

Stat 515: Introduction to Statistics

Chapter 1

Introduction

- William Cipolli
 - You can call me Will
- Graduate Student of Statistics
- BA in Mathematics
- BS in Comp Science
- MS in Statistics
- Previously worked at Travelers Insurance, Pitney Bowes, International Data Corp, Larch Lane Advisors, and FDTC.

FIRST HOMEWORK

- Your first homework assignment, due next class, is to write me a note that includes:
 - Your name
 - Your interests
 - Your major(s) and/or minor(s)
 - Why you're in college and what you want to do
 - What you expect to get out of this class and how you think it will help you in pursuit of your goals

Let's Look at the Syllabus

- Five minutes early is on time
 - If you're late because you got coffee, I want one!
 - I'll be taking attendance and keeping track of questions – this will make up your class participation grade

Cell Phones

- Cell phones should be on silent or vibrate. They should be looked at sparingly and I will answer it if it rings
- Excessive use will result in awkward, silent starring

Let's Look at the Syllabus

- Don't cheat
 - Don't cheat unless it will give you an A in every class for the rest of your college career (it won't)
 - The person you're cheating off may be wrong
- Use the internet!
 - You aren't so unique that no one else has ever come across the problem you're having
 - Google is your friend.

Let's Look at the Syllabus

- Descriptive Stats
- Probability
- Binomial Problems
- Sampling Distributions
- Confidence Intervals & Hypothesis test
- Regression and goodness of fit

Let's Look at the Syllabus

- 3 in-class exams (75 points each)
 - You are allowed to make up one at the end of the semester
- Final Exam (75 points) (Can not exempt)
- Homework (60 points)
- See Syllabus for grading breakdown

Before We Get to Statistics...

- You are all dumb.
- I am dumb
- We are all going to school to learn and become less dumb
- We should not be embarrassed to not understand something at first - it is a sign of intelligence and hard work to ask questions

Importance

- College courses are meant for **advanced** learning and requires you to connect yourself to the information – that means that **YOU** are the most important influence on your success
- You will only get out as much as you put into this class – your effort says a lot about you and the way you approach your work has lasting effects

What I Expect

1. Listen carefully when I lecture and go over examples and ask questions when something doesn't make sense to you
2. Be respectful to me, other students, your parents and the generations of people that didn't have access to an advanced education by being present and not distracting yourself or others

What I Expect

3. Reread the PowerPoint slides as you do your homework. If you aimlessly click around you will not learn anything. You should write out all the work when applicable and ask questions in class if you get stuck
4. Go to the tutoring center or come see me if you're lost or need extra help – all of this builds on itself.

What I Expect

5. Prepare for your lab sessions. Lecture is where I come in to introduce you to the material. Homework is where you **struggle** with and **learn** the material. Lab is where you come in and work.
6. Give and take in your lab sessions – all group members should take part in the lab. Do not be or let your partners be the student that just gets their name slapped on someone else's work.

Summary

- Be on time
- Ask questions: be an active learner
- Give your opinions – they're important
- Be respectful of other students
- Be determined and purposeful
- Keep up with lectures and your homework

STOLEN SLIDES

- The following nineteen slides were stolen from Roy Bower, Master B.



Statistics in the Real World

Sources:

- **Glamour Magazine**
- **University Studies**
- **CDCP**
- **www.cracked.com**



**KEEP
CALM
AND**

**CITE YOUR
SOURCES**

Random Statistics



**The Average Person
Spends 5 Years
Doing This:**

**Waiting in
Line**



1 in 3 Men Don't do This:




**Wash Their
Hands After
Using the
Bathroom**

**43% of Pilots Admit to
Doing This:**



Falling Asleep While Flying

A graphic of a mobile phone with a silver and black finish. The screen is blue and displays text. At the bottom of the phone, there are three icons: a green call icon on the left, a silver square home button in the center, and a red call icon on the right.

**The Average
Person Does
This 1,140
Times Per Year**

**Make A
Phone Call**

On Average, More Than 10 People Are
Killed **by this** Every Year:



Vending
Machines

Facebook Statistics



300: The Average Number of **What** a Teenager Has on Facebook



Friends



facebook

95% of People on Facebook, on Average, do this Daily:

Bloom

Login to Facebook to enjoy the full functionality of Bloom. If you don't want this to happen, go to the normal Facebook login page.

Email:

Password:

Optional: Save my login info to avoid logging in to Facebook again to use this application.

Login

[Forgot your password?](#)

Login!

Security Note: After login, you should never provide your password to an outside application. Facebook does not provide your contact information to Bloom.

Select Photos

Home

Desktop

Pictures

Macintosh HD

Servers

NO NAME

AIM® Buddy Icons

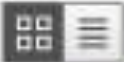
iChat Icons

iPhone Pics

iPhoto Library

Photo Booth

350 Million: The Average Number of...



Select All

Showing 10 photos



IMG_0277.jpg



IMG_0279.jpg



IMG_0280.jpg



IMG_0281.JPG



IMG_0285.JPG

Photos Uploaded in One Day



Selected Photos

0 selected

Use Selected Photos

Cancel

Sex Statistics



20,160 (14): The Average Number of Minutes (days) a Person Spends Doing This:



Kissing

**On Average, Nearly 50% of Men Think
Their **What** is too Small**

No Picture Necessary

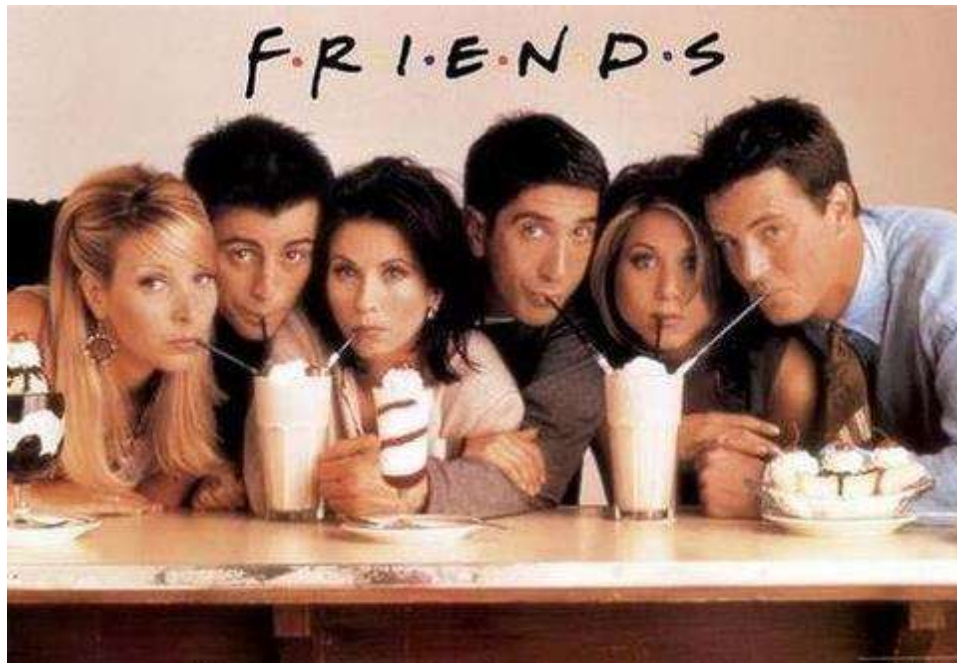
**The Average
Number of Sex
Partners for One
Woman in One
Lifetime???**

4

**The Average
Number of Sex
Partners for One
Man in One
Lifetime???**

7

2 in 3 College Students Have Been in **What Type** of Relationship?



Friends With Benefits



**The Average Age
at which Males
Lose Their
Virginity?**

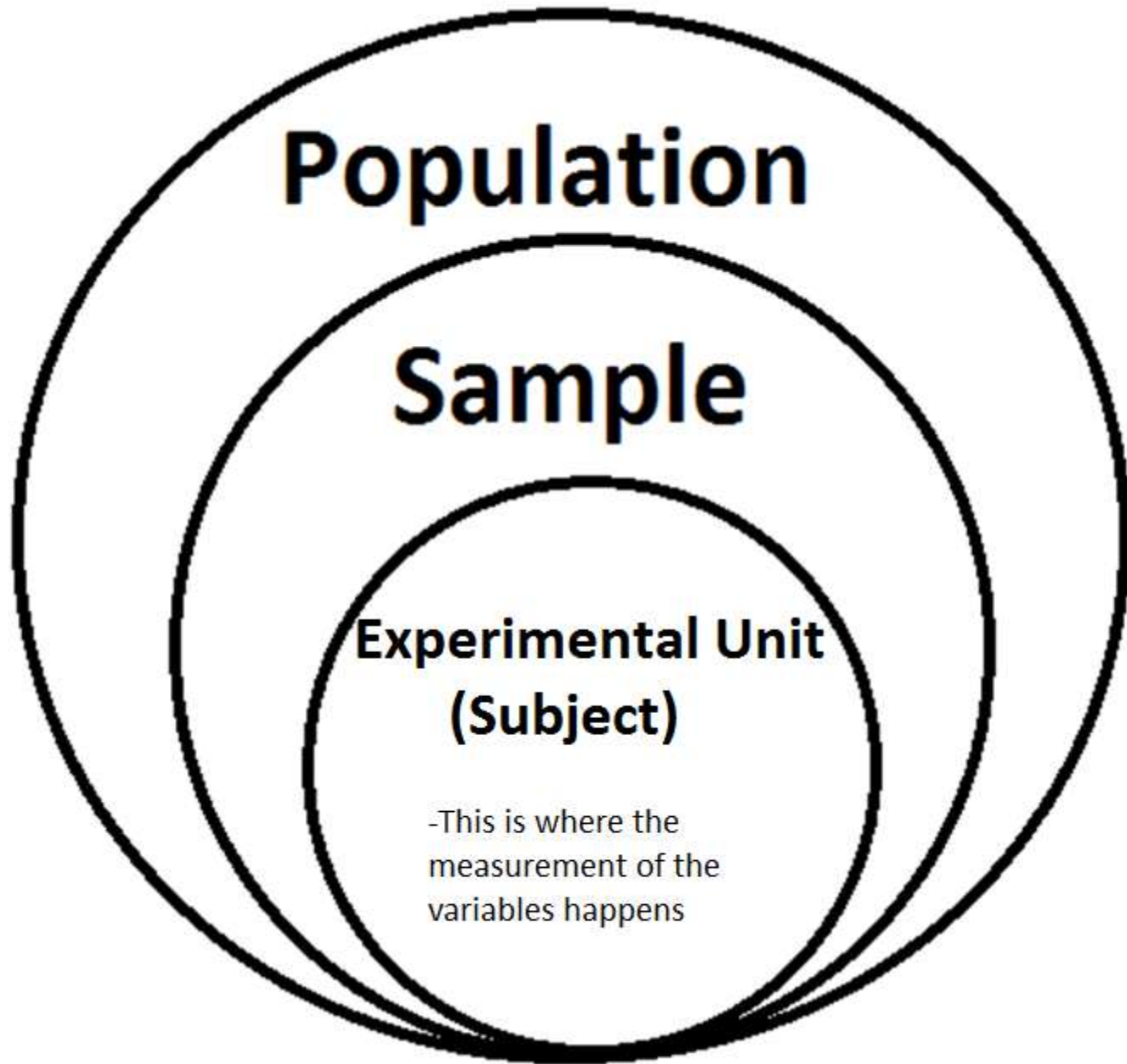
16.9 Years

**The Average Age
at which Females
Lose Their
Virginity?**

17.4 Years

Definitions

- **Population:** the set of all subjects of interest
 - US population, schools in SC, the group we look at
 - Think of this as where we took our sample from
- **Sample:** the set of subjects that we have data for
 - A subset of the population for which we know the variable
- **Experimental Unit (or Subject):** entities that we measure in a study
 - People, schools, the person or thing we look at
- **Variable:** any characteristic that is observed for the subject
 - Height, class size, whatever we're measuring



Definitions 2

- **Statistic:** numerical summary of a sample
 - Mean(\bar{x}), proportion(\hat{p}), median, mode, standard deviation(s), variance(s^2), Q1, Q3, IQR, etc.
 - We use US alphabet letters to denote these
- **Parameter:** numerical summary of a population
 - Mean(μ_x), proportion(ρ), median, mode, standard deviation(σ), variance(σ^2), Q1, Q3, IQR, etc.
 - We usually don't know these values
 - We use Greek letters to denote these

Think About It

- **Statistic** starts with an 's' so it's talking about the sample
- **Parameter** starts with a 'p' so it's talking about the population
- There's an important connection here – a large aspect of statistics is **statistical inference** which is trying to tell information about the parameter, which we don't know, by using the sample statistics which we are able to compute

Why Do We Use Statistics?

- **Descriptive Statistics:** This is when we use statistics to make the leap from massive datasets to what they tell us.
 - Sometimes scientists, politicians, computer engineers etc. have thousands or millions of rows of data in Excel and they want to draw conclusions
 - Here, they could use **descriptive statistics** and charts to summarize thousands of rows with just a few numbers

Why Do We Use Statistics?

- **Descriptive Statistics:**
 - We have a dataset for a **sample** or **population**
 - We have one or more **variables** measured for each **experimental unit (or subject)** that we would like to summarize
 - We use **descriptive statistics** and charts to summarize the data set to identify patterns, trends etc.

Why Do We Use Statistics?

- **Inferential Statistics:** This is when we use the descriptive statistics of a sample to make estimates or predictions about a population
 - Sometimes scientists, politicians, computer engineers etc. have a small **sample** and want to draw conclusions about the larger **population**
 - Here, they could use **descriptive statistics** and statistical methodology to estimate the **population parameters** based off of the **sample statistics** which requires a **measure of reliability** which quantifies the uncertainty of our estimate. Think “plus or minus.”

Why Do We Use Statistics?

- **Inferential Statistics:**

- There's a **population** of interest often too big or too expensive to measure
- We have a **sample**, a subset of the **population**
- We have one or more **variables** measured for each **experimental unit (or subject)** that we would like to summarize
- We use the **descriptive statistics** of the **sample** to make estimates about the **population**
- We report a **measure of reliability** of our inference, which is generally dependent on how large our sample is

Watch These Videos!

- Why Statistics*
 - <https://www.youtube.com/watch?v=yxXsPc0bphQ>
- Target uses statistics
 - <https://www.youtube.com/watch?v=jgsdQxTv5kY>
- Changes in America
 - <https://www.youtube.com/watch?v=ke52L1k9VzQ>
- Wealth Distribution*
 - <https://www.youtube.com/watch?v=krwdJ6DyafQ>

A Great Example of Why we Use Statistics for Watching at Home

- <https://www.youtube.com/watch?v=hVimVzgtD6w>

Example

- The 2012 South Carolina Republican Primary was held on January 21st. Newt Gingrich ended up winning the primary with 244,065 of 603,770 votes, 40.42% of South Carolina Primary voters.
 - **Population:** South Carolina Residents
 - **Sample:** Residents that came out to vote
 - **Experimental Unit(Subject):** Individual Voters
 - **Variable:** Which candidate they prefer
 - **Statistic:** $\hat{p} = \frac{244,065}{603,770} = .4042 = 40.42\%$

Example

- **Population:** South Carolina Residents
- **Sample:** Residents that came out to vote
- **Experimental Unit(Subject):** Individual Voters
- **Variable:** Which candidate they prefer
- **Statistic:** $\hat{p} = \frac{244,065}{603,770} = .4042 = 40.42\%$
 - The statistic 40.42%, a proportion, gives us information that suggests Newt Gingrich was preferred to other presidential hopefuls for all South Carolina Residents based off of those that voted
 - We will talk about how to find a reliability measure later in the semester

Example 2

- A '93 survey of 4,977 found that 3.5% of households surveyed had used a gun for “protection” in the last year
 - **Population:** American households
 - **Sample:** 4,977 households sampled
 - **Experimental Unit(Subject):** Individual households
 - **Variable:** did they've used a gun for “protection?”
 - **Statistic:** $\hat{p} = .035 = 3.5\%$

Example 2

- **Population:** American households
- **Sample:** 4,977 households sampled
- **Experimental Unit(Subject):** Individual households
- **Variable:** Which candidate they prefer
- **Statistic:** $\hat{p} = .035 = 3.5\%$
 - The statistic 3.5%, a proportion, can be applied to the population to give us an idea about the number of households that use their gun for protection in 1992: 3.5% of the population is **1,029,615**
 - We will talk about how to find a reliability measure later in the semester

Example 3

- A poll surveyed 1,772 registered voters, 92 percent of which supported background checks, the Quinnipiac University telephone poll showed with a margin of error of plus or minus 2.3 percentage points
 - **Population:** Registered voters
 - **Sample:** 1,772 registered voters sampled
 - **Experimental Unit(Subject):** Individual registered voters
 - **Variable:** did they support background checks
 - **Statistic:** $\hat{p} = .92 = 92\%$

Example 3

- **Population:** Registered voters
- **Sample:** 1,772 registered voters sampled
- **Experimental Unit(Subject):** Individual registered voters
- **Variable:** did they support background checks
- **Statistic:** $\hat{p} = .92 = 92\%$
 - The statistic 92%, a proportion, can be applied to the population to give us an idea about the number of registered voters that support background checks.
 - The margin of error of 2.3% gives us an idea about the possible error in applying this to the population. In fact, we can expect between 89.7% to 94.3% of the population to support background checks.

Example 4

- CNBC reports that the average price of a gallon of gas at United States gas stations was \$2.63 on Tuesday May 5th, 2015 according to reported prices on GasBuddy.com
 - **Population:** United States gas stations
 - **Sample:** United States gas stations reported on GasBuddy.com
 - **Experimental Unit(Subject):** Individual Stations
 - **Variable:** Price of a gallon of gas
 - **Statistic:** $\bar{x} = 2.63$

Example 4

- **Population:** United States gas stations
- **Sample:** United States gas stations reported on GasBuddy.com
- **Experimental Unit(Subject):** Individual Stations
- **Variable:** Price of a gallon of gas
- **Statistic:** $\bar{x} = 2.63$
 - The statistic \$2.63, a mean, can be applied to the population to give us an idea about the average price of gas at all stations in the United States.
 - We will talk about how to find a reliability measure later in the semester

Example 5

- The Wall Street Journal reports that on average, Panthers fans make 6.6 mistakes per 100 words placing them 19th among other teams, ages ahead of my poor Patriots. These results are based on a sample of 150 comments on each teams' website.
 - **Population:** All Panthers fans
 - **Sample:** Panthers fans that wrote 150 selected comments
 - **Experimental Unit(Subject):** Individual Panthers fans
 - **Variable:** Mean number of grammar mistakes per 100 words
 - **Statistic:** $\bar{x} = 6.6$

Example 5

- **Population:** All Panthers fans
- **Sample:** Panthers fans that wrote 150 selected comments
- **Experimental Unit(Subject):** Individual Panthers fans
- **Variable:** Mean number of grammar mistakes per 100 words
- **Statistic:** $\bar{x} = 6.6$
 - The statistic 6.6 mistakes per 100 words, a mean, can be applied to the population to give us an idea about the average grammar capabilities of Panthers fans.
 - We will talk about how to find a reliability measure later in the semester

Example 6

- Even though Pluto is no longer a full-fledged planet after it was downgraded to a dwarf planet it's still of interest. The gravitational approach allows scientists to use the gravitational approach to estimate Pluto's diameter at 1,471 miles, plus or minus five miles, by taking repeated measurements.
 - **Population:** All possible measurements
 - For now, we assume the population mean is the actual diameter
 - **Sample:** The collection of measurements taken
 - **Experimental Unit(Subject):** Each individual measurement
 - **Variable:** Diameter measurement of Pluto
 - **Statistic:** $\bar{x} = 1,471$

Example 6

- **Population:** All possible measurements
- **Sample:** The collection of measurements taken
- **Experimental Unit(Subject):** Each individual measurement
- **Variable:** Diameter measurement of Pluto
- **Statistic:** $\bar{x} = 1,471$
 - The statistic 1,471, a mean, can be applied to the population to give us an idea about the actual diameter of Pluto
 - The margin of error of 5 miles gives us an idea about the possible error in applying this to the population. In fact, we can expect the actual diameter to be between 1,467 to 1,475 miles.

Types of Variables

- Recall that a **variable** is any characteristic that is observed for the subject – this is what we're interested in describing or making inference on.
- What kinds of variables can you think of about a human?
 - Height, weight, gender, hair color, eye color, age, GPA, ethnicity, origin, favorite color, major, etc.

Types of Variables

- **Qualitative(Categorical):** Observations that belong to a set of categories
 - Examples: gender, hair color, eye color, ethnicity, origin, favorite color, major, etc.
- **Quantitative:** Observations that take on numerical values
 - Examples: Height, weight, age, GPA, etc.

Types of Variables

- **Quantitative:** Observations that take on numerical values
 - **Discrete:** measured by a whole number
 - Examples: Number of books, children, money, etc
 - **Continuous:** measured on an interval
 - Examples: Height, weight, age, GPA, etc.
 - Note: These are often measured as a discrete variable

How to Compare Discrete and Continuous: Continuous

- If you think about time: going from 1 minute to 2 minutes we have to hit all of the times between (seconds, jiffies, etc.)
- If you think of height: growing from 5' to 6' we have to be every height in between 5' and 6' and (inches, cm, mm, etc.)
- If you think of weight: going from 150lbs to 140lbs we have to be every weight between 140 and 150 (oz and g, etc)

How to Compare Discrete and Continuous: Continuous

- **Time:** hours, minutes, seconds, deci-second, jiffy, centi-second, millisecond, microsecond, nano second, planck time unit, etc.
- **Height:** meter, deci-meter, centi-meter, millimeter, micrometer, nanometer, picometer, femtometer, attometer, zeptometer, yoctometer, etc.
- **Weight:** gram, deci-gram, centi-gram, milligram, microgram, nanogram, picogram, femtogram, attogram, zeptogram, yoctogram, etc.

How to Compare Discrete and Continuous: Discrete

- If you think about the number of books, children, money, etc we jump from one number to the next.
 - We can't have half of a book we jump from 0 to 1
 - We can't have 1.5 children we jump from 1 to 2
 - We can't have half a cent we jump from \$0 to \$.01

How to Compare Discrete and Continuous: Discrete

- The big difference here is that we can keep coming up with smaller units for the **continuous** case and we stop at some point for the **discrete** case
 - With children we stop at whole numbers
 - With books we stop at whole numbers
 - With money we stop at pennies

How to Compare Discrete and Continuous

- It should be noted that when we talk about **continuous** variables, we stop somewhere so we are **measuring them discretely**
 - **Example discrete measurements of continuous variables**
 - **Time:** hours, minutes, seconds
 - **Height:** meter, deci-meter, centi-meter, millimeter
 - **Weight:** gram, deci-gram, centi-gram, milligram

Where Does Our Data Come From?

- We often hear reports about polls, research etc. on the news and the sources often a footnote are very important!
- We consider the following:
 1. Published Data
 2. Data from Designed Experiments
 3. Data from an Observational Study

Published Data

- This is just as it sounds. Many people have already completed designed experiments and observational studies and, thanks to the internet, have made their datasets publicly available.
 - <https://www.kaggle.com/>
 - Let's companies and researchers post datasets for other researchers to work with in hopes of finding new patterns
 - <http://www.pewresearch.org/data/download-datasets/>
 - PEW posts many of their datasets for secondary analysis
 - <https://aws.amazon.com/datasets>
 - Even Amazon posts public datasets from many categories

Observational Versus Designed

- An **observational study** measures the response variable without attempting to influence the value of either the response or explanatory variables.
- A **designed study** occurs when a researcher assigns the individuals or subjects into groups and intentionally affects their explanatory variables (think treatments)

Example

- C. Myrray Parkes headed a study of 4,486 men of 55 years of age and older who had their wives die in 1957. For up to nine years, these widowers were tracked and 213 died during the first six months – that's about 5%.
- This experiment is a **observational study**. C Myrray Parkes didn't murder 4,486 women in 1957 just to do this study.

Example

- Note: It was found that 40% above the expected rate for married men of the same age died during the first six months of bereavement. Thereafter, the mortality rate fell gradually to that of married men and remained at the same level.

Example 2

- Herbet Benson, MD headed a study in 2005 to see if intercessory prayer influenced recovery from bypass surgery. There were three groups in the study:
 1. Those being prayed for that didn't know
 2. Those being prayed for that did know
 3. Those not being prayed for
- This is a **designed study** because the researchers assigned different patients to different groups; they controlled who was prayed for and who wasn't instead of just observing and asking the families whether or not they had friends and families praying for the patient.

Example 2

Note: Intercessory prayer itself had no effect on complication-free recovery but knowing they were receiving intercessory prayer was associated with a higher incidence of complications.

Where We Get Our Sample is Important!

- Regardless of which of the three options we get our data from when delivering descriptive or inferential statistics it is paramount that we have a **representative sample**
- The idea is that the sample we use should accurately portray the description of the population we're trying to talk about; we don't want to compare apples to oranges!

Example

- Suppose we want to measure the age of lung cancer instances in smokers and non-smokers and consider the sample made up of smokers in their 80's and 90's and non smokers of any age.
- Overall, the average age of a lung cancer patient is about 70 years old. With the proposed sample above the minimum average age of lung cancer instances is 80, well above the overall average; this could lead to us saying that cigarettes are good for you! We note, however, that people start smoking much earlier than their 80's and 90's so our sample is **not representative**. We say this type of 'non-representative' sample suffers from **selection bias**.

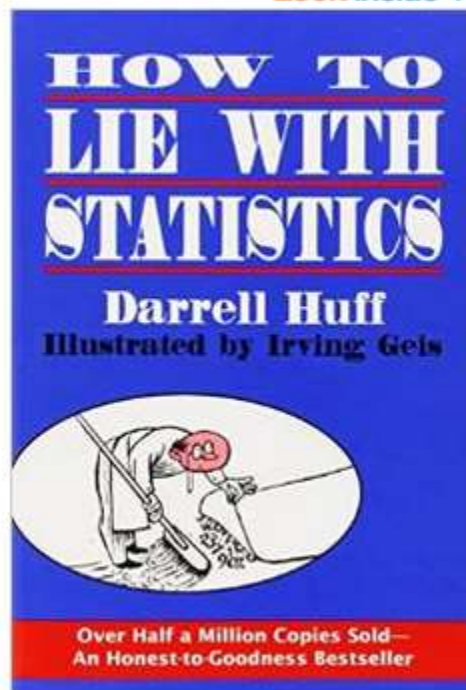
Example

- Suppose we want to measure the age of lung cancer instances in smokers and non-smokers and consider a sample of cigarette smokers and non-cigarette smokers both with wide age ranges and backgrounds that volunteered to be part of the study.
- With the proposed sample above we have comparable experimental units (or subjects) in both groups and because of their wide age ranges and backgrounds we can argue that this sample is **representative** of the population and that the results can be generalized to the public.

[Back to search results for "how to lie with statistics"](#)

[How to Lie with Statistics](#) and over one million other books are available for **Amazon Kindle**. [Learn more](#)

Look inside ↴



↻ Flip to back

How to Lie with Statistics Paperback – October 17, 1993

by [Darrell Huff](#) (Author), [Irving Geis](#) (Illustrator)

★★★★☆ 274 customer reviews

ISBN-13: 978-0393310726

ISBN-10: 0393310728

Edition: Reissue

Buy New

Price: \$10.26 Prime

67 New from \$6.96 | 115 Used from \$2.95

	Amazon Price	New from	Used from
Kindle + +	\$7.99	—	—
▶ Hardcover	—	\$193.83	\$44.10
▶ Paperback	\$10.26	\$6.96	\$2.95
▶ Unknown Binding	—	\$29.95	\$6.99



Get up to 80% Back
When You Sell Us Your Books

[Learn more](#)

Statistical Thinking

- What is the research question?
 - What is the **population** of interest?
 - What is the **variable** of interest?
 - How will the **sample** be selected?
 - How will the data be collected?
-
- Essentially, what's the best way of going about using statistics to solve your problem?

Desired Sample Selection

- **Simple Random Sample** – the sample is chosen in such a way that every subject is equally likely to be selected for the study
 - We prefer this method above all else
 - Problem: Sometimes this isn't feasible
 - We don't always have access to every subject in the population
 - It's not always the case that everyone is willing to participate in the study

Bias

- **Bias** is when the results of a sample are not representative of a population

Sources of Bias

- **Selection Bias** means that the sampling technique favored one part of the population over the other, i.e. parts of the population have no chance of being selected for the sample
 - Experiments using a convenience sample usually suffers from selection bias; consider if I took students at USC as a ‘random’ sample of Americans because I have convenient access to students as a GA. In this case I would be leaving out all non-students from the sample leading to selection bias

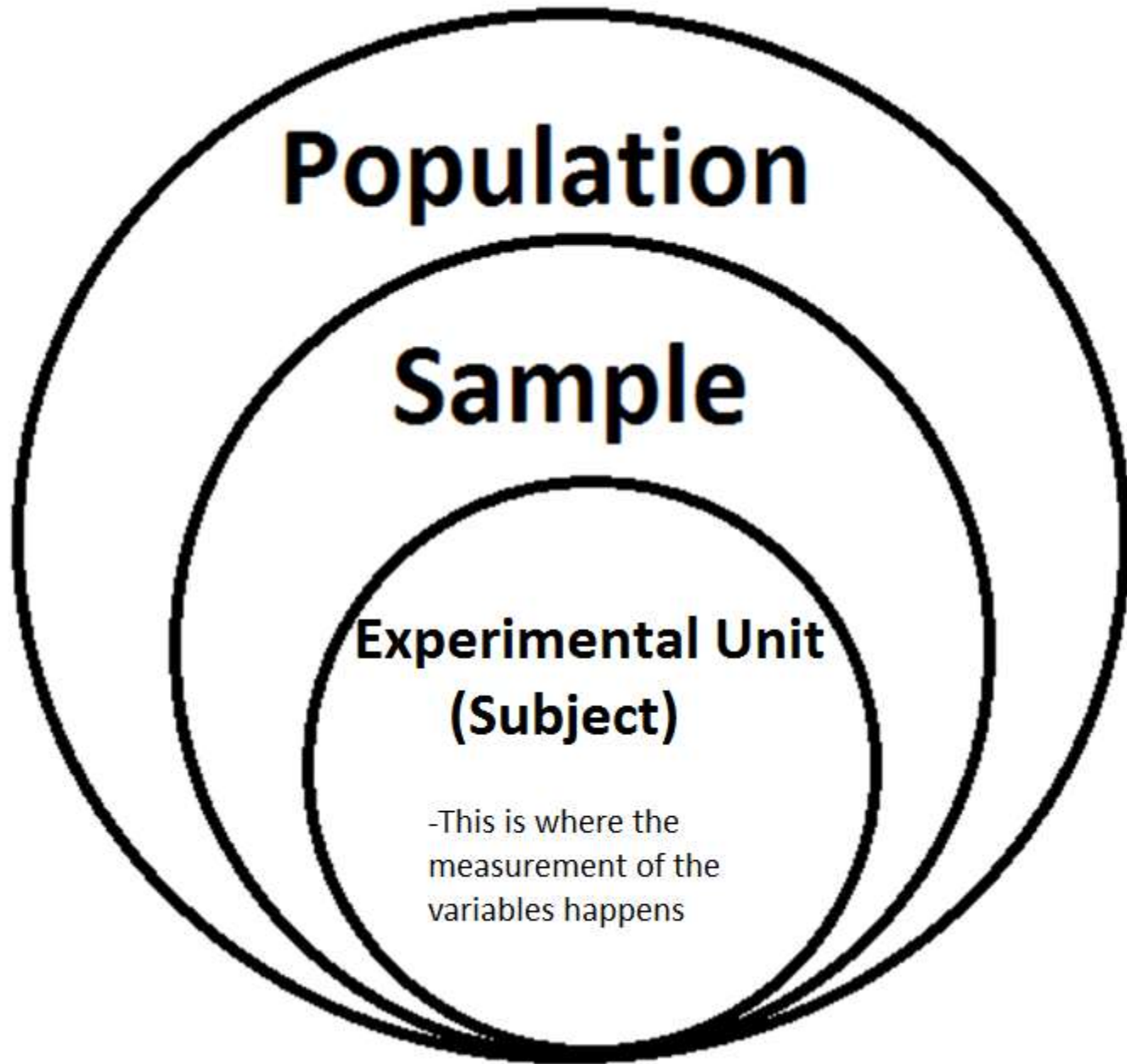
Sources of Bias

- **Nonresponse Bias** occurs when individuals that respond to the survey answer differently than the potential answers of those who did not answer.
 - How many times do you participate in the optional survey at when using the web or calling customer service? Maybe only the people that are angry respond; this leaves all the satisfied customers out of the sample and reflects higher dissatisfaction than it might have otherwise.

Sources of Bias

- **Measurement Error Bias** occurs when we have inaccurate data
 - Often survey data comes from interviewing individuals.
 - Sometimes the content of a survey is sensitive and participants feel the need to lie or omit details in their answer, interviewers can ask leading or confusing questions, or interviewers could have language difficulties
 - Sometimes our tools fail us and can produce faulty measurements
 - Sometimes we don't use the right or 'best' tool for the job

Summaries!



Definitions

- **Population:** the set of all subjects of interest
 - US population, schools in SC, the group we look at
 - Think of this as where we took our sample from
- **Sample:** the set of subjects that we have data for
 - A subset of the population for which we know the variable
- **Experimental Unit (or Subject):** entities that we measure in a study
 - People, schools, the person or thing we look at
- **Variable:** any characteristic that is observed for the subject
 - Height, class size, whatever we're measuring

Definitions 2

- **Statistic:** numerical summary of a sample
 - Mean(\bar{x}), proportion(\hat{p}), median, mode, standard deviation(s^2), variance(s), Q1, Q3, IQR, etc.
 - We use US alphabet letters to denote these
- **Parameter:** numerical summary of a population
 - Mean(μ_x), proportion(ρ), median, mode, standard deviation(σ), variance(σ^2), Q1, Q3, IQR, etc.
 - We usually don't know these values
 - We use Greek letters to denote these

Why Do We Use Statistics?

- **Descriptive Statistics:** This is when we use statistics to make the leap from massive datasets to what they tell us.
 - Sometimes scientists, politicians, computer engineers etc. have thousands or millions of rows of data in Excel and they want to draw conclusions
 - Here, they could use **descriptive statistics** and charts to summarize thousands of rows with just a few numbers

Why Do We Use Statistics?

- **Inferential Statistics:** This is when we use the descriptive statistics of a sample to make estimates or predictions about a population
 - Sometimes scientists, politicians, computer engineers etc. have a small **sample** and want to draw conclusions about the larger **population**
 - Here, they could use **descriptive statistics** and statistical methodology to estimate the **population parameters** based off of the **sample statistics** which requires a **measure of reliability** which quantifies the uncertainty of our estimate. Think “plus or minus.”

Types of Variables

- **Qualitative(Categorical):** Observations that belong to a set of categories
 - Examples: gender, hair color, eye color, ethnicity, origin, favorite color, major, etc.
- **Quantitative:** Observations that take on numerical values
 - Examples: Height, weight, age, GPA, etc.

Observational Versus Designed

- An **observational study** measures the response variable without attempting to influence the value of either the response or explanatory variables.
- A **designed study** occurs when a researcher assigns the individuals or subjects into groups and intentionally affects their explanatory variables (think treatments)

Desired Sample Selection

- **Simple Random Sample** – the sample is chosen in such a way that every subject is equally likely to be selected for the study
 - We prefer this method above all else
 - Problem: Sometimes this isn't feasible
 - We don't always have access to every subject in the population
 - It's not always the case that everyone is willing to participate in the study

Sources of Bias

- **Selection Bias** means that the sampling technique favored one part of the population over the other, i.e. parts of the population have no chance of being selected for the sample
 - Experiments using a convenience sample usually suffers from selection bias; consider if I took students at USC as a ‘random’ sample of Americans because I have convenient access to students as a GA. In this case I would be leaving out all non-students from the sample leading to selection bias

Sources of Bias

- **Nonresponse Bias** occurs when individuals that respond to the survey answer differently than the potential answers of those who did not answer.
 - How many times do you participate in the optional survey at when using the web or calling customer service? Maybe only the people that are angry respond; this leaves all the satisfied customers out of the sample and reflects higher dissatisfaction than it might have otherwise.

Sources of Bias

- **Measurement Error Bias** occurs when we have inaccurate data
 - Often survey data comes from interviewing individuals.
 - Sometimes the content of a survey is sensitive and participants feel the need to lie or omit details in their answer, interviewers can ask leading or confusing questions, or interviewers could have language difficulties
 - Sometimes our tools fail us and can produce faulty measurements
 - Sometimes we don't use the right or 'best' tool for the job