One-sample means with the *t*-distribution

Colby Community College

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:

$$\text{Mean} = \mu \qquad \text{Standard Deviation } (SE) = \frac{\sigma}{\sqrt{n}}$$

Note

It's rare to need to estimate the population mean μ, but somehow know the population standard deviation σ. In most cases σ will need to be estimated.

Conditions to Apply the Central Limit Theorem

Independence: The sample observations must be independent.

Normality: When a sample is small, we also require that the sample observations come from a normally distributed population.

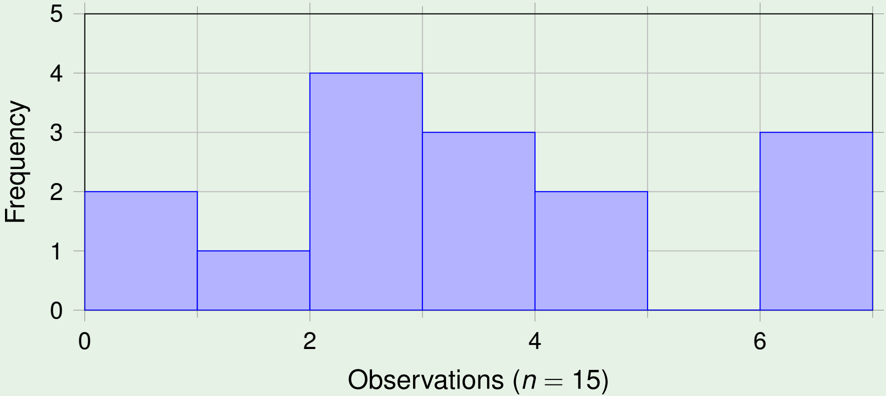
$n < 30$: If the sample size n is less than 30 and there are no clear outliers in the data, then we typically assume the data is from a nearly normal population.

$n \geq 30$: 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 population is not.

Note

In a first course in statistics, you aren't expected to develop perfect judgment on the normality condition.

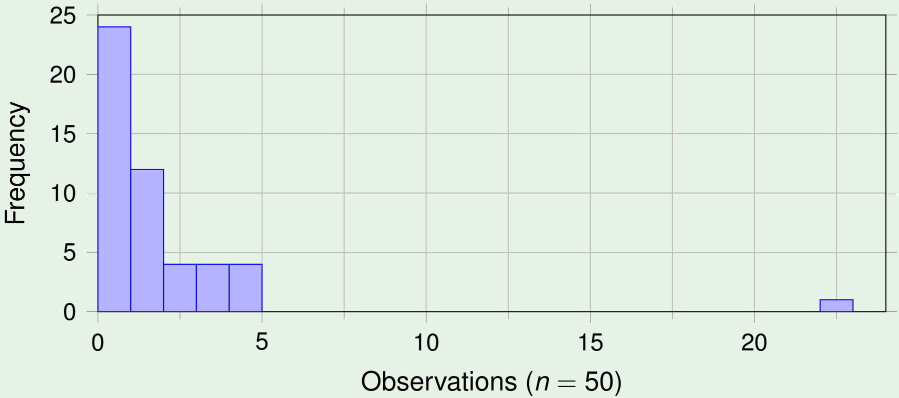
Example 1



Is the normality condition met?

Since there are less than 30 observations, we need to look for *clear* outliers. While there is a gap on the right, the gap is small and 20% of the observations fall in rightmost bar. We can't really call these clear outliers, so the normality condition is reasonably met.

Example 2



Is the normality condition met?

The sample size is greater than 30, so we need to look for an extreme outlier. The gap is more than four times the width of the cluster on the left side, so this is clearly an extreme outlier and the normality condition is not met.

Note

In practice, we cannot directly calculate the standard error for $\bar{x}$, since we do not know the population standard deviation σ.

We can use the sample standard deviation s as the best estimate of σ:

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

Definition

If a population has a normal distribution, then the distribution of

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

is called the **Student *t* distribution** for sample sizes n.

Note

A Student t distribution is commonly called a ***t* distribution**.

Definition

The **degrees of freedom** (or **df**) for a collection of sample data is the number of sample values that can vary after certain restrictions have been imposed on all data values.

When modeling $\bar{x}$ using the *t*-distribution, use:

$$df = n - 1$$

Example 3

If 10 test scores must have mean 80, then their sum must be 800.

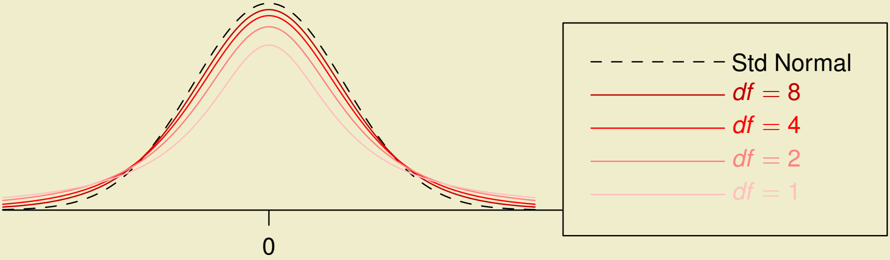
We can freely assign values to the first 9 scores, but the 10th score would need to be:

$$s_{10} = 800 - s_1 - s_2 - s_3 - s_4 - s_5 - s_6 - s_7 - s_8 - s_9$$

Hence 9 degrees of freedom.

Note

The Student *t* distribution changes for different degrees of freedom.



Note

The *t*-distribution has a mean of $t = 0$

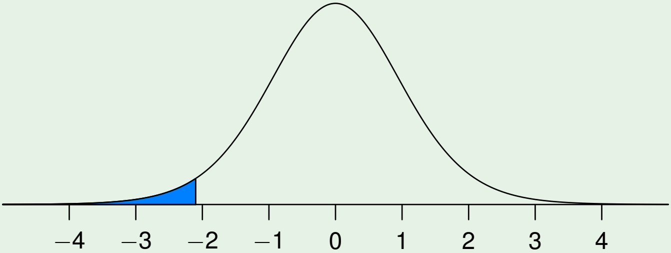
The standard deviation varies with *n*, but is always greater than 1.

Note

As the sample size gets larger, the Student *t* distribution gets closer to the standard normal distribution.

Example 4

The t-distribution with 13 degrees of freedom is shown.

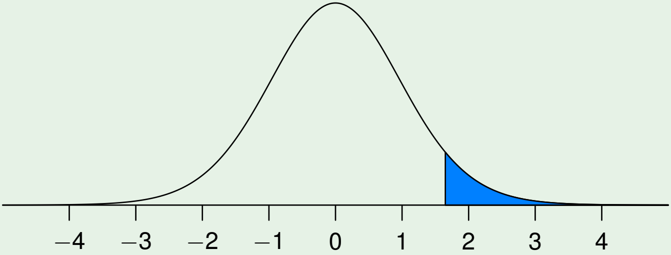


The area to the left of $t = -2.1$ is:

$$P\left(t \leq 2.1\right) \approx 0.0279$$

Example 5

The t-distribution with 20 degrees of freedom is shown.

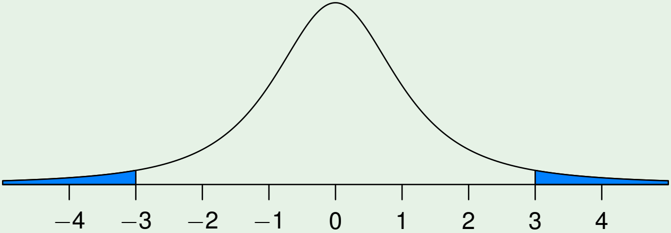


The area to the right of $t = 1.65$ is:

$$P\left(t \geq 1.65\right) \approx 0.0573$$

Example 6

The *t*-distribution with 2 degrees of freedom is shown.



The area more than three units from the mean is:

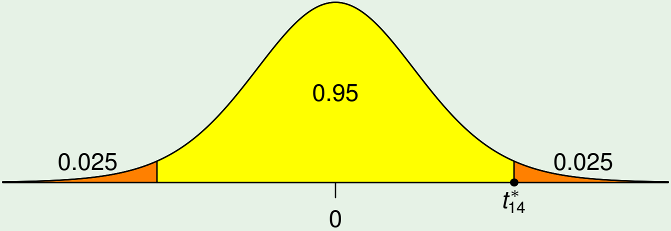
$$P\,(t \leq -3 \text{ or } t \geq 3) \approx 0.0955$$

Example 7

Let us find the critical value, t^*_{df}, corresponding to a 95% confidence level, given that the sample size is $n = 15$.

The degrees of freedom is $df = n - 1 = 15 - 1 = 14$.

We need to find the t-value that has $0.95 + 0.025 = 0.975$ to the left.



Using technology we get $t^*_{14} = 2.145$

Note

The critical value t^*_{df} must be found every time, since the t-distribution changes for different sample sizes.

Confidence Interval For A Single Mean

Prepare: Identify $\bar{x}$, s, n, and what confidence level to use.

Check: Verify the conditions to ensure $\bar{x}$ is nearly normal:

- The sample observations are nearly normal.
- The population is normally distributed or $n \geq 30$.

Calculate: If the conditions hold, compute:

- The degrees of freedom: $df = n - 1$
- The critical value: t^*_{df}
- The standard error: $SE = \dfrac{s}{\sqrt{n}}$
- The confidence interval:

$$\text{point estimate} \pm t^*_{df} \cdot SE \quad \rightarrow \quad \bar{x} \pm t^*_{df} \cdot \frac{s}{\sqrt{n}}$$

Conclude: Interpret the confidence interval in the context of the problem.

Example 8

The weights (in hectograms, hg) of randomly selected girls at birth are

33 28 33 37 31 32 31 28 34 28 33 26 30 31 28

(Based on data from the National Center for Health Statistics.)

The summary statistics for this sample are

$$n = 15 \qquad \bar{x} = 30.9 \qquad s = 2.9$$

Construct a 95% confidence interval for the mean birth weight of girls.

The margin of error is

$$E = t^*_{14} \cdot \frac{s}{\sqrt{n}} = 2.145 \cdot \frac{2.9}{\sqrt{15}} = 1.606126$$

The confidence interval is

$$\begin{array}{ccccc} \bar{x} - E & < & \mu & < & \bar{x} + E \\ 30.9 - 1.606126 & < & \mu & < & 30.9 + 1.606126 \\ 29.3 \text{ hg} & < & \mu & < & 32.5 \text{ hg} \end{array}$$

We are 95% confident that the interval from 29.2 hg to 32.5 hg actually does contain the true value of μ.

Hypothesis Testing For A Single Mean

Prepare: Identify the parameter of interest, list out hypotheses, identify $\bar{x}$, s, n, and what confidence level to use.

Check: Verify the conditions to ensure $\bar{x}$ is nearly normal:

- The sample observations are nearly normal.
- The population is normally distributed or $n \geq 30$.

Calculate: If the conditions hold, compute:

- The degrees of freedom: $df = n - 1$
- The standard error: $SE = \dfrac{s}{\sqrt{n}}$
- The t-score: $t = \dfrac{\bar{x} - \text{null value}}{SE}$
- The p-value.

Conclude: Evaluate the hypothesis test by comparing the p-value to α, and provide a conclusion in the context of the problem.

Example 9

The Cheery Blossom Race is a 10-mile race in Washington, D.C. held every spring.

The average time for all runners who finished the Cherry Blossom Race in 2006 was 93.29 minutes.

We want to determine using data from 100 participants in the 2017 Cherry Blossom Race whether runners are getting faster or slower.

What are the null and alternative hypotheses?

H_0 :The average time was the same in 2007 and 2017

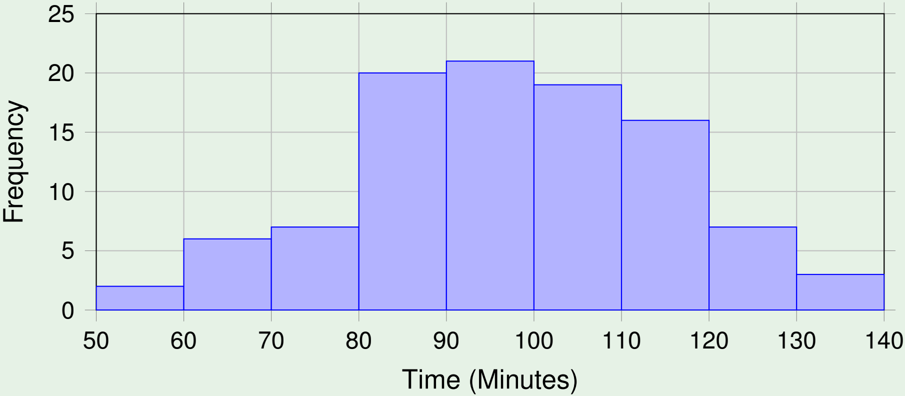
$\mu = 93.29$

H_A :The average time was different in 2017 compared to 2007

$\mu \neq 93.29$

Example 9 (Continued)

The histogram shows the times for 100 of the runners in the 2017 race.



Is the normality condition satisfied?

We have more than 30 observations and there are no outliers, so yes.

Example 9 (Continued)

In 2007, the average time was 93.29 minutes. Our sample of 100 runners from 2017 have a mean of 97.32 minutes and standard deviation of 16.98 minutes.

The degrees of freedom are $df = n - 1 = 100 - 1 = 99$.

The standard error is:

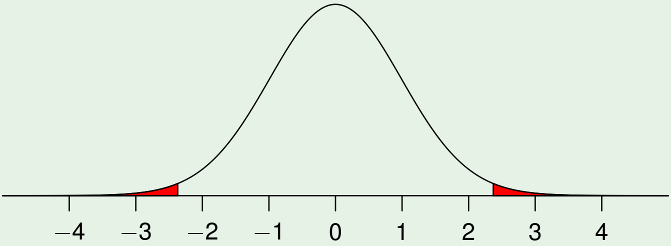
$$SE = \frac{s}{\sqrt{n}} = \frac{16.98}{\sqrt{100}} = 1.70$$

The t-score is:

$$t = \frac{\bar{x} - \text{null value}}{SE} = \frac{97.32 - 93.29}{1.70} = 2.37$$

Example 9 (Continued)

So, we wish to find the area of the tails:



Using technology, we get a *p*-value of 0.02.

Do we reject or fail to reject H_0?

Because $0.02 < 0.05$, we reject the null hypothesis.

Because we reject the null hypothsis, we have evidence that there is a difference in average race times between the 2007 and 2017 races.

Since the average in 2017 (97.32 minutes) is larger than the average in 2007 (93.29 minutes), it is likely that racers in 2017 were slower.