CS555B1 Data Analysis and Visualization

Lecture 4

Statistical inference, t-test, one and two sample tests

Kia Teymourian

Test of Significance (hypothesis test)

Tests of significance are used to assess the evidence provided by the data from a sample about some claim concerning the population.

There are **5 key steps in carrying out any significance test**:

1. Set up the **hypotheses** and **select the alpha level**
2. Select the appropriate **test statistic**
3. State the **decision rule**
4. Compute the **test statistic** and the associated **p-value**
5. State your **conclusion**

Z test

A **Z-test is any statistical test** for which the distribution of the test statistic under the null hypothesis can be approximated by **a normal distribution.**

Because of the **central limit theorem**, many test statistics are **approximately normally distributed** for large samples.

Many statistical tests can be conveniently performed as approximate Z-tests if:

- the sample size is large (**n >=30**)
- the **population variance is known or unknown**

Confidence Interval is: $\bar{x} \pm Z_{CL} * \frac{\sigma}{\sqrt{n}}$ OR $\bar{x} \pm Z_{CL} * \frac{S}{\sqrt{n}}$

If the population variance is unknown (and therefore has to be estimated from the sample itself) and the sample size is not large (n < 30), the Student's t-test may be more appropriate.

n is large and σ is unknown

We still require that
- the data in the sample are taken randomly from the population of interest

We still assume that
- observations are independent from each other
- the distribution of the parameter in the population of interest is normally distributed with a mean of μ and a standard deviation of σ
- the population mean, μ, is unknown and that we are interested in making conclusions about this parameter using data from our sample

The test statistic for hypotheses about the mean μ is the z statistic

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

μ is the value of the population mean under the null hypothesis, **s** is the standard deviation of the variable of interest in sample, **SE is the standard error of the sample mean (denoted)**

n is large and σ is unknown

- z is (even when calculated using s instead of σ) normally with a mean of 0 and a standard deviation of 1.
- Thus the values in the standard normal table can still be used to calculate the associated p-value.

To calculate a confidence interval with a confidence level of C for the population mean, μ:

$$\bar{x} \pm z \,.\, \frac{s}{\sqrt{n}} = \bar{x} \pm z \,.\, SE_{\bar{x}}$$

z is the appropriate critical value corresponding to the confidence level C,
s is the sample standard deviation
n is the sample size

n is small (<30) and σ is unknown

When n is not large (<30), the Central Limit Theorem does not guarantee that the distribution of the sample mean is perfectly normal.

When n is not large(<30) and the standard deviation of the population is not known, the test statistic for hypotheses about the mean μ is the t statistic:

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

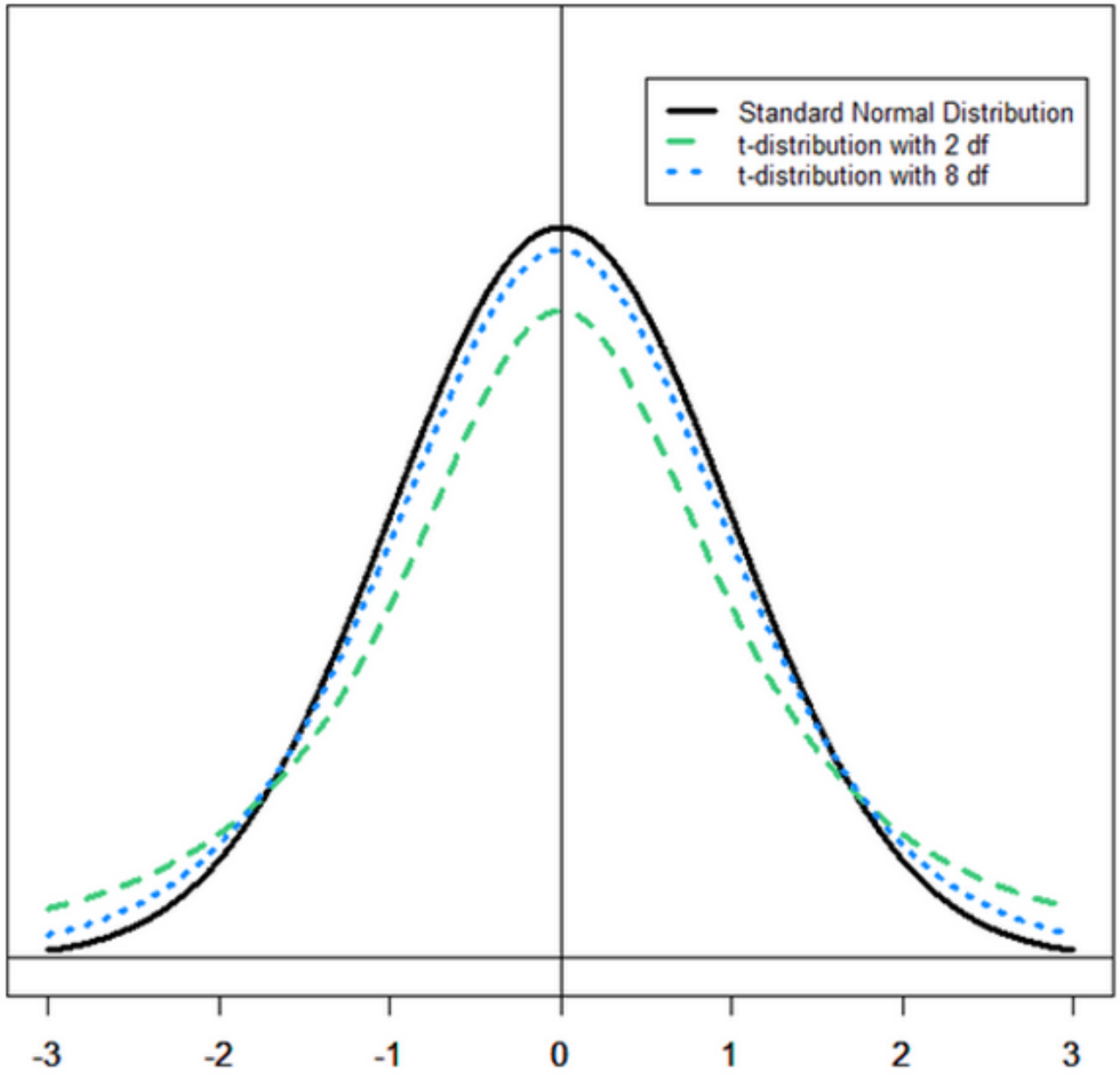
$\bar{x}$ is the sample mean, μ is the value of the population mean under the null hypothesis, s is the sample standard deviation, and n is the number of observations.

This test is known as the **one-sample t-test**.

t-statistic

- t-statistic is not normally distributed
- It is a t-distribution with **n−1 degrees of freedom (df)**
- The shape of the distribution **varies** with the sample size
- The **variability** of the t-distributions is slightly **greater** than that of the standard normal distribution
- As n increases, s is a better and better estimate of σ, and the variability decreases.

Try the example at:
http://rpsychologist.com/d3/tdist/



t-Distribution Critical Values

Based on df and Confidence Level C

Table B. *t*-Distribution Critical Values

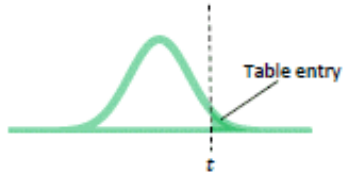


Table entry for p and C is the critical value t with probability p lying to its right and probability C lying between $-t$ and t

	Upper Tail Probability, p											
df	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.895	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z	0.674	0.842	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.090	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence Level, C											

Confidence interval in t-statistic

To calculation confidence interval with a confidence level of C for the population mean, μ:

$$\bar{x} \pm t \cdot \frac{s}{\sqrt{n}} = \bar{x} \pm t \,.\, SE$$

X bar is the sample mean,
t is the appropriate critical value corresponding to the confidence level C with df=n-1,
s is the sample standard deviation, and n is the number of observations.
SE denotes the standard error.

t-statistic may also be used in cases where **n≥30 but the population distribution is not normally distributed.**

When the population is not perfectly normally distributed, the t-distribution provides **an adequate approximation**.

t-statistic: an example

Biologists studying the healing skin wounds measured by the rate at which new cells closed a razor cut made in the skin of a **newt**. The data from **18 newts** were measured.

The **sample mean and standard deviation** of the healing rates were **25.67 and 8.345** micrometers per hour.

Calculate the **95% confidence interval** for the mean healing rate in the population of all newts.

$$\begin{aligned}\bar{x} \pm t \cdot \frac{s}{\sqrt{n}} &= 25.67 \pm 2.110 \cdot \frac{8.345}{\sqrt{18}} \\ &= 25.67 \pm 2.110 \cdot 1.9669 \\ &\approx 25.67 \pm 4.15 \\ &\approx (21.52, 29.82)\end{aligned}$$

n=18

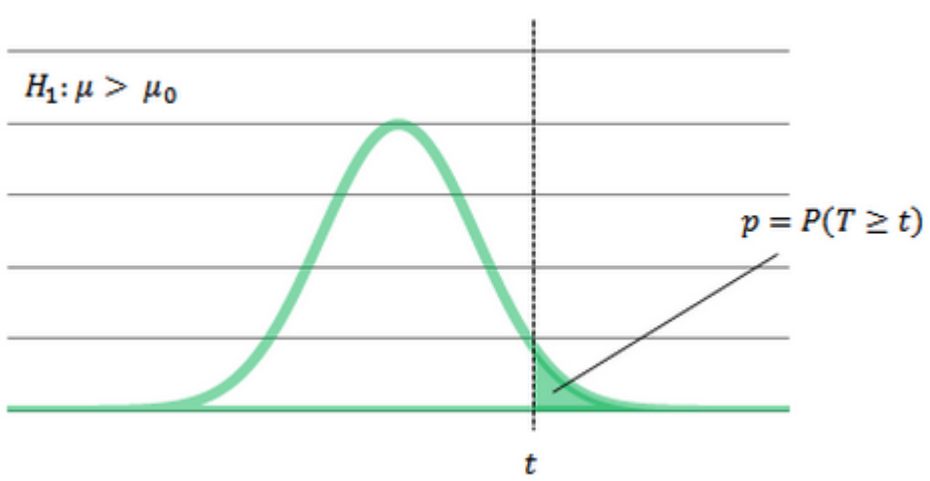
df = n−1=17

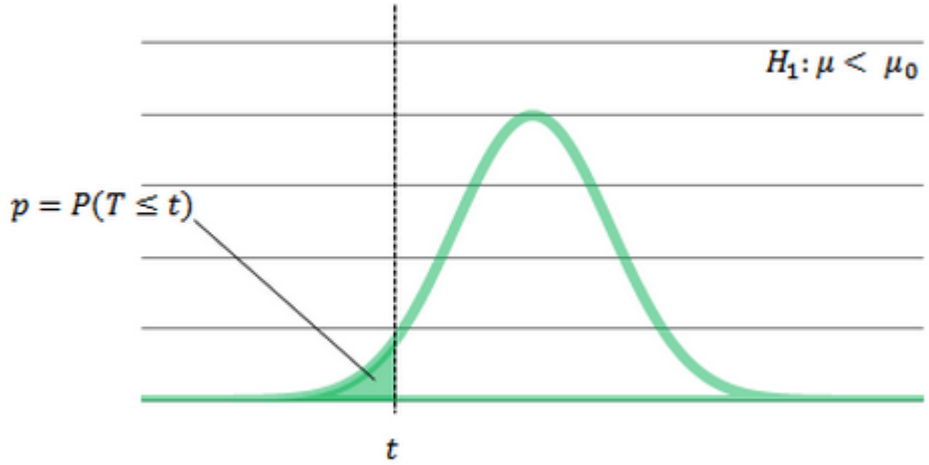
Using Table B, the **critical value is 2.110**.

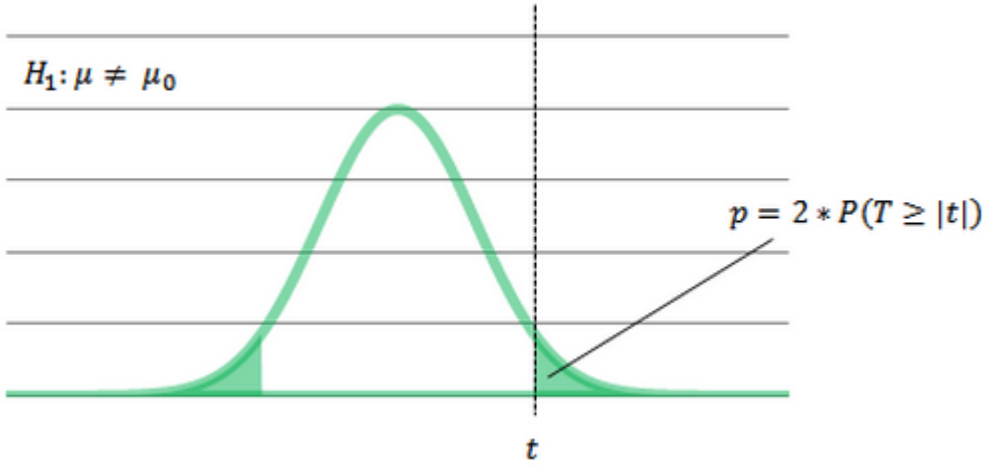
The **95% confidence level** is associated with an upper tail probability of 0.025 (since given the symmetry of the distribution, a total area of **0.05=2 * 0.025** is contained within the left and right tails).

P value calculation in t-statistic

Depending on the **alternative hypothesis**, the p-value is calculated as follows:







t-statistic: a one-sided example

A scientist wishes to test the claim that **great white sharks** average **20 feet** in length.

To test this, he measures **10** great white sharks.

The measurements are **18.1, 23.4, 23.8, 24.1, 22.5, 19, 25.4, 23.1, 16.5, 26.7**.

Are these data good evidence that great white sharks are **longer** than 20 feet in length at the **α=0.10 level of significance**?

t-statistic: a one-sided example

1. Set up the hypotheses and select the alpha level
- H_0:μ=20 (the mean length of great while sharks is 20 ft)
- H_1:μ>20 (the mean length of great while sharks is greater than 20 ft)
- α=0.10

2. Select the appropriate test statistic

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

3. State the decision rule
- Determine the appropriate critical value from the standard t-distribution table associated with a right hand tail probability of **α=0.10 based on df = 9**.
- Using the table, the appropriate critical value is **1.383** as shown below.
- **Decision Rule: Reject H0 if t≥1.383**
- Otherwise, do not reject H0

t-statistic: a one-sided example

4. Compute the test statistic and the associated p-value

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{22.27 - 20}{\frac{3.31}{\sqrt{10}}} \approx \frac{2.27}{1.05} \approx 2.16$$

5. Conclusion

- **Reject H_0 since 2.16 is greater than 1.383**
- We have significant evidence at the **α=0.10** level that the mean length of great white sharks is **longer than 20 feet**
- The sample mean was **22.27 feet**. **We reject the null hypothesis** that the mean length of great white sharks is **20 feet**.

- Calculate p-value

	Upper Tail Probability, p											
df	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.[illegible]	12.[illegible]06	15.895	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.[illegible]03	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.[illegible]82	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.[illegible]32	2.[illegible]76	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.[illegible]71	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.[illegible]47	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.[illegible]65	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.[illegible]60	2.[illegible]06	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587

t test R commands

- Calculating **probabilities from t-statistics**

> p <- **pt**([t statistics], df = [degree of freedom])

It calculates the **area to the left of a given t-statistics**

- Calculating **t-statistics from probabilities**

> t <- **qt**([probability], df = [degree of freedom])

It find the **t-statistic with the specified area to the left**

- **t.test function**

> **t.test**(data$variable, mu=[$\mu_0$], alternative=[alternative], conf.level=[confidence level]

[alternative] = ‘less’, ‘greater’, or ‘two.sided’

NOTE: must use “**two.sided**” option if you want to calculate confidence intervals. Confidence intervals from other options are not correct.

t test R commands

```
> pt(2.76, df=4)
# [1] 0.9745752
> qt(0.975, df=4)
# [1] 2.776445
> qt(0.025, df=4)
# [1] -2.776445
```

Table B. *t*-Distribution Critical Values

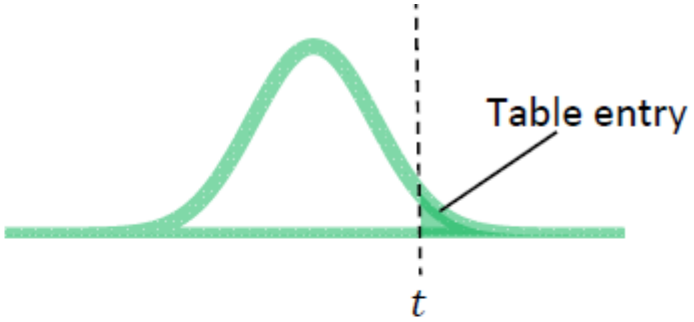


Table entry for p and C is the critical value t with probability p lying to its right and probability C lying between $-t$ and t

	Upper Tail Probability, p											
df	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.895	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610

t test R commands

A scientist wishes to test the claim that great white sharks average **20 feet in length.**

To test this, he measures **10** great white sharks.
The measurements are **18.1, 23.4, 23.8, 24.1, 22.5, 19, 25.4, 23.1, 16.5, 26.7.**

Are these data **good evidence** that great white sharks are longer than **20 feet** in length at the **α=0.10 level of significance**?

```
> shark_len= c(18.1, 23.4, 23.8, 24.1, 22.5, 19, 25.4, 23.1, 16.5, 26.7)
> t <- (mean(shark_len) - 20)/(sd(shark_len)/sqrt(length(shark_len)))
> t
[1] 2.160244
> p <- 1 – pt(t ,  df=length(shark_len) - 1)
> p
[1] 0.02952202
> t.test(shark_len, mu=20, alternative="greater", conf.level=0.9)
```

t-statistic: a two-sided example

The biologists studying the **18 newts** were interested in estimating the mean healing rate for the newts under natural conditions.

In another experiment, the researchers used the same 18 newts and measured the healing rate after applying a **topical therapy that was developed to help speed up healing rates.**

The data from the 18 newts were measured **after the topical treatment**.

The difference between the healing rate for each newt with and without the topical treatment were calculated where positive numbers indicated an increase in the healing rate.

The resulting **mean** of the differences in the **sample was 3.2** and the **standard deviation** of the differences in the sample was **4.5.**

Test whether or not the topical treatment changed the healing rate of the newts at the **α=0.10 level of significance**.

t-statistic: a two-sided example

1. Set up the hypotheses and select the alpha level
- H_0:μ=0 (the mean difference is 0, the tropical treatment has no effect on healing)
- H_1:μ≠0 (the mean difference is not 0, the tropical treatment has an effect on healing)
- α=0.10

2. Select the appropriate test statistic

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

since n is small and the sd of the population in unknown

3. State the decision rule
- Determine the appropriate critical value from the standard t-distribution table associated with a right hand tail probability of α/2=0.1/2=0.05 based on df = 17.
- Using the table, the appropriate critical value is 1.740 as shown below.
- Decision Rule: Reject H_0 if |t|≥1.740
- Otherwise, do not reject H_0

t-statistic: a two-sided example

4. Compute the test statistic and the associated p-value

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{3.2 - 0}{\frac{4.5}{\sqrt{18}}} \approx \frac{3.2}{1.06} \approx 3.02$$

5. **Conclusion**

- Reject H_0 since 3.02 is greater than 1.740
- We have significant evidence at the $\alpha=0.05$ level that the healing rate is faster with the topical cream than without.
- The topical **cream increased the healing rate by 3.2** micrometers per hour.
- **We reject the null hypothesis** that the mean topical cream had no effect on healing rate of the newts.

- Calculate p-value

t-statistic: a two-sided example

Calculate p-value

- The t-statistic lies between the critical values for **0.005 and 0.0025**, the p-value associated with the t-statistic of 3.02 is between 2·0.005 and 2·0.0025 **(since it is a two-sided test)**. That is, the p-value is between 0.005 and 0.010.
- **P value is equal to 0.008 (which indeed is between 0.005 and 0.010).**

	Upper Tail Probability, *p*											
df	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.895	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883

When t-test is not applicable …

In real-world applications, data is **rarely perfectly normally distributed**.

The validity of the procedures involving the t-distribution rely on how **different** the distribution of the data is **from the normal distribution**.

The presence of **outliers or strong skewness** in the distribution **is indicative of severe departure from this assumption**.

As the mean and the standard deviation are **sensitive to outliers and strong skewness**, the procedures based on the **t-distribution** where the sample sizes are small are also greatly influenced by such **departures from normality**.

It is essential to check the distribution of your sample data and **check for strong skewness and outliers before you use t-procedures**.

- If strong skewness exists or there are outliers present, use of the t-procedures is **not recommended** (**especially** when using these procedures when **n<30**).

About the one Sample examples

Studies where the same individual or experimental unit is given more than one treatment and the variable of interest is the change between the treatments is called a **matched pairs design**.

We applied **one sample procedures** (either the one sample t-test or the one sample z-test, as appropriate) to analyze the observed differences.

Since **two measurements** are taken on the same **observational unit** in this type of design, the observations must first be compared and the analysis needs to take place on the differences.

We will go on and discuss **tests for two samples**.

Note that the two-sample tests do not apply to this situation.

Application of that type of analysis to this type of data would result in a loss of efficiency (that is, we'd be less likely to see differences even if differences existed).

Two-sample problems

We are interested in comparing the **distribution of the sample means from two populations**.

We have separate samples from each population and we want to make conclusions about the characteristics of the two samples for a particular parameter.

In the two sample setting, one may be interested in **comparing the centers (the means) or the spreads (the standard deviations)** between the two populations.

Two-sample tests for means are one of the most common situations encountered in statistical practice. It is the focus of our course.

Conditions to use two-sample procedures

In the two-sample setting, conditions necessary for inference include:

1. samples must be **independent** (**not influencing** each other) and **randomly** selected from the two distinct populations of interest
2. the variable of interest must be **measured in the same way** in each of the populations
3. the parameter of interest should be **normally distributed** (or at least have similar shapes and without outliers)

The notation used for two-sample problems

In the two sample tests for means, we are interested in estimating $\mu_1 - \mu_2$ or testing the null hypothesis of no difference between means, $\mu_1 - \mu_2 = 0$ or, $\mu_1 = \mu_2$.

The unknown quantities, the population means, μ_1 and μ_2 , and the population standard deviations, σ_1 and σ_2, are estimated using their corresponding sample statistics. That is, $\bar{x}_1$ and $\bar{x}_2$ estimate μ_1 and μ_2 while s_1 and s_2 estimate σ_1 and σ_2.

Population	Variable	Population		Sample		
		Mean	Standard Deviation	Sample size	Mean	Standard Deviation
1	x_1	μ_1	σ_1	n_1	$\bar{x}_1$	s_1
2	x_2	μ_2	σ_2	n_2	$\bar{x}_2$	s_2

Two-sample Test of Means

The standard deviation of the difference in the sample means, $\bar{x}_1 - \bar{x}_2$, is $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Give that σ_1 and σ_2 are generally unknown, we must estimate the standard deviation of the difference in the sample means by using the sample standard deviations instead of the population standard deviations:

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

This quantity is referred to as the standard error of the difference in sample means.

The two sample t-statistic used for hypothesis testing is calculated by dividing the difference in the sample means by the standard error of the sample means:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{x}_i$ is the sample mean from population i
s_i is the standard deviation of the variable of interest in the sample from population i
n_i is the number of observations in the sample from population i

This represents how far the difference in sample means is from 0 in standard deviations units.

Two-sample Test of Means

The formula above is sometimes written as:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

This is for tests of the null hypothesis: μ1−μ2=d. Since generally we are concerned with testing if the means in the population are different, d is set to 0 which leaves us with the formula presented above

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Two-sample Test: degree of freedom of t-distribution

The t-statistic approximately has a t-distribution. Its **degrees of freedom** has a very complicated formula:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\sqrt{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}}$$

Without access to software, a conservative approach would be to estimate the degrees of freedom by **taking the minimum of the quantities $n_1 - 1$ and $n_2 - 1$.**

- This approach will always result in a p-value that is slightly larger than the actual value.
- Using this approach you will not be far off unless the sample sizes are both quite small and are not equal to each other.
- This approximation becomes better and better as the sample sizes in each of the populations increase.

Two-sample Test: an example

In order to assess **how quickly polyester decays** over time in landfills, a researcher buried strips of the material in the soil for different lengths of time and then tested the force required to break them (as a measure of decay). Lower breaking strength is indicative of decay. Here is the data collected:

Group	Sample Data				
2 weeks	118	126	126	120	129
16 weeks	124	98	110	140	110

Test whether or not the breaking strengths of polyester strips buried for 2 weeks is greater than the breaking strengths of those buried for 16 weeks.

Perform the test at the $\alpha=0.10$ level of significance.

Two-sample Test: an example

The summary statistics are presented in the table:

Group	n	Mean	Standard Deviation
2 weeks	5	123.8	4.60
16 weeks	5	116.4	16.09

1. Set up the hypotheses and select the alpha level
- H_0: $\mu_1=\mu_2$ (the mean breaking strengths are the same)
- H_1: $\mu_1>\mu_2$ (the mean breaking strengths are greater after 2 weeks versus 16 weeks)
- $\alpha=0.10$

2. Select the appropriate test statistic $$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

since n is small and the sd of the population in unknown

Two-sample Test: an example

3. State the decision rule

- Determine the appropriate critical value from the standard t-distribution table associated with a right hand tail probability of $\alpha=0.10$.
- We would choose $5-1=4$ in this case for the degrees of freedom.
- Using the table, the appropriate critical value is 1.533 as shown below.
- Decision Rule: Reject H_0 if $t \geq 1.533$
- Otherwise, do not reject H_0

	Upper Tail Probability, p											
df	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	[illegible]	6.314	12.706	15.895	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	[illegible]	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	[illegible]	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	[illegible]	[illegible]	[illegible]	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869

Two-sample Test: an example

4. Compute the test statistic and the associated p-value

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{123.8 - 116.4}{\sqrt{\frac{4.60^2}{5} + \frac{16.09^2}{5}}} = \frac{7.4}{7.484} = 0.9889$$

5. Conclusion

- **Fail to reject H_0 since 0.9889 is less than 1.53**
- We do not have significant evidence at the α=0.10 level that the mean breaking strength is greater for polyester left to decay for 2 weeks as opposed to 16 weeks.
- The difference in **sample means was 7.4** indicating a higher breaking strength for polyester strips left to decay 2 weeks as opposed to those left to decay for 16 weeks.
- We do not reject the null hypothesis that the mean breaking strength is the same between the two groups.

Two-sample Test: an example

Calculate p-value:

look across the df = 4 row to figure out which values t=0.9889 is between the critical values for 0.2 and 0.15.

The p-value associated with the t-statistic of 0.9889 is between 0.15 and 0.20 (since it is a one-sided test, we don't need to multiply by two).

	Upper Tail Probability, p											
df	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.895	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437

Two-Sample Confidence Interval for Comparison of Means

The confidence interval for the difference in population means is:

$$(\bar{x}_1 - \bar{x}_2) \pm t \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$\bar{x}_i$ is the sample mean from population i
s_i is the standard deviation of the variable of interest in the sample from population i
n_i is the number of observations in the sample from population I
t is the critical value corresponding to the confidence level with df

Confidence intervals are typically of the form:

$$\text{estimate} \pm \text{critical value} \cdot SE_{\text{estimate}}$$

Two-sample Test: Calculating confidence interval

Let’s return to the example on breaking strengths of polyester.

Calculate the 90% confidence interval for the difference in breaking strengths between groups.

The summary statistics are presented in the table below:

Group	n	Mean	Standard Deviation
2 weeks	5	123.8	4.60
16 weeks	5	116.4	16.09

Two-sample Test: Calculating confidence interval

The confidence interval for the difference in population means is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The critical value for the 90% confidence interval can be estimated from the standard t-distribution table associated with a right hand tail probability of α/2=0.10/2=0.05.

We would choose 5−1=4 in this case for the degrees of freedom. Using the table, the appropriate critical value is 2.132.

The confidence interval for the difference in population means:

$$(123.8 - 116.4) \pm 2.132 \cdot \sqrt{\frac{4.60^2}{5} + \frac{16.09^2}{5}} = 7.4 \pm 2.132 \cdot 7.484$$

$$\approx -8.56 \text{ to } 23.36$$

We are 90% confident that the mean difference in breaking strengths is between −8.56 and 23.36.

Pooled two-sample t test

The formulas presented above for two-sample tests for means work regardless of whether or not the sample standard deviations are the same.

When equal variances between groups are assumed, another methodology, called the **pooled two-sample t-test** (or the two-sample tests for equal variances), can be used.

However, the **methodology presented above is preferred** in all cases given its generality and lack of this assumption.

The pooled two-sample t-test is more sensitive to departures from normality.

In cases where the **sample sizes from each population are the same, the pooled two-sample t-test has the same results** as the methodology presented above.

Two-sample t test R commands

- **t.test function**

> **t.test(x, y, alternative=[alternative], conf.level=[confidence level]**

[alternative] = ‘less’, ‘greater’, or ‘two.sided’

var.equal =FALSE (default) should always be used

Be careful: for one-sided tests, order that you put x and y matters!

NOTE: must use **“two.sided”** option if you want to calculate confidence intervals. Confidence intervals from other options are not correct.

Two-sample Test: an example

In order to assess how quickly polyester decays over time in landfills, a researcher buried strips of the material in the soil for different lengths of time and then tested the force required to break them (as a measure of decay). Lower breaking strength is indicative of decay. Here is the data collected:

Group	Sample Data				
2 weeks	118	126	126	120	129
16 weeks	124	98	110	140	110

Test whether or not the breaking strengths of polyester strips buried for 2 weeks is greater than the breaking strengths of those buried for 16 weeks.

Perform the test at the $\alpha=0.10$ level of significance.

Two-sample Test: an example using R

```
> decay = read.csv("decay.csv")
> summary(decay$strength[decay$weeks==2])
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  118.0   120.0   126.0   123.8   126.0   129.0
> summary(decay$strength[decay$weeks==16])
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   98.0   110.0   110.0   116.4   124.0   140.0
```

1. Set up the hypotheses and select the alpha level

- H_0: $\mu_1=\mu_2$ (the mean breaking strengths are the same)
- H_1: $\mu_1>\mu_2$ (the mean breaking strengths are greater after 2 weeks versus 16 weeks)
- $\alpha=0.10$

2. Select the appropriate test statistic

Use t-test as n is small and population sd id unknown

Two-sample Test: an example using R

3. State the decision rule
- Decision Rule: Reject H_0 if $p \leq \alpha$
- Otherwise, do not reject H_0

4. Compute the test t statistic and the associated p-value

> **t.test(decay$strength[decay$weeks==2], decay$strength[decay$weeks==16], alternative="greater", conf.level=0.9)**

t = 0.9889, df = 4.651, p-value = 0.1857

5. Conclusion
- **Fail to reject H_0 since p-value is greater than α**
- We do not have significant evidence at the α=0.10 level that the mean breaking strength is greater for polyester left to decay for 2 weeks as opposed to 16 weeks.
- The difference in sample means was 7.4 indicating a higher breaking strength for polyester strips left to decay 2 weeks as opposed to those left to decay for 16 weeks.
- **We do not reject the null hypothesis** that the mean breaking strength is the same between the two groups.

Two-sample two-sided Test: an example using R

```
> t.test(decay$strength[decay$weeks==2],
decay$strength[decay$weeks==16], alternative="two.sided",
conf.level=0.95)

t = 0.9889, df = 4.651, p-value = 0.3713
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.2789  27.0789
```

Something to know

Though the theory underlying two-sample tests described above is based on the assumption that the samples are taken from populations that are normally distributed, the methodology is actually quite robust to departures from this assumption.

This is particularly true when the sample sizes n_1 and n_2 are equal to each other. In this case, as long as the underlying distributions have similar shapes and the sample sizes are both ≥ 5, the approximation using the above methodology is quite good.

Larger sample sizes are needed when the distributions have different shapes. This procedures lack of sensitivity to skewness and departures of normality make it so widely used in statistics.

z-tests can be used in the setting where n_1 and n_2 are large (both ≥ 30).
In practice, they are rarely used given the **robustness of the t-tests to non-normality.**

R commands: numerical summaries for two-sample tests

Use **aggregate** or **tapply** function to summarize data by population

> **aggregate**(data$variable, by=list(data$population), FUN=[function])

> **tapply**(data$variable, **data$group**, **mean**)

Calculate the mean but group the date based on their group

Can be used for **any function** (summary, sd)

R commands: graphical summaries for two-sample tests

Side by side box plots
> **boxplot**(data$variable~data$group)

Bar graph with confidence intervals
> **install.packages**("**gplots**")
> **attach**(data)

> means <- **tapply**(variable, group, mean)
> lower <- **tapply**(variable, group, function(v) t.test(v)$conf.int[1])
> upper <- **tapply**(variable, group, function(v) t.test(v)$conf.int[2])

> **barplot2**(means, plot.ci=TRUE, ci.l=lower, ci.u=upper, names.arg=[labels])

ci.l = lower bound, ci.u = upper bound , are boundaries of our plot.

R commands: the polyester example

> **attach**(decay)

Side by side box plots
> boxplot(strength~weeks)

Bar graph with confidence intervals
> means <- tapply(strength, weeks, mean)
> lower <- tapply(strength, weeks, function(v) t.test(v)$conf.int[1])
> upper <- tapply(strength, weeks, function(v) t.test(v)$conf.int[2])

> barplot2(means, plot.ci=TRUE, ci.l=lower, ci.u=upper, names.arg=c(“2 weeks”, “16 weeks”))
> abline(h=0)

R Examples on Github

- **Example R Code: White sharks average**
  - https://github.com/kiat/R-Examples/blob/master/R-Example-Program-3.1.R

- **Example R Code: Polyester Data, Two Sample t-test**
  - https://github.com/kiat/R-Examples/blob/master/R-Example-Program-3.R

- **Example R Code: Bar Plot with lower and upper levels**
  - https://github.com/kiat/R-Examples/blob/master/R-Example-Program-3.3.R