Intro to Statistical Reasoning

Yiqiao YIN
Statistics Department
Columbia University
Notes in LaTeX

April 27, 2018

Abstract

This document presents notes from Intro to Statistical Reasoning.

This note is dedicated to Professor Ronald Neath.

Notes

Jan. 16, 2018

Discussion of the heart disease. From such example, we discuss the notions of controlled experiments and observation studies. To conduct such experiment, we discussed the aspirin group (people who take asprin) and placebo group (people who take some pills but the pills, unknown to the people, do nothing to their bodies).

Jan. 18, 2018

Example 0.0.1. Is BMI a valid measure of overweight? Is it a reliable measure? Is it an unbiased measure? Answer: no. It depends.

Definition 0.0.2. A valid measure is one that accurately measures what it claims to measure. A reliable measure is one that gives consistent results if repeated measures taken on same individual.

A bias is a systematic prejudice in one direction. A measure is unbiased if it does not have any bias. Note validity definition has two parts: what you are measuring; and what you are using it for. Is BMI a valid measure of overweight?

Reliability and unbiasedness refer more to instrument used to take measurements. A person who weighs 100 kg takes 6 weights an each of 4 scales. Results:

(1)	106	106	105	104	105	104
(2)	92	94	93	100	93	100
(3)	99	100	97	96	104	103
(4)	100	99	101	99	101	100

(1) is reliable, meaning more consistency. but biased. The reason is consistently 5 kg higher. That is, bias is systematic tendency to error in same direction. Poor reliability just means high variability between measures.

(2) is unreliable and biased. The readings are consistently lower than what we have known to be the truth.

(3) is unbiased but unreliable.

(4) is reliable and unbiased.

Note: Each of these concepts is truly a matter of degree, not an all-or-nothing proposition.

In most real world systems whatever measurement error results from biased and/or unreliable instruments is of less importance, then the natural variability between different individuals. Measurement error means true value minus observed value. Natural variability refers to the differences between the (true) values for different individuals.

Example 0.0.3. Half of subjects are given aspirin and half are given placebo. In the end of studying period, Blood pressures are recorded? What's natural variability? What created variability?

Definition 0.0.4. Natural variability means even within a treatment group, subjects blood pressure (the test or the goal in the experiment) will vary.

Definition 0.0.5. Created variability means the differences between average blood pressure (the goal or tests of the experiment) in aspirin group and average blood pressure in placebo group.

We will say there is statistical evidence of a treatment effect if created variability is big relative to natural variability.

At this point in the course we do not have the tools for a precise definition of "big relative to" but that's okay, the idea should still make sense.

More definitions include validity, reliability bias, measured variability, natural variability, created variability.

Definition 0.0.6. A variable is a quantity we can measure which takes different values for different units.

That being said, political ideology is not a variable. Political party affiliation is. Categorical is contrast to measurement. The primer means qualitative view whilst the latter is quantitative. Categorical can be normal or ordered. Measurement can be discrete or continuous.

Example 0.0.7. Consider your present living space, size in square footage, and number of persons with when you share it. Do you own the property or rent? First two are quantitative. Second one is discrete. First one is continuous.

Definition 0.0.8. X is discrete if we can list the possible values it assumes. X is continuous if it takes values in a measure.

Example 0.0.9. Board game involves two dice and a spinner (6 speed, # 1-6). X is the result of dice roll. Y the angle between spinner and x-axis. X takes values 2, 3, 4, ..., 12, so this is discrete. Y takes values between 0 and 360.

Example 0.0.10. Your intended major: LS, PS, SS, LA. Your attained years in school Fr, So, Jr, Sr, other. They are both categorical. Major is nominal. Years in school is ordinal. These distributions are not always clear.

Example 0.0.11. Survey taker asks years of schooling completed. Then we convert to education level attained HS, HS, HS + college, college, degree. First is measurable and it is discrete. Second is categorical and it would be ordinal.

Jan. 23, 2018

Continue from last class, we discussed measurement error.

Example 0.0.12. Consider a person who weighs 100 kg that steps on a scale. Measure value is true value + measurement error. That is, MV = TV + ME = TV + (MV - TV) = TV + (IA - TV) + (MV - IA). Thus, we have MV = TV + (IA - TV) + (MV - IA).

Thus, two distinct sources of measurement error are (1) first term is due to bias; (2) second term is due to ”unreliability”.

Example 0.0.13. Experiment. Treatment is fever reducing drug (A or B). Response is change to body temperature (the lower the better). Notice that NV refers to different outcomes within each treatment group. CV refers to the difference between the two groups. If $\underline{\text{CV}}$ is $\underline{\text{big}}$ relative to $\underline{\text{NV}}$, then there is evidence of a real difference between the drugs.

The first situation with high CV and low NV is that we have two bell curves A and B with peaks far away from each. The situation low CV and high NV is that we have A and B (two bell-shape curves) with peaks close to each other. In which case is there greater evidence of a treatment effect? The answer to this question is the first situation (peaks far away).

Though its possible that second situation gives stronger statistical evidence of a treatment effect if $\underline{\text{second situation comes with greater sample size}}$ (image second is done with a lot bigger sample size than the former situation).

Today we continue to Chapter 4 of textbook by Uttz [1]. However, in Chapter 25, we will discuss meta-analysis (study on tons of studies, combining results from a lot of studies). Case studies (not covered and there is no role for statistics).

Let us start sample surveys with some terminologies.

Definition 0.0.14. Unit is an individual or object to be measured. Population is the collection of units for which we desire this measure. Sample is the collection of units for which we obtain this measure. If sample is the same as population, let us call it census.

Sampling frame is the list of units from which sample is chosen.

Why sample? Consider a population of size 83 million people. We want to know what percent approve of President Trummp? If I ask 8.5 million people I'll know the exact answer. If I ask 1600 people, I'll be able to say I am 95% confident that Trumps approval rating in NYC is between LB and UB. Here LB and UB is just two numbers that creates a bound (where UB - LB $\leq$ 0.05). The fact that we say exactly 95 percent confident is explained in Chapter 20.

The number (UB-LB)/2, or a 0.025 in the example is called the margin of error. LB = estimate - ME; UB = estimate + ME.

$$\text{95 percent truth is between LB and UB}$$

$$\Leftrightarrow \text{95 percent confident our estimate is within ME of the truth}$$

What I said there was not right? What's wrong with it? It was mentioned that "if I ask 1600 people", but this is not good enough. We should have said "ask 1600 people by random".

Simple random sample (SRS) means we are imagine a giant urn with 8.5 million marbles and shake it enough so that this urn has marbles well mixed enough. This is the idea that we want to think about when we say randomly choose 1600 people.

Other sampling methods:

Example 0.0.15. Population is 76 for this stats class. We have CC = 27, GS = 28, BC = 13, SP = 8. Take a sample of 17 students. In my sample I get CC = 6, GS = 5, BC = 3, and SP = 3. GS is 28/76 = 37% of population but only 5/17 = 29% of sample. SP is 8/16 = 10% of population, but 3/17 = 18% of sample. If a stratified random sample we would take 4 separate SRSs, one from each school. Here the class is the population. Each college is a stratum. The 4 colleges are the strata.

Example 0.0.16. Stratified sampling with proportional sampling sizes: CC = 6, GS = 6, BC = 3, SP = 2. We are guaranteed sample is representative of population at least with respected to "school". This is why stratified sampling is the best one we know.

Moreover, we have cluster sampling. They both involve dividing population into subsets, but how they proceed is exactly the opposite. In stratified sampling we sample from each group. In cluster sampling we sample just a few of the groups (then survey every member of sampled group.

Example 0.0.17. 50 stacks of 10 papers as population. Stratified sample: randomly pick 4 out of 10 from each stack. Cluster sample: randomly pick 2 of the 5 stacks. Look at all 10 papers in selected stacks. Stratified sampling is effective when units are (1) homogeneous within strata, (2) heterogeneous between strata. Cluster sampling is most useful when each cluster is a representative cross section of population. This is usually done for convenience and does not offer precision.

Example 0.0.18. Sample residents of a dorm choose floors at random, and survey everyone on those floors.

There are the following difficulties and disasters. (1) sampling frame is not exactly the population (For example, population is students in this class today; sampling frame is class roster from yesterday); (2) selected individuals cannot be reached; (3) reached but decline to participate. There are statistical methods for dealing with low response rates. Low response rate is a difficulty, if not handled correctly could be disaster. There are, however, some potential disaster as well.

Voluntary response sampling is a horrible idea. Imagine the Chair came to professor indicating he is horrible teacher. If 3 or 4 students showed up and try to help out, that does not disprove the Chair's idea. The reason is that voluntary respondents are never a representative sample. Go to 501 NWC (a building) and ask everyone coming out of that class. That is a voluntary sampling as well.

Jan. 25, 2018

Example 0.0.19. Apple maggots gets inside fruit, runs it. Put plastic bag over apple. The question is does bagging affect the size of the apple at harvest?

Take an apple tree. Early in season identify 30 apples to use. Put a ziploc bag over 10 of them. Put a Baggy over 10 and leave 10 unbagged. Weigh each apple at harvest.

Questions:
(1) What is explanatory variable in this study?
(2) What is the response variable?
(3) What is/are the treatment(s)?
(4) Is this controlled experiment or observational study?
(5) What are some possible confounding variables?
(6) What are the experimental units? What are the observational units?
(7) How should we decide which applies get ziploc or the Baggy or unbagged?

Let us define the following:

Definition 0.0.20. The response variable is the measured outcome.

We are interested in how response differ for different levels of explanatory variable. The experimental units are the smallest basic object to which treatments can be assigned. The observational units are the smallest object on which response variable is measured.

Let us answer them. (1) The explanatory variable is to bagging the apples or not. This is a categorical variable with 3 levels (ziploc, bagging, or unbagged). (2) The response variable is the weight of the apples. This is a quantitative variable. (3) The treatments are the different levels of explanatory variables. See answer (1) for 3 levels. (6) Apples or both.

Example 0.0.21. A study in which observational unit is smaller than experimental unit. Educational studies often experimental units is a class. The observational unit is the students.

Reverse situation is possible too? If treatment is tutoring technique (applied to students in both classes), but only class argues are reliable. (4) This is a controlled experiment.

Definition 0.0.22. A confounding variable is an added variable such that (1) tends to vary for different levels of explanatory variables, and (2) different values of confounding variable. likely to result in different values of response variable., i.e. it is related to the explanatory variable and it affects the response variable.

For the rest of the question, (5) there might be confounding if we only put bags on the lowest fruit and put ziplocs on one and baggies on other. Notice that we need the experiment to be randomized. A proper randomization uses computer program (random number generator) or physical randomization (coin flips, draw marbles from can).

Let us discuss blocking on page 99 of text [1]. Experiment compare 3 oat varieties. The units are plots of land (12 of them). Each variety sown in 4 plots. Response variable is yield random assignment. Suppose the 12 plots are actually 4 plots split into 3 subplots each. Here we might instead sow each variety in one (randomly selected) subplot within each plot. This is called blocking. The plots are called blocks. This experiment uses randomized block design. Here is a particular realized block design vs. a particular complete randomization. Let us say we got (213), (321), (312), (231), and the other four blocks are (212), (133), (321), and (321). RBD (randomized block design) provides protection against confounding effect of blocking variable (further protection, beyond randomization).

Another way to think about it: we will only learn which variety is the best if created variable (different between variety yields) is big relative to natural variable (different yields for some variety). Blocking reduces natural variation. In the disciplined of experimental design, blocking is a variance reduction technique.

Interactions.

An experiment can have more than one explanatory variable.

Example 0.0.23. 18 cokes baked, 2 at each combination of

Time / Temp	335	350	365
32			
35			
38			

The response variable is taste rating. This experiment has two explanatory variables. Both categorical, 3 levels each. Thus there are 9 treatments total. Note balancing 38 min may yield the best cake at 335 degrees but the worst at 365 degrees. The effect time may differ depending on temperature. Equivalently, effect of temperature will differ for each baking time (turning 4 up to 365 may be good idea for 32 min but not 38 min). There is interaction effect between time and temperature.

Jan. 30, 2018

We are discussing observational study and controlled experiment today. For controlled experiment, researchers controls the explanatory variables. Observational studies we merely observe the explanatory variables.

Example 0.0.24. Aspirin and heart attack example introduced earlier. We have CE that is: given half the subjects aspirin, and half placebos (randomized, double blind). Follow for 5 to 10 years.

In a case control study, we take 1000 MI patients, say then find 1000 other similar subjects who did not just have a heart attack. Classify each subject according to aspirin-taking habits. Case-control studies are retrospective. Starting with outcome, add back to past.

A prospective study would follow group of subjects into the future, observe outcomes. Primary advantage of a case-control study is the following. Some advantages to case-control studies:

(1) saves time; completed in time, it takes to round up the subjects, do not have to waste around for actions.;

(2) allows consideration of explanatory variables that couldn't be in a CE;

(3) allows us to study a realistic range of behaviors with respect to explanatory variables;

(4) There is less interference, an OS allows study of real world behavior.

(5) More efficient allocation of resources. Does not require long time tracking of subjects.

(6) Because study participation is less burdensome, get better participation rate so lower selection bias.

(7) Efficiency gain. In a prospective study, you'll need a lot of subjects to observe a small number of heart attacks.

Now we present disadvantages. In a case-controlled study we do our best to account for confounding variables by thoughtful matching. But we can't think of everything. Observational study practically never justifies inference of causation.

A good case-controlled study will use matching including controls to be as otherwise similar to the cases as can reasonably be done. There is no way to use placebos in OS. There is no way to do blinding. Observed levels of explanatory variable not always clearly defined (note that this is the flip side of the coin to one of the strength of case-control studies). The cases in a case-control study may include subjects who wouldn't qualify for a CE in the first place. One more advantage to case-control studies is that it can do data mining! Potential explanatory variables in case-control study are almost unlimited.

Chapter 7

Now we move on to Chapter 7.

Part 1 covers Ch1-6: "Finding Data in Life"; Part 2 covers Ch7-13 in the same book. They introduced two main branches of statistics (data analysis). Descriptive stats: given a data set (which is just a branch of numbers), succinctly describe its main features. The richer area is inferential statistics. Given a data set (survey experiment, or other study), draw conclusions about the larger population from which these data are just a sample.

Part 2 hints at latter but is primarily focused on the former. Inference is taken up in Part 4 (Ch 19-22). Part 3 (Ch 14-18) is concerned with probability.

In descriptive statistics, we distinguish between graphical summaries of data. Draw a picture! Numerical summaries of data replace a long list of numbers with just a few key numbers that tell most of the story.

Example 0.0.25. Given data:

$$76, 81, 73, 77, 78, 83, 81, 75, 83, 85, 77, 80, 83, 77, 82, 87, 80$$

while $n = 17$. They are the heights in inches of NY Knicks basketball team players.

To get started, let us rearrange these numbers:

$$73, 75, 76, 77, 77, 77, 78, 80, 80, 81, 81, 81, 82, 83, 83, 83, 85, 87$$

graphical tools for summarizing data: two of them.

First is stemplot (aka "stem and leaf display").

$$7 :: 3, 5, 6, 7, 7, 7, 8$$

$$8 :: 0, 0, 1, 1, 2, 3, 3, 3, 5, 7$$

It would be more informative if we split the decades in half.

Another version of stemplot of these data would look:

$$7 :: 3,$$

$$7 :: 5, 6, 7, 7, 7, 8$$

$$8 :: 0, 0, 1, 1, 2, 3, 3, 3$$

$$8 :: 5, 7$$

This leads us to introduce histogram.

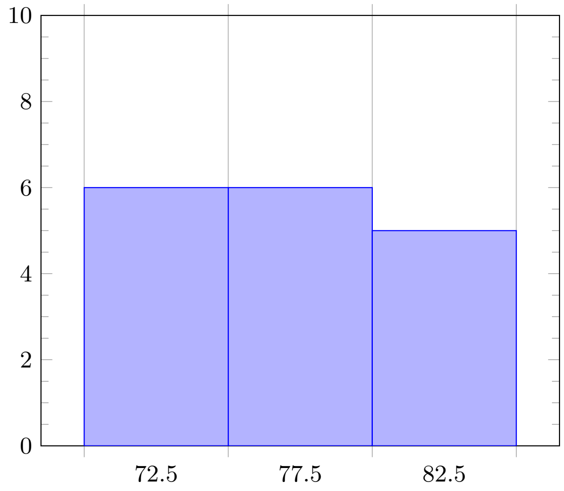
Definition 0.0.26. Histogram (3 steps).

(1) Split the data into ”bins” of equal width (can be anything; thus more flexible then stemplot).

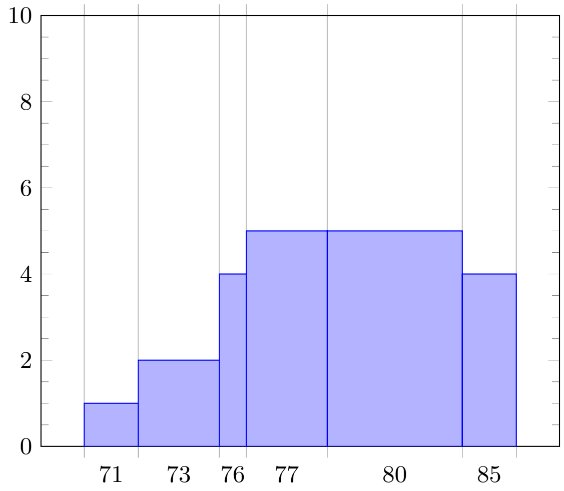
(2) For each bin, we count the number of units in that category;

(3) draw it. (a) Draw x-axis over range of data make marks at bin and points; (b) a block over each bin with height = number of units. Comment: it's often a good idea to define bin endpoints at 0.5's (tailored to the data set).

From the example above, we have



or we can do more bins:



Feb. 1, 2018

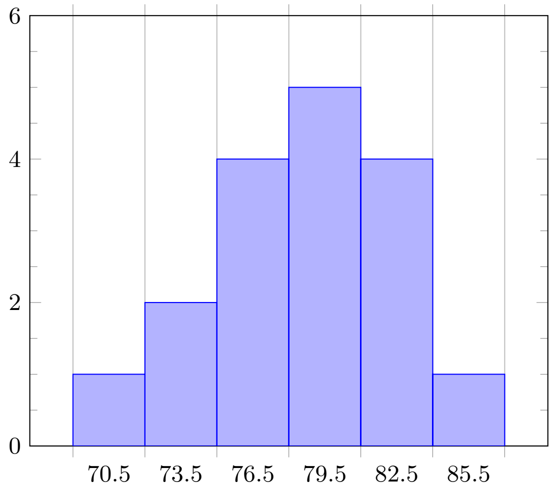
Let us discuss summarizing data from Chapter 7 of text [1].

Example 0.0.27. For $n = 17$, sort values are 73, 75, 76, 77, 77, 77, 78, 80, 80, 81, 81, 82, 83, 83, 83, 85, 87. We want to draw a histogram in three steps:

(1) Split data into bins of equal width; that is, we have bins [77,80), [80,83), and [83,86). However, we can also do: 70.5-73.5, 73.5-76.5, 76.5-79.5, 79.5-82.5, 82.5-85.5, 85.5-88.5.

(2) For each bin count numbers of units in that category;

(3) Draw it.



Next, we can describe a distribution from its histogram:

(1) Location or center of distribution;

(2) The variability or "spread" of the distribution;

(3) The shape of the distribution: (a) unimodel or bimodel, or, (b) symmetric or skewed, if skewed, in which direction, (c) any outline?

Note that (1) unimodel distribution has one hill, (2) bimodel distribution has two hills, and (3) trimodel distribution has three hills.

Next, we can discuss symmetry and skewness.

Definition 0.0.28. A distribution is symmetric if there is a point in the middle of the histogram such that LS and RS are mirror images of each other. There are three: (1) bell-shape curve, not to left or right, this is symmetric distribution, (2) skewed to the right, or positively skewed is a curve with long tail to the right, (3) skewed to the left, or negatively skewed is a curve with long tail to the left. Henceforth, by convention (arbitrarily), the direction of skewness is the side with the long tail. We will say a distribution is approximate symmetric.

Example 0.0.29. The distribution of heights is approximately symmetric/moderately skewed to the left. Left skewness makes sense here. Augmented height is 6'8" or 6'9".

Outliers. None in the data set. An outliers is a value that is far removed from the rest of the data. Outlierness is likewise a matter of degree, there is no hard and fast rule for what's an outliers.

Numerical summaries. Measuring the center of distribution. The mean of a distribution is the augmented value. If the data $x_1, x_2, ..., x_n$, the mean is $\bar{x} = (x_1 + \cdots + x_n)/n$. For example, we can look at the above data set and we can calculate the mean $(73 + 75 + \cdots + 88) = 79.3$, that is, $\bar{x} = 79.3$. The median is the middle value in the sorted data. Note: if n is an even number there are two middle values. Take their average and call that the medium. The median splits the distribution drawn the middle half the values are less than half are greater than the medium. The mean is the center of mass. If the x-axis is a ruler, and each case is little block, stacked at appropriate point. The mean is the point at which the ruler will balance on your finger. In a prefer/roughly symmetric distribution, the mean and median are exactly/approximate equal. In a right skewed distribution the mean is greater than the median. In a left skewed distribution the mean is less than the median. You want to choose median.

Example 0.0.30. Imagine scores for all but one students on an exam range from 20 to 40, but there is one student got 100. The mean will be affected (meaning going up), but the median will stay the same.

Example 0.0.31. Grading on a curve. Grading standard adjusted to reflect the difficulty of exam, exam performance gives measure of that difficulty, which makes more sense. Mean is B+ and work from there. Medium is B+ and work from there.

Measuring variability (spread). The usual measure is the standard deviation. Define the variance by

$$s^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

The standard deviation from the mean is $s = \sqrt{s^2}$.

Feb. 6, 2018

Recall the data set recording heights in inches of Nicks basketball players. The median is 80 inches (6'8"); and the mean is

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = 79.9$$

We can plot points which splits in half is 80. The center of mass (point at which histogram will balance) is 79.9. Note that this plot (dot plot) looks pretty symmetric! To be consistent is approximately referring to median. Left skewness in histogram we drew last time was an artifact of the bin selection.

Measuring variability. (spread). The usual measure is standard deviation. Define the variance by

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

which represents average squared distance from mean. Then the standard deviation is $s = \sqrt{s^2}$.

Example 0.0.32. For basketball data set, we calculate $s = 3.77$.

Discuss the following phenomenon. In a unimodel, roughly symmetric distribution. the mean and standard deviation tell the most of the story. In a skewed distribution its more complicated:

(1) mean can be misleading if there is outliers/long tail

(2) no single-number summary of spread, if the spread to the left is different from spread to the right. In such cases it makes more sense to report the 5-number summary. Min and Max are included. UQ is upper quartile. LQ is lower quartile.

Example 0.0.33. Consider

$$73, 75, 76, 77, 77, 77, 77, 78, 80, 80, 81, 81, 82, 83, 83, 83, 83, 85, 87$$

and we have

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
73.00	77.00	80.00	79.89	83.00	87.00

(1) Draw a box that spans LQ to UQ;
(2) Draw a line at median (will be inside box)
(3) Draw "whiskers" extending LQ to min and UQ to max.

The recipe above actually produces a "skeletal". Basketball, no outliers, thus skeletal box plot is boxplot. In a proper boxplot, the whiskers extend from LQ/UQ to the smallest/largest value that is not an outlier. Outliers are marked with a special symbol.

Definition 0.0.34. The interquartile range is IQR = UQ - LQ.

Definition 0.0.35. A case with with value x is a low outlier if $x < LQ - 1.5 \star IQR$ and high outlier if $x > UQ + 1.5 \star IQR$.

Example 0.0.36. In terms of assets, we record top 10 companies in United States.

$$12.6, 24.1, 45.7, 7.8, 5.1, 7.0, 5.3, 9.4, 13.0, 4.6$$

and we sort

$$4.6, 5.1, 5.3, 7.0, 7.8, 9.4, 13.0, 13.6, 24.1, 45.7$$

Skeletal box plot is graphical depiction of 5-number summary. Boxplot requires entire data set.

Now we move on.

Chapter 8

Chapter 8. Normal Distribution

A nice graph to summarize distribution of sample values is the frequency histogram. A frequency curve extends this idea in two ways. Imagine:

(1) the number of units n goes up to ∞;

(2) the bin width goes down to 0. Together (1) and (2) give us jagged histogram becomes a smooth curve;

(3) The whole thing is rescaled to have total area under curve AUC = 1. The AUC between any two parts represents the proportion (if total AUC = 1) or percentage (if total AUC = 100%).

We will use frequency curves to answer to kinds of questions: (1) given a value x, what percentage of population are less than or equal to x? (2) Given a proportion $0 < p < 1$, or percentage 100p, what is the value such that 100p% of population is less than or equal to that value? The second thing is called a percentile.

If you scored the 70th percentile on a test, you scored higher than 70% of the test takers and 30% scored higher than you. The two question: (1) given a score what's its percentile? (2) Given a desired percentile, what's score is required?

Feb. 8, 2018

Consider the cake recipe tasting ratings. We discuss interaction effect through this example.

-	litter butter	lotta butter
litte sugar	1	2
lotta sugar	2	1

Definition 0.0.37. Interaction: the effect of one explain variable differs depending on level of another.

Confounding occurs when the effect of one variable is indistinguishable from another. Good experimental design protects against confounding, its mainly an issue in observational studies.

Example 0.0.38. Insurance company has paid many claims. The claim payment distribution looks like bell-shape curve with shared area to be the proportion claims that were for an payment between 20k and 30k.

What does the height of the curve represent? Do not worry about it. The insurance example is right-skewed distribution as are many other real world phenomena. Not a concern today. But many other things are remarkably well represented by a unimodal, symmetric, bell-shaped curve called the normal distribution.

Example 0.0.39. Human height (sex-specific). For IQs. Cholesterol Gestation period. A normal distribution is completely characterized by its mean and sd.

If you know the population mean μ the population SD σ that the population distribution is normal. You know everything.

Example 0.0.40. The actual height of product in a 16-ounce cereal box follows a normal distribution with mean 16.3 and SD of 0.2 ounce.

(1) What percent boxes have at least 15.9 ounces?

(2) What percent are between 16.5 and 16.7?

(3) What is the value such that 16% of the boxes have less 84% have more, i.e. what is the 16th percentile value?

Answer:

(1) At least 15.9? 1 - 0.025 = 0.975. While 97.5% of boxes are $\geq$ 15.9.

(2) between 1 and 2 SDs above mean $\Leftrightarrow$ 0.135.

(3) The 16th percentile value for this distribution is 16.1 ounces (16% of boxes weigh less 84% weight more).

Definition 0.0.41. The standardized score for an observed value indicates how many SDs away the mean (and in which direction) this value sits.

Standardized value = (observed value - mean)/std dev = z-score

Example 0.0.42. weight of cereal boxes is approximately normal ($\mu = 16.3, \sigma = 0.2$). What percent weigh loss than 16.0?

We know the answer is between 2.5% and 16%. With Table 8.1, we can do better. The z-score for $x = 16$ is $z = \frac{x-\mu}{\sigma} = (16 - 16.3)/0.2 = -1.5$.

Example 0.0.43. What is the weight such that 90% of boxes contain less and 10% contain more? What is the 90th percentile weight? 84% percentile weight is 16.5 ounce; 97.5 percentile weight is 16.7 ounces. The 90th percentile must be between these values.

Given observed x, we find percentile by

(1) standardized $z = \frac{x-\mu}{\sigma}$;

(2) Look for z in Table 8.1, (settle for closest available value) then read to the right.

Given a percentile find its value by

(1) Look up percentile in Table 8.1, read to the left for corresponding z-score;

(2) Unstandardized if $z = \frac{x-\mu}{\sigma}$, then $x = \mu + \sigma z$;

Example 0.0.44. The 90th percentile z-value is z=1.28. In a normal distribution 90% less than mean 1.28 SD, 10% are greater than this. What value is 1.28 SD above mean? SD is 0.2, 128(0.2) = 0.256. Mean is 16.3, 16.3 + .256 = 16.56. Answer is 16.56 ounces.

Example 0.0.45. What is the 25th percentile value for cereal amount?

Feb. 13, 2018

Example 0.0.46. Amount of cereal in "16-ounce" box takes a normal distribution with mean = 16.3 and variance 0.2. Find the cut=off point where 25% of boxes have less and 75% have more then this amount, i.e. find the 25th percentile amount. Given the percentile find corresponding value by

(1) look up the percentile in Table 8.1, read left for z-score.

(2) Unstandardize

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

hence, looking at page 175 then we know the 25th percentile z-score is $z = -.67$. In a normal distribution 25% of values are more then .67 SDs below mean.

Example 0.0.47. Height in inches of young women takes normal distribution with mean 64.5 and standard deviation 2.5.

(a) Find and interpret the 20th percentile and 80th percentile heights.

(b) What percent of women are between 5'1"?

Answer is: 20th percentile is as if we are looking at z = -0.84. Then

$$x = \mu + \sigma z = 64.5 + 2.5(-.84) = 62.4$$

and 80th percentile is as if I am looking at z = 0.84. Then

$$x = \mu + \sigma z = 64.5 + 2.5(.84)$$

Chapter 9

We discuss plots, graphs, and pictures. Read ourselves. To summarize the distribution of a categorical variable, one can draw

(1) pie chart,

(2) bar chart (like histogram), but categories not numeric ranges on x-axis. Second one is better, particularly for comparing two or more distributions.

Example 0.0.48. Survey "does a HS diploma mean student has learned the basis?" Choose between Yes and No. Results (for different responded groups).

-	Prof	Employees	Parents	Teaches	Students
Yes	22	35	62	73	77
No	76	63	32	26	22
-	98	98	94	99	99

Draw a yes bar and no bar for each group, Put in common display for easy comparison. Think about how else you could have drawn this if your goal was to deceive (not lie). e.g. to magnify difference between groups make range of y-axis not 0-100.

Example 0.0.49. The age-height data. We can plot all 16 observations in a scatter plot. In a histogram we look for

(1) location (center of dist)

(2) variability (how spread at are values)

(3) shape: unimodal? bimodal? symmetric? skewed?

In a scatter plot, we are generally interested in the association between the two variables.

Question:

(1) Are the variables positively associated? or negatively?

(2) If monotone (positive or negative) is it linear or curved? All three display are positively associated.

(3) How strong is the association? If we drew a “best-fitting” line through the plot, how tightly clustered are the points around that line?

(4) What other interesting features? e.g. distinct clusters? outliers?

Note. In bivariate data set outliers may not be unusual x value or unusual y value.

Definition 0.0.50. The variables X and Y are positive (negative) associated if bigger values of y tend to occur with bigger (smaller) values of x.

Feb. 15, 2018

Graphical summary for bivariate data. Two quantitative variables. We can use scatter plot. Plot the paints (x_i, y_i) on axes. Study the association between variables. (1) positive or negative, (2) linear? or more complicated?

Example 0.0.51. Let us suppose we are looking at the population of United States and Violent Crimes. This is a data with three columns: (1) year, (2) US population, and (3) violent crimes. We plot time series data with US population and violent crimes. Apparently both went up as in he year 82' to 92'. Exercises: think about how you could draw the plots if your agenda was to scare people what if you agenda was the opposite?

(a) If you are allowed to do blatantly deceptive things,

(b) if not.

Ignore 82' and start at 83'. This would be allowed. You could skip 84' in the x-axis. It's missing right? 85' is the one after 83. This is definitely not allowed. How about start y-axis at 1.0 million, cap it at 2.0, would make up in crime less then step p. Keep the bounds as they are play around with the “aspect ratio” .

——————————————-

Let us discuss midterm 1 and the review contents. The exam is Tues. Feb. 27 (next week). The exam will cover chapters 1-9, and not include 10 and 11. It will cover homework 1-4. It will cover lecture up till now. Note that:

(1) Table 8.1 will be provided; based on assumptions of normal distribution,

(2) Please bring calculate for computation.

(3) Allow single sheet of notes. Please include all the contents.

(4) Assigned seating.

(5) we will review on Thursday Feb. 22.

(6) Format: there will be a sequence of T/F questions at front of exam. Then the written answer (short answer) and numeric calculations (back end of exam).

——————————————-

Chapter 10

Correlation and Regression

We have bivariate data with two quantitative variables. For example, we let x_i to be the height in inches and y_i to be the weight in pounds of the ith basketball player for $i = 1, ..., n = 16$. In assessing a scatter plot we look at

(1) direction of association (pos, neg, cyclic, or other)

(2) form of association, linear, curved, cyclic,

(3) strength of association.

Numerical summing of bivariate data set will require at least 5 numbers mean and SD of x's mean and SD of y's. We want to distinguish

(1) linear (pos and neg)

(2) strong pos association;

(3) weak pos association.

and we also want the measure of association between x and y. The correlation coefficient measures strength and direction of linear association between two quantitative variables. Let $\bar{x} + \bar{y}$ be the means and $s_x + s_y$ be the SDs. Then

$$r = \frac{1}{n-1} \sum_{i=1}^{n} (\frac{x_i - \bar{x}}{s_x})(\frac{y_i - \bar{y}}{s_y})$$

Note that

(1) always between -1 and 1;

(2) $r > 0$ implies positive association and $r < 0$ negative association.

(3) Moreover, $r = 0$ or $r \approx 0$ has no or weak linear association.

(4) Last, $r = 1$ or $r = -1$ if and only if all the points fall on a straight line with positive (negative) slope.

Example 0.0.52. NY Knicks with n = 16. $\bar{x} = 79.5$ and $s_x = 3.86$, $\bar{y} = 221$ and $s_y = 21.7$. This case we have $r = .8579$.

Let $x^\star$ to be height in centimeters and $y^\star$ to be weight in kg. Reminder: 1 inch = 2.54 cm, 1kg = 2.2 lbs. What is the correlation between $s^\star$ and $y^\star$. This is because $(x_i - \bar{x})$ and s_x get multiplied by the same number if we change the unit. Thus, correlation does not depend on units.

Some caveats about r:

(1) r only measures strength of linear association; There could be case that the data presents two variables with zero correlation but the data can be quadratic. The best fitting line is horizontal.

(2) Recall mean vs median as measure of location; mean is highly sensitive to outliers; and median is resistant to outliers. Correlation is not resistant. If $r \approx +1$ the exact outliers. If $r < 0$ include outliers. A single number summary only tells the right story if the truth is simple. We have linear (directly up), non-linear but monotone (up but smoothly, not at all coming down), and not monotone (up and down).

Feb. 20, 2018

Graphical summing of bivariate data: the scatter plot. Numerical summary of association between two quantitative variables. The correlation coefficient measures strength and direction of linear association between two quant variables

$$r = \frac{1}{n-1} \sum (\frac{x_i - \bar{x}}{s_x})(\frac{y_i - \bar{y}}{s_x})$$

Note r makes no distinction between explain variable and response variable. The correlation between $x + y = corr$ between y and x. Often what were most interested in is understanding relationship of response variable by y given of x often with the specific goal of predicting value of y given value of x.

Example 0.0.53. College admissions $y =$ college GPA and $x =$ SAT.

Example 0.0.54. Golf puffs: X to be distance from hole, and y to be percent of puffs made. Now we have aggregated data with "'unit" have is a bunch of puffs from loft away not just one player/shot. When we model the expected / predicted value of y for each possible value of x we are doing.

Discuss: Regression Analysis. If we assume the regression is a straight line, we were doing linear regression. The equation of the regression line is as follows. We omit graph but there are three types: (1) not appropriate, (2) is good idea, or (3) bad idea. Thus, we have

$$y = a + bx$$

while b is the slope (expected change in y per unit change in x) and a intercept to be predicted y value for case with $x = 0$. If that makes sense, often it does not. Given a data set (x_i, y_i) for $i = 1, ..., n$ How to choose a and b? Can look at scatter plot and pencil in the best-fitting line.

But there is also an objective mathematically optimal solution: principle of least squares (middle of p.211 [1]). The "best line" is that which minimizes the sum of squared vertical distances between the points and the line. Note that this means if we use linear regression to predict y from x, the predicted y-value at x = $\bar{x}$ is going to be $\bar{y}$.

Example 0.0.55. $\bar{x} = 79.5$, $s_x = 3.86$, $\bar{y} = 22.1$, $s_y = 21.7$. Recall 5-number summary. we have min, LQ, med, UQ, max. Then we compute

$$b = r\frac{s_y}{s_x} = 0.86(\frac{21.7}{3.86}) = 4.82$$

and $a = \bar{y} - b\bar{x} = 221 - 482(79.5) = -162.05$. Answer: we have *predicted weight* $= -162.05 + 4.82$ in height. The intercept only has meaningful interpretation if $x = 0$ is within range of plausible x-values.

There are three solutions:

(1) Weight of current players is 221.

(2) look at current roster, find 4 players listed at 6'5". Their weights are 210, 200, 200, 190. Then average is 200.

(3) Plug x into regression equation $y = a + bx = -162.05 + 4.82(77) = 209$.

Discuss:

Solution (1) ignores crucial information of players height. 6'5" is shorter than average, so predicted weight should be lighter than average. An argument for solution 3. It uses all the data, not just 4 players. If the new players height were 6'4". Solution 2 would been impossible anyway.

Feb. 22, 2018

The exam will cover Chapter 1-9 and corresponding lecture and HW 1-4. Chapter 3 is where it started. Focus on Chapter 3-5, and as well as 7-8.

————————————-

Exam:

Part 1: T/F (no more than half)

Part 2: Written answer (both calculation and short response)

————————————-

Chapter 3

Measurement and variables. Categorical versus quantitative variables. Categorical discusses nominal versus ordinal. Quantitative discuss discrete and continuous. A discrete quantitative variable can only assume distinct values, e.g. number of siblings, continuous variables, height, weight, ..., etc..

In measurement, there are a few tricky ideas. A valid measurement is one that actually measures what it purports to. For example, GPA is a valid measure of academic achievement. Is it a valid measurement of intelligence. A reliable measure is one that gives consistent results if applied to same object. An unbiased measurement is one that gives correct results on average.

Example 0.0.56. Device A gives readings that vary between 5% and 15% greater than actual truth. Device B gives readings that vary uniformly between 5% too low and 5% too high. Device C gives readings between 9% and 11% too high.

Answer: Device B is the only one that is unbiased. However, it is not that reliable since 5% error is too much. Device C is fairly reliable, but it is biased since it consistently returns readings higher. Device A is the least useful. It is neither reliable nor unbiased. The range of value read by A is quite high and it is not that consistent.

Variability across measurements result from natural variability that is different values for different individuals. Another source across measurements that is less discussed is measurement error (measured value not exactly true value). A model for measurement error is Measured value = true value + measured error, which we write

$$MV = TV + (MV - TV) = TV + (IA - TV) + (MV - IA)$$

while IA means instrument augmentation for this measurement on this object. We just partitioned ME to two components: IA - TV is due to bias is due to "unreliability".

We can refer to page 57 of text [1] in bold statement. It says **The key to statistical discovery is comparing natural variability to create variability**.

Example 0.0.57. Take 20 tomato plants and they are treated identically with water and sunlight. However, 10 plants get fertilizer and 10 do not. What is the natural variability?

The natural variability is difference in yield within the treatment groups. Creative variability is the difference between the treatment group. If CV is big relative to NV, we have evidence of a fertilizer effect.

Chapter 4

Glossary as follows.

unit: individual or objects to be measurement;

population: collection of units for which we desired this measurement;

sample: collection of units for which we obtain this measurement;

sampling frame: list of units from which sample is drawn;

Example 0.0.58. Chapter 4. Exercise 17. Representative sample of 252082 from at 464 colleges nationwide. The unit is the freshman student at one of these schools. The population are all such students. Sample would be the 252082 students sampled. In a SRS, all the names are thrown in a hat, and 252082 are drawn.

More likely that this was a stratified example. A separate SRS of freshman was taken at each of the school. How would a cluster sample have worked for this survey? Take a random sample from the 464 schools. Then survey every freshman student at the selected schools.

Question: might this way be better? Answer: absolutely not. Not statistically anyway. It may be practically easier to conduct survey this way.

We can discuss sampling frame versus true population. Sampling from may include students who have since dropped out and exclude late registrants.

Example 0.0.59. (2009). Random sample of 702 children, 464 owned a cell phone. Discuss

(a) Estimate proportion of children who own cell phone.

(b) Find the margin of error

(c) Find a range you can be highly confident contains the true proportion of children who own a phone. (Bring calculator)

Answer:

(a) $\frac{464}{702} = 66.1\%$

(b) Recall margin of error: $\frac{1}{\sqrt{n}}$ while n is the sample size. This is 3.8%.

(c) Estimate - ME = .661 - .038 = .623, and also Estimate + ME = .661 + 0.038 = 0.699. The estimate is that we are highly confident that somewhere between 62.3% and 69.9% of US children in 2009 own a phone.

Controlled experiments versus observational study. Whether the explanatory variable is controlled by experiment. Interaction effect: effect of one explanatory variable can differ depending on the level of another. Confounding occurs when the effect of one explanatory variable is indistinguishable from another. The latter is not a bug of observational statements. The former is a feature of two factor experiments.

Example 0.0.60. Volunteers chose one of 3 exercise programs: jogging, swimming, or aerobic dance. The participants can choose one of these three. The researchers found jogging seem to be the most effective one. However, the study is crucially flawed. It may be that the most highly motivated individuals were more likely to choose jogging. Think how this could be done differently. Suppose subjects given a pre-test to assess their motivation. Suppose the results are (higher number means better outcome). Then

$-$	$LowMot.$	$HighMot.$
Jog	1	3
$Swim$	2	3
$Dance$	2	3

This would allow us to look for interaction between motivation and exercise program. We did not review Chapter 7.

Example 0.0.61. Heights of black cherry trees takes $\mathcal{N}(76, 6)$. Questions:
(a) What percent are taller than 85 feet?
(b) What percent are between 68 and 85 feet?
(c) Find the 70th percentile height of black cherry trees.
Answer: (a) 7%; (b) 84%; (c) 79 feet.

Feb. 29, 2018

Midterms to be graded and returned next Thursday, March 8th. Suppose we have a bivariate data set (two quant variables). Look at scatter plot. Compute correlation coefficients and run a regression analysis. Find there exists a statistical relationship between y and x. There are Section 11.3 at least 7 plausible explanations for this. x causes y is one, but there are six others.

(1) Indeed x causes y, ex: x is altitude, and y atmosphere pressure,

(2) It's actually y causes x, from data alone we cannot possibly tell the difference. Ex: the units are small business operation, x is advertising expenditure and y is sales revenue. There are m observational study, find negative association. Note even if association is positive. Is advertising causing sales? or is revenue being reversely understood. If association is negative, only those companies that are struggling have to advertise.

(3) x contributes to y, but there are other causes too. Ex: $n = 50$ students in math class. Let x be hours spent studying and y be the score on exam. I would expect a positive association, but it want to be a perfect one and other contributes factors too.

(4) A confounding variable is the cause. Subjects are heart patients x to be happiness and y to be survival. Confounding variables? Lots. Happier people are more likely to take their pills and exercises. (3) lets y to be score on test and x_1 to be score on pre-test and x_2 to be hours of study. Confounding variables effect of one variable hard to distinguish from that of another (this may happen if multiple explain variables are associated with each other.

(5) Both x and y are caused by some other variables x to be atmosphere pressure and y to be boiling point of water. Both are affected in highly predictable and by altitude. Ex: x to be SAT score and y to be college GPA.

(6) Both x and y change over time. Ex: x to be marriage rate and y to be life expectancy. For years 2000 to 2011 x is decreasing and y is increasing with $r = -.984$.

Homework Problem 11.14 (and 15, 16). Explain why the association exists, do not worry about which of the 7 causes it is (unless you find that helpful). Chapter 12. Relationships between categorical variables. Numerical measures. Quant data correlation regression equation categorical data. This is called a contingency table. This is a retrospective study. Explanatory variable: Smoking or not. Response variable: Pregnancy first cycle.

The appropriate number of comparing the groups (smokers vs. Non). is percent first cycle. For smokers, its $29/100 = .29$ or 29% and for non, its $188/496 = .41$. Can we conclude that smoking has a harmful effect on fertility? Not strictly from those data, being an 0.5. Just to make things consuing, there are multiple ways there things get reported. But only two fundamental quantities, the first guess by many names.

(1) proportion of individual who possess traits, and

(2) the odds of that an individual possess trait.

Ex. population of 1000 individuals with 400 are carriers for some disease. The proportion of carriers is 40%. The percentage is 40%. The probability some is carrier is 40%. The proportion / prob / risk is p, a percentage; while odds are $p/(1-p)$ to 1. If the odds are a to b, the probability is $a/(a+b)$

March 6, 2018

Chapter 12

Chapter 12 considers relationships between categorical variables. These are also bivariate data.

Example 0.0.62. Explain variable is smoker and non-smoker. Response variable is pregnant first circle: yes or no. Hence, we record the following:

-	First Cycle	Later	Total
Smoker	29	71	100
Non	198	288	486
Total	227	359	586

Example 0.0.63. Explain variable: wifes parents divorced? Yes or no. Response variable: couple separated? separated or intact.

WPD	Separated	Intact	Total
Yes	42	292	334
No	72	1092	1164
Total	114	1384	1498

WPD	Separated	Intact	Total
Yes	.126	12.6ˆ	.144
No	.062	6.2%	.066
Total	.076	7.6%	.082

Recall that odds = prob/(1-prob) and prob = odds/(odds+1).

Moreover, we have RR for group 1 relatively to group 2 is

$$\frac{\text{Risk for Group 1}}{\text{Risk for Group 2}}$$

The IR for group 1 vs. group 2 is IR = (RR - 1)×100%.

Definition 0.0.64. The group designated "group 2" in the above formulas is the baseline group. The rise for group 2 is baseline risk.

Example 0.0.65. In the above calculations, we took wife's parents not "divorced" as baseline group.

Interpretation, RR and OR get compared to 1. IR gets compared to 0. IR is positive if and only if Risk higher then baseline; IR negative if and only if risk lower then baseline. A 9.4 per 1000; 17.1 per 1000.

Aside, with lower probability events odds and probability are very close.

$$\text{odds} = \frac{\text{odds}}{1 - \text{prob}} \approx \frac{\text{prob}}{1}$$

$$\text{prob} = \frac{\text{odds}}{\text{odds} + 1} \approx \frac{\text{odds}}{1}$$

Find RR/IR/OR: RR for aspirin relative to placebo,

$$\frac{0.0094}{0.0101} = 0.55$$

$$\text{IR} = \left(\frac{\text{Risk} - B \times \text{Base Risk}}{\text{Base Risk}}\right) \times 100\% = 45\%$$

$$\text{OR} = \left(\frac{0.0095}{0.0174}\right) = .55$$

and we have, though rare, OR is approximately to RR. Risk of MI for aspirin is .55 times the baseline risk, i.e. 45% lower then baseline risk. Some people prefer to only report RR ¿ 1 or IR ¿ 0 or OR ¿ 1. which means defining the baseline group as that with lower risk or RR of placebo relative to aspirin is 1/.55 = 1.82.

Risk statistics can be misleading in three ways, see Section 12.3.

(1) If RR is high, but baseline risk is really low. The RR may sound a lot scarier then it really is.

(2) Risk of esophseal cancer for men is 7.7 out of 100,000. Risk of esophageal cancer for men is 1/198 = 505/100,000.

March 8th, 2018

Risk statistics can be misleading in three ways (Sec 12.3).

(1) If RR is high, but Base Risk is really low, the RR may sound scarier than it really is.

Example 0.0.66. Odds ratio for esophageal cancer for beer drinks to nan is 3. Drinking beer triples the risk of cancer. Drinking beer triples the risk of cancer? First, how much beer? From what to what?

Compute 1/100000 to 3/100000 or 1/10 to 3/10?

(2) Risk over what time period? Risk of esophageal cancer for men is 7.7 out of 100000. Risk of esophageal cancer for men is 1/198.

These statements are both correct. First is based on annual incidence. Second one is life time risk.

(3) Reported risk is not your risk.

Example 0.0.67. Shark attacks per year over number of population equals to risk of shark attack is a pretty useless statistics. If you never swim in ocean, your risk is zero. If your swim in the ocean a lot, it's a lot higher than this.

Chapter 13

Example 0.0.68. Roadside survey of young drivers in 1970 Oklahoma.

-	Yes	No	Total
Male	77%	40&	481
Female	60%	12%	138
Total	93	526	619

RR for M vs F is .16/.116 = 1.38 which is IR = 38%. OR = .16/(1-.16)/(.116/(1-.116)) = 1.45. Pick one: these data prove that drinking and driving is more prevalent among M than F. There is only 600 stops, and less than 200 F drives. The data don't prove anything, it could be random variation in sampling. There is a gender effect in the sample. Can we conclude that this effect holds in the population? To answer is to assess whether or not the sample effect is statistically significant.

If yes, its strong evidence that the effect holds for population. If no, we conclude that no such effect holds for population. Wrong.

If no, we only say the data don't prove any effect one way or the other. Statistical significance depends on two things.

(1) the size of the sample effect. Ex: Is 16% vs. 11.6% a big enough difference to conclude the effect is real?

(2) the size of the study. Ex: Is 600 stops enough to reach such a conclusion? The basic logic goes: Suppose the truth is there is no effect in the population. Ex: Suppose aspirin has no effect can risk of MI. Suppose a couples prospects totally unrelated to wife's parents. Suppose smoking has no effect on pregnancy success. Then ask: how unlikely would be the sample effect, if the true population effect is in!

If the answer is highly unlikely the most plausible explanation is the effect is real. In this case the sample effect is statistically significant.

Chapter 13 There are four steps for hypothesis testing:

(1) Determine null hypothesis and alternative hypothesis; Null and alternative hypothesis are competing claims about the population not the sample.

(2) Collect data, compute the appropriate one-number summary of the data called test statistics;

(3) Determine how unlikely the test statistics value would be if null were true.

(4) Make a decision;

Let us practice:

(1) Generally alternative hypothesis is the thing we are trying to prove. Null is some form of "nothing going on here".

Example 0.0.69. Null: there is no relationship between gender; Alternative: there is relationship between gender and driving.

Example 0.0.70. Null: risk of MI is same for aspirin and placebo groups Alternative: Risk of MI different for aspirin versus placebo.

(2) Test statistics: RR/IR/OR are not ideals as test statistics because they only tell half the story. Measure strength of sample relationship, ut do not reflect sample size at all.

Here "expected" means what we'd expect if null were true. The test statistics should measure how far apart. For each cell in the table, observed cell count and expected cell count = (row total x column total)/Table total.

.	Yes	.	No
M	77	404	481
F	16	122	138
.	93	526	681

Compute

72.3	408.7
20.7	117.3

Compute odds ratio for expected table. Should set 1.
Chi-square

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Observed - Expected})^2}{\text{Expected}}$$

(3) The p-value is the probability, assuming the null hypothesis is true, of observing a test statistics as extreme (or more so) as the value actually observed. The closer the p-value is to zero (one) the stronger (not strong at all) the evidence against null hypothesis. Common convention is for p-value less than 0.5 to be considered statistically significant. p-value less than 0.01 is considered fairly strong evidence against null hypothesis.

March 20th, 2018

Chapter 13

Consider the alcohol among male and female data. We ask is there association between gender and alcohol. Also consider the aspirin and placebo data set. We ask does aspirin really reduce risk of MI?

We apply the Hypothesis Testing with 4 steps.

1. Determine null and alternative hypothesis;
2. Compute test statistics;
3. Find the p-value;
4. Interpret the p-value.

Let us do this one by one.

1. Ex 1. Null: no gender difference; Alternative: there is a difference. Ex 2. Null: rate of MI same for aspirin and placebo; Alternative: different rate of MI.

Generic form (for two-way table): Null: there is no association between row variable and column variable; Alternative: there is an association between row variable and column variable.

2-4. The test statistic is a measure of how different the observed data are from what we would have expect to have observed if the null hypothesis were the truth. The p-value is the probability calculated, assuming the null hypothesis is true, of getting a test stat value as extreme as (as more so) the value observed. If the p-value is close to 0, we have evidence against the null, in favor of alternative hypothesis. If p-value is not close to 0, there is no evidence against null hypothesis.

The test stat for two-way tables would be

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Observed - Expected})^2}{\text{Expected}}$$

and expected here means

$$\text{Expected} = \frac{\text{Row total} \times \text{ Column total}}{\text{Table Total}}$$

Example 0.0.71. Consider

-	Y	N	
M	77	404	481
F	16	122	138
-	93	526	619

then we compute

Exp	
72.3	408.7
20.7	117.3

which is the expectation table. Note that for example, 72.3 is found by computing $481 \times 93/619$.

Using computer, we have $\chi^2 = 1.6$, p=value = 0.2. If there were no difference in drinking and driving between young men and young women, then the probability observing a sample difference like that observed would be 0.2. The data do not prove a relationship between gender and driving and drinking.

Example 0.0.72. Consider

-	Y	N	
Aspirin	103	10933	11037
Placebo	189	10,045	11,034
-	293	21,778	22,071

then we compute expectation table

Exp	
146.5	10,890.5
146.5	10887

Using computer, get $\chi^2 = 25$ and p-value ≈ 0. If there were no association between MI and aspirin (i.e., MI risk were same for aspirin and placebo), the probability of observing a sample difference (9 per 1000 vs. 17 per 1000) would be zero. It would not have happened.

Data provide irrefutable evidence that indeed aspirin reduces risk of MI (at least to male physicians).

Chapter 14

Chapter 14. Probability

Example 0.0.73. Probability roll 2 dice, the probability of getting a total of 8 is 5/36.

Example 0.0.74. The probability of Trump being impeached .64.

Example 0.0.75. Personal probability interpretation (subjective). The probability of an event measure the degree to which one believes the event will happen.

Under either interpretation, probabilities have to follow certain rules in order to be coherent.

Rule 1. If the probability A is p, then the probability of not A is 1-p. $P(A) + P(A^C) = 1$, while C is for complement.

Example 0.0.76. Probability of boy is .51. Probability of girl is .49.

Rule 2. If two events are mutually exclusive then the probability of one or the other occurring is the sum of their individual probability. They can't both happen.

If $P(A \text{ and } B) = 0$ then $P(A \text{ or } B) = P(A) + P(B)$.

Example 0.0.77. Probability getting A is .5. Probability of getting B is .3. Then the probability of B or better is .8. The probability of getting C or lower is .2.

Rule 3. If two events are independent, the probability that both occur is found by multiplying their individual probabilities.

Mutual exclusive: If A occurs, B can't occur.

Independent: A occurring or not conveys no info about B occurring or not.

If A and B are independent, then $P(A \cap B) = P(A)P(B)$.

Example 0.0.78. The pregnant friends: P(both boys) = .51 × .51 = .26. P(both girls) = .49 × .49 = .24.

Example 0.0.79. Probability of A's in both courses is .3? Probability A's in both courses is .3? Yes if you believe they are independent. But it could also reasonably be low or higher.

Rule 4. If the ways in which one event can occur are a subset of the ways in which another event can occur. Then the probability of the subset event is

Example 0.0.80. Probability Trump is impeached is .64. Probability Trump finishes term is .3. Probability of Trump being reelected is .4. The first two are coherent. the last one is not.

March 22, 2018

Example 0.0.81. Lottery game, prob of winning is 0.3. I will buy a ticket, if I win I'll stop. If I lose, I'll buy another one.

Q1. What's the prob I have to buy a third ticket?

Q2. I only have \$12 on me. What's the prob I'll have to norrow \$ from someone?

Q3. What is the "expected value" for number of tickets I end up buying?

First winning ticket, we will compute probability

First winning ticket	Probability
1st	.3
2	.7 (.3) = .21
3	.7 (.7) (.3) = .147
4	.7 (.7) (.7) (.3) = .1029
5	$(.7)^4(.3) = .0720$
6	$(.7)^5(.3) = .0504$

We can answer the questions

(1) We compute

$$\begin{aligned} P &= 1 - P(\text{win the first 2}),\ \text{by Rule 1} \\ &= 1 - (.3 + .21),\ \text{by Rule 2} \\ &= 1 - .51 = .49 \end{aligned}$$

(2) We compute

$$\begin{aligned} P &= 1 - P(\text{ win somewhere in first 12}) \\ &= 1 - (.3 + .21 + .147 + \cdots + .0085 + .0059) \\ &= 1 - .986 = 0.014 \end{aligned}$$

Alternatively, we can also have

$$\begin{aligned} P > \$12 &= P(\text{losses in a row}) \\ &= .7 \times .7 \times \cdots \times .7 \\ &= (.7)^{12} \\ &= .014 \end{aligned}$$

(3) We compute

which is from Section 14.5 in [1]

We will keep trying till we succeed. Assume the probability of success is same on every trial. The trials are independent. 3 Rules you should know.

(1) The probability our first success comes on kth try is

$$(1-p)^{k-1}p$$

(2) We discuss two cases. (a) The probability of no successes after n trials is $(1-p)^n$, (b) the probability of getting a success within the first n tries is

$$p + (1-p)p + (1-p)^2 p + \cdots + (1-p)^{n-1} p = 1 - (1-p)^n$$

(3) The expected value for number of tries required to get the first success is $\frac{1}{p}$.

(4) The most likely outcome is that first success happens on first trial. $(1-p)^{k-1}p$ is maxed at $k = 1$.

Expected value (section 14.6)

If the experiment were repeated a large number of times, the average of the results could be close to expected value. Suppose buy tickets till we win. Note that number of tickets bought. Do this again and again. The average of number of tickets required will be approximately .33. Imagine a gamble costs \$1 to play. Pays \$5 within probability .18. Is this a good bet? Payoff

Payoff	Prob
\$5	.18
\$0	.82

Expected value of payoff = 5(.18) + 0(.82) = .9 < 1. Expected value of net pay of .9 - 1 = -.1

Net pay	Prob
\$4	.18
-\$1	.82

The general form of expected value:

Given a random variable with n possible values, call them $A_1, A_2, ..., A_n$ with probabilities $p_1, p_2, ..., p_n$. Each $p_i \geq 0$, $i = 1, ..., n$ and $p_1 + p_2 + \cdots + p_n = 1$. Then the expected value is

$$\mathbb{E}(\text{value}) = A_1 p_1 + A_2 p_2 + \cdots + A_n p_n = \sum_{i=1}^{n} A_i p_i$$

If you go and compute by using expected value we would get 3.7 and this is under estimate.

Wierd idea: the EV formula works even if there is no finite n. Using calculus techniques you can still evaluate the sum. You'll get 3.33 (geometric series). What you must know is to calculate EV for random variables with small number of possible values. What EV means: the long run average in large number of repetitions.

Example 0.0.82. Deal 5 cards, find the probability getting at least one pair.

$$\begin{aligned} \text{Answer} &= 1 - P(\text{all diff denom}) \\ &= 1 - \left(\tfrac{52}{52}\right)\left(\tfrac{58}{52}\right)\left(\tfrac{44}{52}\right)\left(\tfrac{40}{52}\right)\left(\tfrac{36}{48}\right) \\ &= .493 \end{aligned}$$

What if you didn't believe me? How could you check this for yourself? Deal yourself 1000 poker-hands (Shuffle in between!) Note the percentage of the time you get at least one pair. This will be a good estimate of the true probability.

Even better, write a computer program to do it. This is called the Monte Carlo method, or stochastic simulation. Stochastic is opposite of deterministic, not random. This is based on "random number generator" computer programs.

What sorts of things is this used for in the world today? Economic forecasts, e.g. stochastic model of "yield curve", interest rate vs. term of bond. What this curve looks like today is not how it's going to look like in 5-10 years. There will be changes.

What will my portfolio be worth in 10 years? Generate large number of scenarios using stochastic simulation. We will show best case and worst case outcomes, etc.

Where this approach is most valued is situations where there is too many moving parts to get a handle on

Epidemiology. Simulate multiple scenarios for spread of contagious disease. Recent example where simulation based forecasting get it completely wrong? 2016 election is a notorious example.

March 27th, 2018

Probability.

Long run frequency interpretation (objective). Personal probability interpretation (subject). With latter, there is danger ones personal probabilities are incoherent(internally consistent).

Might violate one or more of the probability rules. Recall the four rules:

Rule 1: The probability an event doesn't occur is 1 - Probability the event does occur. $P(A^C) = 1 - P(A)$.

Rule 2: If A and B are mutually exclusive, then $P(A \text{ and } B) = 1$, and $P(A \text{ or } B) = P(A) + P(B)$.

Rule 3: If A and B are independent, $P(A \text{ and } B) = P(A) + P(B)$. Mutually exclusive means that they can't both occur e.g., roll a die, 6 and "odd number" are mutually exclusive. Independent means that one having occurred gives no information about likelihood of the other. Roll 2 dice, even number on first and even number on second. They are independent. Roll 1 die and "3 or higher" are independent.

Rule 4: If B is a subevent of A, then $P(B) \leq P(A)$.

Example 0.0.83. I have got a weighted dice. The probability are not all 1/6. Probability of a 6 is greater than probability of even numbers. This is not possible. It must be P(even numbers) = P(2 or 4 or 6) which is greater or equal than 6.

Chapter 16

We are starting Chapter 16 today.

It is about the ways people personal probability can violate these rules. (i.e. be incoherent).

Example 0.0.84. Which of the following is most likely?

(a) President Trump is impeached by end of 2018.

(b) President Trump is impeached on charges of colluding with Russia by end of 2018.

(c) President Trump is impeached for illegal activity to sormy daniels affair.

If you answer (b) or (c) you just violated Rule 4. (a) = (b) or (c) or other things. Thus (a) is more likely than (b) or (c). Even when our personal probabilities are coherent. We tend to be (1) optimistic, (2) related to change, (3) over confident in our assessment. Over confident here means. I think A is the most likely outcome. My personal probability of A is .

one professor particularly susceptible to this over confidence is doctors. Once group that is particularly good at callibrating their probabilities is weather forecasts. How were these assessments made? Everyday for a large number of days, forecasters say probability of rain storming is p where p is 0, .05, .1, .2, .3, ... If their probabilities are property callibrated we should find

$$\frac{\text{number of days fain on which it does rain}}{\text{number of days where probability of rain is}} = \text{ actual probability } \approx p$$

Plot of actual versus predicted probability if probability are perfectly callibrated, the points will something on the 45 degree line.

Do something similar with doctors diagnoses. When a doctor says I believe it's pneumonia and the probability I'm right is 90%. It is actually pneumonia in about of those cases 10%.

Chapter 17

Chapter 17.

Example 0.0.85. The right answer is to all 3 examples: it doesn't matter. In every single case, the probability of red is 18/36, for black it's 18/36. Probability of 6 is 2/38. The spins are independent. To believe that patterns exist even in events that are actually independent is called gambler's fallacy. A sequence of independent events means there are no patterns.

Example 0.0.86. Random drug testing.

Suppose a drug screening test has sensitivity of 0.96, specificity of 0.93. The prevalence of drug use is 0.01. Randomly selected individual. What is the probability of positive test result? Given a positive result, what's the probability that person actually is a drug user? Sensitivity = 0.96. If subject on drugs, the probability of test positive is 0.96. Specificity = 0.93. If subject is clean, probability of test negative is 0.93. Prevalence = 0.01. which is 1% of population use drugs. If drug user, probability test positive is 0.96. If test positive, probability drug user is not 0.96. We can solve this using tree diagram

User, 0.01

(1) Test Pos 0.96

(2) Test Neg, 0.04

clean, 0.99

(1) Test Pos, 0.07

(2) Test Neg, 0.93

Thus, we compute

User, Test Pos = 0.0096

User, Test Neg = 0.0004

Clean, Test Pos = 0.0693

Clean, Test Neg = 0.9207

Given a positive test result, the probability subject is indeed

$$\frac{\text{user and test positive}}{\text{test positive}} = \frac{0.0096 + 0.0693}{0.0096 + 0.0693} = 0.12$$

Only 12% of positives are true positive 88% are false positives. Surprised? What's the explanation It's because the probability is so low.

March 29, 2018

Midterm 2 is coming:

Date is April, 3.

Bring calculator and one 8.5 by 11 sheet (both sides) of original handwritten notes.

The exam will cover

Ch 1-7 as background only

Ch 8-15 but not Sec 15.3 - 15.4 (Not Ch 9), 8, 10-14, 15.1-15.2.

HWs 5-7 review Ch 8 Problems on HW 3-4. Some Ch 13 practice problems on Coursework.

Below is the midterm 2 review.

Example 0.0.87. On exam:

There are 30 seats in a room with 5 rows of 6 columns. 28 students, include you and your best friend. What is the probability you sit next to each other?

Answer:

Take a room of 30 seats 5 rows of 6. Use a random number generator. We randomly generate assignment of all 28 students and 2 empty seats.

Note: next to your best friend? Yes or no. Repeat this 1000 times then

$$\frac{\text{number of yeses}}{1000}$$

is an estimate of the desired probabilities.

Example 0.0.88. Heights of black cherry trees is similar to normal distribution mean = 76 feet and SD 6 feet.

(a) What percentage of trees are less than 66 feet?

(b) What percent taller than 84 feet?

(c) What is the 80th percentile height?

Answer:

(a) Compute $z = \frac{x-\mu}{\sigma} = \frac{66-76}{6} = -1.67$ so the tree that is 66 feet height is 1.67 below the mean height. The percentage (check table) that is according to z-value here is what I want. Table 8.1 tells me that this is 5%. Find -1.67 in left column (or closest thing). Read to the right and you get 5%.

(b) $z = \frac{x-\mu}{\sigma} = \frac{84-76}{6} = 1.33$ which means that a tree that is 84 feet would be 1.33 SD above the mean. The probability for trees above this height would be, by symmetric property, would be the same as the area under the curve to the left of the tree, which is 0.09. I can also reason to argue that this is 1 - 0.91 = 0.09. Answer is 9%.

(c) Look for 0.8 in right hand columns and read to the left. $z = 0.84$. Then $x = \mu + z\sigma = 76 + 6(0.84) = 81$. About 80% of the trees are shorter and 20% are taller than 81 feet.

Example 0.0.89. Hospital collected data to studying relationship between y to be patient satisfaction and x = patient age. The satisfaction score is based on questionnaire responses. Age has 0 to 100 in scale in years. We have regression equation

$$y = 119.94 - 1.52x$$

(a) Explain what "-1.52" means here?

(b) Use this model to predict the satisfaction rating given by a 30-year patient

(c) Give an example of a question a Hospital administration might ask, but couldn't answer from this model.

Answer

(a) -1.52 means that per year increase in age the satisfaction score is decreased by 1.52.

Or alternatively, you can state: for each added years of age, we estimate that average patient satisfaction decreases by 1.52 points.

(b) Plug $x = 30$ into the equation $y = 119.94 - 1.52(30) = 74.34$

(c) An example can be: what is the score for satisfaction for a 9-year old or any child (young age).

Note 119.94 and -1.52 were calculated from the data, using formulas at the end of Ch 10.

Example 0.0.90. Risk of concussion for youth soccer players?

-	Concussion		
-	Yes	No	
Soccer	33	67	100
Non-soccer	18	82	100

Calculate and interpret

(a) relative risk

(b) increased risk

(c) odds ratio

(d) Two parts: (i) Do the data prove anything? Specify appropriate null and alternative hypothesis? (ii) Calculate table of expected cell counts, "expected" if null is true.

Answer:

Consider

-	Concussion		
-	Yes	No	
Soccer	A_1	A_2	$A_1 + A_2$
Non-soccer	B_1	B_2	$B_1 + B_2$

(a) RR $= \frac{A_1/(A_1+A_2)}{B_1/(B_1+B_2)} = 33/100/(18/100) = 1.83$ Risk of concussion for soccer players is 1.83 times that of non-soccer players.

(b) increased risk:

$$\text{IR} = (\text{RR} - 1) \times 100\% = (1.83 - 1) \times 100\% = 83\%$$

The risk (prob) of concussions for soccer players is 83% higher than for non-soccer players.

(c) odds ratio

$$\text{OR} = \frac{A_1/A_2}{B_1/B_2} = (33/67)/(18/82) = 2.24$$

The odds of concussion or soccer player are 2.24 times those of non-soccer players.

(d) Two parts: we do (d)(i) and (d)(ii).

(d)(i). The right answers: Null says: there is no association between soccer and concussion risk. Alternative hypothesis is that there is some association between

Null hypothesis: risk of concussion is the same with soccer player as for non-soccer players.

Alternative hypothesis: Risk of concussion for soccer players is different then that of non-soccer.

(d)(ii). For each cell,

$$\frac{\text{(total in that row)(total in column)}}{\text{total in table}}$$

which gives us $(25.5, 74.5; 25.5, 74.5)$. Note

$$\frac{1}{2}\left(\frac{33}{100} + \frac{18}{100}\right) = 25.5$$

Here is what will always hold. In the expected table, the percentages across columns will be same for every row.

Example 0.0.91. Flip a coin, probability heads is 1/2.

Roll a die probability of 6 is 1/6.

Roll 2 dies, probability of 11 is 2/36.

Personal probability interpretation. A person with probability is number between 0 and 1 which measures one's belief about how likely a particular outcome is 0 for impossible,

Example 0.0.92. The probability I get home by 8:30 tonight is 0.6. There are wrong personal probability if they are not coherent. When will it happen? Section 14.5. Independent trials, each is S or F. Probability of success is p.

Three useful formulas.

(1) Probability we get our first success on kth trial is $p(1-p)^{k-1}$.

(2) The probability of no S's in n tries $(1-p)^n$.

(3) Probability at least one S is n tries would be probability first S comes within n tries $= 1-(1-p)^n$.

Example 0.0.93. Large number of columbia undergrad:

21% Barnard

54% CC

25% GS

(a) Probability first student selected is Barnard.

(b) Probability first BC student we get is on our 3rd pick

(c) Probability at least one BC student among first 3 selected.

Answer:

(a) 21%;

(b) $.79 \times .79 \times .21 = .131$

(c) Complement of at least one would be none.

Hence, compute no BC all three picks $= (.79)^3$ then complement would be $1 - (.79)^3 = .507$.

April 4, 2018

Post Midterm 2 session: we cover Ch 16-17, skip Ch 18, and discuss Ch 19 today.

Example 0.0.94. Imagine a population with just 4 individuals. The population is $\{108, 112, 106, 96\}$. We will take a random sample of $n = 2$ of them, and calculate their average. The mean is 105.5, represented μ. Now consider the sample $n = 2$. There are 6 possible samples each has equal probability

Sample	Prob	Sample Mean	$\bar{x}$
1,2	1/6	110	
1,3	1/6	107	
1,4	1/6	102	
2,3	1/6	109	
2,4	1/6	104	
3,4	1/6	101	

and we have population distribution

$$\begin{bmatrix} \text{value} & 96 & 106 & 108 & 112 \\ \text{Freq} & 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}$$

and the sample distribution of the sample mean is

$$\begin{bmatrix} \text{Value of } \bar{x} & 101 & 102 & 104 & 107 & 109 & 110 \\ \text{Prob} & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{bmatrix}$$

The mean of the population is $\mu = 105.5$. The mean for the distribution of possible values of $\bar{x}$ is 105.5, same as the population mean.

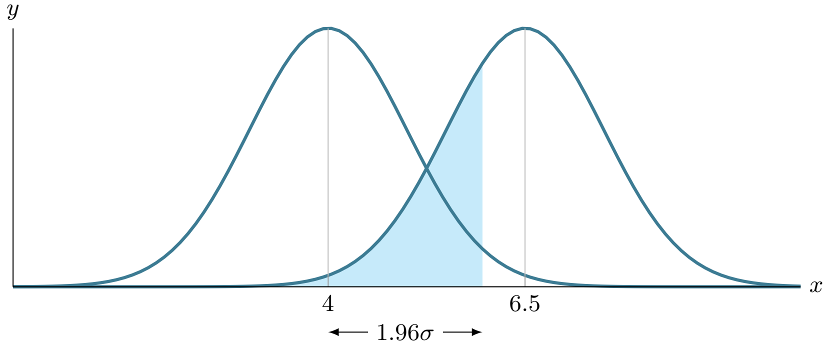
Both of these observations are true in general. The mean of the distribution of $\bar{x}$ is population mean. The Sd of the distribution of $\bar{X}$ is less than population SD. The greater is n, the greater SD. Now imagine the population size is not 4 individuals, it's 1000s or millions. The sample size is not 2 it's 50?

Obviously (even if population entirely known) we can't enumerate all the possible values for $\bar{x}$. But the distribution of possible values is in fact well known and was summarize the distribution by drawing a histogram of the results. Then (rules of sample means, page 416):

(1) The mean of this distribution is mean of the population;

(2) The SD of this distribution is SD of the population divided by $\sqrt{n}$.

(3) The shape of the distribution depends on (1) the shape of the population distribution; (2) the size of the sample, n. (B0x on p. 415). If the population distribution is normal (bell curve). then the distribution of $\bar{x}$ values is normal also... If the population is not normal, then the distribution of $\bar{x}$ values is not normal either. Thus, if population mean is μ. Population SD $= \sigma$. The distribution of possible $\bar{x}$ values given sample of size n looks like a bell-shape curve.



Example 0.0.95. Large population of people average amount of sleep lost night is 7 hours, SD is 2 hours (suppose). Take a random sample of 140 people. What can we expect to get? $\mu = 7$, $\frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{140}} = 0.169$.

The average looks like bell-shape curve. That is, the average value is close to 7 hours, only 5% chance that it's not between 6.66 + 7.34. Now suppose the question was: Did you get 8 hours sleep last night? Yes or No. Suppose 30% did. Lot of population. What can we say about the percent of Yes's in a sample of $n = 140$?

Draw a histogram to summarize the distribution of $\hat{p}$ values. Now the possible values of $\hat{p}$ are $0, 1/n, 2/n, ..., (n-1)/n, 1$. i.e. $\hat{p}$ is a discrete variable. But if n is large enough, it is well approximated by continuous distribution. Suppose n is large. Then (Rules for sample proportion, p.412)

(1) the distribution of the possible values of $\hat{p}$ is approximate a normal bell curve.

(2) The mean of thus distribution is the population proportion, p.

April 10, 2018

Consider a large population of individuals, the proportion of which possess.

Example 0.0.96. Trump approval rate, suppose $p = .4$. Take a random sample of n individuals. Note the sample proportion, call it $\hat{p}$.

Example 0.0.97. Every 1500 registered votes. By the Rule for Sample Proportion p412 which holds when n is sufficiently large

$$\hat{p} \sim \mathcal{N}(\text{mean} - p, \text{SD} = \sqrt{\frac{p(1-p)}{n}}$$

About 95% of possible samples will give $\hat{p}$ between 0.375 and 0.425.

In a random sample there is a 95% chance that

$$p-2\sqrt{\frac{p(1-p)}{n}}<\hat{p}<p+\sqrt{\frac{p(1-p)}{n}}$$

and thus there is a 95% chance that

$$\hat{p}-2\sqrt{\frac{p(1-p)}{n}}<p<\hat{p}+2\sqrt{\frac{p(1-p)}{n}}$$

or

$$\hat{p}-2\sqrt{\frac{p(1-p)}{n}}<p<\hat{p}+2\sqrt{\frac{p(1-p)}{n}}$$

or

$$\text{sample proportion} - 2\text{SD} < \text{population proportion} < \text{ sample proportion } + 2\text{SD}$$

and this gives us 95% chance of this 95% of possible samples. The interval

$$\hat{p}-2\sqrt{\frac{p(1-p)}{n}},\hat{p}+2\sqrt{\frac{p(1-p)}{n}}$$

is called a 95% confidence interval for p.

Margin of error for the estimate $\hat{p}$ would be

$$2\sqrt{\frac{p(1-p)}{n}}$$

while we use $\hat{p}$ instead of p.

While standard error SE would be

$$\sqrt{\frac{p(1-p)}{n}}$$

using $\hat{p}$ instead of p.

The confidence interval is formed by [estimate - ME, estimate + ME].

We are 95% confident that the proportion of voters who approve of job being done by president between 0.382 and 0.432. Based on this survey we believe that between 38% and 43% of voters approve of president trump with 95%. We do not say there is 95% chance that the true proportion is between 0.382 and 0.432. Or that the chance true proportion is between the interval computed is 0.95. There is no "probability" to speak of here. Because there is no random quantity.

When we say 95%, we are referring to shorthand for:

"the interval calculated by a method that in a large number of repeated samples the resulting interval would contain the true proportion about 95% of the time."

April 12, 2018

Imagine a large population of individuals. The proportion of which possess a total of interest is p with interval $0<p<1$. Take a random sample of n of them and calculate

$$\hat{p}=\frac{\#\text{ of success in sample}}{n}$$

Do this a lot of times and make a histogram of difference $\hat{p}$'s you get. It will take a normal curve. Then we have

$$\hat{p}-2.5\text{EP}<p<\hat{p}+2.5\text{EP}$$

For about 25% we would have

$$p < \hat{p} - 2.5\text{EP}$$

For about 2.5%, we will have

$$p > \hat{p} + 2.5\text{EP}$$

The interval

$$[\hat{p} - 2.5\text{EP}, \hat{p} + 2.5\text{EP}]$$

or $\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is called a 95% confidence interval for p. A polling agency does not take a lot of samples, they take one.

Is our sample one of the 95% for which the confidence interval contains the true p? Or one of the 5% that misses it? We cannot even know.

Example 0.0.98. Insurance company collects date on 582 accidents notes that 91 of them involved a teenage driver. What can you say about the percent of all accidents that involve teenage driver? We compute

$$\hat{p} = \frac{91}{582} = 0.156$$

and we have

$$\text{SEP} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(.156 \times .844)}{582}} = 0.015$$

We estimate that 15.6% of accidents involve a teenage driver. We are 95% confident that between 12.6% and 18.6% of accients involve teenage drivers.

Chapter 19

Imagine a large population of individuals, associated with each is a number. Let μ be average value for population. Let σ be SD of the population values. Imagine we took a whole bunch of random samples of size n, calculated the sample mean, call it $\bar{x}$, for each plot a histogram of those $\bar{x}$ - values. It will look like a normal curve with mean = μ (population mean). Then SD = $\frac{\sigma}{\sqrt{n}}$ with population SD / square root of sample size. That's the rule of sample means from Section 19.3. From this result we can conclude for 95% of possible samples, there are

$$\mu - 2\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 2\frac{\sigma}{\sqrt{n}}$$

realized value within 2 SDs of expected value. For 95% of possible samples. There is

$$\bar{x} - 2\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{\sigma}{\sqrt{n}}$$

then we have 95% of confidence interval for the true population mean to be

$$\text{sample mean } \pm 2 \times \text{SEM}$$

while SEM = SD/$\sqrt{n}$. In practice, the population Sd is unknown, so use sample SD in its place.

Example 0.0.99. Food supplement promote weight gain in livestock? 70 cows, the average weight gain was 55 pounds, the SD was 22 pounds. What can we say about the mean weight gain for all such cows?

Our best estimate is 55 pounds (the sample mean). The std error of our estimate is sample error of mean, SEM, would be SD/$\sqrt{n}$ = 22 / $\sqrt{70}$ = 2.63. This means, for 95% confidence, the margin of error is 2(2.63) = 5.26.

Thus, a 95% confidence interval is estimate ± margin of error = sample mean $\pm 2\times$ SEM = $\bar{X} \pm 2\frac{\text{SD}}{\sqrt{n}}$, that is, $55 \pm 5.3 \Rightarrow [55 - 5.3, 55 + 5.3] = [49.7, 60.3]$.

We conclude that we are 95% confident that the mean weight gain for cows on this supplement is between 50 and 60 pounds.

April 17, 2018

Example 0.0.100. Study of 50 parents of children diagnosed ADHD the average stress level (based on PSS index) is 53, SD is 10. Find a 95% CI for the mean stress level of all such parents. Here are the steps Call the following Sample size n = 50, sample mean $\bar{x} = 50$, sample SD = 10. Then
(1) $\text{SEM} = \frac{\text{SD}}{\sqrt{n}} = \frac{10}{\sqrt{50}} = 1.41$
(2) Margin of Error: $\text{ME} = 2 \times \text{SEM} = 2 \times 1.41 = 2.8$
(3) Standard Error: $\text{SE} = \frac{\text{SD}}{\sqrt{n}}$ (4) Use Margin of Error, which is twice Standard Error, to construct 95% confidence interval, $[53 - 2.8, 53 + 2.8]$. (5) Interpretation: we are 95% confident that the mean stress level (PSS index) for parents of ADHD children between 50.2 and 55.8. (6) Can you say anything for population? No. But we can say something about sample population.

Example 0.0.101. Random sample of 58 parents of non-ADHD kids. The average stress level is 45 with a SD of 8. A 95% confidence for the mean stress level of all such parents would be
(1) $\bar{x} \pm 2\frac{\text{SD}}{\sqrt{n}} = 45 \pm 2\frac{8}{\sqrt{58}} = 1.05$
(2) Margin of error would be: $2 \times 1.05 = 2.1$
(3) 95% confidence interval would be: $[45 - 2.1, 45 + 2.1] = [42.9, 47.1]$
(4) Interpret: We are 95% confident that the mean stress level of all such parents is between 42.9 and 47.1.
(5) What can we say about parents of ADHD kids vs. other parents?

	n	$\bar{x}$	SD	95% CI
ADHD	50	53	10	[50, 56]
Other	58	45	8	[43, 47]

We might conclude that the difference between the two means is between 3 and 13, computed from 50 - 47 and 56 - 43, respectively, from above table. This is not wrong. But we can do better. The formula for the std error of a difference between two means is

$$\text{SED} = \sqrt{\text{SEM}_1^2 + \text{SEM}_2^2} = \sqrt{\frac{10^2}{50} + \frac{8^2}{58}} = 1.76$$

A 95% confidence interval for the difference between the two population means is

$$\text{dif in sample means } \pm 2 \times \text{SED} = 8 \pm 2(1.76) \Rightarrow [8 - 3.5, 8 + 3.5] = [4.5, 11.5]$$

and this results corresponds to the answer above that is between 3 and 13. We can get more precise answer if we get use of the formula.

We can now conclude: we are 95% confident that the mean stress level for parents of kids diagnosed with ADHD is between 4.5 and 11.5 points higher.

We have seen now 3 different formulas for standard error.

(1) key words for proportion, percentage, or generally yes or no type of data, we use

$$\text{SEP} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

(2) key words for mean, average, sample mean, standard deviation, we use

$$\text{SEM} = \frac{\text{sample SD}}{\sqrt{n}}$$

(3) key words for difference between, two different sample means, standard deviations, we use

$$\text{SED} = \sqrt{\text{SEM}_1^2 + \text{SEM}_2^2}$$

This is just like Pythagorean theorem. Error in mean 1 implies horizontal move. Error in mean 2 implies vertical move.

Now of these is fool proof, you still gotta think.

Example 0.0.102. Survey of 834 CA voters, 450 favors legalization. What percent of all CA voters favor legalization?

Answer: Check that $\frac{450}{834} = 0.54$ or 54%, which gives us estimate $\hat{p} = 54\%$. The standard error of this estimate is

$$\text{SEP} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.54 \times 0.49}{834}} = 0.017$$

and then using such result, we construct 95% confidence interval to be

$$\hat{p} \pm 2 \times \text{SEP} = 0.54 \pm 2(0.017) \Rightarrow [0.54 - 0.035, 0.54 + 0.035]$$

Data suggests that legalization is favored by a majority right? More than 50%, yes.

Chapter 22

Chapter 22 describes another method for assessing this question. Recall Hypothesis Testing, four steps (p475).

(1) Formulate NH and AH

(2) Collect data, and calculate a one-number summary of how different the data are from what we'd expected if NH were true. i.e. calculate a test statistic χ^2 in a context of Chapter 13, but that's a special case.

(3) Determine how unlikely the data would have been if NH were true, i.e. calculate the p-value.

(4) Interpret the p-value and reach a conclusion.

Example 0.0.103. Do the data (834 surveyed, 450 in favor) prove that a majority of California voters favor legalization?

(1) Let us formulate NU and AH:
NH: proportion who favor is 0.50 ($\leq 50\%$)
AH: proportion who favor is ¿ 0.50

(2) Compute
The test statistics for this situation is given on page p.480. that is,

$$\text{test stat } = \frac{\hat{p} - \text{HV}}{\text{Null SE}}$$

where $\hat{p}$ = sample proportion, HV = hypothesized value, and

$$\text{Null SE} = \sqrt{\frac{\text{HV} \times (1 - \text{HV})}{n}}$$

so we compute

$$\text{Null SE} = \sqrt{\frac{.5 \times .5}{834}} = 0.017$$

and $\hat{p} = \frac{450}{834} = 0.54$. Then we have

$$\text{test stat } = \frac{\hat{p} - \text{HV}}{\text{Null SE}} = \frac{0.54 - 0.50}{0.017} = 2.31$$

(3) p-value
Let us continue next time.

April 19, 2018

Chapter 22 & 23

Hypothesis Testing is a decision-making framework for deciding between two competing claims about the world conclude AH only if there's compelling evidence for H. We seek to "disprove" (to contradict) NH. Either we reject null hypothesis or fail to reject null hypothesis. Absence of evidence against NH. "$\not\Rightarrow$" evidence in favor of NH.

There are four steps:
(1) Formulate NH and AH
(2) Collect data and compute a measure of how different those data are from what we would have expected if NH were true, i.e. compute test statistic
(3) Compute a measure of how unlikely the observed test stat value would have been if NH were true, i.e. compute p-value
(4) Interpret that p-value: explain clearly what that hypothesis means in the context of this problem. State your conclusion.

Definition 0.0.104. The p-value is the probability, supposing NH were true, of a sample outcome as inconsistent with NH (more so) as the outcome actually observed. The lower the p-value, the stronger the evidence against NH.

Example 0.0.105. Survey of 834 votes, 450 in favor. Do the data prove that a majority of voters are in favor? That is, compute

$$\frac{450}{834} = .54 \text{ or } 54\%$$

(1) NH: proportion in favor is 0.5. AH: proportion in favor is greater than 0.5.
(20 For many situations, the appropriate test stat takes the form

$$\text{test stat} = \frac{\text{observed value} - \text{Null EV}}{\text{Null SE}}$$

From Chapter 19, we know if NH were true, the sample proportion $\hat{p} \sim \mathcal{N}(\approx)$ with a mean of $p = 0.5$ and a SD of $(\frac{.5\times.5}{834})^{.5} = 0.0173$. The Null EV = 0.5 and Null SE = 0.0173.

(2) Then we compute

$$\text{test stat} = \frac{0.54 - 0.5}{0.0173} = 2.31$$

which is the z-score that counts the number of standard deviation from observed value to mean. If NH were true, the test stat is a standard normal variable, a z-score.

(3) How unusual would test stat = 2.31 be if NH were true? Consult with t-table and we find out that the p-value is 0.01.

(4) If the true proportion in favor was 0.50, the probability of observing a sample proportion as high as that observed (i.e. 54% of 834 people) would be 0.01, one in a hundred. This is fairly strong evidence that in fact the proportion in favor is greater than 0.5.

Hypothesis Testing: concerning population proportion:

$$\text{test stat} = \frac{\text{obs value} - \text{Null EV}}{\text{Null SE}}$$

Null EV is hypothesis value, and null SE is

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

To find p-value compare the test stat, to std normal distribution (Table 8.1) to see how unusual it is.

Numerical data (quant variables).

Example 0.0.106. All parents' lives are stressful is the mean stress higher for parents of kids with ADHD? Conduct a test of hypotheses to find out. The data (response variable is "PSS") parental stress scale.

	n	$\bar{x}$	SD
ADHD	50	53	10
Non	58	45	8

(1) Let us state NH and AH:

AH: mean PSS for parents of ADHD kids is > than mean PSS for other parents.

NH: replace ">" with "no greater than". Or "=" or "≤"

(2) Test stat:

$$\text{test stat} = \frac{\text{obs value} - \text{null EV}}{\text{Null SE}}$$

and we note

$$\text{observed value} = 53 - 45 = 8$$

the null EV is 0 because there is no difference. Then the null SE is

$$\text{SED} = \sqrt{\text{SEM}_1^2 + \text{SEM}_2^2} = \sqrt{(\frac{10}{\sqrt{50}})^2 + (\frac{8}{\sqrt{53}})^2} = 1.76$$

and then we have test stat

$$\text{test stat} = \frac{8-0}{1.76} = 4.55$$

Note strictly speaking, it is not correct to compare this value to the standard normal distribution. We should be using student's t-distribution.

(3) Note on reading/exam prep: you will not be tested on properties of students t-distribution. You will not be tested on one-sided vs two-sided hypothesis tests, always do one-sided. The observed value is 4.5 standard errors above what we would expect if NH were true. p-value is LOW (close to 0). If mean PSS were no higher for parents of ADHD kids than other parents, the sample difference we observed would have been highly unlikely.

The data provide compelling evidence that indeed mean PSS is higher for parents of ADHD kids than for other.

(4) Interpret the p-value. Write your conclusion. If you are a researcher you are done. If you are a business manager, military officer, etc., you may be compelled to chaos. One course of action over another based on whether you reject null hypothesis or fail to reject null hypothesis.

Decision rule: Reject NH if p-value $< \alpha$ while α is called the level of significance. Suppose we reject NH, but in fact NH is true. This is called Type 1 Error. If we fail to reject NH, when in fact AH is true, that's a Type 2 Error. The more costly is a Type 1 Error the lower we should set α. So set $\alpha = 0.000001$, and practically never make a Type 1 Error.

Remember AH usually is the thing we hope to prove. Setting α too low, we never make Type 1 Errors (good), but we never prove anything ever (bad). Anything that lower the probability of Type 1 Error necessarily increases the probability of Type 2 Error.

Remark 0.0.107. The final exam is Thur, May 10th, 1:10PM regular classroom. You are allowed a calculator. Two 8.5 by 11 sheets (4 sides) of original hand-written notes. The entire course is fair game.

But will heavily emphasize material after 2nd midterm.

Ch 14-17, 19-24;

Hws: 7-10;

Last day of class is Thur. April 26.

Review session optional 2:40 - 3:55 here on Thursday May 3rd.

You will have 150 minutes to complete the exam, we'll start as close to 1:10pm as we can.

April 24, 2018

Chapter 22-24

Hypothesis Testing.

Four steps in total:

(1) Formulate Null and Alternative Hypothesis: AH is the thing were trying to prove. If question says "do the data provide evidence that ... ". The ... is AH! The Null is the border of the AH region.

(2) Collect data and calculate test statistic a measure of how different the data are versus what we'd expect if null were true.

(3) p-value = probability (supposing that Null is true) of a sample at come at least as inconsistent with Null as that observed.

(4) Interpret the p-value. Reach a conclusion. make decision: (i) reject null, (2) fail to reject.

A template for interpreting a p-value. If x, then the probability x would be x. Thus the data provide x evidence that in fact x. Fill in these blanks with the following: adapted to language/subject matter of actual problem. That is, If Null H were true, then the probability of getting a sample result like that observed would be p-value. Thus the data provide p value greater than 0.10 or say no evidence (or p-value greater than 0.05 but less than 0.10, some indication) (or p-value less 0.05 but greater than 0.01, some evidence) (or p-value less than 0.01, fairly strong evidence) (or p-value approximately 0, overwhelming evidence) that in fact Alternative H is true.

Two important formulas for Hypothesis Testing.

If this is a problem about proportions, both look like

$$\text{test stat} = \frac{\text{Obs value} - \text{Null EV}}{\text{Null SE}}$$

Type of Problem	Observed Value	Null EV	Null SE
Test about proportion (or %)	Sample Proportion	Pop'n Proportion in Null (HV)	$\sqrt{\frac{\text{HV}\times(1-\text{HV})}{n}}$
Test about pop'n mean	Sample Proportion	Pop'n mean in NH	SD/$\sqrt{n}$ = SEM
Diff. b/w two means	Diff. b/w sample means	Diff. in pop;n means under NH (often 0)	SED = $\sqrt{\text{SEM}_1^2 + \text{SEM}_2^2}$

Example 0.0.108. bank surveys 230 account holders find 180 satisfied customers. Do the data prove that more than 75% of banks customers are satisfied?

Answer:

(1) State the hypothesis:

NH: Proportion satisfied is 75%
AH: Proportion satisfied is more than 75%
(2) Compute test statistics:
Note that observed value $\hat{p} = \frac{180}{230} = 0.7826$ and also Null EV $= 0.75$ according to the problem. The null standard error is

$$\sqrt{\frac{0.75 \times 0.25}{230}} = 2.7\%$$

and hence test statistics would be

$$\frac{0.7826 - 0.75}{0.027} = 1.14$$

(3) Find p-value to be 0.13 from the table. More, AUC to left is 0.87 and p-value would be 0.13
(4) How to interpret? If in fact 75% of banks customers are satisfied, the probability of getting a sample percent like we observed (78.26%) would be 0.13, not not unusual at all. Thus the data are consistent with the true percent satisfied being 75%.

Example 0.0.109. Random sample of 50 children 7-10. Average sugary beverage per day of 8.2 ounce with SD of 7.2 oz. Random sample of 20 children age 11-13 drink average of 14.5 oz, SD of 10.7 oz. Do the data prove older kids drink more soda on average? If yes, how much more?

Answer:
(1) State the hypothesis:
NH:both age groups have the same mean soda consumption
AH: mean soda per day for older kids is greater than mean soda per day for younger
(2) Compute test statistics:
Observed value $= 14.5 - 8.2 = 6.3$ and that means from the sample the older kids drink 6.3 oz more soda on average per day. Null EV $= 0$ because from the problem we would expect there is no difference (assuming null were true).

$$\text{SE for Diff} = \sqrt{\text{SEM}_1^2 + \text{SEM}_2^2} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{7.2^2}{50} + \frac{10.7^2}{20}} = 2.6$$

and we have standard error. Hence test statistics would be

$$\frac{6.3 - 0}{2.6} = 2.42$$

(3) Find p-value to be very small approximately 0.01 from table. (4) How to interpret? If soda consumption for these two age groups were the same, the probability of observing a sample difference as great as we observed would be 0.01, one chance in 100. The data provide fairly compelling evidence that the older kids are drinking more soda on average. (5) How much more? Let's do a confidence interval. An approximate 95% confidence interval for the difference b/w average soda consumption of older kids versus that of younger kids is

$$\text{estimate} \pm 2 \times \text{SE} \Rightarrow 6.3 \pm 2 \times 2.6 \Rightarrow [1.1, 11.5]$$

and thus we are about 95% confident that the mean soda consumption for 11-13 year olds is b/w 1.1 and 11.6 oz per day greater than that of 7-10 year olds.

Remark 0.0.110. In Example 1 for today we get p-value = 0.13. we would conclude do not reject NH. What if in fact AH were true, i.e. the true percent satisfied is 75%? Then we'd have made a Type II error.
In Example 2, say we conclude reject null and that older kids do drink more soda. What if they don't? We'd would Type I error.

Remark 0.0.111. Three important caveats on Hypothesis Testing:
(1) p-value not close to 0; this would imply that data are consistent with Null H. It implies that data give no strong evidence in favor of AH. This implies that no strong evidence against NH. It does not imply that there is evidence in favor of NU. Absence of evidence against NH does not necessarily imply evidence in favor of NH.
(2) p-value is close to 0; ⇒ difference is statistically significant; this does not imply that the difference is practically important.

Example 0.0.112. Consider an exam on which mean score is 525 and SD is 100. There are 200 students take coaching program. Their average is 535. Does it prove that the program was effective? Compute

$$\text{test stat} = \frac{535 - 525}{100/\sqrt{200}} = 1.41$$

and this gives us p-value about .08.
Moreover, 2000 students take a different coaching program and their average 530. Does it prove that the second program is effective?

$$\text{test stat} = \frac{530 - 525}{\sqrt{100/\sqrt{2000}}} = 4.47$$

and this gives us p-value approximate to 0.
For the first program, the coaching effect was not statistically significant. For the second program, it definitely is effective. But is it practically important? Not really. Just because it is statistically significant does not mean it is practically significant.

April 26, 2018

Example 0.0.113. Standardized exam, mean score is 525, and SD = 100. 200 students take a coaching program. Their average was 535. Do the data prove the program worked?

Answer: (1) Hypothesis:
NH: Mean score in program is equal to 525
AH: Mean score in program is greater than 525

(2) Test statistic:

$$\frac{\text{observed value} - \text{Null EV}}{\text{Null SE}}$$

note that the observed value is 535, Null EV = 525 and we want SEM to be $\text{SD}/\sqrt{n} = 100/\sqrt{200} = 7.07$. and hence we have

$$\frac{535 - 525}{7.07} = 1.41$$

to be our test statistics.

(3) p-value. This gives us a p-value to be 0.08.

(4) Interpret: If the mean score for program participants were 525, the probability of getting a sample mean of 535 or higher would be 0.08, not so unusual. The data do not prove the program is effective.

Example 0.0.114. 2000 candidates take a different program. Their average is 530. Do the data prove this program works? NU and AH same as above. Then we compute

$$\text{Null SE} = \text{SEM} = \frac{\text{SD}}{\sqrt{n}} = \frac{100}{\sqrt{2000}} = 2.24$$

and

$$\text{test stat} = \frac{530 - 525}{2.24} = 2.24$$

and this gives us p-value less than 0.02. Interpret? If the mean score for all students take this program were 525, the probability of getting a sample mean of 530 or higher would be less than 0.02. We have fairly compelling evidence that this program is effective.

Program 1 has 535 average, not proven effective. Program 2 has 530 average, it is effective (we may conclude). That is, program 2 increases the average score. By how much? Let us construct a 95% confidence interval.

$$\text{estimate} \pm 2 \times \text{SE} \Rightarrow 530 \pm 2(2.24) \Rightarrow [525.5, 534.5]$$

and thus we say that we are 95% confident that the mean score of all candidate go through Program 2 is b/w 525.5 and 534.5.

Three caveats about Hypothesis Testing

(1) As of today's class, from Example 1 we have p-value greater than 0.10. $\Rightarrow$ No compelling evidence for AH. $\Rightarrow$ No evidence against Null. Data are consistent with Null being true. $\not\Rightarrow$ Evidence in favor of Null.

The slogan is "absence of evidence that there is an effect does not imply evidence that there isn't one."

(2) p-value is less than 0.05. $\Rightarrow$ The effect is statistically significant. $\Rightarrow$ Strong evidence that there is same effect. $\not\Rightarrow$ Evidence that there is a strong effect.

We rejected the Null that coaching didn't work. We concluded that coaching has some benefit. When we went to quantify that benefit, we found it could be a little as half a point!

Failing to reject Null $\not\Rightarrow$ Accepting Null. Statistical significance $\not\Rightarrow$ Practical significance. Issue small is particularly a concern if sample size is big. Then we can compute

$$\text{test stat} = \frac{\text{observed value} - \text{Null EV}}{\text{Null SE}} = \frac{\bar{x} - \mu_0}{\text{SD}/\sqrt{n}}$$

Even a small value of observed value - Null EV will give a big test stat if n is big. With two few data, even if there is an effect, their may not be enough data to detect it.

(3) The "multiple testing" issue. Google "I fouled millions into thinking chocolate helps weight loss". Here's how. Get a group of volunteers. Half eat a few ounces of dark chocolate per day, half don't. After a period of time. Measure 18 different health outcomes for all the subjects, eg. weight, cholesterol, sodium, blood protein, sleep ability, well being, and 12 others.

Do 18 different tests of Null H. chocolate has no effect; Alternative H. chocolate has beneficial effect. Suppose the truth is chocolate has no effect. The probability that at least one effect is statistically significant = probability at least one Type 1 error = probability at least one p-value is less than 0.05. = 1 - probability none are significant = 1 - probability all p-values greater than 0.5 = $1 - (0.95)^{18}$ = 1 - 0.4 = 0.6.

If chocolate has no effect, we've got 60% chance that at least one outcome is statistically significant. Write a paper about that one. Don't mention the others.

Discuss final .

Review session thursday May 3rd. 2:40 PM. Two things will be on the final:

(1) Give a rigorous interpretation of the p-value in a given context.

(2) Chapter 17 Exercise 19.

Index

References

[1] Jessica, M. Utts, Seeing Through Statistics, 4th Edition