Stat 342: HW 3 Key

Michael Pearce and Ranjini Grove

Spring 2019

1. We are told that the lethal dose for 60% of the rats is $x = 1$. In other words

$$P(X \leq 1) = 0.60$$

From this we can conclude that

$$\begin{aligned} P(\ln(X) \leq \ln(1)) &= 0.60, \\ P(Y \leq 0) &= 0.60, \\ P(Z \leq (0-\mu)/\sigma) &= 0.60. \end{aligned}$$

In other words $-\mu/\sigma = \Phi^{-1}(0.6)$ or $\mu = -0.2533 \times \sigma$.

Similarly, since the lethal dose for 95% of the rats is $x = 4$, we have:

$$\begin{aligned} P(X \leq 4) &= 0.95, \\ P(Y \leq ln(4)) &= 0.95, \\ P(Z \leq (\ln(4)-\mu)/\sigma) &= 0.95. \end{aligned}$$

In other words, $(\ln(4) - \mu)/\sigma = 1.645$ or $\mu = 1.386 - 1.645 \times \sigma$.

We can solve the two equations simultaneously to get $\sigma = 0.996$ and $\mu = -0.252$.

If the manufacturer's specifications are correct, then at dose $x = 2$ the % of rats which should die is

$$\begin{aligned} P(X \leq 2) &= P(Y \leq \ln(2)), \\ &= P(Z \leq (\ln(2)-\mu)/\sigma), \\ &= P(Z \leq 0.949), \\ &= 0.829. \end{aligned}$$

Therefore, if the manufacturer's specs are correct, the number of rats out of 150 which die (call this X) should be a binomial random variable with parameters $n = 150$ and $p = 82.9\%$.

We wish to test the null $H_0 : p \geq 0.885$ against the alternative $H_1 : p < 0.885$. The size 0.05 LRT (based on the normal approximation) says we will reject H_0 *iff*

$$X \leq 150 \times 82.9\% - 1.645\sqrt{150 \times 82.9\% \times 17.1\%}.$$

The cut-off for the size 0.05 test is 119.7. Since our observed X is 95, we find the evidence supports the alternative. The poison is not as effective as the manufacturer claims.

2. a. In HW2, we found the rejection rule to be

$$\sum x_i^2 \leq \sigma_0^2 \times \chi^2_{1-\alpha,n}$$

so the acceptance region is

$$A(\sigma_0) = \{\mathbf{x} : \sum x_i^2 \geq \sigma_0^2 \times \chi^2_{1-\alpha,n}\}$$

and the upper confidence bound for σ is

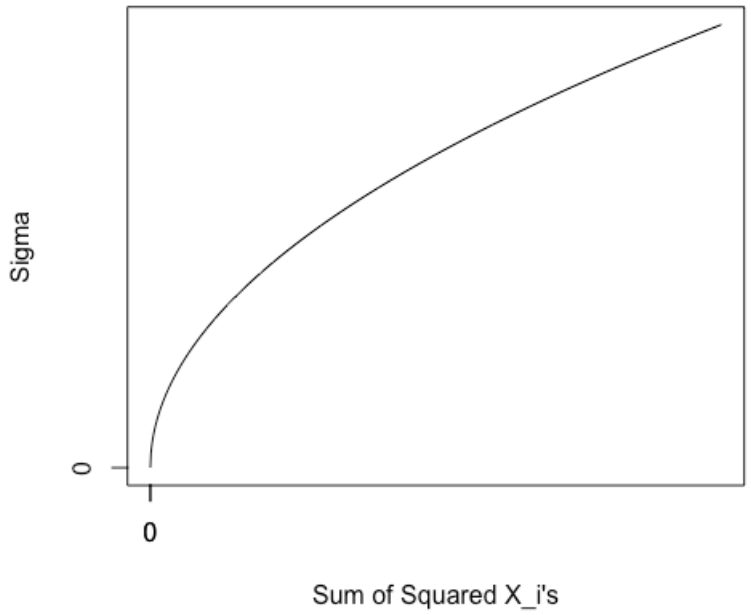
$$I(\mathbf{x}) = \{\sigma : \sigma \leq \sqrt{\frac{\sum x_i^2}{\chi^2_{1-\alpha,n}}}\}$$

Thus,

$$\{\sigma : \sum x_i^2 \geq \sigma^2 \times \chi^2_{1-\alpha,n}\}$$

is a $100(1-\alpha)\%$ confidence set for σ.

b. See the sketch below:



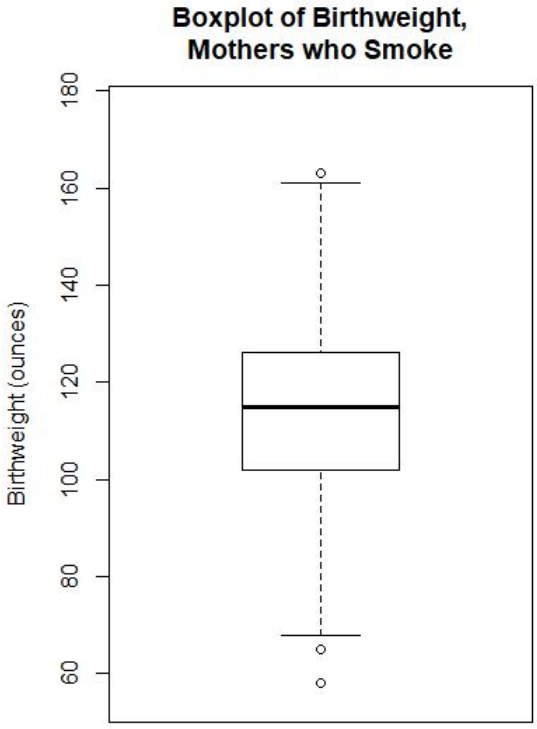
In this plot the line represents $\sigma = \sqrt{\frac{\sum x_i^2}{\chi^2_{1-\alpha,n}}}$ (note that the x and y axis values must be positive). Values below the line will be part of the acceptance region for σ.
See code used to create this plot below:

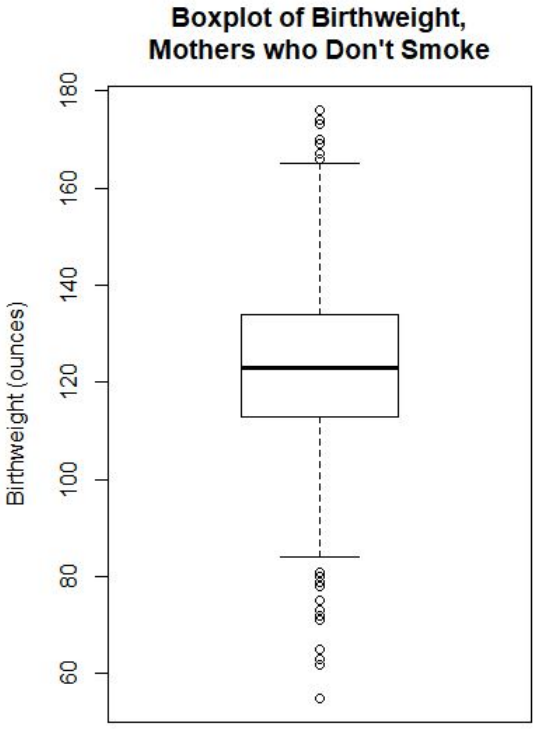
```
 sumxsquared<-seq(0,10,by=0.01)
sigma<-sqrt(sumxsquared/5)

plot(sumxsquared,sigma,type="l",
     xaxt='n',yaxt="n",
     xlab="Sum of Squared X_i's",
     ylab="Sigma")
at1 <- 0
axis(side =1, at1, labels = T)
axis(side =2, at1, labels = T)
```

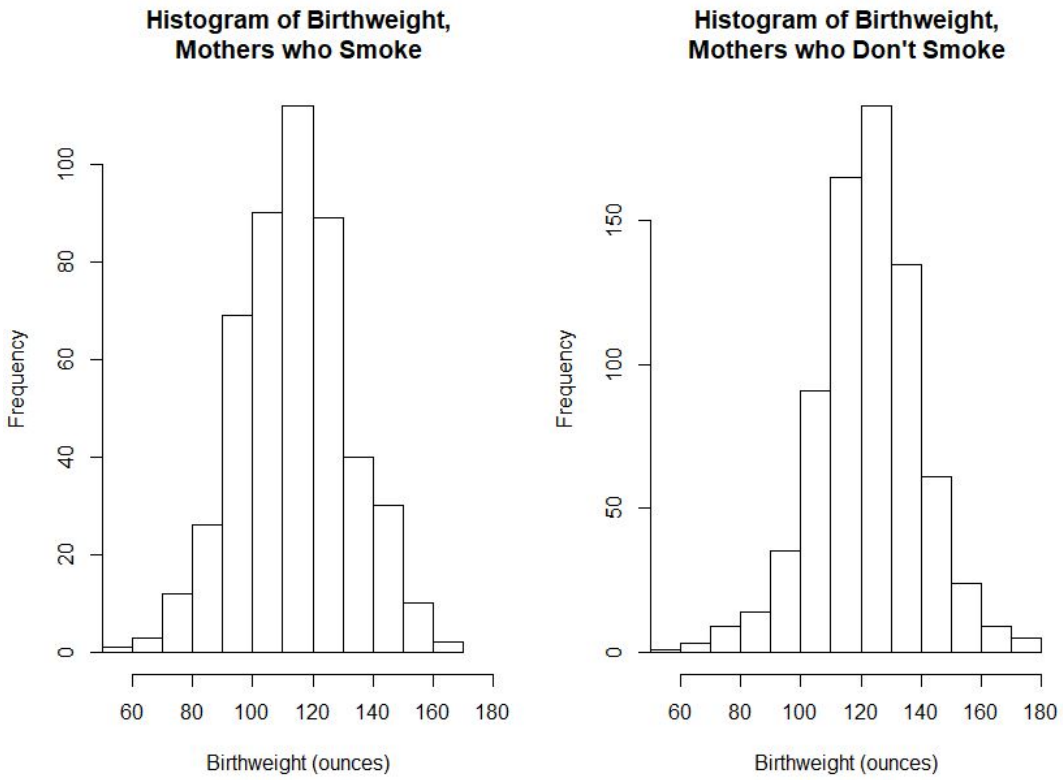
3. a. See the following code, with results below:

```
## Load the data
babiesI <- read.csv("C:/Users/mpp790/Downloads/babiesI.data", sep="")

## Exclude mothers with unknown smoking status
babiesI <- subset(babiesI, babiesI$smoke!=9)

## Exploratory Data Analysis
babiesI_Smoke <- subset(babiesI, babiesI$smoke==1)
babiesI_NoSmoke <- subset(babiesI, babiesI$smoke==0)

# overall summary statistics
table(babiesI$smoke)
summary(babiesI$bwt)
sd(babiesI$bwt)

# summary statistics for each group
summary(babiesI_Smoke$bwt)
sd(babiesI_Smoke$bwt)

summary(babiesI_NoSmoke$bwt)
sd(babiesI_NoSmoke$bwt)

# boxplots
par(mfrow=c(1,2))
boxplot(babiesI_Smoke$bwt, ylim=c(55,176),
        main="Boxplot of Birthweight,\nMothers who Smoke",
        ylab="Birthweight (ounces)")
boxplot(babiesI_NoSmoke$bwt, ylim=c(55,176),
        main="Boxplot of Birthweight,\nMothers who Don't Smoke",
        ylab="Birthweight (ounces)")

# histograms
hist(babiesI_Smoke$bwt, xlim=c(55,176),
     main="Histogram of Birthweight,\nMothers who Smoke",
     xlab="Birthweight (ounces)")
hist(babiesI_NoSmoke$bwt, , xlim=c(55,176),
     main="Histogram of Birthweight,\nMothers who Don't Smoke",
     xlab="Birthweight (ounces)")
```

```
> table(babiesI$smoke)

  0   1
742 484
> summary(babiesI$bwt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   55.0   109.0   120.0   119.5   131.0   176.0
> sd(babiesI$bwt)
[1] 18.20357
>
> # summary statistics for each group
> summary(babiesI_Smoke$bwt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   58.0   102.0   115.0   114.1   126.0   163.0
> sd(babiesI_Smoke$bwt)
[1] 18.09895
>
> summary(babiesI_NoSmoke$bwt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     55     113     123     123     134     176
> sd(babiesI_NoSmoke$bwt)
[1] 17.39869
```







Our dataset contains 742 mothers who didn't smoke during pregnancy and 484 who did. For the former group, the mean baby birth weight was 123 ounces, with a standard deviation of 17.4 and an IQR of 21 ounces. For the latter group, the mean baby birth weight was 114.1 ounces, with a standard deviation of 18.1 and an IQR of 24 ounces.

In the boxplots, we can see that there are likely outliers in each group (which appear as dots). There are two low outliers and one high outlier for the mothers that smoked, but numerous low and high outliers for the mothers that didn't smoke. Outliers may occur for a variety of reasons: Low outliers may be due to premature births; high outliers may be due maternal diabetes or genetic factors. These factors may be correlated with smoking or not, so we should not exclude the outliers in our analysis.

b. I used a difference in means t-test for unequal size and unequal variance. Let μ_S be the population mean birth weight for mothers who smoke, and μ_N be the population mean birth weight for mothers who don't smoke. We will test $H_0 : \mu_S = \mu_N$ vs $H_1 : \mu_S < \mu_N$.

Let $\bar{S}$ be the sample mean birth weight for smoking mothers, $\bar{N}$ be the sample mean birth weight for non-smoking mothers, with sample sizes n_S and n_N, respectively. Then, the sample statistic is

$$T = \frac{\bar{S} - \bar{N}}{\sqrt{\frac{s_S^2}{n_S} + \frac{s_N^2}{n_N}}} = -8.58$$

and the degrees of freedom of the t-distribution is

$$df = \frac{(\frac{s_S^2}{n_S} + \frac{s_N^2}{n_S})^2}{\frac{(s_S^2/n_S)^2}{n_S - 1} + \frac{(s_N^2/n_N)^2}{n_N - 1}} = 1003.2$$

We will reject when $T \leq t_{df,\alpha} = -1.64$. See results from the analysis in R below:

```
        Welch Two Sample t-test

data:  babiesI_Smoke$bwt and babiesI_NoSmoke$bwt
t = -8.5813, df = 1003.2, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -7.222928
sample estimates:
mean of x mean of y
 114.1095  123.0472
```

With a p-value lower than 0.01, we reject the null hypothesis and find that the evidence is consistent with the hypothesis that mean birth weights are lower babies born to smoking mothers than non-smoking mothers.

See code:

```
# ttest
t<-(mean(babiesI_Smoke$bwt)-mean(babiesI_NoSmoke$bwt))/
     sqrt(var(babiesI_Smoke$bwt)/length(babiesI_Smoke$bwt)+
          var(babiesI_NoSmoke$bwt)/length(babiesI_NoSmoke$bwt))
df=(var(babiesI_Smoke$bwt)/length(babiesI_Smoke$bwt)+
       var(babiesI_NoSmoke$bwt)/length(babiesI_NoSmoke$bwt))^2/
  ((var(babiesI_Smoke$bwt)/length(babiesI_Smoke$bwt))^2/(length(babiesI_Smoke$bwt)-1)+
  (var(babiesI_NoSmoke$bwt)/length(babiesI_NoSmoke$bwt))^2/(length(babiesI_NoSmoke$bwt)-1))
pvalue<-pt(t,df)
print(t)
print(df)
print(pvalue)

t.test(babiesI_Smoke$bwt,babiesI_NoSmoke$bwt,
       alternative="less",
       conf.level=0.95)
```

4. Since sample sizes in the two groups are small, it makes sense to calculate an exact p-value using Fisher's exact test approach to compare the probability of "waiting alone" in the groups. Let p_1 denote the probability of waiting alone in the low anxiety group and p_2 in the high anxiety group.

   Let X denote the number who choose to wait alone in the low anxiety group and Y the corresponding number in the high anxiety group. If X is high conditional on $X+Y$, it will support the alternative hypothesis that $p_1 > p_2$. The conditional random variable $X|(X+Y)$ is therefore a reasonable test statistic to calculate a p-value from.

   This conditional distribution $X|X+Y=14$ is a hypergeometric distribution:

$$P(X=x|X+Y=14) \quad = \quad \frac{\binom{14}{x} \times \binom{16}{13-x}}{\binom{30}{13}}, x = 0, 1, 2, \ldots, 13.$$

   (Recall: the classic hypergeometric experiment states the urn has 30 marbles, 14 of which are "blue" (or choose to wait alone). We draw 13 from the urn randomly and count X – the number of "blue" in the sample drawn)

   We observe $X = 9$. Therefore the p value is

$$p - value \quad = \quad \sum_{x=9}^{14} \frac{\binom{14}{x} \times \binom{16}{13-x}}{\binom{30}{13}}.$$

   The exact p value is 0.035. Assuming we are willing to accept a max type 1 error of $\alpha = 0.05$, this data provides significant evidence that those with low anxiety are more likely to wait alone.

5. The theorem says that the null distribution of the p-value (call it $V(\mathbf{X})$)

$$V(\mathbf{X}) = P_{\theta_0}(T(\mathbf{X}) \geq T(\mathbf{x})$$

has a uniform distribution on $(0, 1)$. First note

$$\begin{aligned} V(\mathbf{X}) &= 1 - P_{\theta_0}(T(\mathbf{X}) < T(\mathbf{x})), \\ &= 1 - F_{\theta_0}(T(\mathbf{x})), \end{aligned}$$

where F_{θ_0} is the CDF of $T(\mathbf{X})$ evaluated using $\theta = \theta_0$.

The CDF of $V(\mathbf{X})$ is: for $-\infty < a < \infty$

$$P_{\theta_0}(V(\mathbf{X}) \leq a) = \begin{cases} 0 & a < 0 \\ P_{\theta_0}(1 - F_{\theta_0}(T(\mathbf{X}) \leq a)) & 0 < a < 1 \\ 1 & a \geq 1. \end{cases}$$

We can focus on the part where $0 < a < 1$ to get:

$$\begin{aligned} P_{\theta_0}(V(\mathbf{X}) \leq a) &= P_{\theta_0}(F_{\theta_0}(T(\mathbf{X})) \geq 1 - a), \\ &= P_{\theta_0}(T(\mathbf{X}) \geq F_{\theta_0}^{-1}(1 - a)), \\ &= 1 - F_{\theta_0}(F_{\theta_0}^{-1}(1 - a)), \\ &= 1 - (1 - a), \\ &= a. \end{aligned}$$

Differentiating the CDF above gives the density function

$$\begin{aligned} f_{\theta_0}(a) &= \frac{\partial}{\partial a} a, \\ &= 1, \quad 0 < a < 1. \end{aligned}$$