UNIVERSITY OF HAWAI'I
Office of Biostatistics & Quantitative Health Sciences
JOHN A. BURNS SCHOOL OF MEDICINE

Research Design & Biostatistics

Lecture 2

John J. Chen, Ph.D.
Office of Biostatistics & Quantitative Health Sciences
UH JABSOM

OB/GYN Friday Conference

October 14, 2016

Lecture Note: http://biostat.jabsom.hawaii.edu/Education/training.html

Outline

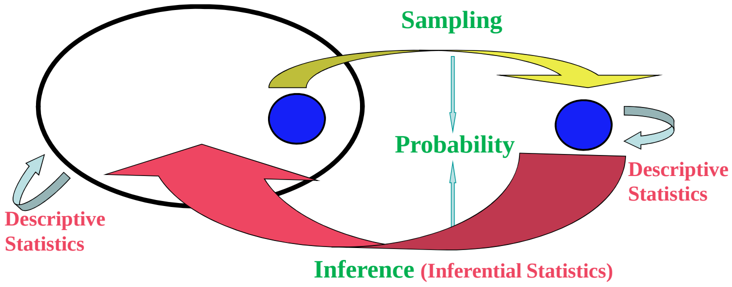
Lecture 1 (08/12/2016)

- Introduction to descriptive biostatistics
  - The goal of statistics
  - Descriptive statistics
  - Different ways of presenting data
- Some research design and data presentation issues

Lecture 2 (10/14/2016)

- Introduction to inferential statistics
- Some commonly used statistical approaches
- Multiple testing problems

The Goal of Statistics



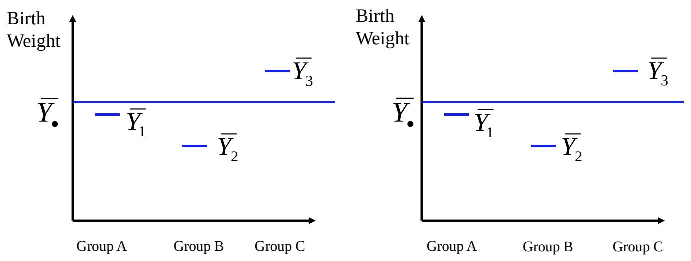
Population
[Parameters]
(μ, σ)

Sample
[Statistics]
($\bar{X}$, S)

Formulation of Research Question

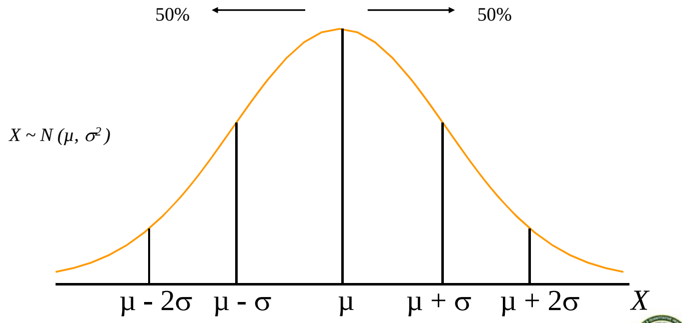
- Null hypothesis & alternative hypothesis
- Specific outcome(s) and how to measure them
- Treatment / control groups
- How to declare success
- Identifying potential sources of variation
- Statistical test to be used
- Potential confounding variables
- Missing data

Effects and Variability



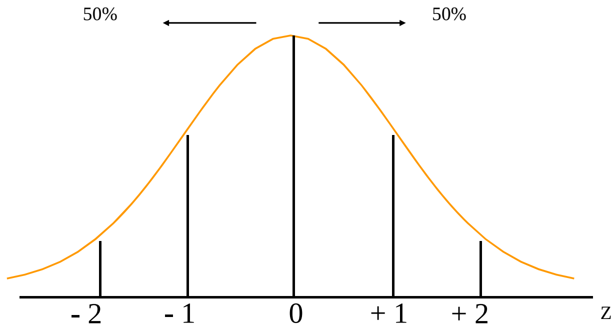
Note: Biological/clinical significance vs. statistical significance

The Normal Distribution



The Normal Distribution

Standard normal distribution: $Z \sim N(\mu = 0, \sigma^2 = 1)$



Given $X \sim N(\mu, \sigma^2)$, we have $Z=(X - \mu)/\sigma$.

AUC For Normal Distribution

The Rule of Thumb:

Within one s.d.:	68.27% (2/3)
Within two s.d.:	95.45% (95%)
Within three s.d.:	99.74% (99%)

Sampling Distribution

The distribution of individual observations versus the distribution of sample means:

Central Limit Theorem

The distribution of sample means (sampling distribution) from a population is approximately normal as long as the sample size is large, i.e.,

$$\bar{X} \sim N(\mu_{\bar{X}}, \sigma^2_{\bar{X}}) \quad \rightarrow \quad Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

1. The population distribution can be non-normal.
2. Given the population has mean μ, then the mean of the sampling distribution, $\mu_{\bar{X}} = \mu$.
3. If the population has variance σ^2, the standard deviation of the sampling distribution, or the standard error (a measure of the amount of sampling error) is $\sigma_{\bar{X}} = s.e.(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.

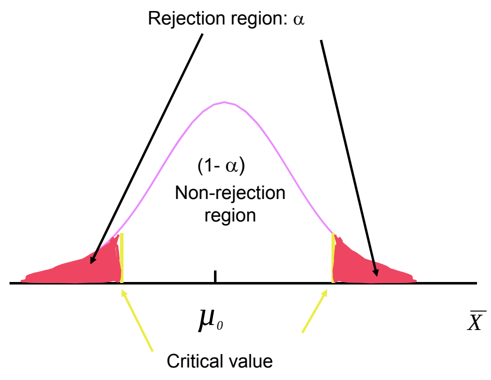
Hypothesis Testing

An Example:

Normal serum creatinine level depends on the population studied. From the literature an OB resident found that one well-established study showed an average sCr of 0.56 (with a standard deviation of 0.15 mg/dL) for 2nd trimester Caucasian pregnant women living on the east coast. But based on her knowledge and experience, she believed that the μ of sCr among Japanese pregnant women in Hawaii seemed different.

She decided to test this by measuring sCr of 49 local Japanese 2nd trimester pregnant women.

Hypothesis Testing (cont.)



Hypothesis Testing

Basic steps of hypothesis testing:

1. State null (H_0:) and alternative (H_1:) hypotheses
2. Choose a significance level, α (usually 0.05 or 0.01)
3. Determine the critical (or rejection) region and the non-rejection region, based on the sampling distribution and under the null hypothesis
4. Based on the sample, calculate the test statistic and compare it with the critical values
5. Make a decision, and state the conclusion

Errors & Power

Type I Error (α) - False positives, errors due to chance

- Reject H_0 when H_0 is true

Type II Error (β) - False negatives

- Don’t reject H_0 when H_1 is true

Power: (1- β) = 1 - P (Type II Error)

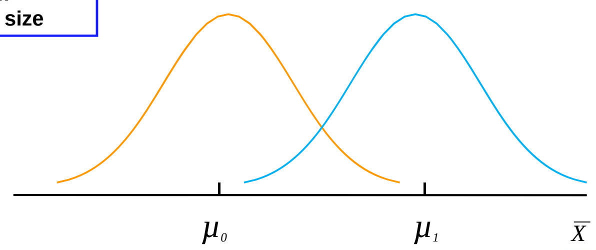
Statistical Decision

Truth

Decision		H_0 True	H_0 False
	Reject H_0	α	1- β
	Not reject H_0	1- α	β

Statistical Decision

Design factors:
- effect difference
- power
- alpha level
- std. dev.
- sample size



Now, what is a p-value?

p-values

Interpretation:

The *p*-value is the probability of obtaining a result as extreme or more extreme than the one observed based on the current sample, given the null hypothesis is true.

Note: "Statistically significant" does not necessarily mean "biologically (or clinically) significant"!!!

Study Design: Power & Sample Size

Five General Design Factors:

1. Effect difference
2. Variability
3. Statistical power ($1- \beta$)
4. α level (Type I error)
5. Sample size

$$\text{Sample Size} = f(DF1, DF2, DF3, DF4)$$

$$\text{Statistical Power} = f(DF1, DF2, DF4, DF5)$$

$$\text{Sample Size} = f(DF1 = 0.62 - 0.56, DF2 = 0.15, DF3 = 0.80, DF4 = 0.05) \approx 49.$$

Hypothesis Testing (cont.)

Example (cont.): Say, the average sCr of the sample of 49 locals is 0.60 mg/dL and the population standard deviation is 0.15 mg/dL (based on the literature).

Step 1. State H_0: and H_1:

$H_0 : \mu_{sCr} = 0.56$ vs. $H_1 : \mu_{sCr} \neq 0.56$

Step 2. Choose a significant level, say, α =0.05.

Step 3. Calculate the test statistic:

$$Z = \frac{\overline{X} - \mu_{sCr}}{\sigma / \sqrt{n}} = \frac{0.60 - 0.56}{0.15 / \sqrt{49}} = 1.87.$$

Hypothesis Testing (cont.)

Step 4. Determine the critical region and the non-rejection region:

The critical value: $\pm$ *1.96.*

The rejection region: $| Z | \geq 1.96$.

The non-rejection region: $| Z | < 1.96$.

Step 5. Make a decision, based on the sample, and state the conclusion: As the test statistic $Z = 1.87 < 1.96$, it is within the non-rejection region. Therefore, we do not reject the null hypothesis. We conclude that there is no evidence that the average sCr among local Japanese 2nd trimester women is different from 0.56 mg/dL.

Confidence Intervals

CIs for μ:

90% CI : $\overline{X} \pm 1.645 \frac{\sigma}{\sqrt{n}}$

95% CI : $\overline{X} \pm 1.960 \frac{\sigma}{\sqrt{n}}$

99% CI : $\overline{X} \pm 2.575 \frac{\sigma}{\sqrt{n}}$

Confidence Intervals

95% Confidence Interval for μ: $\overline{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$

Definition 1: You can be 95% sure that the true mean (μ) will fall within the upper and lower bounds.

Definition 2: 95% of the intervals constructed using sample means, $\overline{x}$, will contain the true mean (μ).

Guinness & The Student’s *t*-Test

- A small sample from normal distribution
- Unknown population standard deviation, σ

$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$ with *n -1* degrees of freedom.

The (Student’s) t-distribution is very similar to normal distribution, with heavier tails.

χ^2 Tests

Expected and observed frequencies are compared

- Goodness of fit of a single variable
- Test of independence of two variables

χ^2 Test of Independence

Observed:

	C1	C2	
R1	A	C	A+C
R2	B	D	B+D
	A+B	C+D	A+B+C+D

Expected:

$$\text{Exp} = \frac{\text{(row total) * (column total)}}{\text{(grand total)}}$$

e.g., $E_{1,1}$=(A+C)*(A+B) / (A+B+C+D)

$$\chi^2 = \sum_{i,j} \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}}$$

d.f. = (r-1)(c-1)=(2-1)*(2-1) = 1*

Tea Tasting & Fisher's Exact Tests

Fisher's Tea Tasting Experiment

Guess Poured First

		Milk	*Tea*	
Poured First	*Milk*			*4*
	Tea			*4*
		4	*4*	*8*

Fisher's Tea Tasting Experiment

Based on hypergeometric distribution, one can calculate the probability of obtaining each table, and the p-value is the sum of all probabilities for tables that give even more evidence in favor of the lady's claim.

Guess Poured First

Poured First		*Milk*	*Tea*
	Milk	0	4
	Tea	4	0

… …

Guess Poured First

	Milk	*Tea*
Milk	3	1
Tea	1	3

Guess Poured First

	Milk	*Tea*
Milk	4	0
Tea	0	4

p-value = $P_{(1,1)}(3) + P_{(1,1)}(4) = 0.229 + 0.014 = 0.243$

Therefore, the experiment did not establish a significant association between the actual order of pouring and the woman's guess.

Sources of Multiplicity

1. Multiple treatments (e.g., multiple comparisons problem)
2. Multiple endpoints (or outcome measures)
3. Multiple measurements over time (e.g., repeated measures problem)
4. Subgroup analyses
5. Interim analyses (e.g., the multiple looks problem)

Consequences of Multiplicity

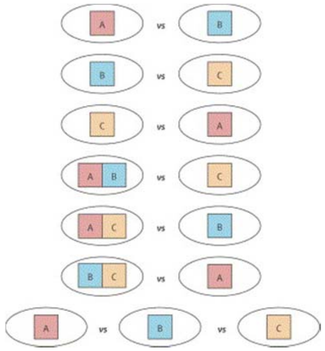
Given a planned alpha=0.05,

m: # of independent tests	1	2	3	4	5	10	20	50	100
Probability of at least one false-positive results	0.05	0.10	0.14	0.19	0.23	0.40	0.64	0.92	>0.99

$$P(\geq 1\,\text{false-positive}) = 1-(1-\alpha)^m.$$

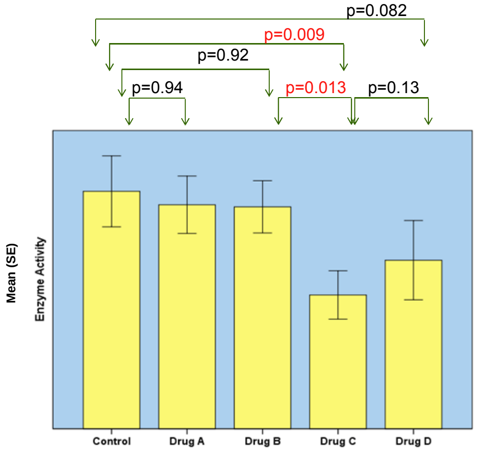
Multiple Treatments

For a three-arm trial, there are at least seven possible comparisons.



Investigators should specify a priori the comparisons intended.

Multiple Treatments



Subgroup Analyses

In a study reviewing 50 reports randomly selected from general medical journals (*JAMA, NEJM, The Lancet, and BMJ*), 70% reported subgroup analyses. Of them, 40% did at least six subgroup analyses, one with 24. Some of "exciting" subgroup analysis results was highlighted in the conclusions.

Pocock et al. (2002). *Statistics in Medicine*; 21:2917

Subgroup Analyses

- Seeking positive subgroup effects, in the absence of overall effects, is purely data-dredging
- Similarly, in a trial with a clear overall effect, subgroup testing can produce false-negative results due to chance and/or lack of power

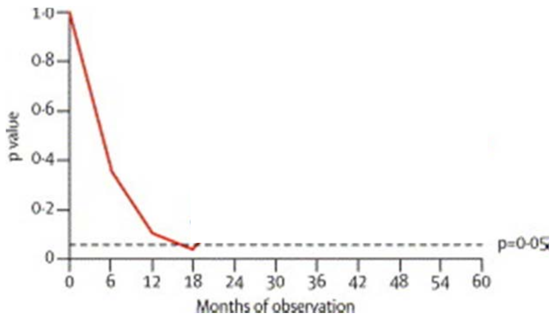
Reasons for Early Stopping

- Superiority of the new treatment
- Inferiority of the new treatment
- Slow accrual
- Poor data quality
- Poor adherence
- Resource deficiencies
- Unacceptable adverse effects
- Fraud
- Emerging information that make the trial irrelevant, unnecessary, or unethical

The Problem of Interim Analyses

- can't be avoided for data monitoring
- investigators may be tempted to do analyses on the main endpoint with accumulating data
- repeated, especially unplanned, interim analyses will increase false-positive rate
- can't use regular statistical approaches

Interim Analysis / Early Stopping

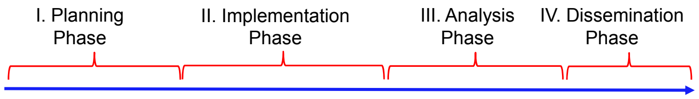


Interim analyses done every 6 months for 5 years.
The p-value is shown for the comparison between the treatment group and control group.

Recommendation on Interim Analysis

- All interim analyses should be planned in advance, including the prespecified statistical stopping method
- Best be done by an independent data safety monitoring committee (DSMC)
- Main goal: make sure overall probability of type I error is controlled
- You have to pay a price with an interim analysis

A Statistician Can Help



I.
- Provide a new and less biased perspective on your study
- Clarify and formalizing the research hypothesis
- Define the primary and secondary outcome variables
- Determine the appropriateness of the research design
- Consider the issues of bias, blinding, stratification, missing data, data and safety monitoring
- Figure out justifiable sample size and statistical power
- Specify a detailed and appropriate statistical analysis plan

II.
- Provide interim analysis for data and safety monitoring
- Conduct data checking for quality control
- Develop or adapt statistical tools for the study

III – IV.
- Execute the statistical analysis plan: descriptive and inferential analyses
- Statistical methods section, TLG, and results interpretation for publications

Common Questions from A Statistician

1. What is the research hypothesis?
2. What is your primary outcome variable?
3. What type of variable is it? How is it measured?
4. How many groups or arms in your study?
5. What is a biological/clinical meaningful difference?
6. How many subjects can you recruit or have access to?
7. Do you expect loss-to-follow-up?
8. How much variation do you expect? Any preliminary data?
9. Are there other variables that might affect the results?

University of Hawaii
Office of Biostatistics & Quantitative Health Sciences
John A. Burns School of Medicine

http://biostat.jabsom.hawaii.edu

RMATRIX U54MD007584
BRIDGES G12MD007601
IDeA Network of Biomedical Research Excellence INBRE III Hawaii State Research and Education Partnership P20GM103466