Just Enough to be Dangerous: Basic Statistics for the Non-Statistician

Thomas Simpson, Research Associate

Office of IR, Clemson University

First Order of Business

Knowing When You Should Do Something Research-y to Answer a Question

#1: What am I trying to say?

- **Who** am I trying to say something about?
  - Answer leads to how to appropriately select or collect data and to whom your findings can (or more importantly, cannot) be generalized.
- **What** are my research questions?
  - Well-defined research questions guide selection of study design & statistical methods.
- Am I trying to **describe a behavior** or **make an inference to future behavior**?
  - 99% of IR is descriptive in nature, & we typically only engage in inference when asked by others. This may lead to bad choices in design, data collection & analysis.

#2: How well do I know my **data**?

- Do I have the correct data available? If not, is it available? If it is, how will I collect it?
- Am I collecting data from a **sample** that will allow me to make an inference about a **population**? Is it representative of the group I'm trying to say something about? If not, why not? (***NOTE:*** IR collects & reports so much data mandated by the State & Feds that we often thinks is good enough to answer questions that were never intended to be answered with the data.)
- Are there **questions I can't answer** with the data I have?

After Taking All of This Into Consideration…

You should **go further** & do something research-y (conduct a study) **if & only if** you have a

GREAT IDEA!!!

Steps in Doing Something Research-y (At Each Step: Ask Friends for Advice)

- **Conduct a literature review.**
- **Write out your research question(s).**
- **Identify your variables** (independent (predictor), dependent (response), confounding (extraneous)).
- **Identify how your variables are scaled/measured** (discrete or continuous; nominal, ordinal, interval or ratio scale).
- **Determine the type of study you need to conduct** (observational/ causal/correlational or experimental).
- **Determine the statistical procedures needed for the study design** (e.g., hypothesis test, regression analysis).

- **Identify your data source(s).**
  - ✓ Are you studying a sample or a population?
  - ✓ Is your data readily available from an existing database?
  - ✓ Will you need to design an instrument (e.g. survey, test) to collect the data? Is the instrument *reliable*? Are the inferences made with the instrument *valid*?
  - ✓ Will you need to use a **random sampling** method (simple random, systematic, stratified, cluster) for data collection?
  - ✓ If collecting or using data on human subjects, have you consulted the Institutional Review Board (**IRB**) concerning policies & procedures at your institution?
- **Collect your data.**
  - ✓ Have the data been collected by random sampling/assignment?
- **Clean your data.**

- **Decide on the best analytical tool (statistical software) for your needs (number-crunching).**
- **Format your data for use by the analytical tool.**
- **Verify that your data meet the assumptions of the statistical procedure used** (e.g., normality, independence).
- **Number-crunch with the analytical tool.**
- **Interpret the output from the analytical tool. Make your inferences.**
- **Summarize your findings.**
- **Share your findings!**

Basic Statistics: Part I

Statistical Vocabulary

(or buzzwords people use a lot but probably don’t understand much)

Where Does Data REALLY Come From?

- At the simplest level, data comes from **subjects** or **objects**. Data is collected **from** subjects or objects. Data is **NEVER** a subject or an object.
- When conducting statistical analysis, data comes from two groups:
  - **Population**: the "large" or "total" group under consideration
    - A number associated with a population is a **parameter**.
  - **Sample**: a subset of the population
    - A number associated with a sample is a **statistic**.
- Most inferential statistical procedures assume that samples are selected **randomly** from populations. In a **random sample**, each subject has the same chance of being chosen & all relevant characteristics of the population are retained in the sample.

Types of Random Samples

- **Simple Random Sample**
- **Stratified Random Sample**
  - Population is separated into two or more groups (strata) & samples are randomly selected from each strata.
- **Cluster Random Sample**
  - Population is separated into two or more groups, some of the groups are randomly selected & ALL subjects in the groups are considered.
- **Systematic Random Sample**
  - Every kth subject in the population is randomly selected.

Instruments for Data Collection from Populations & Samples

- Major data collection instruments in IR are:
  - Institutional databases with student-level data
  - Institutional surveys & tests
  - National surveys & tests (‘standardized’ instruments)
- ‘Standardized’ instruments have 2 important characteristics:
  - **Reliability**: How reliable is the instrument in measuring the *constructs* under consideration? Common measures of reliability are Cronbach’s Alpha, Spearman-Brown Prophecy & KR-21 (Kuder-Richardson).
  - **Validity**: How valid are the inferences that can be made by the instrument? There are many types of validity & threats to it.

What Does Data REALLY Give Us?

- Data gives us the values of **variables** related to some aspect of the subjects. Subjects are **NEVER** variables.
- Variables in the statistical world are no different than variables you remember from Algebra I; they represent a changing set of values, but in statistics, there are two major types of variables:
  - **Qualitative Variables**: non-numerical-valued variables
  - **Quantitative Variables**: numerical-valued variables; quantitative variables are either **discrete** (countable) or **continuous** (measurable)
- Quantitative variables are what we deal with 99% of the time in IR. We even make qualitative variables quantitative on a regular basis (e.g., IPEDS Race Code)

Levels (Scales) of Measurement for Variables

- **Nominal (Categorical) Data**: "name" or "label" data; most qualitative data is nominal; some numerical data is (e.g., SSN)
- **Ordinal Data**: values can be ordered low to high but the numbers(or symbols) are only placeholders & differences in the values are unable to be determined (e.g., Likert scale)
- **Interval Data**: differences in values exist but are meaningless; there is no "0" & arithmetic operations cannot be performed (e.g., SAT scores)
- **Ratio Data**: differences in values are meaningful; there is a "0"; arithmetic operations can be performed (e.g., count data)

Quantitative Research Studies

- **Descriptive Study**: a study which describes what is happening with one variable (univariate) at a time
- **Observational (Correlational) Study**: a study which looks at the relationship between two (bivariate) or more (multivariate) variables. **CORRELATION DOES NOT IMPLY CAUSATION**. At the very best, a correlational study can only lead to weak causal inferences.
- **Experimental Study (Experiment)**: a study where one or more of the (independent & other non-response) variables are manipulated by the researcher and subjects are randomly assigned to an experimental condition
- **Quasiexperimental Study**: an experimental study where random assignment to experimental condition is not possible

Variables in Research Studies

- Research studies involve two types of variables (measured as previously discussed):
  - **Independent (Predictor) Variables (IV's)**: variables that you as the researcher can control or are already a characteristic of the subject; in your research question or hypothesis, the independent variable(s) are affecting or driving the action of the dependent variables(s). IV's are usually denoted with an X or an X with a subscript (e.g., X_1)
  - **Dependent (Response) Variables (DV's)**: variables that you suspect are being affected by or are related to the influence of the independent variable(s). DV's are usually denoted with a Y.
- **IV** → **DV** or **X** → **Y**

Other Variables to Consider in a Research Study

- **Mediating Variables**: variables that change the action of IV’s &/or DV’s
- **Confounding Variables**: variables that may mask, when accounted for by your model/design, the true relationship between IV’s & DV’s
- **Blocking Variables**: variables that may explain some of the phenomena at hand & which you can measure but whose influence you are not directly interested in

Basic Statistics: Part II

Descriptive Statistics

(or 99% of the number-crunching we do in IR)

What Are Descriptive Statistics?

- **Descriptive statistics** describe large amounts of data from variables in a single number or a set of numbers. The classes of descriptive statistics are:
  - **Measures of Central Tendency**: mean, median, mode
  - **Measures of Variability**: range, variance, standard deviation, interquartile range (IQR)
- **Descriptive statistics** can be calculated for (finite) population & sample data. We denote a (finite) population size as **N** & a sample size as **n**.

Measure of Central Tendency: Mode

- **Mode:** the most frequently occurring score in a distribution, the most commonly given single response. It is the only measure of central tendency available for categorical data. When there are two or more most frequently occurring responses (each having an equal number of observations) we say the distribution of that data is bimodal. Beyond two, depending on who you ask, the data are multimodal or have no mode.

Measure of Central Tendency: Median

- **Median:** the value at which half the scores, when placed in order, are above & half are below. The actual values of the scores are irrelevant. Because of this, the median is the best way to report central tendency for scores with a skewed distribution, like income (where there may be outlying values many multiples of the typical values of the data). The median is also the best way to report central tendency for ordinal data. If you have an odd number of scores, the median is simply the number in the middle. If you have an even number of scores, the median is the average of the two numbers in the middle.

Measure of Central Tendency: Mean

- **Mean or Average**: the arithmetic average of a set of scores; you add them all up and divide by the number of scores. Because you can do math with interval or ratio data, the mean is often the best way to report central tendency for these types of variables, but extreme values (outliers) can make the median still a better choice. Means are frequently calculated for ordinal variables in the real world, but may in fact not really tell you anything. The mean can be unduly influenced by outliers.

Measure of Variability: Range

- **Range**: the high score among your data minus the low score. It gives a very rough estimate of the numerical spread of your observations, and can be influenced, like the mean, by outliers.   You cannot calculate the range or any other measure of variability for categorical data, other than the number of different scores or responses reported.

Measure of Variability: Interquartile Range (IQR)

- **Interquartile Range**: Quartiles are the score values at the ¼ (25%) and ¾ (75%) points when the data is put in order, just as the ½ point is the median. The ¼ point is also known as the 25th percentile (Q_1) and the ¾ point is the 75th percentile (Q_3). The interquartile range is derived by subtracting the value of the 25th percentile from the value of the 75th percentile. This improves on the raw range by limiting the relative influence of extreme outlying scores. You are probably familiar in your work with reporting the interquartile range of SAT & ACT scores at your institutions.

Measure of Variability: Variance

- **Variance**: the (adjusted, unbiased) average of the deviation scores. A **deviation** is the squared difference between each score in the distribution & the mean. (The differences are squared to remove negatives and positives, because approximately half the scores will be above the mean, leading to a positive difference score, & half below, leading to a negative. If you just added up all the difference scores without squaring, you would get zero.) Deviation scores become very important in understanding concepts such as correlation. The variance is the sum of the deviation scores divided by n – 1 (rather than n because calculating the mean uses up 1 **degree of freedom**, the number of elements allowed to vary once restrictions have been placed on them, such as calculating the mean). The variance has 'squared' units & is always nonnegative (positive or 0). A variance should only be calculated on interval or ratio-scaled data.

Measure of Variability: Standard Deviation

- **Standard Deviation:** the square root of the variance; it returns the expression of variation to the 'units' of the scores you're trying to describe. All of the things that are true about the variance are also true about the standard deviation.

Other Ideas Usually Considered in Descriptive Statistics

- **Relative Frequency**: the percentage of the distribution of scores that have this value; divide each frequency count by the total number of observed scores or n. Frequency & relative frequency are often depicted graphically as a **histogram**, a bar graph where the values in the distribution are divided into classes of a given size & are represented on the horizontal (x) axis & the frequencies or relative frequencies are represented on the vertical (y) axis & the bars have a height of the frequency or relative frequency for that class. A **frequency polygon** (line graph) can be constructed from a histogram by placing a point in the middle of the top of each bar & 'connecting the dots' from the origin (point (0,0)) to the other points in order and then to the horizontal axis point of the right end of the last bar. The mode will be the highest point on both a histogram & frequency polygon.

- **Cumulative Frequency**: accumulating the total frequency or percentage of scores in the distribution that have this value or lower. This can be done with individual values or classes like in a histogram. An **ogive** is a line graph with either individual values or classes on the horizontal axis, cumulative frequencies on the vertical, and a line connecting the points representing the cumulative frequency for each value or class (located at the end of each class if using classes). The last point in an ogive will have a value of either the population/sample size if using frequencies or 1 or 100% if using relative frequencies/percentages.
- **Frequency Table**: a table that summarizes the distribution using individual values or classes of values, frequencies, relative frequencies, cumulative frequencies & used to construct the related graphs for each.

- **Skewness**: if the high frequencies (the ‘bump’) in the middle of the range of scores on a histogram or frequency polygon is off-center, the data are **skewed**. If the graph tails off to the left, the distribution is **left (negatively) skewed** & mean < median < mode. If the graph tails off to the right, the distribution is **right (positively) skewed** & mode < median < mean. If the distribution is not skewed, it is **symmetric (normally distributed)** & mean = median = mode.

- **Kurtosis**: the ‘heaviness’ or ‘lightness’ of the tails of a distribution. Distributions with a relatively large number of scores in the tails, at the extreme values of the range of scores, are called heavy-tailed or platykurtic (flat, like the bill of a platypus). Some statistical tests are not valid for use with platykurtic data. The most extreme example of heavy-tailed data is a uniform distribution, with the same number of each score in the distribution. Most parametric tests are not valid for uniform data.

- **Percentiles**: a value represents the kth percentile if k% of the values in the distribution are below the value & (100 – k)% of the values are above it.
- **5-Number Summary**: the lowest value, the 25th percentile, the median (50th percentile), 75th percentile & the highest value in a distribution. The 5-number summary is graphed as a **boxplot** or **box-&-whiskers plot**.
- **Z-Scores (Standard Scores)**: used to express any kind of data as a standardized score in relation to its mean; to calculate a z-score you subtract a score from the mean and divide by the standard deviation. If the data are normally distributed, 95% of the scores will have z-scores between -2 & +2 & 99.7% of the scores between -3 & +3. Z-scores are useful to compare scores on variables with very different scales.

Example #1: Descriptive Statistics

- **Two variables:** Average student grades in a class using an e-textbook; average student grades in a class using a print textbook
- **For each variable:**
  - Construct a histogram
  - Find the descriptive statistics (mean, median, mode, range, variance, standard deviation)
  - Find the 25th & 75th percentiles (1st & 3rd quartiles) & a boxplot
  - Find the percentile rank of a given value.
  - Convert all of the values to z-scores, find their descriptive statistics & draw a histogram. What did you find out?

Summary of Excel Procedures

- **Histogram**: Insert → Recommended Charts → All Charts → Histogram
- **Descriptive Statistics**: Data → Data Analysis → Descriptive Statistics (be sure to check 'Display Descriptive Statistics'); may also use commands
- **Quartiles:** =QUARTILE.INC(array name, 1 (for 25th) or 3 (for 75th))
- **Percentile**: =PERCENTILE.INC(array name, kth percentile)
- **Percentile Rank**: =PERCENTRANK.INC(array name, value in array)
- **z-score**: =STANDARDIZE(value, mean, standard deviation)
- **Boxplot**: Insert → Recommended Charts → All Charts → Box & Whisker

Basic Statistics: Part III

Probability & Probability Distributions

(or that stuff you think is rocket science but isn’t)

Probability: The Basics

- **Probability**: the chance that an "event" will occur. Probabilities are expressed as fractions, decimals or percentages & is usually written as P(E), read as "P of E" or the probability that event E occurs.
- **Two types of probability**:
  - **Empirical**: based on actually conducting a probability experiment; number of times the event occurs/number of times the experiment is repeated
  - **Theoretical**: based on determining the sample space of a probability experiment & how many times a certain "event" occurs; number of times the event occurs in the sample space/sample space size

Probability Facts

- The probability of an impossible event is 0 or 0%. The probability of an event certain to occur is 1 or 100%. All other probabilities are between 0 & 1 (or 0% & 100%). There are **NO** negative probabilities & **NO** probabilities above 1 or 100%.
- The probability that an event will **NOT** occur is 1 – the probability that the event occurs, or 1 – P(E). **Not** an event is called a **complementary** event.
- Probability is **NOT** the same thing as **odds**. The odds that an event will occur (odds in favor) is the ratio of the probability that the event will occur and the probability that it will not occur: P(E)/(1 – P(E)) & reduced to a ratio written as a:b (read "a to b"). The odds that an event will not occur is the reciprocal of the odds in favor: (1 – P(E))/P(E) & will reduce to the ratio b:a. **Even odds** occur when the odds in favor & against an event are the same or 1:1. This only happens when P(E) & 1 – P(E) are the same (1/2 or 0.5 or 50%).

One Useful Formula That's Not a Probability But Has Percentages

$$\text{Percent Change/Difference} = \frac{\text{New} - \text{Old}}{\text{Old}} \cdot 100\%$$

- If it is negative, it is a decrease. If it is positive, it is an increase. It can be over 100%.

Probability Distributions

- **Probability Distribution**: a distribution of values & probabilities of a **random variable** (the values of the variable occur at random). The values in a probability distribution follow the rules for probabilities and they add up to 1 or 100%.
- **2 Types of Probability Distributions/Random Variables**:
  - **Discrete**: countable values & bar graph depiction, e.g. Binomial, Hypergeometric, Uniform, Poisson
  - **Continuous**: measurable values & "smooth curve" graph depiction, e.g. Normal, Chi-Square, Gamma

The (Standard) Normal Distribution

- By knowing the properties of the normal distribution, we can determine the position of any score in relation to other scores in the same distribution.
- The normal curve or bell curve is merely a plot of scores along the horizontal axis, and the frequency of each score along the vertical axis (a frequency polygon for an infinite number of possible scores).
- We use a normal distribution described in terms of z-scores. We have seen that the distribution of z-scores has a mean of 0 & a standard deviation of 1, & the same is true for the normal distribution. The frequency of the scores clusters about 0. The distribution is symmetric about the vertical axis (about 0). Below 0, the z-scores are negative & positive above 0. The total area under the curve is 1 or 100% & each half of the curve had 0,5 or 50% of the area. In a continuous distribution, **probability = area under the curve**, so we cannot find the probability of **exactly** a given value.

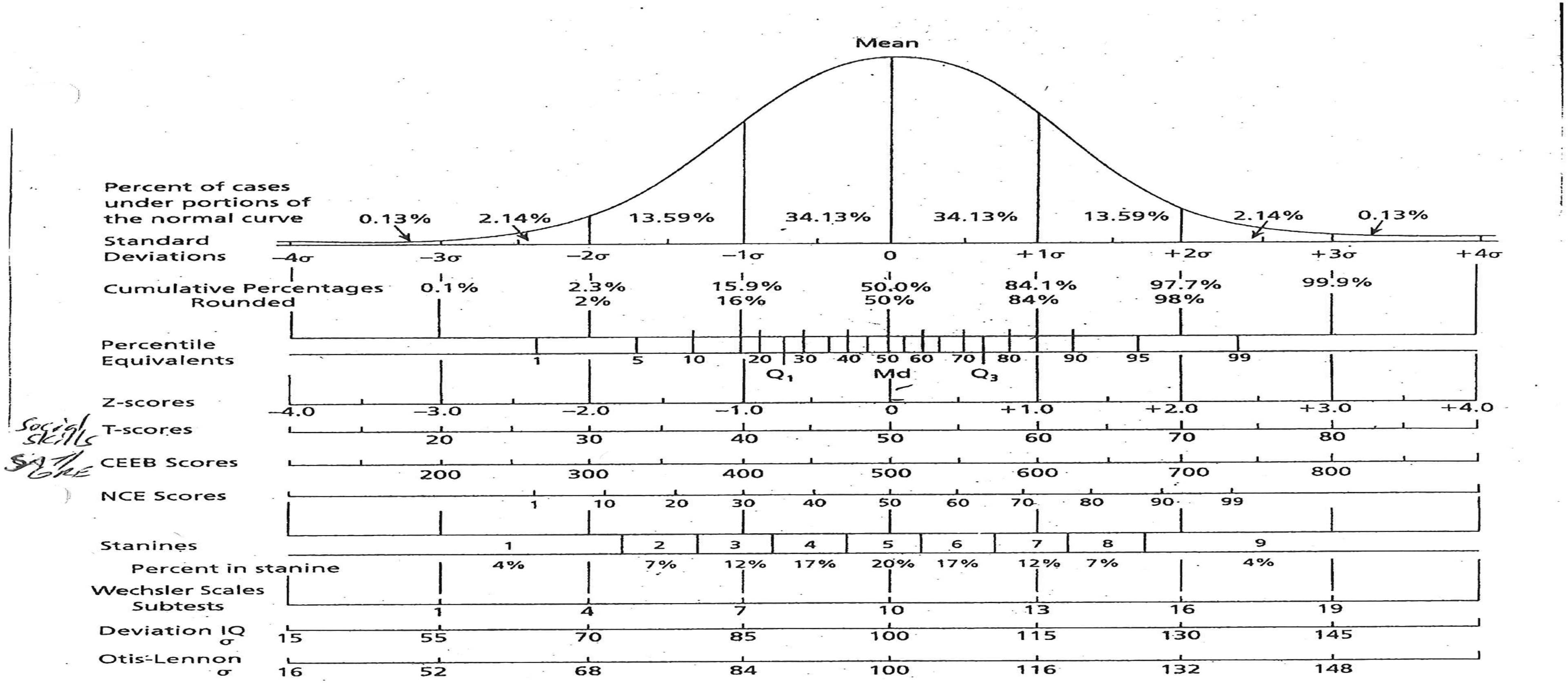

FIGURE 4.3 Normal Curve Distribution and Common Scaled Scores

Source: *Methods of Expressing Test Scores.* Test Service Bulletin, January 1955 by the Psychological Corporation, No. 48. Reproduced by permission of Harcourt Brace Educational Measurement.

Finding Probabilities with the Normal Distribution

- If you know a z-score, you can find the cumulative probability (the probability of less than that number) in Excel by using =NORM.S.DIST(z-score,1).
- If you know a certain cumulative probability & wish to find the z-score, use =NORM.S.INV(probability)

BUT…

What we really want to do is find probabilities or percentages or proportions or percentile ranks for practical situations where we do not have to calculate z-scores!

Example #2: Normal Probabilities

The Wechler-Belview G intelligence test has scores that are normally distributed with a mean of 100 and a standard deviation of 15.

a. Find the percentile rank of a person who scores 120.
b. Find the proportion of people who score above 70.
c. Find the proportion of people who score between 75 & 130.
d. A special program is to be given to students who score in the lowest 20% on the test. What is the cutoff score for the program?
e. A special program is to be given to students who score in the top 5% on the test. What is the cutoff score for the program?

Summary of Excel Procedures

- **Percentile Rank, Probability Below a #:** =NORM.DIST(#, mean, standard deviation, 1)
- **Probability Above a #**: =1 – NORM.DIST(#**,** mean, standard deviation, 1)
- **Probability Between Lower # & Higher #**: =NORM.DIST(higher #, mean, standard deviation, 1) – NORM.DIST(lower #, mean, standard deviation, 1)
- **Cutoff Score for the Lowest %**: =NORM.INV(percentage as a decimal, mean, standard deviation)
- **Cutoff Score for the Highest %**: =NORM.INV(1 – percentage as a decimal, mean, standard deviation)

One Last Topic In Probability: The Sampling Distribution of Sample Means

- **Sampling Distribution of Sample Means**: a distribution of sample means of size n taken from a population with a given mean (μ: mu) & standard deviation (σ: sigma); the mean of the distribution of sample means is the population mean, μ, & the standard deviation of the sample means (known as **the standard error of the mean**) is $\frac{\sigma}{\sqrt{n}}$.

- **The Central Limit Theorem**: as n increases, the sampling distribution of the sample means approaches the normal distribution; the Central Limit Theorem 'kicks in' when n ≥ 30, so you can assume at the very least approximate normality

- **How This Connects to Probability**: you can find the probability that a sample mean is below, above or between different numbers, as well as percentile rank, for a given value using the sample mean as the value under consideration & standard error of the mean, rather than the standard deviation, in a z-score

Basic Statistics: Part IV

Inferential Statistics

(or what is often used & abused)

Inferential Statistics: The Big Picture

Descriptive Statistics for 1 or more samples

+

Probability Distributions

↓

Inferential Statistics

Inferential statistics: used to compare the results of your descriptive statistics against a probability distribution that you have reason to believe is appropriate for the data and research question at hand to see if your snapshot or part of your snapshot is an unusual result.

Confidence Intervals

- **Idea**: We cannot calculate a population mean. The sample mean is an (best linear) **unbiased estimator** of the population mean, but it does have a measurement error (standard error of the mean). If we estimate an unknown number, we will come up with values above & below the number. We construct a **confidence interval** for a population mean using a sample mean, the standard error of the mean & a value from the normal distribution that corresponds to a certain percentage of the time (usually 95%) that we wish for the interval to capture the true value of the population mean. The interval structure is (Sample Mean – Error, Sample Mean + Error) or ($\bar{x} - E$, $\bar{x}$ + E). Confidence intervals can be calculated for many parameters. Excel does not calculate confidence intervals.

Hypothesis Testing: Could This Have Really Happened by Chance?

- **Step 1**: Formulate all hypotheses you're interested in.
  - **Null Hypothesis:** H_o: The observations are the result purely of chance, the hypothesis of no difference or no change. The null hypothesis always has equality as a part of it, e.g., $\mu = 50$, $\mu_1 = \mu_2$
  - **Alternative Hypothesis**: H_a: There is a real effect. The observations are the result of this real effect plus error, or chance variation. The alternative hypothesis determines the 'tail' of the test (related to the tail of the probability distribution used): **2-Tail** (not equal, e.g., $\mu_1 \neq \mu_2$), **Right Tail** (greater than, e.g., $\mu_1 > \mu_2$), **Left Tail** (less than, e.g., $\mu_1 < \mu_2$)

- In the end, there are 2 decisions you can make about your hypotheses:
  - **Reject $\mathbf{H_0}$**: there is statistically significant evidence to support that the null hypothesis is false
  - **Fail to Reject $\mathbf{H_0}$**: there is statistically significant evidence to support that the null hypothesis is not false (i.e., true)
- The good researcher never says absolutely true or false. We support a claim or do not support a claim (a claim is **our** hypothesis, not the statistical one).
- There are 2 errors that can be made in statistical decision making:
  - **Type I error**: The null hypothesis is rejected when it is in reality true.
  - **Type II error**: The null hypothesis is not rejected when it is in reality false.

- **Significance (Alpha: α) Level:** the probability of making a Type I error. The alpha level is usually preset to 0.05 in most educational settings. The lower the level of significance (type I error), the higher the probability that a type II error will be committed. In other words, the harder you make it to reject the null hypothesis, the easier it is to fail to reject the null hypothesis when it is in fact false (Type II error.
- **Beta (β)**: the probability of making a Type II error
- **Power**: 1 – β. The (statistical) **power** of a test is the ability to detect a correct result—to reject the null hypothesis when it's false or not reject it when it's true. That ability can be affected by several aspects of the test—the sample size (power increases with sample size), the specified level of significance (the lower alpha, the higher beta so less power), and making your design more specific to what you want to know (like switching from a two-sided hypothesis to one if you have a good idea that the result will fall in one direction).

- **Step 2**: Identify a test statistic that will assess the evidence against the null hypothesis. This is usually a z, t, X^2 (Chi-Square), F test statistic.
- **Step 3**: Find **the p-value** for your data. This answers the question, "If the null hypothesis were true—if there were actually nothing going on with my data--then what is the probability of observing a test statistic at least as extreme as the one I observed?
- **Step 4**: Compare the p-value to a fixed significance level, called alpha (a). In most research/statistics classes, alpha will always be 0.05. If the p-value is .05 or less, reject the null hypothesis. Otherwise, fail to reject the null hypothesis. When we rule out the null hypothesis, we are agreeing that something else is going on. That something else is the alternative hypothesis. In modern statistical practice, we usually include a 95% confidence interval for the parameter(s) under study to strengthen our case.

Now That You Have Found Statistically Significant Results…

The larger the sample, the more power you have to reject the null hypothesis. Therefore, with very small sample sizes, very small, "real world" insignificant differences between groups will be statistically significant. To combat this problem, **effect size** statistics can be calculated. The effect size is merely the difference between groups: ES $= \bar{x}_1 - \bar{x}_2$. The **standardized effect size**, d, is d $= \dfrac{\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2}{\text{Standard Error}}$.

Cohen’s Convention: d = 0.3, weak effect size; d = 0.5, moderate effect size; d = 0.8, strong effect size.

There is no ‘magic’ level of (standardized) effect size to guide real-world policy decisions. In some high-risk situations, a weak effect size may be evidence enough to make a change; in some low-risk, moderate-reward situations, it may be necessary to look for a higher effect size to justify a change.

There are many other types of effect sizes for the many other types of hypothesis tests where you compare two (or more) groups.

Example #3: Comparing 2 Independent Population Means

- **Situation**: Students in one class used an e-textbook & completed online homework. Students in another class used a print textbook & completed homework ‘by hand’. Students in both classes were taught by the same instructor & completed the same quizzes & tests.
- **Hypothesis**: Students in the e-textbook class scored higher, on average, than the students in the print textbook class.
- **Statistical Hypotheses**

H_0: $\mu_1 = \mu_2$ (e-book class mean = print class mean)

H_a: $\mu_1 > \mu_2$ (e-book class mean > print class mean)

Conducting the Hypothesis Test

- **Check the assumptions for using one of the appropriate tests.**
  - Random selection/assignment?
  - Normality? Look at the boxplots.
- **Test/Test Statistic to Use**: For this research design, we typically use the t-Test for 2 independent samples, based on the Student's t-Distribution (a symmetric distribution that approaches the standard normal as its degrees of freedom increases). For us to use this test, however, we must first determine if the variances of the two groups are equal. This involves using an **F-Test** (a test using a ratio based on the F-Distribution, a positively skewed distribution with degrees of freedom for both numerator & denominator of the ratio)

F-Test for 2 Independent Population Variances

- **Statistical Hypotheses (2-Tail Test)**

  H_0: The variances of the two groups are equal.

  H_a: The variances of the two groups are not equal.

- **F-Test Statistic**: $F = \dfrac{\text{sample standard deviation of first group}}{\text{sample standard deviation of second group}}$

- **Test Conclusions**
  - **Reject** H_0: Conclude that the variances are not equal. Use t-Test: 2-Sample Assuming Unequal Variances (uses Satterthwiate approximation)
  - **Fail to Reject** H_0: Conclude that the variances are equal. Use t-Test: 2-Sample Assuming Equal Variances

Basic Statistics: Part V

Correlation & Regression

(remember…correlation is NOT causation)

Correlational Research

- The **correlation coefficient** is a standardized measure of the slope of the line that expresses the linear relationship between two variables (x, independent variable & y, dependent variable) on a -1 to 1 scale. Pearson correlation coefficients do this for two continuously measured variables, but there are other types as well. The correlation is standardized by a ratio of the two variables' standard deviations so that the unit of measurement of the two variables does not affect the value of the correlation coefficient.
- Correlational research is used to determine (a) if there is a relationship between two or more variables, and if so (b) to determine the strength & direction of the relationship.

- The relationship between variables should generally only be investigated if there is a reason/theory supporting the possibility of a relationship; **don’t go shopping for correlations**.
- Correlation, like regression and other techniques built on correlation, are techniques of mathematical maximization. This means that they can possibly show significant relationships that only exist in the sample of observations used to calculate the correlation or build the model; these relationships can fall apart or change when a new set of data is collected. For this reason, **theory should drive your use of correlational and other mathematical maximization techniques**.
- Sometimes correlations are significant, but make little obvious sense. If your sample size is even modestly large (like 30), you can get statistically significant correlations that may not be practically significant.
- There are significance tests for r & other correlation coefficients.

- If reliability and validity of your instrument are low, error is added, thus reducing correlation. However, good reliability and validity does not assure a strong correlation. Also, it helps if your variables are measured as accurately as possible.
- If the range of your variables is limited or your sample is very homogenous, this can also artificially reduce correlation coefficients. Also, if you are using a variable that is mathematically derived from another variable (such as a passing grade/failing grade group), you are artificially limiting the variation in your data and it would be better to just use the raw grade in your correlation.
- For two interval or ratio variables, the Pearson correlation coefficient is calculated. If at least one variable is only ordinal or the data is not normally distributed, ranks replace the scores and a Spearman correlation is appropriate. Other types of correlation coefficients are the phi (Φ) coefficient, the point-biserial correlation and the tetrachoric correlation.

Interpreting the (Pearson) Correlation Coefficient, r

- **Positive** values of r indicate a positive linear relationship between x & y, i.e., as x increases, y increases or as x decreases, y decreases.
- **Negative** values of r indicate a negative linear relationship between x & y, i.e., as x increases, y decreases or as x decreases, y increases.
- **r = 0** means that there is **no** linear relationship between x & y.
- **Strong Correlation**: Less than or equal to -0.7 or greater than or equal to 0.7.
- **Moderate Correlation**: Between -0.7 & -0.3, exclusive, or between 0.3 & 0.7, exclusive, & can be moderately strong or weak.
- **Weak Correlation**: Between -0.3 & 0.3, inclusive.

Multiple Correlation Coefficient, R, & Coefficient of Determination, $\mathbf{R}^2$

- **Multiple Correlation Coefficient (R)**: Pearson product-moment correlation coefficient for multiple independent variables & one dependent variable. Interpretation of R is the same as for r. An adjusted version is often used to account for the fact that adding independent variables increases R.
- **Coefficient of Determination, $\mathbf{R}^2$:** the amount of variation in the dependent variable that can be explained by the predictor (independent) variable(s). An adjusted version is often used to account for the fact that R^2 automatically increases when more predictor variables are added. The adjusted version accounts for this. We look for high values of R^2 (they range between 0 & 1) in correlational studies.

Regression

- Procedures built on the correlation coefficient include **simple linear regression** (X predicting Y), **multiple regression** (more than one X simultaneously predicting one Y), or **multivariate regression** (one or more X predicting a set of Y). When Y is dichotomous (only has two values, like yes/no, or membership in a group such as degree completers), **logistic regression** predicts the likelihood of having one or the other value. When Y has a pattern that is nonlinear, there are regression models that are connected to associated functional relationships, such as polynomials, exponential, logarithmic and power functions.

- Regression works from a mathematical principle known as **least squares**. Least squares allows you to find the most appropriate line to describe the relationship between Y, the criterion variable, and X (or X1, X2 . .), the predictors, by minimizing the total distance between the actual plotted observations in your data and the line used to predict where they might be. These individual differences between where the points in your data are and where the line predicts they should be are called residuals. You want to minimize the residual effect in a regression model while still explaining a significant amount of the variation in the data.
- Regression models are used to **predict** the value of the dependent variable(s) from the values of the independent variable(s). Predictions outside the range of values of the independent variable(s) that you know are done with great care, necessitating the construction of **prediction intervals** (similar to confidence intervals) for values of the dependent variable predicted by the regression line.

Simple Linear Regression Model

$$\hat{Y} = a + bX$$

where:

- $\hat{Y}$ = the predicted value of the dependent variable, Y
- a = the y-intercept of the line, (0, a)
- b = the slope of the line (negative, positive or 0); the slope of the line corresponds to the sign of the correlation coefficient, r
- X = the value of the independent variable

This is an **estimate** of a probabilistic general linear model.

- **Multiple Regression** models add more x’s (& sometimes combinations thereof, such as interactions) to the equation.
- When you have a set of variables that highly correlate with each other and you pick and choose among them without a clear preexisting theory as to what should have a relationship to what, you can set out to create instability in your model by not controlling for a phenomenon known as multicollinearity—basically that there are multiple linear relationships that all predict your data equally well and you have no way to know the best one. In a perfect world you would have a criterion variable (what you’re trying to predict) that is highly correlated to each predictor but those predictors are not highly correlated with each other, and a model that is driven by theory. But in real life, as in human relationships, many of the variables we use are already related to each other. Always be guided by a strong **theory** of what it is you are trying to formulate.

The ANOVA (Analysis of Variance) Table

Source	**SS** (Sum of Squares, the numerator of the variance)	**df** (degrees of freedom, the denominator of the variance)	**MS** (Mean Square, the variance)	**F**
Regression (or Between or Model)	SSR	k – 1	MSR =	F =
Residual (or Error or Within)	SSE	n – k	MSE =	
Total	TSS	n – 1		

- The F statistic (& its associated p-value) allow us to determine if the regression model is valid; this is an **omnibus** F-test. If the p-value ≤ your chosen significance level, there is statistically significant evidence that the model is valid. If not, there is statistically significant evidence that it is not.
- There are also significance tests for the slope (in simple linear regression) & the predictor coefficients (in multiple regression).
- The independent & dependent variables in a regression model are assumed to have been randomly selected or assigned from a normal population.
- The errors (residuals) in a regression model are assumed to be identically & independently distributed from a normal distribution with a mean of 0 & a constant (error) variance.
- Verifying the assumptions of the model is called **regression diagnostics**.

Example #4: Building Linear Regression Models

- **Situation**: The data are the SAT Verbal scores, SAT Math scores & the GPA of 10 students who have just completed their first year of college.
- We want to:
  - Conduct a full correlation analysis for the data.
  - Construct linear models to predict GPA using SAT Verbal &/or Math scores.
  - Determine if the models are significant.

Summary of Excel Procedures

- **Scatterplot with Line & Regression Model:** Insert → Scatter (or use Add Recommended Charts → Right click on a data point → Add Trendline → Linear → Check Display Equation on Chart
- **Correlation Matrix**: Data → Data Analysis → Correlation
- **Regression Model with ANOVA Table & Related Output**: Data → Data Analysis → Regression (for diagnostic outputs, check the boxes under Residual & Normal Probability)