GREATER JAKARTA • BANDUNG • MALANG

People
Innovation
Excellence

STAT6171001
Basic Statistics

Logic of Hypothesis Testing
Session 9

Raymond Bahana
rbahana@binus.edu

Session Learning Outcomes

Upon completion of this session, students are expected to be able to

- LO 2. Analyze a problem by using the basic concept of descriptive and inferential statistics
- LO 3. Design a descriptive and inferential statistics solution to meet a given set of computing requirements in the context of computer science
- LO4. Produce descriptive and inferential statistics solutions

Topic

- Logic of Hypothesis Testing

Introduction

James Bond case study

- Mr. Bond was given 16 trials: martini shaken or stirred
- Correct = 13 of the trials

Binomial distribution

- $N = 16$, $x = 13$, $\pi = 0.5$
- $P(x = 13) = 0.008544921875$ (0.85%)
- $P(x \geq 13) = 0.0106353759765 6$ (1.06%)

The Probability Value

- In the James Bond example, the hypothesis is that he cannot tell the difference between shaken and stirred martinis.
- The probability value is low (0.0106), thus providing evidence that he can tell the difference.
- The p-value is a number, describes how likely you are to have found a particular set of observations if the null hypothesis were true.
- p-values are used in hypothesis testing to help decide whether to reject the null hypothesis.
- The smaller the p-value, the more likely you are to reject the null hypothesis.

The Null Hypothesis

The Null Hypothesis

- The null hypothesis is a characteristic arithmetic theory suggesting that no statistical relationship and significance exists in a set of given, single, observed variables between two sets of observed data and measured phenomena.
- The hypotheses play an important role in testing the significance of differences in experiments and between observations.
- H_0 symbolizes the null hypothesis of no difference.
- It presumes to be true until evidence indicates otherwise.

source: https://www.sciencedirect.com/topics/earth-and-planetary-sciences/null-hypothesis

Examples of a Null Hypothesis

- A school principal claims that students in her school score an average of seven out of 10 in exams.
- The null hypothesis is that the population mean is 7.0.
- To test this null hypothesis, we record marks of 30 students (sample) from the entire student population of the school (300) and calculate the mean of that sample.
- We can then compare the (calculated) sample mean to the (hypothesized) population mean of 7.0 and attempt to reject the null hypothesis.

Examples of a Null Hypothesis

- The annual return of a particular mutual fund is claimed to be 8%.
- Assume that a mutual fund has been in existence for 20 years.
- The null hypothesis is that the mean return is 8% for the mutual fund.
- We take a random sample of annual returns of the mutual fund for, say, five years (sample) and calculate the sample mean.
- We then compare the (calculated) sample mean to the (claimed) population mean (8%) to test the null hypothesis.

Examples of a Null Hypothesis

For the previous examples, null hypotheses are:

- Example 1: Students in the school score an average of seven out of 10 in exams.
- Example 2: Mean annual return of the mutual fund is 8% per year.

Significance Testing

- A low probability value casts doubt on the null hypothesis.
- So, how low must the probability value be in order to conclude that the null hypothesis is false?
- Although there is clearly no right or wrong answer to this question, it is conventional to conclude the null hypothesis is false if the probability value is less than 0.05 (< 0.05).
- More conservative researchers conclude the null hypothesis is false only if the probability value is less than 0.01 (< 0.01).

Significance Testing

- When a researcher concludes that the null hypothesis is false, the researcher is said to have rejected the null hypothesis.
- The probability value below which the null hypothesis is rejected is called the α level or simply **α** (alpha).
- It is also called the significance level.

Significance Testing

There are two approaches to conducting significance tests.

- favored by R. Fisher
- favored by the statisticians Neyman and Pearson

R. Fisher

- A significance test is conducted, and the probability value reflects the strength of the evidence against the null hypothesis.
- If the probability is below 0.01, the data provide strong evidence that the null hypothesis is false.
- If the probability value is below 0.05 but larger than 0.01, then the null hypothesis is typically rejected, but not with as much confidence as it would be if the probability value were below 0.01.
- Probability values between 0.05 and 0.10 provide weak evidence against the null hypothesis and, by convention, are not considered low enough to justify rejecting it.
- Higher probabilities provide less evidence that the null hypothesis is false

Neyman and Pearson

- This approach is to specify an α level before analyzing the data.
- If the data analysis results in a probability value below the α level, then the null hypothesis is rejected; if it is not, then the null hypothesis is not rejected.
- According to this perspective, if a result is significant, then it does not matter how significant it is.
- Moreover, if it is not significant, then it does not matter how close to being significant it is.
- Therefore, if the 0.05 level is being used, then probability values of 0.049 and 0.001 are treated identically.
- Similarly, probability values of 0.06 and 0.34 are treated identically.

Significance Testing

- The former approach (preferred by Fisher) is more suitable for scientific research.
- The latter is more suitable for applications in which a yes/no decision must be made.

Steps in Hypothesis Testing

1. Specify the null hypothesis.
2. Specify the α level which is also known as the significance level. Typical values are 0.05 and 0.01.
3. Compute the probability value (also known as the p value).
4. Compare the probability value with the α level.

Type I and II Errors

Bhandari, P.(2022). Type I & Type II Errors | Differences, Examples, Visualizations. Scribbr. from https://www.scribbr.com/statistics/type-i-and-type-ii-errors/

Introduction

- In statistics, a **Type I error** is a **false positive conclusion**, while a **Type II error** is a **false negative conclusion**.
- Making a statistical decision always involves uncertainties, so the risks of making these errors are unavoidable in hypothesis testing.
- The probability of making a Type I error is the significance level, or alpha (α), while the probability of making a Type II error is beta (β).
- These risks can be minimized through careful planning in your study design.

Example

- You decide to get tested for COVID-19 based on mild symptoms.
- There are two errors that could potentially occur:
  - Type I error (false positive): the test result says you have coronavirus, but you actually don't.
  - Type II error (false negative): the test result says you don't have coronavirus, but you actually do.

Error in Statistical Decision-Making

- Hypothesis testing starts with the assumption of no difference between groups or no relationship between variables in the population—this is the null hypothesis.
- It’s always paired with an alternative hypothesis, which is your research prediction of an actual difference between groups or a true relationship between variables.

Error in Statistical Decision-Making

Example:

- You test whether a new drug intervention can alleviate symptoms of an auto immune disease.
- In this case:
  - The null hypothesis (H_0) is that the new drug has no effect on symptoms of the disease.
  - The alternative hypothesis (H_1) is that the drug is effective for alleviating symptoms of the disease.
- Then, you decide whether the null hypothesis can be rejected based on your data and the results of a statistical test.

Error in Statistical Decision-Making

- Since these decisions are based on probabilities, there is **always** a risk of making the wrong conclusion.

- If your results show statistical significance, that means they are very unlikely to occur if the null hypothesis is true.
- In this case, you would reject your null hypothesis.
- But sometimes, this may be a Type I error.

- If your findings do not show statistical significance, they have a high chance of occurring if the null hypothesis is true.
- Therefore, you fail to reject your null hypothesis.
- But sometimes, this may be a Type II error.

Error in Statistical Decision-Making

- A Type I error happens when you get false positive results: you conclude that the drug intervention improved symptoms when it actually didn’t.
- These improvements could have arisen from other random factors or measurement errors.

- A Type II error happens when you get false negative results: you conclude that the drug intervention didn’t improve symptoms when it actually did.
- Your study may have missed key indicators of improvements or attributed any improvements to other factors instead.

Error in Statistical Decision-Making

Type I and Type II Error

Null hypothesis is ...	True	False
Rejected	**Type I error** False positive Probability = α	**Correct decision** True positive Probability = $1 - \beta$
Not rejected	**Correct decision** True negative Probability = $1 - \alpha$	**Type II error** False negative Probability = β

Scribbr

How do you reduce the risk of making a Type I error?

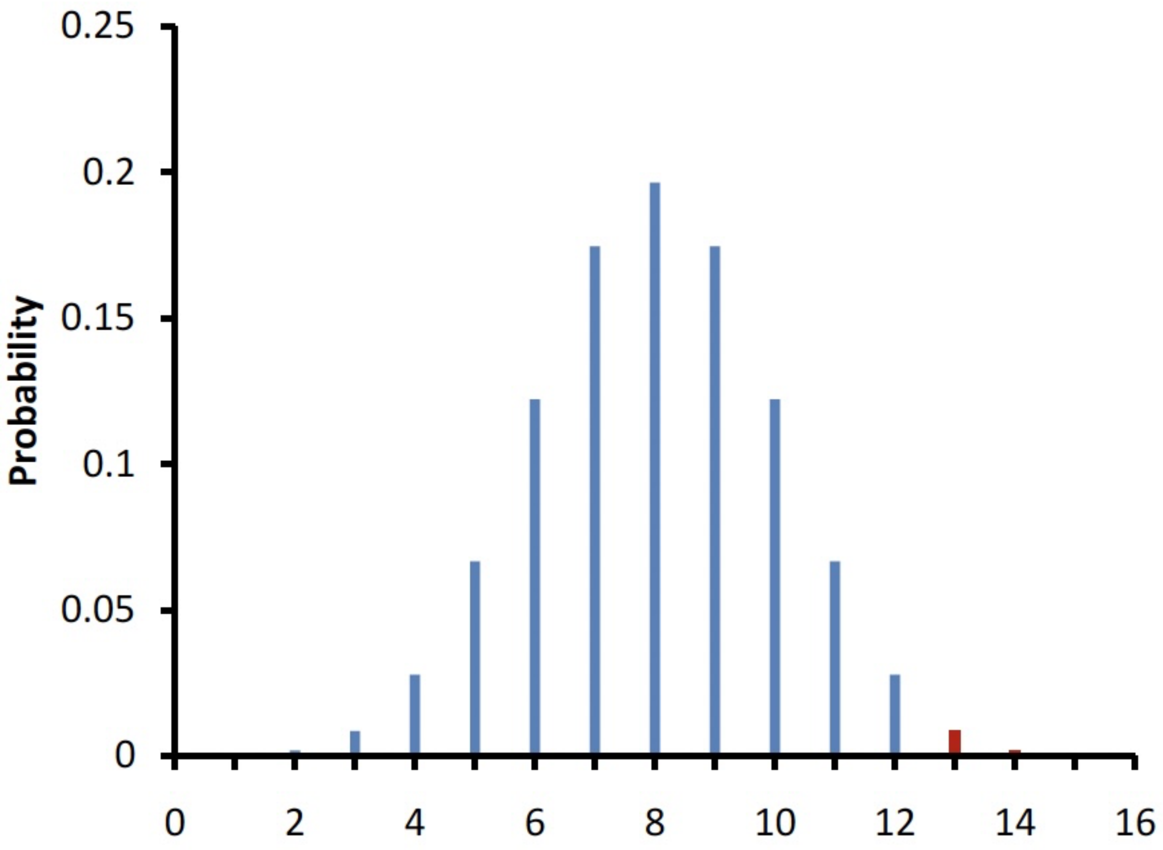
- The risk of making a Type I error is the significance level (or alpha) that you choose.
- That's a value that you set at the beginning of your study to assess the statistical probability of obtaining your results (p value).
- The significance level is usually set at 0.05 or 5%.
- This means that your results only have a 5% chance of occurring, or less, if the null hypothesis is actually true.
- To reduce the Type I error probability, you can set a lower significance level.

How do you reduce the risk of making a Type II error?

- The risk of making a Type II error is inversely related to the statistical power of a test.
- Power is the extent to which a test can correctly detect a real effect when there is one.
- To (indirectly) reduce the risk of a Type II error, you can increase the sample size or the significance level to increase statistical power.

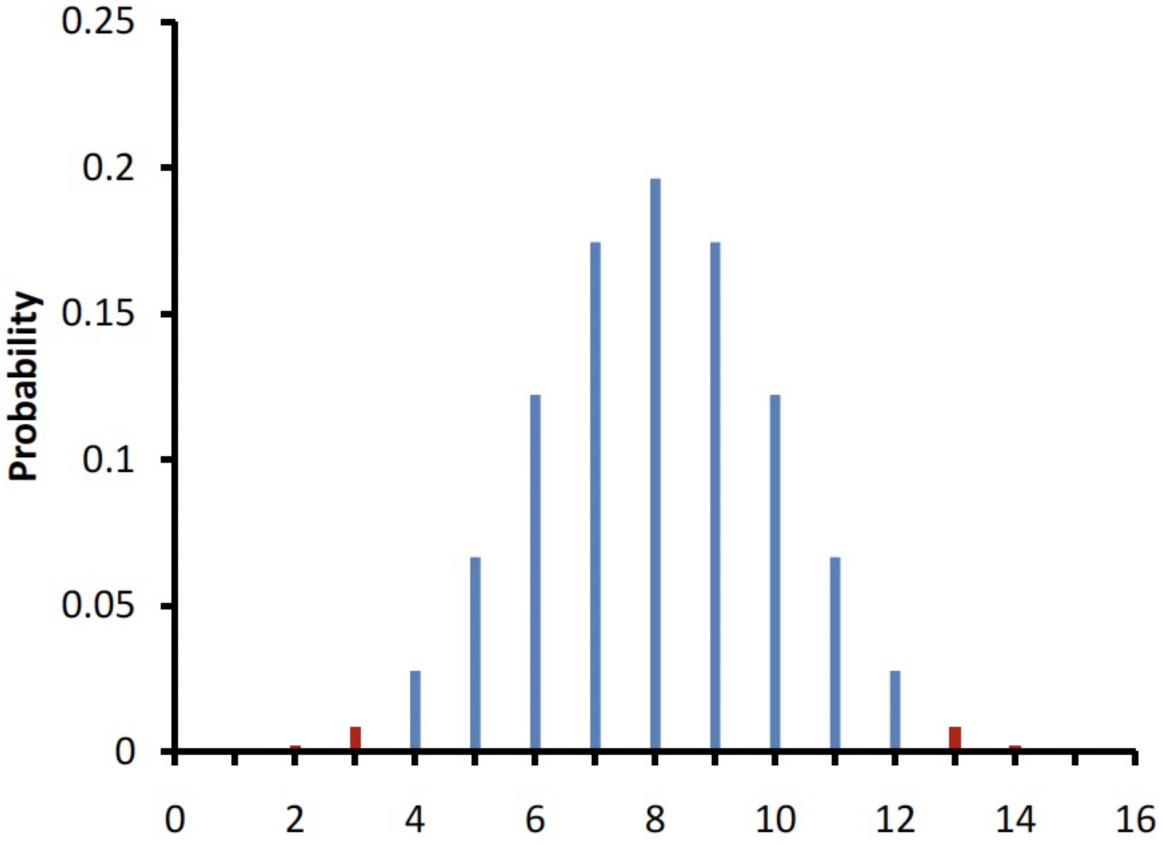
One- and Two-Tailed Tests

One-Tailed Test



- From the binomial distribution, we know that the probability of being correct 13 or more times out of 16 if one is only guessing is 0.0106
- The red bars show the values greater than or equal to 13.
- The probabilities are calculated for the upper tail of the distribution
- A probability calculated in only one tail of the distribution is called a “one-tailed probability.”

Two-Tailed Test



- "What is the probability of getting a result as extreme or more extreme than the one observed"?

- Since the chance expectation is 8/16, a result of 3/13 is equally as extreme as 13/16.
- Since the binomial distribution is symmetric when $\pi = 0.5$, this probability is exactly double the probability of 0.0106 computed previously.
- Therefore, $p = 0.0212$
- A probability calculated in both tails of a distribution is called a two-tailed probability

Introduction

- One and Two-Tailed Tests are ways to identify the relationship between the statistical variables.
- For checking the relationship between variables in a single direction (Left or Right direction), we use a one-tailed test.
- A two-tailed test is used for checking whether the relations between variables are in any direction or not.

Example

- A two-tailed test is appropriate if you want to determine if there is any difference between the groups you are comparing.
- For instance, if you want to see if Group A scored higher or lower than Group B, then you would want to use a two-tailed test.

- If you are only interested in determining if Group A scored higher than Group B, and you are completely uninterested in possibility of Group A scoring lower than Group B, then you may want to use a one-tailed test.

T Test for One Sample

T Test for One Sample

- When the population standard deviation, σ (sigma), is unknown, it must be estimated with the sample standard deviation, *s*.
- By the same token, the standard error of the mean, $\sigma_{\bar{x}}$ then must be estimated with $S_{\bar{x}}$. Under these circumstances, *t* rather than *z* should be used to test a hypothesis or to construct a confidence interval for the population mean.

- Further reading: chapter 13 (Witte, 2017)

What is a One Sample T Test?

- The one sample t test compares the mean of your sample data to a known value.
- For example, you might want to know how your sample mean compares to the population mean.
- You should run a one sample t test when you don’t know the population standard deviation, or you have a small sample size.

Example

- A company wants to improve sales. Past sales data indicate that the average sale was $100 per transaction.
- After training your sales force, recent sales data (taken from a sample of 25 salesmen) indicates an average sale of $130, with a standard deviation of $15.
- Did the training work? Test your hypothesis at a 5% alpha level.

- **Step 1**: Write your null hypothesis statement.
- The accepted hypothesis is that there is no difference in sales, so:

  $H0: \mu = \$100$

Example

- **Step 1**: Write your null hypothesis statement.
- The accepted hypothesis is that there is no difference in sales, so:

  H0: $\mu = \$100$

- **Step 2**: Write your alternate hypothesis.
- This is the one you're testing in the one sample t test.
- You think that there is a difference (that the mean sales increased), so:

  H1: $\mu > \$100$

Example

- **Step 3**: Identify the following pieces of information you'll need to calculate the test statistic.
- The question should give you these items:
  - The sample mean($\bar{x}$) → This is given in the question as $130.
  - The population mean(μ) → Given as $100 (from past data).
  - The sample standard deviation(s) = $15.
  - Number of observations(n) = 25.

Example

- **Step 4**: Insert the items from above into the t score formula.

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

t = (130 – 100) / ((15 / √(25))
t = (30 / 3) = 10

- This is your calculated t-value.

Example

- **Step 5:** Find the t-table value. You need two values to find this:
  1. The alpha level: given as 5% in the question.
  2. The degrees of freedom, which is the number of items in the **sample** (n) minus 1: 25 – 1 = 24.

Critical values of *t* for one-tailed tests

Significance level (α)

Degrees of freedom *(df)*	.2	.15	.1	.05	.025	.01	.005	.001
1	1.376	1.963	3.078	6.314	12.706	31.821	63.657	318.309
2	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327
3	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215
4	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173
5	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893
6	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208
7	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785
8	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501
9	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297
10	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144
11	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025
12	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930
13	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852
14	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787
15	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733
16	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686
17	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646
18	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610
19	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579
20	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552
21	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527
22	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505
23	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485
24	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467
25	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450

Example

- Look up 24 degrees of freedom in the left column and 0.05 in the top row.
- The intersection is 1.711.
- This is your one-tailed critical t-value.
- What this critical value means in a one tailed t test, is that we would expect most values to fall under 1.711.
- If our calculated t-value (from Step 4) falls within this range, the null hypothesis is likely true.

Example

- **Step 6**: Compare Step 4 to Step 5.
- The value from Step 4 does not fall into the range calculated in Step 5, so we can reject the null hypothesis.
- The value of 10 falls into the rejection region (the left tail).
- In other words, it's highly likely that the mean sale is greater.
- The one sample t test has told us that sales training was probably a success.

T Test for Two Independent Samples

T Test for Two Independent Samples

- Must be selected from among the following three possibilities

Nondirectional:	$H_0 : \mu_1 - \mu_2 = 0$ $H_1 : \mu_1 - \mu_2 \neq 0$
Directional, lower tail critical:	$H_0 : \mu_1 - \mu_2 \geq 0$ $H_1 : \mu_1 - \mu_2 < 0$
Directional, upper tail critical:	$H_0 : \mu_1 - \mu_2 \leq 0$ $H_1 : \mu_1 - \mu_2 > 0$

- Further reading: chapter 14 (Witte, 2017)

What is a Two-Sample T-Test?

- A two-sample t-test is used when you want to compare two independent groups to see if their means are different.
- “Independent” implies that the two samples must have come from two completely different populations.
- If you have independent samples, you can use the two-sample t-test.
- On the other hand, if your samples are connected in some way, run a paired samples t-test.
- “Connected” means that you are collecting data twice from the same group, person, item or thing.

Hypotheses

- H_o: The population mean of one group equals the population mean of the other group, or $\mu_1 = \mu_2$
- H_1: The population mean of one group does not equal the population mean of the other group, or $\mu_1 \neq \mu_2$

This test can also be conducted with a **directional** alternate hypothesis:

- H_o: The population mean of one group equals the population mean of the other group, or $\mu_1 = \mu_2$
- H_1: The population mean of one group is greater than the population mean of the other group, or $\mu_1 > \mu_2$

Example

- The following are a few real-life examples where two-sample T-test for independent samples can be used:
  1. Comparing the average test scores of two classes from two different schools
  2. Comparing the average weights of two different or independent groups of people
  3. Determining whether the medication have the same efficacy on two different or independent groups of people
  4. Compare whether the effect of vaccination on two different groups

Example

- One way to measure a person’s fitness is to measure their body fat percentage.
- Average body fat percentages vary by age, but according to some guidelines, the normal range for men is 15-20% body fat, and the normal range for women is 20-25% body fat.
- Our sample data is from a group of men and women who did workouts at a gym three times a week for a year. Then, their trainer measured the body fat.

Example

- The table below shows the data.

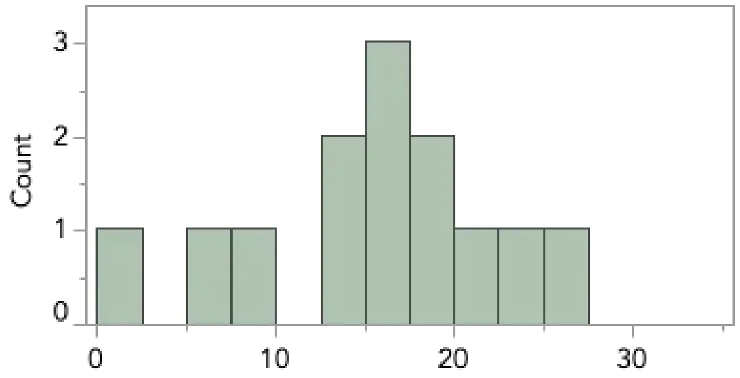
Table 1: Body fat percentage data grouped by gender

Group	Body Fat Percentages				
Men	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		
Women	22.0	16.0	21.7	21.0	30.0
	26.0	12.0	23.2	28.0	23.0

Example

Group=Men

Body Fat Percentage

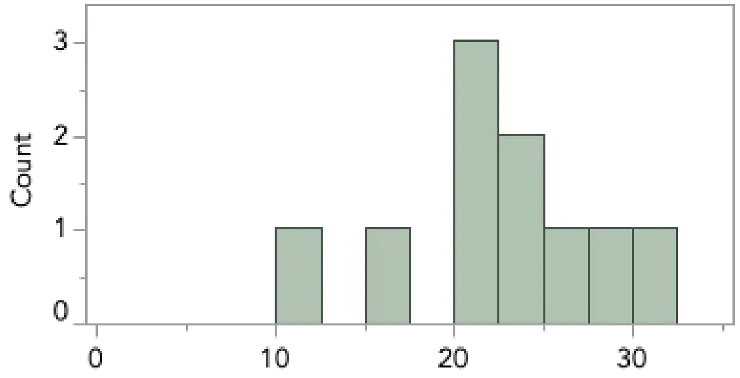


Summary Statistics

Mean	14.95
Std Dev	6.84
Std Err Mean	1.90
Upper 95% Mean	19.08
Lower 95% Mean	10.81
N	13.00

Group=Women

Body Fat Percentage



Summary Statistics

Mean	22.29
Std Dev	5.32
Std Err Mean	1.68
Upper 95% Mean	26.10
Lower 95% Mean	18.48
N	10.00

Example

- For each group, we need the average, standard deviation and sample size.

Table 2: Average, standard deviation and sample size statistics grouped by gender

Group	Sample Size (n)	Average (X-bar)	Standard deviation (s)
Women	10	22.29	5.32
Men	13	14.95	6.84

Example

- We start by calculating our test statistic.
- This calculation begins with finding the difference between the two averages:

  $$22.29 - 14.95 = 7.34$$

- This difference in our samples estimates the difference between the population means for the two groups.
- Next, calculate the pooled standard deviation.

Example

- This builds a combined estimate of the overall standard deviation.
- First, we calculate the pooled variance:

$$s_p^2 = \frac{((n_1-1)s_1^2)+((n_2-1)s_2^2)}{n_1+n_2-2}$$

$$s_p^2 = \frac{((10-1)5.32^2)+((13-1)6.84^2)}{(10+13-2)}$$

$$= \frac{(9\times 28.30)+(12\times 46.82)}{21}$$

$$= \frac{(254.7+561.85)}{21}$$

$$= \frac{816.55}{21} = 38.88$$

$$\sqrt{38.88} = 6.24$$

Example

- Formula: $$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t = \frac{\text{difference of group averages}}{\text{standard error of difference}} = \frac{7.34}{(6.24 \times \sqrt{(1/10+1/13)})} = \frac{7.34}{2.62} = 2.80$$

- The population mean$(\mu_1 - \mu_2) = 0$

Example

- To evaluate the difference between the means, compare the test statistic to a theoretical value from the t-distribution. This activity involves 4 steps:
  1. Decide on the risk we are willing to take for declaring a significant difference. We decide that we are willing to take a 5% risk of saying that the unknown population means for men and women are not equal when they really are. In statistics-speak, the significance level, denoted by α, is set to 0.05.
  2. Calculate a test statistic. The test statistic is 2.80.
  3. Find the theoretical value from the t-distribution based on our null hypothesis which states that the means for men and women are equal.

Example

- To find this value, we need the significance level ($\alpha = 0.05$) and the degrees of freedom.
- The degrees of freedom (df) are based on the **sample** sizes of the two groups.
- For the body fat data, this is:

$$df = n1 + n2 - 2 = 10 + 13 - 2 = 21$$

Example

- The t value with α = 0.05 and 21 degrees of freedom is 2.080.

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05
df							
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086
21	0.000	0.686	0.859	1.063	1.323	1.721	$\boxed{2.080}$
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074

Example

4. Compare the value of our statistic (2.80) to the *t* value.
   Since 2.80 > 2.080, we reject the null hypothesis that the mean body fat for men and women are equal and conclude that we have evidence body fat in the population is different between men and women.

T Test for Two Related Samples

T Test for Two Related Samples

- Selected from among the following three possibilities, where μ_D represents the population mean for all difference scores:

Nondirectional:

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

Directional, lower tail critical:

$$H_0\colon \mu_D \geq 0$$
$$H_1\colon \mu_D < 0$$

Directional, upper tail critical:

$$H_0\colon \mu_D \leq 0$$
$$H_1\colon \mu_D > 0$$

- Further reading: chapter 15 (Witte, 2017)

T Test for Two Related Samples

- The test statistic for the Paired Samples t Test, denoted t, follows the same formula as the one sample t test.

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

Example

- Testing two production lines to see if their outputs are different.
- One-line feeds into a second line, so the second line depends on the first for at least part of the production.

- Comparing test scores for the same group of students before an intensive study session and after the session.
- You're testing the same people twice, so a paired test is needed.

- You are subjecting two different model cars to crashworthiness, using the same equipment.
- Although you're testing different items, they are being subjected to the same conditions and so are paired.

Example

- An instructor wants to use two exams in her classes next year.
- This year, she gives both exams to the students.
- She wants to know if the exams are equally difficult and wants to check this by looking at the differences between scores.
- If the mean difference between scores for students is “close enough” to zero, she will make a practical conclusion that the exams are equally difficult.

Example

- Here is the data:
- If you look at the table, you see that some of the score differences are positive, and some are negative.
- You might think that the two exams are equally difficult.
- Other people might disagree.
- The statistical test gives a common way to make the decision, so that everyone makes the same decision on the same data.

Table 1: Exam scores for each student

Student	Exam 1 Score	Exam 2 Score	Difference
Bob	63	69	6
Nina	65	65	0
Tim	56	62	6
Kate	100	91	-9
Alonzo	88	78	-10
Jose	83	87	4
Nikhil	77	79	2
Julia	92	88	-4
Tohru	90	85	-5
Michael	84	92	8
Jean	68	69	1
Indra	74	81	7
Susan	87	84	-3
Allen	64	75	11
Paul	71	84	13
Edwina	88	82	-6

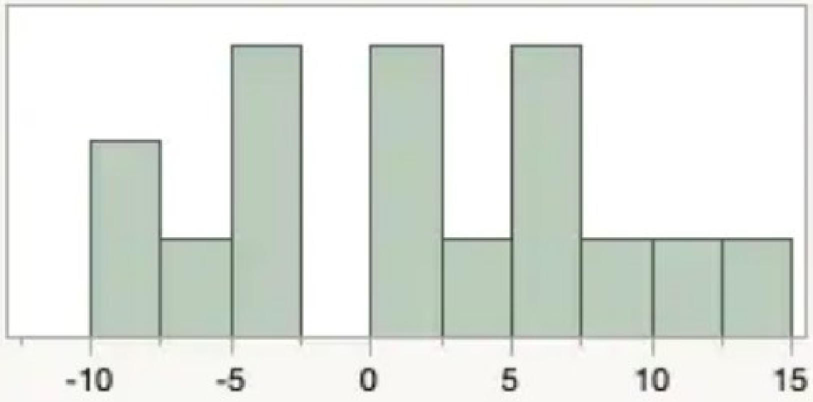
Example

Checking the data

- Let's start by answering: Is the paired t-test an appropriate method to evaluate the difference in difficulty between the two exams?
  1. Subjects are independent. Each student does their own work on the two exams.
  2. Each of the paired measurements are obtained from the same subject. Each student takes both tests.
  3. The distribution of differences is normally distributed. For now, we will assume this is true.
- We decide that we have selected a valid analysis method.

Example

- The figure below shows a histogram and summary statistics for the score differences.



Summary Statistics

Mean	1.3125
Std Dev	7.002083
Std Err Mean	1.7505208
Upper 95% Mean	5.0436467
Lower 95% Mean	-2.418647
N	16

Figure 1: Histogram and summary statistics for the difference in test scores

- From the histogram, the data are roughly bell-shaped, so the idea of a normal distribution for the differences seems reasonable.

Example

- The average score difference is:

$$\overline{x_d} = 1.31$$

- Next, we calculate the standard error for the score difference. The calculation is:

$$\text{Standard Error} = \frac{s_d}{\sqrt{n}} = \frac{7.00}{\sqrt{16}} = \frac{7.00}{4} = 1.75$$

- In the formula above, *n* is the number of students – which is the number of differences.

Example

- Calculate our test statistic as:

$$t = \frac{\text{Average difference}}{\text{Standard Error}} = \frac{1.31}{1.75} = 0.750$$

Example

- To make our decision, we compare the test statistic to a value from the t-distribution. This activity involves four steps:
  1. Decide on the risk we are willing to take for declaring a difference when there is not a difference. For the exam score data, we decide that we are willing to take a 5% risk of saying that the unknown mean exam score difference is zero when in reality it is not. In statistics-speak, we set the significance level, denoted by α, to 0.05.
  2. We calculate a test statistic. Our test statistic is 0.750.
  3. We find the value from the t-distribution.

Example

- To find this value, we need the significance level ($\alpha = 0.05$) and the degrees of freedom.
- The degrees of freedom (df) are based on the population size.
- For the exam score data, this is:

  df = n = 16

Example

- The t value with α = 0.05 and 16 degrees of freedom is 2.120.

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05
df							
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131
16	0.000	0.690	0.865	1.071	1.337	1.746	$\boxed{2.120}$
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074

Example

4. We compare the value of our statistic (0.750) to the t value. Because 0.750 < 2.120, we can reject our idea that the mean score difference is zero.

References

- Witte, R.S.&Witte, J.S. (2017). Statistics (11th ed.). Wiley. ISBN: 978-1119386056.
- Lane, D.M., Scott, D., Hebl, M., Guerra, R., Osherson, D.& Zimmer, H. (2003). Introduction to Statistics. Online edition at https://open.umn.edu/opentextbooks/textbooks/459
- Levine, D.M., Stephan, D.F. & Szabat, K.A. (2017). Statistics for Managers Using Microsoft Excel (8th ed.). Pearson. ISBN: 978-0134566672

Thank you