Mervar - Lab 2

Alexander Mervar

2022-09-14

Loading the Data and Preparing Libraries

```
library(dplyr) # great library for massaging data
library(ggplot2) #library for making great plots
library(ez) # library for doing ANOVA statistics easily

clownData <- read.csv("clowns.csv", header = TRUE, sep="")
head(clownData)
```

```
##   salary      party gender
## 1  34000 republican   male
## 2  31000 republican female
## 3  28000   democrat   male
## 4  29000   democrat female
## 5  30000 republican   male
## 6  23000 republican female
```

Problem 1

What are your independent variables and dependent variables? What are the different levels for each variable?
Independent: The salary of each clown is the independent variable.
Dependent: The political party and gender are the dependent variables for this analysis.
Levels for Each Variable: The levels for the independent value are continuous due to the fact that the salary can be any range of numbers. The levels for the dependent variable are Republican or Democrat for "party" and Male or Female for "gender."

Problem 2

a) What is the mean (R command: mean(x)) and standard deviation (R command: sd(x)) of all 24 clowns' salaries?

```
clownSalaryMean <- mean(clownData$salary)
print(paste("Salary Mean: ", clownSalaryMean))
```

```
## [1] "Salary Mean:  31312.5"
```

```
clownSalarySD <- sd(clownData$salary)
print(paste("Salary SD: ", clownSalarySD))
```

```
## [1] "Salary SD:  4515.22183960761"
```

b) Show a table of the mean of the clowns' salaries broken down separately for Democrats and Republicans. There are many ways to do this in R. One easy way is to use dplyr's "group_by" command, followed by a "summarize" command (another way would be to use the "aggregate" function in R's base package).

```
byPartyMean <- aggregate(salary~party, data=clownData, mean)
print(byPartyMean)
```

```
##        party   salary
## 1   democrat 31041.67
## 2 republican 31583.33
```

c) Show a table of the mean of the clowns' salaries broken down separately for males and females.

```
byGenderMean <- aggregate(salary~gender, data=clownData, mean)
print(byGenderMean)
```

```
##   gender   salary
## 1 female 29416.67
## 2   male 33208.33
```

d) Show a table of the mean of the clowns' salaries broken down separately for all for combinations of gender and party.

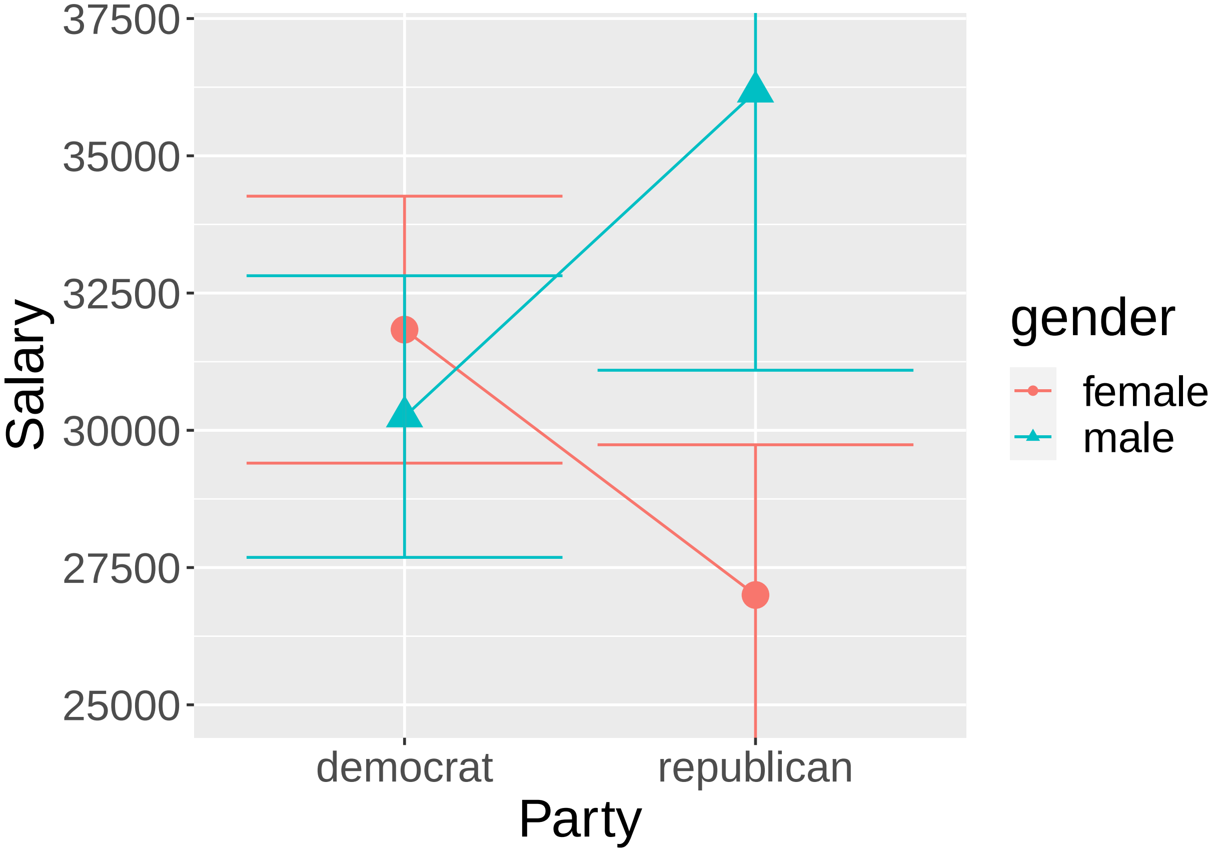
```
byGenderAndParty <- group_by(clownData, gender, party)
byGenderAndPartyMean <- summarize(byGenderAndParty, salary=mean(salary))
print(byGenderAndPartyMean)
```

```
## # A tibble: 4 x 3
## # Groups:   gender [2]
##   gender party      salary
##   <chr>  <chr>       <dbl>
## 1 female democrat   31833.
## 2 female republican 27000
## 3 male   democrat   30250
## 4 male   republican 36167.
```

Problem 3

Make a line graph of the means for male democrats, female democrats, male republicans, and female republicans, with all means shown in a single graph, and error bars showing 95% confidence intervals. For this task I use ggplot, and the sub-command I use to create the error bars is stat_summary(fun.data = mean_cl_normal, geom = "errorbar", position = position_dodge(width = 0.90),width=0.2). Some points will be taken off for ugliness (e.g. if the plot's legend overlap with a line) or imprecision (e.g. if you don't show a legend or label your axes). Plot the political affiliation on the x-axis and the salary on the y-axis. Identify the lines for male and female means by a change in line boldness, symbol shape, color, or dash style. Does it look like there is an interaction between the independent variables? Why or why not?

```
ggplot(clownData,aes(x=party,y=salary,color=gender,group=gender,shape=gender))+coord_cartesian(ylim = c
```



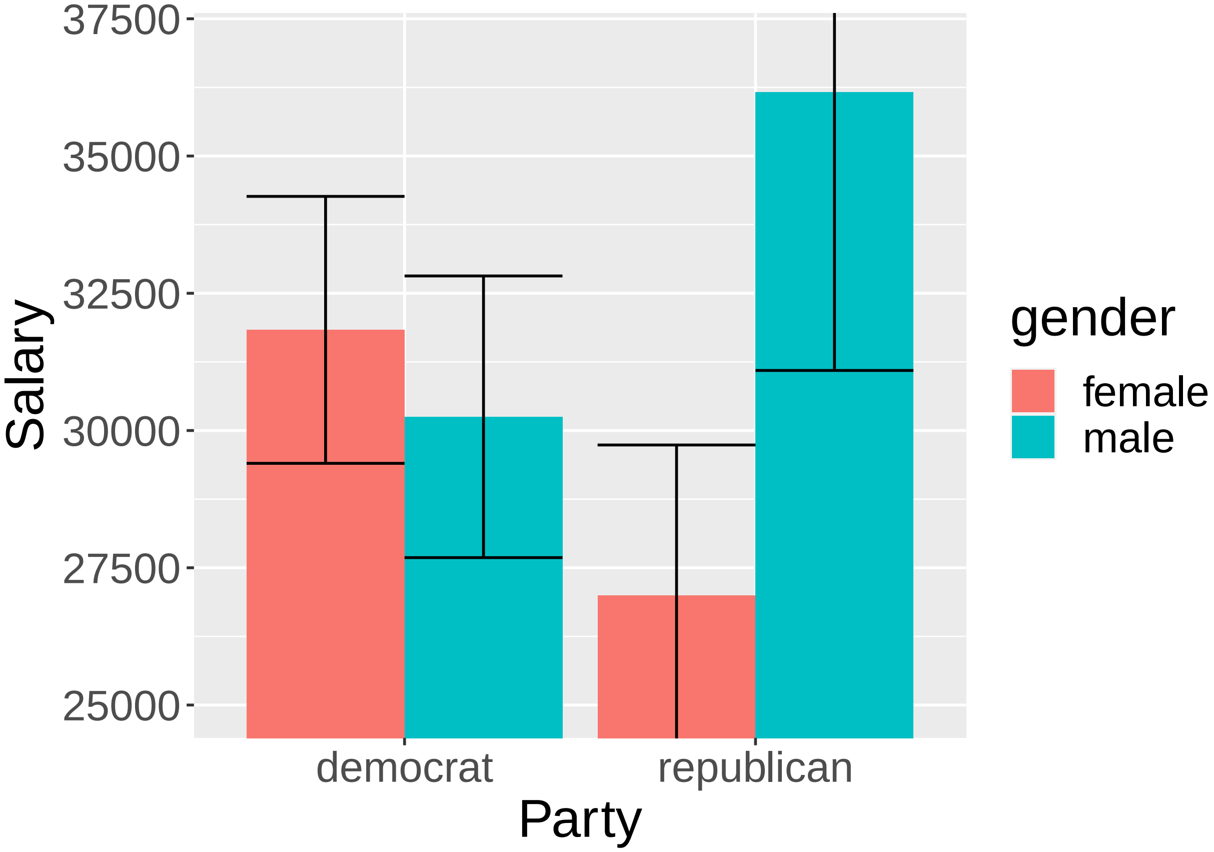
```
#ggplot(testonly,aes(x=schedule,y=hit,color=group,group=group,shape=group))+coord_cartesian(ylim = c(0.
```

There does indeed look like there is an interaction between the independent variables due to the fact that there is an intersection in lines as well as the means for each political party looking similar.

Problem 4

Make a bar graph with 95% confidence interval bars of the same four means as in 3). I use ggplot with geom= "bar" in stat_summary().

```
ggplot(clownData,aes(x=party,y=salary,fill=gender))+coord_cartesian(ylim = c(25000, 37000))+ stat_summa
```



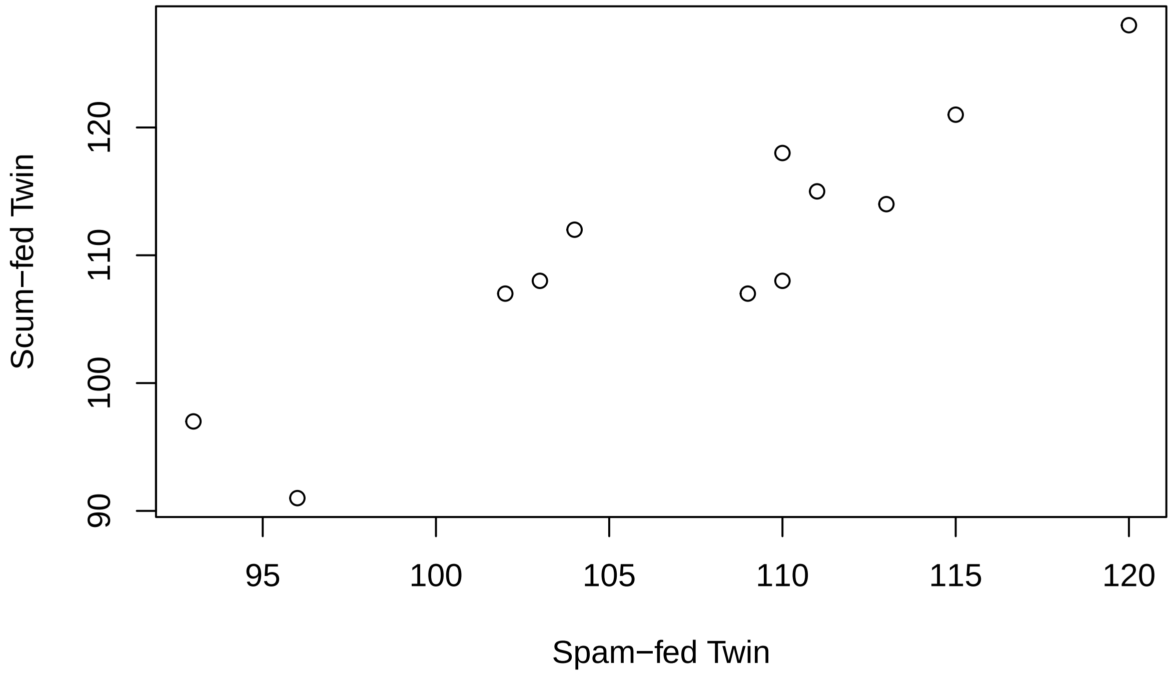
Problem 5

Make a scatterplot of the twins' IQs. Label the X and Y axes "Spam-fed Twin" and "Scum-fed Twin" respectively.

```
spamData <- read.csv("spam.csv",header = TRUE, sep="")
head(spamData)
```

```
##   pair spam scum
## 1    1  104  112
## 2    2  110  108
## 3    3   96   91
## 4    4  113  114
## 5    5  120  128
## 6    6  111  115
```

```
plot(spamData$spam, spamData$scum, xlab = "Spam-fed Twin", ylab = "Scum-fed Twin")
```



Problem 6

Are the IQs for twins from the same pair significantly correlated? Find the correlation coefficient. In R, cor(x,y) will find the correlation between two variables x and y, but if you want R to tell you if the correlation is significant you'll want to use cor.test(x,y).

```
cor.test(spamData$spam, spamData$scum)
```

```
##
##  Pearson's product-moment correlation
##
## data:  spamData$spam and spamData$scum
## t = 6.8975, df = 10, p-value = 4.207e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7006265 0.9745240
## sample estimates:
##       cor
## 0.9090197
```

There is a significant correlation between the spam and scum twin because $p < 0.05$. The correlation coefficient is estimated at 0.9090197.

Problem 7

Is there a significant difference in IQs due to diet differences? Use a paired samples t-test. You could use an R function like "t.test(x,y,paired=TRUE)"

```r
t.test(spamData$spam, spamData$scum, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  spamData$spam and spamData$scum
## t = -2.6386, df = 11, p-value = 0.02305
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -6.1138664 -0.5528003
## sample estimates:
## mean difference
##       -3.333333
```

There is a significant difference due to diet differences. We can reject the null hypothesis because $p < 0.05$.

Problem 8

Is there a significant difference in IQs due to diet differences? Use an independent samples (unpaired) t-test. "t.test(x,y)" would do the trick.

```
t.test(spamData$spam, spamData$scum)
```

```
##
##  Welch Two Sample t-test
##
## data:  spamData$spam and spamData$scum
## t = -0.90756, df = 20.77, p-value = 0.3745
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.976617   4.309951
## sample estimates:
## mean of x mean of y
##  107.1667  110.5000
```

There is NOT a significant difference due to diet differences. We cannot reject the null hypothesis because $p < 0.05$.

Problem 9

You should have gotten opposite answers to the question "Is there a significant difference" in Questions 7) and 8). How is this possible, and why did this happen in the current example? **This difference is due to the fact that a paired t-test accounts for matching between groups. In this case, we have twins so using a paried t-test is appropriate. Therefore, we can conclude there is a significant difference. The unpaired t-test is innapropriate for this use case.**