Statistics: A Brief Overview
Part I

Katherine 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

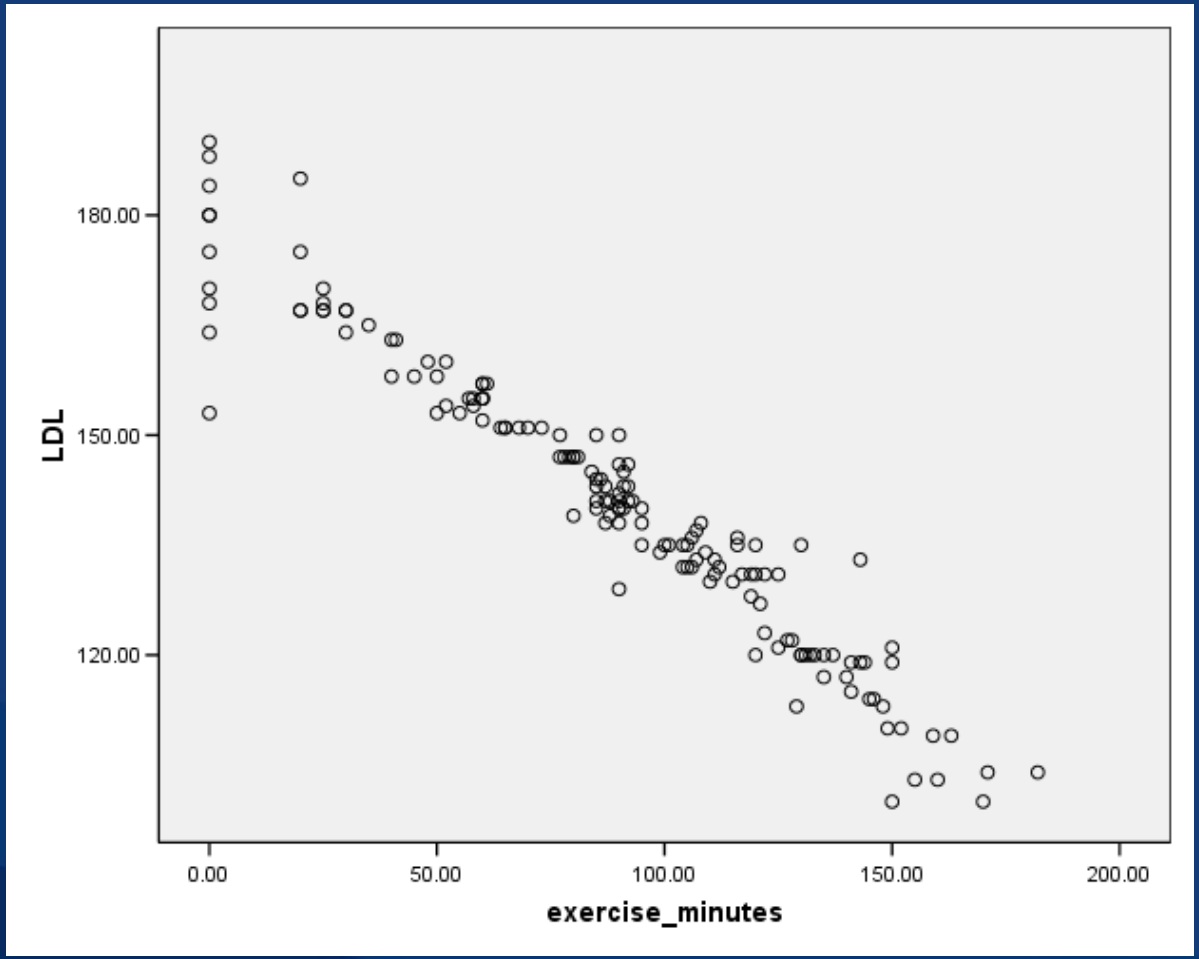
- N (size of sample)
- Mean, Median, Mode (central tendency)
- Standard deviation, 25th and 75th percentiles (variability)
- Minimum, Maximum (range of data)
- Frequencies (counts, percentages)

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

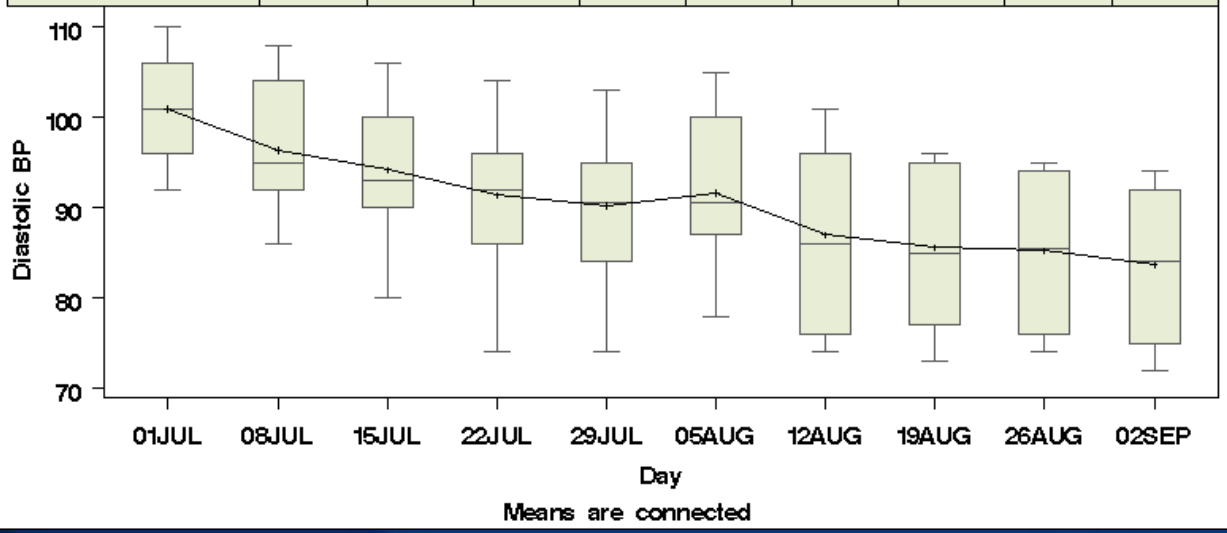


Box-and-Whisker Plot

Box Plot for Diastolic Blood Pressure Over Time

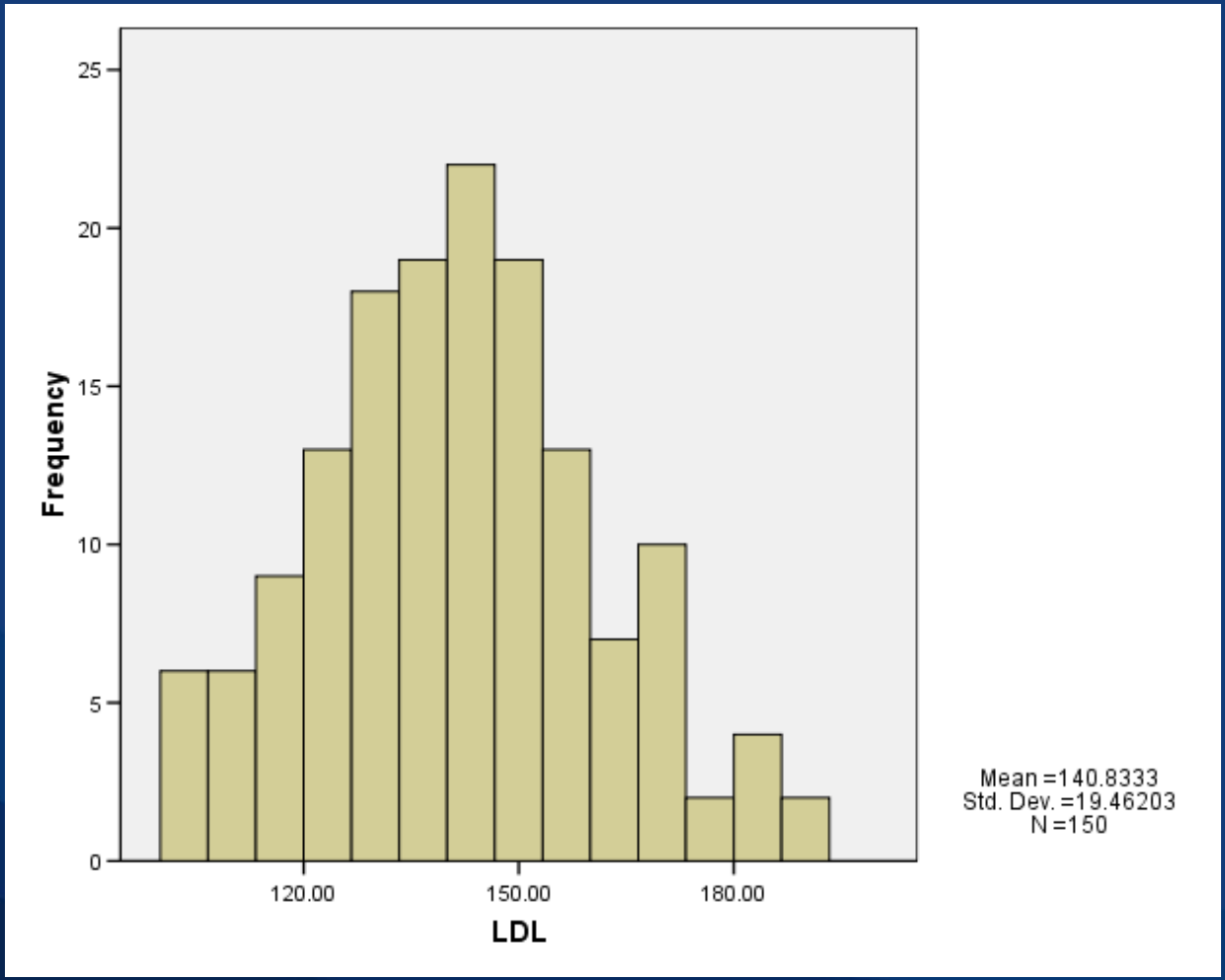
Group= A

Statistics by Day										
N	10	10	10	10	10	10	10	10	10	10
Mean	101	96.4	94.2	91.4	90.3	91.7	87	85.6	85.3	83.7
Std Dev	6.1	7.1	8.0	9.0	8.9	9.0	10.0	9.1	8.8	8.7
Min	92	86	80	74	74	78	74	73	74	72
25th Percentile	96	92	90	86	84	87	76	77	76	75
Median	101	95	93	92	90.5	90.5	86	85	85.5	84
75th Percentile	106	104	100	96	95	100	96	95	94	92
Max	110	108	106	104	103	105	101	96	95	94

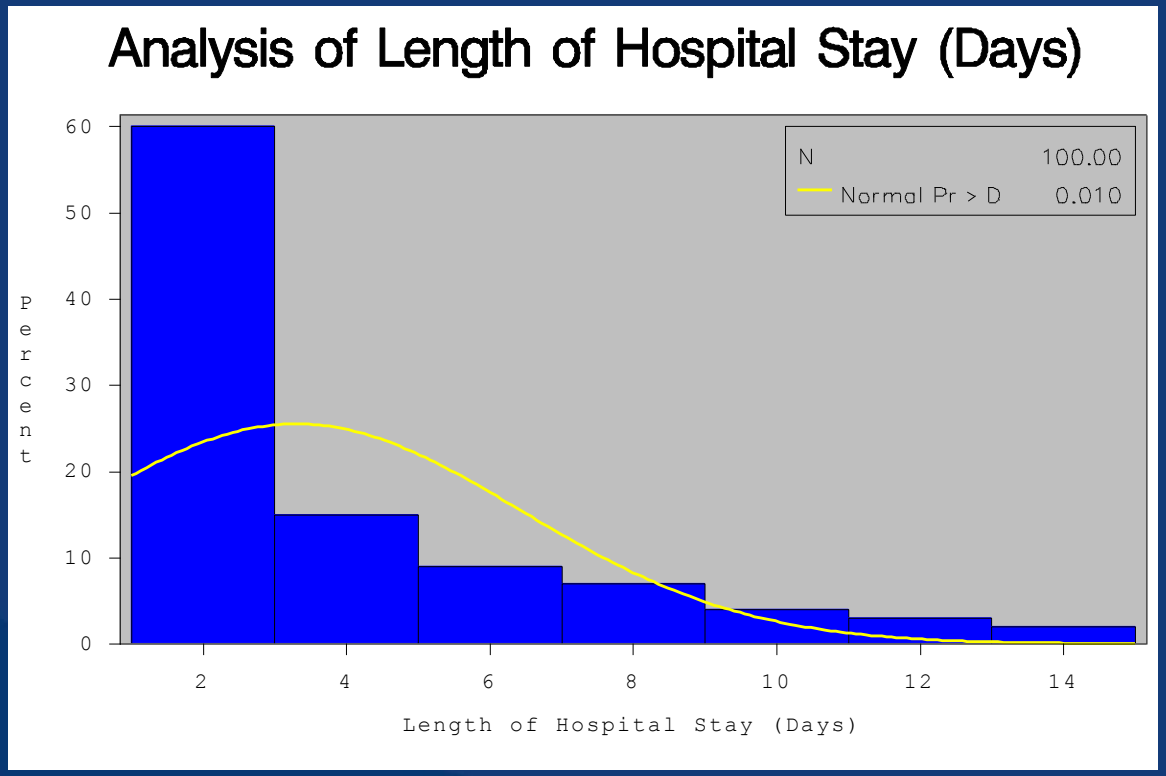


Means are connected

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)

Discrete:

- **Measurement scale consisting of a number of separate values where intermediate values are not permissible**

  (e.g., the number of trauma patients admitted to the hospital in a given day)

Continuous:

- **Data with a potentially infinite number of possible values (e.g., weight, blood pressure)**

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**

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.

One-sample Student's *t*-test

- A one-sample Student's *t*-test is useful for *comparing the mean value* of an experimental group with the mean value of a known norm.

- Example:

  The statewide mean ISS score for patients involved in a motorcycle accident and admitted to a hospital is 10. Does the ISS score of CRMH motorcycle patients differ from the statewide average?

Data Layout for a One-sample Student's *t*-test

Patient_ID	CRMH_ISS
1	4
2	9
3	13
4	3
5	20
6	14
7	4
8	17
9	7
10	15

One-sample Student's *t*-test

- Test Statistic

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$$

- The test statistic is calculated. If the probability of getting the resulting test statistic by chance is <0.05 then the CRMH ISS is considered to be statistically significantly different than the VA statewide ISS.

- Note: Failure to achieve statistical significance ***does not imply that there is no difference***. It should be interpreted that there is not enough evidence to detect a statistically significant difference.

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.

Part II…

- Part II of this course will cover analysis of variance (ANOVA), correlation, regression, and analysis of categorical data / proportions. We will also discuss power analysis.

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
  - mmtenzer@carilionclinic.org
  - 224-5192 (x55192)
- Katherine Shaver, Biostatistician
  - khshaver@carilionclinic.org
  - 224-5197 (x55197)