Homework 5

Rui Gao

2021-12-08

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.

```
#1983-1989
#Set male = 1, female = 0. The mean or successful rate should be 0.5
# H_0: is 0.5, H_a: is not 0.5
binom.test(x = 16, n = 26, p = 0.5, conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  16 and 26
## number of successes = 16, number of trials = 26, p-value = 0.3269
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4057075 0.7977398
## sample estimates:
## probability of success
##              0.6153846
```

The male came up 16 times in 26, giving an estimate of the true proportion of male of 0.62. There is no evidence that the male proportion is not 0.5 . With 95% confidence the true proportion of male is between 0.4 and 0.8.So, it's not male biased.

```r
#1990-2001
#Set male = 0, female = 1. The mean or successful rate should be 0.5.
# H_0: is 0.5, H_a: is not 0.5
#because we know the population mean is 0.5, use score test
prop.test(x = 56, n = 75, p = 0.5,
  conf.level = 0.95, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  56 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:
##  0.6378967 0.8313995
## sample estimates:
##         p
## 0.7466667
```

The female came up 56 times in 75, giving an estimate of the true proportion of male of 0.75. There is no evidence that the male proportion is 0.5 . With 95% confidence the true proportion of male is between 0.64 and 0.83.SO, it's female biased.

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.

H_0: mu = 129, H_a: mu != 129

```r
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
```

The data doesn't justify the suspicions regarding the blood pressure of this community.

We estimate the mean is 133, and with 95% confidence, the mean is between 124 and 141, and 129 is stay in the area.

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.

H_0: M = 5.4, H_a: M != 5.4

```
M_3 = 5.4;
x_3_less_M = rain < M_3;
n_3 = 20;
binom.test(sum(x_3_less_M), n_3, p=0.5);
```

```
##
##  Exact binomial test
##
## data:  sum(x_3_less_M) and n_3
## number of successes = 16, number of trials = 20, p-value = 0.01182
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.563386 0.942666
## sample estimates:
## probability of success
##                    0.8
```

```
median(rain);
```

```
## [1] 5.34
```

```
lower_index <- round(n_3/2 - (qnorm(0.975)*sqrt(n_3))/2)
upper_index <- round(n_3/2 + (qnorm(0.975)*sqrt(n_3))/2 + 1)
c(lower_index, upper_index)
```

```
## [1]  6 15
```

```
x_3_sorted = sort(rain);
x_3_sorted[c(lower_index, upper_index)]
```

```
## [1] 5.305 5.380
```

There is no evidence the population median for this quantity is 5.4. It is estimated the population median for this quantity is 5.34. With 95% confidence, the population median is between 5.305 and 5.38.

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:

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

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

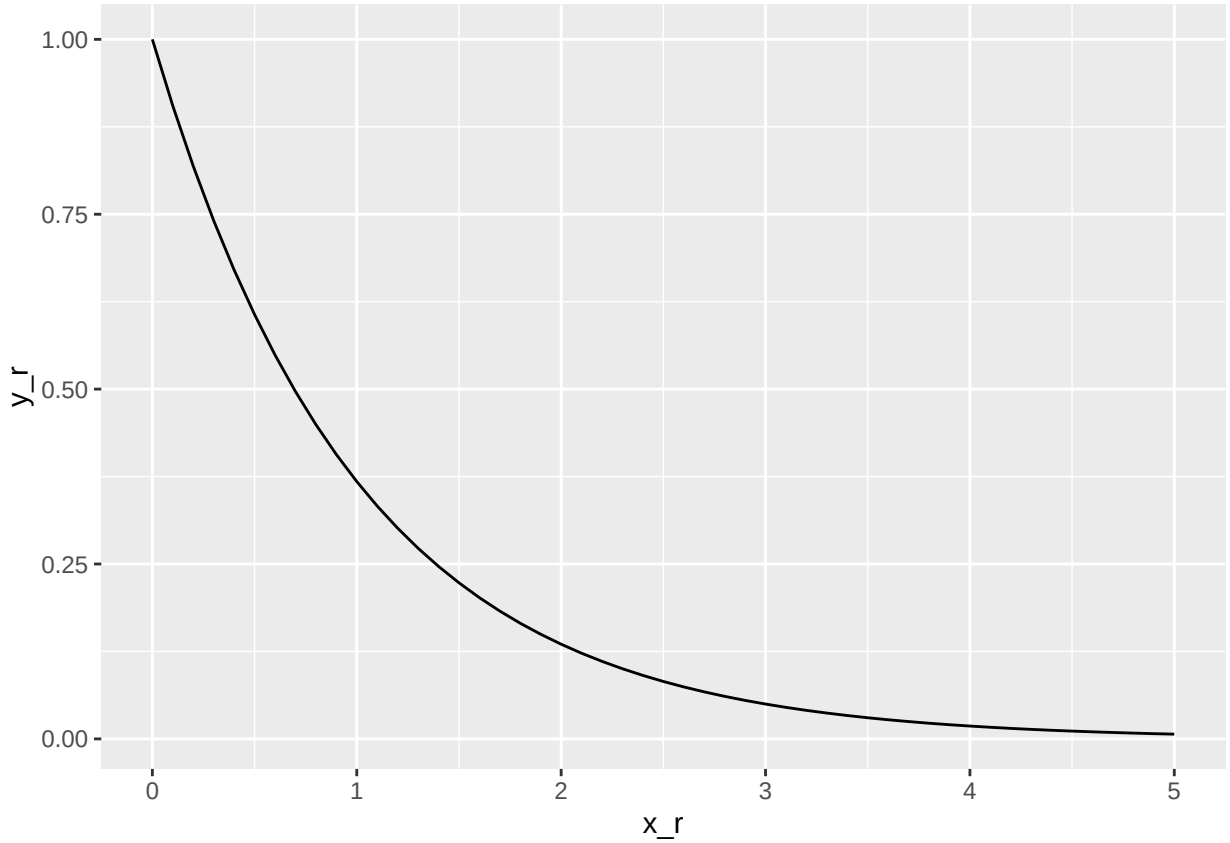
```
exp_sample <- rexp(n = 10, rate = 1)
```

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.

```
library(ggplot2)
x_r <- seq(from = 0, to = 5,by =0.1)
y_r <- dexp(x_r, rate = 1)

qplot(x_r, y_r, 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.**

```
#I don't understand the sentence "Run a t-test on your size 10 sample,"
#do you mean we can create a whatever sample, as long as the n = 10, for ourselves? If so...

set.seed(2021)
x_r_b <- rnorm(n = 10, mean = 5, sd = 10)
x_r_b
```

```
##  [1]   3.775400  10.524566   8.486495   8.596322  13.980537 -14.225695
##  [7]   7.617444  14.155664   5.137719  22.299632
```

```
#Our null and (two-sided) alternative hypotheses, in statistical notation, are:
#H_0: mu = 2, H_a: mu != 2
t.test(x_r_b, mu = 2, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  x_r_b
## t = 2.0175, df = 9, p-value = 0.07442
```

```
## alternative hypothesis: true mean is not equal to 2
## 95 percent confidence interval:
##   1.268223 14.801393
## sample estimates:
## mean of x
##  8.034808
```

There is convincing evidence the population mean is not equal to 2. We estimate the mean is 8.03, and with 95% confidence, the mean is between 1.27 and 14.8.

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).**

```
n_r_b <- 10
x_r_b_bar <- mean(x_r_b)
sd_r_b <- sqrt(sum((x_r_b - x_r_b_bar) * (x_r_b - x_r_b_bar))/(n_r_b))

#df = 9
Z <- (x_r_b_bar - 2) / (sd_r_b/sqrt(length(x_r_b)-1))
P <- 2 * pnorm(abs(Z), mean = 0, sd = 1, lower.tail = FALSE)
```

```
#CI
Z
```

```
## [1] 2.017515
```

```
P
```

```
## [1] 0.04364182
```

```
low <- x_r_b_bar - Z * sd_r_b/sqrt(length(x_r_b))
high <- x_r_b_bar + Z * sd_r_b/sqrt(length(x_r_b))
low
```

```
## [1] 2.309686
```

```
high
```

```
## [1] 13.75993
```

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

Because we used different standard deviation for calculate the z_alpha/2.

e) **Which is more appropriate in real life, where the population standard deviation is usually unknown?** The second, so we are able to own a standard deviation for using.

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.