

## Lab Session: Statistical Inference

In statistics, statistical inference is the process of drawing conclusions from data that is subject to random variation, for example, observational errors or sampling variation. More substantially, the terms statistical inference, statistical induction and inferential statistics are used to describe systems of procedures that can be used to draw conclusions from datasets arising from systems affected by random variation, such as observational errors, random sampling, or random experimentation.

Lab session five is dedicated to reviewing important statistical concepts in

- Confidence intervals
- Hypothesis testing

### Confidence Intervals

---

A confidence interval estimate of a population parameter consists of an interval of numbers produced by a point estimate and an associated confidence level specifying the probability that the interval contains the parameter. Most confidence intervals take the general form

$$\begin{aligned} &\text{point estimate} \pm \text{margin of error} \\ &\text{or} \\ &\text{point estimate} \pm \text{critical value} \times \text{std.dev of the estimate} \end{aligned}$$

where the margin of error is a measure of the precision of the interval estimate. A smaller margin of errors indicates a higher accuracy.

- The parameter of interest, e.g., the population mean, population proportion, the difference in population's means, etc.
- Design of the sample: SRS, stratified, experiments
- Confidence level or a confidence coefficient,  $(1 - \alpha)100\%$ , e.g., 95%, 99%, 90%, 80%, corresponding, respectively, to  $\alpha$  values of 0.05, 0.01, 0.1, 0.2, etc...

In most general terms, for a 95% CI, we say "we are 95% confident that the true population parameter is between the lower and upper calculated values".

Example:

**Construct confidence interval for the population mean using following data when the population standard deviation is 2.**

```
sampleData<-c(3.7, 2.3, 2.8, 4.0, 4.1, 3.3, 2.8, 1.2, 2.6, 3.4, 2.9, 2.5, 3.3, 4.1, 4.8, 3.3, 2.5, 3.1, 4.8, 4.1, 3.1, 3.1, 4.0, 3.7, 5.2, 4.3, 2.2, 1.9, 3.5, 0.7)
z.test(sampleData, sd=2)$"conf.int"
```

output

```
[1] 2.527656 3.959011
attr(,"conf.level")
[1] 0.95
```

**Construct confidence interval for the population mean using following data (population variance is unknown).**

```
sampleData<-c(3.7, 2.3, 2.8, 4.0, 4.1, 3.3, 2.8, 1.2, 2.6, 3.4, 2.9, 2.5, 3.3, 4.1, 4.8, 3.3, 2.5, 3.1, 4.8, 4.1, 3.1, 3.1, 4.0, 3.7, 5.2, 4.3, 2.2, 1.9, 3.5, 0.7)
t.test(sampleData)$"conf.int"
```

output

```
[1] 2.862446 3.624221
attr(,"conf.level")
[1] 0.95
```

## Hypothesis Testing

---

We will introduce you with the statistical hypothesis in R the Z test and t-test. Statisticians use hypothesis testing to formally check whether the hypothesis is accepted or rejected. Hypothesis testing is conducted in the following manner:

1. State the Hypotheses – Stating the null and alternative hypotheses.
2. Formulate an Analysis Plan – The formulation of an analysis plan is a crucial step in this stage.
3. Analyze Sample Data – Calculation and interpretation of the test statistic, as described in the analysis plan.
4. Interpret Results – Application of the decision rule described in the analysis plan.

Hypothesis testing ultimately uses a p-value to weigh the strength of the evidence or in other words, what the data are about the population. The p-value ranges between 0 and 1. It can be interpreted in the following way:

- A small p-value (typically  $\leq 0.05$ ) indicates strong evidence against the null hypothesis, so you reject it.
- A large p-value ( $> 0.05$ ) indicates weak evidence against the null hypothesis, so you fail to reject it.

### One Sample Z-test ( When $\sigma$ is known)

One sample Z-tests can be used to determine if the mean of a sample is different from a particular value when the population variance is known and the population is distributed as Normal distribution ( for small sample sizes typically less than 30 checks for Normality assumption).

**Example:** Suppose the manufacturer claims that the mean lifetime of a light bulb is less than 10,000 hours. In a sample of 30 light bulbs, it was found that they only last 9,900 hours on average. Assume the population standard deviation is 120 hours.

Write the null and alternative hypotheses first:

$\mu$  – mean lifetime of a light bulb

$$H_0: \mu = 10,000_{\text{hours}}$$

$$H_1: \mu < 10,000_{\text{hours}}$$

- Determine whether this is a one-tailed or a two-tailed test. Because the hypothesis involves the phrase "less than", this must be a one-tailed test.
- Specify the  $\alpha$  level:  $\alpha = .05$

```
xbar = 9900      # sample mean
> mu0 = 10000    # hypothesized value
> sigma = 120    # population standard deviation
> n = 30         # sample size
> z = (xbar-mu0)/(sigma/sqrt(n))
> z             # test statistic
[1] -4.5644

#we are interest on lower tail p value as H1:  $\mu < 10,000$ 
pval = pnorm(z)
> pval          # lower tail p-value
[1] 2.5052e-06
```

**Conclusion:** As p-value turns out to be less than the .05 significance level, we reject the null hypothesis at 5% significant level. Thus, there is a piece of evidence to say that the mean lifetime of a light bulb is less than 10,000.

For upper tailed p value and two tail p value use following commands.

```
pval = pnorm(z, lower.tail= FALSE)
> pval          # Upper tail p-value

pval = 2 * pnorm(z)
> pval          # two-tailed p-value
```

## Alternative Solution

Instead of using the textbook formula, we can apply the `z.test` function in the `TeachingDemos` package. It is not a core R package and must be installed and loaded into the workspace beforehand.

```
library(TeachingDemos)  # load TeachingDemos package
> z.test(9900,mu=10000, sd=120/sqrt(30), alternative ="less")
```

## One Sample t-test ( When $\sigma$ is unknown)

The one-sample t-test is one of the useful tests for testing the population mean when  $\sigma$  is unknown. To carry out a one-sample t-test in R, the name of a single vector and the mean with which it is compared is supplied. Unlike one sample z test, the one-sample T-test can be implemented using `t.test()`.

### `t.test()`

Performs one and two sample t-tests on vectors of data.

### Example:

Let us assume a scenario where an investor assumes that the mean of daily returns of stock since inception is at most \$3. Test the investor assumption at .05 significance level?

The sample of 30 days' daily return in \$ are as follows.

3.7, 2.3, 2.8, 4.0, 4.1, 3.3, 2.8, 1.2, 2.6, 3.4, 2.9, 2.5, 3.3, 4.1, 4.8, 3.3, 2.5, 3.1, 4.8, 4.1, 3.1, 3.1, 4.0, 3.7, 5.2, 4.3, 2.2, 1.9, 3.5, 0.7

Write the null and alternative hypotheses first:

$\mu$  – mean of daily returns of a stock

$H_0: \mu \leq \$3$  (at most 3)

$H_1: \mu > \$3$

```
x=c(3.7, 2.3, 2.8, 4.0, 4.1, 3.3, 2.8, 1.2, 2.6, 3.4, 2.9, 2.5, 3.3, 4.1, 4.8, 3.3, 2.5, 3.1, 4.8, 4.1, 3.1, 3.1, 4.0, 3.7, 5.2, 4.3, 2.2, 1.9, 3.5, 0.7)
t.test(x, mu=3, alternative = "greater")
```

Depend on the alternative hypothesis you can choose the option for alternative from following options.

alternative= c("two.sided", "less", "greater")

output

```
data: x
t = 1.3066, df = 29, p-value = 0.1008
alternative hypothesis: true mean is greater than 3
95 percent confidence interval:
 2.926901      Inf
sample estimates:
mean of x
 3.243333
```

According to the output, the p-value is 0.1, which is greater than 0.05%. Therefore, we do not reject  $H_0$  at 5%. Hence there is a piece of evidence to say that the investor assumption about the mean of daily returns of stock since inception is at most \$3 is correct.


## Lab Exercise

- 1) The data set named "rock" contains measurements on 48 rock samples from a petroleum reservoir. Variables in the data set are as follows.

Area	: area of pores space, in pixels out of 256 by 256
peri	: perimeter in pixels
shape	: perimeter/sqrt(area)
perm	: permeability in milli-Darcies

- Load the data set in the package "datasets".
- Carryout a descriptive analysis for the above variables and comment on your findings.
- Construct 95% confidence interval for the variable "area" and interpret your results.
- A researcher claims that the area of pores space is greater than 7000 pixels. Formulate suitable hypotheses to test the researcher's claim. Assuming the area is normally distributed test the validity of the researcher's claim and interpret your results.

2

**18.35 Men of few words?** Researchers claim that women speak significantly more words per day than men. One estimate is that a woman uses about 20,000 words per day while a man uses about 7,000. To investigate such claims, one study used a special device to record the conversations of male and female university students over a four-day period. From these recordings, the daily word count of the 20 men in the study was determined. Here are their daily word counts:<sup>20</sup> 

28,408	10,084	15,931	21,688	37,786
10,575	12,880	11,071	17,799	13,182
8,918	6,495	8,153	7,015	4,429
10,054	3,998	12,639	10,974	5,255

- Examine the data. Is it reasonable to use the  $t$  procedures (assume these men are an SRS of all male students at this university)?
- If your conclusion in part (a) is "Yes," do the data give convincing evidence that the mean number of words per day of men at this university differs from 7,000?