Inference for Two Proportions and One Mean

Stat 120

Sections 6.2, 6.3

Day 18

Tagging Penguins

A study looked at the 10 year survival rate of penguins tagged either with a metal tag or an electronic tag. 20% of the 167 metal tagged penguins survived, compared to 36% of the 189 electronic tagged penguins.

Is there a statistically significant difference in survival rates?

$$H_0 : p_M = p_E \qquad H_A : p_M \neq p_E$$

p = true survival rate

Source: Saraux, et. al. (2011). "Reliability of flipper-banded penguins as indicators of climate change," *Nature*, **469**, 203-206.

Tagging Penguins

20% of the 167 metal tagged penguins survived, compared to 36% of the 189 electronic tagged penguins.

Are the conditions met for using the normal distribution for inference?

a). Yes

b). No

With metal tags, 33 survived and 134 died. With electronic tags, 68 survived and 121 died. All counts are greater than 10.

Pooled Proportion

- We don't know p_M or p_E, so how do we compute the SE for our hypothesis test?
- Assume the two proportions are equal and use one proportion for both groups.
- Our best guess of this one proportion comes from combining data from both groups and computing the overall proportion, called the pooled proportion p.
- Hint: the pooled proportion will always be somewhere in between the two sample proportions.

Tagging Penguins

20% of the 167 metal tagged penguins survived, compared to 36% of the 189 electronic tagged penguins.

33 survived with metal tags and 68 with electronic
Pooled proportion:

$$\hat{p} = \frac{33 + 68}{167 + 189} = 0.2837$$

SE for our test:

$$SE = \sqrt{\frac{0.284(1 - .284)}{167} + \frac{0.284(1 - .284)}{189}} = 0.048$$

Tagging Penguins

20% of the 167 metal tagged penguins survived, compared to 36% of the 189 electronic tagged penguins. The pooled SE is 0.048.

Standardized test stat: $$z = \frac{(0.2 - 0.36) - 0}{0.048} = -3.34$$

P-value: 2 x proportion below -3.34

```
> 2*pnorm(-3.34,0,1)
[1] 0.0008377839
```

Reject the null.

Tagging Penguins

A difference in survival rates as extreme, or more extreme, than 16% would occur by chance only about 0.08% of the time. There is a statistically significant difference (z=-3.34, p=0.0008).

How much do the rates differ?

Compute a 95% CI for the difference…
How do we compute the SE?

We **can't** use the pooled version since we've concluded the proportions differ!

Tagging Penguins

20% of the 167 metal tagged penguins survived, compared to 36% of the 189 electronic tagged penguins.

95% CI for $p_M - p_E$:

$$(0.20 - 0.36) \pm 1.96\sqrt{\frac{0.20 \times 0.80}{167} + \frac{0.36 \times 0.64}{189}} = -0.16 \pm 1.96 \times 0.047$$

$$= (-0.251, -0.069)$$

We are 95% confident that between 6.9% to 25.1% fewer penguins survive when metal tags are used compared to electronic tags.

Test for a Difference in Proportions

$$H_0 : p_1 = p_2$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$$

- If observed counts in the two-way table are at least 10, then the p-value can be computed as the area in the tail(s) of a standard normal beyond *z*. Used pooled proportion for the SE.

Confidence Interval for $p_1 - p_2$

For large enough n_1 and n_2 : $statistic \pm z^* \times SE$

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Inference Using N(0,1)

If the distribution of the sample statistic is normal:

A confidence interval can be calculated by

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

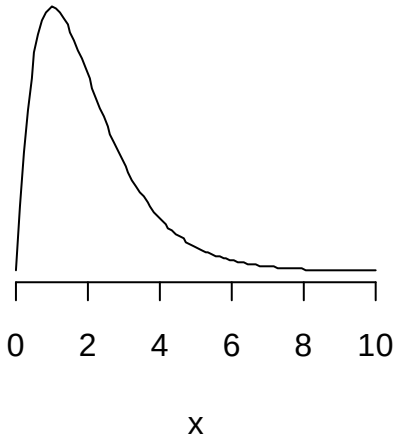
where z^* is a N(0,1) percentile depending on the level of confidence.

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

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

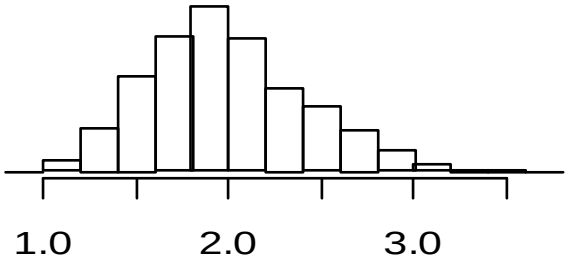
CLT for a Mean

Population

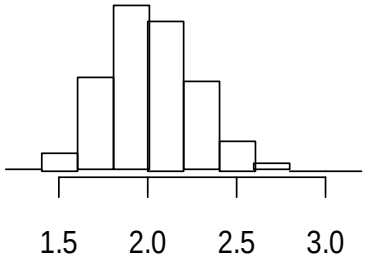


Distribution of Sample Means

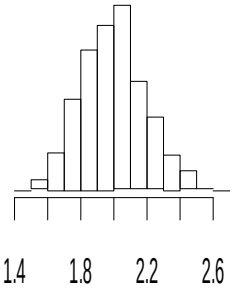
n = 10



n = 30



n = 50



SE of a Mean

The standard error for a sample mean can be calculated by

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

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

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

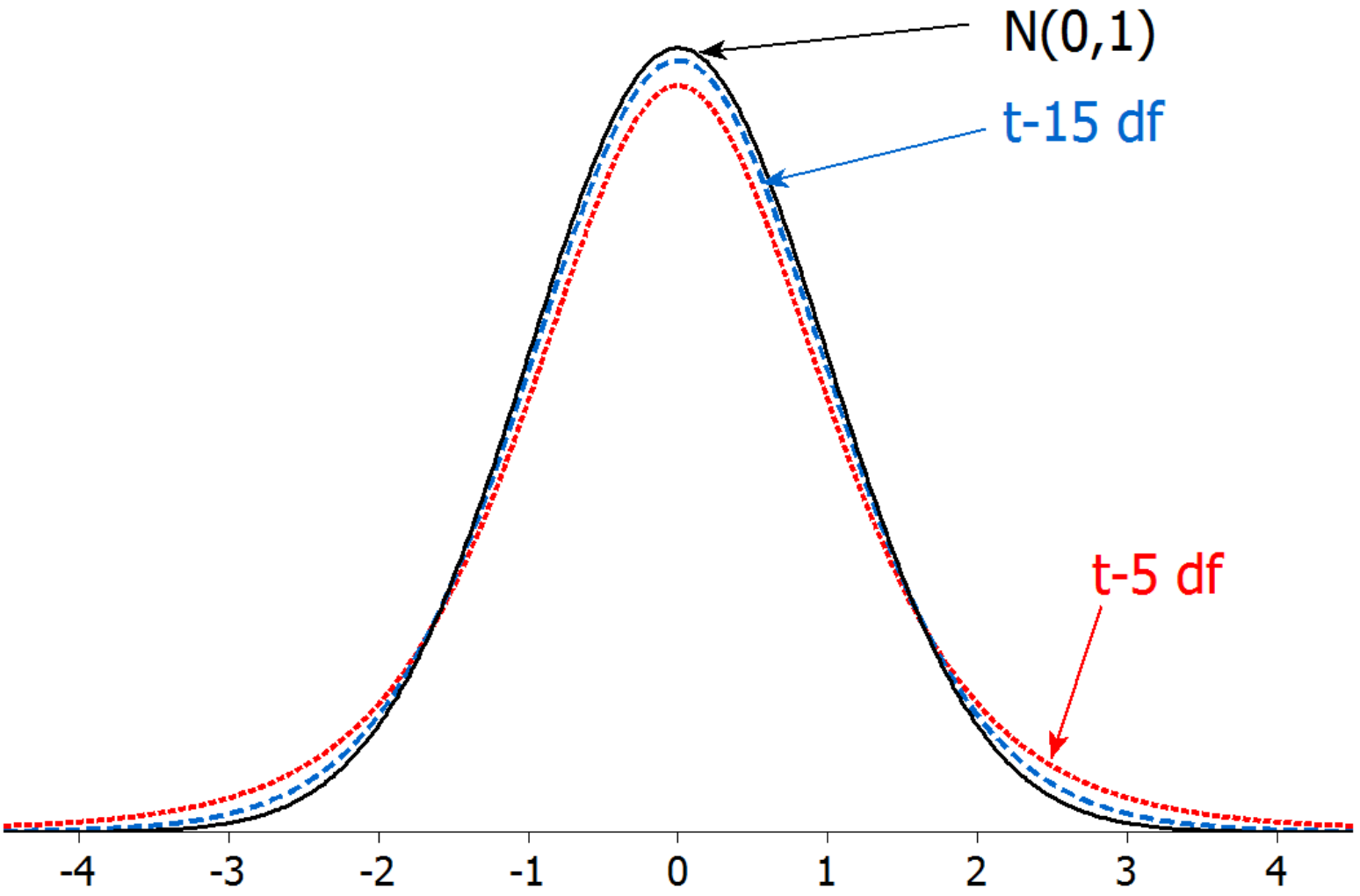
t-distribution

- 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

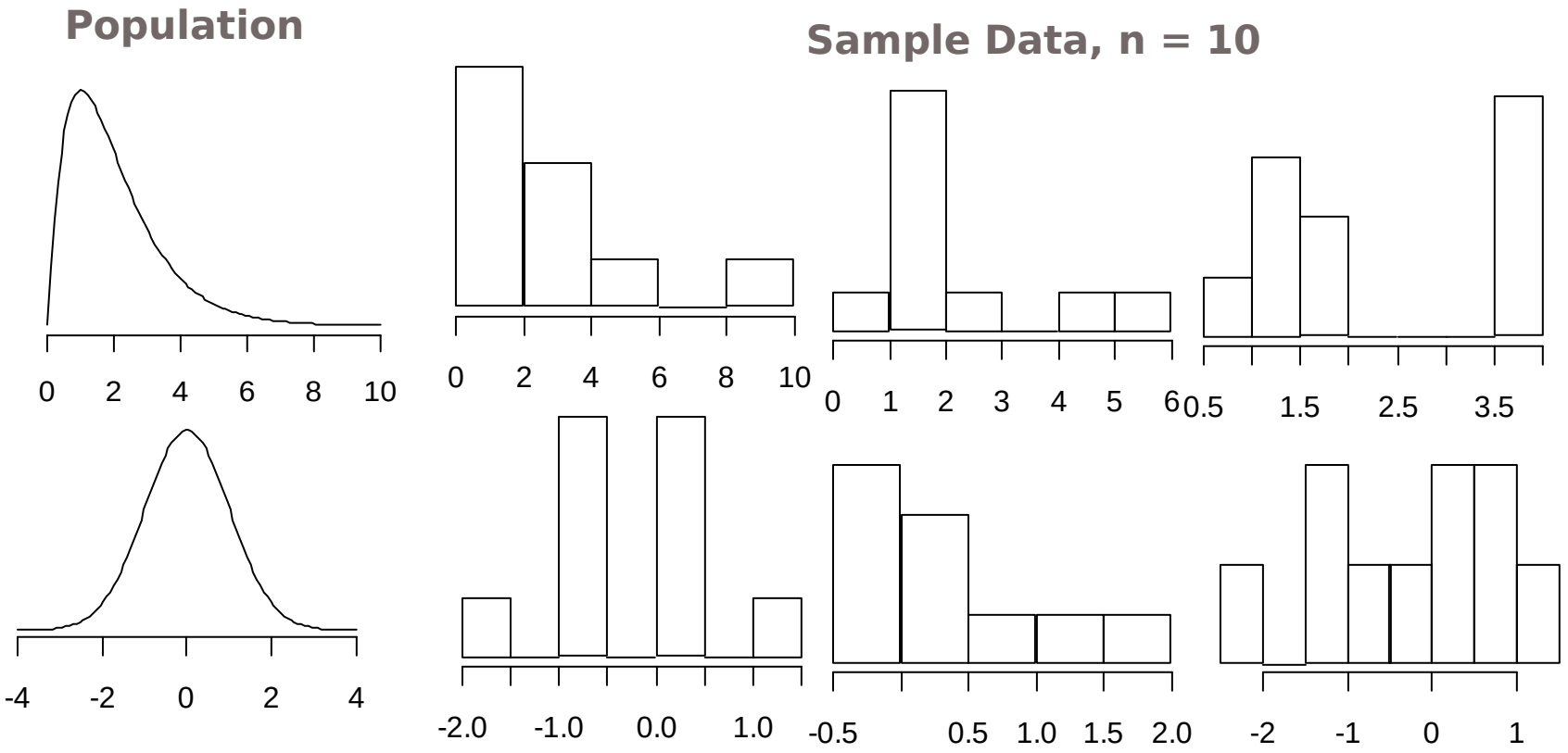


Normality Assumption

- 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!



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

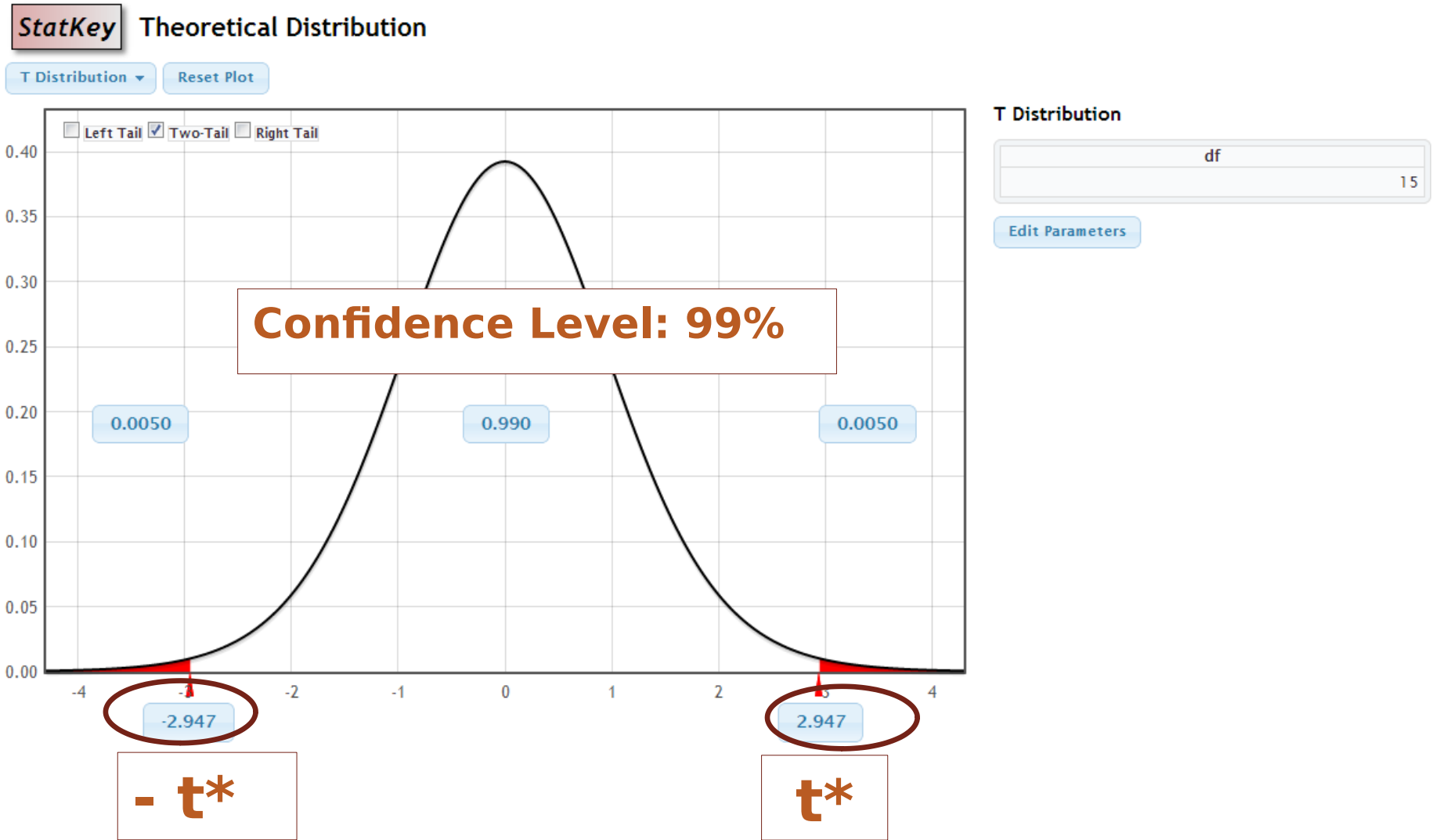
$$\text{sample statistic} \pm t^* \times SE$$

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

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

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

Margin of Error

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

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*.

$$statistic \pm t^* \times SE \qquad \bar{x} \pm t^* \times \frac{s}{\sqrt{n}}$$

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?

A. About 100
B. About 400
C. About 800
D. About 1000

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

Can use z* instead of t* here!

Hypothesis Testing

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

$H_0 : \mu = \mu_0$ $\qquad$ $t = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}}$ $\qquad$ **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:

$$H_0 : \mu = 1000$$

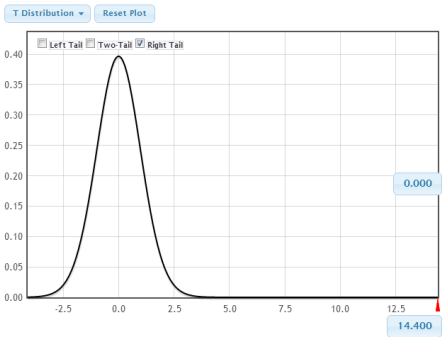
$$H_a : \mu > 1000$$

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:



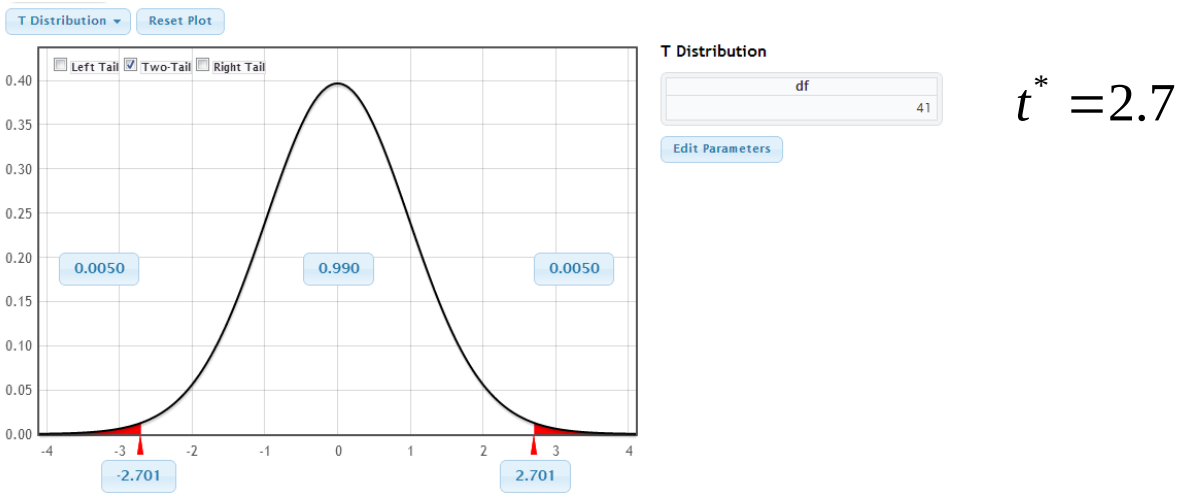
$p - value \approx 0$

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

Central Limit Theorem for a mean: If the sample size is large ($n \geq 30$), then . 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.