

Final Exam Question Bank

Stat 218

For your final exam **50%** of questions will be randomly selected from **this** question bank. The remainder of the exam will be **25%** selected from the **Midterm 1** question bank and **25%** selected from the **Midterm 2** question bank.

There **will not** be a solution key posted. It is your responsibility to discuss your ideas with your group members and / or with Dr. Theobold during office hours prior to the exam.

Golden Ticket

| Scenario | One Categorical Response | Two Categorical Variables | One Quantitative Response | Two Quantitative Variables | Quant. Response and Categ. Explanatory |
|-----------------------|-------------------------------|---|--|--|--|
| Type of plot | Bar plot | Dodged Bar plot, Stacked Bar plot, Filled Bar plot | Dot plot, Histogram, Boxplot | Scatterplot | Faceted Histograms, Side-by-side Boxplots |
| Summary measure | Proportion | Deviation between Observed Counts and Expected Counts (X^2) | Mean or Mean of Differences | Slope or Correlation | Difference in Means |
| Parameter notation | π | $\pi_1, \pi_2, \dots, \pi_k$ | μ or μ_{diff} | Slope: β_1 ; Correlation: ρ | $\mu_1 - \mu_2$ |
| Statistic notation | \hat{p} | $\hat{p}_1, \hat{p}_2, \dots, \hat{p}_k$ | \bar{x} or \bar{x}_{diff} | Slope: b_1 ; Correlation: r | $\bar{x}_1 - \bar{x}_2$ |
| Statistical Method(s) | χ^2 Goodness of fit Test | χ^2 Test of Independence, χ^2 Test of Homogeneity, Permutation Test for X^2 | t -test for One Mean, t -test for Paired Differences, Bootstrap Confidence Interval for One Mean | t -test for β_1 , Permutation Test for β_1 , Bootstrap Confidence Interval for β_1 | t -test for $\mu_1 - \mu_2$, Permutation Test for $\mu_1 - \mu_2$, Bootstrap Confidence Interval for $\mu_1 - \mu_2$ |

Provided Formulas

$$IQR = Q3 - Q1$$

1.5 IQR Rule: above $Q3 + (1.5 \times IQR)$ or below $Q1 - (1.5 \times IQR)$

$$\hat{y} = b_0 + b_1 \times$$

$$\text{Residual} = y - \hat{y}$$

$$R^2 = r^2$$

general formula for a confidence interval: point estimate \pm multiplier \times SE(point estimate)

t-based confidence interval: $\bar{x} \pm t_{df}^* \times SE(\bar{x})$

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$F = \frac{MSG}{MSE}$$

$$\alpha^* = \frac{\alpha}{\# \text{ of comparisons}}$$

Expected Counts for One Categorical Variable

Expected Count = total sample size \times null proportion for group k

Expected Counts for Two Categorical Variables

$$\text{Expected Count} = \frac{(\text{row i total}) \times (\text{column j total})}{\text{total sample size}}$$

Chi-Square Test Statistic

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Final Exam Question Bank

Wild Mushrooms

Q1[17 points] Wild mushrooms, such as chanterelles or morels, are delicious, but eating wild mushrooms carries the risk of accidental poisoning. Even a single bite of the wrong mushroom can be enough to cause fatal poisoning. An amateur mushroom hunter is interested in finding an easy rule to differentiate poisonous and edible mushrooms. They think that the mushroom's gills (the part which holds and releases spores) might be related to a mushroom's edibility. They used a data set of 8124 mushrooms and their descriptions. For each mushroom, the data set includes whether it is edible or poisonous and the spacing of the gills (Broad or Narrow).

Please Note: According to The Audubon Society Field Guide to North American Mushrooms, there is no simple rule for determining the edibility of a mushroom; no rule like “leaflets three, leave them be” for Poisonous Oak and Ivy.

| Class | Broad | Narrow | Total |
|-----------|-------|--------|-------|
| Edible | 3920 | 288 | 4208 |
| Poisonous | 1692 | 2224 | 3916 |
| Total | 5612 | 2512 | 8124 |

(a)[4 pts] Fill in each blank with one of the options in parentheses to best describe the variables collected.

Whether the mushroom is edible or poisonous is the (explanatory / response) _____ and it is (categorical / quantitative) _____.

Gill size (Broad or Narrow) is the (explanatory / response) _____ and it is (categorical / quantitative) _____.

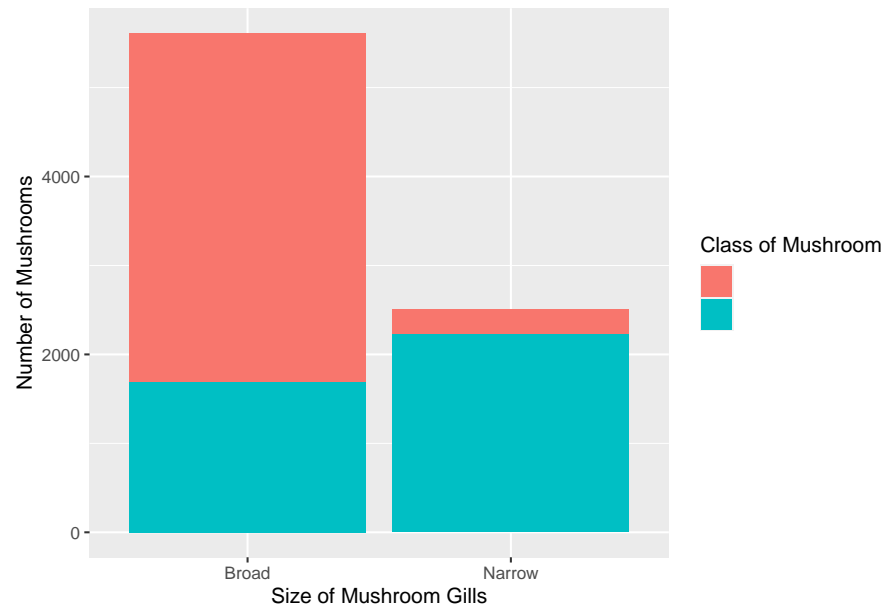
(b)[3 points] Calculate the proportion of mushrooms with a broad gill size that are poisonous. *Leave your value in **unreduced** fraction form.*

$$\frac{\text{_____}}{\text{(notation)}} = \frac{\text{_____}}{\text{(value)}}$$

(c)[3 points] Calculate the proportion of mushrooms with a narrow gill size that are poisonous. *Leave your value in **unreduced** fraction form.*

$$\frac{\text{_____}}{\text{(notation)}} = \frac{\text{_____}}{\text{(value)}}$$

(d)[2 points] Using your answers to (b) and (c), fill in the correct names next to each color, to label the bar chart showing the relationship between gill size (broad or narrow) and whether the mushroom is edible.



(e)[3 points] Based on the plot, describe the relationship between a mushrooms gill size and whether it is edible or not.

(f)[2 points] Suppose the Chi-Squared test resulted in a “significant” p-value. Which of the following would be the correct scope of inference for this study?

- (i) It can be inferred for all mushrooms that gill size causes a mushroom to be poisonous.
- (ii) It can be inferred for all mushrooms that gill size is associated with whether a mushroom is poisonous.
- (iii) It can be inferred for this sample of mushrooms that gill size causes a mushroom to be poisonous.
- (iv) It can be inferred for this sample of mushrooms that gill size is associated with whether a mushroom is poisonous.


Mendelian Genetics

Q2 [18 points]

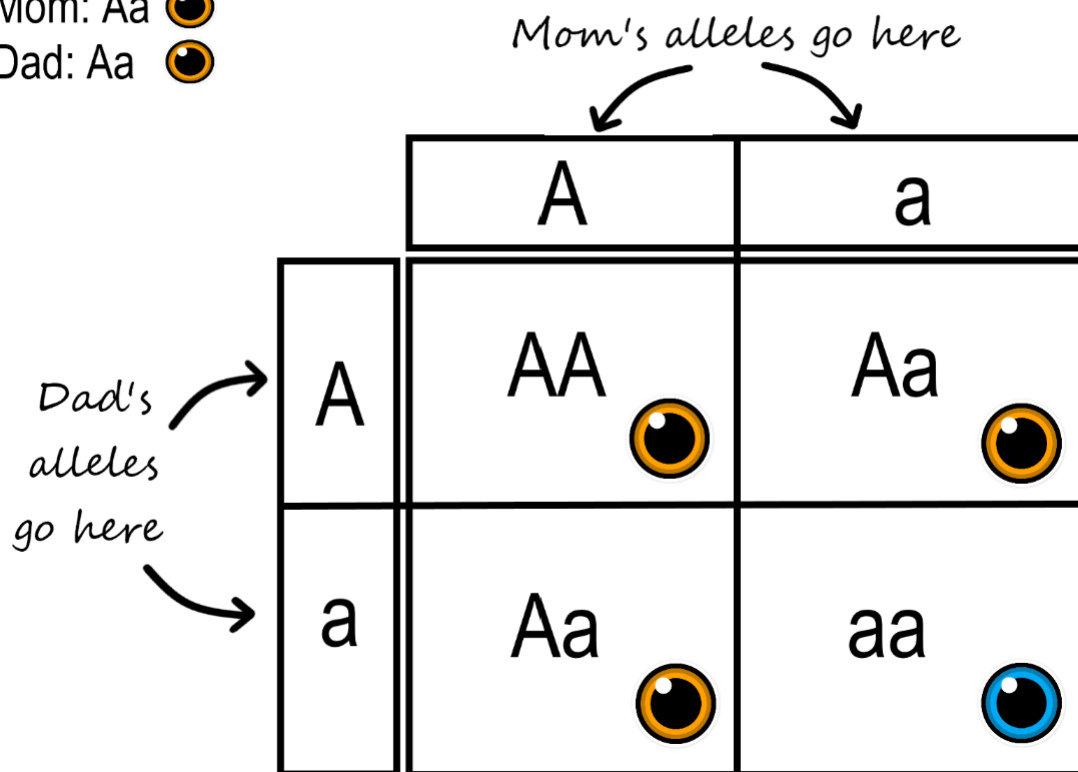
Mendelian inheritance refers to certain patterns of how traits are passed from parents to offspring. These general patterns were established by the Austrian monk Gregor Mendel, who performed thousands of experiments with pea plants in the 19th century. Mendel's discoveries of how traits (such as color and shape) are passed down from one generation to the next introduced the concept of dominant and recessive modes of inheritance.

Mendelian inheritance refers to the inheritance of traits controlled by a single gene with two alleles, one of which may be completely dominant to the other. You can use a Punnett square to determine the expected ratios of possible genotypes in the offspring of two parents.

In the table below, we see an example of eye color inheritance. In this case, both parents are heterozygotes (Aa) for the gene. Half of the gametes produced by each parent will have the A allele, and half will have the a allele, shown on the side and the top of the Punnett square. Filling in the cells of the Punnett square gives the possible genotypes of their children. It also shows the most likely ratios of the genotypes, which in this case is 25% AA, 50% Aa, and 25% aa.

Mom: Aa 

Dad: Aa 



(a) [2 points] When Mendel crossed his pea plants, he learned that tall (T) was dominant to short (t). Suppose in your Biology course you carried out an experiment to test if the plot offspring would follow Mendelian inheritance.

Fill in the cells of Punnett square to give the possible genotypes for plant tallness.

| | T | t |
|---|---|---|
| T | | |
| t | | |

(b) [3 points] If the Mendelian inheritance is true, what proportions would you expect for each of the following genotypes? Insert the corresponding values in each cell.

| TT | Tt | tt |
|--------------|--------------|--------------|
| $\pi_{TT} =$ | $\pi_{Tt} =$ | $\pi_{tt} =$ |

(c) [2 points] Actually, our table could be a bit simpler. Both the TT and Tt genotypes will present as “tall” plants, whereas tt genotypes will present as “short” plants.

Compress your previous table into a new table with only two levels of tallness.

| Tall | Short |
|----------------|-----------------|
| $\pi_{Tall} =$ | $\pi_{Short} =$ |

(d) [3 points] If the table above represents what Mendelian inheritance assumes to be true about tallness under H_0 , state the alternative hypothesis using words.

(e) [2 point] After you cross your plants, you measure the characteristics of the 400 offspring. You note that there are 305 tall pea plants and 95 short pea plants.

Fill in the table summarizing these observed counts.

| | Total |
|--|-------|
| | 400 |

(f) [4 points] Fill in the table below, summarizing the expected counts for these 400 plants.

| | |
|--|-------|
| | Total |
| | 400 |

(g) [4 points] Calculate how far “off” was your observed number of tall and short plants were from what you expected if H_0 was true. Use these values to report the X^2 statistic for your experiment.

Tall:

Short:

X^2 statistic:

(h) [3 points] The p-value associated with your X^2 statistic is 0.5645424. Your Biology textbook suggests you interpret this value as:

The large p-value proves that Mendelian inheritance is true.

What issue(s) to you have with this interpretation?

Seasonal Colds

Q3 [22 points] A local doctor suspects that there is a seasonal trend in the occurrence of the common cold. She estimates that 40% of the cases each year occur in the winter, 40% in the spring, 10% in the summer and 10% in the fall. A random sample of 1000 patient cases was collected, and the number of cold cases for each season was recorded.

A summary table of the observed counts is included below:

| Fall | Spring | Summer | Winter | Total |
|------|--------|--------|--------|-------|
| 165 | 292 | 169 | 374 | 835 |

(a) [4 points] If the doctor's suspicion was correct, what proportions would you expect for each cell? Insert the corresponding values in each cell.

| Fall | Spring | Summer | Winter |
|-----------------------|-------------------------|-------------------------|-------------------------|
| $\pi_{\text{fall}} =$ | $\pi_{\text{spring}} =$ | $\pi_{\text{summer}} =$ | $\pi_{\text{winter}} =$ |

(b) [3 points] If the table above represents what is assumed to be true under H_0 , state the alternative hypothesis using words.

(c) [4 points] Compute the table of expected counts.

| Fall | Spring | Summer | Winter |
|------|--------|--------|--------|
| | | | |

(d) [2 points] What is the summer cold cell's contribution to the X^2 statistic?

(e) [2 points] Evaluate whether the conditions required to use the χ^2 distribution to obtain a p-value are violated.

(f) [3 points] A X^2 statistic of 124 was obtained for these data. Fill in the R code below to find the p-value for this statistic. *Hint:* The `lower.tail` input takes one of two options TRUE or FALSE.

```
pchisq(_____, df = _____, lower.tail = _____)
```


(g) [4 points] Using the code you input above, a p-value of <0.00001 was obtained. Based on this p-value what would you conclude about the Doctor's hypothesis regarding the distribution of colds throughout the year?

Professor Attractiveness

Q4 [21 points] Data were scraped from *ratemyprofessors.com*, a website which allows college and university students to assign ratings to professors and campuses of American, Canadian, and United Kingdom institutions. In order for the rating to be posted, a rater must rate the professor in the following categories: overall quality, level of difficulty, and their hotness.

The Dean of the College of Science and Math would like to know if the difficulty rating of female professor's at Cal Poly is related to how attractive they are perceived to be.

A table of observed counts is shown below.

| Difficulty | Attractive | Unattractive | Total |
|------------|------------|--------------|-------|
| Easy | 76 | 54 | 130 |
| Hard | 45 | 170 | 215 |
| Medium | 51 | 90 | 141 |
| Total | 172 | 314 | 486 |

(a) [3 points] What type of test should be performed with these data to address the research question?

(b) [4 points] Write out the null and alternative hypotheses for the test you stated in (a).

H_0 :

H_A :

(c) [3 points] When analyzing the data, the administrators chose to use a χ^2 distribution to obtain their p-value. Was this an appropriate choice? Why or why not?

(d) [2 points] What χ^2 distribution did the administrators use to find their p-value?

(e) [4 points] A p-value of <0.00001 (X^2 statistic = 50) was found using the distribution stated in (d). Based on the p-value obtained, report your conclusion to the hypothesis test **in the context of the problem**.

Guilty Decisions

Q5[11 points] The effect of guilt on how a decision maker focuses on a problem was investigated in the *Journal of Behavioral Decision Making* (January 2007). A total of 155 volunteer students participated in the study, where each was randomly assigned to one of three emotional states (guilt, anger, or neutral) through a reading / writing task. Immediately after the task, the students were presented with a decision problem (e.g., whether or not to spend money on repairing a very old car). The researchers found that a higher proportion of students in the guilty-state group chose not to repair the car than those in the neutral-state and anger-state groups.

(a)[3 pts] What is the study design? Select one.

- (i) Observational study. The researchers did not take a random sample of students.
- (ii) Observational study. There is no random assignment of students to emotional state.
- (iii) Experiment. The students are a representative sample of all students.
- (iv) Experiment. The students were randomly assigned to emotional state.

(b)[3 pts] Which types of sampling bias may be present in this study? Select all that are present, or if you believe there is no bias present, select option (iv) No bias.

- (i) Selection bias
- (ii) Non-response bias
- (iii) Response bias
- (iv) No bias

(c) [3 pts] The researchers found a difference in the proportion of students in the guilty-state group chose not to repair the car than those in the neutral-state and anger-state groups. Can we conclude that the emotional state group caused a higher proportion of students to not repair the car? Select one.

- (i) Yes, because these data are from a representative sample.
- (ii) No, because the students are volunteers.
- (iii) Yes, because the researchers evened out confounding variables across emotional state group by randomly assigning emotional state.
- (iv) No, because the sample size is not large enough.

(d) [2 pts] Which type of plot would be the **most** appropriate to display the relationship between each level of emotional state and whether students choose to repair the car? Select one.

- (i) Scatterplot
- (ii) Filled bar plot
- (iii) Pie chart
- (iv) Side-by-side boxplot

Distribution of Fish on Blackfoot River

Q6 [19 points] Montana Fish, Wildlife, & Parks personnel have collected data on fish caught on the Blackfoot River (outside Helena, Montana) for the last 25 years. To capture the fish, fisheries biologists use electrofishing equipment to attract the fish to the boat, then dip them out of the water with nets. Each fish's length (in cm) and weight (in grams) is then measured. Once the measurements are taken, the fish is tossed back into the river. Biologists are often working in cold conditions in late autumn or early spring, so some measurement error and missing data are expected.

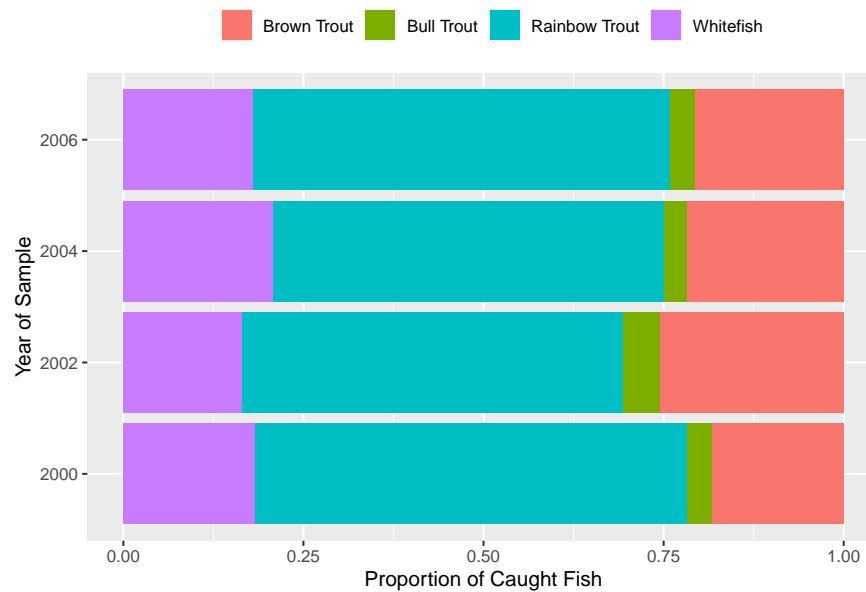
These data are not from a random sample. The goal is to catch all fish within a section of the Blackfoot River every few years to assess the health of the population. Montana Fish, Wildlife, & Parks is interested in analyzing if the prevalence of the different species of fish has stayed relative stable over the years. The dataset consists of 7729 observations, recorded over 4 years.

```
## # A tibble: 7,729 x 5
##   length weight year  section species
##   <dbl>  <dbl> <fct>  <chr>   <chr>
## 1    358    400 2000   Johnsrud Rainbow Trout
## 2    309    290 2000   Johnsrud Rainbow Trout
## 3    302    250 2000   Johnsrud Rainbow Trout
## 4    272    210 2000   Johnsrud Rainbow Trout
## 5    284    230 2000   Johnsrud Rainbow Trout
## 6    268    180 2000   Johnsrud Rainbow Trout
## 7    280    245 2000   Johnsrud Rainbow Trout
## 8    340    380 2000   Johnsrud Rainbow Trout
## 9    267    205 2000   Johnsrud Rainbow Trout
## 10   188     80 2000   Johnsrud Rainbow Trout
## # ... with 7,719 more rows
```

(a) [2 points] Based on the output above, what is the observational unit for this study?

(b) [3 points] Based on the output above, what type of variable is **year**? Given the stated analysis, is this the correct data type for this variable?

(c) [3 points] Based on the bar plot below, describe the relationship between the sampling year and the species of captured fish. Make direct reference to characteristics of the plot!



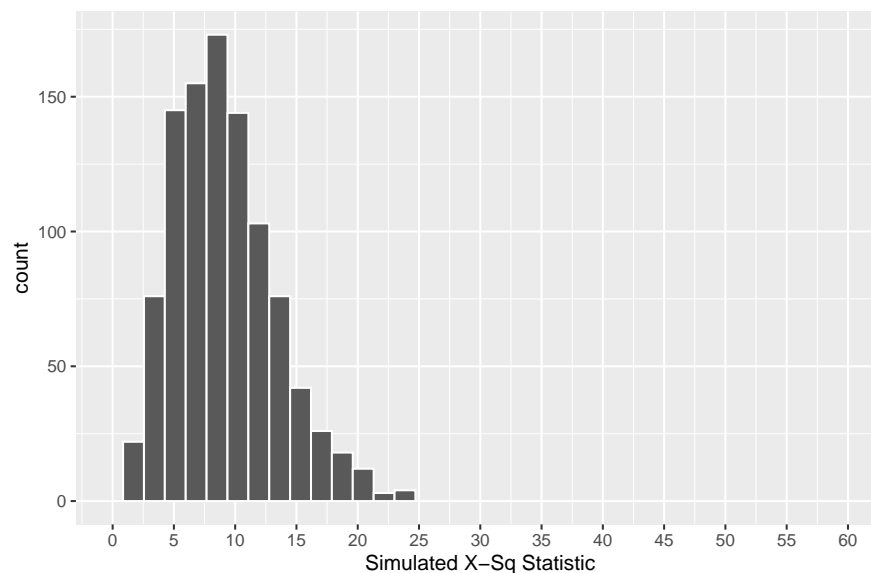
(d) [5 points] The Statistician at Montana Fish, Wildlife, & Parks prefers using simulation-based methods rather than theory-based methods. They used computer simulation to obtain the null distribution of the X^2 statistic shown on the next page.

Fill in the steps necessary to obtain out **one** statistic on the null distribution.

On (#) _____ cards, write _____ and _____ on the cards.

Generate a new sample that could have happened if the null hypothesis was true by:

Calculate and plot the _____ from each permutation / computer simulation.



(e) [2 points] The observed X^2 statistic was 57.42. Use this statistic to draw a line and shade the direction that should be used when calculating the p-value.

(f) [1 point] Approximate the p-value for this hypothesis test.

(g) [3 points] Using the conditions of the simulation-based method used, evaluate if you believe the p-value you obtained in (f) is accurate.

General Concepts

Q7[2 points] Suppose you reject the null hypothesis at the 0.05 level of significance. A colleague had planned to use a 0.01 level of significance instead. Will your colleague also reject the null hypothesis?

- (a) Yes
- (b) No
- (c) Maybe
- (d) Changing a level of significance cannot affect decisions.

Q8[2 points] In hypothesis testing,

- (a) a type II error occurs when you fail to reject a false null hypothesis.
- (b) you do not need to decide on a level of significance before you find a p-value; you can adjust based on the p-value you observe.
- (c) one of the main sources of type I errors is that the sample size was not large enough.
- (d) p-values are the probability that the null hypothesis is false.

Q9 For each of the following, select the single most appropriate analysis for the situation described. You may use an analysis for more than one situation. (2 pts each)

Chi-Square Test of Independence
Simple Linear Regression
Chi-Squared Goodness-of-Fit test
Confidence interval for μ
Hypothesis test for $\mu_1 - \mu_2$

One-Way ANOVA
Chi-Square Test of Homogeneity
Paired t-test
Hypothesis test for μ

- (a) Researchers are interested in investigating how the number of visitors to Yellowstone National Park in a year impacts the local economy in Livingston. To do this they count the number of yearly visitors to Yellowstone and measure the dollars spent by tourists in Livingston for the year.

- (b) A study of honeybees looked at whether the species of honeybees varied by state. Ten states were used in the study, and 100 honeybees were randomly sampled in each state, and 7 different species were seen in the data set.

- (c) An attorney in Boston observes that some judges seem to select juries that contain few women. She collects data on 20 randomly selected juries from each of 10 judges, and the number of women on each jury for each judge.

- (d) Researchers are interested in determining if the yield of a tomato plant differs among three tomato varieties.

- (e) You are interested in deciding if you should rent a new apartment off campus. As this will be your first time living off campus, you are anxious to know the average amount of time it should take you to walk to campus. What is the best **method** to estimate the average time it will take you to walk to campus?

- (f) Matchmaking data scientists are always investigating what characteristics of a person can produce better matches. Data scientists at Tinder are interested in looking into the relationship between someone's sexual orientation and whether they would date someone who is taller than them.
