STAT 518 --- Section 2.1: Basic Inference

Basic Definitions

Population: The collection of all the individuals of interest.

• This collection may be _______ or even ____________.

Sample: A collection of elements of the population.

• Suppose our population consists of a finite number (say, *N*) of elements.

Random Sample: A sample of size *n* from a finite population such that each of the possible samples of size *n* was

Another definition:

Random Sample: A sample of size *n* forming a sequence of

• Note these definitions are equivalent only if the elements are drawn ________ __________________ from the population.

• If the population size is very large, whether the sampling was done with or without replacement makes little practical difference.

<u>Multivariate Data</u>

• Sometimes each individual may have <u>more than one</u> variable measured on it.

• Each observation is then a <u>multivariate</u> random variable (or ____________ ____________)

<u>Example</u>: If the weight and height of a sample of 8 people are measured, our <u>multivariate</u> data are:

• If the sample is random, then the components Y_{i1} and Y_{i2} might not be independent, but the vectors $\underline{X}_1$, $\underline{X}_2$, …, $\underline{X}_8$ will still be independent and identically distributed.

• That is, knowledge of the value of $\underline{X}_1$, say, does not alter the probability distribution of $\underline{X}_2$.

Measurement Scales

• If a variable simply places an individual into one of several (unordered) categories, the variable is measured on a _____________ scale.

<u>Examples</u>:

• If the variable is categorical but the categories have a meaningful ordering, the variable is on the ___________ scale.

<u>Examples</u>:

• If the variable is numerical and the value of zero is arbitrary rather than meaningful, then the variable is on the ______________ scale.

<u>Examples</u>:

• For <u>interval</u> data, the interval (difference) between two values is meaningful, but <u>ratios</u> between two values are not meaningful.

• If the variable is numerical and there is a meaningful zero, the variable is on the __________ scale.

<u>Examples</u>:

• With <u>ratio</u> measurements, the ratio between two values has meaning.

Weaker ←---------------------------------→ Stronger

• Most classical parametric methods require the scale of measurement of the data to be interval (or stronger).

• Some nonparametric methods require ordinal (or stronger) data; others can work for data on any scale.

• A <u>parameter</u> is a characteristic of a population.

<u>Examples</u>:

• Typically a parameter cannot be calculated from sample data.

• A <u>statistic</u> is a function of random variables.

• Given the data, we can calculate the value of a statistic.

<u>Examples of statistics</u>:

Order Statistics

• The *k*-th order statistic for a sample $X_1, X_2, \ldots, X_n$ is denoted $X^{(k)}$ and is the *k*-th smallest value in the sample.

• The values $X^{(1)} \leq X^{(2)} \leq \ldots \leq X^{(n)}$ are called the <u>ordered random sample</u>.

<u>Example</u>: If our sample is: 14, 7, 9, 2, 16, 18
then $X^{(3)} =$

<u>Section 2.2: Estimation</u>

• Often we use a statistic to <u>estimate</u> some aspect of a population of interest.

• A statistic used to estimate is called an <u>estimator</u>.

<u>Familiar Examples</u>:

• The sample mean:

• The sample variance:

• The sample standard deviation:

• These are point estimates (single numbers).

• An interval estimate (confidence interval) is an interval of numbers that is designed to contain the parameter value.

• A 95% confidence interval is constructed via a formula that has 0.95 probability (over repeated samples) of containing the true parameter value.

Familiar large-sample formula for CI for μ:

Some Less Familiar Estimators

• The cumulative distribution function (c.d.f.) of a random variable is denoted by F(*x*):

$$F(x) = P(X \leq x)$$

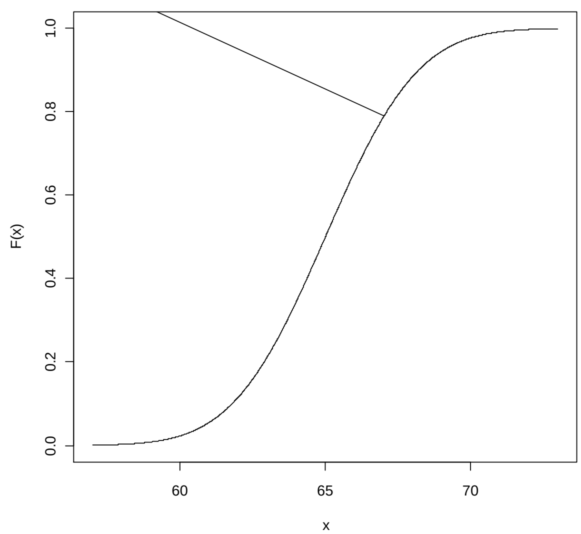
• This is $\int_{-\infty}^{x} f(t)dt$ **when *X* is a continuous r.v.**

Example: If *X* is a normal variable with mean 100, its c.d.f. F(*x*) should look like:

• Sometimes we do not know the distribution of our variable of interest.

• The empirical distribution function (e.d.f.) is an estimator of the true c.d.f. – it can be calculated from the sample data.

Example: Suppose heights of adult females have normal distribution with mean 65 inches and standard deviation 2.5 inches. The c.d.f. of this distribution is:



• Now suppose we do NOT know the true height distribution. We randomly sample 5 females and measure their heights as: 69.3, 66.3, 62.6, 62.9, 67.4

e.d.f.:

• The <u>survival function</u> is defined as 1 – F(*x*), which is the probability that the random variable takes a value greater than *x*.
• This is useful in reliability/survival analysis, when it is the probability of the item surviving past time *x*.

• The Kaplan-Meier estimator (p. 89-91) is a way to estimate the survival function when the survival time is observed for only some of the data values.

<u>The Bootstrap</u>

• The nonparametric bootstrap is a method of estimating characteristics (like expected values and standard errors) of summary statistics.

• This is especially useful when the true population distribution is unknown.

• The nonparametric bootstrap is based on the e.d.f. rather than the true (and perhaps unknown) c.d.f.

<u>Method</u>: Resample data (randomly select *n* values from the original sample, with replacement) *m* times.

• These "bootstrap samples" together mimic the population.

• For each of the *m* bootstrap samples, calculate the statistic of interest.

• These *m* values will approximate the sampling distribution.

• From these bootstrap samples, we can estimate the:

(1) expected value of the statistic
(2) standard error of the statistic
(3) confidence interval of a corresponding parameter

<u>Example</u>: We wish to estimate the 85^{th} percentile of the population of BMI measurements of SC high schoolers.

• We take a random sample of 20 SC high school students and measure their BMI.

• See code on course web page for bootstrap computations: