Econ 300/QAC 201: Quantitative Methods in Economics/Applied Data Analysis

2nd Class **6/8/10**

"The statistics you don't compile never lie."
Stephen Colbert

show updated website with workshop dates and times

bring book to class as we will use it often to look at a problem or a table

Quick review: The most important lessons of the course are in the first three chapters
--Don't be fooled by statistics! Learn to use and interpret them correctly
    --are data collected in a reliable manner?
--Think about how experimental outcomes are influenced by the way the experiment is set up
    --truly random?
    why is randomization a good thing?--Because while it is uncertainty, it is manageable uncertainty
    --truly independent?
        --relation, or correlation, does not mean causation
--We are constantly faced with the need to make decisions under uncertainty
    --comes up in economics in both micro and macro contexts;
    --need to work out sound methods to avoid using bad heuristics/techniques

review confidence interval concept from last time--can make surer statements
about things the more data are collected

also think about confidence interval as indicating by its width how relatively sure different people are about different things and how often you would make a mistake (say at a 95% level; contrast to 99% or 90% levels)

can also have confidence intervals on other things besides proportions, e.g.,
means (average weight) and totals (e.g., population)

Question for the day: do people correctly construct confidence intervals to take account of their uncertainty?

[I asked countries of origin last time to make sure no one would have a homefield advantage on this question]

on a piece of paper, give me your 95 percent confidence interval for the population of South Korea [specifically, the early-2010 (updated as of 3/19/10) population of 48,636,068, according to the U.S. Census on-line international database: http://www.census.gov/ipc/www/idb/ranks.php]
which has been in the news lately regarding bad relations with North Korea

--Ch. 2

use class heights to illustrate simple descriptive graphs and statistics
[show excel spreadsheets]
show frequency graphs for height--compare one in book to ones for class--why different?

go through how to construct various measures of central tendency and dispersion

why have more than one measure of central tendency? Each has advantages. The mean is simpler to calculate and manipulate. But the median is not sensitive to outliers/extreme values. Example: In 1984 the University of Virginia announced that its department of rhetoric and communications graduates' mean starting salary was $55,000 (that would be about $112,000 in current dollars—used the inflation calculator at http://www.westegg.com/inflation/). The outlier, the salary of NBA center Ralph Sampson, did not represent the earning power of a B.A. in speech from UVA (the median salary was not published).

point out how tests will be constructed later in the class which could rigorously test for significant differences in male and female height distributions

[handout on elementary rules of summation]

sample mean: $\overline{X} = \frac{1}{\mathbf{n}} \sum \mathbf{X}$

sample variance: $s^2 = \frac{1}{\mathbf{n}-1} \sum (\mathbf{X} - \overline{\mathbf{X}})^2$

motivate variance measure as using degrees of freedom (n-1) instead of number of observations in sample (n)

sample standard deviation: s

The fundamental descriptive statistics of a sample are sample size (n), mean, and standard deviation.

most observations in a sample will lie within two standard deviations on each side of the mean

show how sample mean is weighted average of subsample means
(note that sample variance is <u>not</u> the weighted average of subsample variances;
rather, one has to know the relative frequencies of x's for each subsample to calculate the sample variance)

consider general linear transformations, again using the height data

add shoes of constant height of 6 inches to everyone: $X' = X + 6''$

$\overline{X}' = \overline{X} + 6''$, $s' = s$

express everyone's height in centimeters instead: $X' = 2.54X$

$\overline{X}' = 2.54\,\overline{X}$, $s' = |2.54|s$

general case: $X' = a + bX$ (where a and bX need to be in the same units, e.g. centimeters); $\overline{X}' = a + b\,\overline{X} + 6''$, $s' = |b|s$

discuss principles of graphs [handouts on rules for good graphs][overheads of bad graphs]

now that you've seen some bad graphs; let's look at some examples of good graphs
a lot of interesting possibilities for interactive graphs now
[go through links on the webpage]

for next class, read as much of Chapter 3 as you can get through; we'll spend 4 lectures total on Chs 3, 4, and 5

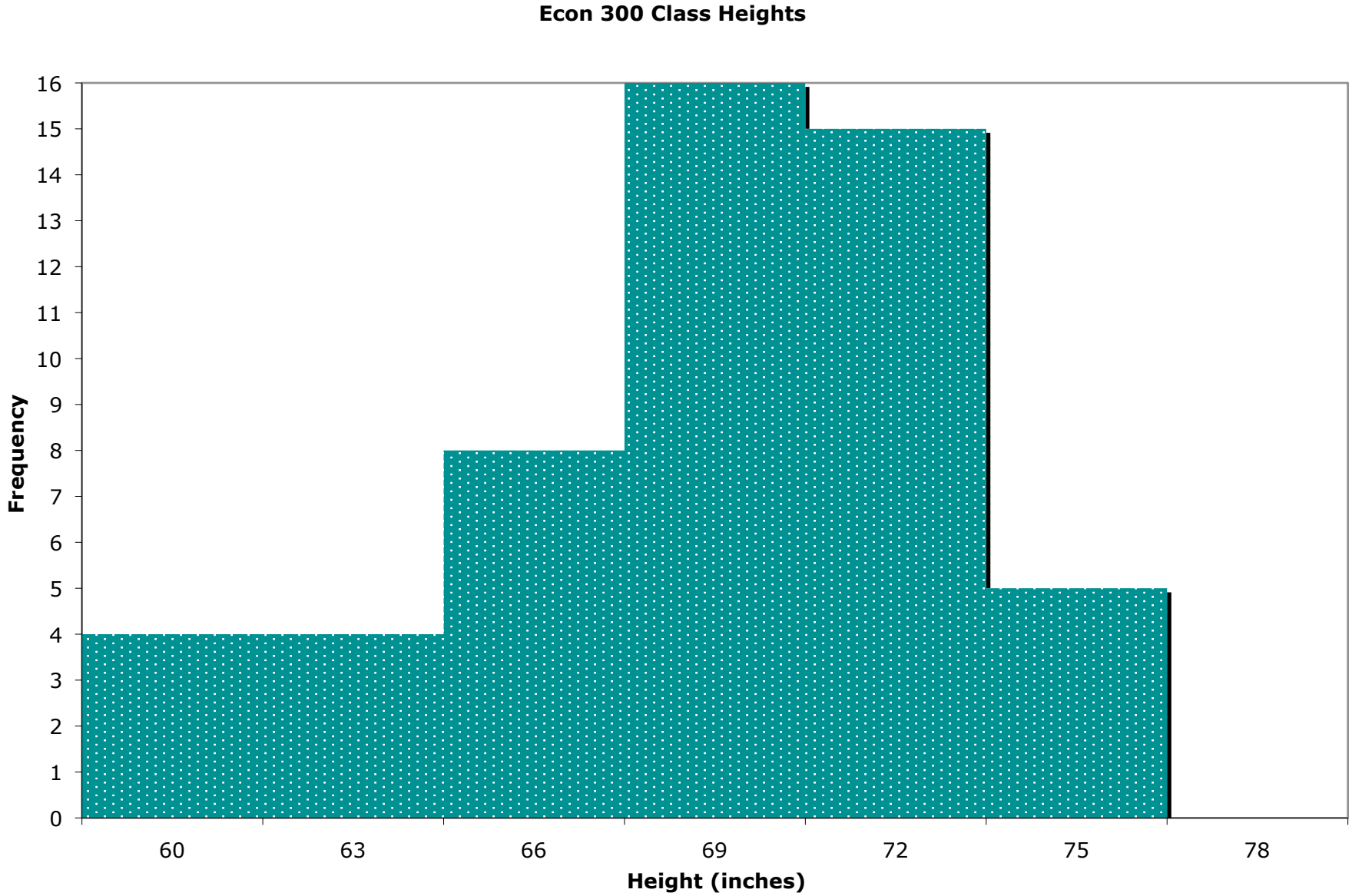
hand out ps#2

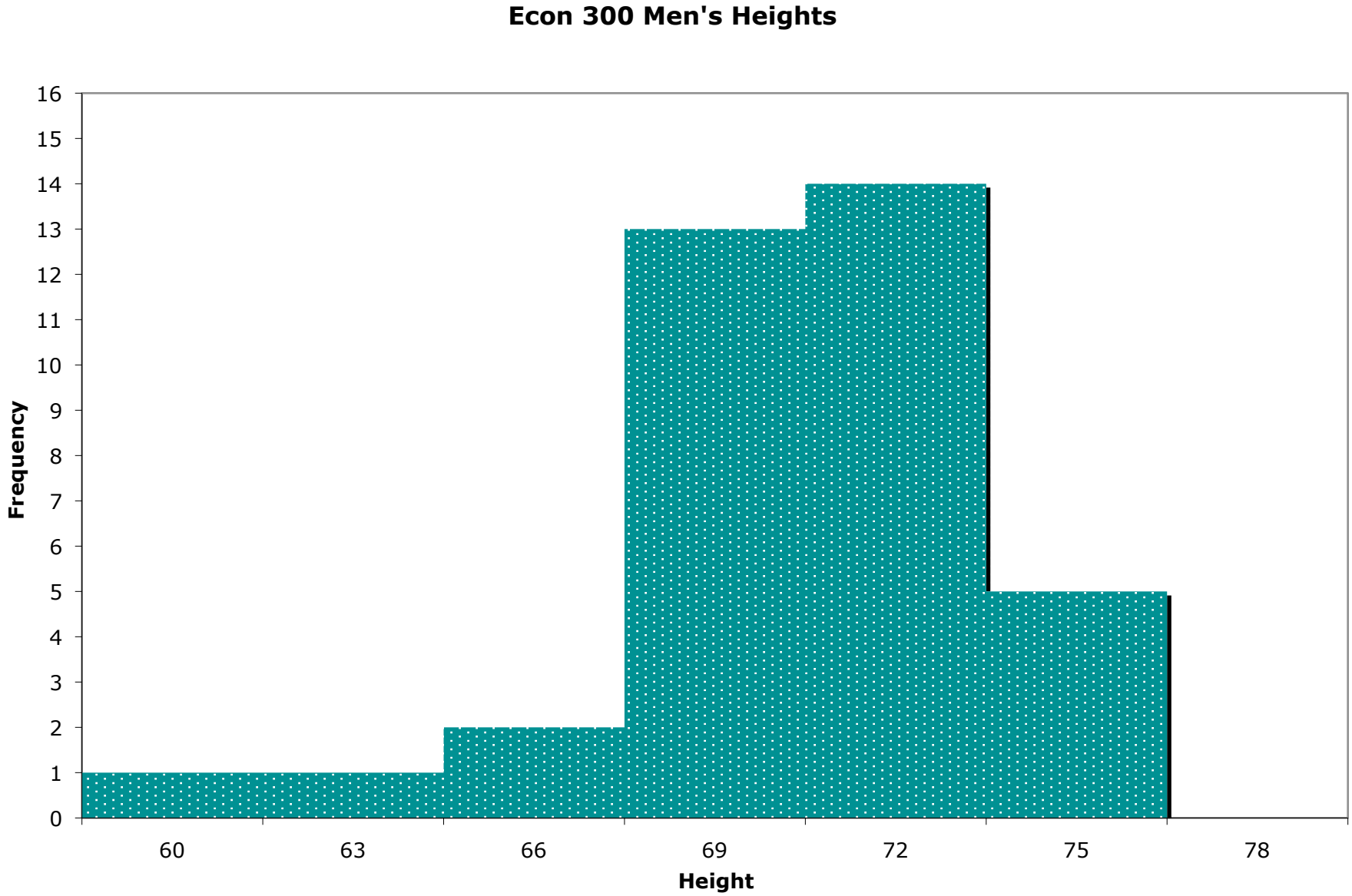
Econ 300/QAC201 Class Height Statistics 2009-10

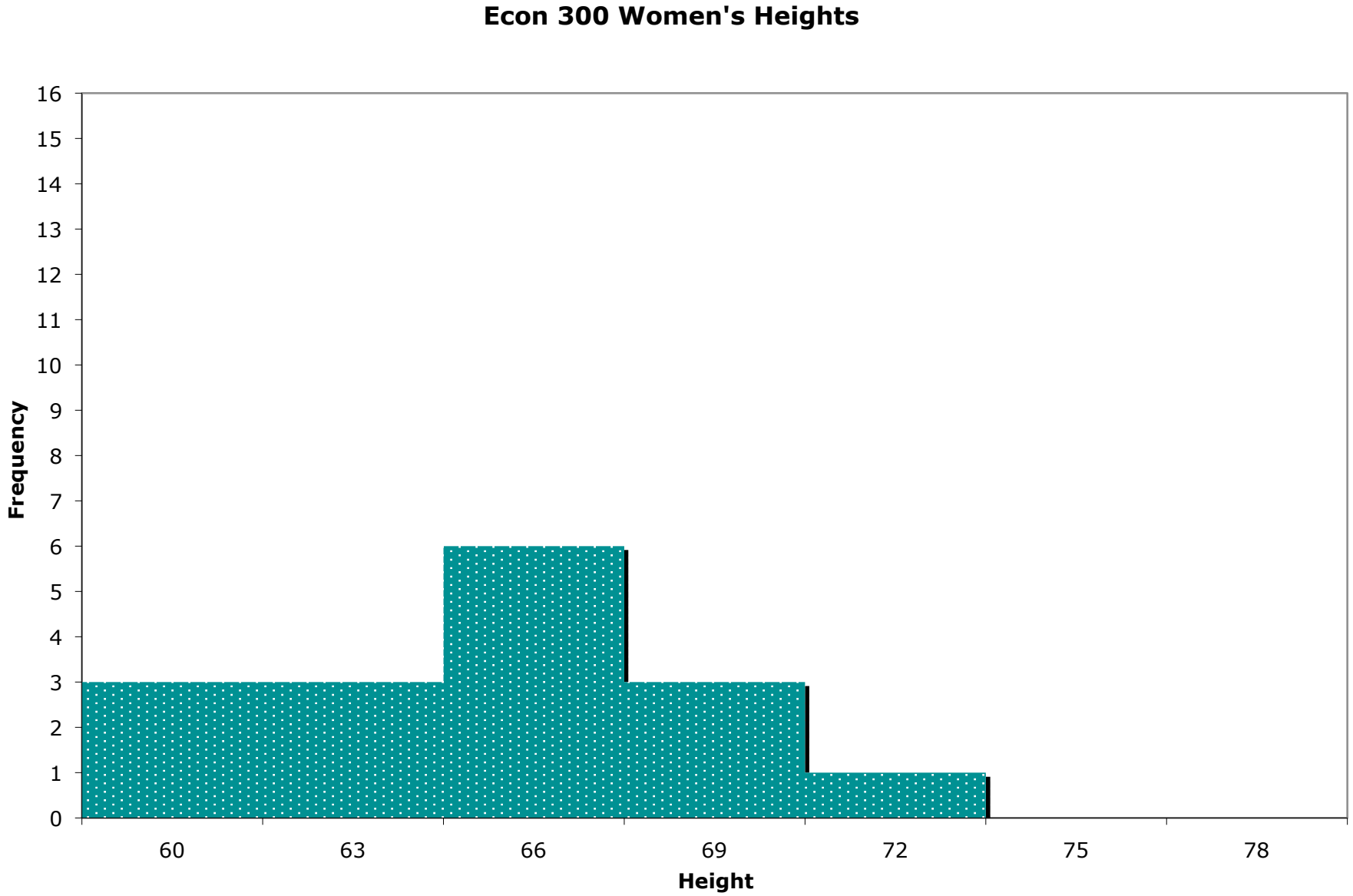
Rank	Sex	Height
1	M	76
2	M	75
3	M	75
4	M	74
5	M	74
6	M	73
7	M	73
8	M	72
9	M	72
10	M	72
11	M	72
12	M	72
13	M	72
14	F	72
15	M	71
16	M	71
17	M	71
18	M	71
19	M	70.5
20	M	70.5
21	M	70
22	M	70
23	M	70
24	F	70
25	M	69
26	M	69
27	M	69
28	M	69
29	F	69
30	M	68
31	M	68
32	M	68
33	M	68
34	M	68
35	M	68
36	F	68
37	M	67
38	M	66
39	F	66
40	F	66
41	F	66
42	F	65
43	F	65
44	F	64.5
45	M	64
46	F	64
47	F	63.5
48	F	63
49	M	61
50	F	61
51	F	60
52	F	59

52	class size
36	number of men (69%)
16	number of women (31%)
76-59=17	class range
69	class median
72	class mode (7)
72-66=6	class inter-quartile range
69	class mean
4	class standard deviation
76-61=15	men's range
72-59=13	women's range
70.5	men's median
65	women's median
68, 72	men's modes (6)
66	women's mode (3)
72-68=4	men's inter-quartile range
68.5-61.5=7	women's inter-quartile range
70	men's mean
65	women's mean
3	men's standard deviation
3.5	women's standard deviation

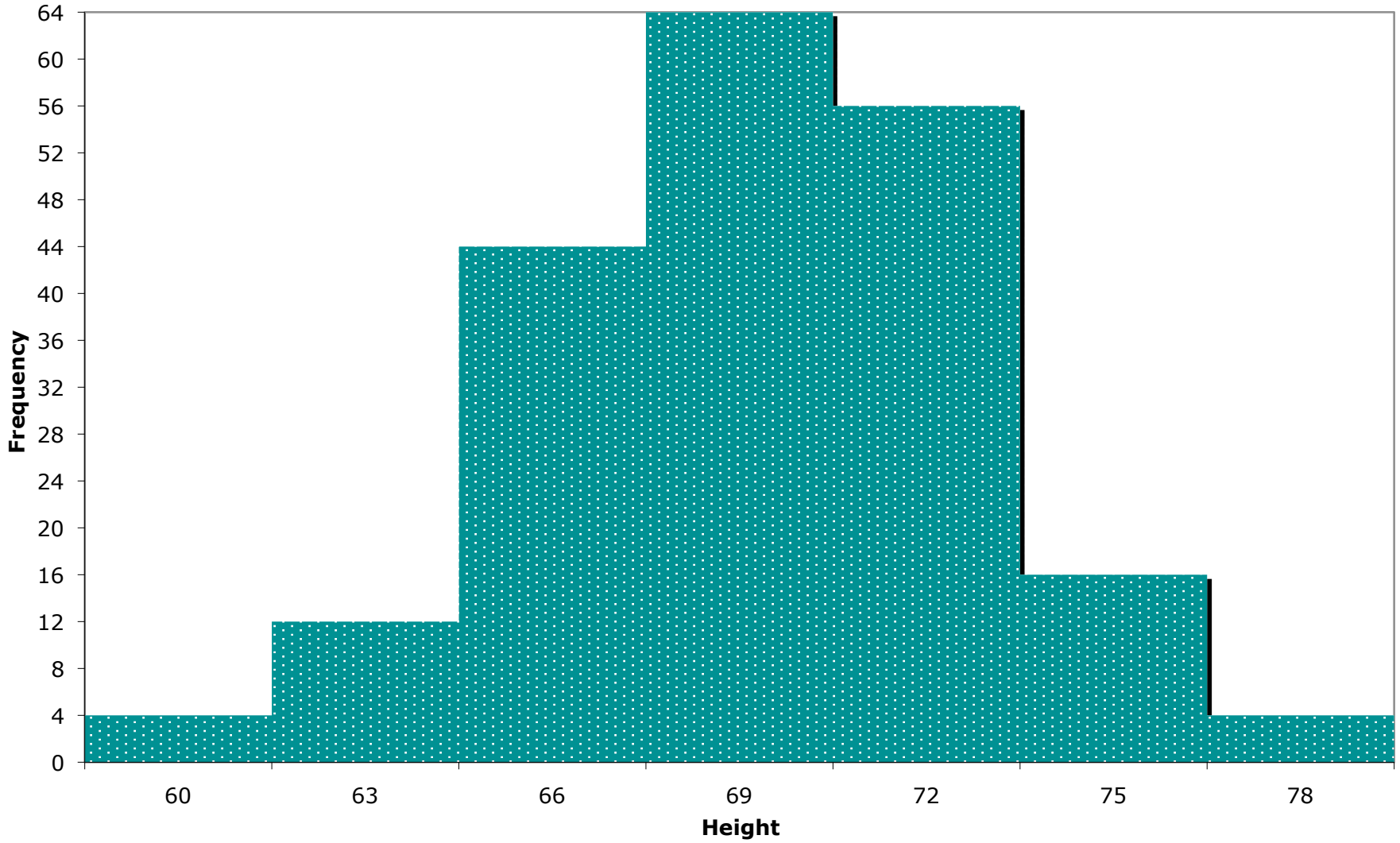
Values	Class F	Men F	Women F	book
60	4	1	3	4
63	4	1	3	12
66	8	2	6	44
69	16	13	3	64
72	15	14	1	56
75	5	5	0	16
78	0	0	0	4


Econ 300 Class Heights
Frequency
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
60
63
66
69
72
75
78
Height (inches)


Econ 300 Men's Heights
Frequency
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
60
63
66
69
72
75
78
Height


Econ 300 Women's Heights
Frequency
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
60
63
66
69
72
75
78
Height

Book p. 29


Frequency
64
60
56
52
48
44
40
36
32
28
24
20
16
12
8
4
0
60
63
66
69
72
75
78
Height

Elementary Rules of Summation Useful for Statistics and Econometrics

1. $\sum_{i=1}^{n} X_i = X_1 + X_2 + \ldots + X_{n-1} + X_n$ (sum from 1 to n, where n is the total number of observations)

2. $\sum_{i=1}^{n} k = n*k$ (where k is any number)

3. $\sum_{i=1}^{n} k*X_i = k* \sum_{i=1}^{n} X_i$ (where k is any number)

4. $\sum_{i=1}^{n} (X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$

5. $\sum_{i=1}^{n} (X_i - Y_i)^2 = \sum_{i=1}^{n} X_i^2 - 2* \sum_{i=1}^{n} X_i*Y_i + \sum_{i=1}^{n} Y_i^2$

One common mistake that is made is to assume that:

$\sum_{i=1}^{n} X_i*Y_i = \sum_{i=1}^{n} X_i * \sum_{i=1}^{n} Y_i$ (this is <u>NOT TRUE!</u>)

If the subscripts and superscripts are left off of the $\sum$ sign, then the summation is implicitly assumed to go over the range 1 to n.

Often multiplication signs are left off as well, so $\sum_{i=1}^{n} k*X_i = \sum_{i=1}^{n} kX_i = k \sum_{i=1}^{n} X_i$, and

$\sum_{i=1}^{n} X_i*Y_i = \sum_{i=1}^{n} X_iY_i$