x_i - \bar{x})\, \bar{x}$$
$$= \sum_{i=1}^{n}(x_i - \bar{x})x_i - \bar{x}\sum_{i=1}^{n}(x_i - \bar{x})$$
$$= \sum_{i=1}^{n}(x_i - \bar{x})x_i$$

Similarly, $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})Y_i$ and $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})Y_i$

c. Deterministic x_i

   i. You can think of X_i and Y_i as being random (i.e. they depend on who you interview), and this is what we did when we derived the population regression parameters. But for simplicity, assume in the estimation that $X_i = x_i$ are

known. That is, we are only considering various samples of n individuals who have the same education levels as the people we sampled today (and incomes that potentially differ from the people we interviewed).

ii. If an estimator is unbiased conditional on these x_i's, it is also unbiased unconditionally. For example, if $E(\hat{\beta}_1 | X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \beta_1$ for every sample of x's then, averaging across all such samples, $E(\hat{\beta}_1) = \beta_1$ as well.

iii. Y_i is still random because $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ and ε_i is random.

L21 Regression Inference (WMS 11.4-9)

Introduction

1. Long lecture (use time efficiently)
2. We've derived estimators $\hat{\beta}_1$, $\hat{\beta}_0$, $\hat{Y}^*$, but so far all we have are point estimates. Are these good estimators (i.e. unbiased and consistent)? What are the margins of errors? To get confidence intervals or do hypothesis tests, we need to know their distributions.
3. Estimator distributions: if $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ (which is plausible, by Central Limit Theorem, if each error term is viewed as the sum total of a lot of smaller, independent factors) then
    a. $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \sim N$
    b. $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim N$
    c. $\hat{\beta}_1 = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})Y_i \sim N$
    d. $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \sim N$
    e. $\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^* \sim N$
    f. $Y - \hat{Y}^* \sim N$
    g. Estimation error $\hat{\varepsilon}_i = Y_i - \hat{Y}_i \sim N$
    h. $\frac{(n-2)s_\varepsilon^2}{\sigma_\varepsilon^2} \sim \chi^2(n-2)$ (essentially because estimating $\hat{\varepsilon}_i$ requires estimating two parameters $\hat{\beta}_0$ and $\hat{\beta}_1$, leaving only $n-2$ pieces of information)
    i. Could compare s_ε^2 from two regressions to see which better explains Y, using F distribution

Slope estimator

1. It can be shown that $E(\hat{\beta}_1) = \beta_1$; unbiased ☺!

$E(\hat{\beta}_1) = E\left[\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})Y_i\right] = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})[\beta_0 + \beta_1 x_i + E(\varepsilon_i)]$

$= \frac{1}{S_{xx}}[0\beta_0 + \beta_1 \sum_{i=1}^{n}(x_i - \bar{x})x_i] = \frac{S_{xx}}{S_{xx}}\beta_1 = \beta_1$

2. $V(\hat{\beta}_1) = V\left[\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})Y_i\right] = \frac{\sigma_\varepsilon^2}{S_{xx}} = \frac{\sigma_\varepsilon^2}{(n-1)s_x^2} \to 0$; consistent ☺!

$= \frac{1}{S_{xx}^2}V[\sum_{i=1}^{n}(x_i - \bar{x})Y_i] = \frac{1}{S_{xx}^2}[\sum_{i=1}^{n}(x_i - \bar{x})^2 V(Y_i) + 0]$

$= \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i - \bar{x})^2(0 + 0 + \sigma_\varepsilon^2) = \frac{\sigma_\varepsilon^2}{S_{xx}} = \frac{\sigma_\varepsilon^2}{(n-1)s_x^2}$

Note: most noisy when incomes more varied (conditional on education); least noisy when education more varied (s_x^2 in denominator)

3. $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma_\varepsilon^2}{S_{xx}}}} \sim N(0,1)$; therefore, $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{s_\varepsilon^2}{S_{xx}}}} \sim t(n-2)$

4. Example
    a. From previous lecture, $n = 5$, $s_\varepsilon^2 = 37$, $S_{xx} = 28$
    b. 95% Confidence interval: $\$8.2k \pm 3.182\sqrt{\frac{37}{28}} = [\$4.5k, \$11.9k]$
    c. Hypothesis Test $H_a: \beta_1 > \$5k$ at $\alpha = .05$ level
        i. Critical value 2.353
        ii. Test statistic $\frac{8.2-5}{\sqrt{\frac{37}{28}}} \approx 2.78$; reject H_0

Intercept estimator

1. It can be shown that $E(\hat{\beta}_0) = \cdots = \beta_0$; unbiased ☺!

   It can be shown that $V(\hat{\beta}_0) = \sigma_\varepsilon^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right) \to 0$; consistent ☺!

    a. [For those curious,

$E(\hat{\beta}_0) = E\left(\frac{1}{n}\sum_{i=1}^{n}Y_i - \hat{\beta}_1\bar{x}\right) = \frac{1}{n}\sum_{i=1}^{n}[\beta_0 + \beta_1 x_i + E(\varepsilon_i)] - \bar{x}E(\hat{\beta}_1)$

$= \frac{n\beta_0}{n} + \beta_1\frac{1}{n}\sum_{i=1}^{n}x_i - \bar{x}\beta_1 = \beta_0$

$V(\hat{\beta}_0) = V(\bar{Y} - \hat{\beta}_1\bar{x})$

$= V(\bar{Y}) + \bar{x}^2 V(\hat{\beta}_1) - 2\bar{x}Cov(\bar{Y}, \hat{\beta}_1)$

$= \frac{\sigma_\varepsilon^2}{n} + \bar{x}^2\frac{\sigma_\varepsilon^2}{S_{xx}} - 0 = \sigma_\varepsilon^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right) \to 0$

Note: $C(\bar{Y}, \hat{\beta}_1) = 0$ because…

$$C\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \frac{1}{S_{xx}}\sum_{j=1}^{n}(x_i-\bar{x})Y_j\right) = \frac{1}{nS_{xx}} C\left(\sum_{i=1}^{n} Y_i, \sum_{j=1}^{n}(x_i-\bar{x})Y_j\right)$$

$$= \frac{1}{nS_{xx}}\left[\sum_{i=1}^{n}(x_i-\bar{x})C(Y_i,Y_i) + (x_i-\bar{x})\sum_{i\neq j} C\left(Y_i,Y_j\right)\right]$$

$$= \frac{1}{nS_{xx}}\left[\sum_{i=1}^{n}(x_i-\bar{x})V(Y_i) + (x_i-\bar{x})\sum_{i\neq j} 0\right]$$

$$= \frac{1}{nS_{xx}}\left[\sigma_y^2\sum_{i=1}^{n}(x_i-\bar{x})\right]$$

$$= \frac{\sigma_y^2}{nS_{xx}}[0] \qquad]$$

2. Note two pieces: small error in identifying (μ_x, μ_y) and larger error in identifying slope (which matters more when $\bar{x}^2$ high).
3. $\frac{\hat{\beta}_0-\beta_0}{\sqrt{\sigma_\varepsilon^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} \sim N(0,1)$; therefore, $\frac{\hat{\beta}_0-\beta_0}{\sqrt{s_\varepsilon^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} \sim t(n-2)$

Prediction estimator

1. $\widehat{(\beta_0+\beta_1}x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$
2. $E\left(\widehat{\beta_0+\beta_1}x_i\right) = E\left(\hat{\beta}_0+\hat{\beta}_1 x_i\right) = \beta_0+\beta_1 x_i$; unbiased ☺!
3. $V\left(\widehat{\beta_0+\beta_1}x_i\right) = \cdots = \sigma_\varepsilon^2\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right) \to 0$; consistent ☺!

   $[V\left(\widehat{\beta_0+\beta_1}x_i\right) = V\left(\hat{\beta}_0\right) + x_i^2 V\left(\hat{\beta}_1\right) + 2Cov\left(\hat{\beta}_0, \hat{\beta}_1 x_i\right)$

   $= \sigma_\varepsilon^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right) + (x_i)^2\frac{\sigma_\varepsilon^2}{S_{xx}} - 2x_i\bar{x}\frac{\sigma_\varepsilon^2}{S_{xx}}$ (since $Cov\left(\hat{\beta}_0,\hat{\beta}_1\right) = Cov\left(\bar{Y}-\hat{\beta}_1\bar{x},\hat{\beta}_1\right) = 0 - \bar{x}\frac{\sigma_\varepsilon^2}{S_{xx}}$) =

   $\sigma_\varepsilon^2\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right) = \sigma_\varepsilon^2\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right)]$

   Note: most precise close to $\bar{x}$; can still make predictions far away from $\bar{x}$, but more noisy
4. $\frac{\widehat{(\beta_0+\beta_1}x_i)-(\beta_0+\beta_1 x_i)}{\sqrt{\sigma_\varepsilon^2\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right)}} \sim N(0,1)$

   $\frac{\widehat{(\beta_0+\beta_1}x_i)-(\beta_0+\beta_1 x_i)}{\sqrt{s_\varepsilon^2\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right)}} \sim t(n-2)$

Error variance

1. $\frac{(4)s_\varepsilon^2}{\sigma_\varepsilon^2} \sim \chi^2(4)$, $s_\varepsilon = 6.1$
2. 95% confidence interval for σ_ε
   a. $.95 = P\left(.48 < \frac{(4)s_\varepsilon^2}{\sigma_\varepsilon^2} < 11.14\right) = P\left(\sqrt{\frac{4\cdot 37}{.48}} > \sigma_\varepsilon > \sqrt{\frac{4\cdot 37}{.48}}\right) = P(17.6 > \sigma_\varepsilon > 3.6)$

3. Test $H_a: \sigma_\varepsilon^2 > 4^2$ at $\alpha = .05$ level
    a. Critical value 9.49 (from Table 6)
    b. Test statistic $\frac{4 \cdot 37}{4^2} = 9.25$; not significant
4. If you had two regressions and wanted to know which has better predictive power (i.e. lower residual error variance) you could compare $\sigma_{\varepsilon A}^2$ and $\sigma_{\varepsilon B}^2$ using F distribution

Review

1. Thanks for a great semester!
2. Thanks TAs!
3. Recommend Econ 388: regression with multiple variables
    a. We're on the brink of knowledge explosion
    b. Also Econ 210 for exploring careers in Economics
    c. For advanced statistics/econometrics, recommend linear algebra (Math 213)
4. Student project findings
    a. Wide variety of projects
    b. Value of statistics for consumers, not just producers
5. Key concepts
    a. We've seen several trees, now let's notice the forest
    b. Pre-midterm, we discussed population distributions (discrete or continuous), including moments such as mean, variance, and covariance.
    c. From sample, we seek to estimate population parameter: μ, σ, p, ρ, $\mu_1 - \mu_2$, $p_1 - p_2$, $\frac{\sigma_1^2}{\sigma_2^2}$, β_0, β_1, $y^* = \beta_0 + \beta_1 x^*$, or θ from another distribution function (e.g. a, b from uniform, θ from exponential), including distributions we haven't encountered yet (e.g. p from geometric, λ from Poisson, β_2 from quadratic regression)
    d. Deriving estimators: MOM and ML
    e. Properties of estimators: bias, efficiency, consistency
    f. Margin of error, confidence intervals, hypothesis tests
    g. Matrix algebra

i. This class has focused on the correlation ρ between two variables, which also underlies linear regression line: slope coefficient $\rho\frac{\sigma_y}{\sigma_x}$ and coefficient of determination ρ^2

ii. Matrix algebra allows us to generalize everything to multiple variables (e.g. $\boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ instead of $\beta = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$

iii. Individual slope coefficient β_1 then reflects *partial* correlation between education and income, holding experience, age, race, gender, etc. fixed.

iv. Controlling for more variable makes stronger case for correlation as causation

6. Deriving estimator distributions

a. Estimators depend on $X_1, X_2, \dots, X_n$, and so are random variables with distributions

b. $E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}(n\mu) = \mu$

c. $V(\bar{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n}(n\sigma^2) = \frac{\sigma^2}{n}$

d. Distribution $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ or $\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$ (by normality of X_i or Central Limit Theorem)

e. CLT also implies that $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$

f. Difference estimators $\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$ and $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0,1)$

g. X_i normal implies $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1)$

and therefore $\frac{S_A^2/\sigma_A^2}{S_B^2/\sigma_B^2} \sim F(n_A-1, n_B-1)$ and $\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim t(n-1)$

h. Similarly, $E(\hat{\beta}_1) = \cdots = \beta_1$, $V(\hat{\beta}_1) = \cdots = \frac{\sigma_\varepsilon^2}{S_{xx}}$

and ε_i normal implies that $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_\varepsilon^2}{S_{xx}}\right)$ or $\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\sigma_\varepsilon^2}{S_{xx}}}} \sim N(0,1)$

i. $(n-2)\frac{S_\varepsilon^2}{\sigma_\varepsilon^2} \sim \chi(n-2)$, implying that $\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{S_\varepsilon^2}{S_{xx}}}} \sim t(n-2)$

7. Matrix algebra True/False question tip

a. Check simple cases (e.g. 1×1, 2×2), but not special cases
b. Example: "All symmetric matrices are idempotent"
    i. Try simple example: not special matrix like identity matrix or $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, but something with no special properties other than symmetry, such as $\begin{pmatrix} 1 & 3 \\ 3 & 2 \end{pmatrix}$ (and show that $\begin{pmatrix} 1 & 3 \\ 3 & 2 \end{pmatrix}\begin{pmatrix} 1 & 3 \\ 3 & 2 \end{pmatrix} = \begin{pmatrix} 10 & 9 \\ 9 & 13 \end{pmatrix} \neq \begin{pmatrix} 1 & 3 \\ 3 & 2 \end{pmatrix}$)
    ii. Can also try really simple examples, such as scalar matrices: $(5)(5) \neq (5)$.
c. T/F questions are not "trick" questions, but be careful. In real world, much of statistics is simply a matter of careful attention to details, and knowing exactly which inferences are legitimate under exactly which circumstances.

8. Example problems
    a. Confidence interval for $\hat{Y}^*$
    b. Hypothesis test for $\hat{p}_1 - \hat{p}_2$