R Notebook

Code

Preview Cmd+Shift+K to preview the HTML file). 1. You attend a scientific conference. One of the presenters showed the following figure: The presenter said the distribution of these 100 fish lengths was approximately normally distributed. However, the presenter did not say what the error bars in the figure meant. 10 cm is the distance between both the sample mean at 20 cm and the upper error bar at 30 cm, and the sample mean at 20 cm and the lower error bar at 10 cm. Note: t0.25 = 1.98 for 99 degrees of freedom (R command: qt(.025, df=99).

a. First assume 10 cm is the standard deviation of the fish lengths. Calculate the standard error of the mean and the 95% confidence interval for the mean.

Hide

```
stdev_fish = 10
n = 100
t_val = qt(0.25, df=99, lower.tail=FALSE)
std_error_fish = stdev_fish/sqrt(n)
confint_upper = 20 + (t_val*std_error_fish)
confint_lower = 20 - (t_val*std_error_fish)
cat("Standard error of the mean:", std_error_fish)
```

```
Standard error of the mean: 1
```

Hide

```
cat("\nThe 95% confidence interval is ", confint_lower, "to", confint_upper)
```

```

The 95% confidence interval is  19.32302 to 20.67698
```

b. Next assume 10 cm is the standard error of the mean. Calculate the standard deviation of the fish lengths and the 95% confidence interval for the mean.

Hide

```
std_error_mean = 10
std_dev = std_error_mean*sqrt(n)
confint2_upper = 20 + (t_val*std_error_mean)
confint2_lower = 20 - (t_val*std_error_mean)
cat("Standard deviation:", std_dev)
```

```
Standard deviation: 100
```

Hide

```
cat("\nThe 95% confidence interval is ", confint2_lower, "to", confint2_upper)
```

```
The 95% confidence interval is  13.23024 to 26.76976
```

c. Finally assume the 95% confidence interval for the mean is (10 cm, 30 cm). Calculate the standard deviation of the fish lengths and the standard error of the mean.

Hide

```
confint3_upper = 30
confint3_lower = 10
std_error_mean3 = (confint3_upper - 20)/t_val
std_dev3 = std_error_mean3*sqrt(n)
cat("Standard deviation:", std_dev3)
```

```
Standard deviation: 147.7157
```

Hide

```
cat("\nStandard error:", std_error_mean3)
```

```

Standard error: 14.77157
```

d. Which of these three assumptions (a, b, or c) do you think is correct? Explain why. I think C is correct because it's confidence interval confirms the initial upper and lower errors bars at 30 and 10cm while the first and second assumptions have very small or large confidence intervals.

2. A friend flips a coin 100 times and gets 40 heads. Assume the flips are independent and identically distributed.

a. Calculate the two-tailed p-value that the coin is fair (i.e., the probability of heads on each flip is 50%). This is the same as calculating that probability that you get 40 or fewer heads or 60 or more heads if the coin is fair.

Hide

```
flips = 100
heads = 40
prob = 0.5
expectation = flips*prob
std_dev4 = sqrt(flips*prob*(1 - prob))
X = (prob-expectation)/(std_dev4)
prob_lower = pnorm(X)
prob_upper = pnorm(-X, lower.tail=FALSE)
prob_value = prob_lower + prob_upper
cat("The p-value is: ", prob_value)
```

```
The p-value is:  4.16275e-23
```

b. Based on your answer to part (a), can you quantify the probability that the coin is fair? If yes, do so. If no, explain why not.
The p-value indicates that if the coin were fair, then the probability of observing such extreme results is extremely lower, however, it doesn't tell us the probability that the coin is fair because p-values only assess the consistency of the data within the null hypothesis, not the likelihood of the process itself.

3. In a genome wide association study (GWAS), genetic variants from across the genome are separately tested for association with a phenotypic trait. For each genetic variant the null hypothesis is that this variant is not associated with the trait.

a. Calculate the probability that the null hypothesis is rejected at a randomly chosen variant.
b. Given that the null hypothesis is rejected at a variant, calculate the conditional probability that this variant is actually associated with the trait.
c. Let's plug some numbers into your answer for part (b). First let's not include a multiple tests correction, let α = .05, π = .8, and γ = .01. Next let's assume there are 100,000 genetic variants and apply the Bonferroni correction, so now α = .05/10$. (look at pdf)

Hide

```
alpha = 0.05
gamma = 0.01
pi = 0.8
prob_nullreject = (gamma*pi)/((alpha*(1-gamma))+gamma*pi)
cat("Probability that the null hypothesis is rejected at randomly chosen variant is ", p
rob_nullreject*100, "%")
```

```
Probability that the null hypothesis is rejected at randomly chosen variant is  13.91304
%
```

Hide

```
alpha2 = 0.05/100000
prob_nullreject2 = (gamma*pi)/((alpha2*(1-gamma))+gamma * pi)
cat("\nThe probability that the variant is actually associated with the trait given the
null hypothesis is rejected at a variant is ", (prob_nullreject2)*100, "%")
```

```


The probability that the variant is actually associated with the trait given the null hy
pothesis is rejected at a variant is  99.99381 %
```

Recall in HW #1 problem #1 we considered the Cleveland Heart dataset (posted on Brightspace). Below I am going to repeat our earlier explanation of this dataset. The data is consecutive patients that were referred for coronary angiography at the Cleveland Clinic between May 1981 and September 1984. Each row is a different patient. No patient had a history or electrocardiographic evidence of prior myocardial infarction or known valvular or cardiomyopathic disease. Here is what I want you to do for this assignment (in R). Make sure you label all figures. a. Make side-by-side boxplots. The one boxplot is of MaxHR for those patients diagnosed with heart disease (Yes for AHD) and the other boxplot is of MaxHR for those patients not diagnosed with heart disease (No for AHD). (Hint: the document "rnb1" in the R folder of the class Brightspace page will be helpful if you have forgotten some R commands.)

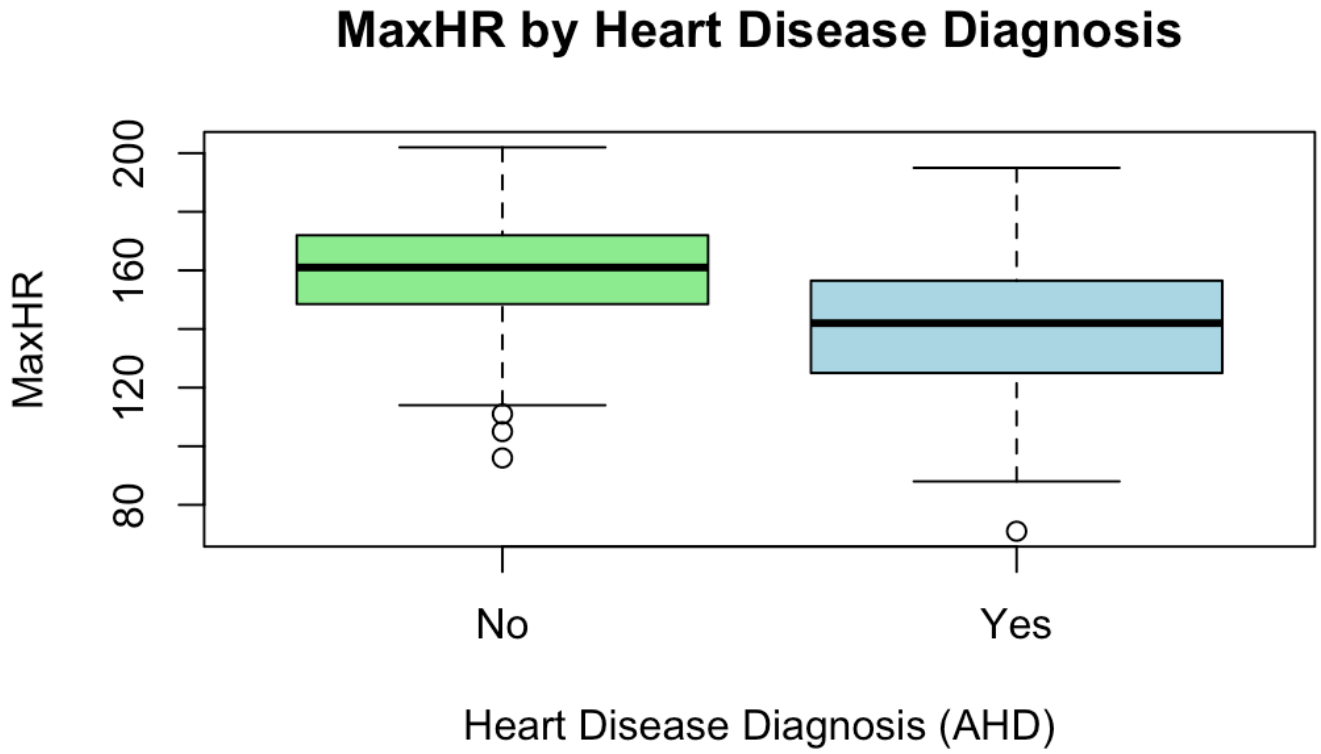
Hide

```
clevheart = read.csv("ClevelandHeart.csv")
dim(clevheart)
```

```
[1] 303  12
```

Hide

```
boxplot(MaxHR ~ AHD, data = clevheart,
        main = "MaxHR by Heart Disease Diagnosis",
        xlab = "Heart Disease Diagnosis (AHD)",
        ylab = "MaxHR",
        col = c("lightgreen", "lightblue"))
```

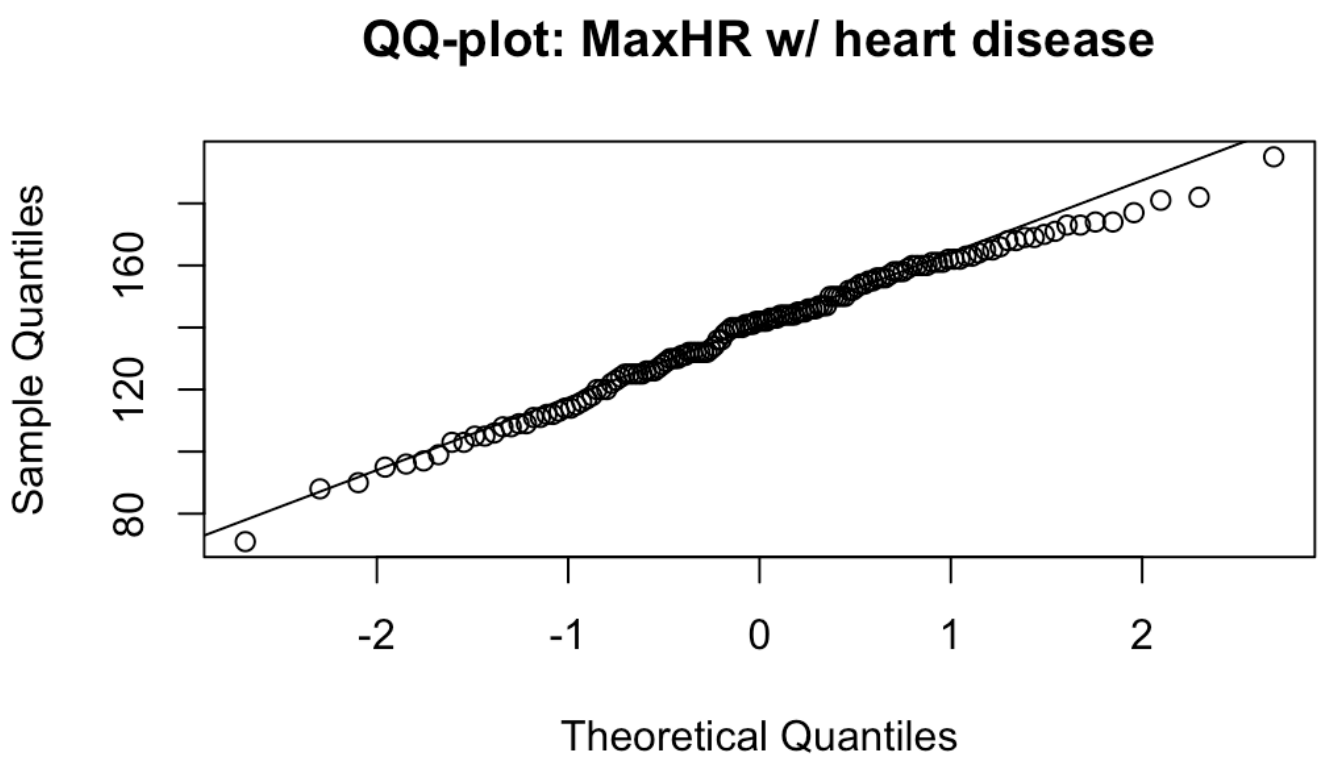


b. Make two separate QQ-plots. The one QQ-plot is of MaxHR for those patients diagnosed with heart disease and the other QQ-plot is of MaxHR for those patients not diagnosed heart disease. Are these two distributions each approximately normally distributed? (Hint: the R command to make a QQ-plot is qqnorm)

Both distributions are approximately normal because they appear to follow the theoretical normal line closely. The tails have sllight deviation, but otherwise the distributions are normal.

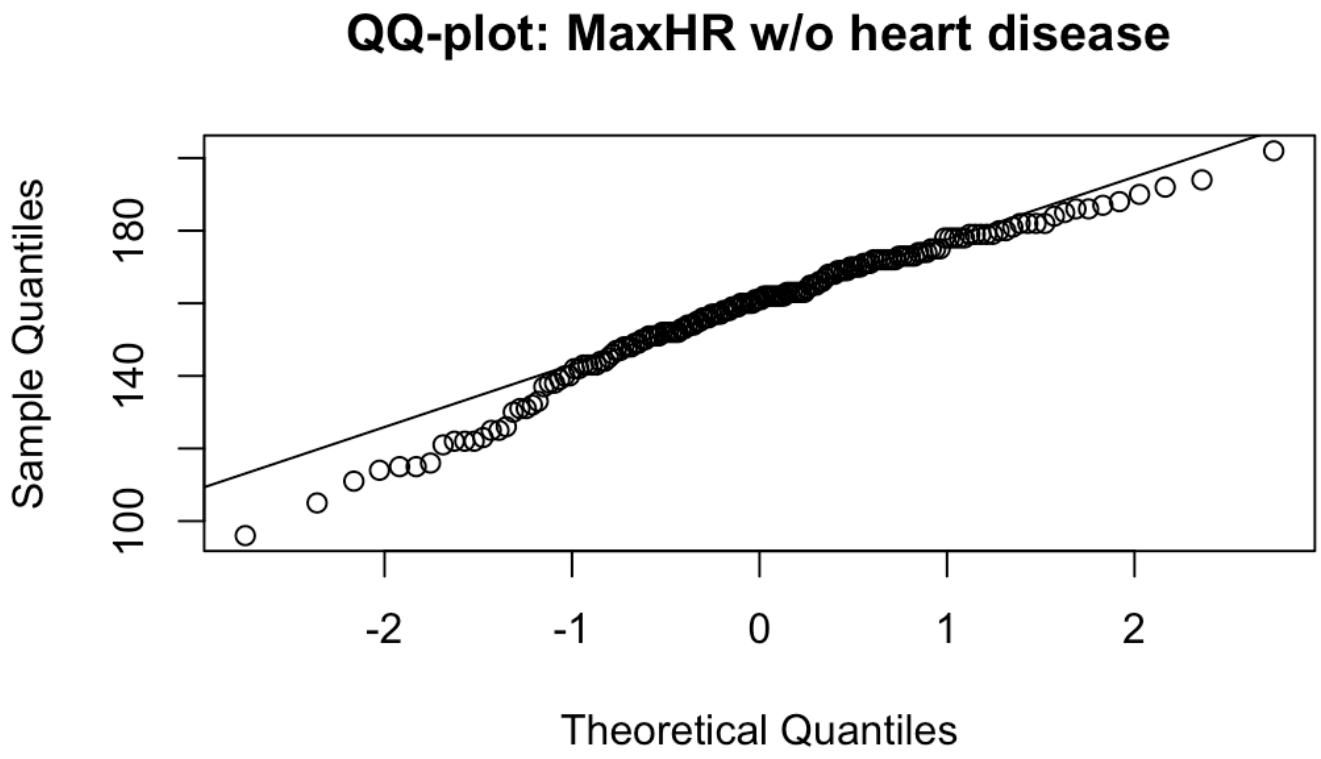
Hide

```
qqnorm(clev$MaxHR[clev$AHD == "Yes"], main = "QQ-plot: MaxHR w/ heart disease")
qqline(clev$MaxHR[clev$AHD == "Yes"])
```



Hide

```
qqnorm(clev$MaxHR[clev$AHD == "No"], main = "QQ-plot: MaxHR w/o heart disease")
qqline(clev$MaxHR[clev$AHD == "No"])
```



c. Do the t-test with significance level $\alpha = 0.05$.

Null hypothesis The population mean MaxHR for those patients diagnosed with heart disease equals the population mean MaxHR for those patients not diagnosed with heart disease.

Alternative hypothesis The population mean MaxHR for those patients diagnosed with heart disease does not equal the population mean MaxHR for those patients not diagnosed with heart disease. Report the p-value, the 95% confidence interval for the difference of the means, and clearly state your conclusion.

Hide

```
ttest_clev <- t.test(MaxHR ~ AHD, data = clev)
print(ttest_clev)
```

```
    Welch Two Sample t-test

data:  MaxHR by AHD
t = 7.8579, df = 272.27, p-value = 9.106e-14
alternative hypothesis: true difference in means between group No and group Yes is not e
qual to 0
95 percent confidence interval:
 14.32900 23.90912
sample estimates:
 mean in group No mean in group Yes
          158.378           139.259
```

The analysis produced a p-value of 9.106e-14 and a 95% confidence interval from 14.32900 - 23.90912. Since the p-value is much smaller than 0.5, we reject the null hypothesis and conclude the mean MaxHR for patients diagnosed with heart disease doesn't equal the population mean MaxHR for patients not diagnosed with heart disease.