Homework 5

Nora Quick

2021-10-27

Instructions

Use the R Markdown version of this file to complete and submit your homework. Items in **bold** require an answer. Make sure you change the author in the header to your own name.

Conceptual Questions

While these questions are labelled "Conceptual" you may, and probably should, use R to answer them.

1. Ruchdeschel et al. Ruckdeschel, Shoop, and Kenney (2005) claim that the sex ratio of Ridley's sea turtle (a very rare and endangered sea turtle) moved from a male biased ratio to a female biased ratio.

They recorded the sex of stranded seas turtles on Cumberland Island. From 1983 to 1989 there were 16 males and 10 females. From 1990 to 2001 there were 19 males and 56 females.

Is there evidence that the sex ratio of Ridley's sea turtles was male biased in the period 1983-1989, and female biased in the period 1990-2001?

For each of this period, conduct an appropriate test, construct a confidence interval and write a summary with your conclusions in the context of the study.

```
old_ratio <- prop.test(x = 16, n = 26, p = 0.50, conf.level = 0.95, correct = FALSE)
old_ratio
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  16 out of 26, null probability 0.5
## X-squared = 1.3846, df = 1, p-value = 0.2393
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4253485 0.7757141
## sample estimates:
##         p
## 0.6153846
```

```
new_ratio <- prop.test(x = 19, n = 75, p = 0.50, conf.level = 0.95, correct = FALSE)
new_ratio
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  19 out of 75, null probability 0.5
## X-squared = 18.253, df = 1, p-value = 1.934e-05
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.1686005 0.3621033
## sample estimates:
##         p
## 0.2533333
```

The confidence interval from 1983 to 1989 is (0.43, 0.78) and the confidence interval from 1990 to 2001 is (0.17, 0.36).

From the confidence intervals above we can conclude that the sex ratio moved from male to female between the first interval of 1983-1989 and the second interval of 1990-2001. The first confidence interval has a median value at about 0.62 showing the male population to be roughly above 0.5 and with a p-value of about 0.24 we fail to reject the null hypothesis meaning that the population is skewed in the favour of males. The second confidence interval has a median value at about 0.26 showing the male population is now below 0.5 and with a p-value of 1.93-05 we can reject the null hypothesis that the population is skewed in the favour of males.

In conclution there is convincing evidence that between 1983-1989 and 1990-2001 the sex ratio of Ridley's sea turtles moved from a male biased ration to a female biased ratio.

2. *(From Ex 6. Chapter 4 Statistical Methods. Freund, R.; Mohr, D; Wilson,W. (2010))*

Average systolic blood pressure of a normal male is supposed to be about 129. Measurements of systolic blood pressure on a sample of 12 adult males from a community whose dietary habits are suspected of causing high blood pressure are (in R ready format):

```r
bp <- c(115, 134, 131, 143, 130, 154, 119, 137, 155, 130, 110, 138)
```

Do the data justify the suspicions regarding the blood pressure of this community?

Conduct an appropriate test, construct a confidence interval and write a summary with your conclusions in the context of the study.

```r
bp <- c(115, 134, 131, 143, 130, 154, 119, 137, 155, 130, 110, 138)

t.test(bp, mu = 129, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  bp
## t = 0.9939, df = 11, p-value = 0.3416
## alternative hypothesis: true mean is not equal to 129
## 95 percent confidence interval:
##  124.142 141.858
## sample estimates:
## mean of x
##       133
```

Justification of t.test: the learnr tutorial specifically uses blood pressure as an example in the t.test section.

The confidence interval of the above blood pressure data is (124, 142) and we find a p-value of 0.34. While the median value of the confidence interval is 133 (above the 129 that it's supposed to be) the p-value indicated that we fail to reject the null hypothesis that the normal male has a blood pressure of about 129. Therefore, there is moderately strong evidence that the blood pressure of a normal male is about 129.

3. *(Adapted From Ex 22. Chapter 4 Statistical Methods. Freund, R.; Mohr, D; Wilson,W. (2010))*

The following data gives the average pH in rain/sleet/snow for the two-year period 2004-2005 at 20 rural sites on the U.S. West Coast. (Source: National Atmospheric Deposition Program).

```
rain <- c(5.335, 5.345, 5.380, 5.520, 5.360, 6.285, 5.510, 5.340,
          5.395, 5.305, 5.190, 5.455, 5.350, 5.125, 5.340, 5.305,
          5.315, 5.330, 5.115, 5.265)
```

Is there evidence the median pH is not 5.4?

Conduct an appropriate test, construct a confidence interval and write a summary with your conclusions in the context of the study.

```
rain <- c(5.335, 5.345, 5.380, 5.520, 5.360, 6.285, 5.510, 5.340,
          5.395, 5.305, 5.190, 5.455, 5.350, 5.125, 5.340, 5.305,
          5.315, 5.330, 5.115, 5.265)

t.test(rain, mu = 5.4, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  rain
## t = -0.40957, df = 19, p-value = 0.6867
## alternative hypothesis: true mean is not equal to 5.4
## 95 percent confidence interval:
##  5.267102 5.489398
## sample estimates:
## mean of x
##   5.37825
```

In this data we look at if the pH level is 5.4 meaning the null hypothesis is that the pH level IS 5.4 so we are trying to disprove this to answer the question above.

The confidence interval from the above data is (5.3, 5.5) which gives us a median value of about 5.4. In addition to this the p-value is about 0.69 which means we would fail to reject the null hypothesis. In conclution there is strong evidence that the pH of the rain IS 5.4.

R Question

This question explores the difference between the Normal distribution and t-distribution as reference distributions for a two sample comparison.

Begin by setting the seed to 1908:

```r
set.seed(1908)
```

Then use `rexp()` to draw a sample of size 10 from an Exponential distribution with rate parameter 1:

```r
exp_sample <- rexp(n = 10, rate = 1)
```

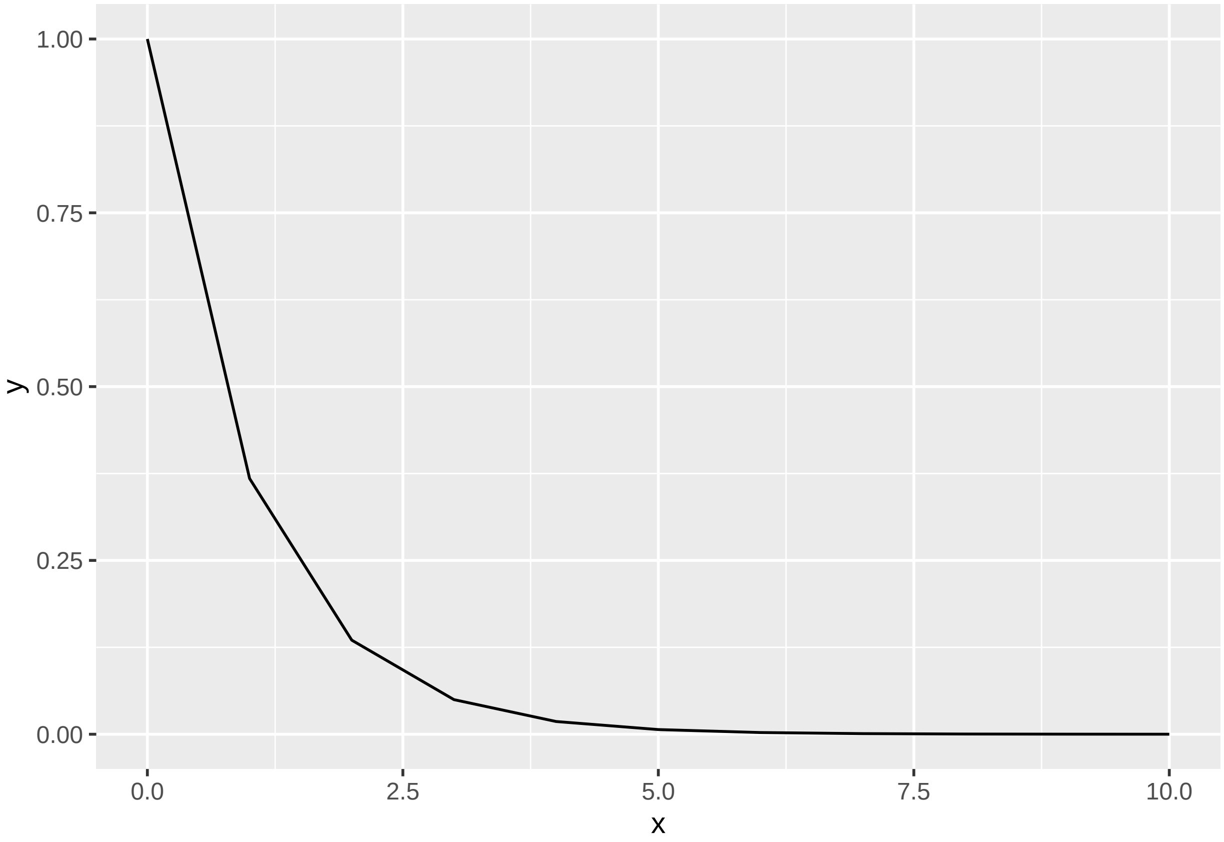
a) It is helpful to be able to picture the Exponential distribution, so follow the steps below to **plot the distribution function curve.**

- First you need to create a vector of x-axis values, called `x`. The function `seq()` creates a sequence, and has as arguments `from`, `to`, and `by`.
- Then you need to find the values of the Exponential(1) distribution at those x-axis values. The function `dexp(x, rate = 1)` gives the value of the Exponential(1) distribution for the values stored as the vector `x`. Store these values as `y`.
- Then use `qplot(x, y, geom = "line")` to create a plot. Remember to load the `ggplot2` package.

```r
library(ggplot2)

x <- seq(from = 0, to = 10, by = 1)
y <- dexp(x, rate = 1)

qplot(x, y, geom = "line")
```



Alternatively, you can use the base-R plotting function `plot(x, y, type="l")`.

b) Run a t-test on your size 10 sample, for a null hypothesis that $\mu = 2$, against a two sided alternative. **Write a non-technical summary that includes an interpretation of the p-value and 95% confidence interval.**

```
t.test(x, mu = 2, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 3, df = 10, p-value = 0.01334
## alternative hypothesis: true mean is not equal to 2
## 95 percent confidence interval:
##  2.771861 7.228139
## sample estimates:
## mean of x
##         5
```

With a p-value of 0.01 and a confidence interval of (2.77, 7.22) there is no evidence that the value equals 2.

c) **Calculate the t-statistic "by hand" using the formula on slide 9 from Module 5 Lecture 1 (in particular, use the sample SD, not population SD) for the same hypotheses as part b), but compute the p-value based on the normal distribution (i.e., assume the t-statistic follows a normal distribution, you can follow examples from Module 4 lab and homework).**

```
    x_bar <- mean(x)
    mu <- 2
    s <- sd(x)
    sq_n <- sqrt(10)

    t <- (x_bar - mu) / (s / sq_n)
    t
```

```
## [1] 2.860388
```

```
    p <- pnorm(t, mean = 0, sd = 1, lower.tail = FALSE)
    p
```

```
## [1] 0.002115616
```

d) **If the test statistic is the same for both tests, why is the p-value different?**

No, the t-value is 2.86 (2.9) so it is close the same t-value as above which was 3, however, the p-value is much different. The p-value is likely different because we did the calculation by hand and therefore we could specify certain aspects of the calculation causing a difference in p-value.

e) **Which is more appropriate in real life, where the population standard deviation is usually unknown?**

We would want to do the calculations by hand. Again, it would allow us more refined data control and when things like standard deviation are not known we are able to calculate them.

References

Ruckdeschel, Carol, C Robert Shoop, and Robert D Kenney. 2005. "On the Sex Ratio of Juvenile Lepidochelys Kempii in Georgia." *Chelonian Conservation and Biology* 4 (4): 858–61.