Lecture 1: Laying the Foundations + Terminology

Chapters 1.1-1.2

Goals for Today

- Go over the syllabus
- Show some fun examples
- Discuss how to evaluate the efficacy of a treatment
- Describe the different kinds of variables we'll consider

What is statistics?

(Direct from text) The general scientific process of investigation can be summed up as follows:

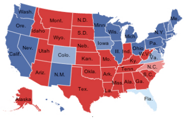
1. Identify the scientific question or problem
2. Collect relevant data on the topic
3. Analyze the data
4. Form a conclusion and communicate it

Statistics concerns itself with points 2 through 4.

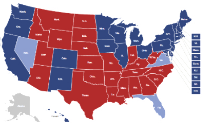
Your Majors

```
                    Biology                  Economics
                         11                          5
                    History      Environmental Studies
                          4                          3
                Mathematics                 Psychology
                          3                          3
Biochem and Molecular Biology                Chemistry
                          2                          2
 International Policy Studies              Linguistics
                          2                          2
                  Undecided               Anthropology
                          2                          1
      Economics/Mathematics Environmental Studies-Hist
                          1                          1
Environmental Studies-Pol Sci                  Physics
                          1                          1
                  Sociology
                          1
```

Example: 2012 Election - Nate Silver's Predictions vs Actual Results



Nate Silver's Map



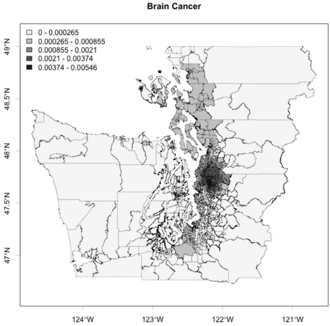
The Actual Map

Example: Brain & Breast Cancer in Western Washington

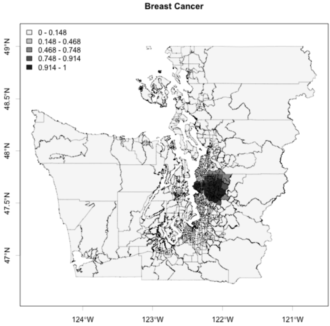
My PhD dissertation involved detecting cancer "clusters": areas of residual spatial variation of disease risk.

We modeled the (Bayesian) probability of cluster membership for each of the $n = 887$ census tracts in Western Washington in 2000, using cancer data from 1995–2005, controlling for age, race, and gender.

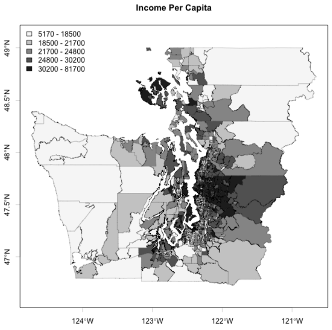
Brain Cancer Controlling for Age, Race, & Gender



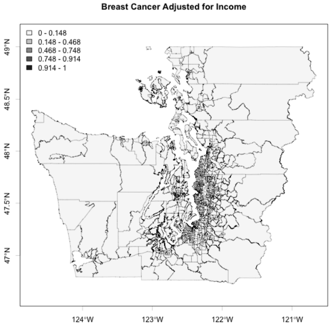
Breast Cancer Controlling for Age, Race, & Gender



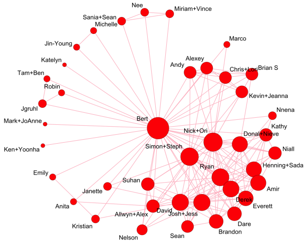
Income per Capita Quintiles



Breast Cancer Adjusted for Income as Well



Example: Social Network Display of a Recent Party I Had



Say we want answer the following questions:

- Does a new kind of cognitive therapy alter levels of depression in patients?
- Or you question the effectiveness of antioxidants in preventing cancer.
- Will reassuring potential new users to a gambling website that we won't spam them increase the sign-up rate?

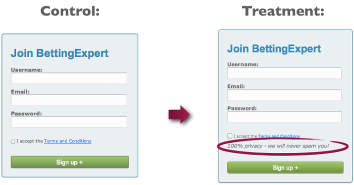
Evaluating the efficacy of a 'treatment'

In all the above cases, you are questioning the efficacy of a treatment/intervention. One way to evaluate the efficacy is via an experiment where you define

- A control group: the "business as usual" baseline group
- A treatment group: the group that receives/is subject to the treatment/intervention

and make comparisons.

Website Experiments



Example of a treatment vs control

Two other examples in the media of late

- Facebook's tinkering with user's emotions (link)
- OkCupid's admission that they experiment on human beings (link)

Variables

A variable is a description of any characteristic whose value may change from one unit in the population to the next:

Data

At its simplest, data are presented in a data table or matrix where (almost always) each

- row corresponds to cases or units of observation/analysis
- column represents the variables corresponding to a particular observation

It is almost always the case that

- n is the number of observations
- p is the number of variables

Data Summaries

Consider the variable "federal spending per capita" in each of the 3,143 counties in the US. One can hardly digest this:

```
   [1]   6.068095   6.139862   8.752158   7.122016   5.130910   9.973062   9.311835  15.439218
   [9]   8.613707   7.104621   6.324061  10.640378   9.781442   8.982702   6.840035  20.330684
  [17]   9.687698  11.080738   7.839761   9.461856   9.650295   7.760627  25.774791  13.948106
   ...
[3121]   7.520731  10.246400   3.106800  17.679572   4.824044   7.247212   8.484211   8.794626
[3129]   9.829593   8.100945  17.090715   4.855849   6.621378  22.587359  10.813260  11.422522
[3137]   9.580265   4.368986   5.062138   6.236968   4.549105   8.713817   6.694784
```

Data Summaries

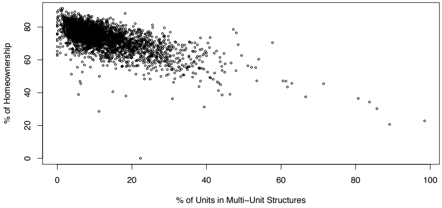
We can't interpret all the data at once; we need to boil it down via summary statistics, single numbers summarizing a large amount of data.

Using the `summary()` command in `R`:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
0.000   6.964   8.669   9.991  10.860 204.600       4
```


Relationships between variables

We can best display the relationship between two variables using a scatterplot AKA bivariate plot:



Relationships between variables

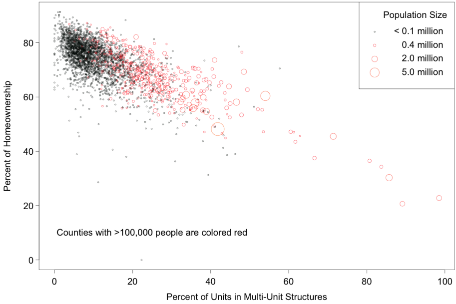
Almost always we are interested in the relationship between two or more variables.

A pair of variables are either related in some way (associated) or not (independent). No pair of variables are both associated and independent.

We can have either a negative association (as the value of one variable increases, the other decreases) or a positive association.

Relationships between variables

We can consider a third variable in the previous plot.



Lecture 2: Sampling and Bias

Chapter 1.3

Goals for Today

- Understand important considerations about data collection, in particular sampling.
- Food for thought about the next lecture: explanatory/response variables and causality.

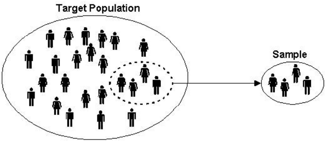
Populations and Samples

We want to make statements about some aspect of a study/target population.

1. What proportion of Oregonians smoke?
2. What are the sexual behaviors of males and female Americans in 1948?
3. What proportion of the Reed community believes they have personally experienced offensive, hostile, or intimidating conduct on campus?

Populations and Samples

It is often not feasible to collect data for every case in the population. If so, we take a sample of cases.



If the sample is representative of the desired population then our results will be generalizabile.

Populations and Samples

So say we take a representative sample of 1000 Oregonians and poll their smoking habits. We can then generalize the results to the entire population of Oregon.

One example of a non-representative sample is a biased sample.

How do we take a representative sample? In its simplest form, you need to randomly sample from the entire population. But this is easier said than done.

Comment on the Representativeness of These Samples:

1. The Royal Air Force wants to study how resistant their airplanes are to bullets. They study the bullet holes on all the airplanes on the tarmac after an air battle against the Luftwaffe (German Air Force).
2. I want to know the average income of Reed graduates in the last 10 years. So I get the records of 10 randomly chosen Reedies. They all answer and I take the average.
3. Imagine it's 1993 i.e. almost all households have landlines. You want to know the average number of people in each household in Portland. You randomly pick out 500 phone numbers from the phone book and conduct a phone survey.
4. You want to know the prevalence of illegal downloading of TV shows among Reed students. You get the emails of 100 randomly chosen Reedies and ask them "How many times did you download a pirated TV show last week?"

Statistics in Society: Alfred Kinsey

In the mid 20th century, biologist/sexologist Alfred Kinsey wanted to study human sexuality.

At the time sexuality was an extremely taboo subject, very little research had been conducted at that point and Kinsey was astonished at the public's general ignorance.

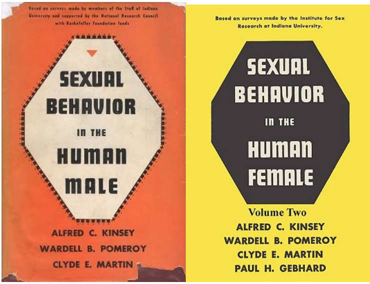
Statistics in Society: Kinsey's Questions/Research Problem

What type of questions was Kinsey interested in? Using his 300 question survey, he hoped to address...

1. What percentage of Americans engaged in premarital and extramarital sex?
2. What were the homosexual tendencies of American males?
3. How common were oral sex and masturbation?
4. ...

Statistics in Society: Kinsey Reports

The results were published two books on human sexual behavior known as the "Kinsey Reports": Sexual Behavior in the Human Male (1948) and Female (1953).



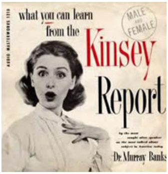
Statistics in Society: Conclusions of Kinsey Reports

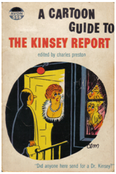
Kinsey claimed, among other things

1. 85% of white men had had premarital sex, 50% had had extra-marital sex
2. Kinsey wrote in 1948 that one in ten white men were more or less, exclusively homosexual for at least three years between the ages of 16 and 55.
3. Kinsey reported that oral sex was very common (70% of couples did it), masturbation was very common (almost 63%/92% of women/men did it)

Statistics in Society: Reaction to Kinsey Reports

Needless to say, people were taken quite aback.





There was also a huge conservative backlash against the reports.

Statistics in Society: Kinsey's Methods

What were his data collection methods? How did he sample his data? Focusing on the male report, my understanding is that

1. He did in fact base his conclusions on a very large sample size of 5300 males.
2. He sought out volunteers to answer his 300 question survey.
3. He recruited new people by asking previous respondents if they knew other people. This led to a large proportion of his sample to include prison populations and male prostitutes.

What could be some issues?

Response of the American Statistical Association

The American Statistical Association criticized the sampling procedure. In particular, John Tukey, one of the most eminent statisticians of the time, said

> *"A random selection of three people would have been better than a group of 300 chosen by Mr. Kinsey."*

Even though the Kinsey Report was groundbreaking and contributed much to the field of sexology by bringing many topics to the forefront, Kinsey's statements were not generalizable to the general public.

Reed Campus Climate Survey

During the 2012-2013 academic year Reed contracted Rankin & Associates Consulting to conduct the Campus Climate Survey to "examine the learning, living, and working environment at Reed College."

On page v and iii of the Executive Summary:
http://www.reed.edu/institutional_diversity/campus_climate.html

Examples of Different Types of Bias:

Moral of the Story

For you:

1. the consumer of statistics: Ask yourself what was the study design?
   - Who is the study population?
   - Who are the respondents and how were they selected?
2. the producer of statistics: think about how you will collect your data beforehand. If you want your results to generalize beyond just your sample to your study population, your sampling scheme has to as representative as feasible.

Explanatory and Response Variables

Example: A medical doctor pours over some his patients' medical records and observes:

$\Longrightarrow$

He then posits the following causal relationship:

- Explanatory variable: sleeping with shoes on
- Response variable: waking up with headaches

What's wrong with hypotheses?

Lecture 3: Observational Studies + Randomized Experiments + Confounding + Simpsons's Paradox

Chapter 1.4

Goals for Today

- We illustrate the difference between
  - an observational study
  - a randomized experiment, where the treatment is assigned at random.
- Introduce the notion of confounding AKA lurking variables
- Discuss Simpson's Paradox (not in textbook).

Going Back to Previous Example

Going back to the study on

$\Longrightarrow$

- The explanatory variable was: sleeping with your shoes on
- The response variable was: waking up with a headache
- The doctor hypothesized a causal relationship

Confounding Variable AKA Lurking Variable

This is an example of confounding. A confounding variable affects both the explanatory and response variable. So if:

Controlling for Potential Confounding

One way to control for (i.e. take into account) confounding is to do an exhaustive search for all such variables. This is not always practical.

Another way is via an experiment where we randomly assign individuals to a treatment or a control group in a randomized experiment.

Back to Shoes and Headaches

So imagine we recruit 10,000 people for our study and randomly assign 5000 people to each of:

- Treatment: sleep with shoes on
- Control: sleep with shoes off

In this table

Group	n	# with headache
Treatment	5000	n_1
Control	5000	n_2
Total	10,000	$n_1 + n_2$

n_1 and n_2 won't be very different.

Observational Studies vs Randomized Experiments

The key word from the study design above was randomly assign.

- Observational studies: a study where researchers have no control over who receives the treatment
- Randomized experiments: a study where researchers not only have control over who receives the treatment, but also make the assignments at random.

Observational Studies vs Randomized Experiments

Conclusion: The study introduced at the end of the last lecture is an observational study, so we cannot conclude that wearing shoes when you sleep causes you wake up with a headache.

Mantra: Correlation is not causation Just because two variables appear to be associated/correlated, does not mean that one is causing the other.

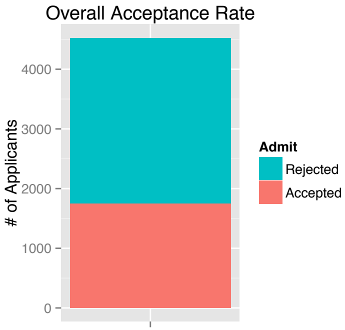
- Spurious correlations: http://www.tylervigen.com/
- Saturday Morning Breakfast Cereal: http://www.smbc-comics.com/?id=3129

Well-Known Example of Confounding

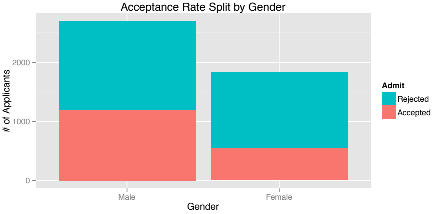
A famous example of an unaccounted for confounding variable having serious repercussions was when the UC Berkeley was sued in 1973 for bias against women who had applied for admission to graduate schools.

Let's consider the $n = 4526$ people who applied to the 6 largest departments.

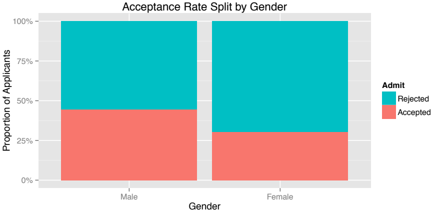
Of the $n = 4526$ applicants:



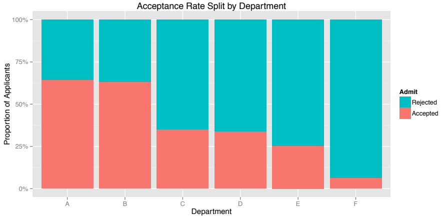
Split the counts by gender:



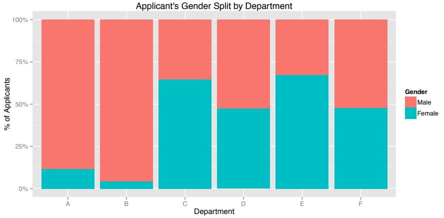
Look at proportions instead of counts:



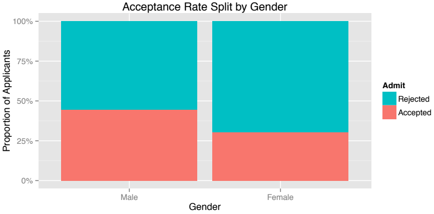
What was the "competitiveness" of departments?



Where were the women applying?



So while in aggregate things looked like this:



You need to account for department!

Bickel et al.'s (1975) Explanation

There was the presence of a confounding variable: competitiveness of applying to the department, which is a function

- number of applicants
- number of available slots

So it wasn't that departments were discriminating against women, rather:

- women tended to apply to departments with high competition and hence lower admission rates, primarily the humanities.
- men tended to apply to departments with low competition and hence higher admission rates, primarily the sciences.

Bickel et al.'s (1975) Explanation

In fact, Bickel et al. found that "If the data are properly pooled...there is a small but statistically significant bias in favor of women."

This was the exact opposite claim of the lawsuit. This is known as Simpson's Paradox.

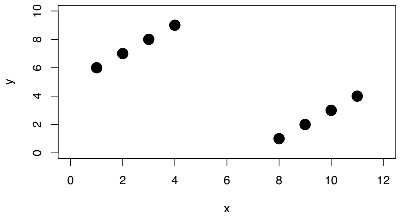
Simpson's Paradox

(From Wikipedia) Simpson's paradox occurs when a trend that appears in different groups of data disappears when these groups are combined, and the reverse trend appears for the aggregate data.

This is due to a confounding variable.

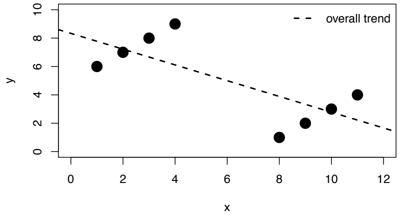
A Graphical Illustration of Simpson's Paradox

Say we have the following points:



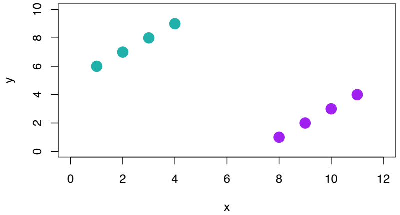
A Graphical Illustration of Simpson's Paradox

Overall, if we fit a single line, the explanatory variable x is negatively related with the outcome variable y:



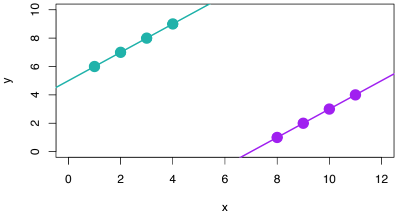
A Graphical Illustration of Simpson's Paradox

But say we consider a confounding variable, in this case color, and fit two separate lines for each group:



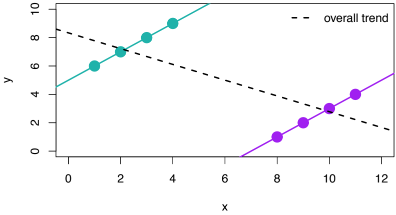
A Graphical Illustration of Simpson's Paradox

The subgroups now exhibit a positive relationship!



A Graphical Illustration of Simpson's Paradox

i.e. the trend in aggregate is the reverse of the trend in the subgroups (teal & purple).



Bickel et al.'s (1975) Conclusion

"The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seem quite fair on the whole, but apparently from prior screening at earlier levels of the educational system."

"Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects."

The original paper can be found here.

Next time

We will discuss

- Specific types of sampling beyond just simple random sampling, as this is not always feasible
- Experimental design: some key principles to keep in mind when evaluating the efficacy of treatments.

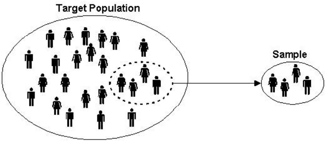
Lecture 4: Sampling Methods + Design of Experiments

Chapter 1.4.2 + 1.5

Goals for Today

- Discuss different types of sampling
- Designing experiments
- Very important example: clinical trials
- Example of my own designed experiment: Fried Chicken Face Off

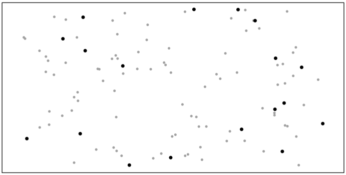
Recall from Lecture 1.3: Population and Samples



If the sample is representative of the desired population then our results are generalizable.

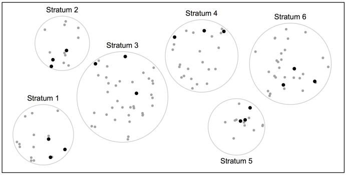
How do we take a representative (i.e. unbiased) sample? You randomly sample from the population.

1. Simple Random Sampling



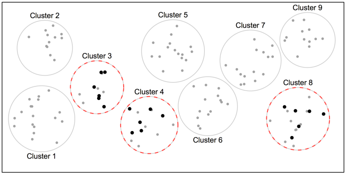
Most granular sampling: Where every individual in the population has the same probability of being sampled. Here, all dots are members of the population, and the bolder dots are sampled.

2. Stratified Sampling



Divide and conquer: The population is divided into strata, and we sample from each strata. For example, each strata could be a census tract in Oregon, and we sample 3 individuals from each strata.

3. Cluster Sampling



Two stage sampling: Very similar to stratified sampling in its process, except that there is no requirement to sample from every cluster. First the clusters in red were chosen at random, and then we sample from them.

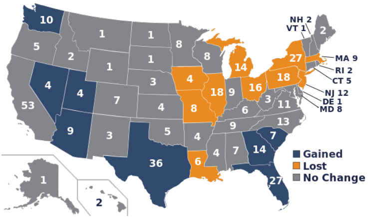
Three Different Types of Sampling

1. Simple random sampling: most granular sampling
2. Stratified sampling: divide and conquer
3. Cluster sampling: two-stage sampling

The mathematics behind the stratified and cluster sampling are more complicated to account for the hierarchies involved. Ex: for stratified sampling use the Horvitz-Thompson estimator.

Statistics in Society: The US Census

The purpose of the decennial US census is congressional apportionment: the 435 seats in the US House of Representatives get distributed to the 50 states in proportion to their population. After the 2010 census:



Statistics in Society: The US Census

President Bill Clinton's administration planned on using sampling in the 2000 census. In an article dated in 1996:

HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS | MOST RECENT

Login | Register Now | Help

The New York Times

U.S.

Search All NYTimes.com

Go

COLLECTIONS > STATISTICAL METHODS

More Like This

U.S. CENSUS BUREAU REJECTS REVISION TO NATION'S TALLY

The Nation: Sample Case; You Fill Up My Census, Even If I...

Lessons From the Election That Shook America

Find More Stories

Statistical Methods

In a First, 2000 Census Is to Use Sampling

By STEVEN A. HOLMES
Published: February 29, 1996

To cut costs and improve accuracy, the Census Bureau said today that it would actually count only 90 percent of the United States population in 2000 and rely on statistical sampling methods to determine the number remaining.

The plans, announced at the Commerce Department, mean that for the first time the official tally of the American population, done every 10 years and used to apportion seats in the House of Representatives, will be based in part on a scientifically determined estimate rather than the actual head count conducted through a mass direct-mail campaign.

Census Bureau officials say the revised method is needed to keep costs down and to avoid a repeat of the 1990 census, which missed record numbers of people that had been traditionally hard to count, mainly members of ethnic and racial minorities.

"What we intend to do to meet our twin goals of reducing costs and increasing accuracy is to make a much greater use of widely accepted scientific statistical methods, and sampling is first and foremost among them," said Martha Farnsworth Riche, the Census Bureau Director.

SIGN IN TO E-MAIL

PRINT

Statistics in Society: The US Census

However, Article I, Section 2 of the US Constitution states: *The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the US, and within every subsequent Term of ten Years...*

As such, the Supreme Court ruled 5-4 in 1999 that

- sampling could not "under any circumstances" be used to reapportion U.S. House seats
- could be used for other purposes such as redrawing state legislative districts or allocating federal funds to cities and states

Statistics in Society: The Census

THE WALL STREET JOURNAL.

U.S. EDITION Friday, May 15, 2009 As of 4:42 PM EDT

Subscribe | Log In

Home | World | U.S. | Business | Tech | Markets | Market Data | Your Money | Opinion | Life & Culture | N.Y. | Real Estate | Management

Seib & Wessel | Politics & Policy | Washington Wire | Economy | Health Law Rollout | WSJ/NBC News Poll | Journal Report | Columns & Blogs

TOP STORIES IN POLITICS

1 of 12 How Serious Are Risks If U.S. Doesn't Act?

2 of 12 White House Wants 'Hard Look' at Syria Weapons Offer

3 of 12 Senators Factor Voters Into Their Syria Equation

Discord Over Military Strik Imperils President's Agend

POLITICS | May 15, 2009, 4:42 p.m. ET

Census Nominee Rules Out Statistical Sampling in 2010

Article | Comments | MORE IN POLITICS & POLICY »

Tweet 1

Email | Print | Save

By TIMOTHY J. ALBERTA

WASHINGTON—President Barack Obama's nominee to head the Census Bureau on Friday ruled out using statistical sampling to adjust the results of the 2010 census, quelling Republican concerns and making his confirmation likely next week.

Robert Groves, director of the University of Michigan's Survey Research Center and a former Census Bureau official, is an expert on statistical sampling, the practice of extrapolating a larger population from a smaller slice of it. Proponents of sampling say it helps produce a more accurate tally of the population, especially when it comes to traditionally undercounted groups, such as minorities living in urban areas.

But many Republican lawmakers insist that sampling violates the Constitution, which calls for an "actual Enumeration" of the population every 10 years. Critics also say the use of sampling would politicize the traditionally nonpolitical Census Bureau.

Available to WSJ.com Subscribers

Life on Wall Street Grows Less Risky

Icahn Throws In the Towel on Dell

Hedge Funds Cut Back on Fees

Lessons of the Rescue: A Drama in Five Acts

Principles Of Designing Experiments

Switching gears...

(Wikipedia) In general usage, design of experiments (DOE) or experimental design is the design of any information-gathering exercises where variation is present, whether under the full control of the experimenter or not.

However, in statistics, these terms are usually used for controlled experiments: experiments where there is a control and treatment group.

Principles Of Designing Experiments

Clinical Trials

To evaluate the efficacy of a drug, they must be subject to a clinical trial. The gold standard for a clinical trial is randomized controlled trial. i.e. randomized control and treatment groups.

Example of Mine: Ezell's Famous Chicken

In Seattle's Central District lies

From Wikipedia: Oprah Winfrey called it her favorite fried chicken. There are a number of photos of her on the wall of the original restaurant proclaiming her love of the chicken. It is also said she has the chicken flown to her in Chicago when she has a craving.

Example of Mine: Ezell's Famous Chicken

One day I was raving about Ezell's Chicken. My friend Nick accused me of being another person "buying into the hype"; that if people were subjected to a blinded taste test, Ezell's would fare no better than KFC. So...

vs

We set up a "Fried Chicken Face Off" where we would have individuals try both kinds of chicken and rate which one they liked more.

Design of Experiment Principles in Place

Goal: Evaluate which kind of chicken, Ezell's or KFC, that people prefer in a blinded taste test. (Not if participant can determine which chicken came from which restaurant.)

Question: What principles of the design of experiments should be put in place to this end?

Design of Experiment Principles in Place

The design principles we put in place:

- Single blinded: The taster doesn't know which (Ezell's or KFC) chicken they are eating, but the server does.
- Randomizing which kind of meat (wing, breast, leg) between tasters. Each taster would try two kinds of meat.
- Controlling for which kind of meat within a taster: i.e. if you eat a KFC wing, you will necessarily eat an Ezell's wing
- Randomizing which order of chicken you eat: KFC first or not

Design of Experiment Principles in Place

The design principles we put in place:

- Controlling for temperature: hence we're picking a place that is central to both Ezell's and KFC given the traveling required.
- Controlling for visual look: We thought blind-folds were a bit excessive
- Controlling for kind of batter: we can't do KFC crispy chicken b/c Ezell's doesn't have that type of batter. This is a limitation of the study b/c some feel the crispy chicken is better, but we have no choice.
- Just one replicate of each kind of meat.

Results

Final score: KFC 8, Ezell's 4.
Some notes:

- Even though people were "blinded", most knew which the two pieces were from KFC.
- People generally felt the chicken meat from Ezell's was better, and this was magnified as the chicken went cold.
- However, they felt the skin was better at KFC. Given that fried chicken is what it is b/c of the skin, people voted for KFC.
- Future metrics need to consider the chicken and the skin separately, as well as the "overall experience" scores. i.e. this face off should be viewed as a pilot study

Caution: Grad Students NOT at Work

Next time

Examining and visualizing numerical data

Lecture 5: Visualizing Numerical Data

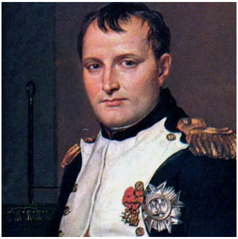
Chapter 1.6 + 1.7

Goals for Today

- Visualizing numerical data
  - Two famous historical examples of data visualization
  - Reed's 2013 entering class
- Histograms
- Measures of Central Tendency: Mean, Median, and Mode
- Measure of Spread: Sample variance and sample standard deviation

Famous Example 1: Napoleon's March on Russia in 1812

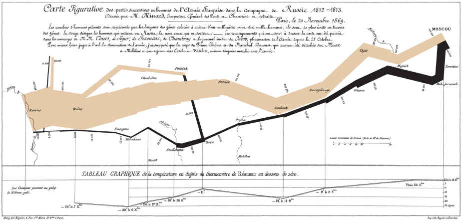
In 1812, Napoleon led a French invasion of Russia, at one point marching on Moscow.



Famous Example 1: Napoleon's March on Russia in 1812

The advance and retreat on Moscow was an unmitigated disaster:

Famous Example 1: Napoleon's March on Russia in 1812



Famous Example 1: Napolean's March on Russia in 1812

Why is this visualization big deal?

On a two-dimensional page, it displays 6 variables (in others words, 6 dimensions of information) at once:

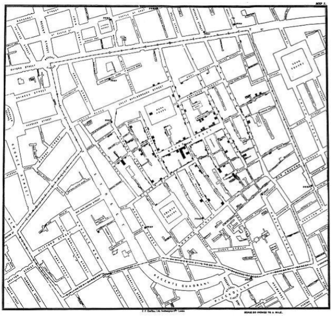
Famous Example 2: 1854 Broad Street Cholera Outbreak

On August 31 1854, an epidemic of cholera began in the Soho neighborhood of London. Over the next three days 127 people near Broad Street had died.

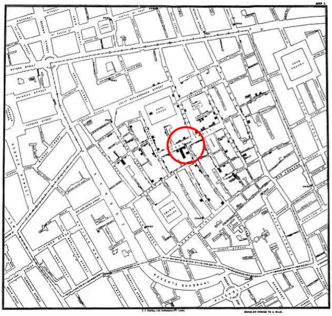
Dr. John Snow, a physician, was a student of the disease. (From Wikipedia) Snow was a skeptic of the then-dominant miasma theory that stated that diseases such as cholera or the Black Death were caused by pollution or a noxious form of “bad air.”

Snow created the following map to investigate:

Famous Example 2: 1854 Broad Street Cholera Outbreak



Famous Example 2: 1854 Broad Street Cholera Outbreak



Famous Example 2: 1854 Broad Street Cholera Outbreak

He identified the source of the outbreak as water from the Broad Street Pump, which was near a cesspit that began to leak.

This led to discovering that cholera was transmitted by food and water being contaminated by fecal matter and not via the air. This was a watershed moment in the emerging field of epidemiology.

Histograms

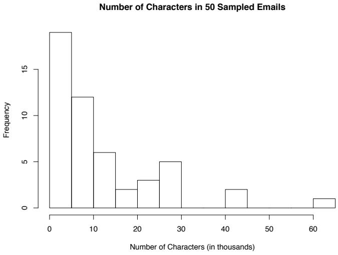
In the openintro package, the email50 dataset contains a random sample of 50 emails, in which researchers try to identify emails as spam. One variable is the # of characters:

```
Characters |  0-4.999  5-9.999  10-14.999  ...  60-64.999
(in 1000's)|
---------------------------------------------------------
Count      |       19       12          6  ...          1
```

So each of the intervals 0-5, 5-10, 10-15, etc. are buckets/bins and we count the number of emails in each bucket/bin.

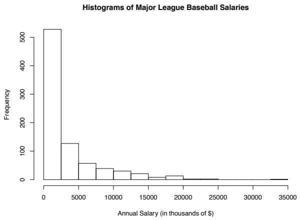
Histograms

Histograms provide a description of the shape of the distribution of data.



Skew and Long Tail

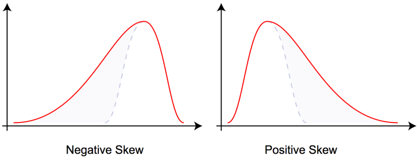
Also in the `openintro` package is MLB salary data in 2010. If we plot a histogram:



The data has a long tail to the right: data is right-skewed. i.e. a small number of players who make a VERY large amount of money.

Trick to Remembering Which Skew is Which

- Long tail to the right: data is right-skewed AKA positively-skewed
- Long tail to the left: data is left-skewed AKA negatively-skewed



Reed's 2013 US-Originating Entering Class

What can we do about skewed data?

http://rpubs.com/rudeboybert/reed2013

Mean

The mean, AKA average, is a common way to measure the center of the data. So for example, the mean of 1, 2, 5, 3, and 7 is

$$\frac{1+2+5+3+7}{5} = 3.6$$

We label the sample mean $\overline{x}$ (pronounced "x bar"):

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

where $x_1, x_2, \ldots, x_n$ are the n observed/sampled values.

Median

The median, however, is the middle number.

Two cases:

- Odd number of values: the median of (1, 3, 5, 8, 10) is 5.
- Even number of values: the median of (1, 3, 5, 8) is the average of the middle two values: $\frac{3+5}{2} = 4$

But why use the median at all?

Mean vs Median: Imaginary Scenario

- Say at company X, there 5 employees: the CEO and everyone else.
- The CEO earns \$1000 an hour, while the others earn \$20, \$21, \$30, and \$40 an hour.
- The employees complain that they are paid too little.
- The CEO counters that the mean hourly salary is $\overline{x} = \frac{20+21+30+40+1000}{5} = 222.20$ an hour, which is really high.

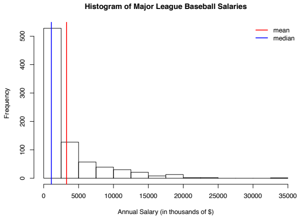
Mean vs Median: Imaginary Scenario

The CEO's extreme salary is inflating the mean. A more appropriate measure is the median hourly salary of 30.

Medians are less sensitive to (i.e. more robust to) outliers than the mean.

Ex: the "median home price" is typically used, because it isn't as sensitive as the mean to the few very expensive houses.

Mean vs Median: Back to MLB Salary Data

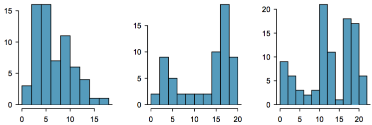


Mode

A mode is the value that appears the most often in a data set. So out of (1, 3, 3, 5, 6), the modal value is 3.

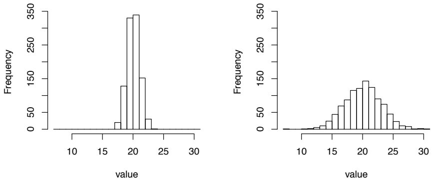
Modes also describe peaks, but this can get subjective.

A distribution can be unimodal, bimodal, or multimodal:



Measure of Spread

Next, consider the following two histograms: Both have mean of about 20. What is the difference between them?



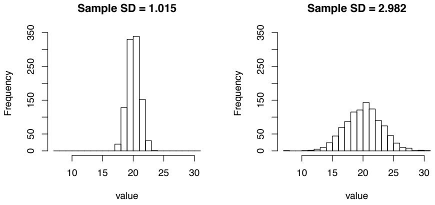
Measure of Spread

We need a measure of spread/variability. The sample variance s^2 is roughly the average squared distance from the mean.

The sample standard deviation s is the square root of the sample variance. The sample standard deviation is useful when considering how close the data are to the mean.

Measure of Spread

Back to example:



How to Compute the Sample Standard Deviation

Read section 1.6.4. The formula really doesn't make much intuitive sense, but is the way it is due to mathematical convenience. Fortunately there is an `R` command that computes it for you: `sd()`

Next Time

- Another simple data visualization tool: boxplots
- Examining/Visualizing Categorical Data

Lecture 6: Visualizing Numerical and Categorical Data

Chapter 1.6+1.7

Goals for Today

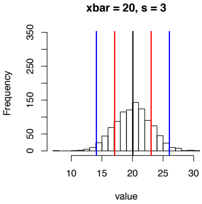
- Rule of thumb for standard deviations
- Population vs sample mean/variance/standard deviations
- Percentiles and Quartiles
- Boxplots
- Piecharts, barplots, mosaicplots

Rule of Thumb for Standard Deviations

If the data distribution is bell-shaped, then

- about $\frac{2}{3}$ of the data will be within one SD of the mean (book says 70%).
- about 95% of the data will be within two SD.

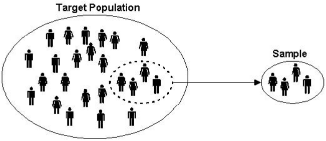
Example



- black line is mean $\overline{x}$
- red lines mark about $\frac{2}{3}$: $[\overline{x} - s, \overline{x} + s] =$ $[20 - 3, 20 + 3] = [17, 23]$.
- blue lines mark about 95%: $[\overline{x} - 2s, \overline{x} + 2s] =$ $[20 - 6, 20 + 6] = [14, 26]$.

Population vs Sample Mean/Variance/Standard Deviation

Recall the notion of taking a representative sample from a study/target population. Say we are interested in the income of the individuals.



Population vs Sample Mean/Variance/Standard Deviation

- The sample mean $\overline{x}$ is the mean income of the 4 sampled people.
- The population mean μ is the mean income of all 24 people in the target population.
- We say $\overline{x}$ estimates μ. If the sample is representative, then $\overline{x}$ estimates μ with high accuracy i.e. it is unbiased.

Population vs Sample Mean/Variance/Standard Deviation

	True Population Value	Sample Value
Mean	μ	$\overline{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s

The sample value is used to estimate the (true) population value.

Percentiles

A percentile (%'ile) indicates the value below which a given %'age of observations fall.

SAT Scores from 2012
http://media.collegeboard.com/digitalServices/pdf/research/SAT-Percentile-Ranks-2012.pdf

So for example, if you scored 700 in critical reading, 95% of college-bound seniors who took the test did worse.

Quartiles

Quartiles split up the data into 4 intervals, each with about one quarter of the data:

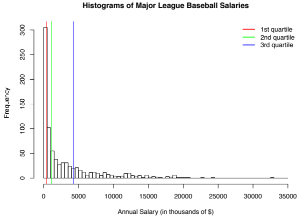
- The lower quartile is the 25th %'ile
- The median is the 50th %'ile
- The upper quartile is the 75th %'ile

The interquartile range (IQR) is another measure of the spread of a sample:

$$\text{IQR} = \text{upper quartile} - \text{lower quartile}$$

MLB Data Quartiles

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
400.0   418.3  1094.0  3282.0  4250.0 33000.0
```



The IQR is (3rd Quartile - 1st Quartile) $= 4250.0 - 418.3 = 3831.7$ i.e the distance between the red and blue line.

Robust Statistics (Chapter 1.6.6)

Robust estimates are statistics where extreme observations (outliers) have less effect on their values, i.e. are more resistant to their effect. The median and IQR are two examples.

Example: Old scoring system in figure skating: drop the highest & lowest scores and then take the average.

Say we have a figure skater who gets judged by countries V-Z:

Country	V	W	X	Y	Z
Score	4.0	5.2	5.2	5.3	6.0

Drop the 4.0 and 6.0, then the final score is: $\frac{5.2+5.2+5.3}{3} = 5.23$

Boxplots

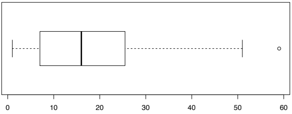
Boxplots are visual summaries of a sample $x_1, \ldots, x_n$ that bring to light unusual values (potential outliers):

Example: # US Forces casualties in the war in Afghanistan for each month from 2008-2009:

7, 1, 7, 5, 16, 28, 20, 22, 27, 16, 1, 3, 14, 15, 13, 6, 12, 24, 44, 51, 37, 59, 17, 17

Boxplots

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.00    7.00   16.00   19.25   24.75   59.00
```



Page 29 of text describes the length of the whiskers: they capture data that is no more than $1.5 \times IQR$ of both ends of the box.

Outliers Are Relatively Extreme

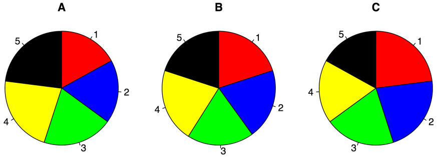
An outlier is an observation that appears extreme relative to the rest of the data.

Why it is important to look for outliers? Examination of data for possible outliers serves many useful purposes, including

- Identifying strong skew in the distribution.
- Identifying data collection or entry errors.
- Providing insight into interesting properties of the data.

Piecharts

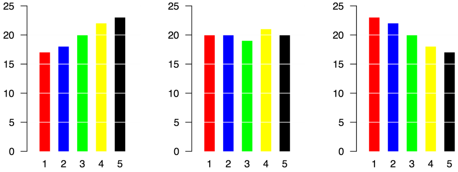
Say we have the following piecharts represent the polling from a local election with five candidates (1-5) at three different time points A, B, an C:



Answer the following questions:

- In the first race, is candidate 5 doing better than candidate 4?
- Who did better between time A and time B, candidate 2 or candidate 4?

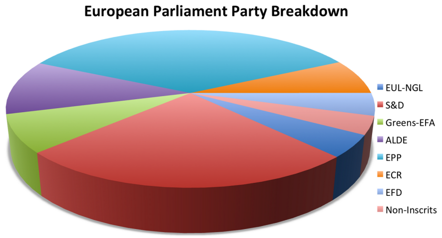
Barplots Instead



Answers:

- Candidate 5 is doing better than 4
- Between A and B, candidate 2 went from about 17% to 20% while candidate went from about 22% to 21%. So candidate 2 did better

3D Piecharts Can Be Deceiving



EEP (teal) has 266 seats, whereas S&D (red) has 190 seats.

Titanic Survival Data

Typing `data(Titanic)` in R loads the survival and death counts, split by each of the following categories:

- Class: 1st, 2nd, 3rd, or crew (4 levels)
- Gender (2 levels)
- Age: Child or adult (2 levels)

i.e. $4 \times 2 \times 2 = 16$ possible groups to consider.

Questions

- What was the effect of class (1st, 2nd, 3rd, crew) on your chances of survival?
- Did the "women and children" first lifeboat policy hold?

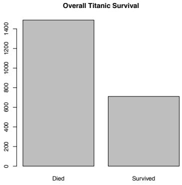
Frequency Table

A table summarizing a single categorial variable is called a frequency table. Overall:

Died	1490
Survived	711
Total	2201

Barplot

Barplots are ways to display categorial variables:



Contingency Table

A table that cross-classifies two categorical variables is a contingency table. Now let's split survival by class: 1st, 2nd, 3rd, and crew.

Before:

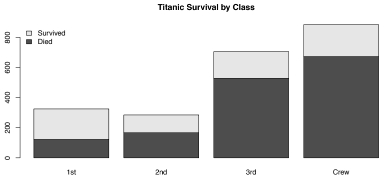
Died	1490
Survived	711
Total	2201

After:

	1st	2nd	3rd	Crew	Total
Died	122	167	528	673	1490
Survived	203	118	178	212	711
Total	325	285	706	885	2201

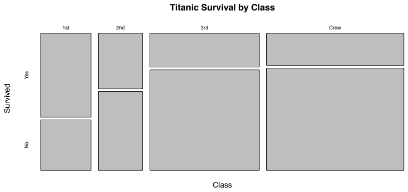
Stacked Barplot

Stacked barplots are one way to display values from a contingency table:



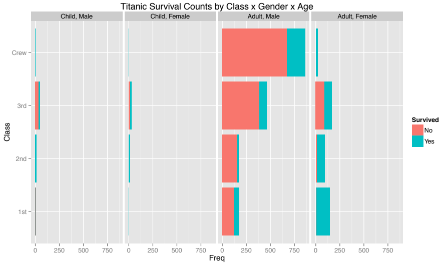
Mosaic Plots

Mosaic plots are similar, but the widths of the bars now reflect proportions:



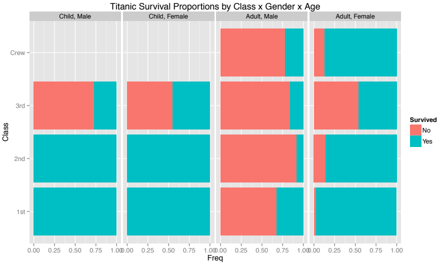
Stacked Barplots

Using the `ggplot2` package, we can plot survivals by class, age, and gender all at once.



Standardized/Normalized Stacked Barplots

Instead of raw counts, we can expand each bar to reflect proportions (i.e. standardize/normalize them).



Lecture 7: Probability

Chapter 2.x

Outcomes

Probability forms the theoretical backbone of statistics. We use probability to characterize randomness.

We often frame probability in terms of a random process giving rise to an outcome.

Typical examples

- Die roll: 6 outcomes
- Coin Flip: 2 outcomes

Disjoint AKA Mutually Exclusive Outcomes

Two outcomes are disjoint (AKA mutually exclusive) if they cannot both occur at the same time.

Die example:

- Rolling a 1 and a 2 are disjoint.
- Rolling a 1 and rolling "an odd number" are not disjoint.

Addition Rule of Probability

If A_1 and A_2 are disjoint outcomes, then

$$P(A_1 \text{ or } A_2) = P(A_1) + P(A_2)$$

Ex: Rolling 1 and 2 are disjoint, so:

$$P(\text{rolling 1 or 2}) = P(\text{rolling 1}) + P(\text{rolling 2}) = \frac{1}{6} + \frac{1}{6}$$

General Addition Rule of Probability

If A_1 and A_2 are two outcomes (not necessarily disjoint), then

$$P(A_1 \text{ or } A_2) = P(A_1) + P(A_2) - P(A_1 \text{ and } A_2)$$

Venn diagram:

General Addition Rule of Probability

Events are just combinations of outcomes. Ex: Deck of cards

- A_1 = event we draw a diamond
- A_2 = event we draw a face card

These two events are not disjoint, as there are 3 diamond face cards. Venn diagram:

General Addition Rule of Probability

$$
\begin{aligned}
P(A_1 \text{ or } A_2) &= P(\text{diamond or a face card}) \\
&= P(\text{diamond}) + P(\text{face card}) - \\
&\quad\; P(\text{diamond AND face card}) \\
&= \frac{13}{52} + \frac{3 \times 4}{52} - \frac{3}{52} = \frac{22}{52} = 42.3\%
\end{aligned}
$$

Sample Space and the Complement of Events

A die has 6 possible outcomes. The sample space is the set of all possible outcomes $S = \{1, 2, \ldots, 6\}$.

Say event A is the event of rolling an even number i.e $A = \{2, 4, 6\}$. The complement of event A is $A^c = \{1, 3, 5\}$ i.e. getting an odd number.

Thm

$$P(A) + P(A^c) = 1$$

Independence

Two processes are independent if knowing the outcome of one provides no useful information about the outcome of the other. Otherwise they are dependent.

Consider:

1. Die rolls
2. You get a movie recommendation from your friend Robin, but then their significant other Sam also recommends it.
3. You compare test scores from two Grade 9 students in the same class. Then same school. Then same school district. Then same city. Then same state.

Independence

We say that events A and B are independent if

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

Ex: Dice rolls are independent:

$$\begin{aligned} P(\text{rolling 1 and then 6}) &= P(\text{rolling 1}) \times P(\text{rolling 6}) \\ &= \frac{1}{6} \times \frac{1}{6} = \frac{1}{36} \end{aligned}$$

Conditional Probability

The conditional probability of an event A given the event B, is defined by

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

Example

Let's suppose I take a random sample of 100 Reed students to study their smoking habits.

	Smoker	Not Smoker	Total
Male	19	41	60
Female	12	28	40
Total	31	69	100

- What is the probability of a randomly selected male smoking?

$$P(S|M) = \frac{P(S \text{ and } M)}{P(M)} = \frac{19/100}{60/100} = \frac{19}{60}$$

- What is the probability that a randomly selected smoker is female?

$$P(F|S) = \frac{P(F \text{ and } S)}{P(S)} = \frac{12/100}{31/100} = \frac{12}{31}$$

Put It Together! Independence and Conditional Prob.

If A and B are independent events, then

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

then

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

i.e. $P(A|B) = P(A)$: the event B occurring has no bearing on the probability of A

Gambler's Fallacy: Roulette

You can bet on individual numbers, sets of numbers, or red vs black. Let's assume no 0 or 00, so that $P(\text{red}) = P(\text{black}) = \frac{1}{2}$.

Gambler's Fallacy: Roulette

One of the biggest cons in casinos: spin history boards.

Let's ignore the numbers and just focus on what color occurred.
Note: the white values on the left are black spins.

Gambler's Fallacy: Roulette

Let's say you look at the board and see that the last 4 spins were red.

You will always hear people say "Black is due!"

Ex. on the 5th spin people think:

$$P(\text{black}_5 \mid \text{red}_1 \text{ and } \text{red}_2 \text{ and } \text{red}_3 \text{ and } \text{red}_4) > P(\text{red}_5 \mid \text{red}_1 \text{ and } \text{red}_2 \text{ and } \text{red}_3 \text{ and } \text{red}_4)$$

Gambler's Fallacy: Roulette

But assuming the wheel is not rigged, spins are independent i.e. $P(A|B) = P(A)$. So:

$$P(\text{black}_5|\text{red}_1 \text{ and } \text{red}_2 \text{ and } \text{red}_3 \text{ and } \text{red}_4) = P(\text{black}_5) = \frac{1}{2}$$

$$P(\text{red}_5|\text{red}_1 \text{ and } \text{red}_2 \text{ and } \text{red}_3 \text{ and } \text{red}_4) = P(\text{red}_5) = \frac{1}{2}$$

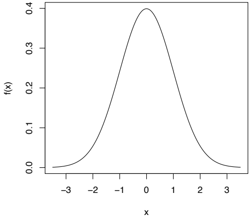
Next Week's Lab

Basketball players who make several baskets in succession are described as having a "hot hand." This refutes the assumption that each shot is independent of the next.

We are going to investigate this claim with data from a particular basketball player: Kobe Bryant of the Los Angeles Lakers in the 2009 NBA finals.

Next Time

Discuss the Normal Distribution



Lecture 8: Normal Distribution

Chapter 3.1

Goals for Today

- Define the normal distribution in terms of its parameters
- Review: $\frac{2}{3}$ / 95% / 99.7% rule
- Standardizing normal observations to z-scores

Normal Distribution

From text page 118:

Many variables are nearly normal, but none are exactly normal. Thus the normal distribution, while not perfect for any single problem, is very useful for a variety of problems.

We will use it in data exploration and to solve important problems in statistics.

Normal Distribution

Normal distributions:

1. are symmetric
2. are unimodal and bell-shaped
3. have area under the curve 1

Normal Distribution

A normal curve can be described by two parameters:

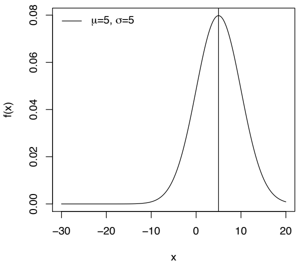
- the mean μ. i.e. the center
- the standard deviation (SD) σ. i.e. the measure of spread

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Recall these were the population mean and the population SD.

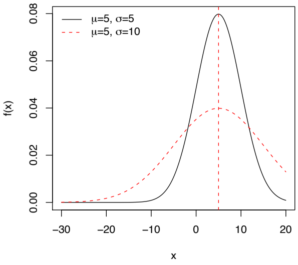
Normal Distribution

μ (mean) specifies the center, σ (standard deviation) the spread.



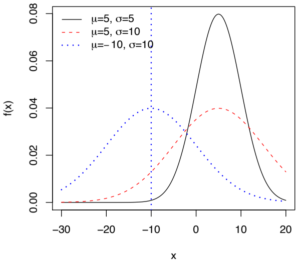
Normal Example

μ (mean) specifies the center, σ (standard deviation) the spread.



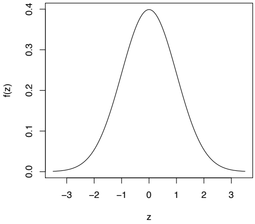
Normal Example

μ (mean) specifies the center, σ (standard deviation) the spread.



Standardized Normal Distribution

If $\mu = 0$ and $\sigma = 1$, this is the standard normal distribution:

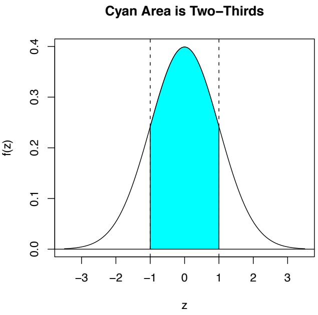


Rules of Thumb

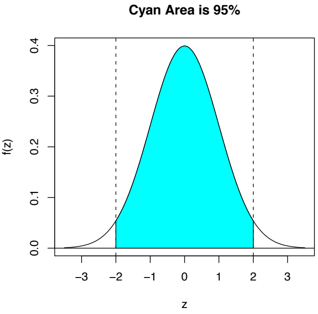
Recall if a distribution is normal, then:

1. Approx. $\frac{2}{3}$'s of the data are within ± 1 SD of the mean
2. Approx. 95% of the data are within ± 2 SD of the mean
3. Also approx. 99.7% of the data are within ± 3 SD of the mean

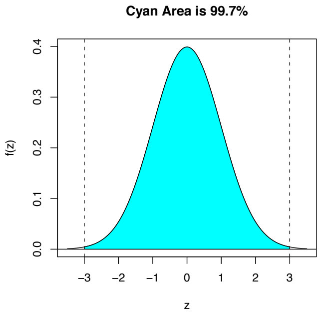
Ex: Standard Normal $\mu = 0$, $\sigma = 1$



Ex: Standard Normal $\mu = 0$, $\sigma = 1$



Ex: Standard Normal $\mu = 0$, $\sigma = 1$

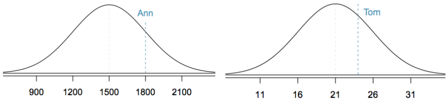


Motivating Example

From text: Say Ann scores 1800 on the SAT and Tom scores 24 on the ACT. Say both tests scores were normally distributed with:

	SAT	ACT
Mean μ	1500	21
SD σ	300	5

Question: Who did relatively better?



z-scores

The z-score AKA standardized observation of an observation x is the number of SD it falls above or below the mean.

The z-score for an observation x that follows a distribution with mean μ and SD σ:

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

z-scores

Why is the z-score $z = \frac{x-\mu}{\sigma}$ called the standardized observation?

1. The observations are centered at μ.
   re-center the x observations to 0 by subtracting μ.
2. The observations have spread σ.
   re-scale the spread of the $x - \mu$ values to be 1 by dividing by σ.

So we can compare observations from any normally distributed data with (μ, σ)

i.e. we've standardized the observations to make them comparable.

Back to Example

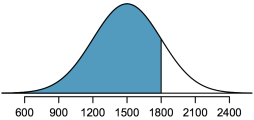
- Ann scored 1800. $z = \frac{1800-1500}{300} = +1$ standard deviation from the mean
- Tom scored 24. $z = \frac{24-21}{5} = +0.6$ standard deviation from the mean

So Ann did relatively better.

Percentiles

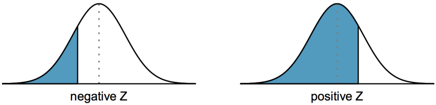
Recall a percentile (%'ile) indicates the value below which a given %'age of observations fall below.

Question: What %'ile is Ann's SAT score of 1800?
i.e. what is the blue shaded area?



Percentiles

Because the total area under the curve is 1, the area to the left of z represents the %'ile of the observation:



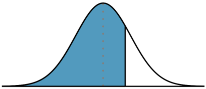
- The blue shaded area on the left plot will be less than 0.5. We have %'iles less than the 50th %'ile.
- The blue shaded area on the right plot will be greater than 0.5. We have %'iles greater than the 50th %'ile.

Normal Probability Table

A normal probability table allows you to:

- identify the %'ile corresponding to a z-score
- or vice versa: the z-score corresponding to a %'ile

The normal probability tables on page 409 represent z-scores and %'iles corresponding to area to the left:



Normal Probability Table

	Second decimal place of Z									
Z	0.00	0.01	0.02	*0.03*	**0.04**	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	*0.6664*	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	**0.7995**	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

- Red case: Given a z-score of 0.43. A lookup tells us the area to the left of z=0.43 is 0.6664, i.e. the 66th %'ile
- Blue case: We want the z-score that is the 80th %'ile. Reverse lookup: the closest value on the table is 0.7995, i.e. a z-score of 0.84.

Back to Ann and Tom

- Since Ann had a z-score of 1.0, her %'ile is 0.8413. (1.0 row, 0.00 column)
  i.e. She did better than 84.13% of SAT test takers.
- Since Tom had a z-score of 0.6, his %'ile is 0.7257. (0.6 row, 0.00 column)
  i.e. He did better than 72.57% of ACT test takers

Next Time

Next time we will:

- Re-iterate the motivation for the normal curve.
- Go over examples using z-scores.
- Evaluating the normal approximation.

Lecture 9: Normal Approximation

Chapter 3.2

Goals for Today

- Discuss how to find %'iles for negative values of z
- Examples
- Evaluating how "normal" certain data are.

Solving Normal Questions

Whenever solving questions of this sort ALWAYS draw a rough picture first and keep in mind:

1. The normal distribution/curve is symmetric
2. The total area under the curve is 1

Normal Probability Tables

Alternatively, whereas

- table on P.409 gives areas to the left of positive values of z.
- table on P.408 gives areas to the left of negative values of z.

I'm only going to give you P.409 table for exams.

Speeding on I-5

The distribution of passenger vehicle speeds traveling on Interstate 5 Freeway (I-5) in California is nearly normal with a mean of 72.6 mph and a standard deviation of 4.78 mph.

a) What percent of passenger vehicles travel slower than 80 mph?
b) What percent of passenger vehicles travel between 60 and 80 mph?
c) How fast to do the fastest 5% of passenger vehicles travel?
d) The speed limit on this stretch of the I-5 is 70 mph. Approximate what percentage of the passenger vehicles travel above the speed limit on this stretch of the I-5.

Speeding on I-5

a) What percent of passenger vehicles travel slower than 80 mph?

Speeding on I-5

b) What percent of passenger vehicles travel between 60 and 80 mph?

Speeding on I-5

c) How fast to do the fastest 5% of passenger vehicles travel?

Speeding on I-5

d) The speed limit on this stretch of the I-5 is 70 mph. Approximate what percentage of the passenger vehicles travel above the speed limit on this stretch of the I-5.

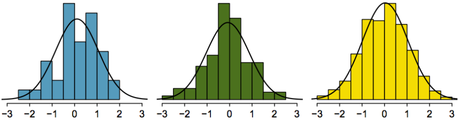
Switching Gears: Normal Approximation

Although we stated that many processes in the physical world look bell-shaped, i.e. roughly normal, we must keep in mind that this is an approximation.

Question: How do we verify normality?

Normal Approximation

What about these ones? How well do the histograms fit to the normal curve?



Normal Probability Plots

Normal probability plots (AKA quantile-quantile plots AKA QQ-plots) are a method for visually displaying how well data fit a normal curve.

The k^{th} $q - quantile$ is the value such that proportion $\frac{k}{q}$ of the observations fall below it. So

- The 4-quantiles are the quartiles.
- The 100-quantiles are the percentiles.

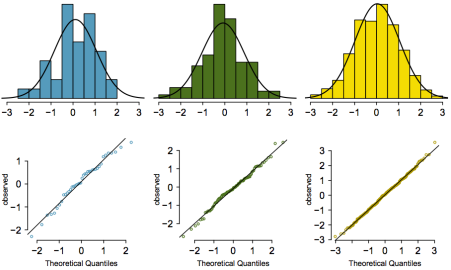
Normal Probability Plots

A normal probability plot compares:

- The observed quantiles of a data set (on the y-axis)
- The theoretical quantiles that are exactly normal (on the x-axis)

The more "normal" the data is, the better the fit.

Normal Probability Plots



Next Time

- Introduce some of the more useful other distributions: Bernoulli, Geometric, Binomial, and Poisson

Lecture 10: Bernoulli and Geometric Random Variables

Chapter 3.3-3.5

Goals for Today

Define

- Bernoulli random variables
- Geometric random variables

Mathematical Definition of a Bernoulli Random Variable

A random variable X is a random process or variable with a numerical outcome.

Random variables are described in terms of their distribution.

Bernoulli Distribution

Say we have an experiment where we define each trial (or instance) to have two possible outcomes of interest. Examples

- Coin flips: heads vs tails
- Medical test (for a disease): positive vs negative
- Rolling a die and getting a 6 vs not getting a 6

In each case we can define the outcomes to be success vs failure. No moral judgement; just labels.

Bernoulli Distribution

Say we have trials where we have two outcomes: either a "success" or a "failure". Classic example: coin flips have $p = 0.5$ of heads, if we define heads as the success.

- probability p of a "success." Denote successes with a "1."
- probability $1 - p$ of a "failure." Denote failures with a "0."

Definition of a Bernoulli Random Variable

If X is a random variable that takes value

- 1 with probability of success p
- 0 with probability of failure $1 - p$

then X is a Bernoulli random variable with mean and standard deviation:

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

Intuition Behind σ

Sample Proportion

Say you repeat n instances of a Bernoulli random variable. You end up with a sample $x_1, \ldots, x_n$

The sample proportion $\widehat{p}$ (p-hat) is the sample mean of these observations. i.e.

$$\widehat{p} = \frac{\#\text{ of successes}}{\#\text{ of trials}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example of Bernoulli Distribution

- A success as rolling a 6.
  So $P(X = 1) = P(\text{success}) = p = \frac{1}{6}$.
- A failure as rolling anything else.
  So $P(X = 0) = P(\text{failure}) = 1 - p = \frac{5}{6}$.

Back to Lecture 3.1: Population vs Sample Values

	True Population Value	Sample Value
Mean	μ	$\overline{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s
Proportion	p	$\widehat{p}$

The sample proportion $\widehat{p}$ is a specific kind of sample mean for Bernoulli random variables, which estimates p, a specific kind of population mean.

Scenario

Question: Say

- the San Francisco Giants have equal probability $p = 0.6$ of winning any game
- games are independent

It's the beginning of the season. What is the probability that they don't win their first game until the 5th game of the season?

For this to happen, there must be 4 loses in the first 4 games AND a win in the 5th game:

$$\begin{aligned}
P(\text{1st W in 5th game}) &= P(\text{4 loses}) \times P(\text{win}) \\
&= (P(\text{loss}))^4 \times P(\text{win}) \\
&= (1-p)^4 \times p \\
&= 0.4^4 \times 0.6 = 0.01536.
\end{aligned}$$

Geometric Random Variables

Geometric Distribution: If the probability of a success in any trial is p, the trials are independent, then the probability of finding the first success on the n^{th} trial is given by

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

Also

$$\begin{aligned}
\mu &= \frac{1}{p} \\
\sigma^2 &= \frac{1-p}{p^2} \\
\sigma &= \frac{\sqrt{1-p}}{p}
\end{aligned}$$

Intuition Behind μ

Think about μ: $\frac{1}{p}$ is the average number of trials we need until the first success.

So compare:

- Say $p = 0.5$. Then $\mu = \frac{1}{0.5} = 2$
- Say $p = 0.001$. Then $\mu = \frac{1}{0.001} = 1000$

In the first case, the probability of a success is lower, so we expect on average it will take more trials until the first success.

Yesterday's Quiz: Placebos

Question 1: Was Dr. Irving Kirsch arguing that anti-depressants are no better than placebos for everyone with depression?

Solution: No, while he argued that anti-depressants were no better than placebo for those with mild to moderate depression, he is of the opinion that there is clinical benefit for those who are severely depressed.

Yesterday's Quiz: Placebos

Question 2: What is Dr. Walter Brown's (bald guy from Yale) criticism of the way the FDA approves anti-depressants?

Solution: That all that is required are two clinical trials where the drug performs better than placebo, regardless of the number of trials with "negative results." Ex: say a drug performs better than placebo in 2 trials, but fails in 998 trials, it will still be approved by the FDA.

Lecture 12: Sampling Distributions & Standard Errors

Chapter 4.1

Goals for Today

Start Chapter 4: Arguably the most important chapter as it goes to the heart of what statistical inference is. Three important definitions today:

1. point estimate
2. sampling distribution
3. standard error

Point Estimates

Definition 1: Point estimates are functions of a random sample of n observations $x_1, \ldots, x_n$. They estimate the value of some unknown population parameter.

Ex: the sample mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \ldots + x_n}{n}$$

is a point estimate of the true population mean μ

Behavior of Point Estimates

Ex: Say we draw a random sample of size $n = 100$ from a large population that is normally distributed with $\mu = 5$ and $\sigma = 2$.

Two Important Questions:

1. Is $\overline{x}$ going to be exactly 5?
2. Say we get $\overline{x} = 5.025$. If we repeat this procedure: i.e. generate a new sample of size $n = 100$ and compute $\overline{x}$), will we get $\overline{x} = 5.025$?

We need to characterize this random error.

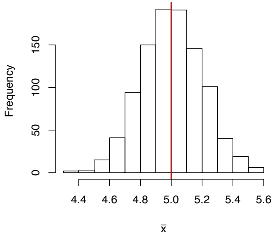
Behavior of Point Estimates

Let's repeat this procedure, say, 1000 times:

1st time	We get $\overline{x} = 4.831$
2nd time	We get $\overline{x} = 5.104$
3rd time	We get $\overline{x} = 4.965$
...	
1000th time	We get $\overline{x} = 4.957$

Sampling Distribution

This histogram is the 1000 instances of $\overline{x}$, where each $\overline{x}$ is based on a sample of $n = 100$. This is the sampling distribution of $\overline{x}$:



Sampling Distributions

Definition 2: the sampling distribution is the distribution of point estimates based on samples of fixed size n.

Every instance of a point estimate can be thought of as a draw from the sampling distribution.

If the sampling is representative (unbiased) then the sampling distribution will be centered around the true population parameter (in our case μ).

Sampling Distributions

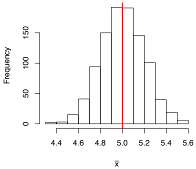
We can define the sampling distributions for any point estimate, not just $\overline{x}$:

- s
- the sample median
- etc.

We will only focus on sample means, including the sample proportion $\widehat{p}$.

Measure of Spread

What about spread? $[4.6, 5.4]$ contains roughly 95% of the data.



$$[\mu - 2SD, \mu + 2SD] = [4.6, 5.4]$$
$\Rightarrow$ length of interval is $4SD = 5.4 - 4.6$
$\Rightarrow$ $SD = 0.2$

Standard Errors

Definition 3: The standard error is the standard deviation of the sampling distribution of a point estimate.

It describes the uncertainty/variability associated with the point estimate. In other words, the "typical" error.

Confusing: the standard error is a specific kind of standard deviation.

Standard Error of $\bar{x}$

Given n independent observations from a population with standard deviation σ, the standard error of the sample mean is

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

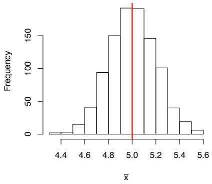
Rule of thumb for independence: You need a simple random sample consisting of less than 10% of the population.

Notice: $\sqrt{n}$ in the denominator: as n increases, SE decreases! This is why sample size matters.

Back to Histogram

Samples were of size $n = 100$ with $\sigma = 2$. We estimated that the SD of the sampling distribution was 0.2. Using the formula:

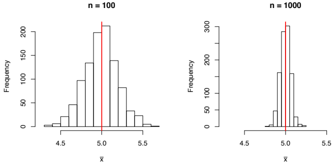
$$SE = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{100}} = \frac{2}{10} = 0.2$$



Standard Error of the Sample Mean $\overline{x}$

Compare 1000 instances of $\overline{x}$ when

- $n = 100$. $SE = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{100}} = 0.2$
- $n = 1000$. $SE = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{1000}} = 0.0632$. Smaller!



Both are "accurate", but the estimates on the right are "more precise."

Repeated Sampling

Popular question: What's up with this "1000" instances? Why would you take 1000 different samples of size n?

Answer: No, in practice you would not sample repeatedly: you do this only once for the largest n possible.

Rather the 1000 instances of $\overline{x}$ is a theoretical exercise to illustrate that $\overline{x}$'s are random and we characterize its randomness by its sampling distribution and its standard error.

Standard Error of the Sample Mean

In this example we knew σ; typically we won't. However, when

- $n \geq 30$
- the distribution of the population is not strongly skewed

we can use the point estimate of σ. i.e. plug in s in place of σ:

$$SE = \frac{s}{\sqrt{n}}$$

Example

Say in you take a simple random sample of 100 runners in a race and you are interested in their ages:

- $\overline{x} = 35.05$
- $s = 8.97$

Assuming that the 100 runners consist of less than 10% of the population, the standard error of $\overline{x}$ is

$$SE = \frac{s}{\sqrt{100}} = \frac{8.97}{10} = 0.897$$

Population Distribution vs Sampling Distribution

Recap

- Point estimates are based on a sample $x_1, \ldots, x_n$ and are used to estimate population parameters.
- The sampling distribution characterizes the (random) behavior of point estimates.
- The standard deviation of a sampling distribution is the standard error: it quantifies the uncertainty/variability of point estimates.

Next Time

- Confidence Intervals
- When quoting survey results, what does: "the results of this survey are estimated to be accurate within 3.1 percentage points, 19 times out of 20" mean?
- Big One: Central Limit Theorem

Lecture 13: Central Limit Theorem + Confidence Intervals

Chapter 4.4 + 4.2

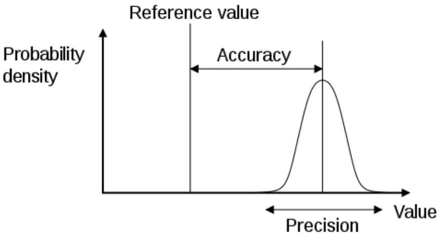
Goals for Today

- Discuss the Central Limit Theorem
- Introduce confidence intervals
- Interpretation

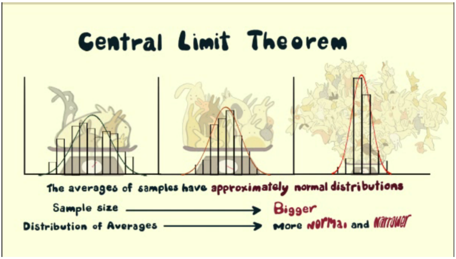
Recap

- Point estimates are based on a sample $x_1, \ldots, x_n$ and are used to estimate population parameters.
- The sampling distribution characterizes the (random) behavior of point estimates (like $\overline{x}$).
- The standard deviation of a sampling distribution is the standard error: it quantifies the uncertainty/variability of point estimates.

Illustrative Image of Sampling Distribution



Central Limit Theorem

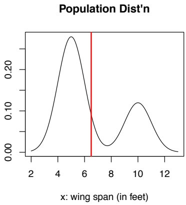


Central Limit Theorem

Question: Why do we care about the CLT?

Answer: We want the sampling distribution of $\overline{x}$ to be Normal irregardless of the shape of population distribution.

Example: The bimodal (population) distribution of dragon wing spans has a mean of 6.5:

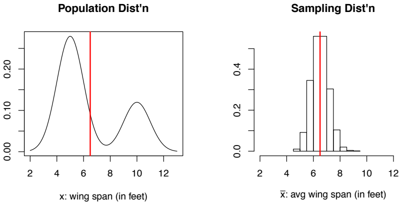


Central Limit Theorem

Question: Why do we care about the CLT?

Answer: We want the sampling distribution of $\overline{x}$ to be Normal irregardless of the shape of population distribution.

Example: The bimodal (population) distribution of dragon wing spans has a mean of 6.5:



Central Limit Theorem

Question: Why do we care that the sampling distribution of $\overline{x}$ is Normal?

Answer: So we can use the Normal table on p.409 of the book to calculate areas/percentiles/probabilities! We call this using the normal model.

	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Definition

For a sample $x_1, \ldots, x_n$ of independent observations, if n is "large" enough to counteract the skew of the population distribution, then the sampling distribution of $\overline{x}$ is approximately Normal with

- mean μ
- SD equal to the $SE = \frac{\sigma}{\sqrt{n}}$

Key: this holds for any population distribution, not just a normally distributed population.

Recall: If we don't know σ, we can plug in its point estimate s if the two conditions are satisfied.

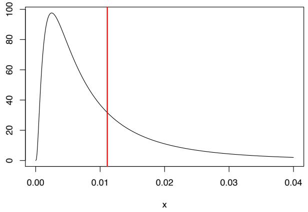
Conditions for the Normal Model

This translates to the following conditions to verify to be able to use the Normal model with s in place of σ, as stated in the book:

1. $n \leq 10\%$ of the population size.
   Comment: To ensure independence.
2. $n \geq 30$.
   Comment: This is a rule of thumb that works for most cases. You might need less, you might need more.
3. The population distribution is not strongly skewed.
   Comment: This is related 2. The larger the n, the more lenient we can be with the skew assumption.

Example of Skew vs n

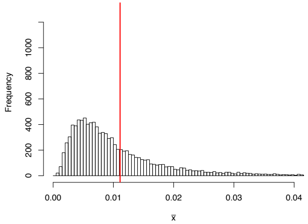
Let's say your observations come from the following very skewed population distribution with mean $\mu = 0.011109$.



This is where your individual observations x_i come from. Now compare 10000 values of $\overline{x}$'s based on different n: 2, 10, 30, 75.

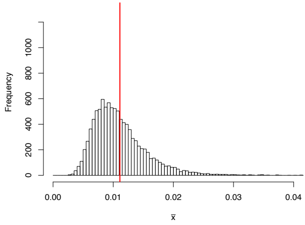
Example of Skew vs n

For 10000 values of $\overline{x}$ based on samples of size $n = 2$, the sampling distribution is:



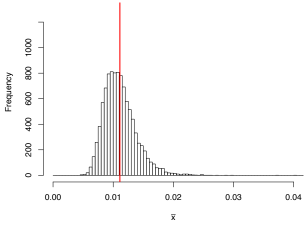
Example of Skew vs n

For 10000 values of $\overline{x}$ based on samples of size $n = 10$, the sampling distribution is:



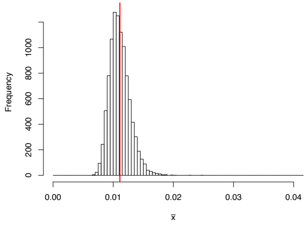
Example of Skew vs n

For 10000 values of $\overline{x}$ based on samples of size $n = 30$, the sampling distribution is:



Example of Skew vs n

For 10000 values of $\bar{x}$ based on samples of size $n = 75$, the sampling distribution is:



i.e. more normal and more narrow

Intuition of a Confidence Interval

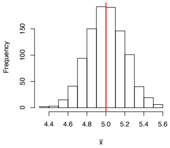
Our Goal: we want estimate a population parameter (e.g. μ). Analogy: imagine μ is a fish in a murky river that we want to capture:

Using just the point estimate:

Using a confidence interval:

Intuition of a Confidence Interval

Recall the example of 1000 instances of $\overline{x}$ based on $n = 100$. Each observation came from a population distribution that was Normal with $\mu = 5$ & $\sigma = 2$.



We observed the sampling distribution

- is centered at μ
- has spread $SE = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{100}} = 0.2$

Intuition of a Confidence Interval

A plausible range of values for the population parameter is called a confidence interval (CI). Since

- the SE is the standard deviation of the sampling distribution
- roughly 95% of the time $\overline{x}$ will be within 2 SE of μ if the sampling distribution is normal

If the interval spreads out 2 SE from $\overline{x}$, we can be roughly "95% confident" that we have captured the true parameter μ.

Intuition of a Confidence Interval

A 95% confidence interval for μ is (no more using rule of thumb $2 \times SD$):

$$\begin{aligned} \bar{x} \pm 1.96 SE &= [\bar{x} - 1.96 SE, \ \bar{x} + 1.96 SE] \\ &= \left[\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \ \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}\right] \end{aligned}$$

If we don't know σ, assuming the conditions hold, plug in s

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}} = \left[\bar{x} - 1.96 \frac{s}{\sqrt{n}}, \ \bar{x} + 1.96 \frac{s}{\sqrt{n}}\right]$$

Confidence Intervals

In general a confidence interval for μ will be

$$\bar{x} \pm z^* SE = [\bar{x} - z^* SE, \ \bar{x} + z^* SE]$$

where the critical value z^* is chosen to achieve the desired confidence.

Ex: For 95% confidence $z^* = 1.96$. For 99% confidence $z^* = 2.58$

Crucial: How to Interpret a Confidence Interval

The confidence interval has nothing to say about any particular calculated interval; it only pertains to the method used to construct the interval:

- Wrong, yet common, interpretation: There is a 95% chance that the C.I. captures the true population mean μ. The probability is 0 or 1: either it does or it doesn't.
- Correct, interpretation: If we were to repeat this sampling procedure 100 times, we expect 95 (i.e. 95%) of calculated C.I.'s to capture the true μ

Illustration: How to Interpret a Confidence Interval

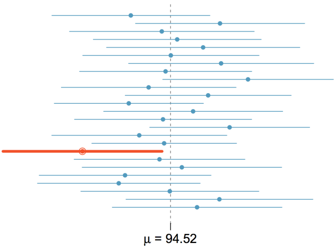
In Chapter 4 there is an example of finish times (in minutes) from the 2012 Cherry Blossom 10 mile run with $n = 16,924$ participants. In this case, we can compute the true population mean $\mu = 94.52$.

Say we take 25 (random) samples of size $n = 100$ and for each sample we compute:

- $\overline{x}$
- s
- and hence the 95% CI: $\left[\overline{x} - 1.96 \times \frac{s}{\sqrt{n}}, \ \overline{x} + 1.96 \times \frac{s}{\sqrt{n}}\right]$

How to Interpret a Confidence Interval

Of the 25 CI's based on 25 different samples of size $n = 100$, one of them (in red) did not capture the true population mean μ:



Political Polls

We polled the electorate and found that 45% of voters plan to vote for candidate X. The margin of error for this poll is ± 3.4 percentage points 19 times out of 20.

What does this mean?

- "19 times out of 20" indicates 95%
- The margin of error of $\pm 3.4\%$ indicates that 95% C.I. is:

$$45 \pm 3.4\% = [41.6, 48.4]$$

Intrepretation: the interpretation is not that there is a 95% chance that $[41.6, 48.4]$ captures the true %'age. Rather, that if we were to take 20 such polls, 19 of them would capture the true %'age.

Next Time

Hypothesis Testing: we can perform statistical tests on population parameters such as μ:

Define:

- Null and alternative hypotheses.
- Testing hypotheses using confidence intervals.
- Types of errors