How can population proportions be estimated using statistics?

Authored by

stats writer

June 29, 2024

Population proportions can be estimated using statistics by taking a representative sample from the population and using that sample to make inferences about the entire population. This is done by calculating the proportion of the sample that possesses a certain characteristic or trait, and then using statistical methods to estimate the proportion of the entire population that possesses the same characteristic. This method allows for a more accurate estimation of population proportions, as it takes into account the variability and potential biases in the population. Additionally, the use of statistical techniques such as confidence intervals and hypothesis testing can provide a level of certainty and accuracy in the estimated population proportions. Overall, the use of statistics allows for a systematic and reliable approach to estimating population proportions.

Statistics - Estimating Population Proportions

A population proportion is the share of a population that belongs to a particular category.

Confidence intervals are used to estimate population proportions.

Estimating Population Proportions

A statistic from a sample is used to estimate a parameter of the population.

The most likely value for a parameter is the **point estimate**.

Additionally, we can calculate a **lower bound** and an **upper bound** for the estimated parameter.

The **margin of error** is the difference between the lower and upper bounds from the point estimate.

Together, the lower and upper bounds define a **confidence interval**.

Calculating a Confidence Interval

The following steps are used to calculate a confidence interval:

Check the conditionsFind the point estimateDecide the confidence levelCalculate the margin of errorCalculate the confidence interval

For example:

Population: Nobel Prize winners**Category**: Born in the United States of America

We can take a sample and see how many of them were born in the US.

The sample data is used to make an estimation of the share of **all** the Nobel Prize winners born in the US.

By randomly selecting 30 Nobel Prize winners we could find that:

6 out of 30 Nobel Prize winners in the sample were born in the US

From this data we can calculate a confidence interval with the steps below.

1. Checking the Conditions

The conditions for calculating a confidence interval for a proportion are:

The sample is randomly selectedThere is only two options:
Being in the categoryNot being in the categoryThe sample needs at least:
5 members in the category5 members not in the category

In our example, we randomly selected 6 people that were born in the US.

The rest were not born in the US, so there are 24 in the other category.

The conditions are fulfilled in this case.

Note: It is possible to calculate a confidence interval without having 5 of each category. But special adjustments need to be made.

2. Finding the Point Estimate

The point estimate is the sample proportion ($\hat{p}$).

The formula for calculating the sample proportion is the number of occurrences (x) divided by the sample size (n):

$$\displaystyle \hat{p} =\frac{x}{n}$$

In our example, 6 out of 30 were born in the US: x is 6, and n is 30.

So the point estimate for the proportion is:

$$\displaystyle \hat{p} = \frac{x}{n} = \frac{6}{30} = \underline{0.2} = 20\%$$

So 20% of the sample were born in the US.

3. Deciding the Confidence Level

The confidence level is expressed with a percentage or a decimal number.

For example, if the confidence level is 95% or 0.95:

The remaining probability ((alpha)) is then: 5%, or 1 - 0.95 = 0.05.

Commonly used confidence levels are:

90% with (alpha) = 0.195% with (alpha) = 0.0599% with (alpha) = 0.01

Note: A 95% confidence level means that if we take 100 different samples and make confidence intervals for each:

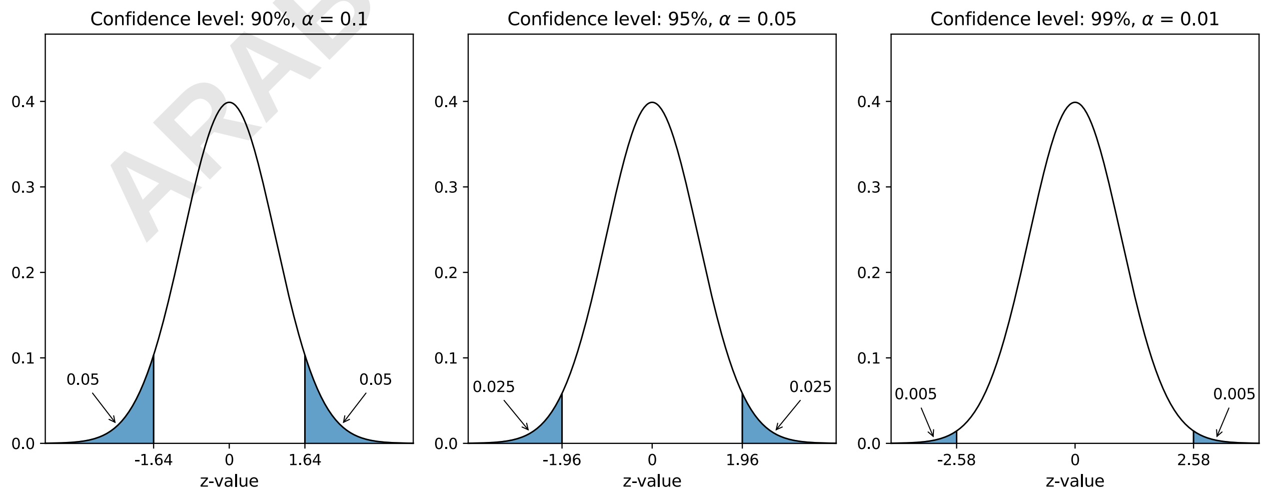
The true parameter will be inside the confidence interval 95 out of those 100 times.

We use the standard normal distribution to find the **margin of error** for the confidence interval.

The remaining probabilities ((alpha)) are divided in two so that half is in each tail area of the distribution.

The values on the z-value axis that separate the tails area from the middle are called **critical z-values**.

Below are graphs of the standard normal distribution showing the tail areas ((alpha)) for different confidence levels.



4. Calculating the Margin of Error

The margin of error is the difference between the point estimate and the lower and upper bounds.

The margin of error (E) for a proportion is calculated with a critical z-value and the **standard error**:

$$\displaystyle E = Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The critical z-value ($Z_{\alpha/2}$) is calculated from the standard normal distribution and the confidence level.

The standard error ($\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$) is calculated from the point estimate ($\hat{p}$) and sample size (n).

In our example with 6 US-born Nobel Prize winners out of a sample of 30 the standard error is:

$$\displaystyle \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.2(1-0.2)}{30}} = \sqrt{\frac{0.2 \cdot 0.8}{30}} = \sqrt{\frac{0.16}{30}} = \sqrt{0.00533..} \approx \underline{0.073}$$

If we choose 95% as the confidence level, the (α) is 0.05.

So we need to find the critical z-value ($Z_{0.05/2} = Z_{0.025}$)

The critical z-value can be found using a Z-table or with a programming language function:

Example

With Python use the Scipy Stats library `norm.ppf()` function find the Z-value for an (α)/2 = 0.025

```
import scipy.stats as stats

print(stats.norm.ppf(1-0.025))
```

Example

With R use the built-in `qnorm()` function to find the Z-value for an (α)/2 = 0.025

```
qnorm(1-0.025)
```

Using either method we can find that the critical Z-value ($Z_{\alpha/2}$) is ($\approx \underline{1.96}$)

The standard error ($\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$) was ($\approx \underline{0.073}$)

So the margin of error ((E)) is:

(displaystyle E = Z_{alpha/2} cdot sqrt{frac{hat{p}(1-hat{p})}{n}} approx 1.96 cdot 0.073 = underline{0.143})

5. Calculate the Confidence Interval

The lower and upper bounds of the confidence interval are found by subtracting and adding the margin of error ((E)) from the point estimate ((hat{p})).

In our example the point estimate was 0.2 and the margin of error was 0.143, then:

The lower bound is:

(hat{p} - E = 0.2 - 0.143 = underline{0.057})

The upper bound is:

(hat{p} + E = 0.2 + 0.143 = underline{0.343})

The confidence interval is:

() or ()

And we can summarize the confidence interval by stating:

The **95%** confidence interval for the proportion of Nobel Prize winners born in the US is between **5.7% and 34.4%**

Calculating a Confidence Interval with Programming

A confidence interval can be calculated with many programming languages.

Using software and programming to calculate statistics is more common for bigger sets of data, as calculating manually becomes difficult.

Example

With Python, use the scipy and math libraries to calculate the confidence interval for an estimated proportion.

Here, the sample size is 30 and the occurrences is 6.

import scipy.stats as stats

```
import math

# Specify sample occurrences (x), sample size (n) and confidence level

x = 6

n = 30

confidence_level = 0.95

# Calculate the point estimate, alpha, the critical z-value, the
standard error, and the margin of error

point_estimate = x/n

alpha = (1-confidence_level)

critical_z = stats.norm.ppf(1-alpha/2)

standard_error = math.sqrt((point_estimate*(1-point_estimate)/n))

margin_of_error = critical_z * standard_error

# Calculate the lower and upper bound of the confidence interval

lower_bound = point_estimate - margin_of_error

upper_bound = point_estimate + margin_of_error

# Print the results

print("Point Estimate: {:.3f}".format(point_estimate))

print("Critical Z-value: {:.3f}".format(critical_z))

print("Margin of Error: {:.3f}".format(margin_of_error))

print("Confidence Interval: ".format(lower_bound,upper_bound))

print("The {:.1%} confidence interval for the population proportion is:".format(confidence_level))
```

print("between {:.3f} and {:.3f}".format(lower_bound,upper_bound))

Example

With R, use the built-in math and statistics functions to calculate the confidence interval for an estimated proportion.

Here, the sample size is 30 and the occurrences is 6.

Specify sample occurrences (x), sample size (n) and confidence level

x = 6

n = 30

confidence_level = 0.95

Calculate the point estimate, alpha, the critical z-value, the standard error, and the margin of error

point_estimate = x/n

alpha = (1-confidence_level)

critical_z = qnorm(1-alpha/2)

standard_error = sqrt(point_estimate*(1-point_estimate)/n)

margin_of_error = critical_z * standard_error

Calculate the lower and upper bound of the confidence interval

lower_bound = point_estimate - margin_of_error

upper_bound = point_estimate + margin_of_error

Print the results

sprintf("Point Estimate: %0.3f", point_estimate)

sprintf("Critical Z-value: %0.3f", critical_z) sprintf("Margin of Error: %0.3f", margin_of_error)

sprintf("Confidence Interval: ", lower_bound, upper_bound)

sprintf("The %0.1f%% confidence interval for the population proportion is:", confidence_level*100)

sprintf("between %0.4f and %0.4f", lower_bound, upper_bound)