Learning Bio- basic terms

Nguyen Thi Tu Van[1] Nguyen Quang Vinh[2]

[1]University of Medicine and Pharmacy of Ho Chi Minh City

[2]Nguyen Tri Phuong Hospital

JICA project - August 3, 2014

Outline

1. Introduction
2. Descriptive Statistics
    - Category
    - Summary measures
3. Inferential Statistics
    - Estimation - Confidence Interval
    - A point estimate - An interval estimate
    - Interpretation a confidence interval
4. Hypothesis testing - p value
    - Hypothesis testing
    - Hypotheses
    - p-value
    - The hypothesis testing procedure
    - Power of Test

Aim

For clinicians, understanding concepts and basic terms in statistics:

- to apply in reading and using information of published medical evidence
- to reinforce basic knowledge biostatistics required for designing research

Statistics - Biostatistics

- Statistics
  - a science (concepts, rules, procedures) - and art of dealing with data
  - to study of uncertainty, to obtain reliable results
- Biostatistics - an application of statistics to:
  - medicine
  - biological sciences (education, psychology, agriculture...)
- Modern society
  - Reading
  - Writing
  - Statistical thinking: to make the strongest possible conclusions from limited amounts of data.

Objectives $\rightleftharpoons$ Statistics

Objectives:

(1) Organize & summarize data

(2) Reach inferences: sample $\rightarrow$ population

$\rightleftharpoons$*Statistics:*

- *Descriptive statistics*$\rightarrow$*(1)*
- *Inferential statistics: drawing of inferences*$\rightarrow$*(2)*
  - *Estimation (point estimate & interval estimate* $\rightarrowtail$ *confidence interval)*
  - *Hypothesis testing* $\rightarrow$*reaching a decision (p value)*
    - *Parametric statistics*
    - *Non-parametric statistics* $<<$ *Distribution-free statistics*
  - *Modeling, Predicting*

Why basic terms?

- empowering belief:
  Statistics = Terms + (Normal distribution + Others)
- and,
  Terms = Basic terms + Others

Descriptive Statistics

- to categorize available data by grouping
- to describe its distribution by calculating summary measures

Outline

1. Introduction
2. Descriptive Statistics
   - Category
   - Summary measures
3. Inferential Statistics
   - Estimation - Confidence Interval
   - A point estimate - An interval estimate
   - Interpretation a confidence interval
4. Hypothesis testing - p value
   - Hypothesis testing
   - Hypotheses
   - p-value
   - The hypothesis testing procedure
   - Power of Test

Grouping

- Raw data needs to be organized - to be presented in a clearer way.
- Data must be grouped into categories, as an example below:

Age (years-old)	Frequency	Relative frequency	Cumulative frequency
Under 20	100	20%	100
20 - 39	350	70%	450
40 or over	50	10%	500

Outline

1. Introduction
2. Descriptive Statistics
   - Category
   - Summary measures
3. Inferential Statistics
   - Estimation - Confidence Interval
   - A point estimate - An interval estimate
   - Interpretation a confidence interval
4. Hypothesis testing - p value
   - Hypothesis testing
   - Hypotheses
   - p-value
   - The hypothesis testing procedure
   - Power of Test

Calculating summary data

Three summary measures:

- Mean (average) = Sum of all data / n
- Median = 50th percentile value
- Mode = Most frequent value(s)

Measures of dispersions

Four measures to show how data is spread around the summary measures (variability)

- Range from minimum to maximum
- Variance = Mean of squared differences of values from a mean
- Standard deviation (SD) = Square root of the variance.
    - SD indicates the mean of differences of values from the mean.
    - It measures absolute dispersion.
- Coefficient of variance (CV) = Ratio of SD to the mean.
    - CV indicates the extent of variability in relation to the mean.
    - It measures relative dispersion.
    - There is no unit to measure CV, therefore they can be compared between any sets of data.

The measures of position

- Two measures of position:
  - Percentile = percent of values at or below
    - Quartile = 25th, 50th, and 75th percentiles
  - Z-score = a standardized value indicating a distance between the value and the mean divided by (in units of) the standard deviation.

Outline

1. Introduction
2. Descriptive Statistics
   - Category
   - Summary measures
3. Inferential Statistics
   - Estimation - Confidence Interval
   - A point estimate - An interval estimate
   - Interpretation a confidence interval
4. Hypothesis testing - p value
   - Hypothesis testing
   - Hypotheses
   - p-value
   - The hypothesis testing procedure
   - Power of Test

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

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

An estimation of these parameters:

- Point estimate
- Interval estimate

Outline

1. Introduction
2. Descriptive Statistics
   - Category
   - Summary measures
3. Inferential Statistics
   - Estimation - Confidence Interval
   - A point estimate - An interval estimate
   - Interpretation a confidence interval
4. Hypothesis testing - p value
   - Hypothesis testing
   - Hypotheses
   - p-value
   - The hypothesis testing procedure
   - Power of Test

A point estimate

Estimator → Parameter

In many cases, 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: In particular, when sampling is from a normal distribution with known variance, an interval estimate for μ may be expressed as: $\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$

Outline

1. Introduction
2. Descriptive Statistics
   - Category
   - Summary measures
3. Inferential Statistics
   - Estimation - Confidence Interval
   - A point estimate - An interval estimate
   - Interpretation a confidence interval
4. Hypothesis testing - p value
   - Hypothesis testing
   - Hypotheses
   - p-value
   - The hypothesis testing procedure
   - Power of Test

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 $\overline{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$, 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 $\overline{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$ contains the population mean, μ
  Example: ...
- E = margin error = maximum error = practical / clinical acceptable error:
  $E = z_{\alpha/2}\sigma_{\bar{x}} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

Outline

1. Introduction
2. Descriptive Statistics
   - Category
   - Summary measures
3. Inferential Statistics
   - Estimation - Confidence Interval
   - A point estimate - An interval estimate
   - Interpretation a confidence interval
4. Hypothesis testing - p value
   - Hypothesis testing
   - Hypotheses
   - p-value
   - The hypothesis testing procedure
   - Power of Test

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

Hypothesis testing for

- mean(s)
  - a single population mean
  - the difference between two population means: unpaired, paired
- proportion(s)
  - a single population proportion
    - unpaired: a small sample, a sufficiently large sample
    - paired
  - the difference of two population proportions
- variance(s)
  - a single population variance
  - the difference of two population variances

Outline

1. Introduction
2. Descriptive Statistics
   - Category
   - Summary measures
3. Inferential Statistics
   - Estimation - Confidence Interval
   - A point estimate - An interval estimate
   - Interpretation a confidence interval
4. Hypothesis testing - p value
   - Hypothesis testing
   - Hypotheses
   - p-value
   - The hypothesis testing procedure
   - Power of Test

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, in general, leads to proof a hypothesis
- It merely indicates whether the hypothesis is supported or not supported by the available data
- When we *fail to reject the H_O*, we do not say that it is true, but that it may be true
- When we speak of "*accepting*" *a H_O*, we have this limitation in mind & do not wish to convey the idea that "accepting" implies proof

Outline

1. Introduction
2. Descriptive Statistics
   - Category
   - Summary measures
3. Inferential Statistics
   - Estimation - Confidence Interval
   - A point estimate - An interval estimate
   - Interpretation a confidence interval
4. Hypothesis testing - p value
   - Hypothesis testing
   - Hypotheses
   - p-value
   - The hypothesis testing procedure
   - Power of Test

p-value

Test statistic $\Longrightarrow$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

Decision rule for a rejection or not the H_O

- α = type I error = level of significance (say, .01, .05, .10)
- β = type II error (say, .05, .10, .20)
- When we reject a $H_O \Longleftrightarrow p < \alpha$, risk of committing a type I error, rejecting a true H_O
- When we fail to reject a H_O: risk of committing a type II error, "accepting" a false H_O

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

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)

Outline

1. Introduction
2. Descriptive Statistics
   - Category
   - Summary measures
3. Inferential Statistics
   - Estimation - Confidence Interval
   - A point estimate - An interval estimate
   - Interpretation a confidence interval
4. Hypothesis testing - p value
   - Hypothesis testing
   - Hypotheses
   - p-value
   - The hypothesis testing procedure
   - Power of Test

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.

Outline

1 Introduction
2 Descriptive Statistics
- Category
- Summary measures

3 Inferential Statistics
- Estimation - Confidence Interval
- A point estimate - An interval estimate
- Interpretation a confidence interval

4 Hypothesis testing - p value
- Hypothesis testing
- Hypotheses
- p-value
- The hypothesis testing procedure
- Power of Test

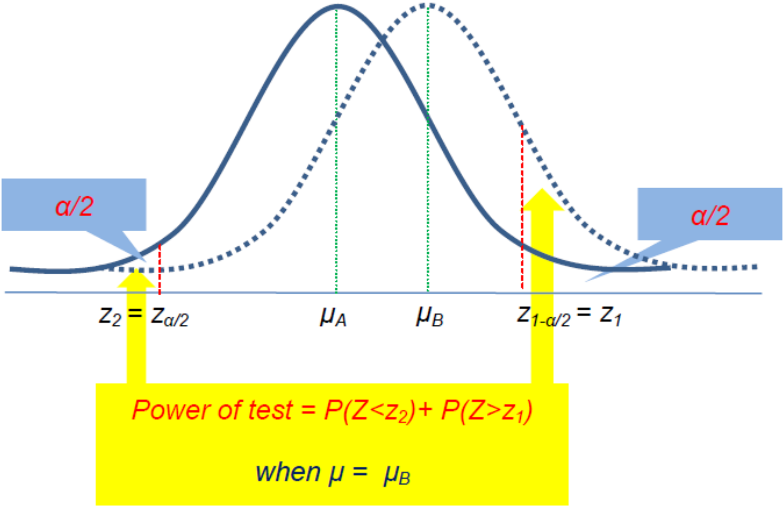
The Power of a Statistical Test *(opt.)*

- The probability of a type II error, β, has remained a phantom:
    - we know it is there,
    - but we don't know what it is
- One thing we can say is that: a wide C.I. for μ means that the corresponding 2-tailed test of Ho versus HA has a large chance of failing to reject a false Ho; that is β is large.

Determining β *(opt.)*

- β = P(fail to reject H_O when H_O is false) $\Longleftrightarrow$ 1 - β = P(rejecting H_O when H_O is false)
- 1 - β represents the probability of making a correct decision in the event that H_O is false
- Since we like β to be small, that is we prefer 1 - β to be large
- The value of 1 - β is referred to as the power of test

Power of test *(opt.)*



Determination of sample size *(opt.)*

- Estimating a confidence interval: for descriptive purposes, investigators just need to consider the type I error.
- Testing a hypothesis: both type I & II errors need to be taken into account.
- When the sample size is fixed, because of some logistical reasons (financial, and/or time constrains):
  - the precision of the available data, and
  - whether this is meaningful for the study objective.

Summary

1. Statistics:
   - Descriptive statistics→organize & summarize data
   - Inferential statistics →drawing of inferences
     - Estimation
     - Hypothesis testing
     - Modeling
2. Estimation - confidence interval, estimator - very helpful as it gives a range of likely values
3. 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. ...