Hypothesis Testing

Hypothesis testing slides are mainly based on chapter 8 of the book "Essentials of Social Statistics for a Diverse Society" Second Edition by Anna Leon-Guerrero, Chava Frankfort-Nachmias , SAGE Publications, Inc, 2010.

See also Chapter 4 of Cohen's book.

Hypothesis testing scenario 1

Assume you are the database administrator of a big information system and you are unhappy with the execution time of a given SQL package.

From historical data (thousands of previous package executions), you know that the average execution time of the package is **83.54** seconds with a standard deviation of **16.36**.

You change the tuning of the database and run the package several times to check the effect.

Questions:

- **Has the new tuning any effect?**
- **Is the new configuration better?**
- **Is the new configuration worse?**

Package exec. time
74
66
88
68
7[illegible]
⋮
79
78
72
86
85
86

Avg = 78.15

Hypothesis testing scenario 2

Assume you are the database administrator of a big information system. The database has just been installed and you are trying two tuning configurations: Conf. **A** and Conf. **B**.

You use a given SQL package to test the execution time for each configuration.

After running several times the SQL package in both configurations you want to take a decision.

Question: **what is the best configuration?**

Conf. A exec. time	Conf. B exec. time
74	69
66	71
88	80
68	88
79	64
68	65
87	74
79	76
78	89
72	68
86	67
85	72
86	

Avg A = 78.15
n = 13

Avg B = 73.58
n = 12

Hypothesis

- **What is an hypothesis?**
  - A proposed explanation for a given phenomenon
  - An assumption about the efficiency of a given component/system
  - A statement about the parameters of a population (**statistical view**)

An hypothesis is a tentative answer!

- Types
  - **Explanatory**: explains the phenomenon, identifies relations and/or causality between variable/elements of the phenomenon
  - **Predictive**: predicts the observation of a phenomenon, anticipates the outcome of an experiment,...

Hypothesis

- **What is an hypothesis?**
  - A proposed explanation for a given phenomenon
  - An assumpti
  - A statement

- Types
  - **Explanatory**
    causality bet
  - **Predictive**: p
    the outcome

- An **hypothesis requires evaluation** to be considered true. It can be **confirmed** or **refuted**.
- True hypothesis means the probability of it being correct is 'high' and the probability of it being incorrect is 'low'.
- Statistics is necessary to quantify the meaning of "high" and "low" and to decide about the validity of the hypothesis.
- Hypotheses are accepted or rejected with some **degree of certainty**

Questions (don't be afraid asking questions)

Research-study questions

- **Exploratory**
  Understand a phenomenon (subject of study) and clarify its features
- **Base-rate**
  Characterize the occurrence patterns of the phenomenon
- **Relational**
  Identify possible relations of the phenomenon under study with other phenomenon
- **Causal**
  Identify cause and effect related to the phenomenon under study

Engineering questions

- **Design and architecture**
  Define the best engineering processes and the best architecture for products
- **Measure and optimization**
  Measure and evaluate figures of merit correctly and use the measurements to optimize products and processes
- **Benchmark and choose**
  Measure to compare and choose among alternatives (components, systems, processes)
- **Verification and validation**
  Confirms that a given implementations works as specified (verification) and solves the intended problem as expected (validation)

Questions (don't be afraid asking questions)

Research-study questions

- **Exploratory**
  Understand a phenomenon (subject of study) and clarify its features
- **Base-rate**
  Characterize t[…] phenomenon
- **Relational**
  Identify possible re[…] phenomenon under[…] phenomenon
- **Causal**
  Identify cause and […] phenomenon under[…]

Engineering questions

- **Design and architecture**
  Define the best engineering processes and the best architecture for products

Examples of questions:

Existing and searching

Does X exist? → Does global warming exist?

Describing and classifying

How is X composed? What are the different types of X? What is X for? What are the properties of X? …

Describing and comparing

What is the difference between X and Y? …

Questions (don’t be afraid asking questions)

Research-study ques

- **Exploratory**
  Understand a phenomeno
  study) and clarify its f
- **Base-rate**
  Characterize the occurrence pat
  phenomenon
- **Relational**
  Identify possible relations of th
  phenomenon under study with
  phenomenon
- **Causal**
  Identify cause and effect relate
  phenomenon under study

intended problem as expected (validation)

Examples of questions:

Frequency and distribution

How frequent does X happen? → How many bugs per thousand lines of code?

Is X more frequent in a given period? What is the average occurrence of X? What is the amount of X occurrences per unit of time?

Process and functioning

How does X work? What is the normal sequence of events of X? How does X produce its outputs? …

Questions (don't be afraid asking questions)

Research-study questions

- **Exploratory**
  Understand a phenomenon (su
  study) and clarify its features
- **Base-rate**
  Characterize the occurrence
  phenomenon
- **Relational**
  Identify possible relations of the phenomenon under study with other phenomenon
- **Causal**
  Identify cause and effect related to the phenomenon under study

Examples of questions:

Are X and Y related? → Is the SW development method related to the number of bugs?

Are the occurrences of X correlated with the occurrences of Y?

Engineering questions

Measure to compare and choose among alternatives (components, systems, processes)

- **Verification and validation**
  Confirms that a given implementations works as specified (verification) and solves the intended problem as expected (validation)

Questions (don't be afraid asking questions)

Research-study questions

- **Exploratory**
  Understand a phenomeno
  study) and clarify its feat
- **Base-rate**
  Characterize the occurren
  phenomenon
- **Relational**
  Identify possible relation
  phenomenon under stud
  phenomenon
- **Causal**
  Identify cause and effe
  phenomenon under study

Engineering questions

Examples of questions:

Causality

Is X the cause of Y? → Is a memory leak the cause of the operating system crash?

What is the effect of X over Y? What is the cause of Y? Does X preclude Y?

Causality-comparison

Is X more relevant as the cause of Y than Z?

Causality-comparison-interaction

Is X or Z a more relevant cause of Y in a given situation than in another (situation)?

Questions (don’t be afraid asking questions)

Research-study questions

- **Exploratory**
  Understand a phenomenon (subject of study) and clarify its features
- **Base-rate**
  Characterize the occurrence patterns of the phenomenon
- **Relational**
  Identify possible relations of the phenomenon under study with other phenomenon
- **Causal**
  Identify cause and effect related to the phenomenon under study

Engineering questions

- **Design and architecture**
  Define the best engineering processes and the best architecture for products
- **Measure and optimization**
  Measure and evaluate figures of merit correctly and use the measurements to optimize products and processes
- **Benchmark and choose**
  Measure to compare and choose among alternatives (components, systems, processes)
- **Verification and validation**
  Confirms that a given implementations works as specified (verification) and solves the intended problem as expected (validation)

Questions (don't be afraid asking questions)

Research-study questions

- **Exploratory**
  Understand a phenomenon (subject of study) and clarify its features
- **Base-rate**
  Characterize the occurrence patterns of th
  phenomenon

Engineering questions

- **Design and architecture**
  Define the best engineering processes and the best a tecture for products
- **asure and optimization**
  asure and evaluate figures of merit orrectly and use the measurements to oducts and processes
- **ark and choose**
  compare and choose among (components, systems, processes)
- **tion and validation**
  hat a given implementations works d (verification) and solves the intended problem as expected (validation)

Examples of questions:

What is the best architecture for X? → What is the best storage subsystem for my database?

What is the most effective process to build X?

Questions (don’t be afraid asking questions)

Research-study questions

- **Exploratory**
  Understand a phenomenon (subject of study) and clarify its features
- **Base-rate**
  Characterize the occurrence patterns of the phenomenon

Engineering questions

- **Design and architecture**
  Define the best engineering processes and the best architecture for products
- **Measure and optimization**
  Mea... and evaluate figures of merit ...use the measurements to ... and processes

...among ...s, processes)

...tion
...ations works ...lves the ...alidation)

Examples of questions:

How to measure figures of merit of X? → How to measure security in my web system?

What is measurement of a figure of merit F of X?

Does the new configuration of X represents an improvement?

Questions (don't be afraid asking questions)

Examples of questions:

Is X better than Y? → Is Oracle 12c faster than IBM DB2?

What is the ranking for X, Y and Z concerning the feature F?

phenomenon

- **Relational**
  Identify possible relations of the phenomenon under study with other phenomenon
- **Causal**
  Identify cause and effect related to the phenomenon under study

neering questions

- **nd architecture**
  best engineering processes and the cture for products
- **and optimization**
  d evaluate figures of merit correctly and use the measurements to optimize products and processes
- **Benchmark and choose**
  Measure to compare and choose among alternatives (components, systems, processes)
- **Verification and validation**
  Confirms that a given implementations works as specified (verification) and solves the intended problem as expected (validation)

Questions (don’t be afraid asking questions)

Research-study questions

- **Exploratory**

- **Relational**
  Identify possible relations of the phenomenon under study with other phenomenon

- **Causal**
  Identify cause and effect related to the phenomenon under study

Engineering questions

- **Design and architecture**
  gineering processes and the
  or products

- **optimization**
  uate figures of merit
  the measurements to
  and processes

- **hmark and choose**
  compare and choose among
  alternat (components, systems, processes)

- **Verification and validation**
  Confirms that a given implementations works as specified (verification) and solves the intended problem as expected (validation)

Examples of questions:

Is X implemented correctly? → Is service A working according to the specification?

Is X really solving the problem it was designed for?

Back to example
Hypothesis testing scenario

Assume you are the database administrator of a big information system and you are unhappy with the execution time of a given SQL package.

From historical data (thousands of previous package executions), you know that the average execution time of the package is **83.54** seconds with a standard deviation of **16.36**.

You change the tuning of the database and run the package several times to check the effect.

Questions:

- **Has the new tuning any effect?**
- **Is the new configuration better?**
- **Is the new configuration worse?**

Package exec. time
74
66
88
68
79
⋮
79
78
72
86
85
86

32 times

Avg = 78.15

Hypothesis testing steps

1. State the hypothesis or claim to be tested
2. Select the criteria for a decision
3. Compute the test statistic
4. Make a decision

Step 1 - State the hypothesis

- **Null hypothesis** (H_0) is a statement about the population parameter (e.g., the population mean) that is assumed to be true.

  This is a provisional answer to the research question or problem under study. For example: **“H_0 - The new configuration has no effect on the execution time of the SQL packaged”**

- **Alternative hypothesis** (H_1) is a statement that directly contradicts the null hypothesis by stating that the actual value of the population is less than, greater than, or not equal to the value stated in the null hypothesis.

  This is what we think is wrong about the null hypothesis. For example: **“H_1 – The execution time of the SQL packaged is different in the new configuration (could be smaller or bigger)”**

Step 1 - State the hypothesis

- **Null hypothesis** (H_0) is a statement about the population paramet

  This

  stud

  exec

- **Alterna**

  contradi

  the popu

  stated in

  This

  "H_1

  new

The decision made in hypothesis testing centers on the null hypothesis H_0

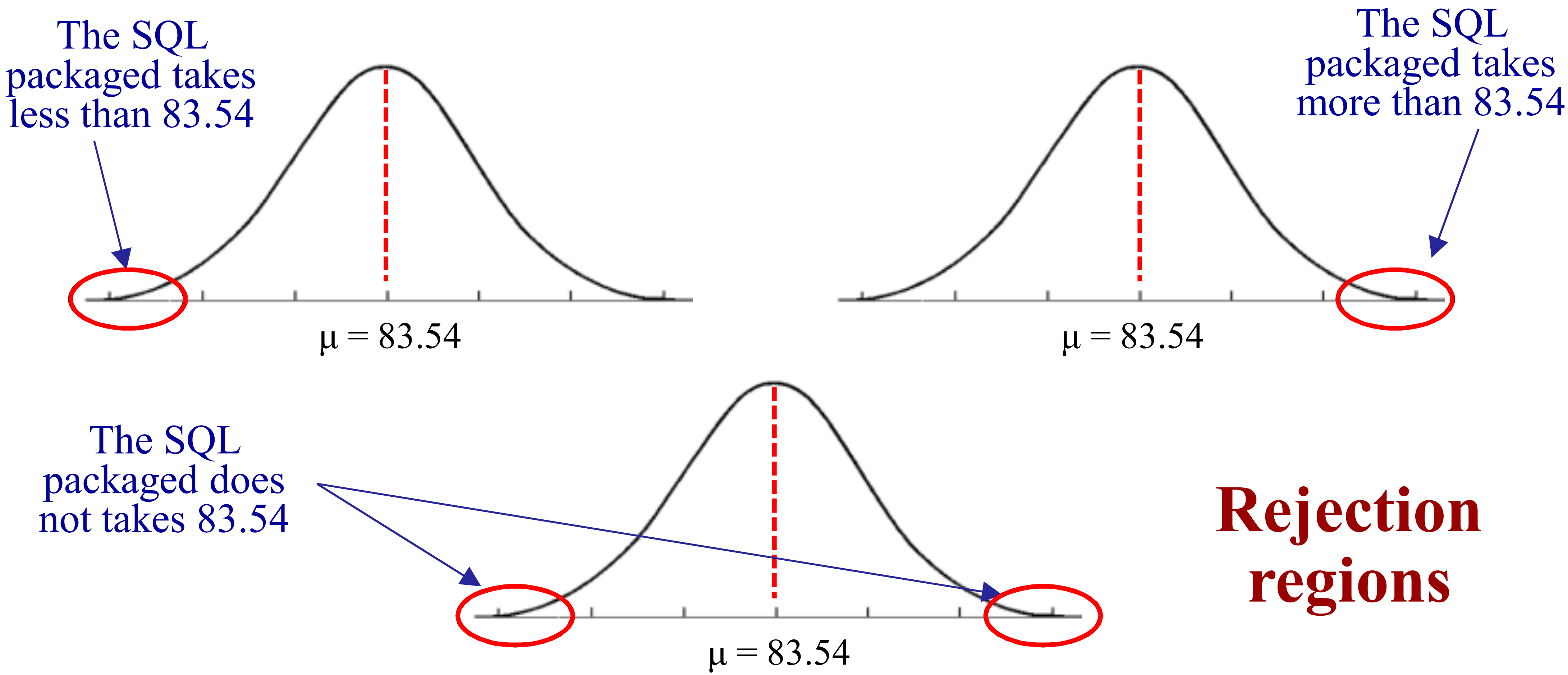
- The idea is to show evidences that H_0 is unlikely, in order to reject the null hypothesis. If failing to do so, the null hypothesis is retained.
- The bias is do nothing. In other words, the burden is put on the researcher to demonstrate that H_0 is not likely to be true. → **The experiments must be defined to collect data to show that H_0 is not true**

Step 2 - Select the criteria for a decision

- To set a criteria means **to state the significance level for the test.**
- **Significance level** refers to a criterion of judgment upon which a decision is made regarding the value stated in a null hypothesis.
- A typical significance level is 5%. This means that when the probability of obtaining a given sample mean is less than 5% <u>if the null hypothesis were true</u>, then we conclude that the sample used to calculate the mean is too unlikely, and so we **reject the null hypothesis**.

Step 2 - Select the criteria for a decision

The alternate hypothesis H_1 establish where to place the level of confidence. For example:



Step 3 – Compute the test statistics

- Select a random sample from the population and measure the sample mean. For example: execute the SQL package n times and measure a mean = 78.15
- To make a decision we need to evaluate how likely this sample outcome is, if the population mean stated by the null hypothesis (83.54) is true.
- **Test statistic** is a formula to determine the likelihood of obtaining sample outcomes if the null hypothesis is true. The value of the test statistic is used to make a decision regarding the null hypothesis.

Step 3 – Compute the test statistics

Test statistic:

Mean of the sample

Mean of the population

$$Z_c = \frac{M - \mu}{\sigma / \sqrt{n}}$$

Standard deviation of the population

Number of elements in the sample

Standard error

Measures how far the sample mean is from the population mean under H_0. The larger the value of $|Z_c|$ the more it will indicate that H_0 is not true.

Step 4 – Make a decision

- The value of the test statistic (Z_c) is the key to make a decision about the null hypothesis. The decision is based on the probability of obtaining a sample mean, given that the value stated in the null hypothesis is true.
- ***P* value** is the probability of obtaining a sample outcome or more extreme, given that the value stated in the null hypothesis is true.
- Example:
  - $P < 5\%$ → reject the null hypothesis (reach significance)
  - $P > 5\%$ → retain the null hypothesis (fail reaching significance)

Example: hypothesis testing using Z test

Assume you are the database administrator of a big information system and you are testing the time of a database system to answer a set of SQL queries.

According to the literature about this system, the average execution time of the package should be **83.54** seconds with a standard deviation of **16.36**.

You would like to check whether this is indeed the case and you ran the queries several times.

Questions:

- → Is the database system working according to the specification?

Package exec. time
74
66
88
68
70
⋮
79
78
72
86
85
86

32 times

Avg = 78.15

Example 1: non-directional (two-tailed)

Step 1- State the hypothesis

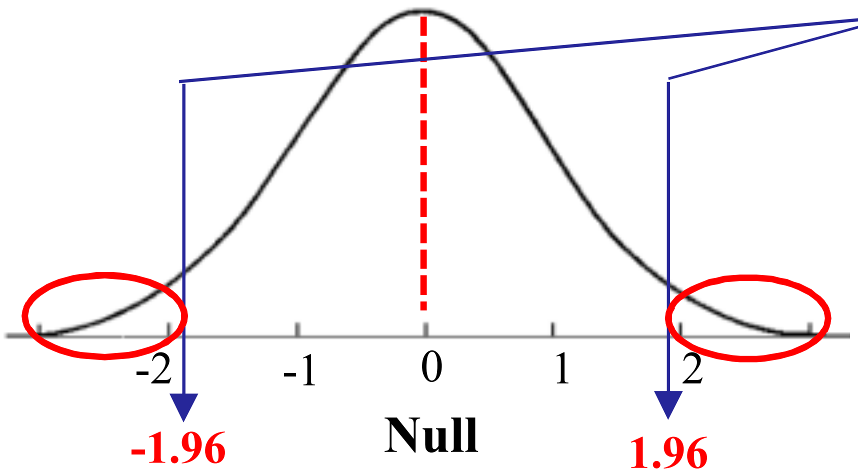
- **H_0** – The database system takes the same execution time as reported. → The average execution time is 83.54
- **H_1** – The execution time is different from that reported in the literature (could be smaller or bigger)

We are testing whether the null hypothesis H_0 is true

Example 1: non-directional (two-tailed)

Step 2 - Set the criteria for a decision

- Consider the level of significance of 5% → $\alpha = 0.05$
- Locate the Z score (in the table for the standard normal distribution) that represents the **critical values**
- A **critical value** is a cutoff value that sets the boundaries beyond which less than 5% of sample means can be obtained is the null hypothesis is true.



Critical values for nondirectional (two-tailed) test with α = 5%

→ Z score = 1.96

Rejection regions

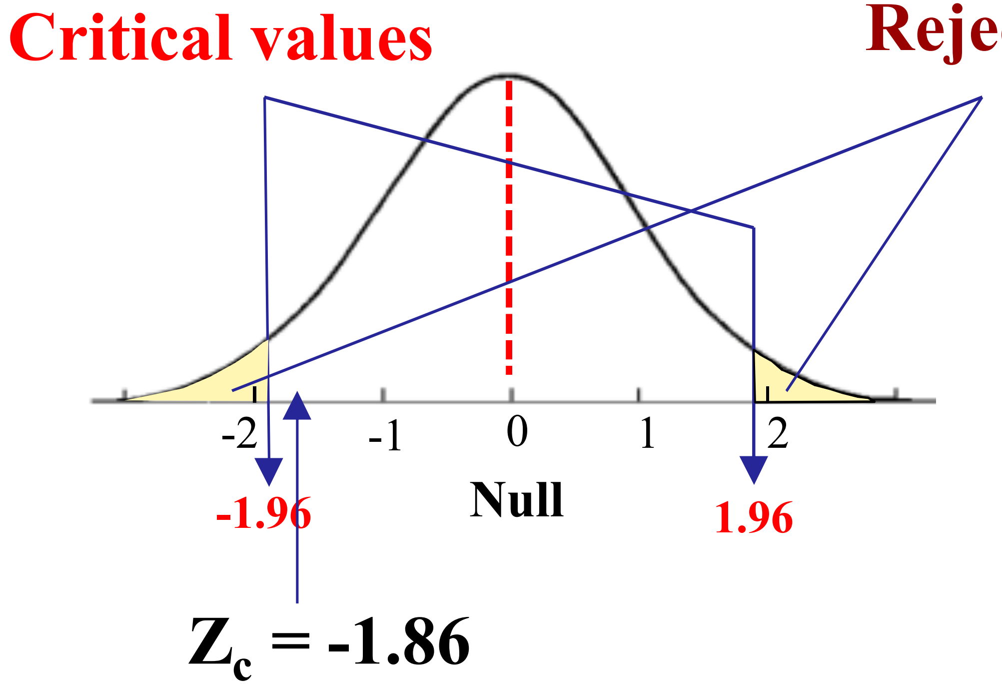
Example 1: non-directional (two-tailed)

Step 3 - Compute the test statistic

Test statistic:

$$Z_c = \frac{M - \mu}{\sigma / \sqrt{n}} = \frac{78.15 - 83.54}{16.36 / \sqrt{32}} = -1.86$$

Example 1: non-directional (two-tailed)
Step 4 - Make a decision



The probability of obtaining $\mathbf{Z_c = -1.86}$ or more extreme is given by the ***P* value**. To obtain ***P* value** for look for 1.86 in the standard normal table. → **the value is 0.0314**

As it is a two-tailed

P = 0.0314 x 2 = 0.0628 → **P = 6.28%**

As *P* > 5%

Retain the null hypothesis
(fail reach significance)

Example2: hypothesis testing using Z test
Directional (one-tailed)

Assume you are the database administrator of a big information system and you are unhappy with the execution time of a given SQL package.

From historical data (thousands of previous package executions), you know that the average execution time of the package is **83.54** seconds with a standard deviation of **16.36**.

You changed the tuning of the database and run the package several times to check the effect.

Question:

- **Is the new configuration better?**
  That is, is the execution time in the new configuration smaller than in the previous one?

Package exec. time
74
66
88
68
7[illegible]
⋮
79
78
72
86
85
86

32 times

Avg = 78.15

Example 2: directional (one-tailed)

Step 1- State the hypothesis

- **H_0** – The new configuration has no effect on the execution time of the SQL packaged. → The average execution time is 83.54
- **H_1** – The execution time of the SQL packaged is smaller in the new configuration

We are testing whether the null hypothesis H_0 is true

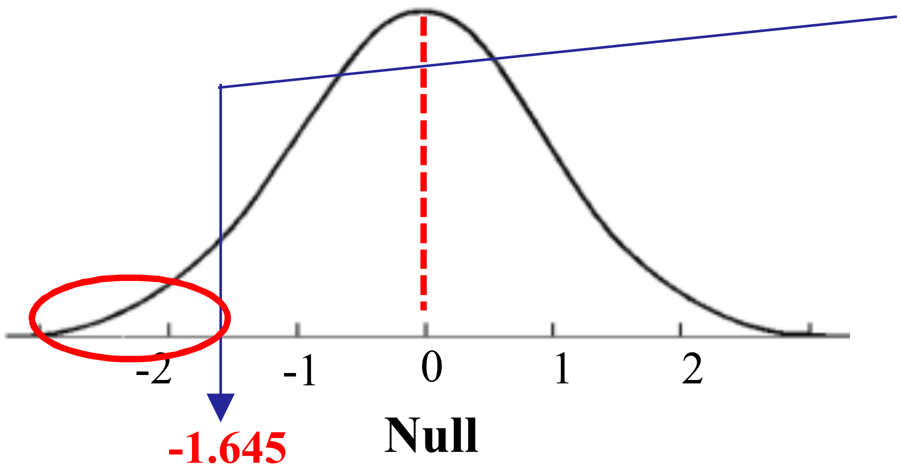
Note that only the alternate hypothesis changed.

Directional or one-tailed tests are hypothesis tests where the alternative hypothesis is stated as greater than (>) or less than (<) the value stated in the null hypothesis

Example 2: directional (one-tailed)

Step 2 - Set the criteria for a decision

- Consider the level of significance of 5% → $\alpha = 0.05$
- Locate the Z score (in the table for the standard normal distribution, one-tailed) that represents the **critical value**
- A **critical value** is a cutoff value that sets the boundaries beyond which less than 5% of sample means can be obtained if the null hypothesis is true.



Critical values for directional (left-tailed) test with $\alpha = 5\%$

→ Z score = -1.645

Example 2: directional (one-tailed)

Step 3 - Compute the test statistic

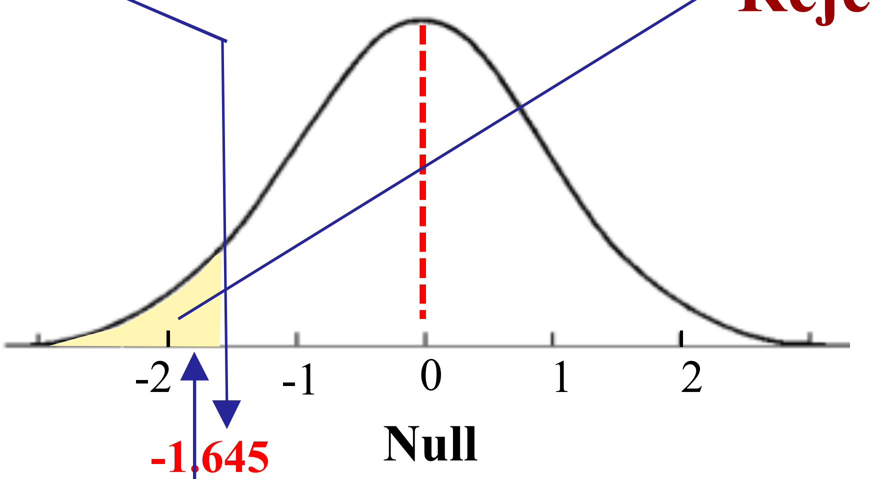
Test statistic:

$$Z_c = \frac{M - \mu}{\sigma/\sqrt{n}} = \frac{78.15 - 83.54}{16.36/\sqrt{32}} = -1.86$$

Example 2: directional (one-tailed) Step 4 - Make a decision

Critical values

Rejection region



Z_c = -1.86

The probability of obtaining **Z_c = -1.86** is given by the ***P* value**. To obtain ***P* value** for look for -1.86 in the standard normal table. → **the value is 0.0314**

P = 3.14% **(do not double in one-tailed)**

This means that the probability of getting at most an average of 78.15 if H_0 is true is 3.14%

As *P* < 5%

Reject the null hypothesis

The exec. time of the SQL packaged is smaller in the new configuration

The example again: **some questions**

Assume you are the database administrator of a big information system and you are unhappy with the execution time of a given SQL package.

From historical data (thousands of previous package executions), you know that the average execution time of the package is **83.54** seconds with a standard deviation of **16.36**.

You change the tuning of the database and run the package several times to check the effect.

Questions:

- **Has the new tuning any effect?**
- **Is the new configuration better?**
- **Is the new configuration worse?**

Package exec. time
74
66
88
68
79
⋮
79
78
72
86
85
86

32 times

Avg = 78.15

The example again: **some questions**

Assume you are the database administrator of a big ... wi... ge.

What should we do if we cannot have a relatively large number of samples?

From historical data (thousands of p... package executions), you know that the avera... execution time of the package is **83.54** seconds with a standard deviation of **16.36**.

You chan... ing of the database and run

What should we do if we don't know the standard deviation of the population?

- **Has the new tuning any effect?**
- **Is the new configuration better?**
-

In this cases we should use the t Test

Package exec. time
74
66
88
68
70
:
79
78
72
86
85
86

13 times

Avg = 78.15

Hypothesis testing using T-test (one sample)

- Follows the same steps as for the Z test
- The critical value comes from the **t table** (considering $n - 1$ degrees of freedom)
- The **test statistics** is now the t-test (similar formula)

Mean of the sample

Mean of the population (assumed)

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

Standard deviation of the sample

Number of elements of the sample

Example 3 - Hypothesis testing using T-test (one sample)

- A professor wants to know if their students are proficient in C programming. The other professors claim that their performance is not that good. Still, the professor wants the class to be able to score above 70 (0-100 scale) on the test (but doesn't want to examine all the students).
- The professor selects 6 students at random from the class and give them a C programming test.
- The six students get scores of 62, 92, 75, 68, 83, and 95.
- **Can the professor have 90 percent confidence that the mean score for the class on the test would be above 70?**

Example 3: **t test** (one sample)

Step 1- State the hypothesis

- **H_0: $\mu = 70$**

  In words: the class knows how to program in C with a proficiency equivalent to 70 in the C programming test

- **H_1: $\mu_1 > 70$**

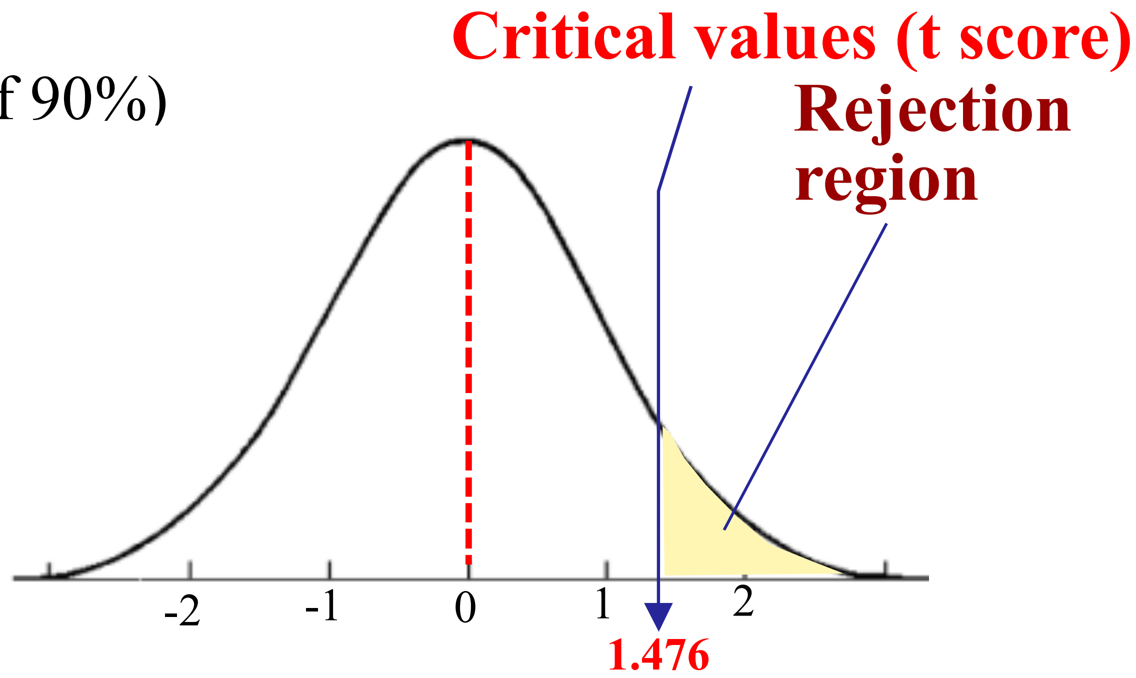
  The class is better on C programming than the score of 70

Example 3: **t test** (one sample)

Step 2 - Set the criteria for a decision

- Consider the level of confidence of 90% → $\alpha = 0.1$
- Locate the **t score** (in the t table for the Student distribution, one-tailed) that represents the **critical value** (for $\alpha = 0.1$ and $n = 6$)
- As the size of the sample is $n=6$, the degree of freedom = $n-1 = 5$
- Look in the t table for:
  - $\alpha = 0.1$ (α level for a conf. of 90%)
  - df = 5 (degree of freedom = 5)

→ t score = 1.476



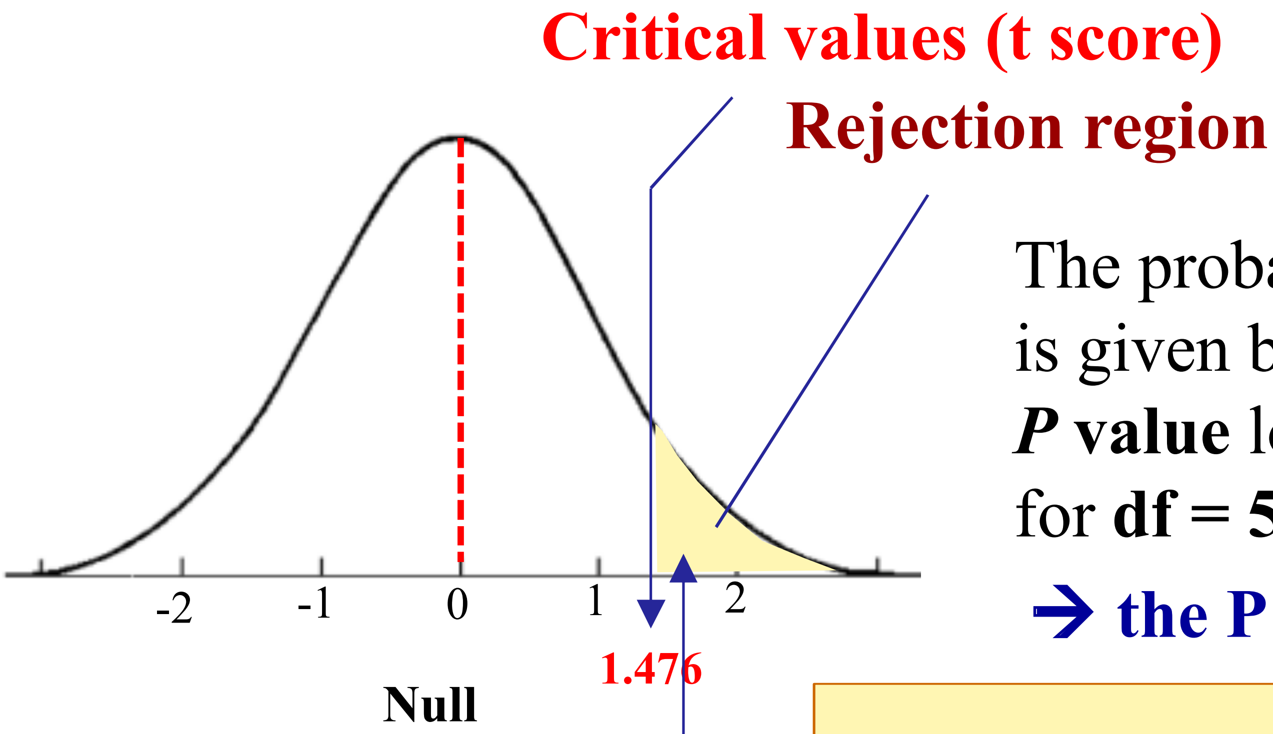
Example 3: **t test** (one sample)
Step 3 - Compute the test statistic

- Average of the sample: 79.17
- Standard deviation of the sample: 13.17

Test statistic:

$$t_c = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{79.17 - 70}{\frac{13.17}{\sqrt{6}}} = 1.71$$

Example 3: **t test** (one sample)
Step 4 - Make a decision



$t_c = 1.71$

The probability of obtaining $\mathbf{t_c = 1.71}$ is given by the ***P* value**. To obtain the ***P* value** look for 1.71 in the **t table**, for **df = 5**

→ **the P value is 0.074 (P = 7.4%)**

Means that the probability of getting an average score of 79.17 if H_0 is true is 7.4%

As *P* < 10% → **Reject the null hypothesis (reach significance)**