HIMANSHU SINGH

Definition. *Statistics* is science of collecting, describing & analysing data.

Definition. *Data* is set of measurements taken on a set of individual sets

Usually, data is stored and presented in a dataset, comprised of variables measured on cases.

Definition. The subjects or objects that we obtain information from are called the CASES or UNITS.

The cases are either the people being studied or the things being studied.

Example. Students in the UT Tyler; engineering majors

Definition. A *variable* is any characteristics that is recorded for each case.

Remark. Generally, each case corresponds to a row in a dataset, and each variable makes up a column.

Example. For each data set below, determine the following.

- What are the cases?

- What are the variables?

CASES

variables

Country	Land Area	Population	Rural	Health	Internet	Birth Rate	Life Expectancy	HIV
Afghanistan	652230	29021099	76	3.7	1.7	46.5	43.9	
Albania	27400	3143291	53.3	8.2	23.9	14.6	76.6	
Algeria	2381740	34373426	34.8	10.6	10.2	20.8	72.4	0.1
American Samoa	200	66107	7.7					
Andorra	470	83810	11.1	21.3	70.5	10.4		
Angola	1246700	18020668	43.3	6.8	3.1	42.9	47	2
Antigua and Barbuda	440	86634	69.5	11	75			
Argentina	2736690	39882980	8	13.7	28.1	17.3	75.3	0.5

VARIABLES

Year	Gender	HigherSAT	SAT	GPA	Sibling	Height	Weight	Exercise	TV	Pulse	Award
Senior	M	Math	1210	3.1	4	71	180	10	1	54	Olympic
Sophomore	F	Math	1150	2.5	2	66	120	4	7	66	Academy
FirstYear	M	Math	1110	2.6	2	72	208	14	5	130	Nobel
Junior	M	Math	1120	3.1	1	63	110	3	1	78	Nobel
Sophomore	F	Verbal	1170	2.7	1	65	150	3	3	40	Nobel
Sophomore	F	Verbal	1150	3.2	2	65	114	5	4	80	Nobel
FirstYear	F	Math	1320	2.8	1	66	128	10	10	94	Olympic
Sophomore	M	Math	1370	3.3	1	74	235	13	8	77	Olympic
Junior	F	Verbal	1100	2.8	2	61		3	6	60	Nobel
FirstYear	F	Math	1370	3.7	7	60	115	12	1	94	Nobel
Sophomore	F	Math	1170	2.1	1	65	140	12	6	63	Olympic
First-year	M	Math	1180		2	63	200	10	5	72	Olympic
Sophomore	M	Math	1150	2.9	3	68	162	12	8	54	Olympic
Junior	F	Verbal	1300	3.1	2	68	135	6	1	66	Nobel
First-year	M	Verbal	1350		1	68	193	9	5	72	Nobel
FirstYear	F	Math	1200	3.9	1	63	110	10	2	59	Olympic
FirstYear	F	Verbal	1200	3	2	63	99	3	15	88	Olympic
Sophomore	M	Verbal	1350	3	2	72	165	7	3	59	Nobel

students are CASES.

variables

Name	Company	Serving	Calories	Fat	Sodium	Carbs	Fiber	Sugars	Protein
AppleJacks	K	1	117	0.6	143	27	0.5	15	1
Boo Berry	G	1	118	0.8	211	27	0.1	14	1
Cap'n Crunch	Q	0.75	144	2.1	269	31	1.1	16	1.3
Cinnamon Toast Crunch	G	0.75	169	4.4	408	32	1.7	13.3	2.7
Cocoa Blasts	Q	1	130	1.2	135	29	0.8	16	1
Cocoa Puffs	G	1	117	1	171	26	0.8	14	1
Cookie Crisp	G	1	117	0.9	178	26	0.5	13	1
Corn Flakes	K	1	101	0.1	202	24	0.8	3	2
Corn Pops	K	1	117	0.2	120	28	0.3	15	1
Crispix	K	1	113	0.3	229	26	0.1	3	2
Crunchy Bran	Q	0.75	120	1.3	309	31	6.4	8	1.3
Froot Loops	K	1	118	0.9	150	26	0.8	12	2
Frosted Mini-Wheats	K	1	175	0.8	5	41	5	10	5
Golden Grahams	G	0.75	149	1.3	359	33	1.3	14.7	2.7
Honey Nut Clusters	G	1	214	2.7	249	46	2.8	17	4
Honey Nut Heaven	Q	1	192	3.7	216	38	3.5	13	4
King Vitaman	Q	1.5	80	0.7	173	17	0.9	4	1.3

cereals are CASES.

A

Remark. Variables are classified as either CATEGORICAL or QUANTITATIVE.

Definition. A *categorical* variable is a variable that divides the case into groups.

Categorical variables are usually, but not always, recorded as words or symbols representing words.

Definition. A *quantitative* variable is a variable that measures a numerical quant for each case.

Note: Numerical operations (+ and averging) are sensi for quantitative variable

WARNING: Sometimes categorical data can be recorded as numbers.

Example.
- What is your favorite number?
- Which of these colors is your favorite: (1) Red, (2) Blue, (3) Green?

Example. Classify each of the following variables from the Student Table data as either categorical or quantitative.

- year in school C
- gender C
- HigherSAT (which is higher ? Math or Verbal?) C
- SAT score Q
- GPA Q
- number of siblings Q
- height Q
- weight Q
- hours of exercise per week Q
- hours of TV per week Q
- pulse rate Q
- award preference (Olympic Gold, Academy Award, or Nobel Prize) C

Data can be used to answer interesting questions.

Example. For each question below, determine the variables and whether each is categorical or quantitative.

Weight (Q, RV) (C, EV)

1. Can eating a cup of yogurt each day change your weight?

can of MD consume/day (Q, RV)

2. Does loud music cause people to consume more cans of Mountain Dew?

C, EV

3. Are lions more likely to attack prey on the night of a full moon?

(Q, RV) (C, EV)

Definition. When we use one variable to help us understand or predict values of another variable, the former is the EXPLANATORY VARIABLE, and the latter is the RESPONSE VARIABLE.

Example. For each part, determine which is the explanatory variable and which is the response variable.

1. Can eating a cup of yogurt each day change your weight?

2. Does loud music cause people to consume more cans of Mountain Dew?

HW 3. Does meditation help reduce stress?

HW 4. Does sugar consumption increase hyperactivity?

5. Are lions more likely to attack prey on the night of a full moon?

HIMANSHU SINGH

Suppose you wanted to determine what proportion of Americans prefer chocolate ice cream to chocolate chip cookie dough ice cream. One way to determine this is to ask every American. Is this practical? How else could you attack this question?

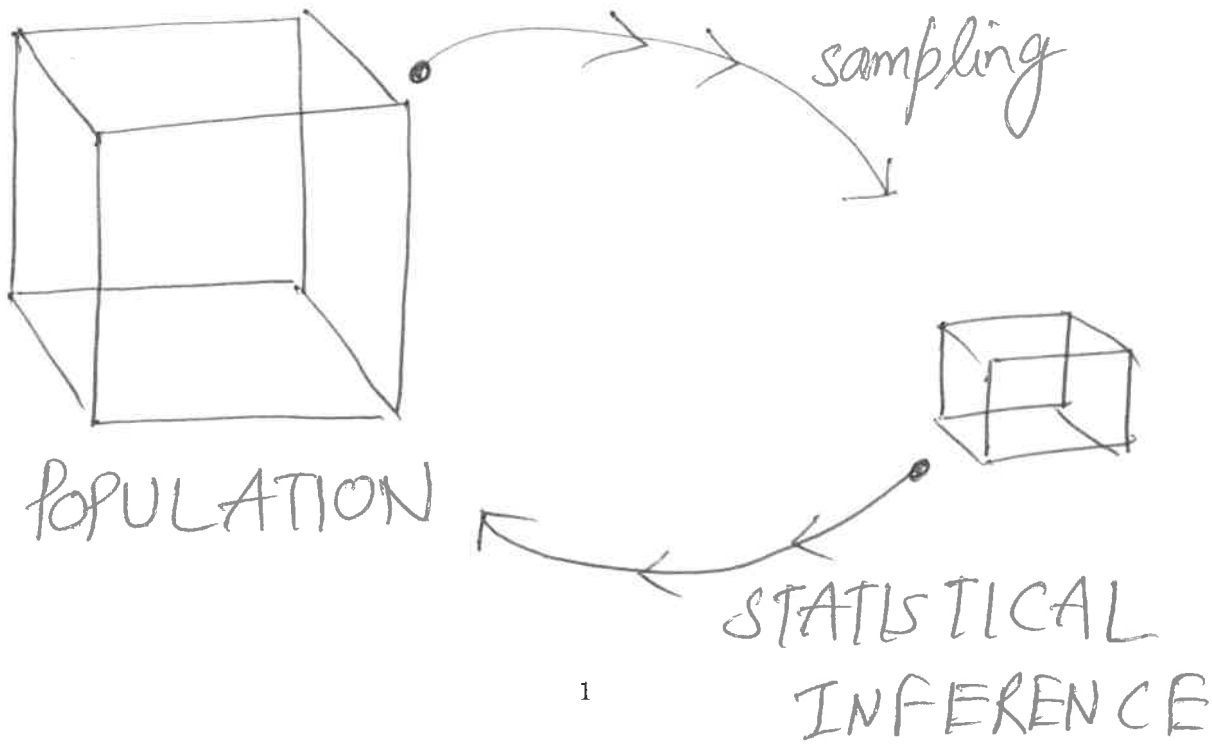
Definition. A *population* includes all individuals / objects of interest.

Definition. A *sample* subset of population for which we have data.

Remark. We usually denote sample size with a lowercase n. For example, if we conducted a survey of 43 Americans to ask whether they preferred chocolate or chocolate chip cookie dough ice cream, then in that case, we would have $n = 43$.

Definition. *Statistical inference* is the process of using data from a sample to gain info about the popl'

Diagram:



Example. Suppose the UT Tyler Office of Student Life wants to investigate what's most important to students. They decide to do this by having all the students in this section of MATH 1342 fill out a questionnaire.

1. What is the population?

   all UT Tyler students

2. What is the sample?

   MATH 1342 students

3. Can the sample data be generalized to make inferences about the population? Why or why not?

   No

Figure 1: Harry S Truman after winning the 1948 presidential election

Example. Suppose we want to know who is the most famous musical act in America. We survey 1,000 residents of assisted living facilities. What do you predict will happen?

They will answer with older musical acts.

Definition. *Sampling bias* occurs when sample selecting meth causes sample to ~~differ from popul^n in some relevant way~~

Remark. If sampling bias exists, we **cannot** trust generalizations from the sample to the population!

Example. "Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we can not dedicate – we can not consecrate – we can not hallow – this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us ? that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion – that we here highly resolve that these dead shall not have died in vain – that this nation, under God, shall have a new birth of freedom – and that government of the people, by the people, for the people, shall not perish from the earth."

We want to know the average length of the words in Lincoln's speech.

1. What is the population? all words in speech.

2. What are the cases? the words

3. What is the variable? Is it categorical or quantitative? word length / Q

4. Select a sample of 10 words from Lincoln's speech that you believe represent the whole speech. List those 10 words here.

5. Calculate the average number of letters for the words in your sample.

6. Actual average length of the words in Lincoln's speech:

7. Moral of the story:

The illness is sampling bias, but the cure is random sampling!

How could we construct a random sample of the words from Lincoln's speech?

↳ number the words → shuffle the number → draw first 10 numbers.
↓
use RNG. // FLIP a coin to determine whether each words goes in the sample.

Only a random sample can be trusted for making generalizations to the population.

Definition. In a Simple random sample, each unit of the population has the same chance of being selected, regardless of the other units chosen for the sample.

Example. The UT Tyler Office of Student Success wants to know the average number of hours students spend studying per week. Which of the following is the best method of sampling?

1. Go to Muntz Library and ask every student there how much they study
2. Email all UT Tyler students asking how much they study and use the replies volunteer bias
3. Ask all students in this section of MATH 1342 and make them respond before leaving class today
4. Stand on the sidewalk by Harvey Lake and ask every student who passes by

Ways To Ensure Your Sample Is Biased

- Sample units based on a trait obviously related to the variable you are studying

- Let people volunteer for the sample

- Word the question in a confusing manner

- Ask a leading question

- Poll people who can't or won't give a true answer

- Offer multiple choice answers but don't list all the choices

HIMANSHU SINGH

Example. Mice eating diets rich in olive oil have better memories and learning abilities than mice who don't eat much olive oil.

People whose cells are resistant to insulin, a major factor in Type II Diabetes, tend to be overweight.

Children whose fathers were over 40 at the time of the child's birth tend to be smarter.

Definition. Two variables are *associated if* values of one variable tend to be related to the values of the other variable.

Definition. Two variables are *causally associated if* changing the value of one variable influences the value of the other variable.

Example. For each headline below, answer the following questions. What is the explanatory variable, and what is the response variable? Does the headline imply a causal association?

1. "Daily Exercise Improves Mental Performance"

EV → daily exercise

RV → mental performance

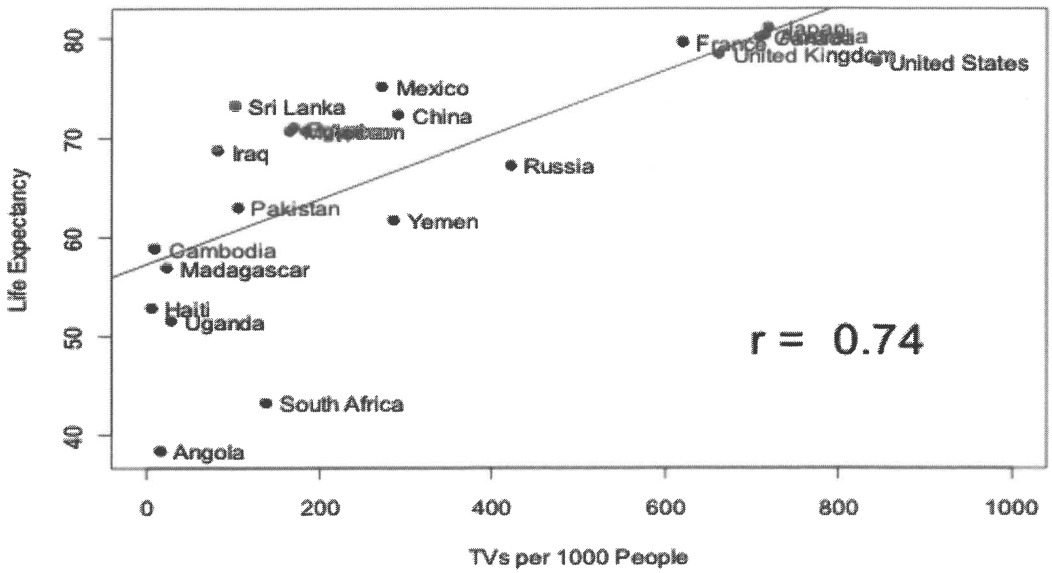
yes ⇒ causal association

2. "Want to Lose Weight? Eat More Fiber!"

EV → eat fiber

RV → weight

yes, ⇒ causal association



TVs / 1000 people & life expectancy are associated

but <u>not</u> causally associated.

(as one increases so does the other)

[THE CAUSE IS WEALTH $$]

Remark. Association does NOT imply causation!

Definition. A *confounding variable*, also known as a *confounding factor* or *lurking variable*, is a 3rd variable that is associated with both of the other variables.

Remark. A confounding variable can offer a plausible explanation for an association between the two variables of interest.

What could be a confounding variable in the last example? WEALTH $$.

Example. For each of the following relationships, identify a possible confounding variable.

1. More ice cream sales in Phoenix have been linked to more deaths by drowning.

   weather, time of year.

2. The total amount of beef consumed and the total amount of pork consumed worldwide are closely related over the past 1000 years.

   popul^n size.

3. People who own a yacht are more likely to buy a sports car.

   $$

4. Air pollution is higher in places with a higher proportion of paved ground relative to grassy ground.

5. People with shorter hair tend to be taller.

   Gender

Remark. Not every third variable is a confounding variable. For example, the journal *Nature* reported that dog ownership is related to longer life.

Remark. A confounding variable must be associated with both the explanatory variable and the response variable.

Definition. A *statistical experiment* is a study in which the researcher actively controls one or more of the EV.

Definition. An *observational study* is one in which the researcher does NOT actively control the value of any variable but simply observes values as they naturally occur.

Remark. There are almost ALWAYS confounding variables in observational studies. For this reason, observational studies can almost NEVER be used to establish causality.

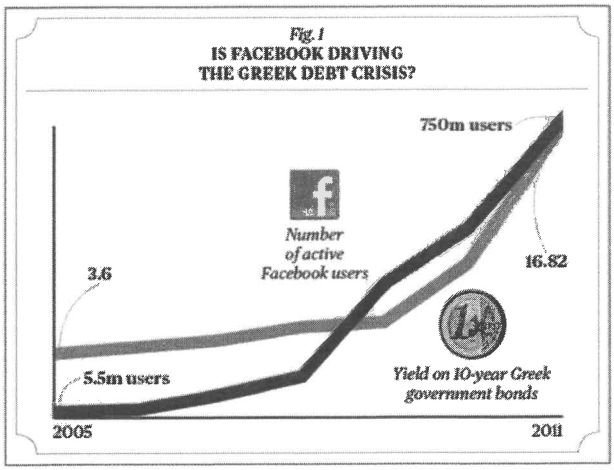
Example. Determine whether each describes a statistical experiment or an observational study.

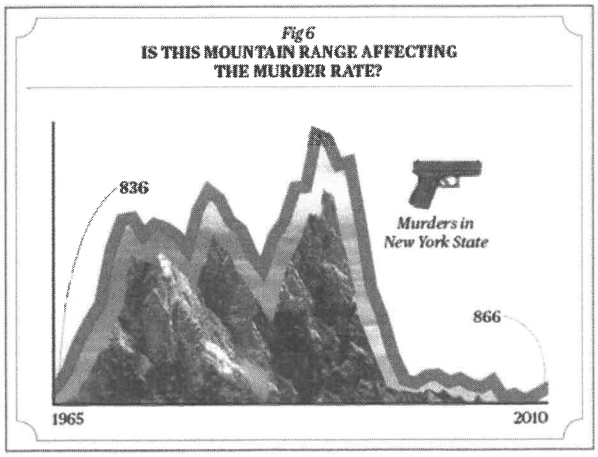
1. To examine whether eating brown rice affects metabolism, we ask a random sample of people whether they eat brown rice and we also measure their metabolism rate.

   observational study

2. To examine whether planting trees reduces air pollution, we find a sample of city blocks with similar levels of air pollution and we then plant trees in half of the blocks in the sample. After waiting an appropriate amount of time, we measure air pollution.

   STATISTICAL EXPERIment





associated but not causally

EXPERIMENT

Three Steps to a Statistical Study

1. take a RANdom sample from popul^n / unless the popul^n is small enough that you can deal with all cases.
2. control CONFOUNDING variables.
3. RANDOMIZE the EV.

Definition. In a *randomized experiment*, the value of EV for each unit is determined randomly, ~~before the RV is~~ measured

Two Fundamental Questions About Data Collection

- Was the sample randomly selected?
- Was the explanatory variable randomly assigned?

Diagram WAS the sample randomly selected?

✓ POSSIBLE to generalize from the sample to popul^n

✗ CAN'T GENeralize from the sample to the popul^n.

WAS THE EV RANDOMly assigned?

✓ possible to make conclusions about CAUSALITY

✗ NOT possible to make conclusions about CAUSALITY

Example. Aspirin

Definition. The values of the explanatory variable which the researcher controls are called TREATMENTS

Definition. In a RANDOMIZED COMPARATIVE EXPERIMENT, we assign cases to different treatment groups and then compare results on the response variables.

Definition. In a MATCHED PAIRS EXPERIMENT, each case gets both treatments in random order, and we examine individual differences in the response variable between the two treatments.

Example. Suppose you want to design an experiment to answer the question "Does alcohol increase reaction time?" Design a randomized experiment to address this question. Assume the participants are 40 college seniors and the response variable is time to react to an image on a screen after drinking either alcohol or water. Be sure to explain how randomization is used in each case.

- Design a randomized comparative experiment with two groups getting two separate treatments.

Number the students. Draw # out of a hat. The FIRST 20 drawn go into the water group; rest go into the alcohol group. Have each drink half a glass of their assigned beverage. Show the image on a screen. Record REACTION TIME.

- Design a matched pairs experiment.

Do everything described above. Drawing numbers out of a hat randomize who gets which treatment first. The next day, have the students come back and each drink the other beverage. Show image on the screen. Record Reaction time

- Which design is better for this situation? Why?

MATCHED PAIRS

PEOPLE have different reaction times naturally and you can compare an individual to himself

Definition. The CONTROL GROUP is the group in an experiment that does not receive the treatment or receives a fake version of the treatment.

A *placebo* is A SUBSTANCE OR TREATMENT THAT HAS NO ~~THERAPEUTIC EFFECT~~ USED AS A CONTROL.

Example. The *placebo effect* occurs when people **believe** they are getting an effective treatment and experience the desired effect regardless of whether the treatment is any good. For example, in the case of a medical study regarding a pill, giving both sets of people pills, fake for one group and real for the other group, makes it easier to see what fraction report results due to the placebo effect.

Definition. A *single-blind experiment* is an experiment IN which the subjects don't know whether they are taking a placebo or not.

Definition. A *double-blind experiment* is an experiment ~~in which neither the~~ researchers nor the subject know who is taking the placebo.

Given that there are almost ALWAYS confounding variables in observational studies, why would anyone ever do an observational study rather than an experiment?

HIMANSHU SINGH

Definition. When working with a categorical variable, the *proportion* in some category, or *relative frequency* of that category, is given by the formula

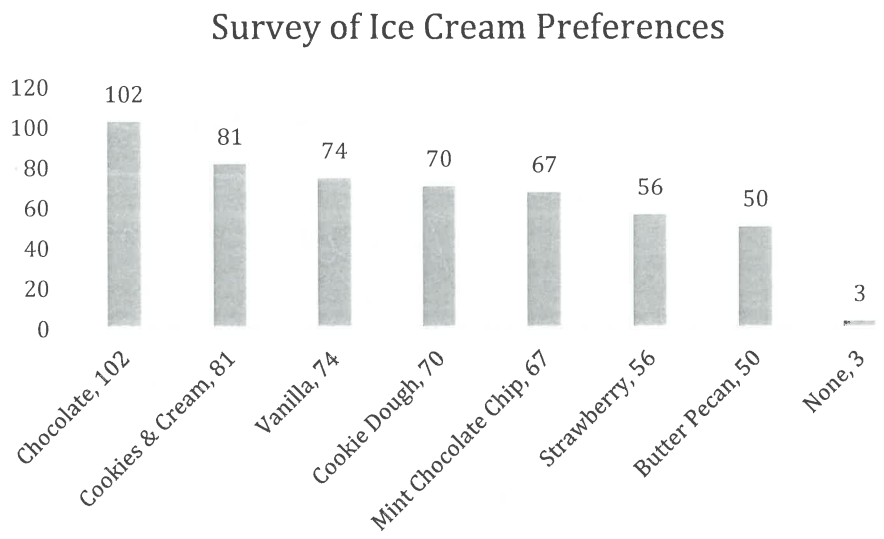
$$\text{proportion} = \text{relative frequency} = \frac{\#\text{ in the category}}{\text{Total number}}.$$

Example. A survey was conducted in which 503 Americans were asked which flavor of ice cream they liked best from a fixed list. The results are recorded below. What is the population?

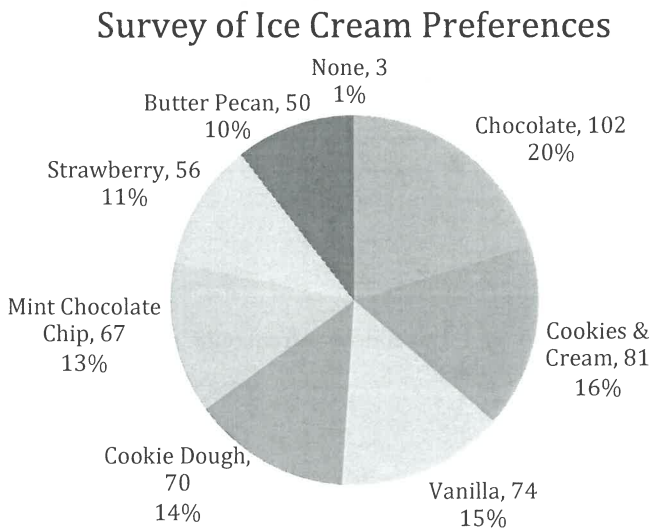
Response	Frequency	Relative Frequency
chocolate	102	102/503
cookies and cream	81	81/503
vanilla	74	...
cookie dough	70	...
mint chocolate chip	67	...
strawberry	56	
butter pecan	50	50/503
none	3	3/503
Total	503	

Remark. A table consisting of just Columns 1 and 2 would be a frequency table. A table consisting of just Columns 1 and 3 would be a relative frequency table.

Bar Chart HISTOGRAM



Pie Chart



Remark. The notation for a sample proportion is $\hat{p}$, which is read "p hat."

The notation for a population proportion is p.

Example. What is the proportion of survey respondents who prefer chocolate ice cream? Use correct notation.

$\hat{p} = 102/503$

Example. Now suppose that our population is students in this section of MATH 1342. What is the proportion of the population that prefers chocolate ice cream to the other flavors on the list? Again, use correct notation.

(do on your own)

EXAMPLE / HINT

$$p = \frac{\text{\#students prefers chocolate ice cream}}{\text{\#students in MATH 1342 (section .001 or .002)}}$$

Definition. A two-way table is used to show the relationship between two categorical variables. The categories for one variable are listed down the side, and the categories for the second variable are listed across the top.

Example. In the survey of 503 Americans, a variable regarding age was also recorded, giving the two-way table below.

Response	Frequency for Age ≤ 30	Frequency for Age > 30	Total
chocolate	59	102 − 59 = 43	102
cookies and cream	49	81 − 49 = 32	81
vanilla	34	40	74
cookie dough	47	23	70
mint chocolate chip	30	37	67
strawberry	27	29	56
butter pecan	10	40	50
none	0	3	3
Total	256	247	

1. What proportion of younger people in the sample prefer chocolate?

$\hat{p}_{Y,CH} = 59/256$

2. What proportion of older people in the sample prefer chocolate?

$\hat{p}_{O,CH} = \frac{43}{247}$

3. What is the difference in the proportion of younger people who prefer chocolate and older people who prefer chocolate?

$① - ② = \frac{59}{256} - \frac{43}{247}$ (approx. 0.06)

4. What is the proportion of people who prefer chocolate ice cream who are younger?

$\hat{p}_{CH,Y} = 59/102$

5. What is the proportion of people in the sample who are older?

$\hat{p}_{O} = 247/503$

HIMANSHU SINGH

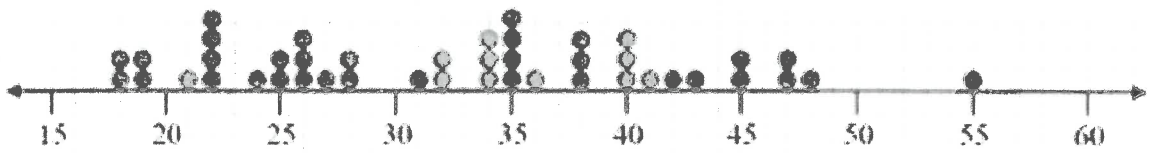
Math 1342 **One Quantitative Variable: Shape and Center** **2.2**

The table below records the amount in dollars spent by a young couple on fast food each week for 44 weeks.

32	22	19	18	45	42	40	43	18	21	55
31	26	22	25	47	40	26	32	22	34	34
28	35	47	26	35	38	35	28	19	38	35
38	36	25	22	45	48	27	34	41	40	24

One way to analyze the properties of quantitative data is to graph it on a dot plot.

Definition. We create a dot plot by using an axis with a scale appropriate for the numbers in the dataset and placing a dot over the axis for each case in the dataset. If there are multiple values that are the same, we stack the dots vertically.

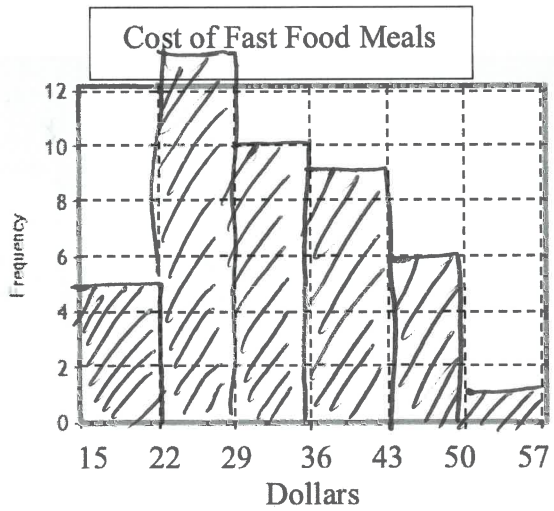


Amount Spent on Fast Food Restaurants during 40 Weeks

Definition. An outlier is an observed value that is notably distinct from the other values in a dataset. Usually, an outlier is much larger or much smaller than the rest of the data values.

Are there any outliers in the data on fast food spending? $55

Another type of graph for displaying a distribution of data is a histogram.

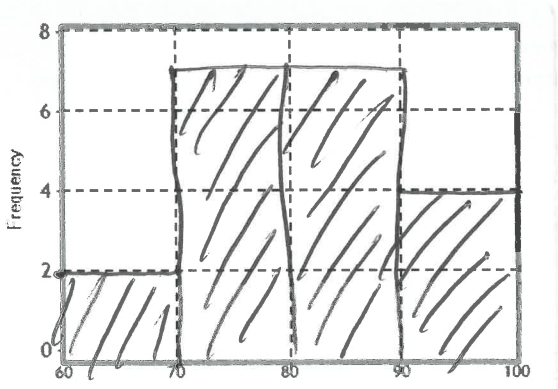


(look ☆ on page 1)

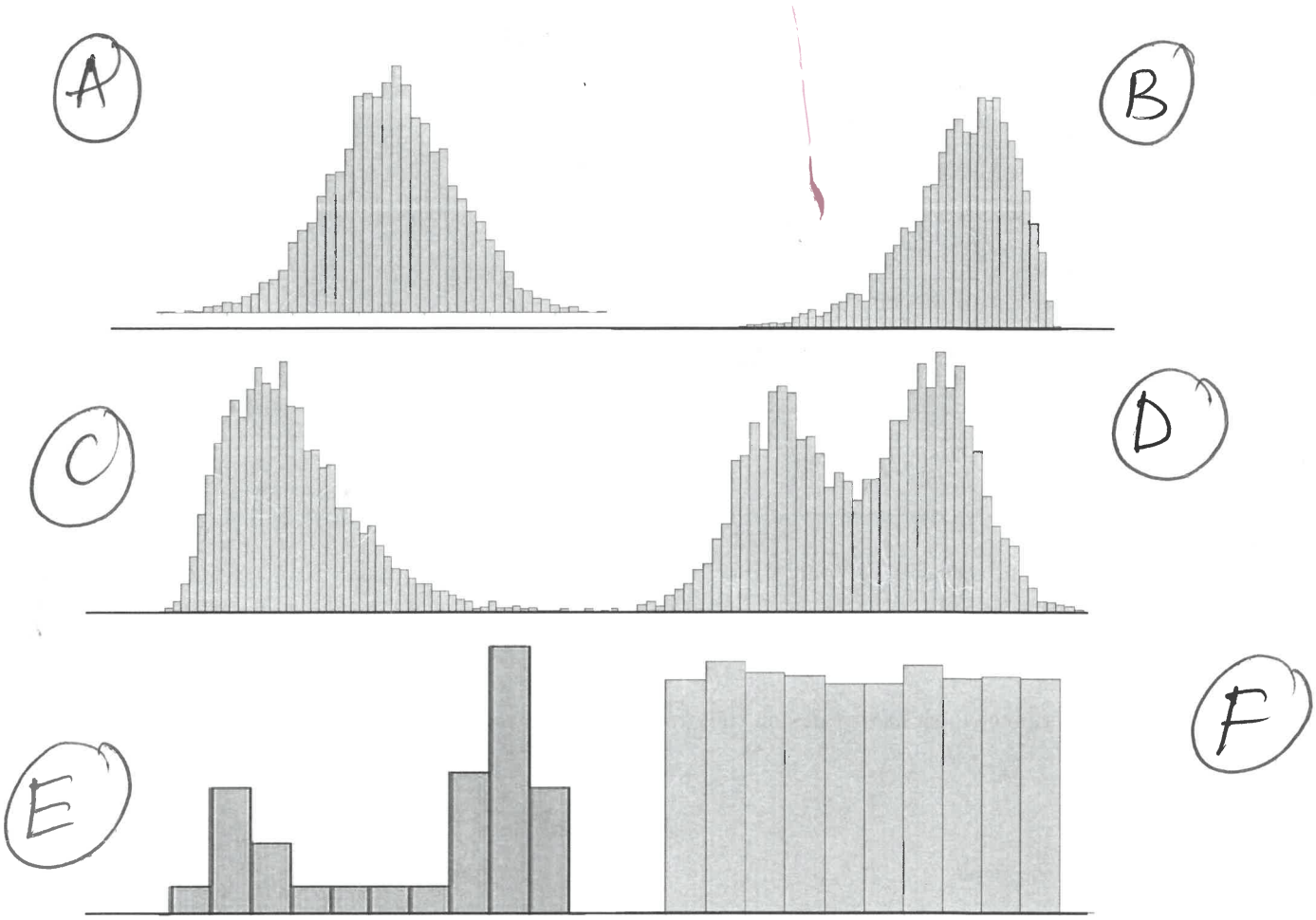
Now consider the table below, which contains values for midterm grades.

85	89	79	78	77	98	86	96	72	80
87	69	98	60	72	95	86	73	79	82

Construct a histogram for the midterm grades data.



Shapes of Histograms



The shape of the histogram is the distribution.

Common Shapes for Histograms

- A distribution is symmetric if the right side and left side are approximately mirror images of one another. (A & F)
- A distribution is skewed right if if the data is piled up on the left and the tail extends relatively far out to the right. (C)
- A distribution is skewed left if if the data is piled up on the right and the tail extends relatively far out to the left. (B & E)
- A distribution is bell-shaped if is the data are symmetric and also shaped like a bell.
- A distribution is bimodal if the data has two humps. (D & E)
- A distribution is uniform if the bars on the histogram are all of the same height. (F)

Example. Determine which histograms on the previous page have each shape.

- symmetric
- skewed to the right
- skewed to the left
- bell-shaped
- bi-modal
- uniform

Recall that when we are working with a sample, we use n to represent the size of the sample, the number of cases in the sample.

Definition. The mean of the data values for a single quantitative variable is given by

$$\text{mean} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

where $x_1, x_2, \ldots, x_n$ are the data values.

- The mean is the average.
- The notation for the mean of a sample is $\bar{x}$, which is read x bar.
- The notation for the mean of a population is the Greek letter μ. (mu, "mew")

Definition. The median of a set of data values for a single quantitative variable, denoted m, is

- the middle entry if an ordered list of the data values contains an odd number of entries OR
- the average of the middle two values if an ordered list contains an even number of entries.

The median splits the data in half.

Example. Return to the data on midterm grades and determine both the mean and the median.

$$\mu = (85 + 87 + \cdots + 79 + 80 + 82)/20 = 82.05$$

[arrange data in increasing order as follows]

60	69	72	72	73	77	78	79	79	80
1	2	3	4	5	6	7	8	9	10
82	85	86	86	87	89	95	96	98	98
11	12	13	14	15	16	17	18	19	20

$$m = \frac{\underset{10}{80} + \underset{11}{82}}{2} = 81$$

Example. Let's instead consider the data on midterm grades with the grades of the students who dropped the course included. Their midterm grades were 25 and 31. What are the mean and median with these grades included?

$$\mu = \frac{25+31+1641}{22} = 77.14$$

25	31	60	69	72	72	73	77	78	79	79	80
1	2	3	4	5	6	7	8	9	10	11	12

82	85	86	86	87	89	95	96	98	98
13	14	15	16	17	18	19	20	21	22

$$m = (\underset{11}{79} + \underset{12}{80})/2 = 79.5$$

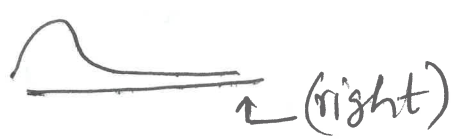
Definition. A statistic is resistant if it is relatively unaffected by extreme values.

The median is resistant, but the mean is not.

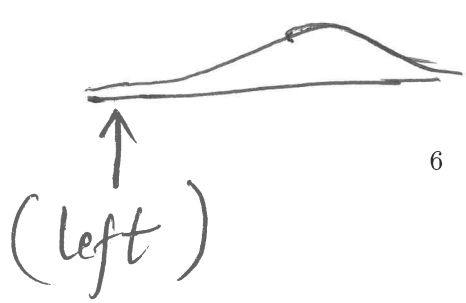
How does the shape of a distribution affect the mean and the median?

- If a histogram is symmetric, then the median is right in the middle and approximately equal to the mean.

- If the data are skewed to the right, the values in the extended right tail pull the mean up but have little effect on the median. Thus the mean is greater than the median.

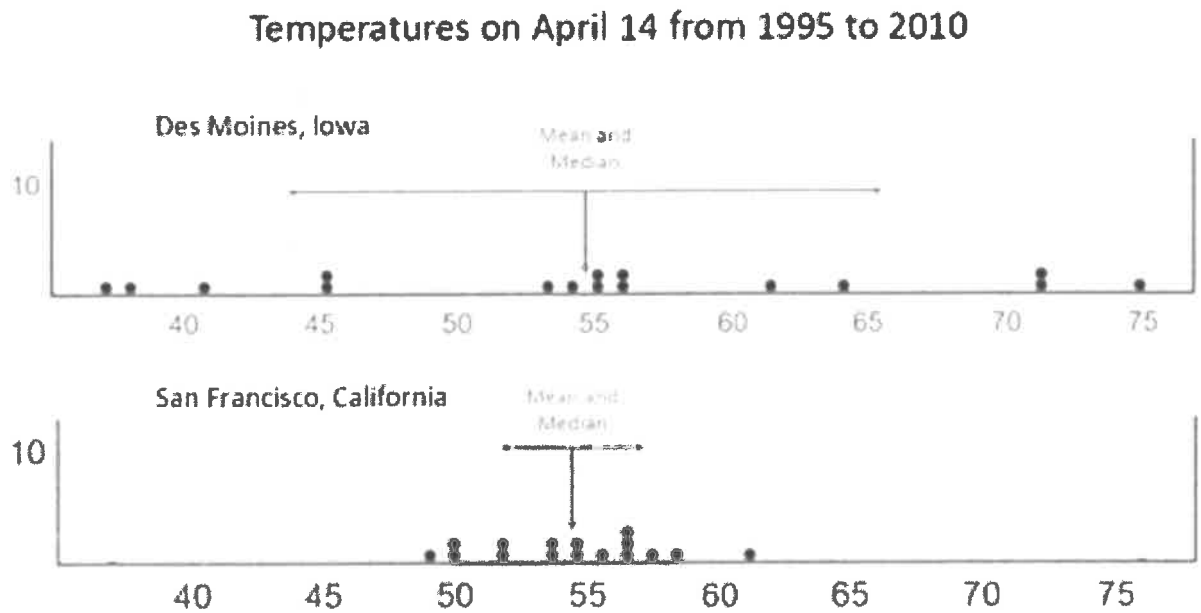


- If the data are skewed to the left, the values in the extended left tail pull the mean down but have little effect on the median. Thus the mean is less than the median.



HIMANSHU SINGH

Knowing just the mean and the median is not enough to describe data. For example, two different bell-shaped distributions can have the same mean and median but very different appearances.



Definition. In a sample, the deviation of an individual data value x from the mean $\bar{x}$ is the difference $x - \bar{x}$.

Example. for data value $x=62$, $x-\bar{x} = 62-55=7$
" " " $x=54$, " $= 54-55=-1$

Example. Consider the dataset $1, 3, 7, 9$. Find the average of the deviations of the values in this dataset.

$\bar{x} = 20/4 = 5$ for $x=1$, $x-\bar{x} = 1-5=-4$
$x=3$, " " $= 3-5=-2$
$x=7$, " " " $=2$
$x=9$, " " " $=4$

average of deviations $= 0/4 = 0$

Definition. Consider a sample with data values $x_1, x_2, \ldots, x_n$ and mean $\bar{x}$. The *standard deviation* for the sample data measures the spread of the data in the sample, is denoted by the letter s, and is given by the formula

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

unit for std. deviation is same as unit of data

- For the sample data on Des Moines temperatures, $s =$ 11.73 °F.
- For the sample data on San Francisco temperatures, $s =$ 3.38 °F.

Definition. Consider measuring a quantitative value on a population. Let $x_1, x_2, \ldots, x_n$ be the values of the variable and μ be the mean. The *standard deviation* for the variable measures the spread of the data, is denoted by the letter σ, and is given by the formula

$$\sigma = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_n-\mu)^2}{n-1}}$$

Remark. The standard deviation gives a rough estimate of the variability of a data value from the mean. The larger the standard deviation, the more variability there is in the data and the more spread out the data are.

$$\bar{X} = \frac{2+4+7+9+12+14}{6} = \frac{48}{6} = 8$$

Example. Consider the sample data 2, 4, 7, 9, 12, 14. Compute the mean and the standard deviation.

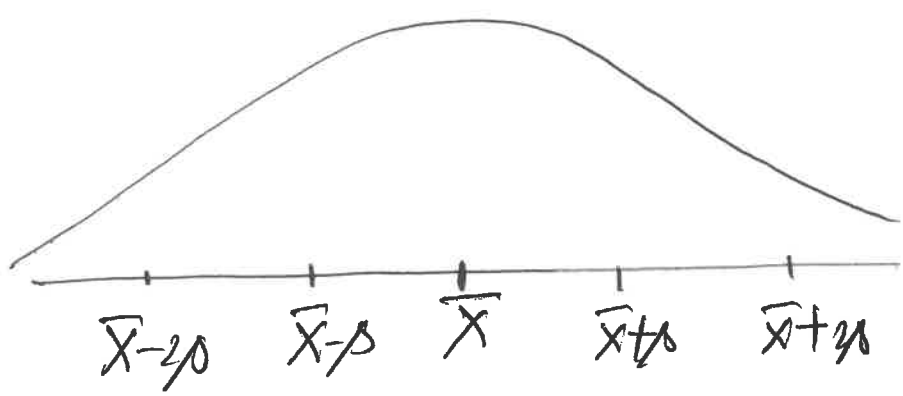
$$s = \sqrt{\frac{(2-8)^2+(4-8)^2+\cdots+(12-8)^2+\cdots}{5}}$$

$$= \sqrt{\frac{36+16+1+1+16+36}{5}} = \sqrt{\frac{106}{5}} \approx 4.60$$

Definition. The nine 95 % says that if a distribution of data is approximately symmetric and bell-shaped, about 95% of the data should fall within two standard deviations of the mean.

This means that about 95% of the data in a sample from a symmetric, bell-shaped distribution should fall in the interval from $\bar{x} - 2s$ to $\bar{x} + 2s$.

Diagram.



Definition. Consider a sample with mean $\bar{x}$ and standard deviation s. Let x be a data value from the sample. The z score for x tells how many standard deviations the value x is from the mean $\bar{x}$ and is given by the formula

$$z - \text{score} = \frac{x - \bar{x}}{s}.$$

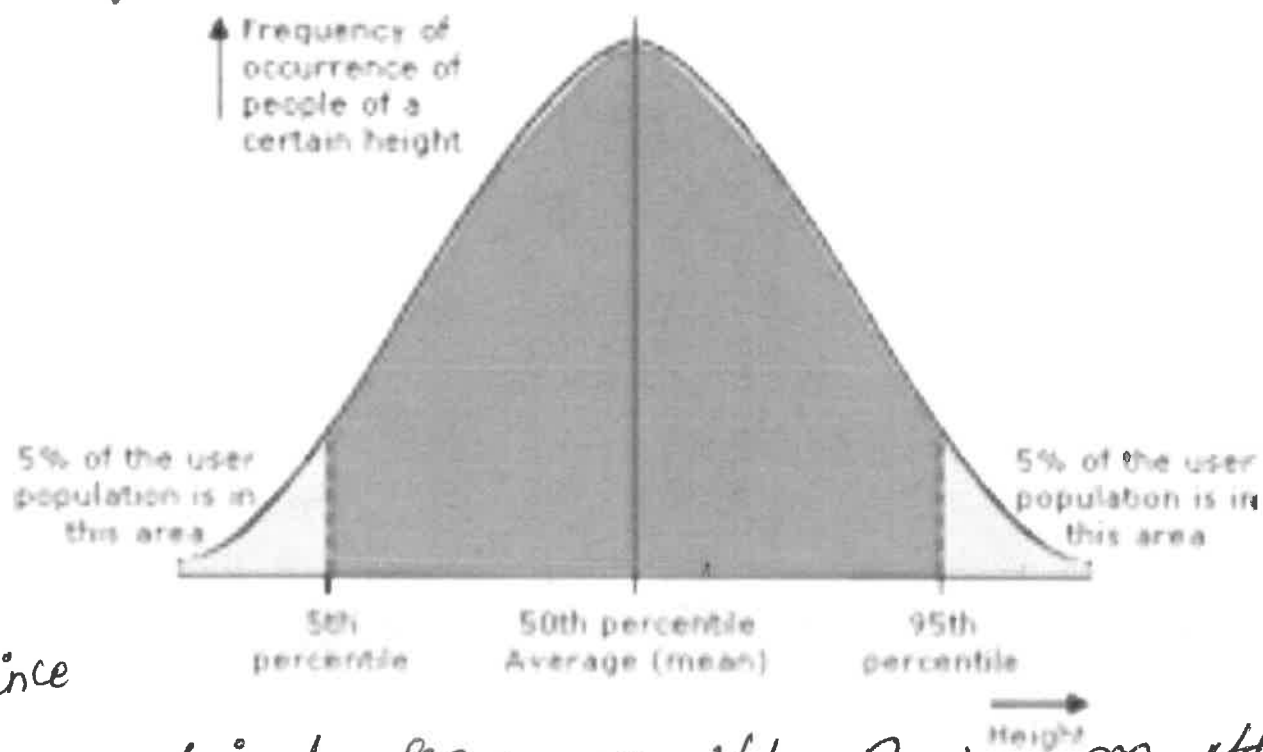
Example. Consider Sample A with mean 100 and standard deviation 10. One of the values in Sample A is 85. Another sample, Sample B, has mean 65 and standard deviation 2. One of the values in Sample B is 60. Which of these sample values is more unusual?

sample-A for $x = 85$ z score $\equiv \frac{85-100}{10} = -1.5$

↳ data point is 1.5 std deviation below the mean.

sample-B for $x = 60$ z-score $\equiv \frac{60-65}{2} = -5/2 = -2.5$

↳ data point is 2.5 std. deviation below the mean.

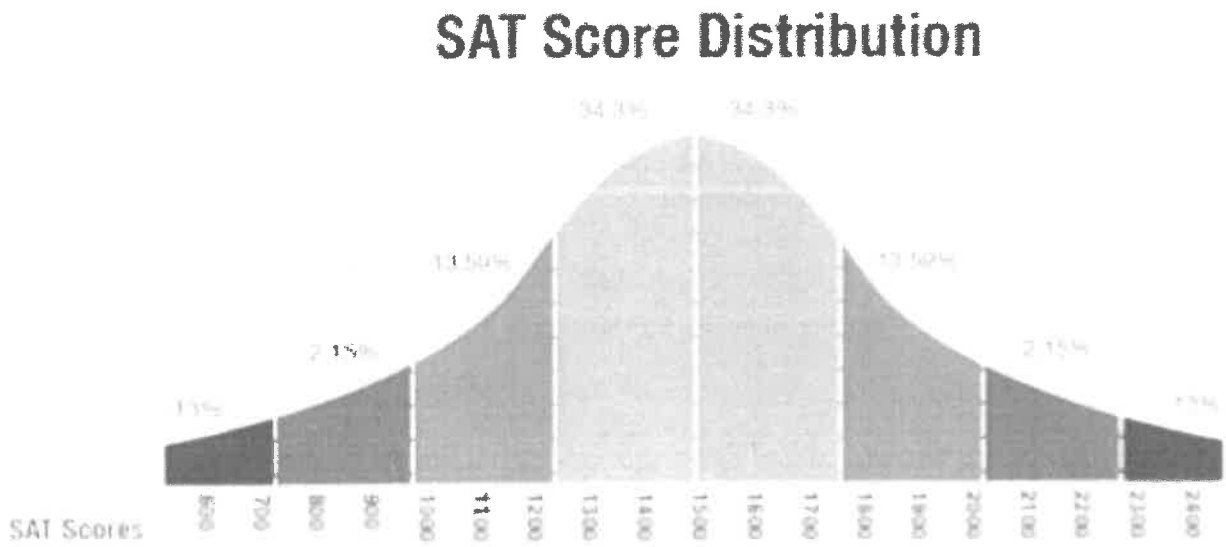


since

∴ the point from sample B is more std. deviation away from the mean, it is more unusual.

Another way to study the spread of a dataset is using percentiles.

Definition. The Pth *percentile* is the value of a quantitative variable which is greater than P percent of the data.



Percentiles of Note

- We can use the median to find the 50% percentile.
- The 25th percentile is called the FIRST QUARTILE and is denoted Q_1.
- The 75th percentile is called the THIRD QUARTILE and is denoted Q_3.

Definition. The five number summary for a dataset is given by the ordered 5-tuple

(minimum, Q_1, median, Q_3, maximum)

The five number summary divides the dataset into fourths. About 25% of the data fall between any two consecutive numbers in the five number summary.

range $(22-4=18)$ IQR $=13-6=7$ odd # of points

Example 0.1. Find the five number summary for the dataset 4, 4, 5, 5, 7, 8, 8, 11, 11, 12, 12, 14, 15, 18, 22.

$(4, 6, 11, 13, 22)$

$Q_1 \rightarrow$ median of values to left of $m=11$ (including $m=11$)

$4, 4, 5, 5, 7, 8, 8, 11 \rightarrow (5+7)/2 = 6$, $Q_1 = 6$

$Q_3 \rightarrow$ median of values to RIGHT of $m=11$ (including $m=11$) $11, 11, 12, 12, 14, 15, 18, 22$

$\rightarrow \frac{12+14}{2} = 13$ and $Q_3 = 13$.

Example 0.2. Find the five number summary for the dataset 1, 3, 8, 9, 12, 15, 18, 22.

$(1, 5.5, 10.5, 16.5, 22)$

$Q_1 \equiv$ ~~left~~ (left of $m=10.5$ & including that)

$(1, 3, 8, 9)$ $\quad \frac{3+8}{2} \equiv 5.5$

$Q_3 \equiv (12, 15, 18, 22)$ $\quad \frac{15+18}{2} = \frac{33}{2} \equiv 16.5$

range $\equiv 22-1 = 21$

IQR $\equiv Q_3 - Q_1 \equiv 11$

$= 16.5 - 5.5$

Definition. The RANGE of a dataset is given by the formula

$$\text{range} = \text{max.} - \text{min.}$$

Definition. The interquartile range, abbreviated IQR, of a dataset is given by the formula

$$\text{IQR} = Q_3 - Q_1.$$

Example. Find the range and IQR for the datasets in Examples 0.1 and 0.2.

- The mean and standard deviation are NOT resistant to outliers. median is resistant
- Use the mean and standard deviation when you want to use ALL of the data. It is best when there are no outliers.
- Use a five number summary when there are outliers and you want information about the main body of the data.

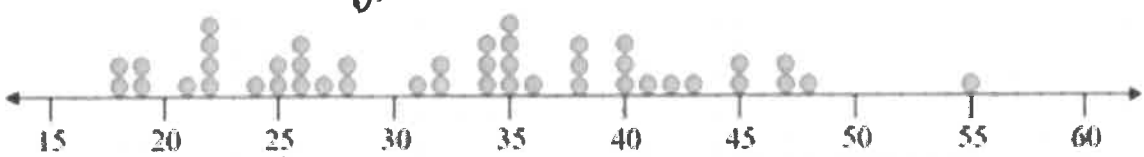
HIMANSHU SINGH

Math 1342 **Boxplots and Quantitaive/Categorical Relationships** **2.4**

Definition. A data value is an OUTLIER if is it either smaller than $Q_1 - 1.5(IQR)$ OR larger than $Q_3 + 1.5(IQR)$

Example. In the data set below, is 55 an outlier? Here $n = 44$, not 40.

(5 number summary) $\equiv$ (18, 25, 34, 40, 55)



Amount Spent on Fast Food Restaurants during 40 Weeks

$m \equiv \frac{34+34}{2} = 34$

$IQR \equiv Q_3 - Q_1 = 40 - 25 = 15$

$[1.5(15) \equiv 22.5]$

Is 55 larger (smaller) than $Q_3 \pm 1.5(IQR)$?

check $Q_3 + 1.5(IQR) = 40 + 22.5 = 62.5$

$\therefore$ 55 is NOT larger than 62.5, 55 is NOT an outlier.

Definition. A *boxplot* includes

- a numerical scale appropriate for the data values,
- a box stretching from Q_1 to Q_3,
- a line that divides the box into two parts and is drawn at the median,
- a line from each quartile to the most extreme data value that is NOT an outlier,
- and each outlier plotted individually with a symbol such as an asterisk or a dot.

$IQR \equiv Q_3 - Q_1 \equiv 56 - 42 = 14$

$1.5\,IQR \equiv 1.5 \times 14 = 21$

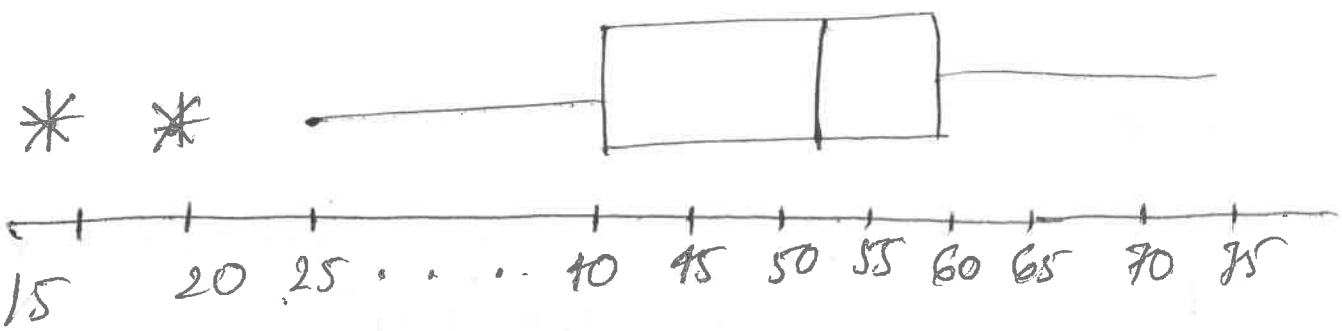
Example. Draw a box plot for the data with five number summary (15, 42, 52, 56, 71) and tails 15, 20, 28, 30, 31, ..., 64, 65, 65, 66, 71.

outlier are smaller than

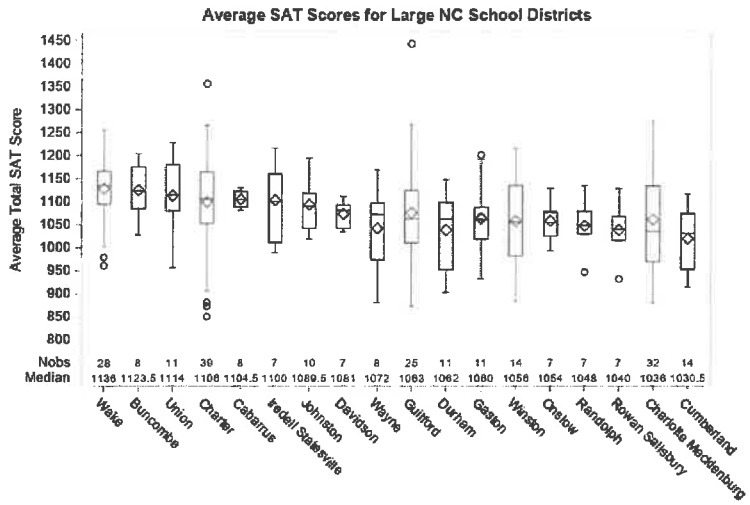
$Q_1 - 1.5\,IQR = 42 - 21 = 21$

$\boxed{\text{OR}}$

greater than $Q_3 + 1.5\,IQR = 56 + 21 = 77$



Side-by-side graphs are used to visualize the relationship between a quantitative variable and a categorical variable. A side-by-side graph includes a graph for the quantitative variable for each group in the categorical variable, all using a common numeric axis. These graphs may be all dotplots, all histograms, or all boxplots.



HW

§ 2.3 / 79, 81, 87, 89, 91, 93
99, 101, 103, 105,
107, 109, 119.

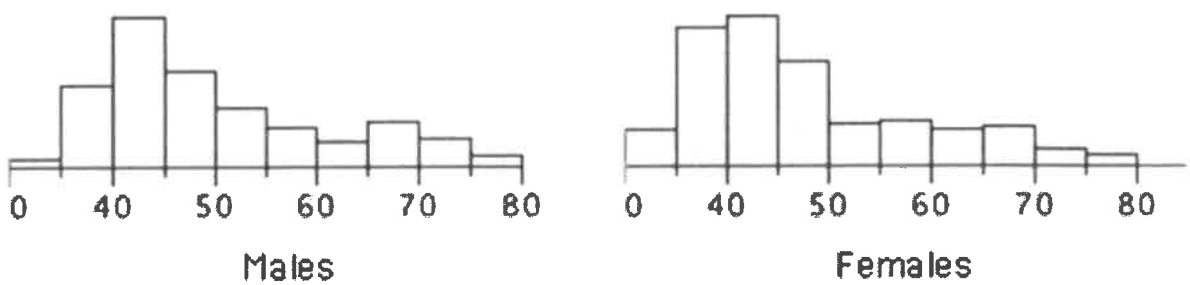
§ 2.4 / 131, 133, 135, 137,
141, 143, 157.

HIMANSHU SINGH

In Section 2.1, we looked at relationships between two categorical variables. In Section 2.4, we looked at relationships between a categorical variable and a quantitative variable. In this section, we will look at relationships between two quantitative variables.

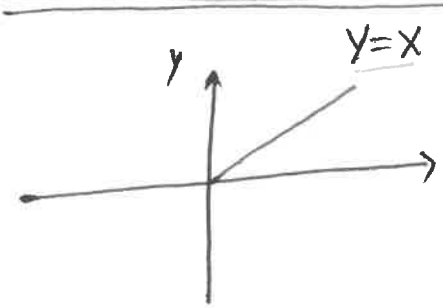
Example. A study was done on the relationship between the ages of husbands and wives. The study considered 282 married couples.

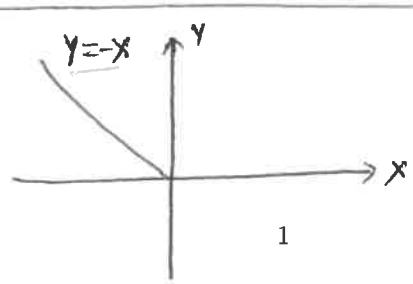
Figure 1. Histograms of spousal ages.



Definition. A SCATTER PLOT is a graph showing the relationship between two quantitative variables.

- A scatterplot includes a pair of axes with the appropriate numerical scales, one for each variable.
- The paired data for each case are plotted as a point on the scatterplot.
- If there are explanatory and response variables, we put the explanatory variable on the horizontal axis and the response variable on the vertical axis.

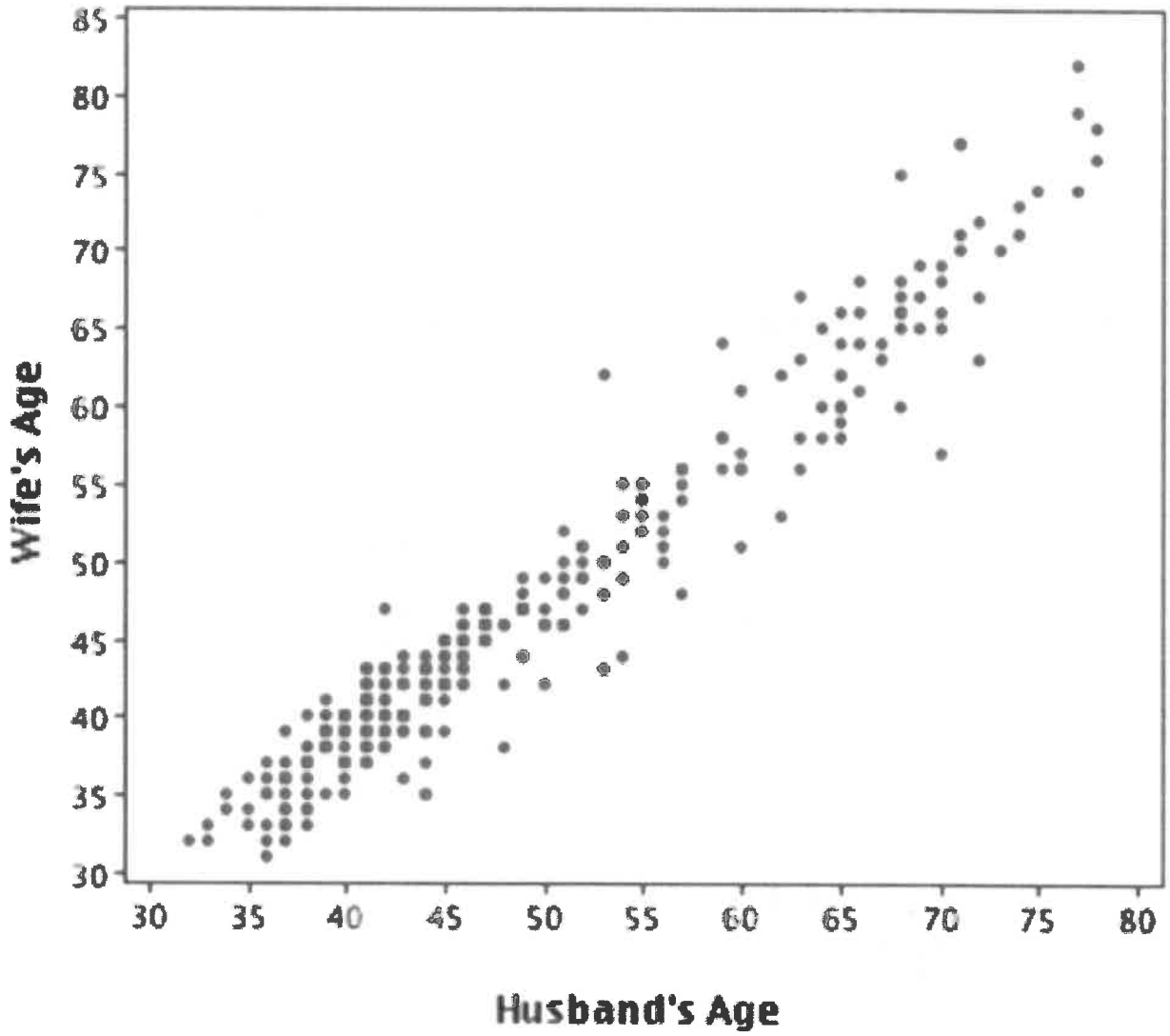




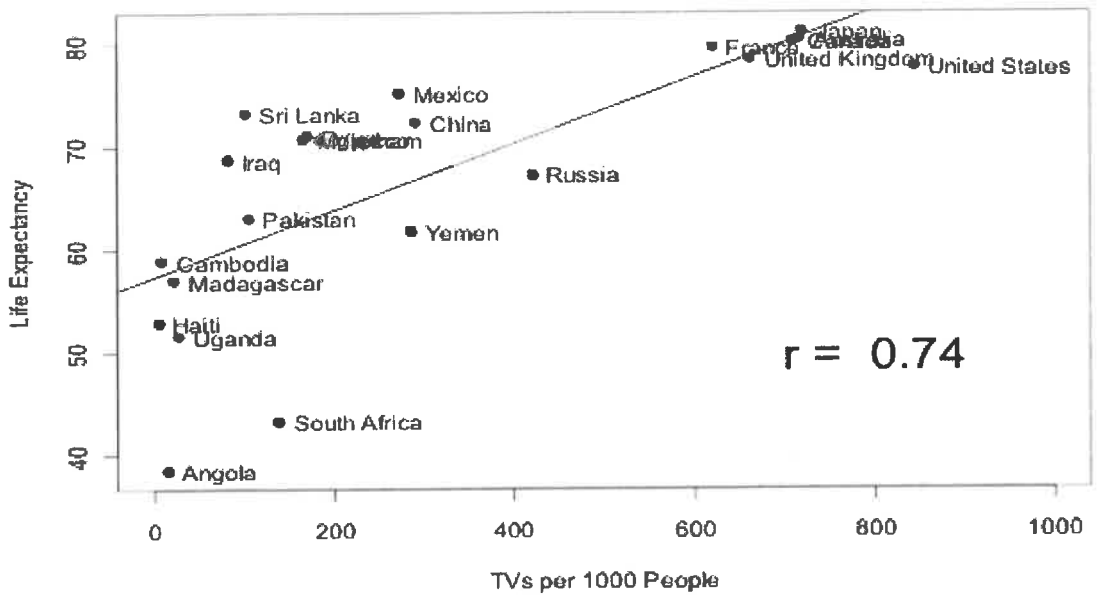
$y = ax \rightarrow$ linear function

$y = ax + b \rightarrow$ affine function

Figure 2. Scatter plot showing wife age as a function of husband age.



What (positive or negative) association you see above?



What (positive or negative) association you see above?

When looking at a scatterplot, we often address the following questions:

- Do the points form a clear trend with a particular direction, are they more scattered about a general trend, or is there no obvious pattern?
- If there is a trend, is it generally more upward or generally downward as we look from left to right? A general upward trend is a positive association, while a general downward trend is a negative association.
- If there is a trend, does it seem to follow a straight line, which we call a linear association, or some other curve or pattern?
- Are there any outlier points that are clearly distinct from a general pattern in the data?

Definition. CORRELATION is a measure of the strength and direction of linear association between two quantitative variables.

$$r = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right)$$

Remark. The correlation between two quantitative variables of a **sample** is denoted r. The correlation between two quantitative variables of a **population** is denoted ρ, which is the Greek lowercase letter rho.

The sample correlation r has the following properties.

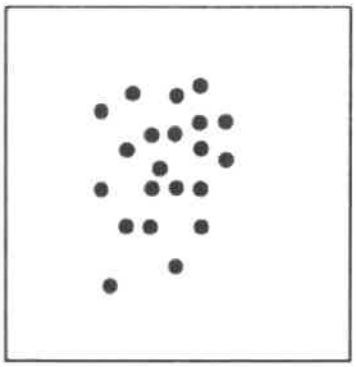
- Correlation is always between -1 and 1, inclusive, so $-1 \leq r \leq 1$.
- The sign of r, positive or negative, indicates the direction of association.
- Values of r close to -1 or 1 show a strong linear relationship, while values of r close to 0 show no linear relationship.
- The correlation r has no units and is independent of the scale of either variable.
- The correlation is symmetric. The correlation between variables x and y is the same as the correlation between y and x.

The population correlation ρ also satisfies these properties.

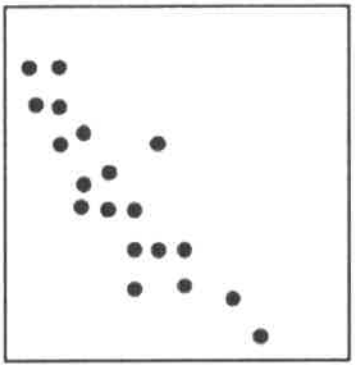
strong positive correlation

$r \approx 0.9$
largest r value of these examples.

moderate positive correlation

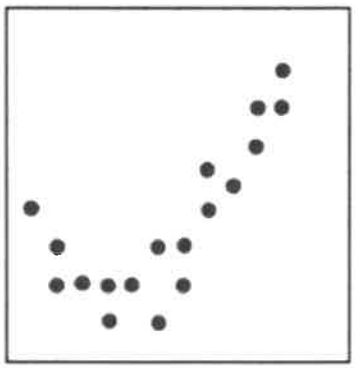


$r \approx 0$ no correlation



moderate (-ve) negative correlation

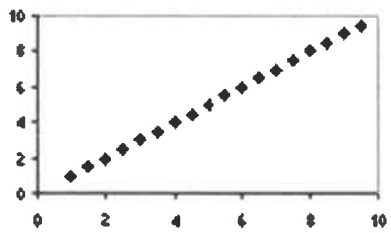
$r \approx -0.9$
strong negative correlation



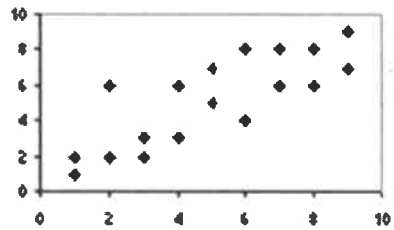
NOT a linear association

Words of Caution

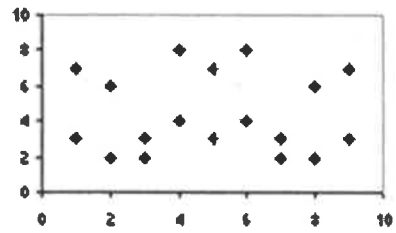
- A strong positive or negative correlations does NOT necessarily imply a cause and effect relationship between the two variables. Correlation does NOT imply causation.
- A correlation near zero does NOT necessarily mean that the two variables are not associated, since the correlation measures only the strength of a **linear** relationship.
- Correlation can be heavily influenced by outliers. Always plot your data!



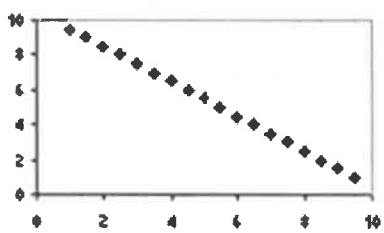
$r = 1$



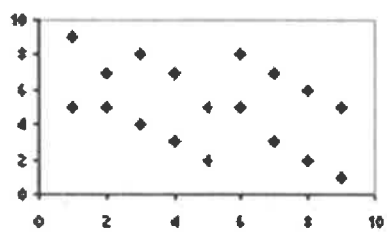
$r = 0.8$



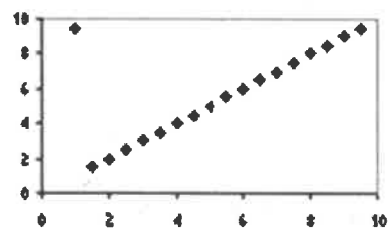
$r = 0$



$r = -1$



$r = -.43$



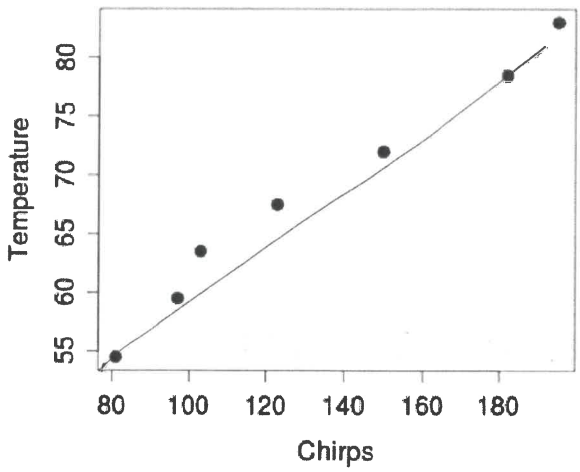
$r = .71$

HIMANSHU SINGH

Math 1342 **Two Quantitative Variables: Linear Regression** **2.6**

In Section 2.5, we investigated the relationship between two quantitative variables. In this section, we discuss how to use one of the variables to predict the other **when there is a linear trend**.

Example. Can you estimate the temperature on a summer evening just by listening to crickets chirp? In the scatterplot below, temperature is measured in degrees Fahrenheit, and chirps are measured in number per minute.



Recall

EV → explanatory Variable

RV → Response Variable

Find a straight line that best fits the data in the scatterplot.

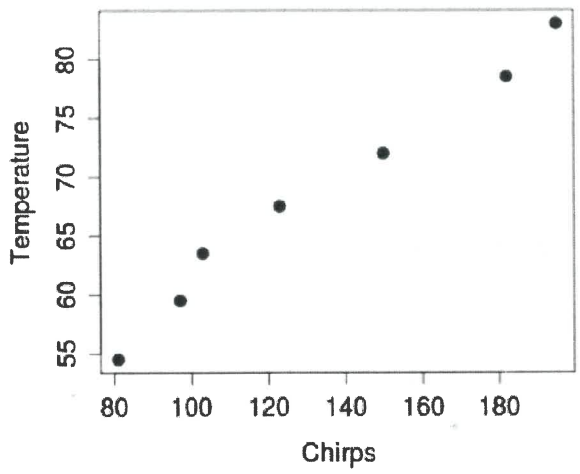
Definition. The process of fitting a line to a set of data is called linear Regression, and the line of best fit is called the regression line

Remark. If x is the explanatory variable, and y is the response variable, then the regression line has equation $\hat{y} = mx + b$, where

- m is the slope,
- b is the y-intercept and
- $\hat{y}$ is the predicted value of the response variable y.

The regression line provides a model of a linear association between two variables, and we can use the regression line on a scatterplot to give a predicted value of the response variable, based on a given value of the explanatory variable.

The regression line for the scatterplot below is $\boxed{\widehat{\text{Temperature}} = 37.7 + 0.23(\text{Chirps}).}$



1. You listen and hear crickets chirping about 140 times per minute. Calculate the predicted value of the outside temperature.

$\widehat{\text{Temp.}} = 37.7 + 0.23(140) = 69.9°F$

2. What is the predicted value of the temperature when the crickets chirp 103 times per minute? (DIY)

Answer → 61.4°F

3. One of the cases in the cricket dataset is 103 chirps per minute when it is 63.5 degrees Fahrenheit. How far is the **predicted** temperature from the **observed temperature**?

difference

Observed − predicted = 63.5 − 61.4 = 2.1°F, so predicted temp. is 2.1°F less than the observed temp.

The experimental observed response value, denoted y, is the response value observed for a particular data point.

The predicted response value, denoted $\hat{y}$, is the response value that would be predicted for a given x value, based on the equation of the regression line.

Remark. The best fitting line is the line that makes the predicted value close to the actual observed values.

Definition. The residual at a data value is the difference between the observed value and predicted value of the response variable.

- The equation is residual $= y - \hat{y}$.
- On a scatterplot, the residual represents the vertical deviation from the line to a data point.
- Points above the line will have positive residuals.
- Points below the line will have negative residuals.
- If the predicted values closely match the observed data values, the residuals will be very close to 0.

Definition. The least square line, also called the line of best fit, is the line that minimizes the sum of the squared residuals.

We will use the terms "regression line" and "least squares line" interchangeably.

Remark. With correlation, it does NOT matter which is the explanatory variable and which is the response variable. With regression, it DOES matter.

When working with the regression line $\hat{y} = mx + b$, the slope m is the change in the predicted value $\hat{y}$ that results from a one unit increase in the explanatory variable x. The intercept b is the predicted value of $\hat{y}$ when the value of the explanatory variable x is 0. In some situations, the case where x is 0 makes no sense.

Example. Recall the regression line $\widehat{\text{Temperature}} = 37.7 + 0.23(\text{Chirps})$. $\hat{y} = mx + b$

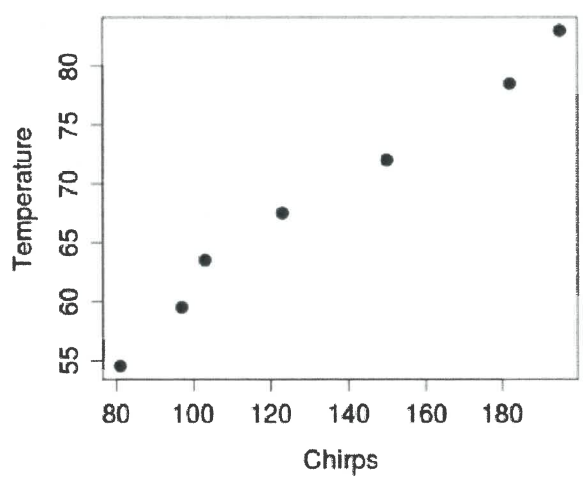
1. What is the slope of the line? Interpret it in context.

$m = 0.23$. For every one unit increase in number of chirps per minute, the temp. increases by $0.23^\circ F$

2. What is the intercept of the line? Interpret it in context.

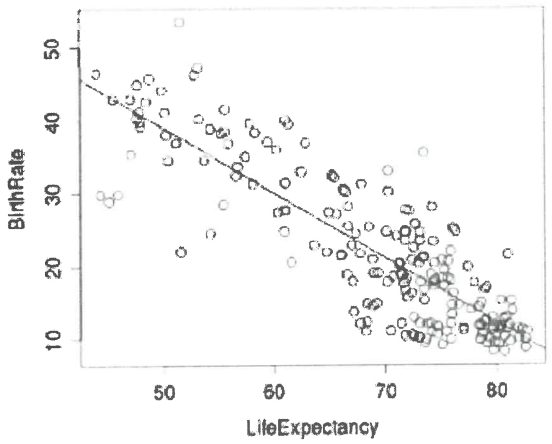
$b = 37.7$. This is the predicted value of the temp. when the value of chirps per minute is 0.

Remark. Do NOT use the regression line to predict far outside the range of values of x available in your data.



Remark. When working with regression, always PLOT YOUR DATA. A regression line should only be used for two quantitative variables with a linear association.

Remark. Outliers can have a strong influence on the regression line. This is especially the case for data points where the value of the explanatory variable is an outlier.



Which of the following is correct?

1. A decrease of 0.89 in the birth rate explains a 1 year increase in predicted life expectancy.
2. Increasing life expectancy by 1 year will cause the birth rate to decrease by 0.89.
3. Both
4. Neither

- Higher values of x may lead to higher (or lower) predicted values of y, but this does NOT mean that changing x will CAUSE y to increase or decrease.
- Causation can only be determined if the values of the explanatory variable x were determine randomly.
- Remember: correlation does NOT imply causation.

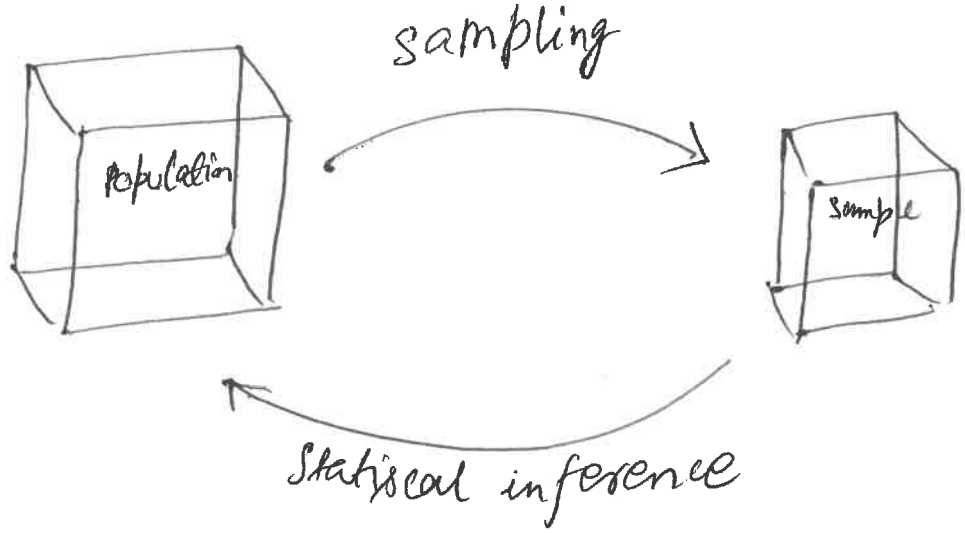
HIMANSHU SINGH

Math 1342 **Sampling Distributions** **3.1**

In Chapter 1, we discussed data collection, methods of obtaining sample data from a population of interest. In this chapter we begin using the information in the sample to understand what might be true about the entire population. If all we see are the data in the sample, what conclusions can we draw about the population? How sure are we about the accuracy of these conclusions? We will answer questions like these.

Definition. statistical inference is the process of drawing conclusions about the entire population based on the information in a sample.

Diagram



Definition. A *parameter* is a number that describes some aspect of a population.

Definition. A *statistic* is a number that is computed from data in a sample.

Remark. We usually have a sample statistic and want to use it to make inferences about the population parameter.

Quantity	Parameter	Statistic
Mean	μ	$\bar{X}$
Standard Deviation	σ	s
Proportion	p	$\hat{p}$
Correlation	ρ	r

$0 \le p \le 1$; $0 \le \hat{p} \le 1$

$-1 \le \rho \le 1$; $-1 \le r \le 1$.

Example. For each of the following, state whether the quantity described is a parameter or a statistic and give the correct notation.

1. Average household income for all households in the U.S., using data from the U.S. census μ

   parameter

2. Correlation between height and weight for players on the 2010 Brazil World Cup team, using data from all 23 players on the roster ρ parameter.

3. Proportion of people who use an electric scooter, using data from a sample of 400 adults $\hat{p}$ statistic

Example. Before the 2016 presidential election, 1143 registered voters were asked who they planned to vote for in the 2016 presidential election.

1. What proportion of voters planned to vote for Trump? $\hat{p} = 0.46$

2. What notation should be use to denote this proportion? $\hat{p}$

Remark. If we only have one sample and we don't know the value of the population parameter, the sample statistic is our best estimate of the true value of the population parameter.

Definition. A point estimate is a value given as an estimate of a parameter for a population.

Agency	Dates	Sample Size	Clinton %	Trump %
NBC News	8/15-8/21	17459	50	42
Reuters	8/13-8/17	1049	41	36
Economist	8/14-8/16	911	47	41
Gravis	8/15-8/15	1143	54	46

- We usually think of a parameter as a fixed value, while the statistic varies from sample to sample, depending on which cases are selected to be in the sample.
- A fundamental question of statistical inference: For a given sample statistic, what are plausible values for the population parameter?
- The key to addressing this question is to understand how the value of the sample statistic varies from sample to sample.

Example. We want to study the number of people in this section of MATH 1342 who currently share their living space with a dog. We will do this by taking samples.

sample $\hat{p}_1 = 3/10$

sample $\hat{p}_2 = 1/10$

sample $\hat{p}_3 = 3/10$

population $p = 9/27$
$= 1/3$

Definition. A sampling distribution is the distribution of sample statistics computed for different samples of the same size from the same population.

A sampling distribution shows us how the sample statistic varies from sample to sample.

Remark. A sampling distribution is a distribution, so we want to describe its center and its shape.

- Center: If samples are randomly selected, the sampling distribution will be centered around the population parameter.
- Shape: For most of the statistics we consider, if the sample size is large enough, the sampling distribution will be symmetric & bell-shaped.

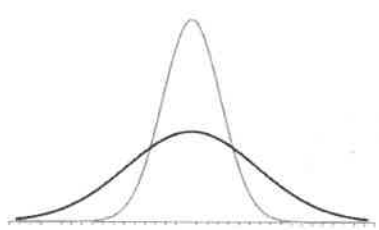
Remark. Be careful!

- If you do not take a random sample, the result of the sampling distribution will not likely be centered around the population parameter
- If you want good results, your sample must be unbiased.

Knowing how much a statistic varies from sample to sample is key in helping us know how accurate a sample is.

Definition. The standard error of a statistic, denoted SE, is the standard deviation of the sample statistic.

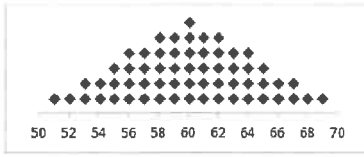
- We think of the standard error as a typical distance between the sample statistics and the population parameter.
- As the sample size increases, the variability of the sample statistics tends to decrease and the sample statistics tend to be closer to the true value of the population parameter.
- For larger sample sizes, you get less variability in the statistics, so less uncertainty in your estimates.
- If the sample were the entire population, then the sample statistic would match the population parameter exactly.



Definition. The 95% rule states that if a distribution of data is approximately symmetric and bell-shaped, about 95% of the data should fall within two standard deviations of the mean.

$\bar{x} \pm 2s$

Example. The figure below shows sample means from samples of size $n = 100$ for a population. Estimate the value of the population mean and estimate the standard error for the sample mean.



95% rule application

$\mu \Rightarrow 60$ in

$\bar{x} \pm 2s = 67, 53.$

$\Rightarrow \boxed{\bar{x} - 2s = 53}$	$\boxed{\bar{x} + 2s = 67}$
$\Rightarrow 60 - 2s = 53$	$60 + 2s = 67$
$\Rightarrow -2s = -7$	$2s = 7$
$\boxed{s = 7/2}$	$\boxed{s = 7/2}$

as this is a sampling distribution,
SE, $s = 7/2$

HIMANSHU SINGH

Example. A report from a Gallup poll in 2011 based on a random sample of $n = 147,291$ adults says, "Forty-five percent of American adults reported getting their health insurance from an employer." How accurate is 45%?

We use the value of a statistic computed from a sample to give an estimate for a parameter of a population. However, since statistics vary from sample to sample, a single estimate is often not sufficient. We need some measure of accuracy associated with our estimate. Thankfully, we can use knowledge about how sample statistics vary to find a **margin of error** for the estimate. This allows us to construct an **interval** that gives a range of plausible values for the population parameter.

Definition. An interval estimate gives a range of plausible values for a population parameter.

Definition. The margin of error is a number that reflects the precision of the sample statistic as an estimate for this parameter.

Remark. One common form for an interval estimate is statistic $\pm$ margin of error

In interval notation, this is (statistic − margin of error, statistic + margin of error).

Example. A report from a Gallup poll in 2011 based on a random sample of $n = 147,291$ adults says, "Forty-five percent of American adults reported getting their health insurance from an employer. The maximum margin of sampling error is ± 1 percentage point." Give an interval estimate for the **proportion** of American adults getting their health insurance from an employer.

$\hat{p} = 0.45$ (statistic)

margin of error $= 1\% = 0.01$

$\rightarrow (\hat{p} - 0.01, \hat{p} + 0.01)$

$= (0.45 - 0.01, 0.45 + 0.01)$

$= (0.44, 0.46)$

DIY (HW)

Example. A 2011 random sample of $n = 755$ U.S. cell phone users ages 18 and older found that the average number of text messages sent or received per day is 41.5 messages, with a margin of error of 12.2. Give an interval estimate for the average number of text messages sent or received per day for U.S. cell phone users ages 18 and older.

Ans: $(29.3, 53.7)$

Poll	Date	Sample	MoE	Clinton (D)	Trump (R)
Final Results	--	--	--	48.2	46.1
RCP Average	**11/1 - 11/7**	--	--	**46.8**	**43.6**
Bloomberg	11/4 - 11/6	799 LV	3.5	46	43
IBD/TIPP Tracking	11/4 - 11/7	1107 LV	3.1	43	42
Economist/YouGov	11/4 - 11/7	3669 LV	--	49	45
LA Times/USC Tracking	11/1 - 11/7	2935 LV	4.5	44	47
ABC/Wash Post Tracking	11/3 - 11/6	2220 LV	2.5	49	46
FOX News	11/3 - 11/6	1295 LV	2.5	48	44
Monmouth	11/3 - 11/6	748 LV	3.6	50	44
NBC News/Wall St. Jrnl	11/3 - 11/5	1282 LV	2.7	48	43
CBS News	11/2 - 11/6	1426 LV	3.0	47	43
Reuters/Ipsos	11/2 - 11/6	2196 LV	2.3	44	39

1. Using the Bloomberg poll, calculate an interval estimate for the proportion of registered voters who planned to vote for each candidate.

for clinton $\hat{p}_C = 46\%$ $(46 - 3.5\%, 46 + 3.5\%)$
$= (42.5, 49.5)$

for Trump $\hat{p}_T = 43\%$ $(43 - 3.5, 43 + 3.5)$
$= (39.5, 46.5)$

2. For each interval, determine whether it contains the true population parameter?

(final results)

yes for both.

- The parameter is fixed.
- The statistic varies depending on the sample.
- The interval " " " " ".

Definition. A CONFIDENCE INTERVAL for a parameter is an interval computed from sample data by a method that will capture the parameter for a specified proportion for all samples. The proportion of all samples whose intervals contain the parameter is the CONFIDENCE LEVEL.

A p% confidence interval will contain the true value of the parameter for p% of all samples.

If you had the sampling distribution, how would you find the margin of error to ensure that intervals of the form

$$\text{statistic} \pm \text{margin of error}$$

capture the parameter for 95% of all samples?

If the sampling distribution is relatively symmetric and bell-shaped, a 95% confidence interval is given by

statistic ± 2 (S.E) standard error

D.I.Y.

In interval notation, this is _____.

In this case,

$$\text{margin of error} = 2 \cdot SE.$$

$\hat{p} = 0.35$

Example. A survey of 2,290 Americans in February 2017 found that 35% consider stimulating jobs the top priority for the president and Congress. The standard error for this statistic is 0.01.

$2SE = 2 \times 0.01 = 0.02$

1. What is the 95% confidence interval for the true proportion of all Americans who considered stimulating jobs the top priority for the president, and Congress at that time?

$(\hat{p} - 2SE, \hat{p} + 2SE) \equiv (0.33, .37)$

2. Interpret the result in context.

We are 95% confident that p, the true proportion of all AMERICANS who considered stimulating jobs to the top priority is in the above interval.

Example. A gas station wanted to calculate how much gasoline they sold on average. Based on a sample of size $n = 9$ randomly selected days, the sample mean was 891 gallons of gasoline with a standard error of 30 gallons.

1. What is the parameter the gas station is trying to estimate? μ (true average amount of gasoline sold in a day in gallons)
2. Find a 95% confidence interval for the parameter.

   $(891 \pm 30) \equiv$ D.I.Y

3. Interpret the 95% confidence interval in context. D.I.Y

Remark. Margin of error, standard error, and standard deviation of a sample are ALL different.

- The margin of error is amount added & subtracted in the confidence interval.
- The standard error is std deviation of sample statistics if we could take many samples of the same size.
- The standard deviation of a sample is std deviation of individual values in the sample.

Some Additional Notes

- The larger the sample size, the smaller the variability in the sample statistics, so the smaller the standard error.
- The confidence interval is built to try and capture a parameter of the population, NOT individual cases.
- A confidence interval gives us information about the population parameter, NOT about means of other samples. Remember that the interval is making a statement about where the population parameter is likely to be found.

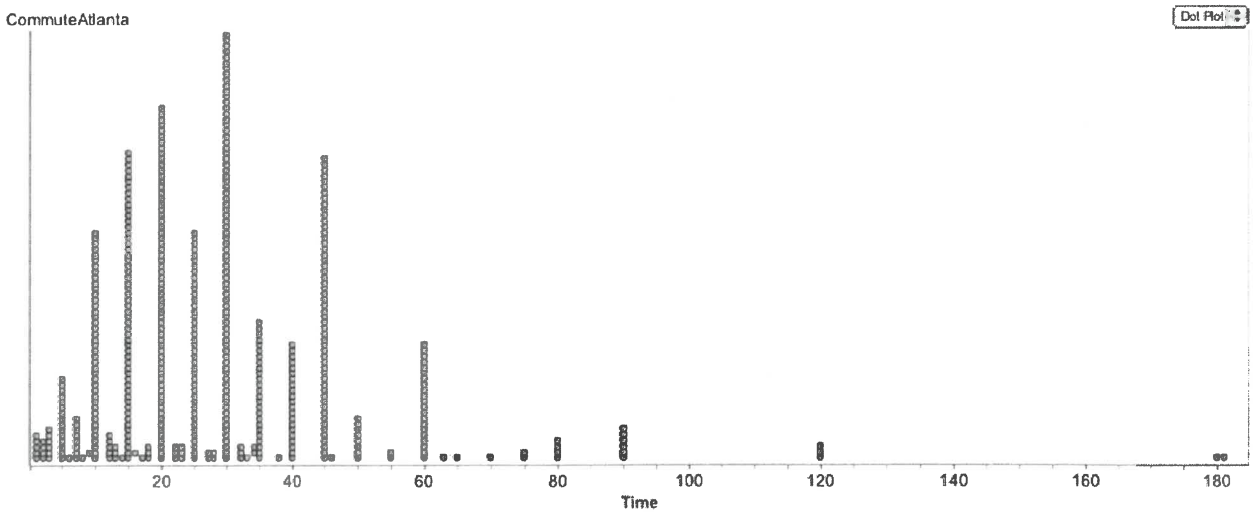
HIMANSHU SINGH

The distributions of sample statistics considered so far in Chapter 3 require us to either already know the value of the population parameter or to have the resources to take thousands of different samples. In most situations, neither of these is an option.

In this section, we introduce a method for estimating the variability of a statistic that uses only the data in the original sample. This technique is called bootstrapping and allows us to approximate a sampling distribution and estimate a standard error using just the information in that one sample.

Example. Suppose we want to know the average commute time for people who live in the Atlanta metropolitan area and commute to work. It is NOT feasible to contact every member of the population and ask about their commute time.

The American Housing Survey contacted a random sample of 500 Atlantans who commute to work. The data collected is recorded in the dotplot below.



- The mean is 29.11 min.
- The standard deviation is 20.72 min.
- Note that this standard deviation is NOT the standard error because standard error does NOT come directly from the original data. (the image above is not sampling distribution)

How can we make the sample data look like data from the entire population? The key idea is to assume for the moment that the population of all commuters in Atlanta is basically just many, many copies of the commuters in our original sample of 500. If we make lots of copies of the sample and then sample repeatedly from this hypothetical population (comprised of copies), we are coming as close as we can to mimicking the process of sampling repeatedly from the population.

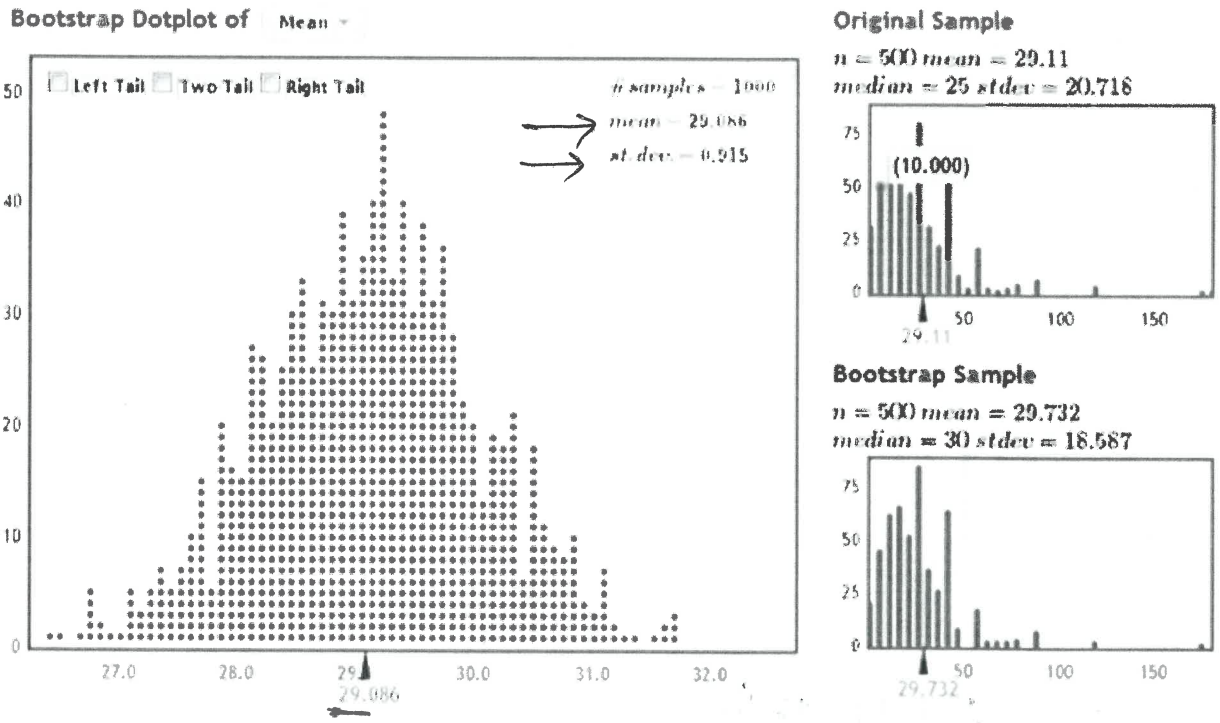
In practice, instead of actually making many copies of the sample and sampling from that, we use a sampling technique that is equivalent. We sample With replacement from the original sample. Sampling with replacement means that once a commuter has been selected for the sample, they will still be available to be selected again.

This is because we're assuming that each commuter represents many fellow Atlantans with a similar commute time. Each sample selected in this way, with replacement from the original sample is called a bootstrap sample.

Recall that the variability of a sample statistic depends on the size of sample. Because we are trying to uncover the variability of the sample statistic, it is important that each bootstrap sample is the same size as the original sample. For the Atlanta commuters, each bootstrap sample will be of size 500

For each bootstrap sample, we compute the statistic of interest, giving us a sample statistic. If we take many bootstrap samples and compute a bootstrap statistic from each, the distribution of these bootstrap statistics will help us understand the distribution of the sample statistic.

The figure below shows a dotplot of the sample means for 1000 samples of size 500, taken with replacement, from the original sample of Atlanta commute times.



This gives a good representation of the bootstrap distribution for mean Atlanta commute times. The distribution is sample statistic centered; not the population parameter.

The variability of bootstrap statistics is similar to the variability of sample statistics if we were to sample repeatedly from the population, so **we can use the standard deviation of the bootsrap distribution to estimate the standard error of the sample statistic.** Thus the standard deviation of these bootstrap statistics is the standard error we need to build a 95% confidence interval for μ,

the true mean commute time among people who live in the ATLANTA metropolitan area commute to work.

statistic ± 2 SE.

Now build the 95% confidence interval. Then interpret it in context.

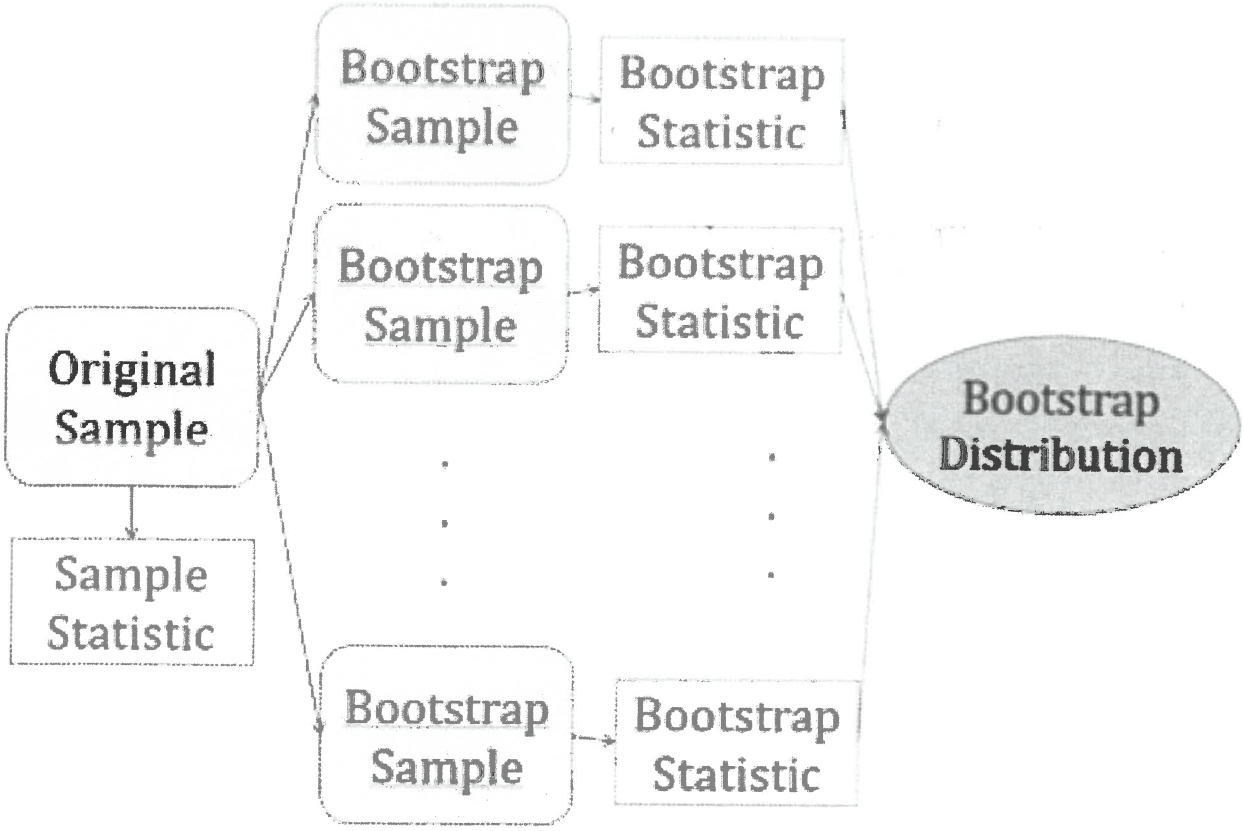
29.086 ± 2(.915)

= (29.086 − 2×0.915, ...) DIY

we are 95% confident that the true value of the mean commute time for people who live in Atlanta metropolitan area and commute to work in the interval (29.086 − 2×0.915, 29.086 + 2×0.915)

Instructions for Generating a Bootstrap Distribution

- Take a random sample from the population.
- Generate bootstrap samples by sampling with replacement from the original sample, making the bootstrap samples the same size as the original sample.
- Compute the statistic of interest, the bootstrap statistic, for each of the bootstrap samples.
- Collect the statistics for many bootstrap samples to create a bootstrap distribution.



Example. Suppose you are going to use bootstrapping to construct a 95% confidence interval for a population parameter. You take a sample of size $n = 5$ with data values $18, 19, 19, 20, 21$. Determine whether each of the following is a possible bootstrap sample.

1. $18, 19, 20, 21, 22$ ✗ as 22 is not in the original sample
2. $18, 29, 20, 21$ ✗ as of not same original sample size
3. $18, 18, 19, 20, 21$ ✓
4. $18, 19, 19, 20, 21$ ✓

- The more bootstrap samples, the more accurate the estimated standard error will be.
- When a bootstrap distribution for a sample statistic is symmetric and bell-shaped we estimate a 95% confidence interval using

  statistic $\pm 2\,SE$.

  where SE denotes the standard error of the statistic estimated from the bootstrap distribution.

- We do NOT use the distribution to get the point estimate for the population parameter of interest.

  We use the original sample. for that.

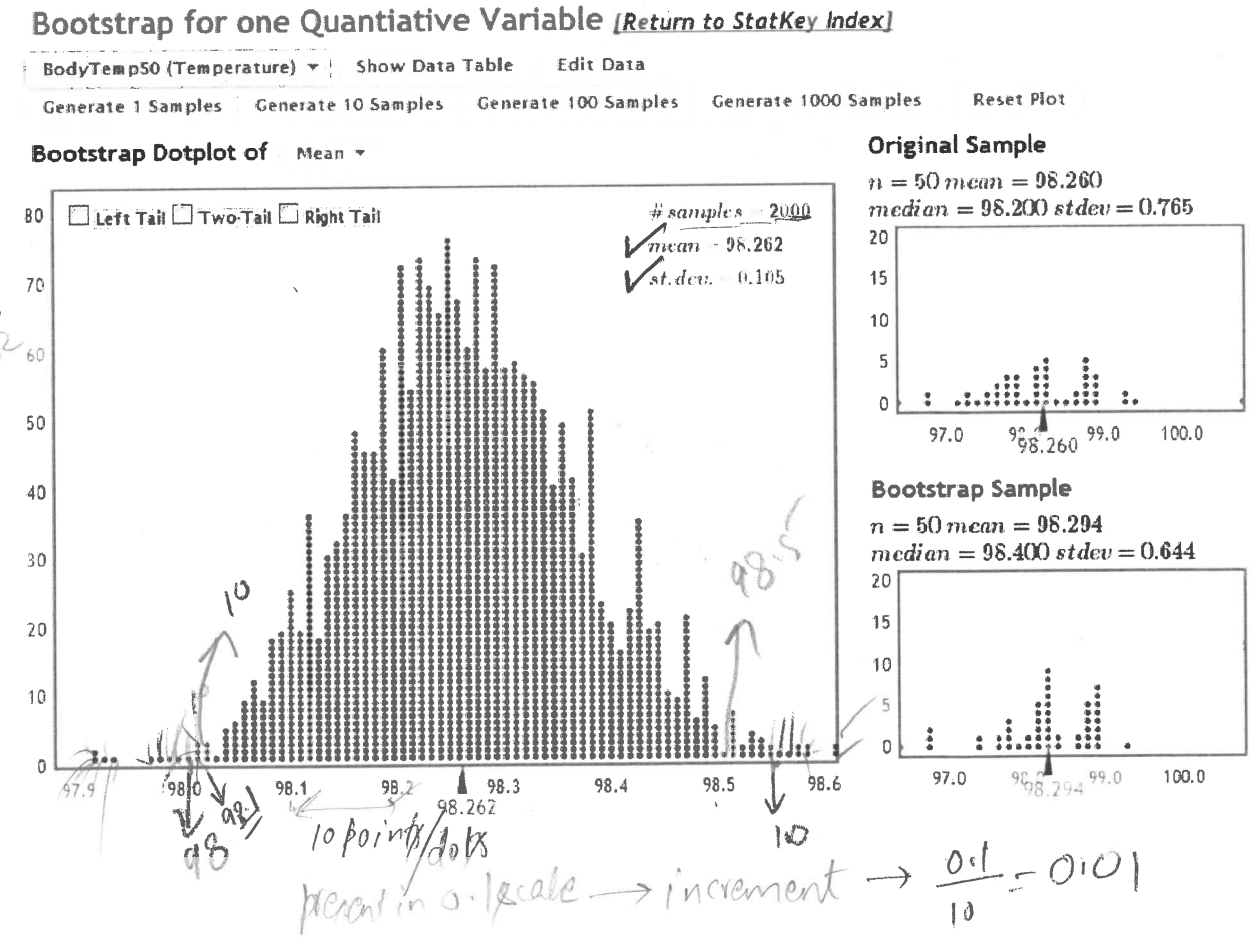
HIMANSHU SINGH

In Section 3.3, we discussed how to use bootstrap distributions to construct a 95% confidence interval for the population parameter of interest. What if we want a different confidence interval, such as a 90% confidence interval or a 99% confidence interval? Well, we shouldn't use the 95% Rule! If the bootstrap distribution is approximately symmetric, we can use percentiles to construct a confidence interval of any level.

Example. Suppose we want to know the average body temperature of humans. We take a random sample of $n = 50$ humans and measure the body temperature of each. We then generate bootstrap samples, also of size 50, by sampling with replacement from the original sample. We generate 2000 of these bootstrap samples. We compute a bootstrap statistic for each bootstrap sample and use these to create a bootstrap distribution. The results are shown below.

0.5% of 2000

$\frac{5}{1000} \times 2000 = 10$



0.5 ← 1%

0.5

10

98.5

98

10

10 points/dots

percent in 0.1 scale → increment → $\frac{0.1}{10} = 0.01$

Find a 95% confidence interval for the average body temperature of humans and interpret it in context.

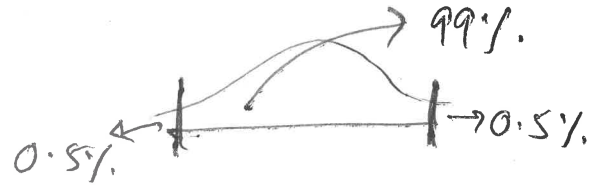
statistic ± 2SE

↓

$98.262 \pm 2 \times 0.105$

$= (98.05, 98.47)$

→ (meaning) we are 95% confident that the average body temp. of all humans is in the given interval.



What if we wanted a different confidence interval?

- Find a 99% confidence interval for the average body temperature of humans and interpret it in context.

we want 99% data capture in our confidence interval & leave 1% on the edges ⇒ 0.5% on both edges.
bootstrap $n = 2000$ ⇒ 1% of $2000 = \frac{1}{100} \times 2000 = 20$.
That is, leave 20 dots on dotplot; 10 dots on both edges.

↳ (98.01, 98.85) → 99.5th percentile
↓ 5th percentile

- Find a 90% confidence interval for the average body temperature of humans and interpret it in context.

(follow the same step as above)
capture → 90% & leave 10% of it on edges.
meaning 5% on both edges. & 5% of 2000 = 100 dots.

so ans (98.1, 98.45) → 95th percentile
↓ 5th percentile

Recall that the pth percentile is the value of a quantitative variable which is greater than p % of the data.

Remark. If the bootstrap distribution is approximately symmetric, we construct a confidence interval by finding the percentiles in the bootstrap distribution so that the proportion of bootstrap statistics between the percentiles matches the desired confidence level.

Remark. If the bootstrap distribution is symmetric AND bell-shaped, the standard error method and the percentile method give approximately the same results for a 95% confidence interval.

Remark. If you want a confidence level other than 95%, your only option is the percentile method.

Returning to the example for average human body temperature, which confidence level gives the widest interval? The narrowest?

narrowest 90% (98.1, 98.45)

95% (98.05, 98.47)

99% → widest (98.01, 98.55)

Remark. In general, the higher the confidence level, the wider the interval.

Remark. You should always look at a plot of the bootstrap distribution. The methods we have discussed for obtaining confidence intervals work only if the bootstrap distribution is smooth and symmetric. f the plot is heavily skewed or contains isolated clumps of data, you should not trust in the intervals it produces.

HIMANSHU SINGH

In statistical inference, we use data from a sample to make conclusions about a population. The two main areas of statistical inference are estimation and testing. In Chapter 3, we discussed how to estimate with a confidence interval. In this chapter, we see how to use a hypothesis test to address situations like the following two examples.

Example. When a drug company wishes to introduce a new drug, they must first convince the Food and Drug Administration that it is a worthwhile addition to the current set of drugs. That is, they must be able to demonstrate that the new drug is better than current drugs, with "better" possibly referring to greater effectiveness, fewer side effects, or quicker response by the patient. There are risks in introducing a new drug. Is it worth it?

Example. When an advertiser is considering launching a new ad campaign, they are first interested in knowing if the new campaign would be an improvement on the current campaign. Will it alienate current customers? Will it reach new customers? Is it worth the cost?

Definition. A statistical test is used to determine whether results from a sample are convincing enough to allow us to conclude something about the population.

In Chapter 3, we use data from a sample to create a confidence interval for a population parameter. In this chapter, we use data from a sample to help us decide between two competing hypotheses about a population. In a statistical hypothesis test, hypotheses of interest are turned into statements about unknown parameters.

Example.

mean survival rate for patients taking new drug is higher than for patients taking other drugs.

Definition. Statistical tests are framed formally in terms of two competing hypotheses. The null hypothesis, denoted H_0 is the claim that there is no effect or difference. The alternate hypothesis denoted H_a is the claim for which we seek significant evidence.

In a hypothesis test, we examine ~~wether~~ whether the sample data provide enough evidence to reject H_0.

Analogy CRIMINAL JUSTICE SYSTEM

$H_0 \longrightarrow$ person is innocent until proven guilty.

$H_a \longrightarrow$ person is guilty.

Example. In 2011, 15.8% of high school students smoked at least one cigarette per month. The CDC now reports that smoking is declining among high school students.

- What is the claim? smoking is declining among high school students.
- What is the population parameter? proportion of high school students who smoke atleast 1 cig./month.
- What are the null and alternative hypotheses?

$H_0 \equiv p = 0.158$ & $H_a \equiv p < 0.158$

Hypotheses are ALWAYS about population parameters.

The null hypothesis is a statement of equality, while the alternative hypothesis contains a range of values. If γ is the population parameter of interest, possible forms for the alternative hypothesis are

- $\gamma >$ value in H_0
- $\gamma <$ value in H_0
- $\gamma \neq$ value in H_0

Example. Suppose the mean GPA of all students graduating from UT Tyler in 1980 was 2.80. The registrar plans to look at records of students graduating in 2018 to see if the mean GPA of graduates has changed.

- What is the claim? mean GPA of UT-Tyler graduates has changed
- What is the population parameter? (μ) mean GPA of UT Tyler graduates.
- What are the null and alternative hypotheses?

$H_0 = \mu = 2.8$

$H_a = \mu \neq 2.8$

Example. A sample of students were given words to memorize and then randomly assigned to either take a 90 minute nap or take a caffeine pill. Two and a half hours later, they were tested on their recall ability. Is there a difference in average word recall between those who slept and those who took caffeine?

- Identify the explanatory and response variables.

EV → student take nap / caffeine pill

RV → score on test of recall ability.

- What is the claim? there is a difference in average word recall b/w who slept and those who consumed caffeine
- What is the population parameter?

μ → mean score on test of recall ability

μ_c → mean score " " " " " after taking caff...

μ_s → " " " " " " after sleeping

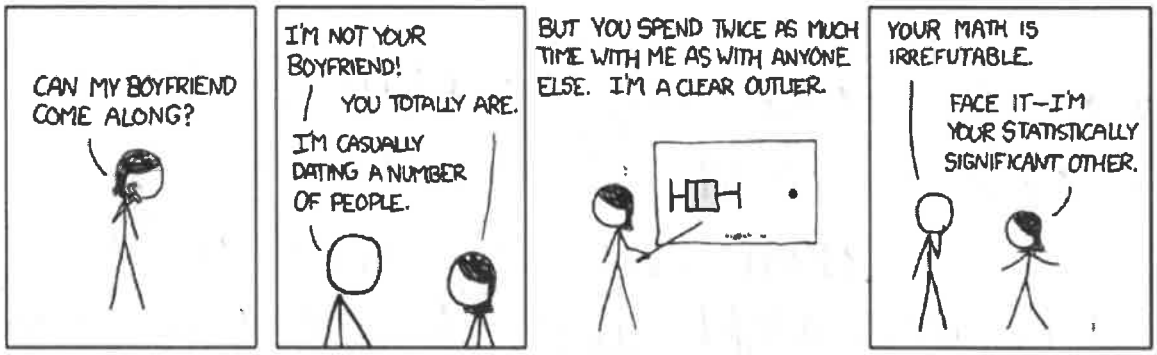
- What are the null and alternative hypotheses?

$H_0 \rightarrow \mu_c = \mu_s$

$H_a \rightarrow \mu_c \neq \mu_s$

Definition. When results as extreme as the observed sample statistic are unlikely to occur by random chance alone, assuming the null hypothesis is true, we say the sample results are Statistically significant.

- If our sample results are statistically significant, we have convincing evidence to reject H_0.
- If our sample results are NOT statistically significant, our test is inconclusive.



HIMANSHU SINGH

In Section 4.1, we saw that a hypothesis test examines whether data from a sample provide enough evidence to reject the null hypothesis. We saw that a sample statistic can be far away from the value corresponding to the null hypothesis for either of two reasons: either because of random chance or because the null hypothesis should be rejected. How can we determine whether there is enough evidence against the random chance option to be convinced that we should reject the null hypothesis? We need to have a sense of what is likely to occur just by random chance.

Example. In 2008, Paul the Octopus predicted 8 World Cup games, and in each game, he chose the eventual winner. Is this evidence that Paul's chance of guessing correctly is really greater than 50%?

- What is the population parameter?

  p (proportion of all games guessed correctly).

- What are the null and alternative hypotheses?

  $H_0 \equiv p = 0.5$

  $H_a \rightarrow p > 0.5$.

  $\hat{p} \rightarrow 8/8 = 1$

- How unusual is it to see a sample statistic as extreme as the observed sample statistic in this scenario if H_0 is true? If it is very unusual, we have statistically significant evidence against the null hypothesis.

- How do we measure how unusual a sample statistic is if H_0 is true?

- To see if a statistic provides evidence against H_0, we need to see what kind of sample statistics we would observe just by random chance if H_0 is true.

- For an octopus picking soccer game winners, there should be a 50% chance of success. What other common activity has 2 possible outcomes, each with probability 0.5? → COIN FLIP

- There is a .37% chance of flipping heads 8 times in a row. Thus there is a .37% chance of Paul picking all 8 winners correctly.

- This data is statistically significant. Thus we reject the null hypothesis.

Head Tail

H ↑ (+) = 2

H ∧ H ∧ H ∧ H ∧ H ∧ H ∧ H ∧ H (∧ → independent)

1 2 3 4 5 6 7 8

1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 → $(\frac{1}{2})^8 = 0.0037$

In Chapter 3, we used a bootstrap distribution to understand how statistics randomly vary from sample to sample. Here we want to understand how statistics randomly vary from sample to sample **if the null hypothesis is true**. Thus, when testing hypotheses, we simulate samples in a way that is consistent with the null hypothesis. We call these randomization samples.

Definition. A randomization distribution is a collection of statistics from samples simulated assuming the null hypothesis is true.

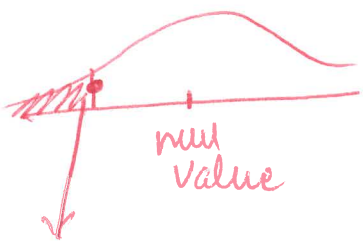
The randomization distribution will be centered at the value indicated by null hypothesis and shows what values of the sample statistic are likely to occur by random chance, **if the null hypothesis is true.**

Definition. The p-value is the proportion of samples that would give a statistic as extreme as or more extreme than the observed sample statistic **if the null hypothesis is true**.

The p-value can be calculated as the proportion of statistics in a randomization distribution that are as extreme or more extreme than the observed sample statistic.

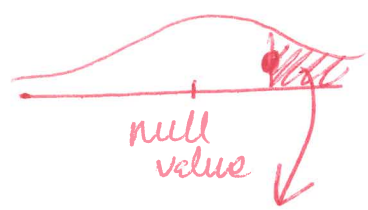
Diagrams (1) one-tailed tests

Ha contains <



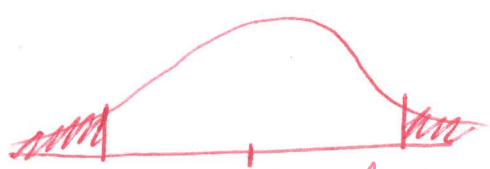
as small or as smaller than the observed sample statistic

Ha contains >



as large as or larger than the observed sample statistic

(2) two-tailed test



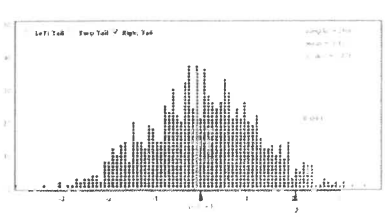
p-value is the sum of those two proportions of the total area

Definition. A one-tailed test contains either a $>$ symbol or a $<$ symbol. A two-tailed test contains a $\neq$ sign.

- If H_a involves a $>$ symbol, then we are conducting a RIGHT-tailed test.
- If H_a involves a $<$ symbol, then we are conducting a LEFT-tailed test.
- If H_a involves a $\neq$ sign, then we are conducting a two-tailed test.

The p-value of a statistic in a one-tail test is the proportion in the tail in the direction specified by H_a. When working with a two-tail alternative, the p-value is twice the proportion in the smallest tail.

Right-Tail or Upper-Tail



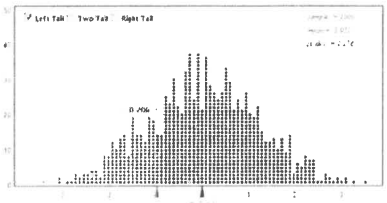
$H_0: \mu = 0$ $\bar{x} = 2$

$H_a: \mu > 0$

p-value corresponding to $\bar{x} = 2$ is the proportion of all statistic in the randomization distribution that are greater than or equal to 2.

p-value $= 0.044$

Left-Tail or Lower-Tail

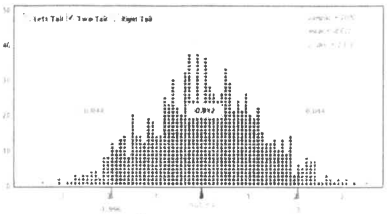


$H_0: \mu = 0$ $\bar{x} = -1$

$H_a: \mu < 0$

p-value corresponding to $\bar{x} = -1$ is the proportion of all statistics in the randomization distribution that are less than or equal to -1; p-value is 0.206

Two-Tailed



$H_0: \mu = 0$ $\bar{x} = -1.996$

$H_a: \mu \neq 0$

p value corresponding to $\bar{x}$-value is the proportion of all statistic that are less than or equal to -1.996 or ≥ 1.996. Here, p-value is

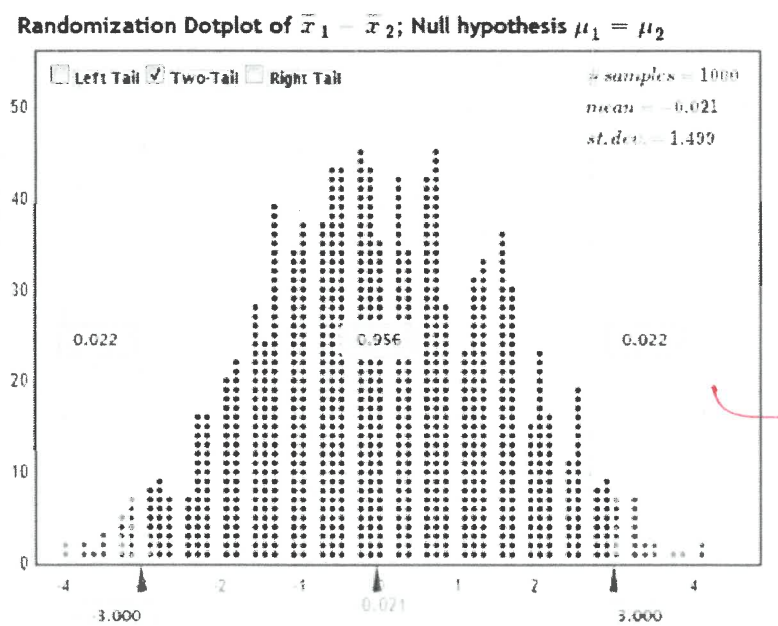
$2(0.044)$

$= 0.088$

(doubled as it's 2-tailed test)

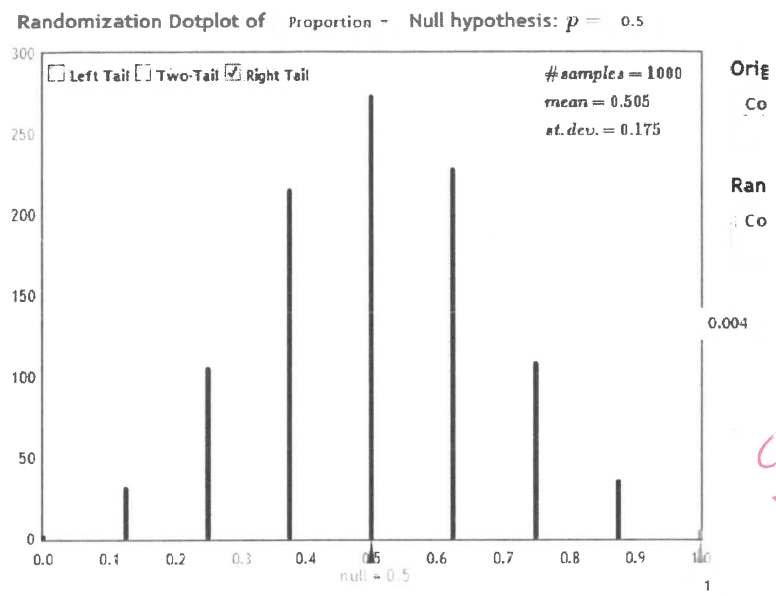
Example. A sample of students were given words to memorize and then randomly assigned to either take a 90 minute nap or take a caffeine pill. Two and a half hours later, they were tested on their recall ability. Is there a difference in average word recall between those who slept and those who took caffeine?

- Identify the explanatory and response variables. EV → whether student took caffiene or a nap.
  RV → score.
- What is the claim?
- What is the population parameter? μ_c → average word recall test after taking caffiene.
  μ_s → same but taking nap.
- What are the null and alternative hypotheses? $H_0 \rightarrow \mu_c = \mu_s$
  $H_a \rightarrow \mu_c \neq \mu_s$.
- Suppose $x_s - x_c = 3$. What is the p-value of this statistic?



→ 2×0.022
$= 0.044$

Example. Returning to our example with Paul the Octopus, where $H_0 : p = 0.5$ and $H_a : p > 0.5$. What kind of test are we conducting? Below is a dotplot of a randomization distribution assuming the null hypothesis is true.



Right

ONE TAILED TEST

Example. In 1980, a random sample of people were asked, "Are you in favor of the death penalty for a person convicted of murder?" This sampling and questioning was repeated in 2010. The results are recorded in the table below.

$\hat{p}_{2010} = \frac{640}{1000} = 0.64$

$\hat{p}_{1980} = \frac{663}{1005} = 0.66$

	Yes	No	Total
1980	663	342	1005
2010	640	360	1000

Did the proportion of Americans who favor the death penalty decrease from 1980 to 2010?

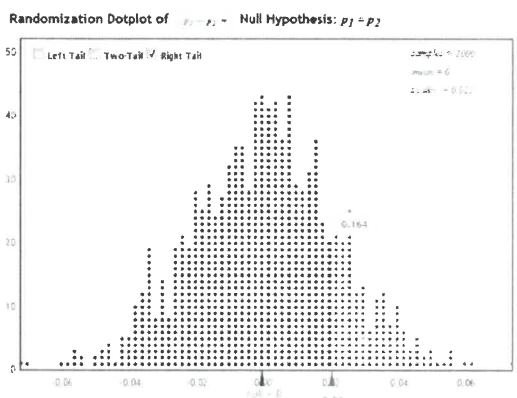
let $p_{1980} \rightarrow$ proportions of AMERICANS who favored the death penalty in 1980

(and) $p_{2010} \rightarrow$ same but who favored it in 2010.

$H_0: p_{1980} = p_{2010} \Rightarrow p_{1980} - p_{2010} = 0$

$H_a: p_{1980} > p_{2010} \Rightarrow p_{1980} - p_{2010} > 0$

as $\hat{p}_{1980} - \hat{p}_{2010} = 0.66 - 0.64 = 0.02$



Original Sample

Group	Count	n	Proportion
Group 1	663	1005	0.66
Group 2	640	1000	0.64
Group 1-Group 2	23	n/a	0.02

Randomization Sample

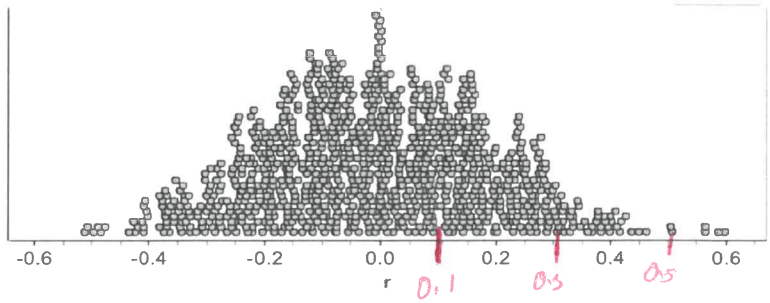
Group	Count	n	Proportion
Group 1	636	1005	0.63
Group 2	667	1000	0.67
Group 1-Group 2	31	n/a	-0.03

p-value corresponding to 0.02 is 0.164 so we do not reject H_0.

If the proportion of Americans supporting the death penalty had not changed from 1980 to 2010, we would see differences this extreme about _____ of the time.

16.4 %

Example. The randomization distribution below is used to test $H_0 : \rho = 0$ against $H_a : \rho > 0$.



Consider the sample statistics $r = 0.1, r = 0.3$, and $r = 0.5$ and the p-values: $0.005, 0.15, 0.35$.

- Match each of the sample statistics with one of the p-values.

r = 0.1	p-value = 0.35
r = 0.3	= 0.15
r = 0.5	= 0.005

- Which sample statistic shows the most evidence for the alternative hypothesis?

r = 0.5

- Which p-value shows the most evidence for the alternative hypothesis?

p-value is 0.005

In Section 4.1, we saw that a hypothesis test involves determining whether sample data provide enough evidence to refute a null hypothesis and support an alternative hypothesis. In Section 4.2, we saw that a p-value measures how unlikely a sample statistic is **if the null hypothesis is true**. In this section we will connect these two ideas.

Definition. The p-value is the proportion of samples that would give a statistic as extreme as or more extreme than the observed sample **if the null hypothesis is true**.

Remark. The smaller the p-value, the stronger the statistical evidence is against the null hypothesis.

If the p-value is small enough, then results as extreme as the observed sample statistic are unlikely to occur by random chance **if the null hypothesis is true**, and we say that sample results are ~~statistically~~ significant. If our sample is statistically significant, we have convincing evidence against H_0.

Question: What is small enough? How small does a p-value have to be for us to achieve statistical significance?

Definition. The significant level, denoted α (alpha), for a test of hypotheses is a boundary below which we conclude that a p-value shows statistically significant evidence against the null hypothesis.

Remark. Common significance levels are $\alpha = 0.05$, 0.01, or 0.1. If a significance level is not specified, we use $\alpha = 0.05$

How to Make a Formal Statistical Decision

Give a significance level α and the p-value from a sample,

- If the p-value is less than α, reject H_0. → results are statistically significant and we have evidence against H_0.
- If the p-value is greater than or equal to α, we fail to reject H_0. → results are not statistically significant and we do not have evidence against H_0

Note that a formal hypothesis test has only two possible conclusions. They are reject H_0 and ~~fail to reject H_0~~. We NEVER accept H_0.

Example. Resveratrol, an ingredient in red wine and grapes, has been shown to promote weight loss in rodents and has recently been investigated in primates, specifically the Grey Mouse Lemur. A sample of lemurs had various measurements taken before and after receiving resveratrol supplementation for 4 weeks.

- A test was conducted to see if mean resting metabolic rate is higher after treatment. The p-value of the sample statistic is 0.013. Use a significance level of $\alpha = 0.05$.
  - Give H_0 and H_a.

$\mu_A \rightarrow$ mean metabolic rate after treatment

$\mu_B \rightarrow$ mean metabolic rate before treatment.

$H_0 \rightarrow \mu_A = \mu_B \rightarrow \mu_A - \mu_B = 0$

$H_a \rightarrow \mu_A > \mu_B$

$\downarrow$

$\mu_A - \mu_B > 0$

  - Is this sample statistic statistically significant?

$p = 0.013 < 0.05 \ (= \alpha)$

yes, $\because p < \alpha$

  - Give a formal conclusion.

We reject $H_0 : \mu_A = \mu_B$.

  - Interpret the conclusion in context.

We reject $H_0 : \mu_A = \mu_B$ as we have evidence against mean resting metabolic rate being the same before & after treatment.

$\mu_B \rightarrow$ mean bodymass before treatment
$\mu_A \rightarrow$ " " " after treatment

- A test was conducted to see if mean body mass is lower after treatment. The p-value of the sample statistic is 0.007. Use a significance level of $\alpha = 0.05$.
  - Give H_0 and H_a.

$H_a \rightarrow \mu_A = \mu_B \rightarrow \mu_A - \mu_B = 0$
$H_0 \rightarrow \mu_A < \mu_B \rightarrow \mu_A - \mu_B < 0$

  - Is this sample statistic statistically significant? $p = 0.007$ & $\alpha = 0.05$

$\therefore p < \alpha$ (yes)

  - Give a formal conclusion.

reject H_0

  - Interpret the conclusion in context. We reject H_0 as we have evidence against mean body mass being same and after treatment.

- A test was conducted to see if mean locomotor activity changes after treatment. The p-value of the sample statistic is ~~0.013~~ 0.98. Use a significance level of $\alpha = 0.05$.
  - Give H_0 and H_a.

$\mu_A \rightarrow$ after treatment
$\mu_B \rightarrow$ before treatment

$H_0 : \mu_A = \mu_B \rightarrow \mu_A - \mu_B = 0$
$H_a : \mu_A \neq \mu_B \quad \mu_A - \mu_B \neq 0$

  - Is this sample statistic statistically significant? $\left.\begin{array}{l} p = 0.98 \\ \alpha = 0.05 \end{array}\right\} p > \alpha$

(No)

  - Give a formal conclusion.

We fail to reject H_0.

  - Interpret the conclusion in context.

We fail to reject $H_0 : \mu_A = \mu_B$ because we do not have evidence against mean locomotor activity being the same before & after treatment

$p_{UV} \rightarrow$ mice proportion in UV developed MS

$p_D \rightarrow$ " " " Vitamin D who developed MS

$p_c \rightarrow$ " " " Control " " "

Example. It is believed that sunlight offers some protection against multiple sclerosis, but the reason why is unknown. Researchers randomly assigned mice to one of three groups. One group received Vitamin D supplements, one group was exposed to UV light, and the third group was a control group. All the mice were injected with proteins known to induce a mouse form of multiple sclerosis, and researchers observed which mice got multiple sclerosis.

- In testing whether UV light provides protection against multiple sclerosis, the UV group is compared to the control group, and the p-value is 0.002.
  - Give H_0 and H_a.

$H_0: p_{UV} = p_c \rightarrow p_{UV} - p_c = 0$

$H_a: p_{UV} < p_c \rightarrow p_{UV} - p_c < 0$

  - Is this sample statistic statistically significant?

$p = 0.02 < \underset{\alpha}{0.05}$ yes as $p < \alpha$

  - Give a formal conclusion.

reject H_0

  - Interpret the conclusion in context.

(DIY)

- In testing whether Vitamin D provides protection against multiple sclerosis, the Vitamin D group is compared to the control group, and the p-value is 0.47.
  - Give H_0 and H_a. $H_0: p_D = p_c$ $H_a: p_D < p_c$ (DIY)
  - Is this sample statistic statistically significant?

$p = 0.47 > 0.05 = \alpha$. No as $p > \alpha$

  - Give a formal conclusion.

fail to reject H_0.

  - Interpret the conclusion in context.

(DIY).

HIMANSHU SINGH

Hypothesis testing is very powerful, as it helps shed light on whether an observed effect is real or just due to random chance. However, statistical significance is not foolproof, and it is possible to make the wrong decision, rejecting a true H_0 or failing to reject a false H_0. This section discusses common pitfalls associated with formal hypothesis testing, along with factors that influence the chances of these errors occurring.

Definition. A Type I Error occurs when we reject a true H_0.

Definition. A Type II Error occurs when we fail to REJECT a false H_0.

	Reject H_0	Fail to reject H_0
H_0 is true	Type I Error	no error
H_0 is false	no error	Type II Error

Example. Suppose we are testing to determine whether a well-known company is lying in its advertising. If there is evidence that the company is lying, the Federal Trade Commission will file a lawsuit against the company.

- Describe H_0 and H_a in words.

$H_0 \rightarrow$ company is innocent; no lying in advertising.

$H_a \rightarrow$ company is guilty; lying in advertising.

- Describe what it means to make a Type I error in this context.

A type I error imply rejecting the presumption that company is innocent when it really is.

- Describe what it means to make a Type II error in this context. A type2 error would be failing to reject the presumption that the company is innocent when in fact the company is guilty.

Example. Suppose we are testing water samples from a contaminated lake to determine whether it is safe for swimming. If the average levels of the contaminant exceed 200 parts per billion, the lake will be deemed unsafe for swimming.

- Give the null and alternative hypotheses. $\mu \longrightarrow$ average

$H_0: \mu = 200$ ppb

$H_a: \mu > 200$ ppb (exceeds)

- Describe what it means to make a Type I error in this context. Type 1 error would occur if we reject $H_0: \mu = 220$, when infact it is true. We would declare the lake is unsafe for swimming when actually it is safe.

- Describe what it means to make a Type II error in this context. Type 2 error would occur if we fail to reject $H_0: \mu = 220$ ppb when in fact $\mu > 220$ ppb. We would declare the lake safe for swimming when it is actually safe for swimming.

- Is a Type I error or a Type II error more dangerous in this context?

$\boxed{\text{Type 2}} \longrightarrow$ $\because$ it let people to swim in unsafe water.

If we make it very hard to reject H_0, we could reduce the chance of making a type 1 error, but then we would increase the chance of making a type 2 error. On the other hand, making it easier to reject H_0 would reduce chance of making type 2 error but increase the chance of making type 1 error.

- To decrease the chance of making a type 1 error we make it harder to reject H_0 by using a lower significance level.
- To decrease the chance of making type 2 error, we make it easier to reject H_0 by using a higher α.

Example. What should we do in the example with the contaminated lake water? Should we set a smaller value of α or a larger value? In this case a type 2 error is more severe. To avoid a type 2 error we should make it easier to reject H_0 Thus choose a higher significance level α.

The significance level α represents the tolerance probability of making a type 1 error.

Remark. A difference that is statistically significant might not have much practical significance.

Because a larger sample size decreases the spread of values we might see just by random chance, a larger sample size makes it easer to to reject H_0 when H_0 is false.

To maximize your chance of finding significant results, you should always use the largest sample size possible. A larger sample size decreases the chance of a type 2 error. In this case, there is no trade-off with Type I errors, because the chance of making a Type I error is α and is not affected by sample size. **A larger sample size is always better!**

How do we choose α?

- Typically, we let $\alpha = 0.05$.
- If a Type I error is a life or death situation or if big money is involved, we let $\alpha < 0.05$.
- If a Type II error is worse than a Type I error, let $\alpha > 0.05$.

We will now discuss some problems that can arise with Confidence Intervals and Hypothesis Tests.

Statistical Versus Practical Significance

- With small sample size ↑, even large differences or effects may not be significant.
- With large sample set size ↑, even a very small difference or effect can be significant.
- A statistically significant result is not always practically significant, especially with large sample sizes.

Example. A weight loss program recruits 10,000 people for a randomized experiment. A difference in average weight loss of 0.5 pounds is found to be statistically significant. Suppose the experiment lasted for a year. Is a loss of 0.5 pounds over a year practically significant?

NO.

Incorrect Results Due To Multiple Testing

When multiple hypothesis tests are conducted, the chance that at least one test incorrectly rejects a true null hypothesis increases with the number of tests. If the null hypotheses are all true, α percent of the tests will yield statistically significant results just by random chance.

Example. Are certain foods in a woman's diet associated with whether she will conceive a boy or a girl? To study this, researchers asked women about their eating habits, including whether or not they regularly ate 133 different foods. For **each** of the 133 foods studied, a hypothesis test was conducted to compare, among women who conceived boys, the proportion who ate that food regularly and, among women who conceived girls, the proportion who ate that food regularly.

Let p_B denote the proportion, among mothers conceive boys, who consume that food regularly. Let p_G denote the proportion, among mothers who conceive girls, who consume that food regularly. State the null and alternative hypotheses.

$H_0: p_B = p_G$

$H_a: p_B \neq p_G$

If there are NO differences in proportion, then all 133 null hypotheses are true. About how many significant differences would we expect to find using the significance level $\alpha = 0.05$?

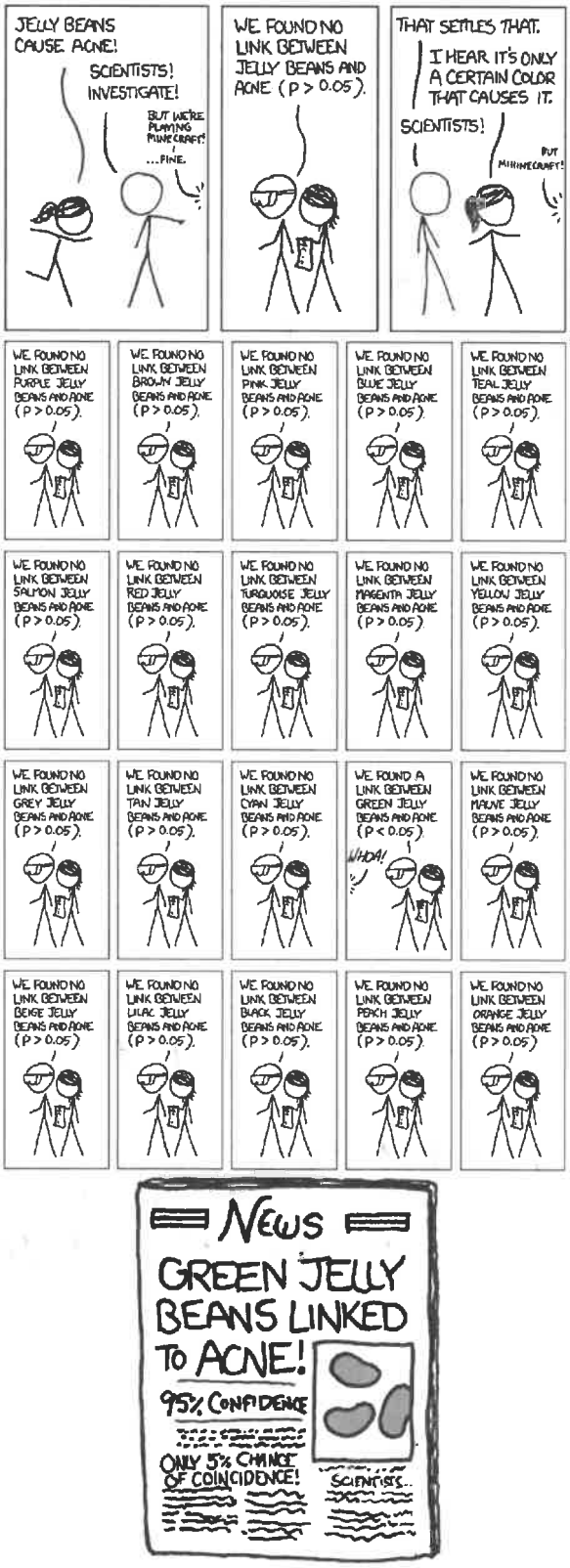
$133 \times 0.05 = 6.65$

A significant difference was found for breakfast cereal, showing that among mothers who conceive boys, cereal is more likely to be consumed regularly than among mothers who conceive girls. This led to the headline "Breakfast Cereal Boosts Chances of Conceiving Boys." How might you explain this headline?

we expect that b/w 6 and 7 of the 133 tests will have statistically significant results even if all 133 H_0 is true.

Publication Bias

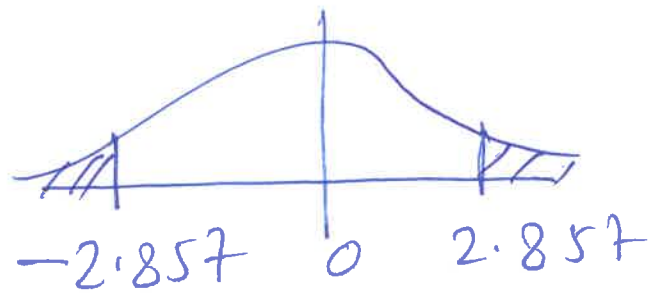
Often, only significant results are published. If many tests are conducted, some of them will be significant just by chance, and it may be only these studies that we hear about. The one study that turns out significant results gets published, and no one knows about all the insignificant results. In the example above, breakfast cereal boosts chances of conceiving a boy was published, but none of the other results were.


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05)
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05)
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

soybeans being alive after 5 days.

say $H_0: p_c = p_0$

$H_a: p_c \neq p_0$



<u>two-tailed test</u>

$$p = P(z > 2.857) + P(z < -2.857)$$
$$= 2P(z > 2.857)$$
$$= 2P(z < -2.857)$$

<u>HW</u> <u>section 5.1</u>

#1-11 odd

15-23 odd

When the distribution of the statistic under H_0 is normal, we compute a **standardized test statistic** using

$$z = \frac{\text{sample statistic} - \text{null value}}{SE}.$$

The p-value for this test is the proportion of a standard normal distribution beyond this standardized test statistic, depending on the direction of the alternative hypothesis.

Example. Does eating organic foods actually offer health benefits? To investigate the answer, a sixteen-year-old girl studied fruit flies. She fed one group conventional soybeans and the other group organic soybeans. After 5 days, the proportion of fruit flies eating organic soybeans still alive is 0.90, while the proportion still alive eating conventional soybeans is 0.84. The standard error for the difference in proportions is 0.021. Conduct a hypothesis test using a 5% significance level.

1. Define the parameters of interest.

$p_c \rightarrow$ proportion of all fruit flies eating conventional soybeans that are still alive after 5 days.

$p_o \rightarrow$ " " " " " " organic " " " " " "

$z = \frac{x-15}{5}$

- Find the area to the left of 12 for a $N(15, 5)$ distribution. $P(X<12)$

$\Rightarrow P\left(\frac{X-15}{5} < \frac{12-15}{5}\right) = P(Z < -3/5) = P(Z < -0.6)$
$= .2743.$

$z = \frac{28-15}{5}$

- Find the area to the right of 28 for a $N(15, 5)$ distribution.

$P(X>28) = P\left(Z > \frac{28-15}{5}\right) = P\left(Z > \frac{13}{5}\right) = P(Z > 2.6)$
$= 1 - P(Z < 2.6) = 1 - .9953 = \text{DIY}.$

Example. Suppose that scores on the verbal portion of the SAT for applicants at a particular university are normally distributed with mean 580 and standard deviation 70. $N(580, 70)$

1. What proportion of applicants have SAT scores above 700?

$P(X>700) = P\left(Z > \frac{700-580}{70}\right) = P\left(Z > \frac{120}{70}\right)$
$= P\left(Z > \frac{12}{7}\right) = P(Z > 1.71) = 1 - P(Z < 1.71)$
$= 1 - 0.9564.$

2. Find the 25th percentile. $\rightarrow 0.25 \Rightarrow P(X<d) = 0.25$ (find d)

we can find 'c' ∋ $P(Z<c) = 0.25$?

From normal distribution chart $c = -0.675$.

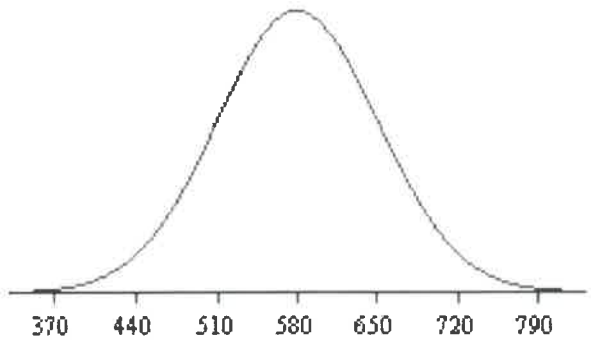
$\left\{ z = \frac{x-\mu}{\sigma} \right.$

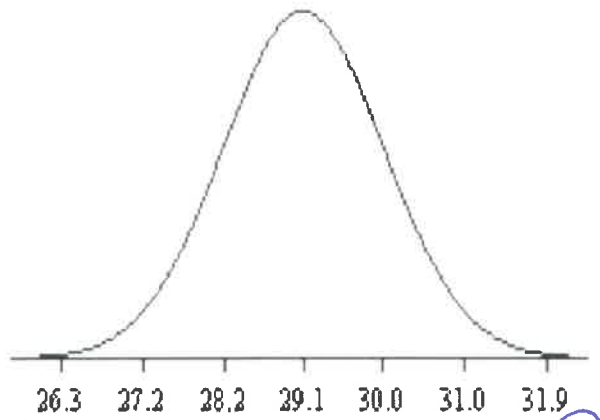
$c = \frac{d-\mu}{\sigma}$

$\Rightarrow -0.675 = \frac{d-\mu}{\sigma} = \frac{d-580}{70} \Rightarrow \boxed{70(-0.675) + 580 = d}$

Central Limit Theorem For random samples with sufficiently large sample size, the distribution of many common sample statistics can be approximated with a normal distribution.

The normal distribution that best approximates a bell-shaped **randomization distribution** has mean equal to the null value of the parameter with standard deviation equal to the SE. This gives N(null value parameter, SE) When the randomization distribution is shaped like a normal distribution, a p-value can be found as the proportion of this normal distribution beyond the observed sample statistic in the direction of the alternative, or twice the smaller tail for two-tailed tests.





$N(580, 70)$ 2/95% rule

$\mu - 2\sigma = 440$ | $\mu + 2\sigma = 720$

$580 - 2\sigma = 440$ | $580 + 2\sigma = 720$

$\Rightarrow \boxed{\sigma = 70}$

$N(29.1, .95)$ 2/95% rule

$\mu - 2\sigma = 27.2$ | $\mu + 2\sigma = 31$

$29.1 - 2\sigma = 27.2$ | $29.1 + 2\sigma = 31$

$\Rightarrow \boxed{\sigma = .95}$

How can we find areas under a normal density curve?

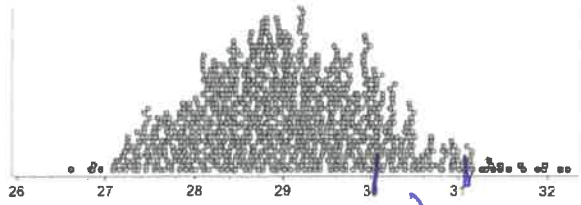
- We can use calculus to determine the area under the curve from a to b by computing $\int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx$.
- We can use technology.
- We can use mathematical tables.

Definition. The standard normal distribution has mean 0 and standard deviation 1 and is thus denoted $N(0,1)$.

To convert from an X value on a $N(\mu, \sigma)$ scale to a Z value on a $N(0,1)$ scale, we standardize values with the z-score

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

What proportion of the data in the distribution below is between 30 and 31?



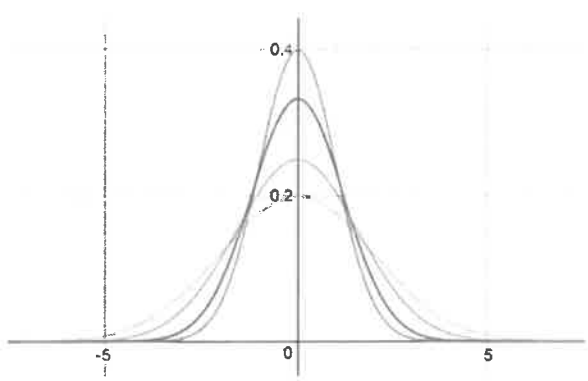
about 1/4 (subjective)

Definition. A normal distribution has a symmetric, bellshaped density curve.

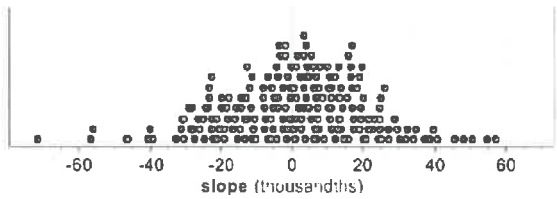
What distinguishes one normal curve from another?

- center of distribution is mean
- spread of " is measured by the standard deviation

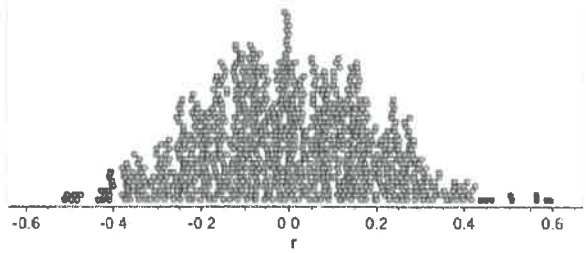
Remark. We use the notation $N(\mu, \sigma)$ to specify that a distribution is normal with mean μ and standard deviation σ. The notation $X \sim N(\mu, \sigma)$ indicates that the distribution of the variable X is represented by a normal distribution with mean μ and standard deviation σ.



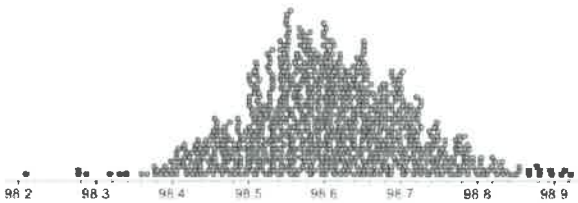
HIMANSHU SINGH



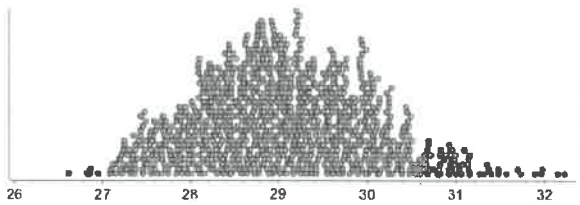
A distribution of tips at a particular restaurant



A distribution of correlation values



A distribution of body temperatures



A distribution of Atlantans' commute times to work

all are bell-shaped curves.

What do all of these distributions have in common?

Definition. A density curve is a theoretical model to describe the distribution of a variable. for each distribution above, the density curve is a bell curve.

Think of a density curve as an idealized histogram satisfying the following.

- The total area under the curve is 1.
- The proportion of the population in any interval is the area over that interval.

HIMANSHU SINGH

Central Limit Theorem For random samples with sufficiently large sample size, the distribution of many common sample statistics can be approximated with a normal distribution.

- The Central Limit Theorem holds for ANY original population distribution although what constitutes a "sufficiently large sample size" varies.
- The more skewed the original distribution is, the larger the sample size has to be for the Central Limit Theorem to hold.
- For quantitative variables that are not very skewed, $n \geq 30$ is usually sufficient.
- For categorical data, counts of at least $n \geq 10$ in each category are sufficient.

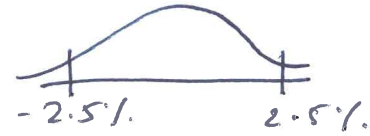
If a bootstrap distribution is bell-shaped, a $P\%$ confidence interval can be found as the interval containing the middle $P\%$ of the normal distribution with mean equal to the observed sample statistic and standard deviation equal to the standard error of the statistic. This means we are using the normal curve ↳ SD.
N(sample statistic, SD).

Our preference is to work with $N(0, 1)$, the standard normal distribution with mean 0 and standard deviation 1. Then we can use our handy charts!

If the distribution for a statistic follows the shape of a normal distribution with standard error SE, we find a confidence interval for the parameter using

$$\text{statistic} \pm z^* SE.$$

where z^* is chosen so that the proportion between $-z^*$ and $+z^*$ in the standard normal distribution is the desired level of confidence.



Find the correct value of z^* to use for building a confidence interval with the specified level of confidence.

- 95% ⇝ 5% left ⇝ 2.5% remaining on L & R (EVEN). SO

$P(Z<-c) = P(Z>c) = 0.025 \Rightarrow c = 1.96$

- 90% ⇝ 10% left ⇝ 0.05 (5% remaining on L & R)

$P(Z<-\cancel{0.05}\ c) = P(Z>\cancel{0.05}\ c) = 0.05 \Rightarrow z^* = 1.645$

- 99%

(DIY) $z^* = 2.576$.

Note that for a larger level of confidence, the normal z^* gets bigger and the interval wider to have a better chance of capturing the true ~~mean commute time~~. value of the parameter.

Confidence Level	80%	90%	95%	98%	99%
z*	1.282			2.326	

(DIY from above)

Table 1: Normal percentiles for common confidence levels

$\hat{p}$

Example. A survey of 1000 air travelers found that 60% prefer a window seat. The sample size is large enough to use the normal distribution, and a bootstrap distribution shows that the standard error is $SE = 0.015$. Use a normal distribution to find and interpret a 99% confidence interval for the proportion of air travelers who prefer a window seat.

Let p denote the proportion of all travellers who prefer a window seat.

$\hat{p} = 0.6$

statistic $\pm z^* SE = 0.6 \pm 2.576 \times 0.15$

↓ (DIY in interval form)

We are 99% confident that the proportion of all air travellers who prefer window seats lies in the interval (0.561, .639).

- **Confidence Intervals:** To build a confidence interval using a standard normal distribution, use

  statistic $\pm$ $z^* SE$

  where z^* is a standard normal endpoint based on the desired confidence level.

- **Hypothesis Tests:** To conduct a hypothesis test using a standard normal distribution, we compute a standardized test statistic using

  $z = (\text{sample statistic} - \text{null value}) / SE.$

  and find a p-value using this test statistic on a standard normal distribution.

Himanshu Singh

Math 1342 **Inference for a Proportion** **6.1**

REFER TO CONTEXT: (page 1-3).

For categorical data, the parameter of interest is often a population proportion p. The Central Limit Theorem states that sample statistics often follow a normal distribution if the sample size is large. This is true for the distribution of sample proportions $\hat{p}$, which will be centered at the population proportion p. Besides the center, the other important quantity we need to describe a normal distribution is the standard deviation, the standard error of the sample proportions. As we have seen, one way to estimate this standard error is to use a bootstrap or randomization distribution. Statistical theory gives us an alternative method for finding the standard error.

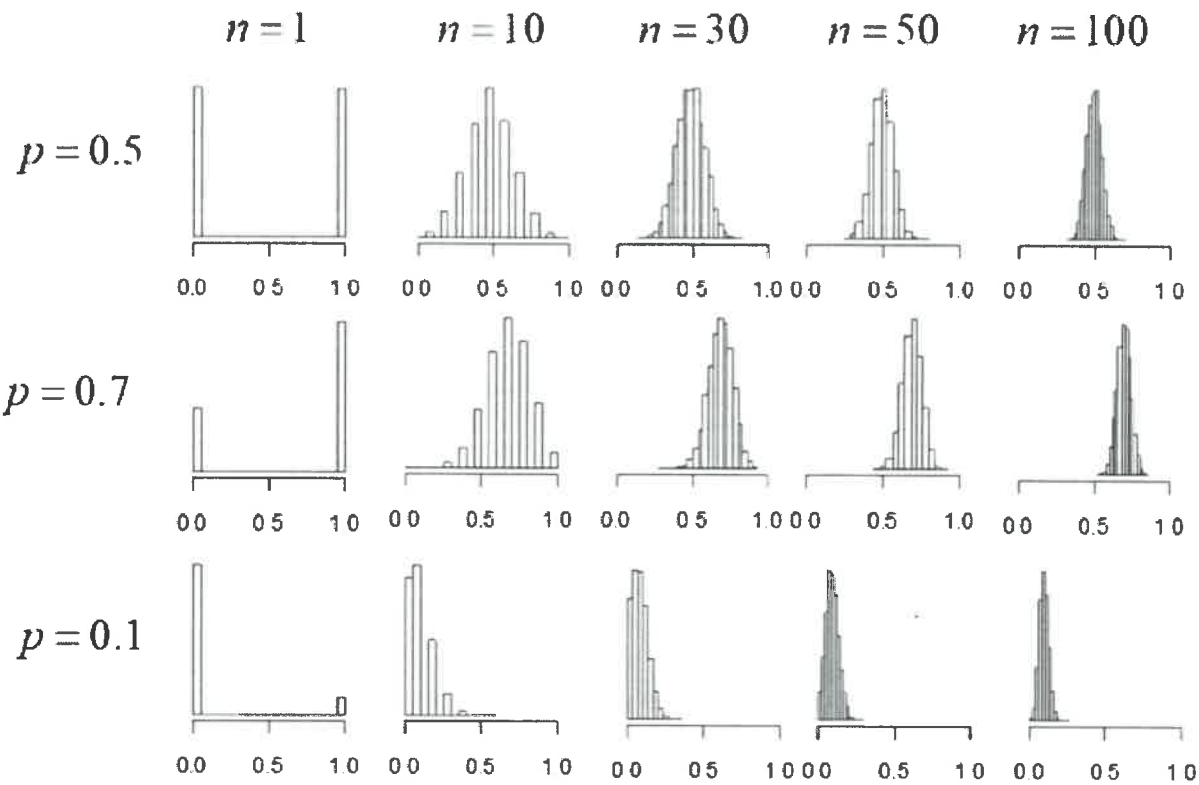
Distribution of a Sample Proportion

When selecting random samples of size n from a population with proportion p, the distribution of the sample proportions

- is centered at _____,
- has standard error given by
- and is reasonably normally distributed if _____ and _____.

Example. The 2010 census reports that, of all of the nation's occupied housing units, 65.1% are owned by the occupants. If we were to take random samples of 100 homes, what would be the standard error of the sampling distribution for $\hat{p}$?

In the formula for standard error, the sample size n is in the denominator. What happens to the standard error as we the sample size increases?



Example. Suppose $n = 100$ and $p = 0.08$. Can we assume that $\hat{p}$ is normally distributed?

The Central Limit Theorem for Sample Proportions

If _____ and _____, then we can approximate the distribution of sample proportions by

Confidence Intervals for a Single Proportion

In order to compute the standard error for the sample proportions, we need to know the population proportion p, but that's exactly the quantity we are trying to estimate with the confidence interval! We can address this issue by using the sample proportion $\hat{p}$ in place of the population proportion p when estimating the standard error for a confidence interval. As long as the sample is large enough for the Central Limit Theorem to apply, a normal distribution with this approximated standard error is still a good model for the sample proportions.

The sample proportion based on a random sample of size n has

If z^* is a standard normal endpoint to give the desired level of confidence and if the sample size is large enough so that

then the confidence interval for a population proportion p is

which, in this case, corresponds to

Example. In March 2011, a random sample of 1000 U.S. adults were asked, "Do you favor or oppose 'sin taxes' on soda and junk food?" A total of 320 adults responded in favor of sin taxes. Give a 95% confidence interval for the proportion of all U.S. adults that favor these sin taxes and interpret your results in context.

Example. When a surgical procedure was tried on 212 patients, 186 recovered and were able to resume a normal life, while the other 26 did not recover. Based on this information, obtain a 98% confidence interval for p, the true proportion of all patients who would recover and be able to resume a normal life after the surgical procedure.

What if we want our $P\%$ confidence interval to have a particular margin of error? We can accomplish this by choosing a large enough sample. How large is large enough? Let $\tilde{p}$ denote our guess for the population proportion. If we can't make a guess at the population proportion, we use $\tilde{p} = 0.5$.

Example. In April 2012, the Gallup Poll reported that in a random sample of 1016 U.S. auldts, only 17% approved of the way Congress was handling its job.

- Use the poll to estimate the proportion of all U.S. adults who approve of the way Congress was handling its job. Use a 99% confidence interval for this estimate.

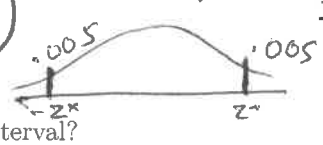
$n = 1016 \quad \hat{p} = 0.17 \quad \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

statistic $\pm z^* SE$

$= 0.17 \pm 2.576 \sqrt{\frac{.17 \times .83}{1016}} = (.14, .2)$

check $n\hat{p} = 1016 \times 0.17 = 172.72 \geq 10$

$n(1-\hat{p}) = 1016 \times 0.83 = 843 \geq 10$



- What is the margin of error in the 99% confidence interval?

$z^* \times SE = 2.576 \sqrt{\frac{.17 \times .83}{1016}} \approx 0.03$

- If the Gallup Poll wants the estimate to be accurate to within ±1%, how large a sample should they use?

given MOE $= 0.01$

$2.576 \sqrt{\frac{.17 \times .83}{n}} = 0.01$

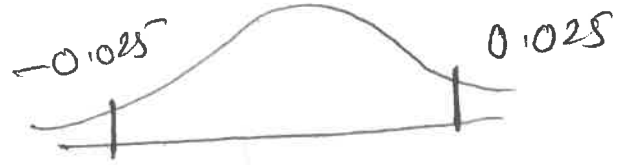
$\Rightarrow \sqrt{\frac{.17 \times .83}{n}} = 0.01 / 2.576$

$\Rightarrow \frac{.17 \times .83}{n} = .000016$

$\Rightarrow n = \frac{.17 \times .83}{.000016}$

$= 8819$ (roughly)

$z^* = 1.96$



Example. A coffee company wants to estimate the true proportion of U.S. adults who drink its brand. How many individuals should be surveyed to be 95% confident of having the true proportion of U.S. adults drinking the brand to a margin of error of .018? $(\hat{p} = 0.5)$

(again same routine)

$z^* SE = 0.018$

$\Rightarrow 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.018$

$\Rightarrow 1.96 \sqrt{\frac{0.5 \times 0.5}{n}} = 0.018$

$\Rightarrow \frac{0.25}{n} = \left(\frac{0.018}{1.96}\right)^2$

$\Rightarrow 0.25/n = .000081$

$\Rightarrow n = .25/.000081$

$= 3086.42$

≈ 3087 individuals

Hypothesis Test for a Single Proportion

For hypothesis testing, we want the distribution of the sample proportions **assuming the null hypothesis is true**. This null hypothesis will be of the form $\cancel{H_0: p = p_0}$. Recall that for hypothesis testing, we need to find the test statistic

$$z = (\text{sample statistic} - \text{null value}) / SE$$

Thus we will again need to find the standard error. Recall that for confidence intervals, we use the standard deviation of a bootstrap distribution as the standard error, while for hypothesis testing we use the standard deviation of a randomization distribution as the standard error. Thus the standard error in a hypothesis test may not be the same as the standard error in a confidence interval.

In conducting a hypothesis test, we want to see if the sample proportion $\hat{p}$ is in an unusual place of a distribution we would expect to see **when H_0 is true**. Since we assume H_0 is true, we use p_0 in place of p for computing the standard error. We have

$$SE = \sqrt{\frac{p_0(1-p_0)}{n}}$$

$H_0: p = p_0$

To test _____ against any alternative hypothesis, we use the standard test statistic

$$Z = \frac{(\text{sample-statistic}) - \text{nullvalue}}{SE} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where $\hat{p}$ is the proportion in a random sample of size n. If the sample size n is large enough that $np_0 \geq 10$ and $n(1-p_0) \geq 10$, the p-value of the test is computed using the standard normal distribution.

$\hat{p} = .549$

Example. Of the 2430 Major League Baseball games played in 2009, the home team won in 54.9% of the games. If we consider the games played in 2009 as a representative sample of all MLB games, is this evidence of a home field advantage in Major League Baseball?

Keyword

$n = 2430$

let p denote the proportion of all MLB games won by the home team. $H_0: p = .5$ & $H_a: p > 0.5$

$$Z = \frac{.549 - .5}{\sqrt{\frac{.5(1-.5)}{2430}}} = \frac{0.049}{\sqrt{\frac{.25}{2430}}} \approx 4.831$$

check $np_0 = 2430(.5) = 1215 \geq 10$
$n(1-p_0) = 2430(.5) = 1215 \geq 10$

[Refer page 1 § 5.2 in lecture notes for this checking]

$p = P(Z > 4.831) = 1 - P(Z < 4.831) = 1 - 1 = 0$

and as $p < \alpha$ $(0 < 0.05)$ $\rightarrow$ $\boxed{\text{REJECT } H_0}$

YES, we have statistical evidence of a home field advantage in MLB

($\alpha = 0.05$ significant level).

$\hat{p} = 110/300$
$= 0.367$

Example. A survey reported that 35% of the U.S. population ages 18 to 74 years of age take vitamin and mineral supplements regularly. Later, an investigator interviewed 300 randomly selected U.S. individuals in that age group and found that 110 took supplements. Do the investigator's findings provide significant evidence to reject $H_0 : p = 0.35$ in favor of $H_a : p > 0.35$. Use $\alpha = 0.025$.

$H_0 : p = 0.35$ $\quad z = (\text{sample statistic} - \text{null value}) / SE$

$H_a : p > 0.35$ (given) $\quad (p_0 = 0.35)$

$$SE = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.35 \times 0.65}{300}} = 0.00275.$$

$$z = \frac{0.367 - 0.35}{0.0275} = 0.61818$$

<u>check</u> $np_0 = 105 \geq 10$; $n(1-p_0) = 195 \geq 10.$

$$\underbrace{P(Z > 0.61818)}_{=p} = 1 - P(Z < \overbrace{0.61818}^{0.62})$$
$$= 1 - 0.7324 = 0.2676 > \underbrace{0.025}_{\alpha}$$

$\therefore p > \alpha$ we fail to REJECT H_0.

The investigator's findings do not provide significant evidence to REJECT $H_0 : p = 0.35$

$p_0 = .25$ & $\hat{p} = 42/200 = .21$

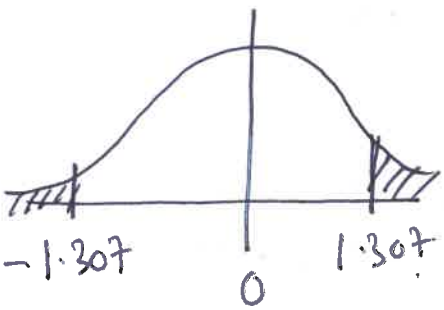
Example. A magazine claims that 25% of its readers are college students. A random sample of 200 readers is taken. It is found that 42 of these readers are college students. Use a 0.10 level of significance to test $H_0 : p = 0.25$ against $H_a : p \neq 0.25$.

$$\left.\begin{array}{l} H_0 : p = 0.25 \\ H_a : p \neq 0.25 \end{array}\right] \text{given}$$

$$z = \frac{0.21 - 0.25}{0.0306} = \frac{-0.04}{0.0306} = -1.307$$

$$SE = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{.25 \times .75}{200}} = 0.0306.$$

Check $np_0 = 200 \times .25 = 50 \geq 10$ and
$n(1-p_0) = 200 \times .75 = 150 \geq 10$.



$$p = P(z < -1.307) + P(z > 1.307)$$
$$= 2P(z < 1.307)$$
$$= 2 \times .0951 = 0.1902 > \alpha \ (0.10)$$

$\therefore p > \alpha$, we fail to REJECT $H_0 : p = 0.25$.

Himanshu Singh

In this section, we consider the distribution of the difference in means between samples taken from two distinct groups. The parameter of interest is $\mu_1 - \mu_2$, where μ_1 and μ_2 represent the "true" means in each of the two groups.

Distribution for a Difference in Two Sample Means

When choosing random samples of size n_1 and n_2 from populations with ~~proportions~~ means μ_1 and μ_2, respectively, the distribution of the differences in the sample means, $\bar{x}_1 - \bar{x}_2$,

- is centered at difference in populations means $\mu_1 - \mu_2$,
- and has standard error estimated by $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

$s_1 \rightarrow$ SD of sample population 1

$s_2 \rightarrow$ SD ..... same but for population 2

The standardized differences in sample means follow a t-distribution with degreess of freedom approximately equal to

$\hookrightarrow$ minimum of $n_1 - 1$ and $n_2 - 1$.

If $n_1 < 30$ or $n_2 < 30$, the t-distribution is only a good approximation if underlying population has a distribution that looks like gaussian / Normal.

Example. Suppose samples of size 100 are drawn from Population A, which has mean 87 and standard deviation 12, and samples of size 80 are drawn from Population B, which has mean 81 and standard deviation 15. Find the standard error of the distribution of differences in sample means, $\bar{x}_1 - \bar{x}_2$.

population A $\longrightarrow \mu_A = 87 \quad \sigma_A = 12 \quad n_A = 100$

population B $\longrightarrow \mu_B = 81 \quad \sigma_B = 15 \quad n_B = 80$

$$SE = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} = \sqrt{\frac{12^2}{100} + \frac{15^2}{\cancel{100}\ 80}}$$

Confidence Interval for a Difference in Means

A difference in sample means based on random samples of sizes n_1 and n_2 has

- sample statistic $\bar{X}_1 - \bar{X}_2$
- and standard error $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

The formula for a confidence interval is (sample statistic) $\pm t^* SE$

which in this case corresponds to $\bar{X}_1 - \bar{X}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Example. At a particular university, 210 first year college students were randomly assigned roommates. Among the 78 students assigned to roommates who brought a video game to college, average GPA after the first semester was 2.84 with a standard deviation of 0.669. Among the 132 students assigned to roommates who did not bring a video game to college, average GPA after the first semester was 3.105 with a standard deviation of 0.625.

How much does getting assigned a roommate who brings a video game to college affect your first semester GPA? Give a 90% confidence interval.

$\mu_V \rightarrow$ 1st sem GPA of all students whose roommate brings video game to college.

$\mu_N \rightarrow$ ..... same but NOT bring video game....

Sample V $n_V = 78$, $\bar{X}_V = 2.84$, $s_V = 0.669$

Sample N $n_N = 132$, $\bar{X}_n = 3.105$, $s_n = 0.625$

$\bar{X}_N - \bar{X}_V = 3.105 - 2.84 = 2.65$.

$SE = \sqrt{\frac{(.669)^2}{78} + \frac{(.625)^2}{132}} \approx 0.0933$. degree of freedom $=$ min $(78-1, 132-1) = 77$

$\Rightarrow t^* = 1.665$.

90% CI $= .265 \pm 1.665(0.0933) = (.1097, .4203)$

(statement) $\rightarrow$ DIY.

Hypothesis Test for a Difference in Means

Our null hypothesis in these tests will be

$H_0: \mu_1 = \mu_2$

The test statistic is $t = \dfrac{(\text{sample statistic}) - \text{null value}}{SE}$

- the statistic is $\bar{X}_1 - \bar{X}_2$
- the null value is 0
- and the standard error is $SE = \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$

If the underlying populations are approximately normal or the sample sizes n_1 and n_2 are greater than 30, we can use a t-distribution to find the p-value for this statistic. The correct number to use for degrees of freedom is

minimum of $n-1$ and $n_2 - 1$

Example. The Pygmalion Effect states that the greater the expectation placed upon people, the better they perform. Teachers were told that certain children, chosen randomly, were expected to be "growth spurters," based on the Harvard Test of Inflected Acquisition, a test that does not actually exist. The response variable is change in IQ over the course of one year. Results are given in the chart below.

	n	$\bar{x}$	s
Control Students	255	8.42	12.0
Growth Spurters	65	12.22	13.3

Does this provide evidence that the Pymalion Effect exists, that merely expecting a child to do better actually *causes* the child to do better?

$\mu_G \rightarrow$ mean change in IQ over 1 year $\forall$ growth spurters

$\mu_C \rightarrow$ . . . . . $\forall$ all control students

$H_0: \mu_G = \mu_C$

$H_a: \mu_G > \mu_C$

$t = \dfrac{\text{sample statistic} - \text{null value}}{SE}$

$$= \frac{(\bar{X}_G - \bar{X}_C) - 0}{\sqrt{\frac{13.3^2}{65} + \frac{12.0^2}{255}}} \approx \frac{12.22 - 8.42}{1.8128}$$

$t \approx 2.0962$ $\quad df = 64$

$$\underbrace{p = P(T > 2.0962)}$$

$$.025 < p < 0.01 < \underbrace{0.05}_{= \alpha}$$

$\because p < \alpha \rightarrow$ REJECT $H_0: \mu_G = \mu_C$. Yes, this provide evidence supporting the existence of the Pygmalion effect.

HIMANSHU SINGH

Math 1342 **Inference for a Mean** **6.2**

For quantitative data, the parameter of interest is often the population mean μ. The Central Limit Theorem says that sample statistics often follow a normal distribution **if the sample size is large**. This is true for the distribution of sample means $\bar{x}$, which will be centered at the population mean μ. Besides the center, the other important quantity we need to describe a normal distribution is the standard deviation, the standard error of the sample means. As we have seen, one way to estimate the standard error is to use a bootstrap distribution or a randomization distribution. Statistical theory gives us an alternate method for finding the standard error. When choosing random samples of size n from a population with mean μ and standard deviation σ, the standard error of the sample means is $= \sigma/\sqrt{n}$

This approximation works well when $n \geq 30$.

Confidence Interval for a Single Mean When σ Is Known

The confidence interval for a single mean when σ is known is statistic $\pm z^* SE$

$$= \bar{x} \pm z^* SE. = \bar{x} \pm z^* (\sigma/\sqrt{n}).$$

Example. Ten vehicles were observed at random for their speed in miles per hour. The average speed was $\bar{x} = 58.2$ miles per hour. We may assume from previous studies that vehicle speed is normally distributed with $\sigma = 3.2$ miles per hour. Construct a 90% confidence interval for the average speed of all vehicles.

$n=10$ $\bar{x} = 58.2$ $\sigma = 3.2$ $z^* = 1.645.$

$$90\% \text{ CI} = 58.2 \pm 1.645 \cdot \left(\frac{3.2}{\sqrt{10}}\right)$$

$$= (56.54, 59.86)$$

98% → z^*score = 2.33

Example. Dr. Vivien McKnight studied 50 successful pregnancies involving natural birth and found that the mean pregnancy term was 274 days. From previous studies, the standard deviation σ of a pregnancy term is known to be 14 days. Construct a 98% confidence interval for μ, the true mean term for successful pregnancies involving natural birth. $n=50$, $\bar{x}=274$, $\sigma=14$

$$\text{statistic} \pm z^* SE = 274 \pm 2.33\left(\frac{14}{\sqrt{50}}\right)$$

$$= \underbrace{(269.39, 278.61)}_{I}.$$

we are 98% confident that μ, the true mean term for successful pregnancies involving natural birth is in the I.

Hypothesis Test for a Single Mean When σ Is Known

Suppose the null hypothesis is $H_0: \mu = \mu_0$. If the population is approximately normal or if $n \geq 30$, the p-value can be computed as the area in the tail(s) beyond z of a standard normal distribution, where z is the test statistic

$$z = \frac{(\text{sample statistic}) - (\text{null value})}{SE}$$

$$= (\bar{x} - \mu)/(\sigma/\sqrt{n})$$

Example. An educator feels that the mean SAT scores of students in his large city are significantly less than 500 and that he can establish this based on sample data. Using national data, he makes the assumption that the data for his city will follow a normal curve and that $\sigma = 70$. He takes a sample of 15 students. The mean score is 455. Conduct a hypothesis test using a 5% significance level.

$n=15$, $\sigma=70$, $\bar{x}=455$, $\alpha=0.05$

$H_0: \mu = 500$

$H_a: \mu < 500$

$$z = \frac{455-500}{\left(70/\sqrt{15}\right)} = -45 \Big/ \frac{70}{\sqrt{15}} = -2.49$$

$$p = P(z < -2.49) = .0064 < 0.05\ (\alpha)$$

zero

$\therefore p < \alpha$ we REJECT $H_0: \mu = 500$.

Example. A machine can be adjusted so that when under control, the mean amount of sugar filled in a bag is 5 pounds. From past experience, the standard deviation of the amount filled is known to be 0.15 pounds. To check if the machine is under control, a random sample of 16 bags was weighed and the mean weight was found to be 5.1 pounds. At a 5% significance level, is the adjustment out of control? Assume a normal distribution of sugar filled in a bag.

$\alpha = 0.05$

$\sigma = 0.15, \quad n = 16, \quad \bar{x} = 5.1$

$H_0 : \mu = 5$

$H_a : \mu \neq 5$ (not equal)

Compute Z-Value $= \dfrac{5.1 - 5}{\left(\dfrac{0.15}{\sqrt{16}}\right)} = \dfrac{0.1}{\left(\dfrac{0.15}{\sqrt{16}}\right)} = 2.67.$

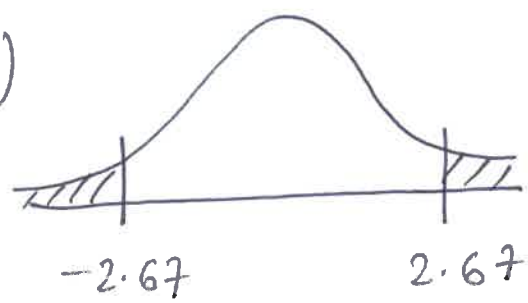
compute p-value.

$p = P(Z > 2.67) + P(Z < -2.67)$

$= 2P(Z < -2.67)$

$= 2(0.0038) = 0.0076$

$< 0.05 \ (\alpha)$



two-tail summation

$\therefore p < \alpha$, REJECT $H_0 : \mu = 5$ and conclude that at 5% significant level, machine is out of control.

QUESTION: what we do if we don't know σ?

When σ is unknown, we can no longer use the formula $SE = \frac{\sigma}{\sqrt{n}}$.

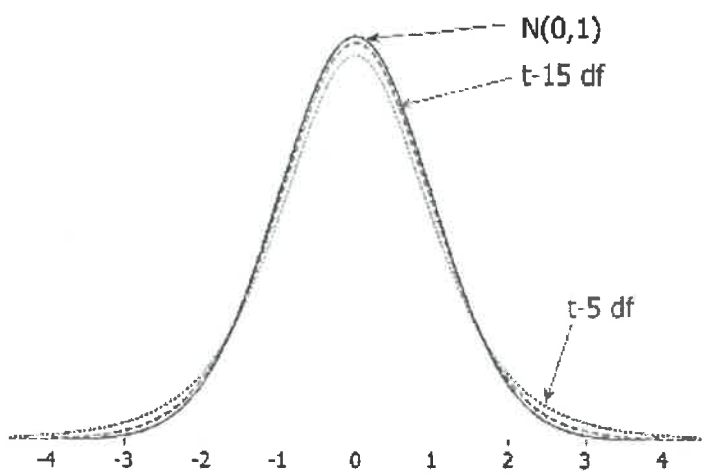
Good news: use 's' sample SD.

Bad news: can't use $N(0,1)$.

Good news: we have a distribution similar to normal distribution, $\rightarrow$ student's t-distribution.

Properties of the t-distribution

- The t-distribution depends on sample size. Technically, it is characterized by the degree of freedom, df, where df = $n-1$.
- The t-distribution is slightly shorter and fatter than the z-distribution.
- As the degrees of freedom increases, the t-distribution approaches the $N(0,1)$.
- The t-distribution is centered at 0.

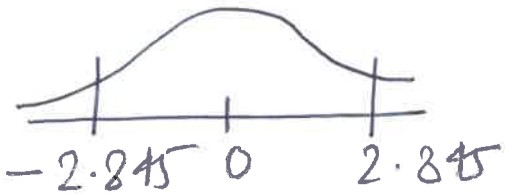


α

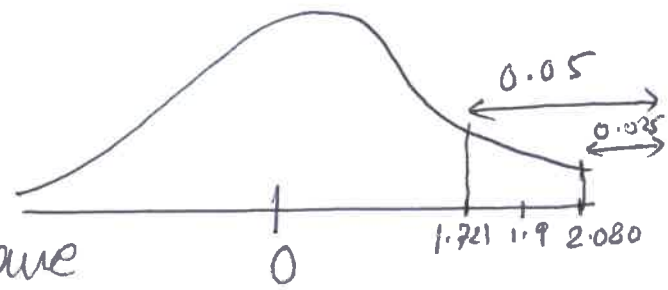
df	0.1	0.05	0.025	0.01	0.005	0.001
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527

- If df = 20, find $P(T > 2.845)$. $P(T > 2.845) = 0.005$

- If df = 20, find $P(T < -2.845)$. $= 0.005$



- If df = 20, find c such that $P(T > c) = 0.1$. 1.325



- If df = 21, find $P(T > 1.9)$. Is this number less than 0.01?

from the figure, we have

$0.025 < P(T > 1.9) < 0.05$

$0.01 <$

NO, $P(T > 1.9)$ is not less than 0.01

We will be using exactly the same equations that we did when studying means with σ known, but everywhere we had a σ, we will substitute a s, and everywhere we had a z^*, we will substitute a t^*.

Confidence Interval for a Single Mean When σ Is Unknown

The confidence interval for a single mean when σ is unknown is

$$\text{statistic} \pm t^* SE$$

$$\bar{x} \pm t^* (s/\sqrt{n})$$

Hypothesis Test for a Single Mean When σ Is unknown

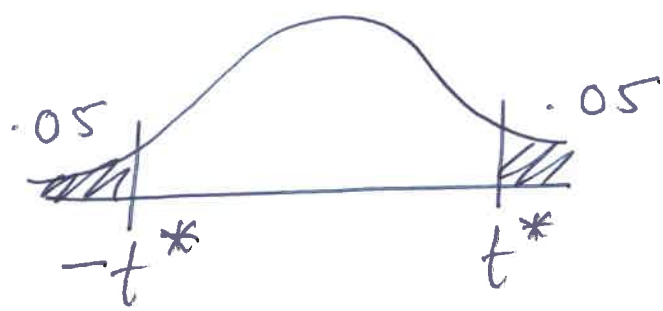
Suppose the null hypothesis is $H_0: \mu = \mu_0$. If the population is approximately normal or if $n \geq 30$, the p-value can be computed as the area in the tail(s) beyond t of a t-distribution with $n-1$ df, where t is the test statistic

$$t = \frac{(\text{sample statistic}) - (\text{nullvalue})}{SE} = \frac{\bar{x} - \mu_0}{(s/\sqrt{n})}$$

Example. Gribbles are small marine worms that bore through wood, and the enzyme they secrete may allows us to turn inedible wood and plant waste into biofuel. A sample of 50 gribbles finds an average length of 3.1 mm with a standard deviation of 0.72 mm. Give a 90% confidence interval for the average length of gribbles.

$n = 50, \quad \bar{x} = 3.1, \quad s = .72$

$t^* = 1.677$



$$\bar{x} \pm t^* (s/\sqrt{n})$$

$$= 3.1 \pm 1.677 \left(\frac{0.72}{\sqrt{50}}\right)$$

$$= (2.9292, 3.2708)$$

Example. The makers of Chips Ahoy! cookies claim that every bag contains 1000 chips. A group of Air Force cadets bought bags of Chips Ahoy! cookies from all over the country to verify this claim. They hand counted the number of chips in 42 bags. They found an average of 1261.6 chips per bag with a standard deviation of 117.6 chips. Consider $H_a : \mu > 1000$ and conduct a hypothesis test using a significance level of 0.01.

$n = 42, \quad \bar{x} = 1261.6 \quad s = 117.6 \quad \alpha = 0.01$

$H_0: \mu = 1000$

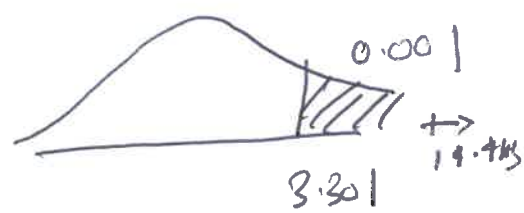
$H_a: \mu > 1000$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1261.6 - 1000}{\frac{117.6}{\sqrt{42}}}$$

$$= 14.4163$$

$p = P(T > 14.4163) < 0.001 < 0.01 = \alpha$

$\therefore p < \alpha$, REJECT $H_0: \mu = 1000$



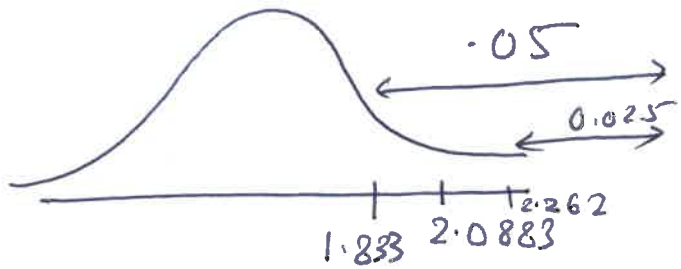
$n=10,\ \bar{X}=10.27,\ s=0.106\quad \alpha=0.05$

set $\rightarrow H_0: \mu=10.2$ & $H_a: \mu\neq 10.2$

Example. The total soluble sugar in a brand of canned soft drink was determined through high performance liquid chromatography on 10 cans of soda. The mean value of the soluble sugar was 10.27 g/100mL with a standard deviation of 0.106g/100mL. Is the true mean soluble sugar significantly different from 10.2g/100mL? Test using $\alpha = 0.05$ and assume a normal population distribution.

$$t=\frac{\bar{X}-\mu_0}{(s/\sqrt{n})}=\frac{10.27-10.2}{\left[\frac{0.106}{\sqrt{10}}\right]}=2.0883.$$

$$p=P(T<-2.0883)+P(T>2.0883)$$
$$=2P(T>2.0883)$$



as

$0.025<P(T<2.0883)<0.05$

↓ multiply by 2

$\Rightarrow$

$$0.05<\underbrace{2P(T<2.0883)}_{p}<0.1$$

$\because p>\alpha$ as $\alpha=0.05$

we fail to REJECT $H_0: \mu=10.2$.

No, the true mean soluble sugar is not significantly different from 10.2g/100ml.

Himanshu Singh

In this section, we consider the distribution of the difference in proportions between samples taken from two distinct groups. The parameter of interest is $p_1 - p_2$, where p_1 and p_2 represent the proportions in each of the two groups.

Distribution for a Difference in Two Sample Proportions

When choosing random samples of size n_1 and n_2 from populations with proportions p_1 and p_2, respectively, the distribution of the difference in the sample proportions, $\hat{p}_1 - \hat{p}_2$,

- is centered at difference in the population proportions $p_1 - p_2$,
- has standard error given by $SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ —— (1)
- and is reasonably normally distributed if $n_1 p_1 \geq 10$, $n_1(1-p_1) \geq 10$, $n_2 p_2 \geq 10$, and $n_2(1-p_2) \geq 10$.

Example. Suppose samples of size 300 are drawn from Population A, which has proportion 0.15, and samples of size 200 are drawn from Population B, which has proportion 0.20.

- Find the standard error of the distribution of differences in the sample proportion, $\hat{p}_A - \hat{p}_B$.

Recall (1) $SE = \sqrt{\frac{.15(.85)}{300} + \frac{.2(.8)}{200}}$

- Determine whether the sample sizes are large enough for the Central Limit Theorem to apply.

$n_A p_A = 300(.15) = 45 \geq 10$ $n_B p_B = 200(.2) = 40 \geq 10$

$n_A(1-p_A) = 300 \times .85 \geq 10$ and similarly $n_B(1-p_B) = 200(.8) = 160 \geq 10$

yes, sample sizes are large enough

Confidence Interval for a Difference in Proportions

Recall from Section 5.2 that when a distribution of a sample statistic is normally distributed, a confidence interval can be found using

sample statistic $\pm$ z^* SE.

We just saw that the standard error for the distribution of the difference in sample proportions, $p_1 - p_2$ is

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

This works great if we know p_1 and p_2, but in practice, we will not know these. Just as we did for the standard error of a single proportion, we substitute the ~~sample proportion~~ to estimate the standard error for the difference in proportions.

If the sample sizes n_1 and n_2 are large enough that $n_1 p_1 \geq 10$, $n_1(1-p_1) \geq 10$, $n_2 p_2 \geq 10$, and $n_2(1-p_2) \geq 10$, then a confidence interval for $p_1 - p_2$ can be computed by

here $(\hat{p}_1 - \hat{p}_2) \pm z^* SE$ where $SE =$

$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Example. How detrimental are metal tags to penguins? A study looked at the 10 year survival rate of penguins tagged either with a metal tag or an electronic tag. At the 10 year mark, 20% of the 167 metal tagged penguins survived, compared to 36% of the 189 electronic tagged penguins.

- What are the two populations of interest?

metal tags

electronic tags.

proportions be denoted by P_M

proportions be denoted by P_E

check $n_E p_E = 189(.36) = 68.04 \geq 10$; $n_E(1-p_E) = 189(1-.36) = 121$; $n_M p_M = 167 \times (.2) = 33.4 \geq 10$

$n_M(1-p_M) = 167 \times .8 = 133.6 \geq 10$

- Give a 90% confidence interval for the difference in proportions.

$z^* = 1.64$

sample statistic $\pm z^* SE$

$\hat{p}_E - \hat{p}_M = .36 - 0.2 = 0.16$ and

$$SE = \sqrt{\frac{\hat{p}_E(1-\hat{p}_E)}{n_E} + \frac{\hat{p}_M(1-\hat{p}_M)}{n_M}} = \sqrt{\frac{.36(.64)}{189} + \frac{.2(.8)}{167}} \approx 0.0467$$

$.16 \pm 1.64(0.0467) \rightarrow (.083, .237)$ we are 90% confident that p_E is between 0.083 & 0.237 higher than p_M

Hypothesis Test for a Difference in Proportions

In Section 5.1, we see that, when a randomization distribution is normal, we can compute a p-value using a standard normal curve and a standardized test statistic of the form

$$z = \frac{(\text{sample statistic} - \text{null value})}{SE}.$$

The sample statistic is computed from the sample data, and the null parameter is specified by the null hypothesis H_0. When comparing proportions between two groups, the null hypothesis is typically $H_0: p_1 = p_2$, or, equivalently, $H_0: p_1 - p_2 = 0$. Thus the null parameter is often equal to 0, and we use $\hat{p}_1 - \hat{p}_2$ as the sample statistic. This gives

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{SE}$$

From the Central Limit Theorem for a difference in proportions, we know that the standard error has the form

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

What should we use for p_1 and p_2? In Section 6.1, when doing a hypothesis test for a single proportion, we solved this issue by substituting p_0, the proportion specified by the null hypothesis. When doing a hypothesis test for a difference in proportions, however, the null hypothesis only specifies that the two proportions are equal, but not the value to which they are equal. We CANNOT just substitute $\hat{p}_1$ and $\hat{p}_2$ from the samples because we are assuming the null hypothesis is true, and that says the proportions must be equal.

If the null hypothesis is really true, then the best way to estimate the common proportion is to combine the two samples into one big sample and find its proportion. That is precisely what we do to compute what is known as a pooled proportion, denoted $\hat{p}$. We combine the two groups into one big combined group with sample size $n_1 + n_2$ and find the sample proportion $\hat{p}$ for that large group.

Recall that the standardized test statistic is

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{SE}$$

With $\hat{p}$ denoting the pooled proportion, the standard error is

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}$$

Example. Two vaccines to prevent the same infection are being compared. Vaccine A was given to 200 people, and 80 of those people became infected. Vaccine B was given to 300 people, and 105 of those people became infected. Does the effectiveness of the vaccines differ? Use a significance level of 10%.

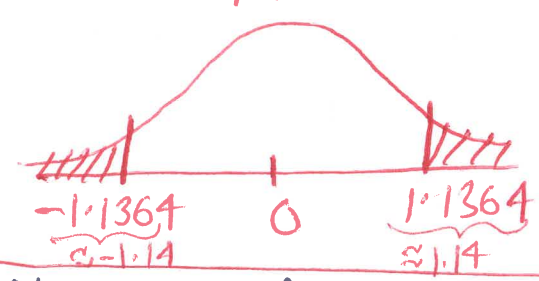
$p_A \rightarrow$ proportion of all people with vaccine A and p_B be the same but for Vaccine B.

$H_0: p_A = p_B$ $\quad \hat{p}_A = 80/200 = .4$ & $\hat{p}_B = \frac{105}{300} = .35$

$H_a: p_A \neq p_B$ $\quad$ pooled proportion $\hat{p} = \frac{80+105}{200+305} = \frac{185}{500} = .37$

$$z = \frac{(\hat{p}_A - \hat{p}_B) - 0}{SE} = \frac{(.4 - .35) - 0}{.044} = \frac{0.05}{.044} = 1.1364$$

$$SE = \sqrt{\frac{.37 \times .63}{200} + \frac{.37 \times .63}{300}} = 0.044$$



$p = P(Z < -1.14) + P(Z > 1.14)$
$= 2P(Z < -1.14)$
$= 2(.1271) = 0.2542 > .1 = \alpha$

$\because p > \alpha$, we fail to REJECT $H_0: p_A = p_B$

we do not have statistically significant evidence that the effectiveness of vaccines differs.

Himanshu Singh

Math 1342 **Paired Difference in Means** **6.5**

In Section 6.4, we considered inference for a difference in means when the data consist of two separate samples. However, in situations such as a matched pairs experiment, the data being compared consist of pairs of data values. Paired data may consist of two measurements on each unit, such as the same units being measured under two different conditions, or measurements on a pair of units that go together, such as twin studies. In this section we see how to handle inferences for a difference in means when the data are paired in some way.

Definition. Data is Paired if the data being compared consists of data values that pair naturally in some way.

Common examples of paired data

- two measurements of same person
- identical twins
- very carefully matched pairs

The key to analyzing paired data is to work with paired differences rather than original data. This helps us to eliminate the variability across different units.

To estimate the difference in means based on paired data, we first subtract to compute the difference for each data pair and compute the mean $\bar{x}_d$, the standard deviation s_d, and the sample size n_d for the sample differences.

Provided the differences are reasonably normally distributed or sample size is large, a confidence interval for the difference in means is given by

statistic $\pm\, t^* SE$ $\qquad \bar{x}_d \pm t^* \left(\frac{s_d}{\sqrt{n}}\right)$

where t^* is a percentile from a t-distribution with $n_d - 1$ degrees of freedom. To test $H_0: \mu_d = 0$ against an alternative hypothesis, we use the t-test statistic

$t = [(\text{sample statistic}) - (\text{null value})] / SE.$

$= (\bar{x}_d - 0) / (s_d / \sqrt{n_d})$

If the differences are reasonably normally distributed or sample size is large, we use a t-distribution with $n_d - 1$ degrees of freedom to compute a p-value for this test statistic.

Example. Scores of 9 students on an exam before and after a special coaching program are given in the chart below.

Student	Before	After	Difference
1	91	82	−9
2	78	80	2
3	47	62	15
4	37	49	12
5	64	55	−9
6	54	73	19
7	43	59	16
8	33	58	25
9	53	43	−10

sum of diff. $= -9 + 2 + \dots + 25 - 10$

5% left

$(\sqrt{9} = 3)$

Find and interpret a 95% confidence interval for the mean increase in exam score after the special coaching program. Assume the differences in means are normally distributed.

$\bar{x}_d = \frac{\text{sum of diff}}{9} = 6.78$ $\qquad s_d = 13.52$

$\bar{x}_d \pm t^* \left(\frac{s_d}{\sqrt{n_d}}\right) = 6.78 \pm 2.306 \left(\frac{13.52}{3}\right)$

$df = 9 - 1 = 8.$

$t^* = 2.306$

Example. Determine whether student scores are higher after the special coaching program. Use a 5% significance level.

$H_0: \mu_d = 0$

$H_a: \mu_d > 0$

$$t = \frac{\bar{X}_d - 0}{\left[\frac{s_d}{\sqrt{n_d}}\right]} = \frac{6.70 - 0}{\left[\frac{13.52}{\sqrt{9}}\right]} = \frac{6.70}{\frac{13.52}{3}} \approx 1.5044$$

also $df = 9 - 1 = 8$

$p = P(T > 1.5044)$ as $\alpha = 0.05$ and

$\alpha < p < 0.1 \Rightarrow$ we fail to REJECT $H_0: \mu_d = 0$

we have no statistically significant evidence that mean exam score is greater after the special coaching program.

Himanshu Singh

Definition. A process is random if its outcome is uncertain.

Definition. The probability of an event is the long run frequency or proportion of times the event occurs.

Like proportions, probabilities are always between 0 and 1.

Definition. A combination of one or more outcomes is an event.

When outcomes are equally likely, we have the following formula for the probability of an event, denoted _____.

$$P(\text{event}) = \frac{\#\text{ outcomes in event}}{\text{total }\#\text{ outcomes in event}}$$

Example. Suppose that we start with a standard deck of 52 playing cards. Each card shows one of four suits (spades, hearts, diamonds, or clubs) and one of 13 denominations (ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, jack, queen, king). We shuffle the cards and draw one at random. Determine the probability of each of the following.

- The card is an ace $P(\text{ace}) = 4/52 = 1/13$
- The card is a club $P(\text{club}) = 13/52 = 1/4$
- The card is the 7 of diamonds $P(\text{7 of diamonds}) = 1/52$
- The card is a face card (jack, queen, king) of any suit $P(\text{face card}) = 12/52 = 6/13.$

Example. From its founding through 2016, the Hockey Hall of Fame has inducted 268 players. The table below shows the number of players by place of birth and position played. If a player is chosen at random, from all player inductees into the Hockey Hall of Fame, let C represent the event of being born in Canada and D represent the event of being a defenseman. Write each of the following questions as a probability expression and find the probability.

	Offense	Defense	Goal	Total
Canada	126	74	33	233
USA	8	5	1	14
Europe	10	5	3	18
Other	2	1	0	3
Total	146	85	37	268

- What is the probability that an inductee chosen at random is Canadian?

P(candain) = 233/268

- What is the probability that an inductee chosen at random is not a defenseman?

P(not D) = P(offense) + P(goal)
= 146/268 + 37/268

- What is the probability that an inductee chosen at random is a defenseman born in Canada?

P(C & D) = 74/268

- What is the probability that an inductee chosen at random is either born in Canada or a defenseman?

P(C or D) = P(C) + P(D) − P(C and D)
= 233/268 + 85/268 − 74/268

- What is the probability that a Canadian inductee plays defense?

P(D if C) = 74/233

- What is the probability that an inductee who plays defense is Canadian?

P(C if D) = 74/85

Definition. The complement rule states that for any event A,

$$P(\text{not } A) = 1 - P(A)$$

Example. Returning to the Hockey Hall of Fame example, what is the probability that an inductee chosen at random is not a defenseman?

$$P(\text{not } D) = 1 - P(D) = 1 - \frac{85}{268}$$

Definition. The additive rule, also known as the inclusion/exclusion principle, states that for any events A and B,

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

Example. Suppose that 31% of the children in the newborn class at a particular daycare have an older brother, 37% of the children have an older sister, and 17% have both an older brother and an older sister. Find the probability that a randomly chosen newborn in the class has either an older brother or an older sister.

$$P(B \text{ or } S) = P(B) + P(S) - P(B \text{ and } S)$$
$$= 0.31 + 0.37 - .17$$

Definition. If two events have no outcomes in common, the events are disjoint.

If events A and B are disjoint, then it is impossible for both A and B to occur. Thus $P(A \text{ and } B) = 0$. Then we have

$$P(A \text{ or } B) = P(A) + P(B) \quad \text{for } A \cap B = \phi$$

Definition. The probability of an event occurring when we assume that some other event has also occurred is a conditional probability

In general, to find the conditional probability of an event B occurring if A occurs, we compute the Probability of B if A or probability of B given A, which is

$$P(B \text{ if } A) = \frac{P(B \text{ and } A)}{P(A)} \xrightarrow{\text{symbolically}} \Rightarrow P(B|A) = \frac{P(B \cap A)}{P(A)}$$

Example. Returning to the daycare example, find the probability of each of the following.

- A child in the newborn class who has a brother also has a sister

$$P(S|B) = \frac{P(S \cap B)}{P(B)} = \frac{.17}{.31}$$

- A child in the newborn class has a brother if they have a sister

$$P(B|S) = \frac{P(B \cap S)}{P(S)} = \frac{0.17}{.37}$$

Definition. The multiplicative rule comes from rearranging the formula for a conditional probability. We see that for two events A and B,

$$P(B|A)\,P(A) = P(A \cap B)$$

Example. Suppose we draw two cards without replacement from a standard 52 card deck. What is the probability that both the cards are diamonds?

$$P(D \,\&\, D) = P(D) \cdot P(D \text{ if } D)$$

$$\Rightarrow P(D \cap D) = P(D) \cdot P(D|D)$$

$$= \frac{13}{52} \cdot \frac{12}{51}$$

$P(B|A) = P(B)$ $P(A|B) = P(A)$

Definition. Two events A and B are independent if _____ and _____.

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

and

Example. Suppose that a set of 3 traffic lights along one section of road operate independently, so there is no communication or special timing between the lights. Since this is a fairly main road, the lights are green with a probability of 0.7 and red with a probability of 0.3. As you go through this stretch of road, find the probability of each of the following.

- All 3 lights are green $P(GGG) = P(G) \cdot P(G) \cdot P(G)$
$= (0.7)^3$

- The first 2 lights are green, but the third is red

$P(GGR) = P(G)P(G)P(R)$
$= 0.7^2 \cdot 0.3$

- At least one of the lights is red RRR | RRG | RGR | GGG

$P(\text{atleast one } R) = 1 - P(\text{not atleast one Red})$
$= 1 - P(GGG) = 1 - (.7)^3$

Example. Suppose that 31% of the children in the newborn class at a particular daycare have an older brother, 37% of the children have an older sister, and 17% have both an older brother and an older sister. Let B be the event that a child in the newborn class has a brother and S be the event that a child in the newborn class has a sister.

- Are the events B and S disjoint? B and S are disjoint $\iff P(B \cap S) = 0$

$P(B \cap S) = .17 \neq 0 \Rightarrow B \cap S \neq \phi$

- Are the events B and S independent? B and S are independent $\iff$
$P(B|S) = P(B)$ and
$P(S|B) = P(S)$

$P(B|S) = \frac{P(B \cap S)}{P(S)} = \frac{.17}{.37} \approx 0.459 \neq .31 = P(B)$

thus B and S are not independent

Recall the formula for the conditional probability of an event B occurring if event A occurs

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

cross multiply

which can be rearranged to

$$\Rightarrow P(B \cap A) = P(A) \cdot P(B|A)$$

Definition. For any two events A and B, the *total probability of* B is given by the formula

$$P(B) = P(B \cap A) + P(B \cap A^c)$$
$$= P(B|A)P(A) + P(B \cap A^c)P(A^c)$$

More generally, if A, B, and C are disjoint events which contain all of the outcomes of another event D, then

$$P(D) = P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C)$$

We can extend this in equivalent ways to more than 3 events.

Example. Consider a particular biology course in which 74% of the pre-med students are also taking organic chemistry, while 32% of the non-pre-med students are taking organic chemistry. If 58% of the students in the biology course are pre-med, what is the probability that a student chosen at random from this biology course is also taking organic chemistry?

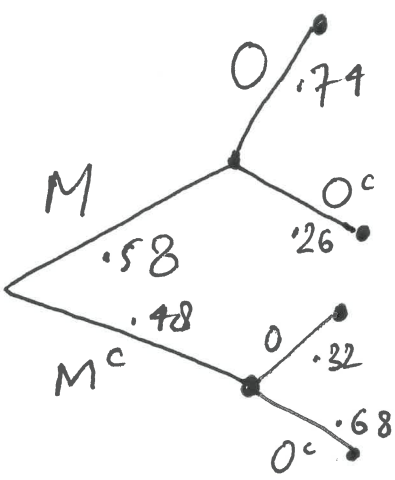
$$P(O) = P(O \text{ and } M) + P(O \text{ and not } M)$$
$$= P(O \text{ and } M) + P(O \text{ and } M^c)$$
$$= P(O|M)P(M) + P(O|M^c)P(M^c)$$
$$= (.74)(.58) + (.32)(1-.58)$$

$M \rightarrow$ biology and pre-med

$O \rightarrow$ biology and Organic chemistry

When the given information is in terms of probabilities for one type of event and conditional probabilities for another, as in in the example above, we can often organize the calculation of the multiplicative rule with a ~~tree diagram~~.

- The initial set of branches shows probabilities from one set of events.
- The second set of branches shows conditional probabilities with initial branch as the condition.
- Multiplying along any set of branches uses the multiplicative rule to find the joint probability fro that pair of events.



$P(M \cap O) = .58 \times .74$

$P(M \cap O^c) = .58 \times .26$

$P(M^c \cap O) = (.42) \times (.32)$

$P(M^c \cap O^c) = (.42) \times (.68)$

Example. The professor of the biology course learns that one of his students is in organic chemistry. What is the probability that the student is pre-med?

$$P(M \mid O) = \frac{P(M \cap O)}{P(O)} = \frac{P(M \cap O)}{P(O \text{ and } M) + P(O \text{ and } M^c)}$$

$$= \frac{(.58)(.74)}{(.58)(.74) + (.42)(.32)}$$

Definition. For any two events A and D, Bayes rule states that

Recall:

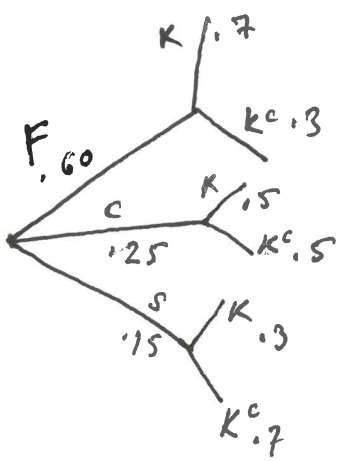
$$P(D|A) = \frac{P(A \cap D)}{P(A)}$$

$$P(A \text{ if } D) = \frac{P(A \cap D)}{P(D)} = \frac{P(D|A)P(A)}{P(D)} = \frac{P(A \cap D)}{P(D \cap A) + P(D \cap A^c)}$$

The last two versions can easily be generalized using the total probability rule to handle more than two events.

$$\frac{P(D|A)P(A)}{P(D|A)P(A) + P(D|A^c)P(A^c)}$$

Example. Slippery Elum is a baseball pitcher who uses three pitches: 60% fastballs, 25% curveballs, and the rest spitballs. Of Slippery's fastballs, 70% are strikes. Of Slippery's curveballs, 50% are strikes. Of Slippery's spitballs, 30% are strikes. If Slippery throws a strike, what is the probability that pitch was a curveball?



$P(F \cap K) = .6 \times .7$

$P(F \cap K^c) = .6 \times .3$

$P(C \cap K) = .25 \times .5$

$P(C \cap K^c) = .25 \times .5$

$P(S \cap K) = .15 \times .3$

$P(S \cap K^c) = .15 \times .7$

$$P(C \text{ if } K) = \frac{P(C \cap K)}{P(K)} = \frac{P(C \cap K)}{P(F \cap K) + P(C \cap K) + P(S \cap K)}$$

$$= \frac{(.25) \times (.5)}{(.6)(.7^3) + (.25)(.5) + (.15)(.3)}$$

Himanshu Singh

Definition. A Random variable denotes a numeric quantity that changes from trial to trial in a random process.

We often use a capital letter to denote a random variable.

Example. $X \longrightarrow$ number showing when a dice is rolled.

$Y \longrightarrow$ number sum when a pair of dice is rolled.

Our aim is to be able to answer probability questions about events determined by a random variable.

Definition. A random variable is DISCRETE if it has a finite set of possible values.

Definition. A random variable is CONTINUOUS if it can take any value within some interval of the real numbers.

For continuous random variables, we use a density curve, where the probability of being in some region is found as the area under the density curve. This is what we did with normal distributions.

For the rest of this section we assume that a random variable is discrete with a relatively small set of possible values.

Definition. A probability function gives the probability for each of its possible values. A probability function assigns a probability, between 0 and 1, to every value of a discrete random variable. The sum of all of these probabilities must be one, that is,

$$\sum p(x) = 1$$

$$\int f(x)\,dx = 1$$

Example. The table below gives the probability function for a random variable X that takes on the values $5, 6, 7, 8$. Verify that $p(x)$ is a probability function.

x	5	6	7	8
$p(x)$	0.1	0.6	0.2	0.1

Recall $\sum p(x) = 1$

here $\sum p(x) = 0.1 + 0.6 + 0.2 + 0.1 = 1$

yes, $p(x)$ is the PROBABILITY FUNCTION.

Example. Suppose that we roll two six-sided dice, and let a random variable X measure the sum of the two rolls.

1. How many possible pairs of rolls are there? $6 \times 6 \longrightarrow 36$

2. How many ways are there to roll an 8? $(3,5)$ $(5,3)$ $(2,6)$ $(6,2)$ $(4,4)$ $\longrightarrow$ 5 ways.

3. Create a table showing the probability function for X.

X	2	3	4	5	6	7	8	9	10	11	12
$p(x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

4. Find $P(X = 3 \text{ or } X = 5)$.

$$P(X=3 \text{ or } X=5) = P(X=3) + P(X=5) \overbrace{- P(X=3 \cap X=5)}^{\text{DISJOINT}}$$
$$= 2/36 + 4/36$$

5. Find $P(X > 4)$.

$$P(X>4) = 1 - P(X \le 4) = 1 - [P(X=2) + P(X=3) + P(X=4)]$$
$$= 1 - \left[\frac{1}{36} + \frac{2}{36} + \frac{3}{36}\right]$$
$$= 1 - \frac{6}{36} = \frac{30}{36}$$

Definition. For a random variable X with probability function $p(x)$, the mean μ is

$$\mu = \sum_x x p(x)$$

- Since this mean is the average value over the "population" of all possible values of X, we use the notation μ to represent the mean.
- In general, we find the mean for a random variable from its probability function by first multiplying each of the possible values by the probability of getting that value and then summing the results.

Definition. The variance of a variable is equal to the square of the standard deviation and is denoted σ^2.

Definition. For a random variable X with probability function $p(x)$ and mean μ, the variance, denoted σ^2, is given by the formula

$$\sigma^2 = \sum_x (x-\mu)^2 p(x)$$

and the standard deviation is σ.

Example. Suppose that the probability function in the table below reflects the possible lifetimes in months after emergence for fruit flies.

x	1	2	3	4	5	6
$p(x)$	0.30	?	0.20	0.15	0.10	0.05

- What proportion fruit flies die in the second month?

$p(2) = 1 - (.3 + .2 + .15 + .1 + .05) = 1 - .8 = .2$

- What is the probability that a fruit fly lives more than four months?

$P(X > 4) = P(X=5) + P(X=6) = 0.1 + .05 = .15$

- What is the mean lifetime for a fruit fly?

$\mu = \sum_x x p(x) = 1 \cdot p(1) + 2p(2) + 3p(3) + 4p(4) + 5p(5) + 6p(6)$

$= 1(.3) + 2(.2) + 3(.2) + 4(.15) + 5(.1) + 6(.05)$

$= .3 + .4 + .6 + .6 + .5 + .3 = .3 + 1 + 1.1 + .3$

$= 2.7$ months.

- What is the standard deviation of fruit fly lifetimes?

Variance $\sigma^2 = \sum_x (x-\mu)^2 p(x)$

$= (1-2.7)^2 p(1) + (2-2.7)^2 p(2) + (3-2.7)^2 p(3) + (4-2.7)^2 p(4) + (5-2.7)^2 p(5) + (6-2.7)^2 p(6)$

take the $\sqrt{\ }$ of variance to have σ. (standard deviation)

HW → 27th Nov 2023 / 1st Dec 2023.

Himanshu Singh

RV → Random Variable

Definition. A Binomial RV counts the number of times that something occurs in a fixed number of independent trials.

- A binomial random variable is a type of discrete RV.
- What the binomial random variable counts and what each trial represents depend on the context.
- On each trial, the outcome that is counted by the random variable is called a success. Any other outcome is called a failure

Conditions for a Binomial Random Variable

For a process to give a binomial random variable, we need the following characteristics:

- A number of trials, n, that is fixed in advance
- A probability of success, p, that does not change from trial to trial
- Independence of the outcomes from trial to trial

A binomial random variable counts the number of success in the n independent trials

Example. In each of the following cases, state whether or not the process describes a binomial random variable. If it is binomial, give the values of n and p.

1. Count the number of times a soccer player scores in five penalty shots against the same goalkeeper, where each shot has a $\frac{1}{3}$ probability of scoring

   binomial with $n=5$ & $p=1/3$
2. Count the number of times a coin lands heads before it lands tails

   not binomial ∵ the number of trials is not fixed in advance
3. Draw 10 cards from the top of a deck and record the number of cards that are aces

   not binomial ∵ p changes from trial to trial
4. Conduct a random sample of 500 registered voters and record whether each voter is Republican, Democrat, or Independent

   not binomial since success is not defined
5. Conduct a random sample of 500 registered voters and count the number that are Republican

   binomial with $n=500$ and $p = \frac{\text{\# registered voters were republican}}{\text{total \# of registered voters}}$
6. Randomly select one registered voter from each state in the US and count the number that are Republican

   not binomial since p changes from trial to trial

Example. Consider the scenario in which we count the number of times a soccer player scores in five penalty shots against the same goalkeeper, where each shot has a $\frac{1}{3}$ probability of scoring. Let X denote the number of times the soccer player scores.

1. What is the probability that the player scores on all 5 kicks?

$P(X=5) = P(SSSSS) = [P(S)]^5 = (1/3)^5.$

2. What is the probability that the player does not score on all 5 kicks?

$P(X=0) = P(FFFFF) = (P(F))^5 = (2/3)^5$

3. What is the probability that the player scores on exactly 4 of the kicks? SSSSF, FSSSS, SFSSS, SSFSS, SSSFS

$P(X=4) = 5 \cdot (1/3)^4 (2/3)^1$

4. What is the probability that the player scores on exactly 1 kick? SFFFF, FSFFF, FFSFF, FFFSF, FFFFS

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

Definition. For any positive integer n, the factorial of n, denoted $n!$, is the product of all positive integers less than or equal to n.

Example. $0! = 1$ (by definition)

$n! = n(n-1)(n-2) \cdots 1$

$1! = 1$

$2! = 2(2-1) = 2$

$3! = 3 \cdot (3-1) \cdot (3-2) = 6$

$4! = 24$

Definition. The binomial coefficient $\binom{n}{k}$, read "n choose k", is given by the formula

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

Example. $$\binom{8}{5} = \frac{8!}{5!\,(8-5)!} = \frac{8!}{5!\cdot 3!} = \frac{8\times 7\times \cancel{6}\times \cancel{5!}}{\cancel{5!}\times \cancel{3!}}$$

$$= 56.$$

Why do we care?

The binomial coefficient $\binom{n}{k}$ gives the number of possible ways to arrange k success among n trials.

Binomial Probabilities

If a random variable X is binomial with n trials and probability of success p, the probability of getting exactly k successes, $P(X = k)$, is given by the formula

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Example. Consider the scenario in which we count the number of times a soccer player scores in five penalty shots against the same goalkeeper, where each shot has a $\frac{1}{3}$ probability of scoring. Let X denote the number of times the soccer player scores. $n=5,\ p=1/3$

1. What is the probability that the player scores on all 5 kicks?

$$P(X=5)=\binom{5}{5}(1/3)^5(2/3)^{5-5}=1\cdot(1/3)^5 1=(1/3)^5$$

2. What is the probability that the player does not score on all 5 kicks?

$$P(X=0)=\binom{5}{0}(1/3)^0(2/3)^5=(2/3)^5$$

3. What is the probability that the player scores on exactly 4 of the kicks?

$$P(X=4)=\binom{5}{4}\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right)^1=5(1/3)^4\left(\frac{2}{3}\right)$$

4. What is the probability that the player scores on exactly 1 kick?

$$P(X=1)=\binom{5}{1}(1/3)^1\left(\frac{2}{3}\right)^4=5(1/3)(2/3)^4$$

Recall the formulas for the mean μ and variance σ^2 of a discrete random variable.

$$\mu=\sum_x xp(x)$$

$$\sigma^2=\sum_x (x-\mu)^2 p(x)$$

Because a binomial random variable is a discrete random variable, these formulas can be used to find the mean μ and variance σ^2 of a binomial random variable. Additionally, because there are n trials and the probability of success is always p, we can simplify these formulas in the case where the discrete random variable is binomial.

Mean, Variance, and Standard Deviation of a Binomial Random Variable

If a random variable X is binomial with n trials and probability of success p, then its mean μ, variance σ^2, and standard deviation σ are given by

$$\mu = np$$
$$\sigma^2 = np(1-p)$$
$$\Rightarrow \sigma = \sqrt{np(1-p)}$$

Example. Consider the scenario in which we count the number of times a soccer player scores in five penalty shots against the same goalkeeper, where each shot has a $\frac{1}{3}$ probability of scoring. Let X denote the number of times the soccer player scores. Find the mean, variance, and standard deviation of X.

$n = 5, \quad p = 1/3$

mean of X, $\mu = 5(1/3) = \frac{5}{3}$

variance of X, $\sigma^2 = 5(1/3)(2/3) = 10/9$

std. dev of X, $\sigma = \sqrt{\frac{10}{9}}$.