STAT 2300 Exam 1

Name: _____

Friday, Feb 14, 2025

Write directly on this exam. You may "show work" by handing in and R script, .Rmd file, or knit Markdown document.

You may use R, the internet, and any reference material. You are not allowed to communicate with anyone - no email, messaging, internet forums, AI, etc. If you happen to find exact copies of the exam questions on a "homework help" website, please bring that to the instructor's attention.

Honor Pledge (10 points)

The work I have submitted represents my own effort. While working on this exam, I did not use generative AI or communicate in any form with individuals other than the instructor.

Signed:

Problems 1-4 all use the data `Quarrels` from the `HistData` library.

Problem 1 (10 points)

a. From the help for the `Quarrels` data set: when was this data created, and who created it?

b. How many observations and how many variables are in this data set?

Solution *The Statistics Of Deadly Quarrels* by Lewis Fry Richardson (1960). Data has 779 observations and 84 variables.

Problem 2 (10 points)

Each row of this data represents a quarrel: two groups at war.

a. The `year` variable represents the year a quarrel began. What are lowest and highest values of `year` in this data set?

b. Find the two years which had the most quarrels begin.

Solution

```
range(Quarrels$year)
```

```
## [1] 1807 1949
```

```
Quarrels |> count(year) |> slice_max(n,n=2)
```

```
##   year  n
## 1 1941 28
## 2 1914 23
```

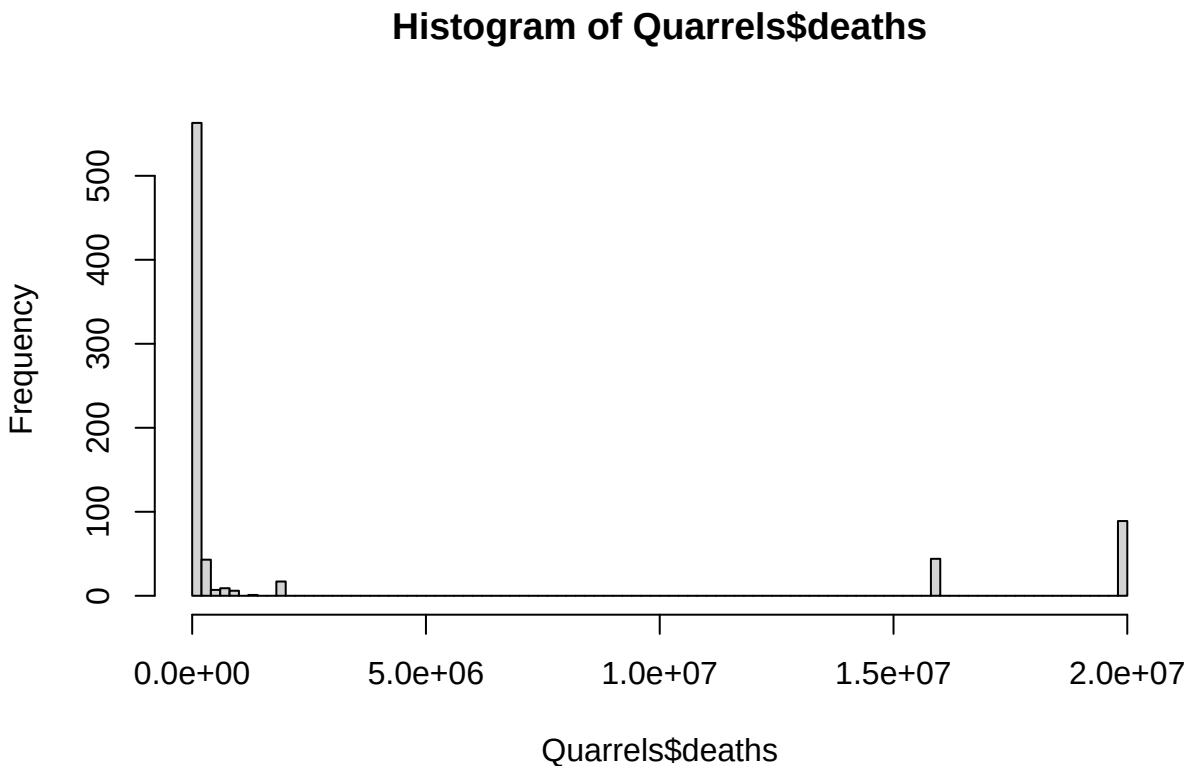
Problem 3 (10 points)

a. Make a histogram of the `deaths` variable from `Quarrels`. What do you observe?

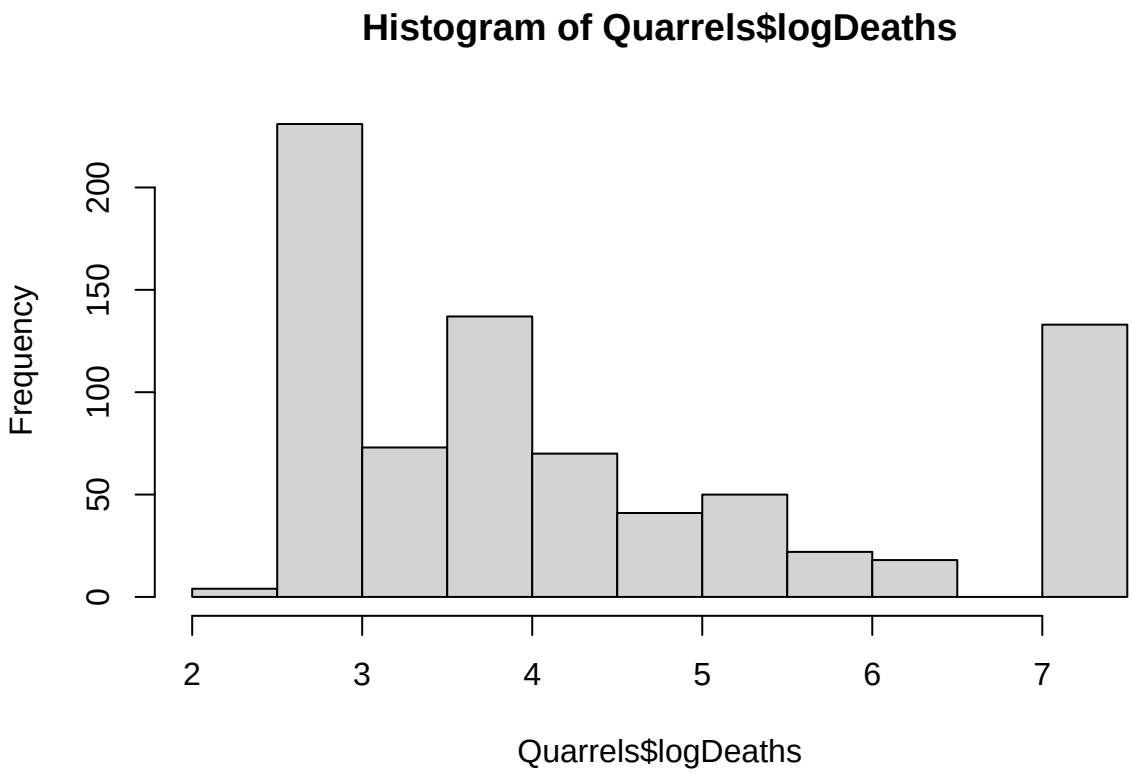
b. Make a histogram of the `logDeaths` variable from `Quarrels`. What do you observe?

Solution

```
hist(Quarrels$deaths, breaks=100)
```




```
hist(Quarrels$logDeaths)
```



`deaths` is extremely right skew, with two huge outliers (these are WWI and WWII). All other quarrels end up in the lowest bin. `logDeaths` is much less skew.

Problem 4 (10 points)

In `Quarrels`, the `prevConflict` variable indicates whether two groups in a quarrel had a previous conflict.

a. Perform a t-test for a difference in `deaths` between groups with a previous conflict and those without. State your conclusion with a p-value.

b. Perform a t-test for a difference in `logDeaths` between groups with a previous conflict and those without. State your conclusion with a p-value.

c. Which of these two tests would you choose, and why?

Solution

```
t.test(deaths ~ prevConflict, data=Quarrels)
```

```
##
##  Welch Two Sample t-test
##
## data:  deaths by prevConflict
## t = -0.21999, df = 364.3, p-value = 0.826
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -1260670  1006993
## sample estimates:
## mean in group 0 mean in group 1
##         3238148         3364986
```

```
t.test(logDeaths ~ prevConflict, data=Quarrels)
```

```
##
##  Welch Two Sample t-test
##
## data:  logDeaths by prevConflict
## t = -1.9906, df = 408.99, p-value = 0.04719
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.47519674 -0.00297978
## sample estimates:
## mean in group 0 mean in group 1
##        4.353187        4.592275
```

Using `deaths`, we fail to reject the null hypothesis: a previous conflict appears to make no difference in deaths (p=0.826). Using `logDeaths`, there is a significantly higher death count among groups with a previous conflict (p=0.047).

Because the deaths data is so skew, the assumptions of t-test are violated and I would choose the results using `logDeaths` instead.

Problem 5 (10 points)

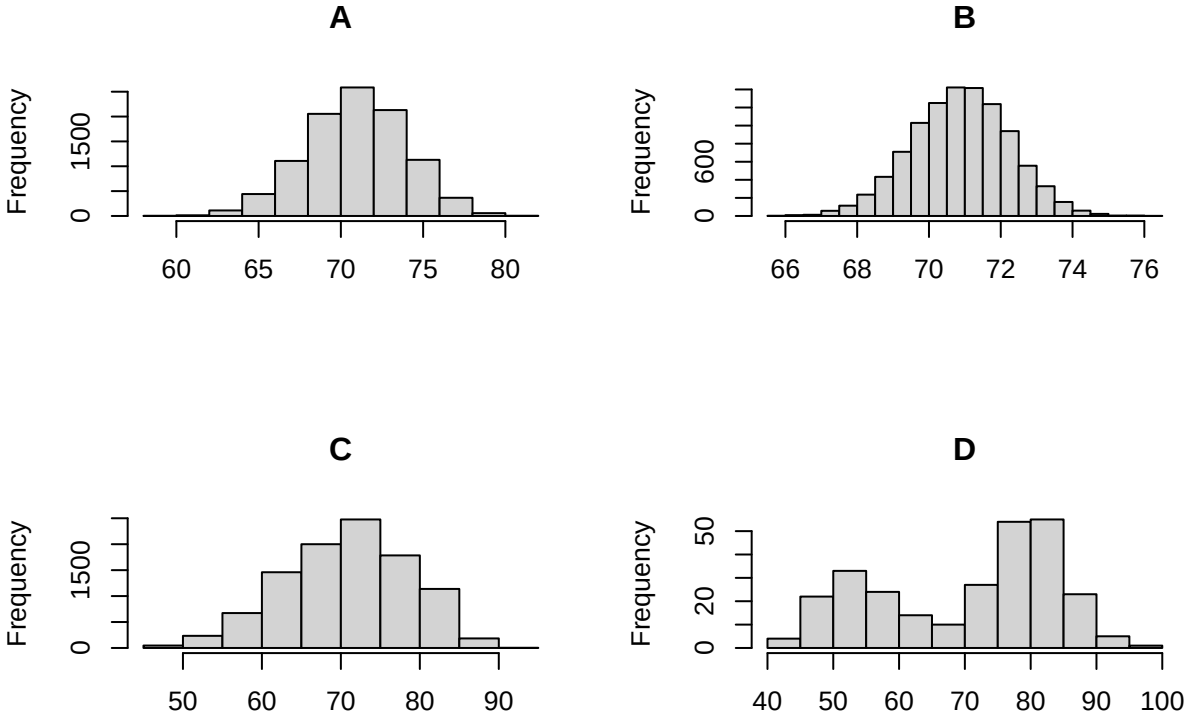
Four histograms are shown below. One is for the original data y from a measurement. The other three are sampling distributions for the mean $\bar{y}$ where the sample size is either $n = 3$, $n = 20$, or $n = 100$.

Which one is y?

Which one is $\bar{y}$ for $n = 3$?

Which one is $\bar{y}$ for $n = 20$?

Which one is $\bar{y}$ for $n = 100$?



Solution

D, C, A, B.

Problems 6-8 use the data `candytuft.csv`, which is available on our course web page at https://turtlegraphics.org/stat2300/data/candytuft.csv

This data is from an experiment run by Charles Darwin in 1876. Darwin took pairs of seedlings, one cross-fertilized and one self-fertilized, and planted them in the same pot. After the plants were fully grown, he recorded their heights.

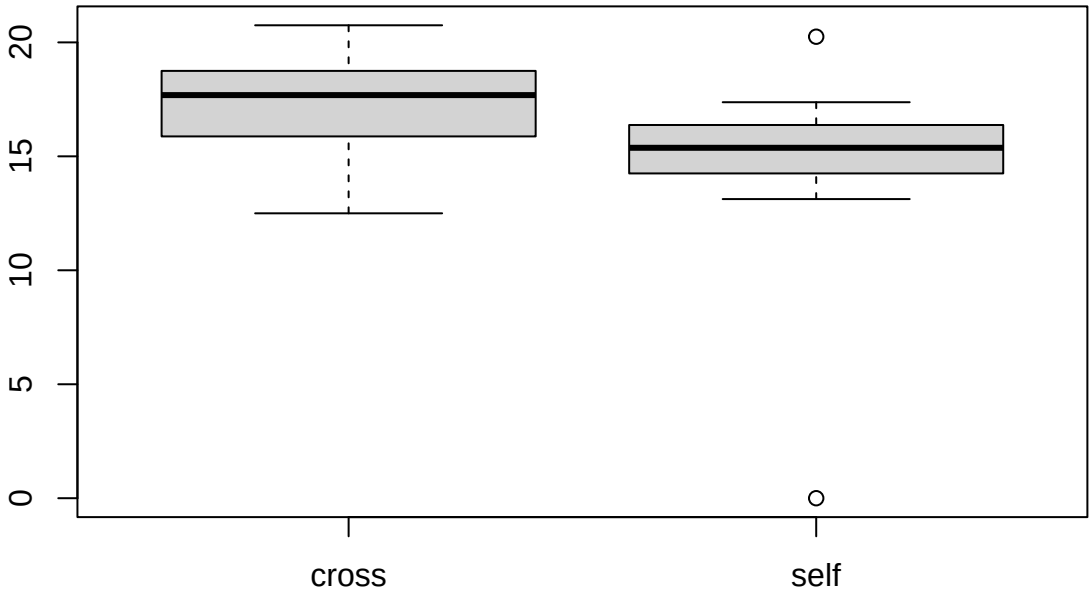
Problem 6 (10 points)

a. Make a side-by-side boxplot of the cross-fertilized and self-fertilized plant heights. Write your code here:

b. Which appears to have produced larger plants, cross-fertilization or self-fertilization?

Solution

```
plants <- read.csv("https://turtlegraphics.org/stat2300/data/candytuft.csv")
boxplot(plants)
```



It appears that cross fertilized plants grew a little taller.

Problem 7 (10 points)

Perform a t-test to determine a difference in heights between cross and self fertilized plants. State the null hypothesis, and then report your results with a p-value.

Solution

```
t.test(plants$cross, plants$self, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  plants$cross and plants$self
## t = 3.0227, df = 29, p-value = 0.005196
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.759931 3.940069
```

```
## sample estimates:
## mean difference
##            2.35
```

The null hypothesis is that cross and self fertilized plants have the same mean height. We reject this, there is a significant difference in height between the cross and self groups, $p = 0.0052$.

Problem 8 (10 points)

One plant in Darwin's experiment died, and had it's height recorded as zero. Remove the pair of plants where one died. How does this affect the p-value of the test?

Solution

```
plantsr <- plants |> filter(self > 0)
t.test(plantsr$cross, plantsr$self, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  plantsr$cross and plantsr$self
## t = 3.5414, df = 28, p-value = 0.001415
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.7304909 2.7350264
## sample estimates:
## mean difference
##        1.732759
```

Removing the outlier caused the p-value to get quite a bit smaller, now it is 0.0014.

Problem 9 (10 points)

Psychology experiments are often criticized for having low power. Say we design an experiment with only 10% power, and will publish if our results are significant at the 0.05 level.

a. If the effect we're looking for is real, what is the probability we will detect it and get to publish?

b. If the effect we're looking for is not real, what is the probability we detect it anyway, and get to publish?

c. What does this say about published papers with low power?

Solution

a. 10%. b. 5%. c. Out of 100 attempts at this sort of experiment, 10 correct ones will be published and 5 incorrect ones. So 1/3 of all the published papers with power this low are incorrect.

Problem X (10 points)

In `Quarrels`, the `international` variable indicates a war between two nations.

a. What percent of these quarrels were international?

b. Which year had the highest number of quarrels that were **not** international?

Solution

```
mean(Quarrels$international)
```

```
## [1] 0.2721438
```

```
Quarrels |> filter(international == 0) |> count(year) |> slice_max(n)
```

```
##   year  n
## 1 1936 19
```