

RHIT - Department of Mathematics

Topics Outline for MA223 Engineering Statistics I - 2005-06

Text and Text Materials

Applied Statistics for Engineers and Scientists, 2nd edition -Jay Devore (Cal Poly) and Nicholas Farnum (Cal State Fullerton)

Topics

1.1: Students should know following definitions:

- 1) population (or, more specifically, target population)
- 2) census
- 3) sample
- 4) univariate, bivariate and multivariate observations
- 5) two types of statistical methods: descriptive, inferential
- 6) process

Distribution Supplement

Since the author does not provide or define the term 'distribution' in 1.2 or elsewhere, we have written a supplement which defines what a (univariate) distribution is and discusses its use as a means of summarizing the useful information in a data set. We suggest students read this supplement before reading section 1.2. Students should understand stationarity.

1.2: This section discusses three methods for visualizing/graphing the distribution of a univariate data set. Students should know how to create (using by-hand or in Minitab) and interpret the following graphical depictions of a univariate distribution:

- 1) Stem-and-leaf
- 2) Dotplot
- 3) Histogram for continuous (equal-width intervals *only*), discrete, and categorical variables
- 4) Students should know how to assess skew of a distribution

1.3: Section 1.3 presents density functions and probability mass functions as mathematical descriptions (models) of the distributions of continuous and discrete data, respectively. From this section, students should know:

- 1) When to describe data as continuous or discrete
- 2) Defining properties of density functions and probability mass functions.
- 3) How to compute proportion of values in a given region.

1.4: After this section is covered, students should

- 1) Know overall visual character of normal distribution; see Fig 1.19
- 2) Know definition of standard normal distribution
- 3) Know how to compute standard normal proportions using standard normal cumulative distribution function and standard normal cumulative distribution table
- 4) Know general definition of cumulative distribution function (CDF) for continuous distributions
- 5) Know how to compute proportions for nonstandard normal distributions using standardization

1.6: In this section, the binomial distribution is informally derived. After this section is presented, students should

- 1) Know when binomial distribution is appropriate model.
- 2) Know how to use formula to compute proportions, for small and large n

2.1: Students should be able to

- 1) Compute the sample mean, sample median, and trimmed mean given a set of data
- 2) Compute the mean for a given density function or probability mass function.

- 3) Compute the mean of the binomial distribution using the shortcut formula
- 4) Compute the median of a continuous distribution given a density function.
- 5) Discuss which measure of center is more appropriate for sample data that is symmetric, positively skewed, negatively skewed, uniform, etc.

2.2: Students should be able to

- 1) Compute the sample variance and standard deviation by hand given a small data set using formula on page 71
- 2) Compute the standard deviation and variance of a distribution.
- 3) Compute the standard deviation and variance of the binomial using the shortcut formulas
- 4) Know and use the empirical rule (page 76)

2.3: Students should

- 1) Know definition of quartiles and be able to compute SAMPLE quartiles from data using following procedure:
 2nd quartile(Q2) = median: sample median equals middle value if n odd; equals average of middle values, n even
 1st quartile(Q1): sample 1st quartile equals median of data less than or equal to sample median
 3rd quartile(Q3): sample 3rd quartile equals median of data greater than or equal to sample median
 These procedures agree with author's: median (bottom page 61), quartiles (bottom page 79)
 ***NOTE: Formula for 1st and 3rd quartiles differs from Minitab; see example 2.11 page 80.
- 2) Know that Interquartile Range (IQR): $Q3 - Q1$ is measure of spread like standard deviation.
- 3) Know that boxplot is graphical depiction of five-number summary (min, max, Q1, Q2, and Q3) and be able to estimate min, max, Q1, Q2, and Q3 from boxplot
- 4) Know and be able to use boxplot outlier criteria (top page 84)
- 5) Know definition of percentiles and be able to compute for standard normal distribution
- 6) Be able to determine the skewness of a distribution from its boxplot

2.4: Students should

- 1) understand use of quantile plot to determine if distribution has specified form, i.e., if quantile plot is linear, then there is a good match between sample data and distribution.
- 2) Students should be able to use normal Quantile plot to assess if data has normal distribution.

3.1: Students should

- 1) Know univariate, bivariate, and multivariate data (from 1.1)
- 2) Know how to construct and interpret a scatter plot.

3.2: Students should

- 1) Know general definition of a correlation coefficient, i.e., "A numerical summary of the strength of relationship between two or more variables in a data set."
- 2) Know formula for Pearson's sample correlation coefficient r
- 3) Know properties and interpretation of r provided on pages 109-110.
- 4) Know definition of population correlation coefficient and properties (page 112).
- 5) Know that correlation/association does not imply causation.
 Example: There is a correlation between gas tank size and mpg. This does not imply that giving your car a smaller gas tank will appreciably increase its mpg.

3.3: Students should

- 1) Be able to identify response variable and predictor or explanatory variable(s) (covariates). The terms dependent and independent variable have somewhat fallen out of favor because in many situations the variables dependent on one another.
- 2) Understand the principle of least squares
- 3) Be shown the fact that the least squares line is computed by solving linear equations derived from 1st derivatives.
- 4) Be able to interpret least squares coefficients, e.g., slope and intercept.
- 5) Be able to use least squares line to predict value of response variable given value of predictor variable.
- 6) Know danger of extrapolation (bottom page 117)

- 7) Know definition of residual (top of page 119)
- 8) Know how to use plot of residuals to assess fit of least squares lines (pages 123-125).
- 9) Know how to compute the coefficient of determination and understand how to interpret it.
- 10) Compute the standard deviation, se , about the least squares regression line and understand how to interpret it
- 11) Computing r , r^2 , se can all be done via Minitab; students should know how to read Minitab output for simple linear regression and determine the values r , r^2 , se from the output

3.5: (Optional: Instructor may choose to cover this section; will not be included in final exam material)
Students should

- 1) Know how to construct and interpret scatterplot matrix (Figure 3.18, pg. 138).
- 2) Know how to interpret and use least squares model with multiple predictors.

4.1: Students should

- 1) Understand the concept of an operational definition and be able to produce/critique operational definitions.
- 2) Understand what a random sample is.

4.2: Students should

- 1) Know that a sample is simply a subset of the population of interest (target population)
- 2) Know that sample size is typically denoted by lower case n .
- 3) Know advantages of using sampling as opposed to examining entire population (census)
- 4) Understand when a sample is representative
- 5) Know the difference between sampling with and without replacement
- 6) Know rationale for random sampling, namely, that it allows us to make statements about the population (using the sample data) with a mathematically specified degree of reliability or confidence.
- 7) Know definition of simple random sampling and be able to tell if sampling procedure will produce simple random samples (SRS's)
- 8) Know key properties of simple random samples:
 - i. Easy to analyze mathematically
 - ii. The larger the size of the sample, the greater the likelihood SRS is representative of population
- 9) Know rationale of stratified sampling and understand what stratified sampling is.
- 10) Be able to recognize when sampling procedure is stratified sampling and identify the strata.
- 11) Be able to determine if stratified sampling will offer significant improvements over simple random sampling.

5.1: Students should:

- 1) Know what a statistic is. (page 191)
- 2) Understand what is meant by the term "statistical inference." (page 192) Statistical inference is essentially the act of making a statement about a population based on examining only a small part of it (the sample), i.e., extrapolating from the sample to the population.
- 3) Know that probability theory is the branch of mathematics concerned with modeling random phenomena, phenomena in which like causes produce unlike effects (outcomes), e.g., tossing a coin. Since similar causes produce dissimilar outcomes, we say the outcomes are determined by (due to) chance. Thus the author calls random phenomena "chance experiments."
- 4) Know that since SRS's are the outcome of a random process; probability theory provides the tools for describing the relationship between the SRS and the population and the relationship between a sample statistic (example: sample mean) and the corresponding population characteristic (population mean).
- 5) Know that outcomes of a random process/chance experiment can be divided into two types:
 - (i) SIMPLE EVENTS -- events which are the individual outcomes of an experiment.
 - (ii) EVENTS -- collections (sets) of simple events
- 6) Know that sample space of a chance experiment, which we denote by "S," is the set of ALL simple events for that experiment.
- 7) Be able to construct and/or interpret a tree diagram depiction of a chance experiment.
- 8) Be able to construct and/or interpret a Venn diagram of a sample space.
- 9) Understand and be able to use set operations "or," "and," and "Complement" to combine events (sets of

simple events) to create new events (sets of simple events).

10) Know the common symbols used to denote the intersection (\cap) and union (\cup) of events

11) Be able to determine when two events are mutually exclusive or disjoint.

5.2: Students should:

1) Know the two INTERPRETATIONS of probability which dominate the field of statistics:

(i) Frequentist (objective) interpretation of probability: The probability of a specific outcome A, $P(A)$, of a chance experiment is defined as the limiting value of the ratio

$$P(A) = \frac{\text{number of times A occurs}}{\text{number of times experiment performed}}$$

as the number of times the experiment is performed goes to infinity. Note that $P(A)$ is a constant of nature whose value is independent of the observer of the experiment and is therefore objective.

(ii) Bayesian (subjective) interpretation of probability: The probability of some statement about the nature of A, $P(A)$, is a measure of the strength of belief that A is true. Since different individuals will generally have different degrees of belief concerning the veracity, the value of $P(A)$ will vary from individual to individual and is therefore subjective.

2) Know that although the subjective interpretation of probability is, for most people, the more natural, the course will use the frequentist interpretation of probability.

3) Know the Probability Axioms, which hold for all interpretations of probability; see page 199.

4) Know and be able to use the addition rule for disjoint events (pg. 200).

5) Know and be able to use the complement rule (pg. 202).

6) Know and be able to use the general addition rule (pg. 203).

5.3: Student should:

1) Understand conditional probability conceptually.

2) Know definition of conditional probability (page 205).

3) Understand independence conceptually.

4) Be able to determine if two events are independent.

5) Know definition of independence.

6) Know how to compute the probability that at least one of two or more independent events occurs (page 207-208).

7) Know how to compute the probability that a system consisting of components IN SERIES and IN PARALLEL will function.

8) Know the difference between mutually exclusive events and independent events.

5.4: Student should:

1) Know that random variable (RV) is a function that assigns numbers to outcomes of the sample space.

2) Know definition of DISCRETE and CONTINUOUS random variables (RVs). (page 213).

3) Know how to define events in terms of RVs; see discussion, page 213.

4) Know that distribution functions (density functions and probability mass functions) are used to assign probabilities to events involving RVs.

5) Know how to compute probabilities of events involving RVs using distribution functions and rules from sections 5.2 and 5.3.

6) Know how to compute the mean, variance, and standard deviation of a random variable. Note that these quantities are simply the mean, variance, and standard deviation of the distribution of the random variable, respectively; see page 216.

Linear Combination Handout

1) Know definition of independent random variables.

2) Know definition of linear combinations of random variables.

3) Know how to compute mean, variance, and standard deviation of INDEPENDENT random variables.

4) Know that linear combination of random variables is normally distributed IF AND ONLY IF those

random variables ARE ALL normally distributed.

5) Know how to use above properties to solve problems involving linear combinations of RVs.

5.5: Sampling Distributions

1) Understand how a sampling distribution is obtained (for sample mean, sample median, sample range); we used the on-line applet on the Rice site (http://www.ruf.rice.edu/~lane/stat_sim/sampling_dist/) to

explore distributions of \bar{x}

2) Understand that when n is large, the distribution of \bar{x} is normal, no matter what the distribution of the original distribution looks like

5.6: Describing Sampling Distributions

1) Determine the distribution of \bar{x} ; discuss the Central Limit Theorem

2) Compute probabilities involving \bar{x}

SKIP CHAPTER 6

Chapter 7: Throughout Chapter 7, students must understand that process data needs to be stationary before constructing a confidence interval on it.

Note: Teaching chapters 7 and 8 together (e.g., introduce confidence interval, then hypothesis test or vice-versa) is worth considering and provides a nice connection between CIs and hypothesis testing.

7.1: (optional per instructor)

7.2: Students should know:

- 1) How to build a confidence interval for one population mean given a large sample
 - 2) Interpret a confidence interval (Figure 7.5, pages 297-298)
 - 3) Determine the sample size n given a confidence level, error bound, and sample standard deviation
- Optional: One sided confidence intervals (or confidence bounds)

7.3: Students should know:

- 1) How to build a confidence interval for a population proportion
NOTE: Do not use book's formula, use Wilson estimation, or should we just way build a confidence interval for a large sample size???
- 2) What conditions to check in order to assure that a confidence interval for the population proportion is "legal"
- 3) Be able to determine the sample size n given a confidence level and error bound, and sometimes knowing an idea of the true population proportion (for means to estimate π)
- 4) How to build a confidence interval for the difference of two population means given large samples; use ideas of linear combinations to establish the sample mean and sample standard deviation of $\bar{x}_1 - \bar{x}_2$
- 5) How to interpret a confidence interval for the difference of two population means

Section 8.4 covers checking the normality condition of data; it is suggested that this section be covered before 8.1; or actually even before 7.4

7.4: Students should know:

- 1) How to build a confidence interval for one population mean given a small sample
- 2) How to read a t-table
- 3) How to check for normally distributed data via Ryan-Joiner normality test in Minitab
- 4) The conditions to check in order for a t confidence interval to be valid (normality, population standard deviations unknown)
- 5) When to use a z critical value versus a t critical value when constructing a confidence interval for one

population mean
Skip prediction intervals
Skip tolerance intervals

7.5: Students should know:

- 1) How to build a confidence interval for the difference of two population means given small samples assuming unequal population variances ONLY
- 2) The difference between dependent (paired) and independent two sample data
- 3) How to build a confidence interval for paired data
- 4) How to check normality conditions on the difference of paired data

7.6: Students should know:

- 1) The rationale/process behind building a bootstrap confidence interval
 - 2) How to use Minitab macros (provided by us: bootmean.mac and boot2mean.mac) to build a bootstrap confidence interval for one population mean and two population means
- Skip other topics in Section 7.6

From Chapter 7, students should know the conditions that must be met to build the various confidence intervals, how to check that those conditions are met, and when to use which confidence interval construction method

In Chapter 7, students should know how to build these intervals in Minitab, and check appropriate conditions in Minitab.

Chapter 8: Throughout Chapter 8, students must understand that process data needs to be stationary before constructing a confidence interval on it.

Understand the relationship between confidence intervals and hypothesis testing

Section 8.4 covers checking the normality condition of data; it is suggested that this section be covered before 8.1; or actually even before 7.4

8.1: Students should know:

- 1) Hypothesis testing lingo (null, alternative, Type I Error, Type II Error, p-values, left/right tailed tests, reject or fail to reject the null hypothesis, etc.) and how to set up a hypothesis test (in accordance with this test, null hypothesis is always $=$)
- 2) How to compute Type I Error
- 3) How to compute Type II Error
- 4) How to compute a p-value; and how to compare the p-value (observed significance level) with α (the given significance level for the test)
- 5) When to reject or fail to reject the null hypothesis according to the p-value and α
- 6) When to use a right tailed, left tailed, or two-tailed test

8.2: Students should know:

- 1) How to run a hypothesis test on a single population mean; they need to know how to check for certain conditions (stationarity for process data, normality), what is the appropriate test to run (z, t, bootstrapping), and how to run it
- 2) How to run a hypothesis test on the difference of two population means with independent samples; they need to know how to check for certain conditions (stationarity for process data, normality), what is the appropriate test to run (z, t with unequal variances only, bootstrapping), and how to run it
- 3) How to run a hypothesis test on the difference of two population means with dependent samples; i.e., paired data; they need to know how to check for certain conditions (stationarity for process data, normality), what is the appropriate test to run (t, bootstrapping), and how to run it

8.3: Optional

8.4: Students should know:

- 1) How to determine if sample data is normally distributed via the Ryan-Joiner test
Skip Chi-squared tests

9.1: Students should know:

- 1) The basic terminology and concepts (why we are analyzing variance instead of comparing means, variance within, variance between, how the test statistic is constructed) of ANOVA
- 2) How to determine the degrees of freedom appropriate for using an F table
- 3) How to read an F table

9.2: Students should know:

- 1) ANOVA assumptions and how to check if these assumptions are met (via Minitab)
- 2) How the test statistic is constructed (via SSE, SST, SSTR, etc), though these values can be obtained from Minitab output or students should know how to read ANOVA Minitab output to determine them
- 3) How to read and fill out an ANOVA table
- 4) How to set up a hypothesis test for ANOVA

SKIP Chapter 10

11.1: Students should know:

- 1) The difference between the sample regression line and the population regression line
- 2) Estimates of the true population slope and intercept
Skip exponential regression

11.2: Students should know:

- 1) Properties of the sampling distribution of the sample slope b
- 2) The basics about how the test statistic for the population slope β is constructed, and why it makes sense; suggested: Rossman/Chance applet:
<http://statweb.calpoly.edu/chance/applets/regcoeff/regcoeff.html>
- 3) How to use Minitab to run a hypothesis test on the population slope β
- 4) How to construct a confidence interval for β
- 5) Regression and ANOVA (Optional)