Hickman_Midterm2

Keith Hickman

November 9, 2017

Problem 1

Boxes of cereal are advertised as having a net weight of 8 ounces. The weights of boxes are assumed to be normally distributed. A new cereal box-filling machine is purchased, and we wish to be sure that on average, it puts at least the correct amount of cereal in the boxes. We randomly select 16 boxes, and find they have weights with sample mean 8.10 ounces and sample standard deviation 0.20 ounces.

(a) Suppose your data included all the box weights. How would you check the normal distribution assumption? Explain what would indicate a violation of this as- sumption. (Note: You do not have to perform the check on the given data.)

There are several ways to check an assumption of normality, including qqnorm or density kernel plots, boxplots, or summary statistics.

(b) Perform a test of the null hypothesis that the average net weight of the boxes of cereal produced by the new machine is less than or equal to 8 ounces, against the alternative that it is greater than 8 ounces. Test at level $\alpha = 0.05$. (Clearly define the parameter you're estimating and the hypotheses you're testing, and give a full and substantive conclusion).

Because this is a one-sample location problem, the parameter we're testing here is the mean μ.
We have a normally distributed variable, and the hypothesis specifies a direction, so we'll perform a one-tailed t-test. The null hypothesis is given as $H_0 : \mu_0 \leq 8$ and the alternative $H_1 : \mu_1 > 8$.

Parameters:

```
n <- 16
s <- .2
mu.0 <- 8
mu.1 <- 8.1
```

Because the alternative hypothesis is "greater than", we use `1-pnorm(t)` to calculate the p-value. The t-statistic and p-value:

```
t <- (mu.1 - mu.0) / (s/sqrt(n))
t
```

```
## [1] 2
```

```
1-pnorm(t)
```

```
## [1] 0.02275013
```

The p-value of .0228 is well under our α of .05, and we can probably safely reject the null, at least for this sample. However, n = 16 is still relatively small, so further testing is probably a good idea. As a side note, what happens if the machine puts too much cereal in the boxes?

Problem 2

To test whether my friend's fish Googly had psychic powers, I wrote R code to display two windows. I entered either Left or Right depending on which way Googly was facing. Then the random number generator in R

selected either the left or the right window, with probability 0.5 for each, in which to display a star. Let p be the probability Googly guesses correctly on a given trial (assume this is constant.) In 80 trials, Googly correctly guessed the window with the star 41 times.

(a) Using mathematical notation, write down null and alternative hypotheses for a one-sided test.

Here, the null hypothesis $H_0 : p \leq .5$, where the alternative hypothesis is $H_1 : P > .5$ where P is the proportion of observation correct selections or guesses.

(b) If the test statistic is the number of correct guesses (41) in 80 trials, write down R code to find the P-value of a one-sided test.

R code for finding the p-value for 41/80 guesses correct:

```
1 - pbinom(40, 80, .5)
```

```
## [1] 0.4555361
```

(c) Even without R, we can see that Googly's success rate was close to its expected value under the null, so the one-tailed P-value will be close to 0.5. State your conclusion about the fish's psychic powers.

The p-value of .45 here doesn't tell us anything that we don't already know - Googly likely doesn't have psychic powers; this is a result we could expect by random methods as well.

*(d) Continued from part (b). If you only known that the R code dbinom(40, 80,0.5) gives the number 0.0889, how would you find the exact P-value of the test? (Hint: use the property of a symmetric probabilty distribution.)

I would multiply the result by 2 giving `2 * (dbinom(40,80,.5))` to obtain both tails of the probability distribution.

```
2 * (dbinom(40, 80, .5))
```

```
## [1] 0.1778558
```

Problem 3

The file snoqualmie14.txt contains the daily precipitation (in inches) in Snoqualmie Falls, WA, for a random sample of 365 days. After saving the file to your computer, you can load it into R by entering the command:

rainfall = scan(file.choose())

and then selecting the file.

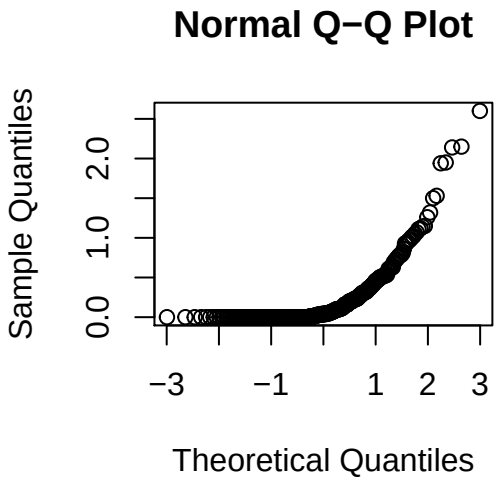
```
rainfall <- read.csv("C:\\Users\\khickman\\Desktop\\Personal\\IUMSDS\\StatsS520\\Module12\\snoqualmie14
## I used read.csv because everytime I knit to PDF, r markdown asks me to choose the file.
rainfall <- rainfall[1]
summary(rainfall)
```

```
##        X0
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0400
##  Mean   :0.2062
##  3rd Qu.:0.2600
##  Max.   :2.6000
```

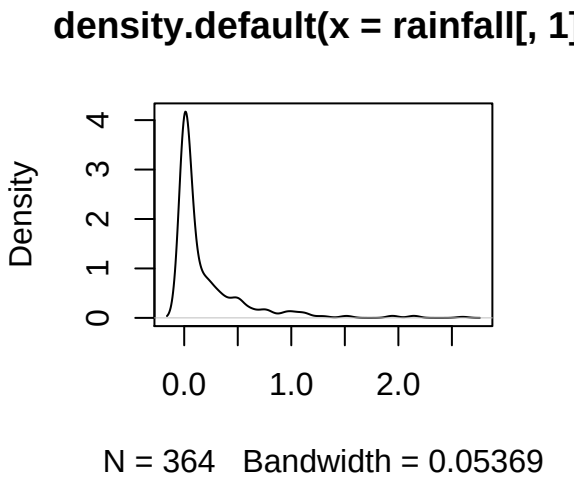
(a) Show that the data does not come from a normally distributed population. Include a graph to support your answer.

We'll use the qqnorm and density kernels to analyze normality for this distribution.

```
qqnorm(rainfall[,1])
```




```
plot(density(rainfall[,1]))
```



This variable is clearly right-skewed and non-normal. The `qqnorm` plot indicates that many values are at or near zero, but that there are a substantial number of higher values. The density kernel also indicates a right-skewed distribution.

(b) The mean annual rainfall in Seattle is 37.7 inches per year. Test (at level $\alpha = 0.05$) the hypothesis that the mean rainfall in Snoqualmie Falls is different from the mean rainfall in Seattle. (Clearly define the parameter you're estimating and the hypotheses you're testing, and give a full and substantive conclusion).

The parameter of interest is the difference in mean rainfall between the two regions. The null hypothesis is that there is no difference in rainfall between the two locales. Let μ_1 be the rainfall in Seattle, and μ_2 be the rainfall in Snoqualmie. Let $\Delta = \mu_1 - \mu_2$. Thus $H_0 : \Delta = 0$. The alternative, which we are trying to show, is stated $H_1 : \Delta \neq 0$.

Additionally, we can assume the variables are independant. We don't know the variance for the Seattle rainfall. Here, we can't do a Welch's or Students t-test, as the populations are neither normal, and we don't know whether they're ID. We also don't know the variance. I think we want to do a two-tailed test. Even without the test, it's clear that with a mean of 37.7 in Seattle and .26 in Snoqualmie, we're way off from the null.

NOT SURE WHAT TO DO HERE WHEN ALL I HAVE IS THE MEAN FOR MY SECOND VAR

```
mean.seattle <- 37.7
mean.sno <- mean(rainfall[,1])
sd.sno <- sd(rainfall[,1]/sqrt(length(rainfall[,1])))
sd.sno
```

```
## [1] 0.01931348
```

```
mean.seattle
```

```
## [1] 37.7
```

```
mean.sno
```

```
## [1] 0.2061538
```

```
delta.hat <- mean.seattle - mean.sno
```

Problem 4

In one year in the United States, 4.247 million babies were born. Of these, 2.173 million were male and 2.074 million were female. With very few exceptions (e.g. identical twins), the sexes of the babies are independent, so we can use the binomial distribution to model the number of babies that are female. Let p be the probability that a random (future) newborn is female.

(a) (2 points) What percentage of the babies were female? (To get credit for this question, you must give your answer as a percentage and you must round appropriately.)

```
p <- (2.074 / 4.247) * 100
p
```

```
## [1] 48.83447
```

Expressed as a percentage, $p = 48.83\%$ of babies were female in that year.

(b) (3 points) Suppose we wish to test the null hypothesis that the probability a baby is female is 50%. Write down null and alternative hypotheses in mathematical notation for this test.

The null hypothesis that $p = .50$ can be stated as follows: $H_0 : p = .50$. The alternative is $H_1 : p \neq .50$

(c) (10 points) Find the P-value, using both Binomial probability and Normal ap- proximation. (Carefully show your work and R codes.)

Here, we are concerned with finding the true value of p. To start, I would like to find the p-value if 2.07 out of 4.24 million babies born were female assuming that the true probability is 50%. The p-value is calculated under the null as follows:

```
pbinom(2074000, 4247000, .5)
```

```
## [1] 0
```

The p-value is essentially 0, indicating that if the true female birth proportion were 50%, this would be an extremely unlikely value. Thus, we can safely reject the null.

(d) (2 points) Using the Central Limit Theorem, find a 95% confidence interval for the probability that a birth is female.

Assuming normality, the 95% confidence interval for a binomial distribution can be stated established by adding and subtracting a standard error term to observed p:

```
p <- .48883
lower_bound <- p - qnorm(.975) * sqrt(p * (1 - p)) / sqrt(4247000)
upper_bound <- p + qnorm(.975) * sqrt(p * (1 - p)) / sqrt(4247000)
upper_bound
```

```
## [1] 0.4893054
```

```
lower_bound
```

```
## [1] 0.4883546
```

Thus, a 95% confidence interval for p runs is very narrow in terms of percentage, mostly because we have such a large n, from 48.83546% to 48.93054%.

(e) (2 points) Explain what this confidence interval means without using the word "confident."

The confidence interval indicates that in 95% of the samples (years?), the proportion of babies born who are female will fall between the upper and lower bounds of 48.84% to 48.93%.

(f) (2 points) The P-value for your test in part (c) is basically zero. From this and your confidence interval, write in a sentence your conclusion about the probability that a random newborn is female.

A very small p-value means that we can safely reject the null, and proceed with an understanding that the actual proportion of babies born who are female is closer to 48.83%