# Types of Data, Descriptive Statistics, and Statistical Tests for Nominal Data

**Patrick Smith, Pharm.D.**
Assistant Professor
University at Buffalo
Buffalo, New York

## Types of Data

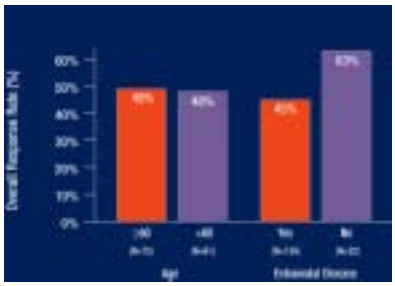
- Qualitative
  - Data are separated into categories with no inherent mathematical meaning (also called a 'categorical' variable')
    - Example: eye color, sex, race
- Quantitative
  - Data are numerical and derived from counts or measurements
    - Discrete: values are limited to integers (example: number of pets)
    - Continuous : value can take on any number (blood pressure)

## Example Dataset

| Obs | Name | Sex | Age | Height | Weight |
|---|---|---|---|---|---|
| 1 | Alfred | M | 14 | 69.0 | 112.5 |
| 2 | Alice | F | 13 | 56.5 | 84.0 |
| 3 | Barbara | F | 13 | 65.3 | 98.0 |
| 4 | Carol | F | 14 | 62.8 | 102.5 |
| 5 | Henry | M | 14 | 63.5 | 102.5 |
| 6 | James | M | 12 | 57.3 | 83.0 |
| 7 | Jane | F | 12 | 59.8 | 84.5 |
| 8 | Janet | F | 15 | 62.5 | 112.5 |
| 9 | Jeffrey | M | 13 | 62.5 | 84.0 |
| 10 | John | M | 12 | 59.0 | 99.5 |
| 11 | Joyce | F | 11 | 51.3 | 50.5 |
| 12 | Judy | F | 14 | 64.3 | 90.0 |
| 13 | Louise | F | 12 | 56.3 | 77.0 |
| 14 | Mary | F | 15 | 66.5 | 112.0 |
| 15 | Philip | M | 16 | 72.0 | 150.0 |
| 16 | Robert | M | 12 | 64.8 | 128.0 |
| 17 | Ronald | M | 15 | 67.0 | 133.0 |
| 18 | Thomas | M | 11 | 57.5 | 85.0 |
| 19 | William | M | 15 | 66.5 | 112.0 |

## Response Rate to Rituximab According to Demographic Factors

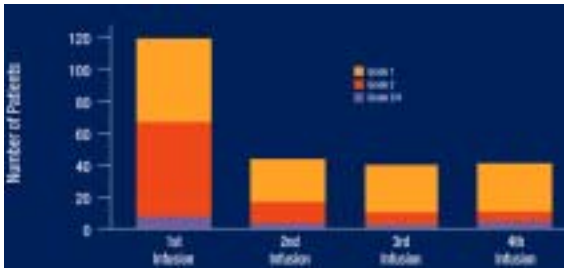
**WHAT TYPES OF VARIABLES ARE THESE ?**



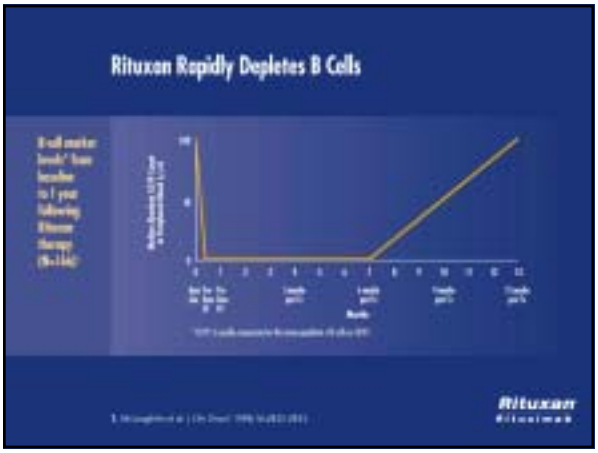
Differences not statistically significant in a multivariate analysis.
McLaughlin P, et al. *J Clin Oncol*. 1998;16:2825-2833

## Infusion-related Reactions in Multicenter Trial (N=166)

**WHAT TYPES OF VARIABLES ARE THESE ?**




McLaughlin P, et al. *J Clin Oncol*. 1998;16:2825-2833



## Transforming Variables

- Sometimes it is more convenient to treat continuous variables as categorical
- Continuous variables can be transformed to categorical variables by using "cut off points“
- For example:
  - Blood pressure can be defined as "*hypertension*" (diastolic blood pressure greater than 90 mmHg), and "*normotension*" (less than or equal to 90 mmHg)
  - Height can be converted into "*short*", *average*" or "*tall*“
  - Age can be converted to *“young”* vs. *“old”*

## Describing a Dataset

You have conducted a study and have the measurements below in 500 patients:

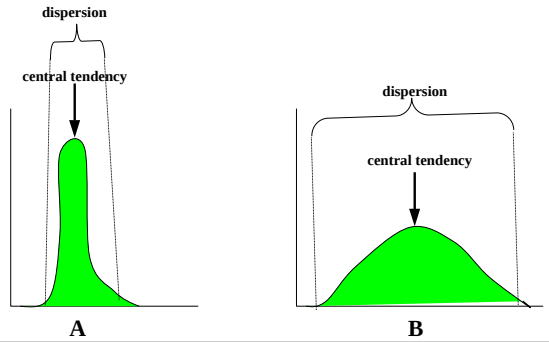
Other than a 10-page list of the dataset, how can you describe and communicate your data ?

| Obs | Name | Sex | Age | Height | Weight |
|---|---|---|---|---|---|
| 1 | Alfred | M | 14 | 69.0 | 112.5 |
| 2 | Alice | F | 13 | 56.5 | 84.0 |
| 3 | Barbara | F | 13 | 65.3 | 98.0 |
| 4 | Carol | F | 14 | 62.8 | 102.5 |
| 5 | Henry | M | 14 | 63.5 | 102.5 |
| 6 | James | M | 12 | 57.3 | 83.0 |
| 7 | Jane | F | 12 | 59.8 | 84.5 |
| 8 | Janet | F | 15 | 62.5 | 112.5 |
| 9 | Jeffrey | M | 13 | 62.5 | 84.0 |
| 10 | John | M | 12 | 59.0 | 99.5 |
| 11 | Joyce | F | 11 | 51.3 | 50.5 |
| 12 | Judy | F | 14 | 64.3 | 90.0 |
| 13 | Louise | F | 12 | 56.3 | 77.0 |
| 14 | Mary | F | 15 | 66.5 | 112.0 |
| 15 | Philip | M | 16 | 72.0 | 150.0 |
| 16 | Robert | M | 12 | 64.8 | 128.0 |
| 17 | Ronald | M | 15 | 67.0 | 133.0 |
| 18 | Thomas | M | 11 | 57.5 | 85.0 |
| 19 | William | M | 15 | 66.5 | 112.0 |

## Describing Data

- Descriptive Statistics
  - Describe and analyze a given group of data without drawing any conclusions or inferences about a larger group
  - Concerned with the presenting, organizing, and summarizing the data
- Measures of central tendency (where do most of the numbers lie? What is the most common value?)
- Measures of dispersion (spread – is everyone about the same age, or is there a wide variation in the cal plots?)

## Central Tendency and Dispersion




## Measures of Central Tendency

- Mean: called the 'average' of a set of data (arithmetic mean)
- Median: the middle value; the value corresponding to the point where 50% of the observations above and below it (when observations are arranged in numerical order)
- Mode: the most common value

## Mean

- Used for continuous data

$$\bar{x} = \left(x_1 + x_2 + \ldots x_n\right) / N$$

- Where $x_1$, $x_2$, and $x_n$ are independent values and N is the total number of observations

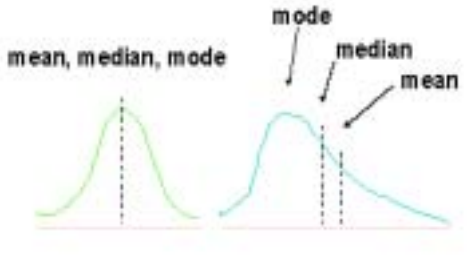
## Median

- 50% of the observations are below the median
- Example: 0,2,9,11,13 → Median = 9
- 8,8,9,10,10 → Median = _____
- 2,4,8,9 → Median = 6
- 1,2,4,87 → Median = _____

## Important Differences Between Means and Medians

- Mean: most commonly used measure, most statistical tests based on testing differences between means
  - The mean can be heavily influenced by the presence of outliers, while the median is not
    - Example dataset: 2,3,9,5,4,0,6,3,4 mean = 4; median = 4
    - If add one outlier (15), new mean = 5.1, median = 4
  - The mean is mathematically useful for making comparisons, while the median is not

## Measures of Central Tendency




Normal Distribution
Mean=median=mode

Right Skewed Distribution
Mean>median>mode

## Uses of Different Measures of Central Tendency and Dispersion

| Type of data | Measure of central tendency | Measure of dispersion |
|---|---|---|
| Categorical | Mode | |
| Discrete/Ordinal | Median<br>Mode | Range<br>IQ Range |
| Continuous | Mean<br>Median<br>Mode | SD<br>Range<br>IQ Range |

## Example 1 Mean, Median, Mode

Respiratory Rates

Group A: 11, 12, 13, 13, 14, 15

Group B: 11, 12, 13, 13, 14, 27

A: mean = 13.0; median = 13; mode = 13

B: mean = 15.0; median = 13; mode = 13

- Mean is most effected by extreme values

## Example 2 Categorical Data

A study evaluates the effect of a new cancer drug on preventing tumor growth. At the end of treatment with either the new drug (A) or the old drug (B), tumor size is measured (assume all started with the same size tumor!)

Tumor size

1 = Small
2 = Medium
3 = Large

Group A: 1, 1, 2, 2, 2, 2, 2, 3

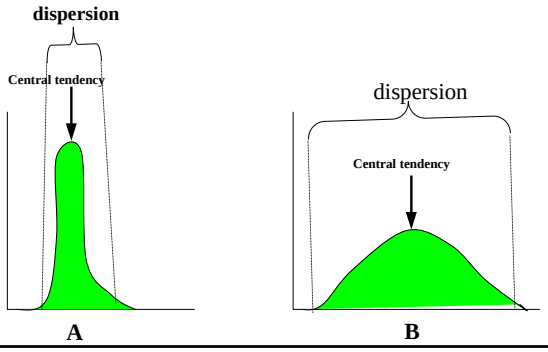
Group B: 1, 2, 2, 3, 3, 3, 3, 3

Mean = NO, does not make sense w/ categorical data
(mean for A = 1.87 → what does this mean?)

Median: A = 2, B = 3

Mode: A = 2, B = 3

## Central tendency and **Dispersion**




## Measures of Dispersion

- How spread out are the data? Is everything pretty close to the mean, or is there huge variability?
- Range**:** – difference between the highest and lowest value; also may be presented as the two extremes in the dataset (example: mean student age = 22; range 17-45)
- Variance**:** – the mean of the 'Sum of squares'
- Standard deviation (SD): the square root of the variance

## Computing Variance

- Problem: we want to compute a number to describe how far off each value in the dataset is from the mean value

1) start by calculating the difference between each observation (data point) and the mean
   - * "*Deviations from the mean*"

2) because there will be + and – numbers, we can not just add them up, so to make everything positive, square each of the values (we only care about the size of the deviation from the mean, not the direction)
   - * "Sum of squares" (SS)

3) Then, take the average* of the SS to get the variance

*the average is computed by dividing the SS by (1-n)

## Variance ($s^2$) Formula

$$s^2 = \frac{1}{1-n}\sum (x_i - \overline{x_i})^2$$

**the minimum number of observations required to compute $s^2$ is 2; thus, the correct average is calculated as (n-1) rather than n.

• Since the variance is computed from the SS, it is no longer in the same units as the original measurement (example: HR goes to $HR^2$)

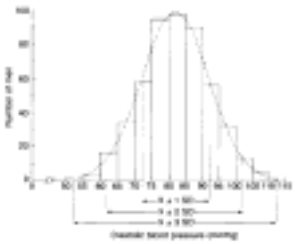
• To standardize our units to be consistent with the original units, we take the square root of the variance (Standard Deviation, s)

$$s = \sqrt{s^2}$$

## Example Calculation of Standard Deviation

- Dataset measures: 3,9,4,7
- Mean = 5.75; Median = 5.5
- SS = $(3\text{-}5.75)^2 + (9\text{-}5.75)^2 + (4\text{-}5.75)^2 + (7\text{-}5.75)^2 = 22.75$
- $S^2 = 22.75/(n\text{-}1) = 22.75/3 = 7.58$
- SD = s = $(7.58)^{1/2} = 2.75$

## Standard Deviation

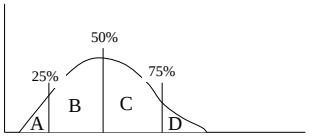


$SD^2$ = variance

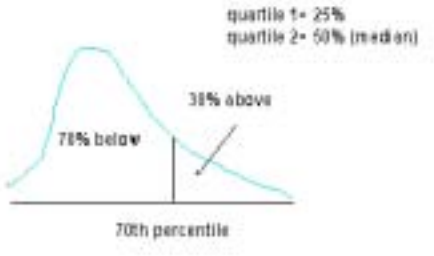
+/- 1 SD = 68% of all data
+/- 2 SD = 95%
+/- 3 SD = 99%

## Measures of Dispersion

- Percentiles/Quartiles
  - Also called the inter-quartile range
  - Divide the distribution of the data into four parts, corresponding to 25%, 50%, and 75% of the distribution
  - Not useful for small datasets



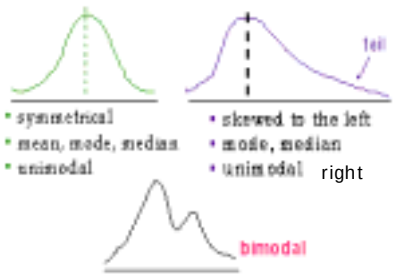

## Percentiles




## Distributions of Data

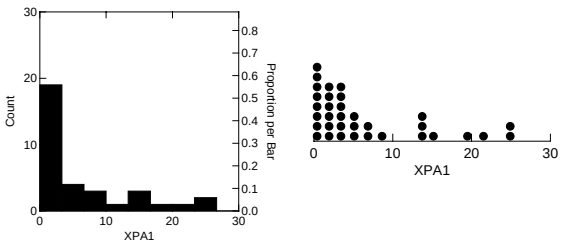
- Define – The shape that the dataset takes on when plotted on a histogram
- Types of Distributions:
- Normal (‘bell shaped curve’)
- Log-Normal – follows a normal distribution after data converted to a log scale
- Skewed
  - Right vs. left
- Unimodal/Bimodal

## Normal and Skewed Distributions




## GRAPHICAL PLOTS

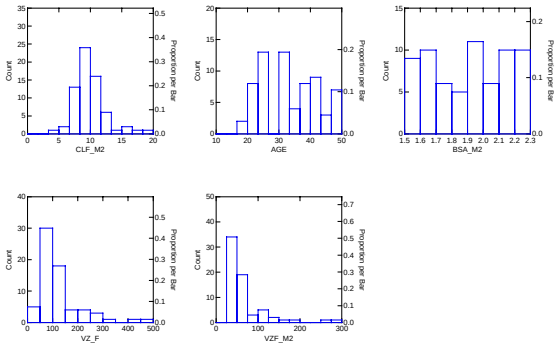
## Histograms and Dot Density Plots to Summarize Data



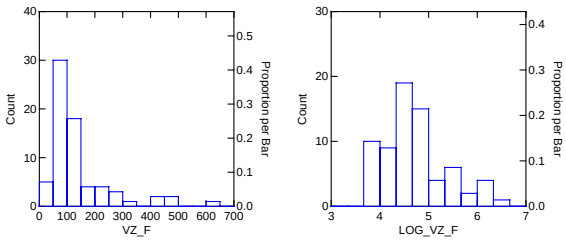
## Stem and Leaf Plots Provide Actual Data

```
0    000001111
 0   22222223333
 0   455
 0   67
 0   8
 1
 1   333
 1   5
 1   9
 2   144
-----------------------------
```

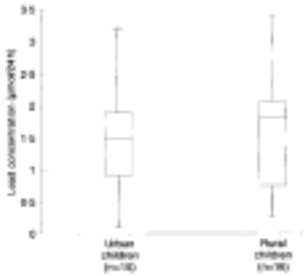
## Example Histograms



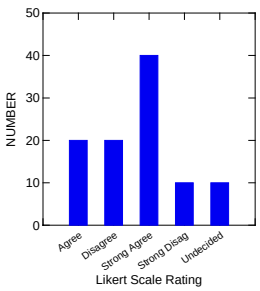

## Log Transformation to Approximate a Normal Distribution




## Box-whisker Plot of Data



## Graphic Display of Categorical Data




**Now that we have described the data, statistics can help us determine, scientifically/mathematically, if there is a real difference between two datasets**

**– or, did an apparent difference occur by chance alone?**

## Inferential Statistics

- Used to determine the probability or likelihood that conclusions based on data from a sample are true, and not due to chance
- The generalization of a conclusion from a set of sample data to a larger group of subjects is strengthened by the method of random sampling

## Random Sampling

- Each member of the population has an equal chance or probability of being selected or included in the sample
- If the sample size is large enough and truly random, should have an excellent description of the larger population in which you are really interested

For more on sampling, see Bland Ch. 3, pages 27-32

## Random Sample vs. Population

| | **Population Parameter** | **Sample Statistic** |
|---|---|---|
| **Measure** | | |
| Mean | $\mu$ (mu) | $\overline{x}$ |
| Standard deviation | $\sigma$ (sigma) | s |
| Variance | $\sigma^2$ (sigma squared) | $s^2$ |

Xbar, s, $s^2$ are computed from a random sample and are estimates of the overall population parameters – in reality it is very difficult to measure the population parameters

## Standard Error of the Mean (SE)

- A measure of variability of the mean of the sample as an estimate of the true value of the population mean
- Quantifies how accurately the true population mean is known
- The larger n, the smaller the SE

$$SE = \frac{SD}{\sqrt{\text{sample size}}}$$

As n → infinity, SE → 0

## Hypothesis Testing

- Used to decide if the results of the *sample* experiment support a hypothesis regarding whole *population*
- $H_0$ = Null hypothesis
  - The hypothesis assuming no effect of a treatment
  - Atorvastatin does not effect HDL levels
- $H_A$ = Alternative hypothesis
  - The alternative to the null hypothesis
  - Atorvastatin does effect HDL levels

## Hypothesis Testing Matrix

| | **$H_0$ True (no difference)** | **$H_0$ False (difference exists)** |
|---|---|---|
| **Reject $H_0$** | Type I Error (also called $\alpha$ error) | Correct Decision |
| **Retain $H_0$** | Correct Decision | Type II Error (also called $\beta$ error) |

Type I error: the error of calling something significant when it is not significant
Type II error: the error of failing to detect significance when it is present

## Probability Value: p value

- Likelihood of an observed result or statistic occurring on the basis of the sampling distribution
- p value can be compared to the $\alpha$ value (usually 0.05)
  - If the p value < $\alpha$ value, then we say that a result is "statistically significant"
- p value can be obtained from the computer or by looking up a value in a table

## Significance Level

- Sample data must differ substantially from what would be expected from the sampling distribution before we can reject $H_0$ and claim accept $H_A$
- Observing results that are so unusual that they could have only occurred by chance 5% of the time
- Significance level of 5% ($\alpha = 0.05$) is established before data is collected

## Power

- The power of a statistical hypothesis test measures the test's ability to reject the null hypothesis when it is actually false - that is, to make a correct decision
- The power of a hypothesis test is the probability of not committing a type II error - t is usually expressed as:
  - Power = 1 – probability of a type II error = 1 - $\beta$
- The maximum power a test can have is 1, the minimum is 0 (ideally we want a test to have high power, close to 1)
- Power is directly related to sample size; the larger the sample size the less likely you are to accept the null hypothesis when it is actually true

## Confidence Intervals

- Often expressed as 95% confidence intervals (CI)
- CIs give us a range of values that represent how likely the sample mean represents the population (a measure of precision of the estimate)
- Results from a sample population with a wider range of values will have broader CIs; increasing n will narrow CI by reducing SE
- There is a 95% probability that the true mean is within ~2 SEs of the sample mean (or, more precisely 1.96)

$$95\%\,\mathrm{CI} = \text{sample mean} \pm 1.96 \times \mathrm{SE}$$

## Confidence Intervals

- May be used as an alternative to computing a p value
- If you have a sample and want to see if a statistically significant difference exists, compute the 95% CI for each
- If the CIs DO NOT overlap, there is less than a 5% chance that the observed difference was due to chance alone

## Confidence Intervals - Example

- Gatifloxacin and hyperglycemia
  - → What is the Ho and HA ?
- Placebo Group : Mean (SE) = 140 (4.7)
- FQ Group: Mean (SE) = 162 (10.2)
- 95% CI: 140 +/- 1.96*4.7
- (130.8 – 149.2)
- (142.0 – 182.0)
- → the 95% CI overlap, we would not reject the null hypothesis

## Statistical Significance vs. Clinical Significance

- Statistical significance: whether or not results of statistical test meet an accepted criterion level (i.e., $p < 0.05$)
  - Impacted by magnitude of effect and sample size
  - Even a trivial difference can have a low p value if the sample size is large enough
- Clinical significance is a matter of judgment (i.e., is the difference between groups of large enough importance from a clinical standpoint)
  - Studies can achieve statistical significance without being clinically significant