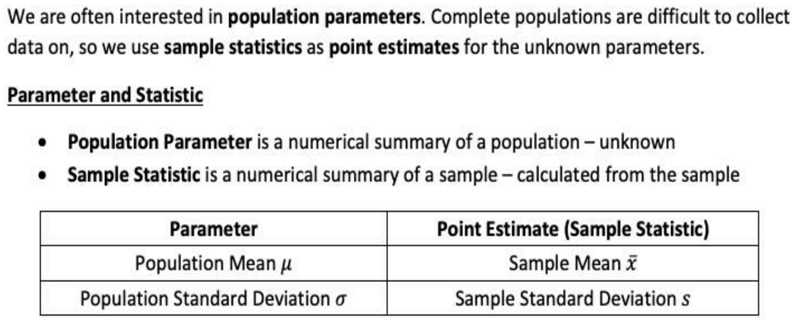
CHAPTER 7

Inference for Numerical Data

7.1 One Sample Means with t-Distribution

The Sampling Distribution of $\bar{x}$



We are often interested in **population parameters**. Complete populations are difficult to collect data on, so we use **sample statistics** as **point estimates** for the unknown parameters.

Parameter and Statistic

- **Population Parameter** is a numerical summary of a population – unknown
- **Sample Statistic** is a numerical summary of a sample – calculated from the sample

Parameter	**Point Estimate (Sample Statistic)**
Population Mean μ	Sample Mean $\bar{x}$
Population Standard Deviation σ	Sample Standard Deviation s

Central Limit Theorem for the Sample Mean:

When we collect a sufficiently large sample of n independent observations from a population with mean μ and standard deviation σ, the sampling distribution of $\bar{x}$ will be nearly normal with

$$Mean = \mu$$

$$Standard\ Error(SE) = \frac{\sigma}{\sqrt{n}}$$

Evaluating the Two Conditions required for modeling $\bar{x}$

Two conditions are required to apply the central Limit Theorem for the sample mean $\bar{x}$:

Independence. The sample observations must be independent. The most common way to satisfy this condition is when the sample is a simple random sample from the population.

Normality. When a sample is small, we also require that the sample observations come from a normally distributed population. We can relax this condition more and more for larger samples.

Central Limit Theorem (C.L.T.)

Sample means will be approximately normally distributed with mean equal to the population mean μ, and standard error equal to $\frac{\sigma}{\sqrt{n}}$. That is,

$$\bar{X} \sim \text{approx.}\, N\left(\mu,\ \frac{\sigma}{\sqrt{n}}\right) \quad \text{or} \quad z = \frac{\bar{x} - \text{mean}}{\text{se}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim \text{approx.}\, N(0, 1)$$

Note: Sample means ($\bar{X}$) will be approximately normally distributed when the sample size is large regardless of the shape of the underlying distribution the data come from.

Parameter	Point Estimate	Sampling Dist. (C.L.T.)	Confidence Interval
p	$\hat{p}$	$\hat{p} \sim \text{approx.}\, N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$	
μ	$\bar{x}$	$\bar{x} \sim \text{approx.}\, N\left(\mu,\ \frac{\sigma}{\sqrt{n}}\right)$	

Note: This confidence interval formula requires the value of ______________________________, and unfortunately we almost always do not know this value.

How to Perform the Normality Check

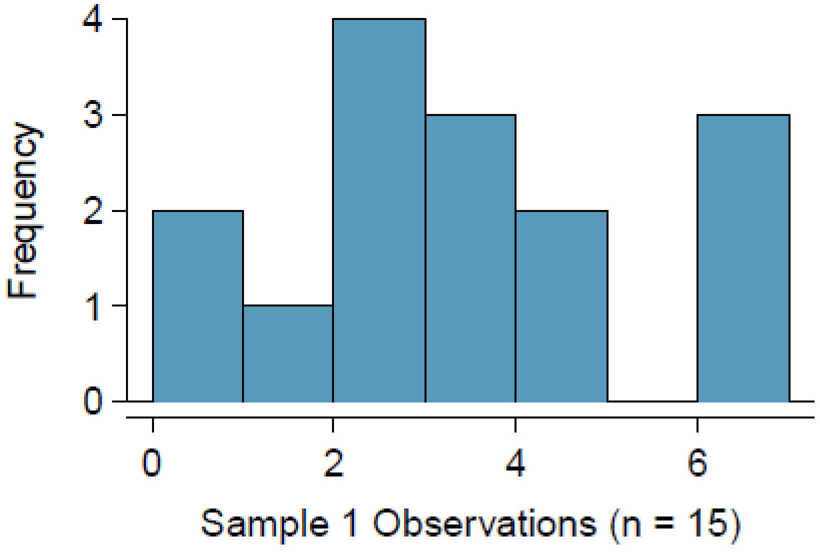
Two Rules of thumb:

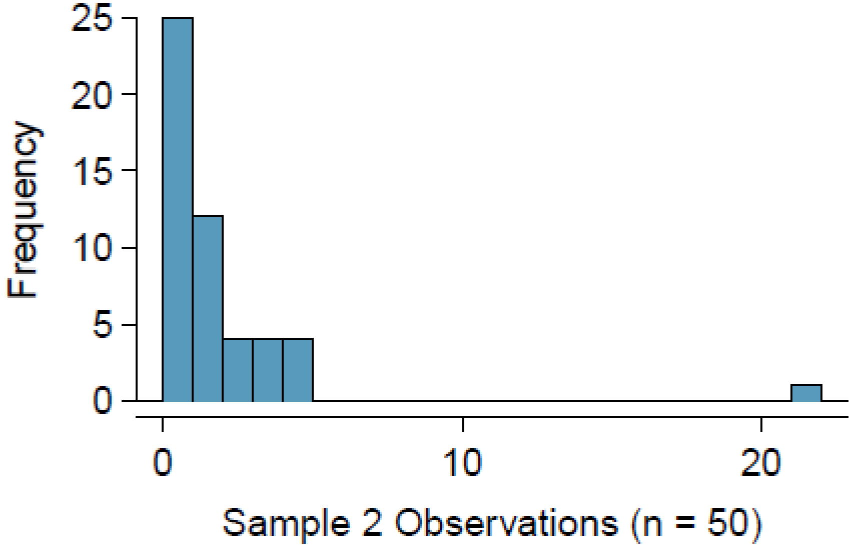
$\boldsymbol{n < 30}$: if the sample size n is less than 30 and there are no clear outlier in the data, then we typically assume the data come from nearly normal distribution to satisfy the condition.

$\boldsymbol{n \geq 30}$: if the sample size n is at least 30 and there are no particularly extreme outliers, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

Example

Consider the following two plots that come from simple random samples from different populations. Their sample sizes are $n_1 = 15$ *and* $n_2 = 50$. Are the independence and normality conditions met in each case?





Introducing the t-distribution

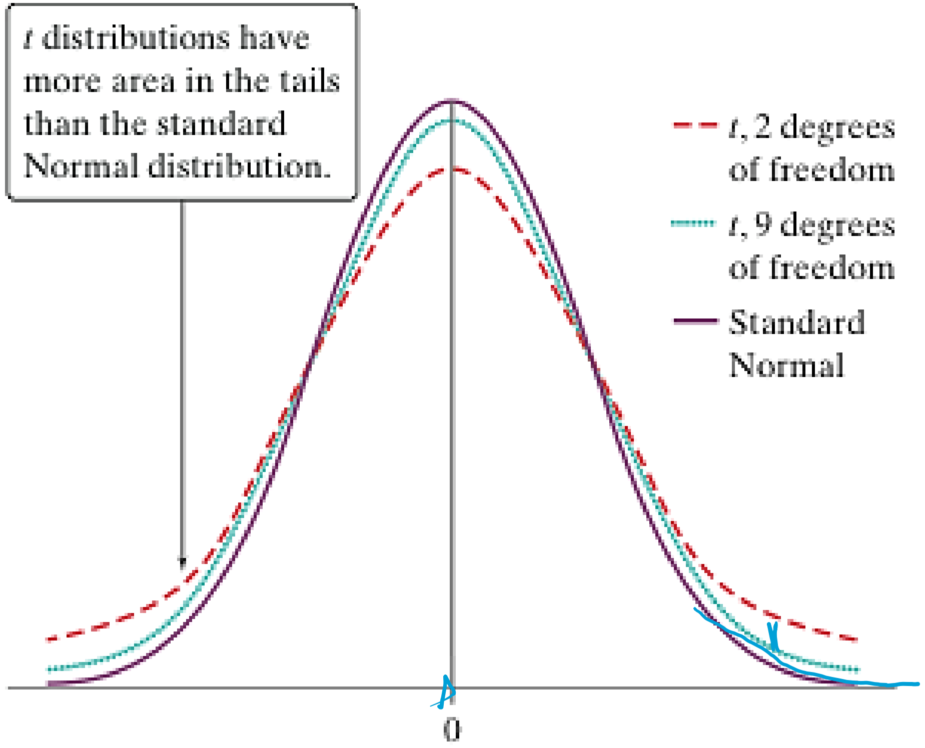
- In practice, we cannot directly calculate the standard error for $\bar{x}$ since we do not know the population standard deviation, σ.
- We replace $\frac{\sigma}{\sqrt{\text{n}}}$ (the standard deviation of the sample mean) by what is called the standard error of the sample mean:

$$SE = \frac{s}{\sqrt{n}}$$

- When we replace σ with s, we move away from the standard normal sampling distribution to a new family of sampling distributions called the ***t-distribution***

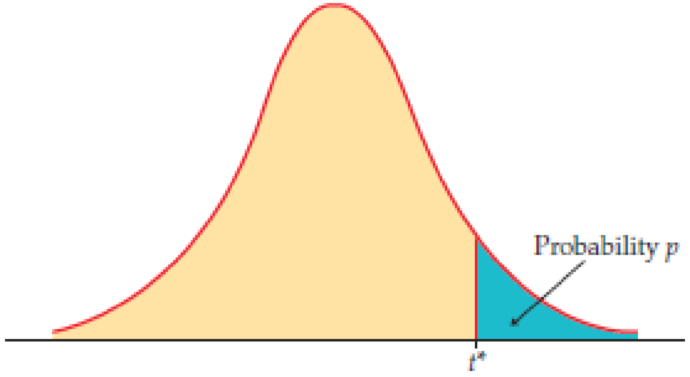
The t-distributions

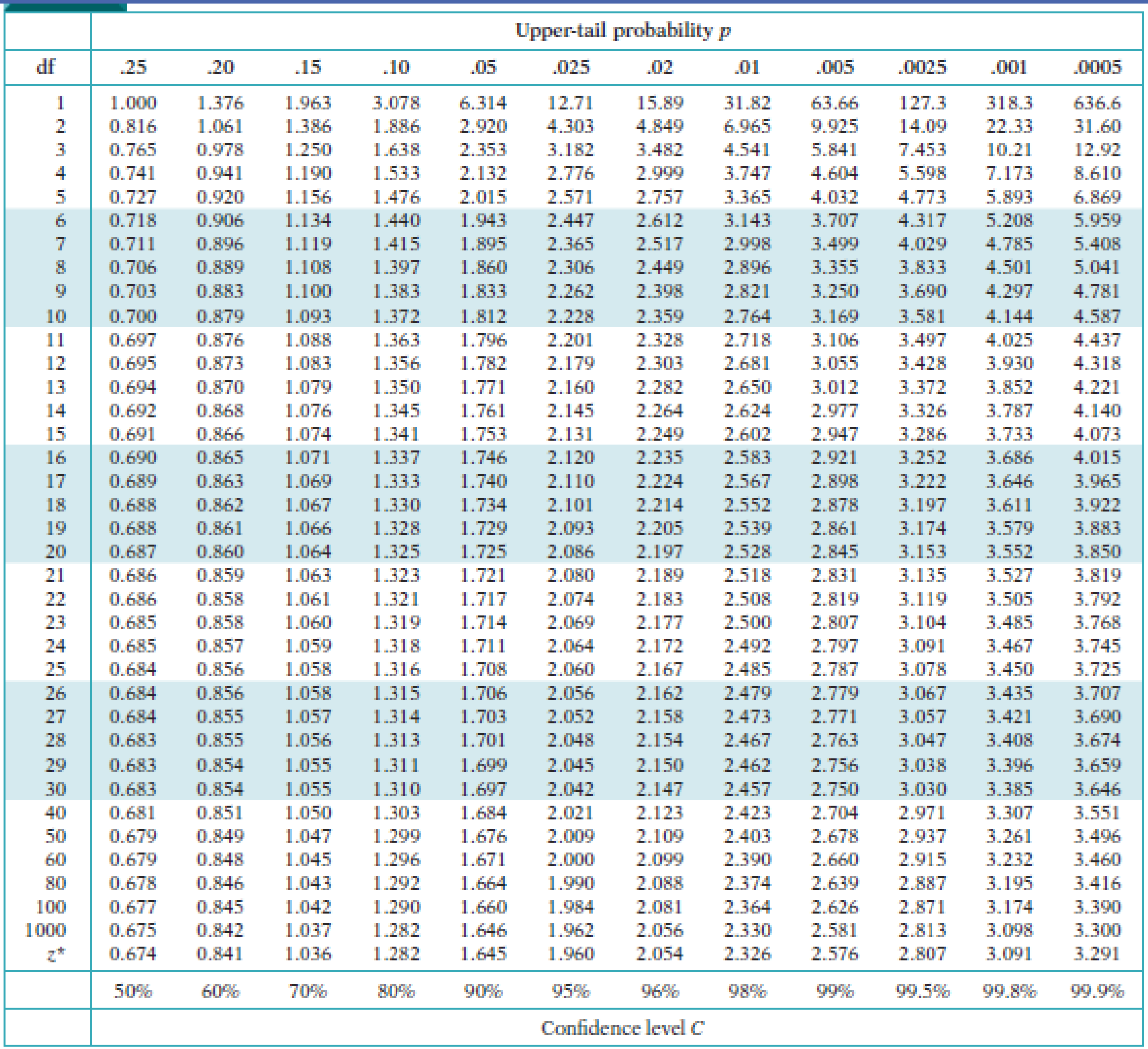
Comparing the density curves of the standard Normal distribution and *t* distributions:



- ✓The density curves of the *t* distributions are similar in shape to the standard Normal curve.
- ✓The spread of the *t* distributions is a bit larger than that of the standard Normal distribution.
- ✓The *t* distributions have more probability in the tails and less in the center than does the standard Normal.
- ✓The *t* distributions is always centered at zero.
- ✓As the degrees of freedom increase, the *t* density curve becomes ever closer to the standard Normal curve.

t-distribution Table





	Upper-tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
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	3.611	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	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

$$\text{Sampling Distribution: } t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\bar{x} - \text{mean}}{\text{estimated se}} \sim \ T \text{ with } df = n - 1$$

		When we know the value σ	When we don't know the value σ
Parameter	Point Estimate	$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$	$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim T \text{ with } df = n - 1$
μ	$\bar{x}$	$\bar{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}}$	

7.1.4 One Sample T Confidence Intervals

One-Sample T Confidence Interval for Mean

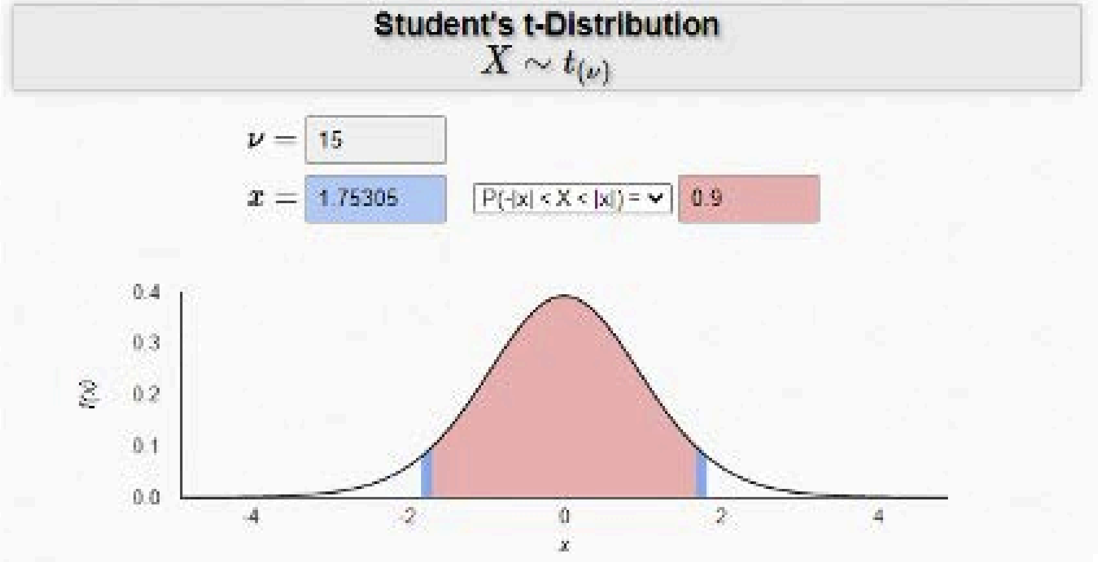
with $df =$ (point estimate ± maring of error)

One-Sample T Confidence Interval for Mean

with $df =$ (point estimate $\pm$ maring of error)

Finding the Critical Value t^* from a T distribution

1) T Distribution Applet (https://homepage.divms.uiowa.edu/~mbognar/applets/t.html)



Student's t-Distribution

$X \sim t_{(\nu)}$

$\nu =$ 15

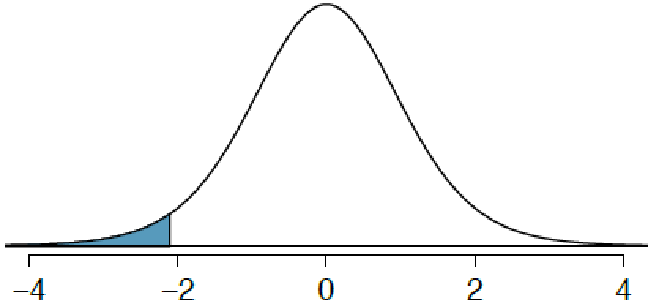
$x =$ 1.75305 P(-|x| < X < |x|) = 0.9

For a 90% Confidence Level with $n = 16$, degrees of freedom: $\nu = 15$

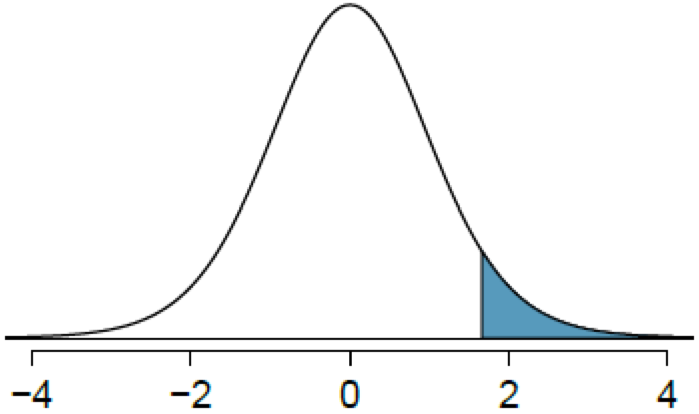
Critical Value: $t^* = 1.75305$

Example

What proportion of the t-distribution with 18 degrees of freedom falls below -2.1?



What proportion of the t-distribution with 20 degrees of freedom falls above 1.65?



One sample t-distribution

The *t* Distributions: Degrees of Freedom

Draw an SRS of size *n* from a large population that has a Normal distribution with mean *μ* and standard deviation *σ*. The **one-sample *t* statistic**

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

has the ***t* distribution** with **degrees of freedom** *df* = *n* – 1.

Example

You randomly choose 16 unfurnished one-bedroom apartments from a large number of advertisements in your local newspaper. You calculate that their mean monthly rent is \$766, and their standard deviation is \$180

a) What is the standard error of the mean

b) What are the degrees of freedom for a one-sample t statistic?

One sample t-distribution Confidence Interval

The One-Sample *t* Interval for a Population Mean

Choose an SRS of size *n* from a population having unknown mean μ. A level *C* **confidence interval** for μ is:

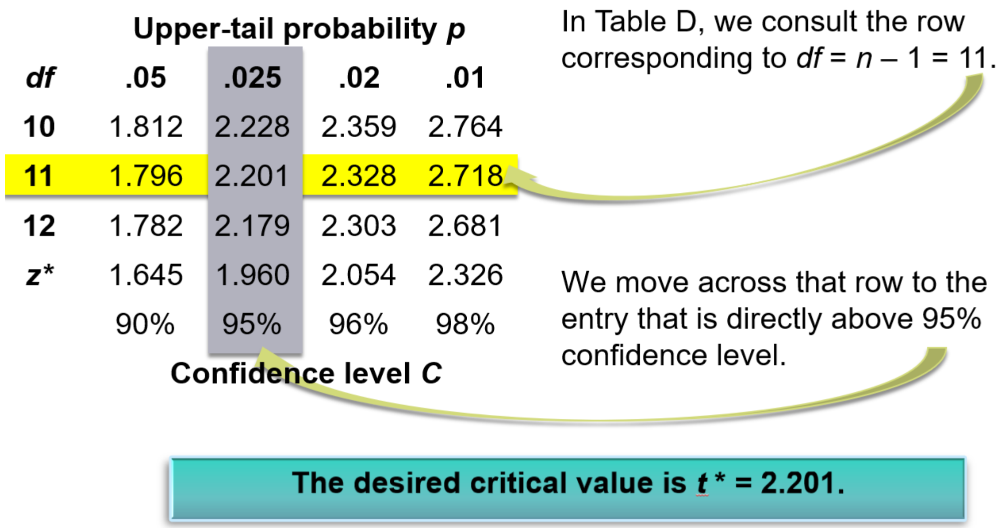
$$\bar{x} \pm t^*_{df} \frac{s}{\sqrt{n}}$$

where *t** is the critical value for the *t*(*n* – 1) distribution.

Example

The Nielson Company is a global information and media company and one of the leading suppliers of media information. In their state-of-the-media report, they announced that U.S. cell phone subscribers average 5.4 hours per month watching videos on their phones. Does this average seem reasonable? We draw the following SRS of size 12 from this population:

11.9, 2.8, 3.0, 6.2, 4.7, 9.8, 11.1, 7.8, 4.0, 5.5, 7.1, 10.2



	Upper-tail probability *p*			
df	**.05**	**.025**	**.02**	**.01**
10	1.812	2.228	2.359	2.764
11	1.796	2.201	2.328	2.718
12	1.782	2.179	2.303	2.681
z*	1.645	1.960	2.054	2.326
	90%	95%	96%	98%
	Confidence level *C*			

In Table D, we consult the row corresponding to *df* = *n* – 1 = 11.

We move across that row to the entry that is directly above 95% confidence level.

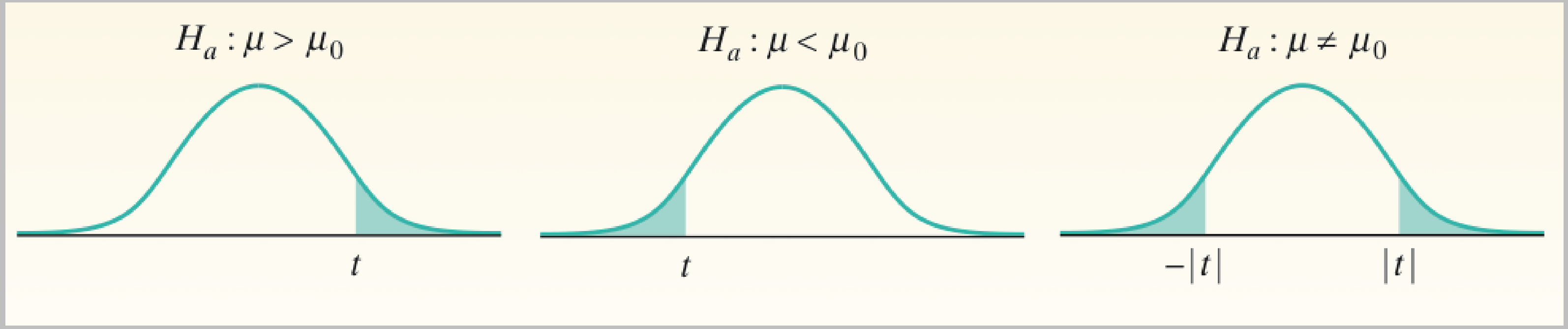
The desired critical value is *t = 2.201.**

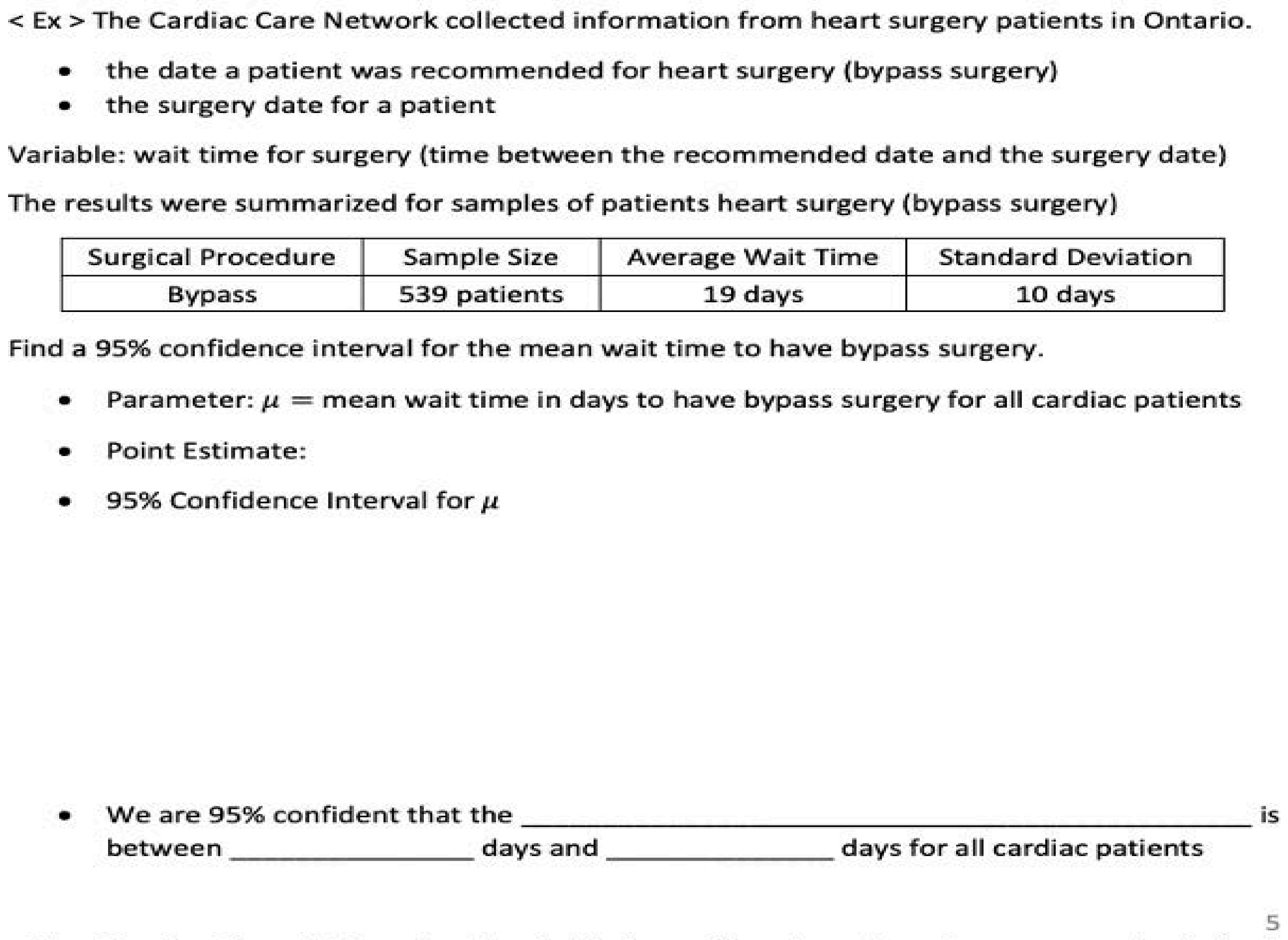
Hypothesis Testing for a Single Mean

Choose an SRS of size *n* from a large population that contains an unknown mean μ. To test the hypothesis $H_0 : \mu = \mu_0$, compute the one-sample *t* statistic:

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

Find the *P*-value by calculating the probability (at degrees of freedom = $n - 1$) of getting a *t* statistic this large or larger *in the direction specified by the alternative hypothesis* H_{a}.





< Ex > The Cardiac Care Network collected information from heart surgery patients in Ontario.

- the date a patient was recommended for heart surgery (bypass surgery)
- the surgery date for a patient

Variable: wait time for surgery (time between the recommended date and the surgery date)

The results were summarized for samples of patients heart surgery (bypass surgery)

Surgical Procedure	Sample Size	Average Wait Time	Standard Deviation
Bypass	539 patients	19 days	10 days

Find a 95% confidence interval for the mean wait time to have bypass surgery.

- Parameter: μ = mean wait time in days to have bypass surgery for all cardiac patients
- Point Estimate:
- 95% Confidence Interval for μ

- We are 95% confident that the ____ is between ______________ days and _____________ days for all cardiac patients

< Ex > The Cardiac Care Network collected information from heart surgery patients in Ontario.

< Ex > A growing concern of employers is time spent in activities like surfing the Internet and e-mailing friends during work hours. The study summarized the findings from a sample of workers in an article “Who goofs off 2 hours a day? Most workers, Survey says.”

The CEO of a company thinks that the average wasted time for their employees is less than the reported 120 minutes and wants to determine it using a hypothesis test.

A sample of ten employees was randomly selected from the company. Each employee is asked about daily wasted time in minutes at work. They would probably have to be guaranteed anonymity to obtain truthful responses. The data are shown below.

108, 112, 117, 130, 111, 131, 113, 113, 105, 128

Do the data provide convincing evidence that the mean wasted time for the employees in this company is shorter than 120 minutes? Conduct a hypothesis test at a 5% significance level.

(a) Let μ = the mean wasted time in minutes during a work day for their employees. State the hypotheses using this parameter μ.

(b) Calculate the t test statistic.

(c) Calculate the p-value.

Practice Problem

Officials in charge of televising an international chess competition in South America want to determine if the average time per move for the top players has remained under five minutes over the last two years. Video tapes of matches which have been played over the two years period are reviewed and a random sample of 31 moves are timed. The sample mean is 4.5 minutes with a standard deviation of 1.15 minutes. Can the officials conclude at $\alpha = 0.05$ that the time per move is still under 5 minutes?

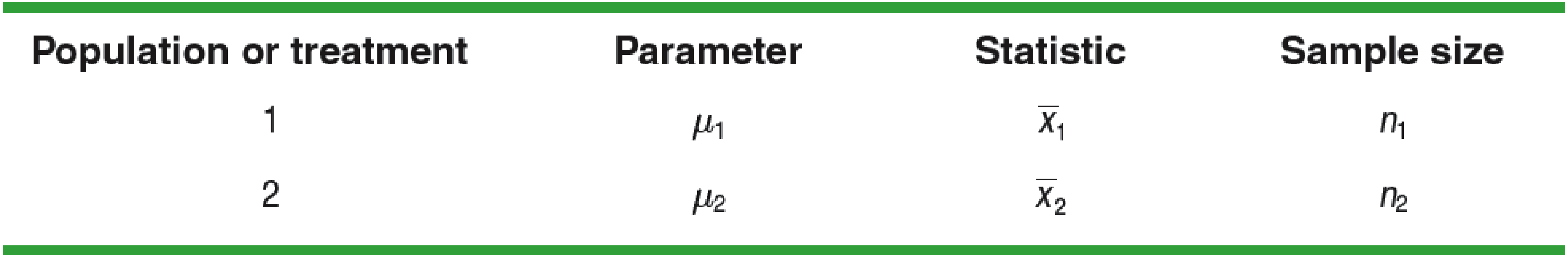
7.3 Difference of Two Means

Two-Sample Problems

What if we want to compare the means of two populations, Population 1 (μ_1) and Population 2 (μ_2)?

The best approach is to take separate random samples from each population and to compare the sample means.

In this case, the parameters μ_1 and μ_2 are the true mean responses for Treatment 1 and Treatment 2, respectively.



Population or treatment	Parameter	Statistic	Sample size
1	μ_1	$\bar{x}_1$	n_1
2	μ_2	$\bar{x}_2$	n_2

The Two-Sample *t* Statistic

When data come from two random samples the statistic $\bar{x}_1 - \bar{x}_2$ is our best guess for the value of $\mu_1 - \mu_2$.

When the two samples are independent of each other, the standard deviation of the statistic $\bar{x}_1 - \bar{x}_2$ is:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Since we don't know the values of the parameters $\sigma_1 and\ \sigma_2$, we replace them in the standard deviation formula with the sample standard deviations.

The result is the **standard error** of the statistic $\bar{x}_1 - \bar{x}_2$: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

The Two-Sample *t* Statistic

We standardize the observed difference to obtain a *t* statistic that tells us how far the observed difference is from its mean in standard deviation units:

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

The two-sample *t* statistic has approximately a *t* distribution. We can use technology to determine degrees of freedom OR we can use a conservative approach, using the smaller of $n_1 - 1$ and $n_2 - 1$ for the degrees of freedom.

Confidence Interval for $\mu_1 - \mu_2$

Two-Sample *t* Interval for a Difference Between Means

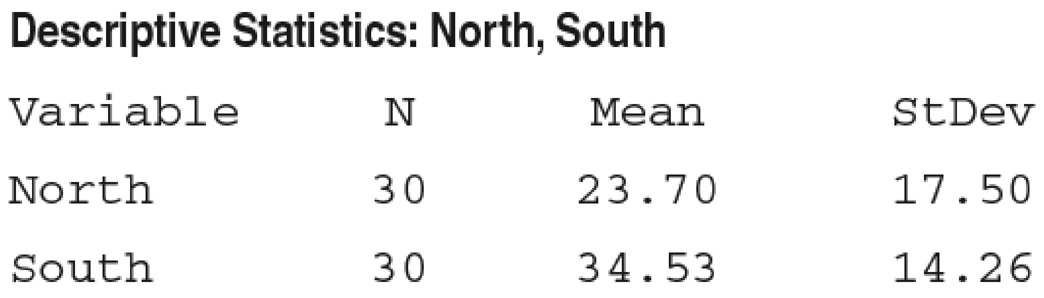
When the Random, Normal, and Independent conditions are met, a level C confidence interval for $(\mu_1 - \mu_2)$ is

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

where t^* is the critical value at confidence level C for the t distribution with degrees of freedom either gotten from technology or equal to the smaller of $n_1 - 1$ and $n_2 - 1$.

Example

“How do the sizes of longleaf pine trees in the northern and southern halves of the forest compare?” To find out, researchers took random samples of 30 trees from each half and measured the diameter at breast height (DBH) in centimeters. Construct and interpret a 90% confidence interval for the difference in the mean DBH for longleaf pines in the northern and southern halves of the Wade Tract Preserve.



Descriptive Statistics: North, South

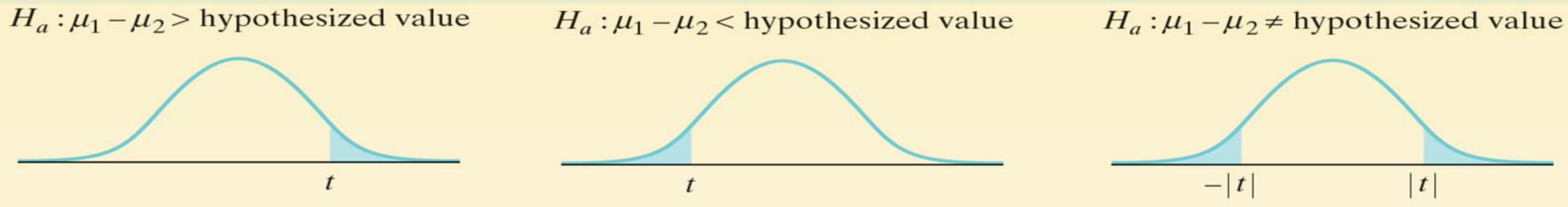
```
Variable     N      Mean      StDev
North        30     23.70     17.50
South        30     34.53     14.26
```

Two-Sample t-test

Suppose the Random Normal, and independent conditions are met. To test the hypothesis $H_0\colon \mu_1 - \mu_2 = 0$, compute the t-statistic

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

Find the p-value by calculating the probability of getting a t-statistic this large or larger in the direction specified by the alternative hypothesisH_a. Use the t-distribution with degree of freedom of the smaller of $n_1 - 1\ and\ n_2 - 1$.



Example

Do different brands of cookies have the same number of chips on average?

Brand 1 = Pantry Essential with $\bar{x}_1 = 18.65, s_1 = 4.13 \; and \; n_1 = 147.$

Brand 2 = Chewy Chips Ahoy with $\bar{x}_2 = 20.09, s_2 = 4.09 \; and \; n_2 = 160.$

1) Find the 95% confidence interval for the difference of population means

Example Cont.

2) State the appropriate $H_0\ and\ H_a$ and carry out the test using a significance level of $\alpha = 0.05$. Give the p-value and interpret the result.

Pooled Two-Sample Procedures

There are two versions of the two-sample *t*-test:

1: Assuming NOT equal variance

- Unequal variance two-sample t-test (what we have been doing)

2: Assuming equal variances

- Pooled two-sample t-test

Pooled Two-Sample Procedures

Suppose both populations are Normal and they have *the* ***same standard deviations***. The **pooled estimator of σ^2** is

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

A level C confidence interval for $\mu_1 - \mu_2$ is

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where the degrees of freedom for t^* are $\boldsymbol{n_1 + n_2 - 2}$**.**

To test the hypothesis H_0: $\mu_1 = \mu_2$ against a one-sided or a two-sided alternative, compute

the pooled two-sample t statistic for the $t(n_1 + n_2 - 2)$ distribution.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Example

Tom Sealack, a supply clerk with the Navy, has been asked to determine if a new battery that has been offered to the Navy (at a reduced price) has a shorter average life than the battery they are currently using. He randomly selects batteries of each type and allows them to run continuously so that he can measure the time until failure for each battery. He randomly selected 35 of the new style of batteries and measured an average life of 700 hours from each battery with a standard deviation of 30. He also randomly selected 35 of the old type of batteries and measured an average life of 710 hours with a standard deviation of 35. Assume that the population standard deviation are equal. Does the data suggest at $\alpha = 0.10$ that the time until failure for the new battery is significantly less than the time until failure for the old battery?

7.4 Power

Type *I and II* Errors

When we draw a conclusion from a significance test, we hope our conclusion will be correct. But sometimes it will be wrong. There are two types of mistakes we can make.

If we reject H_0 when H_0 is true, we have committed a **Type I error.**

If we fail to reject H_0 when H_0 is false, we have committed a **Type II error.**

Type *I and II* Errors

Type *I* error:

It can occur only when the null hypothesis is rejected.

$$p(type\ I\ error) = \alpha$$

Type *II* error:

it occurs when the null hypothesis is false but is not rejected.

$$p(type\ II\ error) = \beta$$

		Truth about the population	
		H_0 true	H_0 false (H_a true)
Conclusion based on sample	Reject H_0	**Type I error (α)**	*Correct conclusion*
	Fail to reject H_0	*Correct conclusion*	**Type II error (β)**

Power

The power of a test is the probability of rejecting $H_0$when H_0 is false (the alternative is true)

$$\text{Power} = 1-\beta$$

Increasing the power

1) Increase α
2) Consider a particular alternative that is farther away from μ_0
3) Increase the sample size.
4) Decrease σ

Calculating the Power of a test

1. State the null and alternative hypotheses, $H_0 \ and \ H_a$
2. Find the critical value (the value(s) for which an area of α is in the tails), and then work backwards from the standardization formula to find the value(s) of the original scale.
3. Calculate the probability of getting a test statistic that falls beyond the value(S) you find in step 2, under the assumption that H_a is true. This is the power of the test.

Example 1

You want to see if a redesign cover of a mail-order catalog will increase sales. A very large number of customers will receive the original catalog, and a random sample of customers will receive the one with the new cover. For planning purposes, you are willing to assume that the sales from the new catalog will be approximately Normal with $\sigma = 50$ dollars and that the mean for the original catalog will be $\mu = 25$ dollars. You decide to use the sample size of n=100. Find the power of the significance test with $\alpha = 0.01$under the alternative that $\mu = 30$.

7.5 Comparing Many Means with ANOVA

ANOVA

- To compare Means of 2 group we use a Z or a T statistic.
- To compare means of 3^+ groups we use a new test called *ANOVA* and a new statistic called *F.*
- ANOVA is used to assess whether the mean of the outcome variable is different for different levels of a categorical variable

Conditions for ANOVA

Generally we must check three conditions on the data before performing ANOVA

1. The observations are independent within and across groups.
2. The data within each group are nearly normal .
3. The variability across group is about equal.

Analysis of Variance (ANOVA)

Step 1: The null and the alternative Hypotheses

H_0: The mean outcome is the same across all categories,

$$\mu_1 = \mu_2 = \ldots = \mu_\mathrm{k},$$

H_a: At least one mean is different than others

Step 2: Test Statistic- F-test

Analysis of variance (ANOVA) is used to test whether the mean outcome differs across 2 or more groups. ANOVA uses a test statistic F, which represent a standardized ratio of variability in the sample means relative to the variability within the group. Compute a test statistic (a ratio)

$$F = \frac{variability\ between\ groups}{variability\ within\ groups} = \frac{MSG}{MSE}$$

Degree of freedom associated with ANOVA

groups: $df_G = k - 1$, where k is the number of groups

total: $df_T = n - 1$, where n is the total sample size

error: $df_E = df_T - df_G$

Step 3: P-value for ANOVA

p-value is the probability of at least as large a ratio between the “between group” and “within group” variability, if in fact the means of all groups are equal. It’s calculated as the area under the F curve, with degrees of freedom df_G and df_E, above the observed F statistic.

Step 4 : Making a decision and conclusion

If the p-value is small enough H_0 is rejected, we conclude that the population means are not equal

Reading ANOVA Table from Software

	Df	Sum of Square SS	Mean Square MS	F-Value	P-Value
Groups	$k-1$	SSG	$MSG = \frac{SSG}{k-1}$	$\frac{MSG}{MSE}$	Tail area above F
Error	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$		
Total	$n-1$	SST			

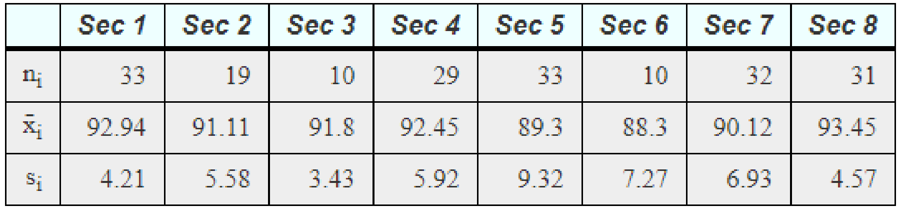
$$SSG = \sum_{i=1}^{n} n_i(\bar{x_i} - \bar{x})^2$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_i\right)^2$$

$$SST = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

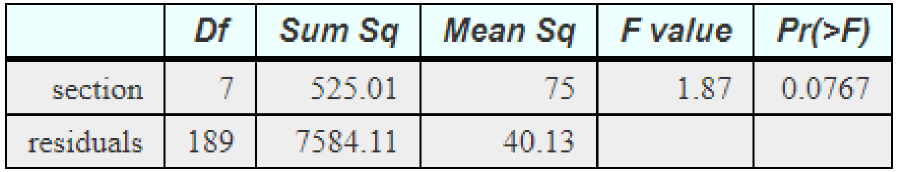
Example 1

A professor who teaches a large introductory statistics class (197 students) with eight discussion sections would like to test if student performance differs by discussion section, where each discussion section has a different teaching assistant. The summary table below shows the average final exam score for each discussion section as well as the standard deviation of scores and the number of students in each section.



	Sec 1	Sec 2	Sec 3	Sec 4	Sec 5	Sec 6	Sec 7	Sec 8
n_i	33	19	10	29	33	10	32	31
$\bar{x}_i$	92.94	91.11	91.8	92.45	89.3	88.3	90.12	93.45
s_i	4.21	5.58	3.43	5.92	9.32	7.27	6.93	4.57

The ANOVA output below can be used to test for differences between the average scores from the different discussion sections.



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
section	7	525.01	75	1.87	0.0767
residuals	189	7584.11	40.13		

Assume that the conditions required for this inference are satisfied. Conduct a hypothesis test to determine if these data provide convincing evidence that the average score varies across some (or all) groups.

Example 1 Cont.

1) Write the hypotheses for evaluating if there is a difference between average scores across sections in your own words.

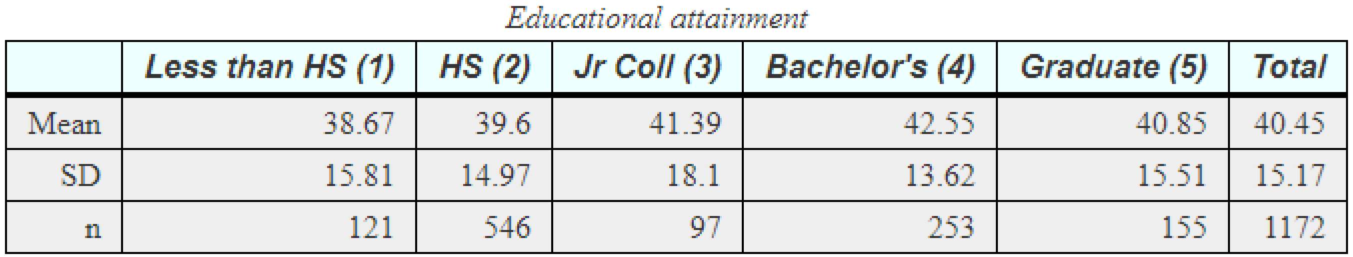
2) What is the test statistic associated with this ANOVA test?

3) What is the p-value associated with this ANOVA test?

4) Interpret the conclusion of the test in the context of the study.

Example 2

The General Social Survey collects data on demographics, education, and work, among many other characteristics of US residents. Using ANOVA, we can consider educational attainment levels for all 1,172 respondents at once. Below are the distributions of hours worked by educational attainment and relevant summary statistics that will be helpful in carrying out this analysis.



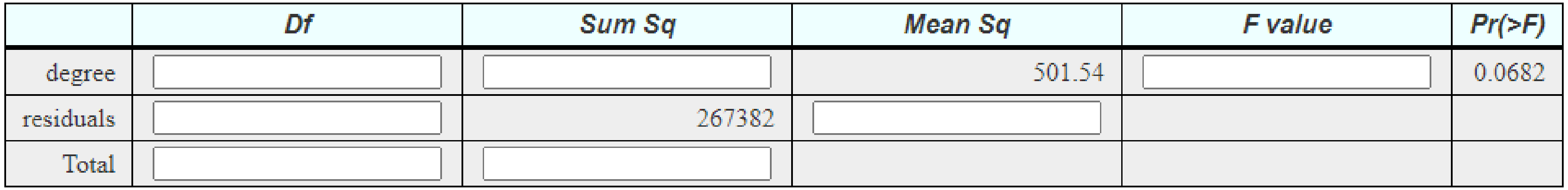
Educational attainment

	Less than HS (1)	HS (2)	Jr Coll (3)	Bachelor's (4)	Graduate (5)	Total
Mean	38.67	39.6	41.39	42.55	40.85	40.45
SD	15.81	14.97	18.1	13.62	15.51	15.17
n	121	546	97	253	155	1172

1) Write hypotheses for evaluating whether the average number of hours worked varies across the five groups.

Example 2 Cont.

2) Below is part of the output associated with this test. Fill in the empty cells *(please be accurate to two decimal places)*.



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
degree			501.54		0.0682
residuals		267382			
Total					

3) What is the value of the test statistic associated with this ANOVA test?

4) What is the p-value associated with this ANOVA test?

5) Interpret the conclusion of the test in the context of the study.