Basic Statistics

Nguyen Thi Tu Van   Nguyen Quang Vinh

JICA project - August 2, 2015

Outline

Aim

For clinicians, understanding concepts in statistics:

- to read and use information of published medical evidence
- to reinforce basic knowledge biostatistics required for designing research

Statistics - Biostatistics

- Statistics
  - a science and art of dealing with data
  - to study of uncertainty, to obtain reliable results
- Biostatistics - an application of statistics to biological sciences: medicine, education, agriculture...
- Modern society - Reading, Writing, Statistical thinking

Objectives ⇌ Statistics

Objectives:
(1) Organize & summarize data
(2) Reach inferences: sample → population
Statistics:

- *Descriptive statistics→(1)*
- *Inferential statistics: drawing of inferences→(2)*
  - *Estimation (point estimate & interval estimate ≡ confidence interval)*
  - *Hypothesis testing →reaching a decision (p value)*
    - *Parametric statistics*
    - *Non-parametric statistics << Distribution-free statistics*
  - *Modeling, Predicting: a combination of estimation and hypothesis testing*

Why estimation?

- Two reasons:
  - Infinite populations: incapable of complete examination
  - Finite populations: cost, time
- In addition, estimation can help not to defer a conclusion until the whole population is observed

Which estimators?

- mean(s):
  - a single population mean: $\bar{x} \to \mu$
  - the difference between two population means - unpaired, paired: $(\bar{x_1} - \bar{x_2}) \to (\bar{\mu_1} - \bar{\mu_2})$
- proportion(s):
  - a single population proportion: $\hat{p} \to p$ or $\hat{p} \to \pi$
  - the difference of two population proportions: $(\hat{p_1} - \hat{p_2}) \to (\pi_1 - \pi_2)$
- variance(s):
  - a single population variance: $s^2 \to \sigma^2$
  - the ratio of two population variances: $\frac{s_1^2}{s_2^2} \to \frac{\sigma_1^2}{\sigma_2^2}$

An estimation of these parameters includes: Point estimate & Interval estimate

A point estimate
Estimator → Parameter

A parameter may be estimated by more than one estimator.

- Example:
  - Sample mean → estimate population mean
  - Sample median → estimate population mean

An interval estimate

- In general, an interval estimate is obtained by the formula: *estimator ± (reliability coefficient) x (standard error)*
- What is different is the source of the reliability coefficient: t, or z.
  - When sampling is from an approximate normal distribution, and/or unknown variance, an interval estimate for μ may be expressed as: $\overline{x} \pm t_{df,\alpha/2} SE$
  - When sampling is from a normal distribution, with known variance, an interval estimate for μ may be expressed as: $\overline{x} \pm z_{\alpha/2}\sigma_{\overline{x}}$

How to interpret the interval

- In repeated sampling $100(1-\alpha)\%$ of all intervals of the form will in the long run include the population mean, μ
- The quantity $(1-\alpha)$, is called the confidence coefficient & The interval $\overline{x} \pm z_{\alpha/2}\sigma_{\overline{x}}$, is called the confidence interval for μ
- *The most frequently used values are: .90, .95, .99, which have associated reliability factors, respectively, of 1.645, 1.96, 2.58*

The practical interpretation

- We are $100(1-\alpha)\%$ confident that the single computed interval $\overline{x} \pm z_{\alpha/2}\sigma_{\overline{x}}$ contains the population mean, μ
  Example: ...
- E = margin error = maximum error = practical / clinical acceptable error:
  $E = z_{\alpha/2}\sigma_{\overline{x}} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

Why hypothesis testing?

- Hypothesis (H.): a statement concerns about some one or more populations
- Testing hypothesis: to aid researcher in reaching a decision concerning a population by examining a sample from that population

Two types of hypotheses

(1) Research Hypotheses:

- The conjecture or supposition
- It may be the results of years of observation
- Research H. leads directly to Statistical H.

(2) Statistical Hypotheses: Hypotheses are stated in such a way that they may be evaluated by appropriate statistical techniques.

- H_O
- H_A

Statistical Hypotheses

- H_O
  - The H_O is the hypothesis that is tested
  - The H_O should contain either $=, \leq, \geq$
    (The statement concerns about some one or more population's parameters in term of equality or inequality)
- H_A
  - What we hope or expect to be able to conclude as a result of the test usually should be placed in the H_A
- The H_O & H_A are complementary
- One-sided vs. Two-sided Hypothesis Tests

Notes

- Neither hypothesis testing nor statistical inference leads to proof a hypothesis.
  It merely indicates whether the hypothesis is supported or not supported by the available data

p-value

Test statistic ⟹p-value

- General formula:
  $Teststatistic = \frac{relevantstatistic - hypothesizedparameter}{S.E.oftherelevantstatistic} \Rightarrow p\text{-}value$
  Example: $z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$
- p-value:
  - Probability of observing the difference if the H_O is true.
  - An expression of how we believe in the H_O.
  - Decision maker.
- p-value overdependence

Type I & Type II error

Conditions under which type I & type II errors may be committed (the four possibilities)		Actual Situation (Truth in the population)	
		H_o false	**H_o true**
The results in the study sample → Conclusion:	**Reject H_o**	Correct decision	*Type I error*
	Fail to reject H_o	*Type II error*	Correct decision

Decision rule for a rejection or not the H_O

- α = risk of committing a type I error = level of significance (say, .01, .05, .10)
  β = risk of committing a type II error (say, .05, .10, .20)
- When $p < \alpha$: we reject H_O, risk of rejecting a true H_O
- When we fail to reject a H_O, risk of "accepting" a false H_O
  - We do not say that it is true, but that it may be true
  - We do not wish to convey the idea that "accepting" implies proof

Testing Hypothesis → Rejected or not rejected H_O

In the testing process, the H_O either is rejected or is not rejected:

- If H_O is not rejected, we will say that the data on which the test is based do not provide sufficient evidence to cause rejection
- If the testing process leads to rejection, we will say that the data at hand are not compatible with the H_O, but are supportive of some other hypothesis & may be designated by H_A (H_A a contradiction statement of H_O)

The Five-Step practical procedure for Hypothesis Testing *(opt.)*

- Step 1: Set up H_O, H_A
  - 1. Data: The nature of the data (classification)
  - 2. Assumptions: The normality of the population distribution, equality of variances, independence of samples...
  - 3. Hypotheses: H_O, H_A
- Step 2: Define the test statistic
  - 4. Test statistic
  - 5. Distribution of the Test Statistic

The Five-Step practical procedure for Hypothesis Testing, *cont. (opt.)*

- Step 3: Define a rejection region: having determined a value for α
  - 6. Decision rule
- Step 4:
  - 7. Calculate the value of the test statistic, and compare it with the acceptance & rejection regions that have already been specified.
  - 8. State our decision: to reject H_O or to fail to reject H_O
- Step 5:
  - 9. Give a conclusion: this statement should be free of statistical jargon & should merely summarize the results of the analysis.

Note

A statistical package helps you only the step 4, but not for the other steps.

Summary

Statistics:

- Descriptive statistics→organize & summarize data
- Inferential statistics →drawing of inferences:
  - Estimation
    - estimator, confidence interval - very helpful as it gives a range of likely values
  - Hypothesis testing
    - p-value, or "surprise-level" value

Exercises

1. A survey of 100 delivery cases in a population gave us the mean of newborn baby weight is 3,050 grams; 95% confidence interval is of (2,950 - 3,150 grams). Please give your explanation for the results.
2. Please give your brief explanations of the p value.
   1. ...
   2. ...
   3. ...