Ch. 1 & 2

ECON 3406

Dr. Josh Martin

"Compared to What?"

- **Applied statistics** involves using numerical data to assess the likelihood that competing claims about reality are valid
  - *(This is my working definition, tailored to this course)*
- Though the field can be complex, at its core, statistical inference relies heavily on **careful comparisons** between groups, treatments, or conditions

Stents and the Risk of Stroke

- Consider a clinical experiment designed to test the effectiveness of **stents** in preventing strokes in patients at risk
  - **Stents:** Small mesh tubes inserted into blood vessels to keep arteries open, aiding recovery after cardiac events and reducing the risk of future heart attacks
  - **Stroke:** A medical condition where blood flow to part of the brain is interrupted or reduced, causing brain cells to die – leading to potential loss of function, disability, or death
  - *Plausible causal pathway:* Stents improve blood flow, potentially reducing stroke risk by preventing blockages or clots

Randomization and Treatment / Control Groups

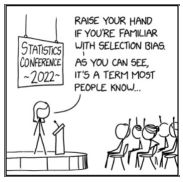
- Patients volunteering for the study were **randomly assigned** to one of two groups to ensure comparability
- **Treatment group:** Received stents **plus** medical management (MM)
  - Medical management includes: medications, controlling risk factors (like blood pressure, cholesterol), and lifestyle modifications (diet, exercise)
- **Control group:** Received only medical management, without stents

Importance of Experimental Design

- Experimental design rests on two key pillars:
  1. **Random assignment of treatment** to groups
- **Randomization** enhances **internal validity** by minimizing selection bias and confounding
  - **Internal validity:** The degree to which a study reliably establishes a cause-and-effect relationship by ruling out alternative explanations
  - **External validity:** How well study findings generalize to other populations or settings
  - **Confounding factors:** Variables that influence both the treatment and the outcome

Selection Bias

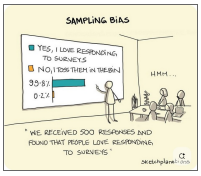
- **Selection bias**: Occurs when the groups being compared differ in ways other than the treatment
  - Often attributable to "selection" into one of the comparison groups
    - Also known as treatment/control group "contamination"
  - "Don't go to hospitals, more people die there than anywhere else"



Selection Bias

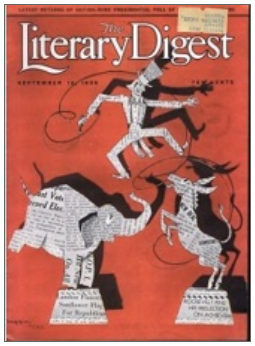
- Example: Those with greater health literacy (HL) demand (and thus receive) a stent
  - Those with greater HL are more likely to change their lifestyle after receiving their stent
  - Thus, “treatment” estimates are partially attributable to lifestyle changes *and* treatment

Sampling Bias is a Form of Selection Bias

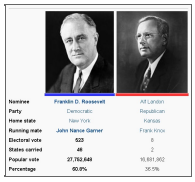


Sampling Bias

- In 1936, The Literary Digest polled about 10 million Americans on the outcome of the presidential election
  - 2.4 million response
  - Prediction: Landon would be overwhelming winner; FDR < 43%

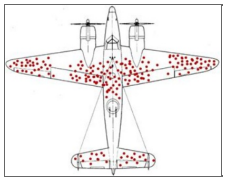


Sampling Bias



- Only polled:
  - its own readers
  - registered automobile owners or telephone users
- No number of observations can eliminate bias

Survivorship Bias



- Survivorship bias: when the outcome is related to the time(s) at which the data is measured
  - Suppose that receiving treatment (a stent) increases one's mobility
  - Thus, "treatment" estimates are partially attributable to reductions in the likelihood that members of the control group have difficulty making it to the doctor to be observed
- "My uncle smokes three packs a day and he's in perfectly good health"

Blinding and placebos

- To circumvent psychological effects among study participants, researchers prefer tha patients not know which group they are in
  - **Blind**: when researchers keep patients uninformed about their treatment
  - **Placebo**: fake treatment
  - **Placebo effect**: when control group participants experience “treatment” effects

Placebo Effect

Placebo Effect

- Placebo “treatment” (= being placed in the control group) can trigger physiological factors, threatening the experimental design by impacting the outcome of interest
  - blood pressure
  - heart rate
  - release of pain-reducing chemicals and stress hormones (endorphins, adrenaline)
- This “contamination” of the control group can, depending upon the outcome studied, lead to a Type II statistical error

Type I and II Statistical Errors

- **Type I (false positive):** Reject a true null (you claim an effect/bias that isn't there)
  - Example: Concluding the coin is biased when it is actually fair (p = 0.5)
- **Type II (false negative):** Fail to reject a false null (you miss a real effect/bias)
  - Example: Concluding the coin is fair when it's actually p = 0.6 heads

Importance of Experimental Design

- Experimental design rests on two key pillars:
  2. Comparison groups
- The control group helps to ensure that the estimated impact of treatment can be attributable to treatment rather than secular trends in unrelated external factors at the time of observation.
  - Control groups act as a reference point against which we can measure the medical impact of stents in the treatment group.

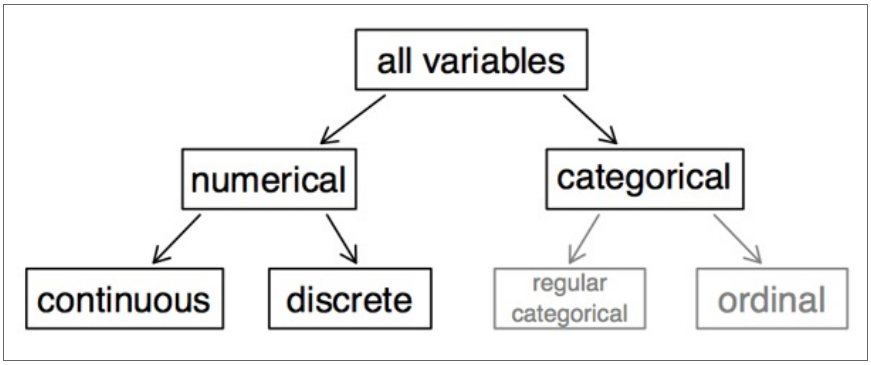
Importance of Experimental Design

- Example: Suppose that money increases longevity and that real incomes are quickly rising during the survey period.
  - Randomization increases the likelihood that the mean change in income will be similar across comparison groups.
  - A control group allows one to “net out” the impact of income on longevity to isolate the treatment effect.

A data table

patient	group	stroke_0_30	stroke_0_365
1	treatment	1	1
2	treatment	1	1
3	treatment	1	1
4	treatment	1	1
5	treatment	1	1
6	treatment	1	1

Types of variables



Types of variables

- **Numerical**: can be measured or counted, amenable to mathematical manipulation
  - **Discrete**: countable, requiring “jumps”
  - **Continuous**: measurable, continuous
- **Categorical**: groups, often described using labels or names
  - **Nominal**: No order
  - **Ordinal**: Order

Classroom survey example

- A survey was conducted on students in an introductory statistics course. Below are a few of the questions on the survey, and the corresponding variables the data from the responses were stored in:
  - gender: What is your gender?
  - intro_extra: Are you an introvert or an extrovert?
  - sleep: How many hours do you sleep at night, on average?
  - bedtime: What time do you usually go to bed?
  - countries: How many countries have you visited?
  - dread: On a scale of 1-5, how much do you dread being here?

Variable types example

- gender: nominal
- sleep: continuous
- bedtime: ordinal
- countries: discrete
- dread: ordinal (potentially discrete)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

Variable type practice

What type of variable is a telephone area code?

a. Continuous
b. Discrete
c. Nominal
d. Ordinal

Variable type practice

What type of variable is a telephone area code?

a. Continuous
b. Discrete
c. **Nominal**
d. Ordinal

A data summary

- “Tabulating” the data into “descriptive statistics” allows us to consider all of the data at once
  - A summary statistic is a single number, based on the sample, that summarizes a large amount of data

	0–30 days		0–365 days	
	stroke	no event	stroke	no event
treatment	33	191	45	179
control	13	214	28	199
Total	46	405	73	378

Data visualtion: Bar chart

Monte Carlo simulations

- Monte Carlo methods are computational algorithms that use repeated **random sampling** from a **probability distribution** to approximate numerical results
  - Concept: exploit randomness to obtain approximate solutions when analytical methods are impractical
- Name: references the Monte Carlo casino in Monaco, where the primary developer of the method was “inspired” by his uncle’s gambling habits

Fair coin?

- Assume you are trying to *prove* that flipping a coin is a fair way to settle a dispute between two friends
  - You decide to run an experiment
  - You have n number of students flip c number of coins
- $P(h) = P(t) = 0.5$ is most likely to be the most frequently occurring unique combination if the coin is fair…
  - … but you are *far* more likely to get a non 50-50% combination

Fair coin?

Fair coin? Takeaways

- More observations increase your confidence in the likelihood that your sample reflects something “true”
- This is accomplished in two ways:
  - Increasing observations per person: each person’s estimate is closer to true value (e.g. reduced variance)
  - Increasing people: unchanged variance, but smoother distribution

More data is more better

- Flipping a random coin is an example of a “data generating process” (DGP)
  - Fluctuation (variance) is part of almost all DGPs
- The DGP is not observable
  - Samples are (ideally) as if you observe small (random) portions of the “true” data

Random fluctuation or real difference?

- Recall that there was an 9 (p.p.) difference in the rate of experiencing a stroke between the treatment and control group
  - Increases in the sample size decrease the likelihood that this is due to statistical chance (variance)
- The magnitude of the difference also increases the plausibility that there are “true” differences in the probability of having a stroke due to treatment
  - 9 (7.8) p.p difference relative to the sample mean of 10.2% (16.2%) after 30 (365) days

Variance Equation

- **Variance** = measures the degree to which observations of a variable *deviate* from one another
  - "How far is the data from the average?"

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

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

Data Example

Tm	Total	Home	Away	Week.1	Week.2	Week.3	Week.4	Week.5
Arizona Cardinals	1138411	575780	562631	70542	63331	63650	63286	71310
Atlanta Falcons	1185384	642432	542952	72291	69879	72441	72204	70016
Baltimore Ravens	1220990	568423	652567	73611	70762	93566	70636	66341
Buffalo Bills	1192970	565564	627406	70542	65601	70316	70636	71427
Carolina Panthers	1175655	640515	535140	70007	70145	62417	77258	59307
Chicago Bears	1076006	527841	548165	59403	71412	66624	59074	59307
Cincinnati Bengals	1191765	530355	661410	66214	73558	66207	77258	66341
Cleveland Browns	1134676	541808	592868	67431	60105	68016	62401	59030
Dallas Cowboys	1401575	836749	564826	67431	93691	93566	80425	67380
Denver Broncos	1142969	515265	627704	68714	74215	62738	82243	72825
Detroit Lions	1153330	584300	569030	66530	64160	63650	66658	
Green Bay Packers	1214251	702030	512221	47236	77827	67322	78335	72842
Houston Texans	1216430	570666	645764	65306	71412	66843	71115	71427
Indianapolis Colts	1161683	526138	635545	65306	77827	66624	66376	60214
Jacksonville Jaguars	1114472	526117	588355	65582	60105	70316	71115	60214
Kansas City Chiefs	1144982	588837	556145	73611	73558	72441	70240	73592
Las Vegas Raiders	1122807	497403	625404	70240	70762	62417	62401	72825
Los Angeles Chargers	1170832	559732	611100	70240	70145	66734	70240	
Los Angeles Rams	1124537	584264	540273	66530	63331		59074	72842
Miami Dolphins	1153651	525150	628501	65582	65601	68658	65291	64628
Minnesota Vikings	1153509	596577	556932	81908	66741	66843	78335	61139
New England Patriots	1147581	517078	630503	66214	64686	80812	71042	64628
New Orleans Saints	1227161	630113	597048	70007	93691	70006	72204	73592
New York Giants	1273642	706231	567411	81908	61841	68016	80425	68306
New York Jets	1224693	631346	593347	71319	65509	80812	82243	61139
Philadelphia Eagles	1194031	606268	587763	47236	69879	70006	64223	

Data Example

team	attendance
Arizona Cardinals	6.7
Atlanta Falcons	6.97
Baltimore Ravens	7.18
Buffalo Bills	7.02
Carolina Panthers	6.92
Chicago Bears	6.33
Cincinnati Bengals	7.01

Mean

```
1 sum(y$attendance) / 32
```

```
[1] 6.895
```

```
1 mean(y$attendance)
```

```
[1] 6.895
```

Variance

```r
deviation <- y$attendance - mean(y$attendance)

variance <- sum(deviation^2) / (32-1)

variance
```

```
[1] 0.1194645
```

```r
var(y$attendance)
```

```
[1] 0.1194645
```

Standard Deviation

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

```r
std_dev <- sqrt(variance)

std_dev
```

```
[1] 0.3456364
```

```r
sd(y$attendance)
```

```
[1] 0.3456364
```

Coefficient of variation

- Coefficient of variation = $\frac{sd(x)}{mean(x)}$
- Heuristic:
  - < 0.10: low variability
  - $0.10 – 0.30$: moderate
  - > 0.30: high

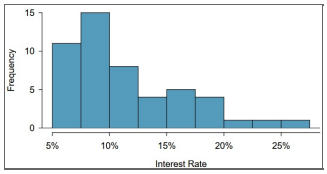
```
1 sd(y$attendance) / mean(y$attendance)
```

```
[1] 0.05012856
```

Distributions

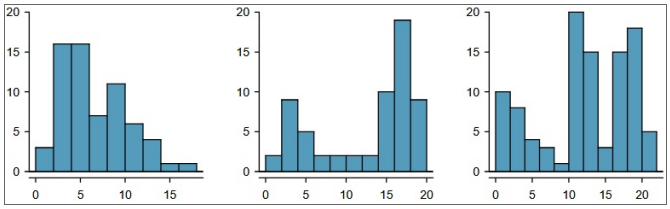
- Analysis of variance is often accompanied by visualizing the a variable's distribution
- **Histogram**: binned counts of numerical variables plotted in as bar chart
  - right (left) skewed: bin density decreases to the right (left)
  - symmetric: roughly equal tails (often called a "Bell Curve")
  - uniform: roughly equal bin density

Distributions: Right-Skewed



Distributions

- A mode is represented by a prominent peak in the distribution
  - Unimodal: one peak
  - Bimodal: two (often opposite) peaks
  - Multimodal: many peaks, no clear pattern



Distributions

```
par(mar = c(4.5, 4.5, 0, 0))
hist(y$attendance, main = "", xlab = "Attendance (per
     cex.axis = 1.25, cex.lab = 1.5)
```

Distributions

```
par(mar = c(4.5, 4.5, 0, 0))
hist(y$attendance, main = "", xlab = "Attendance (per
     cex.axis = 1.25, cex.lab = 1.5, breaks = 8)
```

Distributions

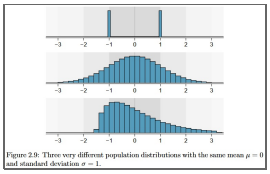


Figure 2.9: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

Outliers can distort inference

Outliers


WOMEN'S 1500M FREESTYLE
ALL-TIME FASTEST TIMES
1. K. LEDECKY 15:20.48
2. K. LEDECKY 15:25.48
3. K. LEDECKY 15:26.27
4. K. LEDECKY 15:27.71
5. K. LEDECKY 15:28.36
6. K. LEDECKY 15:29.51
7. K. LEDECKY 15:29.64
8. K. LEDECKY 15:30.15
9. K. LEDECKY 15:31.82
10. K. LEDECKY 15:34.23
11. K. LEDECKY 15:35.35
12. K. LEDECKY 15:35.65
13. K. LEDECKY 15:35.98
14. K. LEDECKY 15:36.53
15. K. LEDECKY 15:37.34
16. K. LEDECKY 15:37.99
17. K. LEDECKY 15:38.25
18. K. LEDECKY 15:38.81
TYR
USA
KATIE
LEDECKY


Yakult
Nikon
FINA.ORG
Salonpas
1 USA K. LEDECKY
8:04.79
8:07.39
OMEGA
4
8:08.04
Fina
Women's 800m Freestyle

Outliers

- There are two common ways to determine whether something is a statistical outlier

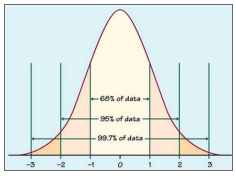
1. Standard deviations relative to the mean

$$z = \frac{x_i - \bar{x}}{sd(x)}$$

2. Quartiles

- Median (aka 50% percentile or Q_2)
- Interquartile range: $Q_3 - Q_1$
- Whiskers: $Q_1 - 1.5 \times IQR$; $Q_3 + 1.5 \times IQR$

Outliers



Outliers

Outliers

Data Transformation

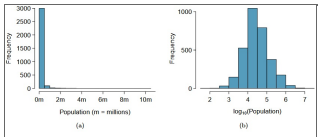


Figure 2.13: (a) A histogram of the populations of all US counties. (b) A histogram of $\log_{10}$-transformed county populations. For this plot, the x-value corresponds to the power of 10, e.g. "4" on the x-axis corresponds to $10^4 = 10{,}000$.

Data Transformation

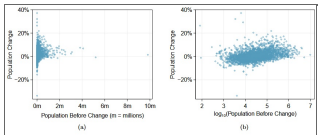


Figure 2.14: (a) Scatterplot of population change against the population before the change. (b) A scatterplot of the same data but where the population size has been log-transformed.

Hypothesis Testing

- Null hypothesis (H_O): Stents were no more effective at preventing strokes than no treatment
- Alternative hypothesis (H_A): Stents were effective and the probability of experiencing a stroke changes due to the stint
- Does a 9% difference in the rate of stroke between the treatment and control group allow us to reject the null hypothesis?
    - No! We know nothing about the “natural” variance in the outcome variable

Using Randomization to Determine Statistical Significance

Using Randomization to Determine Statistical Significance