MEASURES OF ASSOCIATION

POLS 095

Statistical inference and sampling

Traits of the sample
(mean, proportion, etc.)

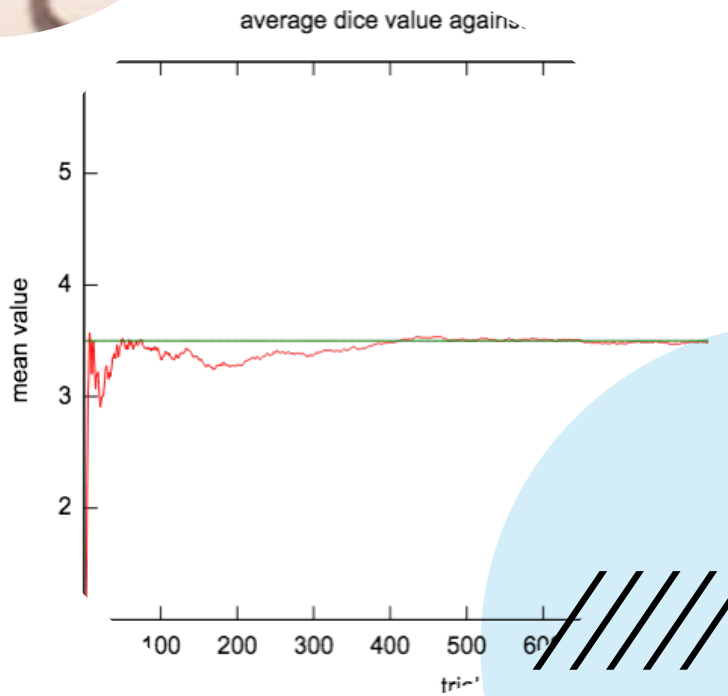
sample statistics

Traits of the population
(mean, proportion, etc.)

population parameters

Same principles

- Probability tells us the likelihood of “chance” outcomes
- This is called the “Law of Large Numbers”

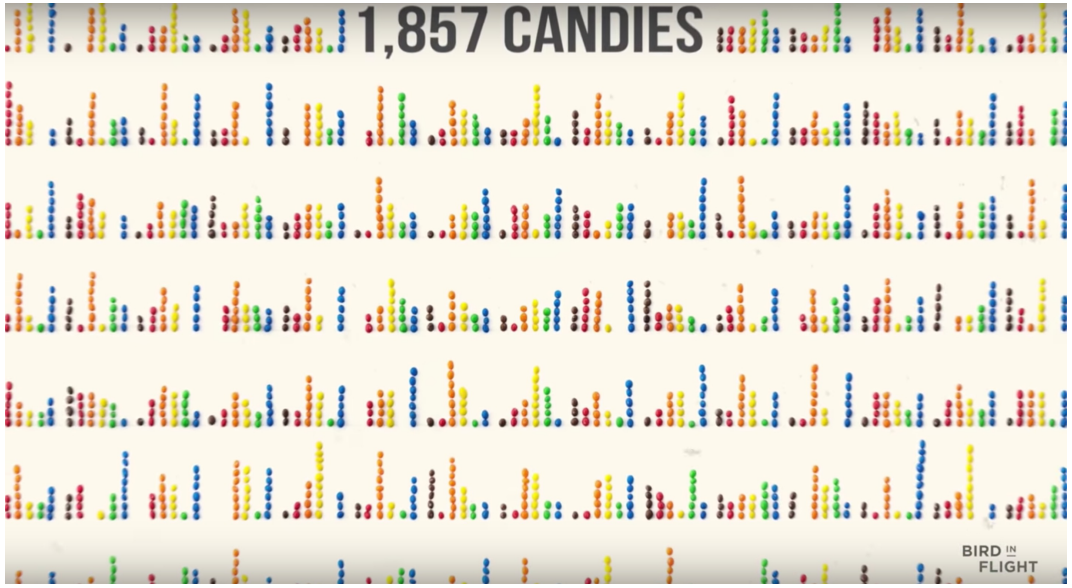


Recall the m&ms

- What percent of all M&Ms are Red? Yellow? Blue? Green?
- Equal?
  - There is a population of m&ms
  - But we don’t know it

Using the law of large numbers

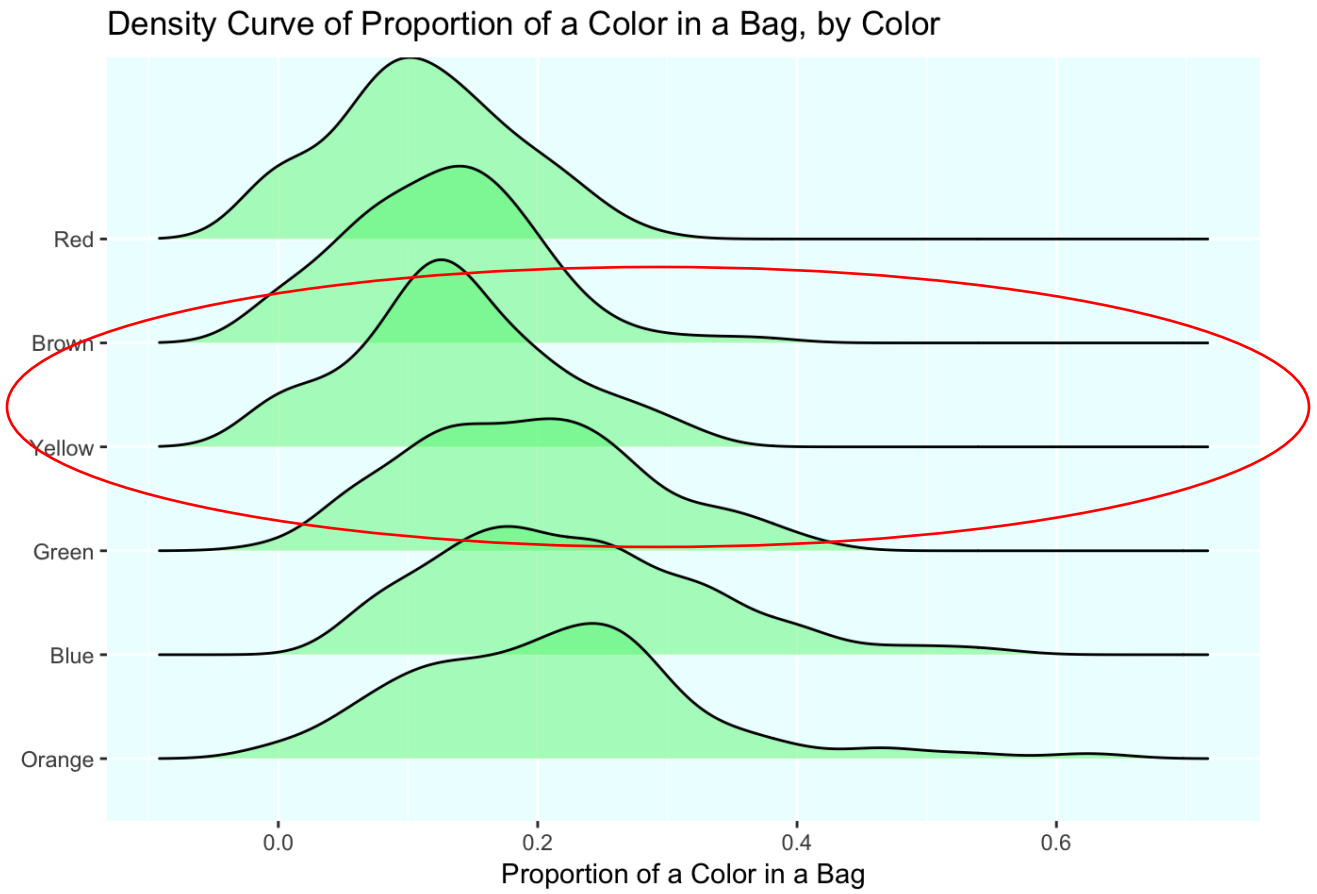
- We can draw samples of m&ms
  - Any one bag is the sample



WHAT IF WE DID THIS A LITTLE DIFFERENTLY?

- **WHAT PROPORTION IS YELLOW?**
  - Pick a pack of M&Ms and record the proportion that is yellow
  - Pick another one
  - Pick another one
  - Pick another one
  - Pick another one
  - And another
  - And another
  - And another
  - And another
  - And another…
  - Just keep doing it

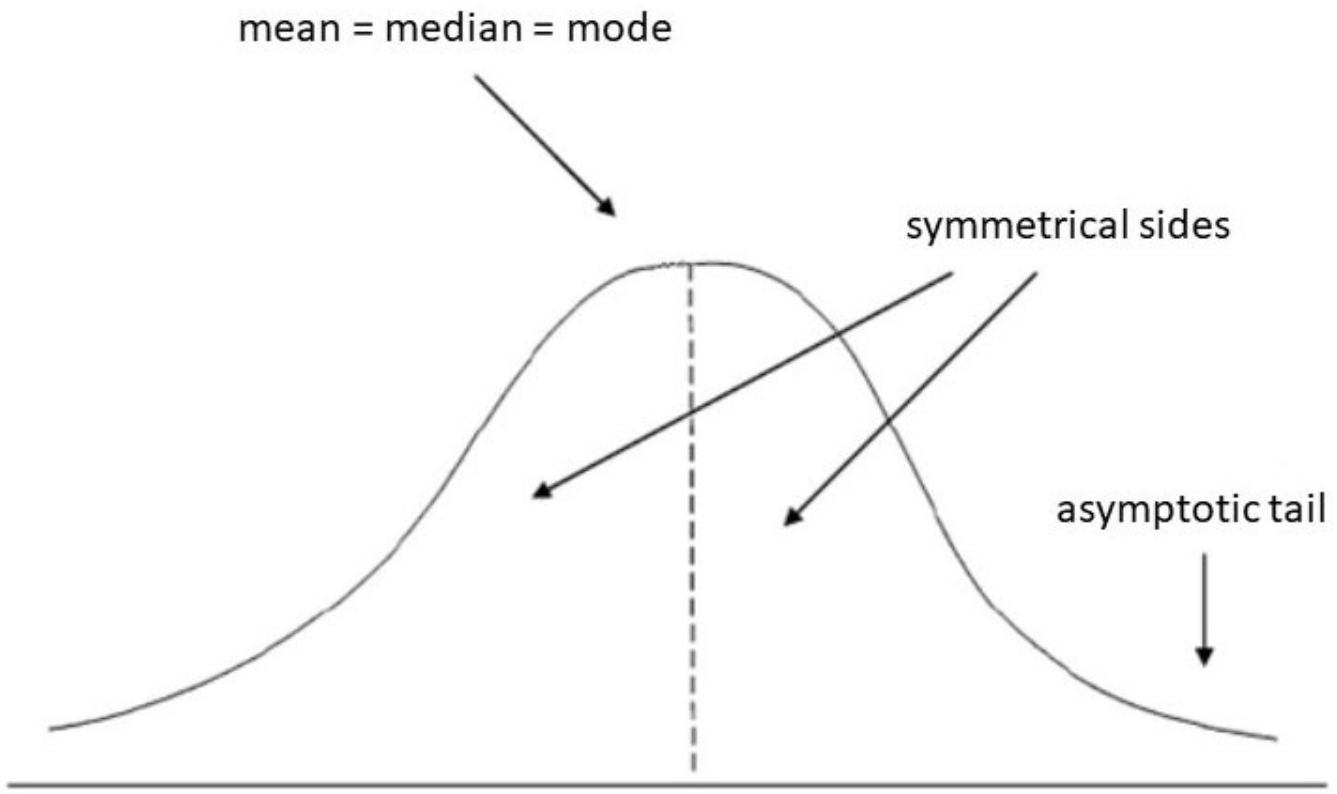
AS YOU KEEP RECORDING DATA….

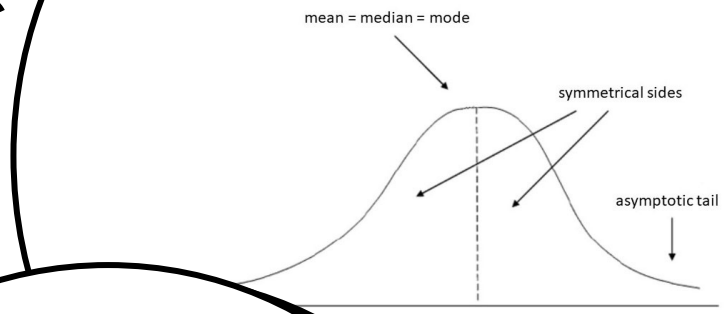


This is called a sampling distribution, a map of a single sample statistic (proportion in this case) for every sample.

Notice that ALL the colors start to have the same shape!

AND THAT SHAPE IS… THE NORMAL DISTRIBUTION





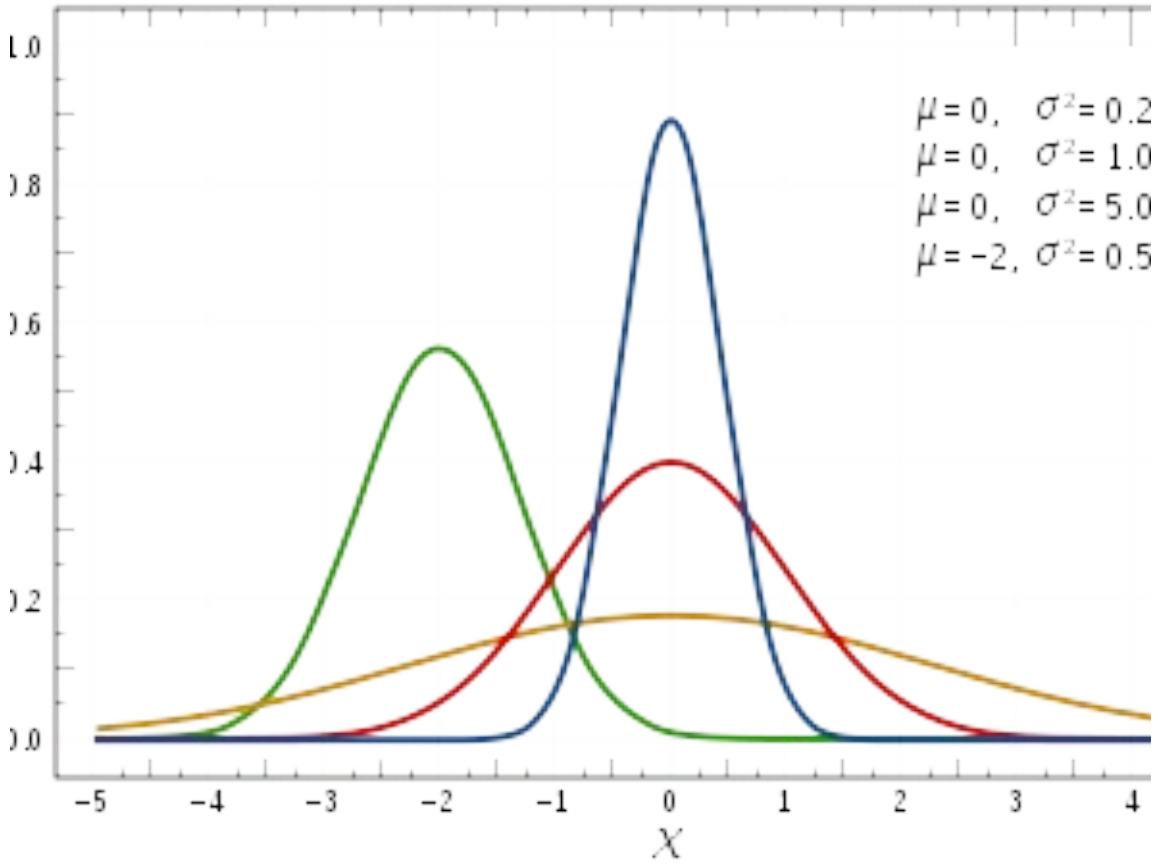
It’s a SUPER TOOL because it has all sorts of good characteristics that help us

It is symmetrical.

The middle of the sampling distribution (the high point) will equal the true population mean, median, and mode.

The asymptotic tail means that VERY few cases will fall at the extremes (i.e., 0% or 100% yellow)

JUST LIKE ALL DATA...

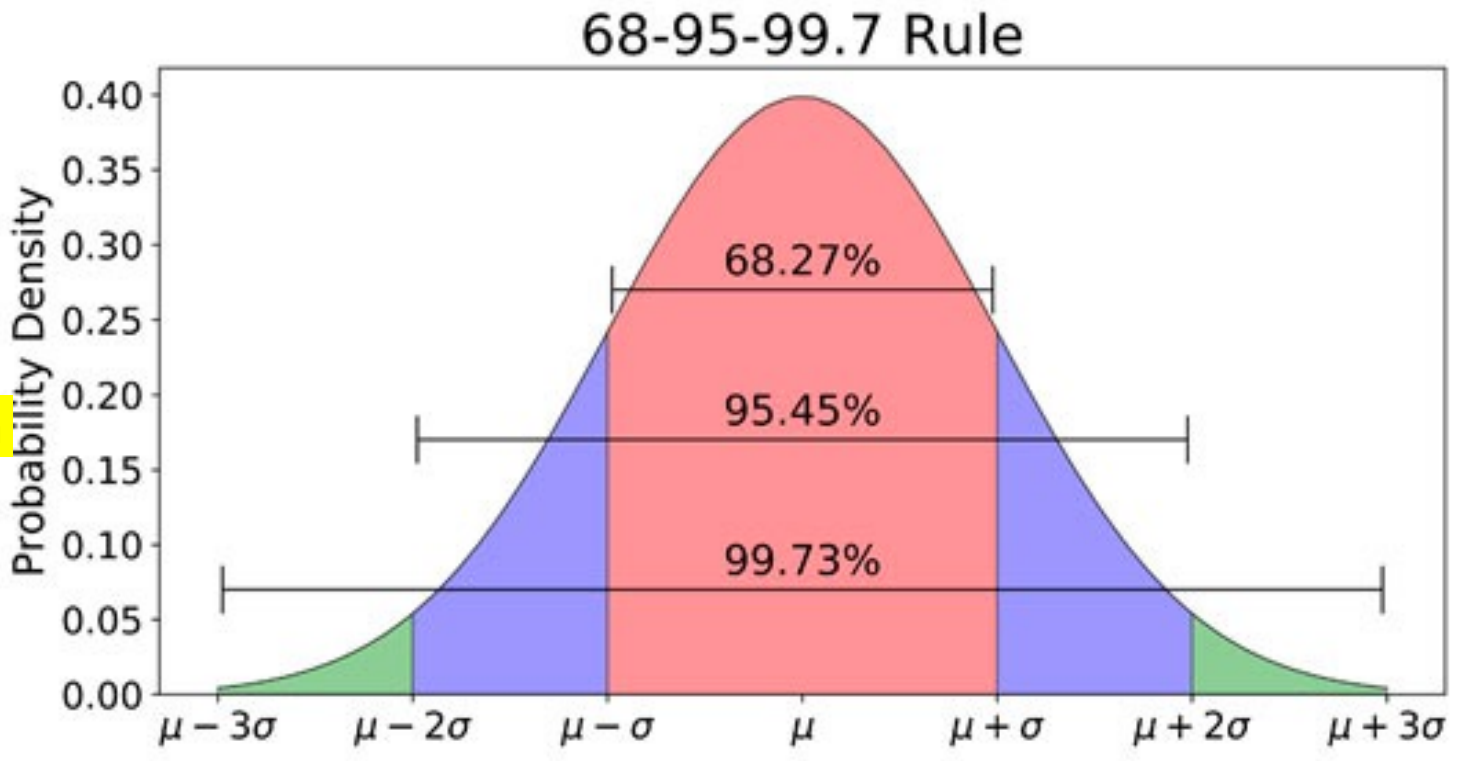


Just like all data…

68% of all samples will fall between one standard error below the mean and one standard error above the mean.

95% will fall between 2 S.E. below the mean and 2 S.E. above the mean.

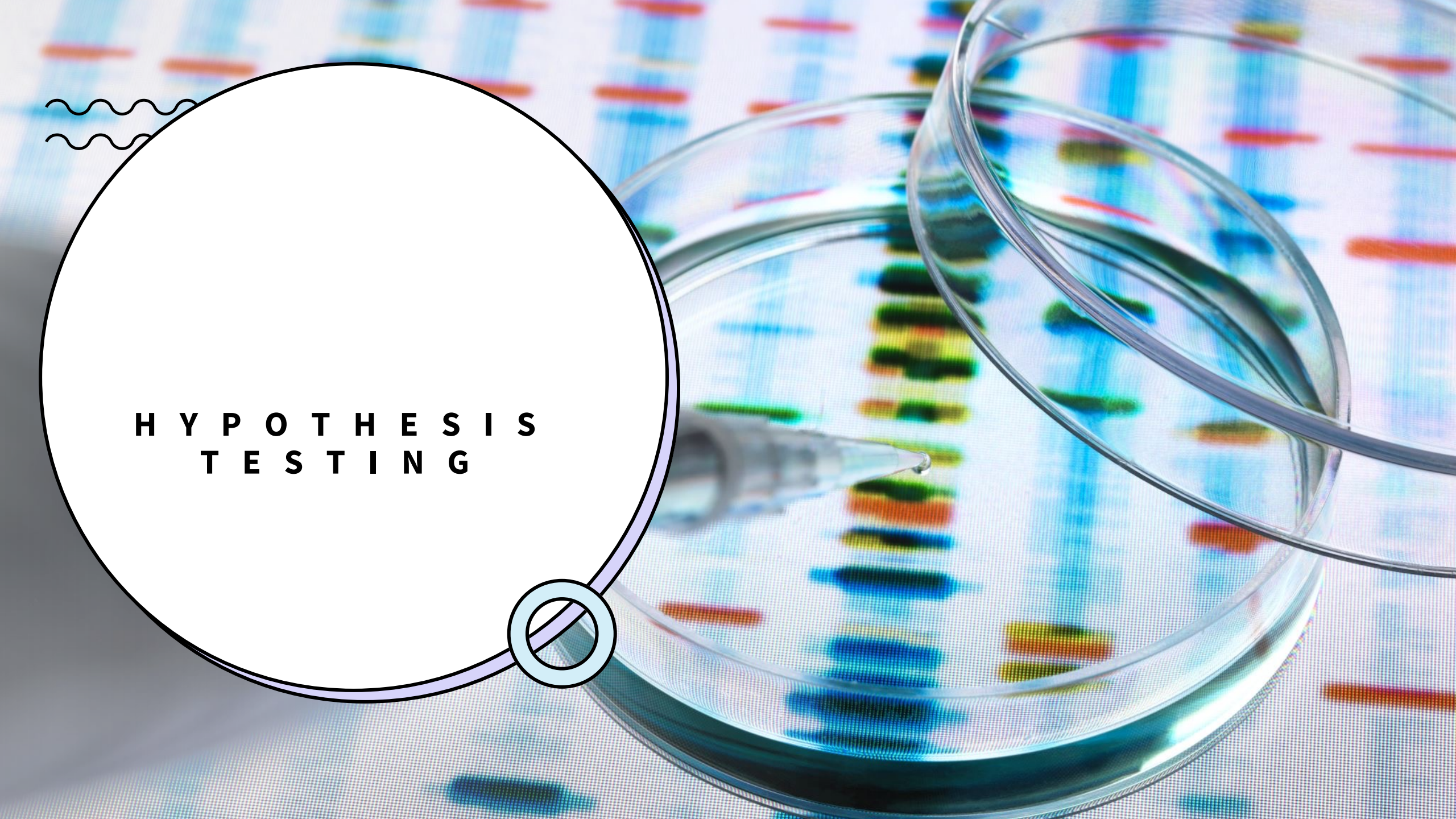
99.7% will fall between 3 S.E. below the mean and 3 S.E. above the mean.



AND THAT IS POWERFUL…

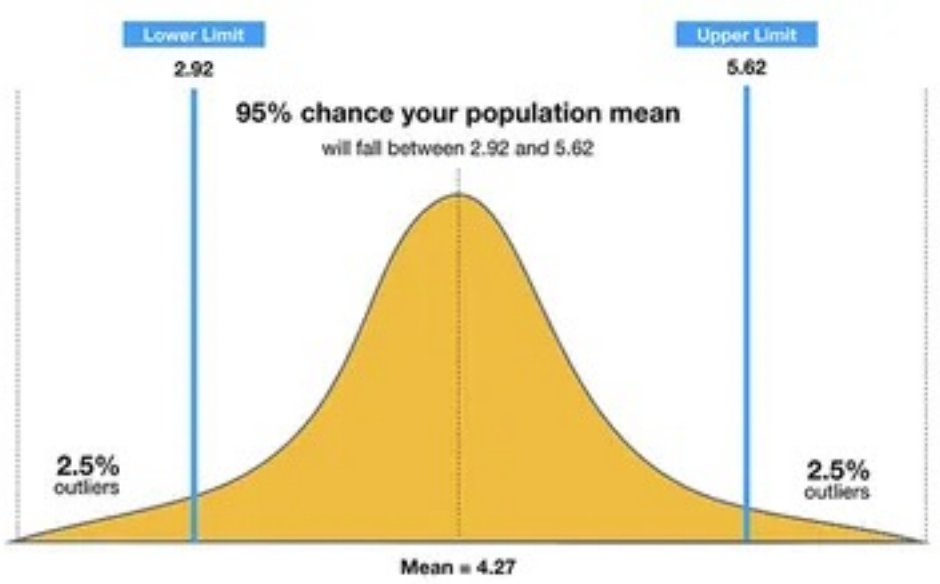
- And we can do it for any population parameter we're interested in….
- If we KNOW the true population mean, we can find the probability of any sample, but we can also reverse the logic and use a sample to infer information about the population.

HYPOTHESIS TESTING



The confidence interval approach

- We did this last week!
- Calculate the interval around the sample mean to calculate “uncertainty”
  - (Accounts for error)



The hypothesis testing approach

- Hypothesis testing or significance testing
- Alternative hypothesis $\boldsymbol{H_A}$
- Null hypothesis $\boldsymbol{H_0}$

"But what if the evidence doesn't favor H_A?"

The hypothesis testing approach

1. Compare parameter to fixed number.
   - $\boldsymbol{H_A}$: $p > 0.50$
   - $\boldsymbol{H_0}$: $p = 0.50$
2. Compare parameter across two groups (say, D vs. R)
   - $\boldsymbol{H_A}$: $\mu_1 - \mu_2 \neq 0$
   - $\boldsymbol{H_0}$: $\mu_1 - \mu_2 = 0$
3. Are two variables related at all?
   - $\boldsymbol{H_A}$: Increases in education lead to increases in income (*X* and *Y*)
   - $\boldsymbol{H_0}$: No association between between *X* and *Y*

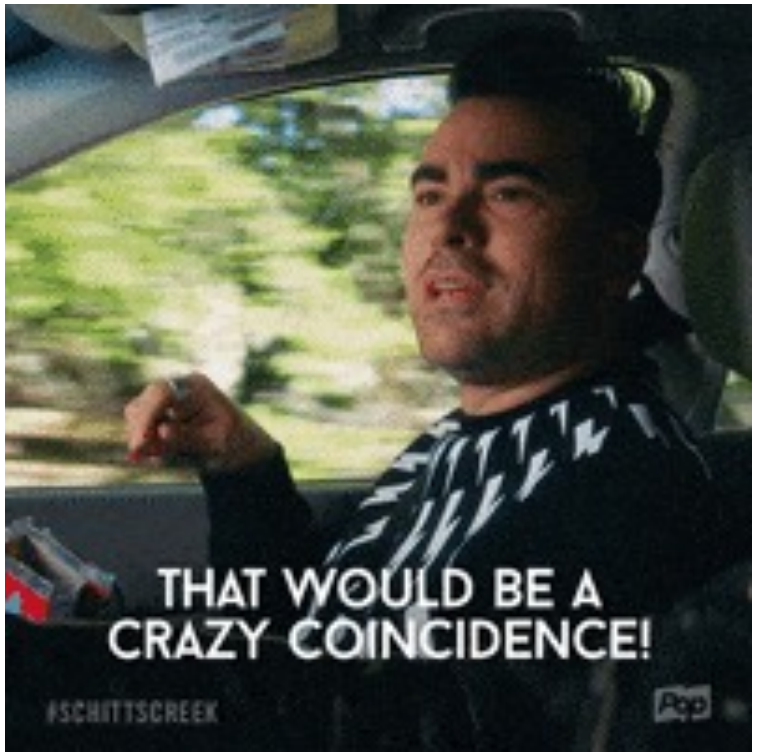
Hypothesis testing

- Reject or fail to reject the null hypothesis
  - DO NOT SAY YOUR HYPOTHESIS IS TRUE!

Statistical significance

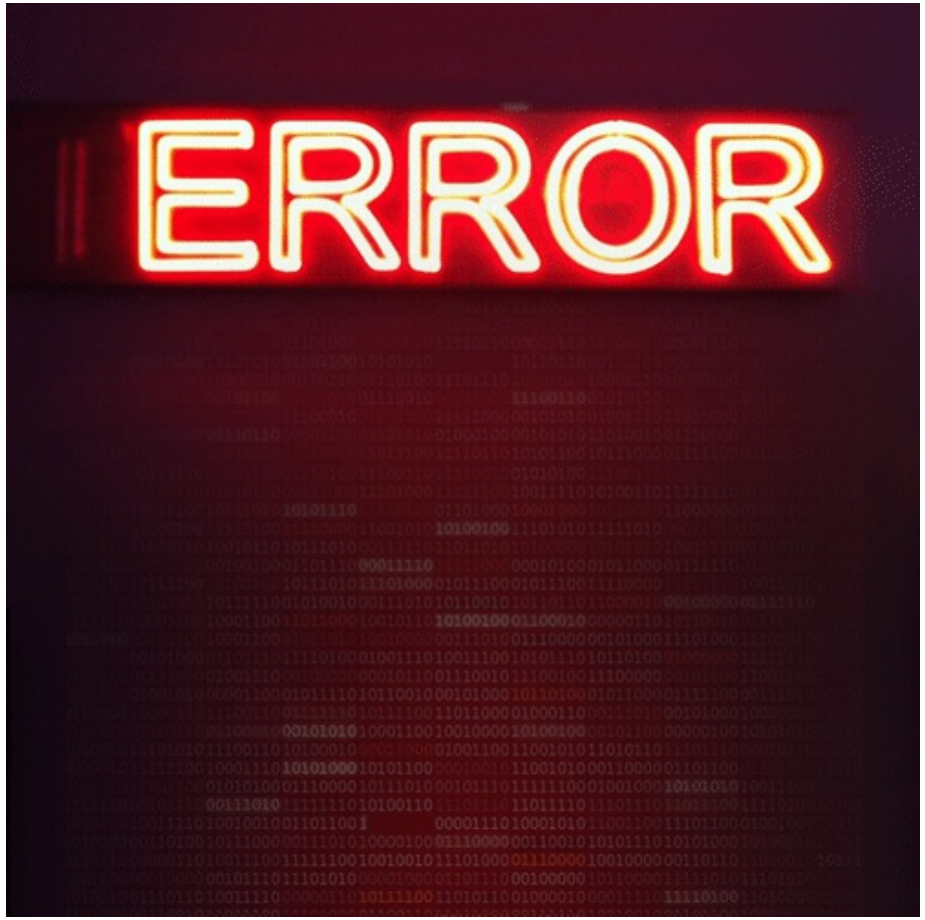
- Important distinction:
  Statistical vs substantive significance

- Significance = confidence
  - It’s “mathematical”



Types of error

- The decision to reject or fail to reject the null can be thought of as being either correct or an error.

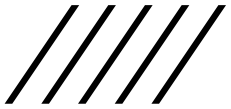


Types of error

Actual situation in population

Researcher decision	H_0 true	H_0 false
Reject H_0	Type I error	Correct inference
Do not reject H_0	Correct inference	Type II error

- In some ways, Type II error isn't quite an "error"
- Errors and The Boy Who Cried Wolf



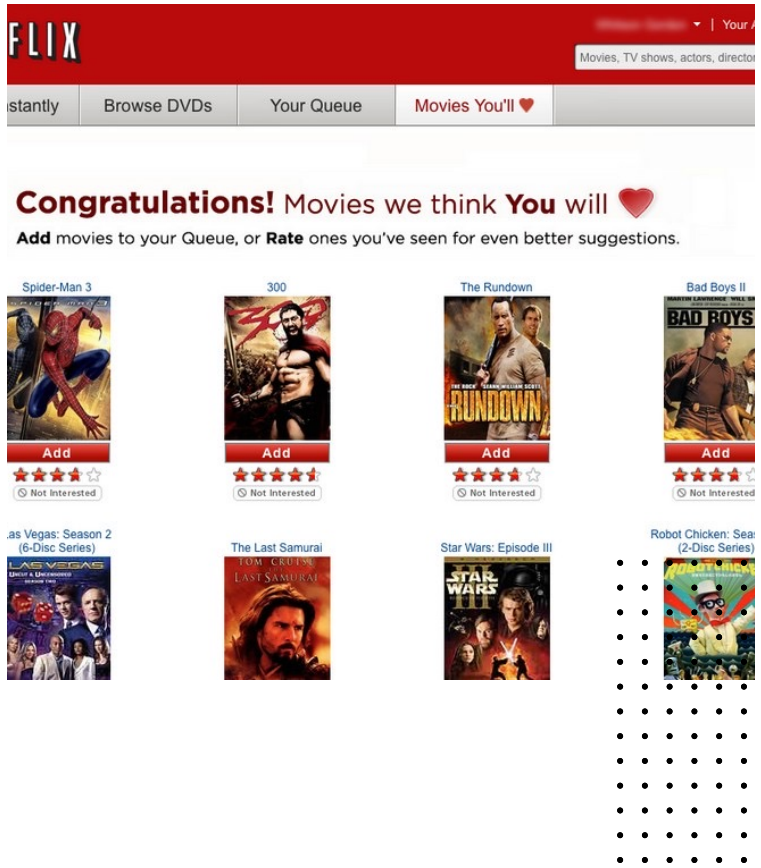
Levels of statistical significance

- Three main levels of statistical significance ($\alpha = significance\ level$)
  - .10 (10% probability of Type I error if $\boldsymbol{H_0}$ true)
  - .05 (5% probability of Type I error if $\boldsymbol{H_0}$ true)
  - .01 (1% probability of Type I error if $\boldsymbol{H_0}$ true)
- Type I and II errors

MEASURES OF ASSOCIATION

MEASURES OF ASSOCIATION

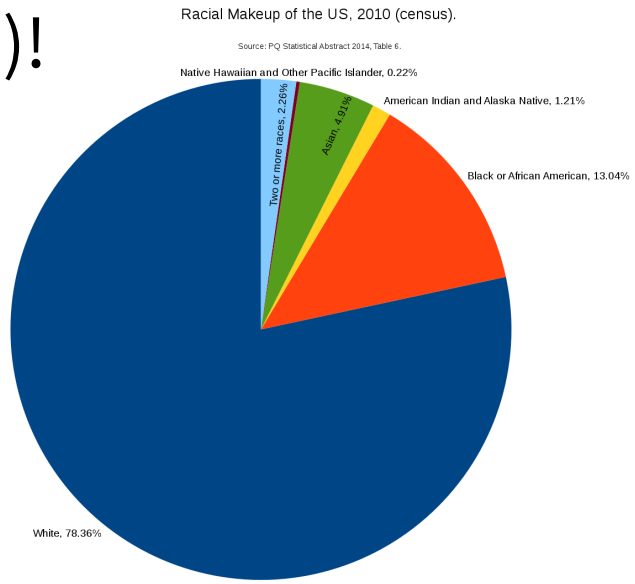
- A measure of association is a single summarizing number that reflects the **strength** of a relationship, indicates the usefulness of **predicting** the dependent variable from the independent variable, and often shows the **direction** of the relationship.



TAKE YOUR BEST GUESS!...

If you know nothing else about a person except that he or she lives in the United States, and I asked you to guess their race/ethnicity, what would you guess?

If you're smart, you will guess the most common race/ethnicity for U.S. residents (e.g., White)!



BUT IF WE ADD MORE INFORMATION…

Now, if we know that this person lives in New York City?

Goal of quantitative analysis:

- Take our best, most reasonable, and informed “guess” at the value of the DV

MEASURES OF ASSOCIATION

- **Measures of association**—a single summarizing number that reflects the strength of the relationship. This statistic shows the **magnitude** and/or **direction** of a relationship between variables.
- **Magnitude**—the closer to the absolute value of 1, the stronger the association. If the measure equals 0, there is no relationship between the two variables.
- For χ^2, use the chi-square distribution to determine "significance" (significance = whether we can fairly certain that a pattern is not occurring randomly)
- **Direction**—the sign on the measure indicates if the relationship is positive or negative. In a **positive relationship**, when one variable is high, so is the other. In a **negative relationship**, when one variable is high, the other is low.

CHI-SQUARE

1. Find expected frequency for each cell in table under null hypothesis: f_e
2. For each cell, subtract expected frequency from observed frequency: $f_o - f_e$
3. Square each & divide by expected: $\frac{(f_o - f_e)^2}{f_e}$
4. Sum all the results: $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$

CHI-SQUARE: EXAMPLE

- Hypothesis:
- In a comparison of *countries*, those with *higher levels of economic development* will be more likely to have *more press freedom* than those that are *less economically developed*.

- Start with the assumption that no relationship exists (null hypothesis)

Press Freedom	High Development	Medium Development	Low Development	Row Totals
Free Press	45% of 75	45% of 58	45% of 62	45% (88)
Partially Free Press	30% of 75	30% of 58	30% of 62	30% (58)
No Free Press	25% of 75	25% of 58	25% of 62	25% (49)
Total	75	58	62	100% (195)

If no relationship exists, we would expect that the same proportion/distribution holds true for every value of the independent variable

EXPECTED FREQUENCY

Press Freedom	High Development	Medium Development	Low Development	Row Totals
Free Press	45% of 75 = 34	45% of 58 = 26	45% of 62 = 28	45% (88)
Partially Free Press	30% of 75 = 23	30% of 58 = 17	30% of 62 = 19	30% (58)
No Free Press	25% of 75 = 19	25% of 58 = 15	25% of 62 = 15	25% (49)
Total	75	58	62	100% (195)

OBSERVED FREQUENCY

Press Freedom	High Development	Medium Development	Low Development	Row Total
Free Press	60% 45	45% 26	27% 17	45% 88
Partially Free Press	28% 21	29% 17	32% 20	30% 58
Not Free Press	12% 9	26% 15	40% 25	25% 49
Total	100% 75	100% 58	100% 62	100% 195

Press Freedom	High Development	Medium Development	Low Development	Row Total
Free Press	60% 45	45% 26	27% 17	45% 88
Partially Free Press	28% 21	29% 17	32% 20	30% 58
Not Free Press	12% 9	26% 15	40% 25	25% 49
Total	100% 75	100% 58	100% 62	100% 195

OBSERVED

Press Freedom	High Development	Medium Development	Low Development	Row Totals
Free Press	45% of 75 = 34	45% of 58 = 26	45% of 62 = 28	45% (88)
Partially Free Press	30% of 75 = 23	30% of 58 = 17	30% of 62 = 19	30% (58)
No Free Press	25% of 75 = 19	25% of 58 = 15	25% of 62 = 15	25% (49)
Total	75	58	62	100% (195)

EXPECTED

For first cell: $f_o - f_e$ = 45-34 = 11

$$\frac{(f_o - f_e)^2}{f_e} = (11)^2/34 = 3.55$$

Press Freedom	High Development	Medium Development	Low Development	Row Total
Free Press	60% 45	45% 26	27% 17	45% 88
Partially Free Press	28% 21	29% 17	32% 20	30% 58
Not Free Press	12% 9	26% 15	40% 25	25% 49
Total	100% 75	100% 58	100% 62	100% 195

OBSERVED

Press Freedom	High Development	Medium Development	Low Development	Row Totals
Free Press	45% of 75 = 34	45% of 58 = 26	45% of 62 = 28	45% (88)
Partially Free Press	30% of 75 = 23	30% of 58 = 17	30% of 62 = 19	30% (58)
No Free Press	25% of 75 = 19	25% of 58 = 15	25% of 62 = 15	25% (49)
Total	75	58	62	100% (195)

EXPECTED

For second cell: $f_o - f_e$ = 26-26 = 0

$$\frac{(f_o - f_e)^2}{f_e} = 0^2/26 = 0$$

- Add them all together….

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \qquad = 19.26$$

And now…

- Is 19.26 big? Small? Significant?
- Compare the chi-square score to a critical value
- Reject null hypothesis of no association if $\chi^2 > \chi^2_{critical}$ for appropriate degrees freedom
  - Degrees of freedom (df) = (r - 1)(c - 1)
    - where r is number of levels (values) in row variable, and c is number of levels (values) in column variable.

Percentage Points of the Chi-Square Distribution

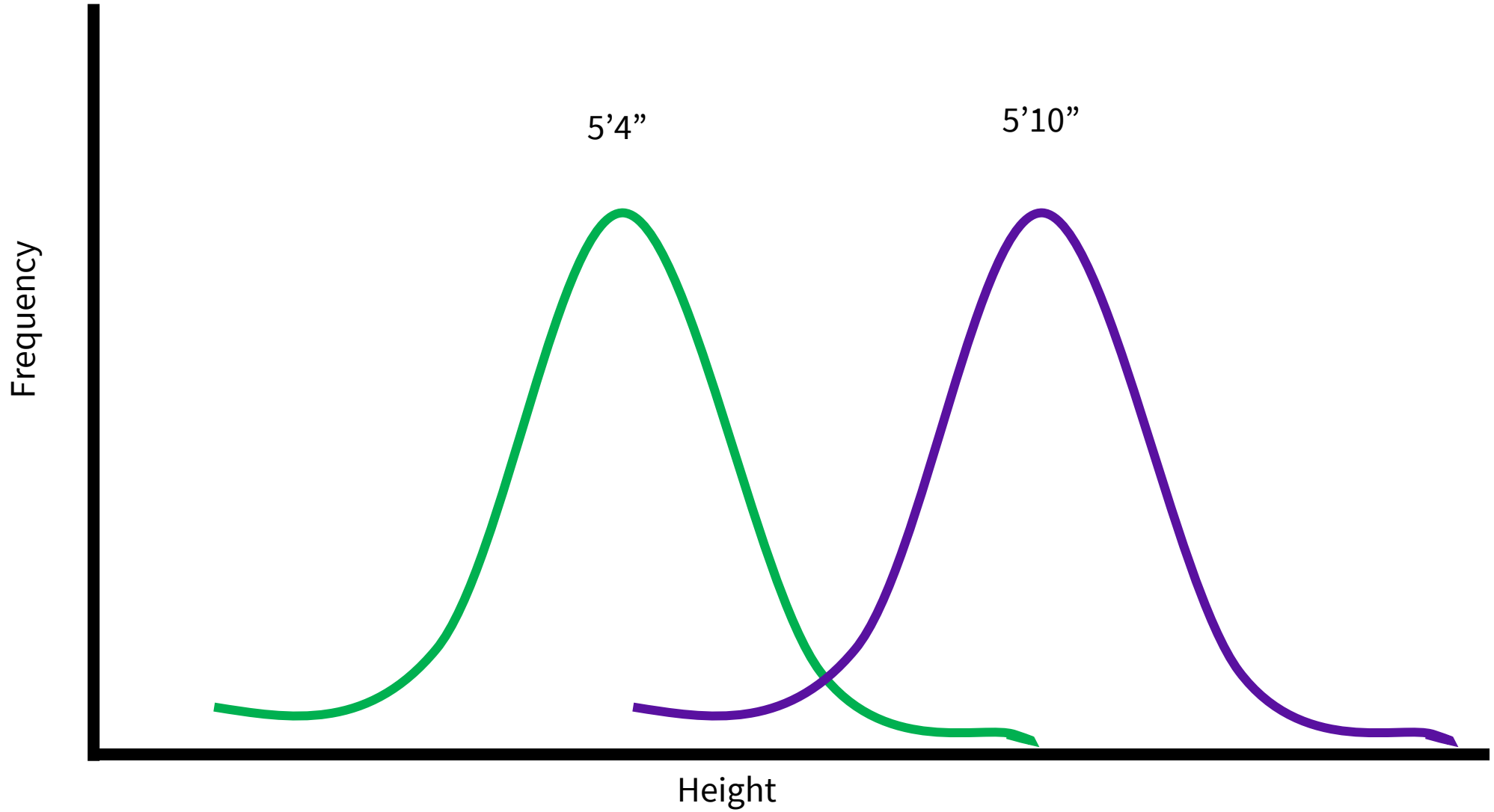
Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

- For us, df = (3 - 1)(3-1) = 4, and we typically use a 95% level of confidence
- $\chi^2_{critical}$ = .711 if we use 95% threshold (.95)
- 19.26 > .711
- Reject H_0? What does that mean?

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

Comparing distributions between groups



IS THERE A RELATIONSHIP BETWEEN GENDER AND PRESIDENTIAL VOTING?

CHI-SQUARE EXAMPLE

Gender and vote in the 2016 US presidential election:

Candidate	Male	Female	Row Total
Clinton	47.2	56.2	52.0
Trump	53.8	43.8	48.0
Total	100.0	100	100.0
Note: Cell entries are column percentages.			

1. What is the unit of analysis?
2. What is a hypothesis we might test regarding this relationship?
3. What is the null hypothesis?
4. What is the level of measurement for vote choice and gender?
5. What is the best tool to evaluate this relationship?

Gender and vote in the 2016 US presidential election:

Hypothesis regarding gender and vote choice in the 2016 US presidential election

- Hypothesis (h_1): In a comparison of voters in 2016, females will be more likely to have voted for Clinton than males.

- Null hypothesis (h_0): There is no relationship between gender and voting for Clinton in 2016.

Gender and vote in the 2016 US presidential election:

Candidate	Male	Female	Row Total
Clinton	47.2	56.2	52.0
Trump	53.8	43.8	48.0
Total	100.0	100	100.0
Note: Cell entries are column percentages.			

Gender and vote in the 2016 US presidential election: hypothetical scenario

Candidate	Male	Female	Row Total
Clinton	?	?	52.0
Trump	?	?	48.0
Total	100	100	100.0
Note: Cell entries are column percentages.			

Gender and vote in the 2016 US presidential election: expectations if there were no relationship

Candidate	Male	Female	Row Total
Clinton	52.0	52.0	52.0
Trump	48.0	48.0	48.0
Total	100.0	100	100.0
Note: Cell entries are column percentages.			

Gender and vote in the 2016 US presidential election: hypothetical scenario

Candidate	Male	Female	Row Total
Clinton	?	?	1269
Trump	?	?	1171
Total	1128	1312	2240

Note: Cell entries are number of respondents.

Gender and vote in the 2016 US presidential election: calculating the expected cell values if gender and presidential vote were unrelated

Candidate	Male	Female
Clinton	(52% of 1128) = 0.52 x 1128 = 586.56	(52% of 1312) = 0.52 x 1312 = 682.24
Trump	(48% of 1128) = 0.48 x 1128 = 541.44	(48% of 1312) = 0.48 x 1312 = 629.76
Note: Cell entries are expectation calculations if these two variables were unrelated.		

Gender and vote in the 2016 US presidential election:

Candidate	Male	Female	Row Total
Clinton	532	737	1269
Trump	596	575	1171
Total	1128	1312	2240
Note: Cell entries are number of respondents.			

Gender and vote in the 2016 US presidential election: calculating the expected cell values if gender and presidential vote were unrelated

Candidate	Male		Female	
Clinton	$f_0 = 532$	$f_e = 586.56$	$f_0 = 737$	$f_e = 682.24$
Trump	$f_0 = 596$	$f_e = 541.44$	$f_0 = 575$	$f_e = 629.76$

Note: Cell entries are the number observed (f_0); and the number expected if there were no relationship (f_e).

Gender and vote in the 2016 US presidential election:

Are the differences between gender and the vote statistically significantly different than if there were no relationship between gender and the vote?

- Chi-square test for tabular association:
  - $X^2 = \Sigma \frac{(f_o - fe)^2}{f_e}$

$$= \frac{(532-596.56)^2}{586.56} + \frac{(737-682.24)^2}{682.24} + \frac{(596-541.44)^2}{541.44} + \frac{(575-629.76)^2}{629.76}$$

$$= \frac{2976.8}{586.56} + \frac{2998.7}{682.24} + \frac{2976.8}{541.44} + \frac{2998.7}{629.76}$$

$$= 5.075 + 5.498 + 4.395 + 4.762$$

$$= 19.73$$

- What do we do with this?
  - We need to compare that 19.73 with some predetermined value, called a *critical value*, of X^2.

Gender and vote in the 2016 US presidential election:

What do we do with a Chi-square value?

- What do we do with our Chi-square (X^2) value of 19.73?
  - We need to compare that 19.73 with some predetermined value, called a *critical value*, of X^2.
  - How do we obtain this critical value? First, we need a piece of information known as the *degrees of freedom (df).*
    - In this case, the *df* calculation is very simple:
      - $df = (r - 1)(c\text{-}1)$, where r is the number of rows in the table and c is the number of columns in the table.
      - So, in our example, $df = (2\text{-}1)(2\text{-}1)=1$.
  - If our X^2 is greater than the critical value then we can reject the null hypothesis of no relationship between gender and the 2016 vote.
  - Next, we find a Chi-square table with critical values.

Gender and vote in the 2016 US presidential election:

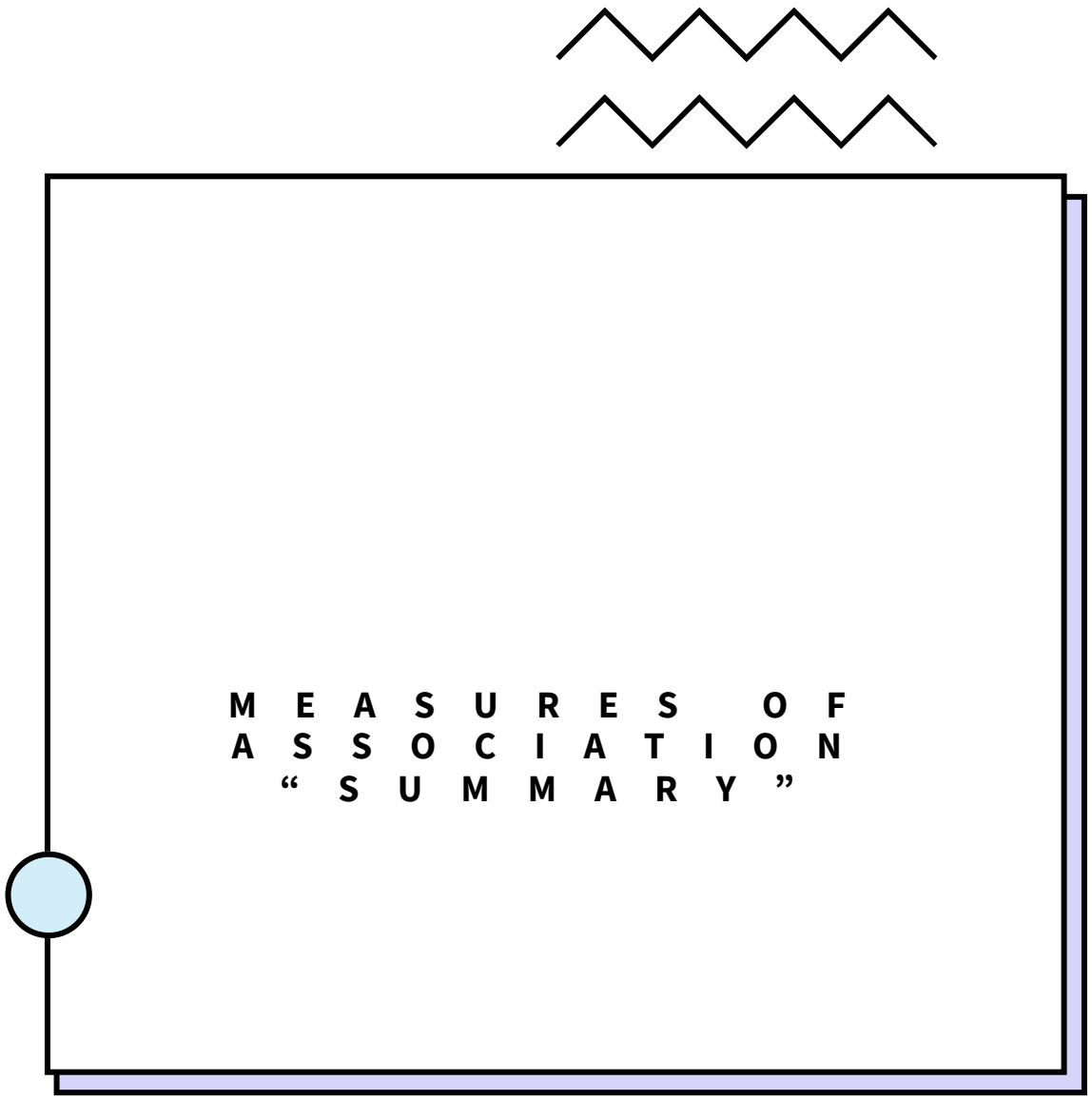
h_0: no relationship between gender and the 2016 vote

h_1: females are more likely to vote for Clinton than males

- $X^2 = 19.73$
- $df = 1$
- Critical value of X^2 at the 95% level of significance (p-value = 0.05)?
  - 3.84
- If $X^2 > X^2_{critical}$ then we reject the null.
  - 19.73 > 3.84, which is needed to achieve a p-value of 0.05.
- We have established that we can reject the null, though we have not yet established a causal relationship between gender and presidential voting.

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38



- **Measures of association**—a single summarizing number that reflects the strength of the relationship. This statistic shows the **magnitude** and/or **direction** of a relationship between variables.
- **Magnitude**—the closer to the absolute value of 1, the stronger the association. If the measure equals 0, there is no relationship between the two variables.
- For χ^2, use the chi-square distribution to determine “significance” (significance = whether we can fairly certain that a pattern is not occurring randomly)
- **Direction**—the sign on the measure indicates if the relationship is positive or negative. In a **positive relationship**, when one variable is high, so is the other. In a **negative relationship**, when one variable is high, the other is low.

DIFFERENCE OF MEANS

When to use a *difference of means* test

The level of measurement of our variables determines which test to use. We use a difference of means test when:

Dependent variable: continuous	Independent variable: categorical

Looking to see if the means are different across values of the independent variable

Example: Length of parliamentary governments

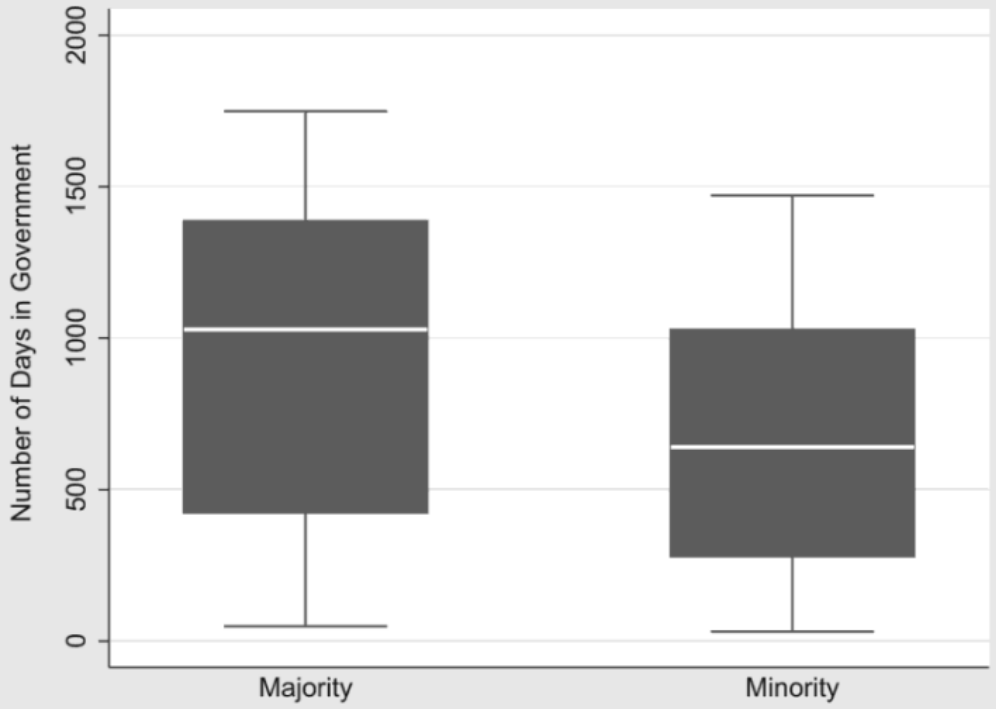
- Parliamentary governments and how long government's last
- What is a parliamentary government?
  - A government in which the lower house of the legislature is the most powerful branch of government and directly selects the head of government.
- Important factors:
  - Whether the party or parties that are "in government" (occupying one or more cabinet posts) are in the majority and can vote out a minority government out of office.
- General theory: majority governments will last longer than minority governments
- Hypothesis: In a comparison of parliamentary democracies, those with majority governments will last longer than will minority governments.
- $H_A : \mu_{majority} - \mu_{minority} \neq 0$
- $H_0 : \mu_{majority} - \mu_{minority} = 0$

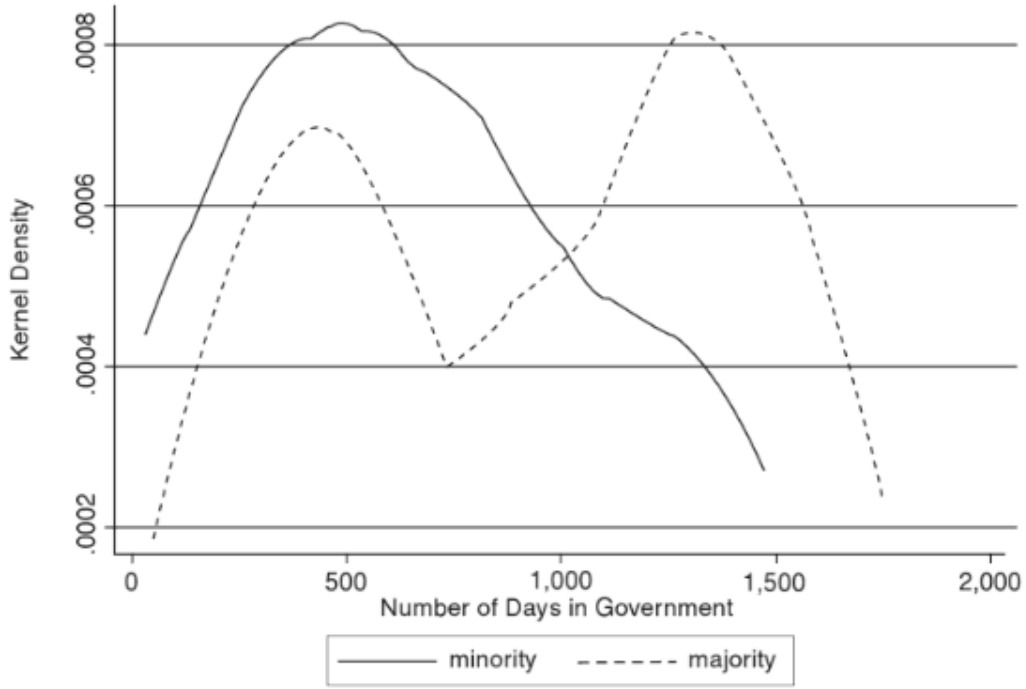
From theory to hypothesis test: The Data

- McDonald and Mendes: “Governments, 1950-1995”
- 21 Western countries
- Independent variable: Government Type
  - “Majority” or “minority”
- Dependent variable: Number of days in government
  - # days each government lasted in office
  - Hypothetical range: 1-1749
  - Actual range: 31-1749

- $H_A : \mu_{majority} - \mu_{minority} \neq 0$
- $H_0 : \mu_{majority} - \mu_{minority} = 0$

Viewing the distribution of this continuous variable





Difference of means test

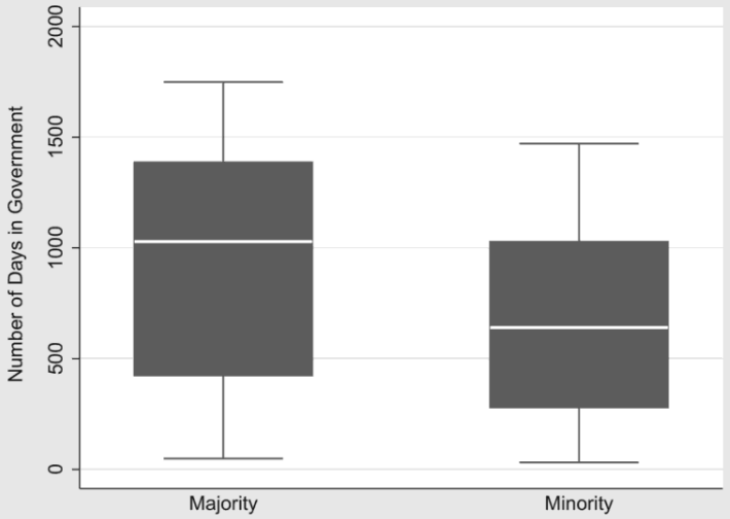
- Compare what we saw in the two figures with what we would expect if there were no relationship between government type and government duration
- We use the *t*-test:

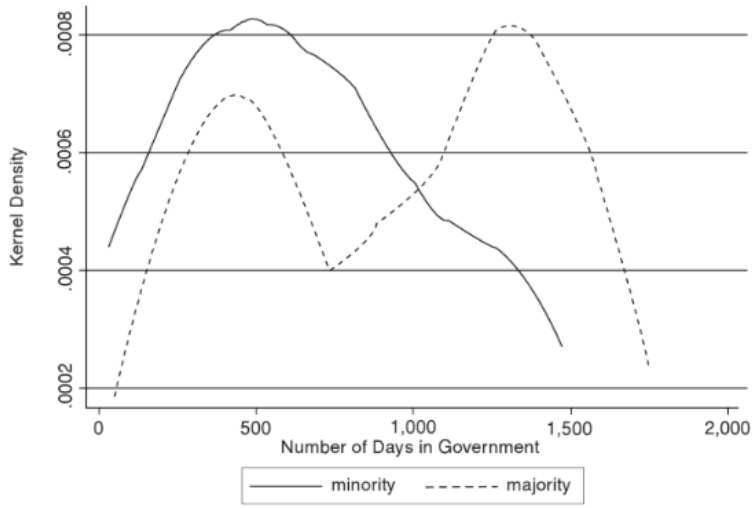
$\bar{Y}_1$ = mean of the dependent variable for the first value of the independent variable
$\bar{Y}_2$ = mean of the dependent variable for the first value of the independent variable

- $t = \frac{\bar{Y}_1 - \bar{Y}_2}{\text{se}(\bar{Y}_1 - \bar{Y}_2)}$

$$\text{se}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right)} \text{ X } \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Viewing the distribution of this continuous variable





Government Type	No. of Observations	Mean Duration	Standard Deviation
Majority	124	930.5	466.1
Minority	53	674.4	421.4
Total	177	853.8	467.1

CALCULATING A *T*-TEST

Calculate the *t*-test:

$t = \frac{\bar{Y}_1 - \bar{Y}_2}{se(\bar{Y}_1 - \bar{Y}_2)}$, the numerator is easy, focus on the denominator: $se(\bar{Y}_1 - \bar{Y}_2)$

- Calculating the standard error of the difference of means:

$$se(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\left(\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\right)} \text{ X } \sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$se(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\left(\frac{(124-1)(466.1)^2+(53-1)(421.4)^2}{124+53-2}\right)} \text{ X } \sqrt{\left(\frac{1}{124}+\frac{1}{53}\right)} = 74.39$$

- n_1 and n_2 are sample sizes
- s_1^2 and s_2^2 are sample variances
- Y_1 = # days in government for majority governments
- Y_2 = # days in government for minority governments

Calculating the t-statistic:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{se(\bar{Y}_1 - \bar{Y}_2)} = \frac{930.5 - 674.4}{74.39} = \frac{256.1}{74.39} = 3.44$$

Government Type	No. of Observations	Mean Duration	Standard Deviation
Majority	124	930.5	466.1
Minority	53	674.4	421.4
Total	177	853.8	467.1

Finding significance

- Calculating the t-statistic:
  - $t = \frac{\bar{Y}_1 - \bar{Y}_2}{se(\bar{Y}_1 - \bar{Y}_2)} = \frac{930.5 - 674.4}{74.39} = \frac{256.1}{74.39} = 3.44$
- This means little on its own, we need to find the degrees of freedom:
  - $n_1 + n_2 - 2 = 124 + 53 - 2 = 175$
- P-value of 0.10 (10% critical value for 90% significance; two-tailed test)
  - Must have a *t*-statistic greater than or equal to 1.66
  - 3.44 > 1.66
- P-value of 0.05 (5% critical value for 95% significance; two-tailed test)
  - *t*-statistic must be greater than or equal to 1.98
  - 3.44 > 1.98

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Confidence intervals for Difference of Means

- Confidence intervals: $\bar{Y} \pm 2(\frac{s}{\sqrt{n}})$
  - Majority: $\bar{Y} \pm 2(\frac{s}{\sqrt{n}})$
    - $\bar{Y} - 2\left(\frac{s}{\sqrt{n}}\right) = 852.75$
    - $\bar{Y} + 2\left(\frac{s}{\sqrt{n}}\right) = 1008.25$
    - (852.75, 1008.25)
  - Minority: $\bar{Y} \pm 2(\frac{s}{\sqrt{n}})$
    - $\bar{Y} \pm 2\left(\frac{s}{\sqrt{n}}\right) = 616.52$
    - $\bar{Y} \pm 2\left(\frac{s}{\sqrt{n}}\right) = 732.28$
    - (616.52, 732.28)

Government Type	No. of Observations	Mean Duration	Standard Deviation
Majority	124	930.5	866.1
Minority	53	674.4	421.4
Total	177	853.8	467.1

Confidence interval as a hypothesis test:

Observed difference ± 2(std. err. of diff. of means)

$= (930.5 - 674.4) \pm 2(74.39)$
$= 256.1 \pm 148.78$
$= (107.32, 404.88)$

The confidence interval does not include 0, so we can be confidence the difference of means is significant. Reject the null. If the confidence interval includes 0, cannot reject the null.

Difference in means test: Example 2

- **Question**: Might more people vote if elections were held on the weekend? Comparison of countries.

	Mean turnout %	Sample size *n*	Standard deviation	Standard error
Held on workday	71.8	18	11.67	?
Not on workday	68.5	40	14.17	?

- Hypotheses (two-sided)
- How do we calculate SEs?

- $H_A : \mu_{workday} - \mu_{weekend} \neq 0$
- $H_0 : \mu_{workday} - \mu_{weekend} = 0$

Difference in means test: Example 2

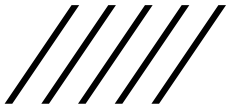
- **Question**: Might more people vote if elections were held on the weekend?

	Mean turnout %	Sample size *n*	Standard deviation	Standard error
Held on workday	71.8	18	11.67	?
Not on workday	68.5	40	14.17	?

- Standard error for each sample mean
  - $SE\ for\ workday\ countries: \frac{\sigma}{\sqrt{n}} = \frac{11.67}{\sqrt{18}} = 2.75$
  - $SE\ for\ nonworkday\ countries: \frac{\sigma}{\sqrt{n}} = \frac{14.17}{\sqrt{40}} = 2.24$

Finding the difference in SEs:

1. Calculate the standard error for each sample mean
2. Square each mean's standard error
3. Add the squared standard errors together
4. Find the square root of them



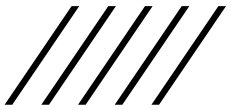
Difference in means test: Example 2

	Mean turnout %	Sample size *n*	Standard deviation	Standard error
Held on workday	71.8	18	11.67	2.75
Not on workday	68.5	40	14.17	2.24

- Calculate *standard error of the difference*.
  - 2. Square each standard error
    - $2.75^2 = 7.57, 2.24^2 = 5.02$
  - 3. Sum the two squared standard errors
    - $7.57 + 5.02 = 12.59$
  - 4. Take the square root of the result
    - $\sqrt{12.59} \approx 3.55$

Finding the difference in SEs:

1. Calculate the standard error for each sample mean
2. Square each mean's standard error
3. Add the squared standard errors together
4. Find the square root of them



Difference in means test: Example 2

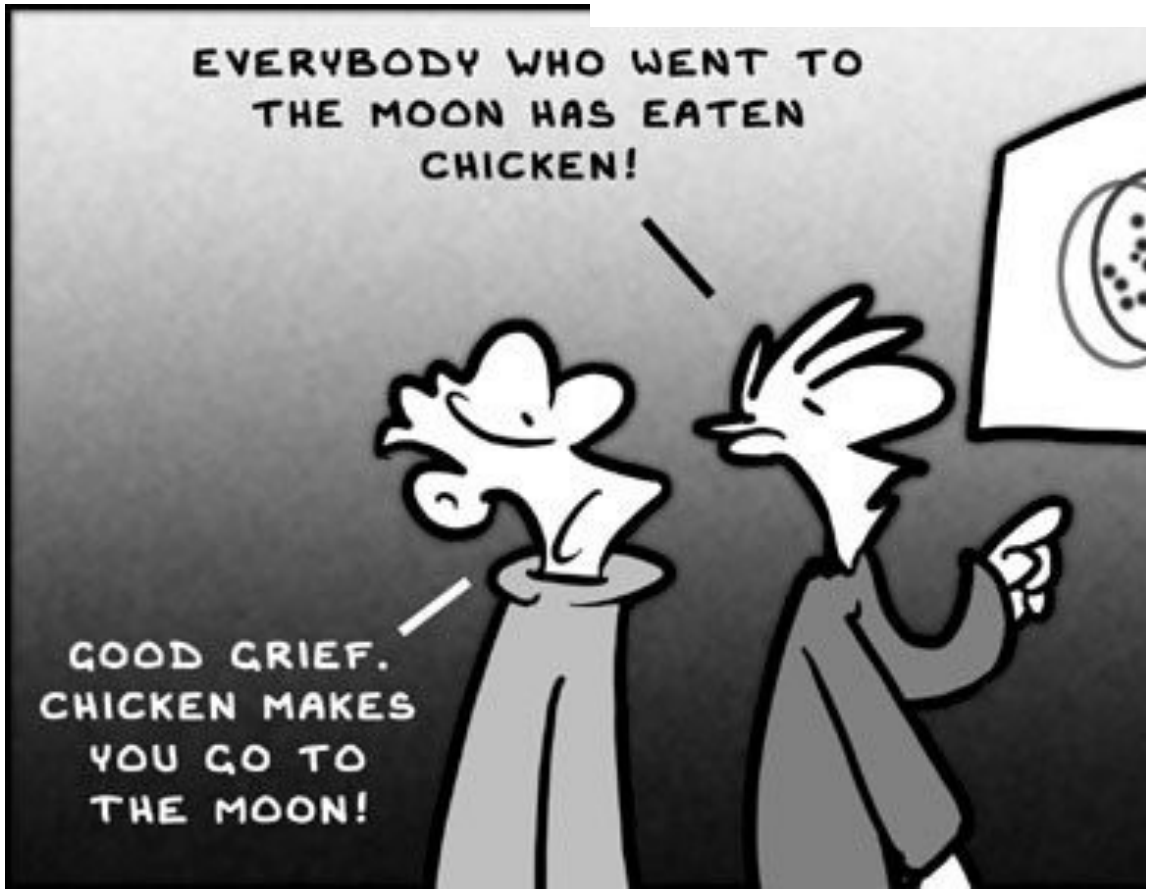
	Mean turnout %	Sample size *n*	Standard deviation	Standard error
Held on workday	71.8	18	11.67	2.75
Not on workday	68.5	40	14.17	2.24

- Standard error of the difference: $SE(\bar{X}_{workday} - \bar{X}_{weekend}) = 3.55$
  - Calculate the t-statistic:
    - $t = \frac{\bar{Y}_1 - \bar{Y}_2}{se(\bar{Y}_1 - \bar{Y}_2)} = \frac{71.8 - 68.5}{3.55} = \frac{3.30}{3.55} = 0.93$
  - Calculate the 95% confidence interval (significance-level 0.05)
    - $(71.8 - 68.5) \pm (2 \times 3.55)$
    - $=3.3 \pm 7.1$
    - Or (-3.8, 10.4)

If interval excludes 0, reject null hypothesis; otherwise “fail to reject” the null. (why 0?)

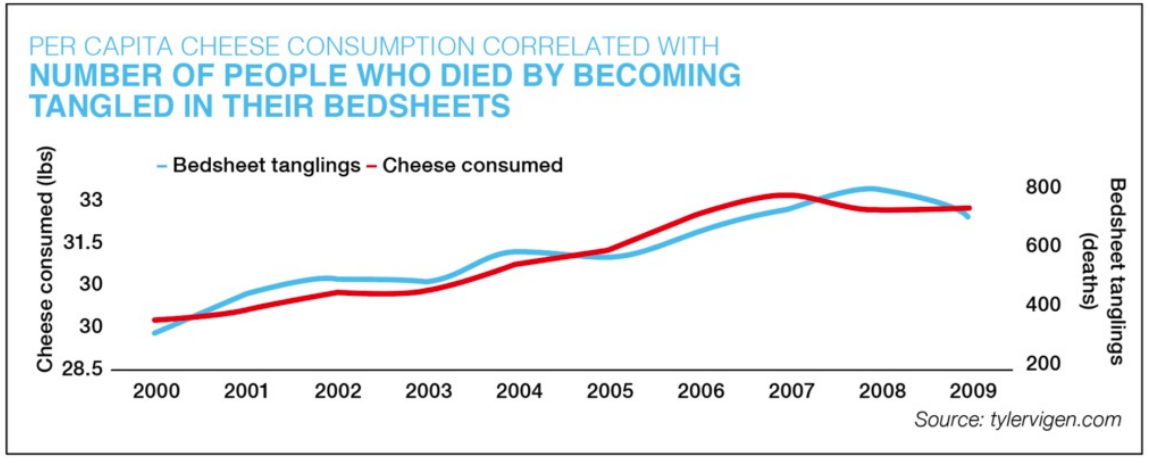
CORRELATION

POLS 095



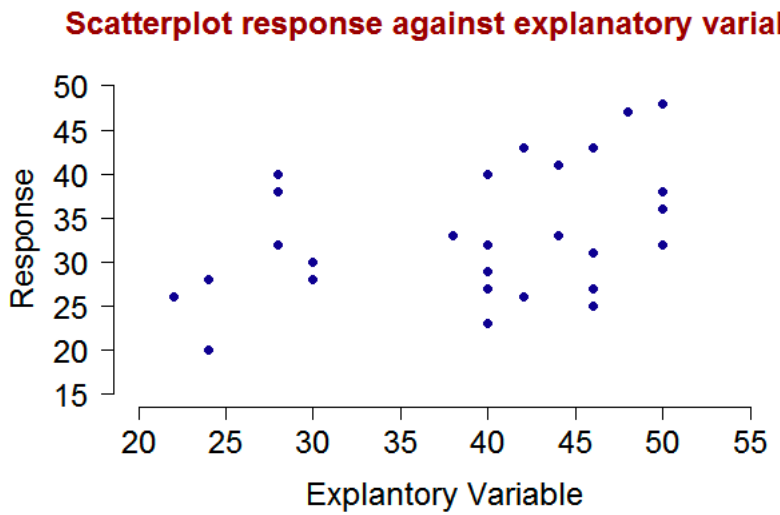
WHAT IS IT?

- Co (“together”) - Relation
- In other words, how much do the variables move together?
- When we have interval level data, we will have SO much more information…. So we can think much more precisely about how the variables change together (or not)



THE SCATTERPLOT

- We can learn a lot by looking at a scatterplot of two variables….



- Independent variable on the horizontal axis, dependent on the vertical axis
- Plot the values for each case in your dataset

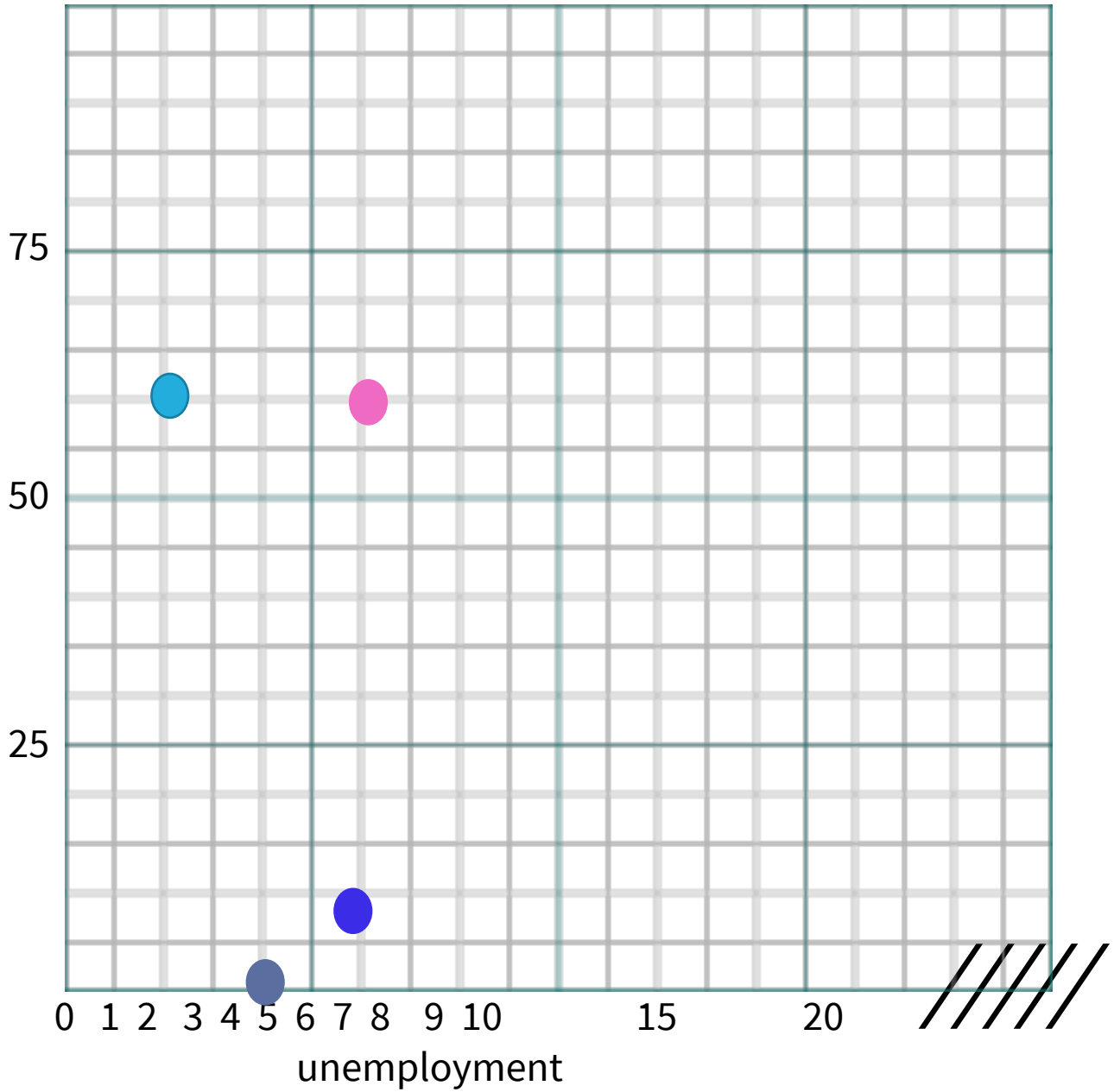
THE SCATTERPLOT

- So let’s say we are looking at the relationship between crime and unemployment in each city, with the hypothesis that:
- In a comparison of cities, those with higher rates of unemployment will have higher rates of crime than those with lower rates of unemployment.

- For each city, we have data on the number of crimes per 1000 people, and we know the rate of unemployment
- IV = unemployment
- DV = crime

PLOTTING

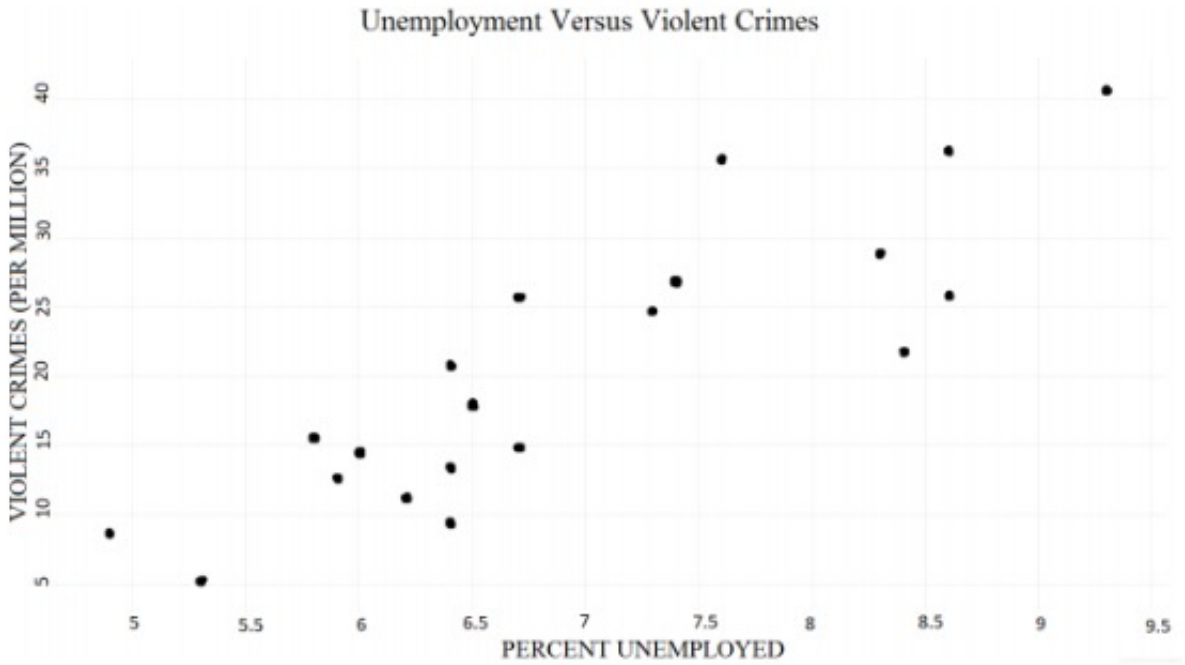
Seattle unemployment = 3% crime = 60
Tucson unemployment = 6% crime = 8
Des Moines unemployment = 4% crime = .13
New Orleans unemployment = 6% crime = 59



WE CAN LOOK AT THE RELATIONSHIP

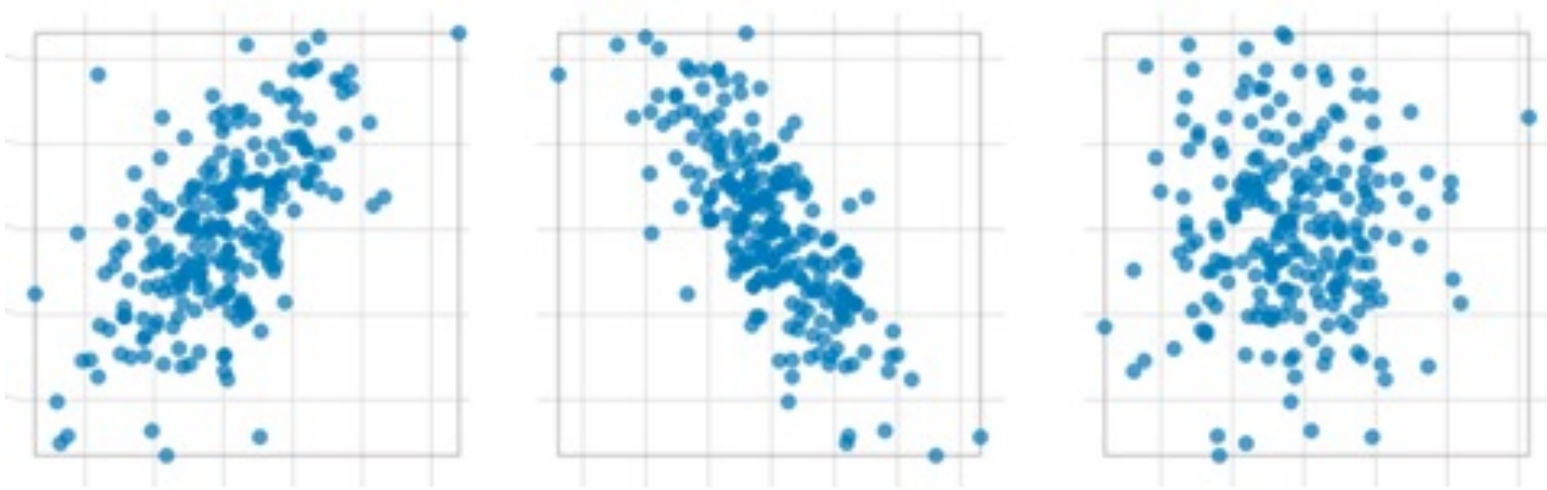
What do we see?

- Lots of variation on both IV and DV
- General upward trend
- Very few points in any clear line



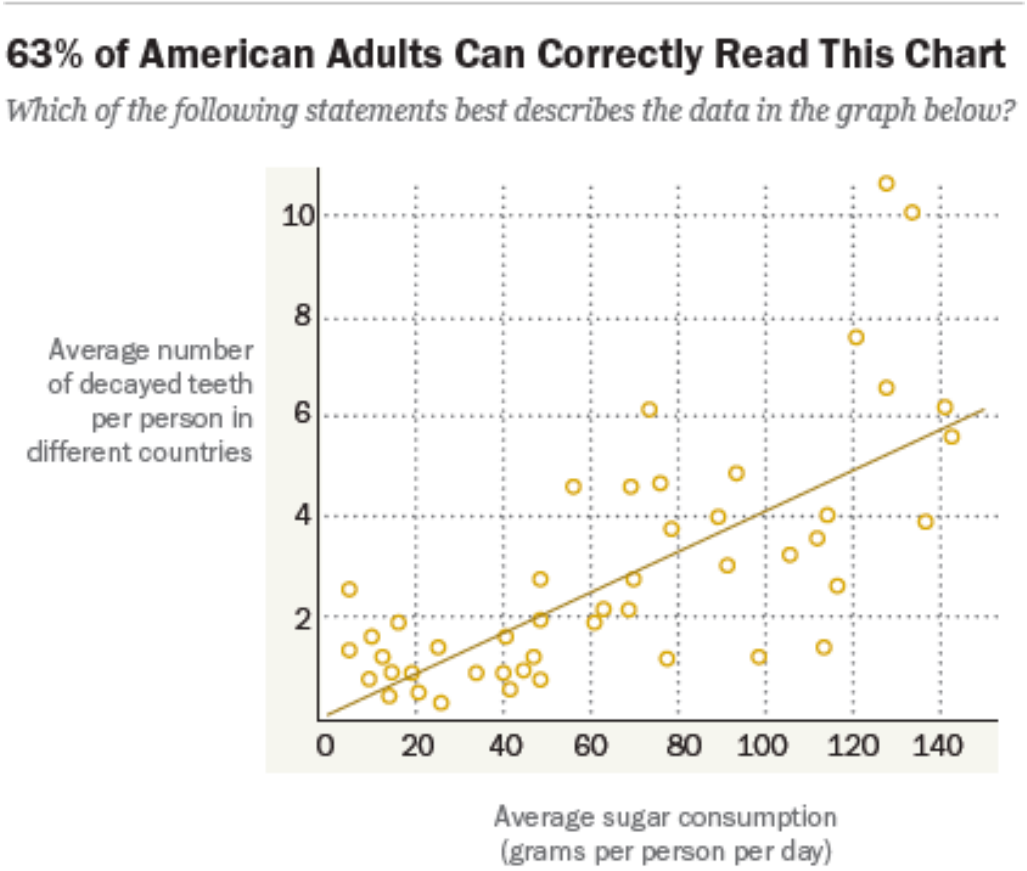
PATTERNS

- General patterns



- If higher values of the IV are associated with higher values of the DV, lower values with lower values, that is a positive relationship
- If higher values of IV are associated with lower values of the DV, that is a negative relationship.
- Or it may be unclear

READING A PATTERN



- This chart indicates that:

(a) In recent years, the rate of cavities has increased in many countries.

(b) In some countries, people brush their teeth more frequently than in other countries.

(c) The more sugar people eat, the more likely they are to get cavities.

(d) In recent years, the consumption of sugar has increased in many countries.

Source: American Trends Panel (wave 6). Survey of U.S. adults conducted Aug. 11-Sept. 3, 2014.

PEW RESEARCH CENTER

PEARSON'S R

Pearson's *r* correlation coefficient is a measure of association for 2 interval-level variables, a single summary number that expresses both the direction and the strength of a relationship.

Number will fall between 1 and -1. 1 indicates a perfect positive relationship, -1 a perfect negative relationship

Pearson's r is symmetrical, meaning that you will get the same result for two variables, regardless of which you theorize is the dependent variable and which the independent variable.

NO, YOU DO NOT NEED TO CALCULATE…

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

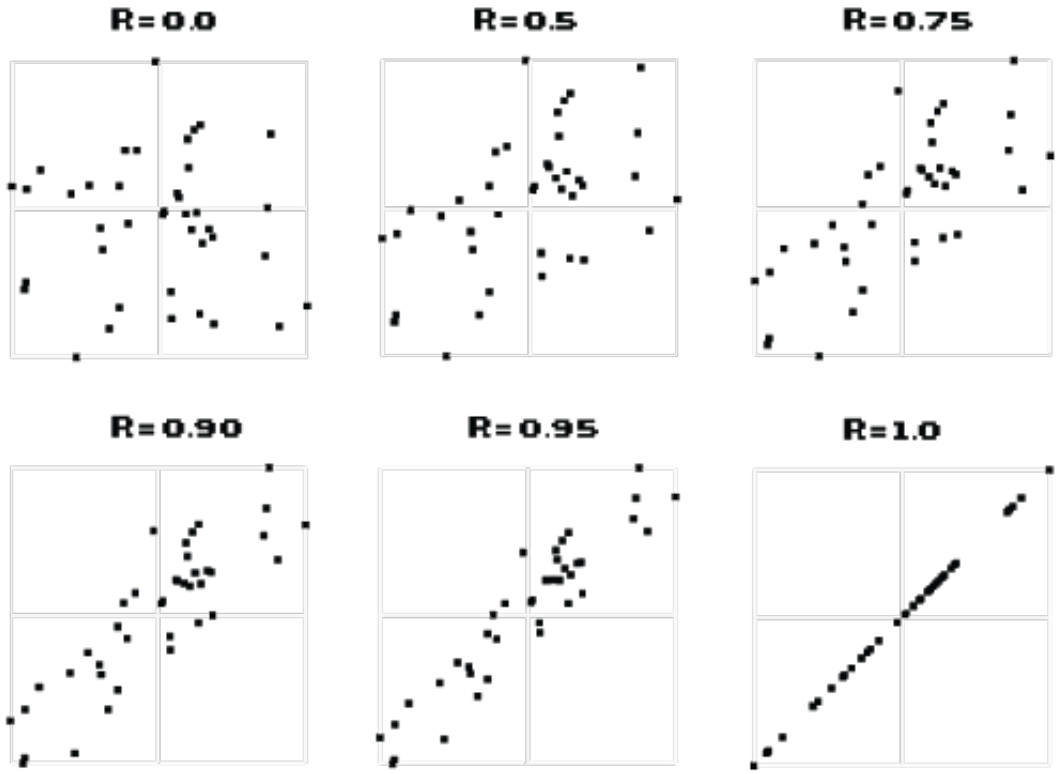
The logic of Pearson's r is fairly simple:
For each observation, it examines how far x is from the mean of x and how far y is from the mean of y, standardized by the standard deviation of x or the standard deviation of y

In other words, it is basically doing the same thing you do with Z-scores – standardizing how far from the mean an observation is.

If an observation is below the mean on both x and y, it will have a positive effect because it is a positive relationship.

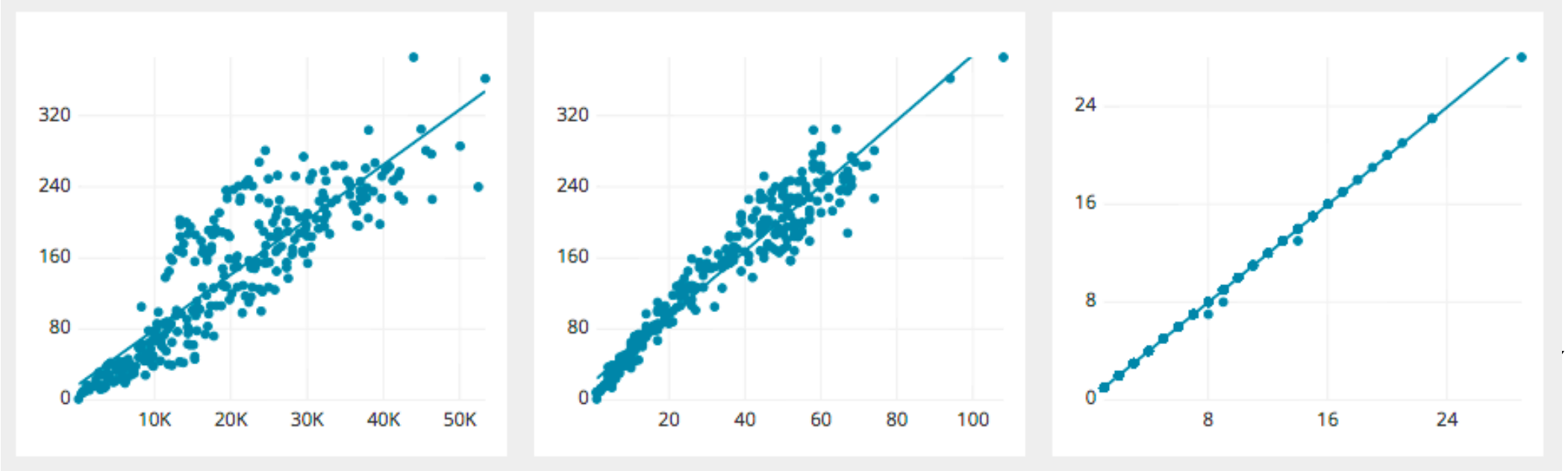
INTERPRETING PEARSON'S R

- The deviation from the mean is telling you how tightly packed together the data are.
- R = 0 means there is no relationship. Higher values indicate a stronger relationship. The sign indicates the direction of the relationship



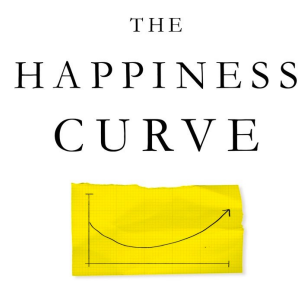
USING THE LAW OF LARGE NUMBERS

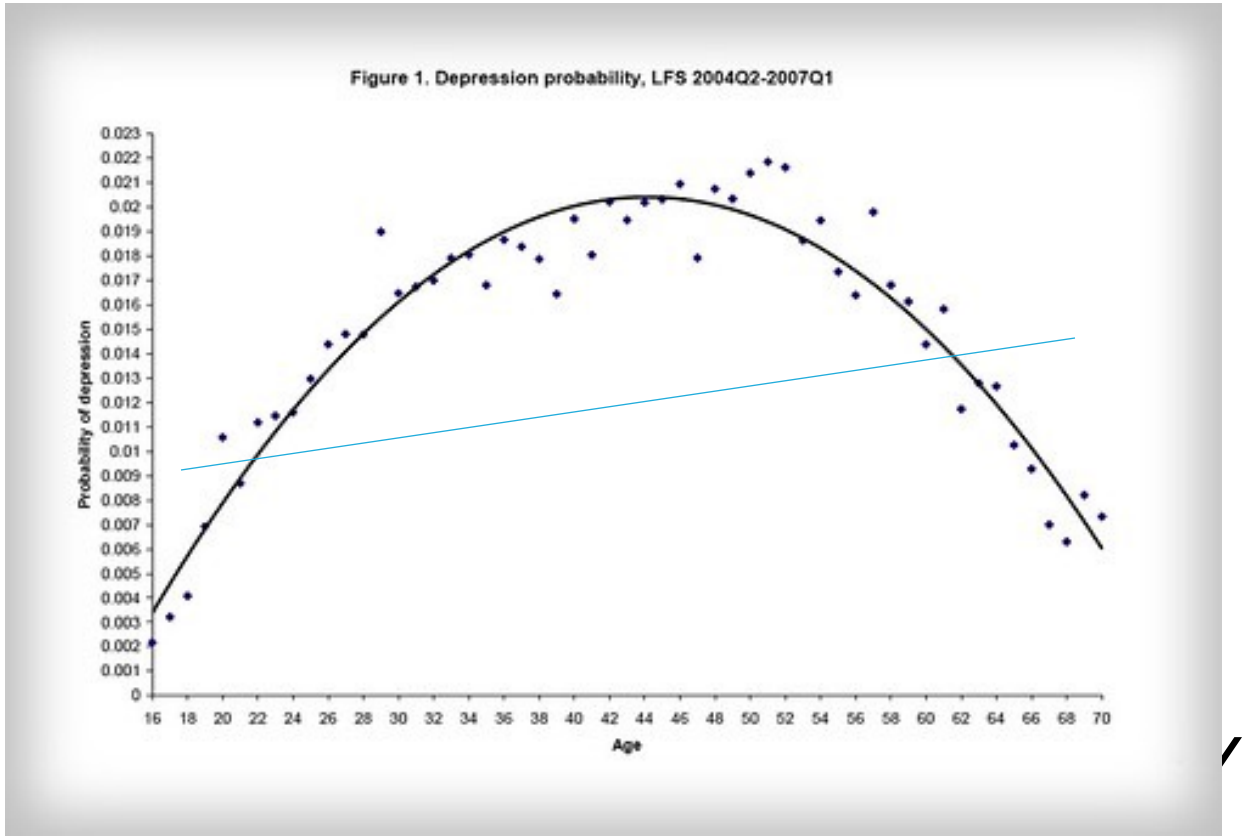
- Importantly, Pearson's r assumes a linear relationship. One good reason to look at your scatterplot first is to ensure that this is a good assumption.



WHAT IF YOUR DATA LOOKS LIKE THIS?

This is the well-known Happiness U-curve
If we assume linearity, we may miss something important…. But we assume linearity a lot!





CORRELATION IS NOT CAUSATION

- It’s a cliché for a reason!
- Establishing that variables change together in patterned ways does not mean that changes in one variable are *causing* changes in the other variable.
- How do we assess causality? We will add controls!