The University of Chicago Booth School of Business
Business Statistics, 41000-81/82, Spring 2011
Instructor: Hedibert F. Lopes
Midterm Exam

Name ____

You may use a calculator.
You may use a one-page-front-back "cheat sheet".
You have 100 minutes to finish the exam.

I pledge my honor that I have not violated the Honor Code during this examination.

Signature ____________________________________

Question	Points	Total
I	10	
II	10	
III	10	
IV	10	
V	10	
VI	10	
VII	10	
VIII	10	
Total	80	

Question I (10): The following table shows the 40 largest GDP in billions of US dollars for 2008, excluding the US.

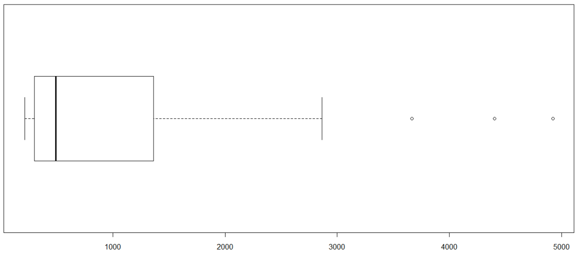
Japan	4,923.76	Sweden	484.55
China	4,401.61	Saudi Arabia	481.63
Germany	3,667.51	Norway	456.23
France	2,865.74	Austria	415.32
UK	2,674.09	Taiwan	392.55
Italy	2,313.89	Greece	357.55
Russia	1,676.59	Iran	344.82
Spain	1,611.77	Denmark	342.93
Brazil	1,572.84	Argentina	326.47
Canada	1,510.96	Venezuela	319.44
India	1,209.69	South Africa	277.19
Mexico	1,088.13	Finland	274.00
Australia	1,010.70	Ireland	273.33
Korea	947.01	Thailand	273.25
Netherlands	868.94	UAE	260.14
Turkey	729.44	Portugal	244.49
Poland	525.74	Colombia	240.65
Indonesia	511.77	Malaysia	222.22
Belgium	506.39	Czech Republic	217.08
Switzerland	492.60	Hong Kong	215.56

a)(2) Describe conceptually what would happen to the sample mean and sample median of the above dataset should the GDP of the US be included (US GDP is around 14 trillion dollars)? Explain the difference.

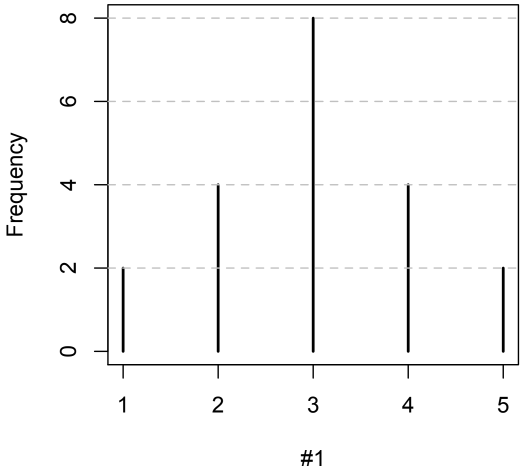
The sample mean would increase because the new point (US) is greater than the current sample mean. The sample median would increase. The old sample median is the average of Switzerland and Sweden. The new sample median is Switzerland.

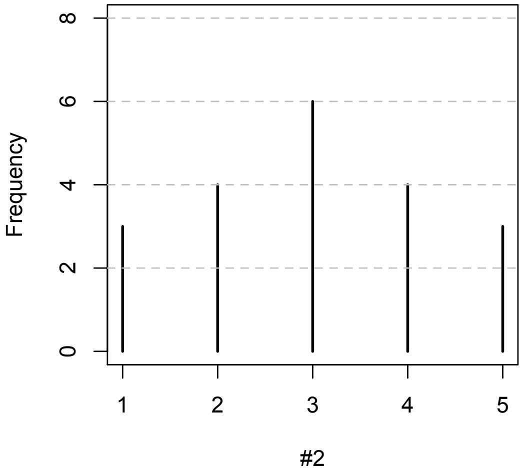
b)(8) Draw the box-plot of the GDPs.

Q_1 = Venezuela 319.44 (#30); Q_2 = 488.575, average of Switzerland 492.60 (#20) and Sweden 484.55 (#21); Q_3 = Canada 1510.96 (#10); IQR = 1191.52. Low outliers: Countries < Q_1 - 1.5 IQR. This threshold is negative, and no countries in the sample have GDP < 0. High outliers: Countries > Q_3 + 1.5 IQR, or above 3298.24. High outliers are Japan, China, Germany.



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) Compute the two sample means.

3 and 3. Visually, both distributions are symmetric around 3. Numerically:
Sample 1 mean = (2 x 1 + 4 x 2 + 8 x 3 + 4 x 4 + 2 x 5) / 20 = 60 / 20 = 3
Sample 2 mean = (3 x 1 + 4 x 2 + 6 x 3 + 4 x 4 + 3 x 5) / 20 = 60 / 20 = 3

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

Neither. Both distributions are symmetric.

c)(4) Does the data from dotplot #1 have larger variance than the data from dotplot #2? Explain.

No, variance of Sample 2 is larger than variance of Sample 1. Visually, both samples are symmetric and have the same range, but Sample 1 has more weight in the middle and Sample2 has more weight at the upper and lower ends. Sample 2 is "flatter".

Question III (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)	P(X)
1	Lip, oral, cavity, pharynx	0.0282	0.0138	**0.0420**
2	Esophagus	0.0536	0.0160	**0.0696**
3	Larynx	0.0166	0.0044	**0.0210**
4	Trachea, lung, bronchus	0.4976	0.3698	**0.8674**
P(Y)		**0.5960**	**0.4040**	

a)(2) Write down the marginal distribution of X and the marginal distribution of Y. You can use the margins of the above table.

See table above.

b)(4) Suppose you were given the information that a person has trachea, lung or bronchus cancer. Is it more likely that the person is a man or a woman? Explain.

More likely a man because $P(Y = 1 \mid X = 4) > \frac{1}{2}$.
$P(Y = 1 \mid X = 4) = P(Y = 1 \,\&\, X = 4) / P(X = 4) = 0.4976 / 0.8674 = 0.5737$

b)(4) Explain conceptually the difference between $P(X=4|Y=1)$ and $P(Y=1|X=4)$.

The differences are in the outcome and the conditioning information. $P(X = 4 \mid Y = 1)$ is the probability that a randomly chosen man has trachea, lung or bronchus cancer. $P(Y = 1 \mid X = 4)$ is the probability that a randomly chosen trachea, lung or bronchus cancer individual is a man.

Question IV (10): Suppose we have returns data for assets X and Y. Let P1, P2 and P3 be portfolios with the following allocation structures:

$$\begin{aligned} P1 &= 0.3X + 0.7Y \\ P2 &= 0.6X + 0.4Y \\ P3 &= 0.9X + 0.1Y \end{aligned}$$

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 3 portfolios minimizes the sample variance? Explain.

P2 minimizes the sample variance.

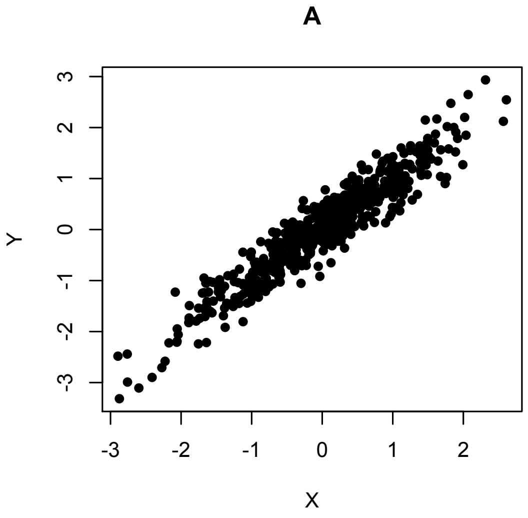
Cov(X, Y) = -0.76 * 0.67 * 0.98 = 0.499
Var(P1) = 0.09 Var(X) + 0.49 Var(Y) + 2 * 0.21 Cov(X, Y) = 0.3014
Var(P2) = 0.36 Var(X) + 0.16 Var(Y) + 2 * 0.24 Cov(X, Y) = 0.0757
Var(P3) = 0.81 Var(X) + 0.01 Var(Y) + 2 * 0.09 Cov(X, Y) = 0.2834

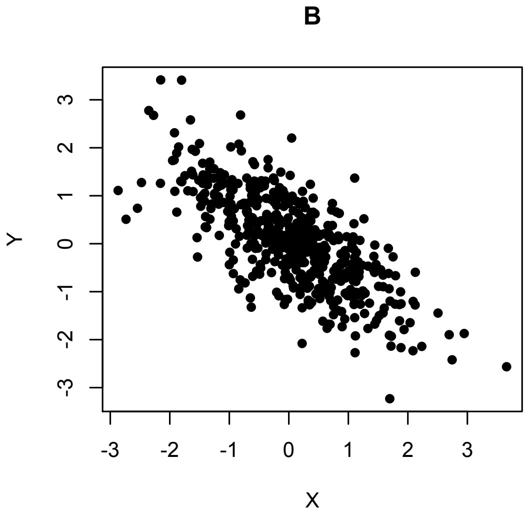
b)(6) Is it riskier to invest in P4=0.5P1+0.5P2 or in P5=0.5P1+0.5P3? Explain.

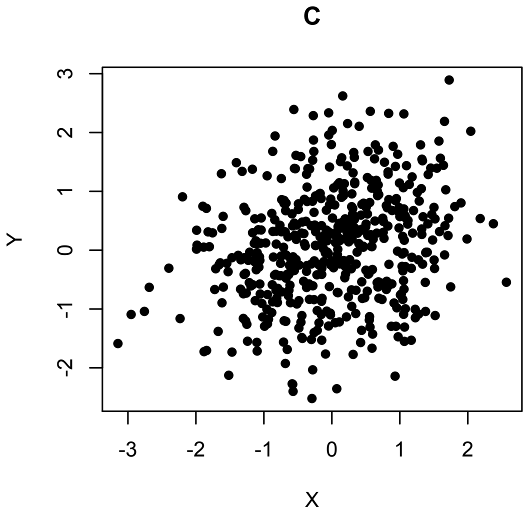
Riskier to invest in P4. Note also that P5 = P2.

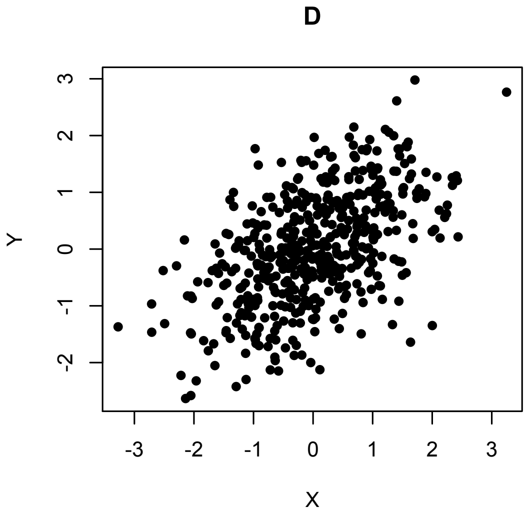
Var(P4) = 0.2025 Var(X) + 0.3025 Var(Y) + 2 * 0.2475 Cov(X, Y) = 0.1344
Var(P5) = Var(P2) = 0.0757

Question V (10): Scatter plots.









Match the sample correlations listed below to the above scatter plots. Notice that once correlation will not match.

a) (2) corr(x,y) = 0.50 [**D**]

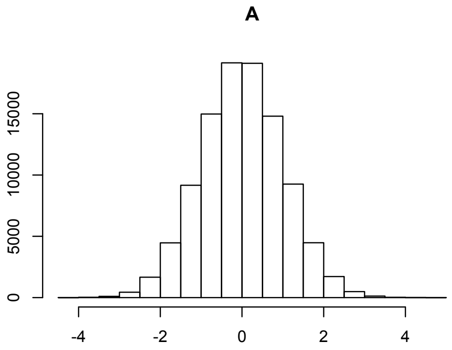
b) (2) corr(x,y) = -0.95 [**no matching plot**]

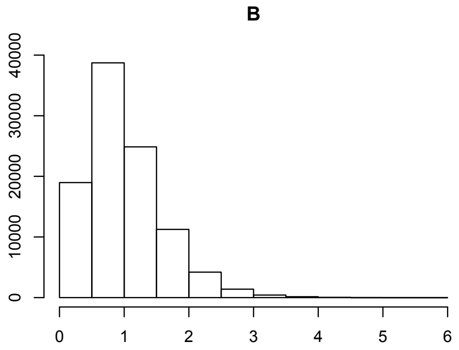
c) (2) corr(x,y) = 0.95 [**A**]

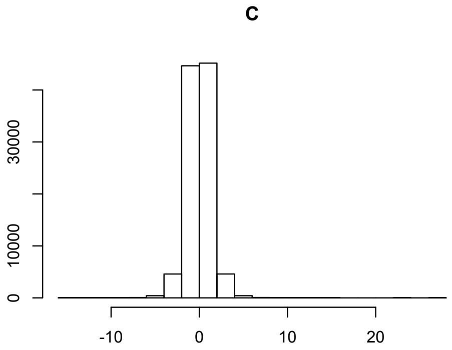
d) (2) corr(x,y) = -0.75 [**B**]

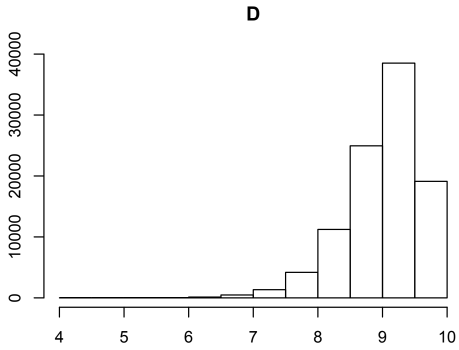
e) (2) corr(x,y) = 0.25 [**C**]

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









a) (2) [**no matching plot**]
skewness = 5.137
excess kurtosis = 4.859

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

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

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

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

Question VII (10): The following table data from a October 2008 GALLUP poll on the presidential campaign. Let X the marital status, Y be gender and Z be candidate.

MARITAL STATUS	GENDER	OBAMA	MCCAIN	OTHERS
MARRIED	MAN	0.0428	0.0520	0.0071
	WOMAN	0.0461	0.0333	0.0186
UNMARRIED	MAN	0.2366	0.1387	0.0326
	WOMAN	0.2470	0.1137	0.0314

a) (2) Derive the joint distribution of X and Z, i.e. P(X,Z)

	Obama	**McCain**	**Others**
Married	**0.0889**	**0.0853**	**0.0257**
Unmarried	**0.4836**	**0.2524**	**0.0640**

Note: Due to rounding, the joint probability distribution does not add up to 1.

b) (4) Compute P(Z=Obama | Y=woman) to P(Z=Obama | Y=man). Comment.

Obama's support is stronger among women than among men.

P(Z=Obama | Y=woman) = P(Z=Obama, Y=woman) / P(Y = woman) = (0.0461 + 0.2470) / (0.0461 + 0.2470 + 0.0333 + 0.1137 + 0.0186 + 0.0314) = 0.5980

P(Z=Obama | Y=man) = P(Z=Obama, Y=man) / P(Y = man) = (0.0428 + 0.2366) / (0.0428 + 0.2366 + 0.0520 + 0.1387 + 0.0071 + 0.0326) = 0.5481

c) (4) Compare P(Z=Obama|Y=man), P(Z=Obama|X=married,Y=man) and P(Z=Obama|X=unmarried,Y=man). Comment.

Obama's support is stronger among unmarried men than among married men.
Obama's support is stronger among unmarried men than among all men.

P(Z=Obama | Y=man) = 0.5481

P(Z=Obama|X=married,Y=man) = P(Z=Obama,X=married,Y=man) / P(X=married,Y=man) = 0.0428 / (0.0428 + 0.0520 + 0.0071) = 0.4200

P(Z=Obama|X=unmarried,Y=man) = P(Z=Obama,X=unmarried,Y=man) / P(X=unmarried,Y=man) = 0.2366 / (0.2366 + 0.1387 + 0.0326) = 0.5800

Question VIII (10): The following cross-tabulation shows household income (in $1000 of dollars) by educational level of the head of the household.

	Under 50	50-75	Over 75	TOTAL
Not HS graduate	0.114	0.019	0.012	**0.145**
HS Graduate	0.161	0.068	0.072	
Some College	0.109	0.063	0.099	**0.271**
Bachelor's Degree or more	0.055	0.056	0.182	**0.293**
TOTAL		**0.206**		

a)(2) What is the probability of a head of household having "some college"?

P(Some college) = 0.109+0.063+0.099 = 0.271

b)(2) If the household head has bachelor's degree or more, what is the probability that his/her income is not over $75,000?

P(NOT Over 75 | Bachelor's Degree or more) = 1 - P(Over 75 | Bachelor's Degree or more) = 1 - P(Over 75, Bachelor's Degree or more) / P(Bachelor's Degree or more) = 1 - 0.182 / (0.055+0.056+0.182) = 0.3788

c)(2) Knowing that a household income is between $50,000 and $75,000, what is the probability that the household head has a bachelor's degree or more?

P(Bachelor's Degree or more | 50-75) = P(Bachelor's Degree or more, 50-75) / P(50-75) = 0.056 / (0.019+0.068+0.063+0.056) = 0.2718

d)(2) What is more likely: i) you make over $75,000 given that you have a bachelor's degree or more or ii) you make under $50,000 given that you are not a high school graduate? Explain probabilistically.

Option (ii) is more likely because P(Under 50 | Not HS graduate) > P(Over 75 | Bachelor's Degree or more).

P(Over 75 | Bachelor's Degree or more) = 1 - P(NOT Over 75 | Bachelor's Degree or more) = 0.6212

P(Under 50 | Not HS graduate) = P(Under 50, Not HS graduate) / P(Not HS graduate) = 0.114 / (0.114+0.019+0.012) = 0.7862

e)(2) Are the educational level and household income independent? Explain your answer

No. Conditional probability does not equal marginal probability. For example, P(50-75) = 0.206, but P(50-75 | Some College) = 0.063 / 0.271 = 0.2325.

Question IX (3): Please, provide a guess for your grade (not including this question, of course!). If your guess is within 3 points of your actual grade, you will be given up to 3 points as bonus. You have nothing to loose and we can use the data later on in class to correlate actual and guessed grades.

Guess your grade =