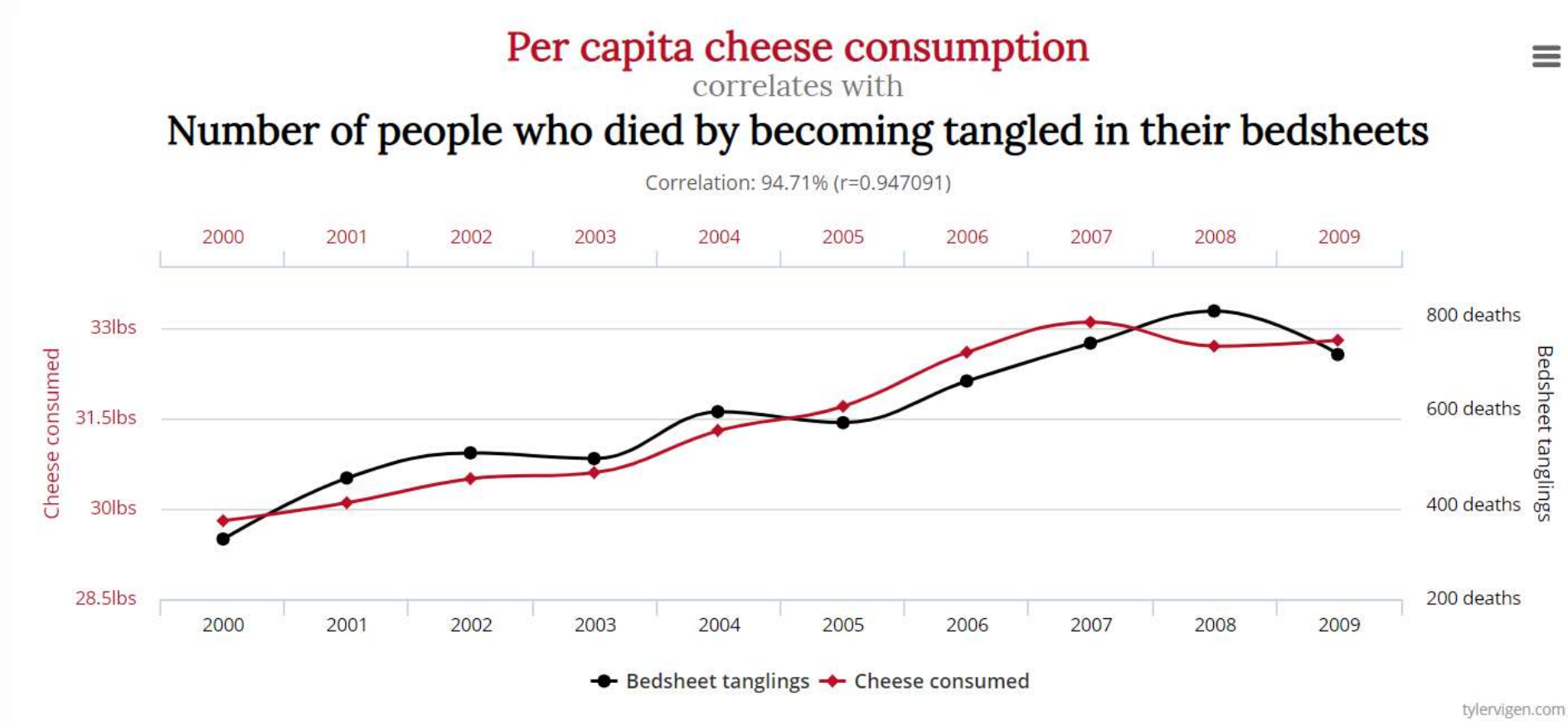
Introduction to Regression

Coming up…

- Lab Assignment 8 due Friday 4/5 11:55pm (extension)
- Peer Evaluation 2 due Tuesday 4/9 11:55 pm
- Problem Set 6 due Wednesday 4/10
- Readiness Assessment 7 Wednesday 4/10
- Don't forget the Project Stage 2 due in ~2 weeks



Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

What's new in Unit 6?

45 secs.

Types of Variable(s) Involved	Population Parameter	Analyses and Ways to Conduct Them (that we know so far): Confidence Interval for the Population Parameter	Hypothesis Test for the Population Parameter *Ho: Pop. Param = #* *Ha: Pop. Param (≠ or < or >) #*
Single **Numerical** Variable	**μ**	CLT Confidence Interval (Unit 3+4) Bootstrap Confidence Interval (Unit 4)	CLT Hypothesis Test (Unit 3+4) Bootstrap Hypothesis Test (Unit 4)
	μdiff	CLT Confidence Interval (Unit 4) Bootstrap Confidence Interval (Unit 4)	CLT Hypothesis Test (Unit 4) Bootstrap Hypothesis Test (Unit 4)
	Median	Bootstrap Confidence Interval (Unit 4)	Bootstrap Hypothesis Test (Unit 4)
Single **Categorical** Variable **(2 levels)**	**p**	CLT Confidence Interval (Unit 5) Bootstrap Confidence Interval (Unit 4)	CLT Hypothesis Test (Unit 5) Bootstrap Hypothesis Test (Unit 4) Randomization Testing (Unit 5-Selecting ball/chips out of bag, rolling dice)
Numerical Response Variable **Categorical Explanatory** Variable **(2 levels)**	**μ1-μ2**	CLT Confidence Interval (Unit 4) Bootstrap Confidence Interval (Unit 5)	CLT Hypothesis Test (Unit 4) Randomization Testing (Unit 1-Shuffling Cards)
	Median1-Median2	Bootstrap Confidence Interval (Unit 5)	Randomization Testing (Unit 1-Shuffling Cards)
Categorical Response Variable **Categorical Explanatory** Variable **(both have 2 levels)**	**p1-p2**	CLT Confidence Interval (Unit 5) Bootstrap Confidence Interval (Unit 5)	CLT Hypothesis Test (Unit 5) Randomization Testing (Unit 1-Shuffling Cards)
Numerical Response Variable **Numerical and/or Categorical Explanatory** Variable(s)	**βi (i=0,1,…,k)**	Regression Coefficient Confidence Interval (Units 6 +7)	Regression Coefficient Hypothesis Test (Units 6 +7)

How could we answer this research question?

- "Is there an association between number of dependents and income?"

Sample Data (Registered Voters)

Gender *(Categorical with 2 Levels)*	Political Affiliation *(Categorical with 3 Levels)*	Voted in 2018? *(Categorical with 2 Levels)*	Age *(Numerical - Continuous)*	Income *(Numerical - Continuous)*	Number of Dependents *(Numerical - Discrete)*
Male	Democrat	Yes	22	$20,000	0
Female	Republican	Yes	39	$40,000	3
Male	Independent	Yes	47	$400,000	0
Male	Republican	Yes	18	$129,000	0
Female	Republican	No	29	$85,000	2
Female	Democrat	No	80	$72,000	1
Female	Democrat	Yes	56	$55,000	0
Male	Independent	Yes	72	$34,000	0
...	...	...	...	...	...

"Is there a linear association between number of dependents and income?"

*Numerical **explanatory** variable*

*Numerical **response** variable*

Simple Linear Regression

***provided necessary conditions are met**

$$\widehat{Income} = \beta_0 + \beta_1(\text{Num. of Dependents})$$

$Ho: \beta_1 = 0$ → No linear association

$Ha: \beta_1 \neq 0$ → Linear Association

Is there more than one way to answer this research question?

- "Is there an association between gender and income?"

https://www.wsj.com/articles/parsing-the-gender-pay-gap-1542917969

Sample Data (Registered Voters)

Gender *(Categorical with 2 Levels)*	Political Affiliation *(Categorical with 3 Levels)*	Voted in 2018? *(Categorical with 2 Levels)*	Age *(Numerical - Continuous)*	Income *(Numerical - Continuous)*	Number of Dependents *(Numerical - Discrete)*
Male	Democrat	Yes	22	$20,000	0
Female	Republican	Yes	39	$40,000	3
Male	Independent	Yes	47	$400,000	0
Male	Republican	Yes	18	$129,000	0
Female	Republican	No	29	$85,000	2
Female	Democrat	No	80	$72,000	1
Female	Democrat	Yes	56	$55,000	0
Male	Independent	Yes	72	$34,000	0
...	...	...	...	...	...

"Is there an association between gender and income?"

*Numerical **response** variable*

*Categorical **explanatory** variable with 2 levels*

https://www.wsj.com/articles/parsing-the-gender-pay-gap-1542917969

Independent Means Hypothesis Test

***provided necessary conditions are met**

$Ho: \mu_{male} - \mu_{female} = 0$ → No association

$Ha: \mu_{male} - \mu_{female} \neq 0$ → Association

"Is there a linear association between gender and income?"

*Numerical **response** variable*

*Categorical **explanatory** variable with 2 levels*

https://www.wsj.com/articles/parsing-the-gender-pay-gap-1542917969

Simple Linear Regression

***provided necessary conditions are met**

$$\widehat{Income} = \beta_0 + \beta_1(\text{Gender: Female})$$

$Ho: \beta_1 = 0$ → No linear association

$Ha: \beta_1 \neq 0$ → Linear Association

What's new in Unit 6? New "observation" that follows a normal distribution under certain conditions.

If $X \sim Bin(n, p)$ and when certain conditions are met...

$X \sim N\left(mean = np, standard\ dev. = \sqrt{np(1-p)}\right)$

When other certain conditions are met...

$\bar{x} \sim N(mean = \mu, standard\ dev./error = \frac{\sigma}{\sqrt{n}})$

When other certain conditions are met...

$\bar{x}_1 - \bar{x}_2 \sim N(mean = \mu_1 - \mu_2, standard\ dev./error = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$

When other certain conditions are met...

$\hat{p} \sim N(mean = p, standard\ dev./error = \sqrt{\frac{p(1-p)}{n}})$

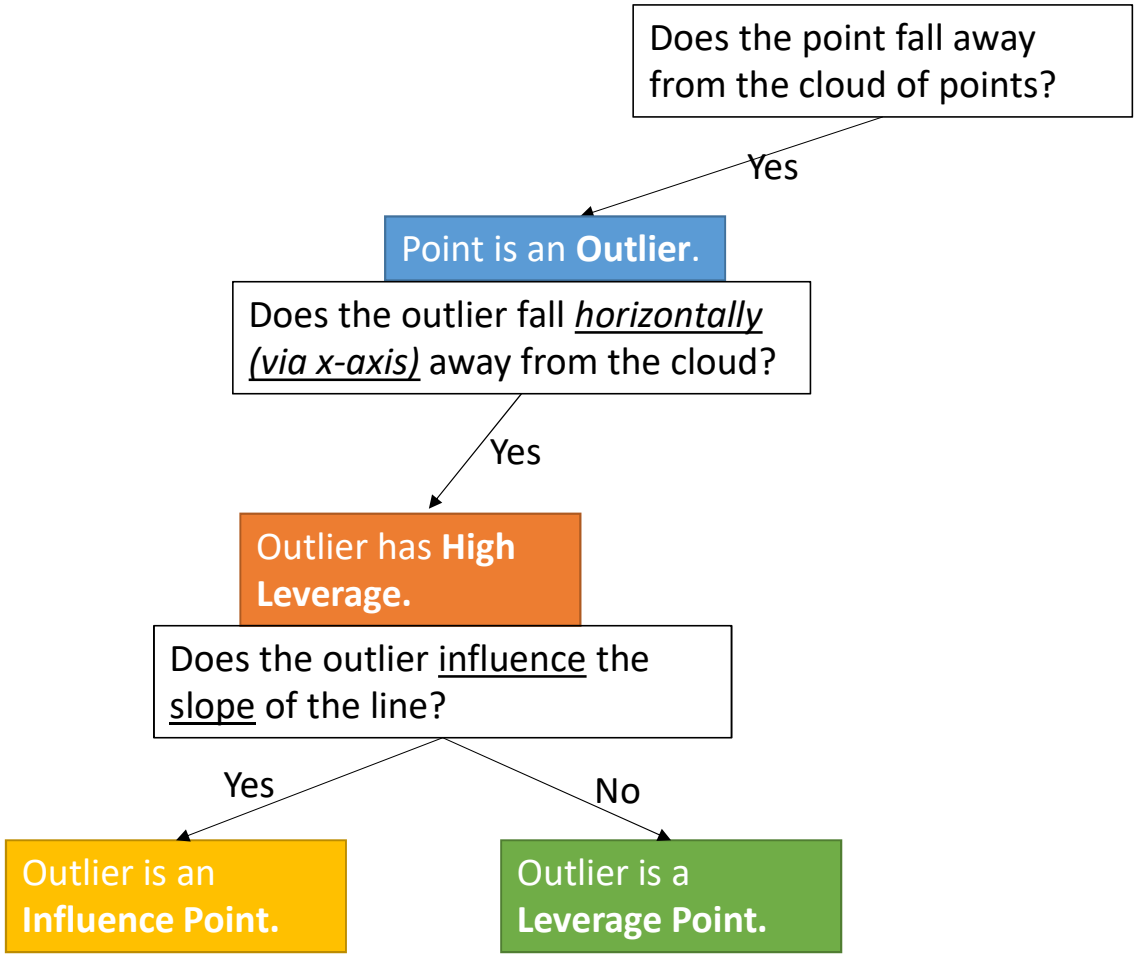
When other certain conditions are met...

$\hat{p}_1 - \hat{p}_2 \sim N\left(mean = p_1 - p_2, standard\ dev/error = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$

When other certain conditions are met...

$b_i \sim N(mean = \beta_i, standard\ dev./error = \_\_\_\_\_)$

Outliers


Does the point fall away from the cloud of points?
Yes
Point is an **Outlier**.
Does the outlier fall *horizontally (via x-axis)* away from the cloud?
Yes
Outlier has **High Leverage.**
Does the outlier influence the slope of the line?
Yes
No
Outlier is an **Influence Point.**
Outlier is a **Leverage Point.**