Statistical Computing with R: Masters in Data Sciences 503, S19 First Batch, SMS, TU, 2021

Shital Bhandary

Associate Professor

Statistics/Bio-statistics, Demography and Public Health Informatics

Patan Academy of Health Sciences, Lalitpur, Nepal

Faculty, Data Analysis and Decision Modeling, MBA, Pokhara University, Nepal

Faculty, FAIMER Fellowship in Health Professions Education, India/USA.

Review Preview

- Hypothesis testing

- Parametric tests for

- One-sample
- Two-samples
- Simple Linear regression to check the results

Hypothesis testing:

- It is part of inferential statistics i.e. taking decision from the sample (random) to population

- It is used to take decision based on statistical tests and models using p-value aka Type I error or alpha error

- It can be done using parametric or non-parametric methods/models

- Parametric means they have certain assumptions on the data (model) and/or errors and we must validate them to accept the results

- Non-parametric means they do not have assumptions about the distribution of the data (model) and errors

Why to use parametric tests?

- Parametric tests are considered “more powerful” than non parametric tests/models as they are based on mean, sd and normal distribution

- They are easy to compute/fit

- Easy to interpret

- Non-parametric tests are considered “less powerful” than parametric tests/models as they are based on median, IQR and non-normal distributions

- They are difficult to compute/fit

- Not so easy to interpret

Two statistical hypothesis:

- Null hypothesis: Equal, same, no difference
- Denoted as: H0
- P-value > 0.05 is needed to accept (fail to reject) it from parametric or non-parametric tests (Goodness-of-fit tests)

- Alternative hypothesis: Not equal, not same, different
- Denoted as H1 or Ha
- P-value <= 0.05 is needed to accept it from parametric or non-parametric tests (Research hypothesis tests!

Commonly used Parametric tests: We will discuss the bold and the red ones!

- **One-sample z-test to compare a hypothesized mean**
- Two-samples z-test to compare means across two groups
  - Pooled variance?
- Two-samples z-test to compare proportions across two groups
  - Chi-square test?

- One-sample t-test
- Two-samples t-test
  - Student
  - Welch
- One-way ANOVA
  - Classical
  - Welch

One-sample z-test:

$$Z = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

This Z follows the standard normal distribution i.e. ~ N(0,1)

Where,

- Z = Zee (normal) test ~ N(0,1)
- Xbar = sample mean
- Mu = Population mean (claim)
- Sigma = Population standard deviation **(must be known a priori!)**
- n = Sample size

Assumptions:

- n >=30 samples
- Test variable ~ Normal distribution

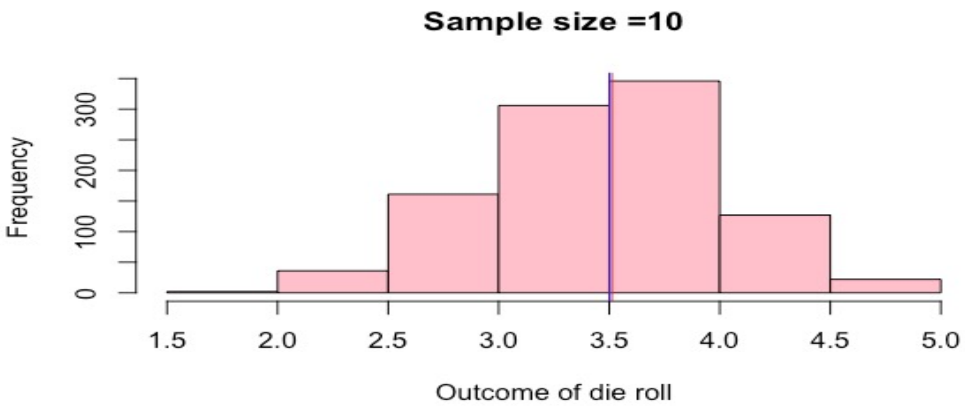
Why n >= 30 is used for z-test?

- The central limit theorem of statistics states that : if we take random sample of certain size for 30 times (or more) and the plot a graph with the mean of these 30 samples (or more) then it will be a bell shaped curve i.e. distribution of means follow the normal distribution

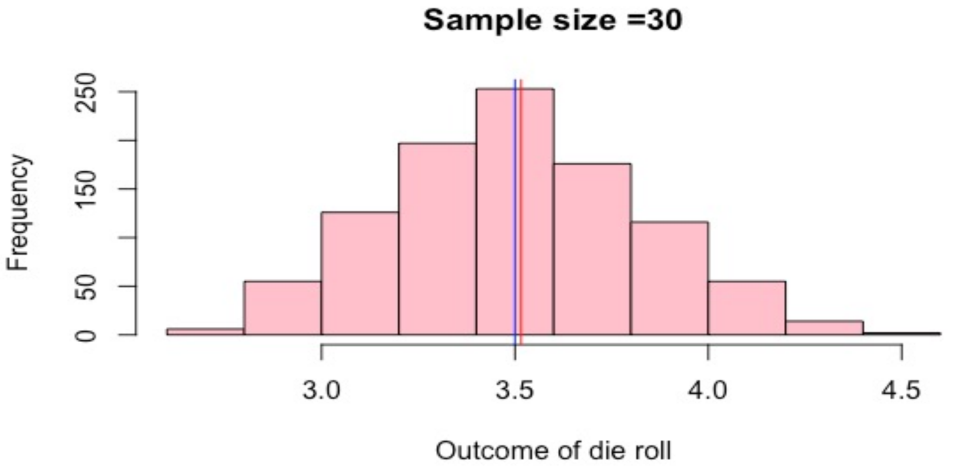
- **The standard deviation of this sampling distribution of means is known as “standard error”**

```
x10 <- c()
k =1000 #Fair die is rolled 1000 times (number of trials)
for ( i in 1:k) {
x10[i] = mean(sample(1:6,10, replace = TRUE))}
hist(x10, col ="pink", main="Sample size =10",xlab ="Outcome of die roll")
abline(v = mean(x10), col = "Red")
abline(v = 3.5, col = "blue")
#(1+2+3+4+5+6)/6 = 3.5
```

Central Limit Theorem: Illustrations



- This graph shows that with 10 random samples from 1000 trials, **means of the samples do not follow the normal distribution**



- This graph shows that with 30 random samples from 1000 trials, **means of the samples follow the normal distribution**

So what? What is its implication?

- Since it will not be possible to take 30 random samples from the population to make our data normal, we check it using “test of normality” on the data we use based on:

- **K-S test for large samples**
- **S-W test for small samples**

- We will “violate” the first assumption of any parametric test if we fail to confirm that the variable under consideration follows the normal distribution or not!

- E.g. We can test the hypothesis that: age of the MDS 503 class is 25 years IFF age ~ ND!

Example 1: Test the claim that mean of the miles per gallon variable is 20 using “mtcars” data in R!

- Hypothesis:

- Null (H_0): μ (mpg) = 20
- **Alternative (H_1): μ (mpg) ≠ 20 (This is a two-tailed test & we use it now!)**
- Alternative (H_1): μ (mpg) > or < 20 (This will be a one-tailed test!)

- We must use population mean in the hypothesis (always!)
- We then use the data from random sample to accept or refute the claim!

No Base R function for 1-sample z-test!

So, we need to define parameters:

```
mu0 <- 20
```

sigma <- 6 (must be known for this test)

```
xbar <- mean(mtcars$mpg)
n <- length(mtcars$mpg)
```

&

Write functions for z-test and p-value:

```
z <- sqrt(n)*(xbar-mu0)/sigma
p_value <- 2*pnorm(-abs(z))
```

Output: mtcars$mpg = 32 cases!

- z
- p_value

- Since, p-value > 0.05, we fail to reject (accept) the null hypothesis

- As sample mean (xbar) = 20.09062

- **We confirm that: Hypothesized mean mpg is 20!**

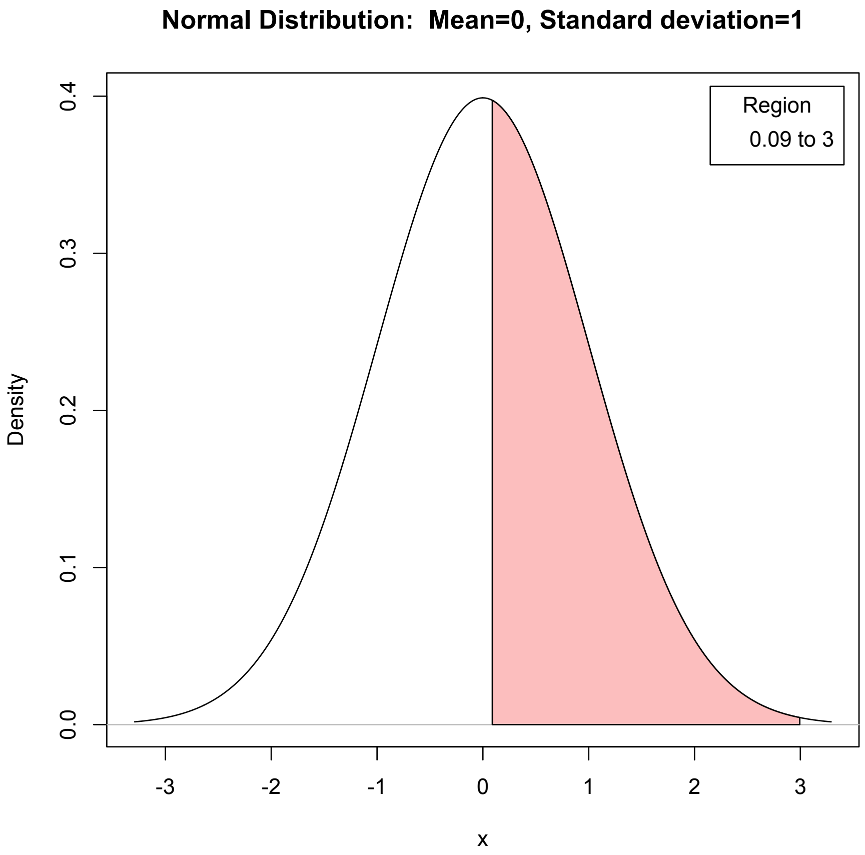
- > z
- [1] 0.08544207

(This is the areas of the standard normal curve or SNC!)

- > p_value
- [1] 0.9319099

(This is the Type I error associated with the z-value in the SNC!)

Illustration of p-value: General & efficient way



Here,

- z = 0.08544207 i.e. positive
- pnorm(z, lower.tail=F) will give 1-tailed p-value for +ve z
- 2 * pnorm(z, lower.tail=F) will give 2-tailed p-value for +ve z

The code below is more efficient as it works for +ve and –ve z values:

- 2 * pnorm(-abs(z)) as “-” sign means lower.tail=F for modulus of the positive and -ve z values!

Why there is no 1-sample z-test function in base R packages?

There is no one-sample z-test in R because:

- T-test can be used for small as well as large samples as t-distribution behaves like z-distribution when n>=30
- Thus, we don’t need one-sample z-test in R!

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

One-sample t-test: Pop. SD not required and can work for both small and large samples!

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

- **t.test(x, mu=?) in the R syntax**
- t.test(mtcars$mpg, mu=20)
  - P-value for mu ≠ 20? (same as z-test?)
- t.test(mtcars$mpg, mu=10)
  - P-value for mu ≠ 10?
- t.test(mtcars$mpg, mu=30)
  - P-value for mu ≠ 30?

One-sample t-test results: n = 32!

```
> t.test(mtcars$mpg, mu=20)

        One Sample t-test

data:  mtcars$mpg
t = 0.08506, df = 31, p-value = 0.9328
alternative hypothesis: true mean is not
equal to 20
95 percent confidence interval:
17.91768         22.26357 (20 lies here!)
sample estimates:
mean of x = 20.09062
```

```
> t.test(mtcars$mpg, mu=10)

        One Sample t-test

data:  mtcars$mpg
t = 9.471, df = 31, p-value = 1.155e-10
alternative hypothesis: true mean is not
equal to 10
95 percent confidence interval:
17.91768         22.26357 (10 does not lie!)
sample estimates:
mean of x = 20.09062
```

Two-independent samples t-test (student):

- It is used to compare means of a dependent variable by grouped independent variable with two categories
- For example, we can compare exam score (dependent variable) by sex of the students i.e. male and female categories!

Assumptions:

- Dependent variable must follow the normal distribution for each category (Test of normality-GOF)

&

- Variance across independent variable categories are homogenous i.e. equal (Test of equal variance-GOF)

Two-independent samples t-test (Welch):

- It is used to compare means of a dependent variable by grouped independent variable with two categories

- For example, we can compare exam score by sex of the students!

Assumptions:

- Dependent variable must follow the normal distribution for each category (Tests of normality)

&

- Variance across independent variable categories are not homogenous i.e. not equal (Test of equal variance)

E2: Compare miles per gallon by automatic and manual gear categories of “am” variable:

- Data = mtcars
- Assumptions check:
- Normality
- Equal variance

Tests of Normality: It is a GOF so we want p-value>0.05!

```
with(mtcars, shapiro.test(mpg[am == 0]))
```

W = 0.97677, p-value = 0.8987 >0.05 ~ Normal Distribution

```
with(mtcars, shapiro.test(mpg[am == 1]))
```

W = 0.9458, p-value = 0.5363 > 0.05 ~ Normal Distribution

E2: Compare miles per gallon by automatic and manual gear categories of “am” variable:

- Data = mtcars
- Assumptions check:
- Normality
- Equal variance

Tests of Group Variance: It is a GOF so we want p-value>0.05!

var.test(mpg ~ am, data = mtcars)

F = 0.38656, num df = 18, denom df = 12, **p-value = 0.06691 > 0.05 so group variance are equal!**

Both assumptions holds true!

We can use two-sample student t-test! So:

- **t.test(mpg ~ am, var.equal = T, data = mtcars)**
- H_0: $\mu_1 = \mu_2$ or H_0: $\mu_1 - \mu_2 = 0$
- H_1: $\mu_1 \neq \mu_2$ **or H_1: $\mu_1 - \mu_2 \neq 0$**
- This is NOT GOF, this is a research hypothesis so we want to accept H1 i.e. p-value<0.05.
- Decision: The reported p-value = 0.000285, which is <0.05 so we accept H_1.
- Conclusion: Milage (mpg) is statistically different among cars with automatic and manual transmission system.

```
        Two Sample t-test

data:  mpg by am
t = -4.1061, df = 30, p-value = 0.000285
alternative hypothesis: true difference in means
between group 0 and group 1 is not equal to 0
95 percent confidence interval:
-10.84837        -3.64151
sample estimates:
mean in group 0  mean in group 1
        17.14737          24.39231
```

Mean Difference: 24.39231 – 17.1437 = 7.245

Check the two-sample student t-test result with simple linear regression model:

- lm = Linear model in R

- **summary(lm(mpg ~ am, data = mtcars))**

- Linear regression gave an estimate of 7.245 (mean difference between category 1 and category 0) i.e. 17.14737 (category=0)   24.39231 (category=1)

- This difference is statistically significant and the p-value is same as given by the two-samples t-test

- **For an independent variable coded as 0 and 1, code = 0 will be the reference category and code = 1 will be the result category so use it wisely!**

```
Coefficients:
             Estimate  Std. Error t value    Pr(>|t|)
(Intercept)   17.147   1.125      15.247     1.13e-15 ***
am[T.1]       7.245    1.764      4.106      0.000285 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ‘ 1
```

Checked with “?mtcars” in R to know am categories:

- When am = 0 (automatic), milage is 17.147 per gallon
- When am = 1 (manual), milage is 17.147+7.245 per gallon **(denoted as “am[T.1]” in the model)**
- **So, we can interpret it as: The milage per gallon is 7.245 unit more for cars with manual gear system than the automatic transmission system because automatic is coded as 0 and manual is coded as 1 in the “am” variable!**

Quick Think!

- Check the attribute of the “am” variable with “?matcars” in R prompt
- Do we need to change the attribute of this variable “as factor”?
- If yes, why?
- If no, why?

- What is a “dummy” variable?
- What is its importance?
- Can we create dummy variable of categorical variables with more than two categories too!
- We will discuss it in next class!

Question/queries?

- So,
- T-test is a “supervised learning” model
- Why?

- Next class
- 1-way ANOVA
- Covariance
- Correlation
- Regression
  - Simple
  - Multiple

Thank you!

@shitalbhandary