GE2262 Business Statistics

Topic 7
Inference for the Proportion

Reference
Levine, D.M., Krehbiel, T.C. and Berenson, M.L., *Business Statistics: A First Course*, Pearson Education Ltd, Chapter 7 & 8 & 9

Outline

- Sampling Distribution of the Sample Proportion
- Confidence Interval Estimate for the Proportion
- Sample Size Determination for the Proportion
- Hypothesis Testing for the Proportion

Will Britain Leave EU?

- Express (UK), 22 June 2016: Analysts at TNS surveyed 2,320 adults across the UK online between June 16-22. The baseline results reveal a 2% lead for the EU leave campaign, with support for Brexit at 43% compared to 41% for Remain
- We do not know the true population of British intending to leave EU until the result of the referendum is announced. A common way to gain insights on election or a referendum campaign is to held a survey and based inference on the proportion of success in a sample taken from the relevant population
- In practice, most of these surveys are inaccurate because
  - The involved sample is biased. It does not truly represent the population
  - Some voters had not honestly revealed their voting intention, or they changed their mind after the survey

Sample Proportion

- Let Y be the number of observations belong to the one of the two levels (e.g. success and failure, yes and no, etc.) of a categorical variable in a random sample of n observations
- The proportion of observations belong to one of the two levels (e.g. success, yes, etc.) in the sample

$$p = \frac{Y}{n}$$

is called the sample proportion

Sample Proportion

Cont'd

- We saw in Topic 3 that Y, obeys a binomial distribution with

$$P(Y = y) = \frac{n!}{y!\,(n-y)!}\pi^{y}(1-\pi)^{n-y}$$

where

$P(Y = y)$ = probability that $Y = y$ events of interest, where y = 0, 1, 2 ,…, n

π = probability of an event of interest, or the population proportion of observations belong to the level of interest

Sampling Distribution of Sample Proportion

- A small enterprise has 4 staff, $N = 4$ (3 males and 1 female)
- Variable of interest: Gender
- Let Y = no. of male staff, π = proportion of male staff = 0.75
- Random samples of size 2 with replacement are taken ($n = 2$)
- As Y obeys a binomial distribution
  - $Y \sim B(2, 0.75)$
  - $\mu = n\pi = 2 \times 0.75 = 1.5$
  - $\sigma = \sqrt{n\pi(1-\pi)}$
    $= \sqrt{2 \times 0.75 \times 0.25} = \sqrt{0.375}$

Sampling Distribution of Sample Proportion

Cont'd

- Sample proportion of male staff, $p = \frac{Y}{n}$

16 possible sample proportions

Respondent	A (M)	B (M)	C (F)	D (M)
A (M)	2/2 = 1	2/2 = 1	1/2 = 0.5	2/2 = 1
B (M)	2/2 = 1	2/2 = 1	1/2 = 0.5	2/2 = 1
C (F)	1/2 = 0.5	1/2 = 0.5	0/2 = 0	1/2 = 0.5
D (M)	2/2 = 1	2/2 = 1	1/2 = 0.5	2/2 = 1

- Probability distribution of p

p	0	0.5	1
$P(p)$	1/16	6/16	9/16

Sampling Distribution of Sample Proportion

Cont'd

- Summary measures for the sampling distribution of sample proportion

$$\mu_p = \sum p_i P(p_i)$$

$$= 0\left(\frac{1}{16}\right) + 0.5\left(\frac{6}{16}\right) + 1\left(\frac{9}{16}\right) = 0.75 = \pi$$

$$\sigma_p = \sqrt{\sum \left(p_i - \mu_p\right)^2 P(p_i)}$$

$$= \sqrt{(0-0.75)^2\left(\frac{1}{16}\right) + (0.5-0.75)^2\left(\frac{6}{16}\right) + (1-0.75)^2\left(\frac{9}{16}\right)}$$

$$= 0.3062$$

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

- We say the sample proportion p is an unbiased estimator of the population proportion π

Sampling Distribution of Sample Proportion

Cont'd

- The exact form of the sampling distribution of p is rather complicated. Instead of using the exact distribution of p, it is common to approximate the sampling distribution by a normal distribution

Sampling Distribution of Sample Proportion

Cont'd

- Suppose you want to estimate the proportion (π) of CityU students who skipped 2 or more classes per week in last semester
- A sample of size n is collected
- You register your data points as categorical observations: Yes, skipped 2 or more classes; or No, skipped 1 or less class
- For subsequent data manipulation, you may code those who skipped 2 or more classes as a 1, and those who skipped 1 or less class as a 0
- Using the numeric coded values, and denotes X_i as the numeric coded value of the ith observed student in the sample, we see that

  $Y = \sum X_i$ = observed number of students who skipped 2 or more classes

  $p = \frac{Y}{n} = \frac{\sum X_i}{n}$ = sample proportion of students who skipped 2 or more classes

  - $\frac{\sum X_i}{n}$ is like the formula for the sample mean, so, a sample proportion is a special case of a sample mean

Sampling Distribution of Sample Proportion

Cont'd

- Since the sampling distribution of p $(= \frac{Y}{n} = \frac{\sum X_i}{n})$ has mean π and standard deviation $\sqrt{\frac{\pi(1-\pi)}{n}}$, then by Central Limit Theorem, sampling distribution of p follows a normal distribution approximately with mean π and standard deviation $\sqrt{\frac{\pi(1-\pi)}{n}}$ for large n

Sampling Distribution of Sample Proportion

Cont'd

- Hence, for large sample size, the distribution of the random variable

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

  is approximately standard normal

- This statistic can be used to obtain confidence intervals, and hypothesis testing for the population proportion
- In practice, "n is large enough" often means that $n\pi \geq 5$ and $n(1-\pi) \geq 5$, that is π cannot be too small or too large

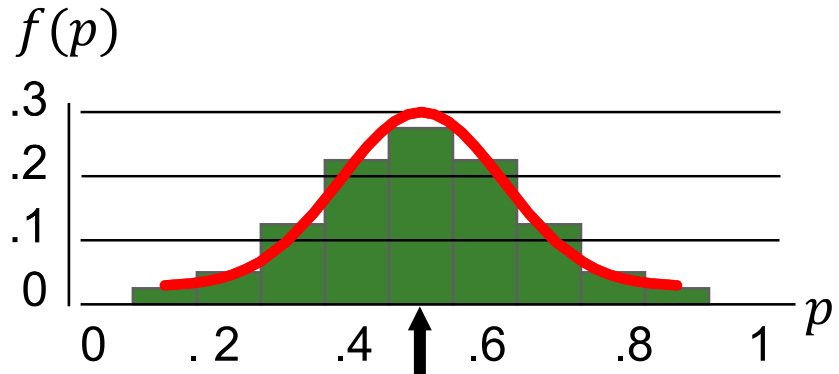
Sampling Distribution of Sample Proportion

Cont'd

- Normal approximation can be used if
  - $n \geq 30$
  - $n\pi \geq 5$
  - $n(1-\pi) \geq 5$

  → Sampling distribution of sample proportion $p \sim N(\mu_p, {\sigma_p}^2)$
- 2 parameters in sampling distribution of sample proportion
  - Mean, $\mu_p = \pi$
  - Variance, ${\sigma_p}^2 = \frac{\pi(1-\pi)}{n}$

Sampling Distribution of p



π = *population proportion*

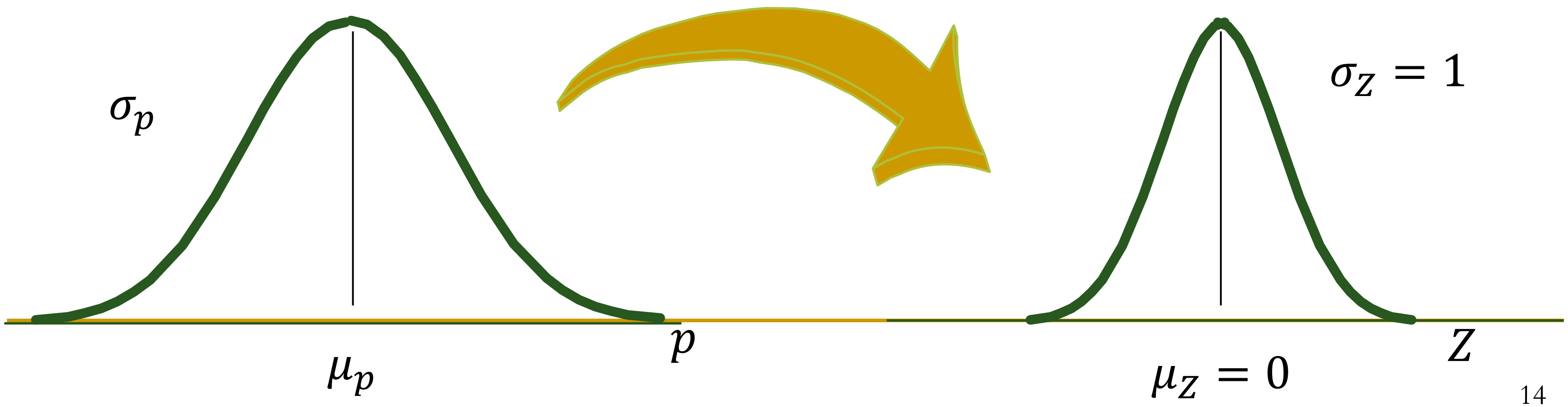
Standardizing Sampling Distribution of Proportion

- Converting the sample proportion p to Z value

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

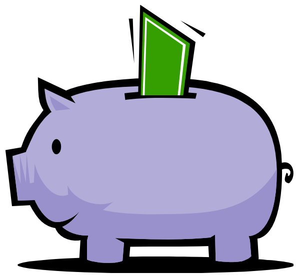
Sampling Distribution of p

Standardized Normal Distribution



Standardizing Sampling Distribution of Proportion – Example

- Suppose that the manager of the local bank determines that 40% of all depositors have multiple accounts at the bank
- If you select a random sample of 200 depositors, what is the probability that the sample proportion of depositors with multiple accounts is less than 0.3 ?



Standardizing Sampling Distribution of Proportion – Example

Cont'd

- Given $\pi =$ population proportion of depositors with multiple accounts $= 0.4$
- As $n = 200 > 30, n\pi = 80 > 5, n(1-\pi) = 120 > 5$
  → The sampling distribution of p follows Normal distribution approximately, i.e. $p \sim N(\mu_p, {\sigma_p}^2)$

$$P(p < 0.3)$$

$$= \mathrm{P}\left(\mathrm{Z} < \frac{0.3-0.4}{\sqrt{\frac{0.4(1-0.4)}{200}}}\right)$$

Standardizing Sampling Distribution of Proportion – Example

Cont'd

- Given π = population proportion of depositors with multiple accounts = 0.4
- As $n = 200 > 30, n\pi = 80 > 5, n(1-\pi) = 120 > 5$
  → The sampling distribution of p follows Normal distribution approximately, i.e. $p \sim N(\mu_p, {\sigma_p}^2)$

$$P(p < 0.3)$$

$$= \mathrm{P}\left(\mathrm{Z} < \frac{0.3-0.4}{\sqrt{\frac{0.4(1-0.4)}{200}}}\right) = P(Z < -2.89)$$

$$= 0.0019$$

Confidence Interval Estimate for the Proportion – Example

Cont'd

For these data, $p = \frac{95}{200} = 0.475$

As $n = 200 > 30, np = 95 > 5, n(1-p) = 105 > 5$

→ The sampling distribution of p follows Normal distribution approximately

95% confidence interval (C.I.) for π

$$p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} = 0.475 \pm 1.96\sqrt{\frac{0.475(1-0.475)}{200}}$$

$$= [0.406, 0.544]$$

We are 95% confident that the population proportion of depositors having RMB deposit account is between 0.406 and 0.544

Determining Sample Size for the Proportion – Example

Cont'd

$$\pi = \frac{22}{10000} = 0.0022$$

$$n = \frac{(Z_{\alpha/2})^2\pi(1-\pi)}{E^2} = \frac{(2.575)^2 0.0022(1-0.0022)}{0.001^2}$$

$$= 14555.28 \cong 14556$$

Round Up

Test of Hypothesis for the Proportion – Exercise

Cont'd

$H_0: \pi = 0.80$
$H_1: \pi \neq 0.80$

$n = 45 > 30$
$np = 39 > 5$
$n(1-p) = 6 > 5$

$\therefore p \sim N$ approximately

At $\alpha = 0.05$

Critical Value = ± 1.96

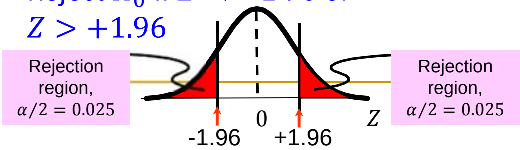
Reject H_0 if $Z < -1.96$ or $Z > +1.96$

$$Z = \frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{\frac{39}{45}-0.80}{\sqrt{\frac{0.80(1-0.80)}{45}}}$$

$$= 1.118$$

At $\alpha = 0.05$, do not reject H_0

There is insufficient evidence that the true proportion of customers to be served within 5 minutes upon their arrival is not 80%



Confidence Interval Estimate for the Proportion

- Since the population proportion, π, is unknown, the standard deviation of p can be estimated by sample standard deviation S_p

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

- Hence, $Z = \frac{p-\pi}{S_p} \sim N(0,1)$ approximately, for large n
- As the population proportion π is unknown, we may verify the "large enough" condition by np and $n(1-p)$

Confidence Interval Estimate for the Proportion

Cont'd

- Conditions
  - The no. of successes, Y, follows Binomial distribution
  - Normal approximation can be used
    - $n \geq 30$
    - $np \geq 5$
    - $n(1-p) \geq 5$

- $100(1-\alpha)\%$ Confidence interval estimate

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

Standard Error, σ_p

Sampling Error, E

Confidence Interval Estimate for the Proportion – Example

- Among the 200 depositors you randomly selected, 95 of them have RMB deposit account at the bank
- Set up a 95% confidence interval estimate for the population proportion of depositors having RMB deposit account at the bank

Confidence Interval Estimate for the Proportion – Example

Cont'd

For these data, $p = \frac{95}{200} = 0.475$

As $n = 200 > 30$, $np = 95 > 5$, $n(1-p) = 105 > 5$

95% confidence interval (C.I.) for π

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

We are 95% confident that the population proportion of depositors having RMB deposit account is between 0.406 and 0.544

Standardizing Sampling Distribution of Proportion – Example

Cont'd

- Given π = population proportion of depositors with multiple accounts $= 0.4$
- As $n = 200 > 30, n\pi = 80 > 5, n(1-\pi) = 120 > 5$
  → The sampling distribution of p follows Normal distribution approximately, i.e. $p \sim N(\mu_p, {\sigma_p}^2)$

$$P(p < 0.3)$$
$$= \mathrm{P}\left(\mathrm{Z} < \frac{0.3-0.4}{\sqrt{\frac{0.4(1-0.4)}{200}}}\right) = P(Z < -2.89)$$
$$= 0.0019$$

Confidence Interval Estimate for the Proportion – Example

Cont'd

For these data, $p = \frac{95}{200} = 0.475$

As $n = 200 > 30, np = 95 > 5, n(1-p) = 105 > 5$

→ The sampling distribution of p follows Normal distribution approximately

95% confidence interval (C.I.) for π

$$p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} = 0.475 \pm 1.96\sqrt{\frac{0.475(1-0.475)}{200}}$$
$$= [0.406, 0.544]$$

We are 95% confident that the population proportion of depositors having RMB deposit account is between 0.406 and 0.544

Determining Sample Size for the Proportion – Example

Cont'd

$$\pi = \frac{22}{10000} = 0.0022$$

$$n = \frac{(Z_{\alpha/2})^2\pi(1-\pi)}{E^2} = \frac{(2.575)^2 0.0022(1-0.0022)}{0.001^2}$$
$$= 14555.28 \cong 14556$$

Round Up

Test of Hypothesis for the Proportion – Exercise

Cont'd

$H_0: \pi = 0.80$
$H_1: \pi \neq 0.80$

$n = 45 > 30$
$np = 39 > 5$
$n(1-p) = 6 > 5$
$\therefore p \sim N$ approximately

At $\alpha = 0.05$

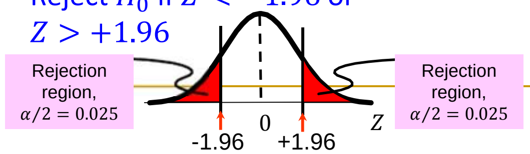
Critical Value = ± 1.96

Reject H_0 if $Z < -1.96$ or $Z > +1.96$

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{\frac{39}{45} - 0.80}{\sqrt{\frac{0.80(1-0.80)}{45}}}$$
$$= 1.118$$

At $\alpha = 0.05$, do not reject H_0

There is insufficient evidence that the true proportion of customers to be served within 5 minutes upon their arrival is not 80%



Confidence Interval Estimate for the Proportion

Cont'd

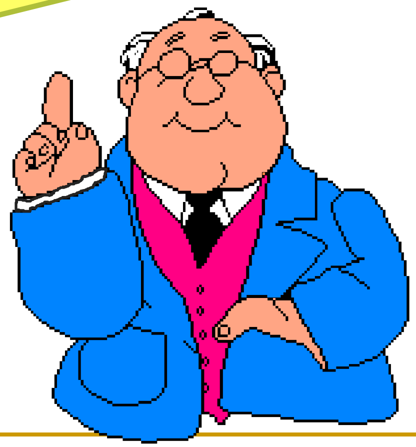
- Special considerations
  - If $p - Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < 0$, we have to replace the lower bound of the confidence interval by **0**
  - If $p + Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} > 1$, we have to replace the upper bound of the confidence interval by **1**
- But why?

Factors Affecting Interval Width (Precision)

- Level of confidence, $(1-\alpha)$
  - $(1-\alpha)\uparrow \;\rightarrow\; |Z\text{-value}|\;\uparrow \;\rightarrow$ width of interval $\uparrow$
- Sample size, n
  - $n\uparrow \;\rightarrow\; \sigma_p \downarrow \rightarrow$ width of interval $\downarrow$
- Sample proportion, p
  - If p increases from 0 to 0.5, then $p(1-p)$ increases from 0 to 0.25, leading to a wider interval
  - If p further increases from 0.5 to 1, then $p(1-p)$ drops from 0.25 to 0, leading to a narrower interval

Intervals extend from

$$p - Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \text{ to } p + Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$



Determining Sample Size for the Proportion

- Sampling error (or margin of error)

$$\mathrm{E} = Z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}$$

- Solving the equation for n gives

$$n = \frac{\left(Z_{\alpha/2}\right)^2 \pi(1-\pi)}{E^2}$$

  - If the computed n is not an integer, round it up to nearest integer

Determining Sample Size for the Proportion – Example

- According to the *Developments in the Banking Sectors* published by Hong Kong Monetary Authority in June 2014, at the end of the first quarter of 2014, 22 credit card lending were found in each 10,000 transactions
- You want to have 99% confidence of estimating the proportion of credit card lending at your bank to within ± 0.001
- What is the minimum sample size being needed?

Determining Sample Size for the Proportion – Example

Cont'd

$$\pi = \frac{22}{10000} = 0.0022$$

$$n = \frac{\left(Z_{\alpha/2}\right)^2 \pi(1-\pi)}{E^2}$$

Standardizing Sampling Distribution of Proportion – Example

Cont'd

- Given π = population proportion of depositors with multiple accounts $= 0.4$
- As $n = 200 > 30, n\pi = 80 > 5, n(1-\pi) = 120 > 5$
  → The sampling distribution of p follows Normal distribution approximately, i.e. $p \sim N(\mu_p, {\sigma_p}^2)$

$$P(p < 0.3)$$

$$= \mathrm{P}\left(\mathrm{Z} < \frac{0.3-0.4}{\sqrt{\frac{0.4(1-0.4)}{200}}}\right) = P(Z < -2.89)$$

$$= 0.0019$$

Confidence Interval Estimate for the Proportion – Example

Cont'd

For these data, $p = \frac{95}{200} = 0.475$

As $n = 200 > 30, np = 95 > 5, n(1-p) = 105 > 5$

→ The sampling distribution of p follows Normal distribution approximately

95% confidence interval (C.I.) for π

$$p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} = 0.475 \pm 1.96\sqrt{\frac{0.475(1-0.475)}{200}}$$

$$= [0.406, 0.544]$$

We are 95% confident that the population proportion of depositors having RMB deposit account is between 0.406 and 0.544

Determining Sample Size for the Proportion – Example

Cont'd

$$\pi = \frac{22}{10000} = 0.0022$$

$$n = \frac{(Z_{\alpha/2})^2 \pi(1-\pi)}{E^2} = \frac{(2.575)^2 0.0022(1-0.0022)}{0.001^2}$$

$$= 14555.28 \cong 14556$$

Round Up

Test of Hypothesis for the Proportion – Exercise

Cont'd

$H_0: \pi = 0.80$
$H_1: \pi \neq 0.80$

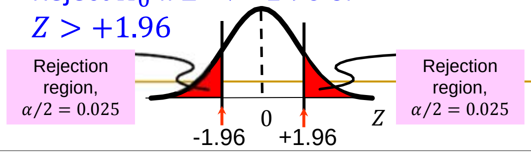
$n = 45 > 30$
$np = 39 > 5$
$n(1-p) = 6 > 5$

$\therefore p \sim N$ approximately

At $\alpha = 0.05$

Critical Value = ± 1.96

Reject H_0 if $Z < -1.96$ or $Z > +1.96$



$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{\frac{39}{45} - 0.80}{\sqrt{\frac{0.80(1-0.80)}{45}}}$$

$$= 1.118$$

At $\alpha = 0.05$, do not reject H_0

There is insufficient evidence that the true proportion of customers to be served within 5 minutes upon their arrival is not 80%

Determining Sample Size for the Proportion

Cont'd

- What should we do if π is unknown?

1. Use p (sample proportion) from some similar studies
   - As p provides the best estimate of π

2. If p also unknown, use 0.5
   - When π = 0.5, $\pi(1 - \pi)$ becomes the largest, i.e. 0.25
   - Hence you can determine a sample size fulfilling the requirement of any other value for the true but unknown π

Test of Hypothesis for the Proportion

- Conditions
  - The no. of successes, Y, follows Binomial distribution
  - Normal approximation can be used
    - $n \geq 30$
    - $np \geq 5$
    - $n(1-p) \geq 5$

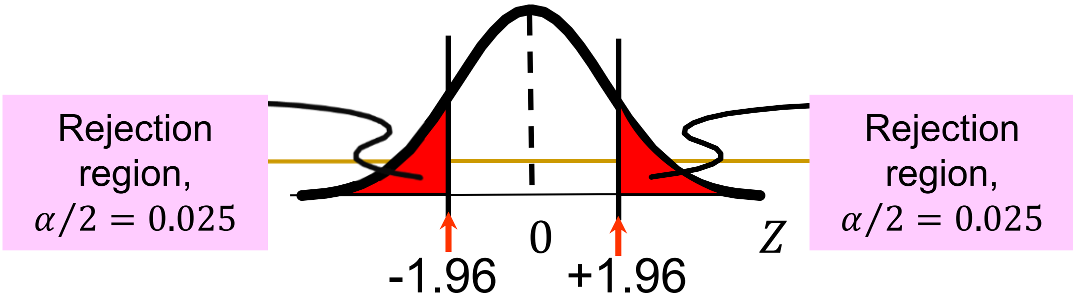
- Test statistic, $Z = \dfrac{p-\pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}}$

Test of Hypothesis for the Proportion – Exercise

- Your bank had the business objective of serving 80% of the customers within 5 minutes upon the time the customer enters the bank
- Of the 45 randomly selected customers, 39 are served within 5 minutes upon their arrival
- Test the claim of the bank at 5% level of significance

Test of Hypothesis for the Proportion – Exercise

Cont'd



Standardizing Sampling Distribution of Proportion – Example

Cont'd

- Given π = population proportion of depositors with multiple accounts = 0.4
- As $n = 200 > 30$, $n\pi = 80 > 5$, $n(1-\pi) = 120 > 5$
  → The sampling distribution of p follows Normal distribution approximately, i.e. $p \sim N(\mu_p, {\sigma_p}^2)$

$$P(p < 0.3)$$
$$= \mathrm{P}\left(\mathrm{Z} < \frac{0.3-0.4}{\sqrt{\frac{0.4(1-0.4)}{200}}}\right) = P(Z < -2.89)$$
$$= 0.0019$$

Confidence Interval Estimate for the Proportion – Example

Cont'd

For these data, $p = \frac{95}{200} = 0.475$

As $n = 200 > 30$, $np = 95 > 5$, $n(1-p) = 105 > 5$

→ The sampling distribution of p follows Normal distribution approximately

95% confidence interval (C.I.) for π

$$p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} = 0.475 \pm 1.96\sqrt{\frac{0.475(1-0.475)}{200}}$$
$$= [0.406, 0.544]$$

We are 95% confident that the population proportion of depositors having RMB deposit account is between 0.406 and 0.544

Determining Sample Size for the Proportion – Example

Cont'd

$$\pi = \frac{22}{10000} = 0.0022$$

$$n = \frac{(Z_{\alpha/2})^2\pi(1-\pi)}{E^2} = \frac{(2.575)^2 0.0022(1-0.0022)}{0.001^2}$$
$$= 14555.28 \cong 14556$$

Round Up

Test of Hypothesis for the Proportion – Exercise

Cont'd

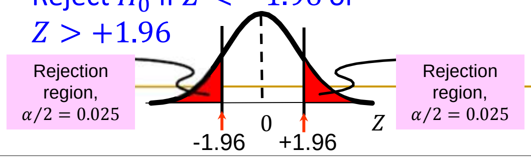
$H_0: \pi = 0.80$
$H_1: \pi \neq 0.80$

$n = 45 > 30$
$np = 39 > 5$
$n(1-p) = 6 > 5$
$\therefore p \sim N$ approximately

At $\alpha = 0.05$

Critical Value = ± 1.96

Reject H_0 if $Z < -1.96$ or $Z > +1.96$



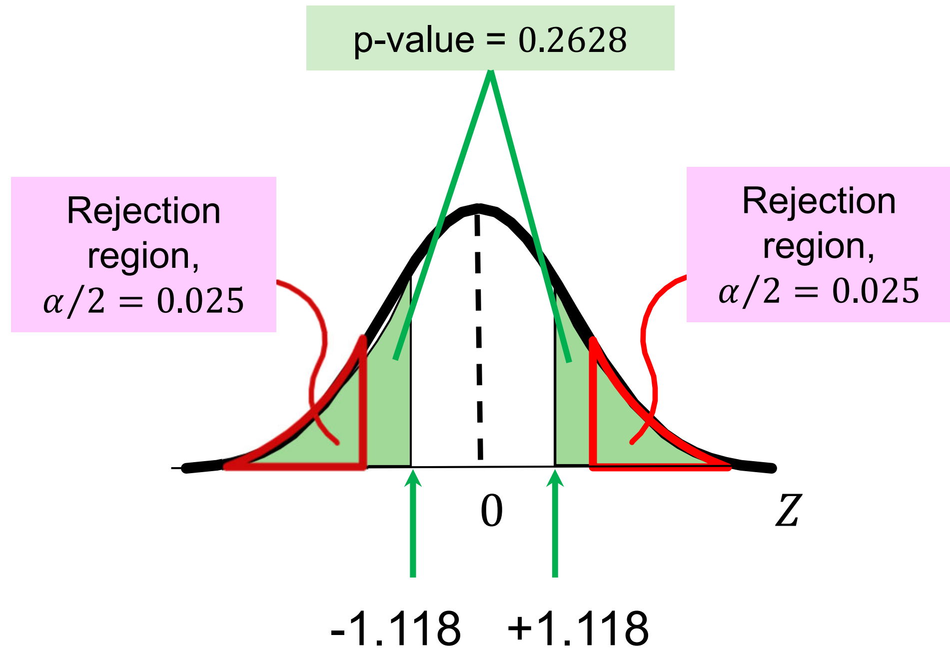
$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{\frac{39}{45} - 0.80}{\sqrt{\frac{0.80(1-0.80)}{45}}}$$
$$= 1.118$$

At $\alpha = 0.05$, do not reject H_0

There is insufficient evidence that the true proportion of customers to be served within 5 minutes upon their arrival is not 80%

Test of Hypothesis for the Proportion – Exercise

Cont'd



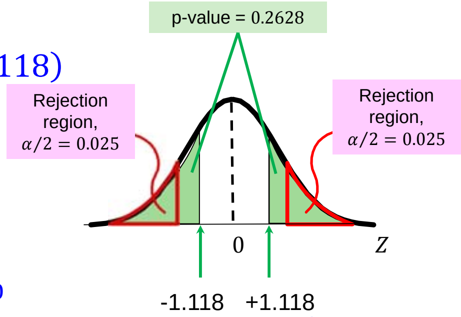
Test of Hypothesis for the Proportion – Exercise

Cont'd

$H_0: \pi = 0.80$
$H_1: \pi \neq 0.80$

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{\frac{39}{45} - 0.80}{\sqrt{\frac{0.80(1-0.80)}{45}}} = 1.118$$

p-value
$= P(Z \leq -1.118) + P(Z \geq 1.118)$
$= 2 \times P(Z \leq -1.118)$
$= 2 \times 0.1314$
$= 0.2628$



As p-value > α, do not reject H_0

There is insufficient evidence that the true proportion of customers to be served within 5 minutes upon their arrival is not 80%

Do Voters Really Vote When They Say They Do?

- On November 8, 1994, a historic election took place in US, in which the Republican Party won control of both houses of Congress for the first time since 1952
- But how many people actually voted?
- On November 28, 1994, *Time* magazine reported that in a telephone poll of 800 adults taken during the two days following the election, 56% reported that they had voted
- But based on information from the Committee for the Study of the American Electorate, in fact, only 39% of American adults had voted
- Could it be the case that the results of the poll simply reflected a sample that, by chance, voted with greater frequency than the general population?

Do Voters Really Vote When They Say They Do?

Cont'd

- Let's suppose that the truth about the population is that only 39% of American adults voted, i.e. $\pi = 39\% = 0.39$
- We can expect in samples of 800 adults, the size used by the *Time* magazine poll, the mean is 0.39 and standard error is 0.017, i.e. $\mu_p = \pi = 0.39$ and $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = 0.017$
- According to the Empirical Rule, we are almost certain that the sample proportion based on a sample of 800 adults should fall within $3 \times 0.017 = 0.051$ of the truth of 0.39
- In order words, if respondents were telling the truth, the sample proportion should be no higher than 44.1% (=39%+5.1%), no where near the reported percentage of 56%

Do Voters Really Vote When They Say They Do?

Cont'd

- We can also find how likely the sample proportion of 0.56 or above to happen
- Given $n = 800$, $\mu_p = 0.39$ and $\sigma_p = 0.017$
- $P(p \geq 0.56) = P(Z \geq 10) \approx 0$
- It is virtually impossible to have such high proportion of voters voted in the election

- The differences between data may be the result of a variety of factors
  - Differences in the respondents' interpretation of the questions
  - Respondents' inability or unwillingness to provide correct information or recall correct information