STAT 333
--- 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 $X_1, X_2, \ldots\ldots, X_n$

•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)

$$\underline{X}_i = (y_{i1}, y_{i2}, \ldots\ldots, y_{ik})$$

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

$\underline{X}_1 = (y_{11}, y_{12})$
$\underline{X}_2 = (y_{21}, y_{22})$
$\underline{X}_3 = (y_{31}, y_{32})$
$\underline{X}_4 = (y_{41}, y_{42})$
$\underline{X}_5 = (y_{51}, y_{52})$
$\underline{X}_6 = (y_{61}, y_{62})$
$\underline{X}_7 = (y_{71}, y_{72})$
$\underline{X}_8 = (y_{81}, y_{82})$

where, y_{i1} : weight , y_{i2}: height
$i=1,2,\ldots\ldots 8$

• 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

Nominal Scale:

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

Examples:

Hair color ,Gender , Nationality, Major

Ordinal Scale:

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

Examples:

Grades of students, Rating of movies, Education level ,
Likerty-Type scale (Strongly agree, agree ,….)

Interval Scale:

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^o
Temperature in F^o

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

Ratio Scale:

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

Examples:

Height , Speed ,Age, Weight loss ,height

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

Weaker ←----------→ Stronger

Nominal ordinal interval ratio

Note:

• 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 of parameter :

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 statistic :

Sample mean
Sample standard deviation (S)
Sample proportion ($\hat{p}$)

Sample median

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 $X^{(3)}$ =

$X^{(5)}$ =

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:

$$(\bar{X} - Z_{1-\frac{\propto}{2}}\frac{s}{\sqrt{n}}, \bar{X} + Z_{1-\frac{\propto}{2}}\frac{s}{\sqrt{n}})$$

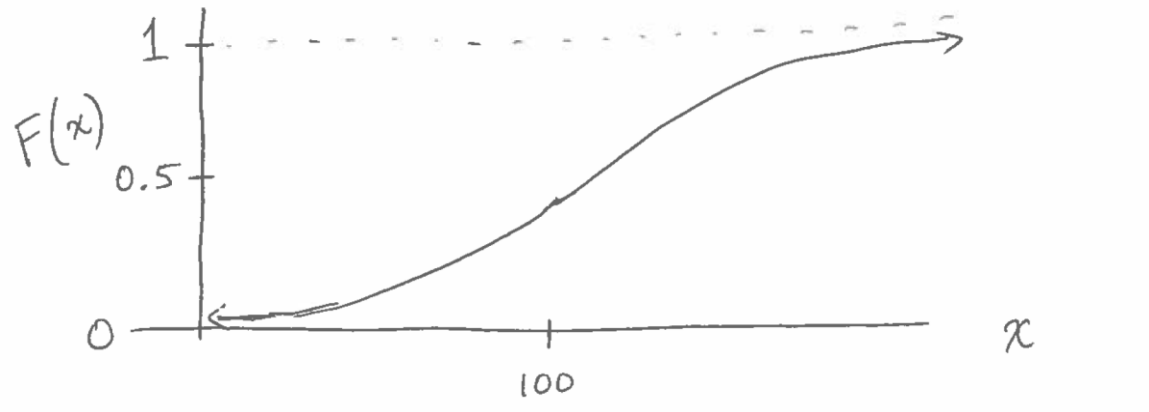
Some Less Familiar Estimators

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

$$\mathbf{F(}x\mathbf{) = P(}X \leq x\mathbf{)}$$

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

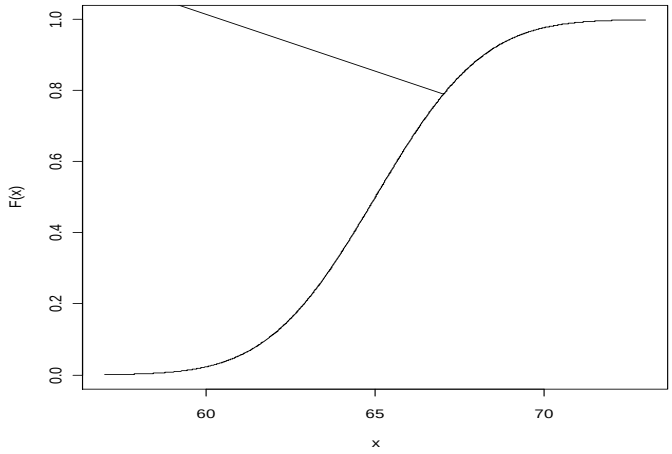
Example:
If X is a normal variable with mean 100, its c.d.f. F(x) should look like:



• 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.

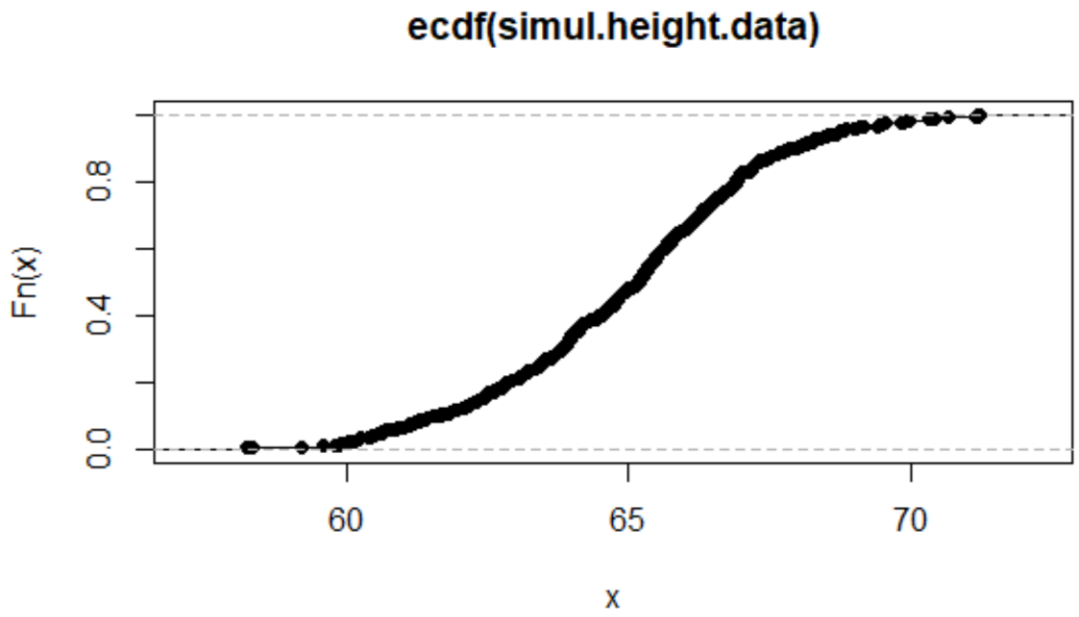
Example: Suppose heights of adult females have normal distribution with mean 65 inches and standard deviation 2.5 inches. The c.d.f. of this distribution is:



R Code:

```
# An example with a simulated data set with LOTS of
observations:
simul.height.data <- rnorm(n=500, mean=65, sd=2.5)
plot(ecdf(simul.height.data))
plot(ecdf(simul.height.data), verticals=TRUE)
plot(ecdf(simul.height.data), verticals=TRUE, do.points=FALSE)
```

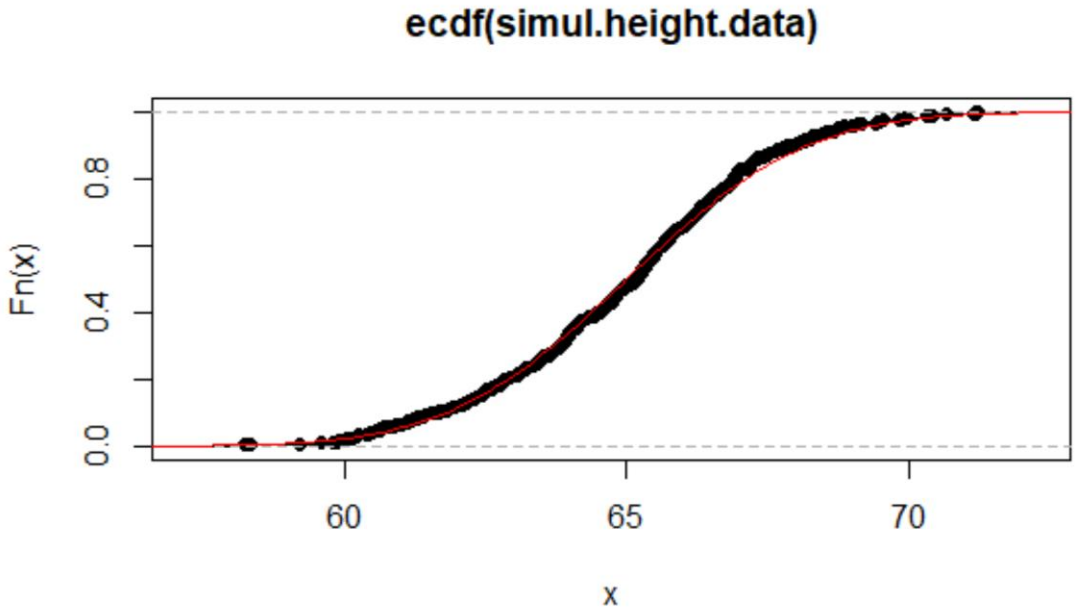
Output:



R Code:

```
gridpts <- seq(55,75,by=0.1)
lines(gridpts,pnorm(gridpts,mean=65,sd=2.5),col='red') # true cdf superimposed in red
```

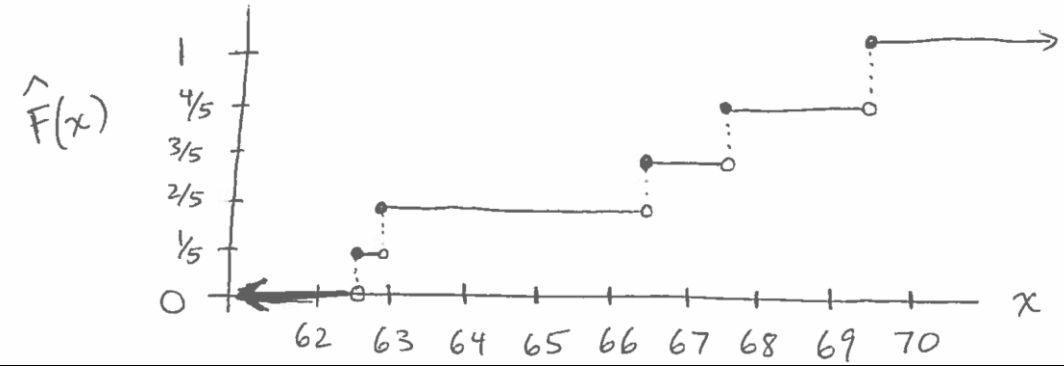
Example



Example :

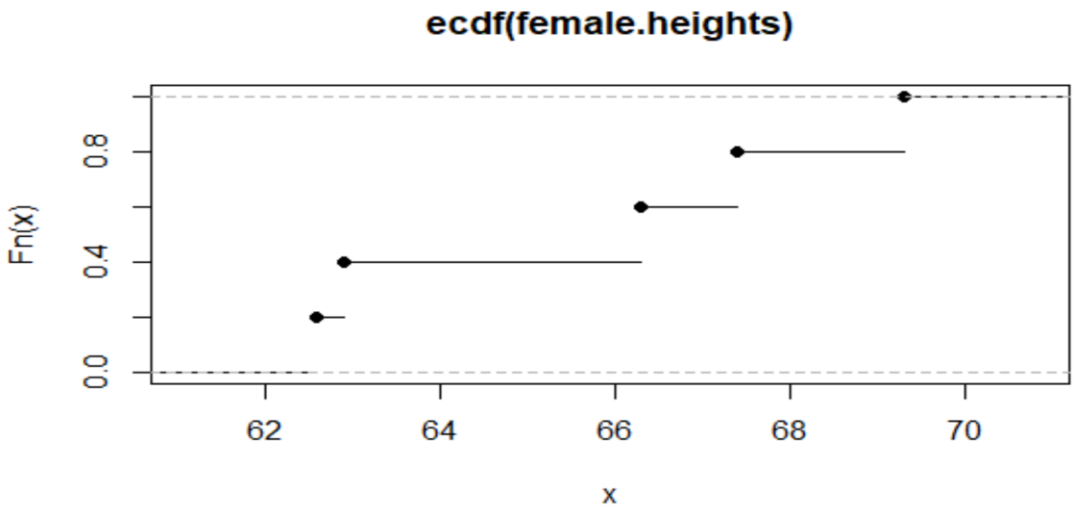
Now suppose we do NOT know the true height distribution.

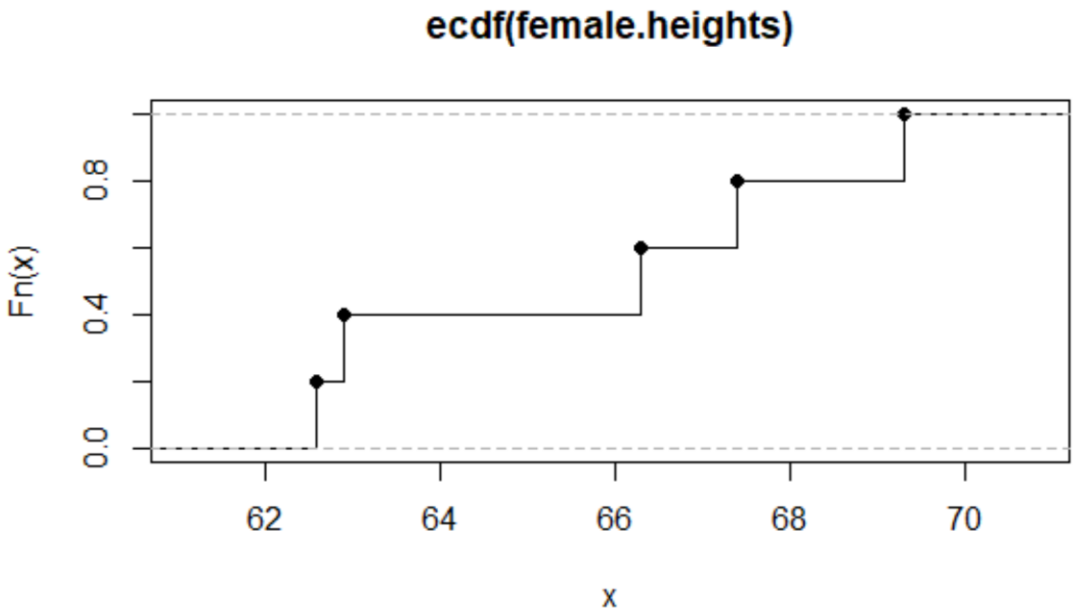
We randomly sample 5 females and measure their heights as: 69.3, 66.3, 62.6, 62.9, 67.4

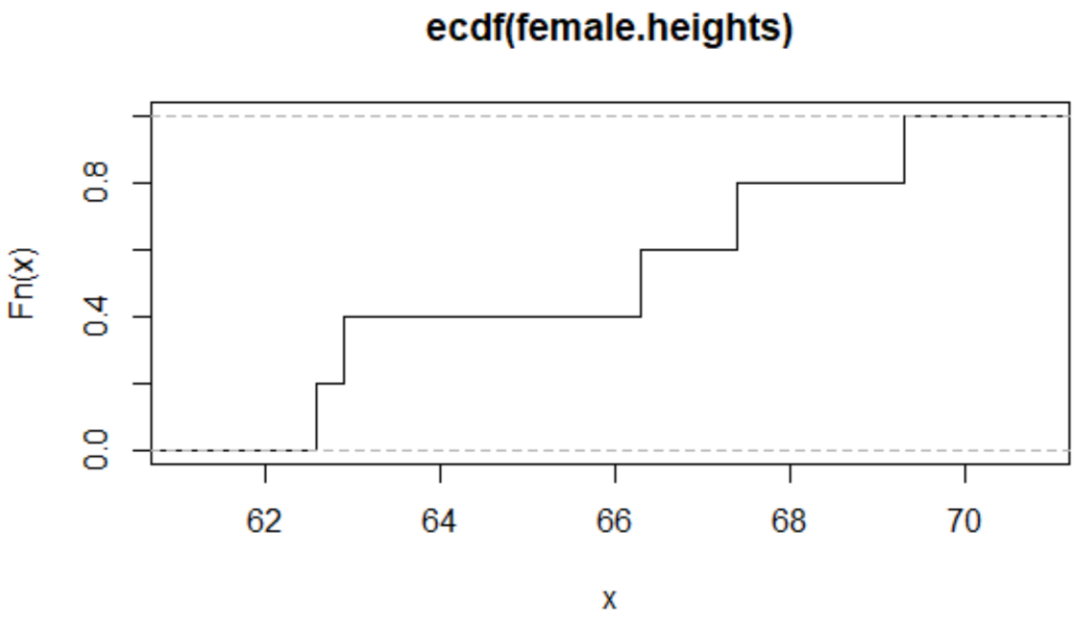


R Code:

```
# Example from class:
female.heights <- c(69.3,66.3,62.6,62.9,67.4)
plot(ecdf(female.heights))
plot(ecdf(female.heights), verticals=TRUE)
plot(ecdf(female.heights), verticals=TRUE, do.points=FALSE)
```







• The **survival function** 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.

Method: 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

Example: 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)

R Code:

```
### Bootstrap example:

# function to calculate the 85th percentile of a sample vector:

perc85 <- function(input.vec){
output <- quantile(input.vec, prob=0.85)
return(output)
}
```

* ادخال البيانات 20 قيمه وهي توزيع

```
binomail

bmi.samp <-
c(21.8,36.6,22.0,24.4,22.2,20.0,19.2,21.6,27.2,28.9,19.4,28.1,18.6,26.6,
20.6,26.7,26.5,25.3,29.6,24.7)
my.n <- length(bmi.samp)
```

*ايجاد عينات

```
#Defining the number of resamples:
my.m <- 1000

*Setting up the matrix to hold bootstrap-sample values

setup.data.matrix <- matrix(bmi.samp, nrow=my.m, ncol=my.n,
byrow=T)

* carrying out the sampling (with replacement):

bootstrap.data.matrix <- apply(setup.data.matrix, 1, sample, size=my.n,
replace=TRUE)

* Transposing to get back to same dimensions as setup.data.matrix

bootstrap.data.matrix <- t(bootstrap.data.matrix)
```

```
# Calculating the sample mean for each of the bootstrap samples

my.85.percs <- apply(bootstrap.data.matrix, 1, perc85)

*Find standard error of this statistic (85th percentile) :

sd(my.85.percs)

*Find 95% Bootstrap interval estimates for population 85th percentile:

lower.upper.CI <- quantile(my.85.percs, probs=c(0.025, 0.975))

print(paste("95% bootstrap interval for 85th percentile: ",
round(lower.upper.CI,2) ))
```

Output:

```
>sd(my.85.percs)
[1] 1.43296


>lower.upper.CI <- quantile(my.85.percs, probs=c(0.025, 0.975))

>print(paste("95% bootstrap interval for 85th percentile: ",
round(lower.upper.CI,2) ))
[1] "95% bootstrap interval for 85th percentile:  26.5"
[2] "95% bootstrap interval for 85th percentile:  30.65"
```