MSAN 504 - Homework 2

Andre Guimaraes Duarte

July 23, 2016

1

We want to verify if the tip amount exceeds 20 percent when the next day's weather is printed on the bill. We are testing the null hypothesis H_0 that the mean tip amount does not differ from normal (20 percent). The alternate hypothesis H_1 is that the mean for weather-inspired tips exceeds 20 percent.

H_0: $\mu = 20$

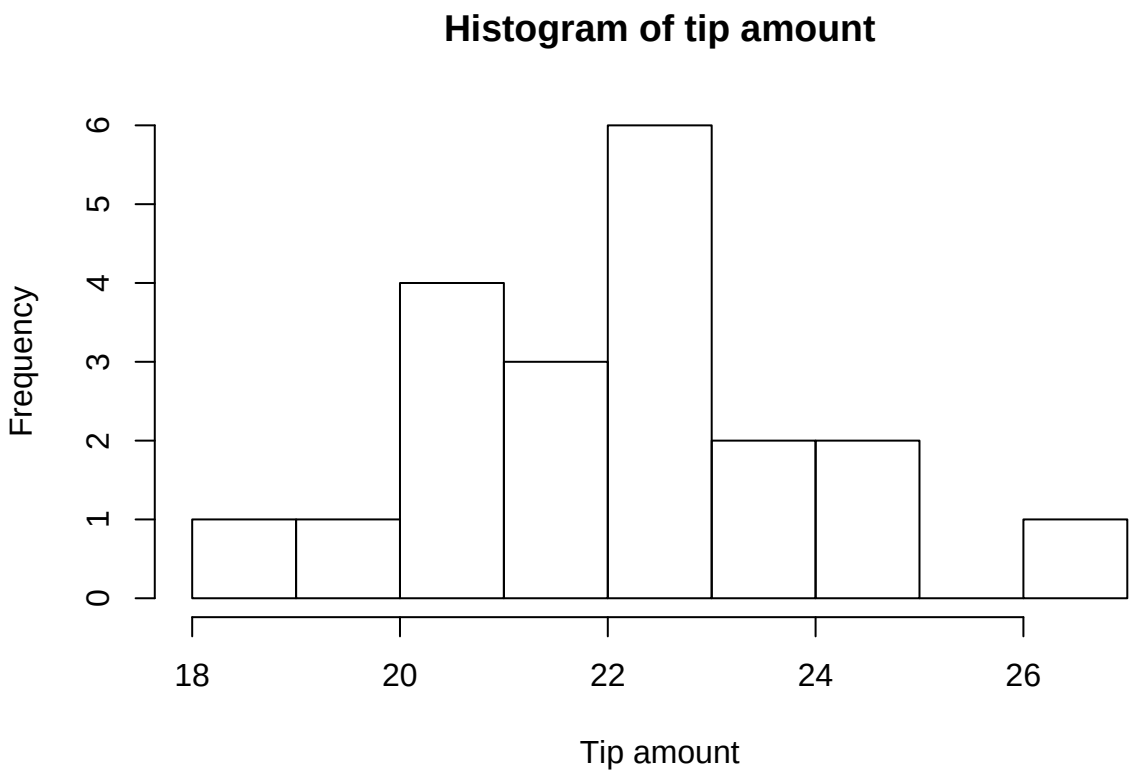
H_1: $\mu > 20$

We only have a sample of 20 tips, so we don't know the variance of the population. The test statistic we use is:

$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$, where $\bar{x}$ is the sample mean, μ_0 is the population mean (20 in this case), s is the sample standard deviation, and n is the sample size. In order to get the p-value of this test statistic, we use Student's t distribution with $n - 1$ degrees of freedom. Aditionally, we are doing a one-sided test.

We use R to plot a histogram of the tips, and calculate the test statistic, as well as the final p-value, with a printed output result.

```
alpha <- 0.06
tips <- c(20.8, 18.7, 19.9, 20.6, 21.9, 23.4, 22.8,
          24.9, 22.2, 20.3, 24.9, 22.3, 27.0, 20.3,
          22.2, 24.0, 21.1, 22.1, 22.0, 22.7)
hist(tips, main="Histogram of tip amount", xlab="Tip amount")
```




```
n <- length(tips)
mu_0 <- 20
x_bar <- mean(tips)
s <- sd(tips)

t <- (x_bar - mu_0)/(s/sqrt(n))
p <- pt(t, df = n-1, lower.tail = F)

if(p < alpha){
  print(paste("p = ", round(p, digits=5),
              " We reject the null hypothesis.", sep = ""))
} else{
  print(paste("p = ", round(p, digits=5),
              " We do not reject the null hypothesis.", sep = ""))
}
```

```
## [1] "p = 4e-05 We reject the null hypothesis."
```

Here, we have $\mu_0 = 20$, $\bar{x} = 22.205$, $s = 1.967$, $n = 20$, and $t = 5.012$.

We can see that we get a p-value $p = 4 \times 10^{-5} < \alpha$. Therefore, we reject the null hypothesis that the tip amount is equal to 20 percent when there is a weather message on the bill, and we accept the alternate hypothesis that the tip amount is greater than 20 percent.

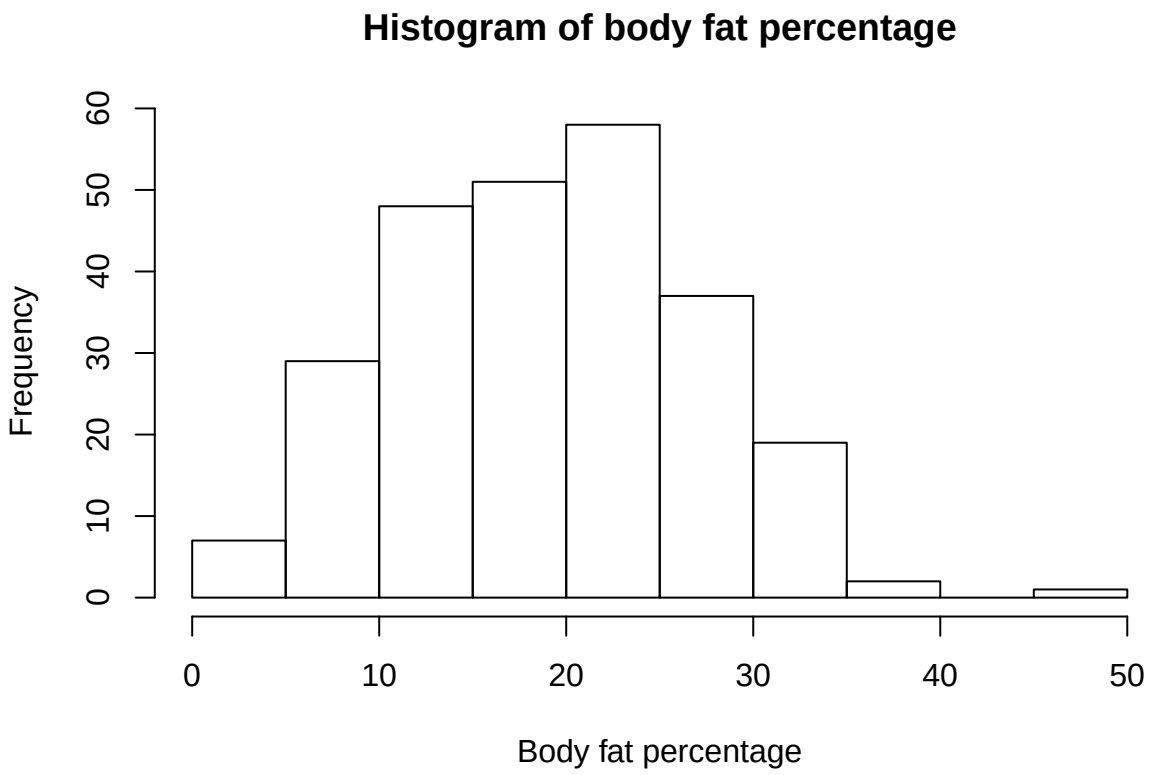
2

We want to compute the 85% confidence interval for the true mean body fat percentage of a male Mormon college student. We don't know the true population variance. We have a sample from BYU students, from which we can calculate the sample size n, the sample mean $\bar{x}$, and the sample standard deviation s. We have $\alpha = 0.15$. Since we don't know the population variance, we will be using a Student's distribution in order to calculate the confidence interval.

In order to get the confidence interval, we use the formula:

$P(\bar{x} - t^*_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t^*_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}) = 1-\alpha$, where $t^*_{\frac{\alpha}{2},n-1}$ is the quantile at $\frac{\alpha}{2}$ for a Student's t distribution with $n-1$ degrees of freedom.

```
bodyfat <- read.csv("BodyFatPercentage.csv")
hist(bodyfat$BODYFAT, main="Histogram of body fat percentage",
     xlab = "Body fat percentage")
```




```
alpha = 0.15
n = length(bodyfat$BODYFAT)
x_bar <- mean(bodyfat$BODYFAT)
s <- sd(bodyfat$BODYFAT)

t_star <- qt(1-alpha/2, df=n-1)
lower <- x_bar - t_star*s/sqrt(n)
upper <- x_bar + t_star*s/sqrt(n)
```

Here, we have $\bar{x} = 18.938$, $s = 7.751$, $n = 252$, and $t^*_{\frac{\alpha}{2},n-1} = 1.444$.

We get the 85% confidence interval `(18.233; 19.644)`, and we have $\bar{x} = 18.938 \in (18.233; 19.644)$.

The confidence interval for my particular sample is `(18.233; 19.644)`. Similarly-constructed intervals, computed over many different random samples, will contain the true mean body fat percentage μ with probability 85%.

3

A *paired t-test* is used to compare the means from two distinct populations when you have two paired samples, meaning that observations in one sample can be associated with observations in the second.

Here, we wish to test whether expert stock pickers outperform the Dow Jones Industrial Average (DJIA). The two samples are paired, since we have the monthly returns for the two samples. The null hypothesis H_0 is that experts do not outperfom the DJIA. The alternate hypothesis H_1 is that the experts have a higher monthly return that the DJIA. In other words, we are testing whether the difference of means is greater than 0. It is a one-sided test. Here, $\alpha = 0.05$.

$H_0 : \mu \leq mean(DJIA)$

$H_1 : \mu > mean(DJIA)$

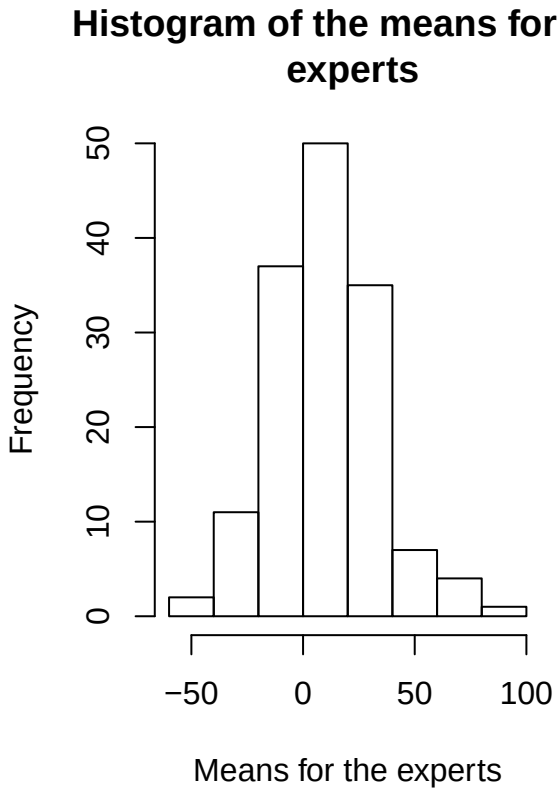
Comparing two means is the same as comparing their difference to 0. In order to perform this test, we first average the monthly return across the stock experts. Then, we calculate the differences between this average and the DJIA for each month. The mean of this measure is stored in the variable $\bar{d}$. We also need the standard deviation of the differences s, and the size of the sample n. With all these values, we can calculate our t-statistic using the formula:

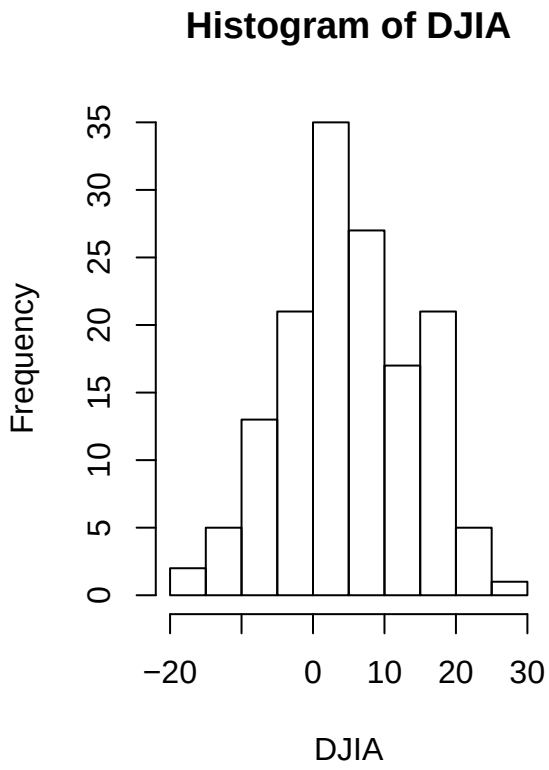
$t = \frac{\bar{d}-0}{\frac{s}{\sqrt{n}}}$

In order to get the p-value of this test statistic, we use Student's t distribution with $n-1$ degrees of freedom.

```
darts <- read.csv("DartsVersusExperts.csv")
darts$ExpertsMean <- rowMeans(darts[,c("Expert..1",
                                       "Expert..2",
                                       "Expert..3",
                                       "Expert..4")])
darts$Differences <- darts$ExpertsMean-darts$DJIA

par(mfrow=c(1,2))
hist(darts$ExpertsMean, main="Histogram of the means for the\n experts",
     xlab="Means for the experts")
hist(darts$DJIA, main="Histogram of DJIA", xlab="DJIA")
```






```
d_bar <- mean(darts$Differences)
s <- sd(darts$Differences)
n <- length(darts$Differences)
t <- d_bar/(s/sqrt(n))

p <- pt(t, n-1, lower.tail = F)

if(p < alpha){
  print(paste("p = ", round(p, digits=5),
              " We reject the null hypothesis.", sep = ""))
} else{
  print(paste("p = ", round(p, digits=5),
              " We do not reject the null hypothesis.", sep = ""))
}
```

```
## [1] "p = 0.00548 We reject the null hypothesis."
```

Here, we have $\bar{d} = 4.408$, $s = 20.742$, $n = 147$, and $t = 2.577$.

We can see that we get a p-value p $= 0.00548 < \alpha$. Therefore, we reject the null hypothesis that the stock experts do not outperform the DJIA, and we accept the alternate hypothesis that their performance is better than the DJIA.

To do the same t-test in R, we would use the following code, and get the output below:

```
t.test(darts$ExpertsMean, darts$DJIA, paired=T, alternative = "greater")
```

```
##
```

```
##  Paired t-test
##
## data:  darts$ExpertsMean and darts$DJIA
## t = 2.5767, df = 146, p-value = 0.005483
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.57621     Inf
## sample estimates:
## mean of the differences
##                4.408163
```

We can see that they are the same results that we got.

4

In this exercise, we are comparing the reading comprehension results by people who used two different types of lenses, Plano and Plus. We want to compute a 99% confidence interval for the true mean difference between the scores. Since the same people used the two lenses and performed the reading comprehension test with both types of lenses, the results are paired, and we can use a paired statistic to test the difference between the scores.

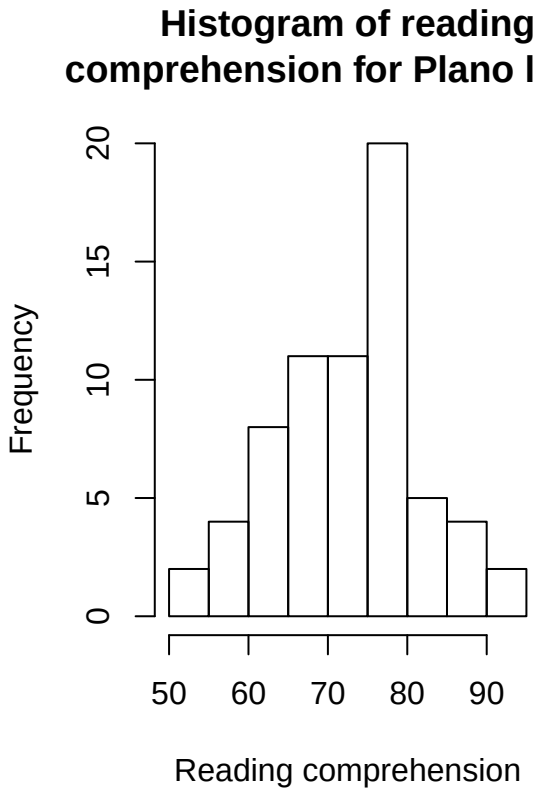
We don't know the true population variance for the difference in the scores. From the samples, we get the difference between the scores using either lenses. We have $\bar{d}$ the mean of these sample differences (Plus - Plano), s the standard deviation for the differences, and n the samples size. We have $\alpha = 0.01$. Since we don't know the population variance, we will be using a Student's distribution in order to calculate the confidence interval.

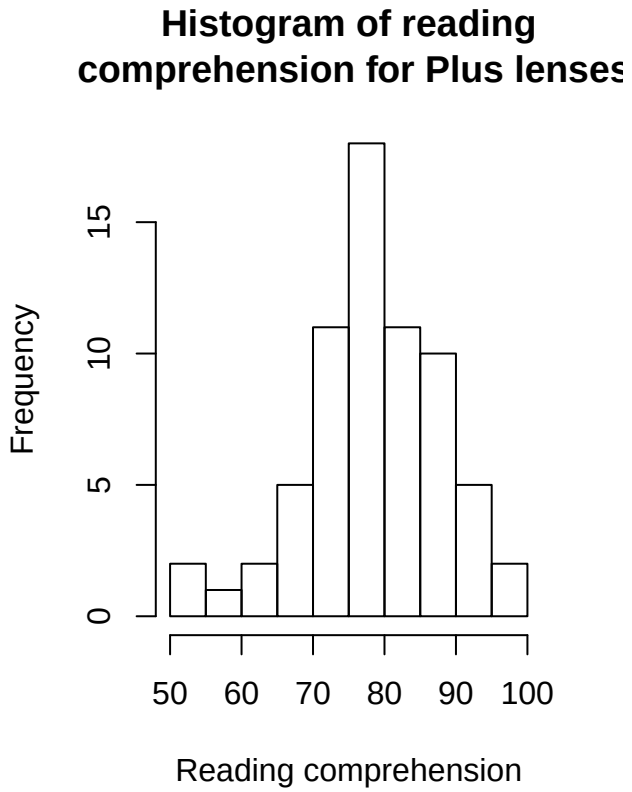
In order to get the confidence interval, we use the formula:

$P(\bar{d} - t^*_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{d} + t^*_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}) = 1-\alpha$, where $t^*_{\frac{\alpha}{2},n-1}$ is the quantile at $\frac{\alpha}{2}$ for a Student's t distribution with $n-1$ degrees of freedom.

```r
lenses <- read.csv("PlanoVersusPlusLenses.csv")
alpha = 0.01

par(mfrow = c(1, 2))
hist(lenses$Comp_Plano, main = "Histogram of reading\n comprehension for Plano lenses",
    xlab = "Reading comprehension")
hist(lenses$Comp_Plus, main = "Histogram of reading\n comprehension for Plus lenses",
    xlab = "Reading comprehension")
```






```
d_bar <- mean(lenses$d)
s <- sd(lenses$d)
n <- length(lenses$d)

t_star <- qt(1 - alpha/2, df = n - 1)
lower <- d_bar - t_star * s/sqrt(n)
upper <- d_bar + t_star * s/sqrt(n)
```

Here, we have $\bar{d} = 5.672$, $s = 13.254$, $n = 67$, and $t^*_{\frac{\alpha}{2},n-1} = 2.652$.

We get the 99% confidence interval `(1.377; 9.967)`, and we have $\bar{d} = 5.672 \in (1.377; 9.967)$. Plus lenses seem to give yield better test scores than Plano lenses.

The confidence interval for my particular sample is `(1.377; 9.967)`. Similarly-constructed intervals, computed over many different random samples, will contain the true mean reading comprehension difference between the two lenses μ with probability 99%.

5

Here, we want to build a 95% confidence interval for the true proportion of times the coin will come up heads. We have the number of trials $n = 34$, the sample probability of the coin coming heads up $\bar{p} = \frac{15}{34} \simeq 0.441$. We have $n\bar{p} = 15 \geq 10$ and $n(1 - \bar{p}) = 19 \geq 10$, which means that we can use the Demoivre-Laplace theorem. In particular, we get

$\bar{X} \sim N(p, \frac{p(1-p)}{n})$, where $X_i \sim Ber(p)$, and $X = X_1 + \ldots + X_n$.

The confidence interval is then found by using the formula:

$P(\bar{p} - Z^*_{\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \leq \pi \leq \bar{p} + Z^*_{\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}) = 1 - \alpha$, where π is the true proportion of the time this coin will come up heads.

```
heads <- 15
n <- 34
p_bar <- heads/n
s <- sqrt(p_bar*(1-p_bar))

alpha = 0.05

z_star <- qnorm(1-alpha/2)
lower <- p_bar - z_star*s/sqrt(n)
upper <- p_bar + z_star*s/sqrt(n)
```

Here, we have $\bar{p} = 0.441$, $s = 0.497$, $n = 34$, and $Z^*_{\frac{\alpha}{2}} = 1.96$.

We get the 95% confidence interval `(0.274; 0.608)`, and we have $\bar{p} = 0.441 \in (0.274; 0.608)$.

The confidence interval for my particular sample is `(0.274; 0.608)`. Similarly-constructed intervals, computed over many different random samples, will contain the true proportion of time this coin will come up heads with probability 95%.

6

In problem #5, we obtained a 95% confidence interval `(0.274; 0.608)` that is symmetrical. At 5% significance, the null hypothesis $H_0 : \pi = 0.75$ would be rejected in favor of its two-sided alternative $H_1 : \pi \neq 0.75$ because $\pi = 0.75 \notin (0.274; 0.608)$.

7

In this problem, we have $\mu = 500$, $\sigma = 100$ (so $\sigma^2 = 10,000$), and $n = 35$. $X_1, \ldots, X_{35}$ is a random sample from the population. The central limit theorem states that

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

We want $P(460 \leq \bar{X} \leq 540)$.

Since, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, then we also have $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$.

We have $P(460 \leq \bar{X} \leq 540) = P(\frac{460-\mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{540-\mu}{\frac{\sigma}{\sqrt{n}}})$

```
mu <- 500
sigma <- 100
n <- 35

lower <- (460 - mu)/(sigma/sqrt(n))
```

```
upper <- (540 - mu)/(sigma/sqrt(n))

prob <- integrate(dnorm, lower, upper)
```

With R, we get $P(460 \leq \bar{X} \leq 540) = P(-2.366 \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq 2.366)$. To find this probability, we integrate the standard normal distribution curve from the lower bound to the upper bound to get $P(460 \leq \bar{X} \leq 540) = 0.982$. In other words, there is a 98.2% that $\bar{X}$ is between 460 and 540.

8

We have $\mu = \mathrm{E}[X] = \int_1^3 xf(x)dx$, and $\mathrm{Var}(X) = \int_1^3 (x - \mathrm{E}[X])^2 f(x)dx$. We can use R to quickly calculate those values for us:

```
f <- function(x){return((x/9)*(x+5/2))}
x_bar <- integrate(f, 1, 3)$value
g <- function(x){return((x-x_bar)^2 * (x+5/2)/9)}
var_x <- integrate(g, 1, 3)$value
sigma <- sqrt(var_x)
```

We get $\mu = 2.074$ and $\mathrm{Var}(X) = 0.328$ ($\sigma = 0.573$). In addition, we have $n = 24$.

The central limit theorem states that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

We want $P(2 \leq \bar{X} \leq 2.15)$.

Since, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, then we also have $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$.

We have $P(2 \leq \bar{X} \leq 2.15) = P(\frac{2-\mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{2.15-\mu}{\frac{\sigma}{\sqrt{n}}})$

```
n <- 24
mu <- x_bar

lower <- (2 - mu)/(sigma/sqrt(n))
upper <- (2.15 - mu)/(sigma/sqrt(n))

prob <- integrate(dnorm, lower, upper)
```

With R, we get $P(2 \leq \bar{X} \leq 2.15) = P(-0.634 \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq 0.65)$. To find this probability, we integrate the standard normal distribution curve from the lower bound to the upper bound to get $P(2 \leq \bar{X} \leq 2.15) = 0.479$. In other words, there is a 47.9% that $\bar{X}$ is between 2 and 2.15.

9

We have $f(x) = \frac{\alpha}{x^{\alpha+1}}$ for $\alpha > 0$ and $x > 1$. If $x < 1$, $f(x) = 0$. The likelihood L is:

$L(\alpha; x_1, \ldots, x_n) = \prod_{i=1}^n f(\alpha; x_i) = \prod_{i=1}^n \frac{\alpha}{x_i^{\alpha+1}}$.

The log-likelihood is:

$$\begin{aligned} l(\alpha; x_1, \ldots, x_n) &= \sum_{i=1}^n \log f(\alpha; x_i) \\ &= \sum_{i=1}^n \log \frac{\alpha}{x_i^{\alpha+1}} \\ &= \sum_{i=1}^n \log \alpha - \sum_{i=1}^n \log x_i^{\alpha+1} \\ &= n \log \alpha - \sum_{i=1}^n (\alpha+1) \log x_i \\ &= n \log \alpha - n(\alpha+1) \sum_{i=1}^n \log x_i \end{aligned}$$

Therefore, $\frac{\partial l}{\partial \alpha} = \frac{n}{\alpha} - n \sum_{i=1}^n x_i$.

$\frac{\partial l}{\partial \alpha} = 0 \Leftrightarrow \frac{n}{\alpha} - n \sum_{i=1}^n x_i = 0 \Rightarrow \hat{\alpha} = \frac{1}{\sum_{i=1}^n \log x_i}$.

We just need to verify that this value is indeed a maximum:

$\frac{\partial^2 l}{\partial \alpha^2} = -\frac{n}{\alpha^2} < 0$. It is indeed a maximum.

10

The probability mass function of a Poisson distribution is:

$p(\lambda; x) = \frac{e^{-\lambda} \lambda^x}{x!}$ for $x \in \mathbb{N}$.

The log-likelihood is:

$$\begin{aligned} l(\lambda; x_1, \ldots, x_n) &= \sum_{i=1}^n \log f(\lambda; x_i) \\ &= \sum_{i=1}^n \log \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \\ &= \sum_{i=1}^n \log e^{-\lambda} \lambda^{x_i} - \sum_{i=1}^n \log x_i! \\ &= \sum_{i=1}^n \log e^{-\lambda} + \sum_{i=1}^n \log \lambda^{x_i} - \sum_{i=1}^n \log x_i! \\ &= \sum_{i=1}^n -\lambda + \sum_{i=1}^n x_i \log \lambda - \sum_{i=1}^n \log x_i! \\ &= -n\lambda + \log \lambda \sum_{i=1}^n x_i - \sum_{i=1}^n \log x_i! \\ &= -n\lambda + \log \lambda n \bar{X} - \sum_{i=1}^n \log x_i! \end{aligned}$$

Therefore, $\frac{\partial l}{\partial \lambda} = -n + \frac{n\bar{X}}{\lambda}$.

$\frac{\partial l}{\partial \lambda} = 0 \Leftrightarrow -n + \frac{n\bar{X}}{\lambda} = 0 \Rightarrow \hat{\lambda} = \bar{X}$.

We just need to verify that this value is indeed a maximum:

$\frac{\partial^2 l}{\partial \lambda^2} = -\frac{n\bar{X}}{\lambda^2} < 0$. It is indeed a maximum.

11

We have $f(x; \theta) = \frac{1}{\theta^2} x e^{-\frac{x}{\theta}}$ for $x > 0$ and 0 otherwise. The log-likelihood is:

$$
\begin{aligned}
l(\theta; x_1, \ldots, x_n) &= \sum_{i=1}^{n} \log f(\theta; x_i) \\
&= \sum_{i=1}^{n} \log \tfrac{1}{\theta^2} x_i e^{-\frac{x_i}{\theta}} \\
&= \sum_{i=1}^{n} \log \tfrac{1}{\theta^2} + \sum_{i=1}^{n} \log x_i + \sum_{i=1}^{n} \log e^{-\frac{x_i}{\theta}} \\
&= -\sum_{i=1}^{n} \log \theta^2 + \sum_{i=1}^{n} \log x_i - \sum_{i=1}^{n} \tfrac{x_i}{\theta} \\
&= -2\sum_{i=1}^{n} \log \theta + \sum_{i=1}^{n} \log x_i - \tfrac{n\bar{X}}{\theta} \\
&= -2n \log \theta + \sum_{i=1}^{n} \log x_i - \tfrac{n\bar{X}}{\theta}
\end{aligned}
$$

Therefore, $\frac{\partial l}{\partial \theta} = -\frac{2n}{\theta} + \frac{n\bar{X}}{\theta^2}$.

$\frac{\partial l}{\partial \theta} = 0 \Leftrightarrow -\frac{2n}{\theta} + \frac{n\bar{X}}{\theta^2} = 0 \Leftrightarrow \hat{\theta} = \theta(2\theta - \bar{X}) = 0 \Leftrightarrow \hat{\theta} = 0$ or $\hat{\theta} = \frac{\bar{X}}{2}$. Since $\hat{\theta}$ cannot be 0 (by definition of the probability density function), we get $\hat{\theta} = \frac{\bar{X}}{2}$.

We just need to verify that this value is indeed a maximum:

$\frac{\partial^2 l}{\partial \theta^2} = \frac{2n}{\theta^2} - \frac{2n\bar{X}}{\theta^3}$. By plugging in our candidate $\hat{\theta}$ into the equation, we get:

$\frac{\partial^2 l}{\partial \theta^2} = \frac{2n}{(\frac{\bar{X}}{2})^2} - \frac{2n\bar{X}}{(\frac{\bar{X}}{2})^3} = -\frac{8n}{\bar{X}^2} < 0$. It is indeed a maximum.

12

We want to test the null hypothesis $H_0 : \rho = 0$ against the alternative hypothesis $H_1 : \rho > 0$, where ρ is the correlation between body fat percentage and age. The plot gives us the knowledge that the one-sided alternative hypothesis should be $\rho > 0$ (there seems to be a positive correlation between the two variables).

```
bodyfat <- read.csv("BodyFatPercentage.csv")
plot(bodyfat$AGE, bodyfat$BODYFAT, main = "Body fat percentage as a function of age",
    xlab = "Age (years) ", ylab = "Body fat percentage (%)")
```


```
rho <- cor(bodyfat$AGE, bodyfat$BODYFAT)
```

Here, we have $\rho = 0.289$.

- Using the perspective of Pearson, we calculate the statistic $t^* = \frac{\rho\sqrt{n-2}}{\sqrt{(1-\rho^2)}}$, and then we use this statistic to calculate a p-value. We compare this statistic to a Student's t distribution with $n-2$ degrees of freedom.

```
alpha <- 0.01
n <- length(bodyfat$AGE)

t_star <- (rho * sqrt(n-2))/(sqrt(1 - rho^2))

pvalue <- pt(t_star, n-2, lower.tail = F)

if(pvalue < alpha){
  print(paste("p = ", round(pvalue, digits=7),
              " We reject the null hypothesis.", sep = ""))
} else{
  print(paste("p = ", round(pvalue, digits=7),
              " We do not reject the null hypothesis.", sep = ""))
}
```

```
## [1] "p = 1.5e-06 We reject the null hypothesis."
```

Here, we have $n = 252$ and $t^* = 4.776$.

We can see that we get a p-value $p = 1.5 \times 10^{-6} < \alpha$. Therefore, using the perspective of Pearson, we reject the null hypothesis that the correlation is 0. Age and body fat percentage have a positive correlation.

- Using the perspective of Fisher, since we are testing $\rho_0 = 0 < 0.55$, we have $Z = \frac{\hat{\rho}-\rho_0}{\sqrt{\frac{1}{n-3}}} \sim N(0,1)$. We then test obtain a p-value using this statistic in a standard normal distribution.

```
alpha <- 0.01
n <- length(bodyfat$AGE)

z <- (rho - 0)/(sqrt(1/(n-3)))

pvalue <- pnorm(z, lower.tail = F)

if(pvalue < alpha){
  print(paste("p-value = ", round(pvalue, digits=7),
              " We reject the null hypothesis.", sep = ""))
} else{
  print(paste("p-value = ", round(pvalue, digits=7),
              " We do not reject the null hypothesis.", sep = ""))
}
```

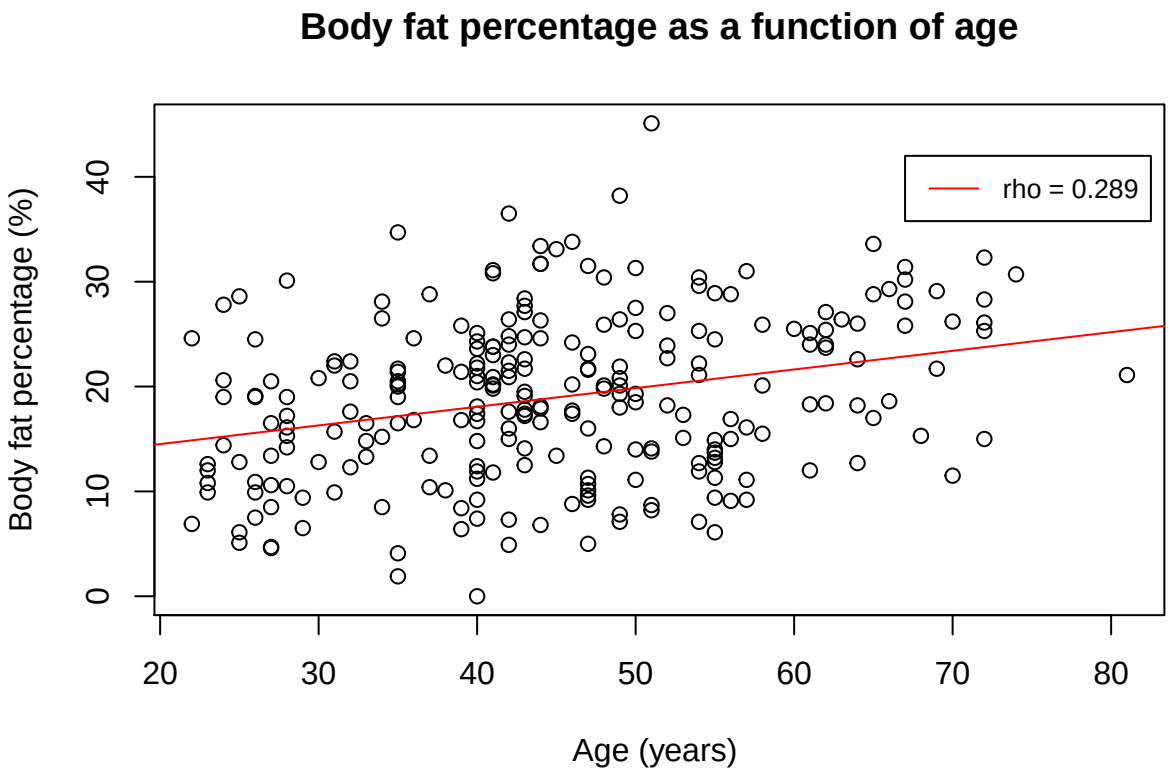
```
## [1] "p-value = 2.5e-06 We reject the null hypothesis."
```

Here, we have $n = 252$ and $z = 4.563$.

We can see that we get a p-value $\text{p} = 2.5 \times 10^{-6} < \alpha$. Therefore, using the perspective of Fisher, we also reject the null hypothesis that the correlation is 0. Age and body fat percentage have a positive correlation.

We can use R to plot the linear regression according to this correlation value.

```
lm1 <- lm(bodyfat$BODYFAT ~ bodyfat$AGE)
plot(bodyfat$AGE, bodyfat$BODYFAT, main = "Body fat percentage as a function of age",
    xlab = "Age (years) ", ylab = "Body fat percentage (%)")
abline(lm1$coefficients[1], lm1$coefficients[2], col = "red")
legend(67, 42, paste("rho = ", round(rho, 3), sep = ""), col = "red", lty = 1,
    cex = 0.85)
```



13

Using the same data set as the previous problem, we now want to compute an approximate 80% confidence interval for the true correlation between age and body fat percentage. Taking the approach suggested by Fisher, we will have to use the Fisher transformation.

The Fisher transformation is given by $F(x) = \frac{1}{2}\log\left(\frac{1+x}{1-x}\right)$.

We also have its inverse, which is important to compute the confidence interval: $F^{-1}(x) = \frac{e^{2x}-1}{e^{2x}+1}$.

Since $|\hat{\rho}| < 0.55$, we don't need to go through the Fisher transformation. In order to compute the confidence interval, we have $\hat{\rho} \pm Z^*_{\frac{\alpha}{2}}\sqrt{\frac{1}{n-3}}$.

```
alpha <- 0.2
n <- length(bodyfat$AGE)
rho_hat <- rho

z_star <- qnorm(1-alpha/2)

lower <- rho - z_star/sqrt(n-3)
upper <- rho + z_star/sqrt(n-3)
```

Here, we have $Z^*_{\frac{\alpha}{2}} = 1.282$.

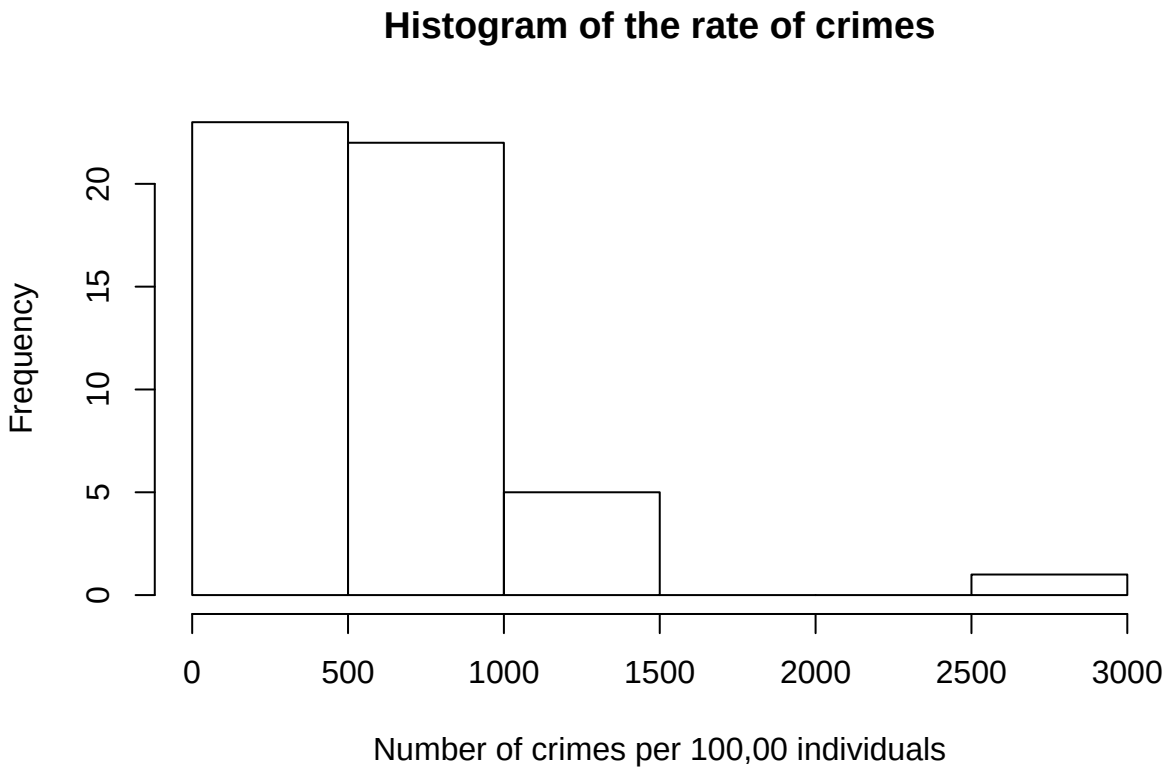
We get the 80% confidence interval `(0.208; 0.37)`, and we have $\hat{\rho} = 0.289 \in (0.208; 0.37)$.

The confidence interval for my particular sample is `(0.208; 0.37)`. Similarly-constructed intervals, computed over many different random samples, will contain the true correlation between age and body fat percentage ρ with probability 80%.

14

Here, we want to test the null hypothesis $H_0 : \sigma^2 = 185,000$, i.e. the true population variance of the rate of crimes (per 100,000 individuals) is equal to 185,000 at the 10% level of significance. The alternate hypothesis is $H_1 : \sigma^2 \neq 185,000$, which is two-sided.

```
crimes <- read.csv("AgrestiFinlayCrime.csv")
hist(crimes$crime, main="Histogram of the rate of crimes",
     xlab="Number of crimes per 100,00 individuals")
```



For a random sample of size n of normal random variables with variance s^2, then $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$, where σ^2 is the desired population variance. The test statistic is $t = \frac{(n-1)s^2}{\sigma_0^2}$, where σ_0 is the value for which we are testing the hypothesis (in this case 185,000).

```
alpha <- 0.1
n <- length(crimes$crime)
sigma2 <- 185000
s2 <- var(crimes$crime)

t <- (n-1)*s2/sigma2
```

```r
p <- pchisq(t, n-1, lower.tail = F)*2 # two-sided

if(p < alpha){
  print(paste("p = ", round(p, digits=5),
              " We reject the null hypothesis.", sep = ""))
} else{
  print(paste("p = ", round(p, digits=5),
              " We do not reject the null hypothesis.", sep = ""))
}
```

```
## [1] "p = 0.7484 We do not reject the null hypothesis."
```

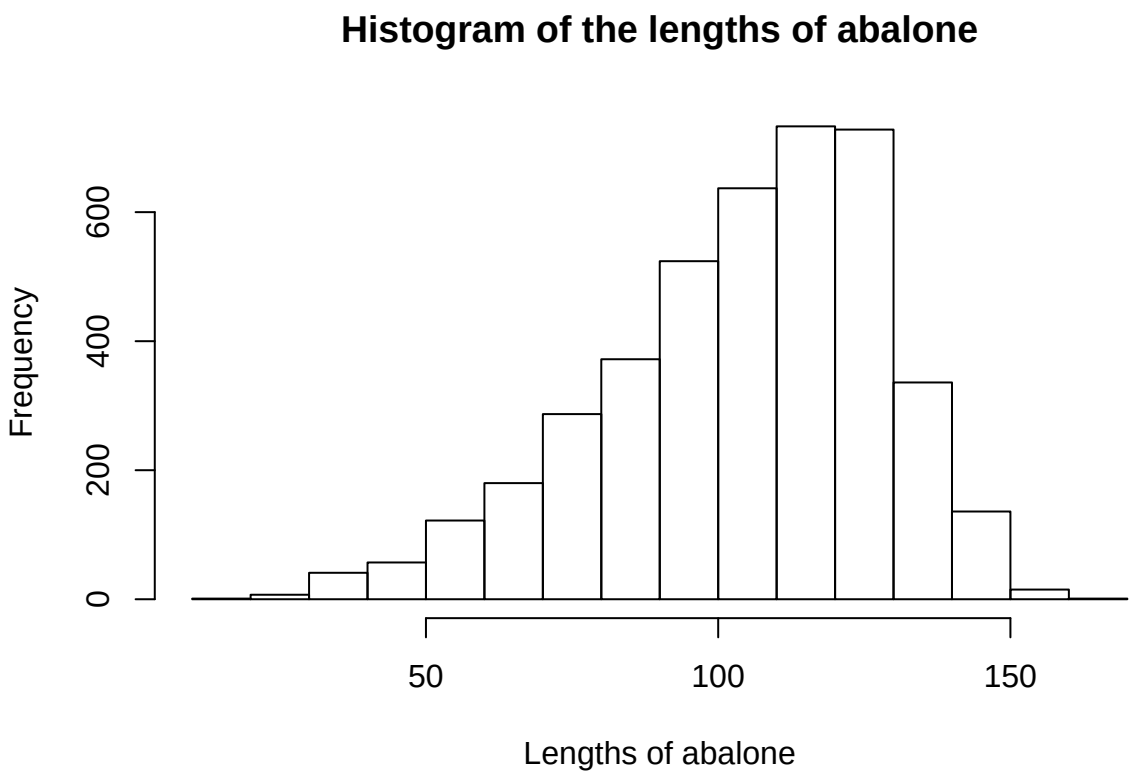
Here, we have $n = 51$, $s^2 = 194569$, $\sigma^2 = 185000$, and $t = 52.586$.

We can see that we get a p-value $p = 0.7484 > \alpha$. Therefore, we do not reject the null hypothesis at 10% level of confidence. At this point in time, there is not enough data to suggest the implausibility of $\sigma^2 = 185,000$, i.e. the true variance being equal to 185,000 is plausible.

15

We wish to compute the 90% confidence interval for the true variance of the lengths of abalone.

```r
abalone <- read.csv("Abalone.csv")
hist(abalone$LENGTH, main="Histogram of the lengths of abalone",
     xlab="Lengths of abalone")
```



For a random sample of size n of normal random variables with variance s^2, then $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$, where σ^2 is the desired population variance. The confidence interval is then found using the formula

$$P(\chi^{2^*}_{\frac{\alpha}{2},n-1} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^{2^*}_{1-\frac{\alpha}{2},n-1}) = 1-\alpha \Leftrightarrow P(\frac{(n-1)s^2}{\chi^{2^*}_{1-\frac{\alpha}{2},n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^{2^*}_{\frac{\alpha}{2},n-1}}) = 1-\alpha$$

```
alpha <- 0.1
n <- length(abalone$LENGTH)
s2 <- var(abalone$LENGTH)

chi_star_left <- qchisq(1-alpha/2, n-1)
chi_star_right <- qchisq(alpha/2, n-1)

lower <- (n-1)*s2/chi_star_left
upper <- (n-1)*s2/chi_star_right
```

Here, we have $\chi^{2^*}_{1-\frac{\alpha}{2},n-1} = 4327.45$ and $\chi^{2^*}_{\frac{\alpha}{2},n-1} = 4026.824$.

We get the 90% confidence interval `(556.702; 598.264)`, and we have $s^2 = 576.892 \in (556.702; 598.264)$.

The confidence interval for my particular sample is `(556.702; 598.264)`. Similarly-constructed intervals, computed over many different random samples, will contain the true variance of the lengths of abalone σ^2 with probability 90%.
