ECO 274 LAB: Inferential Statistics and t-test

Learning Objectives

By the end of this lecture students will be able to learn about the:

- Applications of t-test in applied economics research
- How to interpret t-test results
- Interpretations of p-value and significance level α
- Wilcoxon test to be used to know how to compare two groups under the non-normality assumption
- One sample binomial test
- The power analysis of t-test.

Inferential statistics, as opposed to descriptive statistics, is a branch of statistics defined as the science of drawing conclusions about a population from observations made on a representative sample of that population. In drawing conclusion, t-test is utilized as the individual significance test when normality assumption holds.

One-sample t-test

In economics and finance, the t-test is often used in research when the researcher wants to know if there is a significant difference between the mean of the sample and the population, or whether there is a significant difference between the means of two different groups. There are two types of t-tests: the one-sample t-test and the two samples t-test.

What is a One-sample t-test for the mean?

The one-sample t-test for the mean is used to test whether the population means are equal to the pre-defined (standard/hypothetical) mean (μ) value when the population standard deviation is **unknown,** and the sample size is small.

Assumptions for One Sample Mean t-test

- The parent population from which the sample is drawn should be normal.
- The sample observations should be independent of each other i.e., sample should be random.
- The population standard deviation is unknown.

Hypothesis for the one sample t-test for mean

Let μ_0 denote the hypothesized value for the mean, and $\bar{x}$ denote the sample mean.

Null Hypothesis:

- $H_0 : \bar{x} = \mu_0$ The population means is equal to hypothesized(standard) mean.

Alternative Hypothesis: Three forms of alternative hypothesis are as follows:

- H_a : $\bar{x} < \mu_0$ Sample means is less than the hypothesized mean. It is called the lower-tail test (left-tailed test).
- H_a : $\bar{x} > \mu_0$ Sample means is greater than the hypothesized mean. It is called the Upper tail test (right-tailed test).
- H_a : $\bar{x} \neq \mu_0$ Sample mean is not equal to hypothesized mean. It is called two tail tests.

The formula for the test statistic of one sample t-test for the mean is:

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

where:

$\bar{x}$: observed sample mean

μ_0 : hypothesized population mean

n: sample size

s: sample standard deviation with n-1 degree of freedom

Procedure to perform One Sample t-test for mean.

Step 1: Define both the Null Hypothesis and Alternate Hypothesis.
Step 2: Decide the level of significance α (i.e., alpha).
Step 3: Check the assumptions for the one-sample t-test for the mean using the below function.
Step 4: Calculate the test statistic using R's test () function.
Step 5: Interpret the t-test results.
Step 6: Determine the rejection criteria for the given confidence level and interpret the results to determine whether the test statistic lies in the rejection or non-rejection regions.

	Left-tailed Test	Right-tailed Test	Two-tailed Test
Decision Rule: p-value approach (where α is the level of significance)	If p-value ≤α, then Reject H0	If p-value ≤α then Reject H0	If p-value ≤α then Reject H0

Decision Rule: Critical-value approach	If $t \leq -t\alpha$ then Reject H0	If $t \geq t\alpha$ then Reject H0	If $t \leq -t\alpha$ or $t \geq t\alpha$ then Reject H0

Application in economic analysis

Example of One Sample t-test in R
Twitter Inc. claims that its' economic data scientists spend, on average, 553 hours per month on data cleaning, visualizing, and prediction tasks. We assume that the average monthly hours are less than the company claimed. As a new data scientist, you decided to test the claim by inspecting six randomly selected data scientists' work hours in the Twitter head office in California, and get the following information:

544, 551, 548, 556, 549, 554.

Assume that the hours of work in the head office follow a normal distribution. Test the claim at a 5% level of significance.

Solution: Given data:

sample size (n) = 6

hypothesized mean value ($\mu 0$) = 553

level of significance (α) = 0.05

confidence level = 0.95

Let's solve this example by the step-by-step procedure.

Step 1: Define the Null Hypothesis and Alternate Hypothesis.

let μ be the mean hours/month

Null Hypothesis: the mean hours is equal to 553 hours

$H0 : \mu = 553$

Alternate Hypothesis: the mean hours of Twitter data scientist is less than 553 hours

$Ha : \mu < 553$

Step 2: level of significance (α) = 0.05

Step 3: Let's check the assumptions.

```
# Define given dataset
dataset <- c(546, 551, 548, 556, 549, 554)
#Create qqplot for the dataset
qqnorm(dataset)
qqline(dataset)

# Since the data lies close to the line y=x and has no big deviations from
#the line, it’s fine to consider the sample as coming from a normal
#distribution. We can proceed further with our hypothesis test

# Perform one-sample t-test
t.test( x= dataset,mu=553, alternative = "less",conf.level = 0.95)
## For the two-tailed test, alt = "two.sided"
```

How to interpret t-test results in R?

Let’s see the interpretation of t-test results in R.

data: This gives information about the data set used in the one-sample t-test. In this, we use dataset vectors as data.

t: It is the test statistic of the t-test. In our case, test statistic = -1.5119

df: It is the degree of freedom for the t-test statistic. In our case, df=5

p-value: This is the p-value corresponding to the t-test statistic, i.e., -1.5119, and degree of freedom, i.e., 5. In our case, the p-value is 0.09549.

alternative: It is the alternative hypothesis used for the t-test. In our case, an alternative hypothesis is if the mean is less than 553, i.e., left, tailed.

Ninety-five percent (95%) confidence interval: This gives us a 95% confidence interval for the true mean. Here the 95% confidence interval is [-∞,553.8875].

Sample estimates: These give the sample mean. In our case sample mean is 550.33

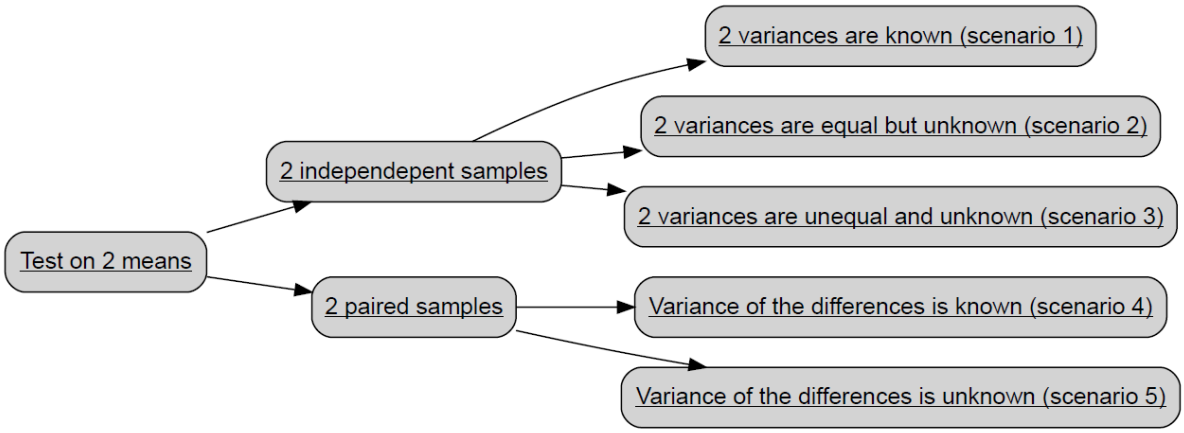
Step 6: Determine the rejection criteria for the given confidence level and conclude the results whether the test statistic lies in the rejection region or non-rejection region.

Conclusion/Inference drawn:

Since the p-value [0.09549] is not less than the level of significance (α) = 0.05, we fail to reject the null hypothesis. This means we do not have sufficient evidence to say that the mean working hours of the economic data scientist on Twitter is different from 553 hours/month.

T-test for two samples: Different versions of the student's t-test

There are several versions of the student's t-test for two samples, depending on whether the samples are independent or paired and depending on whether the variances of the populations are (un) equal and/or (un) known:



Independent samples means that the two samples are collected on different experimental units or different individuals, for instance when we are working on the housing prices from emerging economies and developed economies separately or working on microcredit receivers (household) who have been randomly assigned to a control and a treatment group (and a household belongs to only one group).

The treatment group (also called the experimental group) receives the treatment (microcredit, subsidy, nutritional aids) whose effect the researcher is interested in. The control group receives either no treatment, a standard treatment whose effect is already known, or a placebo (a fake treatment).

Paired samples when measurements are collected on the same experimental units, same individuals. This is often the case, for example in medical studies, when testing the efficiency of a treatment at two different times. The same patients are measured twice, before and after the treatment, and the dependency between the two samples must be taken into account in the computation of the test statistic by working on the differences of measurements for each subject. Other criteria for choosing the appropriate version of the student's t-test are whether the variances of the populations (not the variances of the samples!) are known or unknown and equal or unequal. This criterion is rather straightforward, we either know the variances of the populations or we do not.

Scenario 1: Independent samples with 2 known variances

```
dat1 <- data.frame(
  sample1 = c(0.9, -0.8, 0.1, -0.3, 0.2),
  sample2 = c(0.8, -0.9, -0.1, 0.4, 0.1)
)
dat1
```

```
dat_ggplot <- data.frame(
  value = c(0.9, -0.8, 0.1, -0.3, 0.2, 0.8, -0.9, -0.1, 0.4, 0.1),
  sample = c(rep("1", 5), rep("2", 5))
)

library(ggplot2)

ggplot(dat_ggplot) +
  aes(x = sample, y = value) +
  geom_boxplot() +
  theme_minimal()

boxplot(value ~ sample,
        data = dat_ggplot)
```

The two boxes seem to overlap which illustrate that the two samples are quite similar, so we tend to believe that we will not be able to reject the null hypothesis that the two population means are similar. However, only a formal statistical test will confirm this belief.

```
library(BSDA)

z.test(dat1$sample1,
       dat1$sample2,
       alternative = "two.sided",
       mu = 0,
       sigma.x = 1,
       sigma.y = 1,
       conf.level = 0.95)
```

A note on p-value and significance level α

P-value: The p-value is a probability and as any probability, it goes from 0 to 1. The test is often interpreted using a p-value, which is the probability of observing the result given that the null hypothesis is true, not the reverse, as is often the case with misinterpretations. So, p-value (p) if the probability of obtaining a result equal to or more extreme than was observed in the data.
In interpreting the p-value of a significance test, you must specify a significance level, often referred to as the Greek lowercase letter alpha (a). A common value for the significance level is 5% written as 0.05. In some sense, it gives you an indication of how likely your null hypothesis is. It is also defined as the smallest level of significance for which the data indicate rejection of the null hypothesis.

The significance level α, derived from the threshold of 5% mentioned earlier, is the probability of rejecting the null hypothesis when it is in fact true. In this sense, it is an error (of 5%) that we accept to deal with, in order to be able to draw conclusions. If we would accept no error (an error of 0%), we would not be able to draw any conclusion about the population(s) since we only have access to a limited portion of the population(s) via the sample(s).

Note that we say, “we do not reject the null hypothesis”, and not “we accept the null hypothesis”. This is because it may be the case that the null hypothesis is in fact false, but we failed to prove it with the samples. Also, an important note that statistical significance is not equal to economic (scientific) significance. To this end, a result may be statistically significant (a p-value < α), but of little or no interest from a scientific point of view (because the effect is so small that it is negligible and/or useless for instance).

Scenario 2: Independent samples with 2 equal but unknown variances

```
dat2 <- data.frame(
  sample1 = c(1.78, 1.5, 0.9, 0.6, 0.8, 1.9),
  sample2 = c(0.8, -0.7, -0.1, 0.4, 0.1, NA)
)
dat2
dat_ggplot <- data.frame(
  value = c(1.78, 1.5, 0.9, 0.6, 0.8, 1.9, 0.8, -0.7, -0.1, 0.4, 0.1),
  sample = c(rep("1", 6), rep("2", 5))
)

ggplot(dat_ggplot) +
  aes(x = sample, y = value) +
  geom_boxplot() + theme_minimal()
```

Unlike the previous scenario, the two boxes do not overlap which illustrates that the two samples are different from each other. From this boxplot, we can expect the test to reject the null hypothesis of equal means in the populations. Nonetheless, only a formal statistical test will confirm this expectation.

```
test <- t.test(dat2$sample1, dat2$sample2,
               var.equal = TRUE, alternative = "greater")
test
test$p.value
```

Scenario 3: Independent samples with 2 unequal and unknown variances

```
dat3 <- data.frame(
  value = c(0.8, 0.7, 0.1, 0.4, 0.1, 1.78, 1.5, 0.9, 0.6, 0.8, 1.9),
  sample = c(rep("1", 5), rep("2", 6))
)
dat3


ggplot(dat3) +
  aes(x = sample, y = value) +
  geom_boxplot() +
  theme_minimal()


test <- t.test(value ~ sample,
               data = dat3,
```

```
                var.equal = FALSE,
                alternative = "less"
)
test
```

Scenario 4: Paired samples where the variance of the differences is known

```
dat4 <- data.frame(
  before = c(0.9, -0.8, 0.1, -0.3, 0.2),
  after = c(0.8, -0.9, -0.1, 0.4, 0.1)
)
dat4
dat4$difference <- dat4$after - dat4$before

ggplot(dat4) +
  aes(y = difference) +
  geom_boxplot() +
  theme_minimal()

t.test_pairedknownvar <- function(x, V, m0 = 0, alpha = 0.05, alternative =
"two.sided") {
  M <- mean(x)
  n <- length(x)
  sigma <- sqrt(V)
  S <- sqrt(V / n)
  statistic <- (M - m0) / S
  p <- if (alternative == "two.sided") {
    2 * pnorm(abs(statistic), lower.tail = FALSE)
  } else if (alternative == "less") {
    pnorm(statistic, lower.tail = TRUE)
  } else {
    pnorm(statistic, lower.tail = FALSE)
  }
  LCL <- (M - S * qnorm(1 - alpha / 2))
  UCL <- (M + S * qnorm(1 - alpha / 2))
  value <- list(mean = M, m0 = m0, sigma = sigma, statistic = statistic,
p.value = p, LCL = LCL, UCL = UCL, alternative = alternative)
  # print(sprintf("P-value = %g",p))
  # print(sprintf("Lower %.2f%% Confidence Limit = %g",
  #               alpha, LCL))
  # print(sprintf("Upper %.2f%% Confidence Limit = %g",
  #               alpha, UCL))
  return(value)
}

test <- t.test_pairedknownvar(dat4$after - dat4$before,
                              V = 1
)
test
test$p.value
```

Scenario 5: Paired samples where the variance of the differences is unknown

```
dat5 <- data.frame(
  before = c(9, 8, 1, 3, 2),
  after = c(16, 11, 15, 12, 9)
)
dat5
dat5$difference <- dat5$after - dat5$before

ggplot(dat5) +
  aes(y = difference) +
  geom_boxplot() +
  theme_minimal()
test <- t.test(dat5$after, dat5$before,
               alternative = "greater",
               paired = TRUE)
test
```

The p-value is 0.006 so at the 5% significance level we reject the null hypothesis of the mean of the differences being equal to 0, meaning that we can conclude that the treatment is effective in increasing the running capabilities (because the mean of the differences is greater than 0).

Wilcoxon test

In some cases, the mean is not appropriate to compare two samples, so the median is used to compare them via the Wilcoxon test. Wilcoxon test is used to know how to compare two groups under the non-normality assumption.
The two groups to be compared are either:

- independent, or
- paired (i.e., dependent)

There are actually two versions of the Wilcoxon test:

- The Mann-Whitney-Wilcoxon test (also referred as Wilcoxon rank sum test or Mann-Whitney U test) is performed when the samples are independent (so this test is the non-parametric equivalent to the Student's t-test for independent samples).
- The Wilcoxon signed-rank test (also sometimes referred as Wilcoxon test for paired samples) is performed when the samples are paired/dependent (so this test is the non-parametric equivalent to the Student's t-test for paired samples).

Case 1: Independent samples

For the Wilcoxon test with independent samples, suppose that we want to test whether income at the urban households' level differ between developed and low-income country. We have collected incomes for 24 developed and low-income country's households (12 of each):

```
dat <- data.frame(
  country = as.factor(c(rep("dev", 12), rep("lowInc", 12))),
  income = c(
    19, 18, 9, 17, 8, 7, 16, 19, 20, 9, 11, 18,
    16, 5, 15, 2, 14, 15, 4, 7, 15, 6, 7, 14
  )
)

dat
library(ggplot2)

ggplot(dat) +
  aes(x = country, y = income) +
  geom_boxplot(fill = "darkseagreen1") +
  theme_minimal()
```

We first check whether the 2 samples follow a normal distribution via a histogram and the Shapiro-Wilk test:

```
hist(subset(dat, country == "dev")$income,
     main = "Income for Dev",
     xlab = "Income"
)
hist(subset(dat, country == "lowInc")$income,
     main = "Income for low Income country",
     xlab = "Income"
)

shapiro.test(subset(dat, country == "dev")$income)
shapiro.test(subset(dat, country == "lowInc")$income)
```

The histograms show that both distributions do not seem to follow a normal distribution and the p-values of the Shapiro-Wilk tests confirm it (since we reject the null hypothesis of normality for both distributions at the 5% significance level).

The null and alternative hypotheses of the Wilcoxon test are as follows:

H0: the 2 groups are equal in terms of the variable of interest
H1: the 2 groups are different in terms of the variable of interest

Applied to our research question, we have:

H0: Incomes of developed and low-income countries are equal
H1: Incomes of developed and low-income countries are different (not equal)

```
test <- wilcox.test(dat$income ~ dat$country)
test
```

The p-value is 0.021. Therefore, at the 5% significance level, we reject the null hypothesis, and we conclude that income level is significantly different between developed and low-income countries.

Paired samples

For this second scenario, consider that we administered an nutritional intervention project of child nutritional outcomes in a district of 12 households at the beginning of a project and that we administered a similar intervention at the middle of the project to the exact same households. We have the following data:

```
dat2 <- data.frame(
  Beginning = c(16, 5, 15, 2, 14, 15, 4, 7, 15, 6, 7, 14),
  Middle = c(19, 18, 9, 17, 8, 7, 16, 19, 20, 9, 11, 18)
)

dat2

#We transform the dataset to have it in a tidy format:

dat2 <- data.frame(
  Time = c(rep("Before", 12), rep("Middle", 12)),
  Outcome = c(dat2$Beginning, dat2$Middle)
)
dat2
```

The distribution of the nutritional outcomes at the beginning and middle of the project:

```
# Reordering dat2$Time
dat2$Time <- factor (dat2$Time,
                     levels = c ("Before", "Middle")
)

ggplot(dat2) +
  aes(x = Time, y = Outcome) +
  geom_boxplot(fill = "darkseagreen1") +
  theme_minimal()
```

In this example, it is clear that the two samples are not independent since the same 12 households took the project before and middle of the project. Supposing also that the normality assumption is violated (and given the small sample size), we thus use the Wilcoxon test for paired samples, with the following hypotheses:

H0: outcomes before and at the middle of the project are equal
H1: outcomes before and at the middle of the project are not equal

```
test <- wilcox.test(dat2$Outcome ~ dat2$Time,
                    paired = TRUE)
test
```

The p-value is 0.169. Therefore, at the 5% significance level, we do not reject the null hypothesis that the outcomes are similar before and at the middle of the project.

Power Analysis for t-test

The statistical power of a hypothesis test is the probability of detecting an effect if there is a true effect present to detect. Power can be calculated and reported for a completed experiment to comment on the confidence one might have in the conclusions drawn from the study results. It can also be used to estimate the number of observations or sample size required to detect an effect in an experiment. A power analysis can estimate the minimum sample size required for an experiment, given the desired significance level, effect size, and statistical power.

Power Analysis

Statistical power analysis has four related parts; they are

Effect Size. The quantified magnitude of a result present in the population. Effect size is calculated using a specific statistical measure, such as Pearson's correlation coefficient for the relationship between variables or Cohen's d for the difference between groups. A common measure for comparing the difference in the mean between two groups is Cohen's d measure. It calculates a standard score that describes the difference in terms of the number of standard deviations that the means are different. The large effect size for Cohen's d is 0.80 or higher, as is commonly accepted when using the measure.

Significance. The significance level used in the statistical test, e.g., alpha. The test will calculate a p-value that can be interpreted as to whether the samples are the same (fail to reject the null hypothesis), or there is a statistically significant difference between the samples (reject the null hypothesis). A common significance level for interpreting the p-value is 5% or 0.05.

Statistical Power: The probability of accepting the alternative hypothesis if it is true. We can use the default and assume a minimum statistical power of 80% or 0.8. Power analyses are normally run before a study is conducted. A prospective or a priori power analysis can be used to estimate any of the four power parameters but is most often used to estimate required sample sizes.

Sample Size. The number of observations in the sample. That is, how many observations are required from each sample to at least detect an effect of 0.80 with an 80% chance of detecting the effect if it is true (20% of a Type II error) and a 5% chance of detecting an effect if there is no such effect (Type I error).

One Sample Binomial Test

The one-sample binomial test makes a statistical inference about the proportion parameter by comparing it with a hypothesized value. The methods for estimating the power for such a test are either the normal approximation or the binomial enumeration.
Suppose an experiment has the following characteristics:

- the experiment consists of *n* independent trials, each with two mutually exclusive possible outcomes (which we will call **success** and **failure**)
- for each trial, the probability of success is *p* (and so the probability of failure is 1 – *p*)
- Each such trial is called a **Bernoulli trial**. Let *x* be the discrete random variable whose value is the number of successes in *n* trials. Then the probability distribution function for *x* is called the **binomial distribution**, *B*(*n, p*), whose frequency function (aka probability density function) is:

$$f(x) = C(n, x)p^x (1–p)^{n-x}$$

where $C(n, x) = \frac{n!}{x!(n-x)!}$ and $n! = n(n–1)(n–2)$

$$Mean = np$$

$$Var = np(1–p)$$

A binomial test compares a sample proportion to a hypothesized proportion. The test has the following null and alternative hypotheses:
H_0 : π = p (the population proportion π is equal to some value p)
H_a : π ≠ p (the population proportion π is not equal to some value p)

```
#Binomial
#The binomial random numbers are discrete random numbers.
#They have the distribution of the number of successes in
#n independent Bernoulli trials where a Bernoulli trial results
#in success or failure, success with probability p.

n=1; p=.5                    # set the probability


# Find 10 random values from a sample of 1 with probability of 0.5.
rbinom(10,n,p)               # 10 different such numbers
```

```
n = 10; p=.5
rbinom(1,n,p)

rbinom(5,n,p)                  # 5 binomial number

#The following codes will show 100 binomially distributed random numbers

n=30;p=.25                     # set appropriate prob and number
x=rbinom(100,n,p)             # 100 random numbers
x
hist(x,probability=TRUE,col="darkseagreen1")
```

Applications in economics

We have a microcredit experimental project in which 80% of our credit borrowers are from a group of “No Collateral” and 20% are from “With Collateral.” We wanted to know if we sampled a borrower from this experiment, what would be the probability of selecting 2 borrowers from “With Collateral” and 3 borrowers from “No Collateral” if we are selecting them in groups of 5.

So, getting a “No Collateral” borrower to be a hit for the purposes of this example. So, in this case, p = 0.8, n = 5, and X = 3, which gives us the following:

```
dbinom(3, size = 5, prob = 0.8)
```

So, we would have about a 20.5% chance of selecting 3 “No Collateral” borrower from this group if we sampled them in groups of 5. What if we wanted to know the probability of getting 3 or more “No Collateral” borrower in groups of 5? This would be

Pr[3 or more] = Pr[3] + Pr[4] + Pr[5]

So, we could do the dbinom() function two more times and sum them:

```
dbinom(4, size = 5, prob = 0.8)
dbinom(5, size = 5, prob = 0.8)
# Pr[3 or more] = 0.21 + 0.41 + 0.33 = 0.95. This means that we’d expect to
#get 3 or more “No Collateral” borrower 95% of the time with repeated
#sampling.

##Alternatively

# calculate the number of hits from 3 to 5
xsuccesses <- 3:5

# do each calculation
```

```
probx <- dbinom(xsuccesses, size = 5, prob = 0.8)

# make a table from those two values
probTable <- data.frame(xsuccesses, probx)

# display the tab
```

Sources acknowledged:

- https://statstutorial.com
- https://statsandr.com/blog/
- https://machinelearningmastery.com
- Statistical Analysis with R For Dummies by Joseph Schmuller.
- The Book of R: A First Course in Programming and Statistics by Tilman M. Davies.