

Statistical Methods I

Tamekia L. Jones, Ph.D.

(tjones@cog.ufl.edu)

Research Assistant Professor

Children's Oncology Group Statistics & Data Center

Department of Biostatistics

Colleges of Medicine and Public Health & Health
Professions

Outline of Topics

- I. Descriptive Statistics
- II. Hypothesis Testing
- III. Parametric Statistical Tests
- IV. Nonparametric Statistical Tests
- V. Correlation and Regression

2

Types of Data

- Nominal Data
 - Gender: Male, Female
- Ordinal Data
 - Strongly disagree, Disagree, Slightly disagree, Neutral, Slightly agree, Agree, Strongly agree
- Interval Data
 - Numeric data: Birth weight

3

Descriptive Statistics

- Descriptive statistical measurements are used in medical literature to summarize data or describe the attributes of a set of data
- Nominal data – summarize using rates/proportions.
 - e.g. % males, % females on a clinical studyCan also be used for Ordinal data

4

Descriptive Statistics (contd)

- Two parameters used most frequently in clinical medicine
 - Measures of Central Tendency
 - Measures of Dispersion

5

Measures of Central Tendency

- Summary Statistics that describe the location of the center of a distribution of numerical or ordinal measurements where
 - A distribution consists of values of a characteristic and the frequency of their occurrence
 - Example: Serum Cholesterol levels (mmol/L)

6.8	5.1	6.1	4.4	5.0
7.1	5.5	3.8	4.4	

6

Measures of Central Tendency (contd)

Mean – used for numerical data and for symmetric distributions

Median – used for ordinal data or for numerical data where the distribution is skewed

Mode – used primarily for multimodal distributions

7

Measures of Central Tendency (contd)

Mean (Arithmetic Average)

$$\begin{aligned}\bar{X} &= \frac{\sum X}{n} = \frac{6.8+5.1+6.1+\dots+4.4}{9} \\ &= 48.2/9 \\ &= 5.36\end{aligned}$$

- Sensitive to extreme observations
 - Replace 5.5 with, say, 12.0
 - The new mean = $54.7 / 9 = 6.08$

8

Measures of Central Tendency (contd)

Median (Positional Average)

- Middle observation: $\frac{1}{2}$ the values are less than and half the values are greater than this observation
- Order the observations from smallest to largest
3.8 4.4 4.4 5.0 5.1 5.5 6.1 6.8 7.1
- Median = middle observation = 5.1
- Less Sensitive to extreme observations
 - Replace 5.5 with say 12.0
 - New Median = 5.1

9

Measures of Central Tendency (contd)

Mode

- The observation that occurs most frequently in the data
- Example: 3.8 4.4 4.4 5.0 5.1 5.5 6.1 6.8 7.1
Mode = 4.4
- Example: 3.8 4.4 4.4 5.0 5.1 5.5 6.1 6.1 7.1
Mode = 4.4; 6.1
- Two modes – Bimodal distribution

10

Measures of Central Tendency (contd)

Which measure do I use?

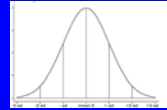


Depends on two factors:

1. Scale of measurement (ordinal or numerical) and
2. Shape of the Distribution of Observations

11

Measures of Central Tendency (contd)

Shape of the distribution

- Symmetric 
- Skewed to the Left (Negative) 
- Skewed to the Right (Positive) 

12

Measures of Dispersion

- Measures that describe the spread or variation in the observations
- Common measures of dispersion
 - Range
 - Standard Deviation
 - Coefficient of Variation
 - Percentiles
 - Inter-quartile Range

13

Measures of Dispersion (contd)

Range = difference between the largest and the smallest observation

- Used with numerical data to emphasize extreme values
- Serum cholesterol example
Minimum = 3.8, Maximum = 7.1
Range = $7.1 - 3.8 = 3.3$

14

Measures of Dispersion (contd)

Standard Deviation

- Measure of the spread of the observations about the mean
- Used as a measure of dispersion when the mean is used to measure central tendency for symmetric numerical data
- Standard deviation like the mean requires numerical data
- Essential part of many statistical tests
- Variance = s^2

15

Measures of Dispersion (contd)

Standard Deviation

6.8 5.1 6.1 4.4 5.0 7.1 5.5 3.8 4.4

Mean = 5.35

n = 9

$$\begin{aligned} s &= \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} \\ &= \sqrt{\frac{(6.8 - 5.35)^2 + (5.1 - 5.35)^2 + \dots + (4.4 - 5.35)^2}{9 - 1}} \\ &= 1.126 \end{aligned}$$

16

Measures of Dispersion (contd)

If the observations have a **Bell-Shaped Distribution**, then the following is always true -

67% of the observations lie between $\bar{X} - 1s$ and $\bar{X} + 1s$
95% of the observations lie between $\bar{X} - 2s$ and $\bar{X} + 2s$
99.7% of the observations lie between $\bar{X} - 3s$ and $\bar{X} + 3s$

The Normal (Gaussian) Distribution

17

Measures of Dispersion (contd)

Coefficient of Variation

- Measure of the relative spread in data
- Used to compare variability between two numerical data measured on different scales
- Coefficient of Variation (C of V) = $(s / \text{mean}) \times 100\%$
- Example:

	Mean	Std Dev (s)	C of V
Serum Cholesterol (mmol/L)	5.35	1.126	
Change in vessel diameter (mm)	0.12	0.29	

18

Measures of Dispersion (contd)

Coefficient of Variation

- Measure of the relative spread in data
- Used to compare variability between two numerical data measured on different scales
- Coefficient of Variation (C of V) = $(s / \text{mean}) \times 100\%$
- Example:

	Mean	Std Dev (s)	C of V
Serum Cholesterol (mmol/L)	5.35	1.126	21%
Change in vessel diameter (mm)	0.12	0.29	241.7%

- Relative variation in Change in Vessel Diameter is more than 10 times greater than that for Serum Cholesterol

19

Measures of Dispersion (contd)

e.g. DiMaio et al evaluated the use of the test measuring maternal serum alphafetoprotein (for screening neural tube defects), in a prospective study of 34,000 women.

Reproducibility of the test procedure was determined by repeating the assay 10 times in each of four pools of serum. Mean and s of the 10 assays were calculated in each of the 4 pools. Coeffs of Variation were computed for each pool: 7.4%, 5.8%, 2.7%, and 2.4%. These values indicate relatively good reproducibility of the assay, because the variation as measured by the std deviation, is small relative to the mean. Hence readers of their article can be confident that the assay results were consistent.

20

Measures of Dispersion (contd)

Percentile

- A number that indicates the percentage of the distribution of data that is equal to or below that number
- Used to compare an individual value with a set of norms
- Example - Standard physical growth chart for girls from birth to 36 months of age
 - For girls 21 months of age, the 95th percentile of weight is 13.4 kg. That is, among 21 month old girls, 95% weigh 13.4 kg or less, and only 5% weigh more than 13.4 kg.
- 50th percentile is the Median

21

Measures of Dispersion (contd)

Interquartile Range (IQR)

- Measure of variation that makes use of percentiles
- Difference between the 25th and 75th percentiles
- Contains the middle 50% of the observations (independent of shape of the distribution)
- Example –
 - IQR for weights of 12 month old girls is the difference between 10.2 kg (75th percentile) and 8.8 kg (25th percentile);
 - i.e., 50% of infant girls at 12 months weigh between 8.8 and 10.2 kg.

22

Hypothesis Testing

- Permits medical researchers to make generalizations about a population based on results obtained from a study
- Confirms (or refutes) the assertion that the observed findings did not occur by chance alone but due to a true association between the dependent and independent variable
- The aim of the researcher is to demonstrate that the observed findings from a study are statistically significant.

23

Hypothesis Testing (contd)

- *Statistical Hypothesis* – a statement about the value of a population parameter
- *Null Hypothesis* (H_0)
 - Usually the hypothesis that the researcher wants to gather evidence against
- *Alternative (or Research) Hypothesis* (H_a)
 - Usually the hypothesis for which the researcher wants to gather supporting evidence

24

Hypothesis Testing (contd)

Example: A researcher studied the relationship between Smoking and Lung cancer.

	Lung Cancer	
	Present	Absent
Smoker	A	B
Non-Smoker	C	D

25

Hypothesis Testing (contd)

H_0 : There is no difference between smokers and nonsmokers with respect to the risk of developing lung cancer. That is, the observed difference (in the sample), if any, is by chance alone.

H_a : There is a difference between smokers and nonsmokers with respect to the risk of developing lung cancer and that the observed difference (in the sample) is not by chance alone.

Conclusion: If the findings of the study are statistically significant, then reject H_0 and fail to reject the alternative hypothesis H_a .

26

Hypothesis Testing (contd)

Test Statistic

- Statistics whose primary use is in testing hypotheses are called test statistics
- Hypothesis testing, thus, involves determining the value the test statistic must attain in order for the test to be declared significant.
- The test statistic is computed from the data of the sample.

27

Hypothesis Testing (contd)

Types of Errors

		Truth	
		H_0 True	H_0 False
Decision	Accept H_0	Correct	Type II error
	Reject H_0	Type I error	Correct

28

Hypothesis Testing (contd)

- **Type I Error**
 - Rejecting the null hypothesis when it is true
 - If H_0 is true in reality and the observed finding of a study is statistically significant, the decision to reject H_0 is incorrect and an error has been made.
- **Type II Error**
 - Failing to reject the null hypothesis when it is false.
 - If in reality H_0 is false and the observed finding of a study is statistically not significant, the decision to accept H_0 is incorrect and an error has been made.

29

Hypothesis Testing (contd)

Alpha (α) = Probability of Type I error; significance level of the test

Beta (β) = Probability of Type II error

Power of a test = $1 - \beta$; probability that a test detects differences that actually exist; typically use 80%

Level of Significance (p-value) in a study:

- Probability of obtaining a result as extreme as or more extreme than the one observed, if the null hypothesis is true
- Probability that the observed result is due to chance alone.
- Most researchers use $p \leq 0.05$ to reject H_0 , and $p > 0.05$ to accept the null hypothesis H_0 and reject the alternative hypothesis H_a .

30

Hypothesis Testing (contd)

One-Sided Test of Hypothesis is one in which the alternative hypothesis is directional (typically includes the '<' symbol or the '>' symbol).

Two-Sided Test of Hypothesis is one in which the alternative hypothesis does not specify departure from the null in a particular direction (typically will be written with the ' \neq ' symbol).

31

One-Sided Test

e.g. Incidence of tuberculosis among Dade county (Miami) residents is known to be no more than 0.0002 (2 cases per 10,000 people). After conducting medical checks, a medical researcher believes that Haitian refugees arriving in Miami have a much higher incidence of tuberculosis. To check this belief, he will test the null hypothesis.

$H_0: \pi = 0.0002$

where π is the proportion of Haitians in Miami who contract TB.

Versus the alternative hypothesis

$H_a: \pi > 0.0002$

because he is interested in detecting whether the true incidence of TB in the Haitian population in Miami is larger than 0.0002.

32

Two-Sided Test

e.g. A researcher would like to determine whether mean age of onset of heart disease in males differs from the mean age for females. The null hypothesis of interest is

$$H_0: \mu_M = \mu_F$$

where μ_M is the mean age of onset of heart disease for males and μ_F is the the mean age of onset for females

Versus the alternative hypothesis

$$H_a: \mu_M \neq \mu_F$$

since she has no reason to believe that one could be higher than the other.

33

One-Sample Tests

One Sample hypothesis tests involve inferences about a single population parameter – based on data from a single sample.

The parameter (mean, proportion) is compared to a single numeric value.

e.g. Hypothesis to test whether the incidence of TB among Haitian refugees is 0.0002.

34

Two-Sample Tests

Two-Sample hypothesis tests involve comparisons of the parameter values between two independent groups .

The parameter (mean, proportion) value is compared between two groups.

Example: Does mean age of onset of heart disease in males differ from the mean age for females?

- Groups: Males versus Females
- Age of onset measured in both groups
- The observations from the sample of males and the sample of females are used to conduct the test

35

Parametric Tests

- Parametric tests are based on assumptions about the distribution of the observed data. (E.g., Normal distribution)
- Hypotheses are formulated in terms of the Mean or the Standard Deviation. Some examples of tests include:

1. **Z-test** (when sample sizes are large, or the population standard deviation is known) used to make inferences about means or proportions
 - Example: Observe serum cholesterol levels among 150 Native Americans in Arizona to study the association with coronary artery disease

36

Parametric Tests (contd)

2. *T-test* (when sample sizes are small $n < 30$, or the population standard deviation is not known and the sample standard deviation is used) to make inferences about means.
- Example: Temperatures of 26 patients were recorded 48 hours after surgery. A researcher is interested in determining if the mean temperatures of the surgical patients are significantly different from the standard normal temperature of 98.6°F.

37

Parametric Tests (contd)

3. *F-test* is used to
- Test hypotheses about a single population standard deviation, or to compare two standard deviations.
 - Compare three or more group means: Analysis of Variance (ANOVA)
 - Example: The four blood groups A, B, O, and AB were studied to compare the quantitative serologic differences among their antigenic structures. Use ANOVA to compare the 4 group means.

38

Parametric Tests (contd)

4. *Chi-Square Test*

- Used for comparing two or more independent proportions within two or more groups – for example when the data are arranged in a 2-by-2 table.
- Example: To assess the possible association between 100% oxygen therapy and the subsequent development of retrolental fibroplasia (RF), a total of 135 premature infants in the intensive care unit were studied.

39

Parametric Tests (contd)

		RF	
		Present	Absent
100% Oxygen	+	36	31
	-	22	46

H_0 : Proportion of infants developing RF is independent of whether they got 100% Oxygen therapy (or proportion developing RF in the +oxygen and –oxygen group are the same)

H_a : Proportion of infants developing RF is dependent on whether they got 100% Oxygen therapy (or proportion developing RF in the +oxygen and –oxygen group are not the same)

40

Nonparametric Tests

Nonparametric tests

- Based on weaker assumptions
- Do not assume a normal distribution
- Called *Distribution-Free Tests*
- Used when the assumption of normality (required for parametric tests) of the data is not met
- Hypotheses may be framed in terms of the Median, quartiles, etc. instead of the Mean scores

41

Nonparametric Tests (contd)

- Sign Test
 - Used to test hypotheses about the Median of a population
 - One-Sample Test
- Wilcoxon Rank-Sum Test
 - Used to compare Medians of two groups
 - Two-Sample Test
- Fisher's Exact test
 - Nonparametric equivalent of the Chi-square test
 - Used when the expected frequencies in a table are small (<5)

42

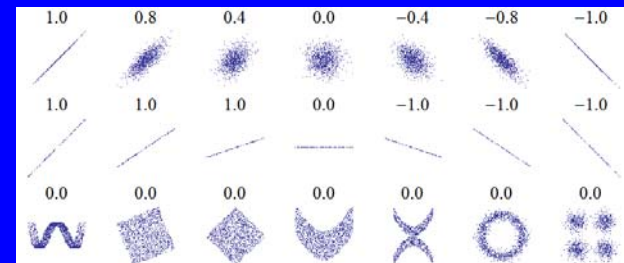
Paired Data

- Paired T-test
 - Tests hypotheses about Mean change in a population
 - Example:
 - Test hypotheses about the mean reduction in weight for a group of subjects in a new weight loss program
 - Test whether the mean difference (change in weight) is greater than zero
- Nonparametric test for paired data – Wilcoxon Signed-Rank Test
 - To test the hypothesis that the medians, rather than the means, are equal in the two paired samples.

43

Association and Prediction

- Correlation Coefficient (r)
 - Measure of the strength of the linear relationship between two variables measured on a numerical scale
 - Ranges from -1.0 to +1.0
 - Positive and negative correlation.
 - Example: Interested in the correlation between cholesterol and triglyceride levels ($r = +1, +0.8, 0, -0.8, -1$)



44

Association and Prediction

- Pearson's Product Moment Correlation Coefficient (parametric)
- Spearman's Rank Correlation Coefficient (nonparametric)
- Coefficient of Determination (r^2)
 $r = 0.8$, $r^2 = (0.8)(0.8) = 0.64$
or 64% of the variation in the values for cholesterol are explained by changes in triglyceride levels

45

Prediction

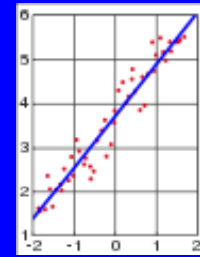
Regression Analysis

- Models the relationship between two or more variables such that one can be expressed in terms of the other variables (mathematical equation)

$$Y = aX + b$$

$$Z = aX + bY + c$$

- Dependent variable/ Response variable
- Independent variable / Explanatory variable
- Linear Regression
 - Example: MCAT (Science) and ACT scores ($Y = aX + b$)



46

Prediction

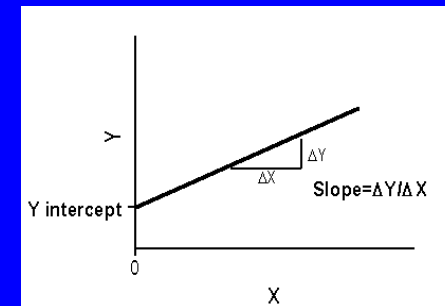
- Simple Linear Regression – only one explanatory variable

Example: MCAT(Science) and ACT scores for 42 medical school applicants ($Y = aX + b$)

- Least squares method used to estimate the regression coefficients 'a' and 'b'
- Regression equation: $\hat{Y} = 0.41X - 161$
- Multiple Regression – two or more explanatory variables

47

Prediction



48

Conclusions

- Summary of some Statistical Methods used in Medical research presented
- The objective is for you to be able to recognize the various methods and understand/interpret the statistical analyses/results presented in published articles – NOT for you to conduct the statistical analyses
- It is highly recommended & encouraged that you work with a trained Biostatistician on your research projects!

49