Statistics: A Brief Overview Part II

Katherine Shaver, M.S.

Biostatistician

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

Picking up where we left off...

- In Part I, we discussed:
  - Descriptive statistics
  - Types of data, dependent vs. independent variables
  - T-tests: one-sample, two-sample (independent), and paired
- Part II:
  - ANOVA, correlation, chi-square, logistic regression, and power analysis

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.
- For each question/method combination, there is also a snapshot of what the dataset would look like.

Analysis of Variance (ANOVA)

- ANOVA is used to compare the means of *three or more groups* and for designs with multiple explanatory (independent) factors.

- Example:

  Do diet type and gender affect LDL levels?

- In our example there are *2 levels of the variable "gender"* and *3 levels of the variable "diet."* This is a 2x3 factorial ANOVA.

Data Layout for a 2x3 Factorial ANOVA

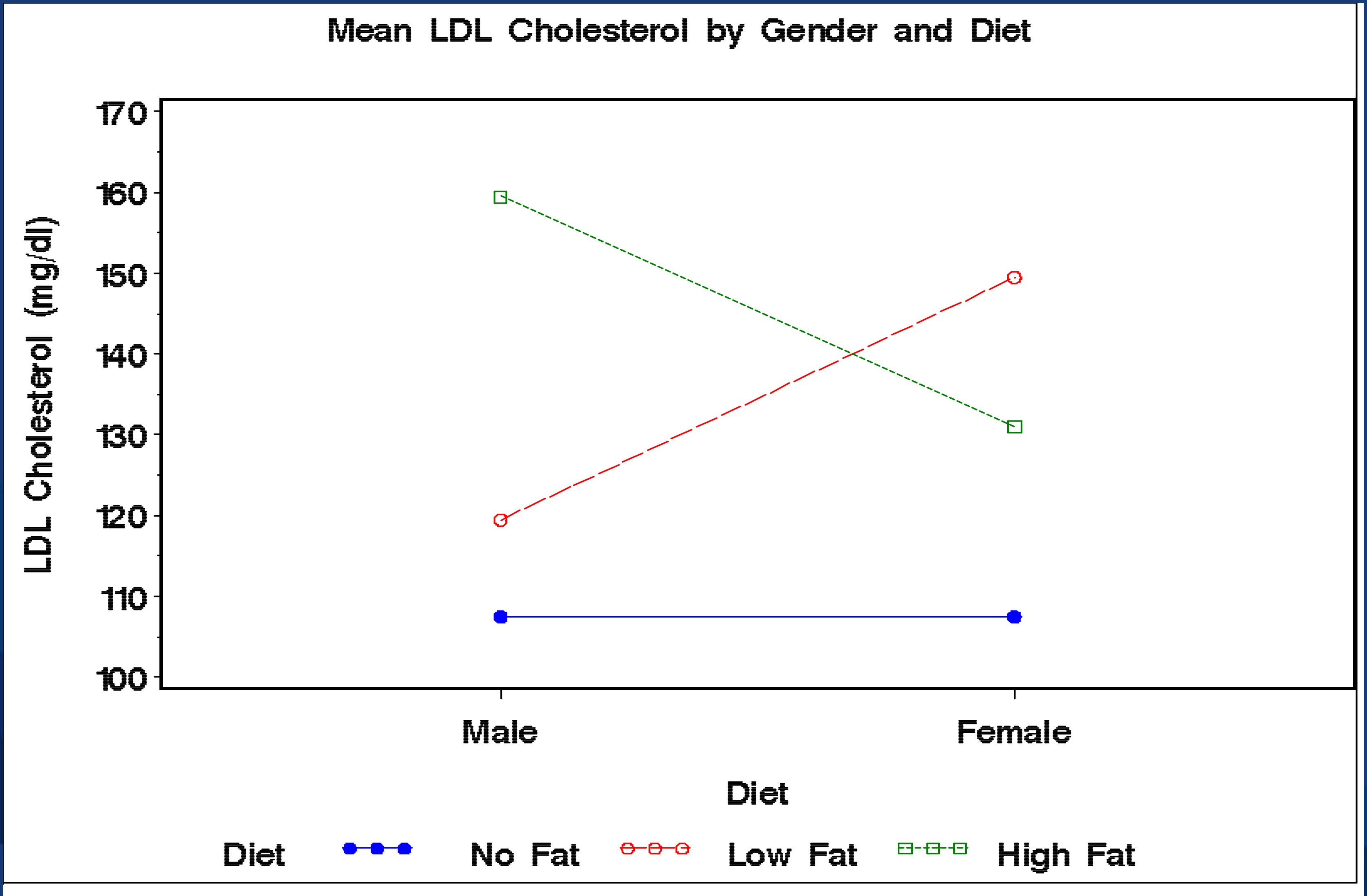
Study_ID	Gender	Diet	LDL
1	Male	No Fat	105
2	Male	No Fat	110
3	Female	No Fat	108
4	Female	No Fat	107
5	Male	Low Fat	120
6	Male	Low Fat	119
7	Female	Low Fat	150
8	Female	Low Fat	149
9	Male	High Fat	157
10	Male	High Fat	162
11	Female	High Fat	130
12	Female	High Fat	132

ANOVA Main Effects

Mean LDL Cholesterol by Gender and Diet

Gender	Male	128.8
	Female	129.3
Diet	No Fat	107.5
	Low Fat	134.5
	High Fat	145.3

ANOVA Interaction



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 usually 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

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
- 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

A Caveat!

- We have presented 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.

Some Things 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

Questions to Consider

- Some questions a biostatistician may ask:
  - Is this a retrospective chart review, an observational study, or a prospective well-controlled randomized clinical trial?
  - What is your primary research question?
  - Are you mainly interested in estimating a parameter or in comparing groups?
  - Is there a single outcome variable that best addresses this research question?
  - How will you measure this outcome variable?

Questions to Consider

- – What demographic, baseline or on-going factors may influence your results (e.g., age, concomitant medications)?
- – How many subjects are available for your study?
- – What is the smallest clinically meaningful difference between two experimental groups?
- – How do you expect the data to vary (e.g., estimate of standard deviation or minimum/maximum value expected)?

Feel free to contact us!

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