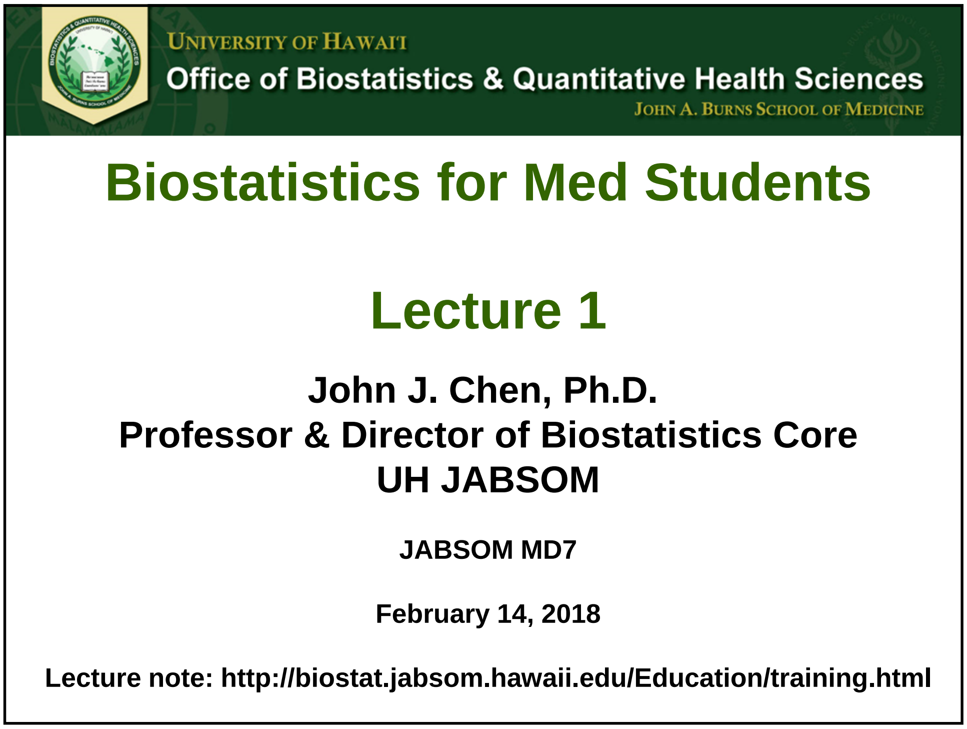


UNIVERSITY OF HAWAI'I

Office of Biostatistics & Quantitative Health Sciences

JOHN A. BURNS SCHOOL OF MEDICINE

Biostatistics for Med Students

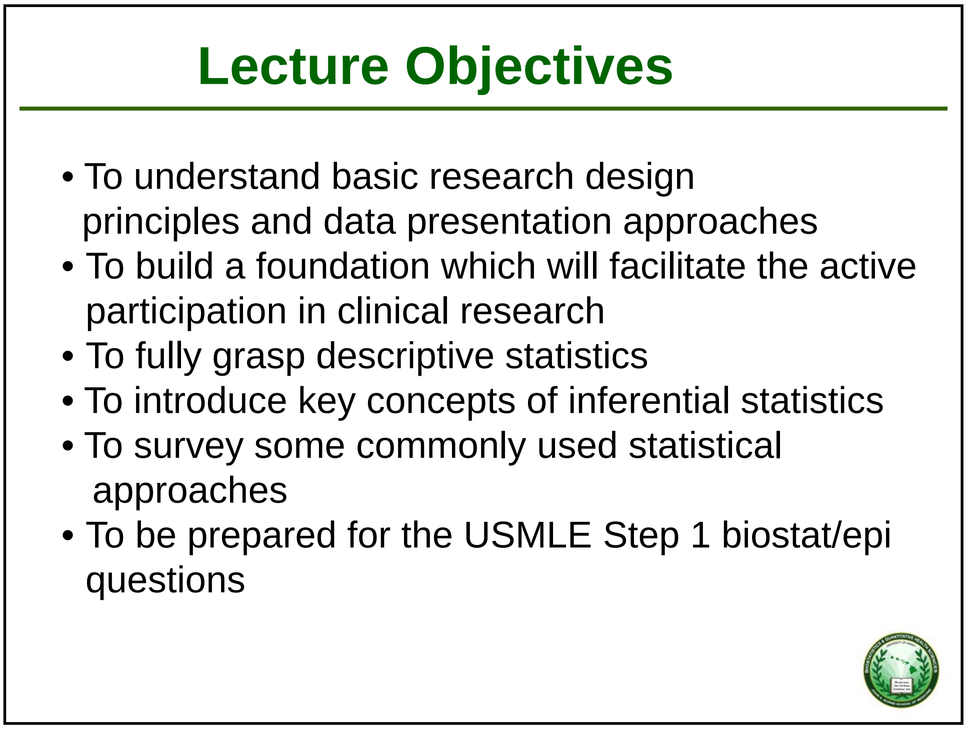
Lecture 1

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

JABSOM MD7

February 14, 2018

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

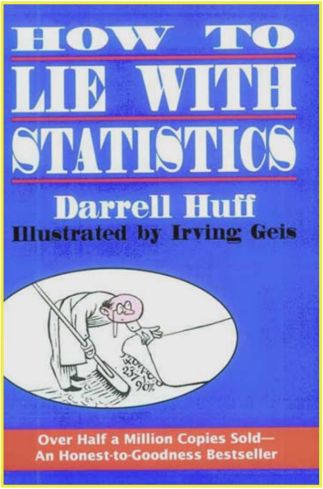
Lecture 1 (02/14/2018)

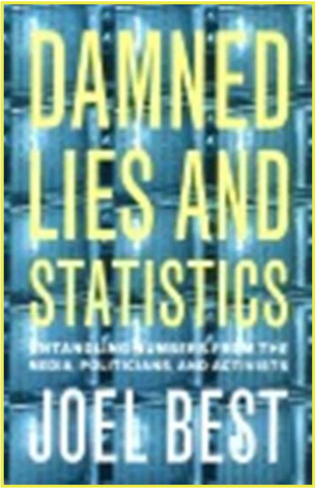
- The goal of statistics
- Introduction to descriptive biostatistics
- Basic research design principles and data presentation approaches

Lecture 2 (02/21/2018)

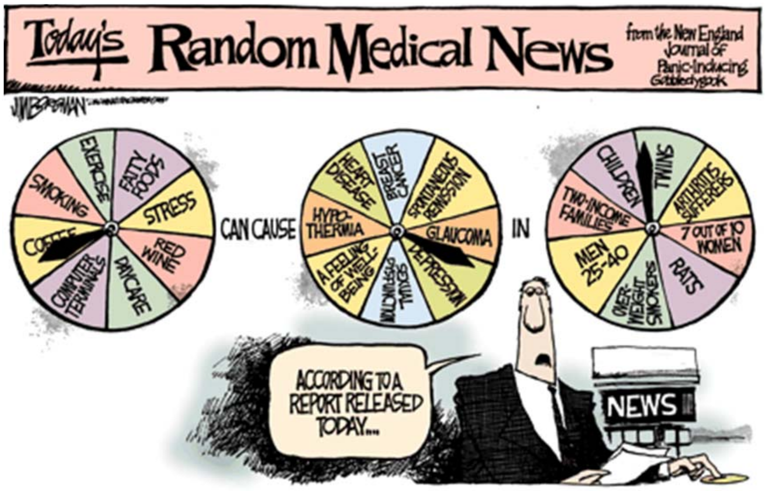
- Introduction to inferential statistics
- Commonly used statistical approaches

Lies, Damned Lies, And Statistics


HOW TO
LIE WITH
STATISTICS
Darrell Huff
Illustrated by Irving Geis
Over Half a Million Copies Sold—
An Honest-to-Goodness Bestseller


DAMNED
LIES AND
STATISTICS
UNTANGLING NUMBERS FROM THE
MEDIA, POLITICIANS, AND ACTIVISTS
JOEL BEST

Medical Research, Media, and Public Health



Definition of Statistics

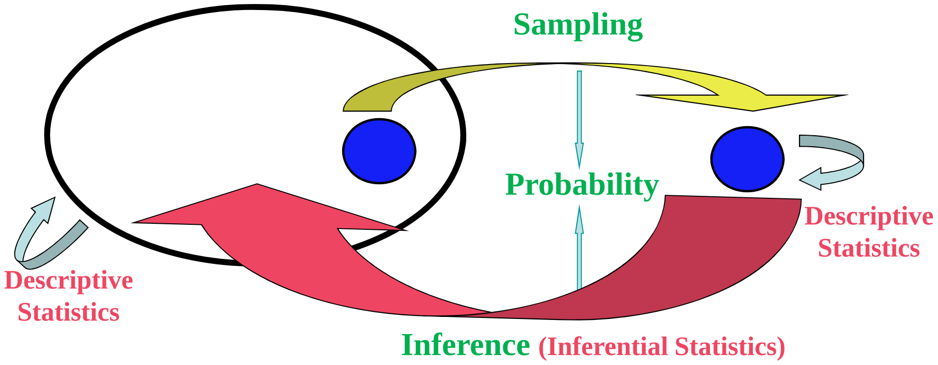
The theory and methodology for research (study) design, and for describing, analyzing, and interpreting information (data) generated from such studies, in which the data is subject to chance variation.

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 2016.

- *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



Sampling

Probability

Descriptive Statistics

Descriptive Statistics

Inference (Inferential Statistics)

Population

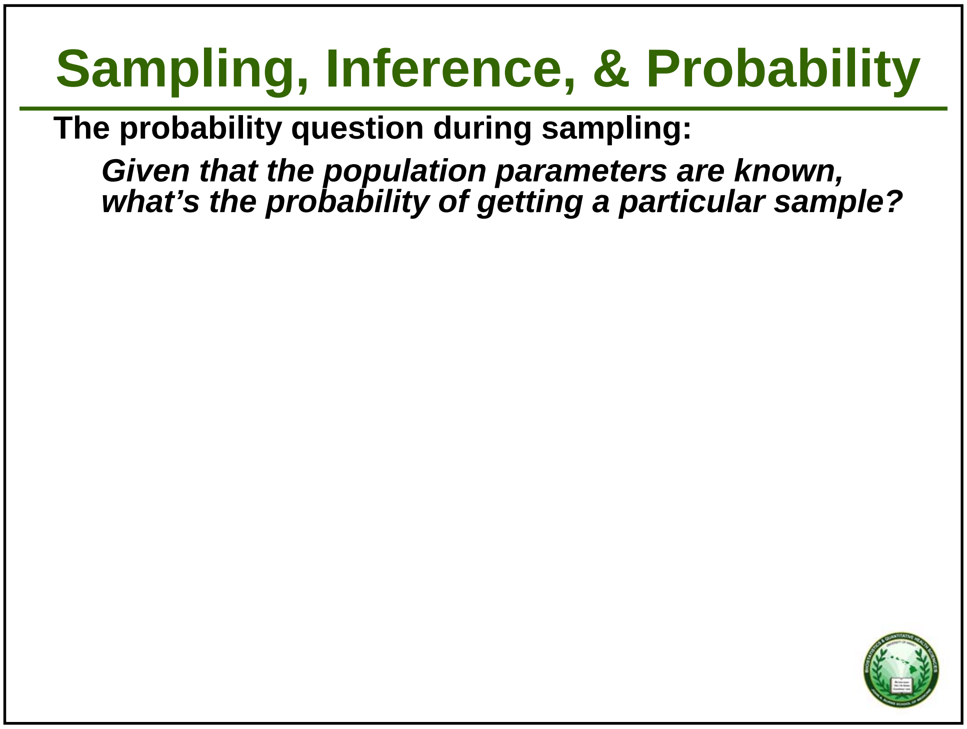
[Parameters]

(μ, σ)

Sample

[Statistics]

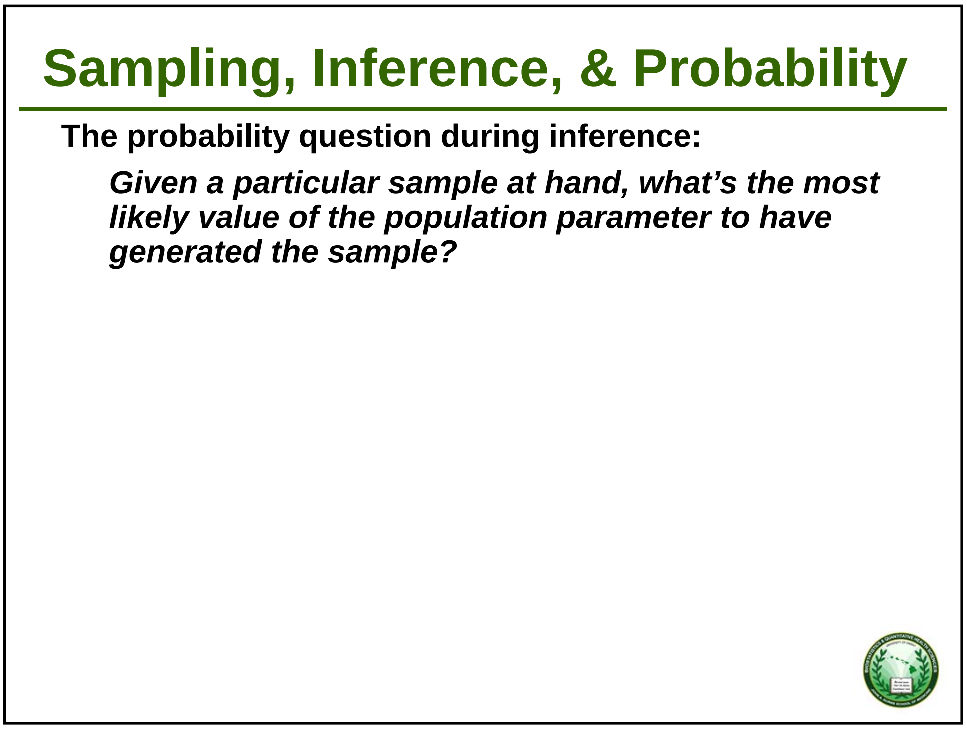
($\bar{X}$, S)



Sampling, Inference, & Probability

The probability question during sampling:

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



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?

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**

Data And Variables

- ***Variable*****:** a characteristic that may differ from one subject to another. For example, age, birth weight, etc.
- ***Data (information)***: the values of the observations recorded for the variables. For example,

Pt. ID	*Mother’s Smoking Status*	*Baby’s Birth Weight (grams)*
101	*None*	*3175*
102	*None*	*3232*
103	*1 pack/day*	*2750*
.	.	.
.	.	.
.	.	.
1001	*1+ pack/day*	*2466*

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

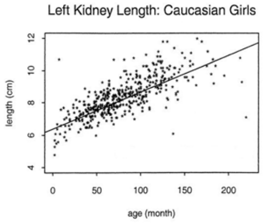
Notes:

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

<u>Major difference:</u> 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**

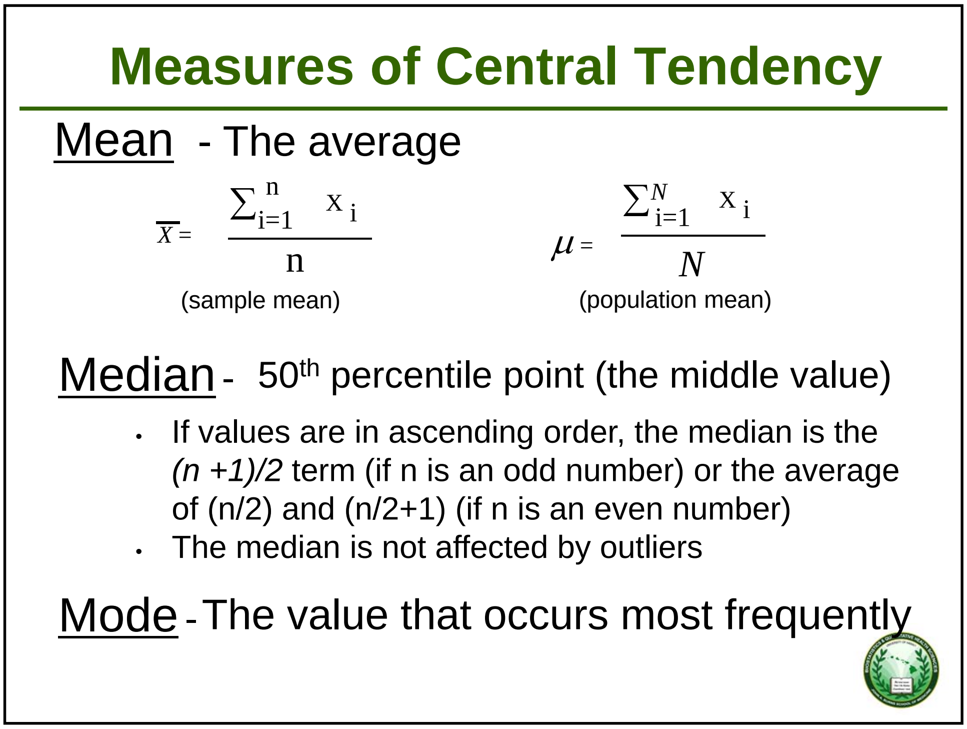
Quantitative data:

- **Measures of central tendency**

  Mean, Median, Mode

- **Measures of variability (dispersion)**

  Standard deviation, Variance, Range, Interquartile range



Measures of Central Tendency

Mean - The average

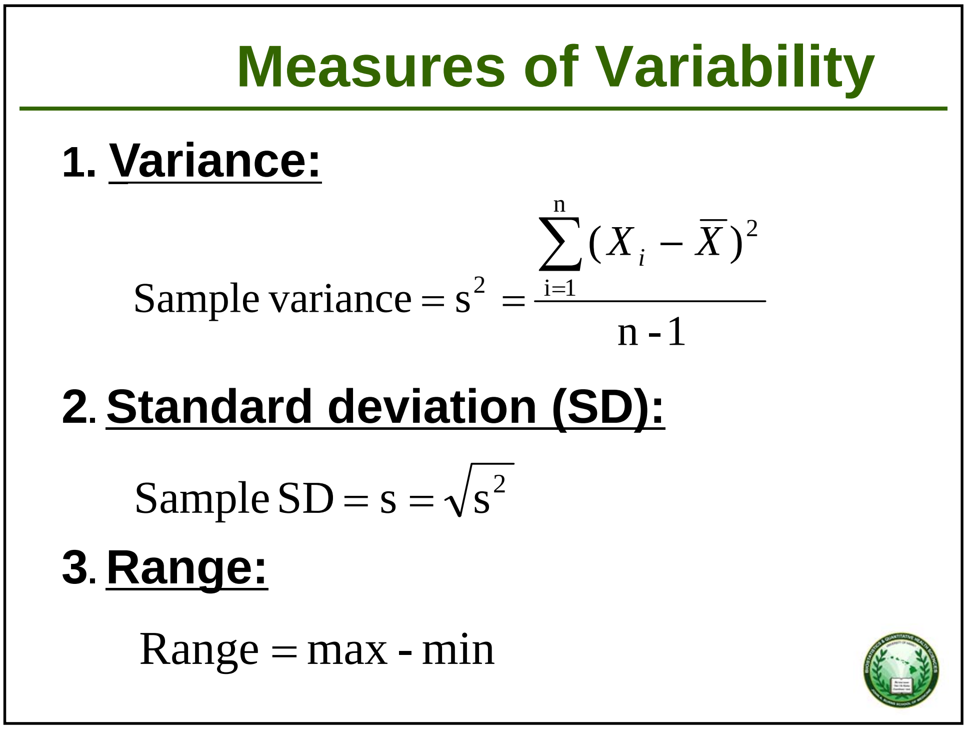
$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

(sample mean) (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):

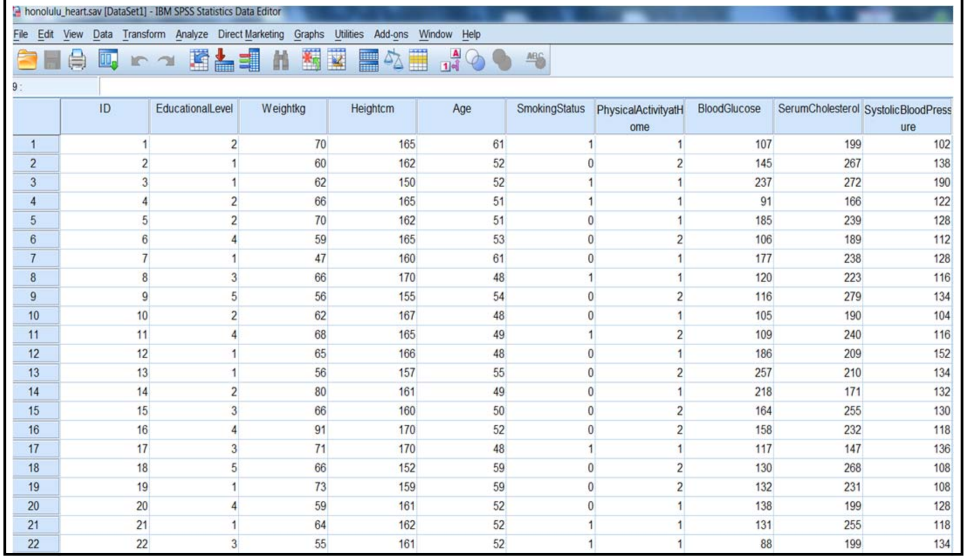
$$\text{Sample SD} = s = \sqrt{s^2}$$

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

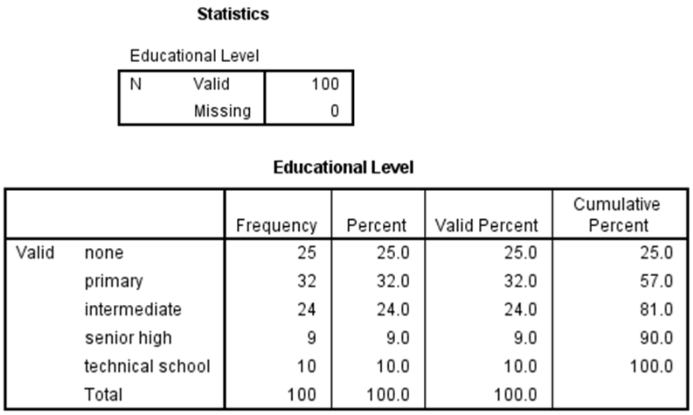
Data Dictionary

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

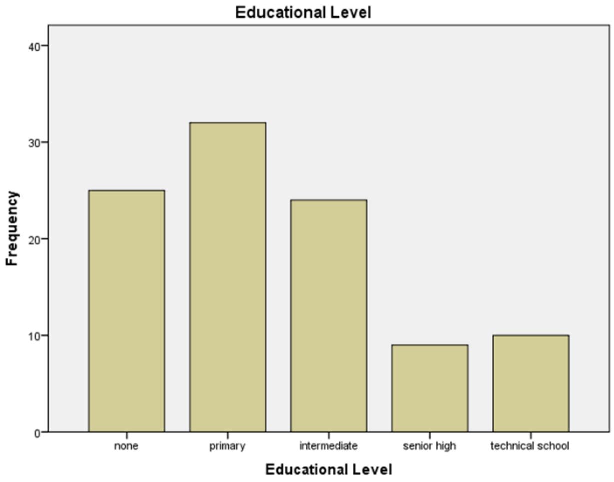
N	Valid	100
	Missing	0

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	

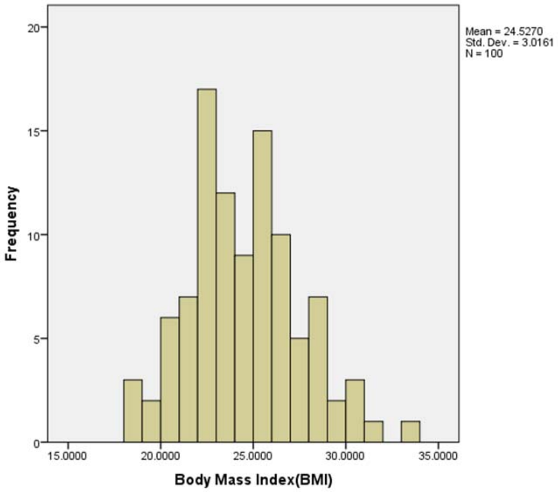
Ways of Presenting Data (cont.)

Bar chart: one categorical variable



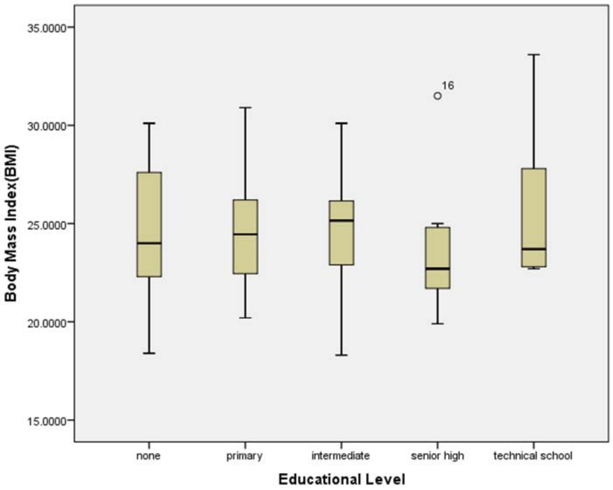
Ways of Presenting Data (cont.)

Histogram: one continuous variable



Ways of Presenting Data (cont.)

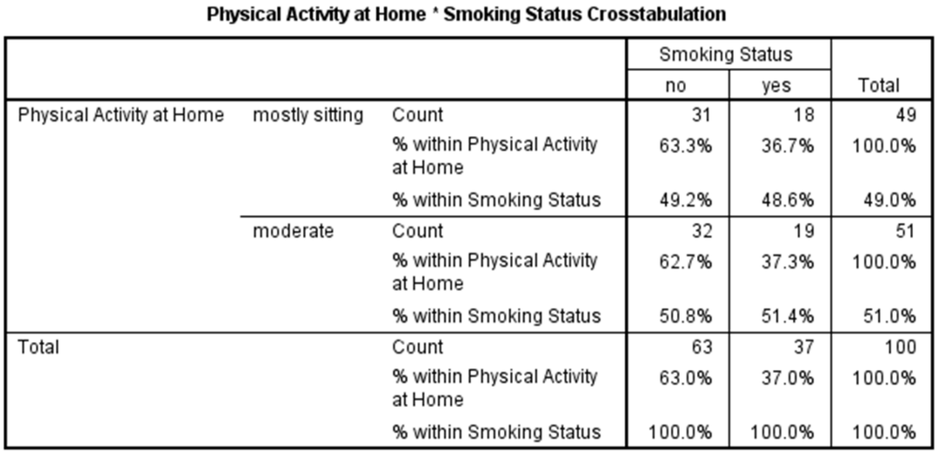
Box plot: one continuous variable (one categorical variable)



24

Ways of Presenting Data (cont.)

Cross-tabulation: two categorical variables

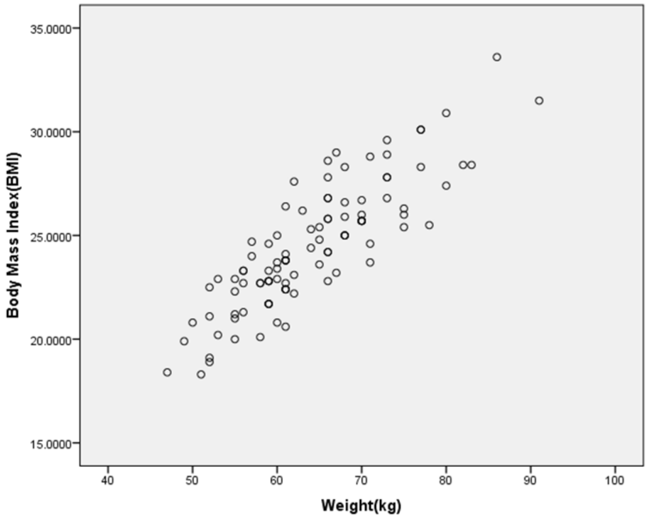


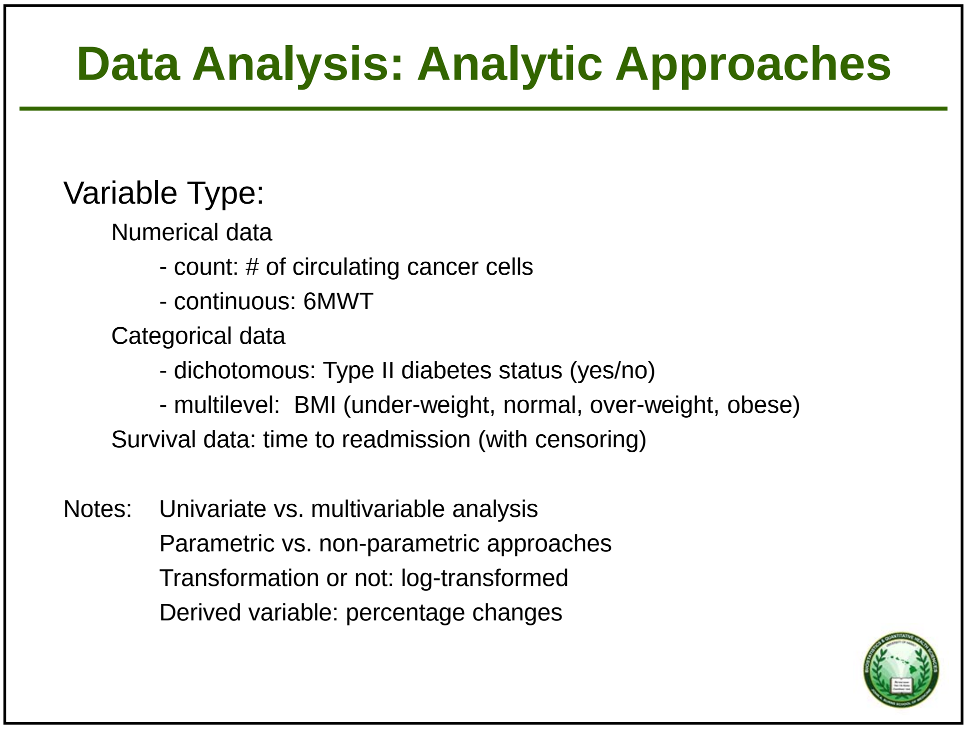
Physical Activity at Home * Smoking Status Crosstabulation

			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 Presenting Data (cont.)

Scatterplot: two continuous variables



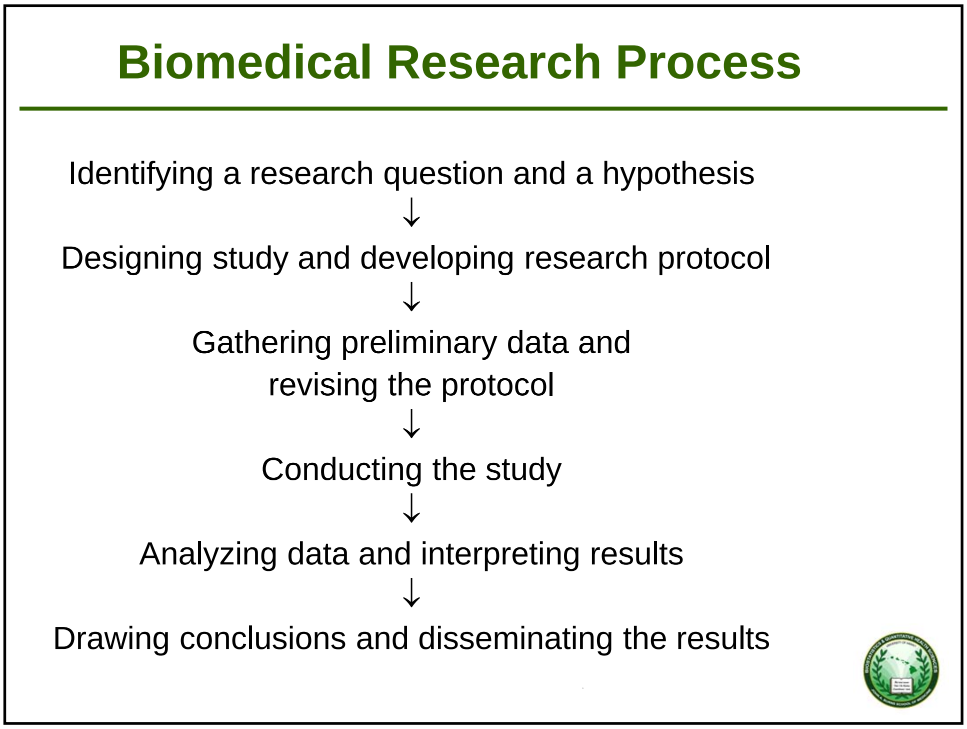


Data Analysis: Analytic Approaches

Variable Type:

- Numerical data
  - - count: # of circulating cancer cells
  - - continuous: 6MWT
- Categorical data
  - - dichotomous: Type II diabetes status (yes/no)
  - - multilevel: BMI (under-weight, normal, over-weight, obese)
- Survival data: time to readmission (with censoring)

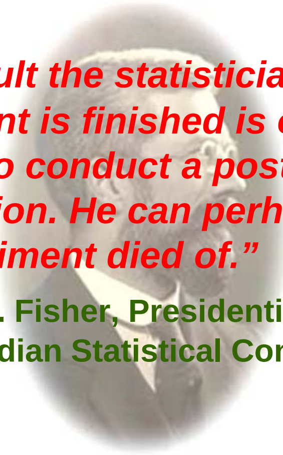
Notes:
Univariate vs. multivariable analysis
Parametric vs. non-parametric approaches
Transformation or not: log-transformed
Derived variable: percentage changes



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

The Importance of Research Design



“To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.”

Sir R.A. Fisher, Presidential Address to the First Indian Statistical Congress (1938)

Basic Principles of Experimental Design

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

Handling A Confounding Variable (Z)

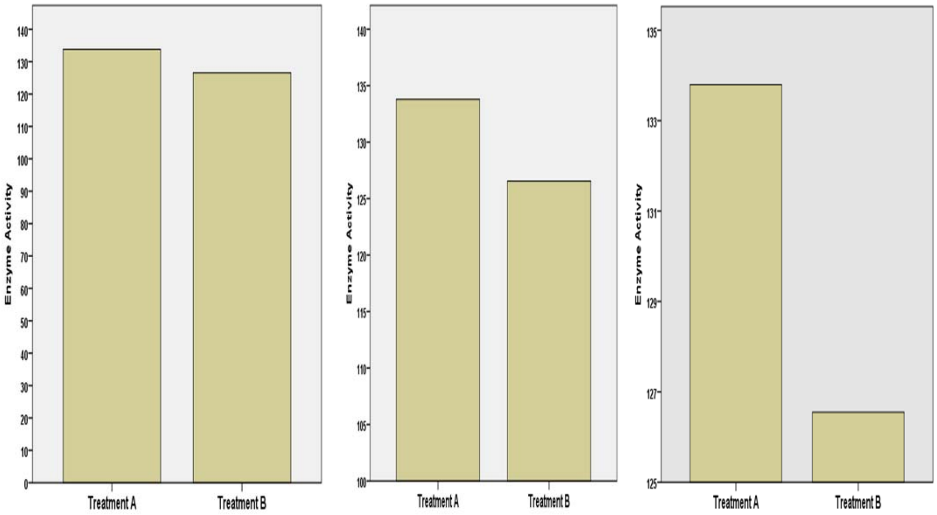
- If you can, fix a variable.
- If you can’t, stratify it.
- If can’t fix or stratify a variable, randomize it.

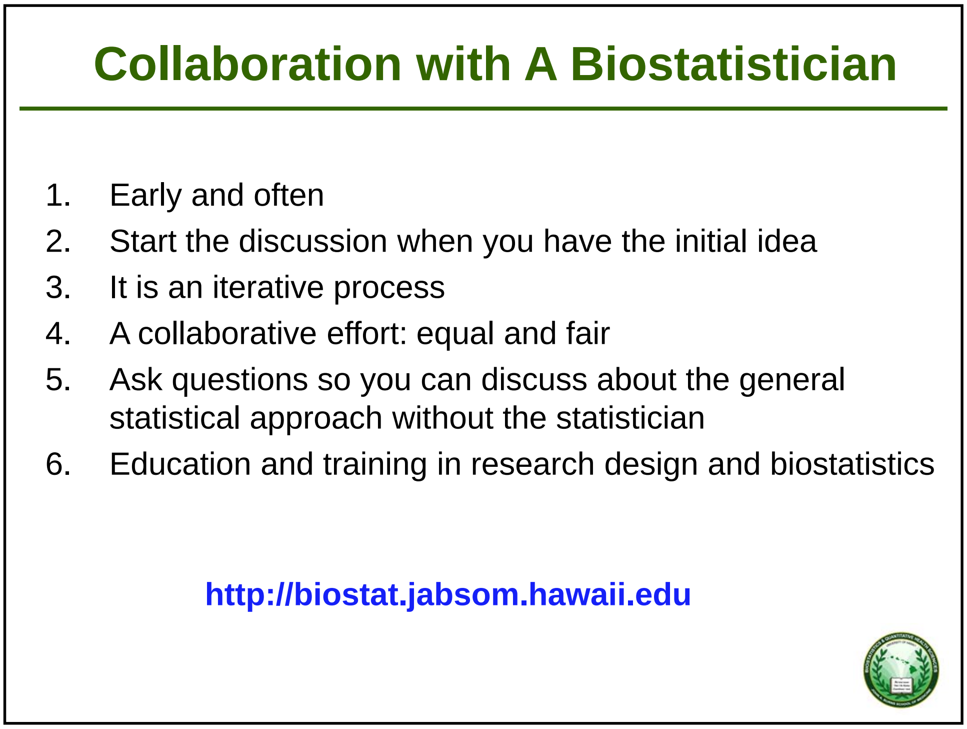
$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

Warning Signs

Data generation **Data consumption**

Different Scales

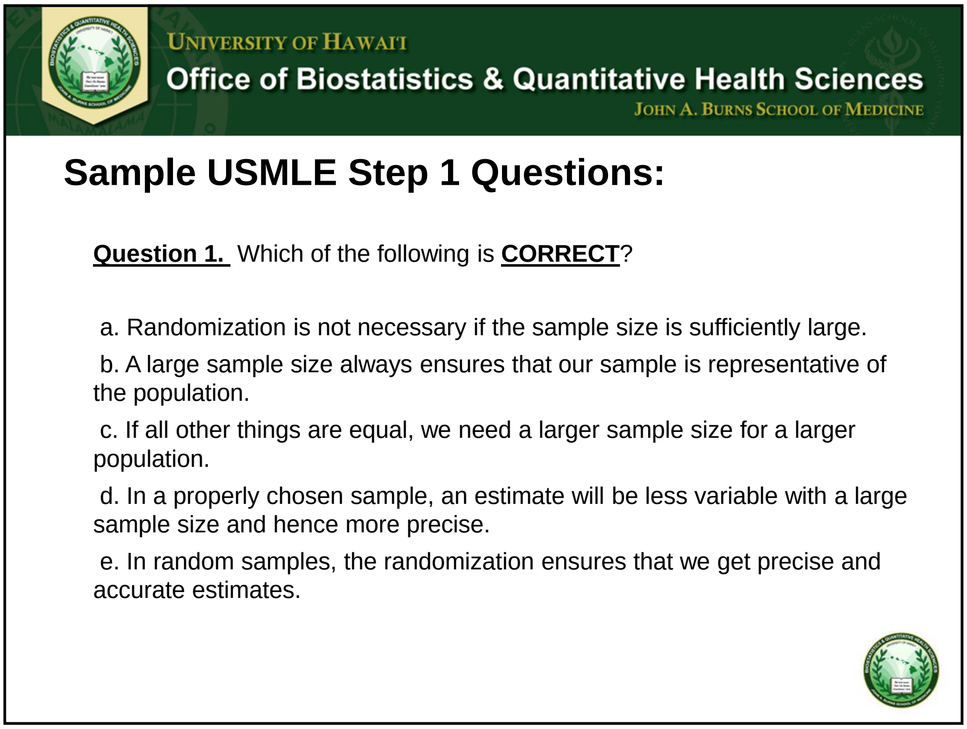




Collaboration with A Biostatistician

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



University of Hawai'i
Office of Biostatistics & Quantitative Health Sciences
John A. Burns School of Medicine

Sample USMLE Step 1 Questions:

Question 1. Which of the following is **CORRECT**?

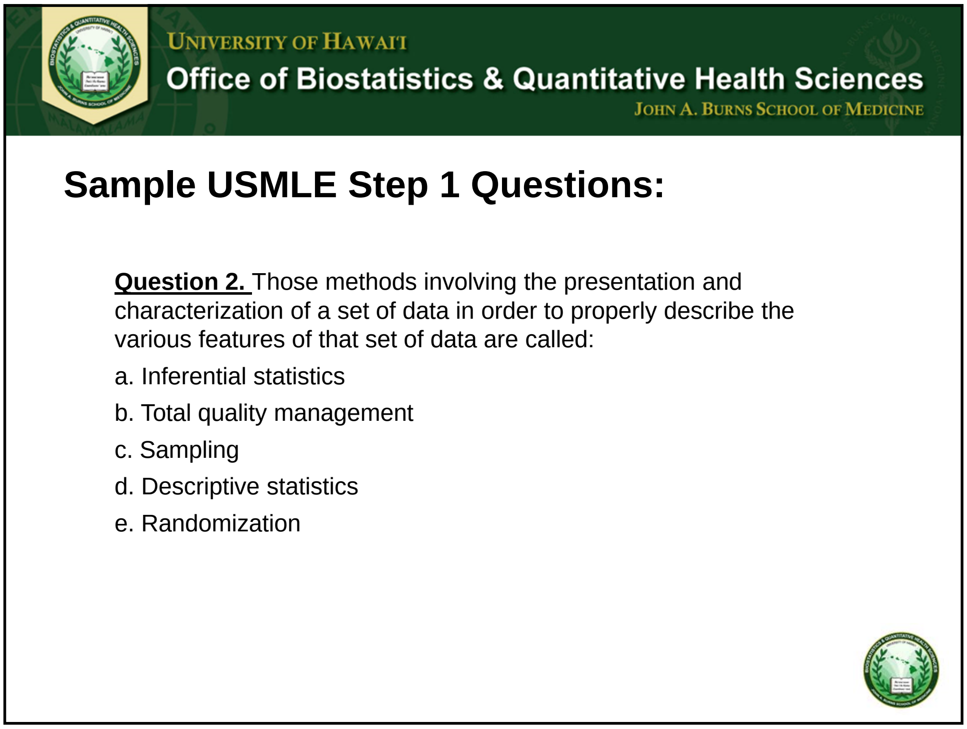
a. Randomization is not necessary if the sample size is sufficiently large.

b. A large sample size always ensures that our sample is representative of the population.

c. If all other things are equal, we need a larger sample size for a larger population.

d. In a properly chosen sample, an estimate will be less variable with a large sample size and hence more precise.

e. In random samples, the randomization ensures that we get precise and accurate estimates.



UNIVERSITY OF HAWAI'I

Office of Biostatistics & Quantitative Health Sciences

JOHN A. BURNS SCHOOL OF MEDICINE

Sample USMLE Step 1 Questions:

Question 2. Those methods involving the presentation and characterization of a set of data in order to properly describe the various features of that set of data are called:

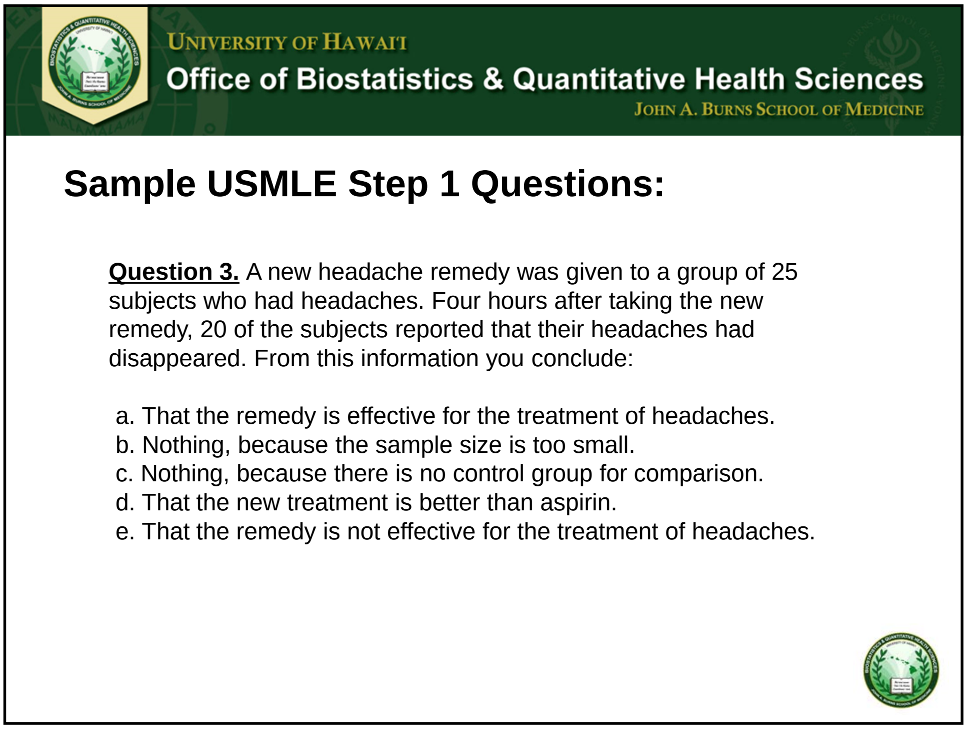
a. Inferential statistics

b. Total quality management

c. Sampling

d. Descriptive statistics

e. Randomization



UNIVERSITY OF HAWAI'I

Office of Biostatistics & Quantitative Health Sciences

JOHN A. BURNS SCHOOL OF MEDICINE

Sample USMLE Step 1 Questions:

Question 3. A new headache remedy was given to a group of 25 subjects who had headaches. Four hours after taking the new remedy, 20 of the subjects reported that their headaches had disappeared. From this information you conclude:

a. That the remedy is effective for the treatment of headaches.
b. Nothing, because the sample size is too small.
c. Nothing, because there is no control group for comparison.
d. That the new treatment is better than aspirin.
e. That the remedy is not effective for the treatment of headaches.

Outline

Lecture 1 (02/14/2018)

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

Lecture 2 (02/21/2018)

- Introduction to inferential statistics
- Commonly used statistical approaches