COLUMBIA
BUSINESS
SCHOOL

Exercises

Fall 2001
B6014: Managerial Statistics

Professor Paul Glasserman
403 Uris Hall

Exercises

Descriptive Statistics

Fall 2001 — Professor Paul Glasserman
B6014: Managerial Statistics — 403 Uris Hall

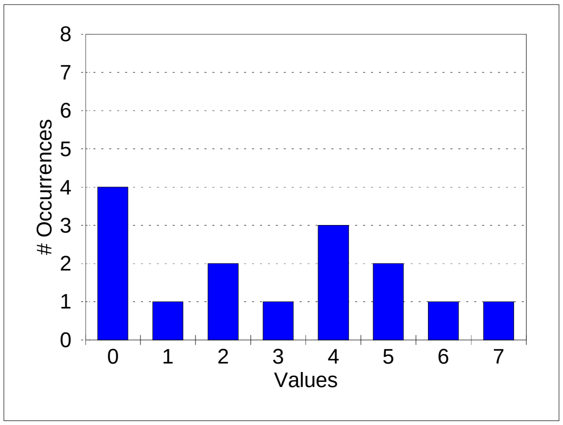


Figure 1: Histogram for Problem 1

1. Find the median of the data in Figure 1.

2. Find the standard deviation of the data in Figure 1.

3. Five students from the 1999 MBA class took jobs in rocket science after graduation. Four of these students reported their starting salaries: \$95,000, \$106,000, \$106,000, \$118,000. The fifth student did not report a starting salary. Choose one of the following:

   (a) The median starting salary for all five students could be anywhere between \$95,000 and \$118,000.

   (b) The median starting salary for all five students is \$106,000.

   (c) The median starting salary for all five students is \$106,500.

   (d) The median starting salary for all five students could be greater than \$118,000.

4. The observations $X_1, \ldots, X_n$ have a mean of 52, a median of 52.1, and a standard deviation of 7. Eight percent of the observation are greater than 66; 7.9% of the observations are below 38. Based on this information, which of the following statements *best* describes the data?

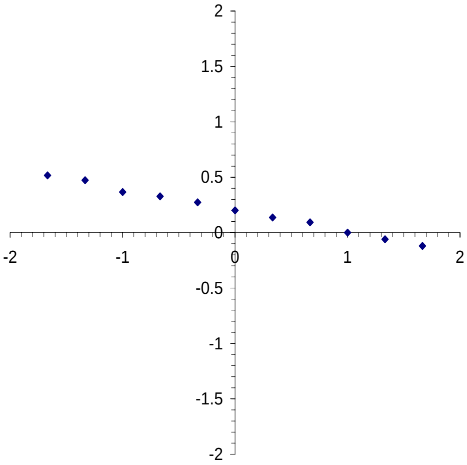


Figure 2: Scatter plot for question 5

(i) The distribution has positive skew.

(ii) The distribution has negative skew.

(iii) The distribution has high kurtosis.

(iv) The distribution conforms to a normal distribution.

5. Consider the data in the scatter plot of Figure 2. The correlation between the X and Y values in the figure is closest to

   (i) 0.2

   (ii) -0.2

   (iii) 1

   (iv) -1

   (v) 0

6. The observations $X_1, \ldots, X_n$ have a mean of 50 and a standard deviation of 7. Which of the following statements is guaranteed to be true according to Chebyshev's rule? (Write "True" or "False" next to each.)

   (i) At least 75% of the observations are between 36 and 64 ____________________

   (ii) At least 80% of the observations are between 34 and 66 ____________________

   (iii) At least 88.9% of the observations are between 31 and 73 ____________________

   (iv) Fewer than 15% of the observations are below 30 ____________________

7. Suppose the observations $X_1, X_2, \ldots, X_n$ have mean 10. Suppose that exactly 75% of the observations are less than or equal to 15. According to Chebyshev's rule, what is the smallest possible value of the population standard deviation of these observations?

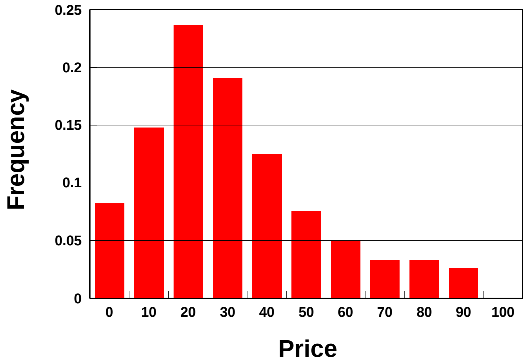


Figure 3: Histogram of bond prices at default, 1974-1995. (Source: Moody's Investor Services.)

8. Which of the following best describes the data in Figure 3? (Base your answer on the appearance of the histogram. You do not need to do any calculations. Select just one statement below and complete the one you select.)

   (a) The mean is greater than the median because ______________________________

   (b) The median is greater than the mean because ______________________________

   (c) The mean and median are roughly equal because ______________________________

9. One proposal that has received little attention from Major League Baseball is to pay pitchers according to the following rule: each pitcher receives a base salary of \$4.25 million, *minus* \$0.25 million times his earned run average (ERA). (A lower ERA is associated with better performance.) If this rule were adopted, what would be the correlation between a pitcher's earnings and ERA? (Assume that the ERA cannot exceed 17, so this rule never results in negative earnings. You may also assume a standard deviation of 1.2 for ERA.)

10. Using the data in Figure 4, answer both (a) and (b) below, providing a numerical value for each.

    (a) The mean of the data in the histogram is ______________________________

    (b) The median of the data in the histogram is ______________________________

11. Cluster Ψ had exams in Finance and Marketing last week. All 60 students in the cluster took both exams. The results were as follows:

    ○ Finance: mean = 25, standard deviation = 2

    ○ Marketing: mean = 75, standard deviation = 12

    ○ Correlation between score in Finance and same student's score in Marketing = 0.84

    Mary, a student in Cluster Ψ, scored a 30 in Finance and a 90 in Marketing. We are interested in comparing her performance on the two exams relative to the rest of the class. In particular, we would like to make a statement about which of her scores ranked higher compared to the other scores on the same exam. Select *one* of the choices below and complete the statement you select.

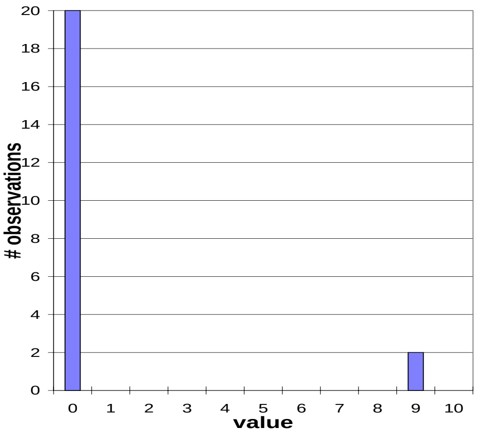


Figure 4: Histogram for Problem 10

(i) Mary's score in Finance probably ranks higher than her score in Marketing because

____

(ii) Mary's score in Marketing probably ranks higher than her score in Finance because

____

(iii) Mary's scores on the two exams probably rank about equally high because

____

(iv) We cannot make any comparison between the two scores because

____

12. Seven students from the 1998 MBA class took jobs in brain surgery after graduation. Five of the students reported their starting salaries: \$55,000, \$90,250, \$90,250, \$95,500, and \$105,000. Choose one of the following:

(a) Based on the information given, the largest possible value of the median starting salary for all seven students is ____________________

(b) Based on the information given, it is not possible to put an upper limit on the median starting salary for all seven students.

Solutions: Descriptive Statistics

1. There are 15 data points in the histogram. Seven are smaller than 3 and seven are greater than 3, so the median is 3.

2. We show four different ways of calculating the standard deviation.

   Method 1. List the full set of observations in a spreadsheet, repeating values as many times as they occur: 0, 0, 0, 0, 1, 2, 2, 3, 4, 4, 4, 5, 5, 6, 7. Apply the function STDEVP to the observations. The result is 2.28. (The function STDEV gives a slightly different answer because it calculates the sample standard deviation rather than the population standard deviation.)

   Method 2. List the full set of observations in a spreadsheet, as in the column labeled "X_i" in Table 1. Calculate the average of these values; this gives 2.867. Now make a column of the *squared* observations, X_i^2, as in the second column of the table. The average of the squared observations is 13.4. The variance is the difference between the average of the squared observations and the square of the average:

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = 13.4 - (2.867)^2 = 5.182.$$

   The standard deviation is the square root, $\sqrt{5.182} = 2.28$.

	X_i	X_i^2	$(X_i - \bar{X})^2$
	0	0	8.218
	0	0	8.218
	0	0	8.218
	0	0	8.218
	1	1	3.484
	2	4	0.751
	2	4	0.751
	3	9	0.018
	4	16	1.284
	4	16	1.284
	4	16	1.284
	5	25	4.551
	5	25	4.551
	6	36	9.818
	7	49	17.084
Mean	2.867	13.4	5.182

Table 1: Solution to Problem 2

Method 3. As in Method 2, list the full set of observations and calculate their mean to get $\bar{X} = 2.867$. For each observation X_i, calculate $(X_i - \bar{X})^2$, the squared distance from the mean; these values are in the third column of Table 1. The average of these squared differences gives the variance

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = 5.182.$$

The standard deviation is the square root, $\sqrt{5.182} = 2.28$.

Method 4. List the *distinct* values observed without repetition, as in the second column of Table 2. For each value, calculate what proportion of the observations had that value, as in the first column of the table. Calculate the weighted average of the values, using the proportions as weights; the result is 2.867. Now calculate the weighted average of the squared values to get 13.4. The difference $13.4 - (2.867)^2 = 5.182$ is the variance. The square root $\sqrt{5.182} = 2.28$ is the standard deviation.

Proportion	Value	Squared Value
4/15	0	0
1/15	1	1
2/15	2	4
1/15	3	9
3/15	4	16
2/15	5	25
1/15	6	36
1/15	7	49
Weighted Avg.	2.867	13.4

Table 2: Solution to Problem 2, Method 4

3. The answer is (b). No matter what the missing salary figure is, 106,000 will be the median. For example, if the missing salary were 50,000, the ordered list would be

$$50,000, \quad 95,000, \quad 106,000, \quad 106,000, \quad 118,000$$

and if the missing salary were 150,000 it would be

$$95,000, \quad 106,000, \quad 106,000, \quad 118,000, \quad 150,000.$$

Either way, the median would be 160,000.

4. The answer is (iii). The data described in the question has the following characteristics: the median is slightly larger than the mean; the right tail has slightly more data than the left tail; *both* tails contain more than three times as many observations as would be predicted by the normal distribution. (Recall that approximately 95% of the values lie within $\pm 2\sigma$ of the mean in the normal case; hence, about 2.5% should be above and 2.5% below.) The most pronounced feature is therefore the last one (heavy tails) which is suggestive of high kurtosis.

5. The answer is (iv), because the points lie almost exactly on a line with negative slope.

6. The answers are T, T, F, T. (i) should be clear because the range given is exactly ±2 standard deviations. (ii) First we find the m that goes with 80%:

$$.80 = 1 - \frac{1}{m^2} \Rightarrow m = 2.236.$$

Thus, Chebyshev tells us that at least 80% are between 34.35 and 65.65. Clearly, at least as many observations must lie between 34 and 66. (iii) To guarantee 88.9% we need a range of ±3 standard deviations, which would be from 29 to 71. The range given only goes as low as 31. (iv) First find how many standard deviations below the mean 30 is:

$$(50 - 30)/7 = 2.86.$$

Chebyshev guarantees that at least

$$1 - \frac{1}{(2.86)^2} = 87.8\%$$

of the observations are within 2.86 standard deviations of the mean. But then at most 12.2% can be below 30.

7. Exactly 25% are greater than 15, so 15 can be at most 2 standard deviations above the mean of 10; i.e., the standard deviation cannot be less than 2.5.

8. (a) because the distribution is skew right.

9. The proposed rule makes Salary $= -0.25 \cdot \text{ERA} + 4.25$, a linear transformation with negative slope. The correlation is therefore -1.

10. (a) The mean is $[(20 \times 0) + (2 \times 9)]/22 = 18/22 = 0.82$. (b) The median is 0. (List the observations from smallest to largest. Since we have an even number of observations, take the two closest to the middle — the 11th and the 12th — and average them. Both are 0, so the median is 0.)

11. The answer is (i) because her score in Marketing is 2.5 standard deviations above the mean whereas her score in Finance is only 1.25 standard deviations above the mean.

12. Even if the two unknown salaries were very large, the median could not be larger than the fourth smallest value, $95,500. For example, if the two unknown salaries were $500,000, the values become

    55,000 90,250, 90,250, 95,500, 105,000, 500,000, 500,000,

    with 95,500 in the middle.

Exercises

Probability and Expected Value

Fall 2001
B6014: Managerial Statistics

Professor Paul Glasserman
403 Uris Hall

1. An *inverse floater* is a type of security whose payments move in the opposite direction of short-term interest rates. The security is ordinarily structured so that no matter how high interest rates rise, the payment cannot be negative, and no matter how low interest rates drop, the payment cannot exceed some specified cap, e.g., 7%. Consider the following specific case: if the prevailing short-term rate is X, the inverse floater pays

$$Y = 100 \times \max(0.07 - X, 0)$$

on $100 of face value. (The notation "$\max(0.07 - X, 0)$" means use whichever is larger, $0.07 - X$ or 0.) Suppose the distribution of the short-term rate X at the next payment date is given by the following table:

x	0.04	0.06	0.07	0.08	0.09
$P(X = x)$	0.30	0.20	0.20	0.15	0.15

Find the expected payment $E[Y]$.

2. A mining company plans to develop two potential gaussite reserves. Each reserve has probability 0.30 of successfully yielding usable gaussite, and the success of each reserve is independent of the other. If either of the two reserves is successful, it will generate $4 million in profit; if both are successful, profits will be $7 million because excess supply will lower prices. If neither is successful, profits will be 0. Let X be the company's profit. Find $E[X]$.

3. The Gourmet Cafe serves the exotic Bernoulli Salmon at lunch and dinner. The number of customers ordering the salmon at lunch and dinner are given by the following distributions:

Lunch demand	0	1	2
probability	0.3	0.5	0.2

Dinner demand	0	1	2
probability	0.2	0.4	0.4

Assume the lunch and dinner demands are independent of each other so the joint distribution of the lunch and dinner demands is given by the following table:

		Lunch			
		0	1	2	
	0	0.06	0.10	0.04	0.2
Dinner	1	0.12	0.20	0.08	0.4
	2	0.12	0.20	0.08	0.4
		0.3	0.5	0.2	

(Each entry of the table is just the product of the marginal probabilities at the end of the corresponding row and the bottom of the corresponding column.)

The chef orders the fish in advance at a cost of $3.50 per serving. Any fish left over at the end of the day is discarded.

(a) What is the expected total demand for the fish in a day?

(b) Suppose the chef orders three servings. What is the breakeven selling price (i.e., the price at which the expected revenue from sales of the fish equals the cost of the fish ordered)? Assume that a customer who would have ordered the fish but finds it sold out simply leaves rather than order something else. (Hint: Expected revenue = price times expected number of units sold.)

4. The Uris & Warren Ratings Agency rates bonds on a simplified scale with just three categories: A, B, and C. The Professors Pension Fund has all its money invested in two bonds, X and Y, both of which are currently rated B. Over the course of the next year the ratings of the bonds may change; the end-of-year value (in millions) of each bond depends on its end-of-year rating as in the following table:

Rating	X Value	Y Value
A	100	100
B	75	75
C	50	50

The joint distribution of the end-of-year ratings of the two bonds is given by the following table:

		Y		
		A	B	C
X	A	0.20	0.05	0
	B	0	0.40	0.10
	C	0	0.05	0.20

(a) Find the probability that bond X will be rated C at the end of one year.

(b) Find the expected value of the year-end value of bond X.

(c) The Pension Fund buys an insurance contract that will pay 20 million if either bond is downgraded to C. If both bonds are downgraded the contract still pays 20; if neither is downgraded the contract pays nothing. Find the expected value of the payoff of the insurance contract.

Solutions: Probability and Expected Value

1. Using the rule $Y = 100 \times \max(0.07 - X, 0)$, list the payments in each scenario as follows:

x	0.04	0.06	0.07	0.08	0.09
$P(X = x)$	0.30	0.20	0.20	0.15	0.15
y	3	1	0	0	0

Now we have $E[Y] = (.30)(3) + (.20)(1) = 1.1$.

2. Consider the possible outcomes and their probabilities:

outcome	probability	payoff
only 1st successful	0.30(1-0.30)	4
only 2nd successful	(1-0.30)0.30	4
both successful	(0.30)(0.30)	7
neither successful	(1-0.30)(1-0.30)	0

Thus,

$$E[X] = (2 \times 0.3 \times 0.7 \times 4) + (0.30 \times 0.30 \times 7) = 2.31$$

3. (a) From the joint distribution of lunch and dinner demands we find that the distribution of total demand is as follows:

Demand	0	1	2	3	4
Prob	0.06	0.22	0.36	0.28	0.08

Now we have that the expected value is

$$0(0.06) + 1(0.22) + 2(0.36) + 3(0.28) + 4(0.08) = 2.10.$$

We could also get this answer by adding the expected lunch demand (0.9) and the expected dinner demand (1.2).

(b) You can't sell fish you don't have: if four people order it, you only sell three. So, the distribution of units sold is this:

Demand	0	1	2	3	4
Units sold	0	1	2	3	3
Prob	0.06	0.22	0.36	0.28	0.08

The expected number sold is

$$0(0.06) + 1(0.22) + 2(0.36) + 3(0.28) + 3(0.08) = 2.02.$$

To break even, the price p needs to be such that $2.02p$ equals the cost of the fish, namely $3 \times 3.50 = 10.50$. The price is therefore $p = 10.50/2.02 = 5.198$ or 5.20.

4. (a) The question asks for the marginal probability that X will be rated C. By summing the third row of the table, we find that this is 0.25.

(b) To find $E[X]$ we need to find the possible values and their probabilities:

Scenario	Value	Prob.
A	100	0.25
B	75	0.50
C	50	0.25

The probabilities are obtained by summing across the rows of the table; this is the marginal distribution of X. We now find

$$E[X] = 100 \cdot 0.25 + 75 \cdot 0.50 + 50 \cdot 0.25 = 75.$$

(c) The insurance contract pays if either bond is downgraded; thus, the scenarios in which the contracts pays are the ones in boxes in the following table:

		Y		
		A	B	C
X	A	0.20	0.05	$\boxed{0}$
	B	0	0.40	$\boxed{0.10}$
	C	$\boxed{0}$	$\boxed{0.05}$	$\boxed{0.20}$

The probability that the contract pays is

$$0 + 0.05 + 0.20 + 0.10 + 0 = 0.35.$$

So, the expected payoff is $0.35 \cdot 20 + (1 - 0.35) \cdot 0 = 7$.

Exercises

Covariance and Correlation

Fall 2001
B6014: Managerial Statistics

Professor Paul Glasserman
403 Uris Hall

1. You manage a retail operation from which you sell to both walk-in and telephone customers. For a particular product, your goal is to set the inventory so that 99% of customers looking or calling for the product find it in stock. Consider the following two scenarios: (i) Days with a lot of walk-in customers are also days with a lot of telephone orders. (ii) Days with a lot of walk-in customers tend to be days with fewer telephone orders and vice-versa. In which scenario would you expect to have to hold more total inventory to meet your service objective? Explain your answer by making reference to the concepts of standard deviation and correlation.

2. You invest \$3 thousand in one stock and your spouse invests \$2 thousand in another. Over the next year, each dollar invested in your pick will increase by X dollars and each dollar invested in your spouse's will increase by Y dollars; X and Y are random variables with the following properties:

   ○ X has a mean of 0.09 and a standard deviation of 0.20.

   ○ Y has a mean of 0.12 and a standard deviation of 0.27.

   ○ The correlation between X and Y is 0.6.

   Your individual earnings are $3X$ thousand, your spouse's individual earnings are $2Y$ thousand and your family earnings are the sum of two.

   (a) What is the expected value of your family earnings?

   (b) What is the standard deviation of your family portfolio earnings?

3. Let X, Y, and Z be random variables. Consider the following statements:

   (i) if $Cov[X,Y] > Cov[Y,Z]$ then $\rho_{XY} > \rho_{YZ}$.

   (ii) if $\rho_{XY} > \rho_{YZ}$ then $\sigma_X > \sigma_Y$.

   (iii) if $\rho_{XY} > 0$ then $Cov[X,Y] > 0$.

   Now pick one of the following:

   (a) all of the above are true

   (b) (i) and (ii) are true

   (c) (i) and (iii) are true

   (d) (ii) and (iii) are true

(e) only (iii) is true

4. Suppose you borrow \$1000 for one year at a variable interest rate tied to the yield on government bonds. As a result, the total interest you will pay (in dollars) is a random variable X_1, having mean 60 and standard deviation 2. You invest the borrowed money. Your earnings on the investment, X_2, have mean 85 and standard deviation 8. Suppose the correlation between your earnings and the interest you pay on the loan is 0.3.

   (a) Your *net* earnings at the end of the year are $Y = X_2 - X_1$. Find the expected value of your net earnings.

   (b) Find the standard deviation of your net earnings.

5. *** A company knows that it will buy 1 million gallons of jet fuel in 3 months, and wants to hedge against a possible price increase. The company chooses to hedge by buying futures contracts on heating oil. Suppose the standard deviation of changes in the price per gallon of jet fuel over 3 months is 0.032, the standard deviation of changes in the futures price per gallon of heating oil is 0.040, and the correlation between the two is 0.8. Also, each heating oil futures contract is for 42,000 gallons.

   (a) What is the standard deviation of the company's unhedged exposure? (Think of the company as holding $-1,000,000$ gallons of jet fuel, because of its anticipated purchase.)

   (b) A simple gallon-for-gallon hedging rule would suggest that the company should buy $1,000,000/42,000 = 23.8 \approx 24$ contracts. What is the standard deviation of the company's exposure under this strategy? (Hint: Let X = change in price of jet fuel and Y = change in futures price (per gallon) of heating oil. Write the exposure in terms of X and Y.)

   (c) Find the number of contract that minimizes the standard deviation of the company's exposure.

Solutions: Covariance and Correlation

1. To meet a 99% service objective, you would need to set the inventory at roughly 2 or 3 standard deviations above the mean. (Mean and standard deviation here refer to the demand for the product.) The standard deviation of the total demand is given by

$$\sigma_{\text{total}} = \sqrt{\sigma^2_{\text{walk-in}} + \sigma^2_{\text{phone}} + 2\sigma_{\text{walk-in}}\sigma_{\text{phone}}\rho},$$

where $\sigma_{\text{walk-in}}$, σ_{phone} are the standard deviations for the individual types of demands and ρ is their correlation. If the demand streams are positively correlated, the variability of total demand will be greater, resulting in a higher inventory requirement. If the demands are negatively correlated, total variability and required inventory will be lower.

2. (a) $E[3X + 2Y] = 3E[X] + 2E[Y] = 3(.09) + 2(.12) = .51$.

(b) $Var[3X + 2Y] = 9\sigma^2_X + 4\sigma^2_Y + 2(3)(2)\sigma_X\sigma_Y\rho = 1.04$. The standard deviation is the square root, 1.02.

3. The answer is (e). In more detail:

(i) False, because σ_X could be larger than σ_Z.

(ii) False, because we don't know anything about the covariances.

(iii) True, because standard deviations are positive so the correlation and covariance always have the same sign.

4. (a) $E[Y] = E[X_2 - X_1] = E[X_2] - E[X_1] = 85 - 60 = 25$.

(b) $Var[Y] = Var[X_2 + (-1)X_1] = \sigma^2_{X_1} + \sigma^2_{X_2} - 2\sigma_{X_1}\sigma_{X_2}\rho = 58.4$. So, $\sigma_Y = \sqrt{58.4} = 7.64$.

5. (a) The standard deviation of the price change for one gallon is 0.032 so the standard deviation for $-1,000,000$ gallons is 32,000.

(b) After buying 24 contracts, the company in effect has a position of $-1,000,000X + 24 \cdot 42,000Y$, using the notation in the hint. The resulting standard deviation is

$$\begin{aligned}
& \sqrt{(-1,000,000)^2\sigma^2_X + (24 \cdot 42,000)^2\sigma^2_Y + 2(-1,000,000)(24)(42,000)\sigma_X\sigma_Y\rho} \\
= \; & \sqrt{(-1,000,000)^2(.032)^2 + (24 \cdot 42,000)^2(.040)^2 + 2(-1,000,000)(24)(42,000)(.032)(.040)(0.8)} \\
= \; & 24,193.
\end{aligned}$$

This is lower than the unhedged standard deviation.

(c) A simple way to find the mimimum is to repeat the calculation in (b) with 24 replaced by several different values and then graph the results; see Figure 5. From this it is clear that the answer is about 15.

For a more systematic approach, let b denote the number of contracts and notice that we want to find the value of b that minimizes

$$(-1,000,000)^2\sigma^2_X + (b \cdot 42,000)^2\sigma^2_Y + 2(-1,000,000)b(42,000)\sigma_X\sigma_Y\rho.$$

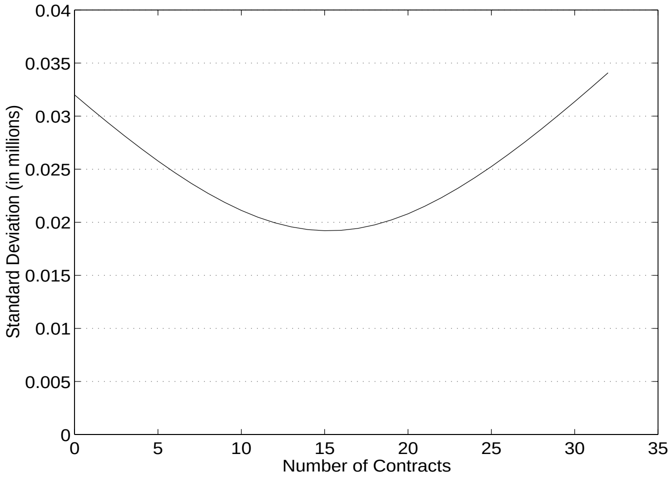


Figure 5: Hedge Effectiveness

This is a quadratic function of b, hence the shape of the graph in Figure 5. Setting the derivative with respect to b equal to zero, we get

$$2b \cdot (42,000)^2 \sigma_Y^2 + 2(-1,000,000)(42,000)\sigma_X \sigma_Y \rho = 0,$$

so the optimal number is

$$b^* = \frac{\sigma_X}{\sigma_Y}\rho \times \frac{1,000,000}{42,000},$$

which works out to be 15.2, or about 15 contracts.

The factor 1,000,000/42,000 is just to get the units right (it converts gallons to contracts). The more interesting factor is $\sigma_X \rho / \sigma_Y$. This is the *minimum variance hedge ratio* and applies whenever we want to hedge X using Y, whatever X and Y may be. If we plug b^* back into the formula for the variance, we get

$$\frac{\text{Variance at optimal hedge}}{\text{Variance with no hedge}} = 1 - \rho^2.$$

Thus, the correlation between the X and Y determines how effectively we can hedge X using Y. If ρ is close to ± 1, we can eliminate almost all the risk in one by hedging with the other.

Exercises

Normal Distribution

Fall 2001
B6014: Managerial Statistics

Professor Paul Glasserman
403 Uris Hall

1. Daily demand for widgets is normally distributed with a mean of 100 and a standard deviation of 15.

   (a) What is the probability that the demand in a day will exceed 125?

   (b) What is the probability that demand will be less than 75? Less than 70?

   (c) How many widgets should be stocked to ensure that with 95% probability all demands will be met?

2. The plant manager of a manufacturing facility is concerned about drug use among plant workers and plans to implement random drug testing. One of the tests to be applied measures the level of factor-X in blood samples. Among recent users of cocaine, the level of factor-X is normally distributed with a mean of 10.0 and a standard deviation of 1.3. Among non-users, the level is normally distributed with a mean of 6.75 and a standard deviation of 1.5.

   The employer plans to send a warning letter to all employees with a factor-X level of x or greater, with x to be determined.

   (a) Find the value of x that will ensure that 90% of recent cocaine users will be sent a warning letter.

   (b) If your answer to part (a) is adopted, what proportion of non-users will also be sent warning letters?

3. A bank finds that the one-day increase in the dollar value of its foreign exchange portfolio is normally distributed with a mean of $1.5 million and a standard deviation of $9.7 million. (A negative increase is a loss.)

   (a) Find the value x such that the probability that the portfolio will lose more than x dollars in one day is 5%.

   (b) For the x you found in part (a), what is the probability that the portfolio will *increase* in value by more than x dollars in one day?

4. Mogul Magazine has recently completed an analysis of its customer base. It has determined that 75% of the issues sold each month are subscriptions and the other 25% are sold at newsstands. It has also determined that the ages of its subscribers are normally distributed with a mean of 44.5 and a standard deviation of 7.42 years, whereas the ages of its newsstand customers are normally distributed with a mean of 36.1 and a standard deviation of 8.20 years.

(a) Mogul would like to make the following statement to its advertisers: "80% of our subscribers are between the ages of ___ and ___." Your job is to fill in the blanks, choosing a range that is symmetric around the mean. (In other words, the mean age of subscribers should be the midpoint of the range.)

(b) What proportion of Mogul's newsstand customers have ages in the range you gave in (a)?

(c) What proportion of all of Mogul's customers have ages in the range you gave in (a)?

Solutions: Normal Distribution

1. (a) Demand $X \sim N(100, (15)^2)$. Standardizing, we get $P(X > 125) = P(Z > [125 - 100]/15) = P(Z > 1.67) = 1 - 0.9525 = 0.0475$. (b) By symmetry of the normal, the probability below 75 is the same as the probability above 125 (both are 25 away from the mean of 100) so the answer is again 0.0475. For 70, standardize to get $P(X < 70) = P(Z < [70 - 100]/15) = P(Z < -2) = P(Z > 2) = 1 - P(Z < 2) = 1 - 0.9772 = 0.0228$. (c) The z-value for 0.95 is 1.645 so we need to set the inventory level 1.645 standard deviations above the mean, at $100 + 1.645 \cdot 15 = 124.67$.

2. (a) Need to find x so that $P(X > x) = 0.90$, where $X \sim N(10, 1.3^2)$. Standardizing, we find that this is the same as $P(Z > (x - 10)/1.3) = 0.90$. From the table, we find that $(x - 10)/1.3$ must be -1.28. Thus, $x = 10 - (1.28)(1.3) = 8.336$.

   (b) Now we need to find $P(Y > x)$, with $x = 8.336$ and $Y \sim N(6.75, 1.5^2)$; i.e., $P(Z > (8.336 - 6.75)/1.5)$, which is $P(Z > 1.0573) = 1 - P(Z < 1.0573) = 1 - .8554 = .1446$.

3. (a) Let X be change in portfolio value, $X \sim N(1.5, (9.7)^2)$. Positive changes are gains, negative changes are losses. Need to find x so that $P(X < -x) = 0.05$. Standardizing, we find that $-x = 1.5 - 1.645(9.7) = -14.456$, so $x = 14.456$. There is only a 5% chance that the portfolio will lose more than 14.456 million dollars.

   (b) Now we need to find $P(X > 14.456)$. Standardizing, this becomes $P(Z > 1.337) = 1 - 0.909 = .091$.

4. (a) For the range from $\mu - x$ to $\mu + x$ to contain the middle 80% of the subscribers, 10% must be abve $\mu + x$ so 90% must be below $\mu + x$. From the normal table, we find that the z-value for 90% is 1.28. Thus, x must be 1.28 standard deviations or $1.28 \cdot 7.42 = 9.5$. The resulting range is from 35 to 54.

   (b) Let Y be $N(36.1, (8.2)^2)$. We need to find

   $$P(35 < Y < 54) = P(Y < 54) - P(Y < 35).$$

   Standardizing, we find that $P(Y < 54) - P(Y < 35)$ equals

   $$P(Z < \frac{54 - 36.1}{8.2}) - P(Z < \frac{35 - 36.1}{8.2}) = P(Z < 2.18) - P(Z < -0.13) = 0.985 - 0.447 = 0.538.$$

   (c) Of the 75% in the subscriber group, 80% are in the range. Of the remaining 25%, 53.8% are in the range. Thus, the overall fraction is $(0.75)(0.80) + (0.25)(.538) = 0.734$.

Exercises

Sampling

Fall 2001 — Professor Paul Glasserman
B6014: Managerial Statistics — 403 Uris Hall

1. Let $X_1, \ldots, X_{10}$ be a random sample from a population with mean 50 and standard deviation 4. Let $\bar{X}$ be the sample mean. Find the expected value and standard deviation of $\bar{X}$.

2. Each visit to Uris.com has a 2% chance of turning into a purchase. Let $\hat{p}$ denote the proportion of visits that turn into purchases from a random sample of 100 visits. Find the expected value and standard deviation of $\hat{p}$.

3. (a) The starting salary among 1999 graduates of the Evian Business School has a standard deviation of \$17,000. If you randomly survey 40 students and average their starting salaries, what is the probability that the average among these students will be greater than the average among all students?

   (b) What is the probability that the average in the sample will exceed the average among all students by more than \$5,000?

   (c) What is the probability that the average in the sample will differ from the average among all students by more than \$5,000?

4. A trader's annual bonus is normally distributed with a mean of \$650,000 and standard deviation \$125,000. Her bonuses are independent from year to year. Find the probability that her average bonus over a five-year period will be less than \$500,000.

5. Fifty-seven percent of students at Calabria Business School support making Introduction to Philanthropy a required course. The school plans to survey 100 students to gauge opinion on this issue. What is the probability that fewer than half of those surveyed will say they support requiring the course?

6. A wireless communication company is considering switching from a per-minute charge to a flat monthly fee for unlimited service. It anticipates that this will result in greater usage and would like to estimate what the average number of minutes per month per customer will be under the new plan. To do this, it offers 1000 customers the flat-fee plan for one month and tracks their usage. From past data, the company knows that the standard deviation of monthly minutes is about 120 minutes; the company expects that this figure will not change much under the new plan. Find the probability that the average among the 1000 test customers will differ from the true mean by more than 5 minutes.

7. A company believes that roughly 35% of consumers rate its brand first in quality. It would like to estimate the proportion more accurately and retains a market research firm to survey n consumers. Using the initial estimate of 35% as a guide, determine how large n must be to ensure that there is only a 5% chance that the estimate from the survey will differ from the true value by more than 3 percentage points.

Solutions: Sampling

1. $\bar{X}$ has expected value 50 and standard deviation $\sigma/\sqrt{n} = 4/\sqrt{10} = 1.26$.

2. $\hat{p}$ has expected value 0.02 and standard deviation $\sqrt{p(1-p)/n} = \sqrt{.02(1-.02)/100} = 0.014$.

3. Let $\bar{X}$ denote the average for the sample and μ the average among all students. Then $\bar{X}$ is (approximately) normally distributed with a mean of μ and a variance of σ^2/n, where σ is given as \$17,000 and $n = 40$.

   (a) $P(\bar{X} > \mu) = 1/2$ because the mean of a normal distribution is also its median.

   (b) $P(\bar{X} - \mu > 5000) = P(Z > 5000/(17000/\sqrt{40})) = P(Z > 1.86) = 1 - 0.9686 = 0.0314$.

   (c) By symmetry, this is twice as large as the answer to (b) and thus 0.0628.

4. Let $\bar{X}$ denote the five-year average. Then $\bar{X}$ is normal with mean 650000 and standard deviation $125000/\sqrt{5} = 55902$. Thus, $P(\bar{X} < 500000) = P(Z < [500000 - 650000]/55902) = P(Z < -2.68) = 0.0037$.

5. Let $\hat{p}$ denote the proportion in favor out of the 100 surveyed. Then $\hat{p}$ is approximately normally distributed with a mean of $p = 57\%$ and a standard deviation of $\sqrt{p(1-p)/n}$, $n = 100$. Thus,

$$\begin{aligned} P(\hat{p} < 0.50) &= P\left(\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < \frac{0.50 - p}{\sqrt{p(1-p)/n}}\right) \\ &= P\left(Z < \frac{.07}{\sqrt{.57(1-.57)/100}}\right) \\ &= P(Z < -1.41) = 0.0793. \end{aligned}$$

6. Let $\bar{X}$ denote the sample mean among the 1000 test customers and let μ denote the true mean under the new plan. Then $\bar{X} \approx N(\mu, \sigma^2/n)$, with σ given as 120 and $n = 1000$. We need to find

$$P(\bar{X} < \mu - 5 \text{ or } \bar{X} > \mu + 5).$$

   By symmetry of the normal, this is the same as

$$2P(\bar{X} > \mu + 5).$$

   Standardizing, this becomes

$$2P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{5}{\sigma/\sqrt{n}}\right) = 2P\left(Z > \frac{5}{120/\sqrt{1000}}\right) = 2P(Z > 1.32) = 2(1 - 0.9066) = 18.68\%.$$

7. Let $\hat{p}$ denote the estimate obtained by the market research firm and let p denote the true proportion. Even though we don't know p, $\hat{p}$, or n, we do know that the error in the

estimate, $\hat{p} - p$ is approximately normally distributed with mean of 0 and a standard deviation of

$$\sigma = \sqrt{\frac{p(1-p)}{n}}.$$

It follows that there is only a 5% chance that the error will be greater than 1.96σ. We would therefore like to choose n so that $0.03 = 1.96\sigma$; i.e.,

$$0.03 = 1.96\sqrt{\frac{p(1-p)}{n}}.$$

Solving for n, we get

$$n = (p(1-p)) \cdot (1.96/0.03)^2.$$

Of course, we still don't know p, but here's where we use the rough prior estimate of 35% as a guide. Substituting this value, we find

$$n \approx (0.35(1-0.35)) \cdot (1.96/0.03)^2 = 971.$$

Exercises

Confidence Intervals

Fall 2001 Professor Paul Glasserman
B6014: Managerial Statistics 403 Uris Hall

1. Let $X_1, \ldots, X_{10}$ be a random sample from a normal population with mean μ and standard deviation 4. Let $\bar{X}$ be the sample mean and suppose $\bar{X} = 48$. Find a 95% confidence interval for μ.

2. Each visit to Uris.com has a probability p of resulting in a purchase. Out of a random sample of 500 visits, 15 result in purchases. Use this to find a 95% confidence interval for p.

3. In a random sample of 40 students from the 1999 class at Evian Business School, the average staring salary is \$110 thousand and the sample standard deviation is \$16 thousand. Find a 95% confidence interval for the population mean.

4. A wireless communication company is considering switching from a per-minute charge to a flat monthly fee for unlimited service. It anticipates that this will result in greater usage and would like to estimate the change in the average number of minutes per month per customer resulting from the new plan. To do this, it offers 850 customers the flat-fee plan for one month and tracks their usage. It also tracks a control group of 700 customers under the old plan. Among the 850 test customers, it finds an average usage of 227 minutes and a sample standard deviation of 135 minutes; in the control group, the corresponding values are 163 and 85. Find a 95% confidence interval for the increase in mean usage.

5. For a t-based confidence interval, what multiplier would you use with a sample size of 15 and a confidence level of 95%? 90%? What confidence level would you get with a multiplier of 2.624?

6. Six months after elections in the Democratic Republic of Urisia, a newspaper reports that two-thirds of French companies with major government contracts have enjoyed increases in the value of their contracts since the election and accuses the government of pro-French policies. Table 3 shows the contract values (in millions of dollars) for the 12 French companies with a major presence in Urisia, before and after the election. Using the t distribution, calculate a 95% confidence interval to assess the difference in average contract size before and after the election. What does this confidence interval suggest?

7. (a) A recent study compared HMOs that provide financial incentives for physicians to reduce referrals with HMOs that do not. The study found that in the first group 12 out of 52 physicians surveyed said they would refer a patient with chest pain upon waking

	Before	After	Difference
	68.3	71.5	3.2
	37.0	32.5	-4.6
	121.4	122.1	0.7
	113.1	119.8	6.6
	117.3	111.2	-6.1
	47.7	44.0	-3.6
	85.7	94.6	8.8
	99.4	103.6	4.2
	108.8	129.4	20.6
	92.0	74.7	-17.2
	63.3	75.1	11.8
	33.9	39.2	5.3
Average	82.3	84.8	2.5
Std Dev	31.6	33.8	9.7

Table 3: Data for Problem 6

to a cardiologist; in the second group, 18 out of 45 physicians surveyed said they would make the referral. Find a 95% confidence interval for the difference.

(b) Repeat the calculation now assuming the proportions are 120 out of 520 and 180 out of 450. Comment on the difference between this interval and the one in part (a).

Solutions: Confidence Intervals

1. $\bar{X} \pm 1.96\sigma/\sqrt{n}$; i.e., $48 \pm 1.96(4/\sqrt{10})$, which is 48 ± 2.48.

2. The sample proportion is $\hat{p} = 15/500 = .03$. Confidence interval:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = .03 \pm 1.96\sqrt{\frac{(.03)(1-.03)}{500}} = .03 \pm .015.$$

3. $\bar{X} \pm 1.96s/\sqrt{n}$, with $\bar{X} = 110$, $s = 16$, and $n = 40$, gives 110 ± 4.96.

4. Test customers: $\bar{X} = 227$, $s_X = 135$, $n_X = 850$. Control: $\bar{Y} = 163$, $s_Y = 85$, $n_Y = 700$. Confidence interval

$$(\bar{X} - \bar{Y}) \pm 1.96\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}} = 164 \pm 11.04.$$

5. Look up $t_{n-1,\alpha/2}$ with $n = 15$ and $\alpha = 5\%$; i.e., $t_{14,.025}$ which is 2.145. For 90% it's 1.761. A multiplier of 2.624 leaves 1% probability in each tail and thus results in a 98% confidence level.

6. This is a matched-pairs setting because we are comparing the same companies before and after the election. The average of the differences is 2.5 and their standard deviation is 9.7. The sample size is 12, so we use the t-multiplier for 11 degrees of freedom and 95% confidence (right-tail probability 0.025) which is 2.201. Putting the pieces together we get a confidence interval of

$$2.5 \pm 2.201\frac{9.5}{\sqrt{12}} = 2.5 \pm 6.0.$$

Comment: This confidence interval is very wide. The uncertainty (as measured by the halfwidth of 6.0) is very large compared with the average increas of 2.5 million. Indeed, we cannot even be confident that the mean changed after the election because the interval crosses zero. (Note that for this interpretation of the confidence interval we need to view the 12 companies as a random sample from a larger hypothetical population.)

7. (a) No incentive: $\hat{p}_X = 18/45 = 0.40$, $n_X = 45$. With incentive: $\hat{p}_Y = 12/52 = 0.23$, $n_Y = 52$. Confidence interval:

$$\hat{p}_X - \hat{p}_Y \pm 1.96\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}} = 0.17 \pm 0.18.$$

(b) Now we have $\hat{p}_X = 0.40$, $n_X = 450$, $\hat{p}_Y = 0.23$, $n_Y = 520$. The same formula gives a confidence interval of 0.17 ± 0.058. We can make two comments comparing (a) and (b). First, the confidence interval in (b) is narrower by a factor of $\sqrt{10}$ because both sample sizes in (b) are 10 times larger. Second, even though the difference in the two proportions is the same in both cases (0.17), this difference is more convincing in (b) because of the narrower interval. This is again a consequence of the larger sample size.

Exercises

Hypothesis Testing

Fall 2001 Professor Paul Glasserman
B6014: Managerial Statistics 403 Uris Hall

1. Let $X_1, \ldots, X_{10}$ be a random sample from a normal population with mean μ and standard deviation 4. Let $\bar{X}$ be the sample mean and suppose $\bar{X} = 48$. Let μ denote the unknown population mean. Test $H_0 : \mu = 45$ vs. $H_1 : \mu \neq 45$ at the 5% level. Test $H_0 : \mu \leq 45$ vs. $H_1 : \mu > 45$ at the 5% level.

2. Each visit to Uris.com has a probability p of resulting in a purchase. Out of a random sample of 500 visits, 15 result in purchases. Is this statistically significant evidence that $p \neq 2\%$?

3. In a random sample of 40 students from the 1999 class at Evian Business School, the average staring salary is \$110 thousand and the sample standard deviation is \$16 thousand. How significant is the evidence that the population mean is greater than \$100 thousand? Give a p-value.

4. A wireless communication company is considering switching from a per-minute charge to a flat monthly fee for unlimited service. It anticipates that this will result in greater usage and would like to estimate the change in the average number of minutes per month per customer resulting from the new plan. To do this, it offers 850 customers the flat-fee plan for one month and tracks their usage. It also tracks a control group of 700 customers under the old plan. Among the 850 test customers, it finds an average usage of 227 minutes and a sample standard deviation of 135 minutes; in the control group, the corresponding values are 163 and 85. Is the increase in mean usage statistically significant?

5. Six months after elections in the Democratic Republic of Urisia, a newspaper reports that two-thirds of French companies with major government contracts have enjoyed increases in the value of their contracts since the election and accuses the government of pro-French policies. Table 4 shows the contract values (in millions of dollars) for the 12 French companies with a major presence in Urisia, before and after the election. Using the t distribution, test whether the difference in average contract size before and after is statistically significant at the 5% level.

6. (a) A recent study compared HMOs that provide financial incentives for physicians to reduce referrals with HMOs that do not. The study found that in the first group 12 out of 52 physicians surveyed said they would refer a patient with chest pain upon waking to a cardiologist; in the second group, 18 out of 45 physicians surveyed said they would make the referral. Is the difference statistically significant?

   (b) Repeat the test now assuming the proportions are 120 out of 520 and 180 out of 450. Comment on the difference between this and part (a).

	Before	After	Difference
	68.3	71.5	3.2
	37.0	32.5	-4.6
	121.4	122.1	0.7
	113.1	119.8	6.6
	117.3	111.2	-6.1
	47.7	44.0	-3.6
	85.7	94.6	8.8
	99.4	103.6	4.2
	108.8	129.4	20.6
	92.0	74.7	-17.2
	63.3	75.1	11.8
	33.9	39.2	5.3
Average	82.3	84.8	2.5
Std Dev	31.6	33.8	9.7

Table 4: Data for Problem 5

Solutions: Hypothesis Testing

1. Test statistic:

$$Z = \frac{\bar{X} - 45}{4/\sqrt{10}} = 2.37.$$

The rejection criterion at 5% is $Z > 1.96$ or $Z < -1.96$, so we reject the null hypothesis. In the second test, the rejection criterion is $Z > 1.645$ so we again reject the null.

2. Set up the test as $H_0 : p = .02$ vs. $H_1 : p \neq .02$. Test statistic:

$$Z = \frac{\hat{p} - .02}{\sqrt{\frac{.02(1-.02)}{500}}} = 1.60,$$

with $\hat{p} = .03$. Note that we use .02 (the value under the null) in calculating the standard error. At 5% significance level, we would reject if $Z > 1.96$ or $Z < -1.96$; clearly, the evidence is therefore not significant at 5%. To measure how significant it is we find a p-value. This is the area outside the range $(-1.60, 1.60)$ under the standard normal, which is 11%.

3. For $H_0 : \mu \leq 100$ vs. $H_1 : \mu > 100$ the test statistic is

$$Z = \frac{\bar{X} - 100}{s/\sqrt{n}} = \frac{110 - 100}{16/\sqrt{40}} = 3.95.$$

The p-value is the area to the right of 3.95 under the standard normal curve which is zero to four decimal places. The evidence is therefore extremely significant.

4. Set up the test as $H_0 : \mu_X - \mu_Y \leq 0$ vs. $H_0 : \mu_X - \mu_Y > 0$, with X the new-plan customers and Y the control group. The test statistics is

$$Z = \frac{\bar{X} - \bar{Y} - 0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}.$$

Using $\bar{X} = 227$, $s_X = 135$, $n_X = 850$, and $\bar{Y} = 163$, $s_Y = 85$, $n_Y = 700$, this evaluates to

$$\frac{227 - 163}{\sqrt{\frac{135^2}{850} + \frac{85^2}{700}}} = 11.35.$$

This is a very large Z value; the null can be rejected at almost significance level; the evidence is highly significant.

5. This is a matched-pairs setting because we are comparing the same companies before and after the election. We need to test $H_0 : \mu_X - \mu_Y = 0$ vs. $H_0 : \mu_X - \mu_Y \neq 0$, with $Y =$ Before and $X =$ After. The average of the differences is 2.5 and their standard deviation is 9.7. The test statistic is $t = 2.5/(9.7/\sqrt{12}) = 0.89$. To reject at 5% signifcance we would need $t > t_{11,.025}$ or $t < -t_{11,.025}$. But $t_{11,.025} = 2.201$ so we cannot reject the null. The test statistic 0.89 tells us that 2.5 is not statistically different from 0 because it is only 0.89 standard errors away from 0.

6. We need to test $H_0 : p_X - p_Y = 0$ vs. $H_0 : p_X - p_Y \neq 0$ with $X =$ no incentive and $Y =$ with incentive. (a) We are given $\hat{p}_X = 18/45 = 0.40$, $n_X = 45$. With incentive: $\hat{p}_Y = 12/52 = 0.23$, $n_Y = 52$. Under the null hypothesis, the two proportions are equal so we first need to calculate the pooled estimate

$$\hat{p}_0 = \frac{18 + 12}{45 + 52} = 0.31.$$

The standard error under the null is therefore

$$\sqrt{\frac{\hat{p}_0(1 - \hat{p}_0)}{n_X} + \frac{\hat{p}_0(1 - \hat{p}_0)}{n_Y}} = 0.094.$$

This results in a test statistic of

$$Z = \frac{\hat{p}_X - \hat{p}_Y}{0.094} = \frac{0.40 - 0.23}{0.094} = 1.80.$$

Because this fails to exceed 1.96, it is not significant at 5% (though it would be significant at 10%).

(b) With the larger sample sizes $\hat{p}_0$ is unchanged but the standard error becomes

$$\sqrt{\frac{\hat{p}_0(1 - \hat{p}_0)}{450} + \frac{\hat{p}_0(1 - \hat{p}_0)}{520}} = 0.030.$$

The test statistic becomes $Z = 0.17/0.030 = 5.71$ which now looks extremely significant. Thus, the difference of 17% looks significant in (b) but not in (a) because (b) is based on a much larger sample size.

Exercises

Regression Analysis

Fall 2001
B6014: Managerial Statistics

Professor Paul Glasserman
403 Uris Hall

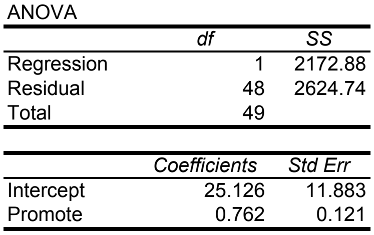
The raw data for all of these exercises can be downloaded in the file Exercises_Regression.xls. It is not necessary to download this file to solve these exercises but you may find it helpful.

The data for Exercises 1-4 is from *Managerial Statistics*, by S.C. Albright, W. Winston, and C. Zappe, published by Duxbury Press.

1. A drugstore chain has collected data on of 50 of its stores. For each store it has values of the following two variables:

   PROMOTE: Store promotional expenditure as a percentage of expenditure
   by leading competitor in the store's area
   SALES: Store sales as a percentage of leading competitor's

   Figure 6 shows part of the output from a regression of SALES against PROMOTE.



ANOVA

	df	*SS*
Regression	1	2172.88
Residual	48	2624.74
Total	49	

	Coefficients	*Std Err*
Intercept	25.126	11.883
Promote	0.762	0.121

Figure 6: Drugstore Regression

   (a) Interpret the coefficient for PROMOTE in words.

   (b) Find the R^2 and standard error of the estimate s_e.

   (c) Find the t-statistic for PROMOTE. What can you conclude from this value?

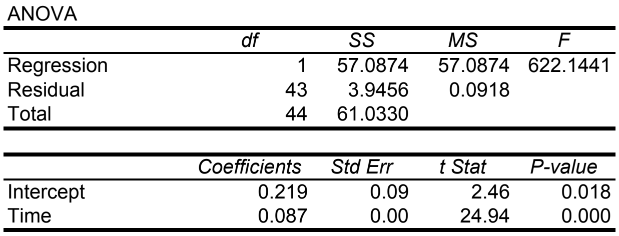
   (d) Find a 95% confidence interval the coefficient for PROMOTE.

2. Figure 7 shows results of a regression of the US minimum wage during 1950-1994 against time. The time variable is measured in years since 1950 and therefore runs from 0 to 44. The wage variable is in dollars per hour and increased from 0.75 to 4.25 over this period.

   (a) Interpret the coefficient for Time in words.

   (b) Find a 95% confidence interval for this coefficient.

   (c) Calculate the R^2. What does this say about the growth in the minimum wage over time?



ANOVA

	df	SS	MS	F
Regression	1	57.0874	57.0874	622.1441
Residual	43	3.9456	0.0918	
Total	44	61.0330		

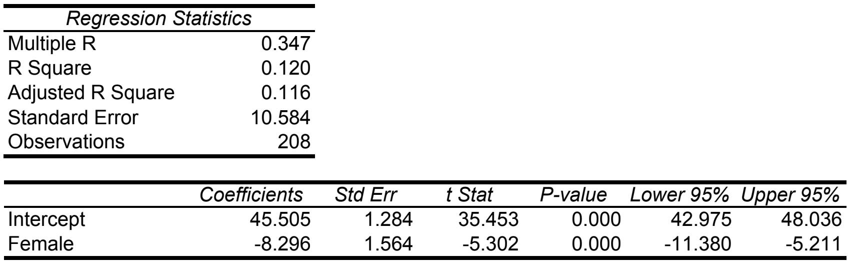
	Coefficients	Std Err	t Stat	P-value
Intercept	0.219	0.09	2.46	0.018
Time	0.087	0.00	24.94	0.000

Figure 7: Regression of minimum wage against time

3. A bank is facing a sex discrimination suit in which it is accused of paying men more than women.[1] Consider the following analysis of 208 employees. For each employee we have the following information:

   SALARY: annual salary in thousands of dollars
   FEMALE: 1 for women, 0 for men
   YrsPrior: years of experience in banking prior to hiring
   YrsExper: years working at this bank

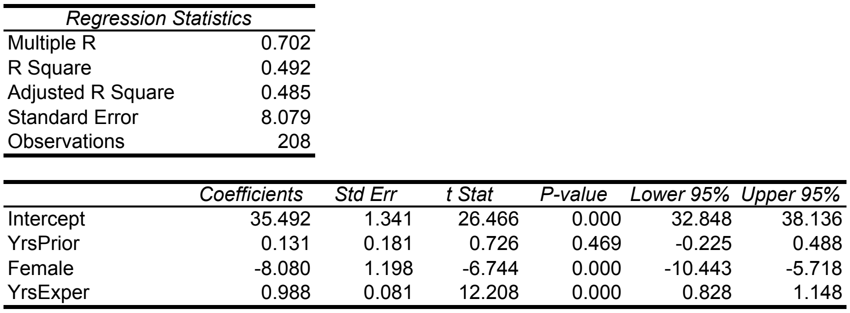
   Figures 8 and 9 show results of regressions based on this data. The bank also has information about the positions held and education backgrounds of the employees, but that is not used in these regressions.



Regression Statistics	
Multiple R	0.347
R Square	0.120
Adjusted R Square	0.116
Standard Error	10.584
Observations	208

	Coefficients	Std Err	t Stat	P-value	Lower 95%	Upper 95%
Intercept	45.505	1.284	35.453	0.000	42.975	48.036
Female	-8.296	1.564	-5.302	0.000	-11.380	-5.211

Figure 8: Regression of bank salaries



Regression Statistics	
Multiple R	0.702
R Square	0.492
Adjusted R Square	0.485
Standard Error	8.079
Observations	208

	Coefficients	Std Err	t Stat	P-value	Lower 95%	Upper 95%
Intercept	35.492	1.341	26.466	0.000	32.848	38.136
YrsPrior	0.131	0.181	0.726	0.469	-0.225	0.488
Female	-8.080	1.198	-6.744	0.000	-10.443	-5.718
YrsExper	0.988	0.081	12.208	0.000	0.828	1.148

Figure 9: Regression of bank salaries

[1] Albright, Winston, and Zappe state that this is real data taken from an actual law suit.

(a) Interpret the coefficient for Female in Figure 8. Interpret the p-value that goes with it.

(b) Interpret the coefficient for Female in Figure 9. Interpret the p-value that goes with it.

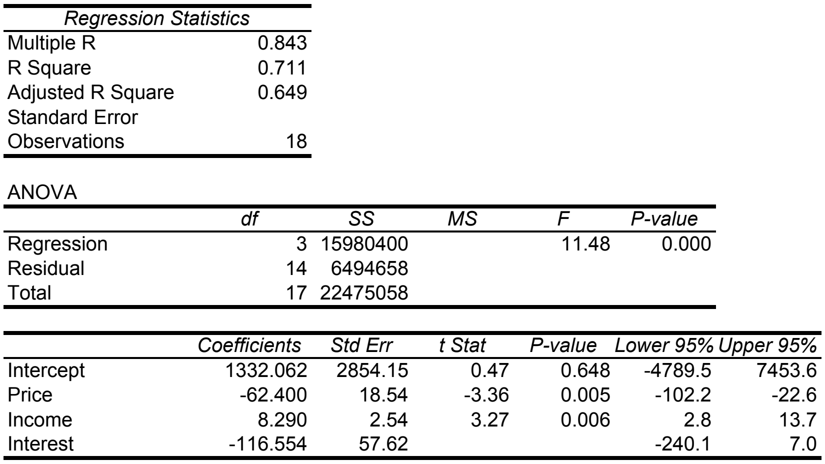
(c) The coefficient for Female has larger magnitude in Figure 8 than Figure 9. Based on information in the regression outputs, can you suggest an explanation for this? What other information in the outputs might temper your conclusion?

(d) You have been hired to help defend the bank. What argument could you offer to defend the bank against the evidence in the figures? What additional regressions might support your argument?

4. Figure 10 shows the output from a regression of the following variables

   QUANTITY: number of cars sold in US, in thousands
   PRICE: index of inflation-adjusted car prices
   INCOME: inflation-adjusted measure of disposable income, in thousands
   INTEREST: prime rate of interest

   We have one value of each of these variables for each year running from 1970 to 1987. QUANTITY is the dependent variable.



Regression Statistics	
Multiple R	0.843
R Square	0.711
Adjusted R Square	0.649
Standard Error	
Observations	18

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	3	15980400		11.48	0.000
Residual	14	6494658			
Total	17	22475058			

	Coefficients	*Std Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1332.062	2854.15	0.47	0.648	-4789.5	7453.6
Price	-62.400	18.54	-3.36	0.005	-102.2	-22.6
Income	8.290	2.54	3.27	0.006	2.8	13.7
Interest	-116.554	57.62			-240.1	7.0

Figure 10: Regression of car sales

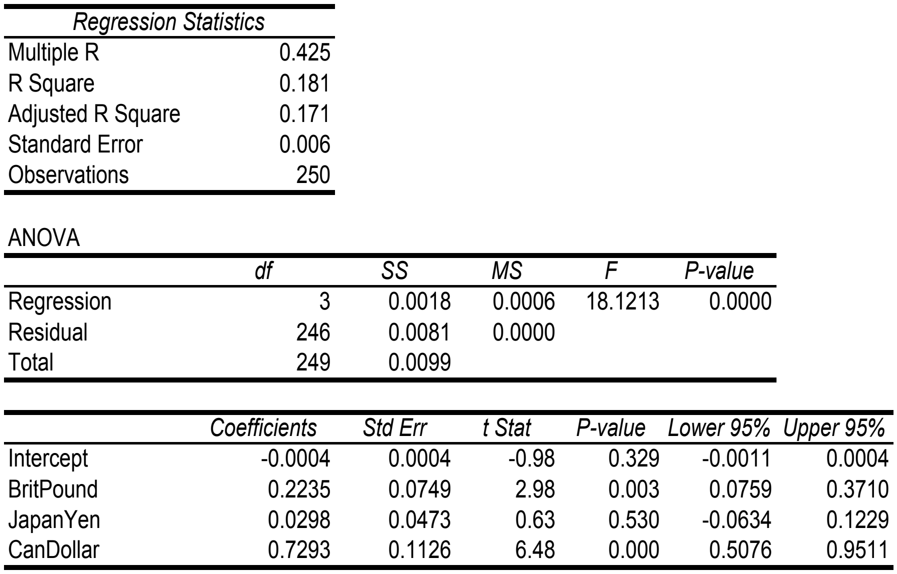
   (a) Find s_e, the standard error of the estimate. Interpret the value you get.

   (b) Find the p-value for INTEREST. (The Excel function TDIST will give the exact value; altneratively, use the z or t table to get an approximate value.) Relate the value you get to the confidence interval for INTEREST.

   (c) Use the regression model to forecast car sales in 2001, assuming a PRICE index level of 285, and INCOME index of 3,110, and prime rate of 9.50% (corresponding to an INTEREST value of 9.5, not 0.095).

5. $* * *$ Figure 11 shows the results of regressing daily returns on the Australian dollar against daily returns on the British Pound, Japanese Yen, and Canadian dollars. The

results are based on 250 days from July 1, 1999 to June 30, 2000. For each currency, the data consists of the daily percentage changes in the number of US dollars per unit of the foreign currency. We may think of these as the percentage changes in the "price" (in US dollars) of the foreign currency.



Regression Statistics	
Multiple R	0.425
R Square	0.181
Adjusted R Square	0.171
Standard Error	0.006
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	3	0.0018	0.0006	18.1213	0.0000
Residual	246	0.0081	0.0000		
Total	249	0.0099			

	Coefficients	*Std Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.0004	0.0004	-0.98	0.329	-0.0011	0.0004
BritPound	0.2235	0.0749	2.98	0.003	0.0759	0.3710
JapanYen	0.0298	0.0473	0.63	0.530	-0.0634	0.1229
CanDollar	0.7293	0.1126	6.48	0.000	0.5076	0.9511

Figure 11: Regression of currencies

(a) State in words the meaning of the coefficients in the regression output.

(b) Suppose you hold 100 million US dollars worth of Australian currency and you want to hedge the possible change in value of this position over the next day. Suppose you can hedge by buying or short selling British Pounds, Japanese Yen, and Canadian dollars. What does the regression suggest you should do, buy or sell? How much of each? [Think about your answer to (a).]

(c) How much "risk" is left after you hedge as in (b)? Be precise about the meaning of your answer, not just the numerical value. What percent is this of the unhedged risk?

(d) Which of the following statements is best supported by the regression output:

(i) Yen should be omitted from the hedge because its coefficient is very small

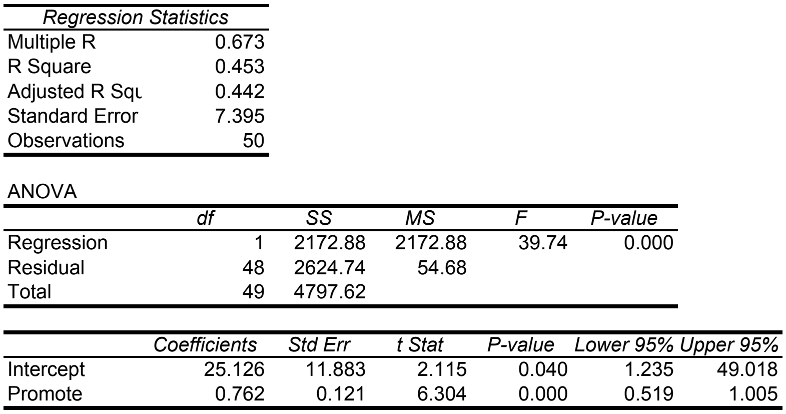
(ii) Yen should be omitted from the hedge because its p-value is rather large

(iii) Yen should be omitted from the hedge both because its coefficient is very small and because its p-value is rather large

(iv) Yen should not be omitted from the hedge

Solutions: Regression Analysis

1. The full output is displayed in Figure 12.



Regression Statistics	
Multiple R	0.673
R Square	0.453
Adjusted R Squ	0.442
Standard Error	7.395
Observations	50

ANOVA

	df	SS	MS	F	P-value
Regression	1	2172.88	2172.88	39.74	0.000
Residual	48	2624.74	54.68		
Total	49	4797.62			

	Coefficients	Std Err	t Stat	P-value	Lower 95%	Upper 95%
Intercept	25.126	11.883	2.115	0.040	1.235	49.018
Promote	0.762	0.121	6.304	0.000	0.519	1.005

Figure 12: Drugstore Regression

(a) An increase in promotional expenditures of 1% of competitors promotional expenditures brings an increase of 0.762% in sales as a percentage of competitor's sales.

(b) $R^2 = SSR/(SSR{+}SSE) = 2172.88/4797.62$ and $s_e = \sqrt{MSE} = \sqrt{SSE/(n-k-1)} = \sqrt{54.68} = 7.395$. See Figure 12.

(c) $t = \hat{\beta}/s_{\hat{\beta}} = 6.304$. Because this is a large t value, we can say that the evidence of a relation between SALES and PROMOTE is statistically significant.

(d) $\hat{\beta} \pm t_{n-k-1,.025} s_{\hat{\beta}} = 0.762 \pm (2.01)(0.121)$. See Figure 12.

2. (a) The coefficient for Time measures the annual rate of increase in the minimum wage. According to the regression output, this rate is estimated at \$0.0867 per year.

(b) $\hat{\beta} \pm t_{43,.025} s_{\hat{\beta}}$; i.e., $0.0867 \pm (2.32)(.0035)$; i.e., $(.0797\,,\,.0937)$.

(c) The R^2 is

$$\begin{aligned} R^2 &= \frac{\text{SSR}}{\text{SSR} + \text{SSE}} \\ &= \frac{57.0874}{57.0874 + 3.9456} \\ &= 93.54\% \end{aligned}$$

This is a high R^2 indicating a good linear fit. This means that the minimum wage grew roughly linearly (i.e., at a constant rate) during 1950-1994.

3. (a) According to Figure 8, the average salary among female employees is \$8,296 less than amonge male employees. The very small p-value indicates that this difference is statistically significant.

(b) According to Figure 9, the average salary among female employees is \$8,080 less than amonge male employees, *once differences in prior years of experience and years with the bank have been accounted for.* The very small p-value indicates that the difference between men and women remains statistically significant *even once these other differences have been accounted for.*

(c) The coefficients for both YrsPrior and YrsExper are positive, indicating that employees with more years of experience tend to have higher salaries. The fact that the difference between average salaries for men and women is smaller when the experience variables are included suggests that women may have lower values of YrsPrior and YrsExper, on average, so that part of the difference observed in Figure 8 is due to differences in experience, rather than differences in sex.

Two considerations temper this argument. First, the confidence intervals for the Female coefficients are very wide compared with the difference across the two regressions; so, the difference many not be meaningful. Second, the p-value for YrsPrior suggests that this variable does not have a statistically significant effect on salary.

(d) To defend the bank, you could argue that these regressions do not take account of other differences between the men and women employed at the bank. For example, women may have been hired with fewer years of education and/or hired into lower-level positions. To support this argument, you would want to run a regression that includes variables for education and job grade to determine whether the salary difference remains significant even after these other factors are considered. (Of course, this would not address the question of whether the initial hiring was discriminatory; rather, it would check whether men and women in comparable positions earn equal pay.)

4. (a) Recall that

$$s_e = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n-k-1}},$$

so

$$s_e = \sqrt{\frac{6494658.12}{14}} = 681.$$

Thus, the s_e is 681,000 cars, and this is roughly the standard error in any forecast we try to make from the model.

(b) First we need the t-stat, which is

$$t = \frac{\hat{\beta} - 0}{s_{\hat{\beta}}} = \frac{-116.554}{57.62} = -2.02.$$

For the p-value, we use TDIST(2.02,14,2) because we have 14 degrees of freedom and we want a 2-sided p-value. The answer is 0.063. Notice that this just fails to be significant at the 5% level; similarly (and equivalently) the 95% confidence interval given in the regression output just straddles 0.

(c) The estimated regression equation is

$$\text{QUANTITY} = 1332.062 - 62.400\text{PRICE} + 8.290\text{INCOME} - 116.554\text{INTEREST}$$

so our forecast is

$$\text{QUANTITY} = 1332.062 - 62.400 \cdot 285 + 8.290 \cdot 3110 - 116.554 \cdot 9.50$$

which is 8223, or 8,223,000 cars.

5. (a) Interpret the coefficient on British Pound as follows: for each percentage point change in the pound, the Australian dollar changes by 0.2235. The interpretation for the others is analogous.

   (b) The coefficients are all positive, which suggests that all four currencies tend to move in the same direction. To hedge a position in Australian dollars, we should therefore take an opposite position in the others. So, we should sell short.

   The interpretation in (a) suggests that we should sell 22.35 million US dollars worth of British pounds, 2.98 million US dollars worth of Yen, 72.93 million US dollars worth of Canadian dollars. If we do this and if, for example, the dollar/pound rate increases by 1% then we lose .2235 million US dollars on the short pound position. The regression predicts that when the US dollar/pound rate goes up by 1%, the US dollar/Australian dollar rate goes up by 0.2235%. So, we earn 0.2235 million on the Australian dollar position, which is offset by the gains on the pound. Of course, in practice these would not offset each other exactly because the regression equation does not hold exactly.

   (c) The residuals of this regression are the hedging errors. Recall that s_e estimates the standard error of the residuals. So, s_e estimates the residual risk. More precisely, the percentage change of the hedged position has a standard deviation of 0.006. On a position of \$100 million, 0.006% means \$6,000. This is the estimated standard deviation of the one-day change in the value of the position.

   Recall that the R^2 gives the ratio of the explained variance to the total variance, and $1-R^2$ gives the ratio of unexplained variance to the total variance. In the current setting, the unexplained variance is how much variance remains after we hedge, and the total variance is the variance without the hedge; similarly,

$$\frac{\text{Var[hedged position]}}{\text{Var[original position]}} = \frac{\text{unexplained variance}}{\text{total variance}} = 1 - R^2.$$

   The ratio of the standard deviations is therefore $\sqrt{1-R^2} = \sqrt{1-0.180} = 0.906$. Hence, we have managed to hedge only about 10% of the risk.

   The spreadsheet Exercises_Regression.xls contains an illustration of this. Using the hedge ratio suggested in (b), it calculates the daily returns on the hedged position. From these values we can calculate the standard deviation over the 250 days to confirm that it is 0.006. We can them compare this with the standard deviation of the unhedged position.

   (d) The best answer is (ii). The magnitude of the coefficient is not relevant to this decision; we could make the coefficient as large or small as we like by changing units (dollars to pennies for example). The p-value suggests that the relation between the Yen and the Australian dollar is not statistically significant once the effect of the other currencies is considered.