STAT 518 --- Section 2.1: Basic Inference

Basic Definitions

Population: The collection of all the individuals of interest.

• This collection may be large **or even** infinite**.**

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 equally likely to have been obtained.

Another definition:

Random Sample: A sample of size n forming a sequence of n independent and identically distributed (iid) random variables $X_1, X_2, \ldots, X_n$.

• Note these definitions are equivalent only if the elements are drawn with replacement **from the population.**

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

Multivariate Data

• Sometimes each individual may have more than one variable measured on it.

• Each observation is then a multivariate random variable (or random vector)

$$\underset{\sim}{X}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ik})$$

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

$$\underset{\sim}{X}_1 = (Y_{11}, Y_{12})$$

weight (Y_{11}), height (Y_{12})

$$\underset{\sim}{X}_2 = (Y_{21}, Y_{22})$$

$$\vdots$$

$$\underset{\sim}{X}_8 = (Y_{81}, Y_{82})$$

• If the sample is random, then the components Y_{i1} and Y_{i2} might not be independent, but the vectors $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_8$ 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** nominal **scale.**

Examples: hair colors of people
Majors of students
genders of people

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

Examples: Ratings of movies
Course Grades of students
Age Groups
Likert-Type Scale (Strongly Agree, Agree, ...)

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

Examples: Temperature in °C
Temperature in °F

• **For interval data, the interval (difference) between two values is meaningful, but ratios between two values are not meaningful.**

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

Examples: Height
Speed
Age
Weight Loss

• With ratio measurements, the ratio between two values has meaning.

nominal ordinal interval ratio

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 parameter is a characteristic of a population.

Examples:
Population mean μ
Population standard deviation σ
Population proportion p
Population median

• Typically a parameter cannot be calculated from sample data.

• A statistic is a function of random variables.

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

Examples of statistics: Sample mean $\bar{X}$
sample standard deviation s

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 ordered random sample.

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

Section 2.2: Estimation

• Often we use a statistic to estimate some aspect of a population of interest.

• A statistic used to estimate is called an estimator.

Familiar Examples:

• The sample mean:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

• The sample variance:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

• The sample standard deviation:

$$S = \sqrt{S^2}$$

• 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:

$$\left(\bar{X} - z_{1-\alpha/2}\frac{S}{\sqrt{n}},\ \bar{X} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right)$$

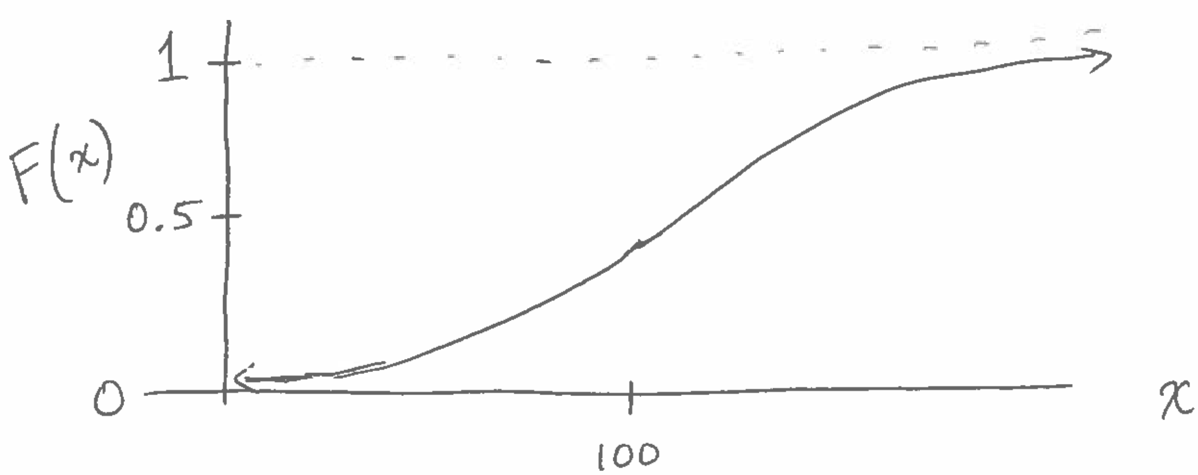
Some Less Familiar Estimators

• 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.

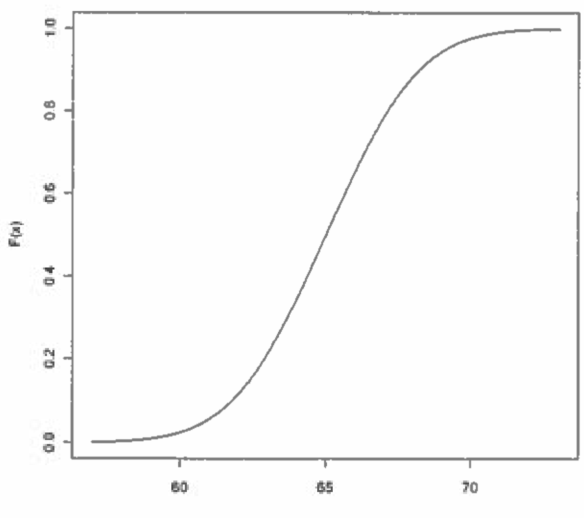
Example: If X is a normal variable with mean 100, its c.d.f. F(x) should look like:



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

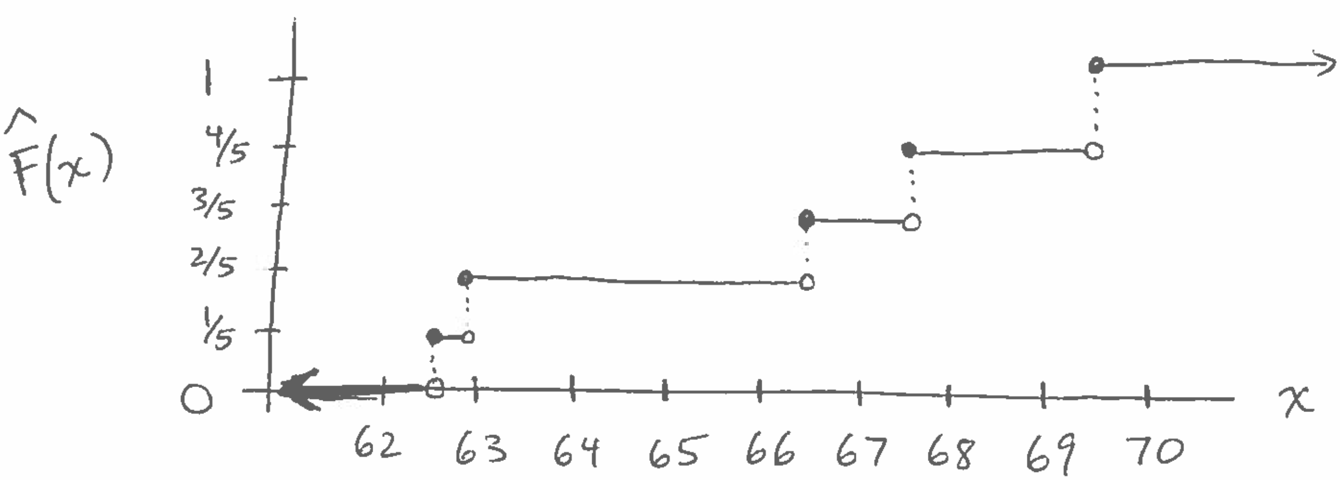
• The empirical distribution function (e.d.f.) is an estimator of the true c.d.f. – it can be calculated from the sample data.

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.

The Bootstrap

• 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 85th 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:**

Estimated standard error of sample 85th percentile is 1.65.

A 95% bootstrap CI for the population 85th percentile is:

(26.6, 30.65)