STAT 421 Homework 3

Lecture 6 Problem 1:
If $\sum_i^n |y_i - \bar{y}|$ had a χ^2 distribution, the degrees of freedom of the chi-squared distribution would be $n-1$. This is because, similar to $\sum_i^n (y_i - \bar{y})^2$, the n terms $y_1 - \bar{y},\ y_2 - \bar{y}, \ldots,\ y_n - \bar{y}$ are subject to one constraint: $\sum_i^n (y_i - \bar{y}) = 0$, and hence have $n-1$ independent terms.

Lecture 6 Problem 2a:
Given that the population variance of some quantity is 1, and that a random sample of size 31 was taken, the probability that the sample variance of the aforementioned quantity will be at most 1.459 can be computed through the use of the chi-squared distribution.

If we let Y denote the unknown quantity, then from the problem description we know that $V[Y] = \sigma_Y^2 = 1$. Furthermore, given that $n = 31$, we are trying to solve $P\left(S_Y^2 \leq 1.459\right)$, where S_Y^2 is the sample variance of the unknown quantity Y. We will manipulate the inside of the previous probability statement in order to get it into the form of something we know how to solve.

$$\begin{aligned} P\left(S_Y^2 \leq 1.459\right) &= P\left((n-1)S_Y^2 \leq (n-1)1.459\right) \\ &= P\left(\frac{(n-1)S_Y^2}{\sigma_Y^2} \leq \frac{(n-1)1.459}{\sigma_Y^2}\right) \end{aligned}$$

If we let $X^2 = \frac{(n-1)S_Y^2}{\sigma_Y^2}$, then from the derivations of lecture 5, it follows that $X^2 \sim \chi^2_{n-1}$. Since $\sigma_Y^2 = 1$ and $n = 31$, it follows that $X^2 \sim \chi^2_{30}$ and hence we can compute the preceding probability.

$$\begin{aligned} P\left(X^2 \leq \frac{30 \cdot 1.459}{1}\right) &= P\left(X^2 \leq 43.77\right) \\ &= 0.9499692 \end{aligned}$$

As computed using R and the chi-square table, the probability that the sample variance of the aforementioned quantity will be at most 1.459 is $0.9499692 \approx 0.95$. The R command for computing this probability is shown below.

```
prob_chi <- pchisq(43.77, 30)
```

Lecture 6 Problem 2b:
If we were to actually observe the sample variance and find it to be smaller than 1.459, we would be able to say that our assumption of $\sigma_Y^2 = 1$ is plausible. We would be able to say that this population variance is a plausible value because, as found in the previous part, assuming that $\sigma_Y^2 = 1$, we see that $P\left(S_Y^2 \leq 1.459\right) \approx 0.95$. Hence, under the assumption that the population variance is one, there is a high probability that the sample variance is at most 1.459. Furthermore, now that we have actually found a sample variance, and it was in fact less than or equal to 1.459, we have no evidence against our claim/assumption that $\sigma_Y^2 = 1$. This doesn't prove that $\sigma_Y^2 = 1$, it just shows that given our assumption that $\sigma_Y^2 = 1$, we have no convincing evidence that $\sigma_Y^2 \neq 1$, and hence it is a plausible value.

Lecture 6 Problem 3:
In this problem, we are running an experiment with a binary factor X, and a response Y such that: $V[Y|X = 1] = V[Y|X = 2]$. Our main goal is to find $P\left(S_1^2 > 2.33 S_2^2\right)$, where S_1^2 and S_2^2 are the sample variances of Y

at $X = 1$ and $X = 2$, respectively, based on sample sizes $n_1 = 21$ and $n_2 = 25$. To do this we will manipulate the inside of the aforementioned probability statement and use our knowledge of the F-distribution.

$$\begin{aligned} P\left(S_1^2 > 2.33 S_2^2\right) &= P\left(\frac{S_1^2}{S_2^2} > 2.33\right) \\ &= P\left(\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} > 2.33\frac{\sigma_2^2}{\sigma_1^2}\right) \end{aligned}$$

If we let $F_0 = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$, then from the derivations of lecture 6, it follows that $F_0 \sim F_{n_1-1,\ n_2-1}$. Since $n_1 = 21$ and $n_2 = 25$, it follows that $F_0 \sim F_{20,\ 24}$ and hence we can compute the preceding probability.

$$\begin{aligned} P\left(F_0 > 2.33\frac{\sigma_2^2}{\sigma_1^2}\right) &= P(F_0 > 2.33) \\ &= 0.02484508 \end{aligned}$$

As computed using R and the F-table, $P\left(S_1^2 > 2.33 S_2^2\right) = 0.02484508 \approx 0.025$. The R command for computing this probability is shown below.

```
prob_f <- pf(2.33, 20, 24, lower.tail=F)
```

Lecture 6 Problem 4:
Given that for a certain random variable X, it is known that $\log(\bar{X}) \sim N(\mu, 1)$, in order to find the p-value for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \square \mu_0$, we must make an assumption and standardize twice. The assumption we have to make is that the X_i's that make up $\bar{X}$ are limited to the following constraint: $X_i > 0$. This assumption needs to be made in order for $\log(\bar{X}) \sim N(\mu, 1)$ to make sense. We can now form an expression for the p-value of this test.

$$\begin{aligned} \text{p-value} &= P(\bar{X} \square \bar{X}_{obs}) \\ &= P(\log(\bar{X}) \square \log\left(\bar{X}_{obs}\right)) \\ &= P(\log(\bar{X}) - \mu \square \log\left(\bar{X}_{obs}\right) - \mu) \\ &= P\left(\frac{\log(\bar{X}) - \mu}{1} \square \frac{\log\left(\bar{X}_{obs}\right) - \mu}{1}\right) \end{aligned}$$

If we let $Z = \frac{\log(\bar{X})-\mu}{1}$, then it follows from the derivations of lecture 4 that $Z \sim N(0, 1)$. Furthermore, under the assumption that H_0 is true, it follows that $\mu = \mu_0$. Hence we can see that, p-value $= P\left(Z \square \log\left(\bar{X}_{obs}\right) - \mu_0\right)$, where Z is a random variable following the standard normal distribution.

Lecture 7 Problem 1a:
In this problem we are testing $H_0 : \mu = 3$ versus $H_1 : \mu > 3$ at $\alpha = 0.05$, where μ is a population mean and the sample variance from a sample of size 21 is found to be 2. With that said, our goal is to find the rejection region of sample mean values. As discussed in lecture 6 and 7, under the assumption that H_0 is true, the statistic $t = \frac{\hat{Y}_{obs}-\mu_0}{S_{obs}/\sqrt{n}} \sim t_{n-1}$, where S_{obs} is the observed sample standard deviation. Furthermore, as discussed in lecture 7, the rejection region for t in the direction of H_1 is $t > t_\alpha$, where t_α is the α percentage point of the t distribution with $n - 1$ degrees of freedom. Thus, in order to find the rejection region for the sample means, we must solve for $\bar{Y}_{obs}$ in the expression $\frac{\hat{Y}_{obs}-\mu_0}{S_{obs}/\sqrt{n}} > t_\alpha$. Since $S_{obs} = \sqrt{2}$, $\mu_0 = 3$, and $n = 21$, it follows from the t table that $t_\alpha = 1.725$. This value can also be computed using R as shown below.

```
alpha <- 0.05
rej_t <- qt(alpha, 20, lower.tail=F)
```

With that said, we can now compute the rejection region for the sample means as shown below.

$$\begin{aligned}
\frac{\hat{Y}_{obs} - \mu_0}{S_{obs}/\sqrt{n}} &> t_\alpha \\
\frac{\hat{Y}_{obs} - 3}{\sqrt{2/21}} &> 1.725 \\
\hat{Y}_{obs} - 3 &> 1.725\sqrt{2/21} \\
\hat{Y}_{obs} &> 1.725\sqrt{2/21} + 3 \\
\hat{Y}_{obs} &> 3.53226
\end{aligned}$$

As computed above, the rejection region for the sample means is $[3.53226, \infty)$.

Lecture 7 Problem 1b:
As computed in the previous problem, the rejection region for the sample means is $[3.53226, \infty)$. Hence if we observe a sample mean of 2.5, which is not in the aforementioned rejection region, we fail to reject the null hypothesis that $\mu = 3$, therefore we have not found significant evidence in favor of the alternative which is $\mu > 3$.

Lecture 7 Problem 1c:
If we observe a sample mean of 2.5, then the p-value, which is defined as the probability of seeing a value of the test statistic more extreme than the one observed, is calculated as $P(t > t_{obs}|\mu_0 = 3)$. Using the same information and t-statistic from part a, we can compute the p-value as shown below.

$$\begin{aligned}
P(t > t_{obs}|\mu_0 = 3) &= P\left(t > \frac{\hat{Y}_{obs} - \mu_0}{S_{obs}/\sqrt{n}}\right) \\
&= P\left(t > \frac{2.5 - 3}{\sqrt{2/21}}\right) \\
&= P\left(t > -1.62019\right) \\
&= 0.93957324
\end{aligned}$$

As computed above, the p-value is 0.93957324. If we were to use the t-table we would see that the p-value would be in the range from 0.9 to 0.95, however, using R pinpoints this exact value. This value is calculated in R as follows.

```
p_val_t <- pt((2.5-3)/sqrt(2/21), 20, lower.tail=F)
```

Lecture 7 Problem 1d:
As computed in the previous problem the p-value for this test is 0.93957 (or $[0.90, 0.95]$ if the t-table is used). Either way, at the significance level $\alpha = 0.05$, since the p-value is greater than the significance level we fail to reject the null hypothesis. Which makes sense as this was the same conclusion we got from part b and we are running the same test.

Lecture 7 Problem 2:
Suppose we obtain a 2-sided confidence interval for a single μ. One common mistaken interpretation of this confidence interval is that it includes a random sample mean with probability $1-\alpha$. The goal of this problem is to find what the observed sample mean would have to be in order for the mistaken interpretation to be true. The observed confidence interval for μ, as computed in lecture 7, is $\bar{Y}_{obs} \pm t_{\frac{\alpha}{2},\, n-1}\frac{S_{obs}}{\sqrt{n}}$. On the other hand, written in probability statement form, the misinterpreted confidence interval is written as

$$P\left(\bar{Y}_{obs} - t_{\frac{\alpha}{2},\, n-1}\frac{S_{obs}}{\sqrt{n}} < \bar{Y} < \bar{Y}_{obs} + t_{\frac{\alpha}{2},\, n-1}\frac{S_{obs}}{\sqrt{n}}\right) = 1 - \alpha$$

If we standardize by subtracting throughout the probability statement by μ_Y, the mean of the distribution of Y, and divide throughout the probability statement by S, the random sample standard deviation, we obtain the following probability statement

$$P\left(\frac{\bar{Y}_{obs} - \mu_Y - t_{\frac{\alpha}{2},\ n-1}\frac{S_{obs}}{\sqrt{n}}}{S/\sqrt{n}} < \frac{\bar{Y} - \mu_Y}{S/\sqrt{n}} < \frac{\bar{Y}_{obs} - \mu_Y + t_{\frac{\alpha}{2},\ n-1}\frac{S_{obs}}{\sqrt{n}}}{S/\sqrt{n}}\right) = 1 - \alpha$$

If we split up the fraction on both sides of the probability statement we obtain the following

$$P\left(\frac{\bar{Y}_{obs} - \mu_Y}{S/\sqrt{n}} - t_{\frac{\alpha}{2},\ n-1}\frac{S_{obs}}{S} < \frac{\bar{Y} - \mu_Y}{S/\sqrt{n}} < \frac{\bar{Y}_{obs} - \mu_Y}{S/\sqrt{n}} + t_{\frac{\alpha}{2},\ n-1}\frac{S_{obs}}{S}\right) = 1 - \alpha$$

If we assume that $S = S_{obs}$, which is impossible since one is random and one is not, then we start making progress as we see that

$$P\left(\frac{\bar{Y}_{obs} - \mu_Y}{S/\sqrt{n}} - t_{\frac{\alpha}{2},\ n-1} < \frac{\bar{Y} - \mu_Y}{S/\sqrt{n}} < \frac{\bar{Y}_{obs} - \mu_Y}{S/\sqrt{n}} + t_{\frac{\alpha}{2},\ n-1}\right) = 1 - \alpha$$

is identical to $P\left(-t_{\frac{\alpha}{2},\ n-1} < t < t_{\frac{\alpha}{2},\ n-1}\right) = 1 - \alpha$ if we choose $\bar{y}_{obs}$ such that $\frac{\bar{Y}_{obs} - \mu_Y}{S/\sqrt{n}} = 0$. Hence by choosing $\bar{y}_{obs} = \mu_Y$, we make the misconception of confidence intervals true.

Alternatively, if we were to assume that σ_Y^2 was known, then we could use z instead of t and end up with the same final result. In this case we wouldn't have to assume that $S = S_{obs}$, which is impossible since one is random and one is not.

Lecture 7 Problem 3:
Given that $t \equiv \frac{(\bar{Y}-\mu)(\sigma/\sqrt{n})}{\sqrt{X^2/(n-1)}} \sim t_{n-1}$, $X^2 \equiv (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, $T \equiv \frac{(\bar{Y}-\mu)(\sigma/\sqrt{n})+\Delta}{\sqrt{X^2/(n-1)}}$ has a distribution called the non-central t-distribution with parameters $n-1$ and Δ, and that the qth quantile of $N(\mu, \sigma)$ is $\mu + z_q\sigma$, where z_q is the qth quantile of the standard normal, in this problem we are looking to show that the $100(1-\alpha)\%$ confidence interval for the qth quantile of $N(\mu, \sigma)$ is

$$\left[\bar{Y} - T_{\frac{\alpha}{2},\ n-1,\ -z_q\sqrt{n}}\frac{S}{\sqrt{n}},\ \bar{Y} - T_{1-\frac{\alpha}{2},\ n-1,\ -z_q\sqrt{n}}\frac{S}{\sqrt{n}}\right]$$

To start off this problem, we know that in the end we are going to want to have $\mu + z_q\sigma$ in the center of a probability statement, thus we must manipulate the definition of T in order to get $\mu + z_q\sigma$ to appear somewhere in the definition. From there we can then isolate the expression in the center of a probability statement we know. First off we know from the definition of T that

$$T = \frac{\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} + \Delta}{\sqrt{\frac{X^2}{n-1}}} \sim T_{\frac{\alpha}{2},\ n-1,\ \Delta}$$

The first step in getting $\mu + z_q\sigma$ to show up in T is to give Δ a common denominator as the fraction of the left. This is done below.

$$T = \frac{\bar{Y} - \mu + \Delta\frac{\sigma}{\sqrt{n}}}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{X^2}{n-1}}} \sim T_{\frac{\alpha}{2},\ n-1,\ \Delta}$$

Furthermore, since $X^2 = (n-1)S^2/\sigma^2$, this implies that $\sqrt{\frac{X^2}{n-1}} = \frac{S}{\sigma}$, it follows that the denominator in the above expression can simplify greatly, as shown below.

$$T = \frac{\bar{Y} - \mu + \Delta\frac{\sigma}{\sqrt{n}}}{\frac{S}{\sqrt{n}}} \sim T_{\frac{\alpha}{2},\ n-1,\ \Delta}$$

Notice that in the above expression we already have many of the key pieces that show up in the solution, furthermore, if we follow the hint and cleverly choose $\Delta = -z_q\sqrt{n}$, we obtain something that is very close to what we are looking for.

$$T = \frac{\bar{Y} - \mu - z_q\sigma}{\frac{S}{\sqrt{n}}} \sim T_{\frac{\alpha}{2},\ n-1,\ -z_q\sqrt{n}}$$

Finally, we can isolate $\mu + z_q\sigma$ in the above expression to obtain.

$$T = \frac{\bar{Y} - (\mu + z_q\sigma)}{\frac{S}{\sqrt{n}}} \sim T_{\frac{\alpha}{2},\ n-1,\ -z_q\sqrt{n}}$$

Now that we have a value of T that we can work with, we will start with a probability statement that we know and work to isolate $\mu + z_q\sigma$ in the center of the probability statement. This is done below.

$$\begin{aligned}
1-\alpha &= P\left(T_{1-\frac{\alpha}{2},\ n-1,\ -z_q\sqrt{n}} < T < T_{\frac{\alpha}{2},\ n-1,\ -z_q\sqrt{n}}\right) \\
&= P\left(T_{1-\frac{\alpha}{2},\ n-1,\ -z_q\sqrt{n}} < \frac{\bar{Y} - (\mu + z_q\sigma)}{\frac{S}{\sqrt{n}}} < T_{\frac{\alpha}{2},\ n-1,\ -z_q\sqrt{n}}\right) \\
&= P\left(T_{1-\frac{\alpha}{2},\ n-1,\ -z_q\sqrt{n}}\frac{S}{\sqrt{n}} < \bar{Y} - (\mu + z_q\sigma) < T_{\frac{\alpha}{2},\ n-1,\ -z_q\sqrt{n}}\frac{S}{\sqrt{n}}\right) \\
&= P\left(-\bar{Y} + T_{1-\frac{\alpha}{2},\ n-1,\ -z_q\sqrt{n}}\frac{S}{\sqrt{n}} < -(\mu + z_q\sigma) < -\bar{Y} + T_{\frac{\alpha}{2},\ n-1,\ -z_q\sqrt{n}}\frac{S}{\sqrt{n}}\right) \\
&= P\left(\bar{Y} - T_{\frac{\alpha}{2},\ n-1,\ -z_q\sqrt{n}}\frac{S}{\sqrt{n}} < \mu + z_q\sigma < \bar{Y} - T_{1-\frac{\alpha}{2},\ n-1,\ -z_q\sqrt{n}}\frac{S}{\sqrt{n}}\right)
\end{aligned}$$

Hence, as seen from above, we have shown that the $100(1-\alpha)\%$ confidence interval for the qth quantile of $N(\mu, \sigma)$ is $\left[\bar{Y} - T_{\frac{\alpha}{2},\ n-1,\ -z_q\sqrt{n}}\frac{S}{\sqrt{n}},\ \bar{Y} - T_{1-\frac{\alpha}{2},\ n-1,\ -z_q\sqrt{n}}\frac{S}{\sqrt{n}}\right]$.

Lecture 8 Problem 1a:
In this problem we are using the data from exercise 2.26, which are shown below.

Type 1		Type 2	
65	82	64	56
81	67	71	69
57	59	83	74
66	75	59	82
82	70	65	79

Furthermore, at $\alpha = 0.05$, and assuming that the population variances are equal, we will test the hypothesis that the mean burning times are different using the rejection region, p-value, and confidence interval.

Since we are assuming equal population variances (i.e $\sigma_1^2 = \sigma_2^2 = \sigma^2$), and these population variances are unknown, it turns out that we are running a 2-sample t-test for the differences between two means. In this case our test statistic is $t = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$, where $S_p = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$. Which, under the assumption that $H_0 : \mu_1 - \mu_2 = 0$ is true, takes the form $t = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$. Our first method of hypothesis testing is the rejection region method. In our case, the rejection region is defined as $|t| > t_{\frac{\alpha}{2},\ n_1+n_2-2}$, where $t_{\frac{\alpha}{2},\ n_1+n_2-2}$ is the $\alpha/2$ percentage point of the t-distribution with $n_1 + n_2 - 2$ degrees of freedom, and $|t|$ is the absolute value of the t-statistic. The calculation of this rejection region is calculated in R below.

```r
# Set significance level:
alpha <- 0.05

# Create the two samples:
type_1 <- c(65, 81, 57, 66, 82, 82, 67, 59, 75, 70)
type_2 <- c(64, 71, 83, 59, 65, 56, 69, 74, 82, 79)

# Find sample information:
n1 <- length(type_1)
n2 <- length(type_2)

ybar_1 <- mean(type_1)
ybar_2 <- mean(type_2)

var_1 <- var(type_1)
var_2 <- var(type_2)

sp <- sqrt(((n1-1)*var_1 + (n2-1)*var_2)/(n1+n2-2))

# Find the rejection region value:
t_crit_up <- qt(alpha/2, n1+n2-2, lower.tail=F)
t_crit_low <- qt(1-(alpha/2), n1+n2-2, lower.tail=F)

# Find the observed test statistic value:
t_obs <- (ybar_1-ybar_2) / (sp*sqrt((1/n1)+(1/n2)))
```

As computed above, the rejection region for this 2-sample t-test is $|t| > 2.100922$. This implies that the rejection region is actually $t < -2.101 \cup t > 2.101$. Furthermore, as computed above, our observed t-statistic value was 0.0480077. Since our observed t-statistic value was outside of this rejection region, we fail to reject the null hypothesis in favor of the alternative. Hence, there is no evidence that $\mu_1 \neq \mu_2$.

We will now do the same analysis, but by calculating a p-value. In this case, since we are testing a two-sided alternative, we have to consider values more extreme than the one observed on both sides of the distribution. However, since the t-distribution is symmetric, we can simply multiply the probability of getting a t-statistic value greater than the one we observed, by two. This calculation is done in R below.

```r
# Find the p-value:
p_val_t <- 2*pt(t_obs, n1+n2-2, lower.tail = F)
```

As calculated above our p-value for this test is 0.9622388. Since $0.9622388 > 0.05$, it follows that we fail to reject the null hypothesis in favor of the alternative. Hence, there is no evidence that $\mu_1 \neq \mu_2$. This is the same conclusion as the rejection region method.

Lastly, we will create confidence intervals for the difference between the two means; $\mu_1 - \mu_2$. As calculated in lecture 8, the $100(1-\alpha)\%$ confidence interval for the difference between two means when the variances are equal is: $(\bar{Y}_1 - \bar{Y}_2) \pm t_{\frac{\alpha}{2},\, n_1+n_2-2} \frac{S_p}{\sqrt{1/n_1+1/n_2}}$. We will compute this interval in R as shown below.

```r
# Find the confidence interval:
ci_upper_1 <- (ybar_1 - ybar_2) + t_crit_up * (sp*sqrt((1/n1)+(1/n2)))
ci_lower_1 <- (ybar_1 - ybar_2) + t_crit_low * (sp*sqrt((1/n1)+(1/n2)))
```

As shown above the 95% confidence interval for the difference between the two means μ_1 and μ_2, assuming equal population variances, was calculated as [-8.5524411, 8.9524411]. Since this interval contains zero, it

follows that we fail to reject the null hypothesis in favor of the alternative. Hence, there is no evidence that $\mu_1 \neq \mu_2$. Again, this is the same conclusion as the rejection region method, and the p-value method.

Lecture 8 Problem 1b:
In this problem, using the same data from part a, we will perform an F-test for the equality of the population variances using the rejection region, p-value, and confidence interval methods, respectively. In particular we are testing the hypotheses $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_1 : \sigma_1^2 \neq \sigma_2^2$. As shown in chapter 2.6 of the textbook, the test statistic for this kind of test is the ratio of the sample variances, in particular $F_0 = \frac{S_1^2}{S_2^2} \sim F_{n_1-1,\ n_2-1}$, where $F_{n_1-1,\ n_2-1}$ is the F-distribution with $n_1 - 1$ numerator and $n_2 - 1$ denominator degrees of freedom, respectively.

It was also shown in chapter 2.6 of the textbook that the rejection region is for this particular alternative hypothesis is $F_0 > F_{\frac{\alpha}{2},\ n_1-1,\ n_2-1}$ and $F_0 < F_{1-\frac{\alpha}{2},\ n_1-1,\ n_2-1}$, where $F_{\frac{\alpha}{2},\ n_1-1,\ n_2-1}$ and $F_{1-\frac{\alpha}{2},\ n_1-1,\ n_2-1}$ denote the upper $\alpha/2$ and lower $1 - \alpha/2$ percentage points of the F-distribution with with $n_1 - 1$ numerator and $n_2 - 1$ denominator degrees of freedom, respectively. These regions are calculated in R below.

```
# Find the rejection region value:
f_crit_up <- qf(alpha/2, n1-1, n2-1, lower.tail=F)
f_crit_low <- qf(1-(alpha/2), n1-1, n2-1, lower.tail=F)

# Find the observed test statistic value:
f_obs <- var_1 / var_2
```

As shown above the rejection region for this F-test is $F_0 > 4.0259942$ and $F_0 < 0.2483859$. Furthermore, since our observed F_0 test statistic was 0.9782168, which is outside of the rejection regions, we fail to reject the null hypothesis in favor of the alternative. Hence, there is no evidence that $\sigma_1^2 \neq \sigma_2^2$.

We will now do the same analysis, but by calculating a p-value. In this case, since we are testing a two-sided alternative, we have to consider values more extreme than the one observed on both sides of the distribution. However, since the F-distribution is not symmetric, we can't simply multiply the probability of getting a F-statistic value greater/less than the one we observed, by two. This calculation is done in R below. Instead, since the statistic $F_1 = \frac{S_2^2}{S_1^2}$ follows the same distribution as F_0, and we observed an F_0 value less than 1, our p-value is equal to the area to the left of our observed F_0 value, and to the right of our observed F_1 value. Since, $n_1 = n_2$ it follows that we could simply multiply the area to the left of our observed F_0 value by two, but this isn't always the case.

```
# Find the p-value:
p_val_f <- pf(f_obs, n1-1, n2-1) + pf(var_2/var_1, n2-1, n1-1, lower.tail=F)
```

As calculated above our p-value for this test is 0.9743665. Since $0.9743665 > 0.05$, it follows that we fail to reject the null hypothesis in favor of the alternative. Hence, there is no evidence that $\sigma_1^2 \neq \sigma_2^2$. This is the same conclusion as the rejection region method.

Lastly, we will create confidence intervals for the ratio of the population variances; σ_1^2/σ_2^2. As calculated in section 2.6 of the textbook, the $100(1-\alpha)\%$ confidence interval for the the ratio of the population variances is: $\frac{S_1^2}{S_2^2} F_{1-\frac{\alpha}{2},\ n_1-1,\ n_2-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2} F_{\frac{\alpha}{2},\ n_1-1,\ n_2-1}$. We will compute this interval in R as shown below.

```
# Find the confidence interval:
ci_upper_2 <- f_obs*f_crit_up
ci_lower_2 <- f_obs*f_crit_low
```

As shown above the 95% confidence interval for the the ratio of the population variances σ_1^2 and σ_2^2, was calculated as [0.2429752, 3.9382952]. Since this interval contains one, it follows that we fail to reject the null

hypothesis in favor of the alternative. Hence, there is no evidence that $\sigma_1^2 \neq \sigma_2^2$. Again, this is the same conclusion as the rejection region method, and the p-value method.

Lecture 8 Problem 2:
In lecture we showed that $SS_{Treatment} = n\sum_i^a(\bar{y}_{i.} - \bar{y}_{..})^2$, but in the book they write $SS_{Treatment} = \frac{1}{n}\sum_i^a y_{i.}^2 - \frac{1}{N}y_{..}^2$. In this problem, we will show that these two expressions are in fact the same.

$$\begin{aligned}
SS_{Treatment} &= n\sum_i^a(\bar{y}_{i.} - \bar{y}_{..})^2 \\
&= n\sum_i^a\left(\bar{y}_{i.}^2 + \bar{y}_{..}^2 - 2\bar{y}_{i.}\bar{y}_{..}\right) \\
&= n\sum_i^a \bar{y}_{i.}^2 + n\sum_i^a \bar{y}_{..}^2 - 2n\sum_i^a \bar{y}_{i.}\bar{y}_{..} \\
&= n\sum_i^a \left(\frac{y_{i.}}{n}\right)^2 + n\sum_i^a \left(\frac{y_{..}}{an}\right)^2 - 2n\sum_i^a \frac{1}{n}y_{i.}\frac{1}{N}y_{..} \\
&= n\sum_i^a \frac{y_{i.}^2}{n^2} + an\frac{y_{..}^2}{(an)^2} - 2\frac{1}{N}y_{..}\sum_i^a y_{i.} \\
&= \frac{1}{n}\sum_i^a y_{i.}^2 + \frac{1}{N}y_{..}^2 - 2\frac{1}{N}y_{..}y_{..} \quad (\text{Since } \sum_i^a y_{i.} = y_{..}) \\
&= \frac{1}{n}\sum_i^a y_{i.}^2 + \frac{1}{N}y_{..}^2 - \frac{2}{N}y_{..}^2 \\
&= \frac{1}{n}\sum_i^a y_{i.}^2 - \frac{1}{N}y_{..}^2
\end{aligned}$$

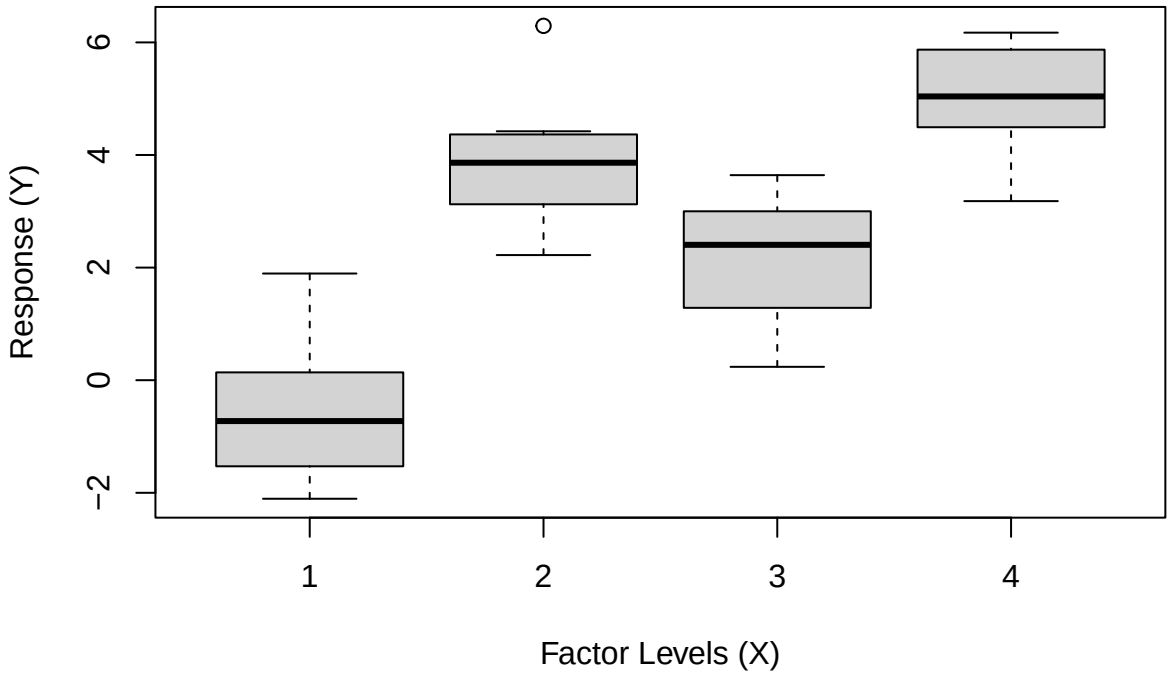
Thus we have shown that $SS_{Treatment} = n\sum_i^a(\bar{y}_{i.} - \bar{y}_{..})^2 = \frac{1}{n}\sum_i^a y_{i.}^2 - \frac{1}{N}y_{..}^2$.

Lecture 8 Problem 3a:
In this problem we are considering data on Y involving a factor X with 4 levels and 10 replications. The data for this experiment is shown below.

```
# Set up an empty matrix:
y = matrix(nrow=4,ncol=10)

# Add the data to the matrix:
y[1,] = c(-2.10552316, 1.89491371, -1.52919682, -0.99265143, -0.45911960,
          1.09271028, -1.54680778, 0.13890677, 0.06240357, -1.09273045)

y[2,] = c(4.25667943, 4.36518096, 4.42108835, 3.77229146, 2.22264903,
          3.95354759, 6.29377745, 3.58501081, 3.12457306, 3.04360597)

y[3,] = c(3.64209745, 2.76932242, 1.46001019, 0.23739519, 0.27629510,
          2.83897173, 2.99999590, 3.54657820, 2.03955378, 1.28515784)

y[4,] = c( 3.18088593, 5.44976665, 5.87116946, 4.01275036, 6.00826692,
           5.19220036, 6.17338313, 4.88846073, 4.49330445, 4.83224707)
```

Now that we have created the data, we will make a comparative boxplot of Y for the 4 levels of X in order to decided if we think X has an effect on Y.

```r
# Create comparative boxplot of Y for the 4 levels of X:
boxplot(t(y), xlab="Factor Levels (X)", ylab="Response (Y)")
```



Before analyzing the presence of treatment effects, it is important to notice that, based on the boxplots, all 4 levels of X seem to have similar spreads. Hence the big assumption that $\sigma_i^2 = \sigma_\epsilon^2$, $\forall i$, is a valid one. Furthermore, after analyzing the boxplots, it seems as if X has an effect on Y, this conclusion is made because it appears as if at least two different treatment factor levels come from a model/population with different μ_i values. For example, factor level 1 and factor level 2 have vastly different centers with no overlap at all. The same claim could be made between factor level 1 and factor level 2 as well.

Lecture 8 Problem 3b:
For each i, the quantity $\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$ is said to be an "estimate of an effect." In this problem we will compute these effect estimates in R.

```r
# Calculate the "effect estimates" for each factor level:
est_effects <- rowMeans(y) - mean(y)
```

Now that these effect estimates have been computed, we will show what the effect estimate is for each treatment factor level.

```r
# Show what the effect estimate is for each treatment factor level:
for (i in 1:4) {
  print(paste("The effect estimate for treatement level", i, "is", est_effects[i]))
}
```

```
## [1] "The effect estimate for treatement level 1 is -3.0961875425"
```

```
## [1] "The effect estimate for treatement level 2 is 1.2613623595"
## [1] "The effect estimate for treatement level 3 is -0.5329402715"
## [1] "The effect estimate for treatement level 4 is 2.3677654545"
```

Lecture 8 Problem 3c:
In this problem, using R for arithmetic, we will perform a 1-way ANOVA on X and Y for testing whether X has an effect on Y. We will use both the rejection region method and the p-value method. Before diving into the calculation in R, I will set up the theory of the test.

In lecture and in the reading we found 4 important results that are key to the 1-way ANOVA F-test, these results are:

$$SS_{Treatment} = n\sum_{i}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2$$

$$SS_E = (n-1)\sum_{i}^{a} S_i^2$$

$$MS_{Treatment} = \frac{SS_{Treatment}}{a-1}$$

$$MS_E = \frac{SS_E}{N-1}$$

Furthermore, under $H_0 : \tau_i = 0,\ \forall i$, it follows that $F \equiv \frac{MS_{Treatment}}{MS_E} \sim F_{a-1,\ N-1}$. Hence, in order to test if there exists a treatment effect from X onto Y, we must run an F-test. We will start by doing this F-test via the rejection region method, the above calculation will be done in R.

```r
# Set up information for 1-way ANOVA F-test:
alpha <- 0.05
n <- 10
a <- 4
N <- a*n

# Find the sample variance for each treatment factor level:
vars <- numeric(a)
for (i in 1:4) {
  vars[i] <- var(y[i,])
}

# Compute sums of squares and mean squares for the F-test:
SS_treat <- n*sum(est_effects^2)
SS_e <- (n-1)*sum(vars)
MS_treat <- SS_treat / (a-1)
MS_e <- SS_e / (N-a)

# Compute the observed F value:
F_obs <- MS_treat / MS_e

# Compute the rejection region of the test:
F_rej <- qf(alpha, a-1, N-a, lower.tail = F)
```

As shown in the lecture 8 derivations, it follows that we will reject H_0 in favor of H_1 iff $F > F_{\alpha,\ a-1,\ N-1}$, where F is our observed F-statistic value, and $F_{\alpha,\ a-1,\ N-1}$ is the α percentage point of an F-distribution with $a-1$ numerator and $N-1$ denominator degrees of freedom, respectively. In our case we found that

$F = 43.5461408$ and that our rejection region is any F value greater that 2.8662656. Since $43.5461408 > 2.8662656 \implies F > F_{\alpha,\ a-1,\ N-1}$, we reject H_0 in favor of H_1. This means that at least two of the μ_i's are different (or one of the τ_i's are nonzero). Hence we have significant evidence at the $\alpha = 0.05$ significance level that X has an effect on Y.

We will now do the same analysis, but by calculating a p-value. In this case, the p-value is the probability of finding F-statistic values greater than the one we observed. This calculation is done in R below.

```
# Compute the p-value of the test:
p_val_F <- pf(F_obs, a-1, N-a, lower.tail = F)
```

As shown in the above calculation, the p-value associated with our data is $4.5785948 \times 10^{-12}$. Since $4.5785948 \times 10^{-12} < 0.05$, it follows that we should reject H_0 in favor of H_1. This means that at least two of the μ_i's are different (or one of the τ_i's are nonzero). Hence we have significant evidence at the $\alpha = 0.05$ significance level that X has an effect on Y.