Unit 8: Final Review

1.Final Exam Review

Sta 104 - Summer 2018, Term 1

Duke University, Department of Statistical Science

Slides posted at *https://www2.stat.duke.edu/courses/Summer18/sta104.001-1/*

1. Housekeeping

2. Review

- When: Wednesday, June 27 from 2:00pm-5:00pm
- What to bring:
  - Scientific calculator (graphing calculator ok, No Phones!)
  - One cheat sheet (can be typed)
- Tables provided, like past exams

1. Housekeeping

2. Review

Some comments on the project:

- Try to tell a story around your one research question.
- Your EDA should include a plot that plots two variables while controlling for a third.
- State hypotheses
- Check conditions
- Describe the process of choosing the final model and don't forget mean squared error and prediction interval.

Final Exam:

- 97 total points
- 37 points short answer questions
- 60 points remaining will be multiple choice and true/fasle

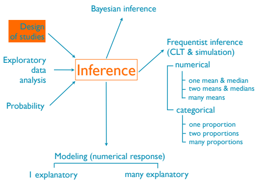
Conditions often relate to the Central Limit Theorem, which states that the sample mean has a nearly normal sampling distribution when certain conditions are met.

- Z-test for sample means
  - Numerical variable, possibly with categorical variable of 2 levels
  - Indepdence, Sample Size
- Z-test for proportions
  - Up to two categorical variables of 2 levels
  - Indepdence, Success/Failure
- T-test for sample means
  - Same as Z test for sample means except for the sample size/skew condition
  - Data need to be nearly normal or sample size large enough to correct for skew.

- ANOVA
  - Numerical variable with at least 1 categorical variable with more than 2 levels
  - Indepdence,constant variance, approximate normality
- Chi-square
  - 1 categorical variable with more than 2 levels: Test of goodness of fit
  - 2 categorical variables, with at least one with more than 2 levels: Test of independence
  - Independence, expected counts greater than 5
- Linear Regression
  - 1 numerical variable plus other explanatory variables
  - Linearity, constant variance, independence, normality of residuals for skew.
- Simulation tests
  - Randomization and bootstrapping
  - No conditions

Clicker question

A recent research study randomly divided participants into groups who were told that they were given different levels of Vitamin E to take daily. Actually, one group received only a placebo pill, and the other received Vitamin E. The research study followed the participants for eight years to see how many developed a particular type of cancer during that time period. Which of the following responses gives the best explanation as to the purpose of the random assignment in this study?



(a) To prevent skewness in the results.

(b) To reduce the amount of sampling variability.

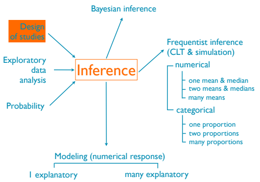
(c) To ensure that all potential cancer patients had an equal chance of being selected for the study.

(d) To produce treatment groups with similar characteristics.

(e) To ensure that the sample is representative of all cancer patients.

Clicker question

A recent research study randomly divided participants into groups who were told that they were given different levels of Vitamin E to take daily. Actually, one group received only a placebo pill, and the other received Vitamin E. The research study followed the participants for eight years to see how many developed a particular type of cancer during that time period. Which of the following responses gives the best explanation as to the purpose of the random assignment in this study?



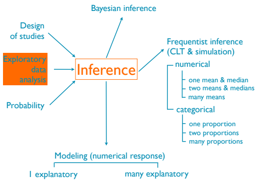
(a) To prevent skewness in the results.

(b) To reduce the amount of sampling variability.

(c) To ensure that all potential cancer patients had an equal chance of being selected for the study.

(d) ***To produce treatment groups with similar characteristics.***

(e) To ensure that the sample is representative of all cancer patients.



Clicker question

Which of the following is the most appropriate visualization for evaluating the relationship between a numerical and a categorical variable?

(a) a mosaic plot

(b) a segmented frequency bar plot

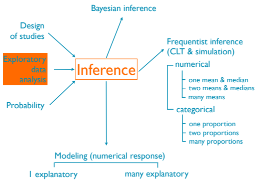
(c) a frequency histogram

(d) a relative frequency histogram

(e) side-by-side box plots

Clicker question

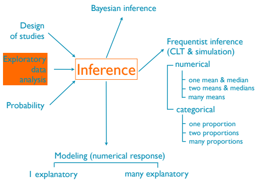
Which of the following is the most appropriate visualization for evaluating the relationship between a numerical and a categorical variable?



- (a) a mosaic plot
- (b) a segmented frequency bar plot
- (c) a frequency histogram
- (d) a relative frequency histogram
- (e) ***side-by-side box plots***

Clicker question

Which of the following is false?



(a) Box plots are useful for highlighting outliers, but we cannot determine skew based on a box plot.

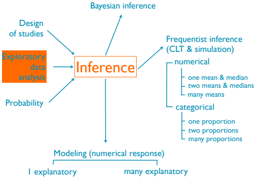
(b) Median and IQR are more robust statistics than mean and SD, respectively, since they are not affected by outliers or extreme skewness.

(c) When the response variable is extremely right skewed, it may be useful to apply a log transformation to obtain a more symmetric distribution, and model the logged data.

(d) Segmented frequency bar plots are “good enough” for evaluating the relationship between two categorical variables if the sample sizes are the same for various levels of the explanatory variable.

Clicker question

Which of the following is <u>false</u>?



(a) ***Box plots are useful for highlighting outliers, but we cannot determine skew based on a box plot.***

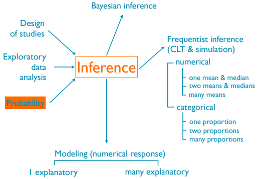
(b) Median and IQR are more robust statistics than mean and SD, respectively, since they are not affected by outliers or extreme skewness.

(c) When the response variable is extremely right skewed, it may be useful to apply a log transformation to obtain a more symmetric distribution, and model the logged data.

(d) Segmented frequency bar plots are “good enough” for evaluating the relationship between two categorical variables if the sample sizes are the same for various levels of the explanatory variable.

Clicker question

Which of the following is false?



(a) If A and B are independent, then having information on A does not tell us anything about B.

(b) If A and B are disjoint, then knowing that A occurs tells us that B cannot occur.

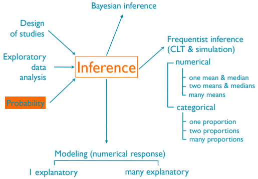
(c) Disjoint (mutually exclusive) events are always dependent since if one event occurs we know the other one cannot.

(d) If A and B are independent, then P(A and B) = P(A) + P(B).

(e) If A and B are not disjoint, then P(A or B) = P(A) + P(B) - P(A and B).

Clicker question

Which of the following is false?



(a) If A and B are independent, then having information on A does not tell us anything about B.

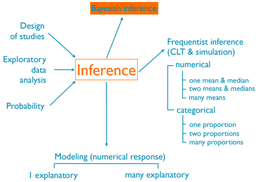
(b) If A and B are disjoint, then knowing that A occurs tells us that B cannot occur.

(c) Disjoint (mutually exclusive) events are always dependent since if one event occurs we know the other one cannot.

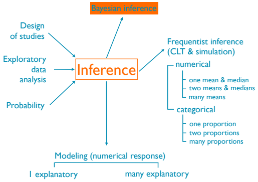
(d) ***If A and B are independent, then P(A and B) = P(A) + P(B).***

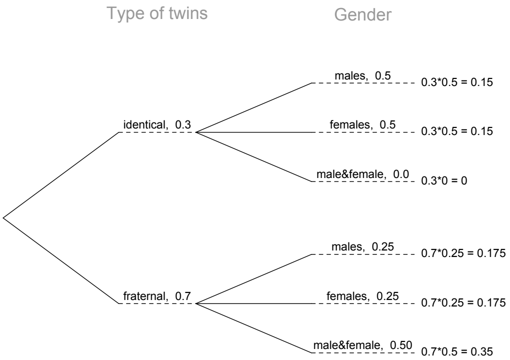
(e) If A and B are not disjoint, then P(A or B) = P(A) + P(B) - P(A and B).

About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the posterior probability that they are identical?

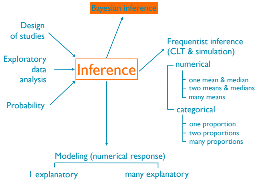


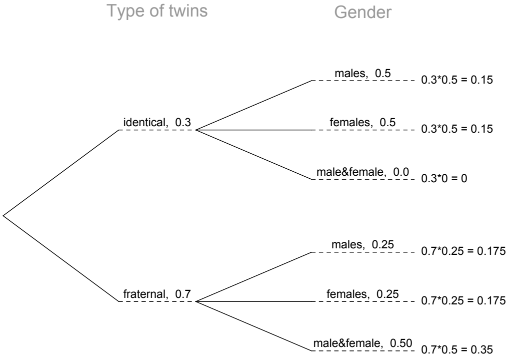
About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the posterior probability that they are identical?





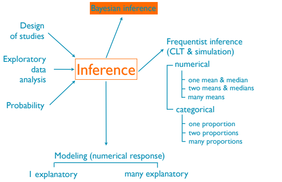
About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the posterior probability that they are identical?

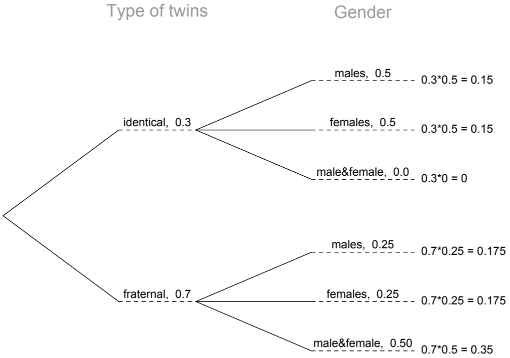




$$P(iden \mid f) = \frac{P(iden \,\&\, f)}{P(f)}$$

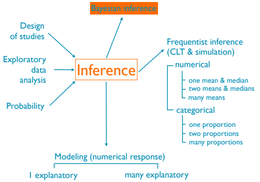
About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the posterior probability that they are identical?

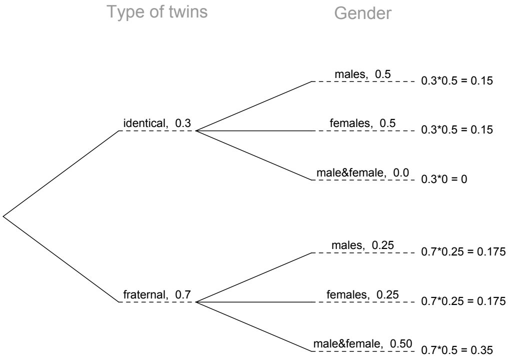




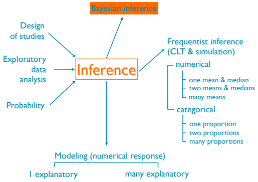
$$P(iden \mid f) = \frac{P(iden \,\&\, f)}{P(f)} = \frac{0.15}{0.15 + 0.175}$$

About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the posterior probability that they are identical?





$$
\begin{aligned}
P(iden \mid f) &= \frac{P(iden \,\&\, f)}{P(f)} \\
&= \frac{0.15}{0.15 + 0.175} \\
&= 0.46
\end{aligned}
$$



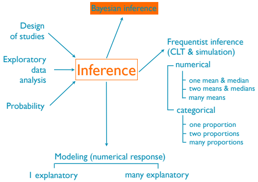
Clicker question

Which of the following is false?

- (a) Suppose you're evaluating 4 claims. If prior to data collection you don't have a preference for one claim over another, you should assign 0.25 as the prior probability to each claim.
- (b) Posterior probability and the p-value are the equivalent.
- (c) One advantage of Bayesian inference is that data can be integrated to the inferential scheme as they are collected.
- (d) Suppose a patient tests positive for a disease that 2% of the population are known to have. A doctor wants to confirm the test result by retesting the patient. In the second test the prior probability for "having the disease" should be more than 2%.

Clicker question

Which of the following is false?



(a) Suppose you're evaluating 4 claims. If prior to data collection you don't have a preference for one claim over another, you should assign 0.25 as the prior probability to each claim.

(b) ***Posterior probability and the p-value are the equivalent.***

(c) One advantage of Bayesian inference is that data can be integrated to the inferential scheme as they are collected.

(d) Suppose a patient tests positive for a disease that 2% of the population are known to have. A doctor wants to confirm the test result by retesting the patient. In the second test the prior probability for "having the disease" should be more than 2%.

Posterior = P(hypothesis | data), p-value ≈ P(data | hypothesis)

Activity:

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 6 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	8:30am section	10:05am section	11:45am section
$p - value$			
$n = 5000$	1:25pm section	3:05pm section	4:40pm section
$p - value$			

Activity:

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 6 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	8:30am section	10:05am section	11:45am section
$p - value$	0.45		
$n = 5000$	1:25pm section	3:05pm section	4:40pm section
$p - value$			

Activity:

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 6 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	8:30am section	10:05am section	11:45am section
$p - value$	0.45		
$n = 5000$	1:25pm section	3:05pm section	4:40pm section
$p - value$	0.04		

Activity:

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 6 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	8:30am section	10:05am section	11:45am section
$p - value$	0.45	0.39	
$n = 5000$	1:25pm section	3:05pm section	4:40pm section
$p - value$	0.04		

Activity:

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 6 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	8:30am section	10:05am section	11:45am section
$p - value$	0.45	0.39	
$n = 5000$	1:25pm section	3:05pm section	4:40pm section
$p - value$	0.04	0.0002	

Activity:

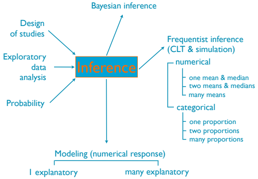
Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 6 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	8:30am section	10:05am section	11:45am section
$p - value$	0.45	0.39	29
$n = 5000$	1:25pm section	3:05pm section	4:40pm section
$p - value$	0.04	0.0002	

Activity:

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 6 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	8:30am section	10:05am section	11:45am section
$p - value$	0.45	0.39	29
$n = 5000$	1:25pm section	3:05pm section	4:40pm section
$p - value$	0.04	0.0002	≈ 0



Clicker question

Which of the following is the best method for evaluating the if the distribution of a categorical variable follows a hypothesized distribution?

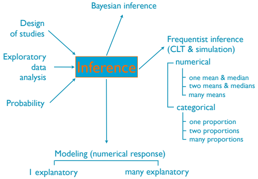
(a) chi-square test of independence

(b) chi-square test of goodness of fit

(c) anova

(d) linear regression

(e) t-test



Clicker question

Which of the following is the best method for evaluating the if the distribution of a categorical variable follows a hypothesized distribution?

(a) chi-square test of independence

(b) ***chi-square test of goodness of fit***

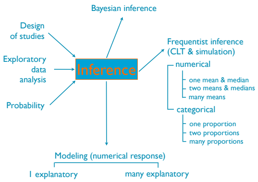
(c) anova

(d) linear regression

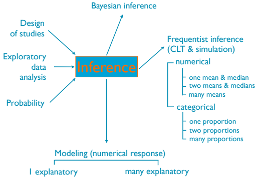
(e) t-test

Clicker question

Which of the following is the best method for evaluating the relationship between a numerical and a categorical variable with many levels?



- (a) z-test
- (b) chi-square test of goodness of fit
- (c) anova
- (d) linear regression
- (e) t-test



Clicker question

Which of the following is the best method for evaluating the relationship between a numerical and a categorical variable with many levels?

(a) z-test

(b) chi-square test of goodness of fit

(c) ***anova***

(d) ***linear regression***

(e) t-test

It is theorized that an important risk factor for breast cancer is age at first birth. An international study was set up to test this hypothesis. Breast-cancer cases were identified among women in selected hospitals in the United States, Greece, Yugoslavia, Brazil, Taiwan, and Japan. Controls were chosen from women of comparable age who were in the hospital at the same time as the cases but who did not have breast cancer. All women were asked about their age at first birth.

The set of women with at least one birth was arbitrarily divided into two categories: (1) women whose age at first birth was less than or equal to 29 years and (2) women whose age at first birth was greater than of equal to 30 years. The following results were found among women with at least one birth: 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than or equal to 30. How can we assess whether this difference is significant or simply due to chance?

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

	Breast Cancer (case)	No Breast Cancer (Controls)	Total
≤ 29			
≥ 30			
Total			

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than 30.

	Breast Cancer (case)	No Breast Cancer (Controls)	Total
≤ 29			
≥ 30			
Total			

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than 30.

	Breast Cancer (case)	No Breast Cancer (Controls)	Total
≤ 29			
≥ 30	683		
Total	3220		

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than 30.

	Breast Cancer (case)	No Breast Cancer (Controls)	Total
≤ 29	2537		
≥ 30	683		
Total	3220		

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than 30.

	Breast Cancer (case)	No Breast Cancer (Controls)	Total
≤ 29	2537		
≥ 30	683	1498	
Total	3220	10245	

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than 30.

	Breast Cancer (case)	No Breast Cancer (Controls)	Total
≤ 29	2537	8747	
≥ 30	683	1498	
Total	3220	10245	

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than 30.

	Breast Cancer (case)	No Breast Cancer (Controls)	Total
≤ 29	2537	8747	11284
≥ 30	683	1498	2181
Total	3220	10245	13465

$$n_{case} = 3220,\ n_{ctrl} = 10245$$

- cases:

$$n_{case} = 3220,\ n_{ctrl} = 10245$$

- cases: 13465 women (hospital patients) with at least one child
- variable(s):

$$n_{case} = 3220, \; n_{ctrl} = 10245$$

- cases: 13465 women (hospital patients) with at least one child
- variable(s): (1) breast cancer status - categorical, (2) age at first birth - categorical
- parameter of interest:

$$n_{case} = 3220,\ n_{ctrl} = 10245$$

- cases: 13465 women (hospital patients) with at least one child
- variable(s): (1) breast cancer status - categorical, (2) age at first birth - categorical
- parameter of interest: $p_{case} - p_{ctrl}$
  - Note: $p_{case} = P(age \geq 30|case)$ and $p_{ctrl} = P(age \geq 30|ctrl)$
- test:

$$n_{case} = 3220, \; n_{ctrl} = 10245$$

- cases: 13465 women (hospital patients) with at least one child
- variable(s): (1) breast cancer status - categorical, (2) age at first birth - categorical
- parameter of interest: $p_{case} - p_{ctrl}$
  - Note: $p_{case} = P(age \geq 30|case)$ and $p_{ctrl} = P(age \geq 30|ctrl)$
- test: compare two population proportion of independent groups
- hypotheses:

$$n_{case} = 3220,\ n_{ctrl} = 10245$$

- cases: 13465 women (hospital patients) with at least one child
- variable(s): (1) breast cancer status - categorical, (2) age at first birth - categorical
- parameter of interest: $p_{case} - p_{ctrl}$
  - Note: $p_{case} = P(age \geq 30|case)$ and $p_{ctrl} = P(age \geq 30|ctrl)$
- test: compare two population proportion of independent groups
- hypotheses: (two-tailed)

$$H_0 : p_{case} = p_{ctrl}$$
$$H_A : p_{case} \neq p_{ctrl}$$

Clicker question

Which of the following is the correct point estimate for this HT?

	BC (Case)	No BC (Controls)	Total
≤ 29	2537	8747	11284
≥ 30	683	1498	2181
Total	3220	10245	13465

(a) $\frac{683}{2181} - \frac{1498}{2181}$

(b) $\frac{683}{13465} - \frac{1498}{13465}$

(c) $\frac{2537}{11284} - \frac{683}{2181}$

(d) $\frac{683}{3220} - \frac{1498}{10245}$

(e) $\frac{683}{2181} - \frac{683}{3220}$

Clicker question

Which of the following is the correct point estimate for this HT?

	BC (Case)	No BC (Controls)	Total
≤ 29	2537	8747	11284
≥ 30	683	1498	2181
Total	3220	10245	13465

(a) $\frac{683}{2181} - \frac{1498}{2181}$

(b) $\frac{683}{13465} - \frac{1498}{13465}$

(c) $\frac{2537}{11284} - \frac{683}{2181}$

(d) $\frac{683}{3220} - \frac{1498}{10245}$ = *0.066*

(e) $\frac{683}{2181} - \frac{683}{3220}$

Clicker question

Which of the following is the correct standard error for this HT?

	BC (Case)	No BC (Controls)	Total
≤ 29	2537	8747	11284
≥ 30	683	1498	2181
Total	3220	10245	13465
$\hat{p}$	0.212	0.146	0.162

(a) $\sqrt{\frac{0.212\times(1-0.212)}{3220}} + \sqrt{\frac{0.146\times(1-0.146)}{10245}}$

(b) $\sqrt{\frac{0.212\times(1-0.212)}{3220} + \frac{0.146\times(1-0.146)}{10245}}$

(c) $\sqrt{\frac{0.162\times(1-0.162)}{3220} + \frac{0.162\times(1-0.162)}{10245}}$

(d) $\sqrt{\frac{0.212\times(1-0.212)}{13465} + \frac{0.146\times(1-0.146)}{13465}}$

(e) $\sqrt{\frac{0.162\times(1-0.162)}{13465} + \frac{0.162\times(1-0.162)}{13465}}$

Clicker question

Which of the following is the correct standard error for this HT?

	BC (Case)	No BC (Controls)	Total
≤ 29	2537	8747	11284
≥ 30	683	1498	2181
Total	3220	10245	13465
$\hat{p}$	0.212	0.146	0.162

(a) $\sqrt{\frac{0.212\times(1-0.212)}{3220}} + \sqrt{\frac{0.146\times(1-0.146)}{10245}}$

(b) $\sqrt{\frac{0.212\times(1-0.212)}{3220} + \frac{0.146\times(1-0.146)}{10245}}$

(c) $\sqrt{\frac{0.162\times(1-0.162)}{3220} + \frac{0.162\times(1-0.162)}{10245}}$ *= 0.0074*

(d) $\sqrt{\frac{0.212\times(1-0.212)}{13465} + \frac{0.146\times(1-0.146)}{13465}}$

(e) $\sqrt{\frac{0.162\times(1-0.162)}{13465} + \frac{0.162\times(1-0.162)}{13465}}$

$$Z = \frac{\hat{p}_{case} - \hat{p}_{ctrl} - 0}{SE} = \frac{0.212 - 0.146}{0.0074} = 8.92$$

$$Z = \frac{\hat{p}_{case} - \hat{p}_{ctrl} - 0}{SE} = \frac{0.212 - 0.146}{0.0074} = 8.92$$

p-value $= P(Z > 8.92) + P(Z < -8.92) \approx 0$

- Confidence level: 98%

- Confidence level: 98%
- Theoretical: Using a critical value based on the Z distr. ($z^\star$):

$$\begin{aligned} &\text{point estimate} \pm ME \\ &= \text{point estimate} \pm z^\star \times SE \end{aligned}$$

- Confidence level: 98%
- Theoretical: Using a critical value based on the Z distr. ($z^\star$):

$$\begin{aligned} &\text{point estimate} \pm ME \\ &= \text{point estimate} \pm z^\star \times SE \end{aligned}$$

For a confidence interval,

$$\begin{aligned} SE &= \sqrt{\frac{\hat{p}_{case}(1-\hat{p}_{case})}{n_{case}} + \frac{\hat{p}_{ctrl}(1-\hat{p}_{ctrl})}{n_{ctrl}}} \\ &= \sqrt{\frac{0.212(1-0.212)}{3220} + \frac{0.146(1-0.146)}{10245}} = 0.008 \end{aligned}$$

- Confidence level: 98%
- Theoretical: Using a critical value based on the Z distr. ($z^\star$):

$$\begin{aligned} &\text{point estimate} \pm ME \\ &= \text{point estimate} \pm z^\star \times SE \end{aligned}$$

For a confidence interval,

$$\begin{aligned} SE &= \sqrt{\frac{\hat{p}_{case}(1-\hat{p}_{case})}{n_{case}} + \frac{\hat{p}_{ctrl}(1-\hat{p}_{ctrl})}{n_{ctrl}}} \\ &= \sqrt{\frac{0.212(1-0.212)}{3220} + \frac{0.146(1-0.146)}{10245}} = 0.008 \end{aligned}$$

$$(0.212 - 0.146) \pm 2.33 \times 0.008 \quad \approx \quad 0.066 \pm 0.0186$$

- Confidence level: 98%
- Theoretical: Using a critical value based on the Z distr. ($z^\star$):

$$\begin{aligned} &\text{point estimate} \pm ME \\ = &\text{ point estimate} \pm z^\star \times SE \end{aligned}$$

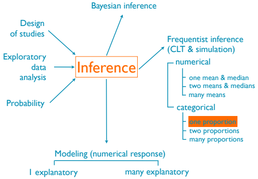
For a confidence interval,

$$\begin{aligned} SE &= \sqrt{\frac{\hat{p}_{case}(1-\hat{p}_{case})}{n_{case}} + \frac{\hat{p}_{ctrl}(1-\hat{p}_{ctrl})}{n_{ctrl}}} \\ &= \sqrt{\frac{0.212(1-0.212)}{3220} + \frac{0.146(1-0.146)}{10245}} = 0.008 \end{aligned}$$

$$\begin{aligned} (0.212 - 0.146) \pm 2.33 \times 0.008 \quad &\approx \quad 0.066 \pm 0.0186 \\ &= \quad (0.0474,\ 0.0846) \end{aligned}$$

Clicker question

$n = 30$ and $\hat{p} = 0.6$. Hypotheses: $H_0 : p = 0.8$; $H_A : p < 0.8$. Which of the following is an appropriate method for calculating the p-value for this test?



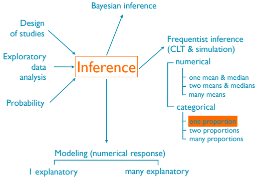
(a) CLT-based inference using the normal distribution

(b) simulation-based inference

(c) exact calculation using the binomial distribution

Clicker question

$n = 30$ and $\hat{p} = 0.6$. Hypotheses: $H_0 : p = 0.8$; $H_A : p < 0.8$. Which of the following is an appropriate method for calculating the p-value for this test?



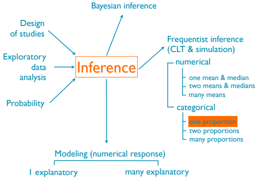
(a) ***CLT-based inference using the normal distribution***

(b) simulation-based inference

(c) ***exact calculation using the binomial distribution***

Clicker question

$n = 30$ and $\hat{p} = 0.6$. Hypotheses: $H_0 : p = 0.8$; $H_A : p < 0.8$. Suppose we wanted to use simulation-based methods. Which of the following is the correct set up for this hypothesis test? Red: success, blue: failure, $\hat{p}_{sim}$ = proportion of reds in simulated samples.



(a) Place 60 red and 40 blue chips in a bag. Sample, with replacement, 30 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.8$.

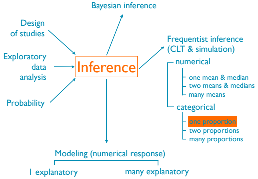
(b) Place 80 red and 20 blue chips in a bag. Sample, without replacement, 30 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.6$.

(c) Place 80 red and 20 blue chips in a bag. Sample, with replacement, 30 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.6$.

(d) Place 80 red and 20 blue chips in a bag. Sample, with replacement, 100 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.6$.

Clicker question

$n = 30$ and $\hat{p} = 0.6$. Hypotheses: $H_0 : p = 0.8$; $H_A : p < 0.8$. Suppose we wanted to use simulation-based methods. Which of the following is the correct set up for this hypothesis test? Red: success, blue: failure, $\hat{p}_{sim}$ = proportion of reds in simulated samples.



(a) Place 60 red and 40 blue chips in a bag. Sample, with replacement, 30 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.8$.

(b) Place 80 red and 20 blue chips in a bag. Sample, without replacement, 30 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.6$.

(c) ***Place 80 red and 20 blue chips in a bag. Sample, with replacement, 30 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where*** $\hat{p}_{sim} \leq 0.6$.

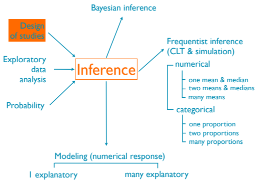
(d) Place 80 red and 20 blue chips in a bag. Sample, with replacement, 100 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.6$.

- Population
- Parameter
- Statistic
- Simple Random Sample
- Stratified Sample
- Cluster Sample
- Multistage Sample
- Experiment
- Observational Study
- Control
- Placebo
- Confounding Variable

ideal experiment	Random assignment	No random assignment	most observational studies
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
most experiments	Causation	Correlation	bad observational studies

Clicker question

A recent research study randomly divided participants into groups who were told that they were given different levels of Vitamin E to take daily. Actually, one group received only a placebo pill, and the other received Vitamin E. The research study followed the participants for eight years to see how many developed a particular type of cancer during that time period. Which of the following responses gives the best explanation as to the purpose of the random assignment in this study?



(a) To prevent skewness in the results.

(b) To reduce the amount of sampling variability.

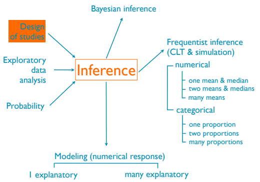
(c) To ensure that all potential cancer patients had an equal chance of being selected for the study.

(d) To produce treatment groups with similar characteristics.

(e) To ensure that the sample is representative of all cancer patients.

Clicker question

A recent research study randomly divided participants into groups who were told that they were given different levels of Vitamin E to take daily. Actually, one group received only a placebo pill, and the other received Vitamin E. The research study followed the participants for eight years to see how many developed a particular type of cancer during that time period. Which of the following responses gives the best explanation as to the purpose of the random assignment in this study?



(a) To prevent skewness in the results.

(b) To reduce the amount of sampling variability.

(c) To ensure that all potential cancer patients had an equal chance of being selected for the study.

(d) ***To produce treatment groups with similar characteristics.***

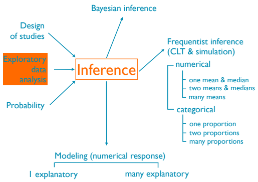
(e) To ensure that the sample is representative of all cancer patients.

Describing Distributions of Numerical Variables:

- *Shape*: skewness, modality
- *Center*: an estimate of a *typical* observation in the distribution (mean, median, mode, etc.)
  - Notation: μ: population mean, $\bar{x}$: sample mean
- *Spread*: measure of variability in the distribution (standard deviation, IQR, range, etc.)
- *Unusual observations*: observations that stand out from the rest of the data that may be suspected outliers

Robust statistics:

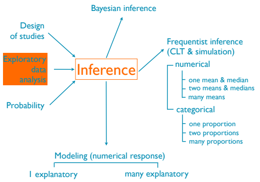
- Mean and standard deviation are easily affected by extreme observations since the value of each data point contributes to their calculation.
- Median and IQR are more robust.
- Therefore we choose median & IQR (over mean & SD) when describing skewed distributions.



Clicker question

Which of the following is false?

- (a) Box plots are useful for highlighting outliers, but we cannot determine skew based on a box plot.
- (b) Median and IQR are more robust statistics than mean and SD, respectively, since they are not affected by outliers or extreme skewness.
- (c) When the response variable is extremely right skewed, it may be useful to apply a log transformation to obtain a more symmetric distribution, and model the logged data.
- (d) Segmented frequency bar plots are “good enough” for evaluating the relationship between two categorical variables if the sample sizes are the same for various levels of the explanatory variable.



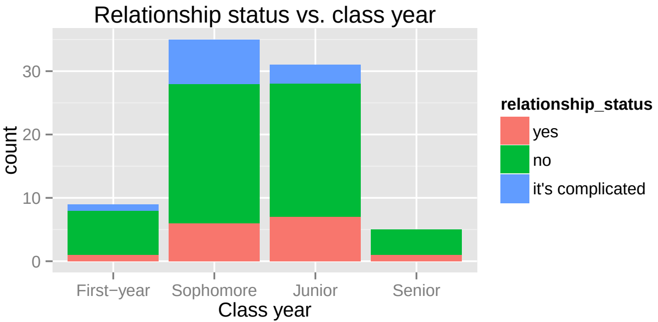
Clicker question

Which of the following is false?

- (a) ***Box plots are useful for highlighting outliers, but we cannot determine skew based on a box plot.***
- (b) Median and IQR are more robust statistics than mean and SD, respectively, since they are not affected by outliers or extreme skewness.
- (c) When the response variable is extremely right skewed, it may be useful to apply a log transformation to obtain a more symmetric distribution, and model the logged data.
- (d) Segmented frequency bar plots are "good enough" for evaluating the relationship between two categorical variables if the sample sizes are the same for various levels of the explanatory variable.

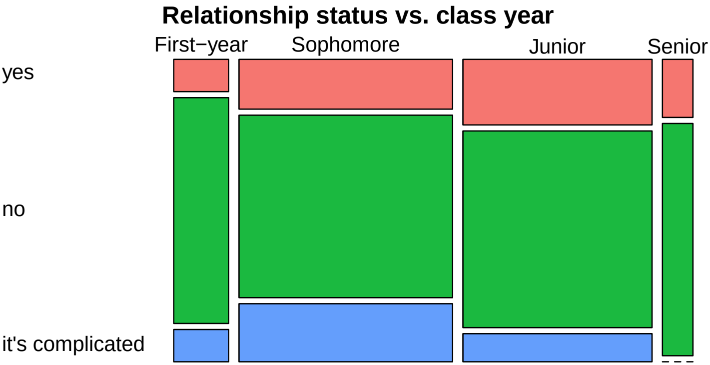
Use segmented bar plots for visualizing relationships between 2 categorical variables

What do the heights of the segments represent? Is there a relationship between class year and relationship status? What descriptive statistics can we use to summarize these data? Do the widths of the bars represent anything?



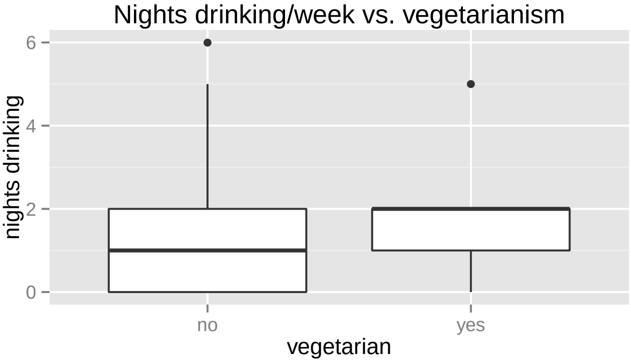
...or use a mosaic plot

What do the widths of the bars represent? What about the heights of the boxes? Is there a relationship between class year and relationship status? What other tools could we use to summarize these data?



Use side-by-side box plots to visualize relationships between a numerical and categorical variable

How do drinking habits of vegetarian vs. non-vegetarian students compare?



Key Ideas:

- Observed differences may be due to random chance
- Test whether difference is significant using simulations

- *Disjoint (mutually exclusive) events* cannot happen at the same time
  - For disjoint A and B: $P(A \ and \ B) = 0$

- *Disjoint (mutually exclusive) events* cannot happen at the same time
  - For disjoint A and B: $P(A \; and \; B) = 0$
- If A and B are *independent events*, having information on A does not tell us anything about B (and vice versa)
  - If A and B are independent:
    - $P(A \mid B) = P(A)$
    - $P(A \; and \; B) = P(A) \times P(B)$

- *Disjoint (mutually exclusive) events* cannot happen at the same time
  - For disjoint A and B: $P(A\ and\ B) = 0$
- If A and B are *independent events*, having information on A does not tell us anything about B (and vice versa)
  - If A and B are independent:
    - $P(A \mid B) = P(A)$
    - $P(A\ and\ B) = P(A) \times P(B)$
- *General addition rule:* P(A or B) = P(A) + P(B) - P(A and B)

- *Disjoint (mutually exclusive) events* cannot happen at the same time
  - For disjoint A and B: $P(A \ and \ B) = 0$
- If A and B are *independent events*, having information on A does not tell us anything about B (and vice versa)
  - If A and B are independent:
    - $P(A \mid B) = P(A)$
    - $P(A \ and \ B) = P(A) \times P(B)$
- *General addition rule:* P(A or B) = P(A) + P(B) - P(A and B)
- *Bayes' theorem:* $P(A \mid B) = \frac{P(A \ and \ B)}{P(B)}$

- Probability trees are useful for organizing information in conditional probability calculations
- They’re especially useful in cases where you know $P(A \mid B)$, along with some other information, and you’re asked for $P(B \mid A)$

- Probability trees are useful for organizing information in conditional probability calculations
- They're especially useful in cases where you know P(A | B), along with some other information, and you're asked for P(B | A)

- Using Bayes' theorem

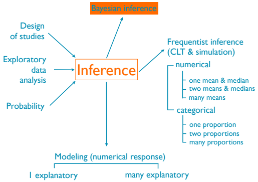
$$P(hypothesis \mid data) = \frac{P(hypothesis \text{ } and \text{ } data)}{P(data)}$$

- Probability trees are useful for organizing information in conditional probability calculations
- They're especially useful in cases where you know P(A | B), along with some other information, and you're asked for P(B | A)

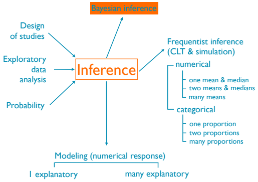
- Using Bayes' theorem

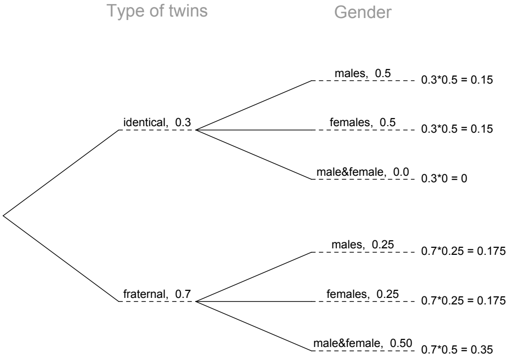
$$
\begin{aligned}
P(hypothesis \mid data) &= \frac{P(hypothesis\ and\ data)}{P(data)} \\
&= \frac{P(data \mid hypothesis) \times P(hypothesis)}{P(data)}
\end{aligned}
$$

About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the posterior probability that they are identical?

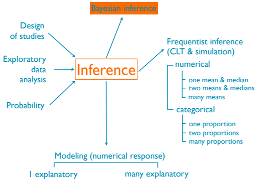


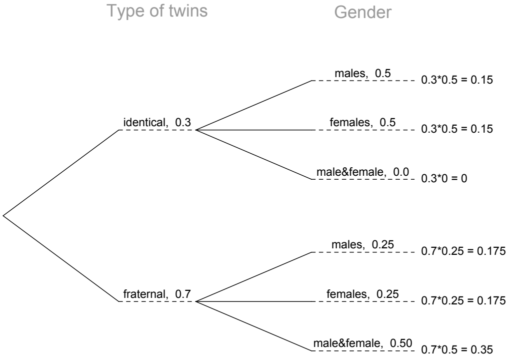
About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the posterior probability that they are identical?





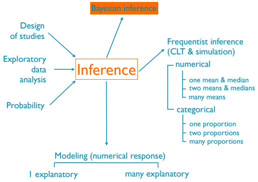
About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the posterior probability that they are identical?

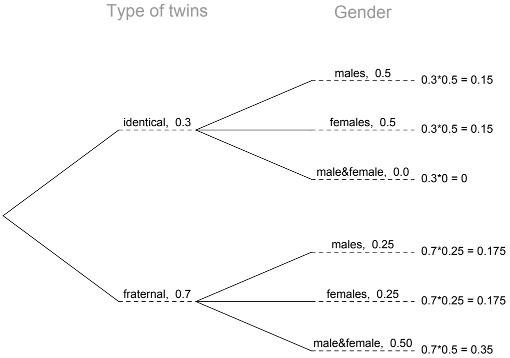




$$P(iden \mid f) = \frac{P(iden \,\&\, f)}{P(f)}$$

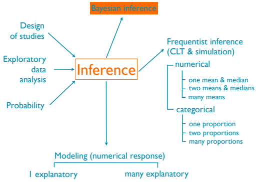
About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the posterior probability that they are identical?

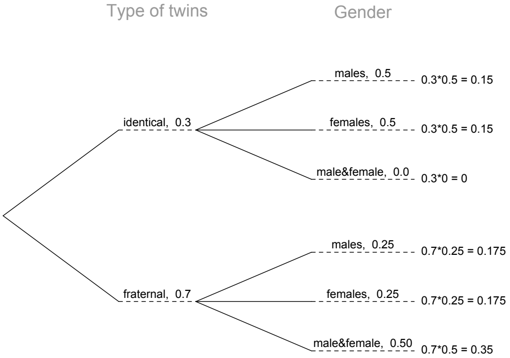




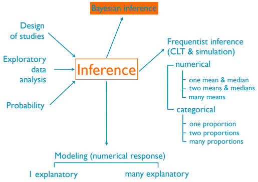
$$P(iden \mid f) = \frac{P(iden \,\&\, f)}{P(f)} = \frac{0.15}{0.15 + 0.175}$$

About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the posterior probability that they are identical?





$$
\begin{aligned}
P(iden \mid f) &= \frac{P(iden \,\&\, f)}{P(f)} \\
&= \frac{0.15}{0.15 + 0.175} \\
&= 0.46
\end{aligned}
$$



Clicker question

Which of the following is false?

(a) Suppose you're evaluating 4 claims. If prior to data collection you don't have a preference for one claim over another, you should assign 0.25 as the prior probability to each claim.

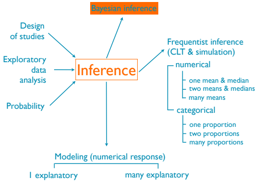
(b) Posterior probability and the p-value are the equivalent.

(c) One advantage of Bayesian inference is that data can be integrated to the inferential scheme as they are collected.

(d) Suppose a patient tests positive for a disease that 2% of the population are known to have. A doctor wants to confirm the test result by retesting the patient. In the second test the prior probability for "having the disease" should be more than 2%.

Clicker question

Which of the following is false?



(a) Suppose you're evaluating 4 claims. If prior to data collection you don't have a preference for one claim over another, you should assign 0.25 as the prior probability to each claim.

(b) ***Posterior probability and the p-value are the equivalent.***

(c) One advantage of Bayesian inference is that data can be integrated to the inferential scheme as they are collected.

(d) Suppose a patient tests positive for a disease that 2% of the population are known to have. A doctor wants to confirm the test result by retesting the patient. In the second test the prior probability for "having the disease" should be more than 2%.

Posterior = P(hypothesis | data), p-value ≈ P(data | hypothesis)

- Two types of probability distributions: discrete and continuous

- Two types of probability distributions: discrete and continuous
- Normal distribution is unimodal, symmetric and follows the 68-95-99.7 rule

- Two types of probability distributions: discrete and continuous
- Normal distribution is unimodal, symmetric and follows the 68-95-99.7 rule
- Z scores serve as a ruler for any distribution

$$Z = \frac{obs - mean}{SD}$$

- Two types of probability distributions: discrete and continuous
- Normal distribution is unimodal, symmetric and follows the 68-95-99.7 rule
- Z scores serve as a ruler for any distribution

$$Z = \frac{obs - mean}{SD}$$

- Z score: number of standard deviations the observation falls above or below the mean

- The *Binomial distribution* describes the probability of having exactly k successes in n independent trials with probability of success p.

$$P(k\ successes\ in\ n\ trials) = \binom{n}{k}\, p^k\, (1-p)^{(n-k)}$$

Note: P (at least one event)= 1 - P(none)

- The *Binomial distribution* describes the probability of having exactly k successes in n independent trials with probability of success p.

$$P(k \; successes \; in \; n \; trials) = \binom{n}{k} \, p^k \, (1-p)^{(n-k)}$$

Note: P (at least one event)= 1 - P(none)

- *Expected Value:* np
- *Standard Deviation:* $\sqrt{np(1-p)}$

- The *Binomial distribution* describes the probability of having exactly k successes in n independent trials with probability of success p.

$$P(k\ successes\ in\ n\ trials) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

Note: P (at least one event)= 1 - P(none)

- *Expected Value:* np
- *Standard Deviation:* $\sqrt{np(1-p)}$
- Shape of the binomial distribution approaches normal when the S-F rule is met

- Sample Statistics vary from sample to sample

- Sample Statistics vary from sample to sample
- CLT describes the shape, center and spread of sampling distributions

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

- Sample Statistics vary from sample to sample
- CLT describes the shape, center and spread of sampling distributions

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

- CLT only applies when independence and sample size/skew conditions are met

- Statistical inference methods based on the CLT require the same conditions as the CLT

- *CI: point estimate ± margin of error*

- Calculate the sample size a priori to achieve desired margin or error

- Statistical inference methods based on the CLT require the same conditions as the CLT

- *CI: point estimate $\pm$ margin of error*

- Calculate the sample size a priori to achieve desired margin or error

  Solve for n:

$$ME = z^* \frac{s}{\sqrt{n}}$$

Hypothesis testing framework:

1. Set the hypotheses.

2. Check assumptions and conditions.

3. Calculate a *test statistic* and a p-value.

4. Make a decision, and interpret it in context of the research question.

$$HT: test\ statistic = \frac{point\ estimate - null}{SE}$$

$$CI: point\ estimate \pm critical\ value \times SE$$

One mean:

$df = n - 1$

HT:

$H_0: \mu = \mu_0$

$T_{df} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

CI:

$\bar{x} \pm t^\star_{df} \frac{s}{\sqrt{n}}$

Paired means:

$df = n_{diff} - 1$

HT:

$H_0: \mu_{diff} = 0$

$T_{df} = \frac{\bar{x}_{diff} - 0}{\frac{s_{diff}}{\sqrt{n_{diff}}}}$

CI:

$\bar{x}_{diff} \pm t^\star_{df} \frac{s_{diff}}{\sqrt{n_{diff}}}$

Independent means:

$df = min(n_1 - 1, n_2 - 1)$

HT:

$H_0: \mu_1 - \mu_2 = 0$

$T_{df} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

CI:

$\bar{x}_1 - \bar{x}_2 \pm t^\star_{df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

- Bootstrapping works as follows:
  1. take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample
  2. calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples
  3. repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics

- Bootstrapping works as follows:
  1. take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample
  2. calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples
  3. repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics
- The XX% bootstrap confidence interval can be estimated by
  - the cutoff values for the middle XX% of the bootstrap distribution,

    OR

  - $point\ estimate \pm t^{\star} SE_{boot}$

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1 - \alpha$	*Type 1 Error, α*
	H_A true	*Type 2 Error, β*	*Power, $1 - \beta$*

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1 - \alpha$	*Type 1 Error, α*
	H_A true	*Type 2 Error, β*	*Power, $1 - \beta$*

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1 - \alpha$	*Type 1 Error, α*
	H_A true	*Type 2 Error, β*	*Power, $1 - \beta$*

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1 - \alpha$	*Type 1 Error, α*
	H_A true	*Type 2 Error, β*	*Power,* $1 - \beta$

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- *Power* of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1 - \beta$

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1 - \alpha$	*Type 1 Error,* α
	H_A true	*Type 2 Error,* β	*Power,* $1 - \beta$

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- *Power* of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1 - \beta$
- In hypothesis testing, we want to keep α and β low, but there are inherent trade-offs.

- *Null Hypothesis:* $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
- *Alternative Hypothesis:* At least on pair of means is different from one another
- *F-statistic:* $F = MSG/MSE$

- *Null Hypothesis:* $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
- *Alternative Hypothesis:* At least on pair of means is different from one another
- *F-statistic:* $F = MSG/MSE$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Between groups	$k-1$	SSG	MSG	F_{obs}	p_{obs}
Within groups	$n-k$	SSE	MSE		
Total	$n-1$	SSG+SSE			

Note: F distribution is defined by two dfs: $df_G = k-1$ and $df_E = n-k$

What significant p-value means here?

To identify which means are different, use t-tests and the Bonferroni correction

- If the ANOVA yields a significant results, next natural question is: "Which means are different?"
- Use t-tests comparing each pair of means to each other,
  - with a common variance (MSE from the ANOVA table) instead of each group's variances in the calculation of the standard error,
  - and with a common degrees of freedom (df_E from the ANOVA table)
- Compare resulting p-values to a modified significance level

$$\alpha^\star = \frac{\alpha}{K}$$

where $K = \frac{k(k-1)}{2}$ is the total number of pairwise tests

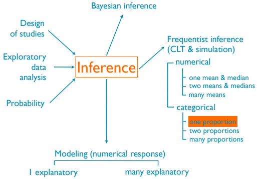
- Question: What is α^*, when df_G is given?

HT vs. CI for a proportion

- Success-failure condition:
  - CI: At least 10 *observed* successes and failures
  - HT: At least 10 *expected* successes and failures, calculated using the null value
- Standard error:
  - CI: calculate using observed sample proportion: $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - HT: calculate using the null value: $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$
- If the S-F condition is not met use *Randomization Test*

Clicker question

$n = 30$ and $\hat{p} = 0.6$. Hypotheses: $H_0 : p = 0.8$; $H_A : p < 0.8$. Suppose we wanted to use simulation-based methods. Which of the following is the correct set up for this hypothesis test? Red: success, blue: failure, $\hat{p}_{sim}$ = proportion of reds in simulated samples.



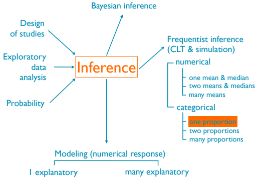
(a) Place 60 red and 40 blue chips in a bag. Sample, with replacement, 30 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.8$.

(b) Place 80 red and 20 blue chips in a bag. Sample, without replacement, 30 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.6$.

(c) Place 80 red and 20 blue chips in a bag. Sample, with replacement, 30 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.6$.

(d) Place 80 red and 20 blue chips in a bag. Sample, with replacement, 100 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.6$.

Clicker question



$n = 30$ and $\hat{p} = 0.6$. Hypotheses: $H_0 : p = 0.8$; $H_A : p < 0.8$. Suppose we wanted to use simulation-based methods. Which of the following is the correct set up for this hypothesis test? Red: success, blue: failure, $\hat{p}_{sim}$ = proportion of reds in simulated samples.

(a) Place 60 red and 40 blue chips in a bag. Sample, with replacement, 30 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.8$.

(b) Place 80 red and 20 blue chips in a bag. Sample, without replacement, 30 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.6$.

(c) ***Place 80 red and 20 blue chips in a bag. Sample, with replacement, 30 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where*** $\hat{p}_{sim} \leq 0.6$.

(d) Place 80 red and 20 blue chips in a bag. Sample, with replacement, 100 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.6$.

For HT where $H_0 : p_1 = p_2$, pool!

As with working with a single proportion,

- When doing a HT where $H_0 : p_1 = p_2$ (almost always for HT), use expected counts / proportions for S-F condition and calculation of the standard error.
- Otherwise use observed counts / proportions for S-F condition and calculation of the standard error.

For HT where $H_0 : p_1 = p_2$, pool!

As with working with a single proportion,

- When doing a HT where $H_0 : p_1 = p_2$ (almost always for HT), use expected counts / proportions for S-F condition and calculation of the standard error.
- Otherwise use observed counts / proportions for S-F condition and calculation of the standard error.

Expected proportion of success for both groups when $H_0 : p_1 = p_2$ is defined as the *pooled proportion*:

$$\hat{p}_{pool} = \frac{total\ successes}{total\ sample\ size} = \frac{suc_1 + suc_2}{n_1 + n_2}$$

Type	Parameter	Estimator	SE	Sampling Dist.
One mean	μ	$\bar{x}$	$s/\sqrt{n}$	t_{n-1}
Two means Paired data	μ_{diff}	$\bar{x}_{diff}$	$s_d/\sqrt{n}$	t_{n-1}
Two means Independent	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	t_{df} for df use $\min\{n_1 - 1, n_2 - 1\}$
One prop	p	$\hat{p}$	**C.I.** $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ **H.T.** $\sqrt{\frac{p_0(1-p_0)}{n}}$	Z
Two prop	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	**C.I.** $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ **H.T.** $\sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_2}}$	Z

Categorical data with more than 2 levels $\rightarrow \chi^2$

- one variable: χ^2 *test of goodness of fit*, no CI
- two variables: χ^2 *test of independence*, no CI

Categorical data with more than 2 levels $\rightarrow \chi^2$

- one variable: χ^2 *test of goodness of fit*, no CI
- two variables: χ^2 *test of independence*, no CI

- χ^2 *statistic:* When dealing with counts and investigating how far the observed counts are from the expected counts, we use a new test statistic called the *chi-square (χ^2) statistic*:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O - E)^2}{E}$$

where k = total number of cells

Important points:

- Use **counts** (not **proportions**) in the calculation of the text statistic, even though we're truly interested in the proportions for inference
- Expected counts are calculated assuming the null hypothesis is true

Important points:

- Use **counts** (not **proportions**) in the calculation of the text statistic, even though we're truly interested in the proportions for inference
- Expected counts are calculated assuming the null hypothesis is true

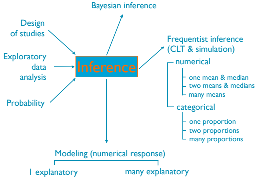
The χ^2 distribution has just one parameter, *degrees of freedom (df)*, which influences the shape, center, and spread of the distribution.

- For χ^2 GOF test: $df = k - 1$
- For χ^2 independence test: $df = (R - 1) \times (C - 1)$

What is the shape of the χ^2 distribution?

Clicker question

Which of the following is the best method for evaluating the if the distribution of a categorical variable follows a hypothesized distribution?



(a) chi-square test of independence

(b) chi-square test of goodness of fit

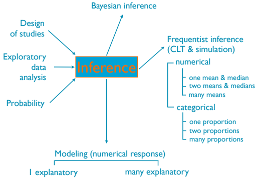
(c) anova

(d) linear regression

(e) t-test

Clicker question

Which of the following is the best method for evaluating the if the distribution of a categorical variable follows a hypothesized distribution?



(a) chi-square test of independence

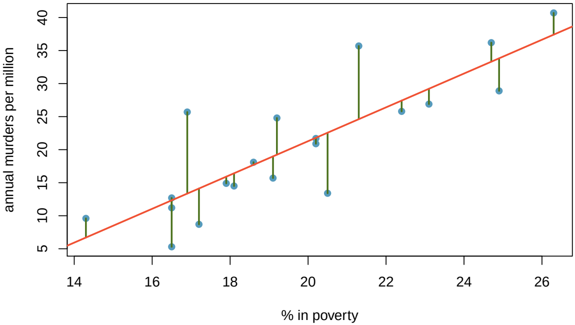
(b) ***chi-square test of goodness of fit***

(c) anova

(d) linear regression

(e) t-test

- Residuals are the leftovers from the model fit, and calculated as the difference between the observed and predicted y:
  $e_i = y_i - \hat{y}_i$
- The least squares line minimizes *squared* residuals:
  - Population data: $\hat{y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x$
  - Sample data: $\hat{y} = b_0 + b_1 x$



- *Slope:* For each unit increase in x, y is expected to be higher/lower on average by the slope.

$$b_1 = \frac{s_y}{s_x} R$$

- *Intercept:* When $x = 0$, y is expected to equal the intercept.

$$b_0 = \bar{y} - b_1 \bar{x}$$

- *Correlation Coefficient:* R measures the strength and direction of the linear association between the two numerical variables

- R^2: percentage of variability in y explained by the model.
- For single predictor regression: R^2 is the square of the correlation coefficient, R.
- For all regression: $R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{error}}{SS_{tot}}$

- Hypothesis testing for a slope: $H_0 : \boldsymbol{\beta}_1 = 0$; $H_A : \boldsymbol{\beta}_1 \neq 0$
  - $T_{n-2} = \frac{b_1 - 0}{SE_{b_1}}$
  - p-value = P(observing a slope at least as different from 0 as the one observed if in fact there is no relationship between x and y
  - Degrees of freedom for the slope(s) in regression is $df = n - k - 1$ where k is the number of slopes being estimated in the model.

- Hypothesis testing for a slope: $H_0 : \boldsymbol{\beta}_1 = 0$; $H_A : \boldsymbol{\beta}_1 \neq 0$
  - $T_{n-2} = \frac{b_1 - 0}{SE_{b_1}}$
  - p-value = P(observing a slope at least as different from 0 as the one observed if in fact there is no relationship between x and y
  - Degrees of freedom for the slope(s) in regression is $df = n - k - 1$ where k is the number of slopes being estimated in the model.
- Confidence intervals for a slope:
  - $b_1 \pm T^{\star}_{n-2} SE_{b_1}$

Important regardless of doing inference

- Linearity $\longrightarrow$ randomly scattered residuals around 0 in the residuals plot – important regardless of doing inference

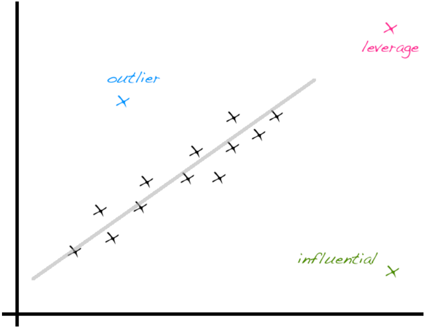
Important regardless of doing inference

- Linearity $\longrightarrow$ randomly scattered residuals around 0 in the residuals plot – important regardless of doing inference

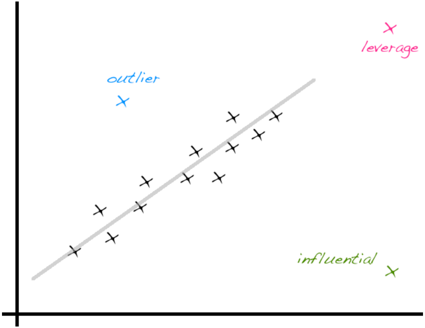
Important for inference

- Nearly normally distributed residuals $\longrightarrow$ histogram or normal probability plot of residuals
- Constant variability of residuals (*homoscedasticity*) $\longrightarrow$ no fan shape in the residuals plot
- Independence of residuals (and hence observations) $\longrightarrow$ depends on data collection method, often violated for time-series data

- *Leverage* point is away from the cloud of points horizontally, does not necessarily change the slope
- *Influential* point changes the slope (most likely also has high leverage) – run the regression with and without that point to determine



- *Leverage* point is away from the cloud of points horizontally, does not necessarily change the slope
- *Influential* point changes the slope (most likely also has high leverage) – run the regression with and without that point to determine

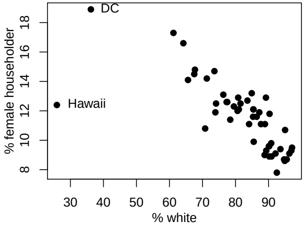


- *Outlier* is an unusual point without these special characteristics (this one likely affects the intercept only)
- If clusters (groups of points) are apparent in the data, it might be worthwhile to model the groups separately.

Clicker question

The scatterplot on the right shows the relationship between percentage of white residents and percentage of households with a female head (where no husband is present) in all 50 US States and the District of Columbia (DC). Which of the below **best** describes the two points marked as DC and Hawaii?

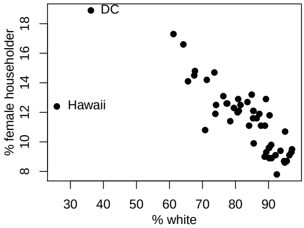
1. Hawaii has higher leverage and is more influential than DC.
2. DC is not an outlier, Hawaii is a leverage point.
3. DC is more influential than Hawaii, but has lower leverage than Hawaii.
4. DC has higher leverage and is more influential than Hawaii.



Clicker question

The scatterplot on the right shows the relationship between percentage of white residents and percentage of households with a female head (where no husband is present) in all 50 US States and the District of Columbia (DC). Which of the below **best** describes the two points marked as DC and Hawaii?

1. ***Hawaii has higher leverage and is more influential than DC.***
2. DC is not an outlier, Hawaii is a leverage point.
3. DC is more influential than Hawaii, but has lower leverage than Hawaii.
4. DC has higher leverage and is more influential than Hawaii.



- Influential points are a subset of outliers since they must be far away from the ‘cloud’.
- High leverage points are a subset of outliers since they are far away from the ‘cloud’ (in the horizontal direction).
- Outlier (Not leverage/influential): An outlier without these above special characteristics (this one likely affects the intercept only). This is a vertical outlier.
- Not all outliers are influential or have high leverage.
- High leverage does not imply influential. Influential does not imply high leverage.

For more details refer to the last two slides!

- All estimates in a MLR for a given variable are conditional on all other variables being in the model.

- *Slope:*
  - Numerical x: *All else held constant*, for one unit increase in x_i, y is expected to be higher / lower on average by b_i units.
  - Categorical x: *All else held constant*, the predicted difference in y for the baseline and given levels of x_i is b_i.

- All estimates in a MLR for a given variable are conditional on all other variables being in the model.
- *Slope:*
  - Numerical x: *All else held constant*, for one unit increase in x_i, y is expected to be higher / lower on average by b_i units.
  - Categorical x: *All else held constant*, the predicted difference in y for the baseline and given levels of x_i is b_i.
- *Categorical Predictors:*
  - Each categorical variable, with k levels, added to the model results in $k-1$ parameters being estimated.
  - It only takes $k-1$ columns to code a categorical variable with k levels as 0/1s.

- Inference for the model as a whole: F-test, $df_1 = k$, $df_2 = n - k - 1$

  H_0 : $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

  H_A : At least one of the $\beta_i \neq 0$

  What conclusion can you draw when your p-value significant or not significant?

- Inference for the model as a whole: F-test, $df_1 = k$, $df_2 = n - k - 1$

  $H_0 : \ \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_k = 0$

  $H_A :$ At least one of the $\boldsymbol{\beta}_i \neq 0$

  What conclusion can you draw when your p-value significant or not significant?

- Inference for each slope: T-test, $df = n - k - 1$
  - HT:

    $H_0 : \ \boldsymbol{\beta}_1 = 0$, when all other variables are included in the model

    $H_A : \ \boldsymbol{\beta}_1 \neq 0$, when all other variables are included in the model
  - CI: $b_1 \pm T^\star_{df} SE_{b_1}$

- When any variable is added to the model R^2 increases.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adjusted R^2 does not increase.

- When any variable is added to the model R^2 increases.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adjusted R^2 does not increase.

Adjusted R^2

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-k-1} \right)$$

where n is the number of cases and k is the number of sloped estimated in the model.

- If the goal is to find the set of statistically predictors of $y \longrightarrow$ use p-value selection

- If the goal is to find the set of statistically predictors of $y \longrightarrow$ use p-value selection
- If the goal is to do better prediction of $y \longrightarrow$ use adjusted R^2 selection

- If the goal is to find the set of statistically predictors of $y \longrightarrow$ use p-value selection
- If the goal is to do better prediction of $y \longrightarrow$ use adjusted R^2 selection
- Either way, can use *backward elimination* or *forward selection*

- If the goal is to find the set of statistically predictors of $y \longrightarrow$ use p-value selection
- If the goal is to do better prediction of $y \longrightarrow$ use adjusted R^2 selection
- Either way, can use *backward elimination* or *forward selection*
- Important to make sure that your explanatory variables are not *collinear*

- If the goal is to find the set of statistically predictors of $y \longrightarrow$ use p-value selection
- If the goal is to do better prediction of $y \longrightarrow$ use adjusted R^2 selection
- Either way, can use *backward elimination* or *forward selection*
- Important to make sure that your explanatory variables are not *collinear*
- We usually prefer simpler (*parsimonious*) models over more complicated ones

Important regardless of doing inference

- Linearity $\longrightarrow$ randomly scattered residuals around 0 in the residuals plot

Important regardless of doing inference

- Linearity $\longrightarrow$ randomly scattered residuals around 0 in the residuals plot

Important for doing inference

- Nearly normally distributed residuals $\longrightarrow$ histogram or normal probability plot of residuals
- Constant variability of residuals (*homoscedasticity*) $\longrightarrow$ no fan shape in the residuals plot
- Independence of residuals (and hence observations) $\longrightarrow$ depends on data collection method, often violated for time-series data

Important regardless of doing inference

- Linearity $\longrightarrow$ randomly scattered residuals around 0 in the residuals plot

Important for doing inference

- Nearly normally distributed residuals $\longrightarrow$ histogram or normal probability plot of residuals
- Constant variability of residuals (*homoscedasticity*) $\longrightarrow$ no fan shape in the residuals plot
- Independence of residuals (and hence observations) $\longrightarrow$ depends on data collection method, often violated for time-series data

Clicker question

Using the p-value approach, which variable would you remove from the model first?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-15342.76	11716.57	-1.31	0.19
hrs_work	1048.96	149.25	7.03	0.00
raceblack	-7998.99	6191.83	-1.29	0.20
raceasian	29909.80	9154.92	3.27	0.00
raceother	-6756.32	7240.08	-0.93	0.35
age	565.07	133.77	4.22	0.00
genderfemale	-17135.05	3705.35	-4.62	0.00
citizenyes	-12907.34	8231.66	-1.57	0.12
time_to_work	90.04	79.83	1.13	0.26
langother	-10510.44	5447.45	-1.93	0.05
marriedyes	5409.24	3900.76	1.39	0.17
educollege	15993.85	4098.99	3.90	0.00
edugrad	59658.52	5660.26	10.54	0.00
disabilityyes	-14142.79	6639.40	-2.13	0.03
birth_qrtrapr thru jun	-2043.42	4978.12	-0.41	0.68
birth_qrtrjul thru sep	3036.02	4853.19	0.63	0.53
birth_qrtroct thru dec	2674.11	5038.45	0.53	0.60

(a) race:other

(b) race

(c) time_to_work

(d) birth_qrtr:apr thru jun

(e) birth_qrtr

Clicker question

Using the p-value approach, which variable would you remove from the model first?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-15342.76	11716.57	-1.31	0.19
hrs_work	1048.96	149.25	7.03	0.00
raceblack	-7998.99	6191.83	-1.29	0.20
raceasian	29909.80	9154.92	3.27	0.00
raceother	-6756.32	7240.08	-0.93	0.35
age	565.07	133.77	4.22	0.00
genderfemale	-17135.05	3705.35	-4.62	0.00
citizenyes	-12907.34	8231.66	-1.57	0.12
time_to_work	90.04	79.83	1.13	0.26
langother	-10510.44	5447.45	-1.93	0.05
marriedyes	5409.24	3900.76	1.39	0.17
educollege	15993.85	4098.99	3.90	0.00
edugrad	59658.52	5660.26	10.54	0.00
disabilityyes	-14142.79	6639.40	-2.13	0.03
birth_qrtrapr thru jun	-2043.42	4978.12	-0.41	0.68
birth_qrtrjul thru sep	3036.02	4853.19	0.63	0.53
birth_qrtroct thru dec	2674.11	5038.45	0.53	0.60

(a) race:other

(b) race

(c) time_to_work

(d) birth_qrtr:apr thru jun

(e) ***birth_qrtr***

Clicker question

Using the p-value approach, which variable would you remove from the model next?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-14022.48	11137.08	-1.26	0.21
hrs_work	1045.85	149.05	7.02	0.00
raceblack	-7636.32	6177.50	-1.24	0.22
raceasian	29944.35	9137.13	3.28	0.00
raceother	-7212.57	7212.25	-1.00	0.32
age	559.51	133.27	4.20	0.00
genderfemale	-17010.85	3699.19	-4.60	0.00
citizenyes	-13059.46	8219.99	-1.59	0.11
time_to_work	88.77	79.73	1.11	0.27
langother	-10150.41	5431.15	-1.87	0.06
marriedyes	5400.41	3896.12	1.39	0.17
educollege	16214.46	4089.17	3.97	0.00
edugrad	59572.20	5631.33	10.58	0.00
disabilityyes	-14201.11	6628.26	-2.14	0.03

(a) married

(b) race

(c) race:other

(d) race:black

(e) time_to_work

Clicker question

Using the p-value approach, which variable would you remove from the model next?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-14022.48	11137.08	-1.26	0.21
hrs_work	1045.85	149.05	7.02	0.00
raceblack	-7636.32	6177.50	-1.24	0.22
raceasian	29944.35	9137.13	3.28	0.00
raceother	-7212.57	7212.25	-1.00	0.32
age	559.51	133.27	4.20	0.00
genderfemale	-17010.85	3699.19	-4.60	0.00
citizenyes	-13059.46	8219.99	-1.59	0.11
time_to_work	88.77	79.73	1.11	0.27
langother	-10150.41	5431.15	-1.87	0.06
marriedyes	5400.41	3896.12	1.39	0.17
educollege	16214.46	4089.17	3.97	0.00
edugrad	59572.20	5631.33	10.58	0.00
disabilityyes	-14201.11	6628.26	-2.14	0.03

(a) married

(b) race

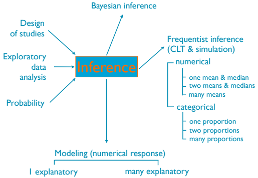
(c) race:other

(d) race:black

(e) ***time_to_work***

Clicker question

Which of the following is the best method for evaluating the relationship between a numerical and a categorical variable with many levels?



(a) z-test

(b) chi-square test of goodness of fit

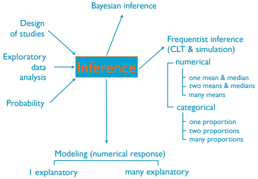
(c) anova

(d) linear regression

(e) t-test

Clicker question

Which of the following is the best method for evaluating the relationship between a numerical and a categorical variable with many levels?



- (a) z-test
- (b) chi-square test of goodness of fit
- (c) ***anova***
- (d) ***linear regression***
- (e) t-test

Example - Breast Cancer & Age

It is theorized that an important risk factor for breast cancer is age at first birth. An international study was set up to test this hypothesis. Breast-cancer cases were identified among women in selected hospitals in the United States, Greece, Yugoslavia, Brazil, Taiwan, and Japan. Controls were chosen from women of comparable age who were in the hospital at the same time as the cases but who did not have breast cancer. All women were asked about their age at first birth.

The set of women with at least one birth was arbitrarily divided into two categories: (1) women whose age at first birth was less than or equal to 29 years and (2) women whose age at first birth was greater than of equal to 30 years. The following results were found among women with at least one birth: 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than or equal to 30. How can we assess whether this difference is significant or simply due to chance?

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

	Breast Cancer (case)	No Breast Cancer (Controls)	Total
≤ 29			
≥ 30			
Total			

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than 30.

	Breast Cancer (case)	No Breast Cancer (Controls)	Total
≤ 29			
≥ 30			
Total			

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than 30.

	Breast Cancer (case)	No Breast Cancer (Controls)	Total
≤ 29			
≥ 30	683		
Total	3220		

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than 30.

	Breast Cancer (case)	No Breast Cancer (Controls)	Total
≤ 29	2537		
≥ 30	683		
Total	3220		

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than 30.

	Breast Cancer (case)	No Breast Cancer (Controls)	Total
≤ 29	2537		
≥ 30	683	1498	
Total	3220	10245	

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than 30.

	Breast Cancer (case)	No Breast Cancer (Controls)	Total
≤ 29	2537	8747	
≥ 30	683	1498	
Total	3220	10245	

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than 30.

	Breast Cancer (case)	No Breast Cancer (Controls)	Total
≤ 29	2537	8747	11284
≥ 30	683	1498	2181
Total	3220	10245	13465

$$n_{case} = 3220,\ n_{ctrl} = 10245$$

- cases:

$$n_{case} = 3220,\ n_{ctrl} = 10245$$

- cases: 13465 women (hospital patients) with at least one child
- variable(s):

$$n_{case} = 3220,\ n_{ctrl} = 10245$$

- cases: 13465 women (hospital patients) with at least one child
- variable(s): (1) breast cancer status - categorical, (2) age at first birth - categorical
- parameter of interest:

$$n_{case} = 3220, \; n_{ctrl} = 10245$$

- cases: 13465 women (hospital patients) with at least one child
- variable(s): (1) breast cancer status - categorical, (2) age at first birth - categorical
- parameter of interest: $p_{case} - p_{ctrl}$
  - Note: $p_{case} = P(age \geq 30|case)$ and $p_{ctrl} = P(age \geq 30|ctrl)$
- test:

$$n_{case} = 3220,\ n_{ctrl} = 10245$$

- cases: 13465 women (hospital patients) with at least one child
- variable(s): (1) breast cancer status - categorical, (2) age at first birth - categorical
- parameter of interest: $p_{case} - p_{ctrl}$
  - Note: $p_{case} = P(age \geq 30|case)$ and $p_{ctrl} = P(age \geq 30|ctrl)$
- test: compare two population proportion of independent groups
- hypotheses:

$$n_{case} = 3220, \; n_{ctrl} = 10245$$

- cases: 13465 women (hospital patients) with at least one child
- variable(s): (1) breast cancer status - categorical, (2) age at first birth - categorical
- parameter of interest: $p_{case} - p_{ctrl}$
  - Note: $p_{case} = P(age \geq 30|case)$ and $p_{ctrl} = P(age \geq 30|ctrl)$
- test: compare two population proportion of independent groups
- hypotheses: (two-tailed)

$$H_0 : p_{case} = p_{ctrl}$$
$$H_A : p_{case} \neq p_{ctrl}$$

Clicker question

Which of the following is the correct point estimate for this HT?

	BC (Case)	No BC (Controls)	Total
≤ 29	2537	8747	11284
≥ 30	683	1498	2181
Total	3220	10245	13465

(a) $\frac{683}{2181} - \frac{1498}{2181}$

(b) $\frac{683}{13465} - \frac{1498}{13465}$

(c) $\frac{2537}{11284} - \frac{683}{2181}$

(d) $\frac{683}{3220} - \frac{1498}{10245}$

(e) $\frac{683}{2181} - \frac{683}{3220}$

Clicker question

Which of the following is the correct point estimate for this HT?

	BC (Case)	No BC (Controls)	Total
≤ 29	2537	8747	11284
≥ 30	683	1498	2181
Total	3220	10245	13465

(a) $\frac{683}{2181} - \frac{1498}{2181}$

(b) $\frac{683}{13465} - \frac{1498}{13465}$

(c) $\frac{2537}{11284} - \frac{683}{2181}$

(d) $\frac{683}{3220} - \frac{1498}{10245}$ = *0.066*

(e) $\frac{683}{2181} - \frac{683}{3220}$

Clicker question

Which of the following is the correct standard error for this HT?

	BC (Case)	No BC (Controls)	Total
≤ 29	2537	8747	11284
≥ 30	683	1498	2181
Total	3220	10245	13465
$\hat{p}$	0.212	0.146	0.162

(a) $\sqrt{\frac{0.212\times(1-0.212)}{3220}} + \sqrt{\frac{0.146\times(1-0.146)}{10245}}$

(b) $\sqrt{\frac{0.212\times(1-0.212)}{3220} + \frac{0.146\times(1-0.146)}{10245}}$

(c) $\sqrt{\frac{0.162\times(1-0.162)}{3220} + \frac{0.162\times(1-0.162)}{10245}}$

(d) $\sqrt{\frac{0.212\times(1-0.212)}{13465} + \frac{0.146\times(1-0.146)}{13465}}$

(e) $\sqrt{\frac{0.162\times(1-0.162)}{13465} + \frac{0.162\times(1-0.162)}{13465}}$

Clicker question

Which of the following is the correct standard error for this HT?

	BC (Case)	No BC (Controls)	Total
≤ 29	2537	8747	11284
≥ 30	683	1498	2181
Total	3220	10245	13465
$\hat{p}$	0.212	0.146	0.162

(a) $\sqrt{\frac{0.212\times(1-0.212)}{3220}} + \sqrt{\frac{0.146\times(1-0.146)}{10245}}$

(b) $\sqrt{\frac{0.212\times(1-0.212)}{3220} + \frac{0.146\times(1-0.146)}{10245}}$

(c) $\sqrt{\frac{0.162\times(1-0.162)}{3220} + \frac{0.162\times(1-0.162)}{10245}}$ = *0.0074*

(d) $\sqrt{\frac{0.212\times(1-0.212)}{13465} + \frac{0.146\times(1-0.146)}{13465}}$

(e) $\sqrt{\frac{0.162\times(1-0.162)}{13465} + \frac{0.162\times(1-0.162)}{13465}}$

$$Z = \frac{\hat{p}_{case} - \hat{p}_{ctrl} - 0}{SE} = \frac{0.212 - 0.146}{0.0074} = 8.92$$

$$Z = \frac{\hat{p}_{case} - \hat{p}_{ctrl} - 0}{SE} = \frac{0.212 - 0.146}{0.0074} = 8.92$$

p-value $= P(Z > 8.92) + P(Z < -8.92) \approx 0$

The following tries to extract the info on outliers in Section 7.3 of the textbook. Quotations from the book are given in quotation marks.

- Definition of outliers including both ‘vertical’ as well as ‘horizontal’ outliers.
- “Outliers in regression are observations that fall far from the “cloud” of points”.
- High leverage points are those that are horizontally removed from the center of the cloud. “Points that fall horizontally away from the center of the cloud tend to pull harder on the line, so we call them points with high leverage”.

- If a leverage point influences the slope of the line, then it is influential. “If one of these high leverage points does appear to actually invoke its influence on the slope of the line (…), then we call it an influential point”.
- Also non-leverage points can be influential, they just need to effect the line of best fit which an extreme vertical outlier can do. “Usually we can say a point is influential if, had we fitted the line without it, the influential point would have been unusually far from the least squares line”.