TD 7

Outils pour l'exploration de données - Confidence Interval and Hypothesis Test.

1 Introduction

A confidence interval describes the amount of uncertainty associated with a sample estimate of a population parameter (e.g., the mean). 95% confidence level means that 95% of the interval estimates are expected to include the population parameter. A simple hypothesis testing is an assumption that specifies the population distribution completely. It examines a random sample from a population to see if the sample data are consistent with the statistical hypothesis. If not, the null hypothesis is rejected. This TD focuses on the steps of processing confidence intervals and performing statistical hypothesis testing in R.

But before starting to work with those topics, it is important to recall a very important concept in statistics: *the central limit theorem*. For any distribution with finite mean and standard deviation, samples[1] taken from that population will tend towards a normal distribution around the mean of the population as sample size increases. Furthermore, as sample size increases, the variation of the sample means will decrease.

2 The central limit theorem

Let n be the sample size, N be the population size (if it is not infinite), μ be the population mean, and σ be the population standard deviation. For virtually all (realistic) situations, two properties of the distribution of the sample mean are[2]:

$$E(X) = \mu$$

$$SD(X) = \frac{\sigma}{\sqrt{n}}$$

(for an infinite population). Thus, $\bar{X}$ stays around μ, with progressively smaller variation as n (the sample size) increases. This result is the essence of statistical inference: that samples can provide information about populations, and that the accuracy of this information increases with an increase in the sample size. This standard deviation, since it is referring to a statistic (X), is sometimes called a standard error.

How can we check this? Simulation is an excellent way.

2.1 Samples from a continuous uniform random distribution

This simulation shows the distribution of samples of sizes 1, 2, 4, ... 32 taken from a uniform distribution[3]. Note, for each sample, we are finding the average value of the sample.

```
##show distribution of sample means of varying size samples
numcases <- 10000 #how many samples to take?
min <- 0 #lowest value
max <- 1
ntimes <- 6
op<- par(mfrow=c(ntimes,1), mar=c(2.5, 2.5,2.5,2.5), mgp=c(3, 1, 0)) #stack ntimes graphs on
top of each other
i2 <- 1 #initialize counters
```

[1] Whenever random samples of a given size are taken repeatedly from a population of scores and a statistic (e.g., the mean) is computed for each sample, the distribution of this computed statistic may be constructed. The resulting distribution is called a sampling distribution (e.g., the sampling distribution of the mean).

[2] http://people.stern.nyu.edu/jsimonof/classes/1305/pdf/clt.class.pdf

[3] http://personality-project.org/r/distributions.html

```
for (i in 1:ntimes) #repeat n times
{ sample=rep(0,numcases) #create a vector
k=0 #start off with an empty set of counters
for (j in 1:i2) # inner loop
{
sample <- sample +runif(numcases,min,max)
k <- k+1 }
x <- sample/k

hist(x, xlim=range(0,1),prob=T ,main=paste( "samples of size", k ),col="blue")
i2 <- 2*i2 } #end of i loop
```

2.2 Example with the `algae` data set

Load the `algae` data set. Plot the histogram for the variable `a1`. From the histogram generated is clear that values of this variable, at least for this sample, does not follow a normal distribution.

```
> load("algae.Rdata")
> attach(algae)
> hist(a1,prob=T,breaks=30)
```

Now, suppose that we would like to infer the mean of `a1` in `algae` based on subsamples — the actual mean of `a1` in `algae` is `16.9235`. Let's take 100 subsample of size 30.

```
> results =numeric(0) # a place to store the results
> for (i in 1:100) { # the for loop --- 100 samples
+ S = sample(a1,30) # take a random sample of size 30
+ results[i]= mean(S) # store the answer
+ }
> hist(results,prob=T)
```

Now, look at the plot of the histogram of the sampling of the mean. What can you observe? Does the sampling distribution of the mean looks like a normal distribution? Is the mean of the sampling distribution of the mean close to the actual mean of `a1` (`16.9235`)?

3 Constructing a confidence interval for a mean

As we have seen in the previous section, for the case of sample means, if several samples or repetitions of same size are taken, the frequency curve of means from various samples will be approximately bell-shaped, like that of the normal distribution[4]. In fact, the mean will be the same as the mean for the population. The standard deviation or the Standard Error of the Mean (SEM) will be:

$$\frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the population and n is the size of the sample. In practice, the population standard deviation is unknown and replaced by sample standard deviation, computed from the data.

Example:
Suppose weight losses for thousands of people in a **population** were bell-shaped with a mean of 8 pounds (3.63 kg) and a standard deviation of 5 pounds (2.27 kg). A **sample** of $n = 25$ people, resulted in a mean of 8.32 pounds and standard deviation of 4.74 pounds.

- population standard deviation = 5 pounds
- sample standard deviation = 4.74 pounds

[4] `http://www.stat.columbia.edu/~madigan/1001-Fall2010/NOTES/p21.pdf`

- standard error of the mean (using population S.D.) = $\frac{5}{\sqrt{25}} = 1$
- standard error of the mean (using sample S.D.) = $\frac{4.74}{\sqrt{25}} = 0.95$

Before we continue, it is important to recall the 68-95-99.7 rule[5] — See Figure 1. Such a rule states that for a normal distribution, nearly all values (99.73%) lie within 3 standard deviations of the mean. About 68.27% of the values lie within 1 standard deviation of the mean. Likewise, about 95.45% of the values lie within 2 standard deviations of the mean. Another way of looking at it: only a small fraction of the observations lie outside this range — Figure 2.

In 95% of all samples, the **sample mean** will fall within 2 standard errors of the **true population mean**. In 95% of all samples, the **true population mean** will fall within 2 standard errors of the **sample mean**. A **95% confidence interval for a population mean** be defined as:

- sample mean $\pm$ 2 standard errors

Tasks: Load the `iris` data set. Build a 95% confidence interval for the mean of the variable `Petal.Length`, given that the specie is `setosa`.

4 Confidence interval: t-interval for the population mean

Important note: the equation in the previous section is used only if there are at least 30 observations in the sample. A 95% confidence interval for population mean based on smaller samples requires a multiplier larger than 2, found from a t-distribution. In fact, when the sample size n is small, the standard error of the mean calculated from the sample can be quite different from that of that population.

As a consequence, the random variable $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ follows a t-distribution with $n-1$ degrees of freedom (**d.f.**), not a z distribution. Especially for small samples, t has more spread than the standard normal z. It is still symmetric about zero and bell-shaped like the z curve.

Calculating the confidence interval when using a t-distribution is similar to using a normal distribution. Let's calculate a 95% confidence interval for the mean of the variable `Petal.Length`, given that the specie is `setosa`. But, this time, we will consider t-distribution.

```
> ## t-distribution ****
> m <- mean(iris[iris$Species=="setosa","Petal.Length"])
> s <- sd(iris[iris$Species=="setosa","Petal.Length"])
> n <- nrow(iris)
> error <- qt(0.975,df=n-1)*s/sqrt(n) ## qt is a quantile function for the t-distribution with
df degrees of freedom.  A quantile function returns for a given probability, the value which
the random variable will be at, or below, with that probability
> T_left <- m-error
> T_right <- m+error
> T_left
[1] 1.433981
> T_right
[1] 1.490019
```

So, the true mean has a probability of 95% of being in the interval between 1.433981 and 1.490019.

[5] `http://en.wikipedia.org/wiki/68-95-99.7_rule`

5 Comparing two means

We can also use inference to compare the mean responses in two groups, each from a distinct population. This is called a two-sample situation, one of the most common settings in statistical applications. One example would be to compare the change in blood pressure for two groups of black men, where one group has been given calcium supplements, the other a placebo, that is, comparing results in an *experiment*. Another example is to compare the average performance of two different machine learning methods applied to solve the same problem.

More specifically, the question is this [1]: given what we know about the variation from replicate to replicate within each sample (the within sample variance), how likely is that our two sample means were drawn from populations with the same average? If the answer is high likely, then we should say that our sample means are not significantly different. If it is rather unlikely, then we should say that our sample means are significantly different.

A possible better way to proceed is to find the probability that the two samples were, in fact, drawn from populations with the same mean. If this probability is very low (e.g., less than 5% or 1% — **level of significance**) then we can be reasonably certain (95% or 99% in the two examples — **confidence level**) that the means are really different from one another. However, it is important to note that we can never be 100% certain: the apparent difference might be just due to random sampling. There are different tests for comparing two sample means. Here, we will use the student's t-test. Basically, in this test the statistics t is the number of standard errors by which two samples are separated.

The outcome of this test is the acceptance or rejection of the null hypothesis (H0). The null hypothesis generally states that: "Any differences, discrepancies, or suspiciously outlying results are purely due to random and not systematic errors". The alternative hypothesis (H1) states exactly the opposite. For example, when comparing the change of blood pressure (example in the beginning of this section), the null hypothesis is: the means are the same. That is, the use of calcium and placebo provide the same results. The differences observed (if any) are purely due to random errors. The alternative hypothesis is: the means are significantly different: the use of calcium supplement do produce a change in blood pressure (so at least one method yields systematic analytical errors).

5.1 Hypothesis testing and sampling distribution

The t-test is a type of hypothesis test or significance test, that is, is a method for testing a claim or hypothesis about a parameter in a population, using data measured in a sample. In this method, we test some hypothesis by determining the likelihood that a sample statistic could have been selected, if the hypothesis regarding the population parameter were true.

The logic of hypothesis testing is rooted in an understanding of the sampling distribution of the mean. We showed characteristics of the mean, which two of which are particularly relevant in this section:

- The sample mean is an unbiased estimator of the population mean. On average, a randomly selected sample will have a mean equal to that in the population. In hypothesis testing, we begin by stating the null hypothesis. We expect that, if the null hypothesis is true, then a random sample selected from a given population will have a sample mean equal to the value stated in the null hypothesis.
- Regardless of the distribution in the population, the sampling distribution of the sample mean is normally distributed. Hence, the probabilities of all other possible sample means we could select are normally distributed. Using this distribution, we can therefore state an alternative hypothesis to locate the probability of obtaining sample means with less than a 5% chance of being selected if the value stated in the null hypothesis is true.

To locate the probability of obtaining a sample mean in a sampling distribution, we must know (1) the population mean and (2) the standard error of the mean (SEM; introduced in the first section). Each value is entered in the test statistic formula computed, thereby allowing us to make a decision.

6 Tasks

Load the data set `Garden` (available on CELENE). These data come from three market gardens. The data shows the ozone concentrations in parts per hundred million (pphm) on ten summer days.

```
> ozone <- read.table("gardens.txt",header=T)
> summary(ozone)
```

Task 1: establish a 95% confidence interval for the three variables in `ozone` — `gardenA`, `gardenB`, `gardenC` — see Section 4.

Example: the use student's t-test to check, at a level of significance of 5%, if the means of the variables `gardenA` and `gardenB` are significantly different.

```
> t.test(gardenA,gardenB)

Welch Two Sample t-test

data:  gardenA and gardenB
t = -3.873, df = 18, p-value = 0.001115
alternative hypothesis:  true difference in means is not equal to 0
95 percent confidence interval:
-3.0849115 -0.9150885
sample estimates:
mean of x mean of y
3 5
```

The p-value calculated for the t statistics is `0.001115` which is much smaller than the level of significance of 5% (0.05) established. Thus, we can reject the null hypothesis (no difference between) in favor of the alternative hypothesis (true difference in the means). In fact, we can state that the mean of `gardenB` (mean = 5.0 p.p.h.m) is significantly greater than that of `gardenA` (mean = 3.0 p.p.h.m). More precisely, you can present the result like this: ozone concentration was significantly higher in `gardenB` (mean = 5.0 p.p.h.m) than `gardenA` (mean = 3.0 p.p.h.m) — p-value=0.0011 (two-tailed), d.f.=18.

Observation: Note that, because the means are significantly different, the *confidence interval* on the difference does not include zero. Indeed, it goes from -3.085 up to -0.915.

Task 2: based on the steps followed in the previous example, use student's t-test to check, at a level of significance of 5%, if the means of the variables `Petal.Length` for `setosa` and `virginica` are significantly different.

Task 3: A social scientist wants to produce statistical evidence that men earn more than women. She records these salaries for a sample of 11 husband-wife pairs (`social.txt` data set on CELENE). Do the data support the scientist's theory? Use a paired (difference) t-test[6] (in `t.test`, set `paired=TRUE`) and discuss the results.

.

References

[1] Michael J. Crawley.

Statiscs: an introduction using R.

Wiley, 2005.

[6] Sometimes we deal with paired data. Paired data usually come from "before and after" scenarios, or any situation where it makes sense to compare a particular measurement in one group to a particular measurement in the other group. A paired difference test uses additional information about the sample that is not present in an ordinary unpaired testing situation, either to increase the statistical power, or to reduce the effects of confounders.

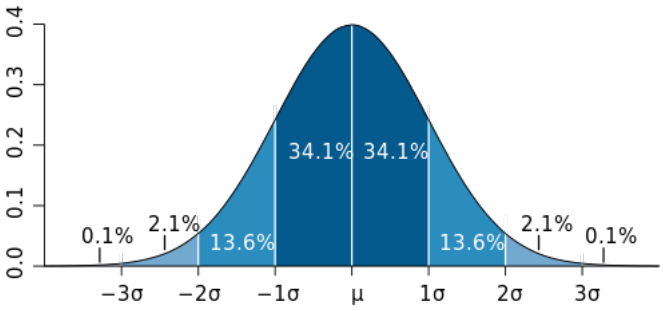

Figure 1: Normal distribution curve that illustrates standard deviations

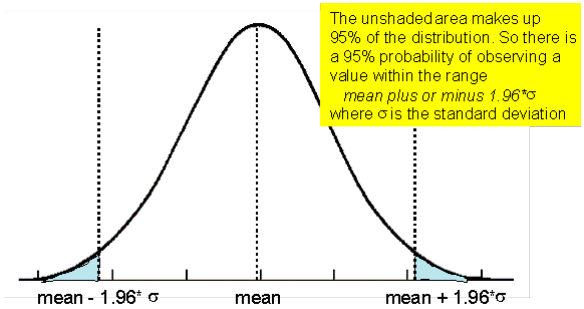

Figure 2: Normal distribution curve: only a small fraction of the observations lie within the blue area