Day 1

Olivia Wu

2024-02-05

Red shows the correct answer or where I should have added more detail

Problem 0.4

a) Observational Units: individual players points
Explanatory: Type of serve (categorical, jump or standard overhand)
Response: Frequency of wins Outcome of a point (categorical, win or lose)

b) Observational Units: Games played in the 2016 and 2017 season of major team sports
Explanatory: Type of team sport (categorical, baseball, football, basketball, and hockey)
Response: Frequency of home wins Outcome of home game (categorical, win or lose)

c) Observational Units: Professional golfers
Explanatory: Gender (categorical, men and women)
Response: Average distance a golf ball is driven and the percentage of drives that hit the fairway (quantitative)

Problem 0.5

a) The observational units in this study are the 85 nutrition experts

b) This is a controlled experiment because the researchers actively gave randomized treatments to the observational units.

c) The response variable is the amount of ice cream each subject scooped, which is quantitative.

d) The explanatory variable is the bowl and spoon size each subject was given, which is categorical. The two levels for bowls are 17 oz and 34 oz bowls. The two levels for spoons are 2 oz and 3 oz spoons.

Problem 0.8

a) The response variable is the predicted number of games an NFL team will win, which is quantitative.

b) The explanatory variables are both quantitative, and they are the average points a team scores per game over an entire season and the points allowed per game.

c) The expected wins increases by 1.5.

d) The expected wins increases by 0.9.

e) They should focus on improving offense because more wins are gained with 1 more point scored than 1 less point allowed.

f) This is an observational study.

Problem 0.11

a) When the bowl size increased, the average amount of ice cream scooped also increased for both spoon sizes.

b) When the spoon size increased, the average amount of ice cream scooped increased for both bowl sizes.

c) Increasing the bowl size resulted in a 16% and 13% increase in ice cream scooped for the 2-oz and 3-oz spoons, respectively. Increasing the spoon size resulted in a 33% and 30% increase in ice cream scooped for the 17-oz and 34-oz bowls, respectively. Overall, increasing the size of the spoon appeared to have a greater effect.

d) The effect of the bowl size appears similar for both spoon sizes.

Problem 0.13

a)

$$\begin{aligned} Wins &= 3.6 + 0.5PF - 0.3PA + \epsilon \\ &= 3.6 + 0.5(27) - 0.3(24.25) + \epsilon \\ &= \boxed{9.825} + \epsilon \end{aligned}$$

b) $y - \hat{y} = 10 - 9.825 = \boxed{0.175}$ wins

c) $11 - 3.04 = \boxed{7.96}$ wins

Problem 0.16

a) Age: negative because the more recent ones should have better technology that allow it to achieve faster speeds safely
Total length: it depends on the layout of the track
Max height: positive because there's more potential energy
Max drop: positive because it contributes greater fall time and acceleration

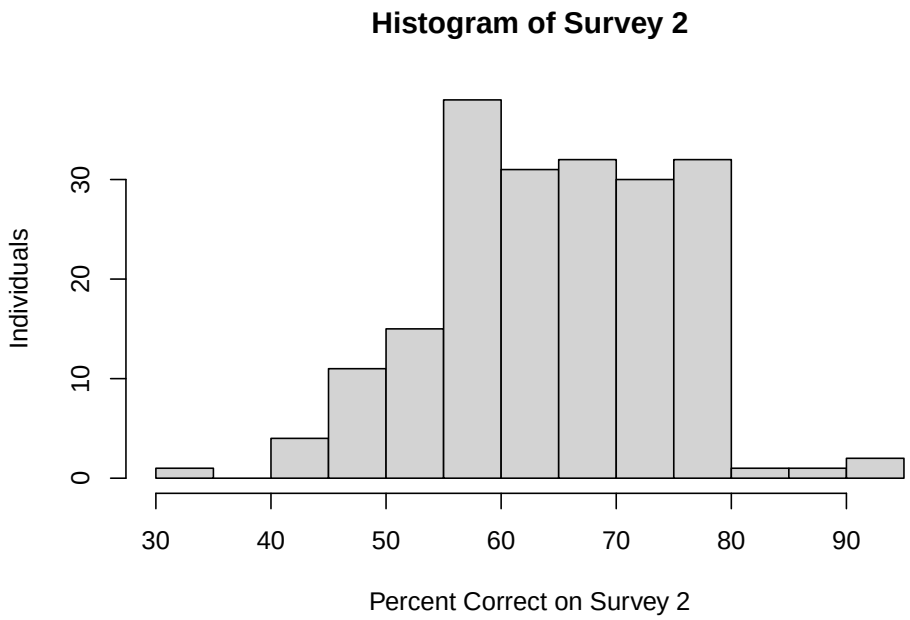
b) I think it is the maximum vertical drop because the roller coaster will get most of its speed from that.

c) The signs of the coefficients are pretty much what I expected, and the value of the coefficient for length is so small that it could have gone either way (pos or neg).

Problem 0.18

a) Histogram

```r
hist(Handwriting$Survey2, main="Histogram of Survey 2",
     xlab = "Percent Correct on Survey 2", ylab = "Individuals")
```



This histogram is roughly normal with a center of about 65%. It shows us that most people correctly identified 55-80% of the handwritings during the second survey.

b) $\hat{y} = \mu + \epsilon$ for $\mu > 0.5$

c) Fitted model

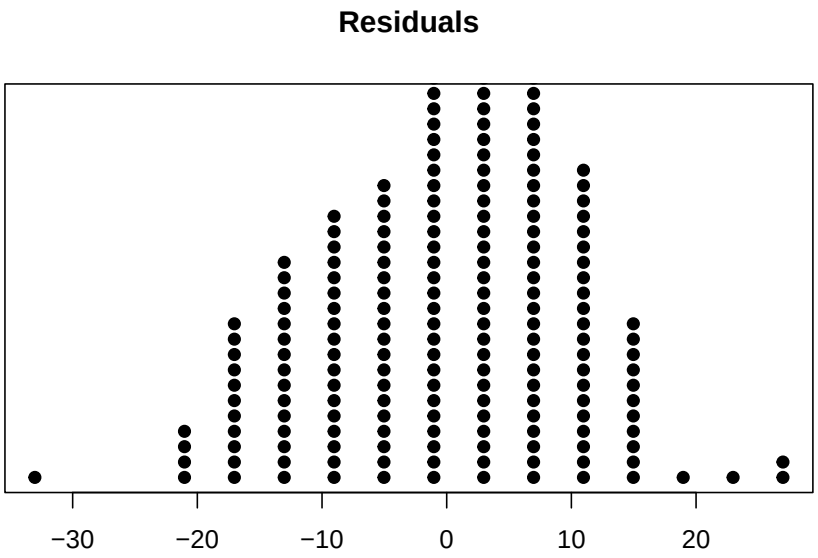
```r
mu <- mean(Handwriting$Survey2, na.rm=TRUE)
```

$\implies \mu = 65.03 \implies \hat{y} = 65.03 + \epsilon$

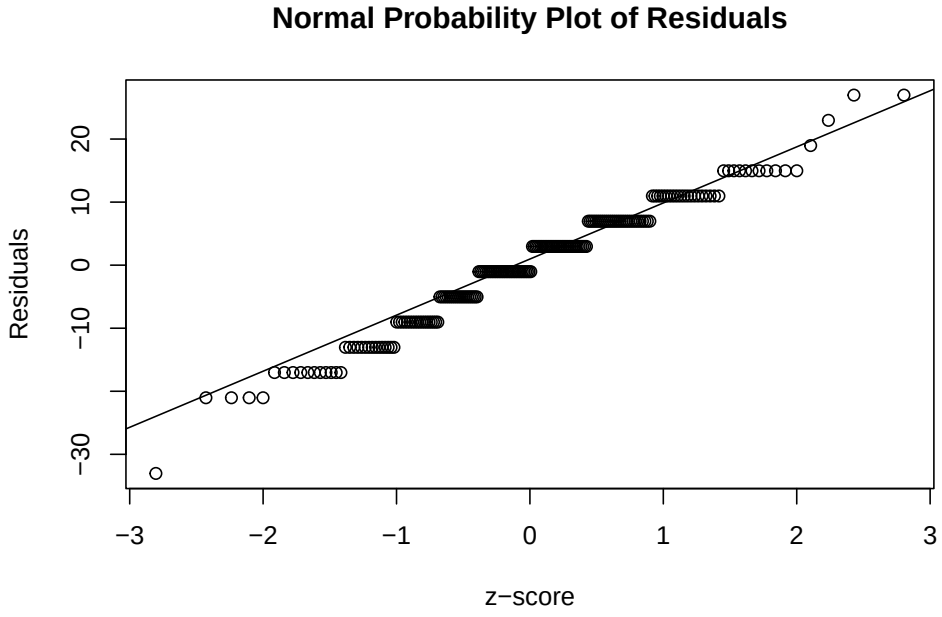
Problem 0.20

a) The dotplot of residuals looks roughly normal, and the normal probability plot looks linear.

```r
resid <- Handwriting$Survey2[!is.na(Handwriting$Survey2)] - mu
stripchart(resid, method="stack", offset=.5, at=0, pch=19, main="Residuals")
```




```
qqnorm(resid, ylab="Residuals", xlab="z-score", main="Normal Probability Plot of Residuals")
qqline(resid)
```



b) **State** H_0: $\mu = 50$
H_a: $\mu > 50$

Plan
Random: given ✓

Normal: $n = 198 \geq 30$✓
Independent: There is likely more than $203 * 10 = 2030$ people in Clark University. ✓

Do

```
t.test(Handwriting$Survey2, mu=50, alternative="greater")
```

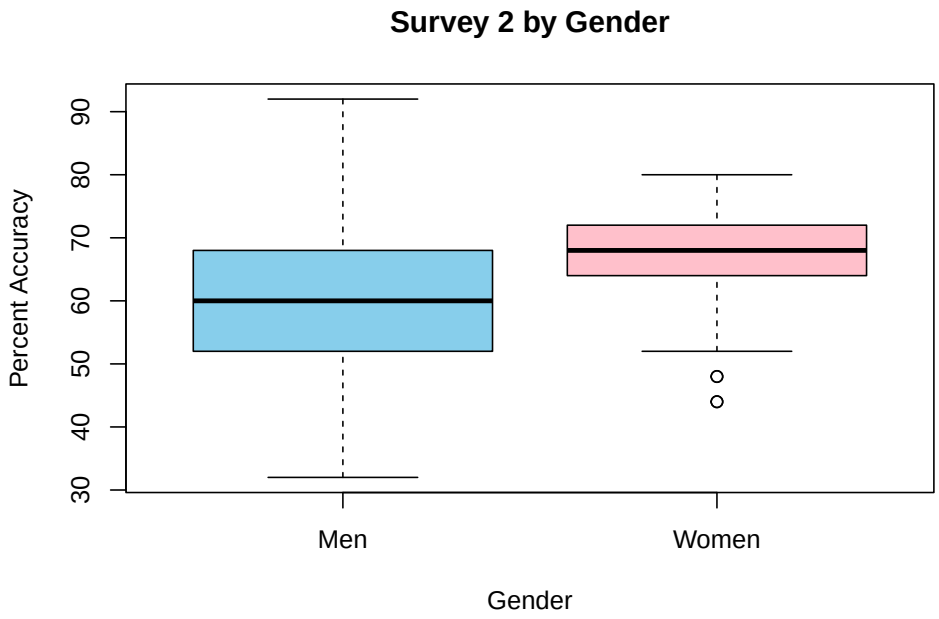
```
##
##  One Sample t-test
##
## data:  Handwriting$Survey2
## t = 21.161, df = 197, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 50
## 95 percent confidence interval:
##  63.85649      Inf
## sample estimates:
## mean of x
##   65.0303
```

Conclude The p-value ≈ 0, so we reject the null hypothesis. There is sufficient evidence to conclude that the subjects are better in gueessing the author's gender than a fair coin.

Problem 0.22

a) These grouped boxplots shows that women tend to have a 70% accuracy and men tend to have a 60% accuracy. They both have symmetric distributions, and there is also less variation among women than men. A model would be $\hat{y} = \mu_i + \epsilon$ for μ_1, the population mean for men, and μ_2, the population mean for women.

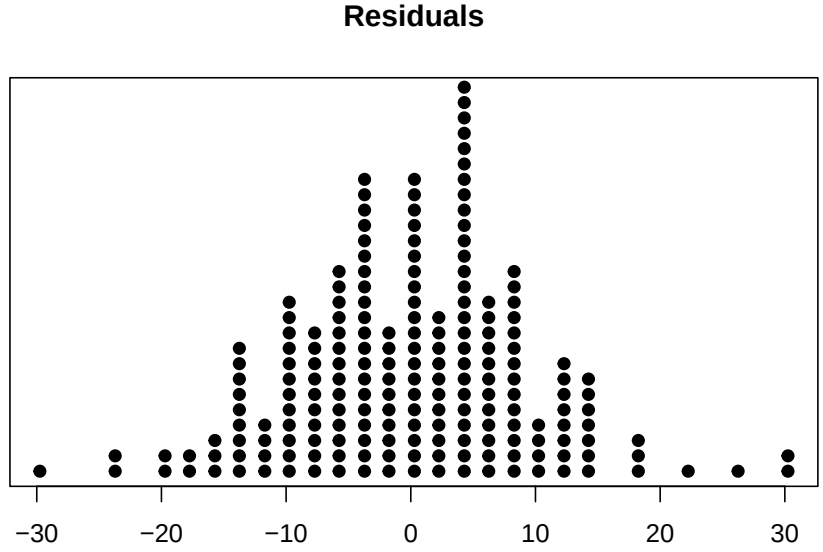
```
boxplot(Survey2~Gender, data=Handwriting,
        col=c("skyblue","pink"),
        main="Survey 2 by Gender",
        xlab="Gender",
        ylab="Percent Accuracy",
        names=c("Men","Women"))
```



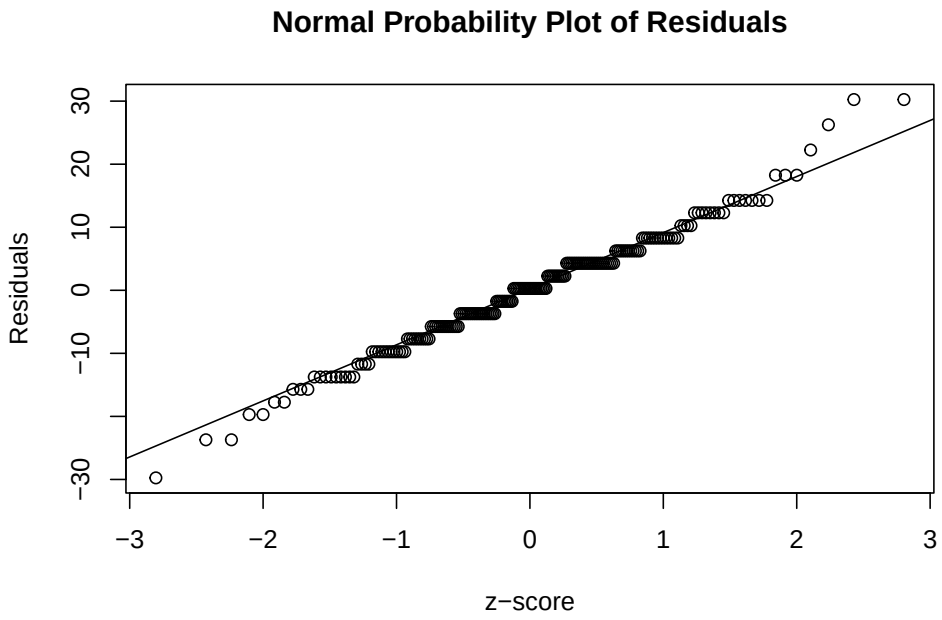
b) The sample mean for women is 67.71 and for men is 61.75. The respective fitted models would be: $\hat{y} = 67.71 + \epsilon$ and $\hat{y} = 61.75 + \epsilon$.

c) The following dotplot looks normal and the normal probability plot looks roughly linear.

```
resid1 <- Handwriting$Survey2[!is.na(Handwriting$Survey2) & Handwriting$Gender %in% 1] - 67.71
resid2 <- Handwriting$Survey2[!is.na(Handwriting$Survey2) & Handwriting$Gender %in% 0] - 61.75
resid <- c(resid1, resid2)
stripchart(resid, method="stack", offset=.5, at=0, pch=19, main="Residuals")
```



```r
qqnorm(resid, ylab="Residuals", xlab="z-score", main="Normal Probability Plot of Residuals")
qqline(resid)
```



State H_0: $\mu_1 - \mu_2 = 0$
H_a: $\mu_1 - \mu_2 < 0$
Plan The residuals look normal, the survey was done with random sampling, and we assume independence.

```r
men <- Handwriting$Survey2[!is.na(Handwriting$Survey2) & Handwriting$Gender%in%0]
wom <- Handwriting$Survey2[!is.na(Handwriting$Survey2) & Handwriting$Gender%in%1]
t.test(x=men,
       y=wom,
       alternative="less",
       mu=0)
```

```
##
##  Welch Two Sample t-test
##
## data:  men and wom
## t = -4.2057, df = 152.27, p-value = 2.213e-05
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -3.610871
## sample estimates:
## mean of x mean of y
##  61.75281  67.70642
```

d) Since the p-value is approximately 0, we reject the null hypothesis. There is a difference in accuracies in Survey 2 between men and women.