Final Exam Review

Introduction to Biostatistics 171:161

The Main Idea

- Statistics are a tool for us to take information from a smaller group and use it to reach conclusions/answer questions about a population.
- Most of what we do is based off of this:
  - We look at study design to make sure that the smaller group is reflective of the population we are interested in.
  - We look at the distribution of our data and chose methods that will give us the most accurate answers.
  - We also look at the questions we are trying to answer and chose methods that best address them.

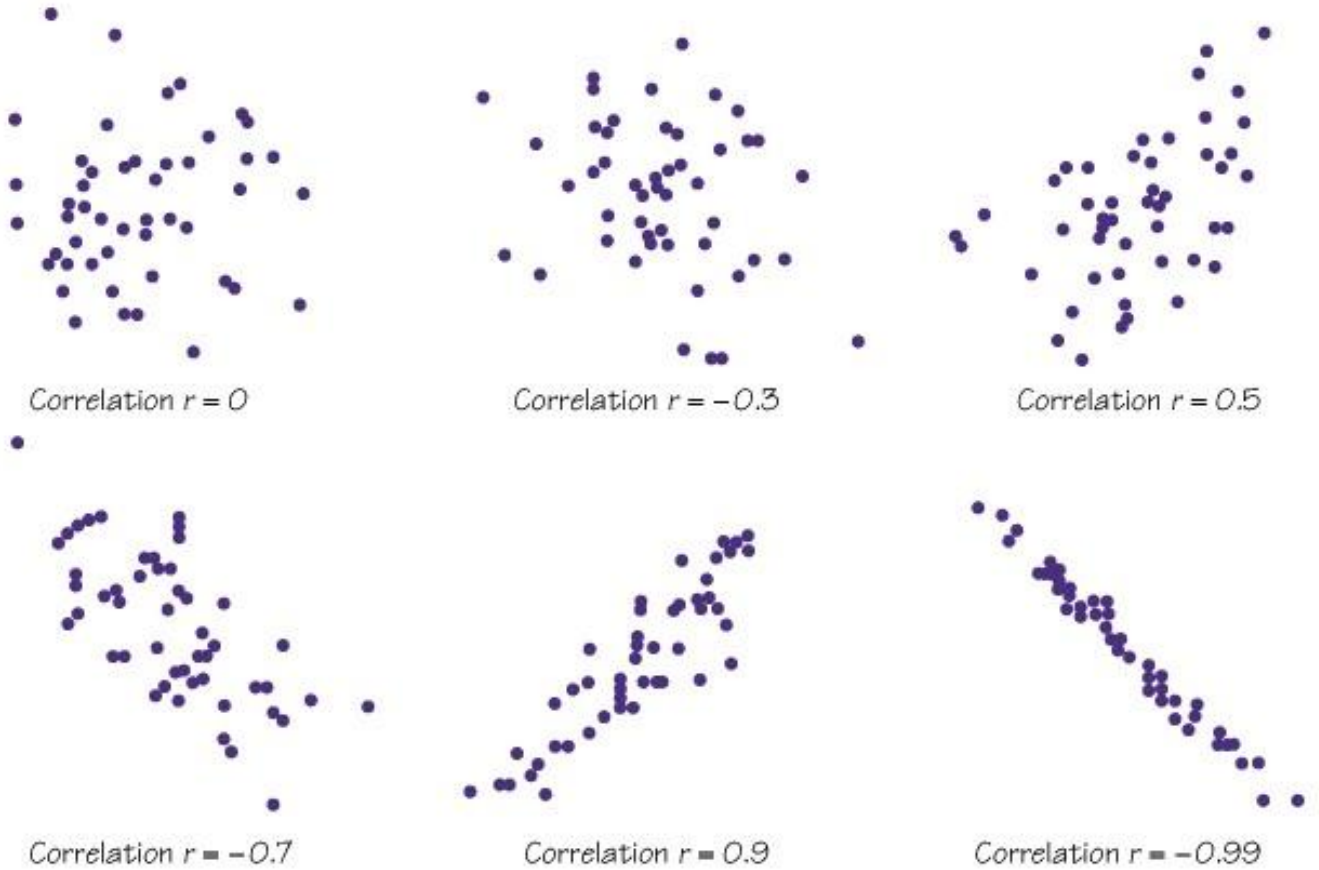
Our Tools and When to Apply them

- The rest of this review will be a summary of the main statistical tools and concepts we have learned in the class.
- We will look at all the tools for continuous data first, and categorical data second.
- The format will look like:
  - An introduction of a tool/method
  - When to use it/when not to use it
  - An example question that hints at that method being appropriate.

Correlation

- Correlation is a point estimate of the association between two continuous (or nearly continuous) variables
- The correlation coefficient is represents by the symbol “r” and takes on values between -1 and 1.
- -1 is a perfect negative correlation, 0 is no correlation, +1 is perfect positive correlation.
- Correlation is symmetric (the r measuring height vs. weight is the same as the r measuring weight vs. height)

Visual of Different Correlation Coefficients



When to use Correlation

- Correlation is used as a summary statistic for the association between two continuous variables, so use it if you are asked to measure the strength of association between two continuous variables.

<u>Example Scenario:</u>

Each day a local ice cream shop keeps track of its sales and the temperature that day. The store manager is interested in seeing if there is a relationship between sales and temperature. How might the manager assess the possibility of a relationship?

Regression

- Regression shares a lot of similarities with correlation
  - Both are often shown visually through scatter plots
  - Both are often used assess the relationship between two continuous variables
- However regression is focused on predicting/modeling one variable based on knowing another.
- The equation has the form of a line:

  $Y = \alpha + \beta X$ X is the predictor, β is the slope, α is the intercept, Y is the response
- Regression equations are not symmetric (the equation where X represents height, Y represents weight will not have the same alpha/beta as the reverse)

When to Use Regression

- Use regression when you want to predict a continuous outcome based off of knowing a related variable.

<u>Example Scenario:</u>

Each day a local ice cream shop keeps track of its sales and the temperature that day. The store manager is interested in projecting his sales over the next week based off of the weather forecast. Is there are way to do this? How?

The One-Sample t-test

- The t-distribution is used over the Normal distribution when population standard deviation is unknown (we use the sample standard deviation as its estimate)
- For hypothesis testing a One-Sample t-statistic has the form:

$$t_{obs} = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}}$$

Where $\bar{x}$ is the sample average, μ_0 is the hypothesized value (zero in paired cases), S is the sample standard dev.

When to use a One-Sample t-test

- Use this t-test when you have 1 group (or 2 paired groups) with continuous, normally distributed data that you wish compare to a hypothesis.

The Two-Sample t-test

- The two sample t-test is an extension of the one sample test. It is used to assess the difference between two the mean or two unpaired groups.

- $t_{obs} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{SD_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ with degrees of freedom = n1 + n2 – 2

- Use the two-sample t-test when comparing two continuous means for normally distributed data. Use Student's test when you are willing to assume equal standard deviations for each group, use Welch's test when you are not.

ANOVA (analysis of variance)

- Think of ANOVA as an extension of the t-test used to assess the difference between the means of 3 or more treatment groups.
- The test statistic is a ratio of variability being explained by a model to the variability that remains unexplained.
- The test statistic follows the F-Distribution, and has the form:

$$F_{obs} = \frac{SSR_0 - SSR_1}{d_1 - d_0} / (\hat{\sigma}^2)$$

- Here $\hat{\sigma}^2$ is found by summing the squares of the residuals and then dividing by (n - #groups)
- SSR1 is the sum of squares of the residuals in the full model
- SSR0 is the sum of squares of the residuals in the null model

Non-Parametric Tests

- Non-parametric tests are done by ranking observations, and comparing the ranks. They are used on continuous data that is highly skewed or contains extremely influential outliers.
  - The Wilcoxon Signed Rank test is the non-parametric version of one-sample t-test on paired data
  - The Wilcoxon Rank Sum test is the non-parametric version of a two-sample t-test on unpaired data
  - Spearman Correlation is a non-parametric approach to quantifying association with correlation.

Binomial Test

The binomial distribution is used for data with binary outcomes.

$$P(X = k) = \binom{n}{k} (p)^k (1-p)^{n-k}$$

-The probability for a certain number of successes(k) out of a certain number of outcomes(n) is found using the above formula.

-Often we need to sum many of these to find a p-value.

-The complement rule is commonly used with binomial probilities

Binomial Example

A study at Johns Hopkins estimated the survival chances of infants born prematurely by surveying the records of all premature babies born at their hospital in a three-year period. In their study, they found 39 babies who were born at 25 weeks gestation, 31 of which survived at least 6 months. If the true survival probability is 50%, how likely is it that 31 or more babies would survive.

This is a binomial setting because:

#1) The outcome is binary (survived/died)

#2) We are assessing a number of successes out of a number of trials (31 survivals out of 39 possibilities)

Pearson’s Chi-Square and Fisher’s Exact Tests

- Pearson’s Chi-Square Test is a way of assessing association between categorical groups typically used on 2x2 contingency tables
- Fisher’s Exact Test is an exact approach used on contingency tables
- See the last review session or the last practice quiz for examples.

Example #1

- A team from Yale School of Medicine took a look at 1,433 people diagnosed with intracranial meningioma, the most commonly diagnosed brain tumor in the United States. Researchers compared these patients to a test group of 1,350 people without tumors. Participants offered self-reported lifetime dental X-ray histories. Researchers then analyzed the different types of X-rays these two groups had undergone. Patients with tumors were more than twice as likely to have had "bitewing" X-rays at least once per year. Bitewings, in which a patient bites down on X-ray film, take photos of the upper and lower back teeth.

Example #1 – Key points

- The outcome here was categorical
- This was a retrospective study
- The article uses an odds ratio to measure an association

Example #2

- In a study of 16 overweight young adults in India, participants were given, in turns, a dose of an extract made from unroasted coffee beans and a placebo, three times a day over 22 weeks. Their diet throughout the study was unchanged, and they were physically active. Between trials, the participants were given a two-week break for their bodies to reset. Though a few participants given the extract only lost 7 pounds, others lost as much as 26 pounds. On average, the subjects lost 17.5 pounds each, and reduced their body weight by 10.5 percent. Body fat also declined by 16 percent, even though the participants were eating an average of 2,400 calories and burning roughly 400.

Example #2 – Key points

- This was a cross-over study design
- These data could be analyzed using a 1-sample t-test testing whether the weight loss could be zero.
- No evidence of outliers, the mean is centered almost perfectly between the minimum and maximum.

Example #3

- Researchers at the University of College London surveyed nearly 8,000 participants over the age of 52. Using a fake aspirin bottle complete with instructions as the testing instrument, researchers asked participants to answer four basic questions, including "What is the maximum number of days you may take this medicine?" and "List three situations for which you should consult a doctor." All the answers could be found on the label. One third of the adults failed to correctly answer all four questions, and one in eight got two or more wrong. Researchers then monitored the volunteers' health for five years. During that time, 621 of the participants died, and people who missed two or more questions were more than twice as likely to have died than those who got the answers correct.

Example #3 – Key points

- Researchers subdivided the participants into 2 groups and assessed the differences between these two groups.
- The outcome was survival, measured as a binary outcome. The researchers used relative risk as a measure of association.

Example #4

- Researchers from Penn State found that increasing the amount of spices in your diet may lower the level of potentially harmful fat in your bloodstream. The experiment compared two groups of healthy, overweight men. One group ate meals seasoned with the special spice blend; the other ate the same meals prepared without the spices. Men who ate the spicy food saw a decrease of one-third in the level of triglycerides (a type of fat linked to heart disease) in their bloodstreams, and 20 percent lower insulin levels overall — even when the meals were high in fat and made with heavy oils.

Example #4 – Key points

- The outcomes measured here are continuous (the difference in triglycerides/insulin)
- There are two groups, a treatment and a control group
- The data could be tested using a two-sample t-test, the researchers reported using a ratio of means.