Course Review Information Statistics 113

(1) Where Do Data Come From?

Statistical studies always involve collecting data for one or more variables of each individual in the study. Two basic ways of collecting data are by an observed study or an (randomized) experiment. A special kind of observed study, called a sample survey is also considered. The sample is collected from a population.

An *individual* is a person, object or entity. A *variable* is a characteristic of an individual. A *data point* is a particular instance of a variable.

In an *experiment*, the *experimenter* decides who is to be given the treatment and who is to be the control. In an *observational study*, it is the *individual* who decides whether or not s/he is to be given the treatment. A *sample survey* is a special kind of observational study where a survey is used to collect information on a representative sample of individuals from a larger group. In fact, all three methods use data collected, the *sample*, to infer something about the *population* from which the sample was collected. A *census* is a sample survey of all individuals in the population.

(2) Samples, Good and Bad

A statistical study is *biased* if one individual in a population is systematically favored over another, often resulting in a sample which is not representative of the population. *Convenience sampling* occurs if easiest to reach individuals are collected. *Voluntary response sampling* occurs if individuals, rather than researchers, decide who is to be in the sample or not. *Simple random sampling* (SRS) involves selecting n units out of N population units where every distinct sample has an *equal* chance of being drawn. SRS produce *representative* samples of population. Table A Random digits.

(3) What Do Samples Tell Us?

A *parameter* is a numerical quantity calculated from a population, whereas a *statistic* is a numerical quantity calculated from a sample. The value of a statistic is used to estimate an unknown parameter.

A statistic which estimates a parameter closely is considered better than one that does not do this, more specifically, a "good" statistic is one which is both a *valid* and *reliable* estimate of a parameter. An estimate is valid if it measures what it is supposed to measure, namely, the parameter; an estimate is invalid, or *biased*, if it does not measure the parameter. Furthermore, an estimate is reliable if, under repeated sampling, closely similar estimate values occur, or, in other words, there is little variability in the estimate values. As an analogy,

a gun shot (statistic) is a "good" one if the gun is not only aimed at the right place (parameter) and not somewhere else, but also each shot lands consistently close to one another.

A "good" sampling method has both small bias and small variability. Bias arises, as discussed earlier, if sampling is *non*-random and, so, to reduce bias it makes sense to sample at random, to take a simple random sample (SRS). (There are still other reasons why bias occurs, but these are discussed later.) Furthermore, variability in a SRS is reduced; that is, the average statistic values are made consistently close to one another (although not necessarily the parameter), by taking large samples, by increasing n.

Margin of error measures variability in sampling method; it says how close the average of a statistic is to a parameter if there is no bias. There are different types of margin of error, but, as discussed in this chapter, the 95% confident margin of error for population proportion parameter p using sample proportion statistic $\hat{p}$ from SRS of size n is roughly $\frac{1}{\sqrt{n}}$. A *(quick method) 95% confidence statement* for p equals $\hat{p} \pm \frac{1}{\sqrt{n}}$ and means 95% of confidence statements will include p under repeated sampling. A more sophisticated 95% confidence statement, called confidence *interval*, for p is given later on.

(4) Sample Surveys in the Real World

Sampling errors and nonsampling errors, which often lead to either bias (validity) or variability (reliability) problems or both, occur when sampling in the real world. *Sampling errors* arise from act of sampling and include, for example, *random* sampling error (measured, previously, by the margin of error) and *undercoverage*. Undercoverage occurs when individuals are left out of a sampling *frame*, a list from which the sample is chosen (such as might occur with the previously discussed convenience sampling and voluntary response sampling methods). *Nonsampling errors* have nothing to do with choosing a sample and occur even in a census and include, for example, nonresponse, processing, response and survey wording errors. *Nonresponse error* occurs when subjects either refuse are cannot be contacted for a survey. *Processing errors* occurs when data is mishandled. *Response errors* occur when subjects give incorrect answers. *Wording errors* occur when survey questions are improperly or incorrectly worded.

In addition to simple random sampling (SRS), other *probability sampling* techniques include stratified sampling, cluster sampling and systematic sampling. *Stratified sampling* involves dividing population into strata and choosing a simple random sample (SRS) from each strata. *Cluster sampling* involves dividing population into clusters, choosing a subset of these clusters at random, and then using either SRS or all items of selected clusters. *Systematic sampling* involves selecting every kth item from population, where first item is chosen at random.

In general, these probability sampling techniques reduce variability (improve reliability) but at the expense of increasing bias somewhat.

(5) Experiments, Good and Bad

Point of both observational studies and designed experiments is to identify variable or set of variables, called *explanatory variables*, which are thought to predict outcome or *response variable*. *Confounding* between explanatory variables occurs when two or more explanatory variables are not separated and so it is not clear how much each explanatory variable contributes in prediction of response variable. *Lurking* variable is explanatory variable not considered in study but confounded with one or more explanatory variables in study.

Confounding with lurking variables effectively reduced in *randomized comparative experiments* where subjects are assigned to treatments at random. Confounding with *a (only one at a time)* lurking variable reduced in observational studies by *controlling* for it by *comparing matched groups*. Consequently, experiments much more effective than observed studies at detecting which explanatory variables *cause* differences in response. In both cases, *statistically significant* observed differences in average responses implies differences are "real", did not occur by chance alone.

(6) Experiments, Good and Bad

Conducting experiments in real world face practical problems. *Placebo effect*, positive or negative influence on patient (or experimenter) not due to drug, is offset by *double-blinding*, if both patients and experimenters do not know which drug is assigned to which patient, until after experiment is over. Experiments suffer from *refusals*, subjects who refuse to participate; *nonadherers*, subjects who do not follow prescribed treatment; and *dropouts*, subjects who do not complete experiment.

Convincing experiments are well designed. Three types designs are described. *Completely randomized design* investigates how one or more explanatory variables (each with two or more levels or treatments) influences a response variable. *Block design* investigates how one explanatory variable influences a response variable, but also, to reduce variability of statistical inference, groups experimental units into homogeneous blocks. An important special case of block design is *matched-pair design* where there are only two experimental units per block.

(7) Data Ethics

Besides honesty, basic ethical requirements for any study of human subjects in U.S. are approval by a review board, informed consent and confidentiality of data. Main purpose of *institutional review board (IRB)* is to be sure subjects in *human* studies are safe. *Informed consent* means asking subjects to agree to participate in a study after first informing them of nature of study and

possible risks and benefits. *Confidentiality* means publishing group data where individual information cannot be determined.

(8) Measuring

The property of a person or thing is *measured* if a numerical value (previously called a numerical data point) is used to represent this property. The result of the measurement is a numerical *variable* which takes different values (also discussed previously). An *instrument* performs the measurement; *units* identifies the type of measurement. Measurement is used throughout statistics. In particular, recall, a "good" statistic is one which is both a *valid* and *reliable* estimate of a parameter and, so it is not surprising measurement can be modelled in the following way:

$$\text{measured value} = \text{true value} + \text{bias} + \text{random error},$$

where, as before, bias is systematic error which either overstates or understates true value and random error is the random variability in the measurement. Notice, a measured value is closer to the true value the smaller the bias (so increasing validity) and also the smaller the random error (so increasing reliability). *Variance* is another way (margin of error was previously used) to measure reliability:

$$\text{variance} = \frac{\text{sum of (each measured value} - \text{average of measured values)}^2}{n-1}.$$

The square root of variance, *standard deviation*, is yet another way to measure random error. If it is difficult to make a (direct) valid measurement of property A, it is often possible to first make an easier measurement of associated property B to then make an (indirect) *predictive valid* measurement of property A instead (as in, for example, randomized comparative experiments).

(9) Do the Numbers Make Sense?

Items which may influence either validity or reliability or both of the study are

- Is context of data valid: is important information missing?
- Are numbers consistent with one another?
- Implausible number: too big or too small?
- Are numbers *too* consistent: agree too well?
- Is the arithmetic correct; in particular, related to percentage change:

$$\%\text{ change } = \frac{\text{amount of change}}{\text{starting value}} \times 100$$

(10) Graphs, Good and Bad

Data from categorical variable and quantitative variable are described in both a graphical and tabular form. Data for *categorical* variable organized into one of several groups (categories) and can only be counted. Data from *quantitative* variable can be added, subtracted, multiplied and divided. Distribution tables, bar graphs (Pareto charts), pie charts and line graphs are discussed in this chapter. *Distribution table* of variable give us values of variable and how often these values occur. *Line graph* is a plot of variable, along y-axis, versus time, along x-axis.

(11) Displaying Distributions with Graphs

Data from quantitative variable is described in both a graphical, in particular *histograms* and *stemplots*, and tabular form. We look at overall pattern and for *deviations* such as *outliers* in these graphs. Overall pattern can be described using *shape*, *center* and *spread*. Some graphs have simple shapes which are *symmetric*, *(right or left) skew*, or *uniform*.

(12) Displaying Distributions with Numbers

We look at important numerical summaries of distributions. Two measures of central tendency are average (or, equivalently, mean) and median. *Mean*, $\bar{x}$, is sum of numbers divided by n. *Median*, M, is middle number in list of numbers arranged smallest to largest and is *located* with the $\frac{n+1}{2}$ rule. Measures of variability are standard deviation (as well as closely related variance) and first and third quartiles. *Standard deviation*, s, is "average distance from mean". *First and third quartiles*, Q_1 and Q_3, are medians of upper half and lower half, respectively, of list of numbers arranged smallest to largest. We look at five-number summary

$$\{\min, \quad Q_1, \quad M, \quad Q_3, \quad \max\},$$

and related boxplot. We also calculate mean, median and SD for grouped data.

(13) Normal Distributions

Histograms of sampled data often *approximated* (or "idealized") by graphs of distributions. Area of a portion of a histogram often approximated by area of equivalent portion of a graph of related distribution. Normal distribution is, by far, most important used in statistical analysis. Many statistical studies related to, say, psychological experiments, economic indicators or scientific measurements are assumed to possess (or, at least, can be fairly well approximated by) a Normal distribution.

We look at percentages and percentiles for Normal distribution. The *68-95-99.7 rule* tells us

- 68% of distribution falls within one SD of mean,

- 95% of distribution falls within two SDs of mean,
- 99.7% of distribution falls within three SDs of mean.

Furthermore, we look at the *standard score* given by

$$\text{standard score} = \frac{\text{observation} - \text{mean}}{\text{standard deviation}}$$

(14) Describing Relationships: Scatterplots and Correlation
We look at scatterplots and linear correlation for paired (bivariate) quantitative data sets. Scatterplot is graph of paired *sampled* data and linear correlation is a measure of linearity of scatterplot. Formula for linear correlation coefficient is

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

(15) Describing Relationships: Regression, Prediction and Causation
We look at drawing a line, called least-squares regression line, on a scatterplot and using it for prediction. We also discuss causation.

(16) The Consumer Price Index and Government Statistics
Generally, inflation of money occurs over time. This means the dollar loses buying power over time; that is, the dollar buys less in the future than now. Consumer Price Index (CPI) measures the buying power of the dollar. Formulas considered include

$$\text{index number} = \frac{\text{value}}{\text{base value}} \times 100$$

and

$$\% \text{ increase} = \frac{\text{value - base value}}{\text{base value}} \times 100$$

which are both closely related to the *fixed market basket price index* and also the consumer price index (CPI), which leads to the following conversion

$$\text{dollars in time B} = \text{dollars in time A} \times \frac{\text{CPI at time B}}{\text{CPI at time A}}.$$

(17) Thinking about Chance
Random means although an individual outcome is unknown, there is still a regular distribution of outcomes in a large number of repetitions. In particular, repeatedly sampling, sample proportion of an outcome will approach and stay close to expected population probability, a number between 0 and 1. This long run result is called *law of averages*. *Personal probability* is also a number between 0 and 1 that expresses an individual's belief in likelihood of an outcome. *Bayes' theorem* is a formal way of adjusting for personal probabilities.

(18) Probability Models

We look at terminology related to *probability models* and then a number of related rules, given below.

- Probability of any event, E, must be between 0 and 1.
- Sum of probability of all outcomes equals 1.
- Probability event does *not* occur equals 1 (one) minus it does occur.
- If no outcomes in common, probability one or other event occurs equals sum of individual probabilities.

We then look at a *sampling distribution*, a probability model of a statistic.

(19) Simulation

It is often difficult to calculate probabilities from probability models. *Simulation* (previously briefly discussed in chapter 3), repeatedly drawing numbers from random numbers table, allows approximation of these probabilities. If outcome of one draw does not depend on outcome of another draw, draws are *independent*; otherwise they are *dependent*.

(20) The House Edge: Expected Values

We discuss *expected value*, or "center", of probability distributions; for example, expected winnings of betting on a roulette game. If exact calculation of expected value not possible, it is often possible to obtain a simulated expected value.

(21) What is a Confidence Interval?

Confidence interval for proportion p from a binomial distribution is

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Related to this, sampling distribution for $\hat{p}$ is approximately normal with a mean of p and a standard deviation of $\sqrt{\frac{p(1-p)}{n}}$. Confidence interval for mean μ is

$$\bar{x} \pm z^* \left(\frac{s}{\sqrt{n}} \right).$$

Related to this, sampling distribution for $\bar{x}$ is approximately normal with a mean of μ and a standard deviation of $\frac{\sigma}{n}$.

For both confidence intervals and sampling distributions, it is assumed a large random sample has been chosen.

(22) What is a Test of Significance?

If chance of observing an outcome sampled from a population with an *assumed* parameter is small, then choice of outcome is unlucky or, more likely, choice

of population parameter is wrong. Chance in this situation is called *P-value*. Procedure of deciding whether population parameter is correct or not is called *test of significance*. Tests of significance involve following ratio,

$$\text{standard score} = \frac{\text{observation} - \text{mean}}{\text{standard deviation}}$$

where, for population proportion p in particular,

$$\text{observation} = \hat{p}, \qquad \text{mean} = p \qquad \text{and} \qquad \text{standard deviation} = \sqrt{\tfrac{p(1-p)}{n}},$$

and, for population mean μ,

$$\text{observation} = \bar{x}, \qquad \text{mean} = \mu \qquad \text{and} \qquad \text{standard deviation} = \tfrac{s}{\sqrt{n}}.$$

Tests of significance can be approximated by simulation.

(23) Use and Abuse of Statistical Inference
Tests of inference and confidence intervals should be used carefully. *Statistical* significance may not necessarily mean *practical* significance. Sample size influences statistical inference: a significant result may be missed if sample size is too small; statistical, but not practical, significance may occur if sample size is too large. Determine statistical inference from one (and only one) SRS because statistical significance will occur by chance alone in one of many SRSs taken.

(24) Two-Way Tables and the Chi-Square Test
We look at two-way tables to determine association of paired qualitative data. We look at marginal distributions, conditional distributions and bar graphs. We also discuss Simpson's Paradox, analogous to lurking variables in paired quantitative data. We perform a chi-square test using statistic

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i},$$

where $E_i = (\text{row total}) \times (\text{column total}) \div (\text{table total})$, which is approximately chi-square, $(r-1)(c-1)$ degrees of freedom, provided expected counts $E_i \geq 1$ and no more than 20% of expected counts are less than 5.