Section 6.2:

Inference for a Mean

Inference Using N(0,1)

If the distribution of the sample statistic is normal:

A confidence interval can be calculated by

$$\boxed{\text{sample statistic} \pm z^{*} \times SE}$$

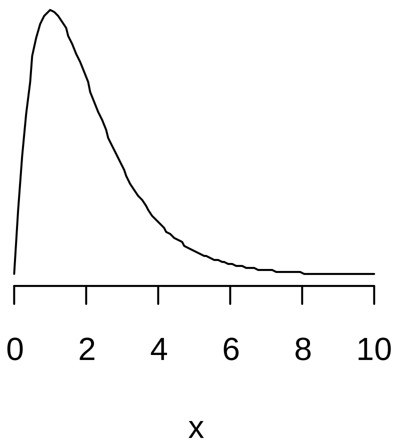
where z^{*} is a N(0,1) percentile depending on the level of confidence.

A p-value is the area in the tail(s) of a N(0,1) beyond

$$\boxed{z = \frac{\text{sample statistic} - \text{null value}}{\text{SE}}}$$

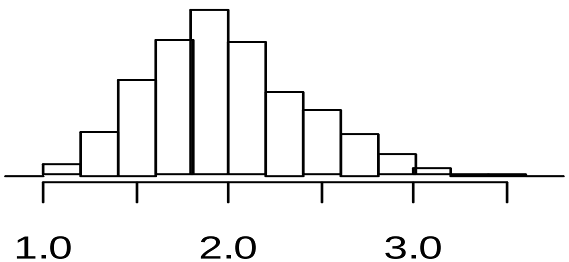
CLT for a Mean

Population

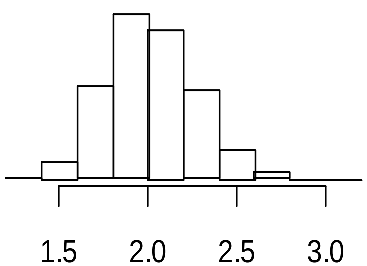


Distribution of Sample Means

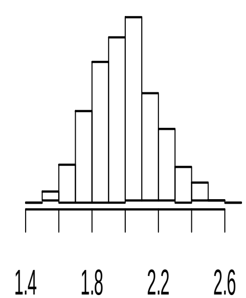
n = 10



n = 30



n = 50



SE of a Mean

The standard error for a sample mean can be calculated by

$$\boxed{SE = \frac{\sigma}{\sqrt{n}}}$$

The standard deviation of the population is

1. σ
2. s
3. $\dfrac{\sigma}{\sqrt{n}}$

The standard deviation of the **sample** is

1. σ
2. s
3. $$\frac{\sigma}{\sqrt{n}}$$

The standard deviation of the **sample mean** is

1. σ
2. s
3. $\dfrac{\sigma}{\sqrt{n}}$

CLT for a Mean

If $n \geq 30$*, then

$$\bar{X} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

*Smaller sample sizes may be sufficient for symmetric distributions, and 30 may not be sufficient for very skewed distributions or distributions with high outliers

Standard Error

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

- We don't know the population standard deviation σ, so estimate it with the sample standard deviation, s

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

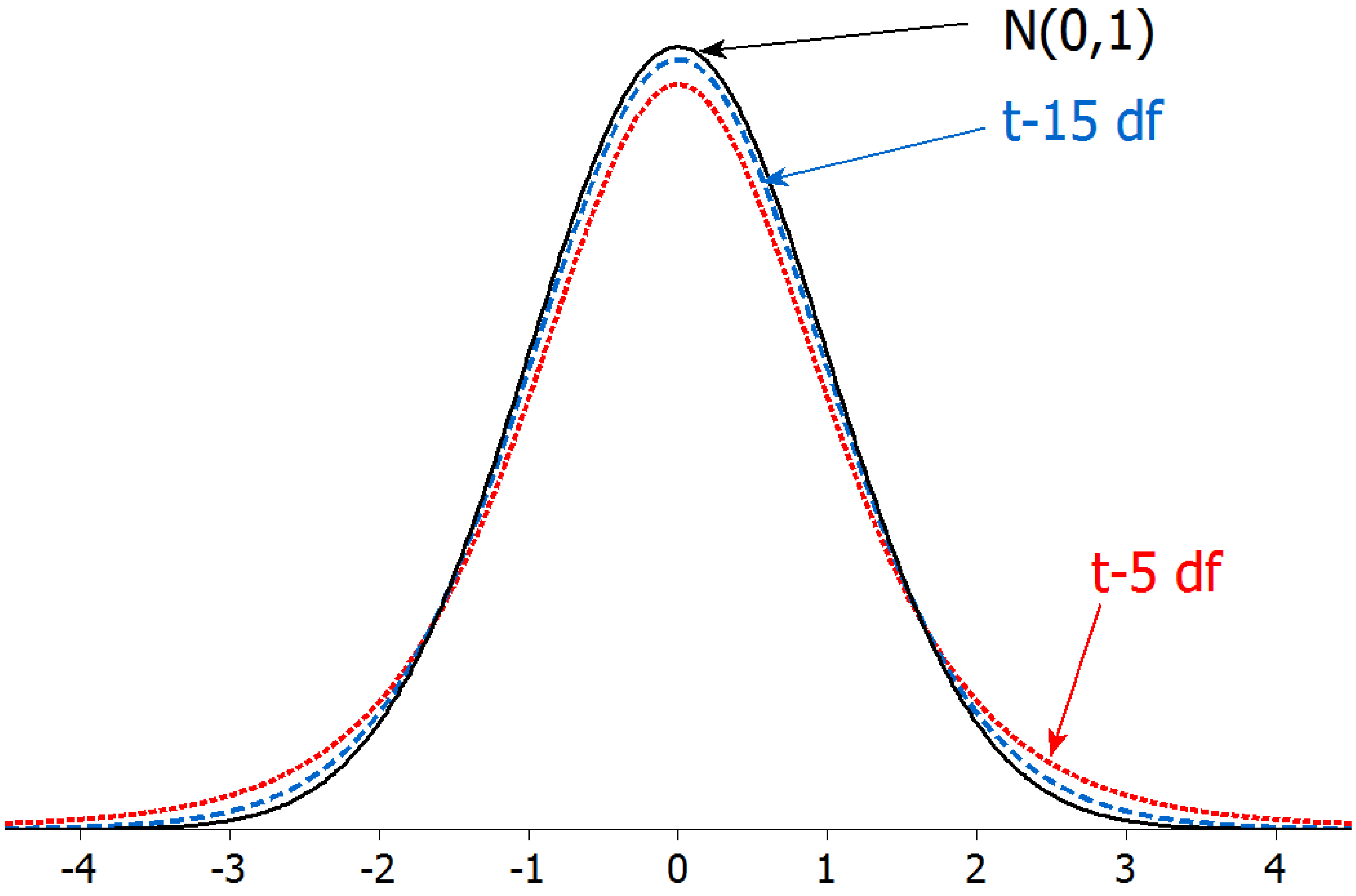
t-distribution

- Replacing σ with s changes the distribution of the z-statistic from a **normal distribution** to a ***t-distribution***
- The t distribution is very similar to the standard normal, but with slightly fatter tails to reflect this added uncertainty

Degrees of Freedom

- The *t*-distribution is characterized by its ***degrees of freedom (df)***
- Degrees of freedom are calculated based on the sample size
- The higher the degrees of freedom, the closer the *t*-distribution is to the standard normal

t-distribution



…But What Are Degrees of Freedom?

- In statistics, the number of **degrees of freedom** is the number of values in the final calculation of a statistic that are free to vary.
- Example:

1. There are 4 values in a particular sample.
2. The sample mean is 5.
3. Three of the values in the sample are 3, 4, and 6.

…Then the fourth value must be 7. Only **three** of the values **(4 – 1)** are free to vary before the fourth is determined!

Normality Assumption

- Using the t-distribution requires an extra assumption: the **data comes from a *normal distribution***

- Note: this assumption is about the original data, not the distribution of the statistic

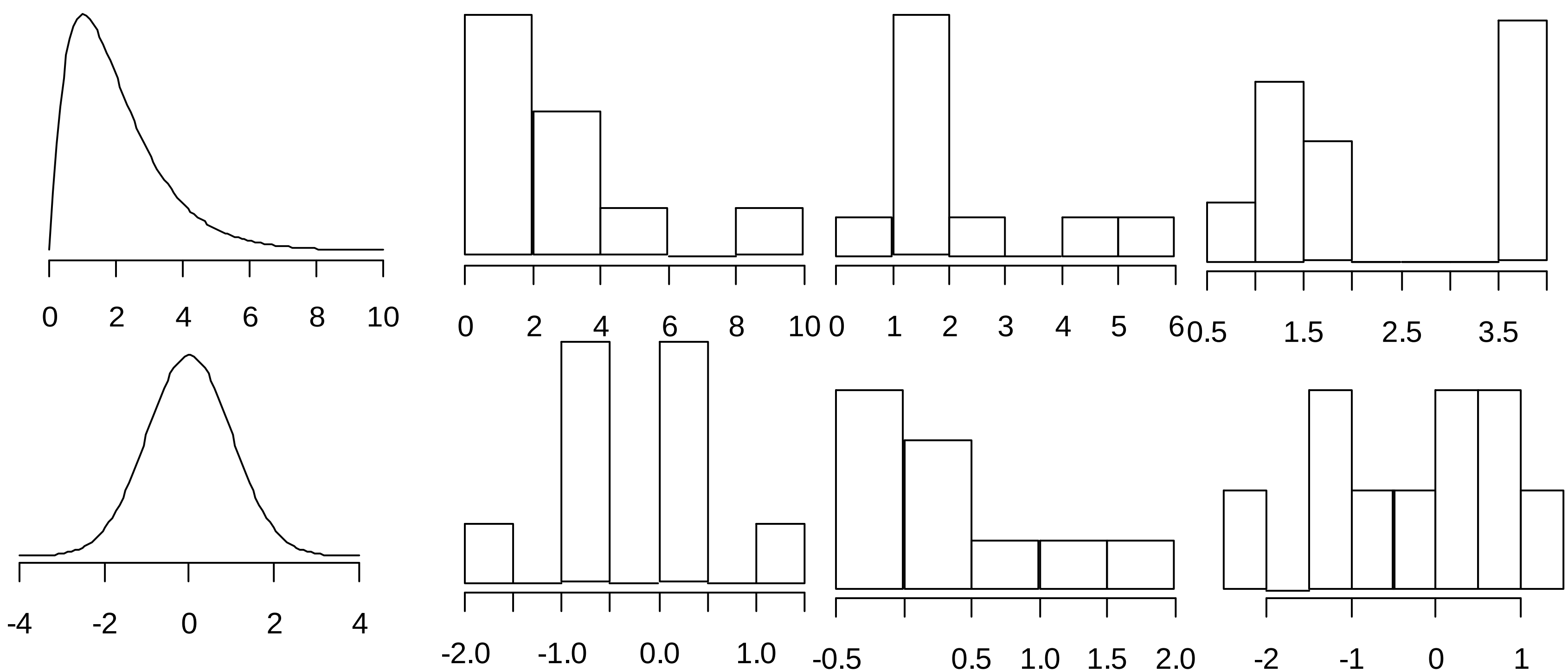
- For large sample sizes we do not need to worry about this, because s will be a very good estimate of σ, and t will be very close to N(0,1)

- For small sample sizes ($n < 30$), we can only use the t-distribution if the distribution of the data is approximately normal

Normality Assumption

- One small problem: for small sample sizes, it is very hard to tell if the data actually comes from a normal distribution!

Population **Sample Data, n = 10**



Small Samples

- If sample sizes are small, only use the *t*-distribution if the data looks reasonably symmetric and does not have any extreme outliers.

- Even then, remember that it is just an approximation!

- In practice/life, if sample sizes are small, you should just use simulation methods (bootstrapping and randomization)

Confidence Intervals

$$\boxed{\text{sample statistic} \pm t^{*} \times SE}$$

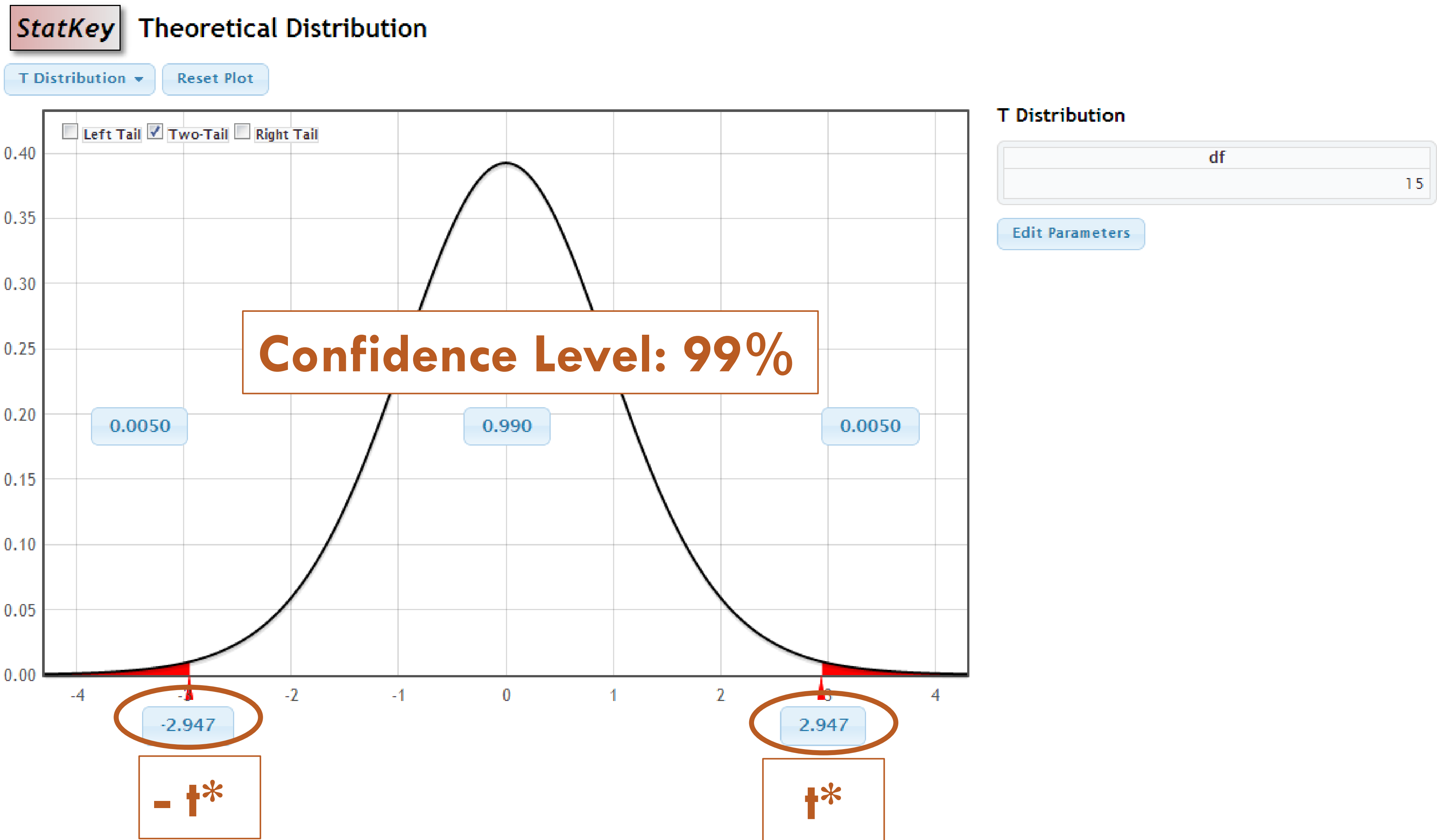
$$\boxed{\bar{X} \pm t^{*} \times \frac{s}{\sqrt{n}}}$$

df = *n* – 1

t^{*} is found as the appropriate percentile on a *t*-distribution with n – 1 degrees of freedom

IF *n* is large or the data is normal

How to Obtain t*



Gribbles

Gribbles are small marine worms that bore through wood, and the enzyme they secrete may allow us to turn inedible wood and plant waste into biofuel

- A sample of 50 gribbles finds an average length of 3.1 mm with a standard deviation of 0.72 mm.
- Give a 90% confidence interval for the average length of gribbles.

A sample of 50 gribbles finds an average length of 3.1 mm with a standard deviation of 0.72 mm. For a 90% confidence interval for the average length of gribbles, what is t*?

StatKey

A. 1.645
B. 1.677
C. 1.960
D. 1.690

A sample of 50 gribbles finds an average length of 3.1 mm with a standard deviation of 0.72 mm. For a 90% confidence interval for the average length of gribbles, what is the standard error.

A. 0.171
B. 0.720
C. 1.960
D. 0.102

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

A sample of 50 gribbles finds an average length of 3.1 mm with a standard deviation of 0.72 mm. For a 90% confidence interval for the average length of gribbles, what is the margin of error?

A. 0.171
B. 0.720
C. 1.960
D. 0.102

Gribbles

statistic $\pm$ *t** $\cdot$ *SE*

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

Margin of Error

$$\text{ME} = t^* \cdot \frac{s}{\sqrt{n}}$$

You can choose your sample size in advance, depending on your desired margin of error!

Given this formula for margin of error, solve for *n*.

Margin of Error

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

- Problem 1: For t^*, need to know n.
  - Solution: Use z^* instead of t^* (they are usually close)
- Problem 2: For s, need data.
  - Solution: estimate s.
    1. Use data from a previous study or similar population
    2. Take a small pre-sample to estimate s
    3. Estimate the range (max – min) and use $s \approx \text{range}/4$
    4. Make a reasonable guess.

Suppose we want to estimate average GPA at a college (where GPA's go from 0 to 4.0), with a margin of error of 0.1 with 95% confidence. How large a sample size do we need?

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

A. About 100

B. About 400

C. About 800

D. About 1000

Hypothesis Testing

$$\boxed{t = \frac{\text{sample statistic} - \text{null value}}{\text{SE}}}$$

$H_0 : \mu = \mu_0$

$$\boxed{t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}}$$

df = *n* – 1

The p-value is the area in the tail(s) beyond *t* in a *t*-distribution with *n* – 1 degrees of freedom,
IF *n* is large or the data is normal

Chips Ahoy!

A group of Air Force cadets bought bags of Chips Ahoy! cookies from all over the country to verify this claim. They hand counted the number of chips in 42 bags.

$$\bar{X} = 1261.6, \quad s = 117.6$$

Source: Warner, B. & Rutledge, J. (1999). "Checking the Chips Ahoy! Guarantee," *Chance*, **12**(1).

Chips Ahoy! Hypothesis Test

1. State hypotheses:

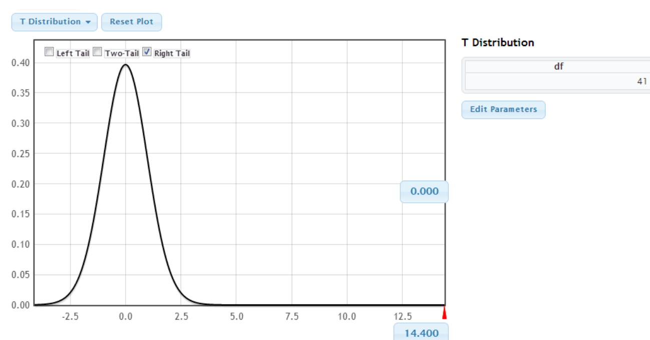
$$H_0 : \mu = 1000$$

$$H_a : \mu > 1000$$

2. Check conditions: $n = 42 \geq 30$

3. Calculate test statistic: $$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{1261.6 - 1000}{117.6/\sqrt{42}} = 14.4$$

4. Compute p-value:



$p - value \approx 0$

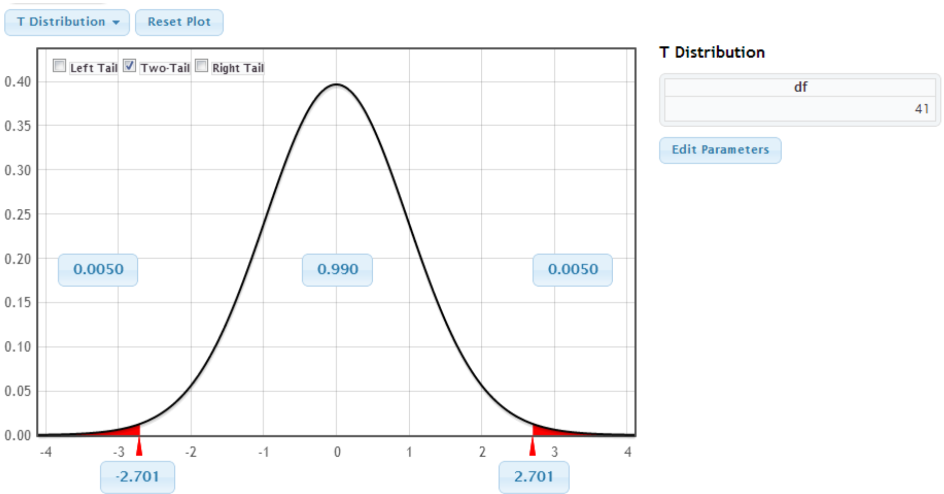
5. Interpret in context:

This provides extremely strong evidence that the average number of chips per bag of Chips Ahoy! cookies is significantly greater than 1000.

Chips Ahoy! Give a 99% confidence interval for the average number of chips in each bag.

1. Check conditions: $n = 42 \geq 30$

2. Find t*:



$$t^* = 2.7$$

4. Compute confidence interval:

$$\bar{X} \pm t^* \times \frac{s}{\sqrt{n}}$$

$$1261.6 \pm 2.7 \times \frac{117.6}{\sqrt{42}}$$

$$(1212.6, 1310.6)$$

5. Interpret in context:

We are 99% confident that the average number of chips per bag of Chips Ahoy! cookies is between 1212.6 and 1310.6 chips.

Which of the following properties is/are necessary for $t = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}}$ to have a *t*-distribution?

a) the data is normal

b) the sample size is large

c) the null hypothesis is true

d) a or b

e) d and c

Summary

- **Standard error** for a sample mean: $\frac{s}{\sqrt{n}}$
- **Central Limit Theorem for a mean:** If the sample size is large ($n \geq 30$), then $\bar{x} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. However, using s in place of σ, ***changes*** the distribution of the sample means ***to a t-distribution***.
  - The t-distribution is characterized by its ***degrees of freedom= n-1***
  - Conditions for the t-distribution: $n \geq 30$ or the data comes from a population that has an approximately normal distribution.