W14 QMS 202 Lecture 2

Dr Boža Tasić

January 20, 2015

Confidence Interval Estimation for the Proportion - Learning Objectives

- In this Chapter we will learn
  - To construct and interpret confidence interval estimates for the proportion.
  - How to determine the sample size necessary to develop a confidence interval estimate for the mean or proportion

Sampling Distribution of the Proportion

- We consider a categorical variable that has only two categories.
- For example:
- the customer prefers your brand OR the customer prefers the competitor's brand,
- voters who support proposition A OR voters who do not support proposition A.
- We are interested in the proportion of the items belonging to one of the categories.
- For example: the customers who prefer your brand OR voters who do not support proposition A.

Sampling Distribution of the Proportion

- The **population proportion**, denoted by π, is the proportion of the population having a characteristic of interest.
- The **sample proportion**, denoted by p, is the proportion of the items in the sample having a characteristic of interest, i.e.,

$$p = \frac{X}{n} = \frac{\#\text{ of items having the characteristic of interest}}{\text{Sample Size}}$$

- The sample proportion, a statistic, is used to estimate the population proportion, a parameter.
- The sample proportion takes on values between 0 and 1, i.e.,

$$0 \leq p \leq 1$$

Sampling Distribution of the Proportion

- The same way the sample mean $\overline{X}$ is an unbiased estimator of the population mean μ, the statistic p is an unbiased estimator of the population proportion π.
- By analogy to the sampling distribution of the mean, whose standard error is $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$, the **standard error of the proportion** σ_p is defined by

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Sampling Distribution of the Proportion

- The **sampling distribution of the proportion** follows the binomial distribution.
- However, we can use the normal distribution to approximate the binomial distribution when

$$n\pi \geq 5 \quad \text{and} \quad n(1-\pi) \geq 5$$

- Assuming sampling with replacement from a finite population or without replacement from an infinite population, if the sample size is big enough to meet the above conditions we will use the normal distribution to estimate the sampling distribution of the proportion.

Sampling Distribution of the Proportion

- Substituting p for $\overline{X}$, π for μ and $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$ for $\frac{\sigma}{\sqrt{n}}$ in

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

  we get

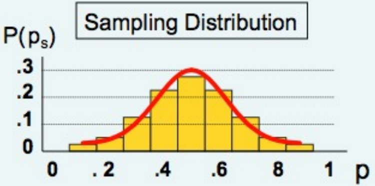
- The equation for finding Z for the sampling distribution of the proportion

$$Z = \frac{p - \pi}{\sigma_p} = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

Sampling Distribution of the Proportion

- Approximated by a normal distribution if:

  - $n\pi \geq 5$ and $n(1-\pi) \geq 5$

  where

  $$\mu_p = \pi \quad \text{and} \quad \sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

  (where π = population proportion)

Sampling Distribution of the Proportion - Example 1

- If the true proportion of voters who support Proposition A is $\pi = 0.4$, what is the probability that a sample of size 200 yields a sample proportion between 0.40 and 0.45?
- We are given $\pi = 0.4$ and $n = 200$. We need to compute $P(0.4 < p < 0.45)$.
- We will solve the problem using the calculator.

Sampling Distribution of the Proportion - Example 1

- We need to compute σ_p first.
- Using $\pi = 0.4$ and $n = 200$ and the formula $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$ we get
- $$\sigma_p = \sqrt{\frac{0.4(1-0.4)}{200}} = 0.03464$$
- **Remember** that you will be using π as μ and σ_p as σ

Using CASIO Calculator to compute $P(0.4 \leq p \leq 0.45)$

1. Select **STAT F5** (DIST) and then **F1**(NORM). Now select **F2**(Ncd) and the following options:
2. Data: **F2** (Variable)
3. Lower: **0.4 EXE**
4. Upper: **0.45 EXE**
5. σ: **0.03464 EXE**
6. μ: **0.4 EXE**
7. Save Res: If you do not want to save results in a list press **F1** (None). If you want to save results in a list press **F2** and type the number of list where you want to save results.
8. Execute: key **EXE** or **F1**(Calc).

- You will get $P(0.4 \leq p \leq 0.45) = 0.42554862$

Confidence Interval Estimation for the Proportion p

- We will now extend the concept of the confidence interval to categorical data.
- We will estimate the proportion of items in a population having a certain characteristic of interest.
- An interval estimate for the population proportion π can be calculated by adding an allowance for uncertainty to the sample proportion p
- Using the fact that the distribution of the sample proportion is approximately normal if the sample size is large, with standard deviation

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

we will estimate this with sample data by

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

Confidence Interval Estimation for the Proportion p

- The confidence interval for the proportion π is determined by

$$p \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

or in the interval form as

- $$p - Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

where

- $p = \frac{X}{n}$ is the sample proportion
- π is the population proportion
- $Z_{\frac{\alpha}{2}}$ is the normal distribution critical value for a probability of $\frac{\alpha}{2}$ in each tail
- n is the sample size
- **Note**: To use this interval estimate, the sample size n must be large enough to ensure that both X and $n - X$ are greater than 5.

Confidence Interval Estimation for the Proportion - Example 1

- A random sample of 100 people shows that 25 are left-handed. Form a 95% confidence interval for the true proportion of left-handers
- We are given $n = 100$ and $X = 25$ which implies $p = \frac{X}{n} = \frac{25}{100} = 0.25$.
- The first thing we need to check is that the sample is big enough, i.e., X and $n - X$ are greater than 5.
- In our case $X = 25 > 5$ and $n - X = 100 - 25 = 75 > 5$. So, the sample is big enough.
- Using $1 - \alpha = 0.95$ we have $\alpha = 0.05$ and hence $Z_{\frac{\alpha}{2}} = 1.96$.

Confidence Interval Estimation for the Proportion - Example 1

- Now we compute

$$p \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} = 0.25 \pm 1.96 \sqrt{\frac{0.25(0.75)}{100}} = 0.25 \pm 1.96(0.0433)$$

- So, the interval is $0.1651 \leq \pi \leq 0.3349$
- **Conclusion**: the true percentage of left-handers in the population is between 16.51% and 33.49%.
- Although the interval from 0.1651 to 0.3349 may or may not contain the true proportion, 95% of intervals formed from samples of size 100 in this manner will contain the true proportion.

Using CASIO Calculator -Example 1

1. Select **STAT**, **F4** (INTR), **F1**(Z) and then **F3**(1-p). Now in the **1- Prop ZInterval** menu select the following options:
2. C-Level: **0.95 EXE**
3. x: **25 EXE**
4. n: **100 EXE**
5. Save Res: If you do not want to save results in a list press **F1** (None). If you want to save results in a list press **F2** and type the number of list where you want to save results.
6. Press key **EXE**

- You will get Left $= 0.1651$ and Right $= 0.3349$, $p = 0.25$ and $n = 100$. This means that the 95% confidence interval estimate is between Left $= 0.1651$ and Right $= 0.3349$.

Confidence Interval Estimation for the Proportion - Example 2

- The operation manager at a large newspaper wants to estimate the proportion of newspapers printed that have a nonconforming attribute, such as excessive ruboff, improper page setup, missing pages, or duplicate pages. A random sample of $n = 200$ newspapers is selected from all the newspapers printed during a single day. In this sample, 35 contain some type of nonconformance. Construct and interpret a 90% confidence interval for the proportion of newspapers printed during the day that have a nonconforming attribute.
- We are given $n = 200$ and $X = 35$ which implies $p = \frac{X}{n} = \frac{35}{200} = 0.175$.
- The first thing we need to check is that the sample is big enough, i.e., X and $n - X$ are greater than 5.
- In our case $X = 35 > 5$ and $n - X = 200 - 35 = 165 > 5$. So, the sample is big enough.

Confidence Interval Estimation for the Proportion - Example 2

- Using $1 - \alpha = 0.90$ we have $\alpha = 0.1$ and hence $Z_{\frac{\alpha}{2}} = 1.645$.
- Now we compute

$$p \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} = 0.175 \pm 1.645 \sqrt{\frac{0.175(0.825)}{200}} =$$

$$= 0.175 \pm 1.645(0.0269)$$

- So, the interval is $0.1308 \leq \pi \leq 0.2192$
- **Conclusion**: the true percentage of the newspapers printed on that day that have some type of nonconformance is between 13.08% and 21.92%.
- Although the interval from 0.1308 to 0.2192 may or may not contain the true proportion, 90% of intervals formed from samples of size 200 in this manner will contain the true proportion.

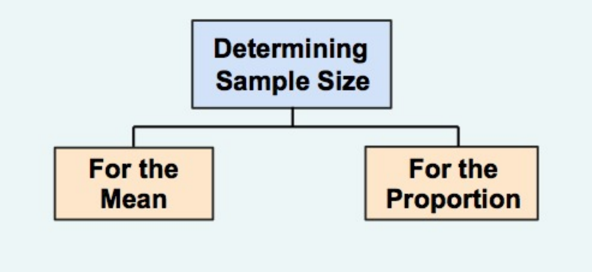
Using CASIO Calculator -Example 2

1. Select **STAT**, **F4** (INTR), **F1**(Z) and then **F3**(1-p). Now in the **1- Prop ZInterval** menu select the following options:
2. C-Level: **0.90 EXE**
3. x: **35 EXE**
4. n: **200 EXE**
5. Save Res: If you do not want to save results in a list press **F1** (None). If you want to save results in a list press **F2** and type the number of list where you want to save results.
6. Press key **EXE**

- You will get Left $= 0.1308$ and Right $= 0.2192$, $p = 0.175$ and $n = 200$. This means that the 90% confidence interval estimate is between Left $= 0.1308$ and Right $= 0.2192$.

Determining Sample Size

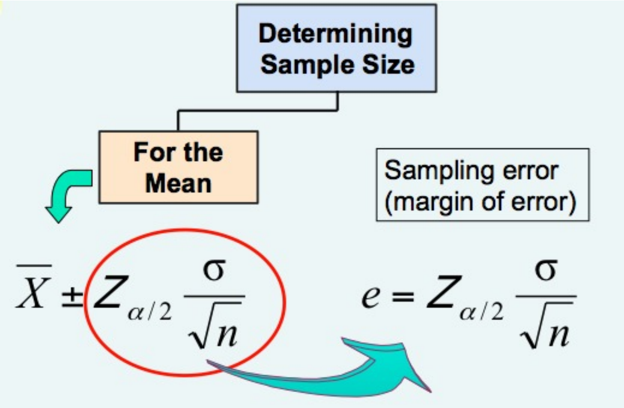
- In the business world, sample sizes are determined prior to data collection to ensure that the confidence interval is narrow enough to be useful in making decisions.
- Determining the proper sample size is a complicated procedure. It is subject to the constrains of budget, time and the amount of allowed sampling error.
- Two cases we will do are



Determining Sample Size

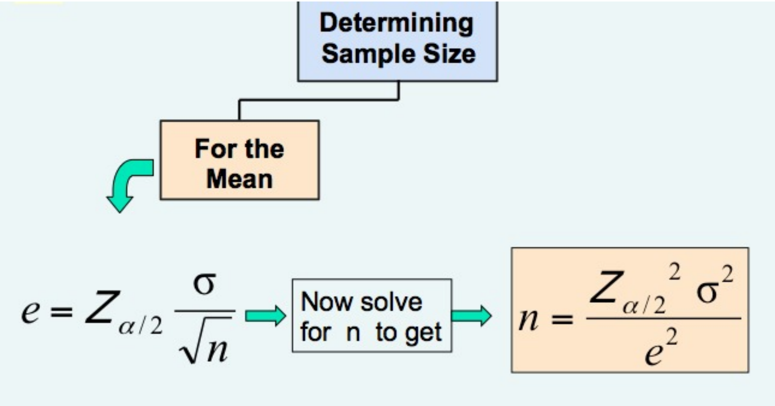
- The required sample size can be found to obtain a desired margin of error e with a specified level of confidence $(1 - \alpha)$
- The **margin of error** is also called **sampling error**. It gives us
  - The amount of imprecision in the estimate of the population parameter
  - The amount added and subtracted to the point estimate to form the confidence interval

Determining Sample Size for the Mean

- To determine the sample size of the mean we use
- 

Determining Sample Size for the Mean

- Now we solve for n



Determining Sample Size for the Mean

- We start with
$$e = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$
Multiplying both sides by $\sqrt{n}$ we get
-
$$\sqrt{n} e = Z_{\frac{\alpha}{2}} \sigma$$
Now dividing both sides by e we get
-
$$\sqrt{n} = \frac{Z_{\frac{\alpha}{2}} \sigma}{e}$$
Finally squaring both sides we get
-
$$n = \frac{Z_{\frac{\alpha}{2}}^2 \sigma^2}{e^2}$$

The Sample Size for the Mean

- The sample size n for the mean is equal to the product of $Z_{\frac{\alpha}{2}}$ squared and σ squared divided by the square of the sampling error e
- $$n = \frac{Z_{\frac{\alpha}{2}}^2 \sigma^2}{e^2}$$

Determining Sample Size for the Mean

- To determine the required sample size for the mean, we must know:
- The desired level of confidence $(1 - \alpha)$, which determines the critical value, $Z_{\frac{\alpha}{2}}$
- The acceptable sampling error, e
- The standard deviation, σ.

Determining Sample Size for the Mean - Example 1

- If $\sigma = 45$, what sample size is needed to estimate the mean within ± 5 with 90% confidence?
- We have the desired level of confidence $(1 - \alpha) = 90$, which determines the critical value, $Z_{\frac{\alpha}{2}} = 1.645$.
- The acceptable sampling error $e = 5$.
- The standard deviation, $\sigma = 45$.
- Now using the formula we get

$$n = \frac{Z_{\frac{\alpha}{2}}^2 \sigma^2}{e^2} = \frac{(1.645)^2 (45)^2}{5^2} = 219.19$$

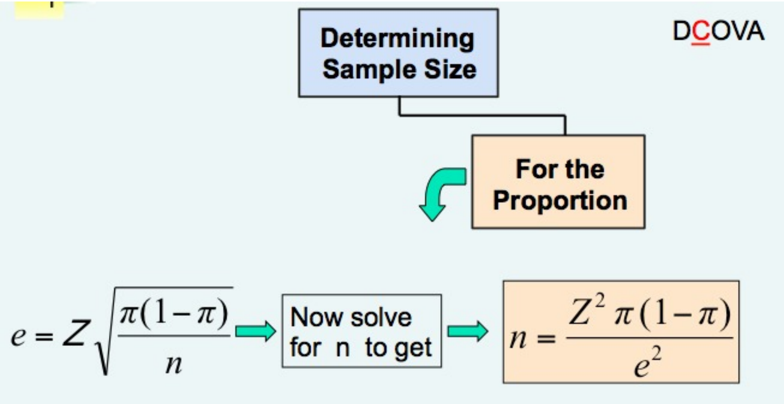
- So, the required sample size is $n = 220$. We always round up.

Determining Sample Size for the Mean

- If σ is unknown, it can be estimated when determining the required sample size. We can
- Select a pilot sample and estimate the standard deviation σ with the sample standard deviation S.
- Use the past data the estimate the standard deviation σ.

Determining Sample Size for the Proportion

- To determine the sample size of the proportion we use



Determining Sample Size for the Proportion

- We start with
$$e = Z_{\frac{\alpha}{2}} \sqrt{\frac{\pi(1-\pi)}{n}}$$
Multiplying both sides by $\sqrt{n}$ we get
- $$\sqrt{n}e = Z_{\frac{\alpha}{2}} \sqrt{\pi(1-\pi)}$$
Now dividing both sides by e we get
- $$\sqrt{n} = \frac{Z_{\frac{\alpha}{2}} \sqrt{\pi(1-\pi)}}{e}$$
Finally squaring both sides we get
- $$n = \frac{Z^2_{\frac{\alpha}{2}} \pi(1-\pi)}{e^2}$$

The Sample Size for the Proportion

- The sample size n for the proportion is equal to the product of $Z_{\frac{\alpha}{2}}$ squared and $\pi(1-\pi)$ divided by the square of the sampling error e
- $$n = \frac{Z_{\frac{\alpha}{2}}^2 \pi(1-\pi)}{e^2}$$

Determining Sample Size for the Proportion

- To determine the required sample size for the proportion, we must know:
- The desired level of confidence $(1-\alpha)$, which determines the critical value, $Z_{\frac{\alpha}{2}}$
- The acceptable sampling error, e
- The true proportion of events of interest, π
- π can be estimated with a pilot sample if necessary
- If you have no prior knowledge or estimate for the population proportion π you should use 0.5 as an estimate of π.

Determining Sample Size for the Proportion - Example 1

- How large a sample would be necessary to estimate the true proportion defective in a large population within ±3%, with 95% confidence? Assume a pilot sample yields $p = 0.12$.
- We have the desired level of confidence $(1 - \alpha) = 95$, which determines the critical value, $Z_{\frac{\alpha}{2}} = 1.96$.
- The acceptable sampling error $e = 0.03$.
- We'll use $p = 0.12$ to estimate π.
- Now using the formula we get

$$n = \frac{Z^2_{\frac{\alpha}{2}}\pi(1-\pi)}{e^2} = \frac{(1.96)^2 0.12(0.88)}{(0.03)^2} = 450.74$$

- So, the required sample size is $n = 451$. We always round up.

Ethical Issues

- A confidence interval estimate (reflecting sampling error) should always be included when reporting a point estimate.
- The level of confidence should always be reported.
- The sample size should be reported.
- An interpretation of the confidence interval estimate should also be provided.