6 One-Sample Confidence Intervals (Cont'd)

MTH 3240 Environmental Statistics

Spring 2020

Objectives

Objectives:

- Determine the sample size needed to keep the margin of error no bigger than a desired value.
- Assess normality of data using graphs, and transform non-normal data to normality.
- Compute and interpret a one-sample z confidence interval for a proportion.

Sample Size Determination

- Planning a study often involves deciding how large the sample size n should be.

Sample Size Determination

- Planning a study often involves deciding how large the sample size n should be.
- **Larger samples** produce **smaller margins of error** in **estimates** of population parameters such as μ.

Sample Size Determination

- Planning a study often involves deciding how large the sample size n should be.
- **Larger samples** produce **smaller margins of error** in **estimates** of population parameters such as μ.
- But larger sample sizes are more expensive and time consuming to collect.

Sample Size Determination

- Planning a study often involves deciding how large the sample size n should be.
- **Larger samples** produce **smaller margins of error** in **estimates** of population parameters such as μ.
- But larger sample sizes are more expensive and time consuming to collect.
- Our goal is to determine the smallest n that's still large enough that the margin of error won't be unacceptably big.

- Suppose we want the **margin of error** in a CI for μ to be **no bigger than** some value B, i.e we want

$$\text{Margin of Error} \ \leq \ B.$$

- Suppose we want the **margin of error** in a CI for μ to be **no bigger than** some value B, i.e we want

$$\text{Margin of Error} \;\leq\; B.$$

- If the population standard deviation σ is known, then for a **95% CI**, we require n to be **large enough** that

$$1.96\,\sigma_{\bar{X}} \;\leq\; B \qquad \text{i.e.} \qquad 1.96\,\frac{\sigma}{\sqrt{n}} \;\leq\; B.$$

- Suppose we want the **margin of error** in a CI for μ to be **no bigger than** some value B, i.e we want

$$\text{Margin of Error} \;\leq\; B.$$

- If the population standard deviation σ is known, then for a **95% CI**, we require n to be **large enough** that

$$1.96\,\sigma_{\bar{X}} \;\leq\; B \qquad \text{i.e.} \qquad 1.96\,\frac{\sigma}{\sqrt{n}} \;\leq\; B.$$

Solving for n gives the **required sample size**:

$$n \;\geq\; \frac{(1.96\,\sigma)^2}{B^2}.$$

- For other confidence levels, replace 1.96 by the appropriate z critical value.

- For other confidence levels, replace 1.96 by the appropriate z critical value.

Sample Size Determination: The margin of error in a $100(1-\alpha)\%$ CI for μ will be no bigger than B if the sample size satisfies

$$n \;\geq\; \frac{(z_{\alpha/2}\sigma)^2}{B^2},$$

which should be *rounded up* to the nearest integer.

- For other confidence levels, replace 1.96 by the appropriate z critical value.

Sample Size Determination: The margin of error in a $100(1-\alpha)\%$ CI for μ will be no bigger than B if the sample size satisfies

$$n \ \geq \ \frac{(z_{\alpha/2}\sigma)^2}{B^2},$$

which should be *rounded up* to the nearest integer.

- In practice, we replace σ by a reasonable **guess**, for example based on a *pilot study* or preexisting studies.

Exercise

Consider a new study of background radiation levels along the Front Range.

Exercise

Consider a new study of background radiation levels along the Front Range.

Suppose we want an **estimate** of the **population mean background radiation level** μ to be **within 0.2 Bq/kg** of the true value (with 95% confidence).

Exercise

Consider a new study of background radiation levels along the Front Range.

Suppose we want an **estimate** of the **population mean background radiation level** μ to be **within 0.2 Bq/kg** of the true value (with 95% confidence).

At how many sites would we need to measure background radiation?

Exercise

Consider a new study of background radiation levels along the Front Range.

Suppose we want an **estimate** of the **population mean background radiation level** μ to be **within 0.2 Bq/kg** of the true value (with 95% confidence).

At how many sites would we need to measure background radiation?

(Use **0.76**, the sample standard deviation from the earlier study, as a **guess** for σ.)

Exercise

Consider a new study of background radiation levels along the Front Range.

Suppose we want an **estimate** of the **population mean background radiation level** μ to be **within 0.2 Bq/kg** of the true value (with 95% confidence).

At how many sites would we need to measure background radiation?

(Use **0.76**, the sample standard deviation from the earlier study, as a **guess** for σ.)

Hint: You should end up with $n = 55.5$, which rounds up to $\boldsymbol{n = 56}$.

Checking Normality of Data

- Many statistical procedures (e.g. the one-sample t procedure) require that the data are a sample from a **normal population** (or that n is large).

Checking Normality of Data

- Many statistical procedures (e.g. the one-sample t procedure) require that the data are a sample from a **normal population** (or that n is large).

  Two commonly used tools to check this assumption are:

  1. **Histograms** (should look bell-shaped)
  2. **Normal probability plots** (the points should hug a line)

- If a **histogram** looks reasonably symmetric and **bell-shaped**, the **normality** assumption is **tenable.**

- If a **histogram** looks reasonably symmetric and **bell-shaped**, the **normality** assumption is **tenable.**
- If the points in a **normal probability plot** follow approximately a **straight line**, the **normality** assumption is **tenable.**

- If a **histogram** looks reasonably symmetric and **bell-shaped**, the **normality** assumption is **tenable.**

- If the points in a **normal probability plot** follow approximately a **straight line**, the **normality** assumption is **tenable.**

  Curved patterns in the **normal probability plot** indicate various types of **non-normality**.

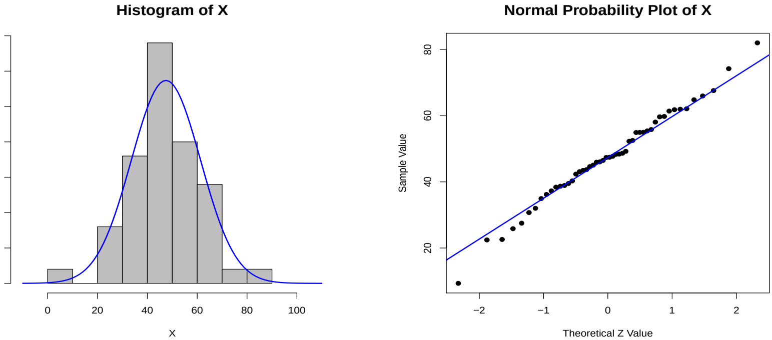


Figure: Histogram of symmetric, approximately normal data (left). Normal probability plot of the same data (right).

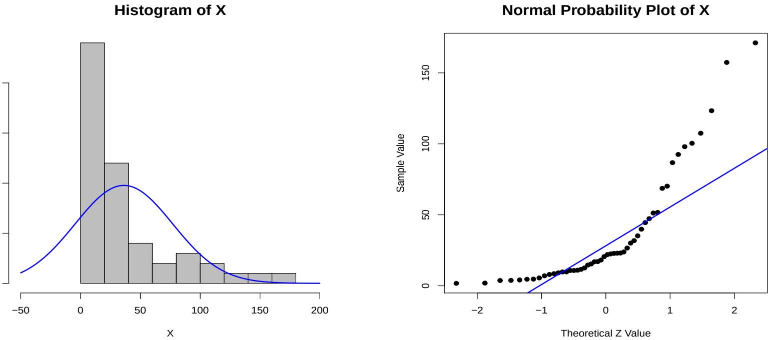


Figure: Histogram of non-normal, right skewed data (left). Normal probability plot of the same data (right).

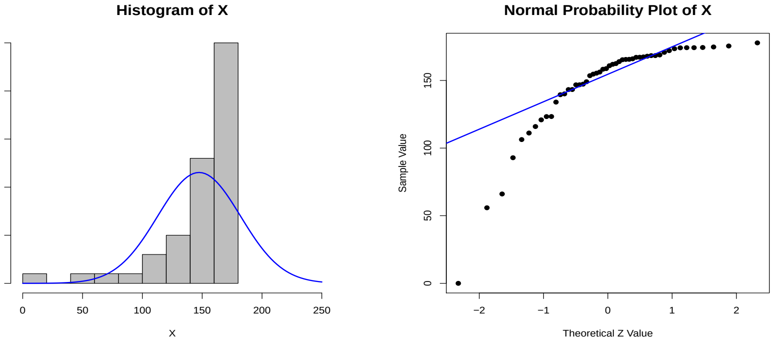


Figure: Histogram of non-normal, left skewed data (left). Normal probability plot of the same data (right).

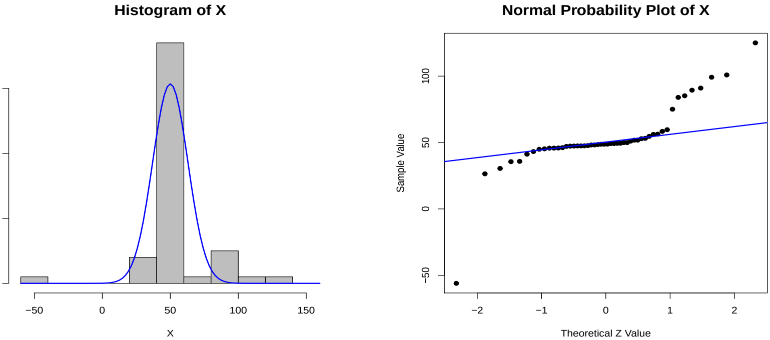


Figure: Histogram of non-normal, ”heavy tailed” data (left). Normal probability plot of the same data (right).

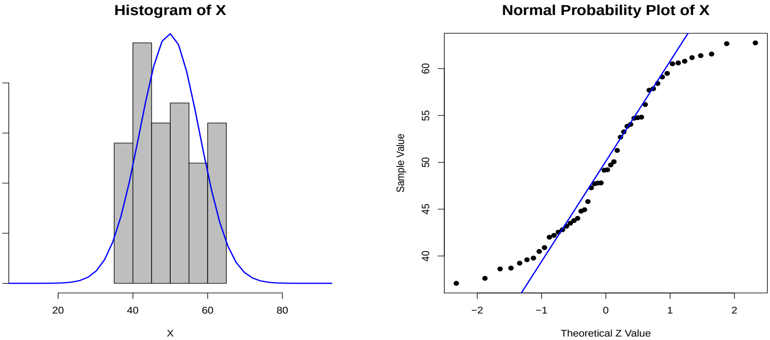
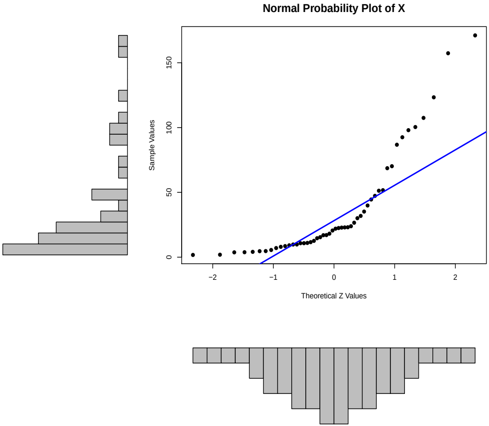


Figure: Histogram of non-normal, ”light tailed” data (left). Normal probability plot of the same data (right).

- A ***normal probability plot*** is a plot of the **observed data values** (y-axis) versus the theoretical values we'd **expect** to get **if** our sample was from a **normal population** (x-axis).

- A ***normal probability plot*** is a plot of the **observed data values** (y-axis) versus the theoretical values we'd **expect** to get **if** our sample was from a **normal population** (x-axis).

  (On the next slide, the observed data (y) follow a right skewed distribution. The theoretical (x) values correspond to an exact normal distribution.)


Normal Probability Plot of X
Sample Values
0
50
100
150
−2
−1
0
1
2
Theoretical Z Values

Transforming Data to Normality

- Most of the statistical procedures that require the normality assumption are ***robust*** to mild departures from normality.

Transforming Data to Normality

- Most of the statistical procedures that require the normality assumption are ***robust*** to mild departures from normality.

  This means they're still **approximately valid**, even when n is **small**, as long as the non-normality isn't severe.

- But if the population is **severely non-normal** (i.e. **skewed**) and n **isn't large**, those procedures **shouldn't be used.**

- But if the population is **severely non-normal** (i.e. **skewed**) and n **isn't large**, those procedures **shouldn't be used.**

  Instead, we can either:

- But if the population is **severely non-normal** (i.e. **skewed**) and n **isn't large**, those procedures **shouldn't be used.**

  Instead, we can either:

  - ***Transform*** the data first, for example by taking their **logs**, so that the **transformed data** are more **normally distributed**, or

- But if the population is **severely non-normal** (i.e. **skewed**) and n **isn't large**, those procedures **shouldn't be used.**

  Instead, we can either:

  - ***Transform*** the data first, for example by taking their **logs**, so that the **transformed data** are more **normally distributed**, or
  - Use a so-called ***nonparametric*** procedure, i.e. one that **doesn't require a normality assumption**.

- **Right skewed** data can be modeled as a sample from a **lognormal** population.

- **Right skewed** data can be modeled as a sample from a **lognormal** population.
- Thus their (natural) **logs** can be treated as a sample from a **normal** population.

- **Right skewed** data can be modeled as a sample from a **lognormal** population.
- Thus their (natural) **logs** can be treated as a sample from a **normal** population.

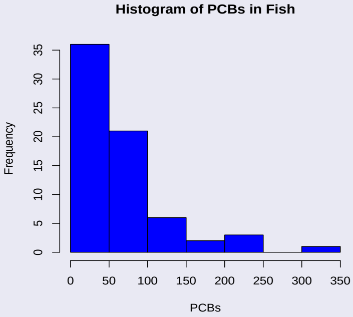
  In this case, we can carry out the statistical procedures on the **logs** of the data.

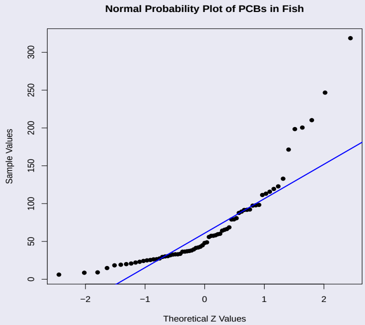
Example

In a U.S. EPA study, Polychlorinated biphenyls (**PCBs**) were measured in fish from $n = 69$ U.S. lakes. There are more than 200 types of PCBs. The data below are measured values of the total sum of all **PCBs** (ppb) found in the fish.

20.0	6.1	25.0	37.4	30.2	20.8	41.4	29.5	24.2	26.3
8.6	36.4	66.4	30.6	25.5	68.6	23.1	43.0	39.5	36.5
26.5	22.1	19.2	33.0	9.1	42.0	48.8	55.9	31.8	60.1
97.3	18.4	27.5	79.0	97.8	44.9	58.2	57.4	57.5	33.6
115.7	14.8	91.6	92.2	37.0	87.7	111.4	48.0	38.1	122.8
113.1	79.2	98.3	33.0	64.2	119.4	80.7	171.4	132.9	91.8
32.6	59.5	200.5	65.4	198.5	89.4	210.3	246.7	318.7	

We want a **95% CI** for the (unknown) **population mean PCB** concentration for U.S. lakes. A **histogram** and **normal probability** plot of the data are below.



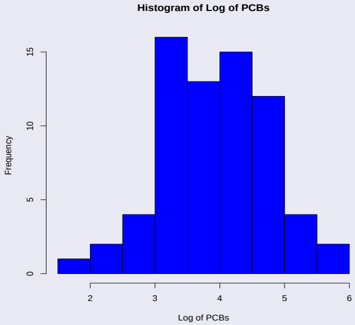


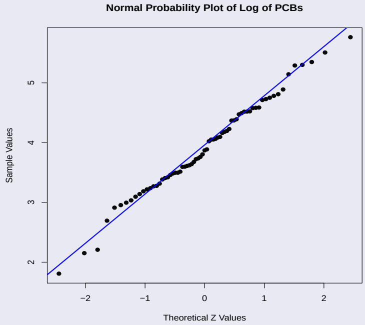
Based on the plots, it would be **unreasonable** to assume the data are a sample from a normal population.

Here are the **logs** of the data.

3.0	1.8	3.2	3.6	3.4	3.0	3.7	3.4	3.2	3.3
2.2	3.6	4.2	3.4	3.2	4.2	3.1	3.8	3.7	3.6
3.3	3.1	3.0	3.5	2.2	3.7	3.9	4.0	3.5	4.1
4.6	2.9	3.3	4.4	4.6	3.8	4.1	4.1	4.1	3.5
4.8	2.7	4.5	4.5	3.6	4.5	4.7	3.9	3.6	4.8
4.7	4.4	4.6	3.5	4.2	4.8	4.4	5.1	4.9	4.5
3.5	4.1	5.3	4.2	5.3	4.5	5.3	5.5	5.8	

A **histogram** and **normal probability plot** of these **log transformed** values are below.





The assumption of a **normal population** appears to be met for the **logs** of the **PCB** concentrations.

The assumption of a **normal population** appears to be met for the **logs** of the **PCB** concentrations.

We'll let μ denote the true (unknown) **population mean *log* PCB** concentration.

The assumption of a **normal population** appears to be met for the **logs** of the **PCB** concentrations.

We'll let μ denote the true (unknown) **population mean *log* PCB** concentration.

The **sample mean** and **standard deviation** of the **log PCB** concentrations are

$$\bar{Y} = \mathbf{3.92} \qquad \text{and} \qquad S = \mathbf{0.80}$$

The assumption of a **normal population** appears to be met for the **logs** of the **PCB** concentrations.

We'll let μ denote the true (unknown) **population mean *log* PCB** concentration.

The **sample mean** and **standard deviation** of the **log PCB** concentrations are

$$\bar{Y} = \mathbf{3.92} \qquad \text{and} \qquad S = \mathbf{0.80}$$

Thus the **estimate** of the (unknown) **population mean *log* PCB** concentration μ is **3.92** (log ppb).

The (estimated) **standard error** is

$$\boldsymbol{S_{\bar{Y}}} \;=\; \frac{S}{\sqrt{n}} \;=\; \frac{0.80}{\sqrt{69}} \;=\; \mathbf{0.10},$$

The (estimated) **standard error** is

$$\boldsymbol{S_{\bar{Y}}} \;=\; \frac{S}{\sqrt{n}} \;=\; \frac{0.80}{\sqrt{69}} \;=\; \mathbf{0.10},$$

so the **95% one-sample t CI for μ** is

$$\begin{aligned} \bar{Y} \;\pm\; t_{0.025,68} S_{\bar{Y}} &= 3.92 \;\pm\; 1.995(0.10) \\ &= 3.92 \;\pm\; 0.20 \\ &= \mathbf{(3.72,\;\; 4.12)} \end{aligned}$$

The (estimated) **standard error** is

$$\boldsymbol{S_{\bar{Y}}} = \frac{S}{\sqrt{n}} = \frac{0.80}{\sqrt{69}} = \mathbf{0.10},$$

so the **95% one-sample t CI for μ** is

$$\begin{aligned} \bar{Y} \pm t_{0.025,68} S_{\bar{Y}} &= 3.92 \pm 1.995(0.10) \\ &= 3.92 \pm 0.20 \\ &= \mathbf{(3.72,\ 4.12)} \end{aligned}$$

We can be 95% confident that the (unknown) **population mean *log* PCB** concentration μ is between **3.72** and **4.12** (**log ppb**).

- When we perform a statistical analysis on **log transformed** data, the results pertain to the **log** measurement scale for the data.

- When we perform a statistical analysis on **log transformed** data, the results pertain to the **log** measurement scale for the data.
- Sometimes it's possible **back-transform** the results to the original scale by taking the **antilog**.

Example (Cont'd)

The sample mean **log PCB** concentration was $\bar{\boldsymbol{Y}} = \mathbf{3.92}$ (**log ppb**).

We can convert $\bar{Y}$ back to the original scale (ppb), so that it's easier to interpret, by taking its **antilog**:

$$\begin{aligned} e^{\bar{Y}} &= e^{3.92} \\ &= \mathbf{50.4} \quad (\textbf{ppb}), \end{aligned}$$

where e is the so-called ***exponential constant***,

$$e = 2.71828.$$

We can take the **antilogs** of the 95% **CI** endpoints

$$(3.72, \;\; 4.12)$$

too, giving the new interval

$$(e^{3.72}, \;\; e^{4.12}) \; = \; \mathbf{(41.3, \;\; 61.6)}.$$

We can take the **antilogs** of the 95% **CI** endpoints

$$(3.72, \;\; 4.12)$$

too, giving the new interval

$$(e^{3.72}, \;\; e^{4.12}) \;=\; \mathbf{(41.3, \;\; 61.6)}.$$

The values **41.3** and **61.6** are easier to interpret because they're measured in **ppb**.

We can take the **antilogs** of the 95% **CI** endpoints

$$(3.72, \;\; 4.12)$$

too, giving the new interval

$$(e^{3.72}, \;\; e^{4.12}) \; = \; \mathbf{(41.3, \;\; 61.6)}.$$

The values **41.3** and **61.6** are easier to interpret because they're measured in **ppb**.

We're (approximately) 95% confident that the true population mean PCB concentration (in **ppb**) is in this range.

One-Sample Z CI for p

- A **categorical** variable is ***dichotomous*** if it takes only **two values**.

One-Sample Z CI for p

- A **categorical** variable is ***dichotomous*** if it takes only **two values.**

  Examples:

One-Sample Z CI for p

- A **categorical** variable is ***dichotomous*** if it takes only **two values.**

  Examples:

  - A random sample of wells from across the state of Iowa is selected, and each tested for the **presence** or **absence** of E. coli.

    The variable, **presence** or **absence** of E. coli, is **dichotomous**.

- **Examples (Cont'd)**:
  - A random sample of biosolids fertilizer specimens is selected from farmlands in Ohio, and each specimen tested (**positive** or **negative**) for salmonella.

    The variable, **positive** or **negative** test result, is **dichotomous**.

- **Examples (Cont'd)**:
  - A random sample of biosolids fertilizer specimens is selected from farmlands in Ohio, and each specimen tested (**positive** or **negative**) for salmonella.

    The variable, **positive** or **negative** test result, is **dichotomous**.
  - A sample of people who didn't participate in decision making involving an environmental assessment of a proposed hog slaughtering facility were asked (**yes** or **no**) whether their reason for not participating was that ”The ultimate decisions were foregone.”

    The variable, **yes** or **no** response, is **dichotomous**.

- We’ll refer to the two values of a **dichotomous** variable (generically) as ***success*** and ***failure***.

- When a **dichotomous** variable is measured on a random sample from a **population** whose **proportion** of **successes** is p, the (point) **estimate** of p is the ***sample proportion***, denoted $\hat{P}$.

Sample Proportion: For a data set of n observations of a dichotomous variable taking values *success* and *failure*,

$$\hat{P} = \frac{\text{Number of successes in the sample}}{\text{Sample size } n}.$$

Example

In a study of bees, a honey solution was sprayed on vegetation along several transects. Bees attracted to the honey were caught with an insect net, and their species later identified.

Example

In a study of bees, a honey solution was sprayed on vegetation along several transects. Bees attracted to the honey were caught with an insect net, and their species later identified.

Among the $n = 1,631$ bees caught, **546** were of the species *Trigona* (*Tetragonula*) *laeviceps*. The **sample proportion** is

$$\hat{P} = \frac{546}{1,631} = 0.33\,,$$

Example

In a study of bees, a honey solution was sprayed on vegetation along several transects. Bees attracted to the honey were caught with an insect net, and their species later identified.

Among the $n = 1,631$ bees caught, **546** were of the species *Trigona* (*Tetragonula*) *laeviceps*. The **sample proportion** is

$$\hat{P} \;=\; \frac{546}{1,631} \;=\; 0.33\,,$$

which is an **estimate** of the (unknown) **population proportion** p.

- A **CI** for p is given by the following.

One-Sample Z CI for p:

$$\hat{P} \;\pm\; z_{\alpha/2}\, S_{\hat{P}}, \qquad \text{where} \qquad S_{\hat{p}} \;=\; \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}.$$

- A **CI** for p is given by the following.

One-Sample Z CI for p:

$$\hat{P} \ \pm \ z_{\alpha/2}\,S_{\hat{P}}, \qquad \text{where} \qquad S_{\hat{p}} \ = \ \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}.$$

- The CI is valid if the sample is from a dichotomous population and the sample size n is *large*.

- The ***margin of error*** is:

Margin of Error: For the one-sample z CI for p, the margin of error is

$$\text{Margin of Error} \;=\; z_{\alpha/2}\, S_{\hat{P}} \;=\; z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}.$$

Example (Cont'd)

In the sample of $n = 1,631$ bees, the **sample proportion** that were *T.* (*T.*) *laeviceps* was

$$\hat{P} = \frac{546}{1,631} = 0.33\,,$$

which is an **estimate** of the **population proportion** p.

The (estimated) **standard error** of $\hat{P}$ is

$$S_{\hat{P}} \;=\; \sqrt{\frac{0.33(1-0.33)}{1,631}} \;=\; \mathbf{0.01}\,,$$

indicating how far off the mark the **estimate** is **expected** to be.

A **95% one-sample** z **CI** for the unknown **population proportion** p that are *T.* (*T.*) *laeviceps* is

$$\begin{aligned}\hat{P} \;\pm\; z_{0.025}\, S_{\hat{P}} &= 0.33 \pm 1.96\,(0.01) \\ &= 0.33 \pm 0.02 \\ &= (\mathbf{0.31},\, \mathbf{0.35}).\end{aligned}$$

A **95% one-sample z CI** for the unknown **population proportion p** that are *T.* (*T.*) *laeviceps* is

$$\begin{aligned}\hat{P} \;\pm\; z_{0.025}\, S_{\hat{P}} &= 0.33 \,\pm\, 1.96\,(0.01) \\ &= 0.33 \,\pm\, 0.02 \\ &= (\mathbf{0.31},\, \mathbf{0.35}).\end{aligned}$$

The **margin of error** is **0.02**, indicating that we **wouldn't expect** the **estimate** to be off the mark by **more** than 0.02.

A **95% one-sample z CI** for the unknown **population proportion p** that are *T.* (*T.*) *laeviceps* is

$$\begin{aligned}\hat{P} \;\pm\; z_{0.025}\, S_{\hat{P}} \;&=\; 0.33 \pm 1.96\,(0.01)\\ &=\; 0.33 \pm 0.02\\ &=\; (\mathbf{0.31},\, \mathbf{0.35}).\end{aligned}$$

The **margin of error** is **0.02**, indicating that we **wouldn't expect** the **estimate** to be off the mark by **more** than 0.02.

We can be **95% confident** that the unknown **population proportion p** is between **0.31** and **0.35**.