140.652 Problem Set 4 Solutions

Problem 1

A special study is conducted to test the hypothesis that persons with glaucoma have higher blood pressure than average. Two hundred subjects with glaucoma are recruited with a sample mean systolic blood pressure of 140mm and a sample standard deviation of 25mm. (Do not use a computer for this problem.)

a. Construct a 95% confidence interval for the mean systolic blood pressure among persons with glaucoma. Do you need to assume normality? Explain.

Let $X_1, X_2, \cdots, X_{200}$ denote the SBP of the 200 individuals. We are given that, $\bar{X} = 140$ and $s = 25$. Since we have a relatively large number of subjects ($n = 200$), we can apply the Central Limit Theorem which states that the distribution of the sample mean of the 200 individuals follows an approximately normal distribution with mean 140 and standard deviation 25. Using the sample variance as an estimate for the population variance, we have

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{140 - \mu}{25/\sqrt{200}} \sim N(0, 1)$$

Thus, we do not need to assume normality for this problem to construct an approximate 95% confidence interval; however, we do need to assume independence among subjects and finite first/second moments (essentially, finite mean and variance).

$$\begin{aligned}
0.95 &= P\left(Z_{0.025} \leq \frac{140 - \mu}{25/\sqrt{200}} \leq Z_{0.975}\right) \\
&= P\left(-140 + Z_{0.025} * \frac{25}{\sqrt{200}} \leq -\mu \leq -140 + Z_{0.975} * \frac{25}{\sqrt{200}}\right) \\
&= P\left(140 - Z_{0.025} * \frac{25}{\sqrt{200}} \geq \mu \geq 140 - Z_{0.975} * \frac{25}{\sqrt{200}}\right) \\
&\overset{(*)}{=} P\left(140 - Z_{0.975} * \frac{25}{\sqrt{200}} \leq \mu \leq 140 + Z_{0.975} * \frac{25}{\sqrt{200}}\right)
\end{aligned}$$

where $(*)$ follows from noting that $N(0, 1)$ is symmetric, so $Z_{0.975} = -Z_{0.025}$. Thus, a 95% confidence interval for mean SBP among persons with glaucoma is:

$$\left(140 - Z_{0.975} * \frac{25}{\sqrt{200}}, 140 + Z_{0.975} * \frac{25}{\sqrt{200}}\right) = (136.54, 143.46)$$

```
140 + c(-1, 1)*qnorm(0.975) *25/sqrt(200)
```

```
[1] 136.5352 143.4648
```

b. If the average systolic blood pressure for persons without glaucoma of comparable age is 130mm. Is there statistical evidence that the blood pressure is elevated?

We want to test whether the average systolic blood pressure (SBP) for persons with glaucoma is statistical different from the average SBP for persons without glaucoma. Let μ_G denote the average SBP for persons with glaucoma. Then, we want to test the hypothesis:

$$H_0 : \mu_G = 130mm$$
$$H_A : \mu_G \neq 130mm$$

To test the null hypothesis, we want to find the probability that we observe a value as or more extreme under the null distribution. Recall that our sample mean and standard deviation were 140 and 25, respectively. The the probability that we observe a value as or more extreme is given by,

$$\begin{aligned} P\left(|\bar{X} - \mu_G| \geq 10\right) &= P\left(\left|\frac{\bar{X} - \mu}{25/\sqrt{200}}\right| \geq \frac{10}{25/\sqrt{200}}\right) \\ &= P\left(|Z| \geq \frac{10}{25/\sqrt{200}}\right) \\ &= P\left(Z \leq -\frac{10}{25/\sqrt{200}}\right) + P\left(Z \geq \frac{10}{25/\sqrt{200}}\right) \\ &\overset{(*)}{=} 2 * P\left(Z \leq -\frac{10}{25/\sqrt{200}}\right) \end{aligned}$$

where Z is a standard normal random variable and $(*)$ follows from the fact that the standard normal distribution is symmetric. Using `R` to evaluate this probability, we get that $P(|\bar{X} - \mu_G| \geq 10)$ is,

```
pnorm(-10/(25/sqrt(200)), lower.tail = TRUE) * 2
```

```
[1] 1.541726e-08
```

With a significance level of $\alpha = 0.05$ and a p-value of 1.54×10^{-8}, we reject the null, suggesting that that blood pressure is indeed elevated among individuals with glaucoma.

Note: A one-sided hypothesis test would be an acceptable homework answer. The two sided test is more conservative.

Problem 2

Consider the previous question. Make a probabilistic argument that the interval

$$\left[\bar{X} - z_{.95}\frac{s}{\sqrt{n}}, \quad \infty\right]$$

is a 95% *lower bound* for μ.

Let us first look at the probability $\mu \in \left[\bar{X} - z_{.95}\frac{s}{\sqrt{n}}, \infty\right]$.

$$\begin{aligned} P\left(\mu \geq \bar{X} - z_{.95}\frac{s}{\sqrt{n}}\right) &= P\left(\mu - \bar{X} \geq -z_{.95}\frac{s}{\sqrt{n}}\right) \\ &= P\left(\frac{\mu - \bar{X}}{s/\sqrt{n}} \geq -z_{.95}\right) \\ &= P\left(-Z \geq -z_{.95}\right) \\ &= P\left(Z \leq z_{.95}\right) \\ &= 0.95 \end{aligned}$$

That is, $\left[\bar{X} - z_{.95}\frac{s}{\sqrt{n}}, \infty\right]$ is a one sided confidence interval that contains 95% of the probability density of the relevant distribution, thus creating a 95% lower bound. We say that this interval is constructed such that in 95% of repeated experiments, this interval would contain the true parameter μ 95% of the time. In addition, we can use the z statistic rather than the t statistic as the sample size is large and we can appeal to asymptotics.

Problem 3

Suppose we wish to estimate the concentration μg/mℓ of a specific dose of ampicillin in the urine. We recruit 25 volunteers and find that they have sample mean concentration of 7.0 μg/mℓ with sample standard deviation 3.0 μg/mℓ. Let us assume that the underlying population distribution of concentrations is normally distributed.

a. Find a 90% confidence interval for the population mean concentration.

Since we only have 25 volunteers, we cannot use the normal distribution to generate a 90% confidence interval. A 90% confidence interval for mean concentration of ampicillin among subjects on this particular dosage is

$$\bar{X} \pm t_{0.95,24} \frac{s}{\sqrt{n}},$$

```
7 + c(-1,1) * qt(0.95, df = 24) * 3/sqrt(25)
```

```
[1] 5.973471 8.026529
```

b. How large a sample would be needed to insure that the length of the confidence interval is 0.5 μg/mℓ if it is assumed that the sample standard deviation remains at 3.0 μg/mℓ?

The length of the confidence interval is given by,

$$\left(\bar{X} + t_{n-1,0.95} \frac{s}{\sqrt{n}}\right) - \left(\bar{X} - t_{n-1,0.95} \frac{s}{\sqrt{n}}\right) = 2 \times t_{n-1,0.95} \frac{s}{\sqrt{n}}$$

Note here that both $t_{n-1}, 0.95$ and $\frac{s}{\sqrt{n}}$ change when n changes, so we cannot plug in 0.5 for the length and solve for n. Thus, the idea here is to use the normal distribution to approximate the necessary sample size and then use that approximation to derive the required sample size under the t-distribution.

Why can we use the normal distribution as an approximation? First note that for the interval given in part (a) with a sample size of 25, we see that the length of the interval is,

$$2 \times t_{24,0.95} \frac{3}{\sqrt{25}} \approx 2.05$$

```
2 * qt(0.95, df = 24) * 3/sqrt(25)
```

```
[1] 2.053058
```

To get a CI with length of 0.5, we need an approximately four-fold reduction in the length. From the equation for the length of a CI (and assuming $t_{n-1,0.95}$ is a constant for now), we need to increase the sample size by approximately $4^2 = 16$ times the current sample size to decrease the width of CI by a factor of 4, as the length is inversely proportional to $\sqrt{n}$. Sixteen times the current sample size gives us $16 \times 25 = 400$. At sample sizes around 400, we can use the Z-statistic as an approximation for the t-statistic.

Solving the equation for the length of a CI using the Z-statistic,

$$0.5 = 2 \times Z_{0.95} \frac{3}{\sqrt{n}} \implies n = \left(\frac{2 * 3}{0.5} \times Z_{0.95}\right)^2$$

```
(2*3/0.5 * qnorm(0.95))^2
```

```
[1] 389.5983
```

we get that we need a sample size of 390 to have a 90% CI with length less or equal to 0.5 (as sample sizes must be integers).

Plugging in $n = 390$ for the length of the 90% CI and adjusting to find the necessary sample size under the t-distribution, we see that for a CI of length at most 0.5, we need a sample size of 392.

```r
# Using n = 390, we see that the length is still larger than 0.5
2*qt(0.95, df = 390-1) * 3/sqrt(390)
```

```
[1] 0.5009354
```

```r
# Using n = 391, we see that the length is still larger than 0.5
2*qt(0.95, df = 391-1) * 3/sqrt(391)
```

```
[1] 0.5002913
```

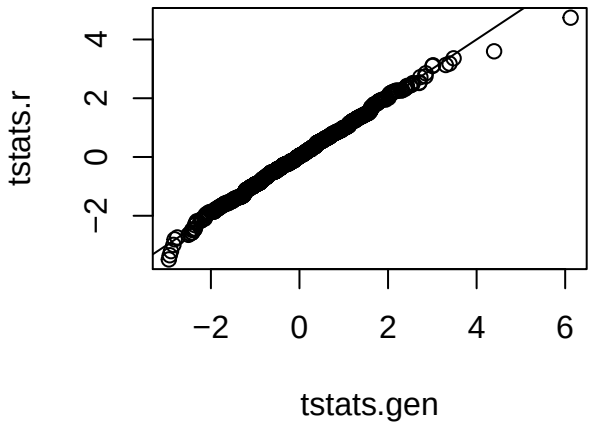
```r
# Using n = 392, the length is finally below 0.5
2*qt(0.95, df = 392-1) * 3/sqrt(392)
```

```
[1] 0.4996497
```

Problem 4

Here we will verify that standard means of iid normal data follow Gossett's t distribution. Randomly generate $1,000 \times 20$ normals with mean 5 and variance 2. Place these results in a matrix with $1,000$ rows. Using two `apply` statements on the matrix, create two vectors, one of the sample mean from each row and one of the sample standard deviation from each row. From these $1,000$ means and standard deviations, create $1,000$ t statistics. Now use R's `rt` function to directly generate $1,000$ t random variables with 19 df. Use R's `qqplot` function to plot the quantiles of the constructed t random variables versus R's t random variables. Do the quantiles agree? Describe why they should.

```r
# Simulate 1000*20 normals with mean 5 and variance 2
simData <- matrix(rnorm(1000*20, mean = 5, sd = sqrt(2)),
                  nrow = 1000, ncol = 20)
# Use apply function to create a vector of sample means and sample variances
means <- apply(simData, 1, mean)
sds   <- apply(simData, 1, sd)
# Create vector of t-statistics
tstats.gen <- (means - 5)/(sds/sqrt(20))
# Use R's rt function to directly generate 1000 t(df = 19) rvs
tstats.r <- rt(1000, df = 19)
# Plot the quantiles
qqplot(tstats.gen, tstats.r)
abline(0,1)
```



They quantiles agree, as expected since standardized means of normal random variables using sample standard deviations follow a t distribution with degrees of freedom equal to $n - 1 = 20 - 1 = 19$.

Problem 5

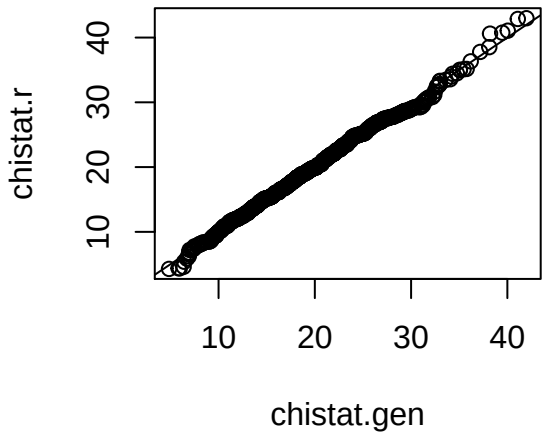
Here we will verify the chi-squared result. Simulate 1,000 sample variances of 20 observations from a normal distribution with mean 5 and variance 2. Convert these sample variances so that they should be chi-squared random variables with 19 degrees of freedom. Now simulate 1,000 random chi-squared variables with 19 degrees of freedom using R's `rchisq` function. Use R's `qqplot` function to plot the quantiles of the constructed chi-squared random variables versus those of R's random chi-squared variables. Do the quantiles agree? Describe why they should.

We will use the same samples generated in problem 4. Recall that

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

```
chistat.gen <- (20 - 1) * sds^2/2      # Generate chi-squared (df = 19) rv
chistat.r   <- rchisq(1000, df = 19)   # Use R's rchisq function

# Plot the quantiles
qqplot(chistat.gen, chistat.r)
abline(0,1)
```



The quantiles of appropriately normalized variances from a normal distribution do agree with the theoretical quantiles (as output by R), thus validating that $(n-1)S^2/\sigma^2$ follows a χ^2 distribution with $n-1$ degrees of freedom.

Problem 6

If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ then we know that $(n-1)S^2/\sigma^2$ is chi-squared with $n-1$ degrees of freedom. You were told that the expected value of a chi-squared is its degrees of freedom. Use this fact to verify the (already known fact) that $E[S^2] = \sigma^2$. (Note that S^2 is unbiased holds regardless of whether or not the data are normally distributed. Here we are just showing that the chi-squared result for normal data is not a contradiction of unbiasedness.)

We want to show that $E[S^2] = \sigma^2$. Recall that if X a chi-squared random variable with n degrees of freedom, then $E[X] = n$, and $(n-1)S^2/\sigma^2$ follows a χ^2 distribution with $n-1$ degrees of freedom. Thus,

$$E[S^2] = E\left[\frac{\sigma^2}{n-1} \cdot \frac{(n-1)S^2}{\sigma^2}\right] = \frac{\sigma^2}{n-1} \cdot E\left[\frac{(n-1)S^2}{\sigma^2}\right] = \frac{\sigma^2}{n-1} \cdot (n-1) = \sigma^2$$

Problem 7

A study of blood alcohol levels (mg/100 ml) at post-mortem examination from traffic accident victims involved taking one blood sample from the leg, A, and another from the heart, B. The results were:

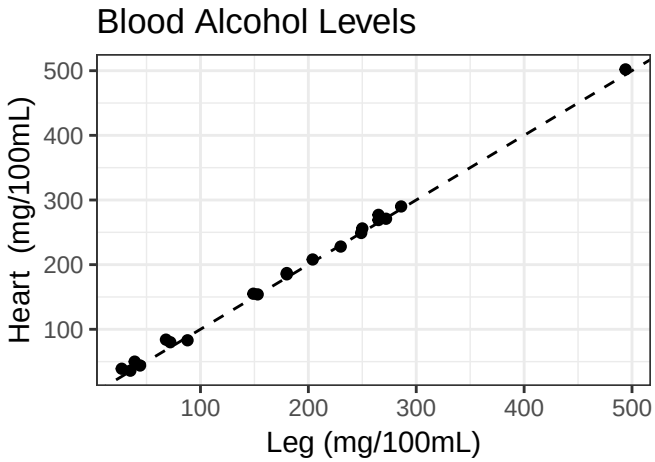
Case	A	B	Case	A	B
1	44	44	11	265	277
2	265	269	12	27	39
3	250	256	13	68	84
4	153	154	14	230	228
5	88	83	15	180	187
6	180	185	16	149	155
7	35	36	17	286	290
8	494	502	18	72	80
9	249	249	19	39	50
10	204	208	20	272	271

a. Create a graphical display comparing a case's blood alcohol level in the heart to that in the leg. Comment on any interesting patterns from your graph.

The data can be displayed multiple ways, but we want to be sure to capture the fact that we are interested in the difference in blood alcohol content of samples from two different locations within individuals. A boxplot of samples in the leg versus samples in the heart gives us information on how the alcohol content of leg samples differs from alcohol content in heart samples at a population level, while we are given paired data. One example to display the data are given below.

```
# Data
A <- c( 44,265,250,153, 88,180, 35,494,249,204,
       265, 27, 68,230,180,149,286, 72, 39,272)
B <- c( 44,269,256,154, 83,185, 36,502,249,208,
       277, 39, 84,228,187,155,290, 80, 50,271)

# Plot data
ggplot(data.frame(Leg = A, Heart = B), aes(x = Leg, y = Heart)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, linetype = 'dashed') +
  labs(title = "Blood Alcohol Levels",
       x = "Leg (mg/100mL)", y = "Heart  (mg/100mL)") +
  theme_bw()
```

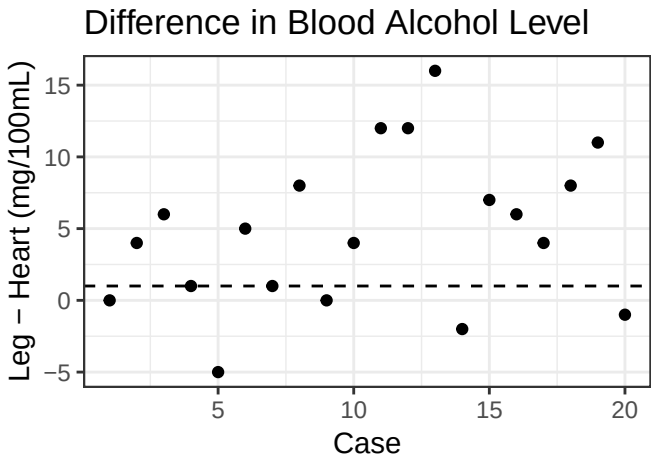


From the plot above, we see that there seems to be a difference between the blood alcohol level between

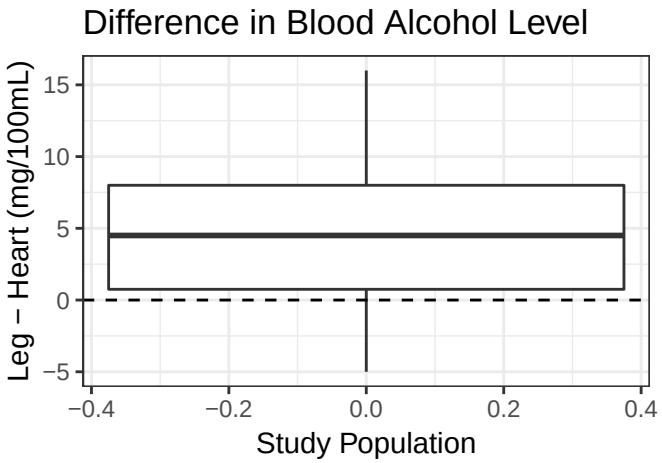
samples from the leg and samples from the heart; samples from the heart tend to have a higher blood alcohol level as indicated by the majority of points being above the $y = x$ line.

b. Create a graphical display of the distribution of the difference in blood alcohol levels between the heart and the leg.

```
# Scatter plot of differences
ggplot(data.frame(case = 1:20, difference = B - A),
       aes(x = case, y = difference)) +
  geom_point() +
  geom_abline(slope = 0, intercept = 1, linetype = 'dashed') +
  labs(title = "Difference in Blood Alcohol Level",
       x = "Case", y = "Leg - Heart (mg/100mL)") +
  theme_bw()
```




```
# Boxplot of differences
ggplot(data.frame(case = 1:20, difference = B - A)) +
  geom_boxplot(aes(y = difference)) +
  geom_abline(slope = 0, intercept = 0, linetype = 'dashed') +
  labs(title = "Difference in Blood Alcohol Level",
       x = "Study Population", y = "Leg - Heart (mg/100mL)") + theme_bw()
```



In comparison to the plot given in part (a), it is clearer in these plots that there is a difference in blood alcohol levels between samples from the leg and samples from the heart.

c. Do these results indicate that in general blood alcohol levels may differ between samples taken from the leg and the heart? Give confidence interval and interpret your results.

Let μ_A and μ_B denote the average blood alcohol levels in leg and heart samples, respectively. With a sample size of 20, we can use a t-test to test,

$$H_0 : \mu_A - \mu_B = 0$$
$$H_A : \mu_A - \mu_B \neq 0$$

```
diff  <- B - A
n     <- length(B - A)

# Using the t-test function in R
t.test(diff)
```

```
    One Sample t-test

data:  diff
t = 4.0367, df = 19, p-value = 0.0007046
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.335275 7.364725
sample estimates:
mean of x
     4.85
```

```
# Deriving the 95% CI from the equation
mean(diff) + c(-1, 1) * qt(0.975, df = n - 1) * sd(diff)/sqrt(n)
```

```
[1] 2.335275 7.364725
```

With a significance level of $\alpha = 0.5$ and a p-value of 7×10^{-4}, we reject the null and conclude that there is indeed a difference in alcohol levels between samples A and samples B. The 95% confidence interval for the average difference in blood alcohol levels is (2.335, 7.365). Since this interval does not contain 0 and is positive, it appears that samples from the heart have higher blood alcohol levels than samples from the leg.

d. Create a profile likelihood for the true mean difference and interpret.

Let X_i for $i = 1, 2, \cdots, n$ be iid random variables denoting the average difference between samples A and B for individual i. By the Central Limit Theorem, we know that, $X_i \sim N(\mu, \sigma^2)$ for some unknown μ and σ^2. The full data likelihood is thus,

$$\mathcal{L}(\mu, \sigma^2 | x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}$$

To get the profile likelihood for μ, we substitute the ML estimate for σ^2 into the data likelihood. Recall from HW3 that the MLE for σ^2 is,

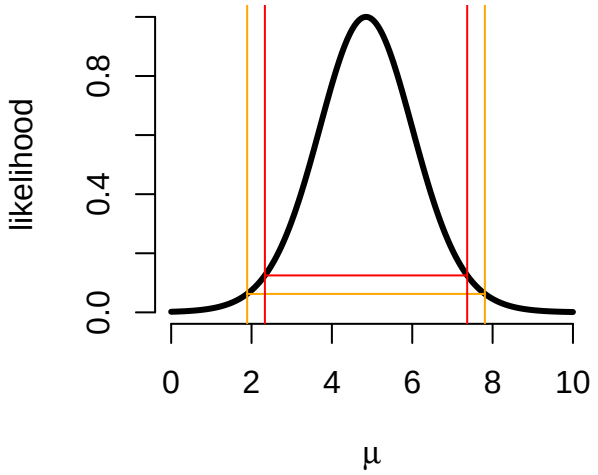
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

The profile likelihood for σ^2 is thus,

$$
\begin{aligned}
\mathcal{PL}(\mu|x_1, x_2, \cdots, x_n) &= \mathcal{L}(\mu, \hat{\sigma}^2|x_1, x_2, \cdots, x_n) \\
&= \left(\frac{2\pi}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right)^{-n/2} \exp\left\{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\frac{2}{n}\sum_{i=1}^{n}(x_i-\mu)^2}\right\} \\
&= \left(\frac{2\pi}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right)^{-n/2} \exp\left\{-\frac{n}{2}\right\} \\
&\propto \left(\sum_{i=1}^{n}(x_i-\mu)^2\right)^{-n/2}
\end{aligned}
$$

```
muVals  <- seq(0, 10, length = 1000)
proflik <- sapply(muVals, function(mu) sum((diff - mu)^2)^(-n/2))
proflik <- proflik / max(proflik)

# Plot
plot(muVals, proflik, type = "l", xlab = expression(mu),
     ylab = "likelihood", frame = FALSE, lwd = 3)
lines(range(muVals[proflik>1/8]), c(1/8,1/8), col = "red")
lines(range(muVals[proflik>1/16]), c(1/16,1/16), col = "orange", cex = 3)
abline(v = range(muVals[proflik>1/8]), col = "red")
abline(v = range(muVals[proflik>1/16]), col = "orange")
```



In the figure above, the 1/8th and 1/16th intervals are colored in red and orange, respectively. More precisely, the 1/8th interval is given by,

```
range(muVals[proflik>1/8])
```

```
[1] 2.332332 7.367367
```

And the ML estimate for the average difference in blood alcohol levels is,

```
muVals[which(proflik == max(proflik))]
```

```
[1] 4.854855
```

With a 1/8th interval of (2.332, 7.367) and an MLE of 4.855, we see that the data supports average differences between 2.5 and 7 given the data. Notice that there is little evidence supporting an effect size of 0 as the likelihood is low at 0.

e. Create a likelihood for the variance of the difference in alcohol levels and interpret.

To get the profile likelihood for σ^2, we substitute the ML estimate for μ into the data likelihood. Recall that the MLE for μ is,

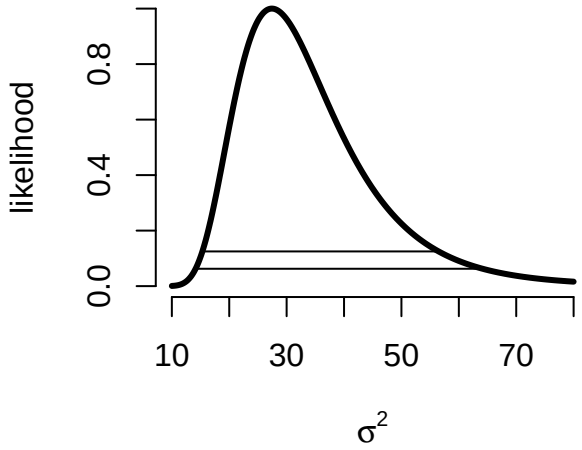
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

The profile likelihood for σ^2 is thus,

$$\mathcal{PL}(\sigma^2|x_1, x_2, \cdots, x_n) = \mathcal{L}(\hat{\mu}, \sigma^2|x_1, x_2, \cdots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}$$

```
sigVals   <- seq(10, 80, length = 1000)
proflik.s <- 1/(sigVals^(n/2))*exp(-sum((diff-mean(diff))^2)/2/sigVals)
proflik.s <- proflik.s/max(proflik.s)

# Plot
plot(sigVals, proflik.s,
     type = "l",
     frame = F,
     xlab = expression(sigma^2),
     ylab = "likelihood", lwd = 3)
lines(range(sigVals[proflik.s >= 1 / 8]), c(1 / 8, 1 / 8))
lines(range(sigVals[proflik.s >= 1 / 16]), c(1 / 16, 1 / 16))
```



The likelihood supports values of σ^2 roughly between 10 and 80. Specifically, the 1/8 likelihood interval is,

```
range(sigVals[proflik.s>1/8])
```

```
[1] 15.32533 56.45646
```

Problem 8

A random sample was taken of 20 patients admitted to a hospital with a certain diagnosis. The lengths of stays in days for the 20 patients were

4, 2, 4, 7, 1, 5, 3, 2, 2, 4

5, 2, 5, 3, 1, 4, 3, 1, 1, 3

a. Calculate a 95% confidence interval (use the method $\overline{X} \pm t$ SE) for the mean length of hospital stay. Is your answer reasonable? What underlying assumptions were required for this method and are they reasonable?

Let $X_1, X_2, \cdots, X_{20}$ be the length of stay for the 20 patients. Recall that the 95% confidence interval is given by,

$$\bar{X} \pm t_{0.975,n-1} \frac{s}{\sqrt{n}}$$

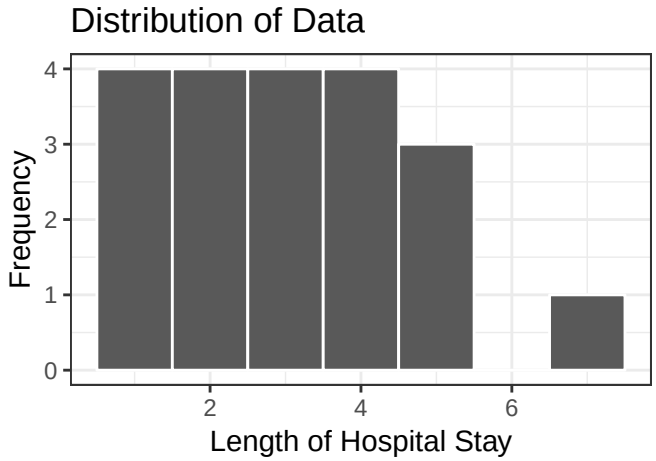
Using `R` to calculate the CI,

```
dat <- c(4,  2,  4,  7,  1,  5,  3,  2,  2,  4,
         5,  2,  5,  3,  1,  4,  3,  1,  1,  3)
mean(dat) + c(-1, 1) * qt(0.975, 20-1) * sd(dat)/sqrt(20)
```

```
[1] 2.327235 3.872765
```

The answer seems reasonable (there doesn't seem to be an obvious reason why it is not, but concluding that the answer is not reasonable is also acceptable). The 95% confidence interval technically assumes that the data are iid normal, but is fairly robust to the normality assumption so long as the data is symmetrically distributed and mound shaped. This assumption may not be reasonable as the data does not seem symmetrically distributed.

```
# Plot distribution of data
ggplot(data.frame(dat = dat), aes(x = dat)) +
  geom_histogram(binwidth = 1, color = 'white') +
  labs(title = "Distribution of Data",
       x = "Length of Hospital Stay", y = "Frequency") +
  theme_bw()
```



b. Calculate a 95% percentile bootstrap interval and interpret.

```
set.seed(1234)
B <- 10000 # Set number of bootstrap simulations

# Resample the data B times, each with n observations
resampled.dat <- matrix(sample(dat, B*n, replace = TRUE), B, n)

# calculate means
resampled.mean <- apply(resampled.dat, 1, mean)

# Calculate 95\% CI
quantile(resampled.mean, c(.0275, .975))
```

```
2.75% 97.5%
 2.45  3.85
```

The bootstrap interval is $[2.45, 3.85]$ (answers may vary slightly). This agrees with the t interval generated in part (a).

Problem 9

Refer to the previous problem. Take logs of the data (base "e").

a. Calculate a 95% confidence interval for the mean of the log length of stay.

```
mean(log(dat)) + c(-1, 1) * qt(0.975, 20-1) * sd(log(dat))/sqrt(20)
```

```
[1] 0.6900811 1.2585628
```

The 95% confidence interval for the average log-length of stay is (0.69, 1.26).

b. Take antilogs (exponential) of the endpoints of the confidence interval found in part a. Explain why that is a 95% confidence interval for the median length of stay if the data is lognormally distributed (lognormally distributed is when the logarithm of the data points has a normal distribution). Technically, under the lognormal assumption, is the confidence interval that you found in this equation also a confidence interval for the mean length of stay?

The exponentiated 95% confidence interval is given by (1.99, 3.52).

```
exp(mean(log(dat)) + c(-1, 1) * qt(0.975, 20-1) * sd(log(dat))/sqrt(20))
```

```
[1] 1.993877 3.520359
```

Now let $X_1, X_2, \cdots, X_n$ be the median length of stay and assume that X_i is lognormally distributed. That is, $\log(X_i)$ follows a normal distribution. To understand why the exponentiated confidence interval is also a confidence interval for the median length of stay, note that

1. We first created a 95% confidence interval for $E[\log(X)]$. Since $\log(X)$ follows a normal distribution, the mean and the median of $\log(X)$ are equal. Thus, the 95% confidence interval for the mean of $\log(X)$ is the same as the 95% confidence interval of the median of $\log(X)$.
2. Next, we exponentiated the 95% confidence interval. Since exp() is a monotonic function, if $a < b$ then $\exp(a) < \exp(b)$. Thus,

$$\exp(\text{median}(\log(X))) = \text{median}(\exp(\log(X))) = \text{median}(X)$$

   and the exponentiated 95% confidence interval is also a confidence interval for the median of the length of stay.

However, under the lognormal assumption, the exponentiated confidence interval is **not** a confidence interval for the mean length of stay. To see this, note that,

1. We first created a 95% confidence interval for the mean of $log(X)$, i.e. the confidence interval has a 95% chance of containing $E[\log(X)]$.
2. When we exponentiate the 95% confidence interval, we exponentiate all values covered by the confidence interval. Thus, if $E[\log(X)]$ is in the confidence interval, we get $\exp(E[\log(X)])$. This means that 95% of the time, our exponentiated confidence interval will contain $\exp(E[\log(X)])$.
3. However, $\exp(E[\log(X)]) \neq E[\exp(\log(X))] = E[X]$. Thus, we cannot conclude whether the exponentiated confidence interval will contain the mean 95% of the time, or equivalently, the exponentiated interval is not a confidence interval for the *mean* length of stay.

Problem 10

Let p denote the unknown proportion of rocks in a riverbed that are sedimentary in type. Suppose that $X = 12$ of a sample of $n = 20$ rocks collected in random locations are found to be sedimentary in type.

a. Plot the likelihood for the parameter p and interpret.

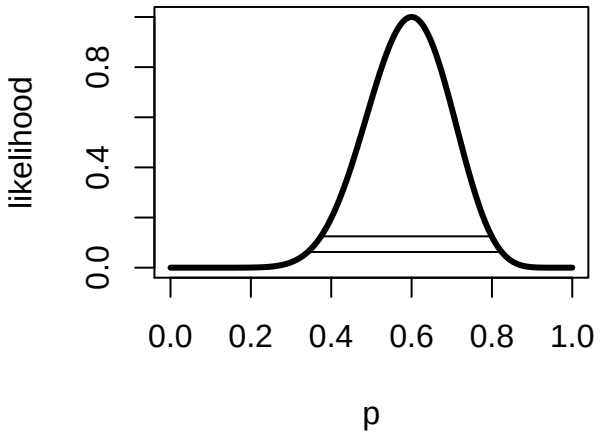
The likelihood is given by,

$$\mathcal{L}(p|X) = \binom{n}{X} p^X (1-p)^{20-X}$$

```
# Define variables
p <- seq(0, 1, by = 0.001)
n <- 20
x <- 12

# Calculate likelihood
likelihood <- choose(n, x) * p^x * (1-p)^(n-x)
likelihood <- likelihood/max(likelihood)

# Plot
plot(p, likelihood, type = "l", xlab = "p",
     ylab = "likelihood", lwd = 3)
lines(range(p[likelihood >= 1/8]), c(1/8, 1/8))
lines(range(p[likelihood >= 1/16]), c(1/16, 1/16))
```




```
# Find values best supported by the data
range(p[likelihood >= 1/8])
```

```
[1] 0.375 0.799
```

The likelihood plot displays value of p by their relative evidence. Values between roughly .4 and .8 seem plausibly supported given the data. Specifically, the 1/8 interval is .38 to .80.

b. From your graphs, determine the value of $\hat{p}$ of p where the curve reaches its maximum. Does this value for the maximum make intuitive sense? Comment in one or two sentences.

From the graphs, the value $\hat{p}$ that maximizes the likelihood curve is given by,

```
p[which(likelihood == max(likelihood))]
```

```
[1] 0.6
```

The value of the maximum makes intuitive sense as it is the most probable value of p given the data.

c. Show that the point that maximizes the binomial likelihood is always X/n.

Suppose we are given data X which follows a binomial distribution. Then, to get the maximum likelihood estimate of $\hat{p}$, we can take the derivative of the log-likelihood, set it equal to zero, and solve for p.

$$\begin{aligned}
\mathcal{L}(p|X) &= \binom{n}{X} p^X (1-p)^{n-X} \\
\log(\mathcal{L}(p|X)) &= \log\binom{n}{X} + X\log p + (n-X)\log(1-p) \\
\frac{\partial}{\partial p}\log(\mathcal{L}(p|X)) &= \frac{X}{p} - \frac{n-X}{1-p}
\end{aligned}$$

Setting the derivative equal to 0 and solving for p, we get

$$\hat{p} = \frac{X}{n}$$

Verifying that $\hat{p}$ indeed maximizes the likelihood by checking that the second derivative is negative,

$$\frac{\partial}{\partial p}\log(\mathcal{L}(X|p)) = -\frac{X}{p^2} - \frac{n-X}{(1-p)^2} < 0$$

Thus, the point that maximizes the binomial likelihood is always $\hat{p} = \frac{X}{n}$.

d. Use the CLT to create a confidence interval for the true proportion of rocks that are sedimentary. Interpret your results.

For a binomial proportion, we can use a $(1-\alpha)\%$ Wald confidence interval given by,

$$\hat{p} \pm Z_{1-\alpha} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Using `R`, we get that the 95% CI is given by,

```
phat <- 12/20
phat + c(-1, 1) * qnorm(0.975) * sqrt(phat*(1-phat)/20)
```

```
[1] 0.3852967 0.8147033
```

which means that 95% of the time, we expect (0.39, 0.81) to contain the true mean.

e. A much larger study is planned and the researchers would like to know how large n should be to have a margin of error on the estimate for the proportion of sedimentary rocks that is no larger than .01 for a 95% confidence interval? Use the fact that $p(1-p) \leq 1/4$. Also try the calculation with the estimate of p from the current study.

The margin of error (MoE) for a 95% confidence interval is given by,

$$\text{MoE} = Z_{0.975}\sqrt{\frac{p(1-p)}{n}}$$

Using the fact that $p(1-p) \leq \frac{1}{4}$, we get that

$$\text{MoE} = Z_{0.975}\sqrt{\frac{p(1-p)}{n}} \leq \frac{Z_{0.975}}{\sqrt{4n}}$$

Setting the margin of error equal to 0.01 and solving for n, we have,

$$n \geq \left(\frac{Z_{0.975}}{2 * 0.01}\right)^2 = 9603.65$$

```r
(qnorm(0.975) / (2*0.01))^2
```

```
[1] 9603.647
```

Since n must be an integer, $n \geq 9,604$.

Rather than using the bound of $\frac{1}{4}$ for $p(1-p)$, we can directly solve for n from the MoE equation above. Using this approach, we get that

$$n = \left(\frac{Z_{0.975}\sqrt{p(1-p)}}{\text{MoE}} \right)^2$$

Substituting 0.01 for MoE and 0.6 for p, we get that $n \geq 9220$.

```r
p <-  0.6
(qnorm(0.975)*sqrt(p*(1-p))/0.01)^2
```

```
[1] 9219.501
```

Problem 11

This problem investigates the performance of the Wald confidence interval.

a. Using a computer, generate 1000 Binomial random variables for $n = 10$ and $p = .3$ Calculate the percentage of times that

$$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$$

contains the true value of p. Here $\hat{p} = X/n$ where X is each binomial variable. Do the intervals appear to have the coverage that they are supposed to?

```r
set.seed(1321)
# Generate binomial r.v.
x <- rbinom(1000, size = 10, prob = 0.3)

# Estimate p and create CI
phat  <- x/10
ub <- phat + qnorm(0.975) * sqrt(phat * (1 - phat)/10)
lb <- phat - qnorm(0.975) * sqrt(phat * (1 - phat)/10)

# get the proportion of times p = 0.3 is within the confidence interval
mean(ub >= 0.3 & lb <= 0.3)
```

```
[1] 0.831
```

No, the interval only contains the truth .83 percent of the time.

b. Repeat the calculation only now use the interval

$$\tilde{p} \pm 1.96\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}$$

where $\tilde{p} = (X+2)/(n+4)$. Does the coverage appear to be closer to .95?

```r
# Estimate p and create CI
ptilde  <- (x + 2) / (10 + 4)
ub <- ptilde + qnorm(0.975) * sqrt(ptilde * (1 - ptilde)/(10+4))
lb <- ptilde - qnorm(0.975) * sqrt(ptilde * (1 - ptilde)/(10+4))

# get the proportion of times p = 0.3 is within the confidence interval
mean(ub >= 0.3 & lb <= 0.3)
```

```
[1] 0.966
```

The coverage for this interval is .966, which is closer to 0.95.

c. Repeat this comparison (parts a. - d.) for $p = .1$ and $p = .5$. Which of the two intervals appears to perform better?

```
set.seed(1321)

# p = 0.1
x <- rbinom(1000, size = 10, prob = 0.1)
phat   <- x/10
ptilde <- (x + 2)/(10 + 4)

# Calculate CI
ub.hat <- phat + qnorm(0.975) * sqrt(phat * (1 - phat)/10)
lb.hat <- phat - qnorm(0.975) * sqrt(phat * (1 - phat)/10)
ub.tilde <- ptilde + qnorm(0.975) * sqrt(ptilde * (1 - ptilde)/(10 + 4))
lb.tilde <- ptilde - qnorm(0.975) * sqrt(ptilde * (1 - ptilde)/(10 + 4))

# get the proportion of times p = 0.1 is within the confidence interval
mean(ub.hat >= 0.1 & lb.hat <= 0.1)
```

```
[1] 0.635
```

```
mean(ub.tilde >= 0.1 & lb.tilde <= 0.1)
```

```
[1] 0.947
```

```
# p = 0.5
x <- rbinom(1000, size = 10, prob = 0.5)
phat   <- x/10
ptilde <- (x + 2)/(10 + 4)

# Calculate CI
ub.hat <- phat + qnorm(0.975) * sqrt(phat * (1 - phat)/10)
lb.hat <- phat - qnorm(0.975) * sqrt(phat * (1 - phat)/10)
ub.tilde <- ptilde + qnorm(0.975) * sqrt(ptilde * (1 - ptilde)/(10 + 4))
lb.tilde <- ptilde - qnorm(0.975) * sqrt(ptilde * (1 - ptilde)/(10 + 4))

# get the proportion of times p = 0.1 is within the confidence interval
mean(ub.hat >= 0.5 & lb.hat <= 0.5)
```

```
[1] 0.889
```

```
mean(ub.tilde >= 0.5 & lb.tilde <= 0.5)
```

```
[1] 0.98
```

The adjusted interval using $\tilde{p}$ performs much better.

Problem 12

Forced expiratory volume FEV is a standard measure of pulmonary function. We would expect that any reasonable measure of pulmonary function would reflect the fact that a person's pulmonary function declines with age after age 20. Suppose we test this hypothesis by looking at 10 nonsmoking males ages 35-39, heights 68-72 inches and measure their FEV initially and then once again 2 years later. We obtain this data.

Person	Year 0 FEV (L)	Year 1 FEV (L)	Person	Year 0 FEV (L)	Year 2 FEV (L)
1	3.22	2.95	6	3.25	3.20
2	4.06	3.75	7	4.20	3.90
3	3.85	4.00	8	3.05	2.76
4	3.50	3.42	9	2.86	2.75
5	2.80	2.77	10	3.50	3.32

a. Create the relevant confidence interval and interpret.

We are interested in the difference in pulmonary function (as measured by FEV). Assume that difference in FEV follows a normal distribution. Then, we can use the `t.test()` function in `R` to get our 95% confidence interval.

```
fev1 <- c(3.22, 4.06, 3.85, 3.50, 2.80,
          3.25, 4.20, 3.05, 2.86, 3.50)
fev2 <- c(2.95, 3.75, 4.00, 3.42, 2.77,
          3.20, 3.90, 2.76, 2.75, 3.32)
diff <- fev2 - fev1
t.test(diff)
```

```
    One Sample t-test

data:  diff
t = -3.0891, df = 9, p-value = 0.01295
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.25465006 -0.03934994
sample estimates:
mean of x
   -0.147
```

Thus, the 95% confidence interval for the average difference in FEV is (-0.255, -0.039) with an estimated difference of -0.147 suggesting that there is a decline in FEV over two years.

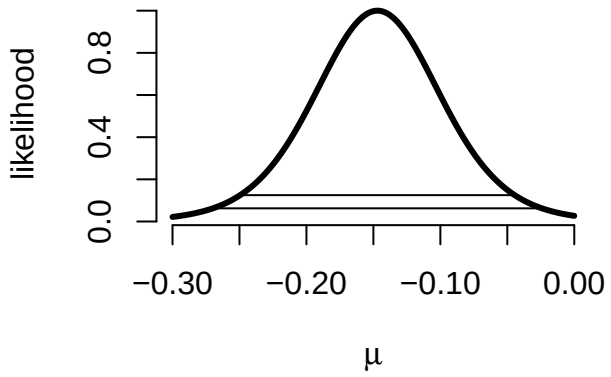
b. Create the relevant profile likelihood and interpret.

Recall from problem 7d that the profile likelihood for the difference of means (assuming that the difference is normally distributed), is given by,

$$\mathcal{PL}(\mu|x_1, x_2, \cdots, x_n) = \mathcal{L}(\mu, \hat{\sigma}^2|x_1, x_2, \cdots, x_n) \propto \left(\sum_{i=1}^{n}(x_i - \mu)^2\right)^{-n/2}$$

```
mu <- seq(-.30, 0, length = 1000)
prof.lik <- sapply(mu, function(mu) sum((diff - mu)^2) ^ (-length(diff)/2))
prof.lik <- prof.lik / max(prof.lik)

# Plot
plot(mu, prof.lik, type = "l", xlab = expression(mu),
     ylab = "likelihood", frame = FALSE, lwd = 3)
lines(range(mu[prof.lik>1/8]), c(1/8,1/8))
lines(range(mu[prof.lik>1/16]), c(1/16,1/16))
```




```
# Get most likely mu
mu[prof.lik == max(prof.lik)]
```

```
[1] -0.1471471
```

```
# Get 1/8th interval
range(mu[prof.lik > 1/8])
```

```
[1] -0.24924925 -0.04474474
```

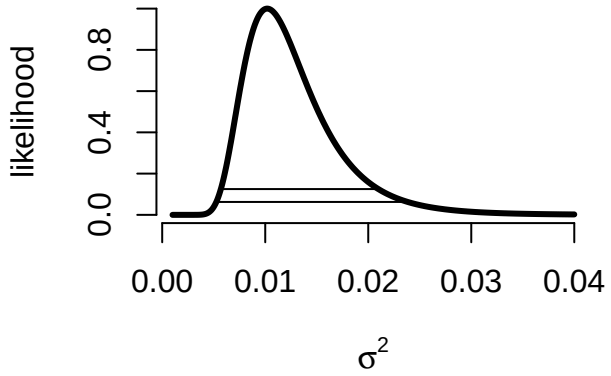
The likelihood shows the relative evidence for various values of μ. The best estimate for μ given the data is -0.147 while a $1/8$ likelihood interval is -0.22 to -0.03. Notice 0 is not a plausibly supported value.

c. Create a likelihood function for the variance of the change in FEV.

Recall from problem 7e that the profile likelihood for the variance of a normally distributed difference is,

$$\mathcal{PL}(\sigma^2|x_1, x_2, \cdots, x_n) = \mathcal{L}(\hat{\mu}, \sigma^2|x_1, x_2, \cdots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}$$

```
sigVals   <- seq(0.001, 0.04, length = 1000)
proflik.s <- 1/(sigVals^(n/2))*exp(-sum((diff-mean(diff))^2)/2/sigVals)
proflik.s <- proflik.s/max(proflik.s)
# Plot
plot(sigVals, proflik.s,
     type = "l",
     frame = F,
     xlab = expression(sigma^2),
     ylab = "likelihood", lwd = 3)
lines(range(sigVals[proflik.s >= 1 / 8]), c(1 / 8, 1 / 8))
lines(range(sigVals[proflik.s >= 1 / 16]), c(1 / 16, 1 / 16))
```



Problem 13

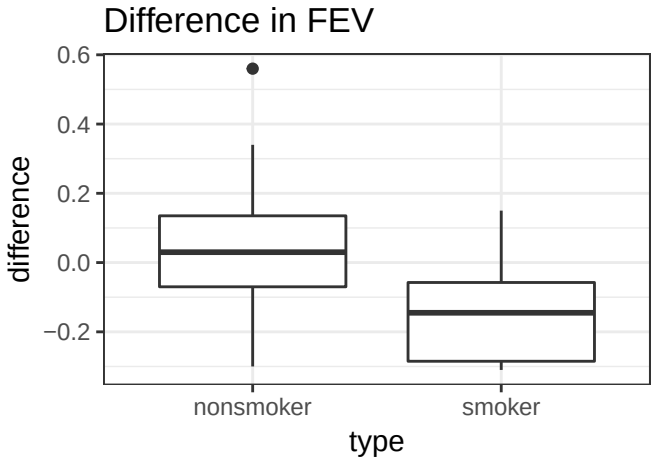
Another aspect of the preceding study involves looking at the effect of smoking on baseline pulmonary function and on change in pulmonary function over time. We must be careful since FEV depends on many factors, particularly age and height. Suppose we have a comparable group of 15 men in the same age and height group who are smokers and we measure their FEV at year 0. The data are given (For purposes of this exercise assume equal variance where appropriate).

Person	FEV Year 0 (L)	FEV Year 2 (L)	Person	FEV Year 0 (L)	FEV Year 2 (L)
1	2.85	2.88	9	2.76	3.02
2	3.32	3.40	10	3.00	3.08
3	3.01	3.02	11	3.26	3.00
4	2.95	2.84	12	2.84	3.40
5	2.78	2.75	13	2.50	2.59
6	2.86	3.20	14	3.59	3.29
7	2.78	2.96	15	3.30	3.32
8	2.90	2.74			

a. Graphically display the distribution of change in pulmonary function for smokers and nonsmokers (from the previous problem).

```
fev1smoker <- c(2.85,3.32,3.01,2.95,2.78,2.86,2.78,2.90,
                2.76,3.00,3.26,2.84,2.50,3.59,3.30)
fev2smoker <- c(2.88,3.40,3.02,2.84,2.75,3.20,2.96,2.74,
                3.02,3.08,3.00,3.40,2.59,3.29,3.32)
diffsmoker <- fev2smoker - fev1smoker

data.frame(type = c(rep("smoker", length(diff)), rep("nonsmoker", length(diffsmoker))),
           difference = c(diff, diffsmoker)) %>%
  ggplot(aes(x = type, y = difference)) +
  geom_boxplot() +
  labs(title = "Difference in FEV") +
  theme_bw()
```



b. Calculate a confidence interval to determine if there is evidence to suggest that the change in pulmonary function over 2 years is the same in the two groups. State your assumptions and interpret your results.

Assume that the difference in FEV follows a normal distribution for both smokers and non-smokers, and that

the two groups have equal variance. Note that the equal variance assumption does not seem valid as smokers seem to have a greater variance in their change in FEV in comparison to non-smokers.

We can calculate the the 95% confidence interval for the difference in average change in FEV using,

$$(\hat{\mu}_S - \hat{\mu}_{NS}) \pm t_{0.975, n_{NS}+n_S-2} * S_p \sqrt{\frac{1}{n_{NS}} + \frac{1}{n_S}}$$

where the pooled sample variance is given by,

$$S_p^2 = \frac{(n_{NS}-1)S_{NS}^2 + (n_S-1)S_S^2}{n_{NS}+n_S-2}$$

```
mu.NS <- mean(diff)
mu.S  <- mean(diffsmoker)
s2.NS <- var(diff)
s2.S  <- var(diffsmoker)
n.NS  <- length(diff)
n.S   <- length(diffsmoker)

s2.pooled <- ((n.NS - 1)*s2.NS + (n.S - 1)*s2.S)/(n.NS + n.S - 2)

# Get average difference
mu.NS - mu.S
```

```
[1] -0.1996667
```

```
# Create CI
(mu.NS - mu.S) + c(-1, 1)*qt(0.975, n.NS + n.S - 2) * sqrt(s2.pooled) * sqrt(1/n.NS + 1/n.S)
```

```
[1] -0.36764246 -0.03169087
```

```
# Check using t.test() function in R
t.test(diff, diffsmoker, var.equal = TRUE)
```

```
    Two Sample t-test

data:  diff and diffsmoker
t = -2.4589, df = 23, p-value = 0.02188
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.36764246 -0.03169087
sample estimates:
  mean of x   mean of y
-0.14700000  0.05266667
```

The average difference for smokers was .05 while it was $-.147$ for non-smokers. The 95% confidence interval for the difference between the two groups is $-.367$ to -0.032, suggesting a difference in the rate of change between smokers and non-smokers as 0 is not within the CI.

Problem 14

In a trial to compare a stannous fluoride dentifrice A, with a commercially available fluoride free dentifrice D, 260 children received A and 289 received D for a 3-year period. The mean DMFS increments (the number of new Decayed Missing and Filled tooth Surfaces) were 9.78 with standard deviation 7.51 for A and 12.83 with

standard deviation 8.31 for D. Is this good evidence that, in general, one of these dentifrices is better than the other at reducing tooth decay? If so, within what limits would the average annual difference in DMFS increment be expected to be?

To compare A and D, let us compare the *annual* change in DMFS due to each fluoride, denoted by A and D, respectively. We are given,

$$\begin{aligned}
\hat{\mu}_{3A} = 9.78 \quad &\Rightarrow \quad \hat{\mu}_A = \frac{9.78}{3} \\
s^2_{3A} = (7.51)^2 \quad &\Rightarrow \quad s^2_A = \frac{(7.51)^2}{9} \\
\hat{\mu}_{3D} = 12.83 \quad &\Rightarrow \quad \hat{\mu}_D = \frac{12.83}{3} \\
s^2_{3D} = (8.31)^2 \quad &\Rightarrow \quad s^2_{3D} = \frac{(8.31)^2}{9}
\end{aligned}$$

Then, the 95% confidence interval for the average difference is given by,

$$(\hat{\mu}_A - \hat{\mu}_D) \pm t_{0.975, n_A+n_D-2} * S_p \sqrt{\frac{1}{n_A} + \frac{1}{n_D}}$$

where the pooled sample variance is given by,

$$S^2_p = \frac{(n_A - 1)S^2_A + (n_D - 1)S^2_D}{n_A + n_D - 2}$$

```
mu.A <- 9.78/3
mu.D <- 12.83/3
s2.A <- (7.51)^2/9
s2.D <- (8.31)^2/9
n.A  <- 260
n.D  <- 289

s2.pooled <- ((n.A - 1)*s2.A + (n.D - 1)*s2.D)/(n.A + n.D - 2)

# Create CI
(mu.A - mu.D) + c(-1, 1)*qt(0.975, n.A + n.D - 2) * sqrt(s2.pooled) * sqrt(1/n.A + 1/n.D)
```

```
[1] -1.4611226 -0.5722107
```

Thus, we expect the interval (-1.46, -0.57) to contain the true average annual difference between dentifrice A and dentifrice D 95% of the time. Note that this interval does not contain 0, suggesting that dentifrice A is more effective than dentifrice D.

Note: Rather than using $t_{0.975, n_A+n_D-2}$ you can use $Z_{0.975}$ since $n_A + n_D - 2$ is large, so $t_{0.975, n_A+n_D-2}$ is approximately equal to $Z_{0.975}.

Problem 15

Suppose that 18 obese subjects were randomized, 9 each, to a new diet pill and a placebo. Subjects' body mass indices (BMIs) were measured at a baseline and again after having received the treatment or placebo for four weeks. The average difference from follow-up to the baseline (followup - baseline) was -3 kg/m^2 for the treated group and 1 kg/m^2 for the placebo group. The corresponding standard deviations of the differences was 1.5 kg/m^2 for the treatment group and 1.8 kg/m^2 for the placebo group. Does the change in

BMI over the two year period appear to differ between the treated and placebo groups? (Show some work and interpret your results.) Assume normality and a common variance.

Since we assume normality and common variance we can use a two sample t-test and the pooled variance estimator to estimate the difference in treatment effect between the diet pill and placebo. Let $\hat{\mu}_D$, s^2_D and $\hat{\mu}_{Pl}$, s^2_{Pl} be the mean and variance of the average difference in BMI over the two year period for the diet pill and placebo, respectively. Then a 95% confidence interval for the difference between the groups is

$$(\hat{\mu}_D - \hat{\mu}_{Pl}) \pm t_{0.975, n_D + n_{Pl} - 2} * S_p \sqrt{\frac{1}{n_D} + \frac{1}{n_{Pl}}}$$

where the pooled sample variance is given by,

$$S_p^2 = \frac{(n_D - 1)S_D^2 + (n_{Pl} - 1)S_{Pl}^2}{n_D + n_{Pl} - 2}$$

```
mu.D <- -3
mu.P <- 1
s2.D <- (1.5)^2
s2.P <- (1.8)^2
n.D  <- 9
n.P  <- 9
s2.pooled <- ((n.D - 1)*s2.D + (n.P - 1)*s2.P)/(n.D + n.P - 2)

# Create CI
(mu.D - mu.P) + c(-1, 1)*qt(0.975, n.D + n.P - 2) * sqrt(s2.pooled) * sqrt(1/n.D + 1/n.P)
```

```
[1] -5.655699 -2.344301
```

The 95% confidence interval for the average change in BMI between those who took the diet pill and those who took the placebo is $(-5.66, -2.34)$. That is, we expect the true mean difference between the two groups will be between $(-5.66, -2.34)$ 95% of the time. Note that the confidence interval for the difference in means does not contain 0. This implies that for with significance level $\alpha = 0.5$, we reject the hypothesis that there is no difference between the two groups.

Problem 16

In a random sample of 100 subjects with low back pain, 27 reported an improvement in symptoms after exercise therapy. Give and interpret an interval estimate for the true proportion of subjects who respond to exercise therapy.

By the Central Limit Theorem we know that the appropriately standardized proportion of subjects with low back pain converges in distribution to a normal random variable. So, we create a Wald 95% Confidence interval using

$$\hat{p} \pm Z_{0.975} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

```
p.hat <- 27/100
# Create CI
p.hat + c(-1, 1) * qnorm(0.975) * sqrt(p.hat * (1 - p.hat)/100)
```

```
[1] 0.1829855 0.3570145
```

The 95% confidence interval is given by $(0.18, 0.36)$. That is, we would expect the interval $(0.18, 0.36)$ to contain the true proportion of subject who report an improvement in symptoms following exercise therapy 95% of the time.

Problem 17

Suppose that systolic blood pressures were taken on 16 oral contraceptive users and 16 controls at baseline and again then two years later. The average difference from follow-up SBP to the baseline (followup - baseline) was 11 $mmHg$ for oral contraceptive users and 4 $mmHg$ for controls. The corresponding standard deviations of the differences was 20 $mmHg$ for OC users and 28 $mmHg$ for controls.

a. Calculate and interpret a 95% confidence interval for the change in systolic blood pressure for oral contraceptive users; assume normality.

Let $\hat{\mu}_C$ and $\hat{\mu}_O$ be the average difference in SBP from baseline to follow-up for controls and oral contraceptive users, respectively. Additionally, let s_C^2 and s_O^2 denote the sample variance in difference in SBP for controls and oral contraceptive users. Because we assume normality, we can create a 95% confidence interval for OC users using the appropriate t-statistic:

$$\hat{\mu}_C \pm t_{0.975,16-1} * \sqrt{\frac{s_C^2}{n_C}}$$

```
mu.C <- 4
mu.O <- 11
s2.C <- (28)^2
s2.O <- (20)^2
n.C  <- 16
n.O  <- 16

# Create CI for OC users
mu.O + c(-1, 1) * qt(0.975, n.O-1) * sqrt(s2.O/n.O)
```

```
[1]  0.3427523 21.6572477
```

The 95% confidence interval of (0.34, 21.67) suggests that there is an increase in SBP among oral contraceptive users following a 2-year period of use as 0 is not within the CI. Plausible ranges for the true mean increase in blood pressure are values between 0.34 and 21.67.

b. Does the change in SBP over the two year period appear to differ between oral contraceptive users and controls? Create the relevant 95% confidence interval and interpret. Assume normality and a common variance.

Under assumptions of normality and common variance, the 95% confidence interval for the change in SBP over the two year period between OC users and controls is given by,

$$(\hat{\mu}_C - \hat{\mu}_O) \pm t_{0.975,n_C+n_O-2} * S_p \sqrt{\frac{1}{n_C} + \frac{1}{n_O}}$$

where the pooled sample variance is given by,

$$S_p^2 = \frac{(n_C - 1)S_C^2 + (n_O - 1)S_O^2}{n_C + n_O - 2}$$

```
s2.pooled <- ((n.C - 1)*s2.C + (n.O - 1)*s2.O)/(n.C + n.O - 2)

# Create CI for difference between controls and OC users
(mu.C - mu.O) + c(-1, 1)*qt(0.975, n.C + n.O - 2) * sqrt(s2.pooled) * sqrt(1/n.C + 1/n.O)
```

```
[1] -24.56829  10.56829
```

Since the 95% confidence interval contains 0, there may not be a difference in the change in SBP over the two year period between OC users and controls.