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

Biostatistics for Med Students

Lecture 1

John J. Chen, Ph.D.
Professor & Director of Biostatistics Core
UH JABSOM

JABSOM MD7

February 13, 2019

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

Lecture Objectives

- To understand basic research design principles and data presentation approaches
- To build a foundation which will facilitate the active participation in clinical research
- To fully grasp descriptive statistics
- To introduce key concepts of inferential statistics
- To survey some commonly used statistical approaches
- To be prepared for the USMLE Step 1 biostat/epi questions

Outline

Lecture 1 (02/13/2019)

- The goal of statistics
- Introduction to descriptive biostatistics
- Basic research design principles and data presentation approaches

Lecture 2 (02/20/2019)

- Introduction to inferential statistics
- Commonly used statistical approaches

Definition of Statistics

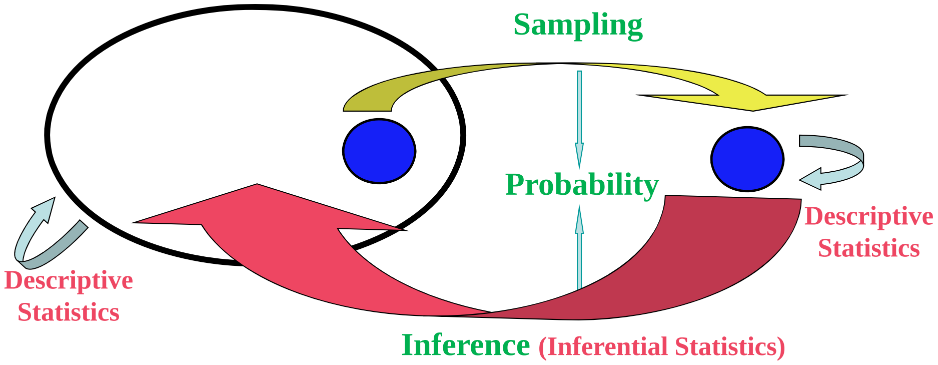
The theory and methodology for research (study) design, and for describing, analyzing, and interpreting information (data) generated from such studies, in which the data is subject to chance variation.

Population & Sample

- *Population*: the set of all subjects of interest having a common observable characteristic. For example, all newborns in US.
- *Sample*: a subset of a population, e.g., all newborns at KMC in 2018.

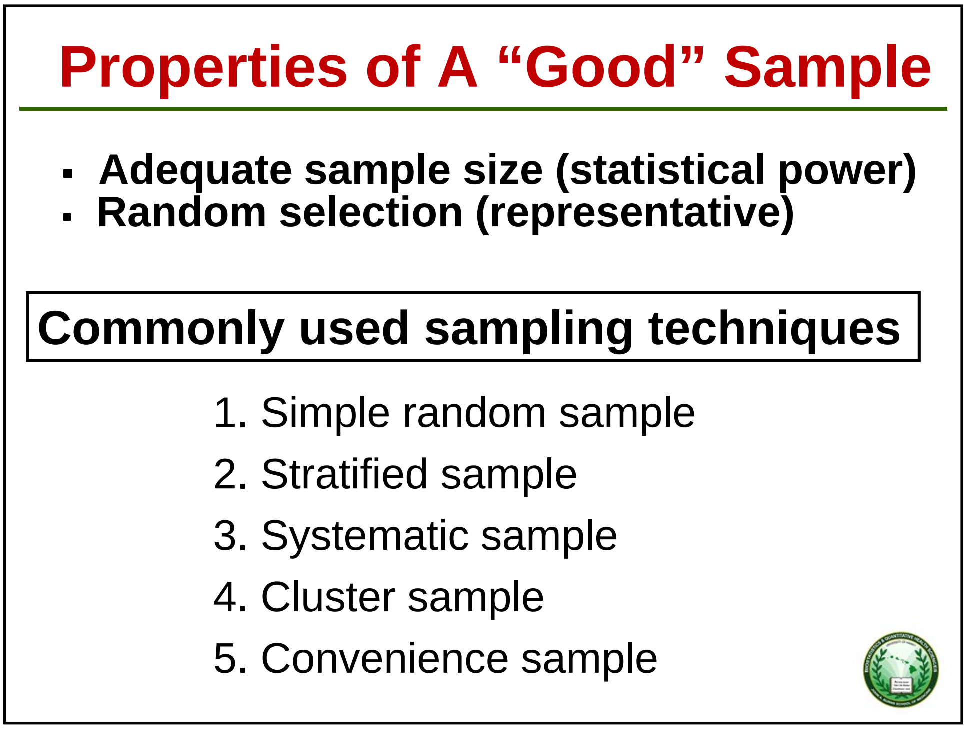
- *Parameter*: a summary measure of the population, e.g., the average birth weight of the above population.
- *Statistic*: a summary measure of the sample, e.g., the average birth weight of the above sample.

The Goal of Statistics



Population
[Parameters]
(μ, σ)

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

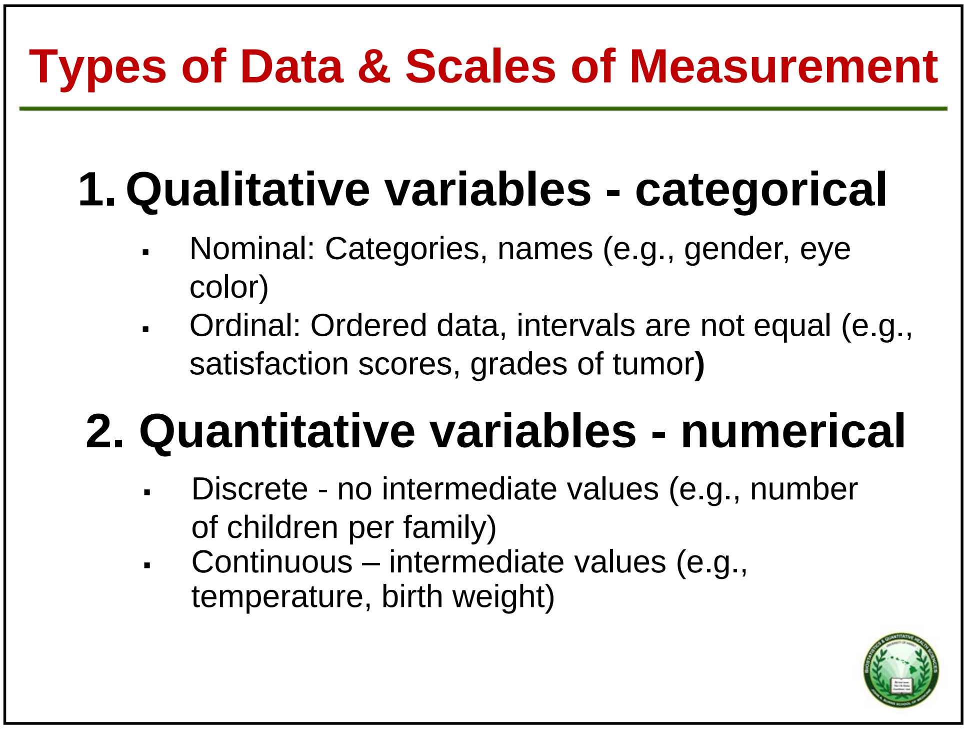


Properties of A "Good" Sample

- **Adequate sample size (statistical power)**
- **Random selection (representative)**

Commonly used sampling techniques

1. Simple random sample
2. Stratified sample
3. Systematic sample
4. Cluster sample
5. Convenience sample



Types of Data & Scales of Measurement

1. Qualitative variables - categorical

- Nominal: Categories, names (e.g., gender, eye color)
- Ordinal: Ordered data, intervals are not equal (e.g., satisfaction scores, grades of tumor**)**

2. Quantitative variables - numerical

- Discrete - no intermediate values (e.g., number of children per family)
- Continuous – intermediate values (e.g., temperature, birth weight)

Types of Variables

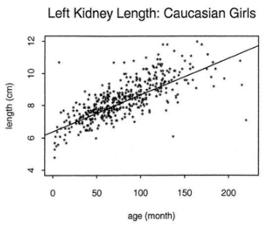
Notes:

Dependent (response) versus

Independent (explanatory) variables

In linear regression analysis:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$



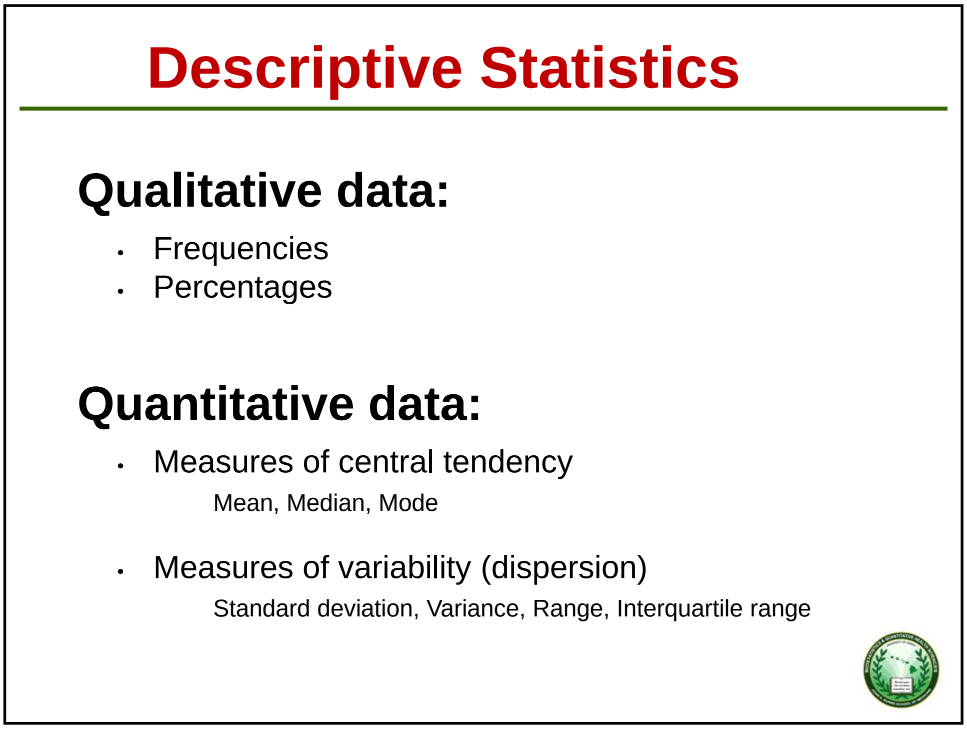
Sources of Data (Types of Studies)

Two major types of investigations:
Surveys versus experiments

Major difference: whether the investigator has control over which subjects enter each study group.

Some examples of survey researches
Prospective (cohort) studies
Retrospective (case-control) studies
Cross-sectional studies

Some examples of experimental studies:
Lab experiments
Clinical trials



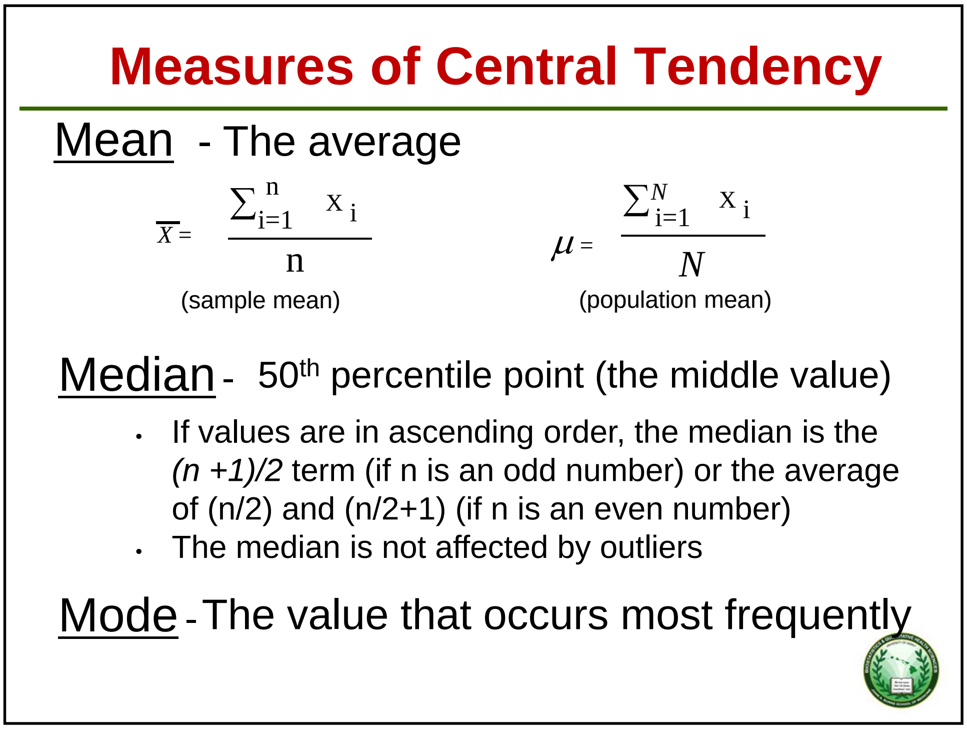
Descriptive Statistics

Qualitative data:

- Frequencies
- Percentages

Quantitative data:

- Measures of central tendency
  Mean, Median, Mode
- Measures of variability (dispersion)
  Standard deviation, Variance, Range, Interquartile range



Measures of Central Tendency

Mean - The average

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

(sample mean)

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

(population mean)

Median - 50th percentile point (the middle value)

- If values are in ascending order, the median is the *(n +1)/2* term (if n is an odd number) or the average of (n/2) and (n/2+1) (if n is an even number)
- The median is not affected by outliers

Mode - The value that occurs most frequently

Measures of Variability

1. Variance:

$$\text{Sample variance} = s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1}$$

2. Standard deviation (SD):

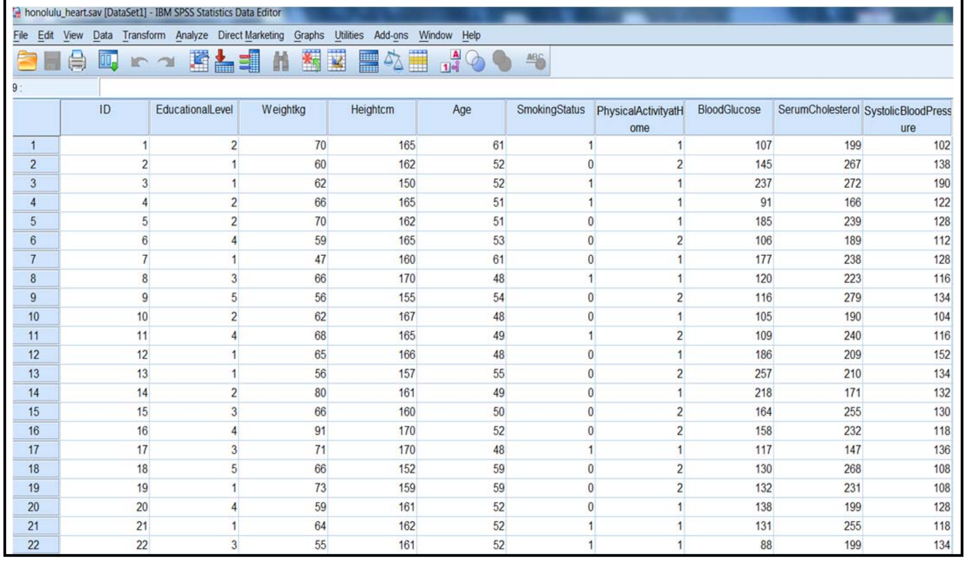
$$\text{Sample SD} = s = \sqrt{s^2}$$

3. Range:

$$\text{Range} = \max - \min$$

Ways of Presenting Data

SPSS: Honolulu Heart Study (partial data)



	ID	EducationalLevel	Weightkg	Heightcm	Age	SmokingStatus	PhysicalActivityatHome	BloodGlucose	SerumCholesterol	SystolicBloodPressure
1	1	2	70	165	61	1	1	107	199	102
2	2	1	60	162	52	0	2	145	267	138
3	3	1	62	150	52	1	1	237	272	190
4	4	2	66	165	51	1	1	91	166	122
5	5	2	70	162	51	0	1	185	239	128
6	6	4	59	165	53	0	2	106	189	112
7	7	1	47	160	61	0	1	177	238	128
8	8	3	66	170	48	1	1	120	223	116
9	9	5	56	155	54	0	2	116	279	134
10	10	2	62	167	48	0	1	105	190	104
11	11	4	68	165	49	1	2	109	240	116
12	12	1	65	166	48	0	1	186	209	152
13	13	1	56	157	55	0	2	257	210	134
14	14	2	80	161	49	0	1	218	171	132
15	15	3	66	160	50	0	2	164	255	130
16	16	4	91	170	52	0	2	158	232	118
17	17	3	71	170	48	1	1	117	147	136
18	18	5	66	152	59	0	2	130	268	108
19	19	1	73	159	59	0	2	132	231	108
20	20	4	59	161	52	0	1	138	199	128
21	21	1	64	162	52	1	1	131	255	118
22	22	3	55	161	52	1	1	88	199	134

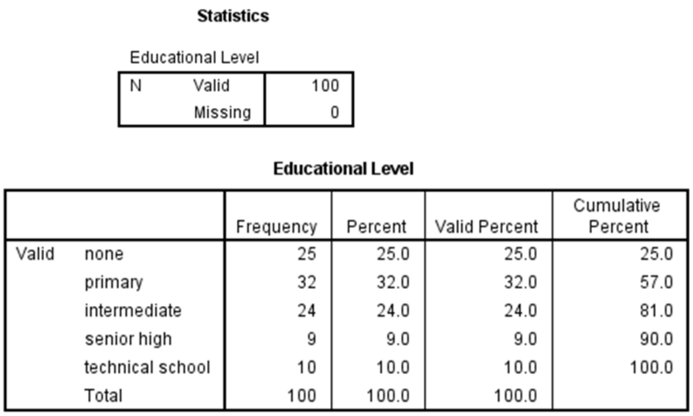
Data Dictionary

An example:

Variable	Education
Description/Label	Education Level
Data Type	Num – Categorical variable
Length	8
Allowable Values	1=none 2=primary 3=intermediate 4=senior high 5=technical school 6=university or above
Notes	Required field. No missing allowed.

Ways of Presenting Data (cont.)

Summary table: one categorical variable



Statistics

Educational Level

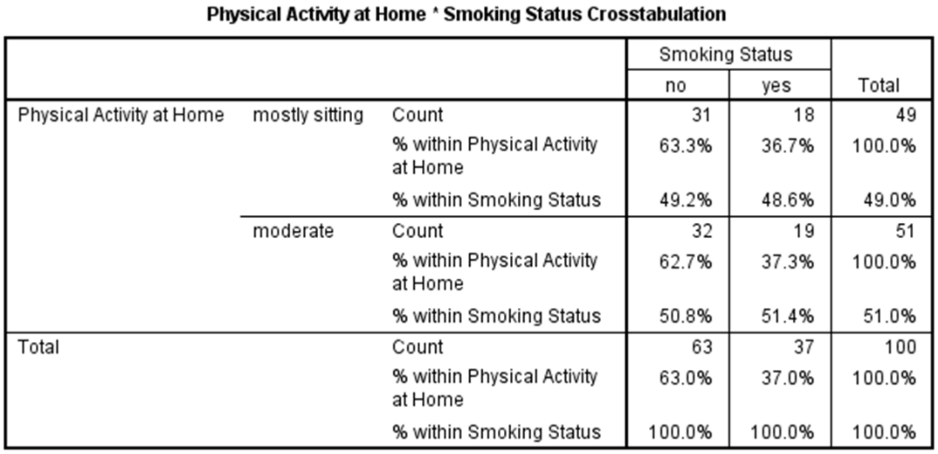
N	Valid	100
	Missing	0

Educational Level

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	none	25	25.0	25.0	25.0
	primary	32	32.0	32.0	57.0
	intermediate	24	24.0	24.0	81.0
	senior high	9	9.0	9.0	90.0
	technical school	10	10.0	10.0	100.0
	Total	100	100.0	100.0	

Ways of Presenting Data (cont.)

Cross-tabulation: two categorical variables

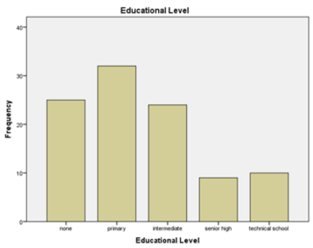


Physical Activity at Home * Smoking Status Crosstabulation

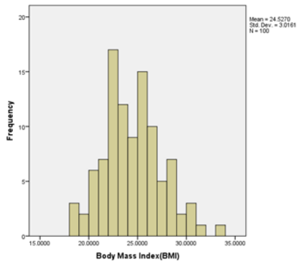
			Smoking Status		Total
			no	yes	
Physical Activity at Home	mostly sitting	Count	31	18	49
		% within Physical Activity at Home	63.3%	36.7%	100.0%
		% within Smoking Status	49.2%	48.6%	49.0%
	moderate	Count	32	19	51
		% within Physical Activity at Home	62.7%	37.3%	100.0%
		% within Smoking Status	50.8%	51.4%	51.0%
Total		Count	63	37	100
		% within Physical Activity at Home	63.0%	37.0%	100.0%
		% within Smoking Status	100.0%	100.0%	100.0%

Ways of Displaying Data

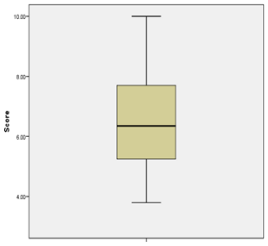
Bar chart



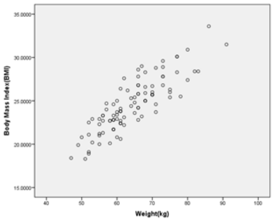
Histogram



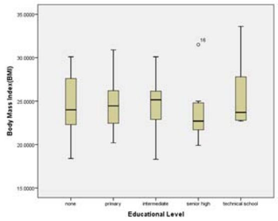
Box Plot



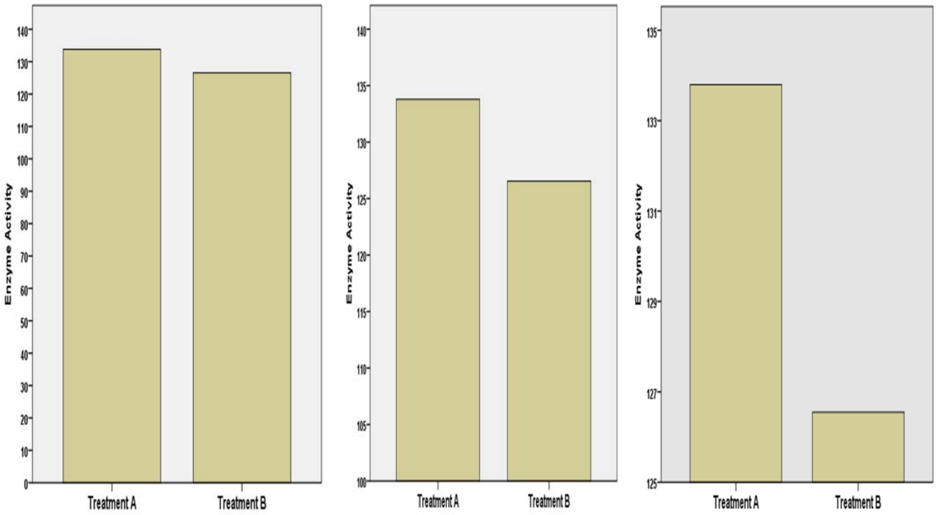
Scatterplot



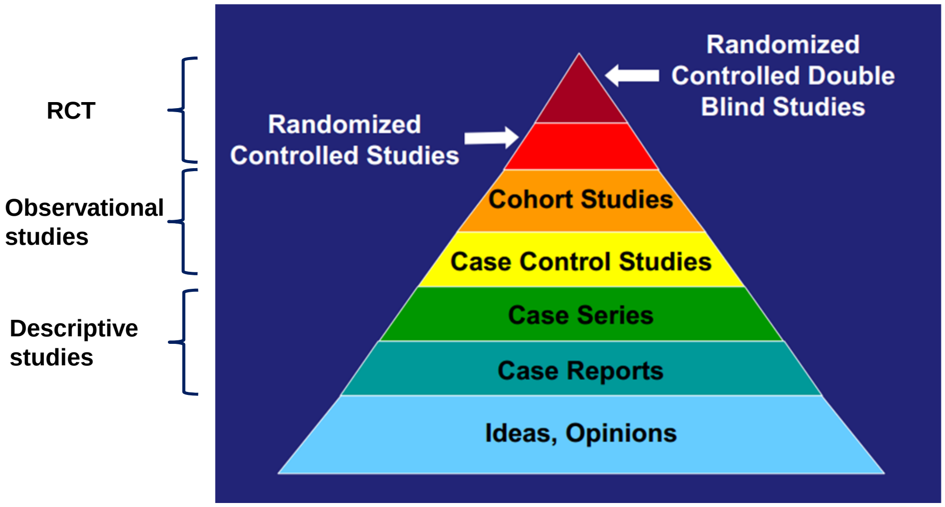
“Box and Bar” Plot



Different Scales



Clinical Research & Scientific Evidence


RCT

Observational studies

Descriptive studies

Randomized Controlled Double Blind Studies

Randomized Controlled Studies

Cohort Studies

Case Control Studies

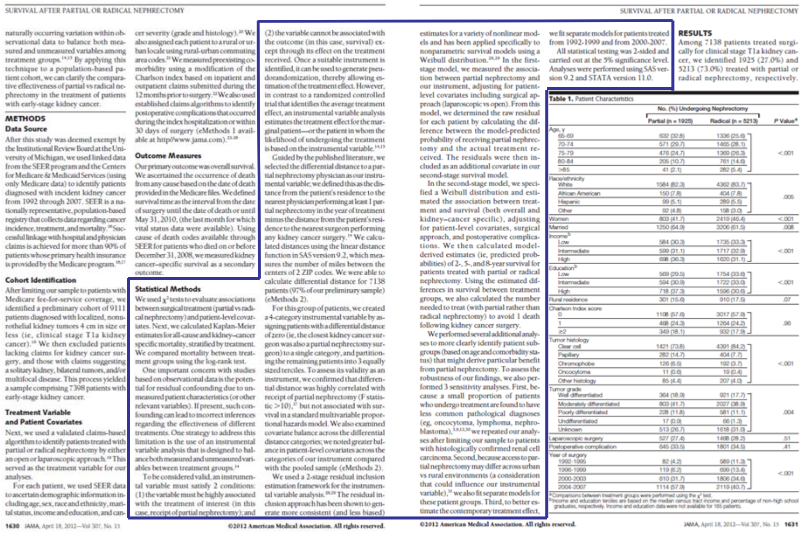
Case Series

Case Reports

Ideas, Opinions

Barnett Kramer (NIH)

Tan et al. Long-term Survival Following Partial vs Radical Nephrectomy Among Older Patients With Early-Stage Kidney Cancer. *JAMA* 2012; 307:1629-1635.



SURVIVAL AFTER PARTIAL OR RADICAL NEPHRECTOMY

naturally occurring variation within observational data to balance both measured and unmeasured variables among treatment groups.[14,15] By applying this technique to a population-based patient cohort, we can clarify the comparative effectiveness of partial vs radical nephrectomy in the treatment of patients with early-stage kidney cancer.

METHODS

Data Source

After this study was deemed exempt by the Institutional Review Board at the University of Michigan, we used linked data from the SEER program and the Centers for Medicare & Medicaid Services (using only Medicare data) to identify patients diagnosed with incident kidney cancer from 1992 through 2007. SEER is a nationally representative, population-based registry that collects data regarding cancer incidence, treatment, and mortality.[16] Successful linkage with hospital and physician claims is achieved for more than 90% of patients whose primary health insurance is provided by the Medicare program.[16,17]

Cohort Identification

After limiting our sample to patients with Medicare fee-for-service coverage, we identified a preliminary cohort of 9111 patients diagnosed with localized, nonurothelial kidney tumors 4 cm in size or less (ie, clinical stage T1a kidney cancer).[18] We then excluded patients lacking claims for kidney cancer surgery, and those with claims suggesting a solitary kidney, bilateral tumors, and/or multifocal disease. This process yielded a sample comprising 7398 patients with early-stage kidney cancer.

Treatment Variable and Patient Covariates

Next, we used a validated claims-based algorithm to identify patients treated with partial or radical nephrectomy by either an open or laparoscopic approach.[19] This served as the treatment variable for our analyses.

For each patient, we used SEER data to ascertain demographic information including age, sex, race and ethnicity, marital status, income and education, and cancer severity (grade and histology).[20] We also assigned each patient to a rural or urban locale using rural-urban commuting area codes.[21] We measured preexisting comorbidity using a modification of the Charlson index based on inpatient and outpatient claims submitted during the 12 months prior to surgery.[22] We also used established claims algorithms to identify postoperative complications that occurred during the index hospitalization or within 30 days of surgery (eMethods 1 available at http//www.jama.com).[23-26]

Outcome Measures

Our primary outcome was overall survival. We ascertained the occurrence of death from any cause based on the date of death provided in the Medicare files. We defined survival time as the interval from the date of surgery until the date of death or until May 31, 2010, (the last month for which vital status data were available). Using cause of death codes available through SEER for patients who died on or before December 31, 2008, we measured kidney cancer–specific survival as a secondary outcome.

Statistical Methods

We used χ^2 tests to evaluate associations between surgical treatment (partial vs radical nephrectomy) and patient-level covariates. Next, we calculated Kaplan-Meier estimates for all-cause and kidney–cancer specific mortality, stratified by treatment. We compared mortality between treatment groups using the log-rank test.

One important concern with studies based on observational data is the potential for residual confounding due to unmeasured patient characteristics (or other relevant variables). If present, such confounding can lead to incorrect inferences regarding the effectiveness of different treatments. One strategy to address this limitation is the use of an instrumental variable analysis that is designed to balance both measured and unmeasured variables between treatment groups.[14]

To be considered valid, an instrumental variable must satisfy 2 conditions: (1) the variable must be highly associated with the treatment of interest (in this case, receipt of partial nephrectomy); and (2) the variable cannot be associated with the outcome (in this case, survival) except through its effect on the treatment received. Once a suitable instrument is identified, it can be used to generate pseudorandomization, thereby allowing estimation of the treatment effect. However, in contrast to a randomized controlled trial that identifies the average treatment effect, an instrumental variable analysis estimates the treatment effect for the marginal patient—or the patient in whom the likelihood of undergoing the treatment is based on the instrumental variable.[14,15]

Guided by the published literature, we selected the differential distance to a partial nephrectomy physician as our instrumental variable; we defined this as the distance from the patient's residence to the nearest physician performing at least 1 partial nephrectomy in the year of treatment minus the distance from the patient's residence to the nearest surgeon performing any kidney cancer surgery.[19] We calculated distances using the linear distance function in SAS version 9.2, which measures the number of miles between the centers of 2 ZIP codes. We were able to calculate differential distance for 7138 patients (97% of our preliminary sample) (eMethods 2).

For this group of patients, we created a 4-category instrumental variable by assigning patients with a differential distance of zero (ie, the closest kidney cancer surgeon was also a partial nephrectomy surgeon) to a single category, and partitioning the remaining patients into 3 equally sized terciles. To assess its validity as an instrument, we confirmed that differential distance was highly correlated with receipt of partial nephrectomy (F statistic >10),[27] but not associated with survival in a standard multivariable proportional hazards model. We also examined covariate balance across the differential distance categories; we noted greater balance in patient-level covariates across the categories of our instrument compared with the pooled sample (eMethods 2).

We used a 2-stage residual inclusion estimation framework for the instrumental variable analysis.[28,29] The residual inclusion approach has been shown to generate more consistent (and less biased) estimates for a variety of nonlinear models and has been applied specifically to nonparametric survival models using a Weibull distribution.[28,29] In the first-stage model, we measured the association between partial nephrectomy and our instrument, adjusting for patient-level covariates including surgical approach (laparoscopic vs open). From this model, we determined the raw residual for each patient by calculating the difference between the model-predicted probability of receiving partial nephrectomy and the actual treatment received. The residuals were then included as an additional covariate in our second-stage survival model.

In the second-stage model, we specified a Weibull distribution and estimated the association between treatment and survival (both overall and kidney–cancer specific), adjusting for patient-level covariates, surgical approach, and postoperative complications. We then calculated model-derived estimates (ie, predicted probabilities) of 2-, 5-, and 8-year survival for patients treated with partial or radical nephrectomy. Using the estimated differences in survival between treatment groups, we also calculated the number needed to treat (with partial rather than radical nephrectomy) to avoid 1 death following kidney cancer surgery.

We performed several additional analyses to more clearly identify patient subgroups (based on age and comorbidity status) that might derive particular benefit from partial nephrectomy. To assess the robustness of our findings, we also performed 3 sensitivity analyses. First, because a small proportion of patients who undergo treatment are found to have less common pathological diagnoses (eg, oncocytoma, lymphoma, nephroblastoma),[3,8,12,30] we repeated our analyses after limiting our sample to patients with histologically confirmed renal cell carcinoma. Second, because access to partial nephrectomy may differ across urban vs rural environments (a consideration that could influence our instrumental variable),[31] we also fit separate models for these patient groups. Third, to better estimate the contemporary treatment effect, we fit separate models for patients treated from 1992-1999 and from 2000-2007.

All statistical testing was 2-sided and carried out at the 5% significance level. Analyses were performed using SAS version 9.2 and STATA version 11.0.

RESULTS

Among 7138 patients treated surgically for clinical stage T1a kidney cancer, we identified 1925 (27.0%) and 5213 (73.0%) treated with partial or radical nephrectomy, respectively.

Table 1. Patient Characteristics

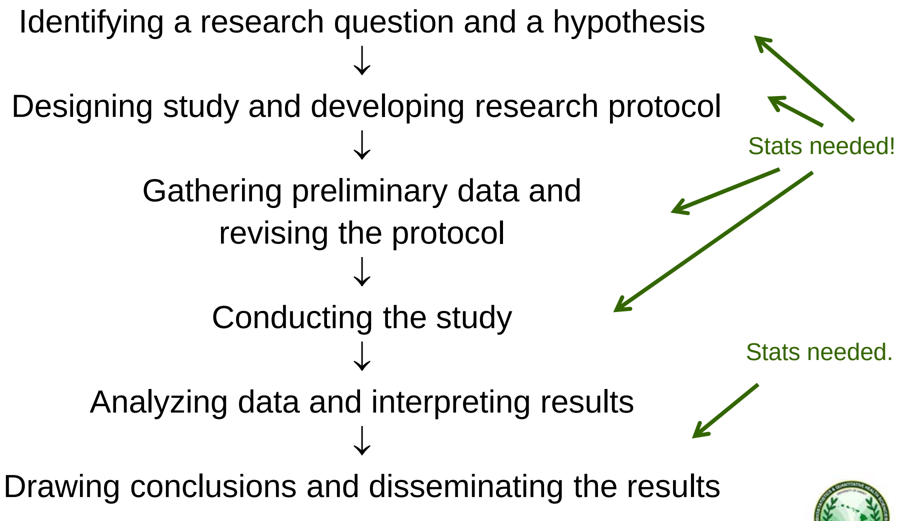
	No. (%) Undergoing Nephrectomy		
	Partial (n = 1925)	Radical (n = 5213)	*P* Value[a]
Age, y			
65-69	632 (32.8)	1336 (25.6)	<.001
70-74	571 (29.7)	1465 (28.1)	
75-79	476 (24.7)	1369 (26.3)	
80-84	205 (10.7)	761 (14.6)	
>85	41 (2.1)	282 (5.4)	
Race/ethnicity			
White	1584 (82.3)	4362 (83.7)	.005
African American	150 (7.8)	404 (7.8)	
Hispanic	99 (5.1)	289 (5.5)	
Other	92 (4.8)	158 (3.0)	
Women	803 (41.7)	2419 (46.4)	<.001
Married	1250 (64.9)	3206 (61.5)	.008
Income[b]			
Low	584 (30.3)	1735 (33.3)	<.001
Intermediate	599 (31.1)	1717 (32.9)	
High	698 (36.3)	1620 (31.1)	
Education[b]			
Low	569 (29.5)	1754 (33.6)	<.001
Intermediate	594 (30.9)	1722 (33.0)	
High	718 (37.3)	1596 (30.6)	
Rural residence	301 (15.6)	910 (17.5)	.07
Charlson Index score			
0	1108 (57.6)	3017 (57.9)	.96
1	468 (24.3)	1264 (24.2)	
≥2	349 (18.1)	932 (17.9)	
Tumor histology			
Clear cell	1421 (73.8)	4391 (84.2)	<.001
Papillary	282 (14.7)	404 (7.7)	
Chromophobe	126 (6.5)	192 (3.7)	
Oncocytoma	11 (0.6)	19 (0.4)	
Other histology	85 (4.4)	207 (4.0)	
Tumor grade			
Well differentiated	364 (18.9)	921 (17.7)	.004
Moderately differentiated	803 (41.7)	2027 (38.9)	
Poorly differentiated	228 (11.8)	581 (11.1)	
Undifferentiated	17 (0.9)	66 (1.3)	
Unknown	513 (26.7)	1618 (31.0)	
Laparoscopic surgery	527 (27.4)	1468 (28.2)	.51
Postoperative complication	645 (33.5)	1801 (34.5)	.41
Year of surgery			
1992-1995	82 (4.2)	589 (11.3)	<.001
1996-1999	119 (6.2)	699 (13.4)	
2000-2003	610 (31.7)	1806 (34.6)	
2004-2007	1114 (57.9)	2119 (40.7)	

[a] Comparisons between treatment groups were performed using the χ^2 test.
[b] Income and education terciles are based on the median census tract income and percentage of non-high school graduates, respectively. Income and education data were not available for 185 patients.

1630 JAMA, April 18, 2012—Vol 307, No. 15

JAMA, April 18, 2012—Vol 307, No. 15 1631

Biomedical 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 and interpreting results

↓

Drawing conclusions and disseminating the results

Stats needed!

Stats needed.

Basic Principles of Experimental Design

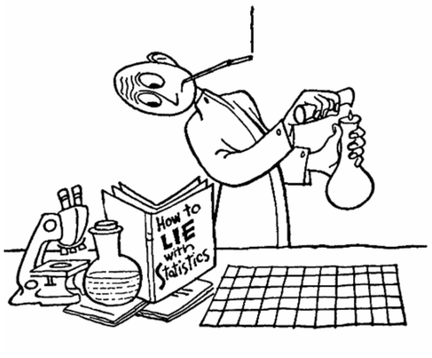
- Replications
- Randomization
- Blocking (stratification)
- Blinding
- Factorial experiments

Handling A Confounding Variable (Z)

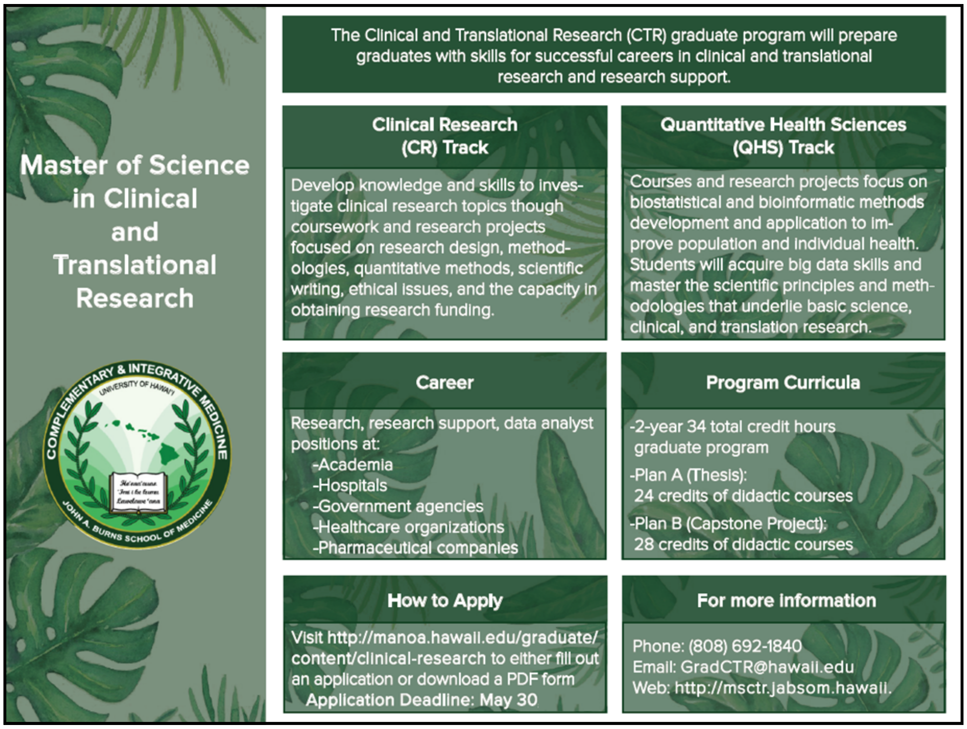
- If you can, fix a variable.
- If you can't, stratify it.
- If can't fix or stratify a variable, randomize it.

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

Warning Signs



Data generation **Data consumption**



Master of Science in Clinical and Translational Research

The Clinical and Translational Research (CTR) graduate program will prepare graduates with skills for successful careers in clinical and translational research and research support.

Clinical Research (CR) Track

Develop knowledge and skills to investigate clinical research topics though coursework and research projects focused on research design, methodologies, quantitative methods, scientific writing, ethical issues, and the capacity in obtaining research funding.

Quantitative Health Sciences (QHS) Track

Courses and research projects focus on biostatistical and bioinformatic methods development and application to improve population and individual health. Students will acquire big data skills and master the scientific principles and methodologies that underlie basic science, clinical, and translation research.

Career

Research, research support, data analyst positions at:

-Academia
-Hospitals
-Government agencies
-Healthcare organizations
-Pharmaceutical companies

Program Curricula

-2-year 34 total credit hours graduate program

-Plan A (Thesis):
24 credits of didactic courses

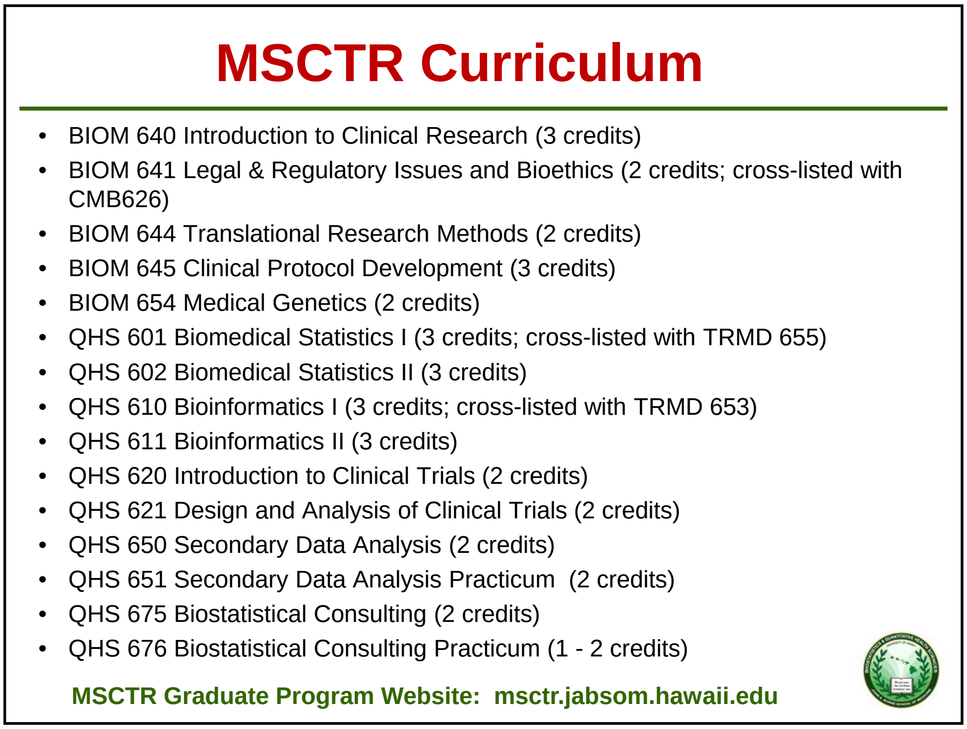
-Plan B (Capstone Project):
28 credits of didactic courses

How to Apply

Visit http://manoa.hawaii.edu/graduate/content/clinical-research to either fill out an application or download a PDF form
Application Deadline: May 30

For more information

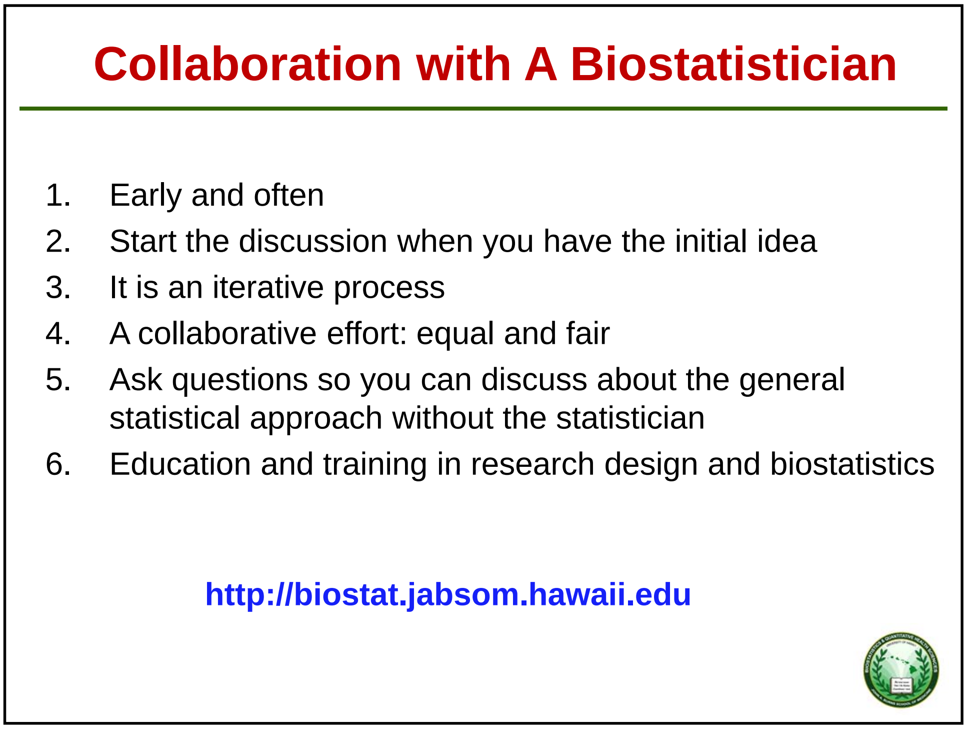
Phone: (808) 692-1840
Email: GradCTR@hawaii.edu
Web: http://msctr.jabsom.hawaii.



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)

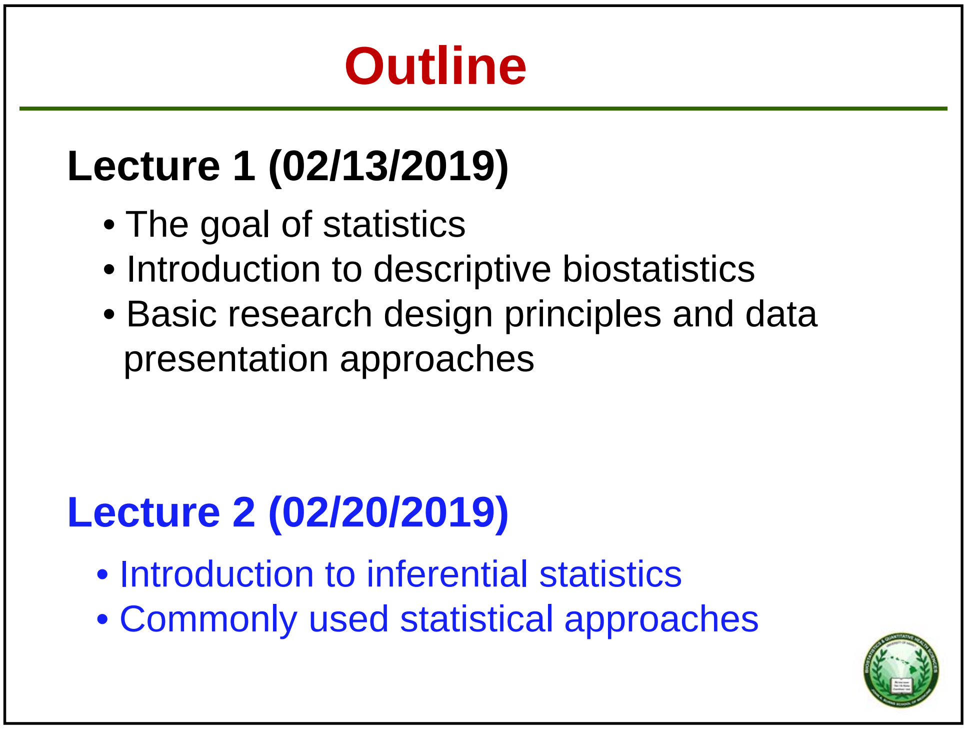
MSCTR Graduate Program Website: msctr.jabsom.hawaii.edu



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
6. Education and training in research design and biostatistics

http://biostat.jabsom.hawaii.edu



Outline

Lecture 1 (02/13/2019)

- The goal of statistics
- Introduction to descriptive biostatistics
- Basic research design principles and data presentation approaches

Lecture 2 (02/20/2019)

- Introduction to inferential statistics
- Commonly used statistical approaches