Chicago Booth **Spring 2013**
Business Statistics 41000-81/82 **Midterm exam**
SOLUTION

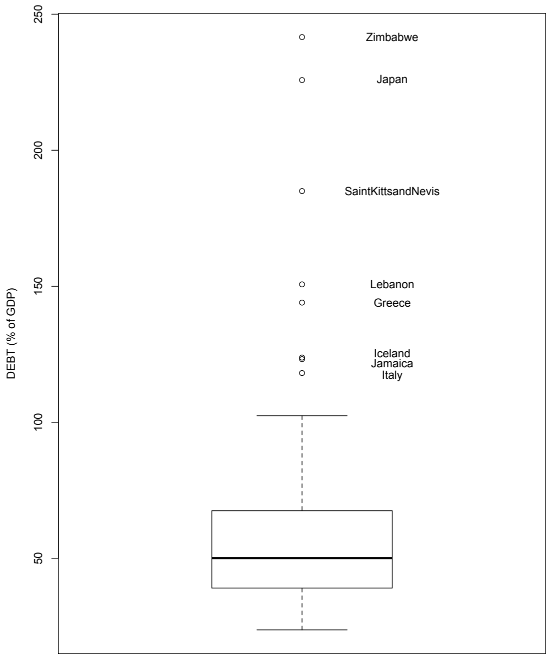
Question I (10): The following table shows public debt (as % of gross domestic product, GDP) for the top 100 countries with largest debts.

	%GDP	Countries	%GDP	Countries	%GDP
Zimbabwe	241.6	Bahrain	59.2	Sweden	40.8
Japan	225.8	United States	58.9	Malawi	40.4
Saint Kitts and Nevis	185.0	Seychelles	58.8	Czech Republic	40.0
Lebanon	150.7	Morocco	58.2	Panama	40.0
Greece	144.0	Bhutan	57.8	Bolivia	39.7
Iceland	123.8	Guyana	57.0	Ethiopia	39.3
Jamaica	123.2	Vietnam	56.7	Bangladesh	39.3
Italy	118.1	Philippines	56.5	Yemen	39.1
Singapore	102.4	Uruguay	56.0	Bosnia and Herzegovina	39.0
Belgium	98.6	India	55.9	Ukraine	38.4
Ireland	94.2	El Salvador	55.0	Switzerland	38.2
Sudan	94.2	Croatia	55.0	Lithuania	36.7
Sri Lanka	86.7	Poland	53.6	Slovenia	35.5
France	83.5	Malaysia	53.1	Romania	34.8
Portugal	83.2	Kenya	50.9	Cuba	34.4
Egypt	80.5	Argentina	50.3	Republic of Macedonia	34.2
Belize	80.0	Pakistan	49.9	Canada	34.0
Hungary	79.6	Tunisia	49.5	Taiwan	33.9
Germany	78.8	Turkey	48.1	South Africa	33.2
Nicaragua	78.0	Norway	47.7	Senegal	32.1
Dominica	78.0	Denmark	46.6	Syria	29.8
Israel	77.3	Aruba	46.3	Guatemala	29.6
United Kingdom	76.5	Latvia	46.2	Papua New Guinea	27.8
Malta	72.6	Finland	45.4	Indonesia	26.4
Austria	70.4	Colombia	44.8	Trinidad and Tobago	26.4
Netherlands	64.6	United Arab Emirates	44.6	Honduras	26.1
Spain	63.4	Costa Rica	42.4	Gabon	25.8
Côte d'Ivoire	63.3	Thailand	42.3	Algeria	25.7
Jordan	61.4	Dominican Republic	41.7	New Zealand	25.5
Cyprus	61.1	Mexico	41.5	Venezuela	25.5
Brazil	60.8	Serbia&Montenegro	41.5	Moldova	25.0
Mauritius	60.5	Slovakia	41.0	Zambia	24.1
Ghana	59.9	Mozambique	40.8	South Korea	23.7
Albania	59.3				

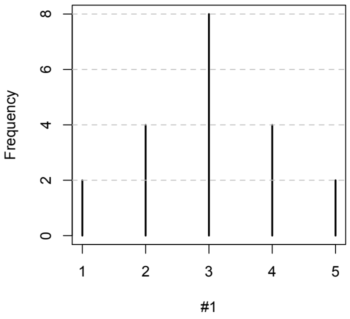
a) (2) Describe conceptually what would happen to the sample mean and sample median of the above dataset should South Korea's debt be replaced by 237.0. Conceptually, the mean is more sensitive to changes in the extremities of the sample while the sample median is more robust or less sensitive.

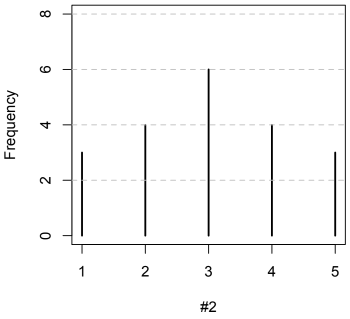
b) (2) Which countries correspond to the three quartiles of debts? Since there are 100 countries, the median is the average of the debts of the middle two countries (Pakistan and Argentina): (50.3+49.9)/2 = 50.1. The 1st quartile is the median of the 1st half of the data. Since there are 50 countries in the 1st half of the data, the median is the average of the debts of the middle two countries (Bosnia and Herzegovina and Yemen): (39.0+39.1)/2=39.05. If you consider only the first 49 observations, than the 1st quartile is 39.0 (Bosnia and Herzegovina). This is equally accepted as a correct answer. The 3rd quartile is the median of the 2nd half of the data. Since there are 50 countries in the 2nd half of the data, the median is the average of the debts of the middle two countries (Netherlands and Austria): (70.40+64.6)/2=67.5. If you consider only the last 49 observations, than the 3rd quartile is 70.4 (Austria). This is equally accepted as a correct answer.

c) (6) Draw the box-plot of the debts.



Question II (10): In the following graphs, the heights are frequencies. For example, there are 2+4+8+4+2=20 observations on the left.





a) (2) Obtain the two sample medians.

Since both distributions are symmetric, both medians (and means) are equal to 3.

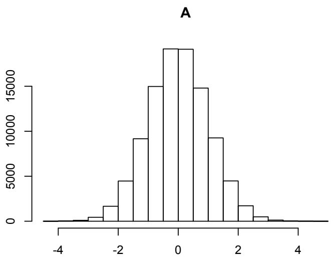
b) (4) Which distribution is more skewed? Explain.

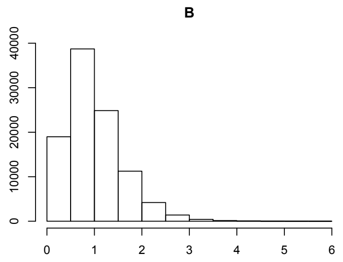
Both are symmetric.

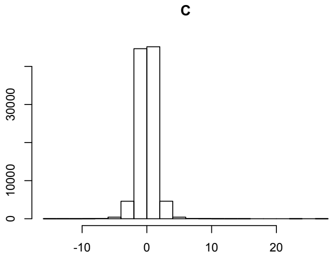
c) (4) Which of the two distributions exhibit more variability? Explain.

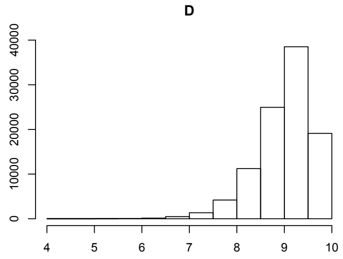
The one on the right is more variable since there are more observations away from the center. For instance, the graph on the right has 6 observations equal to 1 or 5, while the graph on the left has only 5 observations equal to 1 or 5.

Question III (10): Match the sample skewness and sample excess kurtosis. There are 4 histograms and 5 alternatives, therefore one of the alternatives must be wrong.









a) (2) [**D**] skewness = -1.150 and excess kurtosis = 1.930

b) (2) [**C**] skewness = -0.140 and excess kurtosis = 4.859

c) (2) [**B**] skewness = 1.123 and excess kurtosis = 1.781

d) (2) [**A**] skewness = -0.005 and excess kurtosis = -0.003

e) (2) [ ] skewness = 5.137 and excess kurtosis = 4.859

I will consider the pairs (a,c)=c(D,B) and (a,c)=c(B,D) equally correct since I told you in class not to worry about the exact definition of positively or negatively skewed distribution (only symmetric and asymmetric distributions).

Question IV (10): The following table shows the descriptive statistics from 1000 days of returns on IBM and Exxon's stock prices.

	Mean	Standard deviation
IBM	0.0009	0.0157
Exxon	0.0018	0.0224

Here is the covariance matrix

	IBM	Exxon
IBM	0.000247	
Exxon	0.000068	0.00050

a) (2) What is the variance of Exxon?
V(IBM) = 0.000247 (from the covariance matrix) or V(IBM) = (0.0157)**2 = 0.00024649 (square of standard deviation)
b) (4) What is the correlation between IBM and Exxon?
Corr(IBM,Exxon) = 0.000068/(0.0157*0.0224) = 0.1933576 or Corr(IBM,Exxon) = 0.000068/sqrt(0.000247*0.00050) = 0.1934975
c) (4) Consider a portolio that invests 30% in IBM and 70% in Exxon. What are the mean and variance of the portfolio?
P = 0.3IBM + 0.7Exxon
Mean P = 0.3*0.0009+0.7*0.0018 = 0.00153.
Var P = 0.09*0.000247+0.49*0.00050+2*0.3*0.7*0.000068 = 0.00029579.
Do you prefer this portfolio to just investing in IBM on its own?
If the goal is minimizing the variance, then one would prefer IBM. However, if the goal is maximizing the mean, then one would prefer the portfolio.

Question V (10): The following table contains data taken from American Cancer Society's Cancer Prevention Study II (CPS II). CPS II used survey results from about 1.2 million volunteers to find out more about what factors can cause cancer or help prevent it. Statistics from CPS II are used in the United States Surgeon General's Report on the Health Consequences of Smoking and in other government reports. The data are taken from the years 2000 to 2004. The table describes *only* the effects of tobacco use on deaths from cancer. They do not include deaths from other tobacco-related causes such as heart and lung diseases, though there are many deaths caused by those illnesses.

X	Cancer type	Men (Y=1)	Women (Y=2)
1	Lip, oral, cavity, pharynx	0.0282	0.0138
2	Esophagus	0.0536	0.0160
3	Larynx	0.0166	0.0044
4	Trachea, lung, bronchus	0.4976	0.3698

a) (2) Write down the marginal distribution of X and the marginal distribution of Y.

X	1	2	3	4	Y	1	2
Pr(X)	0.0420	0.0696	0.0210	0.8674	Pr(Y)	0.596	0.404

b) (4) Obtain the conditional distribution of Y given that X=1. In plain English, what does this distribution represent?

Y	1	2
Pr(Y|X=1)	0.6714286	0.3285714

Men are twice as likely as women to present lip, oral, cavity and pharynx cancer.

c) (4) You were given the information that a given patient has esophagus cancer. What is the probability the patient is a man? Formulate this question probabilistically and then derive its answer.

Pr(Y=1|X=2) = Pr(Y=1,X=2)/Pr(X=2)= 0.0536/0.0696= 0.7701149 or roughly 78% chance.

Question VI (10): Suppose we have returns data for assets X and Y. Let P1, P2 and P3 be portfolios with the following allocation structures: P1 = 0.3X + 0.7Y; P2 = 0.6X + 0.4Y; and P3 = 0.9X + 0.1Y. The sample means of X and Y are 0.36 and 0.53, respectively. The sample standard deviations of X and Y are 0.67 and 0.98, respectively. Also, the sample correlation between X and Y is -0.76.

a) (4) Which one of the above 3 portfolios maximizes the sample mean? Provide your statistical derivations to support your answer.

Mean P1 = 0.3*0.36 + 0.7*0.53 = 0.479 <=
Mean P2 = 0.6*0.36 + 0.4*0.53 = 0.428
Mean P3 = 0.9*0.36 + 0.1*0.53 = 0.377
Therefore, P1 is the portfolio that maximizes the sample mean.

b) (6) Which one of the above 3 portfolios minimizes the sample variance? Provide your statistical derivations to support your answer.The sample variances of X and Y are 0.4489 and 0.9604, respectively.
The sample covariance between X and Y is -0.499016.
Variance P1 = 0.09*0.4489+0.49*0.9604-2*0.3*0.7*0.499106 = 0.3014
Variance P2 = 0.36*0.4489+0.16*0.9604-2*0.6*0.4*0.499106 = 0.0757 <=
Variance P3 = 0.81*0.4489+0.01*0.9604-2*0.9*0.1*0.499106 = 0.2834
Therefore, P2 is the portfolio minimizes the sample variance.

Question VII (10): The quality of Nvidia's graphic chips is such that the probability that a randomly chosen chip being defective is only 0.1%. You have invented a new technology for testing whether a given chip is defective or not. This test will always identify a defective chip as defective and only "falsely" identify a good chip as defective with probability 1%.

a) (4) Given that the test identifies a defective chip, what is the probability that it is actually defective?
Let us denote D=1 if the chip is defective and D=0 if the chip is not defective.
Similarly, T=1 if the test identifies the chip as defective and T=0 otherwise.
Therefore, we know that Pr(D=1)=0.001. We also know that Pr(T=1|D=1)=1.00 and Pr(T=1|D=0)=0.01.
These information set is enough for the construction of the 2x2 table:

		D		
		0	1	Pr(T)
T	0	0.98901	0.000	0.98901
	1	0.00999	0.001	0.01099
Pr(D)		0.99900	0.001	1.00000

We are interested in computing Pr(D=1|T=1):
Pr(D=1|T=1) = Pr(D=1,T=1)/Pr(T=1) = 0.001/0.01099 = 0.09099181.

b) (4) What percentage of the chips will the new technology identify as being defective?
Pr(T=1) = 0.01099 (or roughly 1% of the chips)

c) (2) Should you advise Nvidia to go ahead and implement your testing device? Explain.
Not really since about 90% of the time a defective is mistakenly signaled by test, despite the fact that 100% of the time non-defective is correctly signaled by the test.