04. HYPOTHESIS TESTING

- Statistical hypothesis is an assumption or a statement which may or may not be true concerning one or more populations.
- The purpose of hypothesis testing is to choose between two conflicting hypotheses about the value of a population parameter.
- A hypothesis test involves two hypothesis:
  - Null hypothesis(H0) : a statement to be tested
  - Alternative hypothesis(Ha): a statement that is an alternative to the null hypothesis
- The hypothesis test is aimed to test if the null hypothesis should be rejected in favor of the alternative hypothesis.
- The criterion for deciding whether to reject the null hypothesis involves a so-called test statistics.

HYPOTHESIS TESTING FOR MEAN: ONE SAMPLE

WHEN POPULATION VARIANCE IS KNOWN

Z TEST

A factory makes tins of soy beans. The desired average weight of a tin is 160g and the weights follows a normal distribution with variance of 15g. Using a sample of 20 cans, let's statistically **test whether the population weight of the tins meets the expectations.**

```
weights=c(165.1,171.5,168.1,165.6,166.8,170.0,168.8,171.1,168.8,173.6,163.5,169.9,165.4,174.4,171.8,166.0,174.6,174.5,166.4,173.8)

library("BSDA")

z.test(x=weights,
       mu = 160,
       sigma.x =15,
       alternative = "two.sided")
```

WHEN POPULATION VARIANCE IS UNKNOWN

T TEST

A factory makes tins of soy beans. The desired average weight of a tin is 160g and the weights follows a normal distribution. Using a sample of 20 cans, let's statistically **test whether the population weight of the tins meets the expectations.**

```
weights <- c(165.1,171.5,168.1,165.6,166.8,170.0,168.8,171.1,168.8,173.6,163.5,169.9,165.4,174.4,171.8,166.0,174.6,174.5,166.4,173.8)

t.test(weights,
       mu = 160,
       alternative = "two.sided")
```

HYPOTHESIS TESTING FOR MEAN: TWO SAMPLES

Case 01: equal variances

EXAMPLE

Body fat percentages of 13 males and 10 females are given in the following variables. We need to check whether body fat percentage of males differs from that of females. Note that the body fat percentages follows a normal distribution.

To check whether the variances are equal, we should use a two sample variance test first, but for this example, let's suppose variances are equal.

```
fat_m <- c(13.3,6.0,20.0,8.0,14.0,19.0,18.0,25.0,16.0,24.0,15.0,1.0,15.0)
fat_w <- c(22.0,16.0,21.7,21.0,30.0,26.0,12.0,23.2,28.0,23.0)

t.test(x = fat_w,
       y = fat_m,
       var.equal = TRUE)
```

Case 02: Unequal variances

- This test is valid for *normally distributes variables* X1 and X2 with unequal variances.

```
fat_m <- c(13.3,6.0,20.0,8.0,14.0,19.0,18.0,25.0,16.0,24.0,15.0,1.0,15.0)
fat_w <- c(22.0,16.0,21.7,21.0,30.0,26.0,12.0,23.2,28.0,23.0)

t.test(x = fat_w,
       y = fat_m,
       var.equal = FALSE)
```

HYPOTHESIS TESTING FOR MEAN: PAIRED SAMPLES

- This test is valid only for *normally distributed data* or *large samples (n>30)*

EXAMPLE

Soil samples that were taken from 15 locations were divided in half and sent to two laboratories to test. The measurements that were observed are given in the following variables. **We want to check whether the two laboratories give the same result.**

```
lab1 <- c(22,18,28,26,13,8,21,26,27,29,25,24,22,28,15)
lab2 <- c(25,21,31,27,11,10,25,26,29,28,26,23,22,25,17)

t.test(x = lab1,
       y = lab2,
       paired = TRUE)
```

HYPOTHESIS TESTING FOR PROPORTION: ONE SAMPLE

Case 01: Large sample

Check the assumptions for proportion test

- A simple random sample of size n is taken
- The conditions for the binomial distribution are satisfied.
- To determine the sampling distribution of p we need to show that $np \geq 5$ and $n(1-p) \geq 5$

If this requirement is true, then the sampling distribution of p is well approximated by a normal curve.

EXAMPLE

The following variable shows the hair colour of 3000 people. Using a sample of 1000 people we are going to **check whether the proportion of black hair is equal to 0.5.**

```
set.seed(10)
Hair_col <- c(rep("black", 1500), rep("brown", 1000), rep("blonde", 500))
sampleP <- sample(Hair_col,1000)
Ptable <-table(sampleP)

prop.test(x = 498,
          n= 1000,
          p=0.5,
          alternative = "two.sided",
          conf.level = 0.95,
          correct = FALSE)
```

Case 02: small samples

binom.test()

```
set.seed(10)
sampleS<- sample(hair_col,10)
Stable <- table(sampleS)
Stable

binom.test(x= length(Stable),
           n= length(sampleS),
           p= 0.5,
           alternative = "two.sided" )
```

HYPOTHESIS TESTING FOR PORPORTION: TWO SAMPLES

- To use this test, the sample must be *large enough*

```
prop.test(x = c(490, 400), n = c(500, 500))
```

HYPOTHESIS TESTING FOR VARIANCE: ONE SAMPLE

- This test is valid only for *normally distributed data*

check whether the variance of the population is 10

```
weights <- c(165.1,171.5,168.1,165.6,166.8,170.0,168.8,
             171.1,168.8,173.6,163.5,169.9,165.4,174.4,
             171.8,166.0,174.6,174.5,166.4,173.8)

library(EnvStats)
varTest(weights,
        sigma =10,
        alternative = "two.sided")
```

HYPOTHESIS TESTING FOR VARIANCE: TWO SAMPLES

- Valid only for *normally distributed samples*

We can actually check whether the variances are equal or not.

```
fat_m <- c( 13.3,6.0,20.0,8.0,14.0,19.0,18.0,25.0,16.0,24.0,15.0,1.0,15.0)
fat_w <- c(22.0,16.0,21.7,21.0,30.0,26.0,12.0,23.2,28.0,23.0)

var.test(x = fat_m,
         y = fat_w)
```

CHECK THE NORMALITY ASSUMPTION

Shapiro-Wilk test

```
shapiro.test(data)
```

Anderson-Darling test

```
library(nortest)
ad.test(data)
```

Kolmogorov-Smirnov test

```
ks.test(data, "pnorm")
```

NON-PARAMETRIC TESTS

- Non-parametric tests are distribution free
- The only assumption holds for these tests is that the data should be an independent random sample

SIGN TEST

Case 01: One sample

- **Compare the true median of a sample with a theoretical value**

The median price of one-bedroom flats in New York in 2008 was 130,000 dollars. We are given a sample of 32 flats (in 1000 dollars) in 2009 and we need to check whether the prices are rising than in 2008.

```
m0 <- 130 # median in 2008
prices <- c(230.00,148.00,126.00,134.62,155.00,157.70,
            160.00,225.00,125.00,109.00,157.00,115.00,
            125.00,225.00,118.00,179.00,176.00,125.00,
            123.00,180.00,151.00,120.00,143.00,170.00,
            190.00,233.00,148.72,189.00,121.00,149.00,
            225.00,240.00)

library(BSDA)
SIGN.test(x = prices,
          md = m0,
          alternative="greater")
```

Case 02: For two paired samples

- **To compare the true median of two paired samples**

Soil samples that were taken from 15 locations were divided in half and sent to two laboratories to test. The measurements that were observed are given in the following variables. Note that no assumptions are made.

```
lab1 <- c(22,18,28,26,13,8,21,26,27,29,25,24,22,28,15)
lab2 <- c(25,21,31,27,11,10,25,26,29,28,26,23,22,25,17)

library(BSDA)
SIGN.test(x = lab1,
          y = lab2,
          alternative = "two.sided",
          conf.level = 0.95)
```

WILCOXON SIGN RANK TEST

Case 01: One sample Sign Rank Test

- Alternative test for sign test which **uses not only the sign but also the rank difference into account**

Let’s apply the Wilcoxon sign rank test for the same flat prices example considered in sign test.

```
m0 <- 130 # median in 2008
prices <- c(230.00,148.00,126.00,134.62,155.00,157.70,
            160.00,225.00,125.00,109.00,157.00,115.00,
            125.00,225.00,118.00,179.00,176.00,125.00,
            123.00,180.00,151.00,120.00,143.00,170.00,
            190.00,233.00,148.72,189.00,121.00,149.00,
            225.00,240.00)

wilcox.test(x = prices,
            md = m0,
            exact = FALSE,
            alternative = "greater")
```

Case 02: Sign Rank Test for Two Paired Samples

```
lab1 <- c(22,18,28,26,13,8,21,26,27,29,25,24,22,28,15)
lab2 <- c(25,21,31,27,11,10,25,26,29,28,26,23,22,25,17)

wilcox.test(x = lab1,
            y = lab2,
            paired = TRUE,
            exact = FALSE)
```

Case 03: Sign rank test for two independent samples

Body fat percentages of 10 males and females are given in the following variables. We need to check whether median body fat percentage of males differs from that of females.

```
fat_m <- c( 13.3,6.0,20.0,8.0,14.0,19.0,18.0,25.0,16.0,24.0)
fat_w <- c(22.0,16.0,21.7,21.0,30.0,26.0,12.0,23.2,28.0,23.0)

wilcox.test(x = fat_m,
            y = fat_w,
            alternative = "two.sided")
```