VOTRE LIEN AVEC CE QUI COMPTE — CONNECTS YOU TO WHAT MATTERS

Estimating a Population Proportion

ADM 2304 – Winter 2016

Objective

- To understand and calculate (confidence) interval estimates for the population proportion **p**.

An interval estimate gives us not only a point estimate, but also a range of possible values (a measure of precision), along with a measure of "confidence".

Standard Normal z-notation

$\mathbf{z}_{\alpha}$ is a number such that Prob [$\mathbf{Z} > \mathbf{z}_{\alpha}$] = α,

For example,

$z_{0.05}$ = 1.645 since Prob(Z > 1.645) = .05.

$\mathbf{z}_{\alpha/2}$ is a number such that

Prob [$\mathbf{Z} < -\mathbf{z}_{\alpha/2}$] = Prob [$\mathbf{Z} > \mathbf{z}_{\alpha/2}$] = α / 2

and Prob [$-\mathbf{z}_{\alpha/2} < \mathbf{Z} < \mathbf{z}_{\alpha/2}$] = 1 - α .

For example, $z_{0.05/2}$ = 1.96 since

Prob(-1.96 < z < 1.96) = .95

Probability Intervals

Based on the Sampling Distribution of the sample proportion **p-hat**, the *probability* that **p-hat** falls inside the interval [$\mathbf{p \pm z_{\alpha/2} * \sqrt{(pq/n)}}$] is $1 - \alpha$.

(the probability is defined over all possible samples)

For example,

Prob(**p-hat** falls inside **[p ± 1.96 * √(pq/n)]**)= 95%.

The interval [$p - z_{\alpha/2} * \sqrt{(pq/n)}$, $p + z_{\alpha/2} * \sqrt{(pq/n)}$] is a fixed (symmetric) probability interval, since it is based on a fixed value of p; moreover, it can be calculated only if p is known or is assumed to be some value, and it is **not** based on sample data (hence, it belongs to a probability analysis and we can calculate the probability that **p-hat** falls inside it).

Confidence Intervals

- The interval [**p-hat ± $z_{\alpha/2}$ * √(pq/n)**] is a 100(1 - α) % **confidence interval**; this is a random interval because the value of **p-hat** changes from sample to sample.
- The confidence interval covers the true value **p**, if and only if **p-hat** falls inside the probability interval **[p ± $z_{\alpha/2}$ * √(pq/n)].**

Use of Standard Error

- The previous confidence interval formula cannot be calculated unless we estimate the unknown value of the standard deviation $\sqrt{(pq/n)}$ by the "standard error":$\sqrt{(\text{p-hat} * \text{q-hat} / n)}$.
- The standard error (SE) is the estimated standard deviation of the sampling dist'n. The standard error of the sample proportion is denoted by **SE(p-hat).**

Confidence Interval to estimate a (Population) Proportion

Based on the "approximately normal" sampling distribution of **p-hat** (sample proportion), the 100(1-α)% confidence interval for estimating the population proportion is:

$$\textbf{p-hat} \pm z_{\alpha/2} * \textbf{SE(p-hat)}$$

where **SE(p-hat)** $= \sqrt{[\text{p-hat} * \text{q-hat} / n]}$ and q-hat = 1 - p-hat,

and $\mathbf{z_{\alpha/2}}$ is the upper $\alpha/2$ point of the standard normal distribution; note that the SDVW text uses the notation **z***.

(We must have a random sample where n/N < 10% but n must be large enough such that both n * p and n * (1 - p) are at least 10. In practice, we check whether n * p-hat and n * q-hat are both at least 10.)

Common CIs

Based on the normal probability table, a 95% confidence interval would be

p-hat ± 1.96 * SE(p-hat)

and a 99% confidence interval would be

p-hat ± 2.575 * SE(p-hat),

since Prob($Z > z_{0.01/2}$) = 2.575.

Terminology of Confidence Intervals

We say that we are 100(1- α)% “confident" that an interval calculated using the formula:

p-hat ± $z_{\alpha/2}$ * √(p-hat * q-hat /n),

contains or covers the value of the population proportion.

Sampling Interpretation

- Confidence intervals are **random** intervals, since they are based on the value of a random quantity (***p-hat***) that varies from sample to sample.
- Given an actual sample of size n, the calculated interval will either contain the true population parameter or not.
- 100(1-α)% of confidence intervals based on all possible samples *will* contain the parameter value. Thus the confidence level is a **collective** characteristic of all intervals, not an individual characteristic. It is incorrect to say that a particular interval contains the true value with probability 95%. Later **conf.xlt** will demonstrate this.
- We can only *hope* that we got one of the good samples where the 95% confidence interval does cover the true parameter value.

Example

A recent poll found 27.4%* supporting the CPC. What is a 95% confidence interval for the true proportion? Assume a sample size of 1000.

p-hat = .274 and **SE(p-hat)** = √ (.274*.726 / 1000) = .0141

The 95% confidence interval for the population proportion supporting the CPC is:

27.4% ± 1.96 * 0.0141 or 27.4**% ± 2.8%**

and the 2.8% is known as the "margin of error".

Such results are usually qualified in the media by the statement: "*The poll is accurate to within 2.8 percentage points, 19 times out of 20.*"

*http://www.threehundredeight.com/

Sample Size Determination

- Suppose we want to estimate the support for the CPC with a margin of error (ME) = .01 using a 95% confidence level. We need to find n such that:

$$\pm z_{\alpha/2}\sqrt{(pq/n)} = \pm ME \text{ or } n = pq\,(z_{\alpha/2} / ME)^2$$

- n is maximized when p = q = .5; therefore, a conservative estimate of n is:

$$(.5*.5)*(1.96/.01)^2 = 9604.$$

- To justify smaller sample sizes, we would need better information on the value of p, e.g.:

$$(.274*.726)*(1.96/.01)^2 = 3900.$$

Other Confidence Intervals

- Usually confidence intervals are calculated as symmetric intervals, with the same margin of error either side of the sample estimate.
- This reflects an *objective* point of view, without any bias in favour of one side or the other.

1-sided Probability Intervals

- Instead of calculating a symmetric **probability** interval
  [$\mathbf{p \pm z_{\alpha/2} * \sqrt{(pq/n)}}$],
  where the probability of **p-hat** falling inside the probability interval is 1 - α, we could calculate non-symmetric probability intervals:
  [$\mathbf{0}$, $\mathbf{p + z_{\alpha} * \sqrt{(pq/n)}}$] or [$\mathbf{p - z_{\alpha} * \sqrt{(pq/n)}}$ **, 1**],
  where again the probability is 1 - α that:
  p-hat < **p** + $\mathbf{z_{\alpha} * \sqrt{(pq/n)}}$ or that
  p-hat > **p** - $\mathbf{z_{\alpha} * \sqrt{(pq/n)}}$, respectively.

- The probabilities are defined over all possible samples of size n.

1-sided Confidence Intervals

The corresponding 1-sided confidence intervals

[**p-hat** - $\mathbf{z_\alpha}$ * $\sqrt{}$**(p-hat * q-hat /n) , 1**] or

[**0**, **p-hat** + $\mathbf{z_\alpha}$ * $\sqrt{}$**(p-hat * q-hat /n)**]

could reflect particular points of view.

- A person opposing the CPC might want to say the level of support is:

  "Less than **p-hat + z_α * √(p-hat * q-hat /n)**",

  placing an upper bound (but no lower bound) on the level of support.

- A person supporting the CPC might want to say the level of support is:

  "Greater than **p-hat - z_α * √(p-hat * q-hat /n)**",

  placing a lower bound (but no upper bound) on the level of support.

Comparisons of CIs

Compare the 2-sided CI:

27.4% ± 1.96* .0141

or **27.4**% ± 2.8% = (24.6%, 30.2%)

with the 1-sided CIs:

“at most .274 + 1.645*.0141”

= (0%, 29,7%], or

“at least .274 - 1.645*.0141”

= [25.1%, 100%)

Summary

- The properties of the sampling distribution allows us to calculate fixed probability intervals for **p-hat**:

$$\mathbf{p \pm z_{\alpha/2} * \sqrt{(pq/n)}}$$

- Keeping in mind the sampling interpretation, we calculate the confidence intervals for p:

$$\mathbf{p\text{-}hat \pm z_{\alpha/2} * \sqrt{(p\text{-}hat * q\text{-}hat / n)}}$$

- A sample where **p-hat** falls inside the probability interval will be one where the confidence interval covers the true value of p.

- These samples occur with probability (1 - α) but we cannot tell whether a particular confidence interval does or does not cover p.

Using Minitab to calculate confidence intervals for a population proportion

- Go first to the **Stat** Menu, and select **Basic Statistics,** then **“1 Proportion”** .
- The column of individual values must contain two different values (numerical or qualitative) or else we need summarized data in terms of “number of trials” **n** and the “number of events” **x**.

Options

- Specify the confidence level and check “test and interval based on normal distribution” if the sample size condition is satisfied (otherwise, Minitab will use the binomial probability calculation).
- For a 2-sided confidence interval, select the ‘**not equal’** “alternative”; for a 1-sided confidence interval, select the ‘**less than’** or the ‘**greater than’** “alternative”. (Next week, we will talk more about “alternative” hypotheses when we introduce hypothesis testing.)