Math 221 Objectives

1 PART I EXPLORING DATA

1.1 DATA

1.1.1 Identify the individuals and variables in a set of data.

1.1.2 Identify each variable as categorical or quantitative.

1.1.3 Identify the units in which a quantitative variable is measured

1.1.4 In situations where one variable explains or influences another, identify the explanatory (independent) and response (dependent) variables.

1.2 DISPLAYING DISTRIBUTIONS

1.2.1 Recognize when use of a pie chart is appropriate.

1.2.2 Make a bar graph of the distribution of a categorical variable.

1.2.3 Interpret pie charts and bar graphs.

1.2.4 Make a histogram of the distribution of a quantitative variable.

1.2.5 Make a stemplot of the distribution of a small set of observations. Round leaves or split stems as needed to make an effective stemplot.

1.3 DESCRIBING DISTRIBUTIONS (QUANTITATIVE VARIABLE)

1.3.1 Assess from a histogram or stemplot whether the shape of a distribution is roughly symmetric, distinctly skewed, uniform, or neither.

1.3.2 Describe the overall pattern by giving numerical measures of center (mean, median, mode) and spread (range, IQR, standard deviation, variance) in addition to a verbal description of shape.

1.3.3 Decide which measures of center and spread are more appropriate for a given set of data: the mean and standard deviation (especially for symmetric distributions) or the five-number summary (especially for skewed distributions).

1.3.4 Recognize outliers and give plausible explanations for them.

1.4 NUMERICAL SUMMARIES OF DISTRIBUTIONS

1.4.1 Find the median M and the quartiles Q_1 and Q_3 for a set of observations.

1.4.2 Find the five-number summary and draw a boxplot; assess center, spread, symmetry, and skewness from a boxplot.

1.4.3 Find the mean $\bar{x}$ and the standard deviation s for a set of observations.

1.4.4 Understand that the median is more resistant than the mean. Recognize that skewness in a distribution moves the mean away from the median toward the long tail.

1.4.5 Know the basic properties of the standard deviation:

- $s \geqslant 0$;
- $s = 0$ only when all observations are identical;
- s increases as the spread increases;
- s has the same units as the original measurements;
- s is pulled strongly up by outliers or skewness.

1.5 DENSITY CURVES AND NORMAL DISTRIBUTIONS

1.5.1 Know that areas under a density curve represent proportions of all observations and that the total area under a density curve is 1.

1.5.2 Approximately locate the median (equal-areas point) and the mean (balance point) on a density curve. Know that the mean and median both lie at the center of a symmetric density curve and that the mean moves further toward the long tail of a skewed curve.

1.5.3 Recognize the shape of Normal curves and estimate by eye both the mean and standard deviation from such a curve.

1.5.4 Use the 68-95-99.7 rule and symmetry to state what percent of the observations from a Normal distribution fall between two points when both points lie one, two, or three standard deviations on either side of the mean.

1.5.5 Find the standardized value (z-score) of an observation. Interpret z-scores and understand that any Normal distribution becomes the standard Normal $\mathrm{N}(0, 1)$ when standardized.

1.5.6 Calculate the proportion of values above a stated number, below a stated number, or between two stated numbers when given that a variable has a Normal distribution with a stated mean μ and standard deviation σ.

1.5.7 Calculate the point having a stated proportion of all values above it or below it when given that a variable has a Normal distribution with a stated mean μ and standard deviation σ.

1.6 SCATTERPLOTS AND CORRELATION

1.6.1 Make a scatterplot to display the relationship between two quantitative variables measured on the same subjects. Place the explanatory variable (if any) on the horizontal scale of the plot.

1.6.2 Describe the direction, form, and strength of the overall pattern of a scatterplot. In particular, recognize positive or negative association and linear (straight-line) patterns. Recognize outliers in a scatterplot.

1.6.3 Judge whether it is appropriate to use linear correlation to describe the relationship between two quantitative variables.

1.6.4 Find the correlation r.

1.6.5 Know the basic properties of correlation:

- r measures the direction and strength of only straight-line relationships;
- r is always a number between -1 and 1;
- $r = \pm 1$ only for perfect straight-line relationships;
- r moves away from 0 toward ± 1 as the straight-line relationship gets stronger.

1.7 REGRESSION LINES

1.7.1 Understand that regression requires an explanatory variable and a response variable. Use a calculator or software to find the least-squares regression line of a response variable y on an explanatory variable x from data.

1.7.2 Explain what the slope b_1 and the intercept b_0 mean in the equation $\hat{y} = b_0 + b_1 x$ of a regression line.

1.7.3 Draw a graph of a regression line when you are given its equation.

1.7.4 Use a regression line to predict y for a given x. Recognize extrapolation and know potential dangers when using.

1.7.5 Find the slope and intercept of the least-squares regression line from the means and standard deviations of x and y and their correlation.

1.7.6 Use r^2 (coefficient of determination) the square of the correlation, to describe how much of the variation in one variable can be accounted for by a straight-line relationship with another variable.

1.7.7 Recognize outliers and potentially influential observations from a scatterplot with the regression line drawn on it.

1.7.8 Calculate the residuals and plot them against the explanatory variable x. Recognize that a residual plot magnifies the pattern of the scatterplot of y versus x. (Optional)

1.8 CAUTIONS ABOUT CORRELATION AND REGRESSION

1.8.1 Understand that both r and the least-squares regression line can be strongly influenced by a few extreme observations.

1.8.2 Recognize possible lurking variables that may explain the observed association between two variables x and y.

1.8.3 Understand that even a strong correlation does not mean that there is a cause-and-effect relationship between x and y.

1.8.4 Give plausible explanations for an observed association between two variables: direct cause and effect, the influence of lurking variables, or both.

2 PART II INFERENCE

2.1 SAMPLING

2.1.1 Identify the population in a sampling situation.

2.1.2 Recognize bias due to voluntary response samples and other inferior sampling methods.

2.1.3 Use software or a table of random digits to select a simple random sample (SRS) from a population.

2.1.4 Recognize the presence of undercoverage and nonresponse as sources of error in a sample survey. Recognize the effect of the wording of questions on the responses.

2.2 EXPERIMENTS

2.2.1 Recognize whether a study is an observational study or an experiment.

2.2.2 Recognize bias due to confounding of explanatory variables with lurking variables in either an observational study or an experiment.

2.2.3 Identify the factors (explanatory variables), treatments, response variables, and individuals or subjects in an experiment.

2.2.4 Outline the design of a completely randomized experiment using a diagram. The diagram in a specific case should show the sizes of the groups, the specific treatments, and the response variable.

2.2.5 Use software or a Table of random digits to carry out the random assignment of subjects to groups in a completely randomized experiment.

2.2.6 Recognize the placebo effect. Recognize when the double-blind technique should be used.

2.2.7 Explain why randomized comparative experiments can give good evidence for cause-and-effect relationships.

2.3 PROBABILITY

2.3.1 Recognize that some phenomena are random and that probability describes the long-run regularity of random phenomena.

2.3.2 Understand that the probability of an event is the proportion of times the event occurs in very many repetitions of a random phenomenon.

2.3.3 Use basic probability rules to detect illegitimate assignments of probability: Any probability must be a number between 0 and 1, and the total probability assigned to all possible outcomes must be 1.

2.3.4 Use basic probability rules to find the probabilities of events that are formed from other events. Know and use appropriately the probability that an event does not occur is 1 minus its probability; and if two events are disjoint, the probability that one or the other occurs is the sum of their individual probabilities.

2.3.5 Find probabilities in a discrete probability model by adding the probabilities of their outcomes. Find probabilities in a continuous probability model as areas under a density curve.

2.3.6 Use the notation of random variables to make compact statements about random outcomes, such as $P(\overline{X} \leqslant 4) = 0.3$. Be able to interpret such statements.

2.4 SAMPLING DISTRIBUTIONS

2.4.1 Identify parameters and statistics in a statistical study.

2.4.2 Recognize the fact of sampling variability: a statistic will take different values when you repeat a sample or experiment.

2.4.3 Interpret a sampling distribution as describing the values taken by a statistic in all possible repetitions of a sample or experiment under the same conditions.

2.4.4 Interpret the sampling distribution of a statistic as describing the probabilities of its possible values.

2.5 THE SAMPLING DISTRIBUTION OF A SAMPLE MEAN

2.5.1 Recognize when a problem involves the mean $\bar{x}$ of a sample and understand that $\bar{x}$ estimates the mean μ of the population from which the sample is drawn.

2.5.2 Use the law of large numbers to describe the behavior of $\bar{x}$ as the size of the sample increases.

2.5.3 Find the mean and standard deviation of a sampling mean distribution from an SRS of size n when the mean μ and standard deviation σ of the population are known.

2.5.4 Understand that $\bar{x}$ is an unbiased estimator of μ and that the variability of $\bar{x}$ about its mean μ gets smaller as the sample size increases.

2.5.5 Understand that $\bar{x}$ has approximately a Normal distribution when the sample is large (central limit theorem). Use this Normal distribution to calculate probabilities that concern $\bar{x}$.

2.6 GENERAL RULES OF PROBABILITY

2.6.1 Use Venn diagrams to picture relationships among several events.

2.6.2 Use the general addition rule to find probabilities that involve overlapping events.

2.6.3 Understand the idea of independence. Judge when it is reasonable to assume independence as part of a probability model.

2.6.4 Use the multiplication rule for independent events to find the probability that all of several independent events occur.

2.6.5 Use the multiplication rule for independent events in combination with other probability rules to find the probabilities of complex events.

2.6.6 Understand the idea of conditional probability. Find conditional probabilities for individuals chosen at random from a table of counts of possible outcomes.

2.6.7 Use the general multiplication rule to find $P(A \text{ and } B)$ from $P(A)$ and the conditional probability $P(B|A)$.

2.7 BINOMIAL DISTRIBUTIONS

2.7.1 Recognize the binomial setting: (i.e. a fixed number n of independent success-failure trials with the same probability p of success on each trial.)

2.7.2 Recognize and use the binomial distribution of the count of successes in a binomial setting.

2.7.3 Use the binomial probability formula to find probabilities of events involving the count x of successes in a binomial setting for small values of n.

2.7.4 Find the mean and standard deviation of a binomial count X.

2.7.5 Recognize when it is appropriate to use the Normal approximation to a binomial distribution. Use the Normal approximation to calculate probabilities that concern a binomial count X. (Optional)

2.8 CONFIDENCE INTERVALS

2.8.1 State in nontechnical language what is meant by "95% confidence" or other statements of confidence in statistical reports.

2.8.2 Calculate a confidence interval for the mean μ of a Normal population with known standard deviation σ, using the formula $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$.

2.8.3 Understand how the margin of error of a confidence interval changes with the sample size and the level of confidence.

2.8.4 Find the sample size required to obtain a confidence interval of specified margin of error when the confidence level for the mean and other information are given.

2.8.5 Identify sources of error in a study that are not included in the margin of error of a confidence interval, (such as undercoverage, nonresponse, or voluntary sample).

2.9 SIGNIFICANCE TESTS

2.9.1 State the null and alternative hypotheses in a testing situation when the parameter in question is a population mean μ.

2.9.2 Explain in nontechnical language the meaning of the P-value when you are given the numerical value of P for a test.

2.9.3 Identify Type I and Type II errors for a given situation.

2.9.4 Calculate the one-sample z test statistic and the P value for both one-sided and two-sided tests about the mean μ of a Normal population.

2.9.5 Assess statistical significance at standard levels α, either by comparing P with α or by comparing z with standard Normal critical values.

2.9.6 Draw the appropriate conclusion for a hypothesis test (test of significance) and interpret the meaning of the results.

2.9.7 Recognize that significance testing does not measure the size or importance of an effect. Explain why a small effect can be significant in a large sample and why a large effect can fail to be significant in a small sample.

2.9.8 Recognize that any inference procedure acts as if the data were properly produced. The z confidence interval and test require that the data be an SRS from the population.

3 PART III (INFERENCE ABOUT VARIABLES)

3.1 RECOGNITION

3.1.1 Recognize when a problem requires inference about population means (quantitative response variable) or population proportions (usually categorical response variable).

3.1.2 Recognize from the design of a study whether one-sample, matched pairs, or two-sample procedures are needed.

3.1.3 Based on recognizing the problem setting, correctly choose among the one- and two-sample t procedures for means and the one- and two-sample z procedures for proportions.

3.2 INFERENCE ABOUT ONE MEAN

3.2.1 Verify that the procedures are appropriate in a particular setting. Check the study design and the distribution of the data and take advantage of robustness against lack of Normality.

3.2.2 Recognize when poor study design, outliers, or a small sample from a skewed distribution make the t procedures risky.

3.2.3 Use the one-sample t procedure to obtain a confidence interval at a stated level of confidence for the mean μ of a population.

3.2.4 Carry out a one-sample t test for the hypothesis that a population mean μ has a specified value against either a one-sided or a two-sided alternative. Use software to find the P-value or t-distribution table to get an approximate value.

3.2.5 Recognize matched pairs data and use the t procedures to obtain confidence intervals and to perform tests of significance for such data.

3.3 COMPARING TWO MEANS

3.3.1 Verify that the two-sample procedures are appropriate in a particular setting. Check the study design and the distribution of the data and take advantage of robustness against lack of Normality.

3.3.2 Give a confidence interval for the difference between two means. Use software if you have it. Otherwise, use the two-sample t statistic with conservative degrees of freedom and t Distribution Table.

3.3.3 Test the hypothesis that two populations have equal means against either a one-sided or a two-sided alternative. Use software if you have it. Otherwise, use the two-sample t test with conservative degrees of freedom and t Distribution Table.

3.3.4 Know that procedures for comparing the standard deviations of two Normal populations are available, but that these procedures are risky because they are not at all robust against non-Normal distributions.

3.4 INFERENCE ABOUT ONE PROPORTION

3.4.1 Verify that you can safely use either the large-sample or the plus four z procedures in a particular setting. Check the study design and the guidelines for sample size.

3.4.2 Use the large-sample z procedure to give a confidence interval for a population proportion p. Understand that the true confidence level may be substantially less than what is asked for unless the sample is very large and the true p is not close to 0 or 1.

3.4.3 Use the plus four modification of the z procedure to give a confidence interval for p that is accurate even for small samples and for any value of p.

3.4.4 Use the z statistic to carry out a test of significance for the hypothesis $H_0 : p = p_0$ about a population proportion p against either a one-sided or a two-sided alternative. Use software or the Standard Normal Table to find the P-value.

3.5 COMPARING TWO PROPORTIONS

3.5.1 Verify that you can safely use either the large-sample or the plus four z procedures in a particular setting. Check the study design and the guidelines for sample sizes.

3.5.2 Use the large-sample z procedure to give a confidence interval for the difference p1 - p2 between proportions in two populations based on independent samples from the populations. Understand that the true confidence level may be less than what is asked for unless the samples are quite large.

3.5.3 Use the plus four modification of the z procedure to give a confidence interval for p1 - p2 that is accurate even for very small samples and for any values of p1 and p2.

3.5.4 Use a z statistic to test the hypothesis $H_0 : p_1 = p_2$ that proportions in two distinct populations are equal. Use software or Standard Normal Table to find the P-value.

4 PART IV INFERENCE OF CATEGORICAL DATA

4.1 GOODNESS OF FIT (Multiple proportions - multinomial)

4.1.1 Recognize the difference between a Goodness of Fit situation and a Two-Way (contingency) Table problem

4.1.2 State the null and alternative hypotheses for a given Goodness of Fit or Multinomial problem.

4.1.3 Find the expected values in a Multinomial or Goodness of Fit problem

4.2 CATEGORICAL DATA

4.2.1 Find the marginal distributions of both variables by obtaining the row sums and column sums when given a two-way table of counts.

4.2.2 Express any distribution in percents by dividing the category counts by their total.

4.2.3 Describe the relationship between two categorical variables by computing and comparing percents. Including comparing the conditional distributions of one variable for the different categories of the other variable.

4.2.4 Understand and explain Simpson's Paradox

4.3 TWO CATEGORICAL VARIABLES (TWO-WAY TABLES)

4.3.1 Understand that the data for a chi-square test must be presented as a two-way table of counts of outcomes.

4.3.2 Use percents to describe the relationship between any two categorical variables, starting from the counts in a two-way table.

4.4 INTERPRETING CHI-SQUARE TESTS

4.4.1 Locate the chi-square statistic, its P-value, and other useful facts (row or column percents, expected counts, terms of chi-square) in output from your software or calculator.

4.4.2 Use the expected counts to check whether you can safely use the chi- square test.

4.4.3 In a goodness of fit situation or a specific two-way table, explain what the null and alternative hypotheses are testing.

4.4.4 If the test is significant, compare percents, compare observed with expected cell counts, or look for the largest terms of the chi-square statistic to see what deviations from the null hypothesis are most important.

4.5 DOING CHI-SQUARE TESTS BY HAND

4.5.1 Calculate the expected count for any cell from the observed counts in goodness of fit situation and with a two-way table. Check whether you can safely use the chi-square test in each case.

4.5.2 Calculate the term of the chi-square statistic for any cell, as well as the overall statistic.

4.5.3 Give the degrees of freedom of a chi-square statistic. Make a quick assessment of the significance of the statistic by comparing the observed value with the degrees of freedom.

4.5.4 Use the chi-square critical values in the Chi-Square Table or use software to approximate the P-value of a chi-square test.

5 INFERENCE OF REGRESSION

5.1 PRELIMINARIES

5.1.1 Make a scatterplot to show the relationship between an explanatory and a response variable.

5.1.2 Use a calculator or software to find the correlation coefficient and the equation of the least-squares regression line.

5.2 RECOGNITION

5.2.1 Recognize the regression setting: a straight-line relationship between an explanatory variable x and a response variable y.

5.2.2 Inspect the data to recognize situations in which inference is not safe: a nonlinear relationship, influential observations, strongly skewed residuals in a small sample, or nonconstant variation of the data points about the regression line.

5.3 INFERENCE USING SOFTWARE OUTPUT

5.3.1 Explain in any specific regression setting the meaning of the slope of the population regression line.

5.3.2 Understand software output for regression. Find in the output the slope and intercept of the least-squares line, their standard errors, and the regression standard error.

5.3.3 Use that information to carry out a test for a significant linear relationship ($H_0 : \beta = 0$).

5.3.4 Calculate confidence intervals for β (optional).

6 ONE WAY ANALYSIS OF VARIANCE: COMPARING SEVERAL MEANS

6.1 RECOGNITION

6.1.1 Recognize when testing the equality of several means is helpful in understanding data.

6.1.2 Recognize that the statistical significance of differences among sample means depends on the sizes of the samples and on how much variation there is within the samples.

6.1.3 Recognize when you can safely use ANOVA to compare means. (Check the data production, the presence of outliers, and the sample standard deviations for the groups you want to compare.)

6.2 INTERPRETING ANOVA

6.2.1 Explain what the null hypothesis is in an ANOVA test.

6.2.2 Locate the F statistic and its P-value on the output of analysis of variance software.

6.2.3 Find the degrees of freedom for the F statistic from the number and sizes of the samples. Use the F distribution Table to approximate the P-value.

6.2.4 If the test is significant, use graphs and descriptive statistics to see what differences among the means are most important.

1 Statistical Thinking

1.1 Data Production

1.1.1 Data basics:

1.1.1.1 Individuals (subjects).

1.1.1.2 Variables: categorical versus quantitative, units of measurement, explanatory versus response.

1.1.1.3 Purpose of study.

1.1.2 Data production basics:

1.1.2.1 Observation versus experiment.

1.1.2.2 Simple random samples.

1.1.2.3 Completely randomized experiments.

1.1.3 Beware: really bad data production (voluntary response, confounding) can make interpretation impossible.

1.1.4 Beware: weaknesses in data production (for example, sampling students at only one campus) can make generalizing conclusions difficult.

1.2 Data Analysis

1.2.1 Always plot your data. Look for overall pattern and striking deviations.

1.2.2 Add numerical descriptions based on what you see.

1.2.3 Beware: averages and other simple descriptions can miss the real story.

1.2.4 One quantitative variable:

1.2.4.1 Graphs: stemplot, histogram, boxplot.

1.2.4.2 Pattern: distribution shape, center, spread. Outliers?

1.2.4.3 Density curves (such as Normal curves) to describe overall pattern.

1.2.4.4 Numerical descriptions: five-number summary or $\bar{x}$ and s.

1.2.5 Relationships between two quantitative variables:

1.2.5.1 Graph: scatterplot.

1.2.5.2 Pattern: relationship form, direction, strength. Outliers? Influential observations?

1.2.5.3 Numerical description for linear relationships: correlation, regression line.

1.2.5.4 Beware the lurking variable: correlation does not imply causation.

1.2.6 Beware the effects of outliers and influential observations.

2 The Reasoning of Inference

2.1 Inference uses data to infer conclusions about a wider population.

2.2 'When you do inference, you are acting as if your data come from random samples or randomized comparative experiments. Beware: if they don't, you may have "garbage in, garbage out."

2.3 Always examine your data before doing inference. Inference often requires a regular pattern, such as roughly Normal with no strong outliers.

2.4 Key idea: "What would happen if we did this many times?"

2.5 Confidence intervals: estimate a population parameter.

2.5.1 95% confidence: I used a method that captures the true parameter 95% of the time in repeated use.

2.5.2 Beware: the margin of error of a confidence interval does not include the effects of practical errors such as undercoverage and nonresponse.

2.6 Significance tests: assess evidence against H_0 in favor of H_a.

2.6.1 P-value: If H_0 were true, how often would I get an outcome favoring the alternative this strongly? Smaller P = stronger evidence against H_0.

2.6.2 Statistical significance at the 5% level, $P < 0.05$, means that an outcome this extreme would occur less than 5% of the time if H_0 were true.

2.6.3 Beware: $P < 0.05$ is not sacred.

2.6.4 Beware: statistical significance is not the same as practical significance. Large samples can make small effects significant. Small samples can fail to declare large effects significant.

2.6.5 Always try to estimate the size of an effect (for example, with a confidence interval), not just its significance.

2.7 Choose inference procedures by asking "What parameter?" and "What study design?" See the Statistics in Summary overview.