Salahaddin University-Erbil
Statistics Department
Second Stage
2023 - 2024

Bio Statistics Course

Chapter One

Hazhar T. A. Blbas

Statistics Department - MSc at UCF in 2014

Statistics Department – PhD Students at SUE

Founder and CEO of STAT Office for Statistical Data Analysis and Training

FB: SOSDAT

Definitions

Statistics

is the science of conducting studies to collect, organize, summarize, analyze, and draw conclusions from data.

Biostatistics

is the application of statistics to a wide range of topics in biology.

Applications of Biostatistics

- **Public health**, including epidemiology , health services research, nutrition and environmental health

- Design and analysis of clinical trials in medicine, genomics, population genetics and statistical genetics.

- Ecology, ecological forecasting

- Biological sequence analysis

Null and Alternative Hypotheses

Convert the research question to null and alternative hypotheses

- The **null hypothesis (*H_0*)** is a claim of "no difference in the population"

- The **alternative hypothesis (*H_a*)** claims "H_0 is false"

- The first step in the procedure is to state the hypotheses null and alternative forms. The null hypothesis (abbreviate "H ") is a statement of no difference. The alternative hypothesis ("H sub a") is a statement of difference.

- **The null hypothesis** is a statement that you want to test. In general, the null hypothesis is that things are the same as each other, or the same as a theoretical expectation. For example, if you measure the size of the feet of male and female chickens, the null hypothesis could be that the average foot size in male chickens is the same as the average foot size in female chickens.
- **The alternative hypothesis** is that things are different from each other, or different from a theoretical expectation. For example, one alternative hypothesis would be that male chickens have a <u>different</u> average foot size than female chickens.

Two types of Error

Type I Error

In a hypothesis test, a type I error occurs when the null hypothesis is rejected when it is in fact true (**We reject Ho while Ho is True**); that is, Ho is wrongly rejected.

For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; i.e.

$\boldsymbol{Ho}$**:** there is no difference between the two drugs on average.

Type II Error

In a hypothesis test, a type II error occurs when the null hypothesis Ho is not rejected when it is in fact false (**We accept Ho while Ho is False)**.

For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; i.e.

Ho: there is no difference between the two drugs on average.

Alpha (α) and Beta (β) Errors

Test Result	Truth	
	H_0 Correct	**H_0 wrong**
Accept H_0	OK $(1-\alpha)$	**Type II (β Error)**
Reject H_0	**Type I (α Error)**	OK $(1-\beta)$

α ≡ probability of a Type I error

β ≡ Probability of a Type II error

Result Possibilities

H_0: Innocent

Jury Trial			Hypothesis Test		
	Actual Situation			Actual Situation	
Verdict / Judgment	Innocent	Guilty	Decision	H_0 True	H_0 False
Innocent	Correct	Error	Accept H_0	$1 - \alpha$	Type II Error (β)
Guilty	Error	Correct	Reject H_0	Type I Error (α)	Power ($1 - \beta$)

False Positive

False Negative

Result Possibilities

H_0: Healthy Person (No Covid19)

Doctor Trial			Hypothesis Test		
	Actual Situation			Actual Situation	
Doctor	No Covid19	Covid19	Decision	H_0 True	H_0 False
No Covid19	Correct	Error	Accept H_0	$1 - \alpha$	Type II Error (β)
Covid19	Error	Correct	Reject H_0	Type I Error (α)	Power ($1 - \beta$)

False Positive

False Negative

Reference: Hazhar Blbas

Explanation of the Type I and Type II error

a) H0: The person is healthy **(No Covid19)**

H1: The person is unhealthy **(Covid19)**

b) A Type I error is a false positive. It has been decided that the person has **Corona Virus's** when she/he dose not have.

c) A Type II error is a false negative. It has been decided that the person is healthy, when they actually have **Corona Virus's** disease.

d) A Type I error would require more testing, resulting in time and money lost. A Type II error would mean that the person did not receive the treatment they needed. A Type II error is much worse.

Result Possibilities

H_0: No Covid19

Doctor Trial		
	Actual Situation	
Doctor	No Covid19	Covid19
st	Correct	Error
Covid19	Error	Correct

Hypothesis Test		
	Actual Situation	
Decision	H_0 True	H_0 False
Accept H_0	$1 - \alpha$	Type II Error (β)
Reject H_0	Type I Error (α)	Power ($1 - \beta$)

False Positive

False Negative

Reference: Hazhar Blbas

Example

a) H0: The person is healthy

H1: The person has Alzheimers

b) A Type I error is a false positive. It has been decided that the person has Alzheimer's disease when he/she doesnt.

c) A Type II error is a false negative. It has been decided that the person is healthy, when he/she actually has Alzheimer's disease.

d) A Type I error would require more testing, resulting in time and money lost. A Type II error would mean that the person did not receive the treatment they needed. A Type II error is much worse.

e) The power of this test is the ability of the test to detect patients with Alzheimer's disease. In this case, the power can be computed as 1− P(Type II error) = 1− 0.08 = 0.92.

Type I and Type II errors cannot happen at the same time

1. Type I error can only occur if H_0 is true
2. Type II error can only occur if H_0 is false
3. There is a tradeoff between type I and II errors. If the probability of type I error (α) increased, then the probability of type II error (β) declines.
4. When the difference between the hypothesized parameter and the actual true value is small, the probability of type two error (the non-rejection region) is larger.
5. Increasing the sample size, n, for a given level of α, reduces β

Significance Level

$\alpha \equiv$ probability of a Type I error

α = Pr(reject H_0 | H_0 true)

(the "|" is read as "given")

Although α is also called the "size of the critical region".

$\boldsymbol{\alpha = 0.05}$

$\boldsymbol{\alpha = 0.01}$

Power of Test

$\beta \equiv$ probability of a Type II error

β = Pr(accept H_0 | H_0 false)

(the "|" is read as "given")

The power of a statistical hypothesis test measures the test's ability to reject the null hypothesis ***Ho*** when it is actually false - that is, to make a correct decision.

$1 - \beta =$ "**Power**" $\equiv$ probability of avoiding a Type II error

$1 - \beta$ = Pr(accept H_0 | H_0 false)

Type I error vs Type II error

Basis for comparison	Type I error	Type II error
Definition	Type 1 error, in statistical hypothesis testing, is the error caused by rejecting a null hypothesis when it is true.	Type II error is the error that occurs when the null hypothesis is accepted when it is not true.
Also termed	Type I error is equivalent to false positive.	Type II error is equivalent to a false negative.
Meaning	It is a false rejection of a true hypothesis.	It is the false acceptance of an incorrect hypothesis.
Symbol	Type I error is denoted by α.	Type II error is denoted by β.
Probability	The probability of type I error is equal to the level of significance.	The probability of type II error is equal to one minus the power of the test.

Type I error vs Type II error

Basis for comparison	Type I error	Type II error
Reduced	It can be reduced by decreasing the level of significance.	It can be reduced by increasing the level of significance.
Cause	It is caused by luck or chance.	It is caused by a smaller sample size or a less powerful test.
What is it?	Type I error is similar to a false hit.	Type II error is similar to a miss.
Hypothesis	Type I error is associated with rejecting the null hypothesis.	Type II error is associated with rejecting the alternative hypothesis.
When does it happen?	It happens when the acceptance levels are set too lenient.	It happens when the acceptance levels are set too stringent.

Type of the T-test

- **One-sample t-test** compares one sample mean with a hypothesized value
- **Paired sample t-test** (dependent sample) compares the means of two dependent variables
- **Independent sample t-test** compares the means of two independent variables
  - Equal variance
  - Unequal variance

One Sample t-test

Assumptions

1- Scale Data

2- Randomization

3- Normality

Steps of One Sample T-Test

1.State the null hypothesis $H_0: \mu = \mu_0, \mu \geq \mu_0 \; or \; \mu \leq \mu_0$

the alternative hypothesis $H_1: \mu \neq \mu_0, \mu < \mu_0 \; or \; \mu > \mu_0$

2.Choose a significance level

$$\propto = \propto_0 \;\; (often\ we\ take\ \propto_0 = 0.05\ or\ 0.01)$$

3.Determine the critical region

4.Compute the $Z = \frac{(\bar{X} - \mu_0)}{\sigma/\sqrt{n}}$

5.Make a decision, reject the null hypothesis if the test statistic Z computed in step 4 falls in the rejection region for the test; otherwise, do not reject the null hypothesis.

One Sample t-test

❖ Hypothesis Testing:

- Unknown Parameters Requires t-test
- Comparison of One Sample Mean to a Specific Value

$$t = \frac{M - \mu_0}{s_M} = \frac{M - \mu_0}{s / \sqrt{n}}$$

- **Assumptions:** dependent variable is scale, Randomization, Normal Distribution

❖ We can calculate one sample t-test by hand and SPSS

Five steps to find one-sample t-test for a population mean **(Slide 1 of 5)**

- **Step 1**: State hypotheses

The null hypothesis is H_0: $\mu = \mu_0$ (the real mean equals some proposed theoretical constant μ_0);

The alternative hypothesis is one of the following:

H_a: $\mu \neq \mu_0$	H_a: $\mu < \mu_0$	H_a: $\mu > \mu_0$
(Two Tailed)	(Left Tailed)	(Right Tailed)

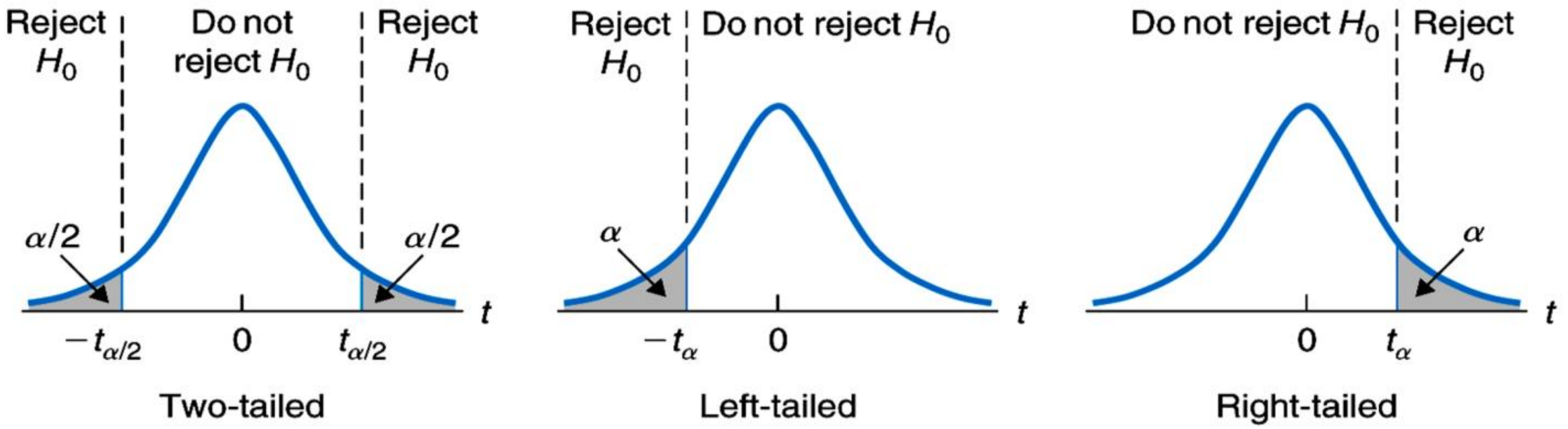
- **Step 2** Decide on the significance level, α

Five steps to find one-sample t-test for a population mean **(Slide 2 of 5)**

- **Step 3** The critical values are

$\pm t_{\alpha/2}$	$-t_{\alpha}$	$+t_{\alpha}$
(Two Tailed)	(Left Tailed)	(Right Tailed)

df = n - 1.



Five steps to find one-sample t-test for a population mean **(Slide 3 of 5)**

- Finding Critical Values

A portion of the t distribution table

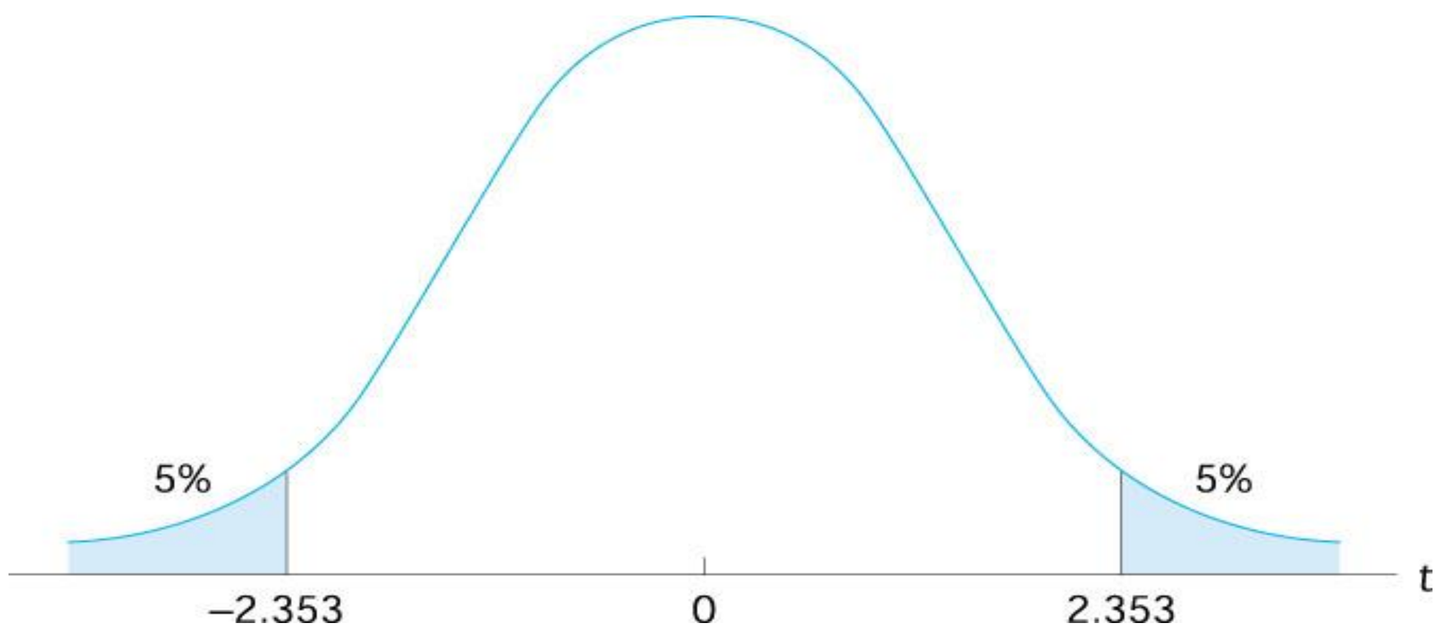
	Proportion in one tail					
	0.25	0.10	0.05	0.025	0.01	0.005
	Proportion in two tails combined					
df	0.50	0.20	0.10	0.05	0.02	0.01
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707

For example, alpha = .05 and number of sample size=6 for the two tails: df =n-1= 6-1 =5. Table says 2.571.

Five steps to find one-sample t-test for a population mean **(Slide 4 of 5)**

- Finding Critical Values

  The t-distribution for df = 3, 2-tailed α = 0.10



Five steps to find one-sample t-test for a population mean **(Slide 5 of 5)**

- **Step 4** Compute the value of the test statistic

$$t = \frac{M - \mu_0}{s_M} = \frac{M - \mu_0}{s / \sqrt{n}}$$

- **Step 5** If the value of the test statistic falls in the rejection region (if absolute value of sample is greater than critical value), reject H_0, otherwise do not reject H_0.

Summary of hypothesis-testing

- The null hypothesis H_0: $\mu = \mu_0$

Type	Conditions	Test Statistic
z-test	μ_0 is known σ is known	$z = \frac{M - \mu_0}{\sigma_M} = \frac{M - \mu_0}{\sigma / \sqrt{n}}$
t-test	μ_0 is hypothesized or predicted σ is unknown	$t = \frac{M - \mu_0}{s_M} = \frac{M - \mu_0}{s / \sqrt{n}}$

Example 1:

Testing whether light bulbs have a life of 1000 hours at $\alpha = 0.05$

800, 750, 940, 970, 790, 980, 820, 760, 1000, 860

Assumptions: dependent variable is scale, Randomization, Normal Distribution

- **Step 1** State hypotheses
  - Null hypothesis is **H_0: $\mu = 1000$**.
  - Alternative hypothesis is **H1: $\mu \neq 1000$**.

- **Step 2** Set alpha. **$\alpha = .05$**

- **Step 3** Determine the critical value. Looking for alpha = .05, two tails with **df = 10-1 = 9**. Critical value= **2.262**.

- **Step 4** Calculate the test statistic

What is the mean of our sample?

$$M = \frac{\sum xi}{n} = 867 \qquad ;(\bar{X} = M)$$

What is the standard deviation for our sample of light bulbs?

$$\text{Standard Deviation of Sample}\,(S) = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}} = 96.73$$

$$t = \frac{M - \mu_0}{\frac{S}{\sqrt{n}}} = \frac{867 - 1000}{30.59} = -4.35$$

- **Step 5** State decision rule, if absolute value of sample is greater than critical value, reject null.

We reject the null hypothesis (Test Statistics=|-4.35| > Critical value=|2.262|) that the bulbs were drawn from a population in which the average life is 1000 hrs.

The difference between our sample mean (867) and the mean of the population (1000) is SO different that it is unlikely that our sample could have been drawn from a population with an average life of 1000 hours

- **SPSS Steps**:-

Click Analyze, Compare Means, One-Sample T Test. Select light bulb (name of variables) and put it in the Test Variables box. Type 1000 in the Test Value box. Click OK. You get the output on this slide

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
BULBLIFE	10	867.0000	96.7299	30.5887

One-Sample Test

	Test Value = 1000					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
BULBLIFE	-4.348	9	.002	-133.0000	-202.1964	-63.8036

Because the p-value (Sig. (2-tailed)) is less than .05, we reject H0. So, it’s significant.

Example 2:

The mean emission of all engines of a new design needs to be **below** 20 ppm if the design is to meet new emission requirements. Ten engines are manufactured for testing purposes, and the emission level of each is determined.

15.6, 16.2, 22.5, 20.5, 16.4, 19.4, 16.6, 17.9, 12.7, 13.9

Does the data supply sufficient evidence to conclude that type of engine meets the new standard, with alpha=0.05?

Assumptions: dependent variable is scale, Randomization, Normal Distribution

Step 1 State hypotheses

H0: Emissions are equal to (or greater than) 20ppm;

H1: Emissions are less than 20ppm (One-Tailed Test)

Step 2 Set alpha. α = .05

Step 3 Determine the critical value. Looking for alpha = .05, one tails with df = 10-1 = 9. Critical value= -1.833.

Step 4 Calculate the test statistic

M =17.17 ; SD = 2.98 ; S_M =0.942 ; t statistic = -3.00

Step 5 Decision

State decision rule, we reject H0 because the absolute of test statistics is greater than critical value (t= |-3| > critical value=|-1.833|, it means the emissions are not equal to 20 ppm (emissions are less than 20 ppm).

Example 3:

An outbreak of Salmonella-related illness was attributed to ice cream produced at a certain factory. Scientists measured the level of Salmonella in 9 randomly sampled batches of ice cream. The levels (in MPN/g) were:

0.593 0.142 0.329 0.691 0.231 0.793
0.519 0.392 0.418

Is there evidence that the mean level of Salmonella in the ice cream is greater than 0.3 MPN/g, at Alpha = 0.05?

Step 1 State hypotheses

H0: $\mu = 0.3$

Ha: $\mu > 0.3$

Step 2 Set alpha. $\alpha = .05$

Step 3 Determine the critical value. Looking for alpha = .05, one tails with df = 9-1 = 8. Critical value= 1.860.

Step 4 Calculate the test statistic

M =0.456 ; SD =0.213 ; S_M =0.071 ; t-statistic =2.197

Step 5 Decision

Since, t-statistics is greater than critical value, reject Ho, there is moderately strong evidence that the mean Salmonella level in the ice cream is above 0.3 MPN/g.

Homework 1:

We want to test whether a new headache medicine provides a relief time equal to or different from the standard of 100 minutes.

90	93	93	99	98	100	103	104	99	102

Homework 2:

You are conducting an experiment to see if a given therapy works to reduce test anxiety. A standard measure of test anxiety is known to produce a $\mu = 20$. In the sample you draw of 81 the mean $M = 18$ with $s = 9$.

Homework 3:

Chapter 7 - Page 218 Fundamental Biostatistics-Rosner

Cardiology A topic of recent clinical interest is the possibility of using drugs to reduce infarct size in patients who have had a myocardial infarction within the past 24 hours. Suppose we know that in untreated patients the mean infarct size is 25 ($ck - g - EQ/m2$). Furthermore, in 8 patients treated with a drug the mean infarct size is 16 with a standard deviation of 10. Is the drug effective in reducing infarct size?

Homework 4:

Chapter 7 - Page 218 Fundamental Biostatistics-Rosner

- **Cardiovascular Disease, Pediatrics** Suppose the mean cholesterol level of 10 children whose fathers died from heart disease is 200 mg/dL and the sample standard deviation is 50 mg/dL

Test the hypothesis that the mean cholesterol level is higher 175 in this group than in the general population.