Where we’ve been

4/28/2021

By the end of the semester, you should be able to:

1. Understand how the way that data is collected affects what you can learn from it
2. Use statistical software to summarize this data numerically and visually
3. Build statistical models of the data. Understand which models are better and why
4. Make predictions about what kind of data you would expect to see in the future
5. Ask questions about the data, and make statistical inferences to answer them
6. Present these results in a transparent way to others
7. Understand the claims that others make from data and be able to critique them.

Main points

1. What is sampling? Why can we use samples to reason about populations? How does the process of sampling change our inferences?
2. Descriptives statistics as compression. Deciding what statistic is appropriate when.
3. What is null hypothesis testing? What do the two outcomes mean?
4. The Central Limit theorem: When it holds and what it means
   The Normal Distribution: Critical values, Z-scores, confidence intervals
5. The t-distribution and t-tests: Proportions, means, paired vs. unpaired
6. Linear regression: residuals, least-squares, assumptions and how to test them, interpreting models, how variables affect each-other
   The Generalized Linear model: Selecting the best model, transforming variables, logistic regression

Do **Americans** think the speech was good?

100 PSYCHOLOGY 1915-2015

Good: 55%
Bad : 10%

31 votes cast

Ipsos

Good: 83%
Bad : 10%

498 votes cast

StarHerald

Good : 76%
Bad : 24%
2510 votes cast

Each of these polls is a **sample**

But I want to make an inference to the **population**

When I draw a conclusion about the population from a sample, I make an **inference**.

The way I collect my sample can lead me to different inferences.

Which of these samples is the best?

What is sampling?

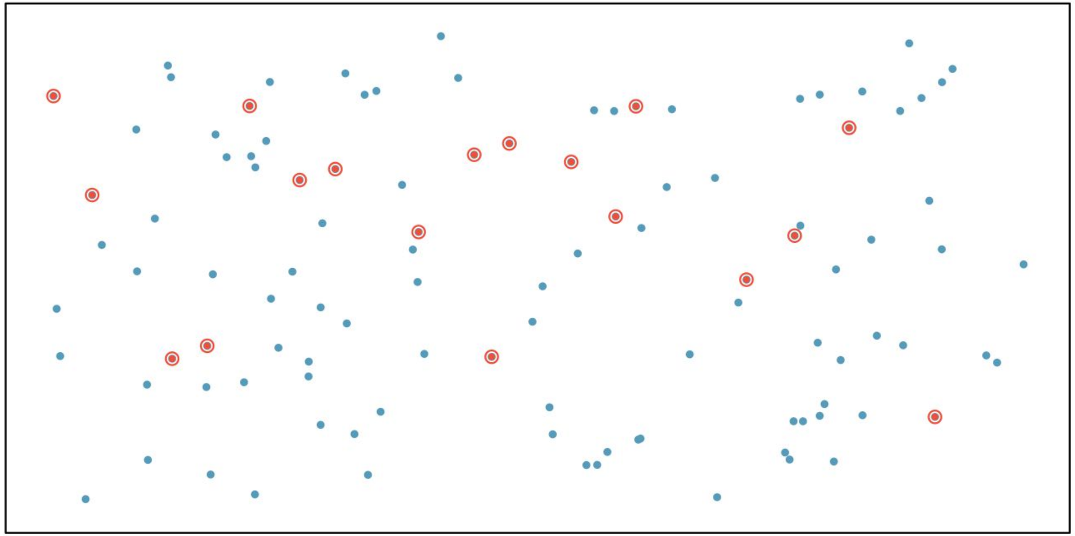
Why is bigger better?

Small samples are more **variable**.

There are 100 dots here, and 18 of them are red.

If I draw 3 dots, **more than half** the time 0 will be red.

If I draw 50 dots,
less than **1 out of 100 billion times** 0 will be red



For random samples, larger samples are more **representative**

Why can we use samples to reason about populations?

When you taste a spoonful of soup and decide the spoonful you tasted isn’t salty enough, that’s **exploratory analysis**

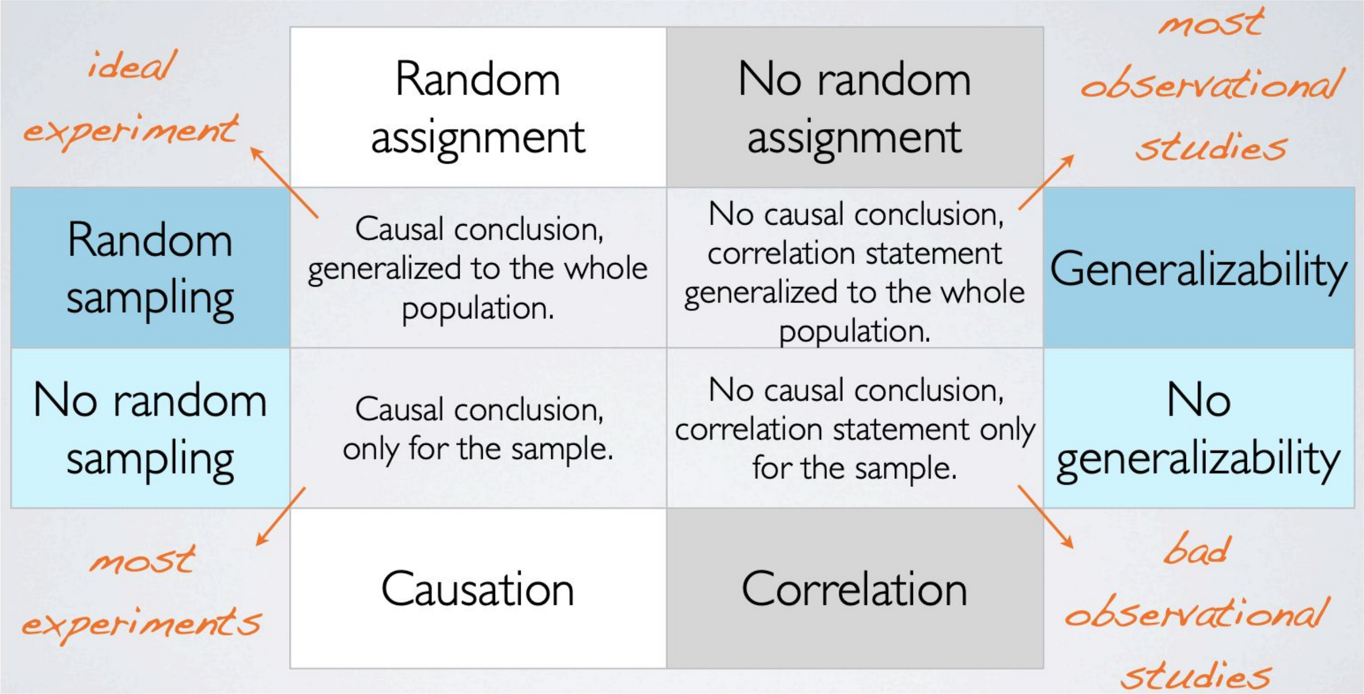
If you generalize and conclude that your entire soup needs salt, that’s an **inference**

For your inference to be valid, the spoonful you tasted (the **sample**) needs to be **representative** of the entire pot (the **population**)

If the soup is not well stirred, it doesn't matter how large a spoon you have, it will still not taste right. If the soup is well stirred, a small spoon will suffice to test the soup.

Thanks Mine Çetinkaya-Rundel

How does the process of sampling change our inferences?



	Random assignment	No random assignment	
Random sampling	Causal conclusion, generalized to the whole population.	No causal conclusion, correlation statement generalized to the whole population.	Generalizability
No random sampling	Causal conclusion, only for the sample.	No causal conclusion, correlation statement only for the sample.	No generalizability
	Causation	Correlation	

ideal experiment

most observational studies

most experiments

bad observational studies

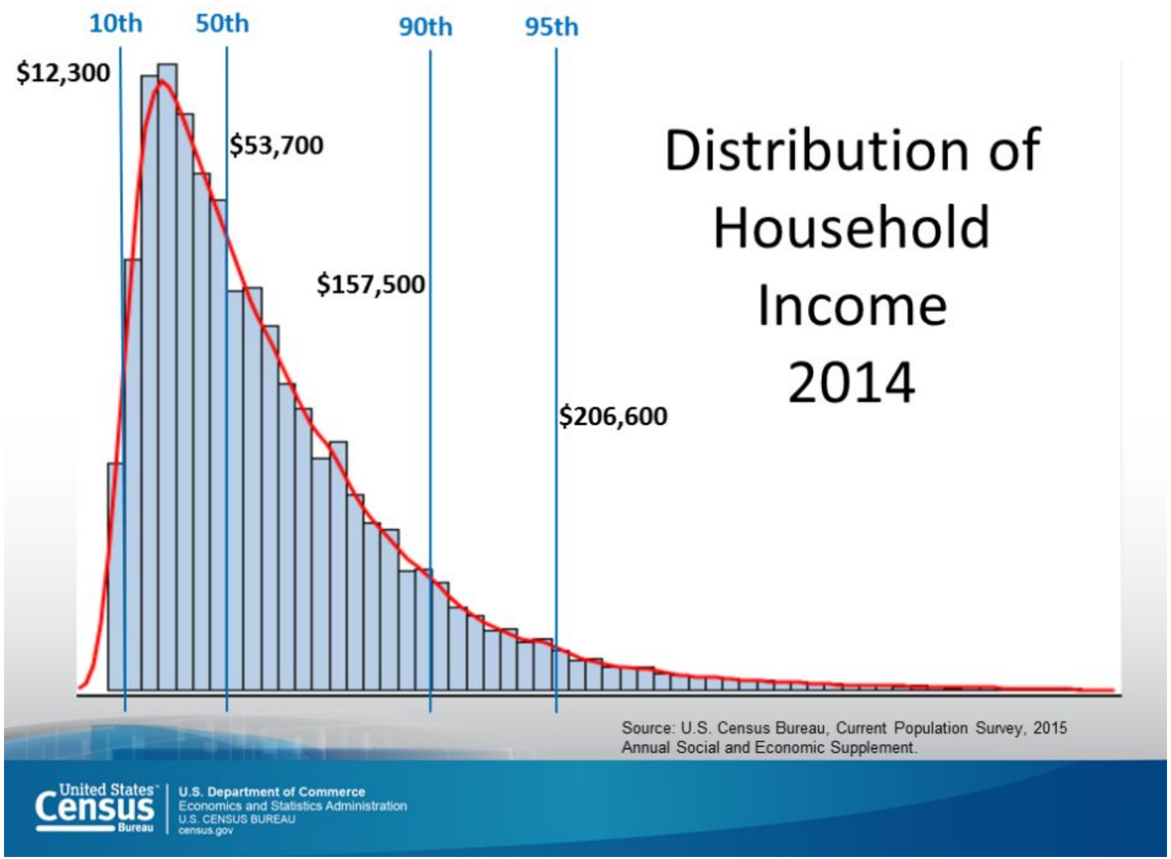
Descriptives statistics as compression

What's the difference between .mp3 and .FLAC?
.jpeg and .png?

.mp3 and .jpeg are **lossy compression** -- they make data smaller by throwing some of it away.

Central tendency is a kind of lossy compression: **What one number is the most representative of my data**?

Deciding what statistic is appropriate when


Distribution of Household Income 2014

Source: U.S. Census Bureau, Current Population Survey, 2015 Annual Social and Economic Supplement.

United States Census Bureau

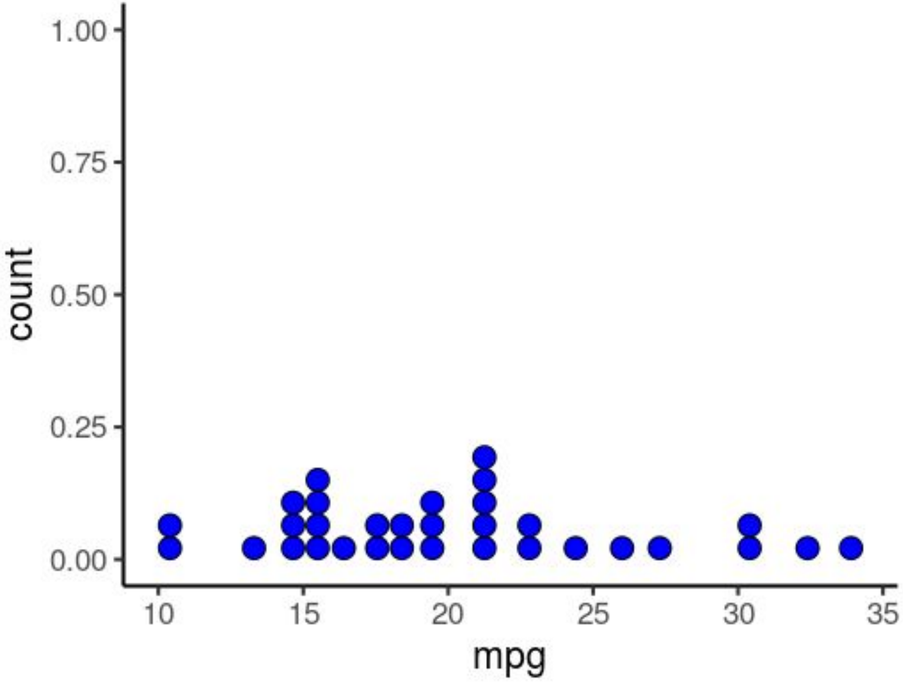
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Median: $53,700

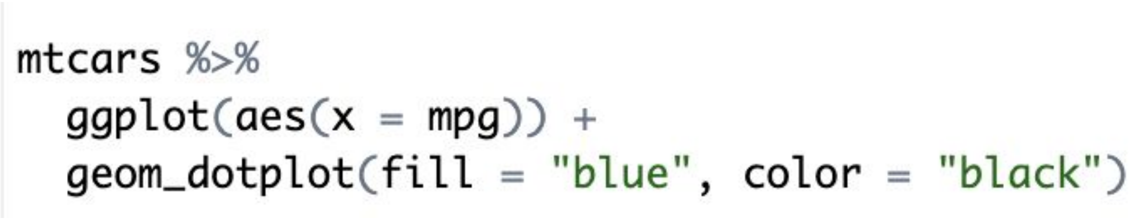
Mean: $75,738

Deciding what statistic is appropriate when

A good visualization makes your intuitions when seeing the data match the results of your statistical analyses



Dot plots make it easy to see where most of the data is.



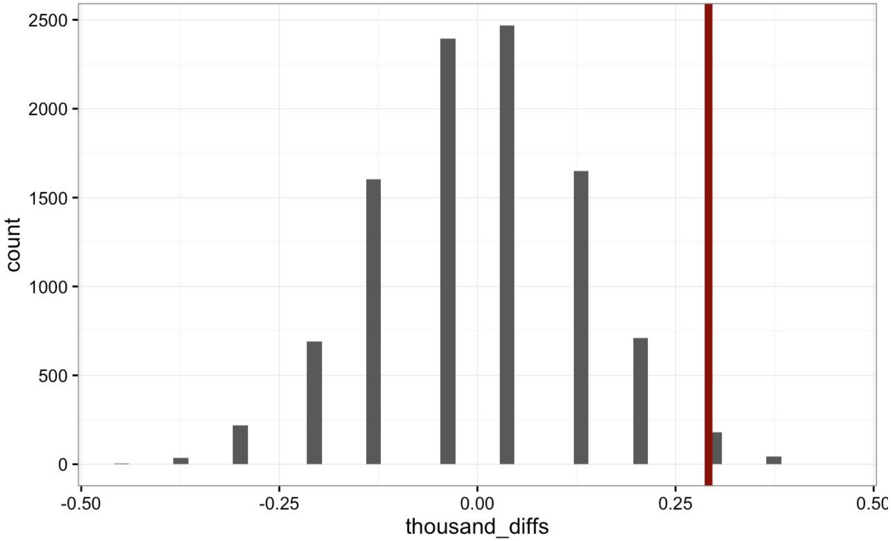
```
mtcars %>%
  ggplot(aes(x = mpg)) +
  geom_dotplot(fill = "blue", color = "black")
```

What is null hypothesis testing?

1. “There is nothing going on” (**Null Hypothesis**)
   The *process* of promotion is independent of gender
   We observed results that *look* dependent due to chance

2. “There is something going on” (**Alternative Hypothesis**)
   The *process* of promotion is dependent of gender
   We observed results that *look* dependent because they *are dependent*

What is null hypothesis testing?



If promotion is independent of gender, we should see a difference like the one we observed l*ess than 1% of the time*.

What do the two outcomes mean?

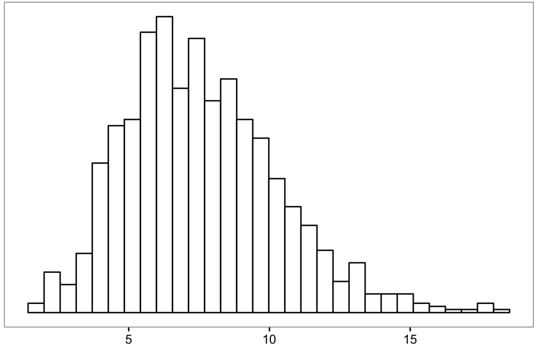
		Inference	
		Do not reject H_0	**Reject H_0 in favor of H_A**
Truth	**H_0 True**	Correct	Type I Error
	H_A True	Type II Error	Correct

Increasing our standard of evidence yields fewer Type I Errors, but more Type II Errors.

You can’t avoid this!

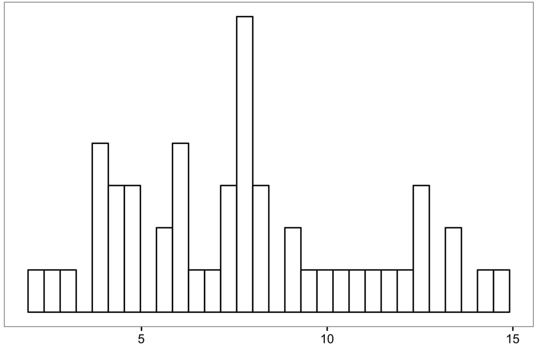
You just have to decide how important each type of error is.

The Central Limit theorem: When it holds and what it means

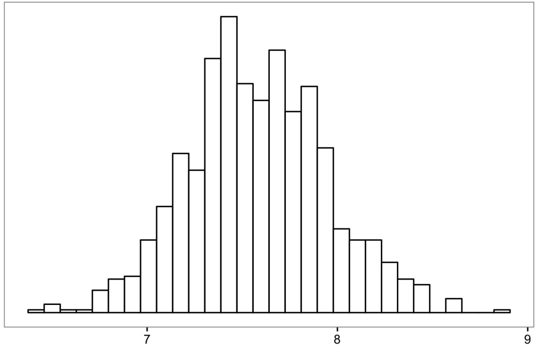


When I draw **independent samples** from the population, as sample size **approaches infinity,** the distribution of means approaches normality

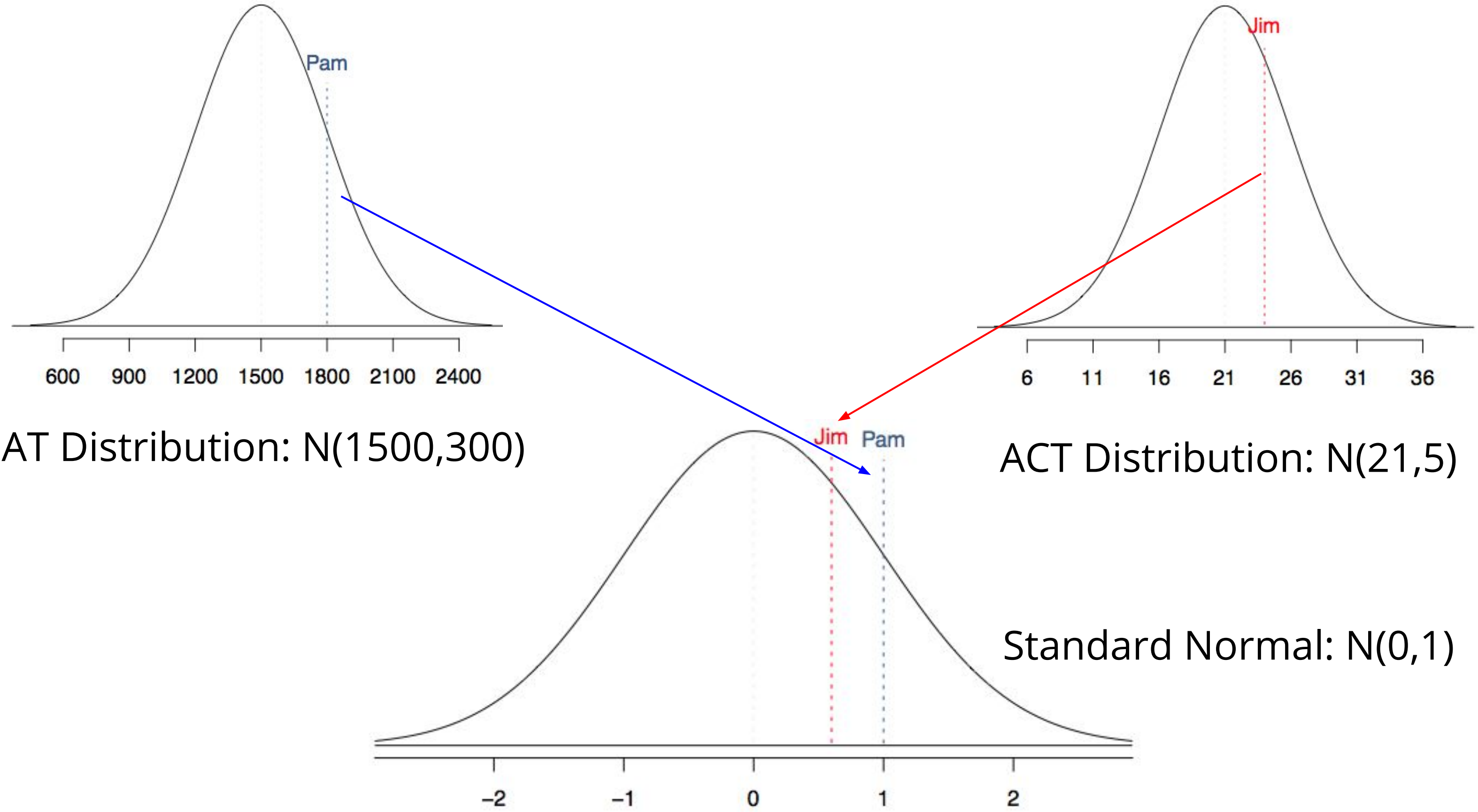
Many statistical methods we use leverage this relationship
(t-test, linear regression, ANOVA, etc)



Take the mean,
Repeat many times...



The Normal Distribution: Z-scores



SAT Distribution: N(1500,300)

ACT Distribution: N(21,5)

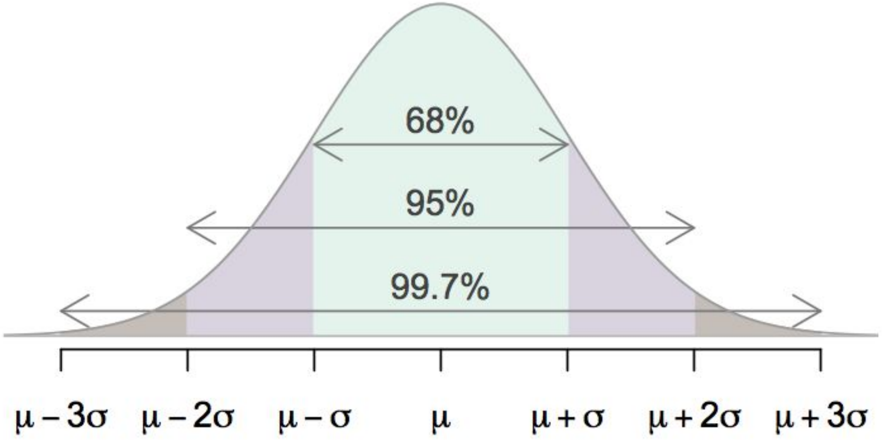
Standard Normal: N(0,1)

The Normal Distribution: Critical values

For nearly normally distributed data,

- about 68% falls within 1 SD of the mean,
- about 95% falls within 2 SD of the mean,
- about 99.7% falls within 3 SD of the mean.

It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but these occurrences are very rare if the data are nearly normal.



The Normal Distribution: Confidence intervals

A plausible range of values for the population parameter is called a *confidence interval*.

Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

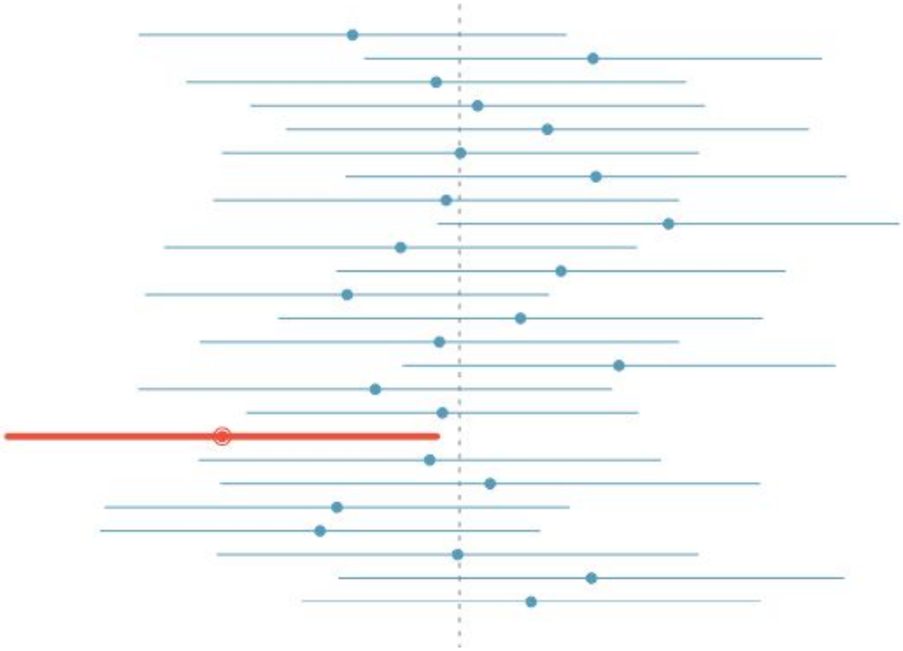
We can throw a spear where we saw a fish, but we'll probably miss. If we toss a net, we have a good chance of catching it.

If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

The Normal Distribution: Confidence intervals

Suppose we took many samples and built a confidence interval from each sample using the equation *point estimate ± 2 x SE*.

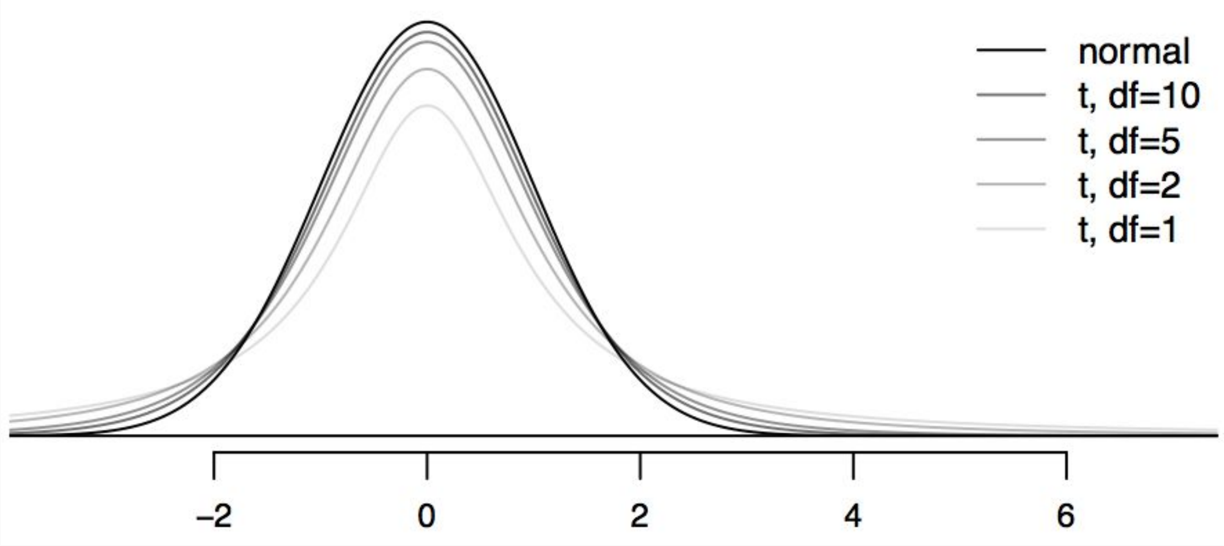
Then about 95% of those intervals would contain the true population mean (*μ*).



The t-distribution and t-tests

Centered at zero like the standard Normal (*z*-distribution).

Has only one parameter: **degrees of freedom (df)**



What happens as df increases? **Approaches the Normal (z)**

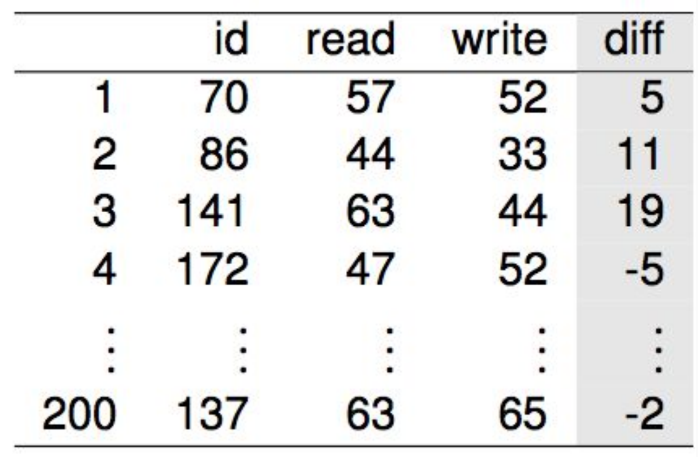
The t-distribution and t-tests: paired vs. unpaired

Two sets of data are ***paired*** if each data point in one set depends on a particular point in the other set.

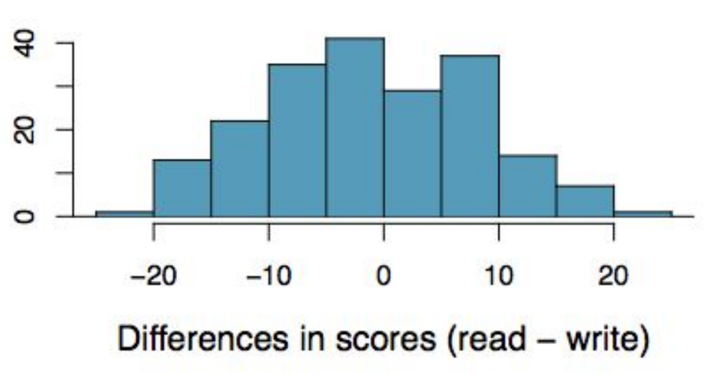
To analyze paired data, we first compute the difference between in outcomes of each pair of observations.

diff = read - write

Note: It’s important that we always subtract using a consistent order.



	id	read	write	diff
1	70	57	52	5
2	86	44	33	11
3	141	63	44	19
4	172	47	52	-5
⋮	⋮	⋮	⋮	⋮
200	137	63	65	-2



The t-distribution and t-tests: paired vs. unpaired

The test statistic for inference on the difference of two small sample means ($n_1 < 30$ and/or $n_2 < 30$) mean is the *T* statistic.

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

$$\text{point estimate} = \bar{x}_1 - \bar{x}_2$$

$$\text{null value} = 0$$

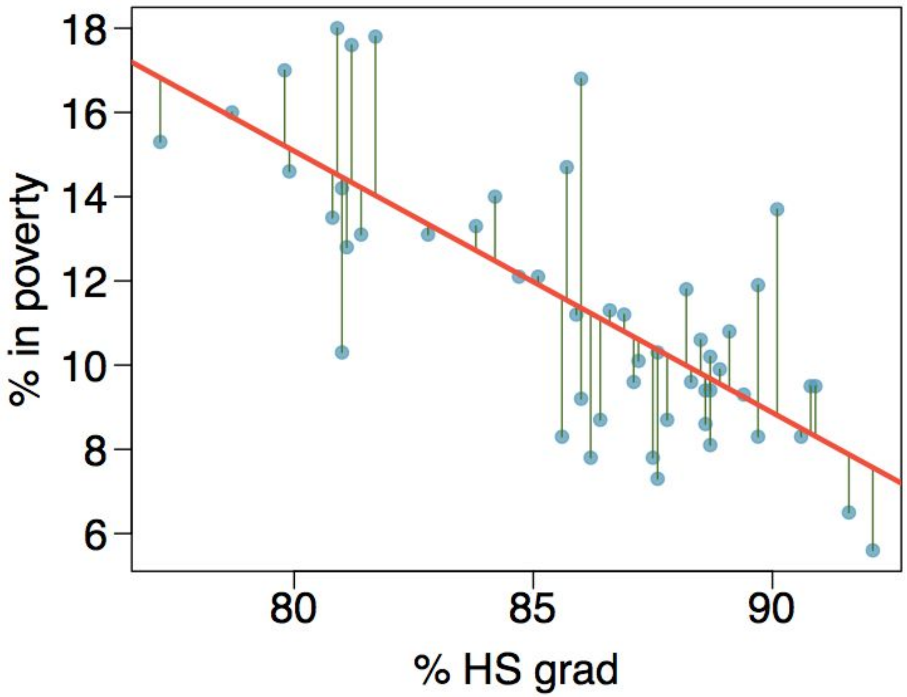
where $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ and $df = min(n_1 - 1, n_2 - 1)$

Note: the true *df* is actually different and more complex to calculate (it involves the variance in each estimate relative to its size). But this is close.

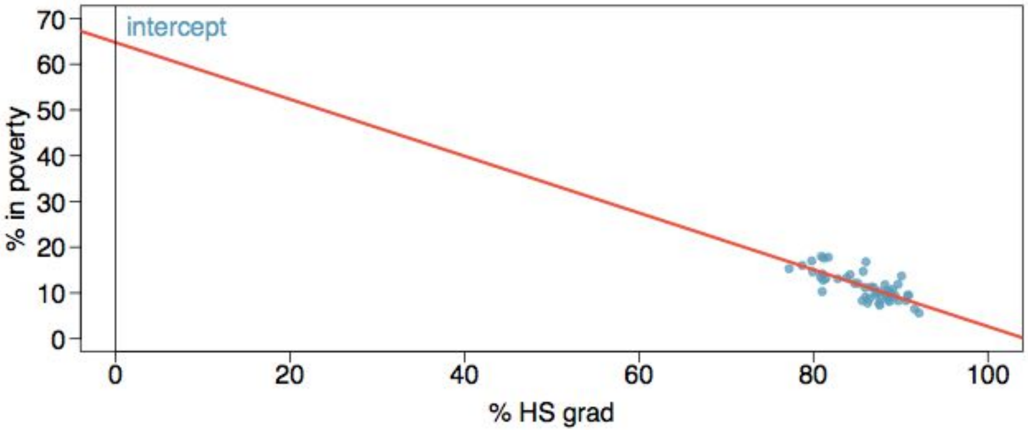
How do figure out that we want line (a)?

We want to find the line that minimizes the **residuals**: the distances between each point and the line.

A **regression** model is a model that says that your data is composed of two things:
(1) A best-fit line, and
(2) the residuals between each point and the line.



Do you believe this inference?



Since there are no states in the dataset with no HS graduates, the intercept is of no interest, not very useful, and also not reliable since the predicted value of the intercept is so far from the bulk of the data.

Conditions for least-squares regression

1. **Linearity**: The relationship between two variables must be linear

2. **Nearly normal residuals**: The errors between the line and the data are assumed to be drawn from a nearly-normal distribution

3. **Constant variability**: Assume that data are approximately equally variable at all ranges of x and y

4. **No extreme outliers**: Data points very far away from the rest can exert undue influence on the model parameters

Extrapolation: What could go wrong?

Analyzing the slope of the regression line

	estimate	std.error	t-value	p-value
(Intercept)	9.0867	6.9203	1.3130	0.1950
twin_a	0.9074	0.0700	12.956	0.0000

We always use a **t-test** in inference for regression.

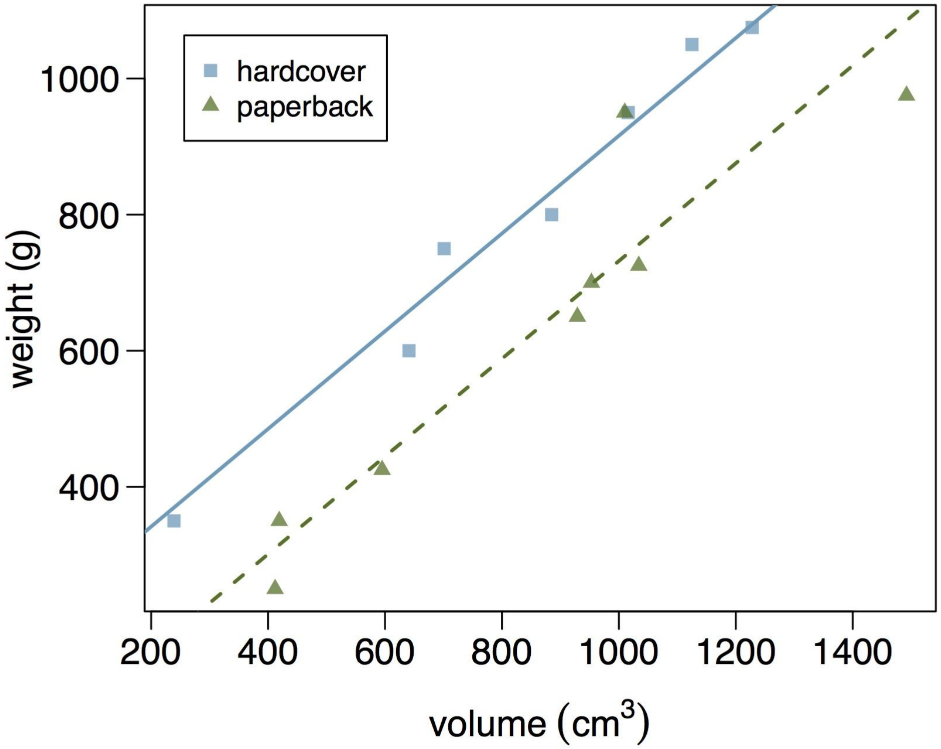
Remember: test statistic *T = (point estimate - null value) / SE*

Point estimate: b_1 is the observed slope. SE_{b1} is the standard error of the slope.

Degrees of freedom of the slope is $df = n - 2$, where n is the sample size.

Remember: we lose 1 degree of freedom for each parameter we estimate, and in simple linear regression we estimate 2 parameters, β_0 and β_1.

Visualizing the model



The Linear Model

```
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   197.96284   59.19274   3.344 0.005841 ** 
volume          0.71795    0.06153  11.669  6.6e-08 ***
cover:pb     -184.04727   40.49420  -4.545 0.000672 ***
```

$$\widehat{weight} = 197.96 + .72volume - 184.05cover{:}pb$$

For **hardcover** books: plug in **0** for cover

$$\widehat{weight} = 197.96 + .72volume - 184.05 \times \mathbf{0}$$

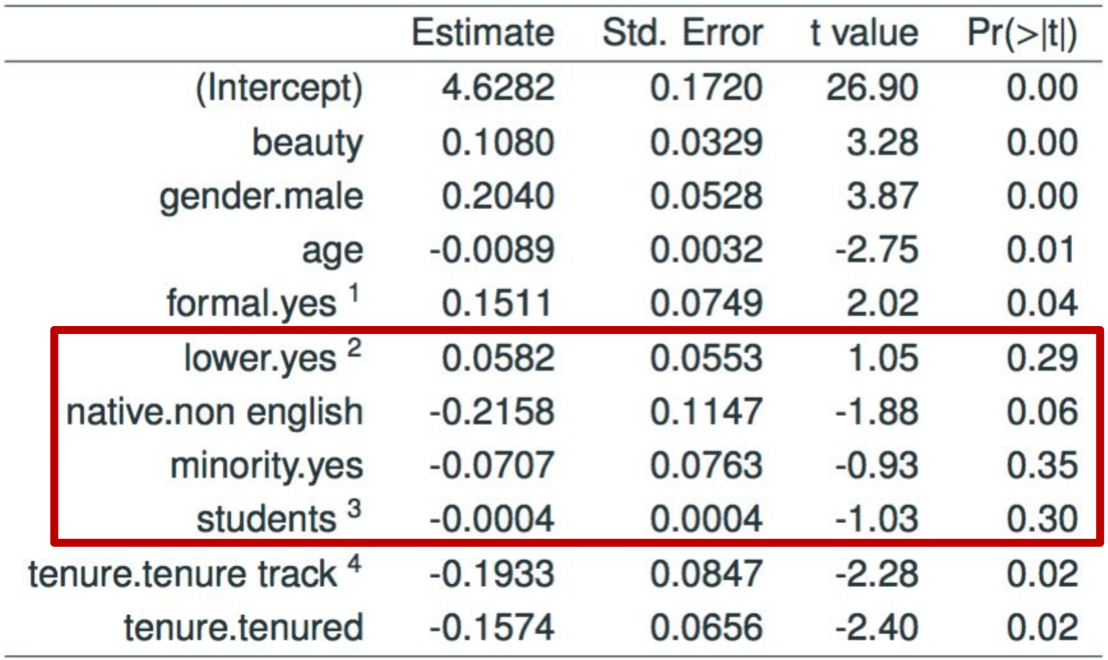
$$\widehat{weight} = 197.96 + .72volume$$

For **softcover** books: plug in **1** for cover

$$\widehat{weight} = 197.96 + .72volume - 184.05 \times \mathbf{1}$$

$$\widehat{weight} = 13.91 + .72volume$$

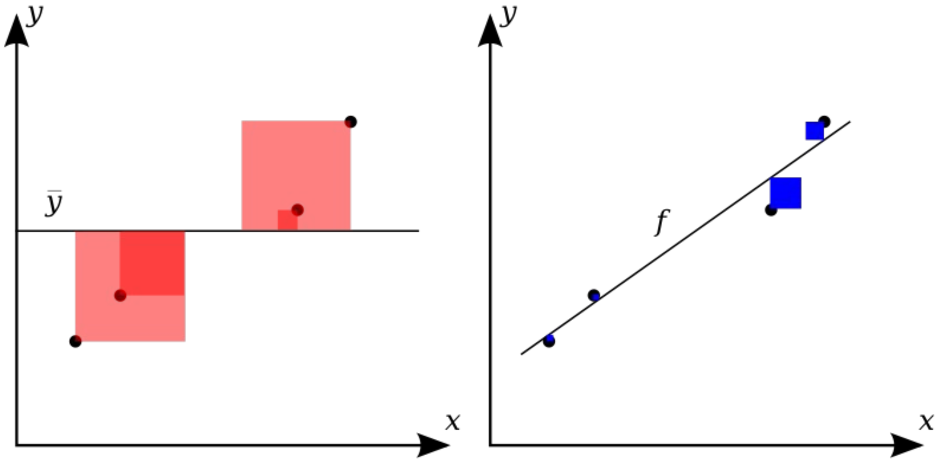
Which explanatory variables do not look like reliable predictors?



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6282	0.1720	26.90	0.00
beauty	0.1080	0.0329	3.28	0.00
gender.male	0.2040	0.0528	3.87	0.00
age	-0.0089	0.0032	-2.75	0.01
formal.yes [1]	0.1511	0.0749	2.02	0.04
lower.yes [2]	0.0582	0.0553	1.05	0.29
native.non english	-0.2158	0.1147	-1.88	0.06
minority.yes	-0.0707	0.0763	-0.93	0.35
students [3]	-0.0004	0.0004	-1.03	0.30
tenure.tenure track [4]	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

Model selection criteria: adjusted R^2

$$R^2 = 1 - \frac{SS_{resid}}{SS_{total}}$$



$$R^2_{adj} = 1 - \frac{SS_{resid}/(n-1)}{SS_{total}/(n-p-1)}$$

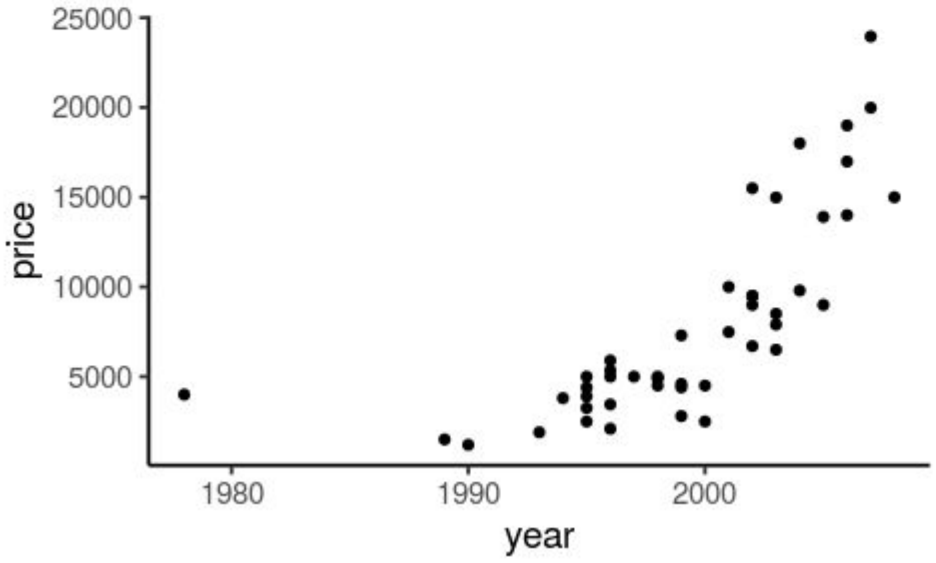
n data points

p parameters

Truck prices

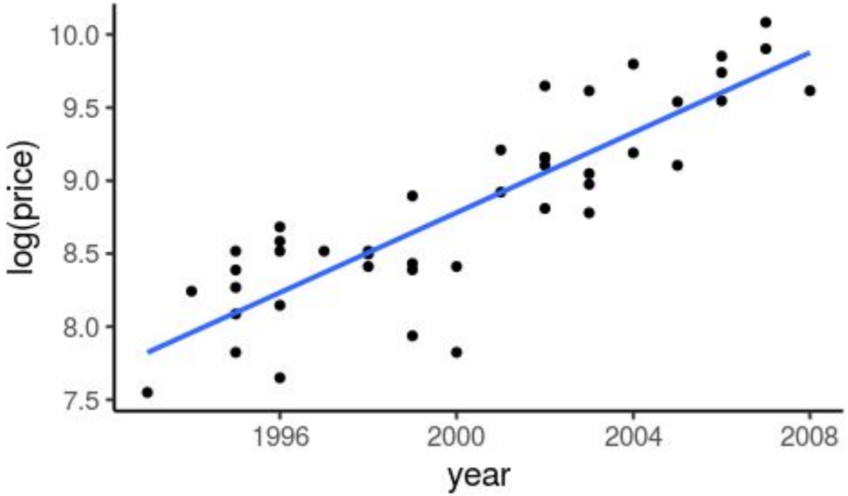
The scatterplot below shows the relationship between year and price of a random sample of 43 pickup trucks.

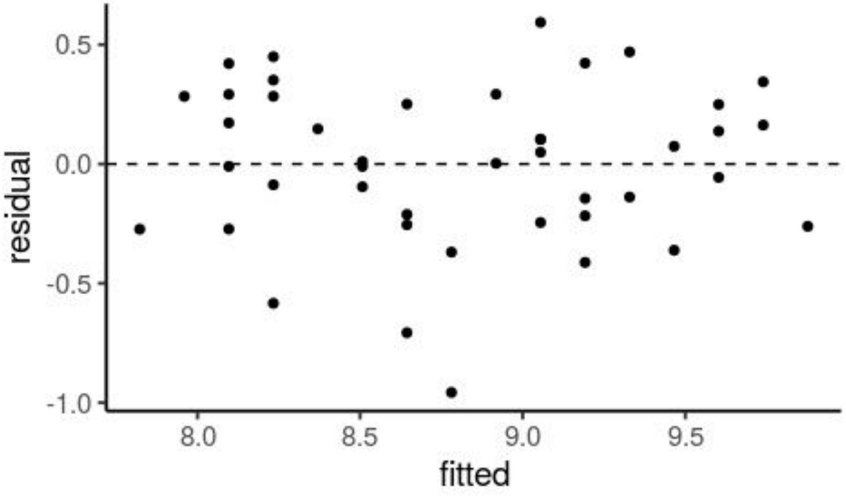
What is the relationship between these two variables?



A Log-transform of price

Model: $log(\widehat{price}) = b_0 + b_1 year$





The linear regression model

The regression models you've already seen are a special case of the Generalized Linear Model (GLM)

1. A probability distribution describing the outcome:

$$y_i = \text{Normal}(p_i, \sigma^2)$$

constant variance

Normal residuals

2. A linear model:

$$\eta = \beta_0 + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2 + \cdots + \beta_n \mathrm{x}_n$$

linearity

3. A link function that relates the linear model to the parameter of the outcome distribution

$$p = \eta$$

The logistic regression model

Logistic regression is another instantiation of the General Linear Model

1. A probability distribution describing the outcome:

$$y_i = \text{Binomial}(p_i)$$

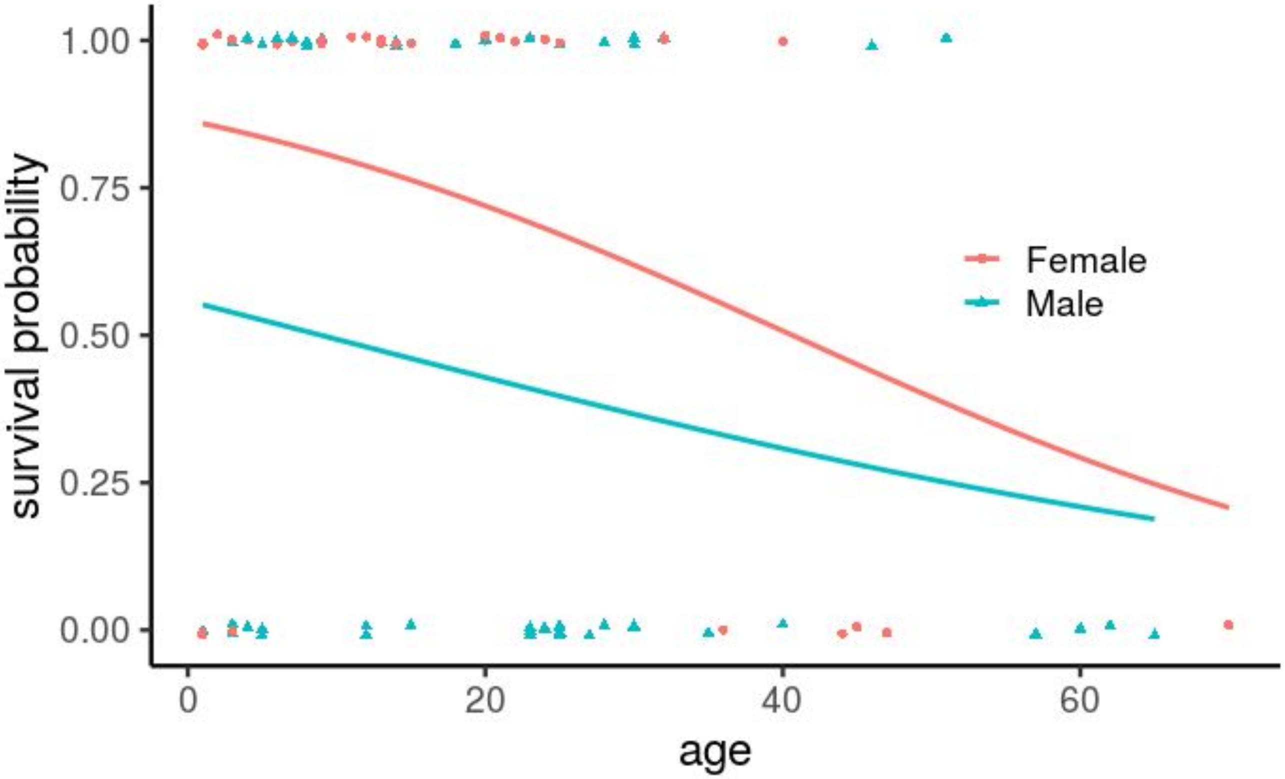
2. A linear model:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

3. A link function that relates the linear model to the parameter of the outcome distribution

$$\text{logit}(p) = n$$

Plotting the model


survival probability
age
Female
Male