Statistical Methods

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

August 27, 2019

Topics

Objectives

Objectives:

- Calculate and interpret a CI for a population proportion p when n is large.
- Carry out a hypothesis test for a population proportion p when the sample size is large.

CI for a Population Proportion p

- Consider a population of **successes** and **failures**, and let p denote the ***population proportion*** of successes.

CI for a Population Proportion p

- Consider a population of **successes** and **failures**, and let p denote the ***population proportion*** of successes.
- Suppose our goal is to **estimate** p using a random sample from the population.

- The **point estimator** of p is the ***sample proportion***, denoted $\hat{P}$.

- The **point estimator** of p is the ***sample proportion***, denoted $\hat{P}$.

Sample Proportion:

$$\hat{P} = \frac{\text{\# of Successes in the Sample}}{\text{Sample Size}} = \frac{X}{n}$$

where

$$X = \text{The number of successes in the sample.}$$

Example

A random sample of $n = 10$ patrons at a restaurant were asked whether they smoke cigarettes (Yes or No). Here are the data.

Yes	No	Yes	No	Yes	No	No	Yes	No	No

The **sample proportion** of smokers is

$$\hat{p} = \frac{4}{10} = 0.4.$$

We'd **estimate** that the true (unknown) proportion p that smokes in the population is 0.4, or 40%.

- The **sampling error** of the sample proportion is

$$\textbf{Sampling Error} \;=\; \hat{P} - p\,.$$

- The **sampling error** of the sample proportion is

$$\textbf{Sampling Error} \ = \ \hat{P} - p\,.$$

- It's preferable to estimate p using a **confidence interval** because its **width** will reflect how big the **sampling error** might be.

- The **sampling error** of the sample proportion is

$$\textbf{Sampling Error} \;=\; \hat{P} - p\,.$$

- It's preferable to estimate p using a **confidence interval** because its **width** will reflect how big the **sampling error** might be.

- To derive the CI, we'll need the **sampling distribution** of $\hat{P}$.

Sampling Distribution of the Sample Proportion $\hat{P}$

- The numerator X in

$$\hat{P} = \frac{X}{n},$$

is the **number of successes** among the n individuals in the sample.

Sampling Distribution of the Sample Proportion $\hat{P}$

- The numerator X in

$$\hat{P} = \frac{X}{n},$$

  is the **number of successes** among the n individuals in the sample.

  X can be regarded as a **binomial**(n, p) random variable.

Sampling Distribution of the Sample Proportion $\hat{P}$

- The numerator X in
$$\hat{P} = \frac{X}{n},$$
is the **number of successes** among the n individuals in the sample.

  X can be regarded as a **binomial**(n, p) random variable.

  Thus $E(X) = np$ and $V(X) = np(1-p)$, and so ...

Sampling Distribution of the Sample Proportion $\hat{P}$

- The numerator X in
$$\hat{P} = \frac{X}{n},$$
is the **number of successes** among the n individuals in the sample.

  X can be regarded as a **binomial**(n, p) random variable.

  Thus $E(X) = np$ and $V(X) = np(1-p)$, and so ...
$$E(\hat{P}) = \frac{1}{n}E(X) = p$$

Sampling Distribution of the Sample Proportion $\hat{P}$

- The numerator X in

$$\hat{P} = \frac{X}{n},$$

is the **number of successes** among the n individuals in the sample.

X can be regarded as a **binomial**(n, p) random variable.

Thus $E(X) = np$ and $V(X) = np(1-p)$, and so ...

$$E(\hat{P}) = \frac{1}{n}E(X) = p$$

and

$$V(\hat{P}) = \frac{1}{n^2}V(X) = \frac{p(1-p)}{n}.$$

Mean and Variance of $\hat{P}$: For a **random sample** from a population of **successes** and **failures** whose proportion of successes is p, the **sampling distribution of $\hat{P}$** has mean

$$E(\hat{P}) \;=\; p\,,$$

and variance and standard deviation

$$V(\hat{P}) \;=\; \frac{p(1-p)}{n} \qquad \text{and} \qquad SD(\hat{P}) \;=\; \sqrt{\frac{p(1-p)}{n}}.$$

- Because $E(\hat{P}) = p$, $\hat{P}$ is an **unbiased** estimator of p.

- Because $E(\hat{P}) = p$, $\hat{P}$ is an **unbiased** estimator of p.
- The standard deviation

$$SD(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$$

is called the ***standard error of $\hat{P}$***. It represents a **typical deviation** of $\hat{P}$ away from p.

- Because $E(\hat{P}) = p$, $\hat{P}$ is an **unbiased** estimator of p.
- The standard deviation

$$SD(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$$

  is called the ***standard error of*** $\hat{P}$. It represents a **typical deviation** of $\hat{P}$ away from p.
- The **standard error** will be **small** if either:
  1. The population proportion p is close to 0 or 1, or
  2. The sample size n is **large**

- The following is a consequence of the **Normal Approximation to the Binomial**:

Proposition

Normality of $\hat{P}$: For a **random sample** from a population of **successes** and **failures** whose proportion of successes is p, **if n is large**,

$$\hat{P} \sim \mathrm{N}\left(p, \ \sqrt{\frac{p(1-p)}{n}}\right) \qquad \text{(approximately).}$$

Thus the **standardized** version of $\hat{P}$ follows a **standard normal** distribution, i.e.

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathrm{N}(0, 1) \qquad \text{(approximately).}$$

One-Sample z CI for p (7.2)

- To derive **95% CI for p**, note that

$$
\begin{aligned}
0.95 &\approx P\left(-1.96 < \frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}} < 1.96\right) \\
&\vdots \\
&= P\left(\hat{P} - 1.96\sqrt{\frac{p(1-p)}{n}} < p < \hat{P} + 1.96\sqrt{\frac{p(1-p)}{n}}\right)
\end{aligned}
$$

- Thus with probability 0.95, the (unknown) population proportion p will lie in the interval

$$\hat{P} \pm 1.96\sqrt{\frac{p(1-p)}{n}}.$$

- Thus with probability 0.95, the (unknown) population proportion p will lie in the interval

$$\hat{P} \pm 1.96\sqrt{\frac{p(1-p)}{n}}.$$

But this interval involves the unknown p. To get around this, we'll plug in the estimate $\hat{P}$ for p.

- Thus with probability 0.95, the (unknown) population proportion p will lie in the interval

$$\hat{P} \pm 1.96\sqrt{\frac{p(1-p)}{n}}.$$

But this interval involves the unknown p. To get around this, we'll plug in the estimate $\hat{P}$ for p.

This leads to the following CI.

Large-Sample 100$(1-\alpha)$% Confidence Interval for p:

$$\hat{P} \ \pm \ z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\,.$$

This is called the ***one-sample z confidence interval for p***.

It's valid when the sample is from a population of successes and failures and n is large.

- In practice, n **is large** enough for the **one-sample** z **CI for** p to be valid as long as

$$n\hat{P} \;\geq\; 10 \qquad \textbf{and} \qquad n(1-\hat{P}) \;\geq\; 10.$$

  i.e. as long as there are at least **10 successes** and at least **10 failures** in the sample.

Example

A June, 2016 Marist poll of $n = 516$ adult Americans found that **284** (or **55%**) oppose legalizing the sale of human organs for transplant purposes.

Example

A June, 2016 Marist poll of $n = 516$ adult Americans found that **284** (or **55%**) oppose legalizing the sale of human organs for transplant purposes.

The **sample proportion** is

$$\hat{p} = \frac{284}{516} = \mathbf{0.55},$$

Example

A June, 2016 Marist poll of $n = 516$ adult Americans found that **284** (or **55%**) oppose legalizing the sale of human organs for transplant purposes.

The **sample proportion** is

$$\hat{p} = \frac{284}{516} = \mathbf{0.55},$$

and this is the **point estimate** of p, the true (unknown) proportion of all Americans that oppose legalizing the sale of organs.

Now we'll estimate p using a **95% CI**:

Now we'll estimate p using a **95% CI**:

$$\hat{P} \pm z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \;=\; 0.55 \pm 1.96\sqrt{\frac{0.55(1-0.55)}{516}}$$

Now we'll estimate p using a **95% CI**:

$$\begin{aligned}\hat{P} \pm z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} &= 0.55 \pm 1.96\sqrt{\frac{0.55(1-0.55)}{516}} \\ &= 0.55 \pm 0.04\end{aligned}$$

Now we'll estimate p using a **95% CI**:

$$\begin{aligned}\hat{P} \pm z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} &= 0.55 \pm 1.96\sqrt{\frac{0.55(1-0.55)}{516}} \\ &= 0.55 \pm 0.04 \\ &= \mathbf{(0.51,\ \ 0.59)}\end{aligned}$$

Now we'll estimate p using a **95% CI**:

$$\begin{aligned} \hat{P} \pm z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} &= 0.55 \pm 1.96\sqrt{\frac{0.55(1-0.55)}{516}} \\ &= 0.55 \pm 0.04 \\ &= \mathbf{(0.51,\ \ 0.59)} \end{aligned}$$

This gives a range of estimates of p, and we can be **95% confident** that p is in the interval somewhere.

Large Sample Test for a Population Proportion p (8.4)

- Suppose we have a random sample of size n from a population of **successes** and **failures**.

Large Sample Test for a Population Proportion p (8.4)

- Suppose we have a random sample of size n from a population of **successes** and **failures**.
- We'll see how to use the sample to decide if the **population proportion** of **successes** p is different from some **hypothesized value** p_0.

Large Sample Test for a Population Proportion p (8.4)

- Suppose we have a random sample of size n from a population of **successes** and **failures**.
- We'll see how to use the sample to decide if the **population proportion** of **successes** p is different from some **hypothesized value** p_0.

  The appropriate test is called the ***one-sample*** z ***test for*** p.

- The **null hypothesis** is that the population proportion p is equal to p_0:

Null Hypothesis:

$$H_0 : p = p_0$$

- The **alternative hypothesis** will depend on what we're trying to "prove":

Alternative Hypothesis: The alternative hypothesis will be one of

1. $H_a : p > p_0$ (**one-sided, upper-tailed**)
2. $H_a : p < p_0$ (**one-sided, lower-tailed**)
3. $H_a : p \neq p_0$ (**two-sided, two-tailed**)

depending on what we're trying to verify using the data.

One-Sample Z Test Statistic for p:

$$Z = \frac{\hat{P} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

One-Sample Z Test Statistic for p:

$$Z = \frac{\hat{P} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Z measures how many standard errors $\hat{P}$ is away from $\boldsymbol{p_0}$.

One-Sample Z Test Statistic for p:

$$Z = \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

- Z measures how many standard errors $\hat{P}$ is away from p_0.
- $\hat{P}$ is an estimator of the unknown population proportion p, so ...

One-Sample Z Test Statistic for p:

$$Z = \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

- Z measures how many standard errors $\hat{P}$ is away from $\boldsymbol{p_0}$.
- $\hat{P}$ is an estimator of the unknown population proportion $\boldsymbol{p}$, so ...
  1. Z will be approximately **zero** (most likely) if $p = p_0$.
  2. It will be **positive** (most likely) if $p > p_0$.
  3. It will be **negative** (most likely) if $p < p_0$.

1. **Large positive** values of Z provide **evidence against H_0 in favor of**
   $\boldsymbol{H_a : p > p_0}$.
2. **Large negative** values of Z provide **evidence against H_0 in favor of**
   $\boldsymbol{H_a : p < p_0}$.
3. **Large positive *and* large negative** values of Z provide **evidence**
   against H_0 in favor of
   $\boldsymbol{H_a : p \neq p_0}$.

- Recall that **if n is large**,

$$\frac{\hat{P} - p}{\sqrt{p(1-p)/n}} \sim \mathsf{N}(0,1) \qquad \text{(approximately).}$$

- Recall that **if n is large**,

$$\frac{\hat{P} - p}{\sqrt{p(1-p)/n}} \sim \text{N}(0,1) \qquad \text{(approximately).}$$

- It follows that **if H_0 is true** (so $p = p_0$),

$$\frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim \text{N}(0,1) \qquad \text{(approximately).}$$

- Recall that **if $\boldsymbol{n}$ is large**,

$$\frac{\hat{P} - \boldsymbol{p}}{\sqrt{\boldsymbol{p}(1-\boldsymbol{p})/n}} \sim \text{N}(0,1) \qquad \text{(approximately).}$$

- It follows that **if $\boldsymbol{H_0}$ is true** (so $p = p_0$),

$$\frac{\hat{P} - \boldsymbol{p_0}}{\sqrt{\boldsymbol{p_0}(1-\boldsymbol{p_0})/n}} \sim \text{N}(0,1) \qquad \text{(approximately).}$$

Sampling Distribution of the Test Statistic Under H_0:
If Z is the one-sample Z test statistic, then when

$$H_0 : p = p_0$$

is true,

$$Z \sim \mathsf{N}(0, 1).$$

- The $N(0,1)$ curve gives us:

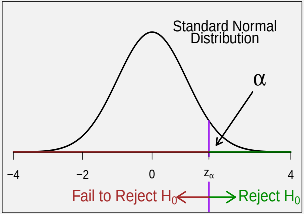
- The $N(0,1)$ curve gives us:
  - The ***rejection region*** as the **extreme 100α% of z values** (in the direction(s) specified by H_a).

- The $N(0,1)$ curve gives us:
  - The ***rejection region*** as the **extreme 100α% of z values** (in the direction(s) specified by H_a).
  - The ***p-value*** as the **tail area(s) beyond the observed z value** (in the direction(s) specified by H_a).

Rejection Region: The **rejection region** is the **set of z values** in the tail of the $N(0,1)$ curve:

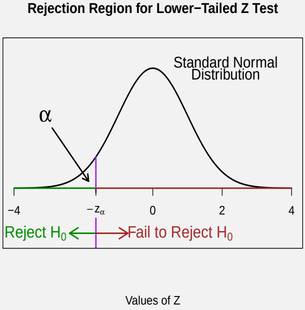
1. To the **right of** z_α if the alternative hypothesis is $H_a : p > p_0$:

Rejection Region for Upper−Tailed Z Test

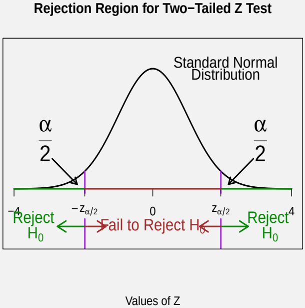


Values of Z

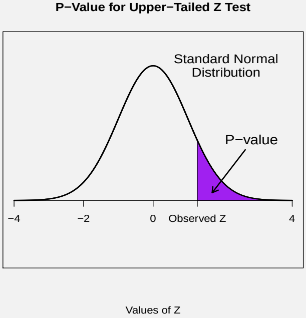
2. To the **left of** z_α if the alternative hypothesis is $H_a : p < p_0$:



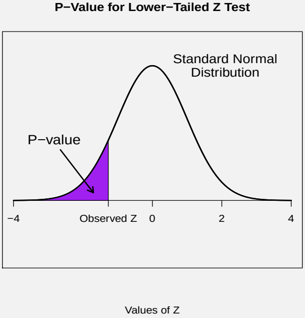
3. To the **left of** $-z_{\alpha/2}$ **and right of** $z_{\alpha/2}$ if the alternative hypothesis is $H_a : p \neq p_0$:



1. **P-value** = Area to the **right** of the observed z if the alternative hypothesis is $H_a : p > p_0$.

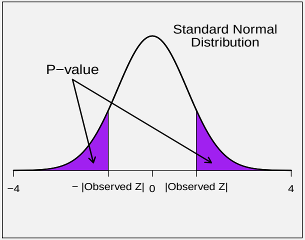


1. **P-value** = Area to the **left** of the observed z if the alternative hypothesis is $H_a : p < p_0$.



1. **P-value** = Area to the **left** of $-|z|$ **and right** of $|z|$ if the alternative hypothesis is $H_a : p \neq p_0$.

P−Value for Two−Tailed Z Test



Values of Z

- In practice, n **is large** enough for the **one-sample z test for p** to be valid as long as

$$np_0 \;\geq\; 10 \qquad \textbf{and} \qquad n(1-p_0) \;\geq\; 10.$$

Example

A new vaccine is meant to prevent meningitis in infants.

Example

A new vaccine is meant to prevent meningitis in infants.

In a clinical trial, the vaccine was administered to **710** infants. Of these, **121** experienced appetite loss (a side effect).

Example

A new vaccine is meant to prevent meningitis in infants.

In a clinical trial, the vaccine was administered to **710** infants. Of these, **121** experienced appetite loss (a side effect).

Is there statistically significant evidence that the **true proportion** of infants experiencing appetite loss from the vaccine is **greater than 0.135**, the proportion that experience this side effect from competing medications?

We'll test the **hypotheses**

$$\begin{aligned} H_0 : p &= 0.135 \\ H_a : p &> 0.135 \end{aligned}$$

where p is the true (unknown) population proportion that experiences appetite loss from the new vaccine.

We'll test the **hypotheses**

$$\begin{aligned} H_0 : p &= 0.135 \\ H_a : p &> 0.135 \end{aligned}$$

where p is the true (unknown) population proportion that experiences appetite loss from the new vaccine.

The **sample proportion** is

$$\hat{P} = \frac{121}{710} = \mathbf{0.170}.$$

Thus the observed **test statistic** is

$$z = \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Thus the observed **test statistic** is

$$\begin{aligned} z &= \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}} \\ &= \frac{0.170 - 0.135}{\sqrt{0.135(1-0.135)/710}} \end{aligned}$$

Thus the observed **test statistic** is

$$
\begin{aligned}
z &= \frac{\hat{P}-p_0}{\sqrt{p_0(1-p_0)/n}} \\
&= \frac{0.170-0.135}{\sqrt{0.135(1-0.135)/710}} \\
&= \mathbf{2.73}.
\end{aligned}
$$

Thus the observed **test statistic** is

$$\begin{aligned} z &= \frac{\hat{P}-p_0}{\sqrt{p_0(1-p_0)/n}} \\ &= \frac{0.170-0.135}{\sqrt{0.135(1-0.135)/710}} \\ &= \mathbf{2.73}. \end{aligned}$$

Thus the **sample proportion** that experiences appetite loss, $\hat{P} = 0.170$, is **2.73 standard errors above 0.135.**

Thus the observed **test statistic** is

$$\begin{aligned} z &= \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}} \\ &= \frac{0.170 - 0.135}{\sqrt{0.135(1-0.135)/710}} \\ &= \mathbf{2.73}. \end{aligned}$$

Thus the **sample proportion** that experiences appetite loss, $\hat{P} = 0.170$, is **2.73 standard errors above 0.135.**

The **p-value** is the **probability** that we'd get a z value this far above zero by chance **if** the **population proportion** p was **0.135.**

From the **upper tail** area of the N$(0,\ 1)$ distribution, to the **right** of **2.73**,

$$\textbf{p-value} \ = \ \mathbf{0.0032}.$$

Using a **level of significance** $\boldsymbol{\alpha = 0.05}$, the **decision rule** is

Reject H_0 if p-value < 0.05.
Fail to reject H_0 if p-value ≥ 0.05.

Using a **level of significance** $\boldsymbol{\alpha = 0.05}$, the **decision rule** is

Reject H_0 if p-value < 0.05.
Fail to reject H_0 if p-value ≥ 0.05.

Because $\mathbf{0.0032 < 0.05}$, we **reject** $\boldsymbol{H_0}$.

Using a **level of significance** $\alpha = 0.05$, the **decision rule** is

$$\begin{array}{l} \text{Reject } H_0 \text{ if p-value} < 0.05. \\ \text{Fail to reject } H_0 \text{ if p-value} \geq 0.05. \end{array}$$

Because $\mathbf{0.0032 < 0.05}$, we **reject** $\boldsymbol{H_0}$.

There's **statistically significant evidence** that the population proportion of infants p that experience appetite loss from the vaccine is greater than 0.135.

Using a **level of significance** $\alpha = 0.05$, the **decision rule** is

Reject H_0 if p-value < 0.05.
Fail to reject H_0 if p-value ≥ 0.05.

Because $\mathbf{0.0032 < 0.05}$, we **reject H_0.**

There's **statistically significant evidence** that the population proportion of infants p that experience appetite loss from the vaccine is greater than 0.135.

The observed result cannot easily be explained by chance variation (sampling error).