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

P-value & Other Statistical Strangeness

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

Family Medicine Resident Research Session
March 13, 2019

Outline

- What exactly is p-value?
- Re-introduce some key concepts related to hypothesis testing
- Statistical strangeness: a few examples

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

Email: jjchen@hawaii.edu

The Value of P-value

ASA News
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®
732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

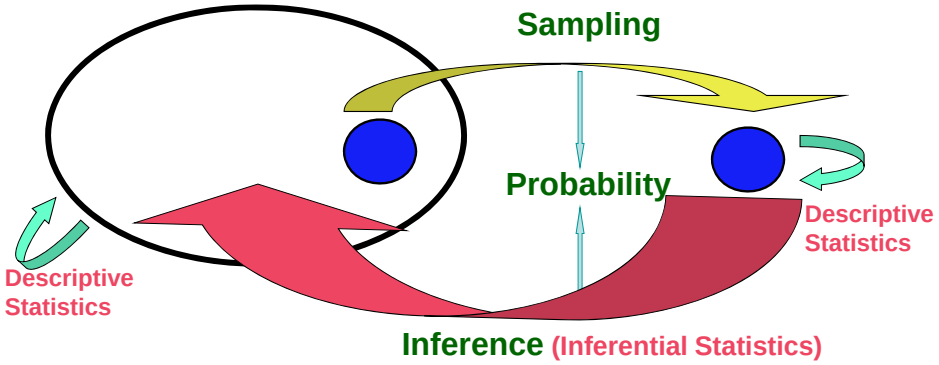
AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

"To address misconceptions and misuse of the p-value."

The Goal of Statistics



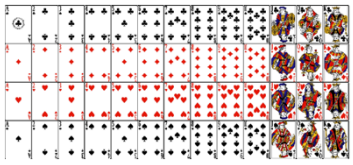
Population
[Parameters]
(μ, σ)

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

Sampling, Inference, & Probability

The probability question during sampling:

Given that the population parameters are known, what's the probability of getting a particular sample?



Parameters:
4 suits
13 ranks

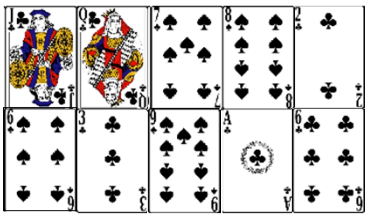
P(A full-house hand)= 3,744 / 2,598,960 = 0.00144.

Sampling, Inference, & Probability

The probability question during inference:

Given a particular sample at hand, what's the most likely value of the population parameter to have generated the sample?

To check whether the proportion of black cards in the deck is 50%.



Based on the information from the few cards, one makes a decision (inference) about the deck. Different sample sizes, say, 2, 5, 10, carry different amount of evidence.

Standard Research Process

Identifying a research question and a hypothesis

↓

Designing study and developing research protocol

↓

Gathering preliminary data and
revising the protocol

↓

Conducting the study

↓

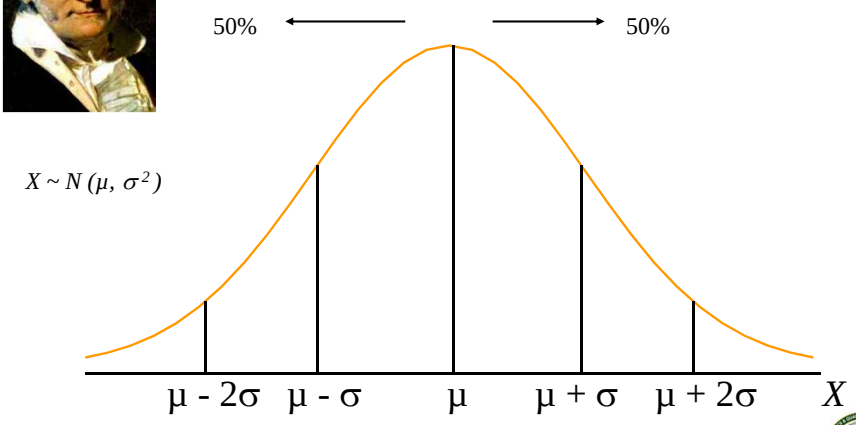
Analyzing data analysis and interpreting results

↓

Drawing conclusions and disseminating the results

The Normal Distribution

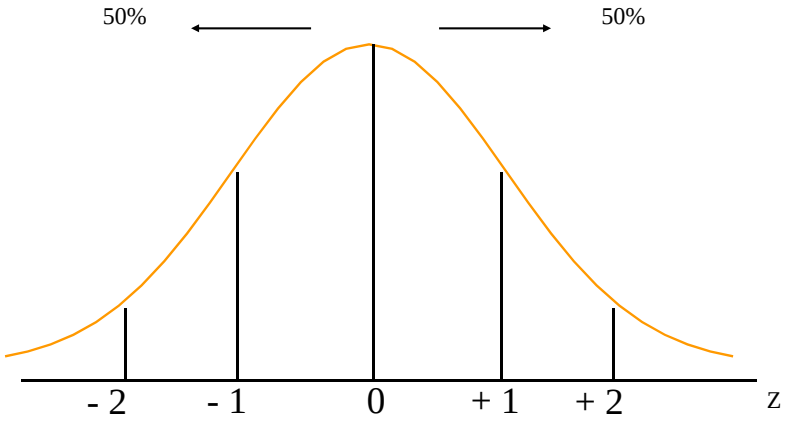
Carl Friedrich Gauss (1777-1855)



We usually are interested area under the curve (AUC).

Standard Normal Distribution

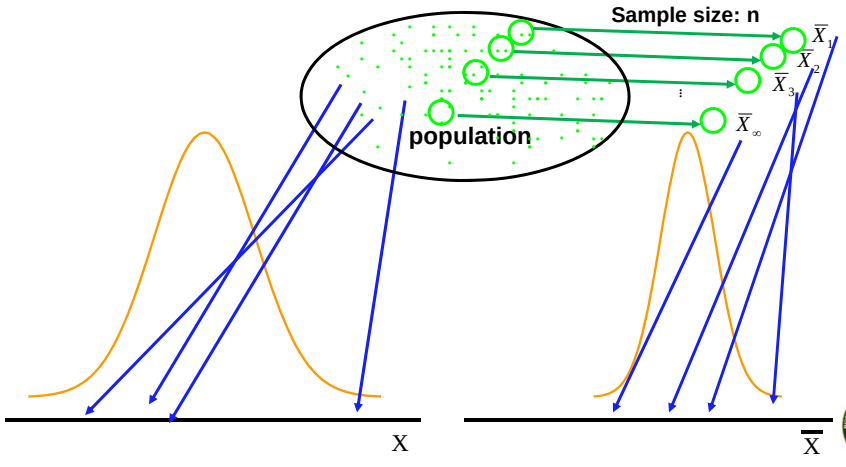
Standard normal distribution: $Z \sim N(\mu = 0, \sigma^2 = 1)$



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

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

$$\sigma_{\bar{X}} = s.e.(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

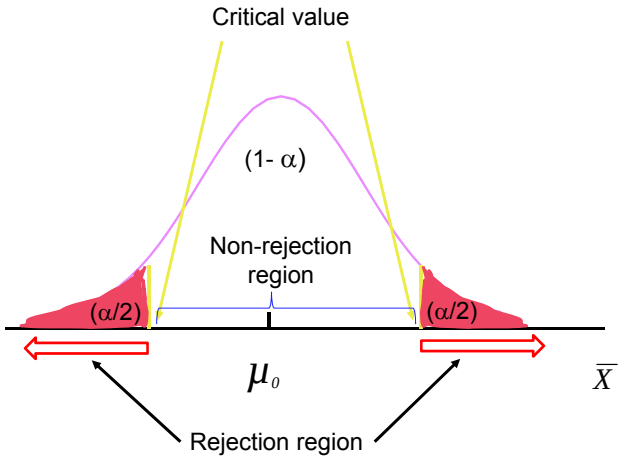
Hypothesis Testing

An Example:

Normal serum creatinine level may depend on the population being studied. From the literature a FM resident found that one well-established study showed an average sCr of 0.73 mg/dL (with a standard deviation of 0.15 mg/dL) among Caucasian women living on the east coast. But based on her knowledge and experience, she believed that the μ of sCr among Japanese women in Hawaii should different.

After discussing with her mentors and her biostatistician, she decided to assess this by measuring sCr of 49 local Japanese 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, and Statistical Decision

Type I Error (α) - False positives
- 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)

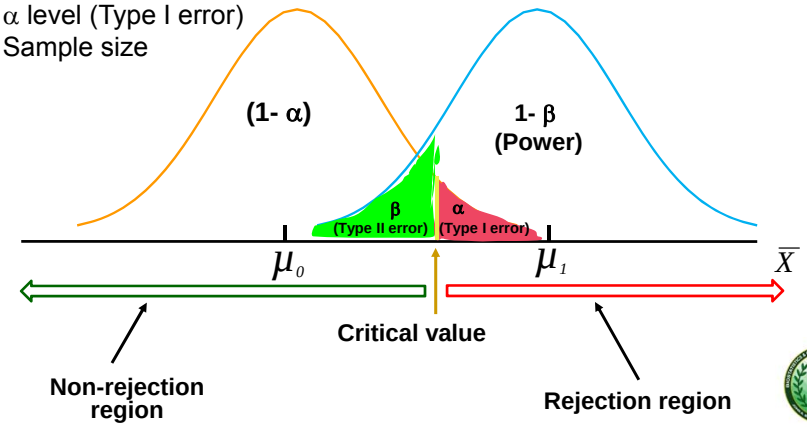
Decision	Truth: H_0 True	Truth: H_0 False
Reject H_0	α	1- β
Not reject H_0	1- α	β

Statistical Design & Decision

General Design Factors:

1. Effect difference
2. Variability
3. Statistical power (1- β)
4. α level (Type I error)
5. 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"!!!

Hypothesis Testing (cont.)

Example (cont.): Say, the average sCr of the sample of 49 locals is 0.68 mg/dL and the population standard deviation is 0.15 mg/dL (based on the literature).

Step 1. State H_0: and H_1:

$H_0 : \mu_{sCr} = 0.73$ vs. $H_1 : \mu_{sCr} \neq 0.73$

Step 2. Choose a significant level, say, $\alpha = 0.05$.

Step 3. Calculate the test statistic:

$$Z = \frac{\bar{X} - \mu_{sCr}}{\sigma / \sqrt{n}} = \frac{0.68 - 0.73}{0.15 / \sqrt{49}} = -2.33.$$

Hypothesis Testing (cont.)

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

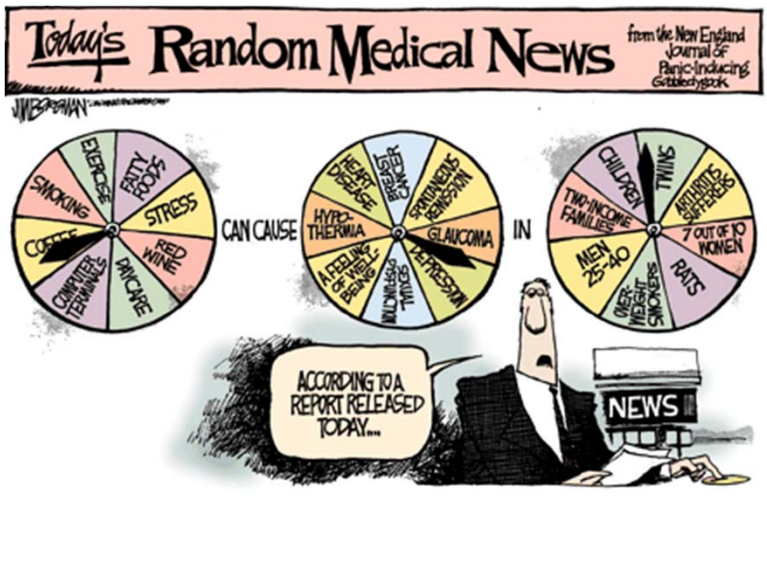
The critical value: $\pm$ *1.96.*

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

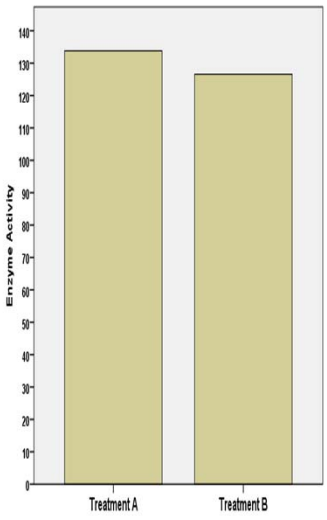
The non-rejection region: $|Z| < 1.96$.

Step 5. Make a decision, based on the sample, and state the conclusion: As the test statistic $|Z| = 2.33 > 1.96$, it is within the rejection region. Therefore, we reject the null hypothesis. We conclude that there is statistical evidence that the average sCr among local Japanese women is different from 0.73 mg/dL, with a p-value=0.02.

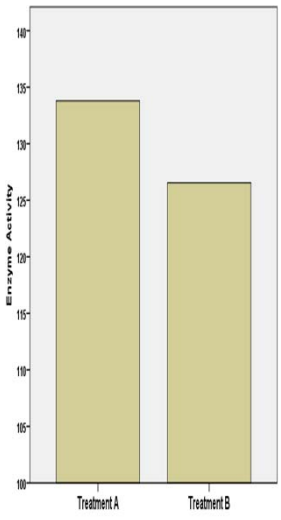
Medical Research, Media, and Public Health



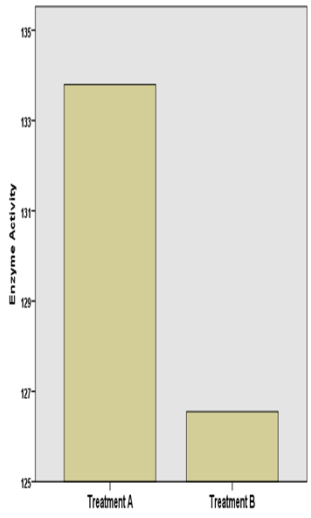
Enzyme Activity Studies



Study #1



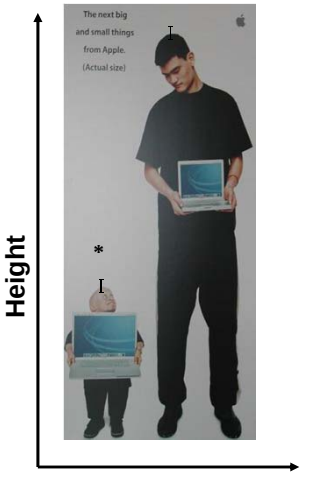
Study #2



Study #3

Yao and Mini-Me: The Height Study

Appropriate Level of Replicates:
Technical vs Biological



Mini-me Yao
(Each measured 5 times. Error bar represents SEM. *: P<0.05)

Guinness & The Student's *t*-Test

William S. Gosset (1876 – 1937)

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

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$ with n -1 degrees of freedom.

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

Two-Sample Student's t-Tests

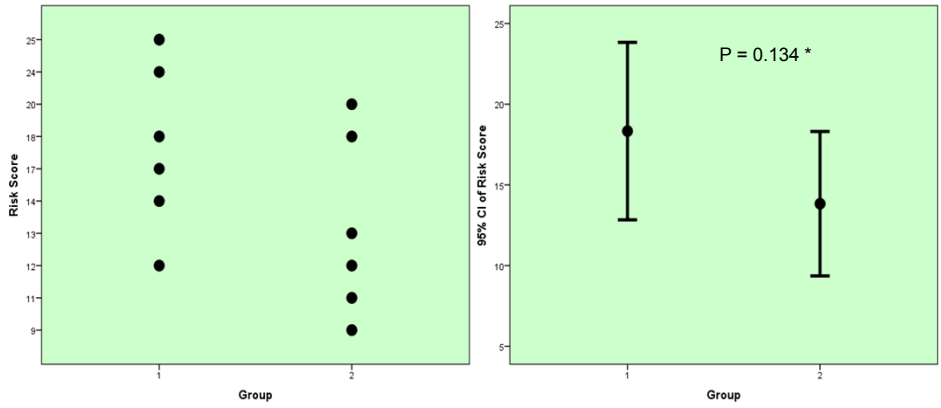
	Group	Risk_Score	Num
1	1	17	1
2	1	25	2
3	1	18	3
4	1	24	4
5	1	14	5
6	1	12	6
7	2	11	1
8	2	18	2
9	2	13	3
10	2	20	4
11	2	12	5
12	2	9	6

Study of reduction of diabetes risk score before (group 1) and after (group 2) the intervention of six patients.

Group Statistics

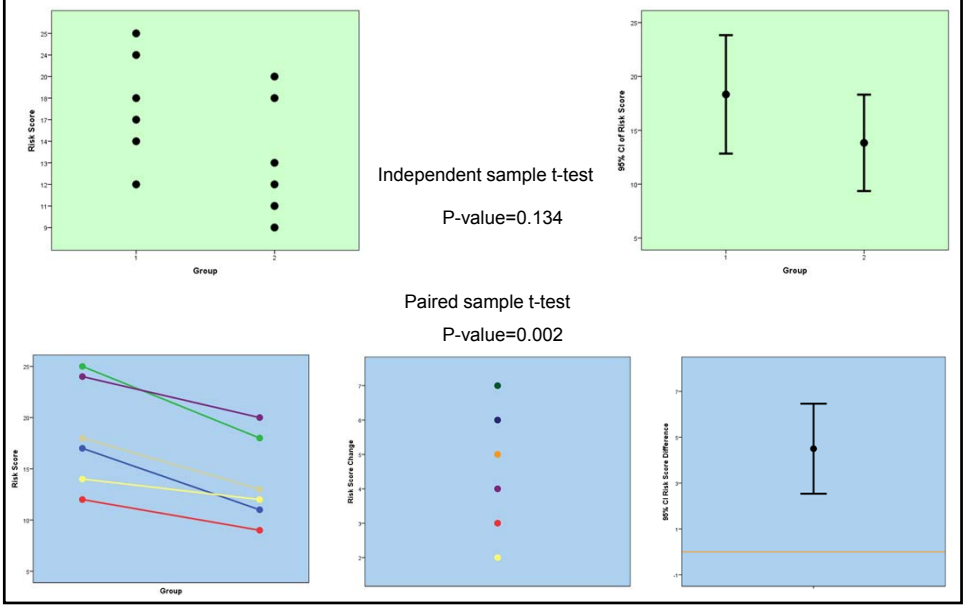
	Group	N	Mean	Std. Deviation	Std. Error Mean
Risk_Score	1	6	18.33	5.241	2.140
	2	6	13.83	4.262	1.740

Two-Sample Student t-tests



*: P-value was based on two independent sample t-test.

Two-Sample t-tests: Unpaired versus Paired



Sources of Multiple Testing

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)

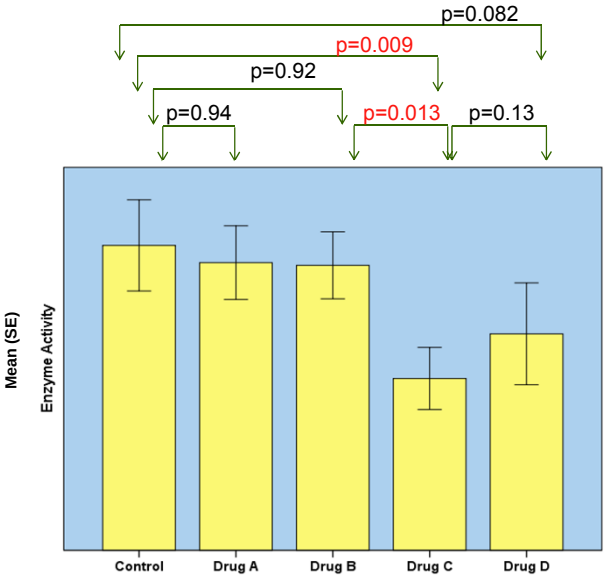
Consequences of Multiplicity

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

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

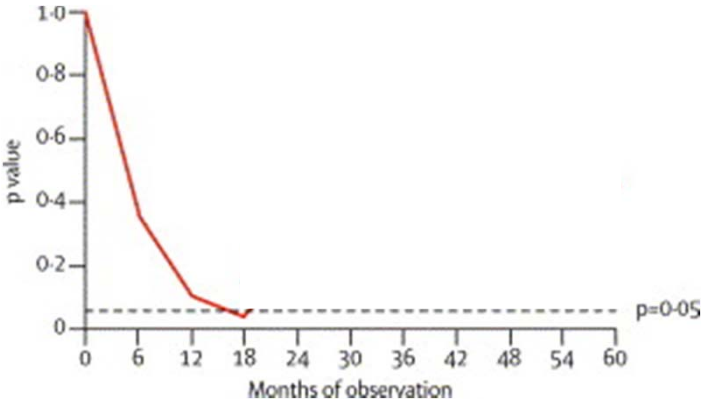
Multiple Treatments



With Bonferroni adjustment for multiple comparisons:

Control vs. A:	1.00
Control vs. B:	1.00
Control vs. C:	0.054
Control vs. D:	0.50
B vs. C:	0.078
C vs. D:	0.78

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

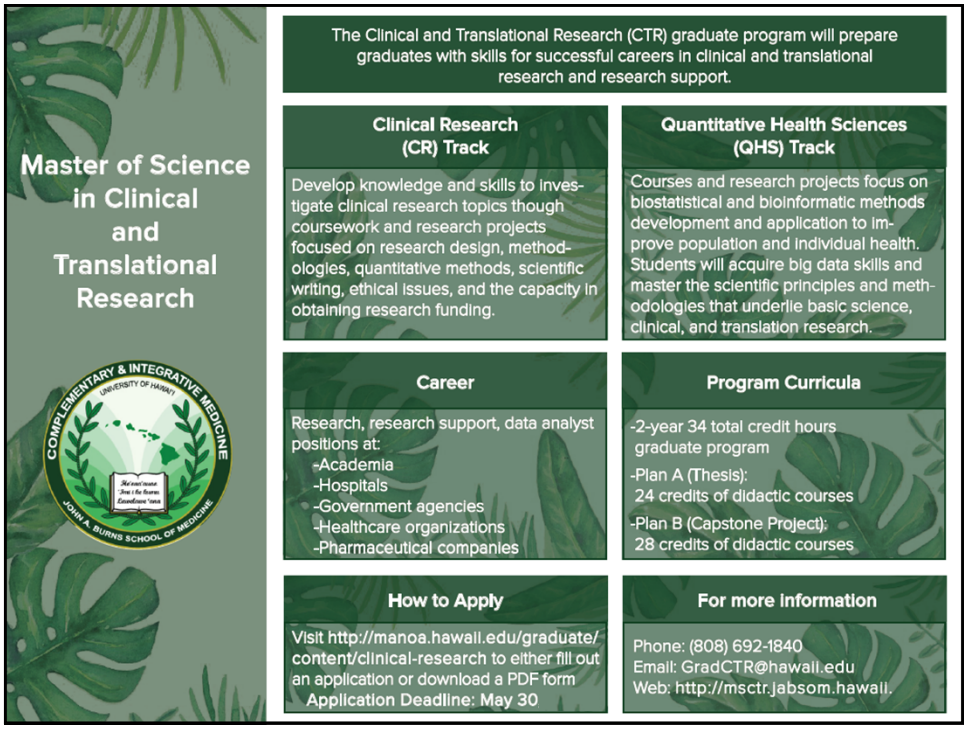
- All interim analyses should be planned in advance, including the pre-specified statistical stopping method
- Best be done by an independent data and safety monitoring committee (DSMC)
- Main goal: make sure overall probability of type I error is controlled
- You have to pay a price with an interim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

Simpson's Paradox

Kidney Stone Treatment (Charig et al. 1986. BMJ)

Treatment	Open Surgery	Percutaneous Nephrolithotomy
Success Rate	78% (273/350)	**83% (289/350)**

Stone size \ Treatment	Open Surgery	Percutaneous Nephrolithotomy
Small stones	**93% (81/87)**	87% (234/270)
Large stones	**73% (192/263)**	69% (55/80)

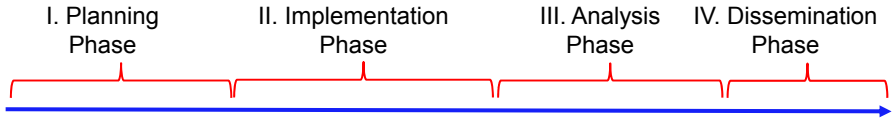


MSCTR Curriculum

- BIOM 640 Introduction to Clinical Research (3 credits)
- BIOM 641 Legal & Regulatory Issues and Bioethics (2 credits; cross-listed with CMB626)
- BIOM 644 Translational Research Methods (2 credits)
- BIOM 645 Clinical Protocol Development (3 credits)
- BIOM 654 Medical Genetics (2 credits)
- QHS 601 Biomedical Statistics I (3 credits; cross-listed with TRMD 655)
- QHS 602 Biomedical Statistics II (3 credits)
- QHS 610 Bioinformatics I (3 credits; cross-listed with TRMD 653)
- QHS 611 Bioinformatics II (3 credits)
- QHS 620 Introduction to Clinical Trials (2 credits)
- QHS 621 Design and Analysis of Clinical Trials (2 credits)
- QHS 650 Secondary Data Analysis (2 credits)
- QHS 651 Secondary Data Analysis Practicum (2 credits)
- QHS 675 Biostatistical Consulting (2 credits)
- QHS 676 Biostatistical Consulting Practicum (1 - 2 credits)

MSCTR Graduate Program Website: msctr.jabsom.hawaii.edu

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

Collaboration with A Biostatistician

1. Early and often
2. Start the discussion when you have the initial idea
3. It is an iterative process
4. A collaborative effort: equal and fair
5. Ask questions so you can discuss about the general statistical approach without the statistician

JABSOM Office of Biostatistics & QHS

http://biostat.jabsom.hawaii.edu

NIH U54MD007601

NIH U54MD007584