Homework 5 Solution

Charlotte Wickham

2017-11-12

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

1. Chi-square goodness of fit with estimated parameters

Cornhole is a popular lawn game in the US, where players throw a bean bag at a wooden platform with a hole in it. A bag in the hole scores 3 points, while one on the platform scores 1 point.

An avid cornhole analyst has observed n = 100 experienced players, and recorded the number of misses before they get a bag in the hole. These 100 observations can be read into R:

```
Y <- c(0L, 1L, 0L, 9L, 0L, 0L, 4L, 0L, 0L, 1L, 0L, 2L, 2L, 1L, 10L,
3L, 0L, 13L, 0L, 0L, 0L, 43L, 3L, 0L, 6L, 1L, 11L, 0L, 0L, 0L,
3L, 0L, 3L, 1L, 0L, 0L, 2L, 1L, 1L, 1L, 2L, 1L, 1L, 1L, 0L, 0L,
0L, 0L, 0L, 16L, 9L, 2L, 15L, 1L, 4L, 3L, 0L, 18L, 1L, 3L, 0L,
0L, 4L, 3L, 9L, 0L, 1L, 19L, 1L, 2L, 0L, 2L, 6L, 0L, 0L, 0L,
2L, 3L, 0L, 8L, 41L, 2L, 1L, 2L, 22L, 0L, 6L, 17L, 17L, 0L, 0L,
6L, 7L, 0L, 0L, 9L, 1L, 0L, 15L, 1L)
Y
```

(FYI the `L` just forces these numbers to be of integer type, it's not essential for this problem).

The analyst is curious if these values are consistent with being drawn from a Geometric distribution.

a. The geometric distribution has probability mass function:

$$P(Y = y) = (1 - p)^y p,$$

where p is an unknown parameter of the distribution. If $Y_1, \ldots, Y_n$ is an i.i.d sample from a Geometric(p) distribution then, a good estimate of p is

$$\hat{p} = \frac{1}{\overline{Y}}$$

Use the data to estimate p.

```
(phat <- 1/mean(Y))
```

```
## [1] 0.2463054
```

b. Tabulate the observed numbers of misses into the categories: 0, 1, 2, 3, 4, 5, 6+

```
Y <- factor(ifelse(Y >= 6, "6+", Y), levels = c(0:5, "6+"))
O <- table(Y)
pander(O)
```

0	1	2	3	4	5	6+
38	18	10	8	3	0	23

c. Find probabilities for each category above using the Geometric distribution with the estimated parameter

(you can use `dgeom()`).

```
probs <- dgeom(0:5, prob = phat)
(probs <- c(probs, 1 - sum(probs)))
```

```
## [1] 0.24630542 0.18563906 0.13991515 0.10545329 0.07947958 0.05990333
## [7] 0.18330418
```

d. Find the expected counts for each category using the probabilities found above.

```
n <- length(Y)
(E <- probs * n)
```

```
## [1] 24.630542 18.563906 13.991515 10.545329  7.947958  5.990333 18.330418
```

e. Check the condition for the Chi-square approximation to be appropriate.

```
all(E > 5)
```

```
## [1] TRUE
```

f. Calculate the value of the Chi-square statistic.

```
(X <- sum((O - E)^2/E))
```

```
## [1] 19.28735
```

g. Check your calculation of the test statistic by running `chisq.test(x = O, p = probs)`, where `O` comes from part (b) and `probs` from part (c).

```
chisq.test(O, p = probs)$statistic
```

```
## X-squared
##  19.28735
```

h. What distribution should this statistic be compared to? Find the p-value for the test that the number of misses are consistent with a Geometric distribution. What would you conclude?

```
1 - pchisq(X, df = length(O) - 1 - 1)
```

```
## [1] 0.001699035
```

i. Why does the `chisq.test()` run in (g) return the wrong p-value?

It doesn't know we estimated a parameter and therefore uses the wrong degrees of freedom.

2. Data Analysis

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

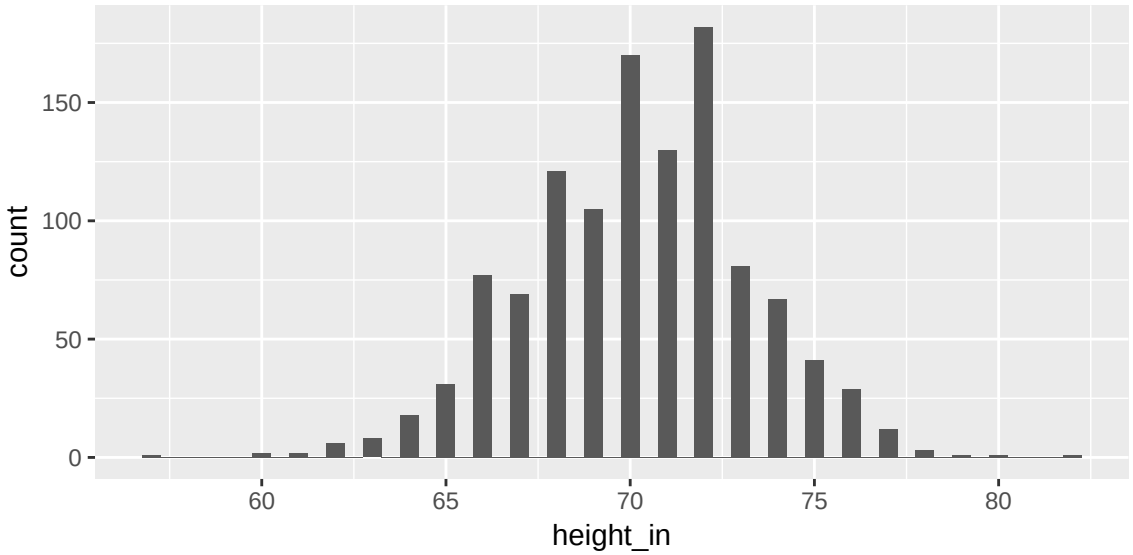
I found the following as the start of an example in a textbook: "The heights of male adults between the ages 20 and 62 in the US is nearly normal with mean 70.0 inches and standard deviation 3.0 inches."

a) Create histogram of the heights (in inches) of the male respondents. Describe the distribution, is there anything unusual about it?

```
download.file("http://st551.cwick.co.nz/data/brfss.rds",
  "brfss.rds", mode = "wb")
brfss <- read_rds("brfss.rds")
males <- filter(brfss, sex == "male")
males <- males %>%
  mutate(height_in = height_m * 39.3701)
```

```
ggplot(males, aes(x = height_in)) +
  geom_histogram(binwidth = 0.5)
```

```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```



While the distribution is roughly unimodal and symmetric, there does appear to be some effect of rounding. It seems more men report even heights (e.g. 68, 70, 72) than odd heights.

b) Conduct a t-test of variance, with the null hypothesis $H_0 : \sigma^2 = 3^2$, where σ is the population variance of heights in inches for male respondents to the BRFSS survey. Write your conclusion in the form of a statistical summary including a point estimate and confidence interval.

```
males <- males %>%
  mutate(height_devations = (height_in - mean(height_in, na.rm = TRUE))^2)
n <- sum(!is.na(males$height_devations))
(t_test_var <- t.test(males$height_devations, mu = (n-1)/(n) * 3.0^2))
```

```
##
##  One Sample t-test
##
## data:  males$height_devations
## t = 1.0995, df = 1157, p-value = 0.2718
## alternative hypothesis: true mean is not equal to 8.992228
## 95 percent confidence interval:
##   8.625698 10.293161
## sample estimates:
## mean of x
##   9.45943
```

There is no evidence that the variance of the height US males is not 9 inches^2 (t-test of variance, two-sided p-value $= 0.27$). It is estimated the variance of the height of US males is 9.5 inches^2. With 95% confidence the variance in height of US males is between 8.6 and 10.3 inches^2.

c) Test the hypothesis that the heights of the male respondents in the BRFSS survey come from a Normal$(70, 3^2)$ distribution.

Write your conclusion in the form of a statistical summary (there is no need for a point estimate or confidence interval).

```
ks.test(x = males$height_in, y = pnorm, mean = 70,
  sd = 3)
```

```
## Warning in ks.test(x = males$height_in, y = pnorm, mean = 70, sd = 3): ties
## should not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  males$height_in
## D = 0.12004, p-value = 6.439e-15
## alternative hypothesis: two-sided
```

There is convincing evidence the heights from US males do not follow a Normal distribution with mean 70 inches and variance 9 inches^2 (Kolmogorov-Smirnov test, p-value < 0.0001).

Really, this is telling us about the heights of US males, as they self report based on the question asked. In particular, it seems they are reporting in whole inches (essentially creating a discrete variable), and from the histogram above tend to round up to the nearest 2 inches (at least if they are under 76 inches). A better way to ask if the answers of these respondents is consistent with the claim that heights are Normal, might be to use a Chi-square goodness of fit test, using these reported height in bins of at least 2 inches.

(*Erin and Chuan*: An early typo had "N(70, 3.3)", e.g a variance of 3.3, anyone with that value, should still get full credit).

3. One-sided K-S tests

In lecture we saw an example where two one-sided K-S tests gave conflicting results.

The setup of that example was:

True population: $Y \sim N(0, 1)$ **Null Hypothesis:** $Y \sim N(0, 100)$

Replicate the example:

a) Draw a sample of size 20 from the true population.

```
y <- rnorm(n = 20, sd = 1)
```

b) Conduct tests of the null hypothesis with one-sided lesser, one-sided greater, and two-sided alternatives.

```
ks_lower <- ks.test(y, pnorm, mean = 0, sd = sqrt(100), alternative = "less")
ks_greater <- ks.test(y, pnorm, mean = 0, sd = sqrt(100), alternative = "greater")
ks_two <- ks.test(y, pnorm, mean = 0, sd = sqrt(100))

pander(
  matrix(c(ks_lower$p.value, ks_greater$p.value, ks_two$p.value)),
  row.names = c("Lesser", "Greater", "Two-sided"),
  caption = "P-values from the K-S testn under the three alternatives")
```

Table 2: P-values from the K-S testn under the three alternatives

Lesser	0.0004203
Greater	0.0004596
Two-sided	0.0008405

c) Plot the ECDF of the sample, along with the CDF of the hypothesized distribution. Indicate on your plot

where the test statistic for each test comes from.

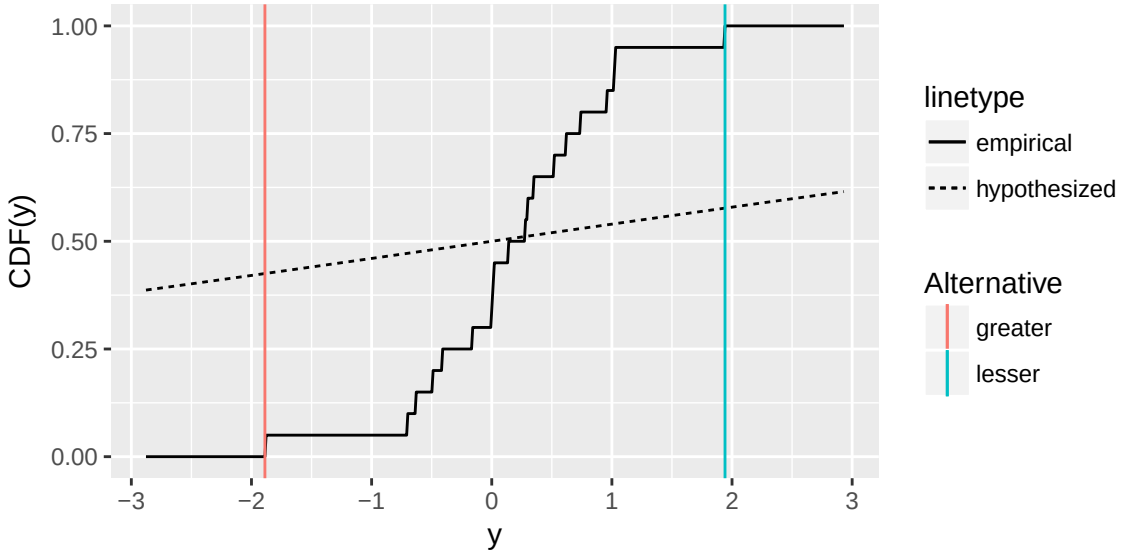
```
ecdf_y <- ecdf(y)

x <- seq(from = min(y) - 1, to = max(y) + 1, by = 0.01)

fhat <- ecdf_y(x)
f_0 <- pnorm(x, mean = 0, sd = sqrt(100))

x_less <- x[which.max(fhat - f_0)]
x_greater <- x[which.max(f_0 - fhat)]

# and add them to our picture
ggplot(mapping = aes(x = x, y = fhat)) +
  geom_line(aes(linetype = "empirical")) +
  geom_line(aes(y = f_0, linetype = "hypothesized")) +
  geom_vline(aes(color = "lesser", xintercept = x_less)) +
  geom_vline(aes(color = "greater", xintercept = x_greater))  +
  labs(color = "Alternative", x = "y", y = "CDF(y)")
```



Test statistics are the difference between the empirical and hypothesized distributions at the indicated y values.

Any indication is worth credit.

d) What properties of the true and hypothesized distributions leads to the contradiction?

In this case while both distributions are centered at the same place, the hypothesized distribution has a much larger spread than the true population, this means there is somewhere on the left of the center where the hypothesized CDF is almost guaranteed to be a lot higher than the ECDF of the data, and also somewhere on the right of the center where the hypothesized CDF is almost guaranteed to be a lot lower than the ECDF of the data.

e) In your own words, describe why this suggests one-sided K-S tests are hard to interpret.

In a one-sided KS test, if the null is rejected, we can't be sure if this is truly because the population CDF is strictly less than the hypothesized one (or greater than in the greater alternative) because it could just reflect that the population CDF has a different spread to the hypothesized one.