The University of Zambia
School of Health Sciences
Department of Biomedical Sciences

Biostatistics, Epidemiology and Research Methods

Introduction to Statistics: Types of Variables, Central Tendency and Measures of Dispersion

S. M. Munsaka, PhD

4th June 2020

Learning Objectives

By the end of this lecturer you should understand:

- *The statistical goals of “estimation” and “inference”*
- *The importance of independent data, and random sampling*
- *Why variability is important*
- *The different types of variables/data (categorical, continuous, discrete and ordinal) and how to summarize each*
- *Measures of dispersion (variance, standard deviation and standard error)*

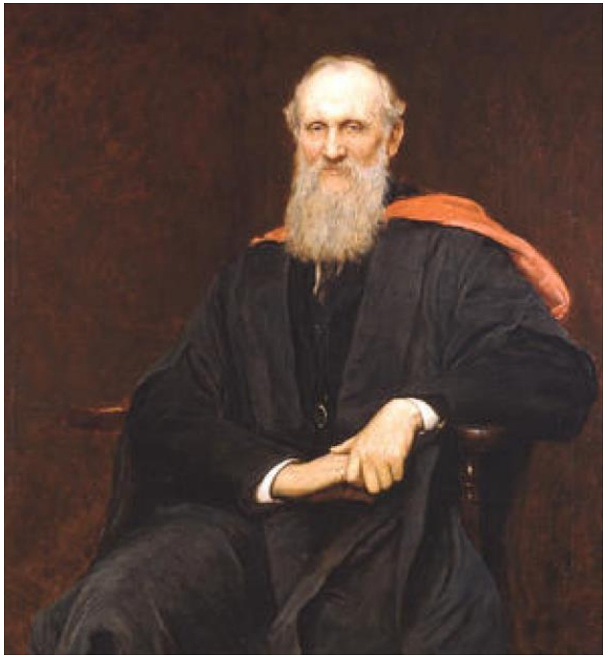
Introduction to statistics

- Statistic or datum: a measured or counted fact or piece of information presented as a figure
  - E.g. height, weight, age, etc
- Statistics or data (plurals) are collected from
  - Experiments/measurements
  - Records
  - Surveys
- Applications: all walks of life including medicine and public health (Biostatistics)
- Statistics is the science of figures; field concerned with the collection of data, classification, summarizing, interpretation, drawing inferences, testing hypothesis, making recommendations etc

Introduction to statistics

- Biostatistics: statistics applied to biological sciences (medicine and public health)

“I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be.” Lord Kelvin, 1883



Lord Kelvin, 1883

What is Statistics

- Statistics assumes there is an *unknown, true* value out there
  - Called the *parameter*
  - We'd have to measure everyone (the "*population*") to find out what it is.

    Usually, this is not possible.

- Instead we collect a *sample* and use our sample to *estimate* the population parameter

Choosing a sample

- Best is *random sample:* every member of population has equal chance of being included
  - In reality, truly random sample may be impossible

- *Independent* observations best
  - Knowing the result for one individual should not give you any information about another individual
    - For example: If a person has a genetic
  - Avoid family members, household contacts (unless this is part of the design)
  - Each person counted once

- If your data are not independent, must consult statistician.

What is Statistics

- Statistics assumes there is an *unknown, true* value out there
- Instead we collect a *sample* and use our sample to *estimate* the population parameter

- Then we use statistics to figure out
  - how close our estimate likely is to the parameter
  - if we can rule out certain values of the parameter

“Estimation”

“Inference”

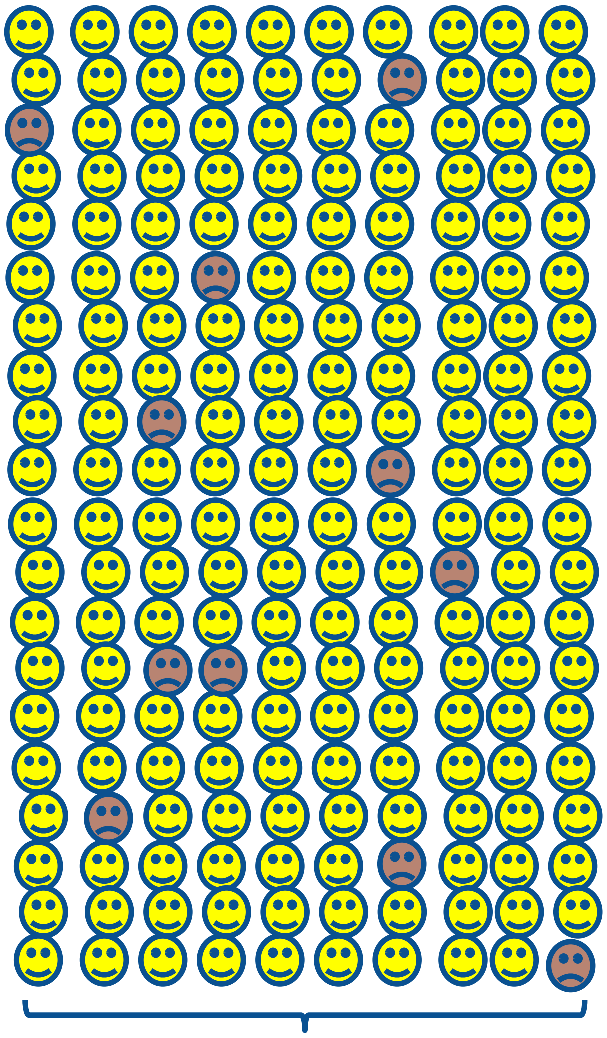
Example: Estimation

Suppose we want to know the prevalence of a certain disease in a specific population

Truth:

- Population is 200 people
- 10 people in population have disease
- Prevalence of Disease is 10/200 = 5%

If we could test all 200 people, we would know the true prevalence (*the “parameter"*).



☺ Healthy Person

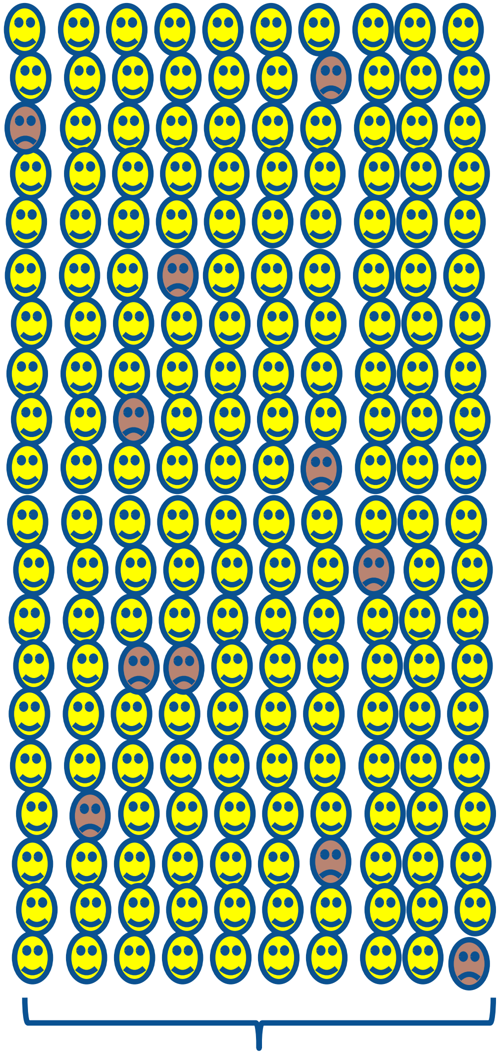
☹ Diseased Person

Example: Estimation

We cannot measure all 200, so we take a *sample*

Sampl
e

In this case, 2/20 in our sample have disease, so our *estimate* for the prevalence is 10%



Populatio
n

☺ Healthy Person

☹ Diseased Person

Introduction
B: Types of Studies

- Many studies are one of two types:
  - **Observational:** You observe data but don’t actively intervene
    - E.g., compare mortality in breast cancer patients choosing radical mastectomy versus lumpectomy
  - **Randomized clinical trials** (RCT): Randomly assign patients to treatments
    - E.g., randomly assign half the patients to radical mastectomy and half to lumpectomy

Introduction
B: Types of Studies

- Big difference between observational studies and randomized trials:
  - In observational studies, patients choosing versus not choosing treatment likely to differ in ways other than just treatment
    - May be more health-conscious (better diet, more exercise, etc., visit doctor more often, etc.)
  - In randomized trials, randomization makes treatment groups comparable with respect to other factors like diet, exercise, etc.

Introduction
B: Types of Studies

- If see difference in mortality between groups:
  - In observational study, can’t conclude treatment caused difference in mortality

    Maybe health-conscious behavior caused difference

  - In clinical trials, can conclude that treatment caused difference in mortality

Randomization makes groups comparable in ways other than treatment received (e.g., health-conscious behavior)

Introduction
B: Types of Studies

- Two common observational studies:
  - Cohort study: Study one group (cohort) with risk factor (e.g., smoking) & another without
    - Which group has more disease over next 5 years?
    - Prospective study: looking from now to future
  - Case/control study: Compare group with disease (cases) to group of controls
    - Which group engaged in risk factor more?
    - Retrospective: looking backward in time

Introduction
B: Types of Studies

- Usual sequence:
  - Epidemiologists analyzing observational data identify risk factor
    - Cholesterol and heart disease
    - Smoking and lung cancer, etc.
  - Other observational studies and animal studies confirm relationship
  - Intervention developed to alter risk factor
  - RCT definitively answers whether intervention alters mortality/heart attacks, etc.
- Observational studies & clinical trials often complement each other
  - Consistent results bolster evidence

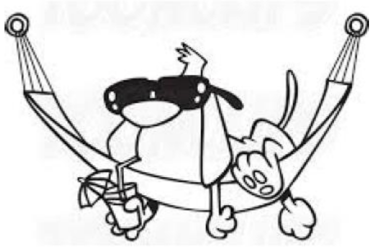
Why measurement

- Measurements supply us with numbers used in data analysis.

- No matter how profound are the theoretical formulation, how sophisticated the experimental designs are, or how elegant the analytical techniques, you **cannot compensate for po measures**.

27°F

- Units of measurement are equally importan



27°C

Describing Data

- *Descriptive statistics* (describing sample):
  - What proportions of the 2 groups of women developed CVD?
  - What is the average Viral Load among HIV+ patients who are taking ARVs and among those not taking ARVs?

- *Inferential statistics* (inferring about population):
  - Do we have evidence that an observed difference was not just due to chance?

Describing Data

- Who is in your study?
  - Summary of the patients in your sample
    - Demographics: Age, Sex (Male / Female)
    - Disease or Health State
    - Other factors possibly-related to disease or health
  - If you are comparing groups, summarize groups separately
  - Often “Table 1” in a paper
- What were their outcomes?

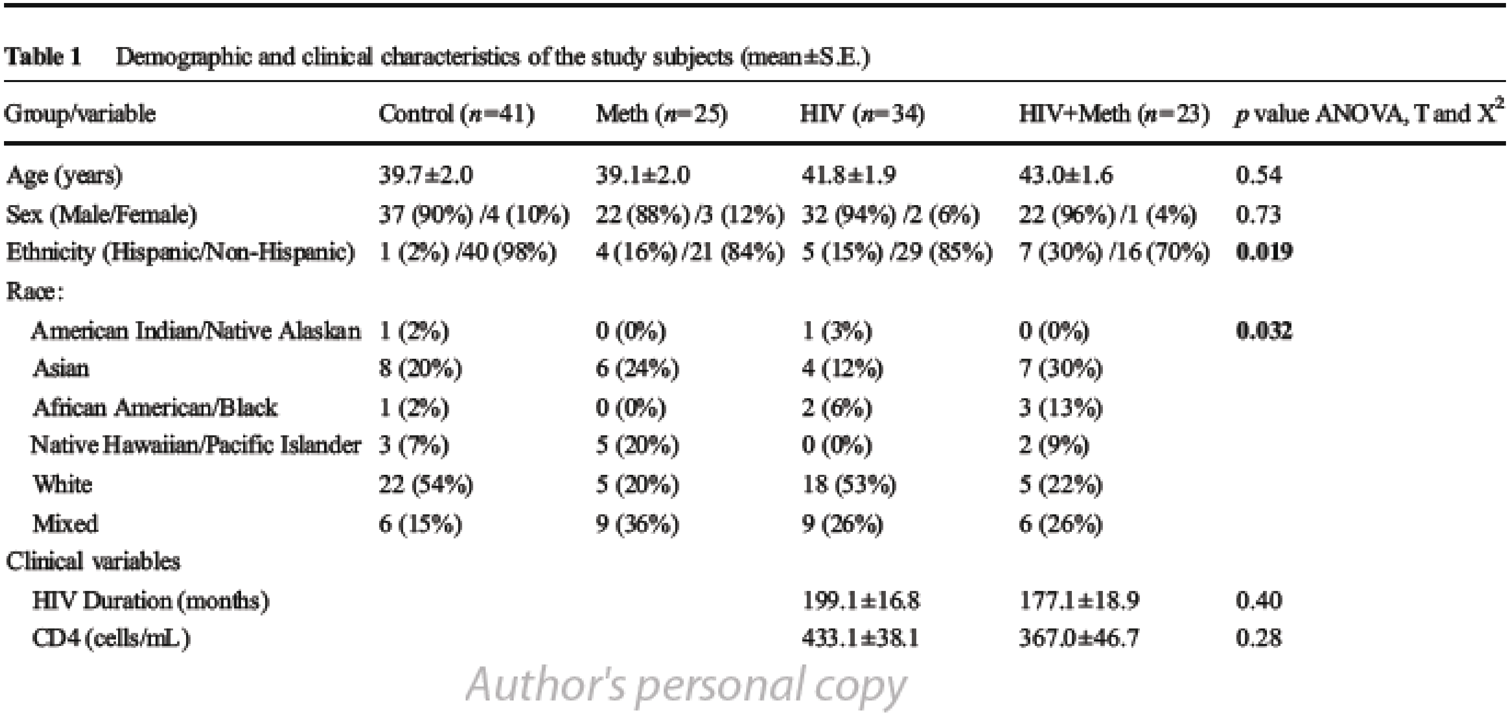
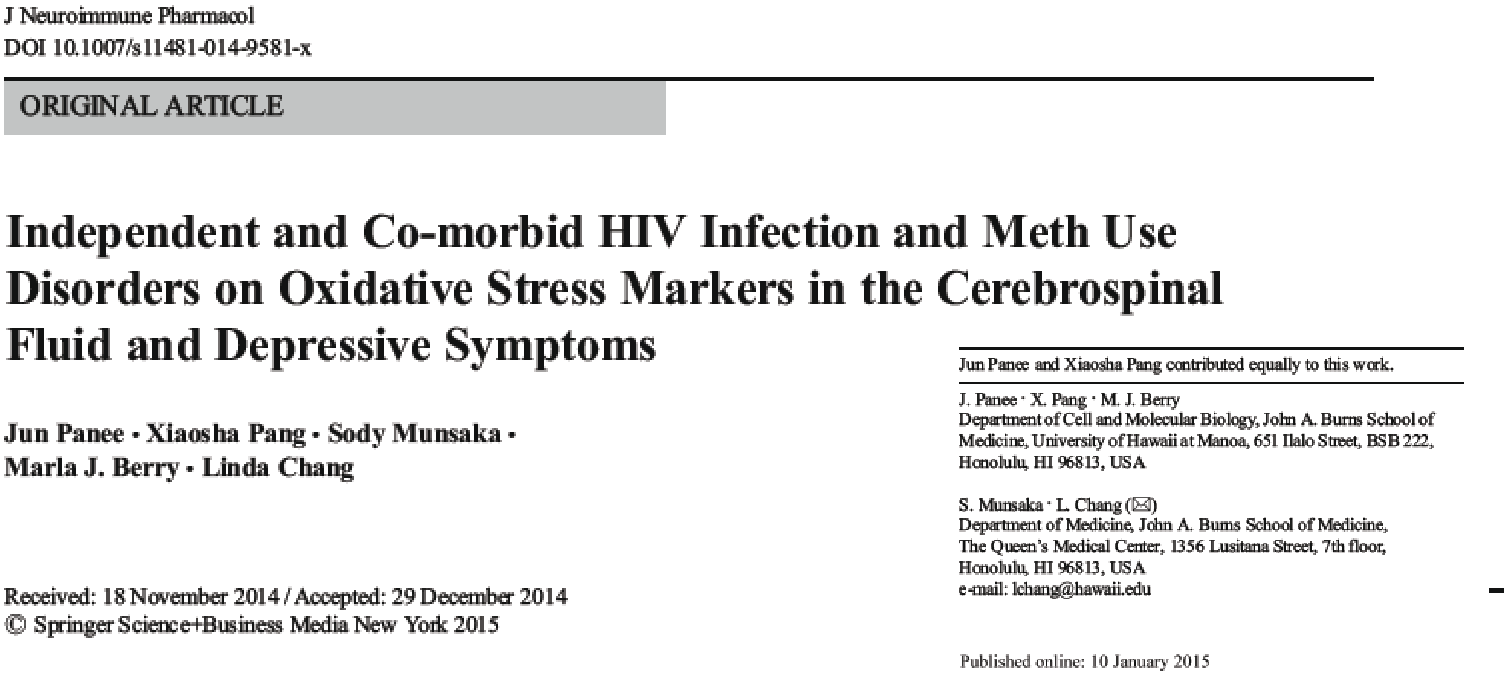


Table 1 Demographic and clinical characteristics of the study subjects (mean±S.E.)

Group/variable	Control (*n*=41)	Meth (*n*=25)	HIV (*n*=34)	HIV+Meth (*n*=23)	*p* value ANOVA, T and X^2
Age (years)	39.7±2.0	39.1±2.0	41.8±1.9	43.0±1.6	0.54
Sex (Male/Female)	37 (90%) /4 (10%)	22 (88%) /3 (12%)	32 (94%) /2 (6%)	22 (96%) /1 (4%)	0.73
Ethnicity (Hispanic/Non-Hispanic)	1 (2%) /40 (98%)	4 (16%) /21 (84%)	5 (15%) /29 (85%)	7 (30%) /16 (70%)	**0.019**
Race:					
American Indian/Native Alaskan	1 (2%)	0 (0%)	1 (3%)	0 (0%)	**0.032**
Asian	8 (20%)	6 (24%)	4 (12%)	7 (30%)	
African American/Black	1 (2%)	0 (0%)	2 (6%)	3 (13%)	
Native Hawaiian/Pacific Islander	3 (7%)	5 (20%)	0 (0%)	2 (9%)	
White	22 (54%)	5 (20%)	18 (53%)	5 (22%)	
Mixed	6 (15%)	9 (36%)	9 (26%)	6 (26%)	
Clinical variables					
HIV Duration (months)			199.1±16.8	177.1±18.9	0.40
CD4 (cells/mL)			433.1±38.1	367.0±46.7	0.28



J Neuroimmune Pharmacol
DOI 10.1007/s11481-014-9581-x

ORIGINAL ARTICLE

Independent and Co-morbid HIV Infection and Meth Use Disorders on Oxidative Stress Markers in the Cerebrospinal Fluid and Depressive Symptoms

Jun Panee · Xiaosha Pang · Sody Munsaka · Marla J. Berry · Linda Chang

Received: 18 November 2014 / Accepted: 29 December 2014

Jun Panee and Xiaosha Pang contributed equally to this work.

J. Panee · X. Pang · M. J. Berry
Department of Cell and Molecular Biology, John A. Burns School of Medicine, University of Hawaii at Manoa, 651 Ilalo Street, BSB 222, Honolulu, HI 96813, USA

S. Munsaka · L. Chang (✉)
Department of Medicine, John A. Burns School of Medicine, The Queen's Medical Center, 1356 Lusitana Street, 7th floor, Honolulu, HI 96813, USA
e-mail: lchang@hawaii.edu

Published online: 10 January 2015

Describing Data

Why is this important?

- Results may be specific to the type of people included in your study; narrow group or different types included?
- Credibility of Results influenced by
  - How many people
  - Who is included
  - Are groups comparable on demographics and (known) risk factors?

Comparability of Groups especially important for observational data

Describing Data

- How to describe your data depends on what kind of data it is
  - Categorical
    - Nominal: order doesn’t matter (Gender)
    - Ordinal: categorical but ordered (Educational Degree)
  - Discrete:
    - Order and magnitude matter, but possible values can be listed (Number of Seizures)
  - Continuous:
    - Data can take on any value in an interval (Systolic Blood Pressure)

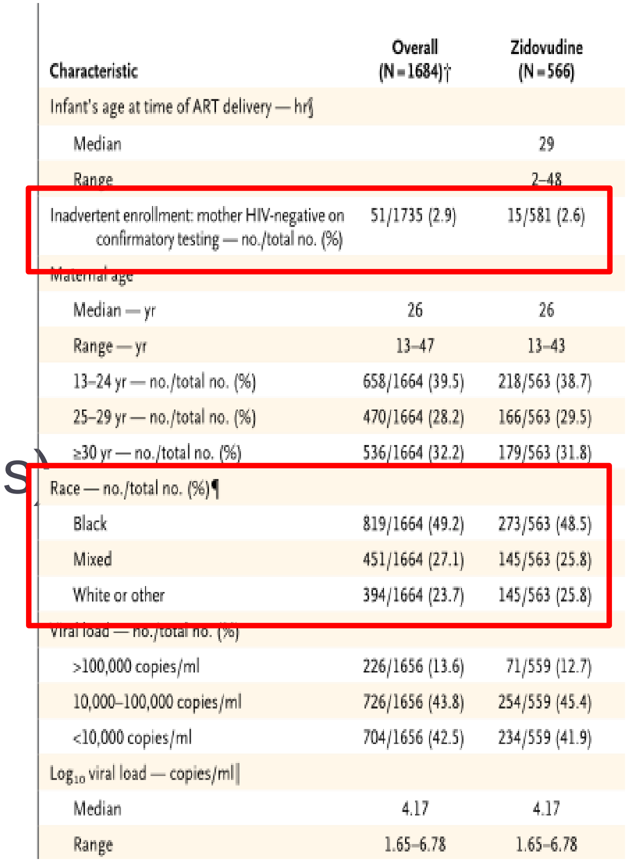
(Continuous data are measured, discrete data are counted)

Ordinal data

Data can be ordered e.g. severity of symptoms

Describing Data
Categorical Data

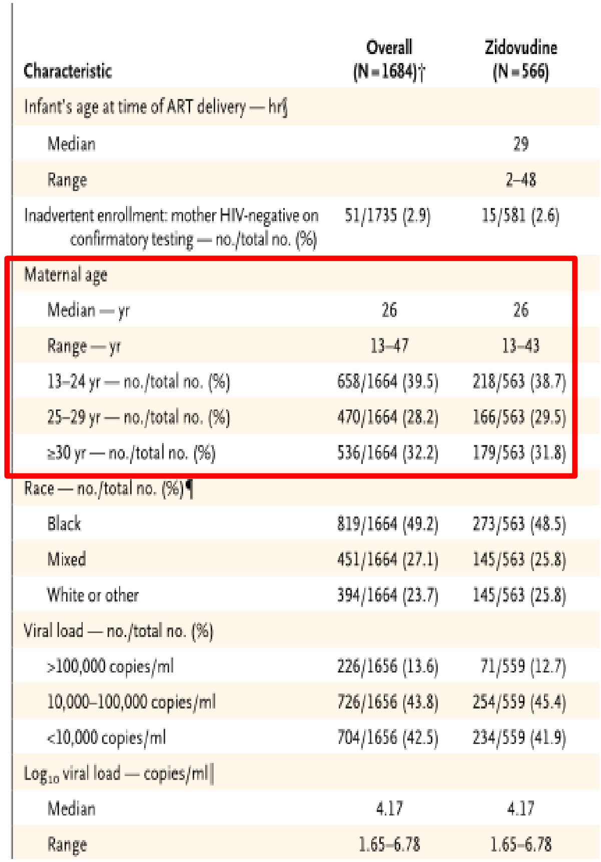
- Summarize with:
  - Number of people in each category
  - % of people in each category
- Optional:
  - Confidence Interval
  - P-value (if comparing 2 or more groups)



Characteristic	Overall (N=1684)†	Zidovudine (N=566)
Infant's age at time of ART delivery — hr§		
Median		29
Range		2–48
Inadvertent enrollment: mother HIV-negative on confirmatory testing — no./total no. (%)	51/1735 (2.9)	15/581 (2.6)
Maternal age		
Median — yr	26	26
Range — yr	13–47	13–43
13–24 yr — no./total no. (%)	658/1664 (39.5)	218/563 (38.7)
25–29 yr — no./total no. (%)	470/1664 (28.2)	166/563 (29.5)
≥30 yr — no./total no. (%)	536/1664 (32.2)	179/563 (31.8)
Race — no./total no. (%)¶		
Black	819/1664 (49.2)	273/563 (48.5)
Mixed	451/1664 (27.1)	145/563 (25.8)
White or other	394/1664 (23.7)	145/563 (25.8)
Viral load — no./total no. (%)		
>100,000 copies/ml	226/1656 (13.6)	71/559 (12.7)
10,000–100,000 copies/ml	726/1656 (43.8)	254/559 (45.4)
<10,000 copies/ml	704/1656 (42.5)	234/559 (41.9)
Log_{10} viral load — copies/ml‖		
Median	4.17	4.17
Range	1.65–6.78	1.65–6.78

Describing Data
Discrete Data

- Few unique numbers– treat like categorical
- Medium unique numbers– make into categories
- Many unique numbers– treat like continuous
    - Example: age, CD4 count

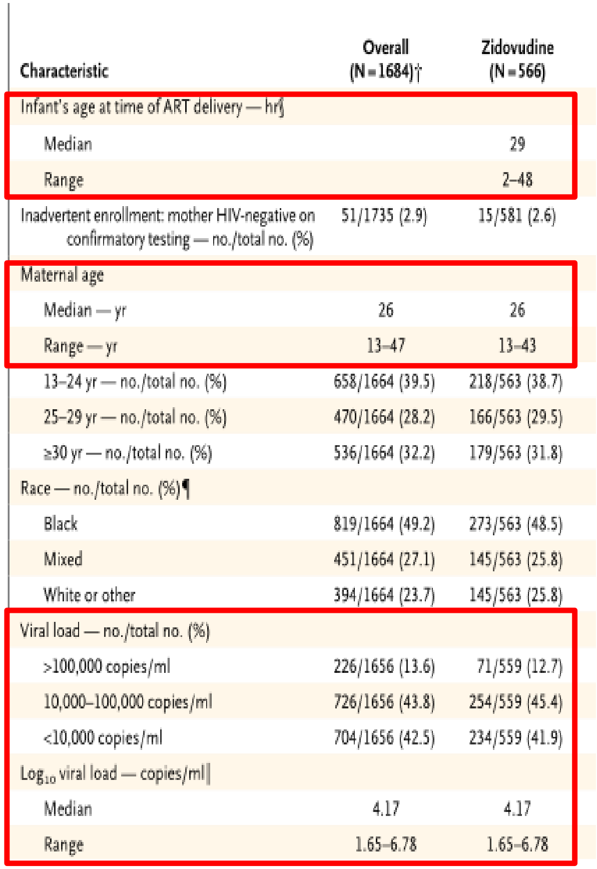


Characteristic	Overall (N=1684)†	Zidovudine (N=566)
Infant's age at time of ART delivery — hr§		
Median		29
Range		2–48
Inadvertent enrollment: mother HIV-negative on confirmatory testing — no./total no. (%)	51/1735 (2.9)	15/581 (2.6)
Maternal age		
Median — yr	26	26
Range — yr	13–47	13–43
13–24 yr — no./total no. (%)	658/1664 (39.5)	218/563 (38.7)
25–29 yr — no./total no. (%)	470/1664 (28.2)	166/563 (29.5)
≥30 yr — no./total no. (%)	536/1664 (32.2)	179/563 (31.8)
Race — no./total no. (%)¶		
Black	819/1664 (49.2)	273/563 (48.5)
Mixed	451/1664 (27.1)	145/563 (25.8)
White or other	394/1664 (23.7)	145/563 (25.8)
Viral load — no./total no. (%)		
>100,000 copies/ml	226/1656 (13.6)	71/559 (12.7)
10,000–100,000 copies/ml	726/1656 (43.8)	254/559 (45.4)
<10,000 copies/ml	704/1656 (42.5)	234/559 (41.9)
Log_{10} viral load — copies/ml‖		
Median	4.17	4.17
Range	1.65–6.78	1.65–6.78

Describing Data

Continuous Data

- Need to Describe:
  - CENTER of data
    - Mean (average)
    - Median
  - SPREAD / VARIABILITY of data
    - Range or Minimum/Maximum
    - Inter-quartile Range (IQR)
    - Standard Deviation or Variance



Characteristic	Overall (N=1684)†	Zidovudine (N=566)
Infant's age at time of ART delivery — hr§		
Median		29
Range		2–48
Inadvertent enrollment: mother HIV-negative on confirmatory testing — no./total no. (%)	51/1735 (2.9)	15/581 (2.6)
Maternal age		
Median — yr	26	26
Range — yr	13–47	13–43
13–24 yr — no./total no. (%)	658/1664 (39.5)	218/563 (38.7)
25–29 yr — no./total no. (%)	470/1664 (28.2)	166/563 (29.5)
≥30 yr — no./total no. (%)	536/1664 (32.2)	179/563 (31.8)
Race — no./total no. (%)¶		
Black	819/1664 (49.2)	273/563 (48.5)
Mixed	451/1664 (27.1)	145/563 (25.8)
White or other	394/1664 (23.7)	145/563 (25.8)
Viral load — no./total no. (%)		
>100,000 copies/ml	226/1656 (13.6)	71/559 (12.7)
10,000–100,000 copies/ml	726/1656 (43.8)	254/559 (45.4)
<10,000 copies/ml	704/1656 (42.5)	234/559 (41.9)
Log_{10} viral load — copies/ml‖		
Median	4.17	4.17
Range	1.65–6.78	1.65–6.78

Why do we need both the center and the variability?

The variability puts the magnitude into context

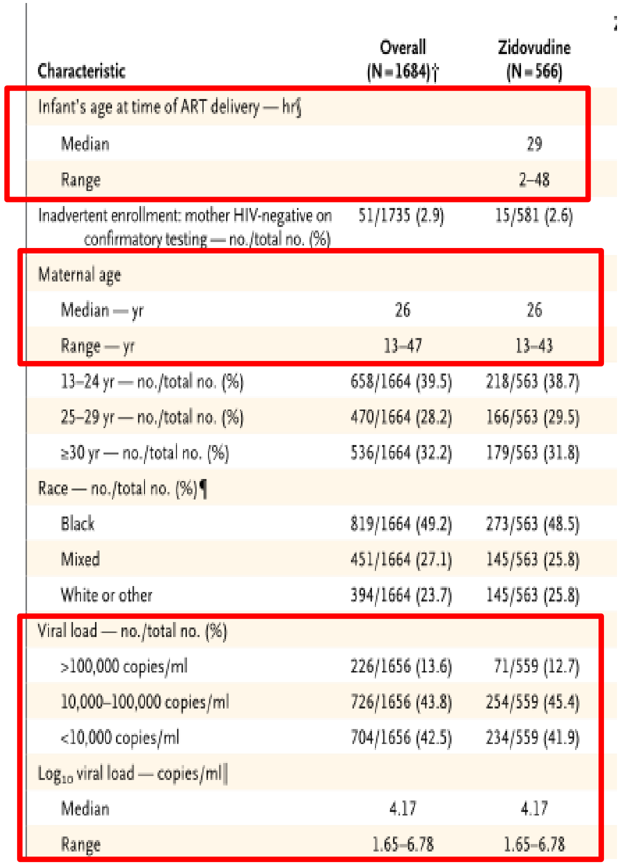
- If I tell you I have a drug that on average reduced days with influenza symptoms by 2 days among those who took it, do you know enough?
  - What if it reduced days of symptoms by 2 for everyone?
  - What if it reduced days of symptoms by 8-10 days for a few people and increased days of symptoms for everyone else?

Describing Data

Continuous Data

- First Question: Logarithmic (log) scale?
- To decide, which statement makes more sense?
  - “The value has doubled since Baseline”
  - “The value has increased by 10 since Baseline”

Use log-scale for any variables where you think in terms of doubling/ fold-change.



Characteristic	Overall (N=1684)†	Zidovudine (N=566)
Infant's age at time of ART delivery — hr§		
Median		29
Range		2–48
Inadvertent enrollment: mother HIV-negative on confirmatory testing — no./total no. (%)	51/1735 (2.9)	15/581 (2.6)
Maternal age		
Median — yr	26	26
Range — yr	13–47	13–43
13–24 yr — no./total no. (%)	658/1664 (39.5)	218/563 (38.7)
25–29 yr — no./total no. (%)	470/1664 (28.2)	166/563 (29.5)
≥30 yr — no./total no. (%)	536/1664 (32.2)	179/563 (31.8)
Race — no./total no. (%)¶		
Black	819/1664 (49.2)	273/563 (48.5)
Mixed	451/1664 (27.1)	145/563 (25.8)
White or other	394/1664 (23.7)	145/563 (25.8)
Viral load — no./total no. (%)		
>100,000 copies/ml	226/1656 (13.6)	71/559 (12.7)
10,000–100,000 copies/ml	726/1656 (43.8)	254/559 (45.4)
<10,000 copies/ml	704/1656 (42.5)	234/559 (41.9)
Log_{10} viral load — copies/ml‖		
Median	4.17	4.17
Range	1.65–6.78	1.65–6.78

Describing Data
Continuous Data

- Need to Describe:
  - CENTER of data
    - Mean (average)
      - Add up data and divide by number of observations
    - Median
      - Put data in order and find the “middle” one

Quiz Question 1:

Question 1.

You collect 5 data points:

1.0, 2.5,1.5,2.0,28.0

Among the choices below, what would best describe the measure of center?

A) Mean

B) Median

C) Proportion

D) Standard Deviation

Quiz Question 1: Answer

Data: 1.0, 2.5,1.5,2.0,28.0

- Mean: (1.0+2.5+1.5+2.0+28.0)/5=35/5=7

- Median: Put data in order then find middle:
  - 1.0,1.5,2.0,2.5,28.0
  - Median=2

In this example, outlier (28.0) caused mean to be much larger than median

Describing Data
Continuous Data

- Need to Describe:
  - CENTER of data
    - Mean (average)
      - Add up data and divide by number of observations
    - Median
      - Put data in order and find the “middle” one

Median is often a better choice, especially if there are any outliers or extreme values in your data

Median also unchanged by log scale

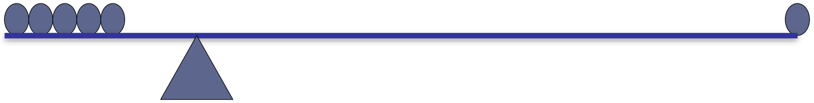
If you take the median of logged values, you get the same answer as if you take the log of the median.

If you have values below a lower limit of detection (or above an upper limit), can still compute Median

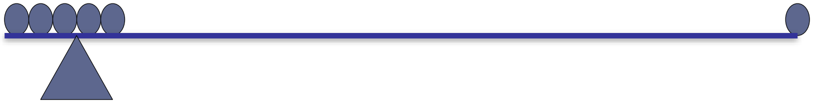
II. Summarizing Data
B. Continuous Data

When the distribution is skewed or has outliers, the median is more meaningful than the mean

- If there are high outliers, mean will be too high
- If there are low outliers, mean will be too low

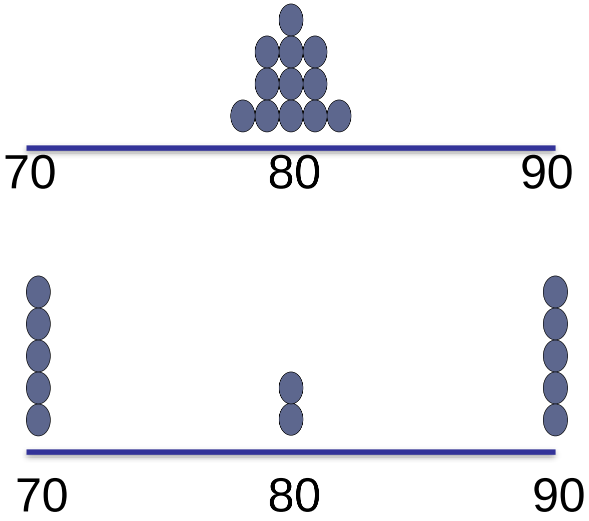


- Median will be just right
  - Half of the data are larger, and half smaller, than the median



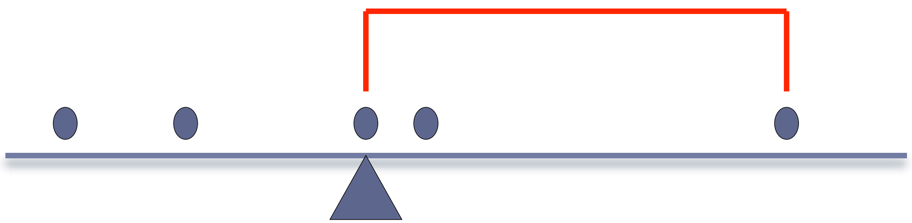
Describing Data
B. Continuous Data

- Another feature of continuous data is how spread out they are
- Two datasets with the same center can be concentrated or more spread out



Describing Data
B. Continuous Data

- One way to measure how spread out the data are is to average the squared deviations (distances) from the mean, called the *variance*

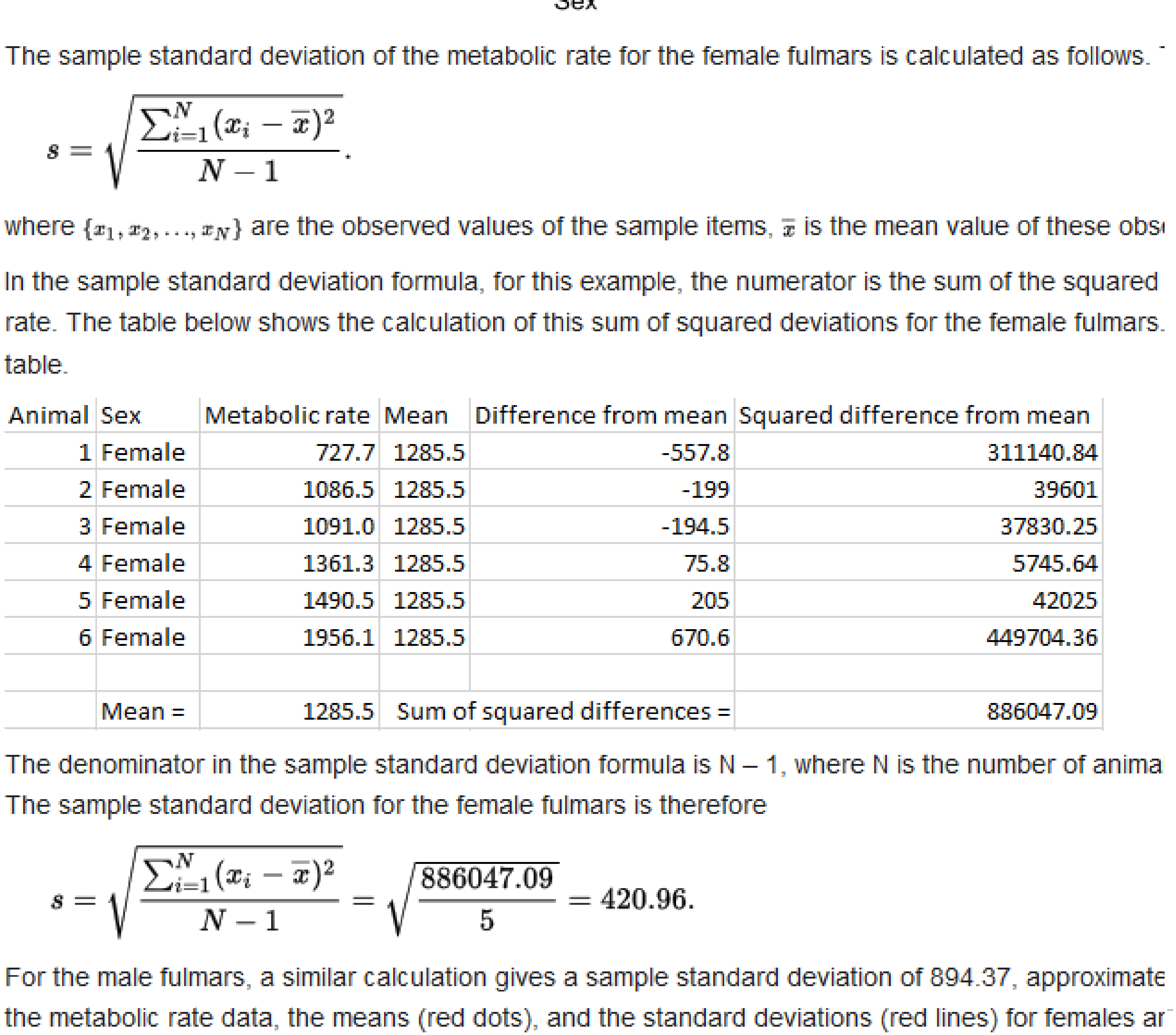


Deviation from mean

Square these and average

The closer the points are to the mean, the smaller the varian

- Closely related is the *standard deviation*, the square root of the variance
- Slightly easier to interpret than variance
  – Units are same as data (e.g., feet, yards, etc.)



Sex

The sample standard deviation of the metabolic rate for the female fulmars is calculated as follows.

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}.$$

where $\{x_1, x_2, \ldots, x_N\}$ are the observed values of the sample items, $\overline{x}$ is the mean value of these obs

In the sample standard deviation formula, for this example, the numerator is the sum of the squared rate. The table below shows the calculation of this sum of squared deviations for the female fulmars. table.

Animal	Sex	Metabolic rate	Mean	Difference from mean	Squared difference from mean
1	Female	727.7	1285.5	-557.8	311140.84
2	Female	1086.5	1285.5	-199	39601
3	Female	1091.0	1285.5	-194.5	37830.25
4	Female	1361.3	1285.5	75.8	5745.64
5	Female	1490.5	1285.5	205	42025
6	Female	1956.1	1285.5	670.6	449704.36
	Mean =	1285.5	Sum of squared differences =		886047.09

The denominator in the sample standard deviation formula is N – 1, where N is the number of anima The sample standard deviation for the female fulmars is therefore

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}} = \sqrt{\frac{886047.09}{5}} = 420.96.$$

For the male fulmars, a similar calculation gives a sample standard deviation of 894.37, approximate the metabolic rate data, the means (red dots), and the standard deviations (red lines) for females ar

Describing Data
Continuous Data

- Need to Describe:
  - SPREAD / VARIABILITY of data
    - Range or Minimum/Maximum
      - Range is the distance between highest and lowest
    - 1st / 3rd quartile or 25%ile / 75%ile
      - Describes the values where:
        - 25% of data is below and 75% above
        - 75% of data is below and 25% above
    - Inter-quartile Range (IQR)
      - Difference between 1st and 3rd quartiles
      - Describes where middle half of data fall

These three work well with median

    - Standard Deviation or Variance
      - "average" distance of each data point from mean

works well with mean

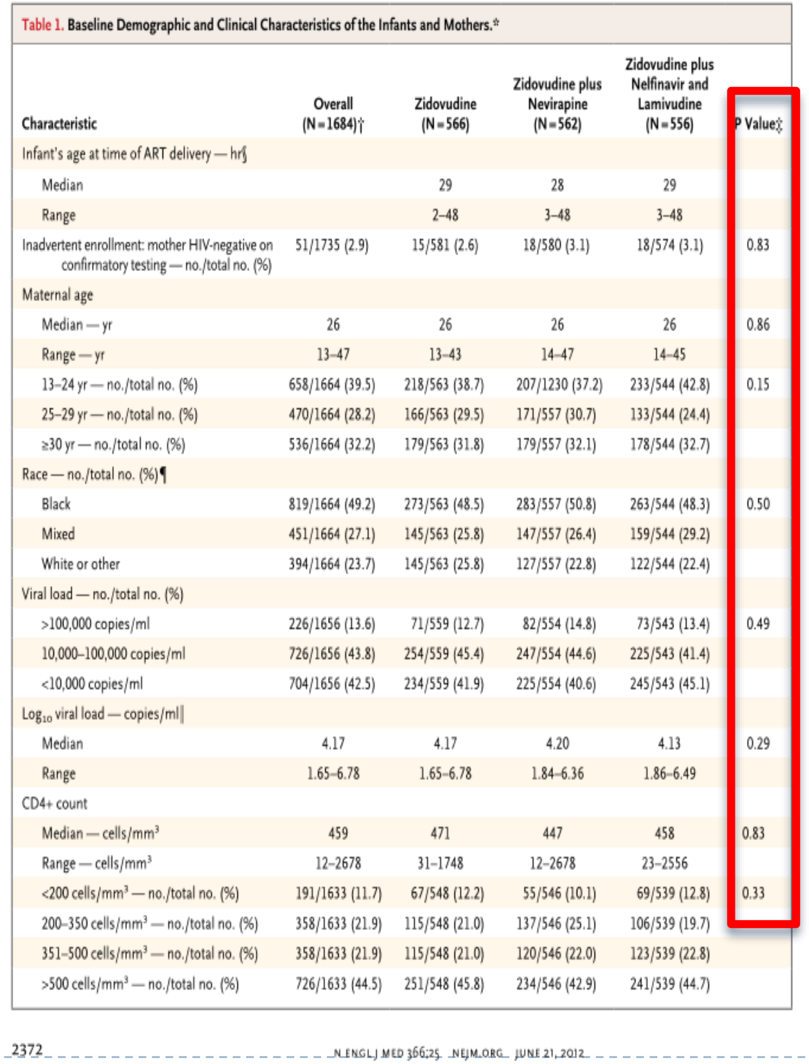


Table 1. Baseline Demographic and Clinical Characteristics of the Infants and Mothers.*

Characteristic	Overall (N=1684)†	Zidovudine (N=566)	Zidovudine plus Nevirapine (N=562)	Zidovudine plus Nelfinavir and Lamivudine (N=556)	P Value‡
Infant's age at time of ART delivery — hr§					
Median		29	28	29	
Range		2–48	3–48	3–48	
Inadvertent enrollment: mother HIV-negative on confirmatory testing — no./total no. (%)	51/1735 (2.9)	15/581 (2.6)	18/580 (3.1)	18/574 (3.1)	0.83
Maternal age					
Median — yr	26	26	26	26	0.86
Range — yr	13–47	13–43	14–47	14–45	
13–24 yr — no./total no. (%)	658/1664 (39.5)	218/563 (38.7)	207/1230 (37.2)	233/544 (42.8)	0.15
25–29 yr — no./total no. (%)	470/1664 (28.2)	166/563 (29.5)	171/557 (30.7)	133/544 (24.4)	
≥30 yr — no./total no. (%)	536/1664 (32.2)	179/563 (31.8)	179/557 (32.1)	178/544 (32.7)	
Race — no./total no. (%)¶					
Black	819/1664 (49.2)	273/563 (48.5)	283/557 (50.8)	263/544 (48.3)	0.50
Mixed	451/1664 (27.1)	145/563 (25.8)	147/557 (26.4)	159/544 (29.2)	
White or other	394/1664 (23.7)	145/563 (25.8)	127/557 (22.8)	122/544 (22.4)	
Viral load — no./total no. (%)					
>100,000 copies/ml	226/1656 (13.6)	71/559 (12.7)	82/554 (14.8)	73/543 (13.4)	0.49
10,000–100,000 copies/ml	726/1656 (43.8)	254/559 (45.4)	247/554 (44.6)	225/543 (41.4)	
<10,000 copies/ml	704/1656 (42.5)	234/559 (41.9)	225/554 (40.6)	245/543 (45.1)	
Log_{10} viral load — copies/ml‖					
Median	4.17	4.17	4.20	4.13	0.29
Range	1.65–6.78	1.65–6.78	1.84–6.36	1.86–6.49	
CD4+ count					
Median — cells/mm³	459	471	447	458	0.83
Range — cells/mm³	12–2678	31–1748	12–2678	23–2556	
<200 cells/mm³ — no./total no. (%)	191/1633 (11.7)	67/548 (12.2)	55/546 (10.1)	69/539 (12.8)	0.33
200–350 cells/mm³ — no./total no. (%)	358/1633 (21.9)	115/548 (21.0)	137/546 (25.1)	106/539 (19.7)	
351–500 cells/mm³ — no./total no. (%)	358/1633 (21.9)	115/548 (21.0)	120/546 (22.0)	123/539 (22.8)	
>500 cells/mm³ — no./total no. (%)	726/1633 (44.5)	251/548 (45.8)	234/546 (42.9)	241/539 (44.7)	

2372 N ENGL J MED 366;25 NEJM.ORG JUNE 21, 2012

Here, p-value tells you if there is evidence groups are different– see Statistics part II tomorrow!

Summarizing Data
B. Continuous Data

One useful graph is the *boxplot*

To make a boxplot, first compute three quartiles

- Order data from smallest to largest
- *First quartile* is data point x such that at least 25% of values are ≤ x and at least 75% are ≥ x
- For *second quartile (median)*, at least 50% of values are ≤ x and at least 50% are ≥ x
  - Middle observation if odd number of values
  - Average middle two if even number of values
- For *third quartile*, at least 75% of values are ≤ x and at least 25% are ≥ x

Summarizing Data
B. Continuous Data

- E.g., 20,2,9,3,1,25,17,22,23,12,9,5,6,16
- Order 14 data points from smallest to largest:

1,2,3,5,6,9,9,12,16,17,20,22,23,25

- To compute median: 50% of 14 is 7, so need at least 7 points ≤ x and 7 points ≥ x

Summarizing Data
B. Continuous Data

1,2,3,5,6,9,9 | 12,16,17,20,22,23,25

Median: average 9 and 12: (9+12)/2=10.5

- First and third quartiles are a little trickier: 5.25 and 19.25 **

- *Interquartile range* (difference between 1st and 3rd quartiles)=19.25-5.25=14

**Surprising number of methods for computing quantiles: 4 and 18.5 would be totally acceptable, just a different method

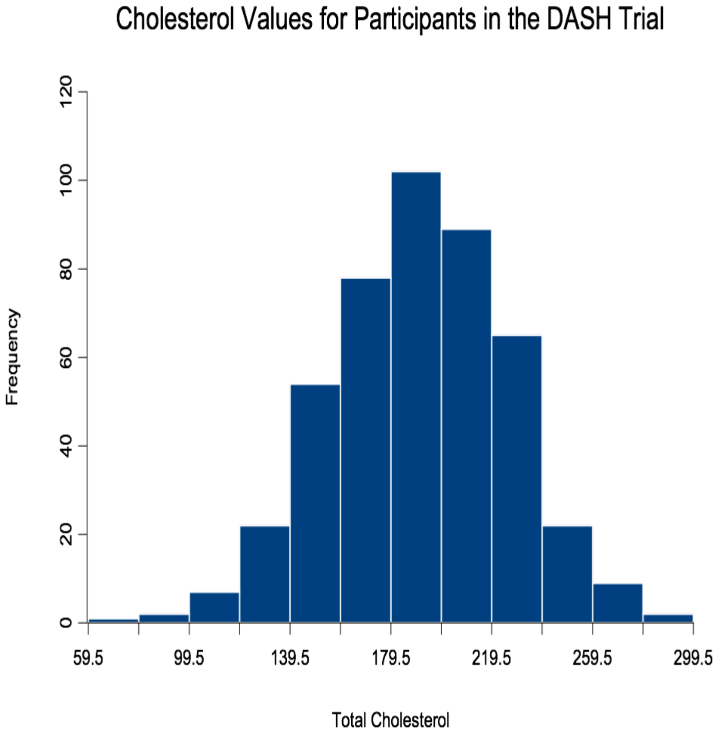
II: Summarizing Data
B. Continuous Data

- Boxplot is box with sides at first and third quartiles
  - Median is line within the box
  - Whiskers drawn at closest points within 1.5 times the interquartile range of the first and third quartiles
  - Observations outside whiskers (outliers) are shown as lines above or below whiskers

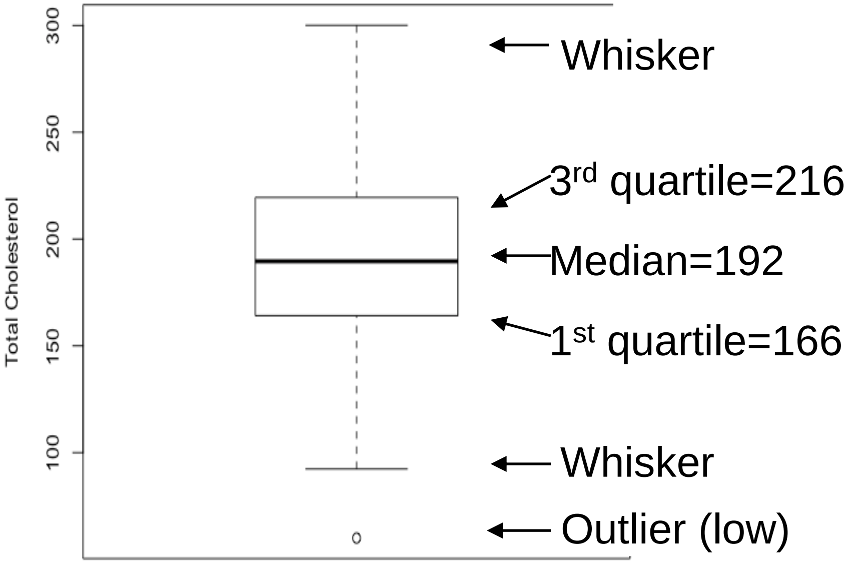
Describing Data

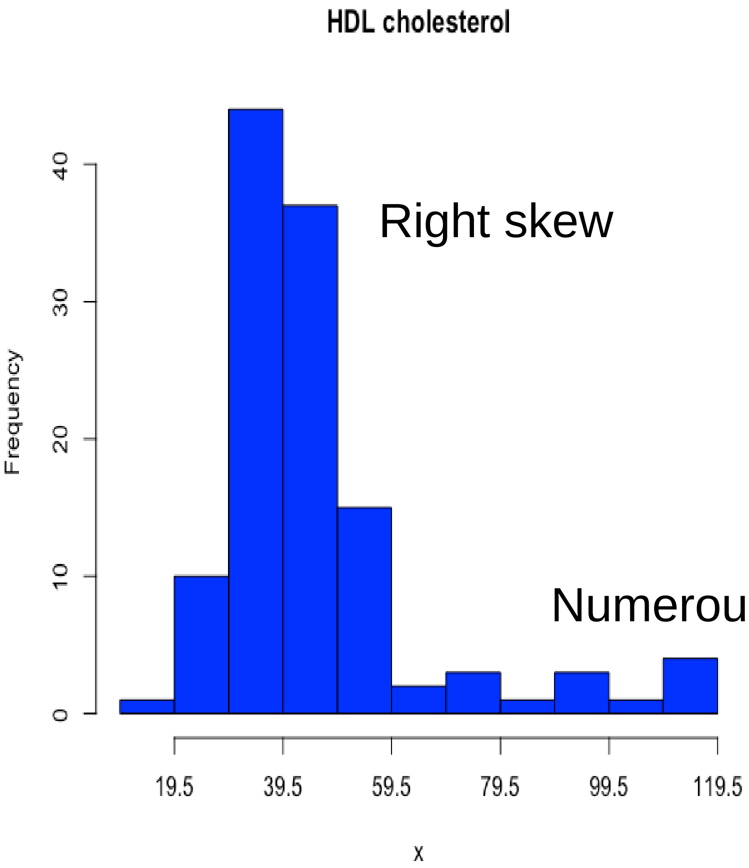
Continuous Data — Graphical

Histogram:



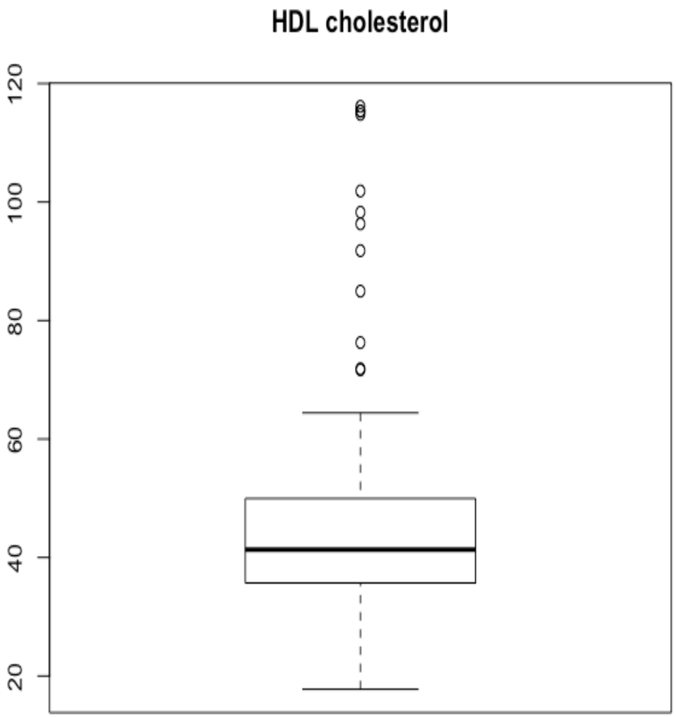
Boxplot:





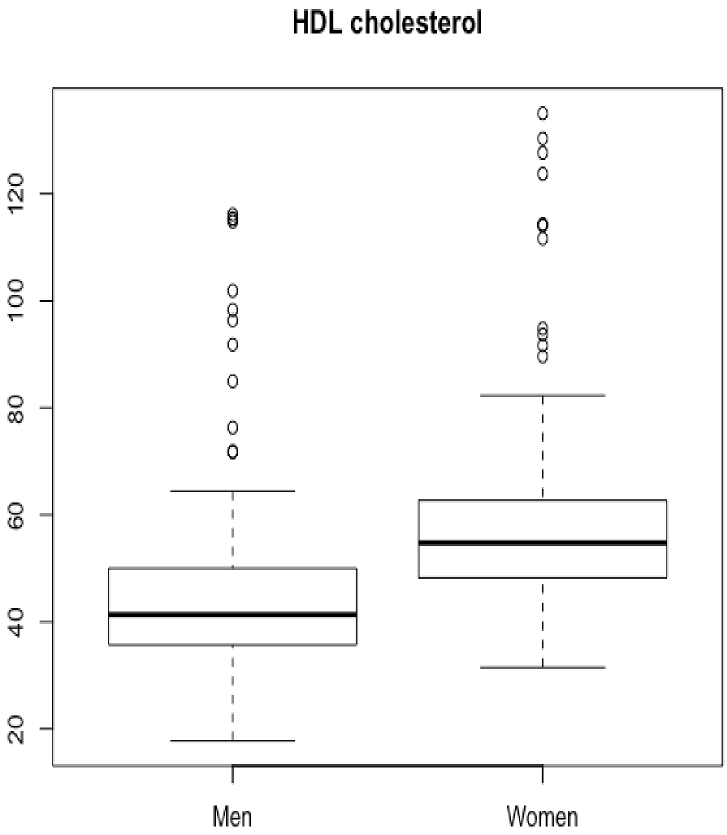
Right skew

Numerous high ou

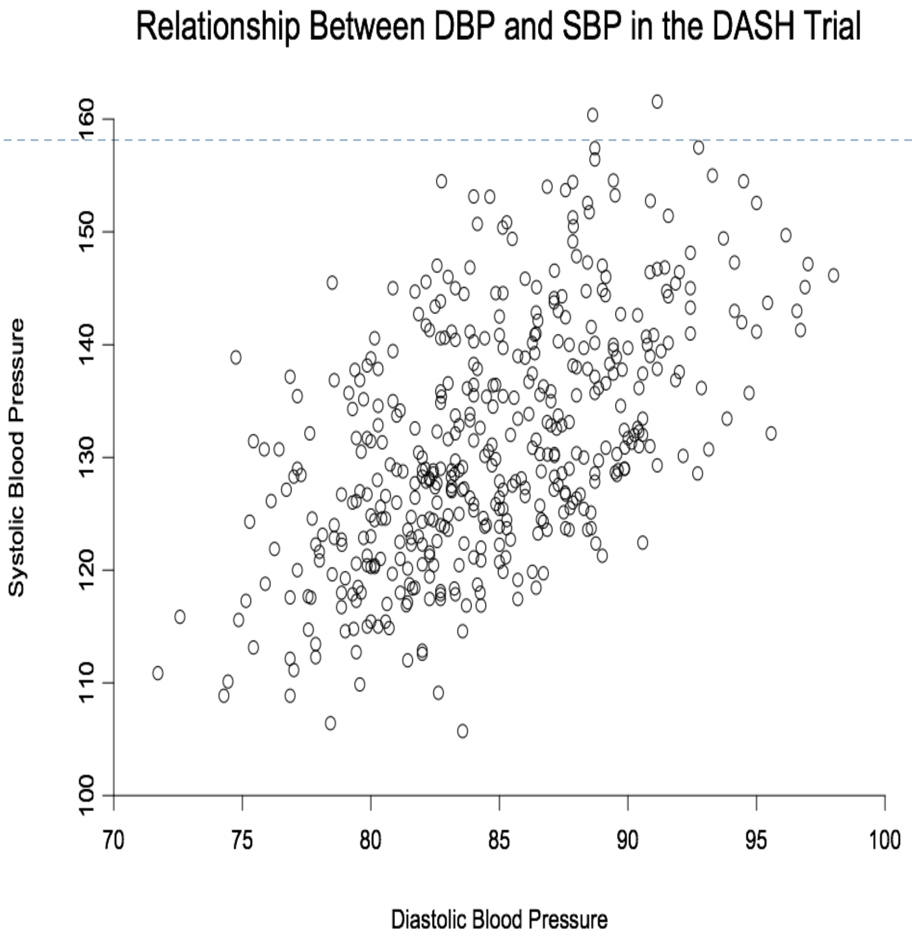


Describing Data
B. Relationships between variables

- Much of statistical analysis is about examining the relationship between 2 or more variables
  - Do high values on one assay predict high values on a gold standard assay?
  - Are there significant differences in immune markers among people who are mono-infected with HIV or TB or who are co-infected?
  - Do people with high LDL (“bad”) cholesterol tend to have low HDL (“good”) cholesterol?
- Plots are often a good place to start for describing relationships between variables

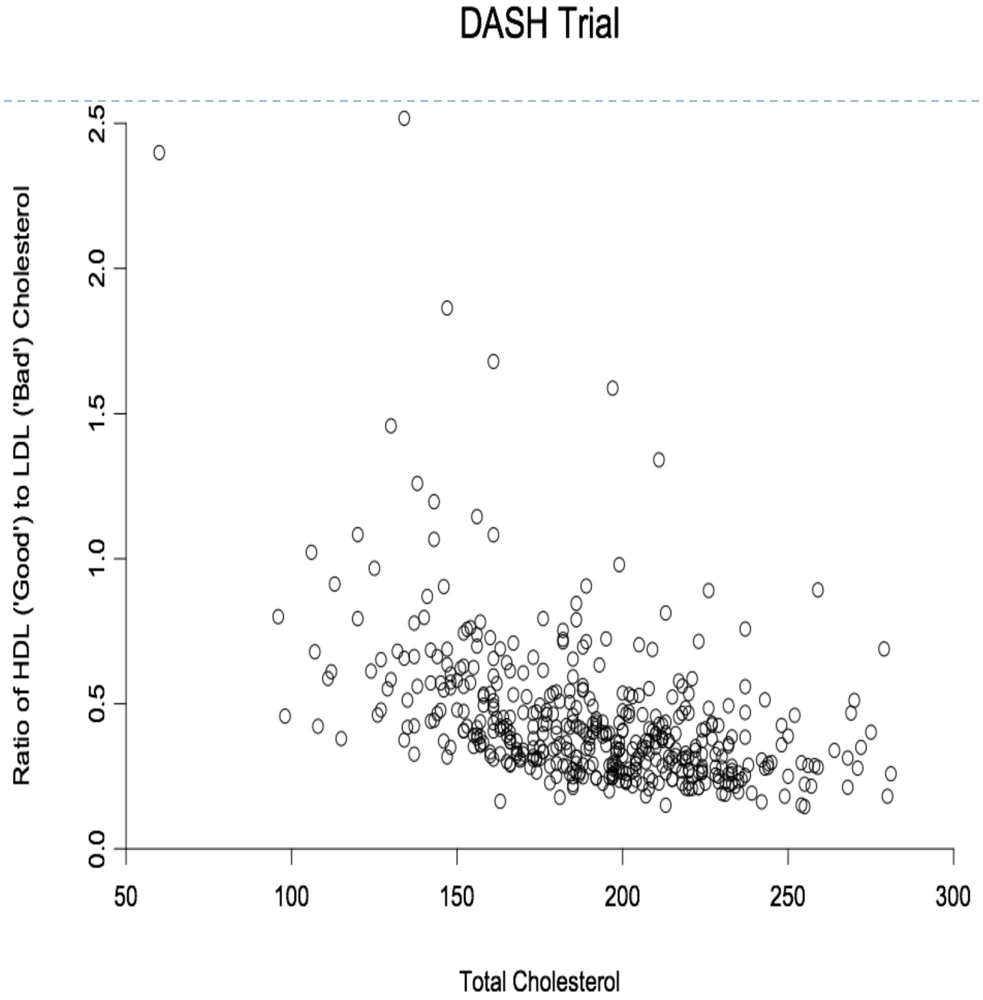


Side-by-side boxplot shows women appear to have higher HDLs thar



Each circle is (DBP, SBP) pair for one person

As move to right, circles tend to get higher:
People with higher diastolic blood pressure tend
to have higher systolic blood pressure as well



As move to right, circles tend to get lower: People with higher total cholesterol tend to have a lower HDL/LDL ratio

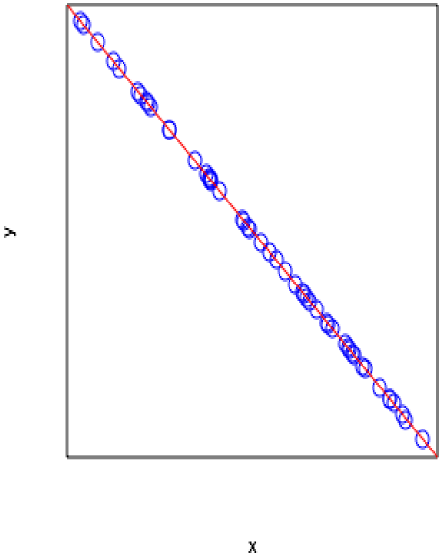
Summarizing Data
B. Continuous Data: Correlation

- Common measure of strength of linear relationship between 2 variables: the (Pearson) correlation coefficient
- To calculate correlation coefficient:
  - First subtract means from each variable
  - Then form cross product and average (but use n-1 instead of n)
  - Divide by the product of standard deviations

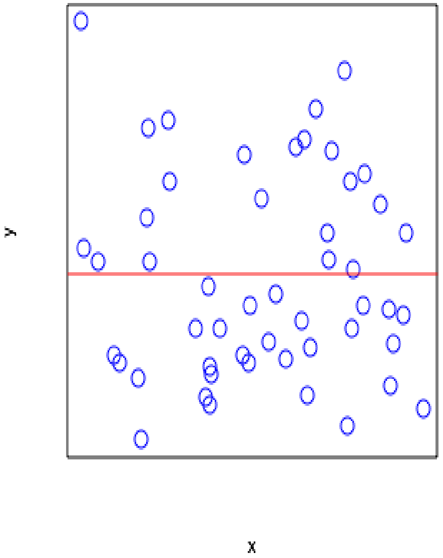
II: Summarizing Data
B. Continuous Data: Correlation

- Correlations are unit-free
  - Get the same answer whether you use feet, inches, meters, etc.
- Correlations range from -1 to 1:
  - -1 means points lie perfectly on negatively sloped line
  - +1 means points lie perfectly on positively sloped line
  - 0 means no linear relationship between the two variables (could be nonlinear relationship)

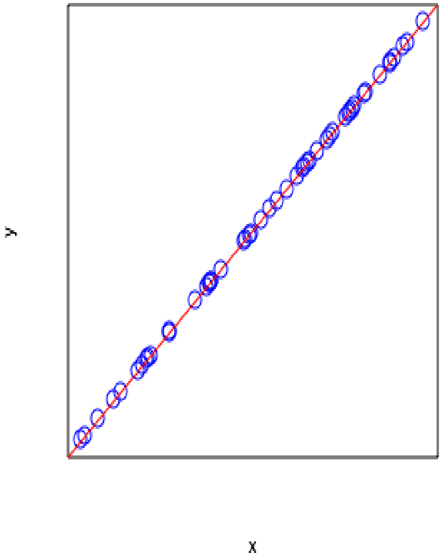
Correlation=-1



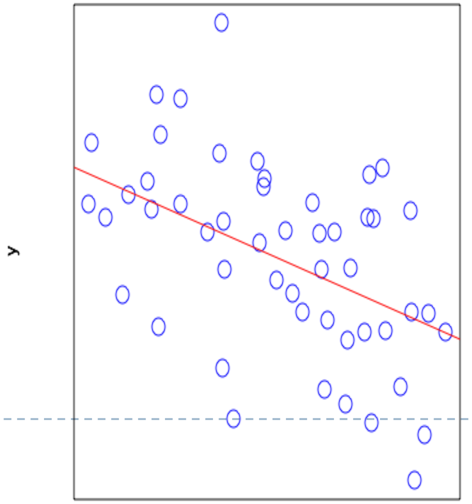
Correlation=0



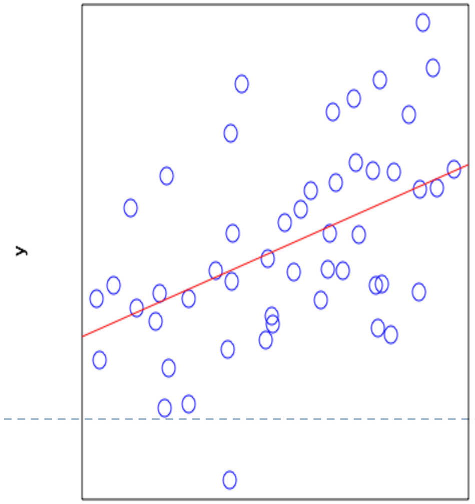
Correlation=+1



Correlation=-.5



Correlation=+.5



Quiz Question 2

I have developed a new assay and I want to compare it to the gold-standard assay. I run 100 samples on both assays, plot the results on a scatterplot and estimate the correlation. I find that the correlation is very close to 1.

This means:

A) The new assay gives the same numeric results as the old assay

B) High values on the new assay are associated with high values on the old assay, but the values won’t necessarily be the same.

C) High values on the new assay are associated with low values on the old assay

Summarizing Data
B. Continuous Data: Correlation

Notes:

- Correlation does not imply that x **causes** y
  - Maybe a third variable causes both
  - Correlation just tells you if the variables high values of one variable tend to go with high (or low) values of another variable

- Even Correlation=1 Does not tell you that the two are the same
  - If assay 1 always returns 1/2 the value of assay 2 —> correlation=1

- “Pearson’s Correlation Coefficient” (presented here) is quite sensitive to outliers, similar to the mean
  - Can use rank-based alternative (Spearman’s Correlation) analogous to median

Summary

Types of Measurements (Variables)

- Categorical (Nominal)
  - Mutually exclusive and exhaustive
  - No particular order
  - E.g. profession, Disease condition, Sex
    - Binary (dichotomous): two outcomes
      - E.g. Sex; male and female, HIV status; Positive and negative

- Ordinal
  - Categorical measurements (variables) that can be ordered
    - E.g. severity of symptoms, age ranges

- Continuous (interval)
  - Numerical variable with many possible values, no upper limit

0
No Hurt
1
Hurts
Little Bit
2
Hurts
Little More
3
Hurts
Even More
4
Hurts
Whole Lot
5
Hurts
Worst

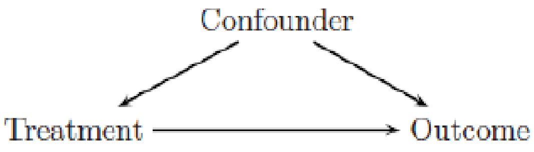
- Discrete variables e.g. 1, 2,3,4 etc
  - Has the most statistical information
    - Egg age, blood pressure, blood glucose levels, CD4, viral load etc

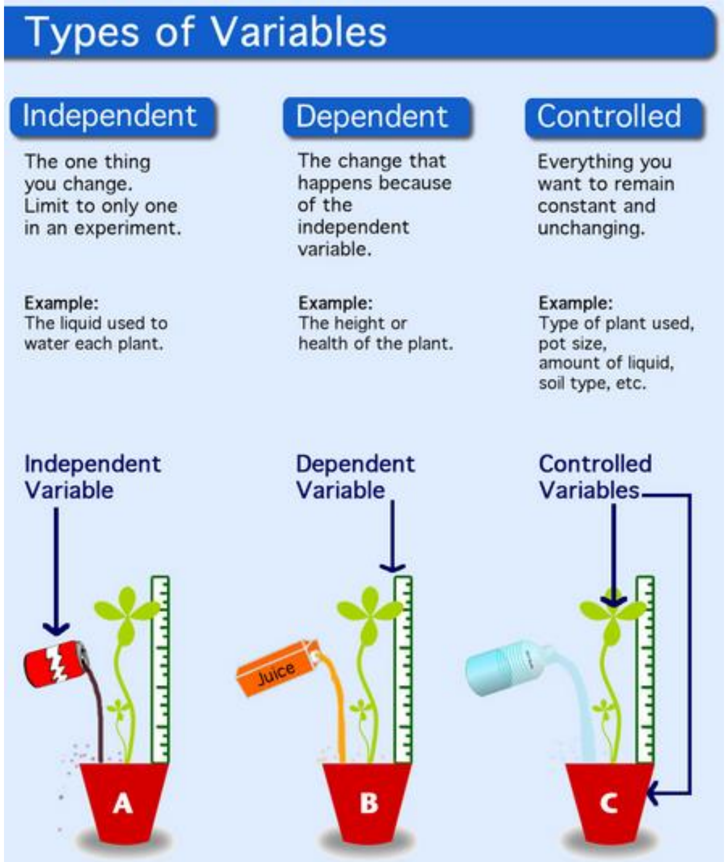
Types of Measurements (Variables)

- Dependent variable (Experimental or Response variable)
  - The variable that you measure
  - The **dependent variable** responds to the **independent variable**. It is called **dependent** because it "depends" on the **independent variable**.
  - E.g. outcomes; death, CD4, viral load, blood glucose

- Independent variable (predictor or descriptor variable)
  - Not controlled by experimenter
  - E.g. race, sex, HIV status

- Adjustment variable (cofounder)
  - A variable that can affect both the dependent and independent variable




Types of Variables

Independent

The one thing you change. Limit to only one in an experiment.

Example: The liquid used to water each plant.

Dependent

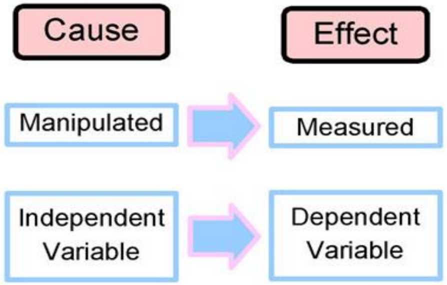
The change that happens because of the independent variable.

Example: The height or health of the plant.

Controlled

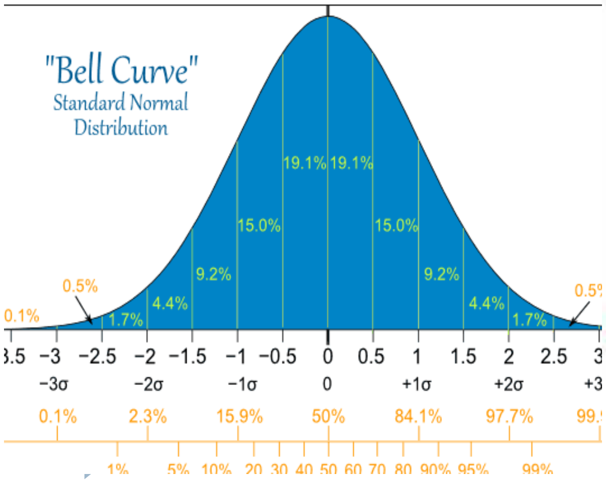
Everything you want to remain constant and unchanging.

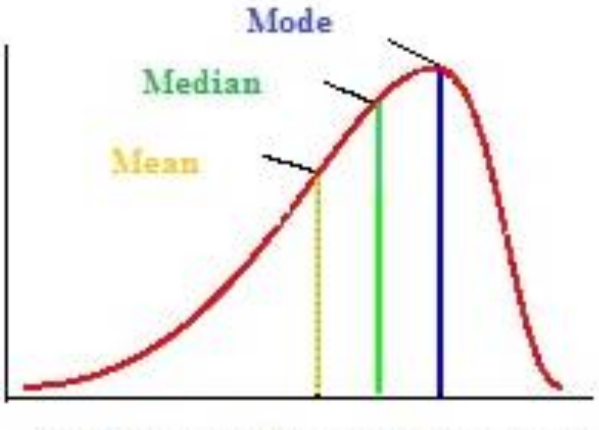
Example: Type of plant used, pot size, amount of liquid, soil type, etc.



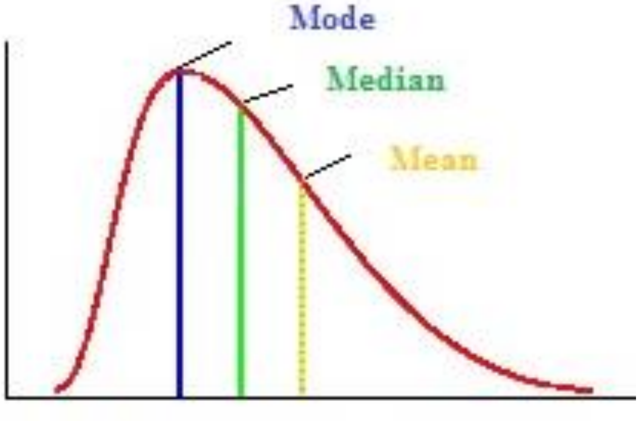
Central Tendency

- Repeated observations give us information about
  - Average or central value
  - Variation i.e. how other values are dispersed about the central value
  - The shape of the distribution





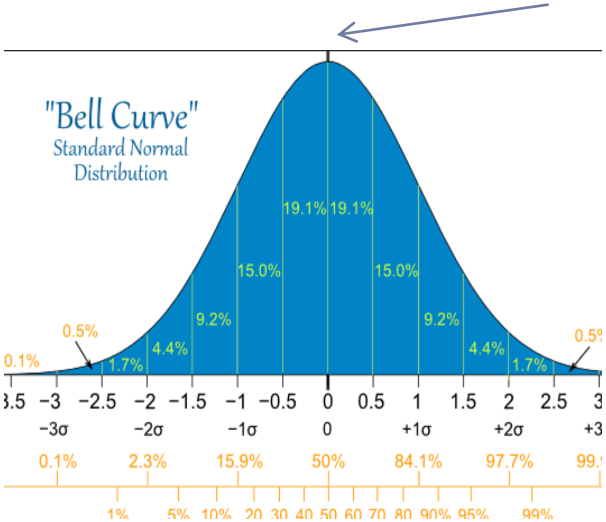
Left-Skewed (Negative Skewness)



Right-Skewed (Positive Skewness)

Measures of Central Tendency

- Mean
  - The mean (average or arithmetic mean) is the sum of all measurements divided by the number of measurements
  - Is the most commonly used measure of central tendency
  - Excel formula: = avg (highlight measurements) + enter
  - Indicates the central point (for normally distributed data)

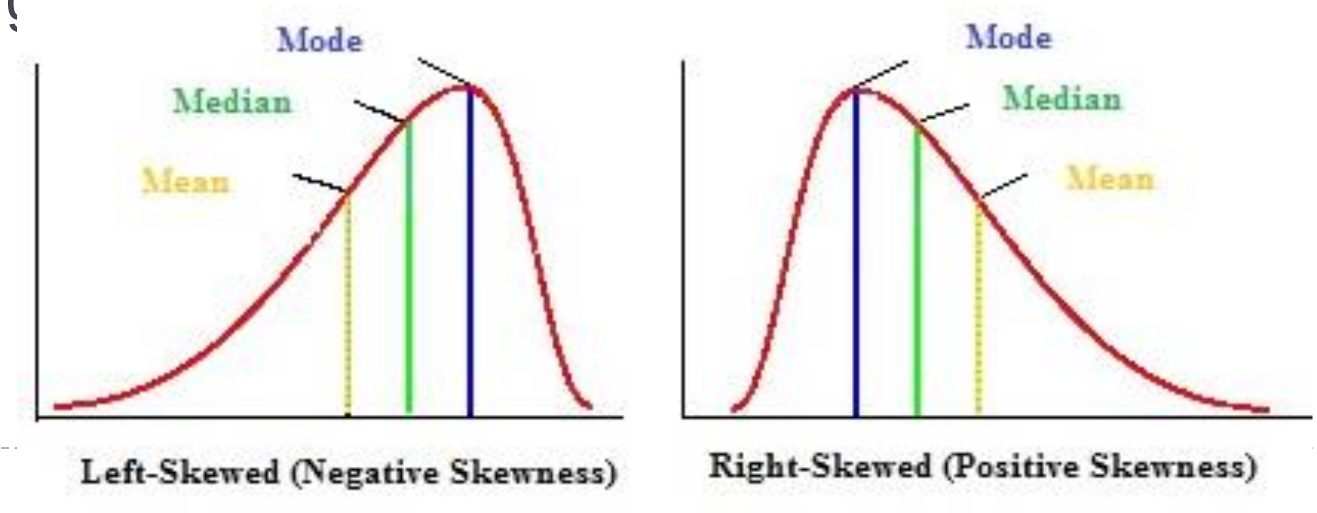


Mean/average

$$\bar{X} = \frac{\Sigma X}{n}$$

Measures of Central Tendency

- Median
  - When observations are arranged in descending or ascending order, the middle observation is known as the median
  - The median is a better indicator of central tendency for skewed data i.e. where one measurement/observation is much larger or smaller than the rest
  - E.g



Measures of Central Tendency

- Mode
  - The mode is the most frequent measurement/observation in a series.
  - Least used in medical statistics
  - E.g. 3, 5, 7, 7, 7, 8, 8, 10, 11, 12,

Measures of dispersion

- Range
  - Max, min

- Variance and standard deviation

- Interquartile range