1. Parametric Statistics: Traditional Approach

1.1 Definition of parametric statistics:

Parametric statistics assume that the variable(s) of interest in the population(s) of interest can be described by one or more mathematical unknowns. Some types of parametric statistics make a stronger assumption—namely, that the variable(s) have a certain distribution. To illustrate, consider a simple problem: do male and female US college students differ in average height? One type of parametric approach is to assume that four mathematical quantities can describe height in the population of college students—the mean for females, the mean for males, the standard deviation for females, and the standard deviation for males. A second parametric approach might assume that height follows a normal distribution in both sexes.

1.2 Key Terms:

1.2.1 Population:

For learning purposes, two types of populations can be distinguished in parametric statistics. Officially, statisticians never make this distinction, so the terminology given below is not standard. The two types of population are:

1.2.1.1 Specific population:

Specific populations consist of definable individual observations (e.g., people, rats, cities) where each individual could be identified and enumerate, although it usually would be impractical to do so. Examples of specific populations would be: mule deer living in Rocky Mountains; US college sophomores; American victims of incest. The height example would have two specific populations, the population of all US male college students and the population of all US female college students.

1.2.1.2 Hypothetical population:

Hypothetical populations do not exist, hence the term “hypothetical.” They are imaginary populations of infinite size and are used whenever parametric statistics assume that the variables are distributed in a certain form. In the height example, the hypothetical population of US female students would produce a perfectly smooth histogram whereas the specific population of US female college students would produce a histogram with some small irregularities. When statisticians refer to the “population,” they usually speak about the hypothetical population.

1.2.2 Sample:

A sample is a finite group of observations taken from a specific population. There are numerous strategies for sampling. Among them are:

1.2.2.1 Random Sampling.

In a random sample, each observation in the specific population stands the same equal chance of being selected into the sample. For obvious reasons, completely random samples are encountered very seldom in psychology. In the height example, it would be necessary to cover many different campuses to obtain a completely random sample.

1.2.2.2 Stratified Random Sampling.

In stratified random sampling, the specific population is divided into two or more groups (or *strata*). Each individual within a strata stands the same, equal chance of being sampled, but one or more strata are deliberately oversampled. A classic example would be a medical study of ethnic differences in hypertension that would sample an equal number of white Americans and African-Americans.

1.2.2.3 Selected Sampling.

Selected samples occur when predefined criteria are used to sample individuals from the specific population. For example, only children with IQ scores less than 70 are sampled. Selected samples have the advantage of increasing statistical power (i.e., the ability to detect effects that are really present) but the disadvantage of limited generalization.

1.2.2.4 Convenience Sampling.

Convenience sampling occurs when the individuals are not truly representative of the specific population but are easy to collect. It is the most frequently encountered sampling technique in psychology. Convenience sampling always makes the assumption that those variables on which the sample is unrepresentative of the specific population are uncorrelated (or have very low correlations) with those variables that are actually measured and studied.

For the height example, a convenience sample would be to take a certain number of University of Colorado men and women. This strategy makes the reasonable assumption that attending the University of Colorado, as opposed to other colleges, is not associated with sex differences in height. The assumption would be violated if, for example, tall women would preferentially choose Colorado over other institutions.

1.2.3 Parameter

A parameter is a mathematical unknown in the population, either the specific or the hypothetical population. It is customary to denote parameters with Greek letters. Some frequently encountered parameters are μ (population mean), σ (population standard deviation), and ρ (population correlation).

The height example could have four parameters: μ_m and σ_m (the mean and standard deviation for male college students) and μ_f and σ_f (the respective parameters for females).

1.2.4 Statistic

There are two types of statistics used in parametric statistics. They are:

1.2.4.1 Descriptive statistics.

A descriptive statistic is an estimate of a population parameter, almost always obtained through sample data For example, the sample mean is used to estimate the population mean. Usually, Roman letters are used to denote descriptive statistics such as $\bar{X}$ (sample mean), *s* (sample standard deviation), and *r* (sample correlation).

In the height example, the estimates of the four parameters might be: $\bar{X}_m$ (the sample mean for males which is an estimate of μ_m, the population mean for males), and s_m, $\bar{X}_f$, and s_f (the three statistics for the other population parameters).

1.2.4.2 Test Statistics.

A test statistic is used to make inferences about one or more descriptive statistics. Usually, a test statistic does not directly measure a population parameter, although in some cases it may be mathematically manipulated to do so. Either Roman or Greek characters are used for test statistics. Examples of test statistics would be using a *t* test statistic to test whether two sample means differ, using an *F* test statistic to test whether two or more sample means differ, and using a χ^2 test statistic to test whether two or more sample proportions differ. Unfortunately for the beginning statistics student, it is customary to drop the word "test" from test statistic. E.g., most textbooks call the *t* test statistic the *t* statistic.

In the height example, one could use the four descriptive statistics ($\bar{X}_m, s_m, \bar{X}_f$, and s_f) and the sample sizes to construct a *t* statistic that could give information about average height differences between males and females.

1.2.5 Probability density function.

The ***probability density function*** is also referred to as ***pdf*** or simply ***density function***. The pdf is a mathematical function used to describe two important phenomena: (1) the distribution of a variable(s) in the hypothetical population; and (2) the distribution of test statistics. Mathematically, the pdf fulfills two conditions: (1) it gives the relative frequency (or probability) of observing a value as a function of that value; and (2) the area under the curve between two values gives the probability of randomly selecting a number between those two values. The unknowns in a pdf are parameters.

The probability density function most often encountered in behavioral research is the normal probability density function (i.e., the normal curve). If *f*(*X*) is the probability of observing a value within a tiny interval of *X*, then the normal probability density function of *X* is

$$f(X) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right].$$

Here, the quantity π is the constant 3.14 and the term "exp" denotes the exponential function (i.e., the natural exponent, *e*, raised to the power of the stuff in the parentheses). The two parameters for the normal curve are μ (population mean) and σ (population standard deviation. Other important probability density functions are the Poisson, binomial, multinomial, lognormal, exponential, and Weibull.

Important probability density functions for test statistics are the *t* pdf (for the *t* test statistic), the *F* pdf (for the *F* test statistic), and the χ^2 pdf (for the χ^2 test statistic). The pdf for a test statistic is called the ***sampling distribution*** of the statistic.

1.2.6 Probability distribution function

The ***probability distribution function*** is also referred to as the ***distribution function***, ***cumulative distribution function***, or ***cdf***. Like the pdf, the cdf is used for both hypothetical populations and for test statistics. The probability distribution function is the mathematical integral of the probability density function. (Hence, conversely, the probability density function is the derivative of the probability distribution function.) The cdf gives the probability of observing a particular value of *X* that falls in between two values, say, X_1 and X_2. So, for example, the probability distribution for the normal curve is

$$\text{Prob}(X_1 < X < X_2) = \int_{X_1}^{X_2} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right) dX .$$

For example, if *X* denotes IQ scores which are assumed to be normally distributed with a mean of 100 and a standard deviation of 15, then the probability distribution function gives the probability of randomly picking an IQ score that falls somewhere in the interval between X_1 and X_2. If $X_1 = 102$ and $X_2 = 118$, then this normal probability distribution function gives the probability of randomly selecting an IQ score between 102 and 118 or

$$\text{Prob}(102 < X < 118) = \int_{102}^{118} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right) dX .$$

Test statistics have their own cdf's. Hence, the cdf for a *t* test statistic gives the area under the curve for a specific *t* distribution.

1.2.7 Sampling distribution.

The ***sampling distribution*** is the probability density function of a test statistic. For completely perverted reasons (i.e., to deliberately confuse the beginning graduate student), statisticians always say "the sampling distribution of the *t* statistic," and never call it what it really is—the probability density function of the *t* test statistic.

1.2.8 Standard error.

For the same perverted reasons, statisticians refer to the standard deviation of the probability density function of a test statistic as the ***standard error*** of the statistic.

Putting 1.2.7 together with 1.2.8, statisticians define the standard error as the standard deviation of the sampling distribution of a statistic.

1.2.9 Mathematical Model

A mathematical model of data consists of one or more mathematical equations that gives the predicted or expected value for a particular variable of an observation. For example, suppose that we assumed that the average height for males was 8 centimeters (cm) greater than the mean height for females. Then one could construct a mathematical model that states the following: "If the observation is a female then the predicted height is an unknown, say μ_f. If the observation is a male, then the predicted height is $\mu_f + 8$." The value for an observation predicted by the mathematical model is called the ***predicted value*** of the observation.

1.2.10 Error or Residual

The terms ***error*** and ***residual*** are synonymous in statistics. Error (or a residual) is the mathematical difference between the actual observed value for an observation and the observation's predicted value (i.e., value predicted by a mathematical model).

1.3 Steps in parametric statistics

1.3.1 Estimate the parameter(s)

Given a sample, one or more mathematical formula are used to obtain descriptive statistics of the parameters and to then test hypotheses about these descriptive statistics.

1.3.1.1 Methods for estimation

There are several different methods that can be used to estimate parameters. No single method is "best," so the choice depends largely on the nature of the problem. For the techniques in this course, all the methods often give the same answers or when they do not, they give very similar answers.

1.3.1.1.1 The Sum of Least Squared Error

The sum of least squared error is most often referred to as simply "least squares." Least squares is the most frequently encountered criterion for estimation in the behavioral sciences, so let's take some time to explore it in detail.

Least squares begins with the observed value for the first observation in the sample. It develops a ***predicted value*** for this observation based on a ***mathematical model***. The ***mathematical model*** is defined as one or more equations based on population parameters and other aspects of the observed data. Least squares then subtracts the predicted value from the observed value giving what is called the ***prediction error*** (or more often, ***error*** or ***residual***) for the first observation. It then squares this prediction ***error***. Usually, the observed value for the first observation is denoted algebraically as Y_1 (the subscript denotes the order of the observation or 1 in this case)

and the predicted value is denoted by the same symbol with a hat (^) placed over it-- $\hat{Y}_1$. The prediction error is usually denoted as e_1. Hence, algebraically, the prediction error for the first observation is

$$e_1 = \left(Y_1 - \hat{Y}_1\right)$$

and the squared error for the first observation is

$$e_1^2 = \left(Y_1 - \hat{Y}_1\right)^2 .$$

Least squares then goes to the second observation and computes its error and its squared error. Least squares continues with this procedure for all the observations in the sample. The final step is to add up all the squared errors giving the ***sum of squared errors*** or ***SSE***.

Algebraically, the following two equations are identical ways of writing ***SSE***:

$$SSE = e_1^2 + e_2^2 + \ldots + e_N^2 = \sum_{i=1}^{N} e_i^2$$

and

$$SSE = (Y_1 - \hat{Y}_1)^2 + (Y_2 - \hat{Y}_2)^2 + \ldots + (Y_N - \hat{Y}_N)^2 = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 .$$

We are now in a position to define the sum of least squared error criterion for parameter estimation. ***Least squares statistics*** are those estimates of the population parameters that minimize the sum of squared prediction errors. Mathematically, least squares statistics are those parameter estimates than generate all the $\hat{Y}$ s so that *SSE* is minimized.

To illustrate consider the females in the height example and suppose we would like to find the least squares statistic for μ_f, the mean height for the population of female college students.

We can write the model by letting the predicted value for height be the mean or

$$Y_i = \hat{Y}_i + e_i = \mu_f + e_i .$$

What now is the descriptive statistic that best estimates the parameter μ_f in terms of minimizing $\sum_{i=1}^{N} e_i^2$? We will eschew all the mathematics that proves the obvious—the best estimate of μ_f is $\bar{Y}$, the mean height in the sample of US college women.

1.3.1.1.2 Maximum Likelihood(*)

Maximum likelihood is another method for estimating parameters. Maximum likelihood estimators are those that maximize the probability of observing the data when the variable(s) in the hypothetical population is(are) assumed to have a certain probability density function. For example, if the variable in the hypothetical population is assumed to follow a normal distribution, then the maximum likelihood estimator of the population mean, μ, is that numerical value that maximizes the probability of observing

the sample data. Like least squares, the maximum likelihood estimator of μ is the sample mean.

1.3.1.1.3 Bayesian Estimation(*)

Baysian estimators are yet another type of estimators. These estimators assume that a descriptive statistics is itsself sampled from a hypothetical population of descriptive statistics that has a certain distribution called the ***prior*** distribution. It them finds the most likely estimator given the data. Like least squares and maximum likelihood, the Baysian estimator of μ is the sample mean.

1.3.1.1.4 Robust Estimators(*)

Robust estimators are a whole class of estimators that are relatively insensitive to departures from the assumptions that go into the mathematical model behind their estimation. An example of a robust estimator would be a "trimmed mean" which deletes, say, the highest three scores and the lowest three scores and then takes the mean of the remaining data points. There is no specific criterion(a) developed to determine whether an estimator is "robust" or not "robust."

1.3.1.2 Criteria for "Good" Estimators(*)

Not all estimators of a population parameter are "good" estimators. Statisticians have established four properties of "good" estimators. They are:

1.3.1.2.1 Unbiasedness(*)

Technically, an unbiased estimator is one whose expected value equals the population parameter. To see what this statement means, examine the sample mean height for males as an estimator of the population mean height for males. Suppose that we were to draw a random sample of *N* observations from the specific population of college males, compute the sample mean, and then enter this sample mean into a data set. We then gathered another sample of *N* observations from the specific population, calculated its sample mean, and add this new sample mean to the data set. We continue with this process until we gathered an infinite number of sample means in the data set. If the mean of all the sample means equaled the mean in the specific population, then the sample mean is an unbiased estimator of the population mean.

1.3.1.2.2 Consistency(*)

A consistent estimator is one whose value gets closer and closer to the population parameter as the sample size increases.

1.3.1.2.3 Efficiency (*)

Efficiency is a relative concept. Given two different estimators of a population parameter, the estimator with the greater efficiency is the one with the smaller standard error for its sampling distribution. Return to the definition of unbiasedness in 1.3.1.2.1. Suppose for each sample drawn, we calculated two statistics, the sample mean (which is

entered into one data set) and the sample median (which is entered into a second data set). We could then calculate the mean and standard deviation of all the sample means as well as the mean and standard deviation of all the sample medians. We would find that both means equal the population mean. This indicates that both the sample mean and the sample median are unbiased estimators of the population mean. However, we would find that the standard deviation of the sample means is smaller than the standard deviation of the sample medians. This indicates that the sample mean is a more efficient estimator than the sample median.

1.3.1.2.4 Sufficiency(*)

A sufficient estimator is one that cannot be improved upon by using any other aspects of the data other than the actual numbers that went into the calculation of the statistic. For example, suppose that we used the sample mean for females to estimate the population mean height for females. Then the sample mean would be a sufficient estimator if no other aspect of the data (e.g, eye color, GPA, astrological sign, etc.) could improve upon estimating the population mean.

1.3.2 Test hypotheses

The second major step in parametric statistics is to test hypotheses and make inferences about the descriptive statistics. This area is referred to as ***inferential statistics*** and it uses test statistics. Inferential statistics are defined as that group of test statistics used to determine the relative probability that a mathematical model of the hypothesis is true or false. The logic behind this is on the complicated side and will be described later in the course. For now, let us consider the height example to obtain an initial taste of the logic behind hypothesis testing.

Inferential statistics assume that the hypothesis can: (1) be stated in precise mathematical terms (i.e., a ***mathematical model***); and (2) the mathematical model is developed without looking at the data. The height example asks whether the average height differs for male and female college students. Looking at the hypothetical population, there are three logical possibilities: (1) mean female height exceeds mean male height or $\mu_f > \mu_m$; (2) mean female and mean male height are the same or $\mu_f = \mu_m$; and (3) mean female height is less than mean male height or $\mu_f < \mu_m$. For reasons that will become clear later in the course (hopefully!), possibilities (1) and (3) are not mathematically precise because they cannot specify (or be mathematically manipulated to specify) a concrete number. For example, possibility (1) states that the female mean is greater than the male mean but does not specify the concrete number of units separating the two means. Possibility (2), however, can be manipulated to give a concrete number—if $\mu_f = \mu_m$, then $\mu_f - \mu_m = 0$.

Hence, the mathematical model that can be tested is $\mu_f - \mu_m = 0$. Logically, it seems fairly simple to test this. If the sample mean is the best estimator of the population mean, then just substitute the sample means into the equation, giving $\overline{Y}_f - \overline{Y}_m = 0$. Consequently, a good test statistic would be the difference between the two

sample means. Let δ equal this quantity so that the equation for the test statistic becomes $\delta = \overline{Y}_f - \overline{Y}_m$. If $\delta = 0$, then the hypothesis of no average difference in height between males and females is supported by the data.

But there is an obvious problem with this logic. Sample means may be the ***best*** estimators of the population means but sample means will ***not always exactly equal*** the population means. In fact, the likelihood that a sample mean equals the population mean to the nth decimal place is vanishingly small. Consequently, we expect δ to be small and close to 0, but it will very seldom equal 0 exactly. So not how do we test the mathematical model?

The answer is that we calculate the probability density function or the sampling distribution of δ. (Remember that statisticians call the probability density function of a test statistic the ***sampling distribution*** of the statistic.) The sampling distribution of δ will provide us with the probability of observing δs that are close to 0 and the probabilities of observing δs that are far away from 0. Hence, we can compare the observed δ from the data with its sampling distribution and have some reasonable, albeit probabilistic, way of determining whether the mathematical model is plausible.

2. Parametric Statistics: The Chick and Gary Approach

2.1 Definition of Parametric Statistics

Chick and Gary's definition of parametric statistics is the same as that given in 1.1 above.

2.2 Key Terms

Chick and Gary's approach contains all the definitions of the traditional approach with one minor twist. Chick and Gary distinguish two types of mathematical models whereas above, in 1.2.9, we gave a generic definition of a mathematical model. The two types of mathematical models defined by C&G are termed the ***compact model*** and the ***augmented model***. Note that the two terms are ***relative***. That is, the same mathematical model that is the augmented model in one situation may be the compact model in another situation. The technical definitions are:

2.2.1 Compact Model

A compact mathematical model is any model that gives the predicted values for all data observations in *p* unknown population parameters.

2.2.2 Augmented model

An augmented mathematical model is any model that gives the predicted values for all data observations in the same *p* unknown population parameters as the compact model plus one or more additional population parameters.

2.3 Steps in C&G's approach to parametric statistics.

The steps are in C&G's approach are the same as those in the traditional approach with a different twist, again involving the compact and augmented mathematical models. The steps in C&G's approach are:

2.3.1 Estimate the parameters of the compact model.

This follows all the conventions of the traditional approach except that it is applied only to the compact model.

2.3.2 Estimate the parameters of the augmented model.

Again, this follows all the conventions of the traditional approach but the conventions are applied only to the augmented model.

2.3.3 Assess the relative fit of the two models.

The ***fit*** of a model to data is determined by the extent to which the predicted values generated by the mathematical model agree with the observed values of the data. The better the predicted values agree with the observed values, the better the fit of the model. C&G stress model comparisons by asking the following question, "How much better does the augmented model fit the data than the compact model?" If the augmented model fits the data much better than the compact model, then one must conclude that the extra parameters in the augmented model are important. If the augmented model does not fit the data much better than the compact model, then the extra parameters in the augmented model are not very important.

3. An example of the traditional and the C&G approach.

The original problem in the height example was to determine whether the mean heights for male and female US were identical. For the present, we will make no assumptions about the distribution of height in either males or females. Instead, we will assume that the specific population of US male students has an unknown mean height of μ_m and an unknown standard deviation of σ_m and that the specific population of US female college students has an unknown mean height of μ_f and an unknown standard deviation of σ_f. We will further assume that the specific population that consists of both male and female students has an overall mean of μ and a standard deviation of σ. We may obtain estimates of all six quantities-- μ, μ_m, μ_f, σ, σ_m, and σ_f—from the sample data, giving respectively $\overline{Y}$, $\overline{Y}_m$, $\overline{Y}_f$, s, s_m, and s_f.

In the traditional approach, we would construct a precise mathematical model that relates the two means. In section 1.3.2, we saw that a good mathematical model would be

$\mu_f = \mu_m$ so that $\mu_f - \mu_m = 0$. Thus, we could calculate the test statistic $\delta = \overline{Y}_f - \overline{Y}_m$ which should be close to 0 if the mathematical model is correct. We would then go to a statistician and ask him/her to construct the sampling distribution (i.e., pdf) of δ for us. We would then plot out the sampling distribution. The possible values of δ, ranging from negative infinity to positive infinity would be on the horizontal axis and the frequency of δ (i.e., relative probability or relative likelihood) would constitute the vertical axis. We would then take the observed value of δ and use the graph to find the relative probability of observing this δ. Of course, the value of δ would be large and negative and the relative probability of observing a large, negative δ would be very small. Hence, we would conclude that the mathematical model $\mu_f = \mu_m$ is very poor which, translated into English, means that US college men and US college women do not have the same average height.

In C&G's approach, we would develop two mathematical models, a compact model and an augmented model. A good compact model would be

$$Y_i = \mu + e_i .$$

In other words, the predicted height for any observation, male or female, is the overall population mean. In the augmented model, we would like to have one additional population parameter to reflect sex differences in height. We could do this by constructing another variable in the data set to reflect sex. Let us denote this variable as *X* and code it in the following way: if the observation is a male then *X* = +1; if the observation is a female, then *X* = -1. We could then write an augmented model as:

$$Y_i = \mu + \beta X_i + e_i .$$

Note how the augmented model looks exactly like the compact model with the addition of one extra population parameter, β. This parameter quantifies sex differences. If β is large and positive, then college men are taller than college women. If β is large and negative then college women are taller than college men. If β is close to 0, then the average height of college men equals that of college women.

The problem now is to see whether the augmented model, $Y_i = \mu + \beta X_i + e_i$, fits the data better than the compact model, $Y_i = \mu + e_i$. To do this, we follow the steps in parametric statistics outlined above in section 1.3. First, we must obtain a "good" estimates of the population parameters μ in the compact model and μ and β in the augmented model. Let us use the least squares criterion for estimation because that is the criterion used most often in psychological research. So we take the problem to a statistician who tells us the following: In the compact model, the best estimate of μ is $\overline{Y}$, the overall sample mean. In the augmented model, the best estimate of μ is still $\overline{Y}$, the overall sample mean, and the best estimate of β is *b* which equals

$$\frac{\sum_{i=1}^{N} X_i(Y_i - \overline{Y})}{\sum_{i=1}^{N} X_i^2}.$$

(NOTE: In estimating these quantities, it is assumed that the number of males in the sample equals the number of females. And for the moment, let us "punt" on the issue of how the statistician arrived at these results.)

Now that we have obtained the parameter estimates, we use inferential statistics to test the hypothesis about sex differences in height. There are two ways of doing this, both of which result in the same mathematical result but express the information in different ways.

The first is to use *b*, the estimate of β, as a test statistic. Re-examine the equations for the compact and augmented model:

$$Y_i = \mu + e_i$$

and

$$Y_i = \mu + \beta X_i + e_i .$$

Obviously, if there are no sex differences then $\beta = 0$ and the augmented model equals the compact model. The only other two logical possibilities for β are $\beta > 0$ (men are taller than women) or $\beta < 0$ (women are taller than men). However, these two possibilities are untestable because they are mathematically imprecise—they do not provide a specific number for β.

Consequently, we want to test the hypothesis that $\beta = 0$ using the estimate of β, i.e. *b*, as the test statistic. Once again, we cannot simply look at *b* and see if it is 0 or not because of sampling error. Hence, we take the problem to our favorite statistician and request a plot of the sampling distribution of *b*. This would plot all the mathematical possibilities of *b* on the horizontal axis and the relative likelihood of *b* on the vertical axis. The observed value of *b* will be large and positive. (Recall that $X = 1$ for males but $X = -1$ for females.) The relative likelihood of observing a large positive value of *b* would be very small according to the plot. Hence, we would conclude that it is very unlikely that the population parameter β is really 0. Men are indeed taller than women,

A second, mathematically equivalent way of testing whether mean height in males equals that in females is to compute another test statistic based on the prediction errors. If the model $Y_i = \mu + e_i$ is true then the quantity β in the augmented model, $Y_i = \mu + \beta X_i + e_i$, should be 0 and the prediction errors in the augmented model should equal those in the compact model.

To do this, we would estimate μ in the compact model $Y_i = \mu + e_i$, calculate the predicted value for each observation (which would equal μ), and then calculate the residual (i.e., error) for each observation in the sample. We could then calculate the sum of squared errors for the compact model. Let us denote the sum of squared errors for the compact model as SSE_C. Then

$$SSE_C = e_{C1}^2 + e_{C2}^2 + ... + e_{CN}^2 = \sum_{i=1}^{N} e_{Ci}^2$$

(The subscript *C* is used to signify that the errors have been computed from the compact model.)

We would then estimate μ and β from the augmented model, $Y_i = \mu + \beta X_i + e_i$ and calculate the predicted value for each observation in the sample. Here, the predicted value would equal μ + β if the observation is male but μ - β if the observation is female. We would then calculate the prediction error (i.e., residual) for each observation. Finally, we would calculate the sum of squared prediction errors, say SSE_A, for the augmented model:

$$SSE_A = e_{A1}^2 + e_{A2}^2 + ... + e_{AN}^2 = \sum_{i=1}^{N} e_{Ai}^2 .$$

If the augmented model is really a better model, then SSE_A should be much smaller than SSE_C. If the augmented model is not that much better then SSE_A should be roughly equal to SSE_C. So let us develop a test statistic based on the two sums of squared error, SSE_A and SSE_C.

At first glance, a good test statistic might be the simple difference between the two or $SSE_C - SSE_A$. If the augmented model is not better than the compact model then $SSE_C - SSE_A$ should be close to 0. But if the augmented model is superior to the compact model, then $SSE_C - SSE_A$ should be large and positive. There are two problems, however, with $SSE_C - SSE_A$ that make this a poor test statistic—(1) it depends on sample size; the larger the sample size, the larger the sum of squared error; and (2) it depends on the measurement metric; for example, sum of squared error will be larger for height measured in centimeters than for height measured in inches.

To get around these problems, C&G suggest that the difference $SSE_C - SSE_A$ be divided by SSE_C. Hence, the test statistic is

$$\frac{SSE_C - SSE_A}{SSE_C} .$$

This is an important test statistic that C&G call *PRE* for the *P*roportional *R*eduction in *E*rror. Mathematically, it can be shown that *PRE* has a lower limit of 0 (when $SSE_A = SSE_C$)[1] and an upper limit of 1 (when $SSE_A = 0$). Hence, if there is no mean difference in height, then the squared error for the augmented model will equal than for the compact model and *PRE* should be close to 0. If the augmented model is superior, then SSE_C should be greater than SSE_A and PRE should be positive.

Hence, we could take *PRE* to our favorite statistician and request the sampling distribution of *PRE* under the assumption that the augmented model is really the same as the compact model. (Remember, we cannot find the distribution of *PRE* under the assumption that the augmented model is *better* than the compact model because, once again, we would have to come up with a specific number for *b* in the augmented model.) The statistician would then give us the sampling distribution of *PRE.* This would plot the value of *PRE*, ranging from 0 to 1, on the horizontal axis and the relative likelihood of

[1] For technical reasons, SSE_A must always be less than SSE_C. Hence, $SSE_C - SSE_A$ can never be negative.

observing that value on the vertical axis. Because men are taller than women, the actual value of *PRE* would be large and it would be very unlikely to observe a large value of PRE under the hypothesis that men and women have the same height.

4. Four Important Descriptive Statistics and Their Formula

Both the traditional and the C&G approach use four basic descriptive statistics. Almost all other formula can be expressed as a function of one or more of these statistics. The four statistics are:

4.1 Arithmetic Mean

The arithmetic mean is simply the arithmetic average. It is computed by summing the scores on a variable over all observations and then dividing by the number of observations. For a specific population, the formula is

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}.$$

The sample mean is an unbiased estimator of the population mean. Hence,

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}.$$

The hat (^) over the μ denotes an estimator of μ. Thus, the English meaning of the symbol $\hat{\mu}$ is "an estimate of the population mean."

There are types of means other than the arithmetic mean (the geometric mean and the harmonic mean). However, these are seldom encountered in traditional parametric statistics, so the term "the mean" is always taken as the arithmetic mean.

The arithmetic mean is a measure of central tendency. That is, it answers the question, "Around which number are the scores clustered?"

4.2 Variance

The variance is a measure of variability. It answers the question, "How dispersed are the scores around their central tendency?" There are two types of variance.

4.2.1 Variance of a Specific Population

The variance of a specific population is the average squared deviation from the mean. One would first calculate the arithmetic mean or μ for the specific population. For each observation in the specific population, one would then calculate the deviation from the mean. For the ith observation, the deviation from the mean is simply $X_i - \mu$. One would then square each deviation from the mean giving $(X_i - \mu)^2$ for the ith observation. The variance of the specific population is then the arithmetic average of these squared deviations. Consequently, the formula for the variance of a specific population is

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}.$$

4.2.2 Sample Variance as an Estimator of the Variance of a Specific or Hypothetical Population

Because it is usually impractical to enumerate all the observations in a specific population and because it is impossible to calculate any parameter for a hypothetical population, sample data are used to estimate the population variance. When sample data are used to estimate the population variance the definition of a variance changes slightly. Here, the variance is defined as the sum of the squared deviations from the mean divided by the degrees of freedom left in the data.

We have encountered the sum of squared deviations from the mean above in discussing the variance of a specific population. Here, we just note some important terminology. The sum of squared deviations from the mean is usually just called ***the sum of squares*** and is abbreviated as ***SS***.

We have not encountered the term ***degrees of freedom*** before. In English, the degrees of freedom for an estimator is the amount of information required over and above the estimator to figure out all the scores in a sample. For example, suppose we had 6 scores and calculated the mean. Then, given that we know the mean, how many of the six scores would we have to know before we could figure out the remaining scores? The answer is 5. Once we know the mean, then we only need to know any 5 scores to figure out the remaining score. Consequently, once the mean is known the data have 5 degrees of freedom left. In general, if a mean is calculated from *N* scores, then there will be *N* - 1 degrees of freedom left in the data.

Consequently, the formula for using sample data to estimate a population variance becomes

$$\hat{\sigma}^2 = s^2 = \frac{SS}{df} = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N - 1}.$$

Once again, the hat (^) over σ^2 denotes an estimator of σ^2; translated into English, the notation $\hat{\sigma}^2$ means "an estimate of the population variance." Because every individual squared deviation from the mean must be positive, the sum of squared deviations from the mean must also be positive. Likewise, the degrees of freedom must always be positive. Consequently, the variance will always be positive (or 0, if all the scores are the same number). If you ever calculate the variance and arrive at a negative number, then you must have made a computational error.

Closely related to the variance is its sibling statistic, the standard deviation. The standard deviation is simply the square root of the variance. Consequently, because $\sqrt{\sigma^2} = \sigma$, the symbol σ is used to denote a population standard deviation. For analogous reasons, the symbol *s* is often used to denote the estimate of a population standard

deviation from sample data. Because a variance is always positive, its standard deviations is always taken as the positive root.

4.3 Covariance

The covariance is a statistic that measures the extent to which two variables vary together. If we let *X* denote one variable and *Y* denote the other variable, then the formula for the covariance in a specific population is

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N}.$$

Unlike the mean (where μ customarily denotes the population parameter) and the variance (where σ^2 customarily denotes the population parameter), there is no conventional Greek letter used to denote the covariance.

When sample data are used to estimate the population covariance, the denominator is replaced by the degrees of freedom or *N* - 1. Thus, the formula for the estimate of a population covariance is

$$\hat{\mathrm{cov}}(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N-1}.$$

There are two attributes of a covariance—(1) the sign of the covariance and (2) the absolute value of the covariance. The sign of a covariance indicates the direction of the relationship between *X* and *Y*. A positive covariance means that high scores on *X* are associated with high scores on *Y* and, conversely, that low scores on *X* are associated with low scores on *Y*. An example of a positive covariance would be the relationship between height and weight. People who are taller than average also tend to weigh more than the average person. Similarly, people who are smaller than average tend to weigh less than the average person.

A negative covariance denotes an inverse relationship. Here high scores on *X* are associated with high scores on *Y* while, conversely, low scores on *X* are associated with low scores on *Y*. An example of a negative covariance would be the relationship between grades and the amount of time spent partying. Folks who party a lot, and presumably do not have as much time to study, tend to have low grades while students who sacrifice play for study will tend to get higher grades.

The absolute value of a covariance is a measure of the magnitude of the relationship. A covariance of 0 denotes no statistical association between *X* and *Y*. As the covariance departs from 0 in either a positive or negative direction, the relationship between *X* and *Y* becomes stronger.

4.4 Correlation.

Although most major statistical procedures operate with the covariance, the covariance is a very poor descriptive statistic for communication among scientists. The

reason is that the telling someone that "the covariance is 20" is akin to telling him/her that the price of an ice cream cone is 20 herns without specifying how much a hern is. If a hern is close to an English pound, then it is a very expensive ice cream cone (about $35). But if a hern is close to an Italian lira, then the ice cream cone is quite the bargin (about 6 cents). Similarly, a covariance of 20 might indicate a very strong relationship or one that is hardly worth caring about—it all depends on the measurement units for *X* and *Y*.

To overcome this difficulty, statisticians convert a covariance into the statistical equivalent of a common currency, the correlation coefficient[2]. The advantage of the correlation coefficient is that it has all the properties of a covariance (i.e., the sign denotes direction and absolute value denotes strength of association) but it has a mathematical lower bound of -1.0 and a mathematical upper bound of 1.0. When a correlation equals -1.0 or 1.0, then one can perfectly predict the value of *Y* from any *X* value. That is, there are no prediction errors for all observations. As the correlation coefficient gets closer to 0, the statistical relationship between the two variables becomes weaker. Like a covariance of 0, a correlation of 0 denotes no statistical association between two variables.

Consequently, it is possible to compare the magnitude of correlations while it is not possible to compare the magnitude of covariances. For example, if the correlation between variables *X* and *Y* is .20 while the correlation between variables *A* and *B* is .35, then it is possible to state that the variables *A* and *B* are more strongly associated than variables *X* and *Y*. The same could not be made for covariances.

The Greek letter rho or ρ usually denotes a population correlation. The estimate of the population correlation is given by the Roman equivalent, *r*, although the uppercase *R* is frequently encountered. The formula for the population correlation is

$$\rho_{XY} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

For sample data, the sample statistics are replaced into the above equation, giving

$$\hat{\rho}_{XY} = r_{XY} = \frac{\hat{\text{cov}}(X,Y)}{s_X s_Y}.$$

One important property about correlations is that the square of the correlation gives the proportion of variance in one variable attributable or predicted by the other variable. For example, if the correlation between height and weight is .60, then 36% of the variance in weight is predictable from knowing height. Similarly, 36% of the variance in height is predictable from knowing weight. It is essential never to think of the phrases "variance atributable" or "variance predicted by" in causal terms. A correlation coefficient only denotes a statistical relationship. The statistical association may or may not be causal, but a correlation coefficient alone is not sufficient to deteremine causality.

---

[2] There are actually several different types of correlation coefficients. The one outlined here is officially called the Pearson product-moment correlation after a famous statstician, Karl Pearson, who developed many of its principles around the turn of the century.