Chapter 18: Inferences About Means

The Sampling Distribution of the Sample Mean

When random samples of size n are taken from a population with mean μ and standard deviation σ, the sampling distribution of the sample mean $\bar{y}$ with sample size n has mean $\mu_{\bar{y}} = \mu$ and standard deviation $\sigma_{\bar{y}} = SD(\bar{y}) = \dfrac{\sigma}{\sqrt{n}}$.

- If the population has a Normal distribution, then the sampling distribution of $\bar{y}$ with sample size n will be exactly Normally distributed, regardless of the sample size n.
- For any population, if the sample size is large enough (usually $n \geq 30$), then by the Central Limit Theorem (CTL), the sampling distribution of $\bar{y}$ with sample size n will be approximately Normally distributed.

When making inferences about μ by building confidence intervals and conducting hypothesis tests using sample data, the population standard deviation σ is typically unknown. (σ known $\rightarrow$ use Z)

In this case, we estimate σ with the sample standard deviation s and then use the **standard error**

$$SE(\bar{y}) = \frac{s}{\sqrt{n}}$$

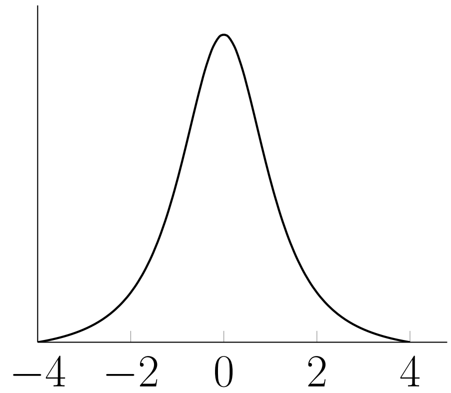
$$\frac{\bar{y} - \mu}{\frac{s}{\sqrt{n}}}$$

to estimate the standard deviation $\sigma_{\bar{y}} = SD(\bar{y})$.

For large sample sizes, computing P-values and margins of error using $SE(\bar{y})$ with a normal model still works fairly well. $\rightarrow$ use t for this course

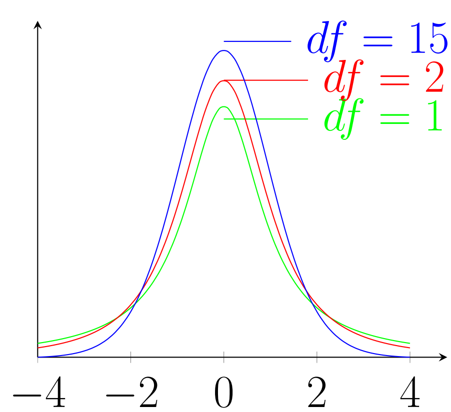
For smaller sample sizes, we require a new collection of models called the **Student's *t*-Models**.

<u>Student's t Distribution</u>

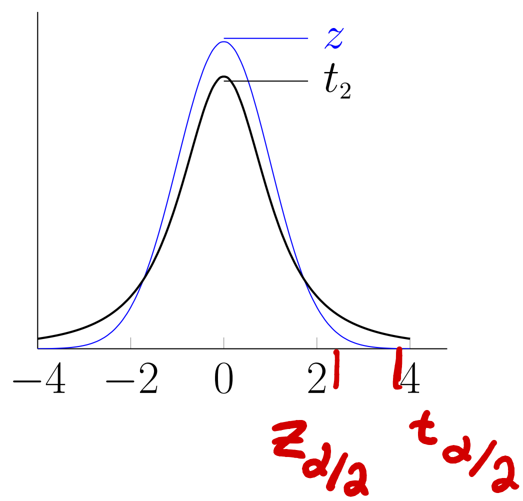


A t-distribution has an associated t-curve, which has the following properties:

- The total area under a t-curve is 1.
- A t-curve extends infinitely in both directions, getting very close to, but never touching, the horizontal axis.
- A t-curve is bell-shaped and symmetric about $t = 0$.
- There are infinitely many t-distributions/t-curves. Each one is identified by its number of **degrees of freedom**.
- We will denote the t-model with degrees of freedom df by t_{df}.



- t-curves have more spread than the standard normal curve.
- t-curves have thicker tails than the z-curve, that is, a t-curve does not approach the horizontal axis as quickly as the z-curve does.
- As the number of degrees of freedom increases, the t-curves look increasing like the standard normal curve. Spread decreases
- For $df \geq 2001$, a t-curve and the z-curve are essentially indistinguishable, but when $df \geq 30$, they are almost identical.
- To find critical values for confidence intervals or P-values for hypothesis tests using a t-distribution, we can use software or a t-table.



$t_{\alpha/2} > z_{\alpha/2}$

Using a t-Table

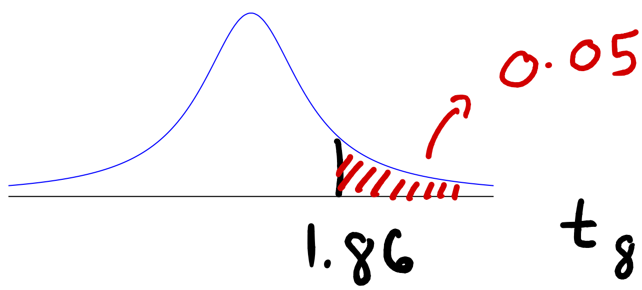
- The table provides either upper-tailed or two-tailed areas/probabilities as the column headings.
- The table only gives certain degrees of freedom as the row headings. If the degrees of freedom you require are not on the table, use the next lower df value. (This also applies to the χ^2-table.)
- The t critical values are given in the interior of the table.
- The final row gives the critical values from the z-distribution and label it $\infty\, df$. (t-model with infinite degrees of freedom).

use for $df > 1000$

Example: In each of the following, find the relevant critical value.

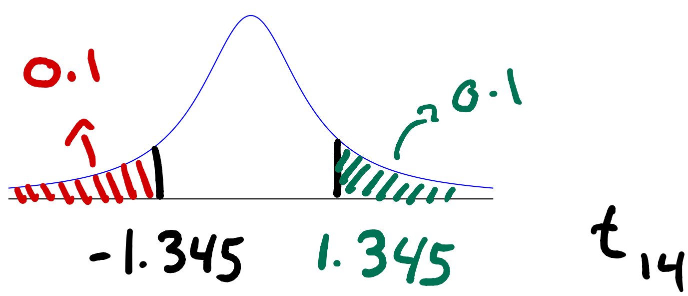
a) $df = 8$

$$P(t_8 > 1.86) = 0.05$$



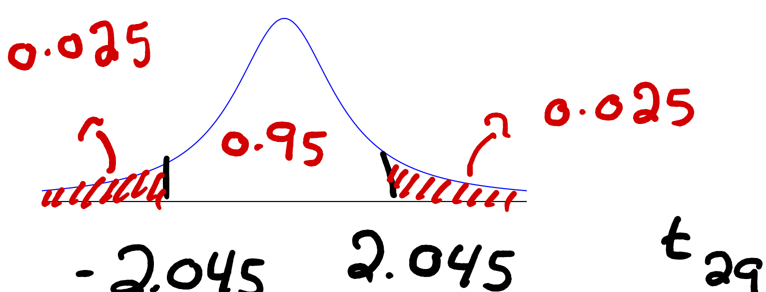
b) $df = 14$

$$P(t_{14} < -1.345) = 0.1$$



c) $df = 29$

(for z-curve, $z^* = 1.96$)

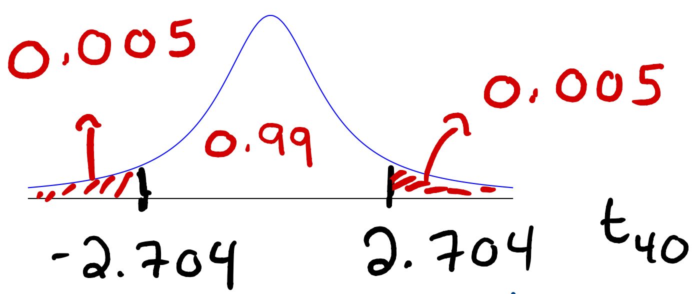


critical value for 95% CI for df = 29

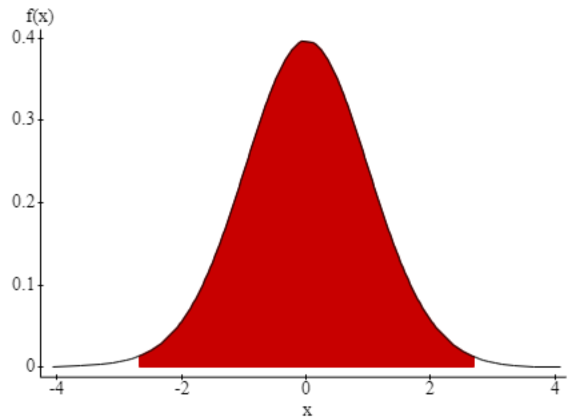
not on t table

d) $df = 42$

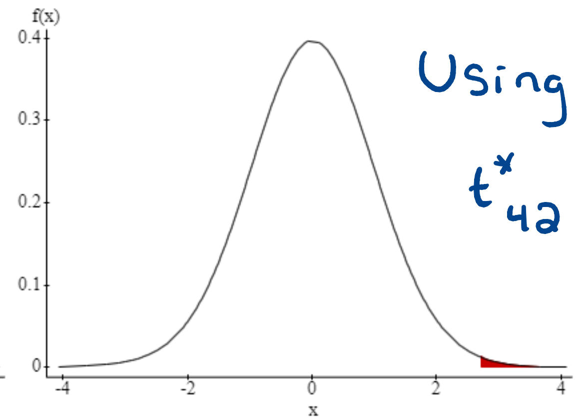
use df = 40



Critical value for 99% CI using t table



T Distribution
DF:42
P(-2.6981≤X≤2.6981) = 0.99



T Distribution
DF:42
P(X≥2.6980662) = 0.005

Using StatCrunch

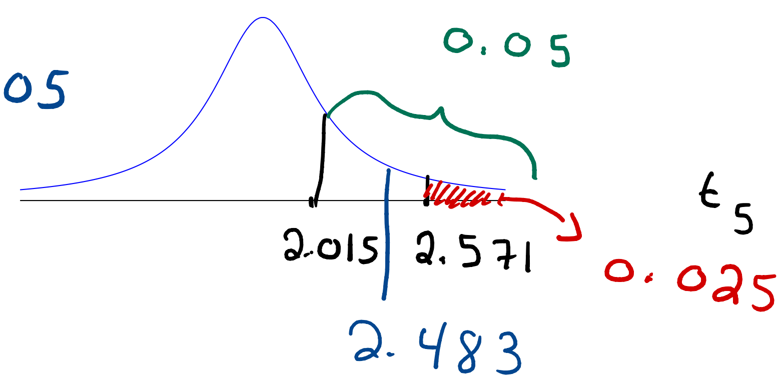
$t^*_{42} = 2.698$

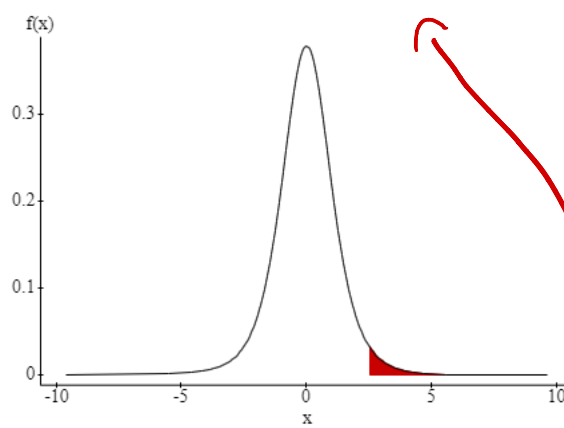
Example: In each of the following, use the t-table to find an interval in which the probability lies and compute the probability using software.

a) $P(t_5 > 2.483)$

$0.025 < P(t_5 > 2.483) < 0.05$

$= 0.028$

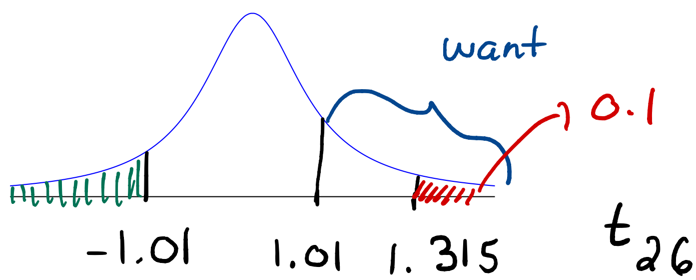


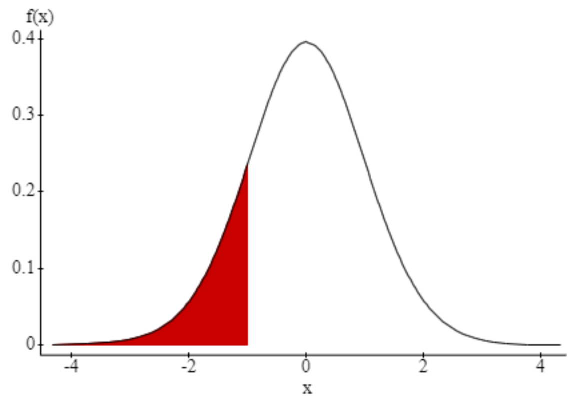


T Distribution
DF:5
P(X≥2.483) = 0.02781806

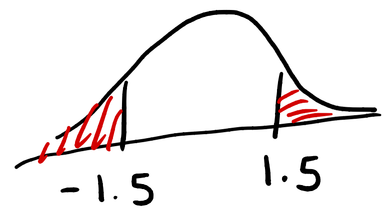
b) $P(t_{26} < -1.01)$ > 0.1

$= 0.161$





T Distribution
DF:26
P(X≤-1.01) = 0.16090147



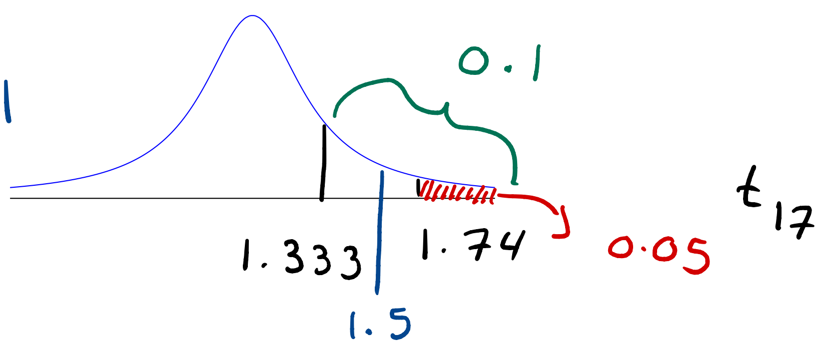
two-tailed test

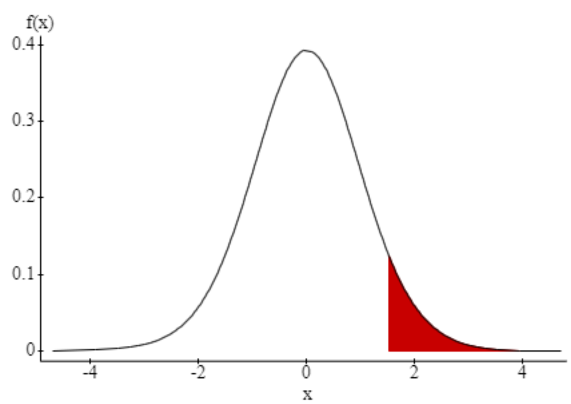
c) $2P(t_{17} > 1.5)$

$0.05 < P(t_{17} > 1.5) < 0.1$

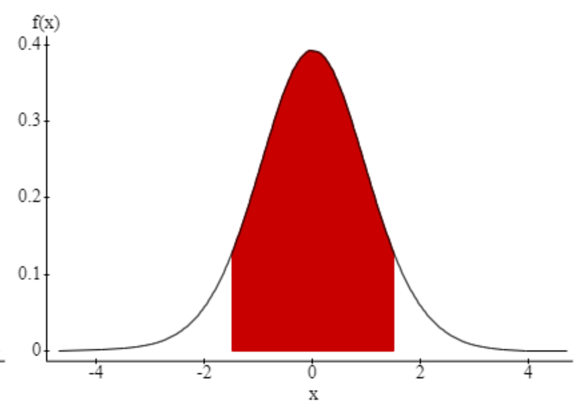
$0.1 < 2P(t_{17} > 1.5) < 0.2$

$= 0.152$





T Distribution
DF:17
P(X≥1.5) = 0.07597852



T Distribution
DF:17
P(-1.5≤X≤1.5) = 0.84804295

$2P(t_{17} > 1.5) = 2(0.0759\ldots)$
$= 0.152$

$2P(t_{17} > 1.5)$
$= 1 - 0.848\ldots = 0.152$

Assumptions and Conditions:

a) **Independence Assumption:** values in sample must be independent of each other.

   i) **Randomization Condition:** data obtained from a simple random sample from the population or from a properly randomized experiment.

   ii) **10% Condition:** sample size should be less than 10% of population size (when drawn without replacement).

b) **Normal Population Assumption:** we require either

   i) A large sample size ($n \geq 30$), or

   ii) **Nearly Normal Condition:** The data comes from an approximately Normal population (distribution is unimodal and reasonably symmetric). $\rightarrow$ make histogram of data

Note: Beware of outliers in the data, especially when n is small, as they can substantially change your conclusions. If you find outliers, consider doing the analysis twice, once with the outliers and once without the outliers, to observe how much they affect the results.

Key Fact:

When these assumption are made and conditions are met, the standardized sample mean

$$t = \frac{\bar{y} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}$$

follows a Student's t-model with $n - 1$ degrees of freedom, where n is the sample size.

One-Sample t-Interval for the Mean when σ is Unknown

Recall: a confidence interval has the form

$$\text{point estimate} \pm \text{margin of error}$$
$$= \text{point estimate} \pm \big(\text{critical value} \times \text{standard error of the estimate}\big)$$

When the population standard deviation σ is unknown and when the relevant assumptions are made and conditions are met, a $100(1-\alpha)\%$ confidence interval for the population mean μ is

$$\bar{y} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$$

where t^*_{n-1} is the critical value corresponding to the $100(1-\alpha)\%$ confidence level based on $n-1$ degrees of freedom.

new sample, new s

- Increasing the sample size, decreases the margin of error, which makes the confidence interval narrower. $t^*_{n-1} \uparrow$
- Increasing the confidence level, increases the margin of error, which makes the confidence interval wider.

Note: When interpreting a confidence interval, keep in mind:

- The confidence interval is about the **population** mean, not the individual values or the sample mean.
- Avoid making probability statements about a particular interval. A particular interval either includes μ or it does not. The population mean μ does not vary.
- The interval we have computed does not set the standard for the other intervals. ↪ don't compare other samples to this
- The confidence level tells us the percentage of all possible samples of size n that result in an interval that contains μ.

Example: A random sample of 21 washing machines was obtained and the length (in minutes) of the wash cycle of each one was recorded. The sample had a mean of $\bar{y} = 37.8$ minutes and a sample standard deviation of $s = 5.9$ minutes. Assume the population distribution of all wash cycle times is approximately Normal.

a) Give a 99% confidence interval for the true mean wash cycle time.

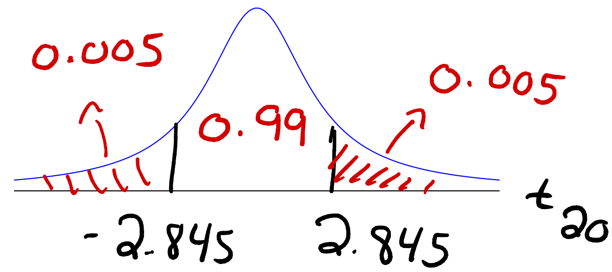
$$\bar{y} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$$

$$= 37.8 \pm (2.845)\left(\frac{5.9}{\sqrt{21}}\right)$$

$$= 37.8 \pm 3.6629$$

$$= (34.14, 41.46) \rightarrow \text{wider than } 90\% \text{ CI}$$

df = 20



b) Give a 90% confidence interval for the true mean wash cycle time.

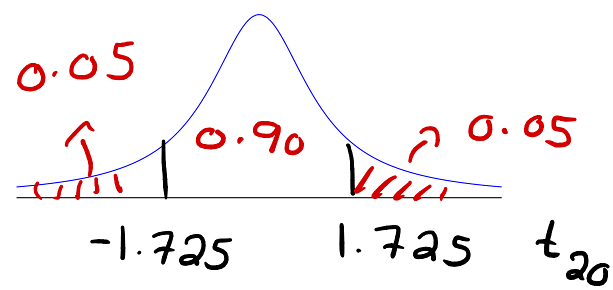
$$\bar{y} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$$

$$= 37.8 \pm (1.725)\left(\frac{5.9}{\sqrt{21}}\right)$$

$$= 37.8 \pm 2.2209$$

$$= (35.58, 40.02)$$

df = 20



c) Give a 90% confidence interval for the true mean wash cycle time, if the sample size was actually 57.

df = 56 not on table

$\hookrightarrow$ use df = 50

$t^*_{50} = 1.676$

$$\bar{y} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$$

$$= 37.8 \pm (1.676)\left(\frac{5.9}{\sqrt{57}}\right)$$

$$= 37.8 \pm 1.3098$$

$$= (36.49, 39.11) \rightarrow \text{narrower than } n = 21$$

Choosing your Sample Size

Sometimes we want to know what sample size to choose in order to have a specific margin of error with a specific confidence level.

largest acceptable ME

To do this, we take the formula for the margin of error

$$ME = t^*_{n-1} \frac{s}{\sqrt{n}}$$

and solve for n

$$n = \left(\frac{t^*_{n-1} s}{ME} \right)^2$$

Problems: since we have not taken a sample, we do not have a value for s or the number of degrees of freedom.

Solutions:

- guess the value of s, conduct a small pilot study, or use a previous study.
- use the z^* critical value (or t^* with $df = \infty$) corresponding to the given confidence level in place of t^*_{n-1}.

Therefore, to find the minimum sample size to construct a $100(1-\alpha)\%$ confidence interval with a margin of error of at most ME, we use a sample of size

$$n \overset{\geq}{=} \left(\frac{z^* s}{ME} \right)^2$$

$n \uparrow$ $ME \downarrow$

$z^* \uparrow$ $n \uparrow$

Note: When you use this formula, always round up (to an integer) at the end to find n!

Example: Suppose that scientists studying infant health want to estimate the mean head circumference of all infants. How large a sample should they take if they want to be 95% confident that the estimate is within 0.5 cm of the true population mean? Suppose that in a previous sample, the sample standard deviation was $s = 2.1$ cm.

$z^* = 1.96$

$s = 2.1$

$ME = 0.5$

$$n = \left(\frac{z^* s}{ME}\right)^2 = \left(\frac{1.96\,(2.1)}{0.5}\right)^2 = 67.766$$

$\therefore$ They should sample 68 infants.

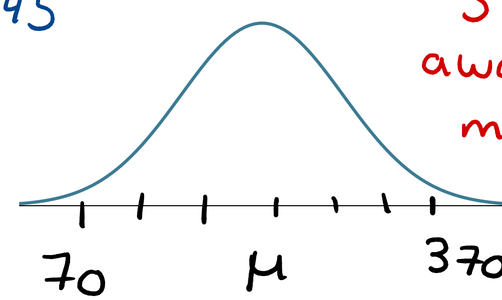
Example: Suppose that we want to estimate the mean cost of textbooks per semester for U of A students to within \$8 of the true population mean. Also suppose that the amount spent on textbooks per semester by U of A students is Normally distributed and that most students spend between \$70 and \$370 per semester. What is the minimum sample size we should use to be 90% confident of attaining the level of accuracy mentioned above?

3 S.D. away from mean

$z^* = 1.645$

$s = 50$

$ME = 8$



$$s \approx \frac{370-70}{6} = 50$$

$$n = \left(\frac{z^* s}{ME}\right)^2 = \left(\frac{1.645\,(50)}{8}\right)^2 = 105.704$$

$\therefore$ We should sample 106 students.

Hypothesis Test for a Population Mean: One-Sample t-Test

A hypothesis test for a population mean μ, when σ is unknown, has five steps:

1. **Assumptions/Conditions:**
   - σ is unknown.
   - Individuals in sample must be independent of each other.
   - Data collected using randomization.
   - If $n < 30$, then the population is approximately Normally distributed.

2. **Hypotheses:**

$$H_0 : \mu = \mu_0$$

choose one

$$H_A : \begin{cases} \mu \neq \mu_0 & \text{(two-tailed test)} \\ \mu < \mu_0 & \text{(lower-tailed test)} \\ \mu > \mu_0 & \text{(upper-tailed test)} \end{cases}$$

3. **Test Statistic:**

$$t_0 = \frac{\bar{y} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$$

   When the conditions are met and the null hypothesis is true, this statistic follows a Student's t-model with $n - 1$ degrees of freedom.

4. ***P*-value:** Compute one of the following using software or the t-table with $df = n - 1$:

Test	P-value
Two-tailed Test	$2P(t_{n-1} > \|t_0\|)$
Lower-tailed Test	$P(t_{n-1} < t_0)$
Upper-tailed Test	$P(t_{n-1} > t_0)$

$P(t > t_0)$ with $df = n-1$

If using a t-table, we will likely only be able to find a range in which the P-value lies.

5. **Conclusion:** Given a significance level α,

- if P - value $\leq \alpha$, we reject H_0 at level α
- if P- value $> \alpha$, we do not reject H_0 at level α

Example: A manufacturer claims that, on average, a gallon of its paint will cover 400 square feet of surface area. (Assume the coverage areas of the cans of paint made by this manufacturer are approximately Normally distributed.) To test this claim, a random sample of ten 1-gallon paint cans of white paint were used to paint ten identical areas using the same type of equipment. The areas (in square feet) covered by these ten cans were:

310	311	412	368	447
376	303	410	365	350

a) Do the data present sufficient evidence to indicate that the average coverage area differs from 400 square feet? Use $\alpha = 0.05$.

Note: $\bar{y} = 365.2$ and $s = 48.417$

$n = 10$, $df = 9$

1. **Assumptions/Conditions:**

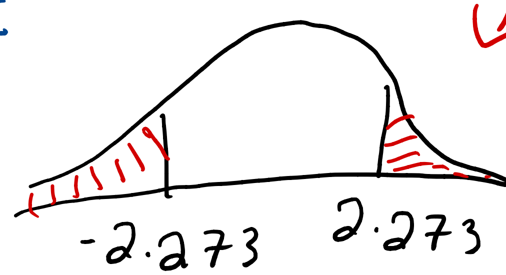
- σ unknown
- random sample (assume independence)
- $n = 10 < 30$, but population is approx. Normally distributed.

2. **Hypotheses:**

$$\left.\begin{array}{l} H_0: \mu = 400 \\ H_A: \mu \neq 400 \end{array}\right\} \text{two-tailed test}$$

3. **Test Statistic:**

$$t_0 = \frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{365.2 - 400}{\frac{48.417}{\sqrt{10}}} = -2.273$$



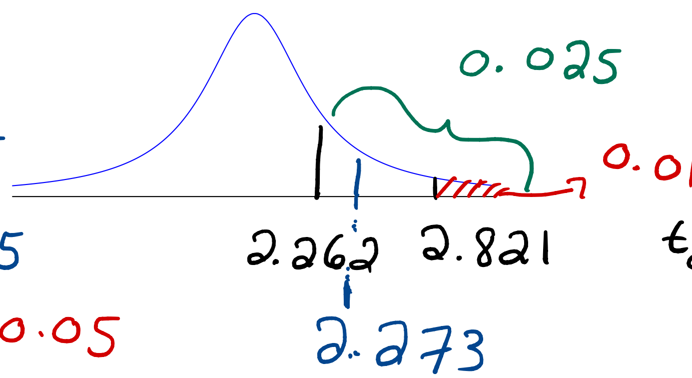
4. ***P*-value:** $df = 9$

Using t-table,

$0.01 < P(t_9 > 2.273) < 0.025$

$0.02 < 2P(t_9 > 2.273) < 0.05$

$= p\text{-value}$ $\quad \alpha = 0.05$



Using stat crunch: $p\text{-value} = 0.0491 \leq 0.05$

Hypothesis test results:

Variable	Sample Mean	Std. Err.	DF	T-Stat	P-value
Coverage area	365.2	15.310708	9	-2.2729191	0.0491

5. **Conclusion:**

P-value = 0.0491
$\alpha = 0.05$

Since P-value $\le \alpha$, we reject H_0 at the 0.05 significance level, that is, there is sufficient statistical evidence to conclude that the average coverage area differs from 400 square feet.

- not reject H_0 if $\alpha = 0.01$ or $\alpha = 0.02$

Note: The P-value is the smallest α level at which we can reject H_0.

- for $\alpha = 0.0491$,
  P-value $= 0.0491 \le \alpha = 0.0491 \to$ reject H_0
- choosing any smaller α level, we would not reject H_0

b) Construct a 95% confidence interval for the true mean coverage area of all cans of paint made by this manufacturer.

$df = 9$

$t^*_9 = 2.262$

$$\bar{y} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$$

$$= 365.2 \pm (2.262)\left(\frac{48.417}{\sqrt{10}}\right)$$

$$= 365.2 \pm 34.633$$

$$= (330.57, 399.83) \not\ni 400$$

$H_0: \mu = 400 \to$ reject

(if $H_0: \mu = 350 \to$ not reject
$350 \in (330.57, 399.83)$)

The Relationship Between Confidence Intervals and Hypothesis Tests

Confidence intervals and hypothesis tests examine data from different perspectives:

- Hypothesis tests begin with a proposed parameter value and ask if the data are consistent with that value. If the data are inconsistent with the proposed value of the parameter, we reject the null hypothesis as the proposed value no longer seems plausible.
- A confidence interval begins with the data and finds an interval of plausible values where the parameter may fall.

Key Fact: A $100(1-\alpha)\%$ confidence interval contains all the null hypothesis values that would **not** be rejected by a two-tailed hypothesis test with significance level α.

90% CI

95% CI, $\alpha = 0.05$

$\alpha = 0.1$

Example: Employers are often concerned about the amount of time that employees spend each day making personal use of company technology. Suppose that the CEO of a large company wants to determine if the average amount of time per day spent on personal use of company devices for his employees is more than 75 minutes. A random sample of 10 employees was selected and asked about their daily personal use of company devices. Their responses were:

66	70	75	88	69
89	71	71	63	86

a) Do the data present sufficient evidence that the mean for this company is more than 75 minutes? Use $\alpha = 0.05$.

Note: $\bar{y} = 74.8$ and $s = 9.45$ $n=10$, $df = 9$

1. **Assumptions/Conditions:**

- σ unknown
- random sample (assume independence)
- $n < 30$, assume population approx. Normally distributed.

2. **Hypotheses:**

$$H_0: \mu = 75$$
$$H_A: \mu > 75$$

upper-tailed test

3. **Test Statistic:**

$$t_0 = \frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{74.8 - 75}{\frac{9.45}{\sqrt{10}}} = -0.067$$

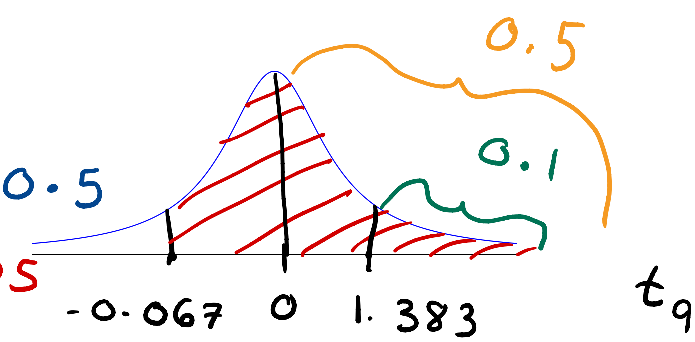
upper-tailed test

4. ***P*-value:** $df = 9$

Using t-curve,

P-value $= P(t_9 > -0.067) > 0.5$

> 0.05



Using StatCrunch, p-value $= 0.526 > 0.05$

Hypothesis test results:

Variable	Sample Mean	Std. Err.	DF	T-Stat	P-value
Personal Use	74.8	2.9881246	9	-0.066931612	0.526

5. **Conclusion:**

Since P-value > α, we do not reject H_0 at the 0.05 significance level, that is, there is not enough statistical evidence to conclude that the mean time spent on personal use of company technology is more than 75 minutes per day at this company.

b) Construct a 95% confidence interval for the mean time spent on personal use of company technology per day at this company.

$$\bar{y} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$$

$df = 9$

$t^*_9 = 2.262$

$$= 74.8 \pm (2.262)\left(\frac{9.45}{\sqrt{10}}\right)$$

$$= 74.8 \pm 6.7597$$

$$= (68.04, 81.56)$$

$\therefore$ we are 95% confident that $\mu \in (68.04, 81.56)$