# Statistics: A Brief Overview

Katherine H. Shaver, M.S.
Biostatistician
Carilion Clinic

CARILION CLINIC

# Statistics: A Brief Overview Course Objectives

- Upon completion of the course, you will be able to:
  - Distinguish among several statistical applications
  - Select a statistical application suitable for a research question/hypothesis/estimation
  - Identify basic database structure / organization requirements necessary for statistical testing and interpretation

# What Can Statistics Do For You?

- Make your research results credible
- Help you get your work published
- Make you an informed consumer of others' research

# Categories of Statistics

- Descriptive Statistics
- Inferential Statistics

# Descriptive Statistics

Used to Summarize a Set of Data

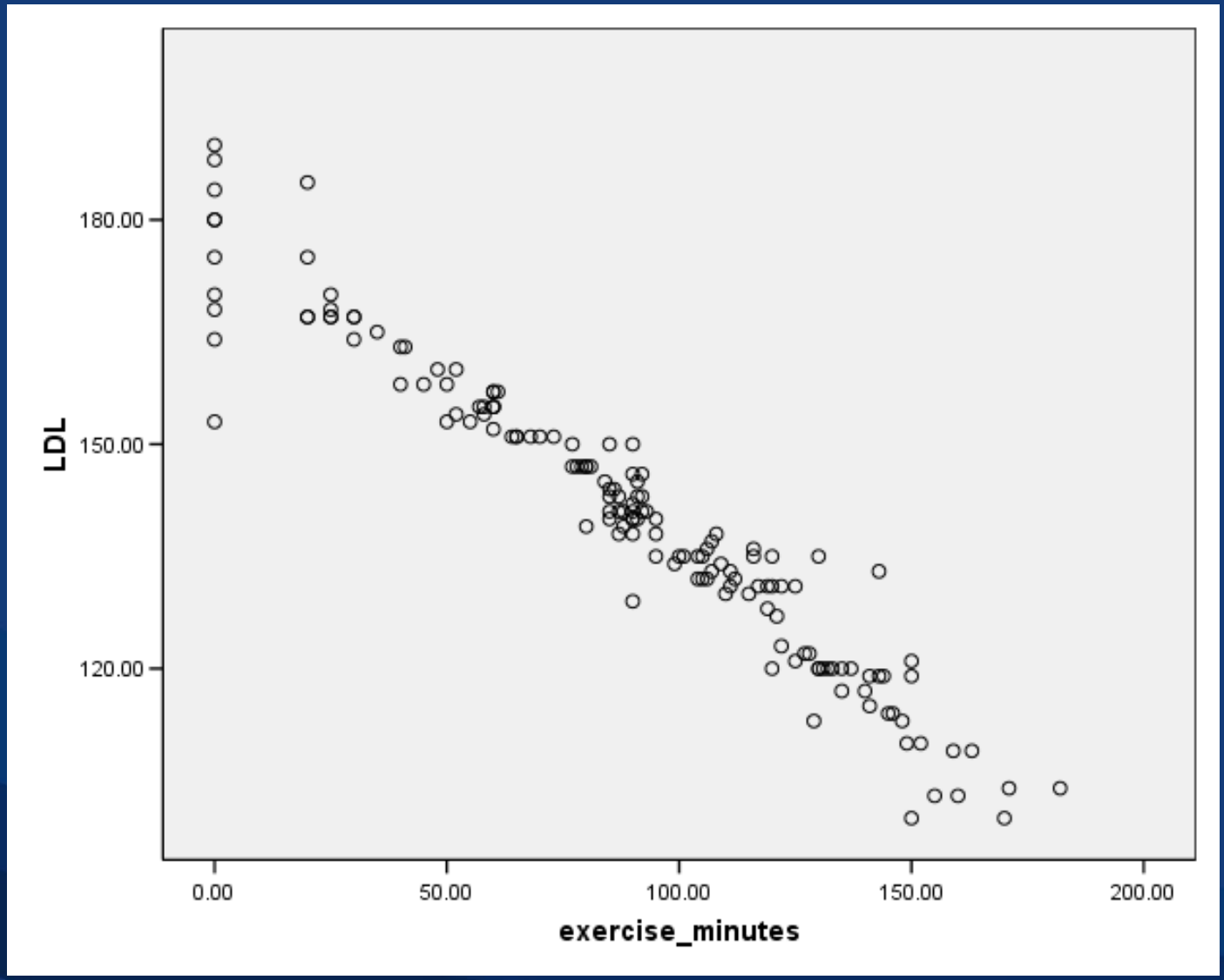
- N (size of sample)
- Mean, Median, Mode (central tendency)
- Standard deviation, 25$^{th}$ and 75$^{th}$ percentiles (variability)
- Minimum, Maximum (range of data)
- Frequencies (counts, percentages)

# Example of a Summary Table of Descriptive Statistics

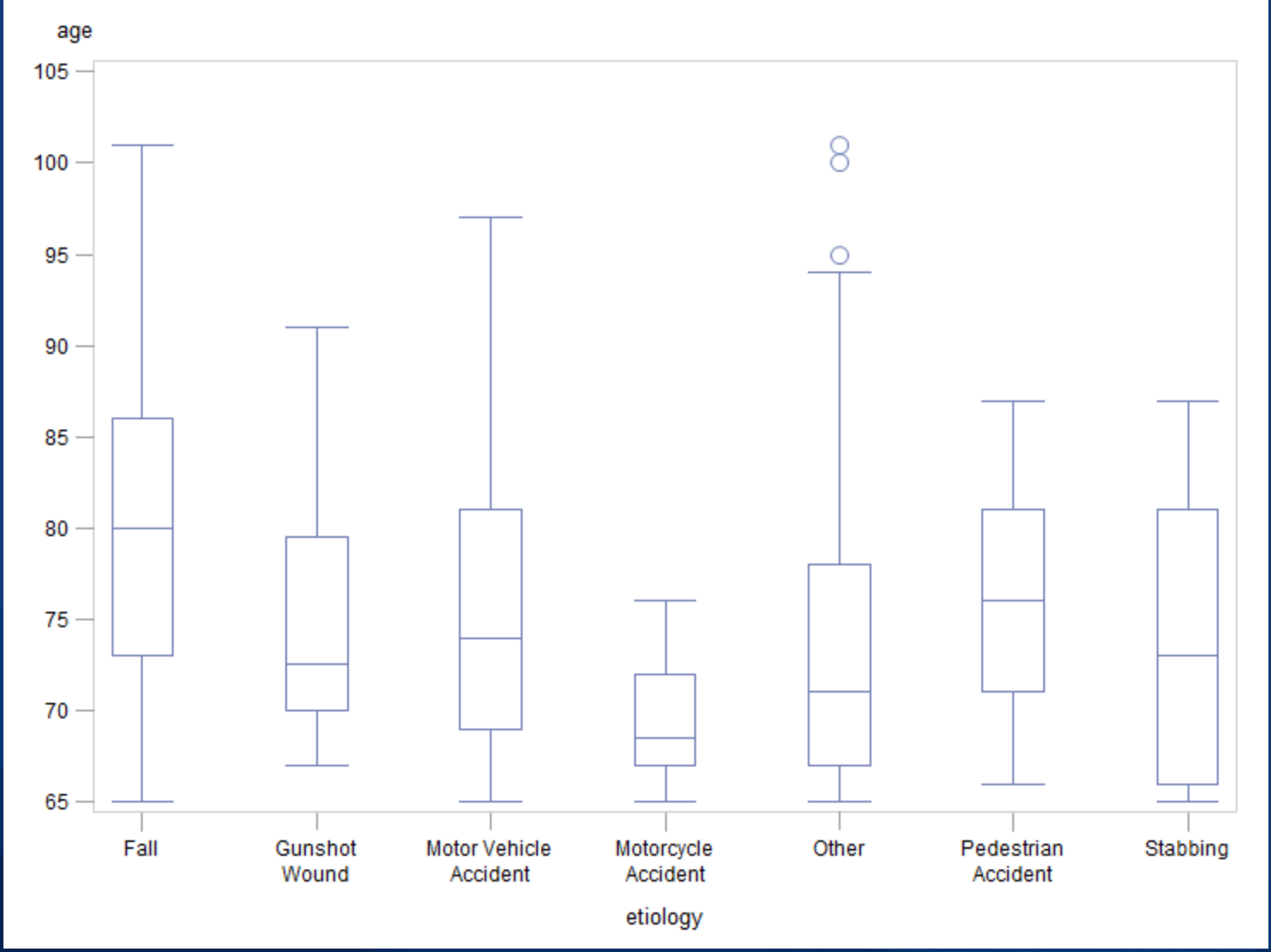
## Summary Statistics for LDL Cholesterol

| Lab Parameter | N | Mean | Std. Dev. | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|---|---|---|
| LDL Cholesterol (mg/dl) | 150 | 140.8 | 19.5 | 98 | 132 | 143 | 162 | 195 |

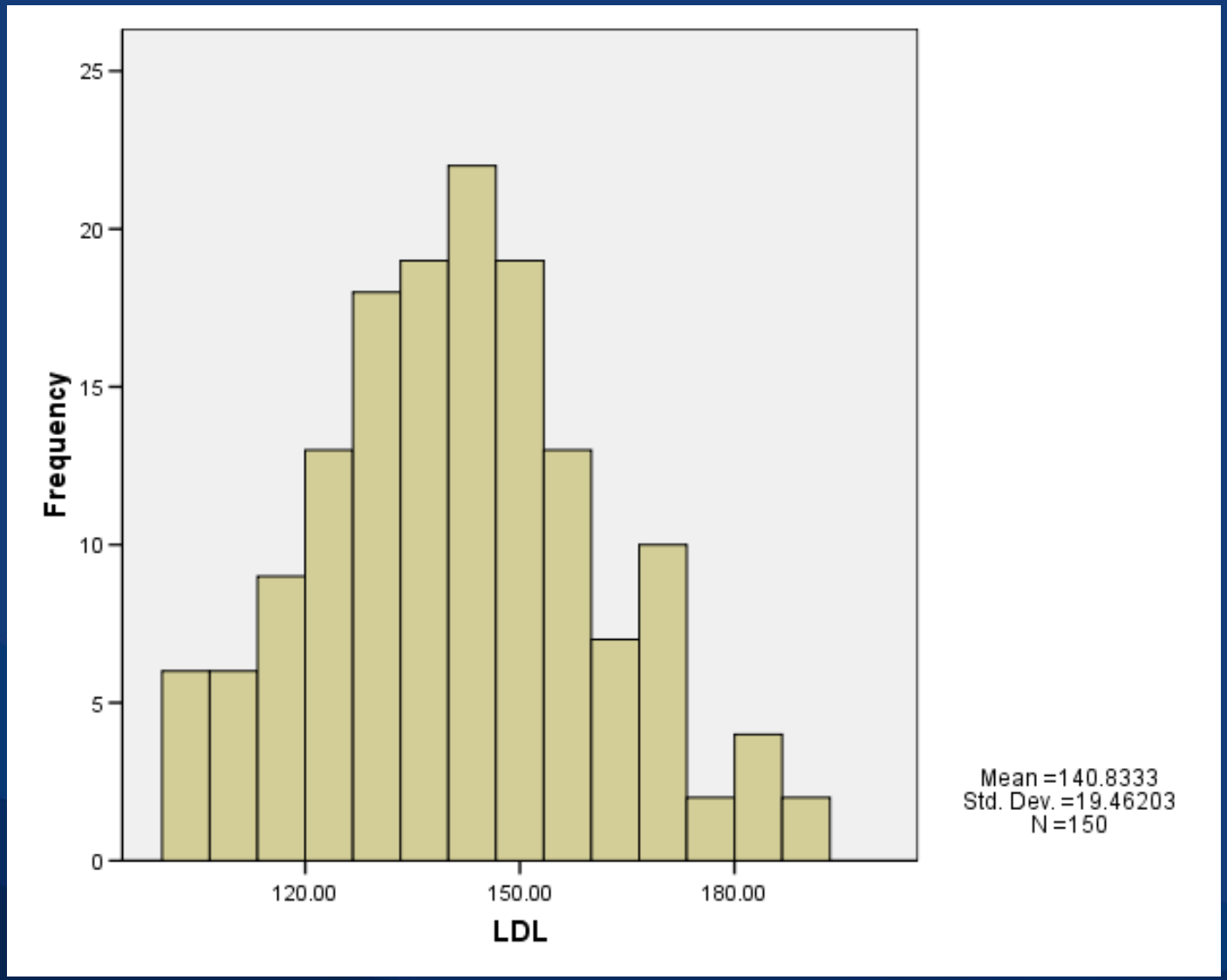
# Scatterplot



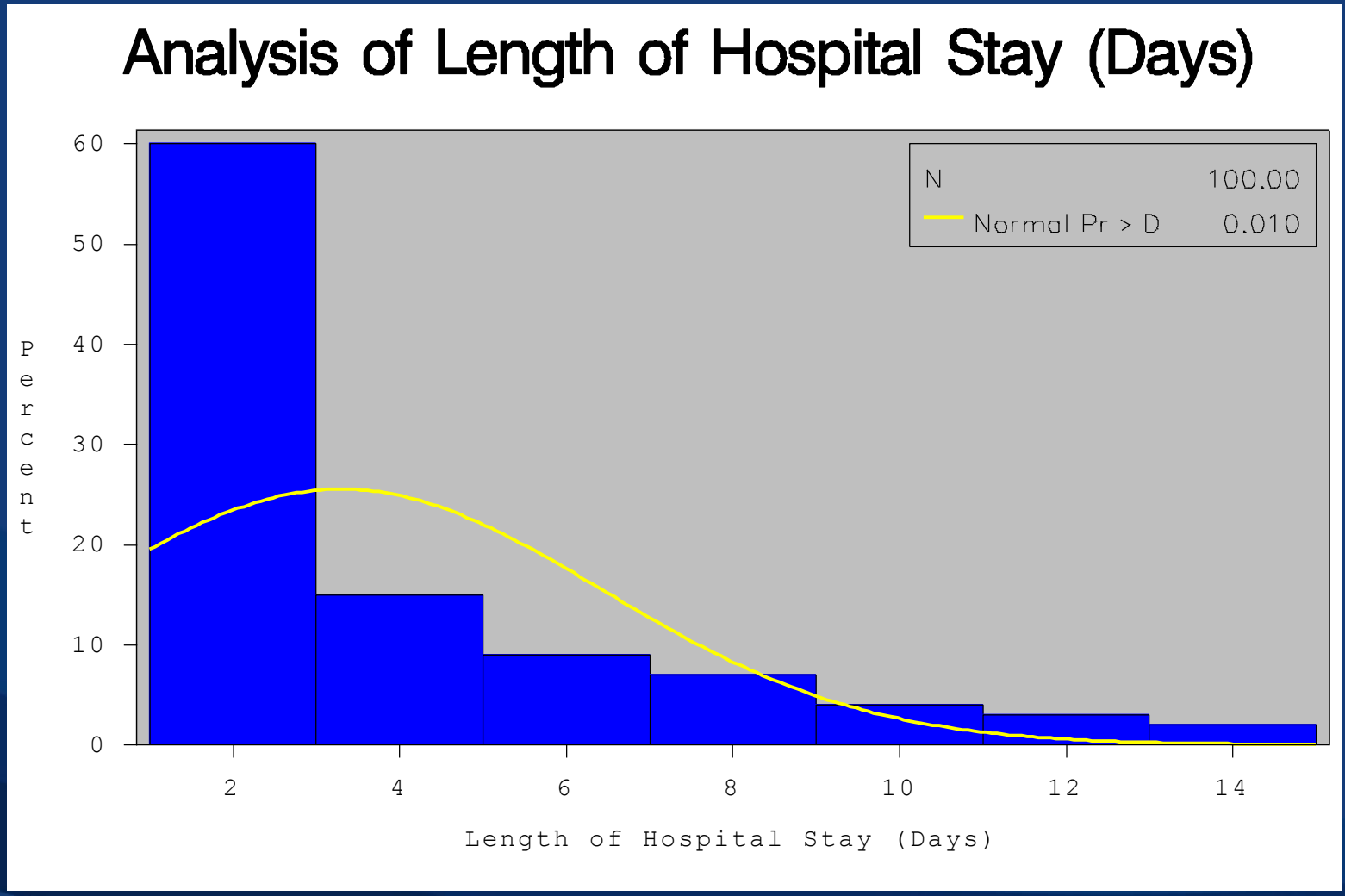
# Box-and-Whisker Plot



# Histogram


25
20
15
10
5
0
Frequency
120.00
150.00
180.00
LDL
Mean =140.8333
Std. Dev. =19.46203
N =150

# Length of Hospital Stay – A Skewed Distribution



# Inferential Statistics

- Data from a random sample used to draw conclusions about a larger, unmeasured population

- Two Types
  - *Estimation*
  - *Hypothesis Testing*

# Types of Inferential Analyses

- Estimation – calculate an approximation of a result and the precision of the approximation [associated with confidence interval]

- Hypothesis Testing – determine if two or more groups are statistically significantly different from one another [associated with p-value]

# Types of Data

The type of data you have will influence your choice of statistical methods…

# Types of Data

Categorical:

- Data that are labels rather than numbers

  - Nominal – order of the categories is arbitrary (e.g., gender)

  - Ordinal – natural ordering of the data. (e.g., severity of pain rated as: None, Mild, Moderate, Severe, Very Severe)

# Types of Data (cont'd)

Continuous:

- Data with a potentially infinite number of possible values (e.g., weight, blood pressure)
- Quanititative; positions along continuous number lines
  - Interval vs. ratio – what's the difference?

# Dependent vs. Independent Variable

- ***Dependent Variable***: variable you believe may be influenced / modified by treatment or exposure.
- May represent variable you are trying to predict.
- Sometimes referred to as response or outcome variable.

# Dependent vs. Independent Variable

- ***Independent Variable***: variable you believe may influence outcome measure.
- Often referred to as predictor or explanatory variable.

# Selection of Appropriate Statistical Method for Hypothesis Testing

| | | Predictor (Independent) Variable | |
|---|---|---|---|
| | | Categorical | Continuous |
| Outcome (Dependent) Variable | Categorical | **Chi-square, Fisher's Exact Test** | **Logistic Regression** |
| | Continuous | **t-test, ANOVA** | **Correlation, Linear Regression** |

# Power Analysis

- Statistical power – the ability to detect an effect that actually exists
  - If your results are significant, then you had enough power.
- Is a blend of science and art

# Power Analysis

- What do you need to know before a power analysis can be conducted?
  - How will you analyze the data?
  - Estimates appropriate to the statistical technique; two main sources:
    - Published literature
    - Educated guesses (really, it's OK!)

# Power Analysis

- T-test example
  - Estimates of group means, variability
- Sometimes we need to approach it from another direction
  - Constraints of total number of patients available, time, funding
  - Multiple scenarios showing what various sample sizes will “buy” you in terms of statistical power

# Sample Questions, Example Datasets

- The following slides contain examples of research questions that are answered using hypothesis testing.
- Each question is matched with an appropriate statistical method. (Today's presentation will only cover t-tests and the rest will be covered in Part II of this course.)
- For each question/method combination, there is also a snapshot of what the dataset would look like.

# Two-Sample Student's *t*-test for Independent Samples

- A 2-Sample Student's *t*-test for independent samples is used to compare the means of two *different* groups.

- Example:

  Is there a difference in the time to therapeutic threshold of a drug in lower weight patients compared to higher weight patients?

# Data Layout for 2-Sample *t*-Test for Independent Samples

| Group | Time_in_Hours |
|---|---|
| < 100 kg | 5.5 |
| < 100 kg | 12 |
| < 100 kg | 6 |
| < 100 kg | 22 |
| < 100 kg | 38 |
| >= 100 kg | 18 |
| >= 100 kg | 42 |
| >= 100 kg | 36 |
| >= 100 kg | 20 |
| >= 100 kg | 14 |

# 2-Sample *t*-Test for Independent Samples

- Group 1: < 100 kg
- Group 2: >= 100 kg

- Test Statistic $$t = \frac{\left(\bar{x}_1 - \bar{x}_2\right)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- The test statistic is calculated. If the probability of getting the resulting value by chance is <0.05 then the means of the two groups are considered to be statistically significantly different from one another.

# Paired *t*-test

- For comparing pre and post data from the same subject

| Patient_ID | Pre_Score | Post_Score |
|---|---|---|
| 1 | 5 | 7 |
| 2 | 4 | 5 |
| 3 | 3 | 3 |
| 4 | 7 | 6 |
| 5 | 7 | 8 |
| 6 | 2 | 5 |

# Paired *t*-test

- Calculate the mean and standard deviation of the difference

- Test statistic $$t = \frac{\overline{D}}{S_d / \sqrt{n}}$$

- The test statistic is calculated and if the probability of getting the resulting value by chance is <0.05 then there is a statistically significant change from pre to post.

# Correlation

- Used to examine linear relationships between two continuous variables.
- Isn’t usually the primary statistical technique of a study
- Example:

  What is the relationship between dietary cholesterol intake and LDL?

# Correlation

- The correlation coefficient, "*r*", ranges from -1.00 to 1.00.
  - The *number* indicates the strength of the relationship
    - Values closer to -1 or 1 indicate a stronger relationship
  - The *sign* indicates the nature of the relationship
    - A positive *r* indicates a direct relationship
    - A negative *r* indicates an inverse relationship

# Correlation

- Two types of relationships can be identified with correlation:
  - ⇧⇧ or ⇩⇩ As the value of one variable increases, the value of the other variable increases. Likewise, as the value of a variable decreases, the value of the other variable decreases.
  - ⇧⇩ or ⇩⇧ As the value of one variable increases, the value of the other variable decreases. Likewise, as the value of a variable decreases, the value of the other variable increases.

***ALWAYS REMEMBER:***

Correlation does not equal causation!

# Data Layout for Correlation

| Study_ID | Avg_daily_chol | LDL |
|---|---|---|
| 1 | 305 | 135 |
| 2 | 212 | 127 |
| 3 | 397 | 148 |
| 4 | 200 | 105 |
| 5 | 195 | 119 |
| 6 | 461 | 164 |
| 7 | 479 | 162 |
| 8 | 354 | 155 |
| 9 | 288 | 130 |

# Correlation

- These data are statistically significantly correlated with an *r* of 0.94.
- This indicates a very strong positive relationship: as average daily cholesterol intake increases, LDL increases.

# Categorical Data

- So far we have considered situations where our dependent variable was continuous.
- What if our variable of interest is categorical?
- Chi-square and logistic regression – two very commonly used techniques at Carilion

# Chi-square / Fisher's Exact Test

- Use when both the *predictor (independent) and the outcome (dependent) variable are categorical*

- Often used to compare proportions of two groups.

# Chi-square / Fisher's Exact Test

- The easy hand-calculation of the chi-square statistic contributed to its popularity in the era before computers.
- However, chi-square does not work well with small sample sizes or sparse data.
- Fisher's exact test is a good alternative for 2x2 tables regardless of the sample size.

# Chi-square / Fisher's Exact Test

- Example:

  Do patients with staph aureus who receive an infectious disease consult have a lower 60-day mortality compared to staph aureus patients who do not?

# Data Layout

| Study_ID | ID_consult | Alive |
|---|---|---|
| 1 | Yes | Yes |
| 2 | Yes | No |
| 3 | Yes | Yes |
| 4 | Yes | Yes |
| 5 | Yes | Yes |
| 6 | Yes | Yes |
| 7 | No | No |
| 8 | No | No |
| 9 | No | No |
| 10 | No | No |
| 11 | No | No |
| 12 | No | Yes |

# 2x2 Table Analysis

The FREQ Procedure

Table of ID_consult by Alive

| | | Alive | | |
|---|---|---|---|---|
| | | No | Yes | Total |
| ID_consult | | | | |
| **No** | **Frequency** | **5** | **1** | **6** |
| | *Row Pct* | *83.33* | *16.67* | |
| | *Col Pct* | *83.33* | *16.67* | |
| **Yes** | **Frequency** | **1** | **5** | **6** |
| | *Row Pct* | *16.67* | *83.33* | |
| | *Col Pct* | *16.67* | *83.33* | |
| | | | | |
| **Total** | **Frequency** | **6** | **6** | **12** |

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 5.3333 | 0.0209 |
| Likelihood Ratio Chi-Square | 1 | 5.8221 | 0.0158 |
| Continuity Adj. Chi-Square | 1 | 3 | 0.0833 |
| Mantel-Haenszel Chi-Square | 1 | 4.8889 | 0.027 |
| Phi Coefficient | | 0.6667 | |
| Contingency Coefficient | | 0.5547 | |
| Cramer's V | | 0.6667 | |

WARNING: 100% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

| Fisher's Exact Test | |
|---|---|
| Cell (1,1) Frequency (F) | 5 |
| Left-sided Pr <= F | 0.9989 |
| Right-sided Pr >= F | 0.04 |
| | |
| Table Probability (P) | 0.039 |
| Two-sided Pr <= P | 0.0801 |

# Logistic Regression

- Logistic regression is used to *predict a single outcome (dependent) variable from two or more predictor (independent) variables*.
- The outcome must be binary (yes/no). Predictors can be any type of data – categorical or continuous.

# Logistic Regression

- Example:

  What are the predictors of mortality in elderly trauma patients?

# Data Layout

| Study _ID | Age | Gender | ISS | Alive |
|---|---|---|---|---|
| 1 | 65 | F | 16 | Yes |
| 2 | 68 | F | 20 | Yes |
| 3 | 90 | M | 13 | Yes |
| 4 | 78 | M | 12 | No |
| 5 | 82 | M | 22 | No |
| 6 | 77 | F | 19 | Yes |
| 7 | 66 | M | 15 | Yes |
| 8 | 94 | F | 18 | No |
| 9 | 73 | M | 11 | Yes |

# Logistic Regression

- Regression should be considered an exploratory technique. *One regression analysis cannot confirm anything.*
- Rule of thumb for sample size – minimum of 20 cases per predictor variable
- More is almost always better!

# Logistic Regression

- Logistic regression analysis includes:
  - Creation of develop, test, and validate datasets (best practice if you have enough data)
  - EDA to describe and understand data
  - Stepwise techniques to reduce number of predictors (use with caution!)
  - Interaction variables

# Logistic Regression

- Results tell you:
  - Which predictors, if any, were statistically significant
  - The overall strength / predictive ability of your model
  - Odds ratios

# A Caveat!

- We have discussed only a few commonly used statistical methods.  If your research question does not quite fit one of the methods discussed here, don’t try to force it.
- There are many variations of these methods, and there are numerous other methods not mentioned in this presentation that are appropriate for almost any research situation.
- Contact a biostatistician for assistance.

# A Biostatistician Can Help You With:

- Study design
- Choosing outcome variables and how they are measured
- Choosing appropriate statistical methodology
- Power and sample size calculation
- Helping to choose data sources
- Helping to design data collection forms
- Data cleaning, derivations, and analysis
- Interpretation of results
- Helping to write method and results sections of a document

# Feel free to contact us!


- Mattie Tenzer, Director, Health Analytics Research Team (HART)
- Min Wang, Manager, HART
- Katherine H. Shaver, Biostatistician, HART