

Foundations and Basics of Statistical Tests and Data Analysis

Lance Ford, PhD

Assistant Professor of Research, Biostatistics

Department of Biostatistics and Epidemiology

Hudson College of Public Health

Lance-Ford@ouhsc.edu

Invited COBRE Lecture

September 30, 2022

Objectives

1. Introduce the foundations of hypothesis testing and statistical inference
2. Describe types of data
3. Review most common statistical methods seen in the literature
4. Discuss how to select proper statistical test for different applications

What is the purpose of statistics?

Sample vs. Population

- Population describes the hypothetical (and usually) large number of people to whom you wish to generalize
- Sample describes those individuals who are in the study (fraction of the population)
 - The study is only generalizable to the type of patients who are in the study

Example: if our sample contains only adults, then can't necessarily generalize results to population of children

Hypothesis Testing

Hypotheses

- Null hypothesis: H_0
 - Typically, a statement of no treatment effect
 - Assumed true until evidence suggests otherwise
 - Example: H_0 : No difference in mean DBP between treatment groups
- Alternative: H_A
 - Reject null hypothesis in favor of alternative hypothesis
 - Often two-sided
 - Example: H_A : mean DBP differs between treatment groups

Type I and Type II Errors

- Errors associated with hypothesis testing:

Truth

Concluded *Association* *No Association*

Association

Correct!

True positive

Power ($1 - \beta$)

False positive

Type I error

Alpha (α)

No Association

False negative

Type II error

Beta (β)

Correct!

True negative

$1 - \alpha$

Sensitivity and Specificity

- How Type I and Type II errors relate to sensitivity and specificity

Gold Standard

Test

Association

No Association

Association

True positive

False positive

a

b

No Association

False negative

True negative

c

d

$$\text{Sensitivity} = \frac{a}{a+c}$$

$$\text{Specificity} = \frac{b}{b+d}$$

Example: Type I and Type II Errors

Truth




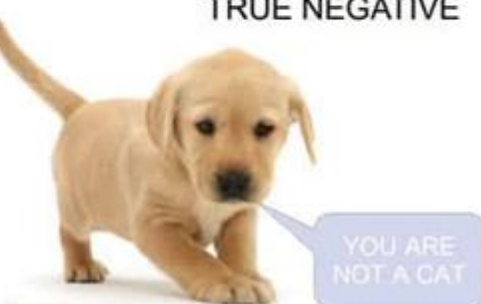
Concluded

Cat

Not a Cat

Cat

Not a Cat

<p>TRUE POSITIVE</p>  <p>YOU ARE A CAT</p>	<p>FALSE POSITIVE</p>  <p>YOU ARE A CAT</p> <p>TYPE I ERROR</p>
<p>FALSE NEGATIVE</p>  <p>YOU ARE A DOG</p> <p>TYPE II ERROR</p>	<p>TRUE NEGATIVE</p>  <p>YOU ARE NOT A CAT</p>

Significance Level

- Significance level: α
 - Probability of a Type I error
 - Probability of a false positive
 - Example: If the effect on DBP of the treatments do not differ, what is the probability of incorrectly concluding that there is a difference between the treatments?
 - Typically chosen to be 5%, or 0.05

Statistical Power

- Power: $1 - \beta$ ($1 - \beta$)
 - Probability of detecting a true treatment effect
 - Power = (1 - probability of a false negative)
 - = (1 - probability of Type II error)
 - = $(1 - \beta)$ = probability of a true positive
 - Example: If the effects of the treatments do differ, what is the probability of detecting such a difference?
 - Typically chosen to be 80-99%

P-value

- The probability of obtaining a difference at least as extreme as that obtained, provided the two groups are really equal (null hypothesis is true)
- Conditional probability that is calculated assuming the null hypothesis is true

Statistical Significance

- **Statistical Significance**: If the p-value of the calculated statistic is less than the alpha set in advance by the researcher (usually 0.05), then we can conclude the groups are different.
- $P \text{ value} \leq \alpha$ implies statistical significance.
- Statistical significance does not necessarily mean clinical significance

Clinical Significance

- **Clinical Significance**: Smallest clinically important difference between two treatments
- Based on clinical judgment of the magnitude of the difference
- Clinician should consider the side effects, long-term complications, and other costs of the two treatments

Example: Clinical vs. Statistical Significance

Consider a randomized comparison between 2 treatments for cholesterol after 1 year of therapy

- Treatment 1 mean LDL: 134 mg/dL
 - Treatment 2 mean LDL: 132 mg/dL

 - Mean difference of 2 mg/dL, p-value = 0.036
- 1) Is there a statistically significant difference between the two groups?
 - a) What was the null hypothesis? What was the alternative hypothesis?
 - b) What is your conclusion?

 - 2) Is there a clinically significant difference between the two groups?
 - a) What factors do you need to determine when deciding?

Clinical vs. Statistical Significance

- Not all statistically significant differences are clinically significant!
- For an effect to be clinically significant (conclusive), the estimate should be statistically different
- Confidence intervals can address both clinical and statistical significance

Confidence Intervals and Estimation

Confidence Interval

An interval estimate consisting of a range of values (with a lower and upper bound) constructed to have a specific probability (the confidence) of including the population parameter with repeated sampling.

Example 1

▪ Experimental Event Rate (EER):
480/800 patients (60%)

▪ Control Event Rate (CER):
416/800 patients (52%)

▪ Absolute Benefit Increase (ABI):
 $|60\% - 52\%| = 8\%$

95% CI: 3% to 13%

p-value=0.001

Is the difference statistically significant?

- p-value $< \alpha$ (0.05)
- 95% CI doesn't contain 0 (the null value)

Is the difference clinically significant?

- Smallest clinically important difference is 15%
- Not clinically significant (CI $< 15\%$)

Braitman LE . Confidence Intervals Assess Both Clinical Significance and Statistical Significance. Annals of Internal Medicine, 1991;114: 515-517.

Relation between the CI and Sample Size

The width of the confidence interval reflects the precision of our estimation.

Holding the mean, alpha, and standard deviation constant:

- Increasing the sample size decreases the width of the confidence interval (tighter CI), and vice versa

Why? The larger the sample the more confident we are of the point estimate (mean, proportion, etc.)

Example 2

Experimental (EER): 15/25 patients (60%)

Control (CER): 13/25 patients (52%)

ABI: $|60\% - 52\%| = 8\%$

95% CI: -19% to 35%

$p = 0.57$

Not statistically significant ($p\text{-value} > \alpha$)

Clinical significance inconclusive

Example 3

Experimental (EER): 15/25 patients (60%)

Control (CER): 8/25 patients (32%)

ABI: $|60\% - 32\%| = 28\%$

95% CI: 1.5% to 54.5%

$p=0.047$

Statistically significant ($p\text{-value} < \alpha$)

Clinical significance inconclusive (need larger n)

Recommendations/Interpretation of CIs

- Use 95% confidence intervals (corresponds with an α of .05)
- If the 95% CI of the difference between two groups *includes* the value of 0, then the result is not statistically significant at the .05 level
 - If 0 is outside the 95% CI, we have statistical significance
- For an odds ratio, if 95% CI of the ratio includes the value of 1, then the result is not statistically significant at the .05 level, and vice versa
 - If 1 is outside the 95% CI, we have statistical significance

Test your knowledge!

Researchers measure cholesterol levels in a sample of patients in New Zealand and Asia and find the following results:

Region	Sample Size	Mean Cholesterol	Standard Deviation
New Zealand	100	5.4	1.2
Asia	150	4.9	1.3

They calculate the mean difference and 95% CI for the true difference in mean cholesterol levels between the 2 populations and find: mean difference 0.5 mmol/L and **95% CI: (0.18-0.82)**.

The 95% CI for the true difference in mean cholesterol levels between 2 populations suggests that:

- There is no statistically significant difference in mean cholesterol levels between the 2 populations.
- There is a statistically significant higher mean cholesterol level in the Asian population as compared to the New Zealand population.
- There is a statistically significant higher mean cholesterol level in the New Zealand sample as compared to the Asian sample.
- There is a statistically significant higher mean cholesterol level in the New Zealand population as compared to the Asian population.

Estimation (CIs) vs. Hypothesis Testing (p-values)

- Significant p-value not the same as clinical significance
- P-value may not give all information needed for interpretation: how large is effect?
- Confidence intervals give more information
- Some authors display both p-values and CIs

Types of Data

Data

- The raw material of statistics
- The recordings of the measurements taken on characteristics

We can think of them as:

- numbers that result from the taking of a measurement
- numbers that result from the process of counting
- categorical levels

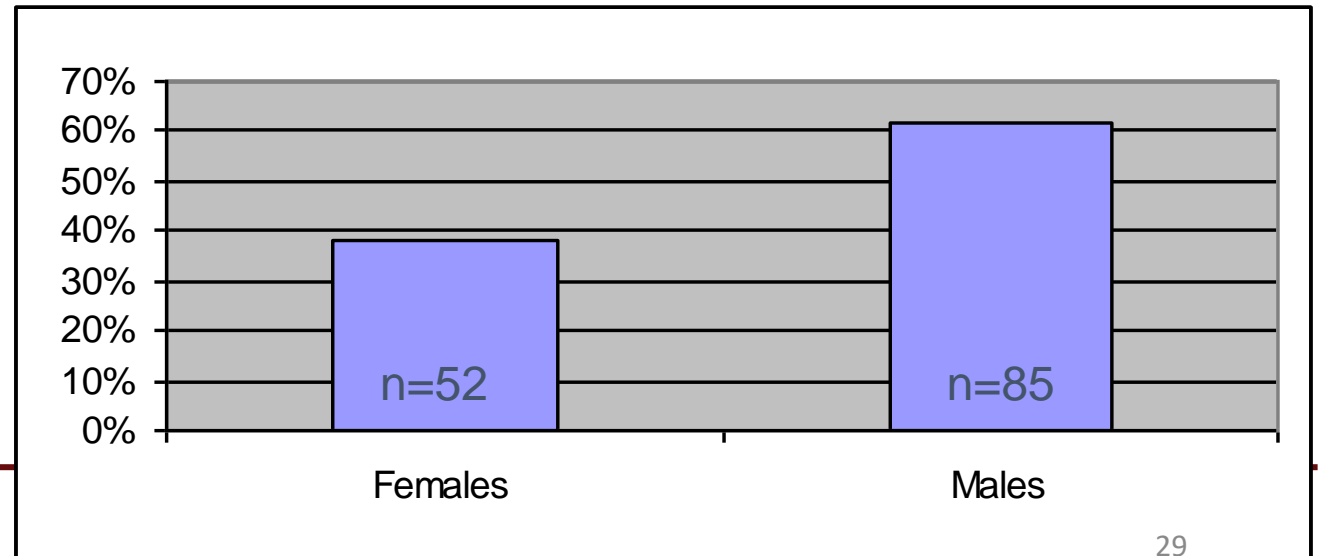
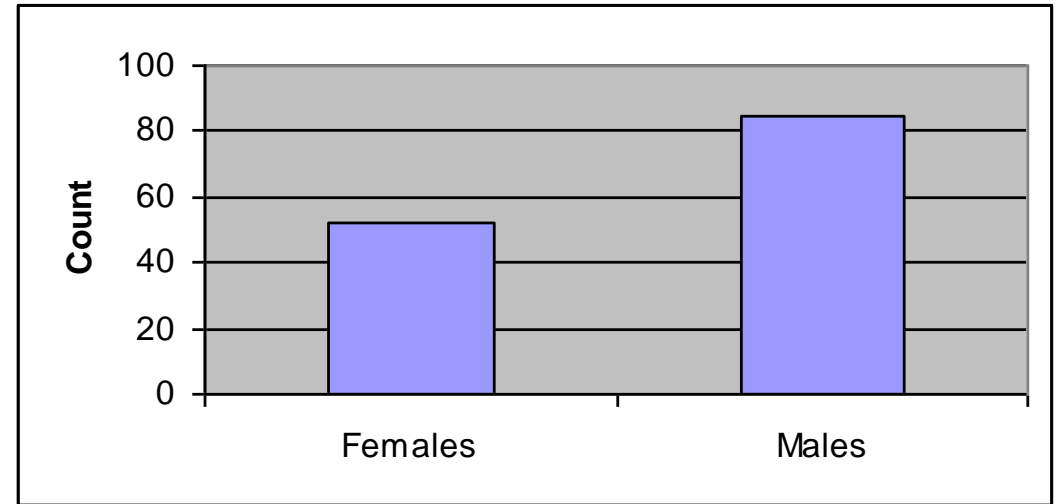
Qualitative / Categorical Variables

- Organized: observations that have the same attributes are in the same category.
- We can count the number of persons, places, or things belonging to various categories.
 - Example: Smoking status (never, quit, current); educational level (<HS, HS, >HS)

Dichotomous variable: categorical variable that can take on only two values, such as dead/alive, disease/not-diseased, or yes/no.

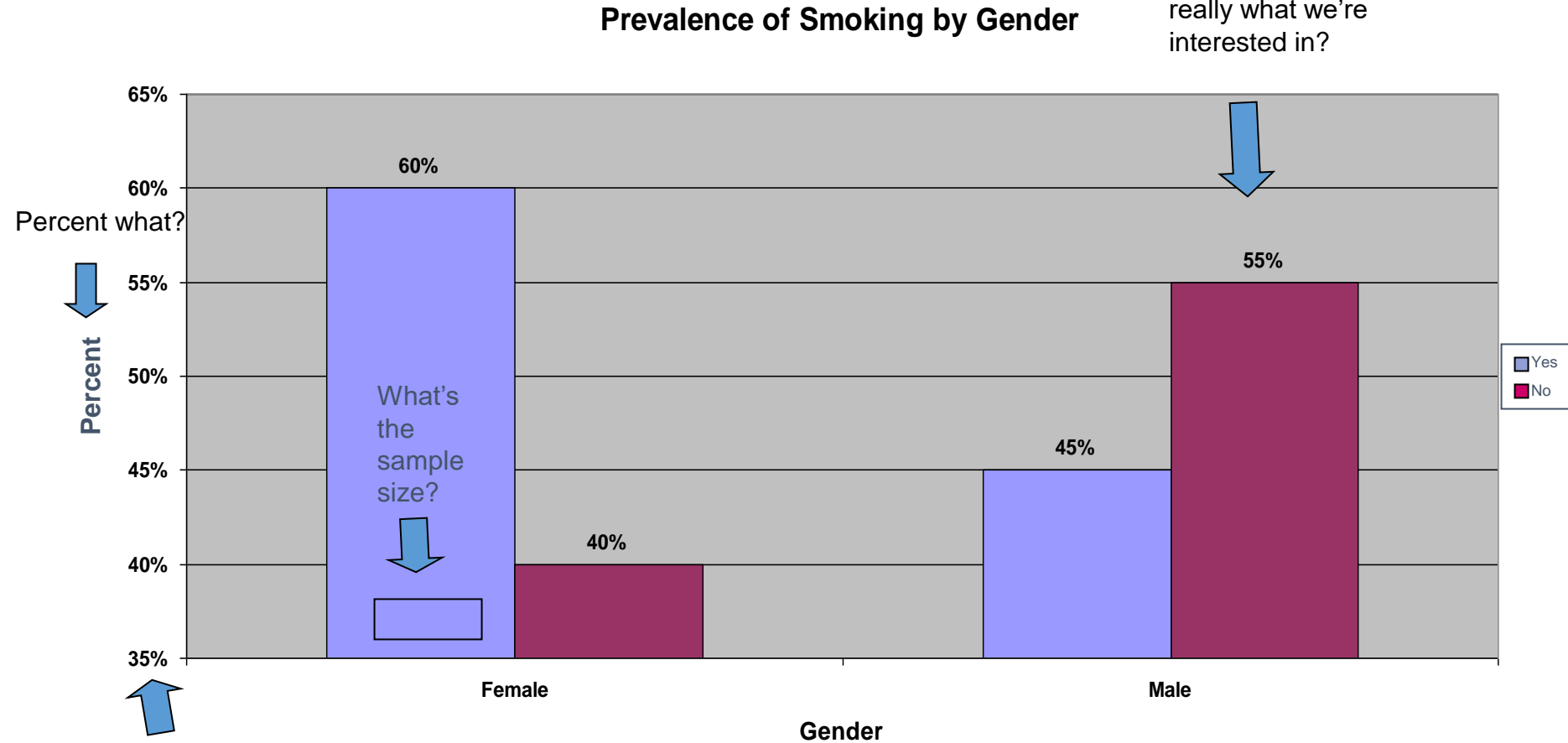
Displaying Categorical Data

Gender	Frequency (count)	Relative Frequency
Female	52	38% (52/137)
Male	85	62% (85/137)
Total	137	100% (137/137)



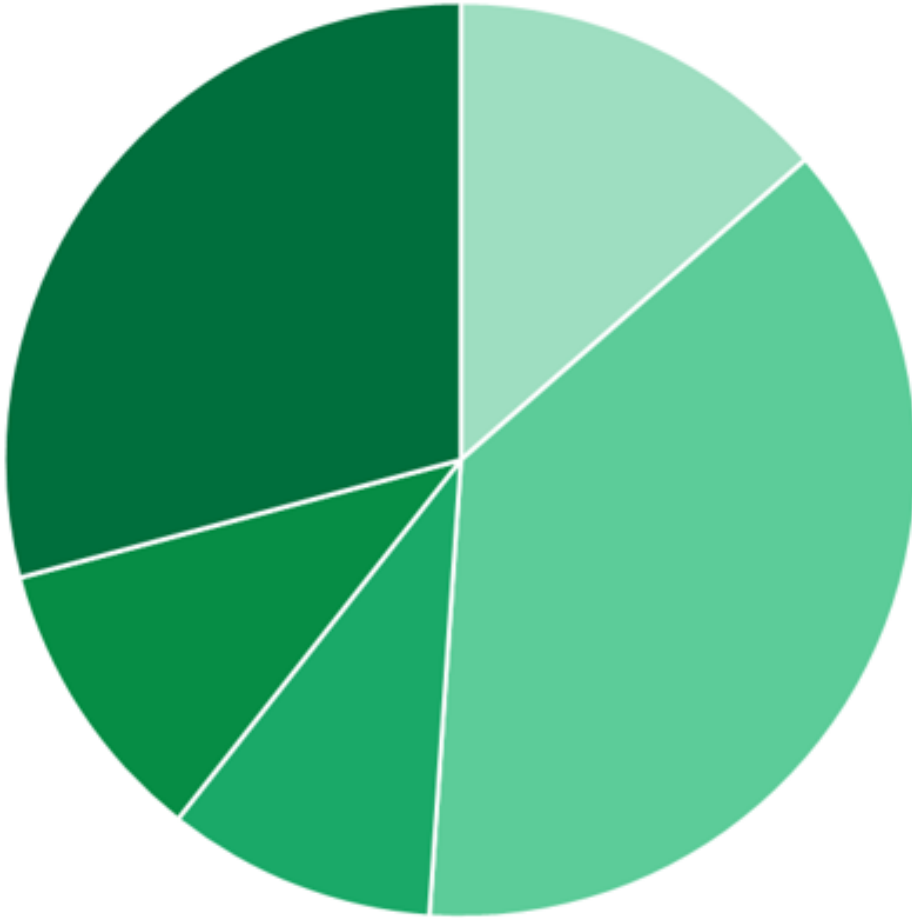
Bad Graph!

Do the red bars give us any extra information? Is it really what we're interested in?



Y-axis doesn't start at zero. This distorts the graph and makes it seem that there are more than double the percent of female smokers than male smokers.

Pie Charts vs. Donut Charts



Quantitative Variables

Discrete:

A variable characterized by gaps or interruptions in the values that it can assume. Typically, the values are “countable”.

Example: the number of missing teeth, number of children per family, number of patients seen in one day

Quantitative Variables

Continuous:

- Variable that does not possess gaps or the interruptions characteristic of a discrete random variable.
- It can assume any value within a specified relevant interval of values.

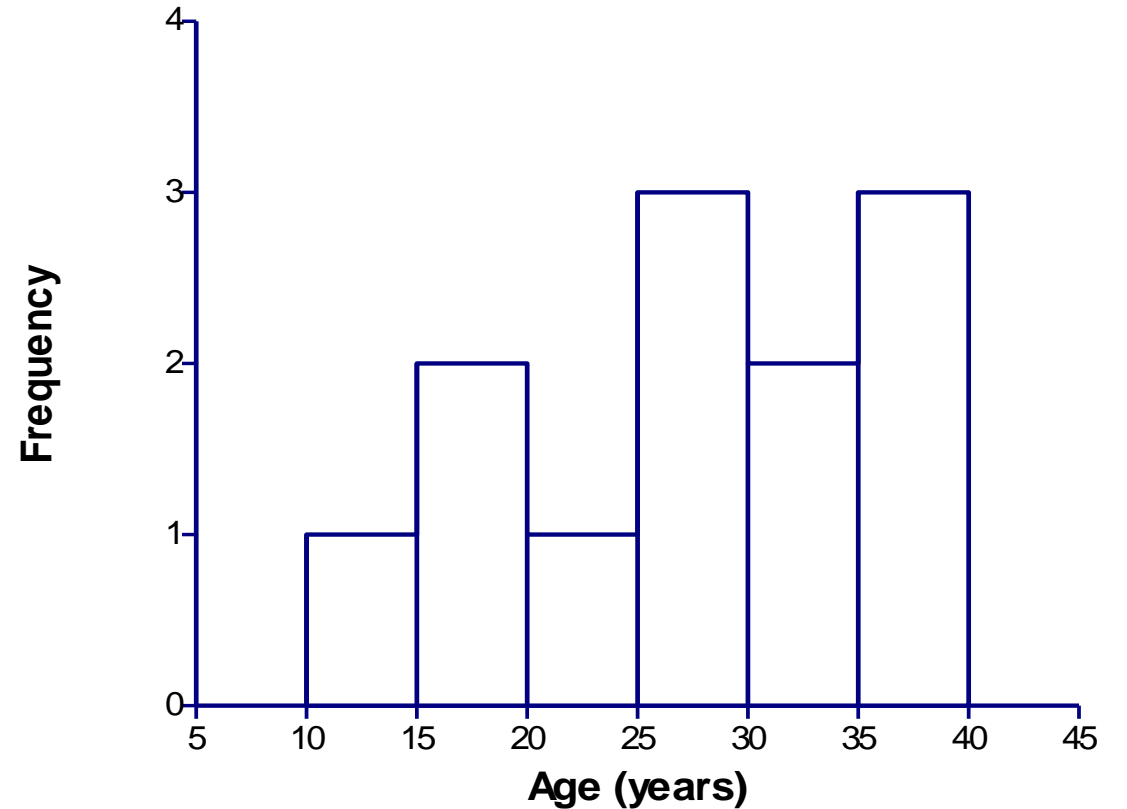
Example, height- no matter how close together the observed heights of 2 people are, we can theoretically find a person whose height falls somewhere in between.

Displaying Quantitative Variables

Tabular Display:

Class interval (years)	Frequency	Relative Frequency
10-14	1	8%
15-19	2	17%
20-24	1	8%
25-29	3	25%
30-34	2	17%
35-39	3	25%
Total	12	100%

Graphical Display:



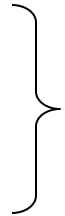
Levels of Measurement

- **NOMINAL SCALE**
- **ORDINAL SCALE**



Qualitative

- **INTERVAL SCALE**
- **RATIO SCALE**



Quantitative

Nominal Scale

- Consists of "naming" observations or classifying them into various mutually exclusive and collectively exhaustive categories

- Lowest measurement scale

Examples: eye color, zip code, blood type

Ordinal Scale

- Consists of qualitative observations that are not only different from category to category but can be ranked according to some criterion

Examples: Income (low, med, high); education level (<HS, HS, >HS); Likert scale (“extremely dislike”, “dislike”, “neutral”, “like”, “extremely like”)

Interval Scale

- If it is not only possible to order measurements, but also the distance between any two measurements is known
- Truly quantitative
- No true zero -- zero point is arbitrary

Examples: BMI, credit score, others?

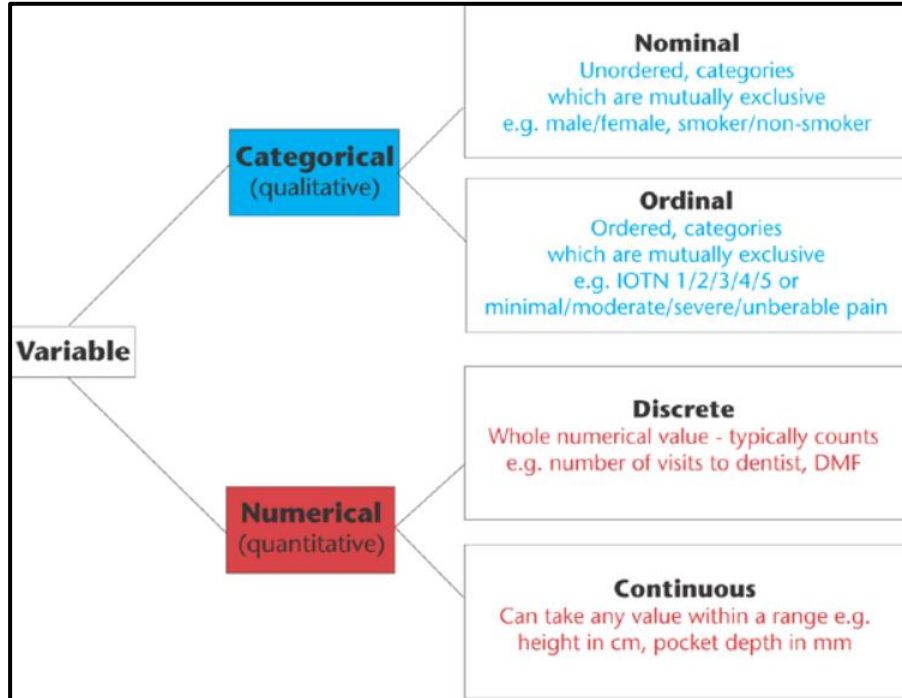
Ratio Scale

- This scale is characterized by the fact that equality of ratios as well as equality of intervals may be determined.
- Fundamental to the ratio scale is a true zero point.
- Zero point represents absolute absence of the characteristic being measured (statements such as that one number is twice as much as the other number make sense)

- Highest level of measurement

Examples: Dose amount, flow rate, pulse

Summarize Levels of Measurement

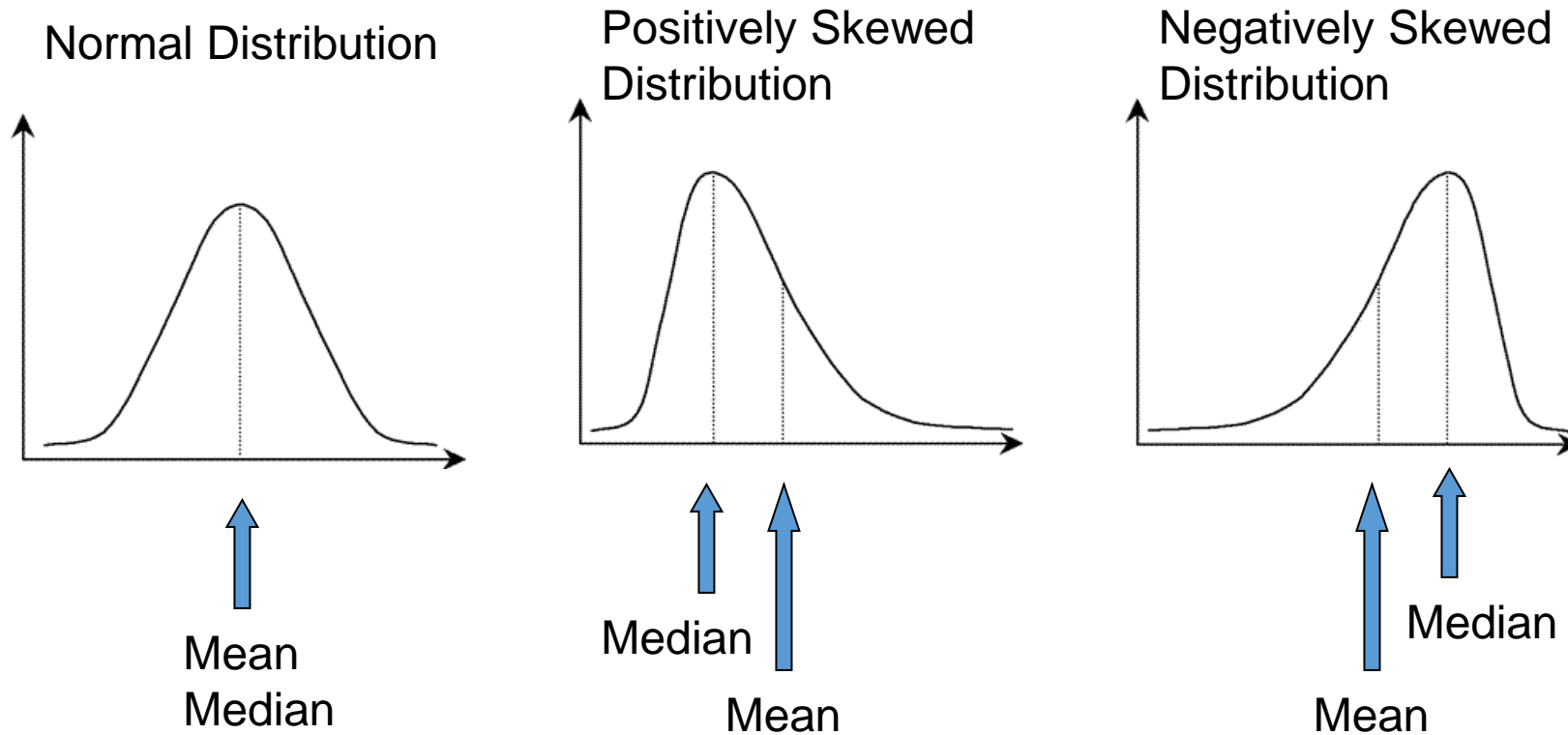


OK to compute....	Nominal	Ordinal	Interval	Ratio
Frequency distribution	Yes	Yes	Yes	Yes
Median and percentiles	No	Yes	Yes	Yes
Add or subtract	No	No	Yes	Yes
Mean, standard deviation, standard error of the mean	No	No	Yes	Yes
Ratios, coefficient of variation	No	No	No	Yes

Summary Statistics for Quantitative Data

- *Central tendency of a distribution* is the score near the center of the distribution. It is a typical and representative score value.
 - A single value that is considered typical of the set of data as a whole
 - 2 most commonly used measures:
 - Mean
 - Median
- *Variability* is the degree to which the measurements in a distribution differ from one another.
 - Often called:
 - Variation
 - Spread
 - Scatter

Measures of Central Tendency



Measures of Dispersion - Range

- One way to measure the variation in a set of values is to compute the range
- Difference between smallest and largest values;
Range = Largest - Smallest
- Usefulness of the range is limited since it only takes into account 2 values. It is a poor measure of dispersion.
- Main advantage is the simplicity of computation

Measures of Dispersion - Variance

- When values lie close to their mean, the dispersion is less than when they are scattered over a wide range.
- The **variance** allows us to measure the dispersion relative to the scatter of the values about their mean.
- Standard deviation = $\sqrt{\text{variance}}$

Recommendations for reporting central tendency and variability...

- If your data is (roughly) normally distributed, report:

Mean and Standard Deviation (SD)

- If your data isn't normally distributed, report:

Median and Interquartile Range (IQR)

Where $IQR = Q3 - Q1$ (the 3rd quartile minus the 1st quartile)

Test your knowledge!

A study wishes to assess birth characteristics in a population. For the following variables, describe the appropriate measurement scale or type:

- A. discrete
 - B. continuous
 - C. ordinal
 - D. nominal
 - E. dichotomous
-
- a. _____ Birthweight in grams
 - b. _____ Birthweight classified as low or high
 - c. _____ Type of delivery classified as cesarean, natural, induced

Selecting Proper Statistical Tests for Different Applications

&

Common Statistical Methods in the Literature

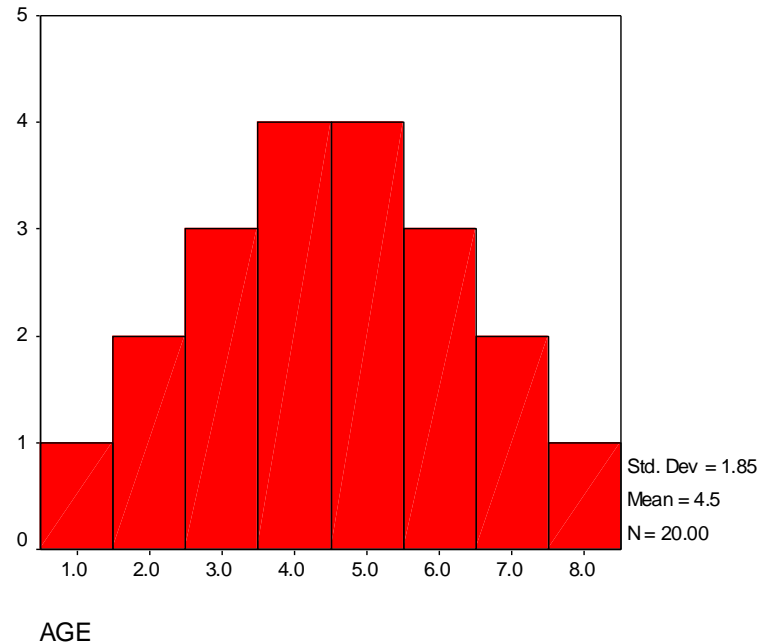
Objective #3 Review most common statistical methods seen in the literature

Objective #4 Discuss how to select proper statistical test for different applications

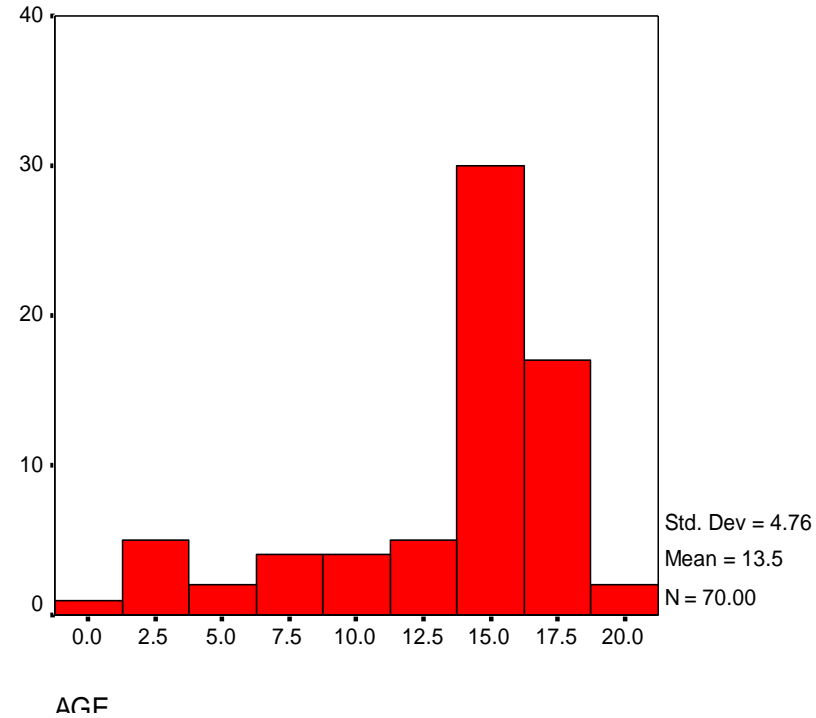
Choosing the correct statistical method...

- Is the outcome categorical, quantitative, or a survival distribution?
- If the outcome is quantitative, is it normally distributed?
- How many study groups do you have?
- Are the study groups independent or dependent (matched or repeated)?

Assessing the distribution



Approximately normal
distribution



Non-normal/skewed
distribution

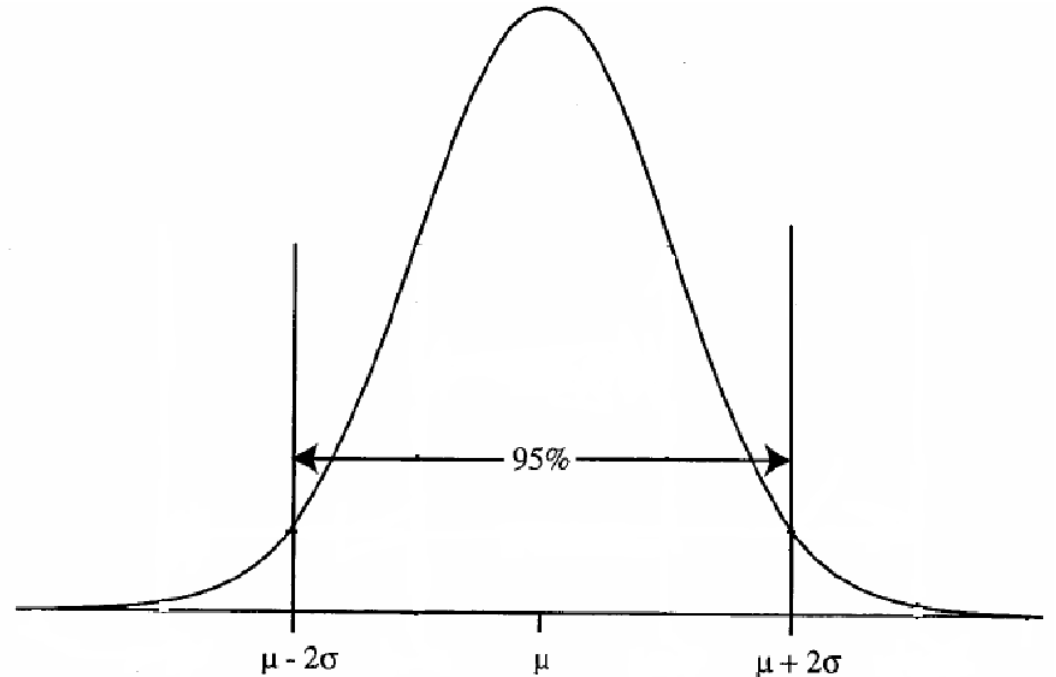
The Normal Distribution

Characteristics of the normal distribution:

1. Symmetrical about its mean μ (mirror image)
2. Mean, median, and mode are all equal (bell shaped)
3. The area between -1 standard deviation and +1 standard deviation from the mean is approximately 68% of total area under the curve
 - ± 2 standard deviation is about 95% of the total area under the curve
 - ± 3 standard deviation is about 99.7% of the total area under the curve

The Normal Distribution

So, 95% of our data values are between -2 and 2 σ 's from the mean



- Therefore, it's common practice that if we want to evaluate whether a value is unusually high or low, we usually see if it is within ± 2 standard deviations from the mean

Approximately Normal Distribution or Large Sample Size (>30)

- T-test
 - Compare 2 independent means
- Paired t-test
 - Compare 2 dependent means
- ANOVA
 - Compare 3 or more independent means
- Repeated Measures ANOVA
 - Compare 3 or more dependent means

Example: T-test used in publications

Table 1—Clinical characteristics and baseline laboratory data of the control and intervention groups

Characteristics	Control group	Intervention group	P
n	55	55	
Age (years)	54.7 ± 9.4	53.5 ± 8.8	0.507
Sex (M/F)	32/18	35/16	0.623
BMI (kg/m ²)	23.9 ± 3.1	24.4 ± 3.4	0.493
Diabetes duration (years)	6.6 ± 5.7	7.0 ± 6.3	0.751
Diagnosis of hypertension (n)	13	17	0.420
Systolic blood pressure (mmHg)	128.5 ± 17.0	124.7 ± 15.8	0.999
Diastolic blood pressure (mmHg)	77.0 ± 9.7	77.5 ± 8.7	0.254
HbA _{1c} (%)	7.19 ± 1.17	7.59 ± 1.43	0.133
Fasting plasma glucose (mg/dl)	136.4 ± 32.3	136.0 ± 35.0	0.826
Total cholesterol (mg/dl)	180.9 ± 28.9	188.8 ± 30.10	0.231
Triglyceride (mg/dl)	136.8 ± 94.0	154.7 ± 98.1	0.358
HDL (mg/dl)	47.9 ± 13.2	47.7 ± 11.0	0.925
Blood urea nitrogen (mg/dl)	16.2 ± 5.2	15.2 ± 3.8	0.393
Creatinine (mg/dl)	0.9 ± 0.3	0.9 ± 0.2	0.498

Data are means ± SD.

Kwon, Hyuk-Sang, et al. "Establishment of blood glucose monitoring system using the internet." *Diabetes care* 27.2 (2004): 478-483.

Approximately Normal Distribution

- Linear Regression
 - Describes the relationship between an explanatory variable (independent) and a continuous outcome variable (dependent)
 - Independent variable: categorical or continuous
 - How well does variable X predict variable Y?
 - Multiple linear regression - used to include multiple independent variables
- Correlation
 - Describes the strength of a linear relationship between two continuous variables
 - Pearson correlation coefficient (r)
 - $-1 \leq r \leq 1$

Nonparametric tests

- Nonparametric tests should typically be used if:
 - the variable (or a transformed version) does not have an approximately normal distribution
 - the distribution is unknown and cannot rely on large sample (>30) theory

Nonparametric tests

Parametric	Nonparametric
One-sample t-test	Sign Test (ordinal data)
Paired t-test	Signed-Rank Test
t-test: 2 independent samples	Mann-Whitney Test Wilcoxon Rank Sum Test
Pearson Correlation	Spearman Correlation
ANOVA	Kruskal-Wallis 1-way ANOVA

Objective #3 Review most common statistical methods seen in the literature

Objective #4 Discuss how to select proper statistical test for different applications

Chi-square Test

- Used to compare proportions between two or more populations
 - If the groups are independent - a general chi-square is appropriate
 - If the groups are dependent - a McNemar chi-square is appropriate
- Usually data that's presented in a 2x2 table is compared using Chi-square test
- **If expected cell counts are too small** - a Fisher's Exact test is appropriate

Example: Chi-square test used in publications

Table 1—Clinical characteristics and baseline laboratory data of the control and intervention groups

Characteristics	Control group	Intervention group	P
n	55	55	
Age (years)	54.7 ± 9.4	53.5 ± 8.8	0.507
Sex (M/F)	32/18	35/16	0.623
BMI (kg/m ²)	23.9 ± 3.1	24.4 ± 3.4	0.493
Diabetes duration (years)	6.6 ± 5.7	7.0 ± 6.3	0.751
Diagnosis of hypertension (n)	13	17	0.420
Systolic blood pressure (mmHg)	128.5 ± 17.0	124.7 ± 15.8	0.999
Diastolic blood pressure (mmHg)	77.0 ± 9.7	77.5 ± 8.7	0.254
HbA _{1c} (%)	7.19 ± 1.17	7.59 ± 1.43	0.133
Fasting plasma glucose (mg/dl)	136.4 ± 32.3	136.0 ± 35.0	0.826
Total cholesterol (mg/dl)	180.9 ± 28.9	188.8 ± 30.10	0.231
Triglyceride (mg/dl)	136.8 ± 94.0	154.7 ± 98.1	0.358
HDL (mg/dl)	47.9 ± 13.2	47.7 ± 11.0	0.925
Blood urea nitrogen (mg/dl)	16.2 ± 5.2	15.2 ± 3.8	0.393
Creatinine (mg/dl)	0.9 ± 0.3	0.9 ± 0.2	0.498

Data are means ± SD.

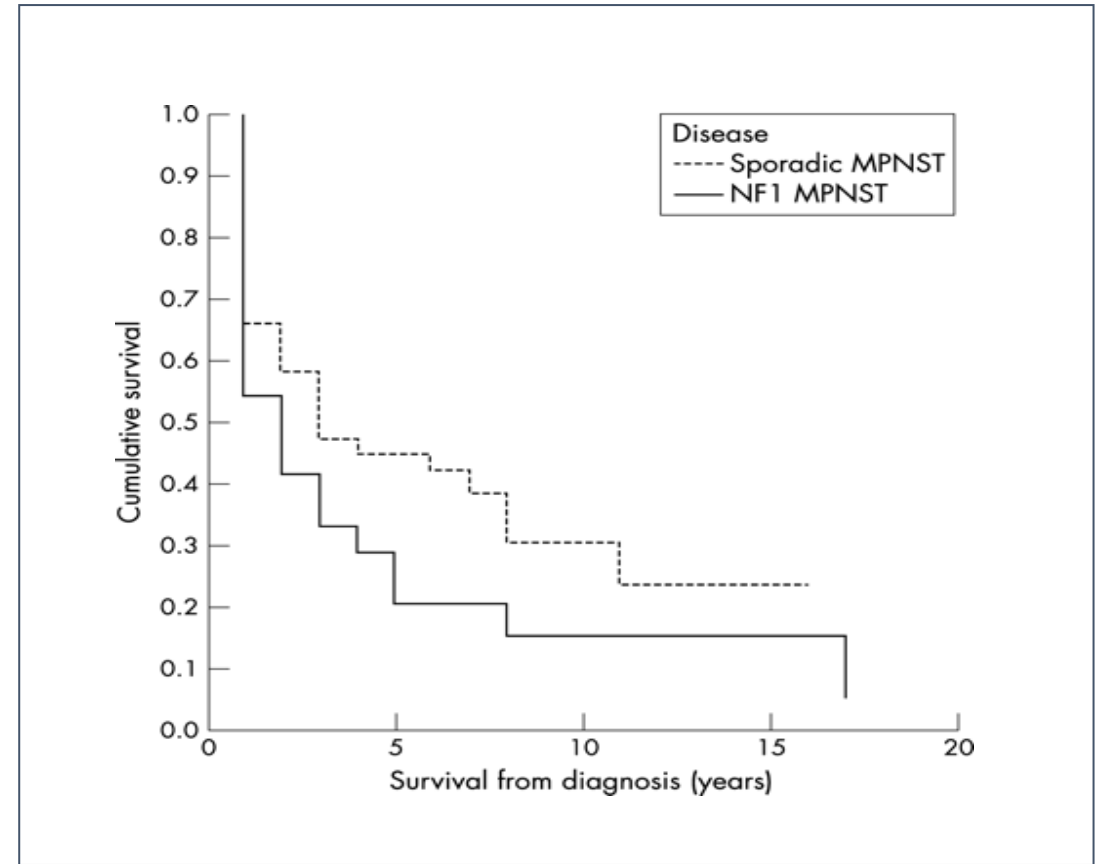
Kwon, Hyuk-Sang, et al. "Establishment of blood glucose monitoring system using the internet." *Diabetes care* 27.2 (2004): 478-483.

Logistic Regression

- Used to predict dichotomous outcome from an explanatory (independent) variable
- Independent variable: categorical or continuous
- Modeling concept similar to linear regression
- Interpretation deals with log odds and odds ratios
- Multiple logistic regression used to include multiple independent variables

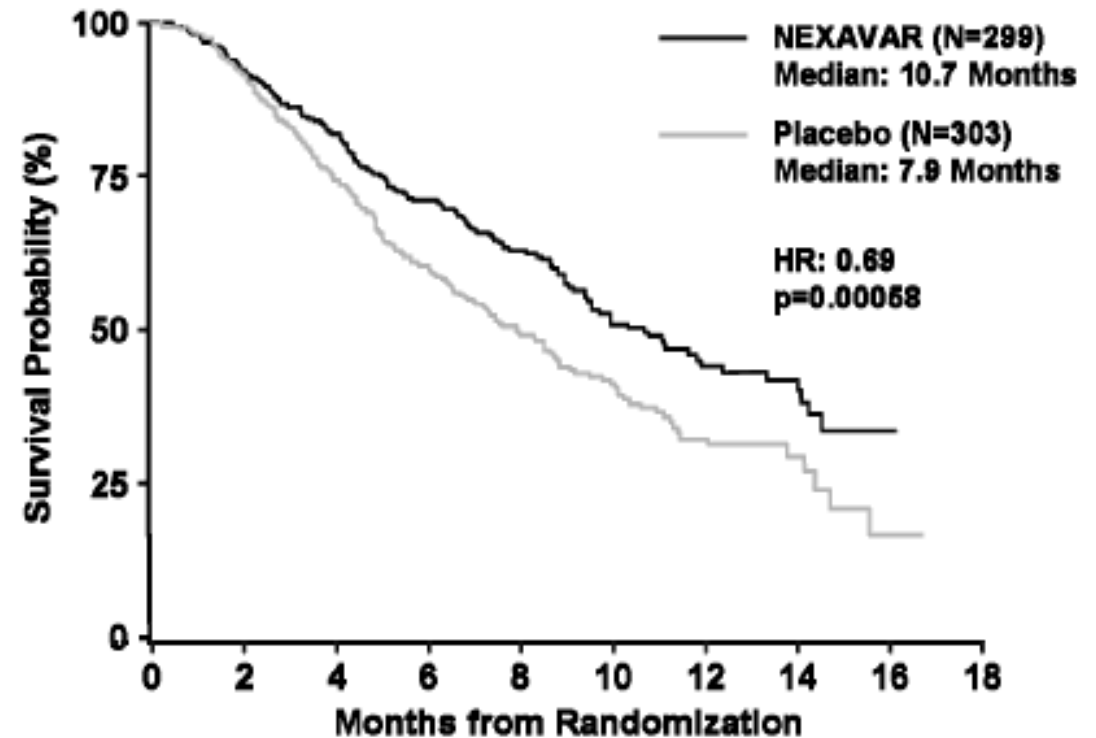
Survival Analysis: Estimation

- Time to event data, censored data
- Kaplan-Meier curves
 - Graphs that illustrate the survivorship function for different groups



Survival Analysis: Comparisons

- Log-rank test
 - Nonparametric method for statistically comparing survival distributions
- Cox proportional hazards model:
 - regression model for time-to-event outcome data;
 - continuous or categorical independent factors



Outcome Variable	Explanatory Variable	
	Categorical	Continuous
Categorical	Chi-square, Logistic Regression	Logistic Regression
Continuous	ANOVA, t-test, Linear Regression	Linear Regression/ Correlation

Objective #3 Review most common statistical methods seen in the literature

Objective #4 Discuss how to select proper statistical test for different applications

Review

- Defined key terms for hypothesis testing
- Reviewed the definition of confidence intervals and estimation
- Looked at the types of data you might encounter
- Introduced the most common statistical methods seen in the literature
- Discussed how to choose the appropriate statistical test based on your data type

Foundations and Basics of Statistical Tests and Data Analysis

Lance Ford, PhD (Lance-Ford@ouhsc.edu)
Assistant Professor of Research, Biostatistics

Sara Vesely, PhD (Sara-Vesely@ouhsc.edu)
Associate Dean of Academic Affairs
David Ross Boyd Professor, Biostatistics

Kai Ding, PhD (Kai-Ding@ouhsc.edu)
Associate Professor, Biostatistics

Chao Xu, PhD (Chao-Xu@ouhsc.edu)
Assistant Professor, Biostatistics

