NYU STERN
NEW YORK UNIVERSITY LEONARD N. STERN SCHOOL OF BUSINESS

IOMS Department

Statistics and Data Analysis

Professor William Greene
Phone: 212.998.0876
Office: KMC 7-90
Home page: http://people.stern.nyu.edu/wgreene
Email: wgreene@stern.nyu.edu
Course web page: http://people.stern.nyu.edu/wgreene/Statistics/Outline.htm

Assignment 1

Notes:
(1) The data sets for this problem set (and for the other problem sets for this course) are all stored on the home page for this course. You can find links to all of them on the course outline, at the bottom with the links to the problem sets themselves.

(2) In the exercises below (and in the other problem sets), the initials HOG refer to the textbook *Basic Statistical Ideas for Managers*, by Hildebrand, Ott and Gray.

Part I. Describing Data

1. Consider the following values:

```
     20 11 14 12 17 14 10 23 15 11 17 10 18 18 13 18
```

   Find the mean, median, and mode for these data.

2. This is Exercise HOG 2.1, page 23. Thedata are available on the course website as **HOG-Ex0201.mpj**.An automobile manufacturer routinely keeps records on the number of finished (passing all inspections) cars produced per eight-hour shift. The data for the last 28 shifts are

```
366 390 324 385 380 375 384 383 375 339 360 386 387 384 379 386 374
366 377 385 381 359 363 371 379 385 367 364
```

   a. What is the average number of finished cars per shift based on these data?
   b. Construct a histogram for these data.
   c. The data above are the results from observing 28 eight hour shifts. You are about to observe a 29th. What would be a good guess of how many will be observed? Explain.
   d. Suppose the 29th shift was expected to be a very productive one – with large output. What would be a good guess of the number of finished cars on a very good day? Explain.

3. Which of the two samples in each set has the higher standard deviation. You can tell by looking at the data. It is not necessary to do any computation to answer this question. Explain your reasoning for each answer.

Set 1 Sample A: 16, 16, 16, 16, 16
Sample B: 15, 16, 16, 16, 16
Set 2 Sample A: 20, 25, 25, 25, 30
Sample B: 15, 25, 25, 25, 35
Set 3 Sample A: 20, 20, 30, 40, 40
Sample B: 20, 25, 30, 35, 40

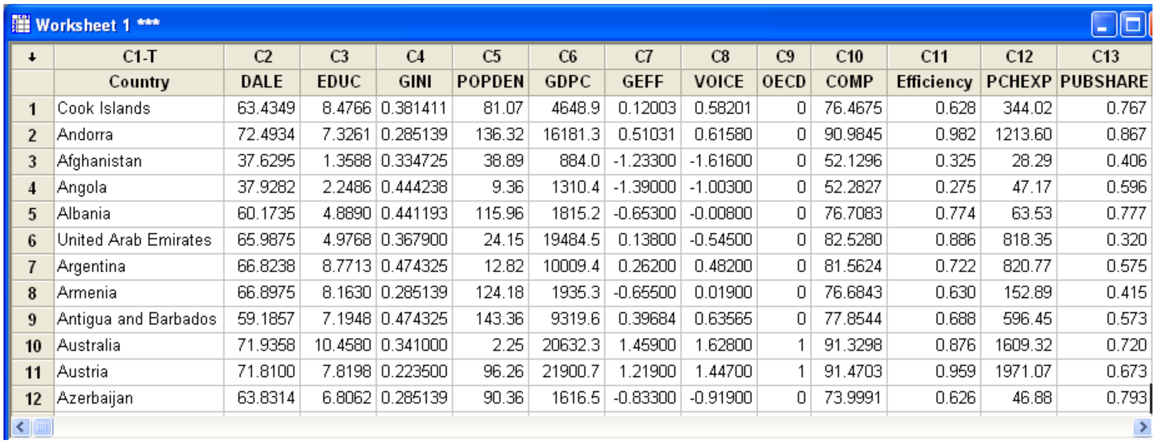
4. This is exercise HOG, problem 2.23, page 42. The data file is **HOG-Ex0222.mpj** on the course outline. Data on 60 telephone operators in terms of number of call requests processed in a workdaywere analyzed using Minitab.

```
Descriptive Statistics: Cleared
Variable     N     Mean        SE Mean     StDev
Cleared      60    794.23      4.42        34.25
Minimum      Q1          Median      Q3          Maximum
601.00       789.00      799.00      807.75      844.00
Data Display
Cleared
797 794 817 813 817 793 762 719 804 811 837 804 790 796 807 801
805 811 835 787 800 771 794 805 797 724 820 601 817 801 798 797
788 802 792 779 803 807 789 787 794 792 786 808 808 844 790 763
784 739 805 817 804 807 800 785 796 789 842 829
```

a. Calculate the "mean plus-or-minus 1 standard deviation" interval used in the Empirical Rule discussed in class.

b. Of the 60 scores in "cleared," 51 fall within the 1 standard deviation interval.How does this result compare with the theoretical value of the Empirical Rule?

5. (Application) The data file **WHO-HealthStudy.mpj** (a Minitab project file) contains a famous data set. These data were used in the World Health Organization's 2000 comparison of the health care systems in 191 countries – nearly the entire world – that was widely discussed in the popular press (including on the front page of the New York Times) If you've seen Michael Moore's movie *Sicko*, or seen the trailer, there is a point at which he takes out a study on a clipboard and shows you how the United States ranked 37th in the world in "health care." These are part of the data that were used to do the study. The extract of the data file in the Minitab project contains 12data columns, as shown below for the first few countries



Worksheet 1 ***

↓	C1-T	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
	Country	DALE	EDUC	GINI	POPDEN	GDPC	GEFF	VOICE	OECD	COMP	Efficiency	PCHEXP	PUBSHARE
1	Cook Islands	63.4349	8.4766	0.381411	81.07	4648.9	0.12003	0.58201	0	76.4675	0.628	344.02	0.767
2	Andorra	72.4934	7.3261	0.285139	136.32	16181.3	0.51031	0.61580	0	90.9845	0.982	1213.60	0.867
3	Afghanistan	37.6295	1.3588	0.334725	38.89	884.0	-1.23300	-1.61600	0	52.1296	0.325	28.29	0.406
4	Angola	37.9282	2.2486	0.444238	9.36	1310.4	-1.39000	-1.00300	0	52.2827	0.275	47.17	0.596
5	Albania	60.1735	4.8890	0.441193	115.96	1815.2	-0.65300	-0.00800	0	76.7083	0.774	63.53	0.777
6	United Arab Emirates	65.9875	4.9768	0.367900	24.15	19484.5	0.13800	-0.54500	0	82.5280	0.886	818.35	0.320
7	Argentina	66.8238	8.7713	0.474325	12.82	10009.4	0.26200	0.48200	0	81.5624	0.722	820.77	0.575
8	Armenia	66.8975	8.1630	0.285139	124.18	1935.3	-0.65500	0.01900	0	76.6843	0.630	152.89	0.415
9	Antigua and Barbados	59.1857	7.1948	0.474325	143.36	9319.6	0.39684	0.63565	0	77.8544	0.688	596.45	0.573
10	Australia	71.9358	10.4580	0.341000	2.25	20632.3	1.45900	1.62800	1	91.3298	0.876	1609.32	0.720
11	Austria	71.8100	7.8198	0.223500	96.26	21900.7	1.21900	1.44700	1	91.4703	0.959	1971.07	0.673
12	Azerbaijan	63.8314	6.8062	0.285139	90.36	1616.5	-0.83300	-0.91900	0	73.9991	0.626	46.88	0.793

Assignment 1

The variables in the file are

DALE = disability adjusted life expectancy
EDUC = average years of education
GINI = a measure of income inequality (low numbers are bad)
POPDEN = the population density, people per square kilometer
GDPC = per capita gross domestic product (country income)
GEFF = World Bank measure of the effectiveness of the government
VOICE = World Bank indicator of how democratic the country is
OECD = an indicator of whether the country is in the OECD. (OECD is the United Nations Organization for Economic Cooperation and Development. Notwithstanding its lofty title, it is mainly a group of the world’s wealthiest countries.)
COMP = an equally weighted average of survey results on five objectives (Health, Health distribution, Responsiveness, Responsiveness in Distribution, Fairness in financing).
EFFICIENCY = estimated overall efficiency (WHO/Paper 30, based on COMP)
PCHEXP = per capita health care expenditure (public)
PUBSHARE = proportion of total health expenditure paid for by the government

a. Let’s compare the incomes of the 30 OECD countries with the incomes of the 161 other countries. A box-plot will be useful. Use Graph -> Boxplot (with groups), then graph variable GDPC with group variable OECD. Note that OECD = 1 is the OECD and OECD = 0 is the other countries. What do you find?

b. Present a description of the variables DALE and GDPC using the tools discussed in class. (Descriptive statistics will include means and standard deviations and medians. Graphical tools include histograms and box plots.)

c. Does higher income buy higher life expectancy? Produce a scatter plot of DALE (on the Y axis) against GDPC (on the X axis). What do you find?

d. Does education produce higher income? Produce a scatter plot of EDUC (on the X axis against GDPC. What do you find? What conclusion do you draw?

e. Do higher levels of education appear to be associated with higher life expectancy?

Part II. Unconditional and Conditional Probability

6. An icosahedron is a regular geometric figure that has 20 geometrically identical faces.These can be marked with numbersand used like dice. Suppose that you have one of these, and that it is marked with the integers 1, 2, 3, ..., 20. If you roll this object exactly once, find

a. the probability that the top face has an even number;
b. the probability that the top face has a number greater than 7;
c. the probability that the top face has a number at most 4;
d. the probability that the top face has a number less than 4.

7. In a fit of boredom, Stanley decides that he will flip a coin 80 times.

a. What is the probability that the first flip will be heads?
b. What is the probability that the first flip will be tails?
c. What is the probability that the second flip will be heads?
d. What is the probability that the 43rd flip will be tails?
e. What is the probability that the total number of heads, out of 80 flips, will be an even number?

Assignment 1

8. A secretary has left you four index cards involving phone messages. You will have to call back Johnson, Ortega, Green, and Baker. If you shuffle these cards into randomorder,

   a. What is the probability that the names will end up in alphabetical order, starting with Baker?

   b. What is the probability that Johnson will precede Green? This should be interpreted as having Johnson *anywhere* before Green, meaning either one turn ahead or two turns ahead or three turns ahead.

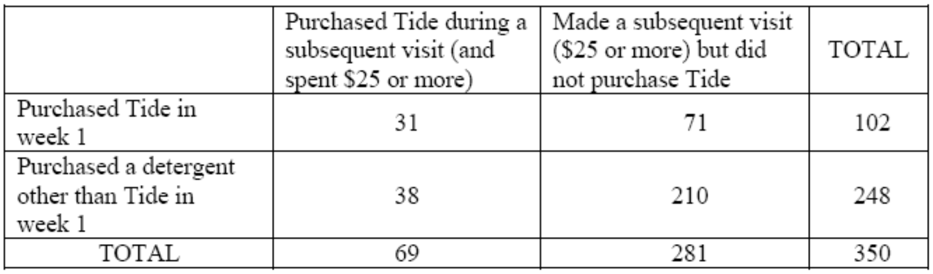
   c. What is the probability that Johnson and Green will be on consecutive cards?

9. Liz Waters, the manager of Food City Supermarket, has all sorts of data on the store's computer. The customers use "frequent shopper" scan tags at each visit, and most of the items are priced by scanning bar codes. Liz is exploring whether coupons can be used to entice consumers to change brands, in particular, will a coupon for Tide detergent cause consumers to purchase Tide. Over the four-week study period, each customer who buys *any* detergent is given a coupon for $1 off the next purchase of Tide. The display belowis limited to those customers who, during the four-week study period

   * purchased detergent during the first week

   * visited the store a second time during the four weeks and made a purchase of at least $25

A summary of the transactions from the study:



	Purchased Tide during a subsequent visit (and spent $25 or more)	Made a subsequent visit ($25 or more) but did not purchase Tide	TOTAL
Purchased Tide in week 1	31	71	102
Purchased a detergent other than Tide in week 1	38	210	248
TOTAL	69	281	350

There were no instances in which a customer bought two different detergent brands on the same visit. Suppose that one of the customers is selected at random. Find

   a. the probability that the customer bought Tide in week 1.

   b. the probability that the customer bought Tide in week 1 and also purchased Tide again during the study period.

   c. the conditional probability that the customer bought Tide at a subsequent visit, given that he or she bought Tide in week 1.

   d. the conditional probability that the customer bought Tide at a subsequent visit, given that he or she bought a non-Tide detergent in week 1.

10. Suppose that, hypothetically, 88% of all people being tried for burglary are in fact guilty of the crime. Suppose that 6% of innocent people are convicted at trial and that74% of guilty people are convicted at trial.

   a. If a person is convicted at trial, what is the probability that the person really is guilty of the crime?

   b. If a person is acquitted at trial, what is the probability that the person really is innocent of the crime?

11. The customer service office of Garsett Bank receives complaints regarding transactions at its two off-site ATMs. We'll identify these sites as *A* and *B*. We know the following:

Site *A* generates 70% of all the ATM activity, and site *B* generates 30%.
The proportion of transactions that lead to a complaint is 0.006.
Among the complaints received by the customer service office, 45% are related to site *B*.

Based on these facts, find the site-specific complaint rates. That is find P(complaint|*A*) and P(complaint | *B*).

Part III. Expected Value, Covariance and Correlation

12. (This is Exercise HOG 4.11, page 148.) An investment syndicate is trying to decide which of two $2,000,000 apartment houses tobuy. An advisor estimates the following probabilities for the two-year net returns (inthousands of dollars):

Return:	-50	0	50	100	150	200	250
Probability for house 1:	.02	.03	.20	.50	.20	.03	.02
Probability for house 2:	.15	.10	.10	.10	.30	.20	.05

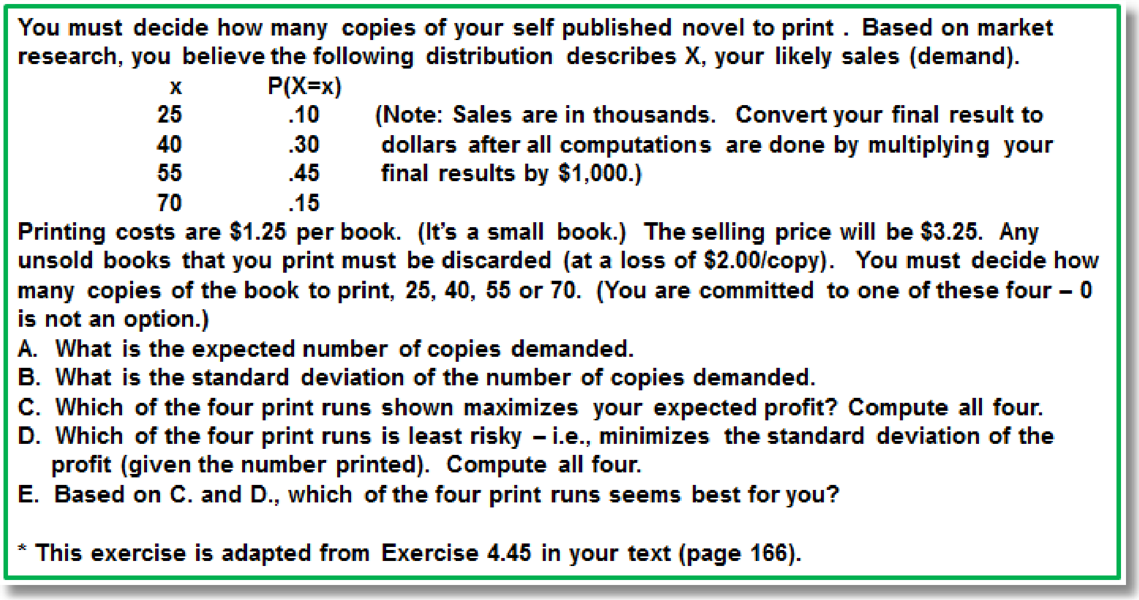
a. Calculate the expected net return for house 1 and for house 2.
b. Calculate the respective variances and standard deviations.

13. The following (completely fictitious) table shows the probabilities of music CD sales per minute of Tower Records in a given month (back when there was a Tower Records, and back when people actually bought music CDs) associated with the random distribution of the number of concerts scheduled in that month. It appears that the two random variables may be correlated. The following investigates. (Note that the zeros at certain points in the table are by themselves suggestive.) We note that the values in the table are retrospective – we have simply tabulated observed results over a long history. This would be in contrast to the case in which we examined specific months in which there were, say, 0 concerts scheduled, or 1, in which the values in the table would be estimates of conditional probabilities, not joint probabilities.

a. What is the expected number of CD sales per minute?
b. What is the variance of the number of concerts per month?
c. Before doing the calculation, what sign do you expect for the covariance of the two variables?
d. What is the covariance of the CD sales and number of concerts?
e. What is the correlation of the two variables.
f. What is the expected number of CD sales per minute in a month in which there are two major concerts scheduled. Same for three major concerts. Are the two values the same?

CDs Sold Per Minute							
		0	1	2	3	4	Total
Concerts per Month	0	.02	.05	.04	.01	.00	.12
	1	.02	.04	.06	.03	.00	.15
	2	.01	.03	.30	.04	.02	.40
	3	.00	.03	.10	.12	.08	.33
	Total	.05	.15	.50	.20	.10	1.00

Part IV. Conditional Expected Value and Variance



You must decide how many copies of your self published novel to print . Based on market research, you believe the following distribution describes X, your likely sales (demand).

x	P(X=x)
25	.10
40	.30
55	.45
70	.15

(Note: Sales are in thousands. Convert your final result to dollars after all computations are done by multiplying your final results by $1,000.)

Printing costs are $1.25 per book. (It's a small book.) The selling price will be $3.25. Any unsold books that you print must be discarded (at a loss of $2.00/copy). You must decide how many copies of the book to print, 25, 40, 55 or 70. (You are committed to one of these four – 0 is not an option.)

A. What is the expected number of copies demanded.

B. What is the standard deviation of the number of copies demanded.

C. Which of the four print runs shown maximizes your expected profit? Compute all four.

D. Which of the four print runs is least risky – i.e., minimizes the standard deviation of the profit (given the number printed). Compute all four.

E. Based on C. and D., which of the four print runs seems best for you?

* This exercise is adapted from Exercise 4.45 in your text (page 166).