統計方法 Statistical methods

FALL 2016

Homework 2, due: 2016/10/19

1. (Irizarry RA, Love MI (2015): Data Analysis for the Life Sciences, page 30 Exercises) For these exercises, we will be using the following dataset:

```
url <-
"http://ghuang.stat.nctu.edu.tw/course/statmethods16/files/data/femaleControlsPopulation.csv"
filename <- basename(url)
download.file(url, destfile=filename)
x <- unlist( read.csv(filename) )
```

Here `x` represents the weights for the entire population.

a. What is the average of these weights?
b. After setting the seed at 1, `set.seed(1)` take a random sample of size 5. What is the absolute value (use `abs`) of the difference between the average of the sample and the average of all the values?
c. After setting the seed at 5, `set.seed(5)` take a random sample of size 5. What is the absolute value of the difference between the average of the sample and the average of all the values?
d. Why are the answers from b and c different?
   - A) Because we made a coding mistake.
   - B) Because the average of the `x` is random.
   - C) Because the average of the samples is a random variable.
   - D) All of the above.
e. Set the seed at 1, then using a for-loop take a random sample of 5 mice 1,000 times. Save these averages. What percent of these 1,000 averages are more than 1 ounce away from the average of `x`?
f. We are now going to increase the number of times we redo the sample from 1,000 to 10,000. Set the seed at 1, then using a for-loop take a random sample of 5 mice 10,000 times. Save these averages. What percent of these 10,000 averages are more than 1 ounce away from the average of `x`?
g. Note that the answers to e and f barely changed. This is expected. The way we think about the random value distributions is as the distribution of the list of values obtained if we repeated the experiment an infinite number of

times. On a computer, we can't perform an infinite number of iterations so instead, for our examples, we consider 1,000 to be large enough, thus 10,000 is as well. Now if instead we change the sample size, then we change the random variable and thus its distribution.

Set the seed at 1, then using a for-loop take a random sample of 50 mice 1,000 times. Save these averages. What percent of these 1,000 averages are more than 1 ounce away from the average of x?

h. Use a histogram to "look" at the distribution of averages we get with a sample size of 5 and a sample size of 50. How would you say they differ?
   - A) They are actually the same.
   - B) They both look roughly normal, but with a sample size of 50 the spread is smaller.
   - C) They both look roughly normal, but with a sample size of 50 the spread is larger.
   - D) The second distribution does not look normal at all.

i. For the last set of averages, the ones obtained from a sample size of 50, what percent are between 23 and 25?

j. Now ask the same question of a normal distribution with average 23.9 and standard deviation 0.43.

The answer to i and j were very similar. This is because we can approximate the distribution of the sample average with a normal distribution.

2. (Irizarry RA, Love MI (2015): Data Analysis for the Life Sciences, page 38 Exercises (parts)) For these exercises, we will be using the following dataset:

```
url <-
"http://ghuang.stat.nctu.edu.tw/course/statmethods16/files/data/mice_pheno.csv"
filename <- basename(url)
download.file(url, destfile=filename)
dat <- na.omit( read.csv(filename) )
```

a. If a list of numbers has a distribution that is well approximated by the normal distribution, what proportion of these numbers are within one standard deviation away from the list's average?

b. What proportion of these numbers are within two standard deviations away from the list's average?

c. What proportion of these numbers are within three standard deviations away from the list's average?

d. Define `y` to be the weights of males on the control diet. What proportion of the mice are within one standard deviation away from the average weight (remember to use `popsd` for the population sd)?

e. What proportion of these numbers are within two standard deviations away from the list's average?

f. What proportion of these numbers are within three standard deviations away from the list's average?

g. Note that the numbers for the normal distribution and our weights are relatively close. Also, notice that we are indirectly comparing quantiles of the normal distribution to quantiles of the mouse weight distribution. We can actually compare all quantiles using a `qqplot`. Which of the following best describes the qq-plot comparing mouse weights to the normal distribution?

- A) The points on the qq-plot fall exactly on the identity line.
- B) The average of the mouse weights is not 0 and thus it can't follow a normal distribution.
- C) The mouse weights are well approximated by the normal distribution, although the larger values (right tail) are larger than predicted by the normal. This is consistent with the differences seen between question c and f.
- D) These are not random variables and thus they can't follow a normal distribution.

h. Create the above qq-plot for the four populations: male/females on each of the two diets. What is the most likely explanation for the mouse weights being well approximated? What is the best explanation for all these being well approximated by the normal distribution?

- A) The CLT tells us that sample averages are approximately normal.
- B) This just happens to be how nature behaves. Perhaps the result of many biological factors averaging out.
- C) Everything measured in nature follows a normal distribution.
- D) Measurement error is normally distributed.

i. All the above exercises relate to the normal distribution as an approximation of the distribution of a fixed list of numbers or a population. But, keep in mind that the central limit applies to averages of random variables. Let's explore this concept.

We will now take a sample of size 25 from the population of males on the chow diet. The average of this sample is our random variable. We will use the `replicate` to observe 10,000 realizations of this random variable. Set the seed at 1, generate these 10,000 averages. Make a histogram and qq-plot of these 10,000 numbers against the normal distribution. What is the average of the distribution of the sample average?

j. What is the standard deviation of the distribution of sample averages?

k. According to the CLT, the answer to exercise i should be the same as `mean(y)`. You should be able to confirm that these two numbers are very close. Which of the following does the CLT tell us should be close to your answer to exercise j?

- A) `popsd(y)`
- B) `popsd(avgs) / sqrt(25)`
- C) `sqrt(25) / popsd(y)`
- D) `popsd(y) / sqrt(25)`

l. In practice we do not know σ (`popsd(y)`) which is why we can't use the CLT directly. This is because we see a sample and not the entire distribution. We also can't use `popsd(avgs)` because to construct averages, we have to take 10,000 samples and this is never practical. We usually just get one sample. Instead we have to estimate `popsd(y)`. As described, what we use is the sample standard deviation. Set the seed at 1, using the `replicate` function, create 10,000 samples of 25 and now, instead of the sample average, keep the sample standard deviation. Look at the distribution of the sample standard deviations. It is a random variable. The real population SD is about 4.5. What proportion of the sample SDs are below 3.5?

3. (Irizarry RA, Love MI (2015): Data Analysis for the Life Sciences, page 46 Exercises (parts))

a. The CLT is a result from probability theory. Much of probability theory was originally inspired by gambling. This theory is still used in practice by casinos. For example, they can estimate how many people need to play slots for there to be a 99.9999% probability of earning enough money to cover expenses. Let's try a simple example related to gambling.

Suppose we are interested in the proportion of times we see a 6 when rolling `n=100` die. This is a random variable which we can simulate with `x=sample(1:6, n, replace=TRUE)` and the proportion we are interested in can be expressed as an average: `mean(x==6)`. Because the die rolls are

independent, the CLT applies.

We want to roll `n` dice 10,000 times and keep these proportions. This random variable (proportion of 6s) has mean `p=1/6` and variance `p*(1-p)/n`. So according to CLT `z = (mean(x==6) - p) / sqrt(p*(1-p)/n)` should be normal with mean 0 and SD 1. Set the seed to 1, then use replicate to perform the simulation, and report what proportion of times `z` was larger than 2 in absolute value (CLT says it should be about 0.05).

b. For the last simulation you can make a qqplot to confirm the normal approximation. Now, the CLT is an *asympototic* result, meaning it is closer and closer to being a perfect approximation as the sample size `n` increases. In practice, however, we need to decide if it is appropriate for actual sample sizes. Is 10 enough? 15? 30?

In the example used in exercise a, the original data is binary (either 6 or not). In this case, the success probability `p` also affects the appropriateness of the CLT. With very low probabilities, we need larger sample sizes for the CLT to "kick in".

Run the simulation from exercise a, but for different values of `p` and `n`. For which of the following is the normal approximation best?

- A) p=0.5 and n=5
- B) p=0.5 and n=30
- C) p=0.01 and n=30
- D) p=0.01 and n=100

4. (Irizarry RA, Love MI (2015): Data Analysis for the Life Sciences, page 70 Exercises (parts)) For these exercises we will load the babies dataset from `babies.txt`. We will use this data to review the concepts behind the p-values and then test confidence interval concepts.

```
url <- "http://ghuang.stat.nctu.edu.tw/course/statmethods16/files/data/babies.txt"
filename <- basename(url)
download.file(url, destfile=filename)
babies <- read.table("babies.txt", header=TRUE)
```

This is a large dataset (1,236 cases), and we will pretend that it contains the entire population in which we are interested. We will study the differences in

birth weight between babies born to smoking and non-smoking mothers.

First, let's split this into two birth weight datasets: one of birth weights to non-smoking mothers and the other of birth weights to smoking mothers.

```
bwt.nonsmoke <- filter(babies, smoke==0) %>% select(bwt) %>% unlist
bwt.smoke <- filter(babies, smoke==1) %>% select(bwt) %>% unlist
```

Now, we can look for the true population difference in means between smoking and non-smoking birth weights.

```
library(rafalib)
mean(bwt.nonsmoke) - mean(bwt.smoke)
popsd(bwt.nonsmoke)
popsd(bwt.smoke)
```

As we did with the mouse weight data, this assessment interactively reviews inference concepts using simulations in R. We will treat the `babies` dataset as the full population and draw samples from it to simulate individual experiments. We will then ask whether somebody who only received the random samples would be able to draw correct conclusions about the population.

We are interested in testing whether the birth weights of babies born to non-smoking mothers are significantly different from the birth weights of babies born to smoking mothers.

a. Set the seed at 1 and obtain two samples, each of size `N = 25`, from non-smoking mothers (`dat.ns`) and smoking mothers (`dat.s`). Compute the t-statistic (call it `tval`).
b. We can summarize our data using a t-statistics because we know that in situations where the null hypothesis is true and the sample size is relatively large, this t-value will have an approximate standard normal distribution.

   Because of the symmetry of the standard normal distribution, there is a simpler way to calculate the probability that a t-value under the null could have a larger absolute value than `tval` – the p-value. Choose the simplified calculation from the following:

   - A) 1-2*pnorm(abs(tval))
   - B) 1-2*pnorm(-abs(tval))

- C) 1-pnorm(-abs(tval))
- D) 2*pnorm(-abs(tval))

c. Our estimate of the difference between babies of smoker and non-smokers: `mean(dat.ns) - mean( dat.s)`. If we use the CLT, what quantity would we add and subtract to this estimate to obtain a 99% confidence interval?

d. If instead of CLT, we use the t-distribution approximation, what do we add and subtract (use `2*N-2` degrees of freedom)?

e. Why are the values from c and d so similar?
- A) Coincidence.
- B) They are both related to 99% confidence intervals.
- C) `N` and thus the degrees of freedom is large enough to make the normal and t distributions very similar.
- D) They are actually quite different, differing by more than 1 ounce.

f. Since we have the full population, we know what the true effect size is (`mean(bwt.nonsmoke) - mean(bwt.smoke)`) and we can compute the power of the test for true difference between populations.

Set the seed at 1 and take a random sample of `N = 5` measurements from each of the smoking and nonsmoking datasets. What is the p-value (use the t-test function)?

g. Set the seed at 1, then use the replicate function to repeat the code used in exercise f 10,000 times. What proportion of the time do we reject at the $\alpha = 0.05$ level (the power)?

h. Note that, not surprisingly, the power is lower than 10%. Repeat the exercise above for samples sizes of 30, 60, 90 and 120. Which of those four gives you power of about 80%?

i. Repeat problem h, but now require an α level of 0.01. Which of those four gives you power of about 80%?