Politecnico di Bari

Politecnico di Bari

Dipartimento di Ingegneria Elettrica e dell'Informazione

Corso di Laurea Triennale in Ingegneria dei Sistemi Medicali

DIPARTIMENTO DI INGEGNERIA ELETTRICA E DELL'INFORMAZIONE

Bioinformatics and Big Data Analytics

Statistical Hypothesis Testing

Eng. Nicola **Altini**, *Ph.D. Student*
Eng. Giacomo Donato **Cascarano**, *Ph.D. Student*
Prof. *Eng*. Vitoantonio **Bevilacqua**, *Ph.D.*

INDUSTRIAL INFORMATICS LAB
LABORATORIO DI INFORMATICA INDUSTRIALE

Anno Accademico 2019/2020

apulian bioengineering company

Hypothesis Testing

- A statistical hypothesis, sometimes called confirmatory data analysis, is a hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables.
- A statistical hypothesis test is a method of statistical inference. Commonly, two statistical data sets are compared, or a data set obtained by sampling is compared against a synthetic data set from an idealized model.
- A hypothesis is proposed for the statistical relationship between the two data sets, and this is compared as an alternative to an idealized null hypothesis that proposes no relationship between two data sets.

Hypothesis Testing

- The comparison is deemed statistically significant if the relationship between the data sets would be an unlikely realization of the null hypothesis according to a threshold probability—the significance level. Hypothesis tests are used when determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance.
- The process of distinguishing between the null hypothesis and the alternative hypothesis is aided by considering two conceptual types of errors. The first type of error occurs when the null hypothesis is wrongly rejected. The second type of error occurs when the null hypothesis is wrongly not rejected.

Hypothesis Testing

- The most common approach to hypothesis testing is to establish a set of two mutually exclusive and exhaustive hypotheses about the true value of the parameter in question.
- Then, our sample statistics will be used to support one or the other of the hypothesized alternatives.
- H_0: **Null Hypothesis**
  - What we assume is true to begin with.
  - Typically that there is no difference/effect/relationship etc.
- H_1: **Research (Alternative) Hypothesis**
  - What we aim to gather evidence of.
  - Typically that there is a difference/effect/relationship etc.

Hypothesis Testing

- **Exhaustive hypotheses sets**. Let θ stand for any particular parameter we wish, our hypotheses set can be expressed:

$$H_0: \theta = a$$
$$H_1: \theta \neq a$$

- **Truncated hypotheses sets**. Let θ stand for any particular parameter we wish, our hypotheses set can be expressed:

$$H_0: \theta = a$$
$$H_1: \theta < a$$

- Or:

$$H_0: \theta = a$$
$$H_1: \theta > a$$

- When using such a truncated or curtailed or *one-sided* hypotheses set we must be absolutely certain, usually on logical grounds, that the third and omitted possibility has a zero probability of occurence.

Hypothesis Testing

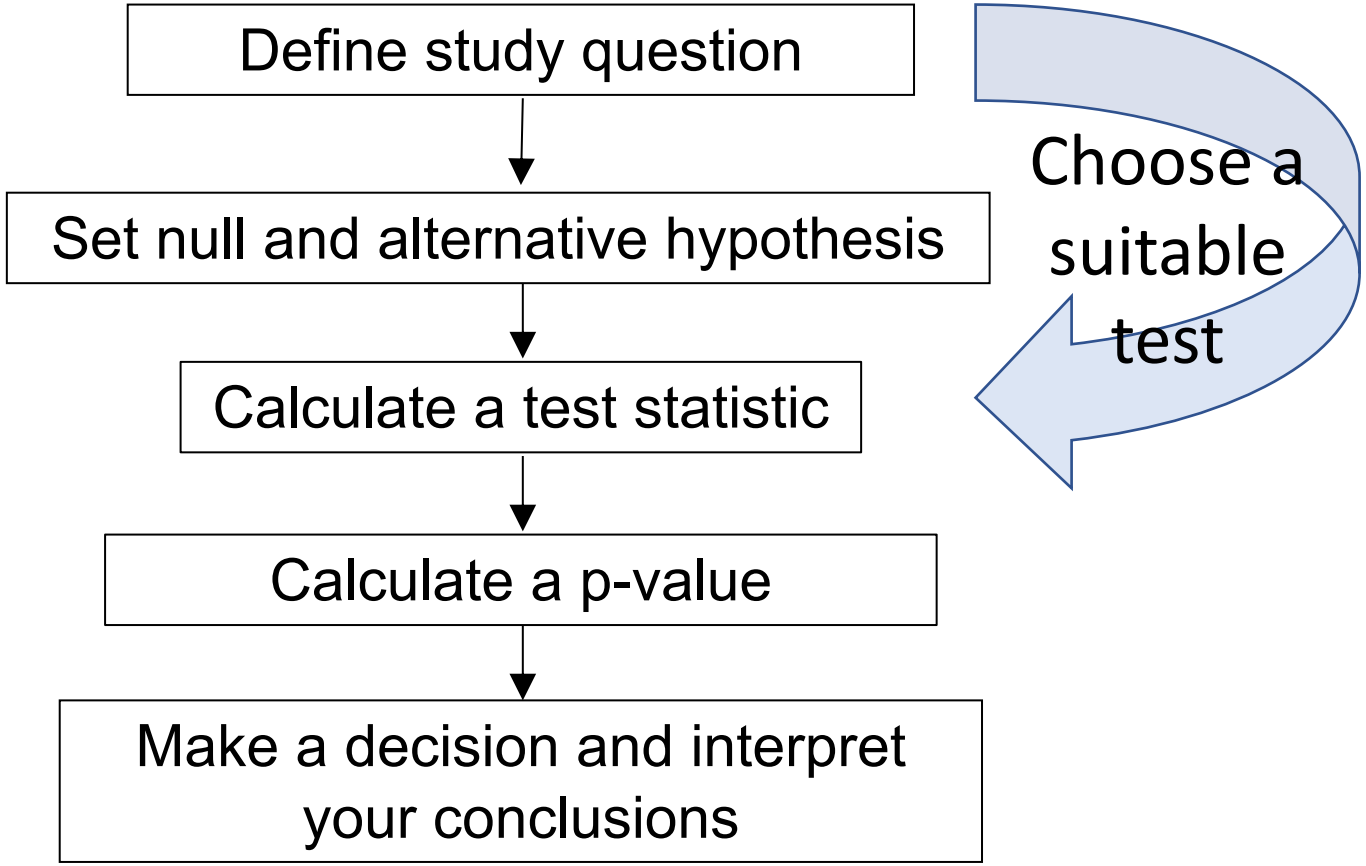
- **Test of a mean difference between paired measures**.
- Very often, we measure a set of objects on two different variables, say x and y.
- We have data in the form of n pairs of observations, one of each pair being a measurement on variable x, and the other a measurement on variable y.
- We can create a derived variable D defined as the difference between the paired values on variable x and y; i.e., $D = x - y$. Consequently, a difference score $D_i = x_i - y_i$ can be determined for each individual object, or pair of objects, whichever the case may be. Furthermore, we can determine the mean of the n differences scores, $\overline{D}$, as well as their standard deviation s_D.
- We can then establish the hypotheses set:

$$H_0: \mu_D = 0$$
$$H_1: \mu_D \neq 0$$

- In which the hypothesis H_0 corresponds to the working assumption that the mean of the population of difference scores is zero.

Steps for Hypothesis Testing

Set the p-value significance threshold



Differential gene expression analysis

- Based on a count table, we want to detect differentially expressed genes between different conditions.
  - How can we detect genes for which the counts of reads change between conditions more systematically than as expected by chance?
- We would like to use statistical testing to decide whether, for a given gene, an observed difference in read counts is significant, that is, whether it is greater than what would be expected just due to natural random variation.
- Null hypothesis H_0:
  - the gene g is not differentially expressed between the conditions
- Alternative hypothesis H_1:
  - the gene g is differentially expressed between the conditions

Hypothesis Testing

- How to quantify the difference?
- The statistical tests do not give a simple answer of whether the hypothesis is true or not. What a statistical test determines is how likely that null hypothesis is to be true.
- After a test statistic is computed, it is often converted to a p-value. Then the difference is quantified in terms of the p-value.
- If the p-value is small then the null hypothesis is deemed to be untrue and it is rejected in favour of the alternative.
- The p-value is the probability of seeing a result as extreme or more extreme than the observed data, when the null hypothesis is true.
- It is a usual convention in biology to use a critical p-value of 0.05.

Type of errors in tests

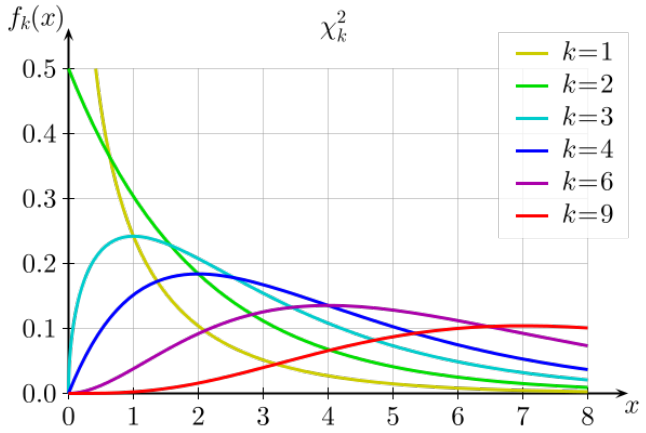
Table of error types		Null hypothesis (H_0) is	
		True	**False**
Decision about null hypothesis (H_0)	**Don't reject**	Correct inference (true negative) (probability = 1 - α)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = 1 - β)

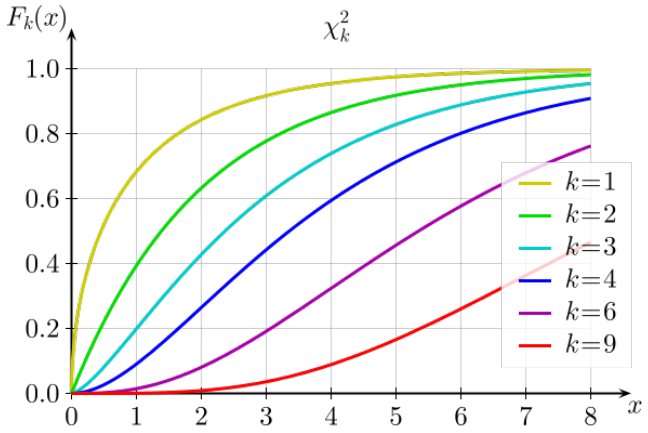
Type of errors in tests

- **Type I error** (false positive – FP): concluding that H_0 is false when H_0 is true.
  - The probability of committing a **Type I error** is simply equal to the significance level α which we use as our criterion for judging whether our sample statistic deviates an unlikely amount from the hypothesized value.
- **Type II error** (false negative – FN): concluding that H_0 is true when H_0 is false.
  - The probability of committing a **Type II error**, which we designate with β, depends on a number of factors, including:
    1. The true value of the parameter in question
    2. The significance level α we use to evaluate our working hypothesis H_0 and whether we use a one-tailed or two-tailed test
    3. The standard deviation σ of the sampled population
    4. The size of our sample n

Chi-squared distribution

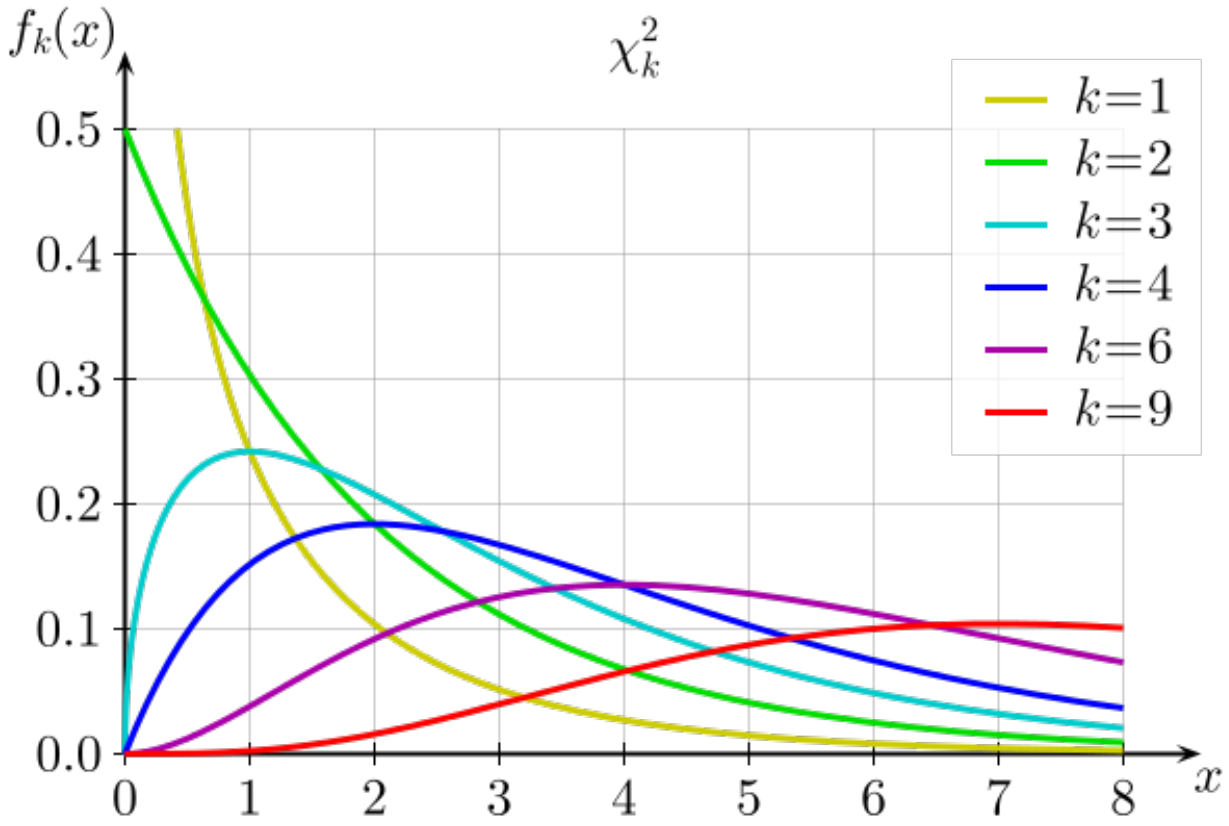
- The chi-square distribution (also chi-squared or χ^2-distribution) with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables.
- The chi-square distribution is a special case of the gamma distribution and is one of the most widely used probability distributions in inferential statistics, notably in hypothesis testing and in construction of confidence intervals.





Chi-squared distribution

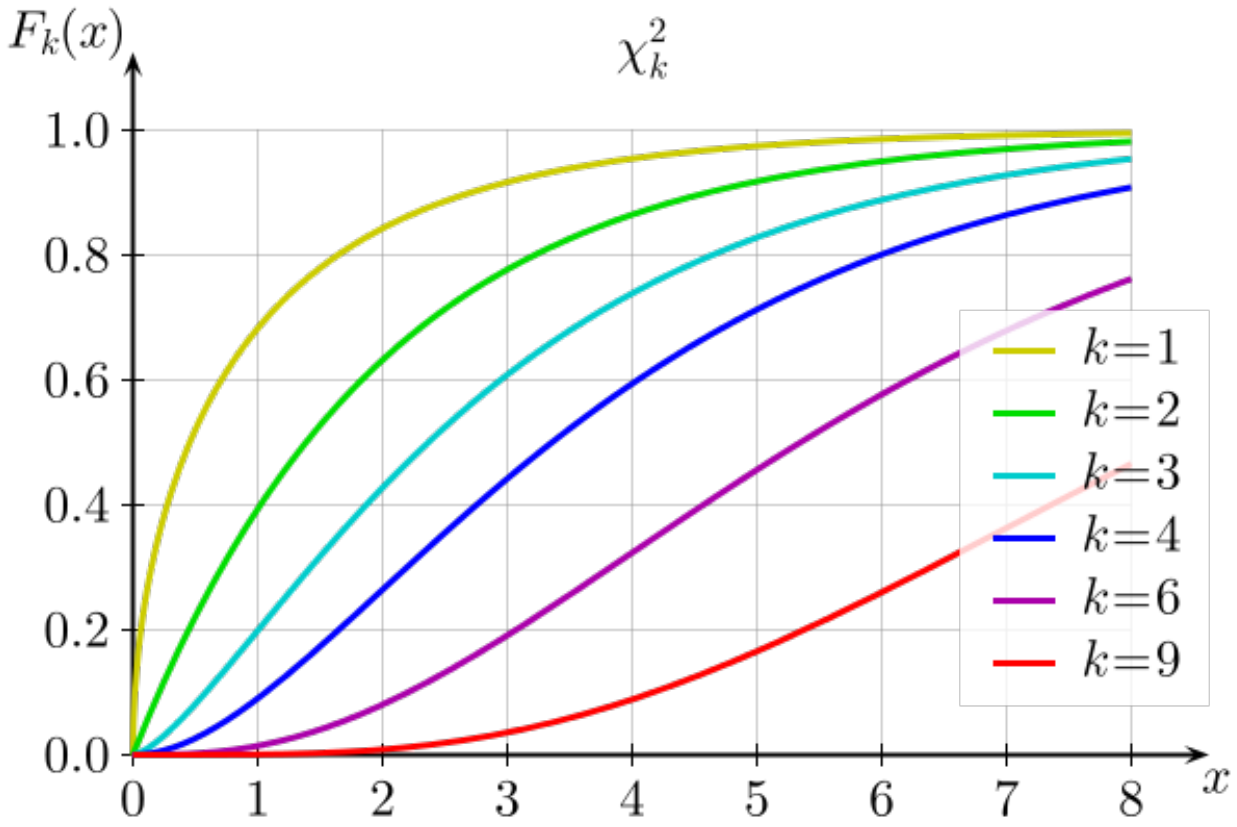
Probability density function



Notation	$\chi^2(k)$ or χ^2_k
Parameters	$k \in \mathbb{N}_{>0}$ (degrees of freedom)
Support	$x \in (0, +\infty)$ if $k = 1$, otherwise $x \in [0, +\infty)$
PDF	$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$
CDF	$\frac{1}{\Gamma(k/2)} \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$
Mean	k
Median	$\approx k\left(1 - \frac{2}{9k}\right)^3$
Mode	$\max(k-2, 0)$
Variance	$2k$

Chi-squared distribution

Cumulative distribution function



Notation	$\chi^2(k)$ or χ^2_k
Parameters	$k \in \mathbb{N}_{>0}$ (degrees of freedom)
Support	$x \in (0, +\infty)$ if $k = 1$, otherwise $x \in [0, +\infty)$
PDF	$\frac{1}{2^{k/2}\Gamma(k/2)}\, x^{k/2-1} e^{-x/2}$
CDF	$\frac{1}{\Gamma(k/2)}\, \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$
Mean	k
Median	$\approx k\left(1 - \frac{2}{9k}\right)^3$
Mode	$\max(k-2, 0)$
Variance	$2k$

Chi-squared distribution in MATLAB

- `X = chi2inv(P,V)` computes the inverse of the chi-square cdf with degrees of freedom specified by `V` for the corresponding probabilities in `P`.
- `P` and `V` can be vectors, matrices, or multidimensional arrays that have the same size. A scalar input is expanded to a constant array with the same dimensions as the other inputs. The degrees of freedom parameters in `V` must be positive, and the values in `P` must lie in the interval `[0 1]`.

Chi-squared distribution in MATLAB

- The inverse chi-square cdf for a given probability p and ν degrees of freedom is:

$$x = F^{-1}(p \mid \nu) = \{x: F(x \mid \nu) = p\}$$

- where

$$p = F(x \mid \nu) = \int_0^x \frac{t^{\frac{\nu-2}{2}} e^{-\frac{t}{2}}}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})} dt$$

- and $\Gamma(\cdot)$ is the Gamma function. Each element of output `X` is the value whose cumulative probability under the chi-square cdf defined by the corresponding degrees of freedom parameter in `V` is specified by the corresponding value in `P`.

Chi-squared test

- The chi-squared test is used when we want to see if two categorical variables are related.
- The test statistic for the Chi-squared test uses the sum of the squared differences between each pair of observed (O) and expected values (E).

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

```
function c = compute_chi(obs,expt)
        c = sum((obs - expt).^2 ./ expt);
end
```

Chi-squared test

- **Null Hypothesis**
- $H_0: O = E$ (There is no statistically significant difference between observed and expected frequencies.)

- **Alternative Hypothesis**
- $H_1: O \neq E$ (There is a statistically significant difference between observed and expected frequencies.)

Chi-squared test in MATLAB

```
c = compute_chi(obs,expt);
p = chi_table(.05,df);
if c > p
    disp('H0 is false')

function C = chi_table(p,df)
    C = chi2inv(1-p,df);
end
```

z-test

- The z-test is a parametric hypothesis test used to determine whether a sample data set comes from a population with a particular mean. The test assumes that the sample data comes from a population with a normal distribution and a known standard deviation.
- The test statistic is

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

- Where $\bar{x}$ is the sample mean, μ is the population mean, σ is the population standard deviation, and n is the sample size.
- Under the null hypothesis, the test statistic has a standard normal distribution.

z-test in MATLAB

- `h = ztest(x,m,sigma)` returns a test decision for the null hypothesis that the data in the vector `x` comes from a normal distribution with mean m and a standard deviation `sigma`, using the z-test. The alternative hypothesis is that the mean is not `m`. The result `h` is 1 if the test rejects the null hypothesis at the 5% significance level, and 0 otherwise.

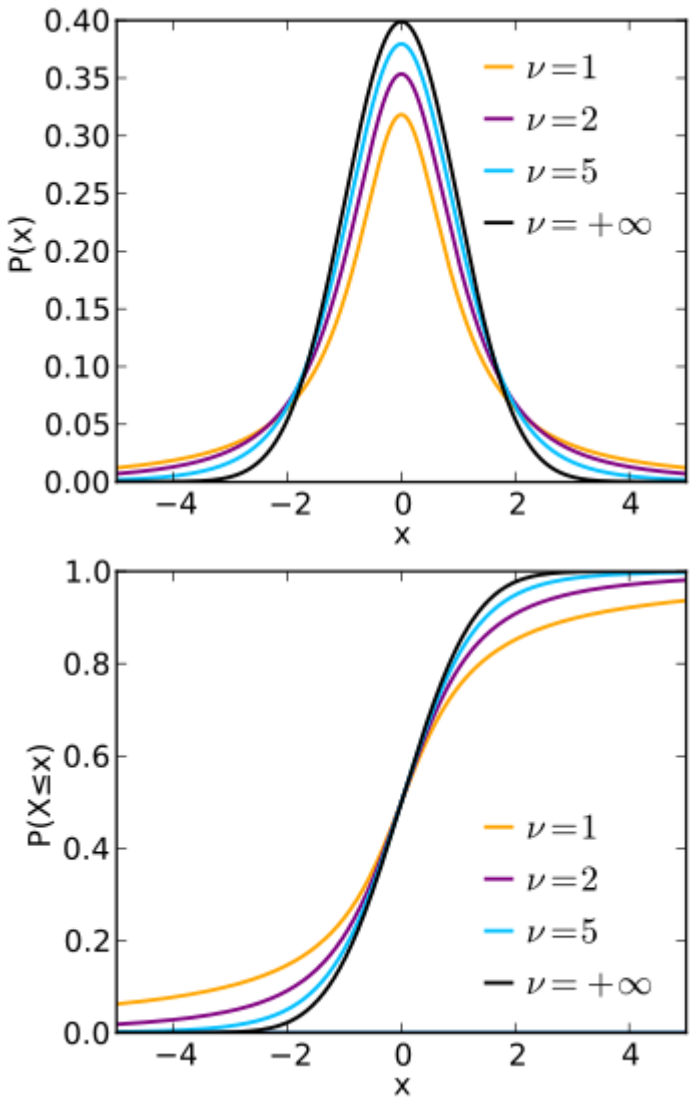
Student's t distribution

- If we take a sample of n observations from a normal distribution, then the t-distribution with $\nu = n - 1$ degrees of freedom can be defined as the distribution of the location of the sample mean relative to the true mean, divided by the sample standard deviation, after multiplying by the standardizing term $\sqrt{n}$.
- In this way, the t-distribution can be used to construct a confidence interval for the true mean.
- The t-distribution is symmetric and bell-shaped, like the normal distribution, but has heavier tails, meaning that it is more prone to producing values that fall far from its mean.
- This makes it useful for understanding the statistical behavior of certain types of ratios of random quantities, in which variation in the denominator is amplified and may produce outlying values when the denominator of the ratio falls close to zero.

Student's t distribution

PDF

CDF



As $\nu \to +\infty$, the *Student's t* distribution tend to a *normal* distribution

Parameters	$\nu > 0$ degrees of freedom (real)
Support	$x \in (-\infty, \infty)$
PDF	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
CDF	$\frac{1}{2}+x\Gamma\left(\frac{\nu+1}{2}\right)\times \frac{{}_2F_1\left(\frac{1}{2},\frac{\nu+1}{2};\frac{3}{2};-\frac{x^2}{\nu}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}$ where ${}_2F_1$ is the hypergeometric function
Mean	0 for $\nu > 1$, otherwise undefined
Median	0
Mode	0
Variance	$\frac{\nu}{\nu-2}$ for $\nu > 2$, ∞ for $1 < \nu \leq 2$, otherwise undefined

One-sample t-test

- The one-sample t-test is a parametric test of the location parameter when the population standard deviation is unknown.
- The test statistic is

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

- Where $\bar{x}$ is the sample mean, μ is hypothesized the population mean, s is the sample standard deviation, and n is the sample size.
- Under the null hypothesis, the test statistic has Student's t distribution with $n-1$ degrees of freedom.

One-sample t-test in MATLAB

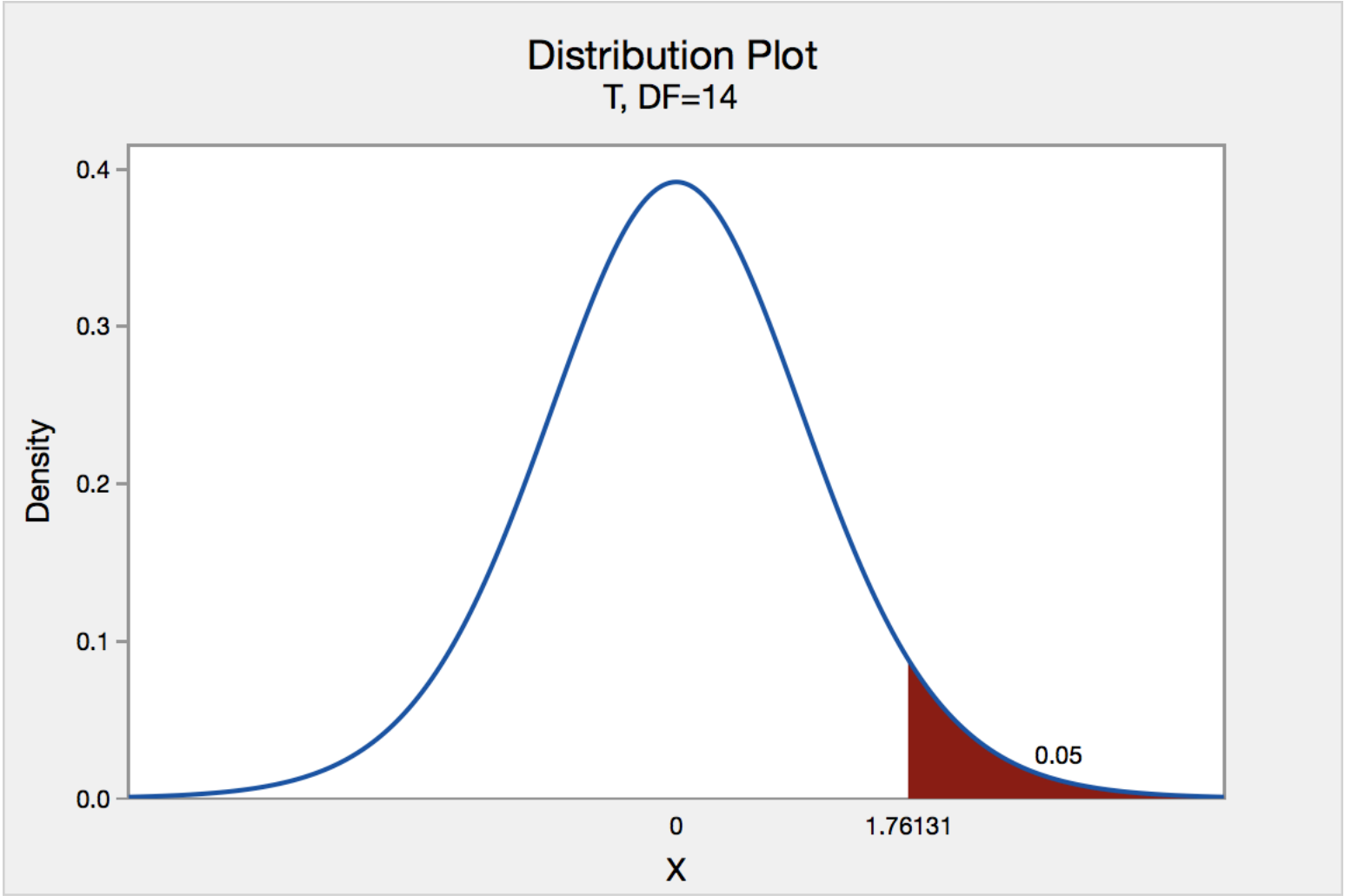
- `h = ttest(x)` returns a test decision for the null hypothesis that the data in `x` comes from a normal distribution with mean equal to zero and unknown variance, using the one-sample t-test. The alternative hypothesis is that the population distribution does not have a mean equal to zero. The result `h` is 1 if the test rejects the null hypothesis at the 5% significance level, and 0 otherwise.
- `h = ttest(x,m)` returns a test decision for the null hypothesis that the data in `x` comes from a normal distribution with mean `m` and unknown variance. The alternative hypothesis is that the mean is not `m`.

One-sample t-test in MATLAB

- 'Tail' — Type of alternative hypothesis
- Type of alternative hypothesis to evaluate, specified as the comma-separated pair consisting of 'Tail' and one of the following.
  - 'both' Test the alternative hypothesis that the population mean is not m.
  - 'right' Test the alternative hypothesis that the population mean is greater than m.
  - 'left' Test the alternative hypothesis that the population mean is less than m.

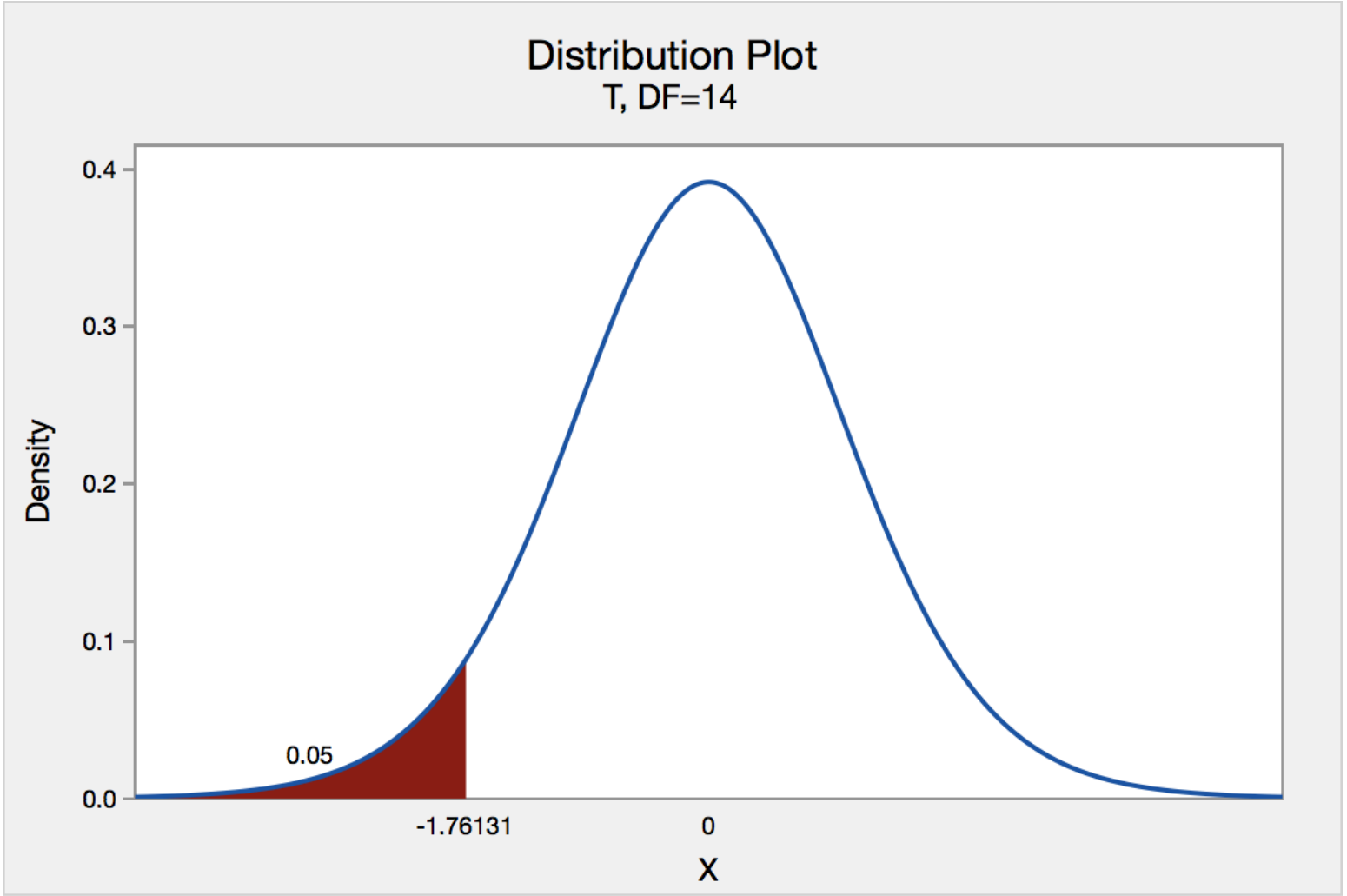
One-tailed vs two-tailed t-test

Right-Tailed t-test



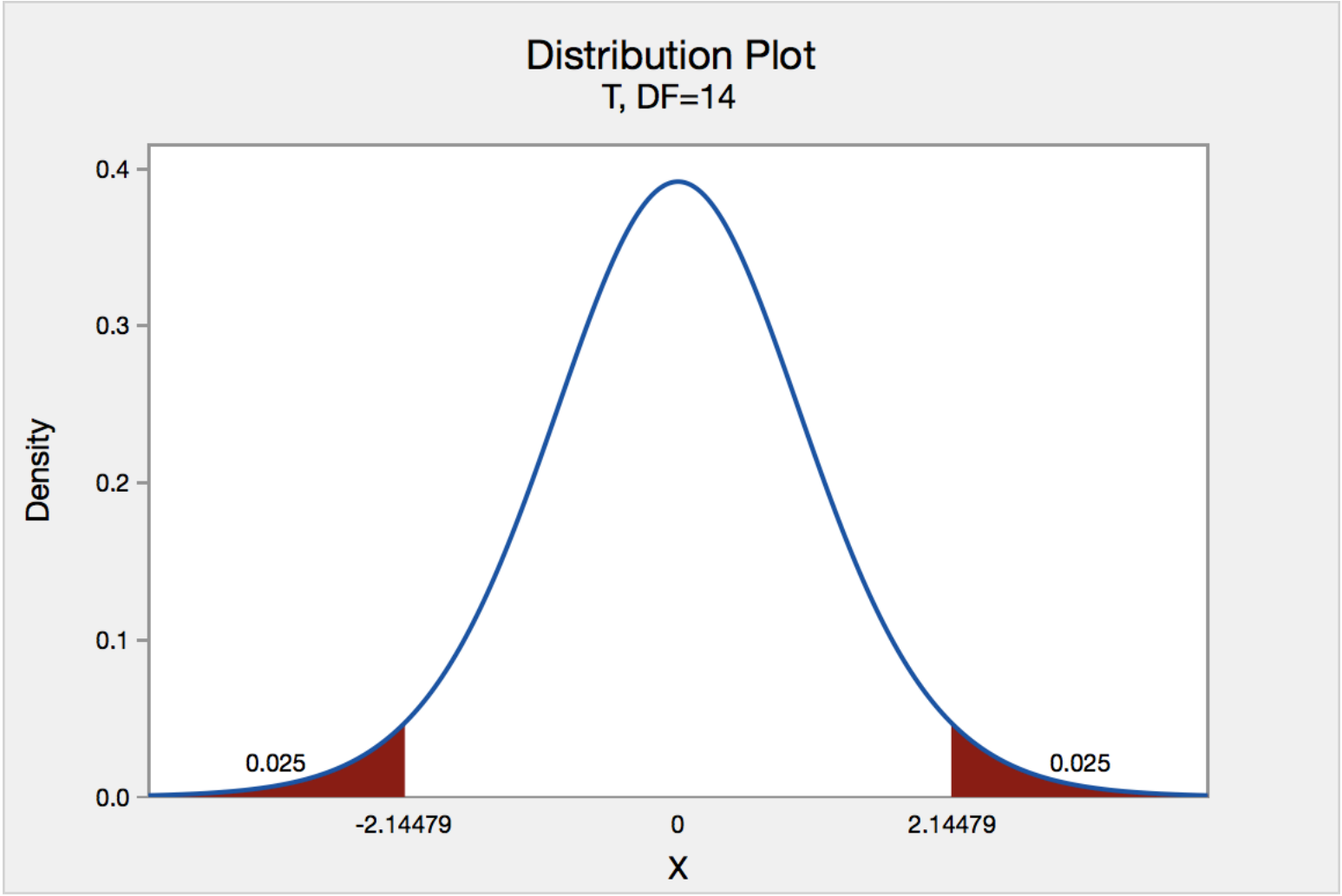
One-tailed vs two-tailed t-test

Left-Tailed t-test



One-tailed vs two-tailed t-test

Two-Tailed t-test



Two-Sample t-test

- The two-sample t-test is a parametric test that compares the location parameter of two independent data samples.
- The test statistic is

$$t = \frac{x - y}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

- Where $\bar{x}$ and $\bar{y}$ are the sample means, s_x and s_y are the sample standard deviations and n and m are the sample sizes.

Two-Sample t-test

- In the case where it is assumed that the two data samples are from populations with equal variances, the test statistic under the null hypothesis has Student's t distribution with $n + m - 2$ degrees of freedom, and the sample standard deviations are replaced by the pooled standard deviation

$$s = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$$

- In the case where it is not assumed that the two data samples are from populations with equal variances, the test statistic under the null hypothesis has an approximate Student's t distribution with a number of degrees of freedom given by Satterthwaite's approximation.
- This test is sometimes called Welch's t-test.

References

- Sam Kash Kachigan. *Multivariate Statistical Analysis: A Conceptual Introduction*. Radius Press, 1991.
- MATLAB Documentation
  - Z-test.
    URL: https://it.mathworks.com/help/stats/ztest.html
  - T-test.
    URL: https://it.mathworks.com/help/stats/ttest.html
  - Two-sample t-test.
    URL: https://it.mathworks.com/help/stats/ttest2.html
  - Chi square inverse cumulative distribution function.
    URL: https://it.mathworks.com/help/stats/chi2inv.html
- Wikipedia
  - Chi-squared distribution.
    URL: https://en.wikipedia.org/wiki/Chi-squared_distribution
  - Chi-squared test.
    URL: https://en.wikipedia.org/wiki/Chi-squared_test