# MSIT 431 Probability and Statistical Methods

Chapter 7 Inference for Distributions

Dongning Guo

Fall 2017

## Chapter 7 Inference for Distributions

**7.1 Inference for the Mean of a Population**

**7.2 Comparing Two Means**

**7.3 Other Topics in Comparing Distributions**

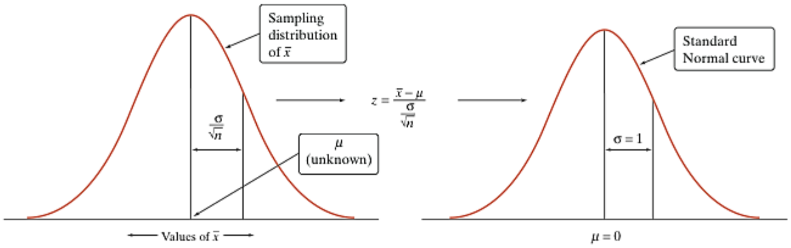
# 7.1 Inference for the Mean of a Population

- The *t* distributions
- One-sample *t* confidence interval
- One-sample *t* test
- Matched pairs *t* procedures
- Robustness of the *t* procedures

## The *t* Distributions

When the sampling distribution of $\overline{x}$ is close to Normal, we can find probabilities involving $\overline{x}$ by standardizing:

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$


Sampling distribution of $\bar{x}$
$\frac{\sigma}{\sqrt{n}}$
$\mu$ (unknown)
Values of $\bar{x}$
$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$
Standard Normal curve
$\sigma = 1$
$\mu = 0$

When we don't know *σ*, we can estimate it using the sample standard deviation $s_x$. What happens when we standardize?

$$?? = \frac{\overline{x} - \mu}{s_x/\sqrt{n}}$$

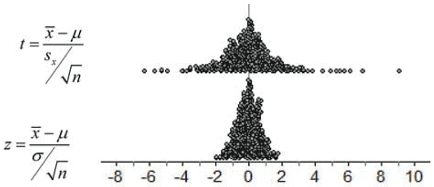
**This new statistic does *not* have a Normal distribution!**

# The *t* Distributions

When we standardize based on the sample standard deviation $s_x$, our statistic has a new distribution called a ***t* distribution.**

The *t* distribution has a shape similar to that of the standard Normal curve in that *it is symmetric with a single peak at 0.*

However, it has *more area in the tails*.



Like any standardized statistic, $t$ tells us how far $\bar{x}$ is from its mean $\mu$ in standard deviation units.

There is a different *t* distribution for each sample size, specified by its **degrees of freedom (*df*).**

# The *t* distributions

**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**

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

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

The density function (df=v)

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

If v is odd,

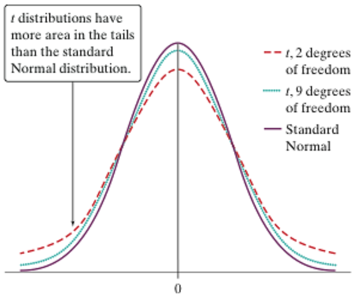
$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} = \frac{(\nu-1)(\nu-3)\cdots5\cdot3}{2\sqrt{\nu}(\nu-2)(\nu-4)\cdots4\cdot2}.$$

If v is even,

$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} = \frac{(\nu-1)(\nu-3)\cdots4\cdot2}{\pi\sqrt{\nu}(\nu-2)(\nu-4)\cdots5\cdot3}.$$

# The *t* Distributions

When comparing the density curves of the standard Normal distribution and *t* distributions, several facts are apparent:



- ✓ 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.
- ✓ As the degrees of freedom increase, the *t* density curve becomes ever closer to the standard Normal curve.

We can use Table D in the back of the book to determine critical values *t** for *t* distributions with different degrees of freedom.

# Using Table D

Suppose you want to construct a 95% confidence interval for the mean *μ* of a Normal population based on an SRS of size *n* = 12. What critical *t** should you use?

| | **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.**

# One-Sample *t* Confidence Interval

The **one-sample *t* interval for a population mean** is similar in both reasoning and computational detail to the one-sample *z* interval for a population mean.

**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:

$$\bar{x} \pm t^* \frac{s_x}{\sqrt{n}}$$

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

The **margin of error** is:

$$t^* \frac{s_x}{\sqrt{n}}$$

This interval is exact when the population distribution is Normal and approximately correct for large *n* in other cases.

# Example

A manufacturer of high-resolution video terminals must control the tension on the mesh of fine wires that lies behind the surface of the viewing screen. The tension is measured by an electrical device with output readings in millivolts (mV). A random sample of 20 screens has the following mean and standard deviation:

$$\bar{x} = 306.32 \text{ mV} \qquad \text{and} \qquad s_x = 36.21 \text{ mV}$$

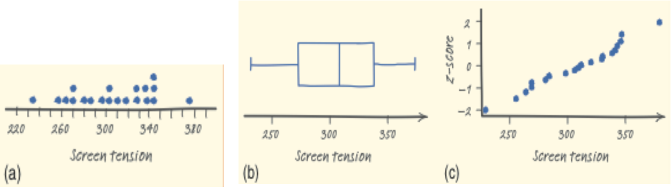
We want to estimate the true mean tension *μ* of all the video terminals produced this day at a 90% confidence level.

# Example

If the conditions are met, we can use a one-sample *t* interval to estimate *μ*.

✓ **Random:** We are told that the data come from a random sample of 20 screens from the population of all screens produced that day.

✓ **Normal:** Since the sample size is small ($n < 30$), we must check whether it is reasonable to believe that the population distribution is Normal. Examine the distribution of the sample data.



These graphs suggest approximate Normality of the population.

✓ **Independent:** Because we are sampling without replacement, we must assume that at least 20(20) = 400 video terminals were produced this day.

# Example

We are told that the mean and standard deviation of the 20 screens in the sample are

$$\bar{x} = 306.32 \text{ mV} \qquad \text{and} \qquad s_x = 36.21 \text{ mV}$$

| | Upper-tail probability *p* | | |
|---|---|---|---|
| ***df*** | **.10** | **.05** | **.025** |
| **18** | 1.130 | 1.734 | 2.101 |
| **19** | 1.328 | 1.729 | 2.093 |
| **20** | 1.325 | 1.725 | 2.086 |
| | ?% | 90% | 95% |
| | | **Confidence level *C*** | |

Since $n = 20$, we use the *t* distribution with *df* = 19 to find the critical value.

From Table D, we find $t^* = 1.729$.

Therefore, the 90% confidence interval for *μ* is:

$$\bar{x} \pm t^* \frac{s_x}{\sqrt{n}} = 306.32 \pm 1.729 \frac{36.21}{\sqrt{20}}$$
$$= 306.32 \pm 14$$
$$= (292.32,\ 320.32)$$

We are 90% confident that the interval from 292.32 to 320.32 mV captures the true mean tension in the entire batch of video terminals produced that day.

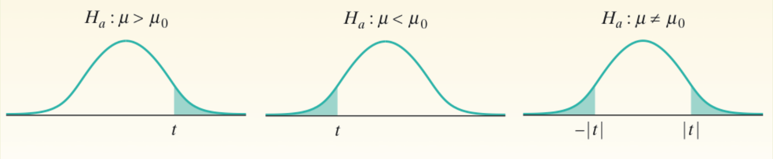
# The One-Sample *t* Test

**One-Sample *t* Test**

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:

$$t = \frac{\bar{x} - \mu_0}{s_x / \sqrt{n}}$$

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* $H_a$.



These *P*-values are exact if the population distribution is Normal and are approximately correct for large *n* in other cases.

# Example

The level of dissolved oxygen (DO) in a stream or river is an important indicator of the water's ability to support aquatic life. A researcher measures the DO level at 15 randomly chosen locations along a stream. Here are the results in milligrams per liter:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 4.53 | 5.04 | 3.29 | 5.23 | 4.13 | 5.50 | 4.83 | 4.40 |
| 5.42 | 6.38 | 4.01 | 4.66 | 2.87 | 5.73 | 5.55 | |

A dissolved oxygen level below 5 mg/l puts aquatic life at risk.

We want to perform a test at the $\alpha = 0.05$ significance level of:

$$H_0: \mu = 5$$
$$H_a: \mu < 5$$

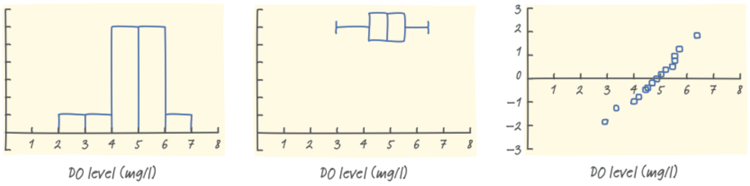
where *μ* is the actual mean dissolved oxygen level in this stream.

# Example

If conditions are met, we should do a one-sample *t* test for *μ*.

✓ **Random:** The researcher measured the DO level at 15 randomly chosen locations.

✓**Normal:** We don't know whether the population distribution of DO levels at all points along the stream is Normal. With such a small sample size (*n* = 15), we need to look at the data to see if it's safe to use *t* procedures.

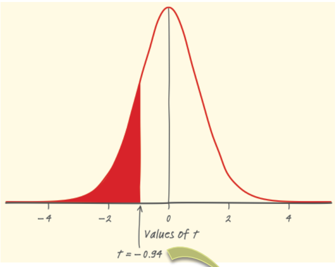


The histogram looks roughly symmetric; the boxplot shows no outliers; and the Normal probability plot is fairly linear. With no outliers or strong skewness, the *t* procedures should be pretty accurate even if the population distribution isn't exactly Normal.

# Example

The sample mean and standard deviation are $\bar{x} = 4.771$ and $s_x = 0.9396$

**Test statistic** $t = \dfrac{\bar{x} - \mu_0}{s_x / \sqrt{n}} = \dfrac{4.771 - 5}{0.9396 / \sqrt{15}} = -0.94$



Upper-tail probability *p*

| *df* | .25 | .20 | .15 |
|---|---|---|---|
| 13 | .694 | .870 | 1.079 |
| 14 | .692 | .868 | 1.076 |
| 15 | .691 | .866 | 1.074 |
| | 50% | 60% | 70% |

Confidence level *C*

***P*-value** The *P*-value is the area to the left of *t* = –0.94 under the *t* distribution curve with *df* = 15 – 1 = 14.

The *P*-value is between 0.15 and 0.20. Since this is greater than our $\alpha$ = 0.05 significance level, we fail to reject $H_0$. We don't have enough evidence to conclude that the mean DO level in the stream is less than 5 mg/l.

Since we decided not to reject $H_0$, we could have made a Type II error (failing to reject $H_0$ when $H_0$ is false). If we did, then the mean dissolved oxygen level *μ* in the stream is actually less than 5 mg/l, but we didn't detect that with our significance test.

# Matched Pairs *t* Procedures

- Comparative studies are more convincing than single-sample investigations. Study designs that involve making two observations on the same individual, or one observation on each of two similar individuals, result in **paired data.**

- When paired data result from measuring the same quantitative variable twice, as in the job satisfaction study, we can make comparisons by analyzing the differences in each pair. If the conditions for inference are met, we can use one-sample *t* procedures to perform inference about the mean difference $\mu_d$.

# Robustness of *t* Procedures

A confidence interval or significance test is called **robust** if the confidence level or *P*-value does not change very much when the conditions for use of the procedure are violated.

**Using the *t* Procedures**

Except in the case of small samples, the condition that the data are an SRS from the population of interest is more important than the condition that the population distribution is Normal.

- *Sample size at least 15*: The *t* procedures can be used except in the presence of outliers or strong skewness.
- *Sample size less than 15*: Use *t* procedures if the data appear close to Normal. If the data are clearly skewed or if outliers are present, do not use *t*.
- *Large samples*: The *t* procedures can be used even for clearly skewed distributions when the sample is large, roughly $n \geq 40$.

# Inference for non-normal distributions

What if the population is clearly non-Normal and your sample is small?

- If the data are skewed, you can attempt to **transform** the variable to bring it closer to Normality (e.g., logarithm transformation). The *t*-procedures applied to transformed data are often quite accurate for even moderate sample sizes.

- A distribution other than a Normal distribution might describe your data well. Many non-Normal models have been developed to provide inference procedures too (not covered here).

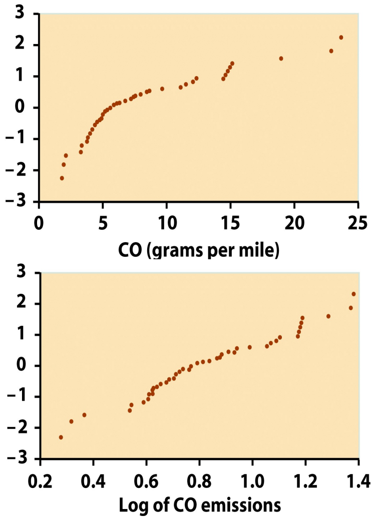
- You can always use a **distribution-free ("nonparametric")** inference procedure (see Chapter 15) that does not assume any specific distribution for the population. However, such a procedure is usually less powerful than distribution-driven tests (e.g., *t*-test).

# Transforming Data

The most common transformation is the **logarithm (log),** which tends to pull in the right tail of a distribution. The data values must all be *positive* in order to use the log transformation.

Instead of analyzing the original variable *X*, we first compute the logarithms and analyze the values of log *X*.

However, we cannot simply use the confidence interval for the mean of the logs to deduce a confidence interval for the mean *μ* in the original scale.



**Normal quantile plots for 46 car CO emissions**

# 7.2 Comparing Two Means

- Two-sample problems
- The two-sample *t* procedures
- Robustness of the two-sample *t* procedures
- Pooled two-sample *t* procedures

## Two-Sample Problems

- What if we want to compare the means of some quantitative variable for two populations, Population 1 and Population 2?
- Our parameters of interest are the population means $\mu_1$ and $\mu_2$. The best approach is to take separate random samples from each population and to compare the sample means.
- Suppose we want to compare the average effectiveness of two treatments in a completely randomized experiment. The parameters $\mu_1$ and $\mu_2$ are the true mean responses for Treatment 1 and Treatment 2, respectively. We use the mean response from each of the two groups to make the comparison. Here's a table that summarizes this situation:

| Population or treatment | Parameter | Statistic | Sample size |
|---|---|---|---|
| 1 | $\mu_1$ | $\bar{x}_1$ | $n_1$ |
| 2 | $\mu_2$ | $\bar{x}_2$ | $n_2$ |

# The Two-Sample *t* Statistic

When data come from two random samples or two groups in a randomized experiment, 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 $\bar{x}_1 - \bar{x}_2$ : $\sqrt{\frac{s_1{}^2}{n_1} + \frac{s_2{}^2}{n_2}}$

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:

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

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.

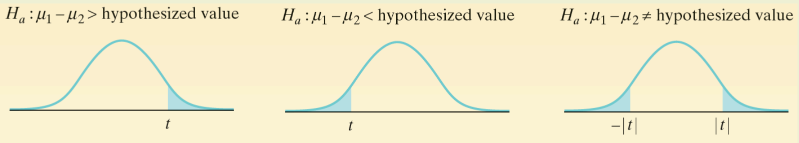
# Two-Sample *t* Test

**Two-Sample *t* Test for the Difference Between Two Means**

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

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

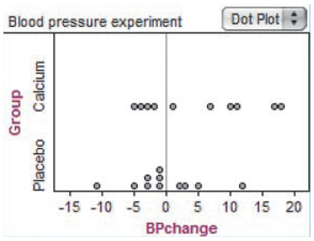
Find the $P$ - value by calculating the probability of getting a *t* statistic this large or larger in the direction specified by the alternative hypothesis $H_a$. Use the *t* distribution with degrees of freedom approximated by technology or the smaller of $n_1 - 1$ and $n_2 - 1$.


$H_a : \mu_1 - \mu_2 >$ hypothesized value
$t$
$H_a : \mu_1 - \mu_2 <$ hypothesized value
$t$
$H_a : \mu_1 - \mu_2 \neq$ hypothesized value
$-|t|$
$|t|$

# Example

Does increasing the amount of calcium in our diet reduce blood pressure? Examination of a large sample of people revealed a relationship between calcium intake and blood pressure. The relationship was strongest for black men. Such observational studies do not establish causation. Researchers therefore designed a randomized comparative experiment. The subjects were 21 healthy black men who volunteered to take part in the experiment. They were randomly assigned to two groups: 10 of the men received a calcium supplement for 12 weeks, while the control group of 11 men received a placebo pill that looked identical. The experiment was double-blind. The response variable is the decrease in systolic (top number) blood pressure for a subject after 12 weeks, in millimeters of mercury. An increase appears as a negative response. Here are the data:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Group 1 (calcium): | 7 | −4 | 18 | 17 | −3 | −5 | 1 | 10 | 11 | −2 | |
| Group 2 (placebo): | −1 | 12 | −1 | −3 | 3 | −5 | 5 | 2 | −11 | −1 | −3 |



# Example

We want to perform a test of:

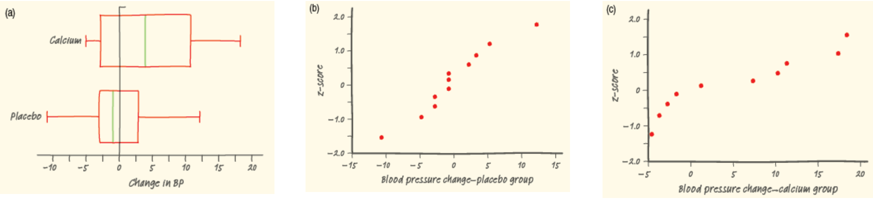
$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 > 0$$

where $\mu_1$ = the true mean decrease in systolic blood pressure for healthy black men like the ones in this study who take a calcium supplement, and $\mu_2$ = the true mean decrease in systolic blood pressure for healthy black men like the ones in this study who take a placebo. We will use $\alpha = 0.05$.

If conditions are met, we will carry out a two-sample *t*-test for $\mu_1 - \mu_2$.

✓**Random:** The 21 subjects were randomly assigned to the two treatments.

✓**Normal:** Boxplots and Normal probability plots for these data are below:



The boxplots show no clear evidence of skewness and no outliers. With no outliers or clear skewness, the *t* procedures should be pretty accurate.

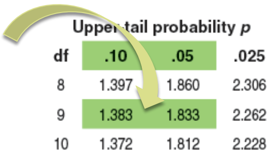
✓**Independent:** Due to the random assignment, the two samples may be regarded as independent of each other.

# Example

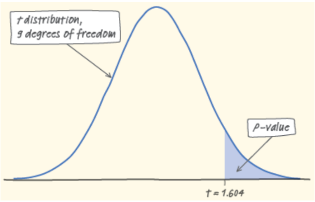
Since the conditions are satisfied, we can perform a two-sample *t*-test for the difference $\mu_1 - \mu_2$.

**Test statistic**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{[5.000 - (-0.273)] - 0}{\sqrt{\frac{8.743^2}{10} + \frac{5.901^2}{11}}} = 1.604$$



|  | Upper-tail probability *p* |  |  |
|---|---|---|---|
| df | .10 | .05 | .025 |
| 8 | 1.397 | 1.860 | 2.306 |
| 9 | 1.383 | 1.833 | 2.262 |
| 10 | 1.372 | 1.812 | 2.228 |



***P*-value** Using the conservative *df* = 10 – 1 = 9, we can use Table B to show that the *P*-value is between 0.05 and 0.10.

Because the *P*-value is greater than $\alpha = 0.05$, we fail to reject $H_0$. The experiment provides some evidence that calcium reduces blood pressure, but the evidence is not convincing enough to conclude that calcium reduces blood pressure more than a placebo. Assuming $H_0$: $\mu_1 - \mu_2 = 0$ is true, the probability of getting a difference in mean blood pressure reduction for the two groups (calcium – placebo) of 5.273 or greater just by the chance involved in the random assignment is 0.0644.

# 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

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

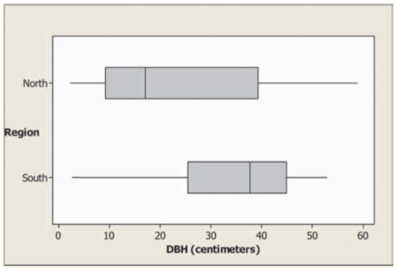
where $t^*$ is the critical value at confidence level $C$ for the $t$ distribution with degrees of freedom either gotten from technology or equal to the smaller of $n_1 - 1$ and $n_2 - 1$.

# Example

The Wade Tract Preserve in Georgia is an old-growth forest of longleaf pines that has survived in a relatively undisturbed state for hundreds of years. One question of interest to foresters who study the area is "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. Comparative boxplots of the data and summary statistics from Minitab are shown below. 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
```



# Example

Our parameters of interest are $\mu_1$ = the true mean DBH of all trees in the southern half of the forest and $\mu_2$ = the true mean DBH of all trees in the northern half of the forest. We want to estimate the difference $\mu_1 - \mu_2$ at a 90% confidence level.

We should use a two-sample *t* interval for $\mu_1 - \mu_2$ if the conditions are satisfied.

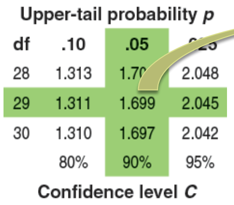
✓**Random:** The data come from random samples of 30 trees, one from the northern half and one from the southern half of the forest.

✓**Normal:** Skewness seen in the boxplots gives us reason to believe that the population distributions of DBH measurements may not be Normal. However, since both sample sizes are at least 30, we are safe using *t* procedures.

✓**Independent:** Researchers took independent samples from the northern and southern halves of the forest.

# Example

Since the conditions are satisfied, we can construct a two-sample *t* interval for the difference $\mu_1 - \mu_2$. We'll use the conservative df = 30 – 1 = 29.



Upper-tail probability *p*

| df | .10 | .05 | .025 |
|---|---|---|---|
| 28 | 1.313 | 1.70 | 2.048 |
| 29 | 1.311 | 1.699 | 2.045 |
| 30 | 1.310 | 1.697 | 2.042 |
| | 80% | 90% | 95% |

Confidence level *C*

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (34.53 - 23.70) \pm 1.699 \sqrt{\frac{14.26^2}{30} + \frac{17.50^2}{30}}$$

$$= 10.83 \pm 7.00 = (3.83,\ 17.83)$$

We are 90% confident that the interval from 3.83 to 17.83 centimeters captures the difference in the actual mean DBH of the southern trees and the actual mean DBH of the northern trees. This interval suggests that the mean diameter of the southern trees is between 3.83 and 17.83 cm larger than the mean diameter of the northern trees.

# Robustness again

The two-sample *t* procedures are more robust than the one-sample *t* methods, particularly when the distributions are not symmetric.

**Using the *t* Procedures**

Except in the case of small samples, the condition that the data are SRSs from the populations of interest is more important than the condition that the population distributions are Normal.

- *Sum of the sample sizes less than 15*: Use *t* procedures if the data appear close to Normal. If the data are clearly skewed or if outliers are present, do not use *t*.
- *Sum of the sample size at least 15*: The *t* procedures can be used except in the presence of outliers or strong skewness.
- *Large samples*: The *t* procedures can be used even for clearly skewed distributions when the sum of the sample sizes is large.

# Details of the *t* approximation

The exact distribution of the two-sample *t* statistic is not a *t* distribution. The distribution changes as the unknown population standard deviations change. However, an excellent approximation is available.

**Approximate Distribution of the Two-Sample *t* Statistic**

The distribution of the two-sample *t* statistic is very close to the *t* distribution with degrees of freedom given by:

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

This approximation is accurate when both sample sizes are 5 or larger.