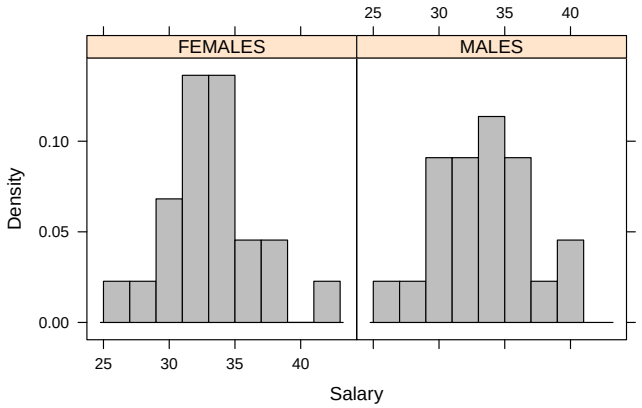
STAT 217: Homework 1

You may work in groups or on your own for this homework - your choice. If you work in a group, turn in one for your group.

1. *Data were collected to investigate differences in salaries between men and women university professors in America (Reference: Academe, Bulletin of the American Association of University Professors, year unknown). Men and women university professors from all over the country were randomly selected for the study. Specifically, they were interested in if men's salaries were higher than women's salaries. The units are thousands of dollars ($1K) The questions on this quiz are based on these data. The observed difference in means was* $\overline{y}_M - \overline{y}_F = 0.23182$ *thousands of dollars ($231.8).*

   (a) Below are summary statistics and histograms for the two groups. Are the equal variance and distributional assumptions met to run a parametric test?

```
##    .group  min   Q1 median    Q3  max  mean    sd  n missing
## 1 FEMALES 25.5 31.2  33.25 34.95 41.2 33.20 3.601 22       0
## 2   MALES 25.3 30.9  33.75 35.25 40.5 33.43 3.546 22       0
```



   i. Yes, the histograms are similar in shape.
   ii. No, the histograms are not similar in shape.
   iii. Yes, the histograms are approximately unimodal and symmetric and have approximately equal spread.
   iv. No, the histograms do not meet the nearly normal condition.

   (b) A test was run and the output is below.

```
##
##  Two Sample t-test
##
## data:  Salary by Gender
## t = -0.2151, df = 42, p-value = 0.4154
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##   -Inf 1.581
## sample estimates:
## mean in group FEMALES   mean in group MALES
##                 33.20                 33.43
```

   Was the test parametric or non-parametric? Circle one.

   (c) What is the distribution of the test statistic under the null hypothesis: $\mu_F - \mu_M = 0$?

   (d) Sketch this distribution. Draw a vertical line at the test statistic and shade in the area that is your p-value. Write your p-value next to the picture.

(e) How would you answer the research question?

i. There is strong evidence against the null hypothesis that there is no difference in the mean salaries paid to male and female professors at American Universities. Therefore, we conclude that based on these data male professors at these universities in America have higher average salaries as their female co-workers.

ii. We conclude that based on these data, male professors at these universities in America receive on average the same salaries than their female co-workers.

iii. There is no evidence of a difference in the mean salaries paid to male and female professors at American Universities.

iv. There is strong evidence against the null hypothesis that there is no difference in the mean salaries paid to male and female professors at American Universities (included in this study). Therefore, we conclude that based on these data, male professors at these universities in America receive on average the same salaries than their female co-workers.

(f) What is the scope of inference for this study?

i. Non-causal to the sample

ii. Since professors were randomly selected, inference extends to all professors worldwide. Since subjects were not randomly assigned to gender, we cannot infer a causal relationship between gender and salary.

iii. Professors were randomly selected from American universities, so inference extends all American university professors. Subjects were not randomly assigned to gender, so we cannot infer a causal relationship between gender and salary.

iv. Causal to the population

2. In 2004, a random sample of 30 25-year-old permanent Bozeman residents was obtained. Each subject was asked to rate their political tendency on a scale of 1-10 (1 being very liberal and 10 being very conservative). Evey five years, they were asked again to re-rate their political tendency on the same scale. That is, ratings of political tendancy for each subject were recorded at three time points (25 yrs, 30 yrs, and 35 yrs). Researchers are now interested in if the average political tendancy score is the same across the three age groups. Assess the independence assumption for this problem.

(a) The independence assumption is met because the subjects were selected randomly.

(b) The independence assumption is met because it is a sample of Bozeman residents.

(c) The independence assumption is not met because the three responses given by one individual are likely to be more similar to each other than to responses given by other people in the sample.

(d) The independence assumption is not met because Bozeman is a college town.

3. Forty subjects were recruited to participate in a clinical trial to assess the effectiveness of a new type of eye drop to alleviate dryness. The participants were randomly divided into four groups, and each of the groups were assigned to receive 1, 2, 3, or 4 eyedrops at regularly scheduled time intervals. After two weeks participants were given questionnaire to assess dryness. Reported in the table below are the standardized scores where higher scores indicate greater problems with dryness.

(a) Write the hypotheses to test whether reported eye dryness depends on the number of drops.

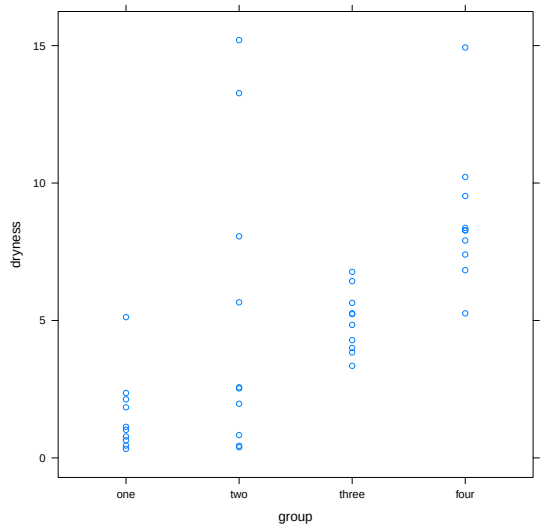
i. $H_0 : \bar{x}_{one} = \bar{x}_{two} = \bar{x}_{three} = \bar{x}_{four}$ vs H_A: At least one difference in the true mean reported dryness scores among the four groups

ii. $H_0 : \mu_{one} = \mu_{two} = \mu_{three} = \mu_{four}$ vs H_A: At least one difference in the true mean reported dryness scores among the four groups

iii. $H_0 : \mu_{one} = \mu_{two} = \mu_{three} = \mu_{four}$ vs H_A: At least one difference in the average reported dryness scores among the four groups in our sample

iv. $H_0 : \mu_{one} = \mu_{two} = \mu_{three} = \mu_{four}$ vs H_A: $\mu_{one} \neq \mu_{two} \neq \mu_{three} \neq \mu_{four}$

(b) The following output shows a plot of the data and some summary statistics. Draw short horizontal lines at the group averages on the plot. What is the average dryness for each of the groups?




```
##   .group  min    Q1 median    Q3   max  mean    sd  n missing
## 1    one 0.33 0.675  1.075 2.058  5.12 1.580 1.429 10       0
## 2    two 0.39 1.115  2.550 7.460 15.20 5.092 5.403 10       0
## 3  three 3.35 4.070  5.035 5.545  6.77 4.964 1.121 10       0
## 4   four 5.26 7.528  8.280 9.240 14.93 8.701 2.580 10       0
```

(c) One person who received two eyedrops reported an eye dryness of 15.20. What is the residual error for this observation? Be sure to show your calculations. Also label $e_{2,10}$ on the plot above.

(d) Draw a long horizontal line at the grand mean, $\bar{\bar{y}}$ on the plot above.

(e) Identify the single mean model with correct notation.

i. $y_{ij} = \mu + \epsilon_{ij}$, $i = 1, 2, 3, 4, 5$ $j = 1, 2, 3, 4, 5$
ii. $y_{ij} = \mu_j + \epsilon_{ij}$, $i = 1, 2, 3, 4, 5$ $j = 1, 2, 3, 4, 5$
iii. $y_{ij} = \mu + \epsilon_{ij}$, $i = 1, 2, ..., 10$ $j = 1, 2, 3, 4$
iv. $y_{ij} = \mu_j + \epsilon_{ij}$, $i = 1, 2, ..., 10$ $j = 1, 2, 3, 4$

(f) Idenfity the separate means model with correct notation.

i. $y_{ij} = \mu + \epsilon_{ij}$, $i = 1, 2, 3, 4, 5$ $j = 1, 2, 3, 4, 5$
ii. $y_{ij} = \mu_j + \epsilon_{ij}$, $i = 1, 2, 3, 4, 5$ $j = 1, 2, 3, 4, 5$
iii. $y_{ij} = \mu + \epsilon_{ij}$, $i = 1, 2, ..., 10$ $j = 1, 2, 3, 4$
iv. $y_{ij} = \mu_j + \epsilon_{ij}$, $i = 1, 2, ..., 10$ $j = 1, 2, 3, 4$

(g) I fit a linear model below. Label each as the cell means model or the reference coded. Briefly justify why you know.

Model 1:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5800     0.9894   1.597   0.1190
grouptwo      3.5120     1.3992   2.510   0.0167
groupthree    3.3840     1.3992   2.419   0.0208
groupfour     7.1210     1.3992   5.089 1.14e-05
```

Model 2:

```
Coefficients:
           Estimate Std. Error t value Pr(>|t|)
groupone     1.5800     0.9894   1.597    0.119
grouptwo     5.0920     0.9894   5.147 9.59e-06
groupthree   4.9640     0.9894   5.017 1.42e-05
groupfour    8.7010     0.9894   8.795 1.71e-10
```

(h) A ONE-WAY ANOVA model was fit and the ANOVA table is below. Fill in the blanks.

```
Analysis of Variance Table

Response: dryness
          Df  Sum Sq   Mean Sq  F value    Pr(>F)
group     ___ 253.75    _____    8.6412    0.0001884
Residuals ___ ______    9.788
```

(i) Would you choose the single mean model or the separate means model and why? Use the above ANOVA table to answer the question.

i. The separate means model because the group mean square is large relative to the residual mean square error.

ii. The separate means model because the group mean square is small relative to the residual mean square error.

iii. The single mean model because the group mean square is large relative to the residual mean square error.

iv. The single mean model because the group mean square is small relative to the residual mean square error.

(j) Based on the results of the statistical test, what is your conclusion?

i. Dryness score does not appear to differ across the number of drops.

ii. The true mean dryness scores for three and four groups are higher, but the true mean dryness scores for groups one and two are the same.

iii. There is no evidence of at least one difference among the true mean dryness scores across the number of drops (p-value=0.0001884 from F-stat= 8.6412 on ___ and __ df).

iv. There is strong evidence of at least one difference among the true mean dryness scores across the number of drops (p-value=0.0001884 from F-stat= 8.6412 on __ and __ df).

4. **Extra Credit**: Draw a picture of a scenario where a single mean model would be appropriate.