Homework 4 Solution

Charlotte Wickham

2017-10-27

```
library(tidyverse)
library(pander)
```

Note this homework is due **Friday** at midnight.

Submit your answers on canvas.

1. Exact Binomial test

As we saw in lecture `binom.test()` performs an exact binomial test.

a. Use `binom.test()` to perform a two-sided test of the null hypothesis that $p = 0.6$ when you observe $X = 8$ successes in $n = 15$ trials. What is the reported p-value?

```
(b_test <- binom.test(x = 8, n = 15, p = 0.6))
```

```
##
##  Exact binomial test
##
## data:  8 and 15
## number of successes = 8, number of trials = 15, p-value = 0.6075
## alternative hypothesis: true probability of success is not equal to 0.6
## 95 percent confidence interval:
##  0.2658613 0.7873333
## sample estimates:
## probability of success
##              0.5333333
```

The p-value is 0.61.

b. Confirm that the p-value from (a) agrees with the method discussed in class: use `dbinom()` to find the probabilities of each outcome $X = 0, 1, 2, \ldots, 15$ and add up the probabilities of outcomes as likely as or less likely than $X = 8$ when $p = 0.6$.

```
x <- 0:15
probs <- dbinom(x, size = 15, prob = 0.6)
```

x	P(X = x)
0	**1.074e-06**
1	**2.416e-05**
2	**0.0002537**
3	**0.001649**
4	**0.00742**
5	**0.02449**
6	**0.06121**
7	**0.1181**
8	**0.1771**
9	0.2066

x	P(X = x)
10	0.1859
11	**0.1268**
12	**0.06339**
13	**0.02194**
14	**0.004702**
15	**0.0004702**

```
(p <- sum(probs[probs <= probs[x == 8]]))
```

```
## [1] 0.6074645
```

The value from summing up the probabilities from the Binomial(15, 0.6) is 0.61, and agrees with the p-vlaue from `binom.test()`.

c. Based on the p-value from part (a), would you reject the null hypothesis $H_0 : p = 0.6$ vs. $H_A : p \neq 0.6$ at level $\alpha = 0.05$?

Since the 0.61 $>$ 0.05, we fail to reject the null hypothesis that $p = 0.6$.

d. In class we found a rough confidence interval by inverting the exact test (Q8 from the worksheet on Oct 13th). Use `binom.test()` and a sequence of p_0 to emulate this procedure in R to find a 95% confidence interval for p up to 2 decimal places of precision.

```
# every 2 digit p between 0 and 1
p_0s <- seq(0, 1, 0.01)

p_values <- map_dbl(p_0s, ~ binom.test(x = 8, n = 15, p = .x)$p.value)

# Which tests fail to reject the null at alpha = 0.05?
p_values > 0.05
```

```
##   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
##  [34]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [45]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [56]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [67]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [78]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [89] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [100] FALSE FALSE
```

```
# Which values do these correspond to?
(in_interval <- p_0s[p_values > 0.05])
```

```
##  [1] 0.30 0.31 0.32 0.33 0.34 0.35 0.36 0.37 0.38 0.39 0.40 0.41 0.42 0.43
## [15] 0.44 0.45 0.46 0.47 0.48 0.49 0.50 0.51 0.52 0.53 0.54 0.55 0.56 0.57
## [29] 0.58 0.59 0.60 0.61 0.62 0.63 0.64 0.65 0.66 0.67 0.68 0.69 0.70 0.71
## [43] 0.72 0.73 0.74 0.75 0.76 0.77
```

The confidence interval is (0.30, 0.77).

2. Data Analysis

Using the same `brfss` data as in HW #3.

a) Estimate the proportion of US residents who are in the overweight or obese categories based on body mass index (BMI $= \frac{\text{weight in kg}}{(\text{height in m})^2} > 25$). **Write a statistical summary of your findings**

```
brfss <- read_rds("brfss.rds")
brfss <- brfss %>%
  mutate(bmi = (weight_kg) / (height_m^2),
    desired_loss = weight_kg - wtdesire_kg)

# table() will ignore NA's by default
table(brfss$bmi >= 25)
```

```
##
## FALSE  TRUE
##  1175  1683
```

Either find $\hat{p}$ and confidence interval "by hand" with:

$$\hat{p} = \frac{1683}{1683 + 1175} = \frac{1683}{2858} = 0.589$$

then CI with

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.589 \pm 1.96\sqrt{\frac{0.589(1-0.589)}{2858}} = (0.571, 0.607)$$

Or use R (ignore test part of the output and just get CI)

```
(prop_test_bmi <- prop.test(x = sum(brfss$bmi >= 25, na.rm = TRUE),
          n = sum(!is.na(brfss$bmi)),
          correct = FALSE))
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  sum(brfss$bmi >= 25, na.rm = TRUE) out of sum(!is.na(brfss$bmi)), null probability 0.5
## X-squared = 90.295, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5707266 0.6067815
## sample estimates:
##         p
## 0.5888733
```

```
ci <- prop_test_bmi$conf.int
estimate <- prop_test_bmi$estimate

# Exact test is also OK to use here
binom.test(x = sum(brfss$bmi >= 25, na.rm = TRUE),
          n = sum(!is.na(brfss$bmi)))
```

```
##
##  Exact binomial test
##
## data:  sum(brfss$bmi >= 25, na.rm = TRUE) and sum(!is.na(brfss$bmi))
```

```
## number of successes = 1683, number of trials = 2858, p-value <
## 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.5705758 0.6069886
## sample estimates:
## probability of success
##              0.5888733
```

Statistical Summary

It is estimated that 59% of US residents fall in the overweight or obsese categories based on BMI.

With 95% confidence between 57% and 61% of US residents fall in the overweight or obsese categories based on BMI.

b) Is the median desired weight loss greater than zero for US females? What is a likely range for the median desired weight loss. **Conduct the appropriate analyses and write a statistical summary of your findings** (Hint: if observed values take the exact value of the hypothesized median, you will need to drop them before proceeding with the sign test.)

$H_0 : M = 0$ versus $H_A : M > 0$ where M is the population median desired weight loss for females.

```
females <- filter(brfss, sex == "female")

# any exact zeros?
sum(females$desired_loss == 0, na.rm = TRUE)
```

```
## [1] 335
```

```
# yes, 335, remove just for p-value calculation
females_sub <- filter(females,
  desired_loss != 0,
  !is.na(desired_loss))
```

Find sample proportion that are less than hypothesized median:

```
x_less_0 <- sum(females_sub$desired_loss < 0)
n_sub <- length(females_sub$desired_loss)
(phat_loss <- x_less_0/n_sub)
```

```
## [1] 0.0392302
```

Then test hypothesis $H_0 : p = 0.5$ now with a lesser alternative (if the median is much greater than 0, we expect to see a proportion much less than 0.5, since the distribution is shifted to the right):

```
# binom.test() also OK
(sign_test <- prop.test(x = x_less_0,
  n = n_sub,
  p = 0.5,
  alternative = "less"))
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  x_less_0 out of n_sub, null probability 0.5
## X-squared = 1145.5, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is less than 0.5
## 95 percent confidence interval:
##  0.00000000 0.04928736
```

```
## sample estimates:
##         p
## 0.0392302
```

```
p_val_sign_test <- sign_test$p.value
```

The sample median is a good point estimate of the population median, but now be sure to **include the zero cases**:

```
(est_med <- median(females$desired_loss, na.rm = TRUE))
```

```
## [1] 6.802721
```

A 95% confidence interval can be found using the formula from class:

$$\left(\left(\frac{n - z_{1-\alpha/2}\sqrt{n}}{2} \right)^{\text{th}} \text{ smallest observation}, \left(\frac{n + z_{1-\alpha/2}\sqrt{n}}{2} + 1 \right)^{\text{th}} \text{ smallest observation} \right)$$

In this case, the observations:

```
n <- sum(!is.na(females$desired_loss))
(lower <- round((n - qnorm(0.975)*sqrt(n)) / 2))
```

```
## [1] 803
```

```
(upper <- round((n + qnorm(0.975)*sqrt(n)) / 2 + 1))
```

```
## [1] 884
```

We need to sort the (non-missing) observations first, then pull out the corresponding observations:

```
sorted_loss <- sort(females$desired_loss[!is.na(females$desired_loss)])
(ci_med <- sorted_loss[c(lower, upper)])
```

```
## [1] 4.988662 6.802721
```

Statistical Summary

There is convincing evidence the median desired weight loss for US resident females is greater than zero (Sign test, one-sided p-value < 0.0001).

It is estimated the median desired weight loss for US resident females is 6.8 kgs.

With 95% confidence, the median desired weight loss for US resident females is between 5 and 6.8 kgs.

c) A colleague suggests testing that median desired loss is zero with a Wilcoxon Signed Rank test. Write a two to three sentence argument for why this is not appropriate in this case.

The Wilcoxon Signed Rank test is a poor test of the population median for skewed distributions becasue it is neither assymptotically exact nor consistent. The population distbrution of desired loss is likely very skewed (evidence of this is seen in the sample histogram) so we would not expect good performance from the wilcoxon rank sum test here.

3. The Approximate Binomial test

In lecture we saw both the "Exact Binomial Test" and the "Approximate Binomial Test". Recall you can get the approximate test p-value in R with `prop.test()` and the exact p-value from `binom.test()`.

We also saw a rough guideline, that if $np > 5$ and $n(1-p) > 5$, the approximate test should be a good approximation the the exact test.

Your task for this question is to investigate this guideline through simulation. Perform simulations to assess the Type I error rate of the two procedures for a range of np (i.e. set p and vary n, or set n and vary p). **Does the guideline seem reasonable?**

A good solution includes simulations for a range of np, a table of Type I errors rates for both the Exact and Approximate tests and a short summary of their observations.

Charlotte's simulations:

Table 2: Rejection rate for the Binomial tests with $\alpha = 0.05$

Sample size, n	Exact test	Approximate test	n * p
100	0.0149	0.0929	0.5
500	0.0424	0.0424	2.5
1000	0.0358	0.0358	5
5000	0.0445	0.0552	25
10000	0.05	0.0587	50

Looks like the rejection rates are pretty close for both tests at all sample sizes. The tests seem further from the ideal rejection rates for smaller np (mostly because the test statistic only takes on a few values in these cases), but by $np = 5$, both seem close enough to 0.05 to be reasonable to use.