SAUDER School of Business | ROBERT H. LEE GRADUATE SCHOOL

STATISTICS Preparation Session

Self-Assessment Test (or Let's Play "Twenty Questions")

Prepared by Jonathan Berkowitz, 2013

Q1. For each of these six variables from a (hypothetical) survey of an MBA cohort, decide whether it would be more appropriately considered as a categorical or a measurement/quantitative variable or neither.

- Country of birth
- Date of birth
- Distance, in km., from home to work
- Type of residence/home
- Amount of money you have currently in your pocket, wallet, or purse
- Satisfaction with the admissions process, using a 0-10 scale

Q2. Suppose that data on the following three variables were collected during a final exam in a large undergraduate Statistics course.

- Scores on the exam (0 to 100%)
- Time to complete the final exam, if there were no time limit and you could stay as long as you wanted!
- The kind of headache (migraine or cluster or tension) developed by the students

For each variable, decide which would be:
a) ... the single most appropriate numerical summary from these choices: proportion, mean or median.
b) ... an appropriate graphical summary.

Q3. Six items about computing numerical summaries...
a) A study was made of the age of entering first-year university students. Which of the following is most likely to be the standard deviation? Explain why, in one sentence only.

A. 1 month B. 1 year C. 5 years

b) The mean age of five people in a room is 30 years. One of the people, whose age is 50 years, leaves the room. What is the mean age of the remaining four people in the room?

A. 40 years B. 30 years C. 25 years D. Not enough information to determine

c) The median age of five people in a room is 30 years. One of the people, whose age is 50 years, leaves the room. What is the median age of the remaining four people in the room?

A. 40 years B. 30 years C. 25 years D. Not enough information to determine

d) A survey of a sample of 400 hospitals showed that the number of hospital beds ranged from 150 to 750. A histogram showed that number of beds is approximately bell-shaped. A good estimate of the standard deviation of the number of beds is:

A. 2.5 B. 50 C. 100 D. 600 E. Not enough information to determine

e) Which of the following is least affected by extreme outliers?

A. Range B. Third quartile C. Mean D. Variance E. All are equally affected

f) A person cycles to a point 10 kilometres away at 20 kph and returns over the same distance at 60 kph. What is his average speed for the round trip?

Q4. The salaries (in $) of secondary school classroom teachers in the United States ten years ago gave the following descriptive statistics:

Min = 31,200 ; Q1 = 37,400 ; Median = 40,000 ;Q3 = 48,400 ; Max = 57,200

a) Find the range and interquartile range.
b) Are there any outliers, as defined by inner fences? Explain.
c) Predict the direction of skewness for this distribution. Explain.

Q5. Suppose each salary in Q4 was increased by $3000 for the following year. For each of the four statistics in the left-hand column, write the **number** of the phrase in the right-hand column that states how each statistic would change for this sample of employees.

- Median ___
- IQR ___
- Range ___
- Standard Deviation ___

1. Will remain unchanged
2. Will increase by $3000
3. Will be multiplied by $3000
4. Will increase by $\sqrt{\$3000}$

Q6. A brokerage firm gathered information on how their clients were investing for retirement. Based on age, clients were categorized according to where the largest percentage of their retirement portfolio was invested and shown in the table below.

	Age 50 or Younger	**Over Age 50**	**Total**
Mutual Funds	30	34	**64**
Stocks	37	45	**82**
Bonds	19	23	**42**
Total	86	102	**188**

a) What percentage of clients are over age 50 and invest in mutual funds?
b) Of the clients over age 50, what percentage invest in mutual funds?
c) Of the clients who invest in mutual funds, what percentage is over age 50?
d) What percentage of clients is over age 50?
e) Does it appear that mode of investment is independent of age?

Q7. At a well-known business school the grade point averages (GPA) of its 1000 undergraduates are normally distributed with mean 2.84 and standard deviation 0.40.
a) What percentage of the undergraduates have GPAs below 2.00 (i.e. “on probation”)?
b) What GPA will be exceeded by only 20% of the student body?
c) Compute the lower and upper quartiles, and the interquartile range for this distribution.

Q8. A company has two manufacturing plants, one that uses low-tech machines and another that uses high-tech machines. From recent history, the number of defects per week observed at each plant is normally distributed with the following parameters.

Low-tech:	Mean = 15	SD = 3
High-tech:	Mean = 10	SD = 1

Last week, the low-tech plant produced ten defects, while the high-tech plant produced eight defects. Which plant – low-tech or high-tech – performed better relative to past performance? Explain why.

Q9. A fast food restaurant always asks their customers whether they want pickle slices on their burgers, and if so, how many. The distribution of the number of slices a customer requests has a mean of 1.3 and a standard deviation of 0.5.
a) What is the distribution of the total daily demand for pickle slices assuming 400 customers per day? Give the name of the distribution, the mean, the variance and the standard deviation of the total daily demand.
b) To satisfy the daily pickle demand of 400 customers with a probability of 0.99, how many total pickle slices would be needed?

Q10. An important part of the customer service responsibilities of a telephone company relates to the speed with which troubles in residential service can be repaired. Suppose that past data indicate that there is a probability of 0.70 that service troubles can be repaired on the same day they are reported.
a) Suppose the company receives 100 trouble calls on a particular day. What is the approximate chance that 80% or more will receive same-day repairs.
b) Suppose it is also known that the repair time for a trouble call has a mean of 480 minutes and a standard deviation of 250 minutes. A random sample of 400 trouble calls was taken and the repair times recorded. Compute the probability that the mean of the 400 repair times is less than 500 minutes.

Q11. An established clothing retailer is interested in customer response to a proposed new logo. A survey randomly samples 100 customers; 55 of them say they would prefer the new logo to the previous one. However, the retailer will only change its logo if it is convinced that the newly designed logo is preferred by the majority (i.e. more than half) of its customers.
a) Construct a 95% confidence interval for the true proportion of the customers who prefer the new logo over the previous one.
b) How large a sample *n* would you need to estimate *p*, the proportion of people who prefer the newly designed logo over the previous one, with margin of error 0.05 with 99% confidence? Use the guess $\hat{p}$ = 0.5 as the value for *p*.

Q12. You are the new Operations Manager of the local public transportation company and are especially interested in the reliability of bus service. You plan, on a monthly basis, to take a random sample of major bus stops and observe whether the buses depart on time or late and how late they are. (Buses never leave early since, if they arrive early, they wait until their departure will be exactly on time.)
a) The first month, you gather a random sample of 121 bus departures from a variety of times of day, days of the week, routes and locations. The sample has an average lateness of departure of 6.4 minutes with a standard deviation of 1.8 minutes. Construct a 95% confidence interval for the average lateness of departures for the entire bus system this month.
b) Five years ago, the system-wide mean lateness of departure was known to be 6.8 minutes. Using a 5% level of significance and the sample results of part a), carry out a hypothesis test to decide whether the system is improving; that is, whether the mean lateness has decreased from five years ago. In your solution, provide: null and alternative hypotheses, test statistic, approx. P-value, and conclusion. The conclusion should be one clearly worded sentence that the bus company management can understand.

Q13.

a) What is the correlation coefficient for the following three points in the X-Y plane? (STOP AND THINK!)

X	1	3	5
Y	4	3	2

b) An American study found that the correlation between two-year-old children's heights (measured in inches) and their weights (measured in pounds) was 0.46. What would the correlation coefficient be if you converted their heights to centimetres and weights to kilograms? (One inch = 2.54 cm and 1 pound = 0.454 kg.)

c) An economist studied salaries of 321 bank employees with five or less years of employment in a national bank. He found that the relationship between years of service and salary was linear and that the regression equation predicting salary (in thousands of dollars) was: Salary = 21.5 + 3.1 * Years.

He concludes that employees with 10 years of service should make an average salary of $52,500. Is his conclusion correct? If not, say why.

d) In part c) the economist has used the regression equation to make a prediction. Which of these numbers best measures the precision of this prediction?

A. The slope of the line
B. The standard deviation of y
C. The standard deviation of x
D. The square of the correlation coefficient
E. The ratio of the two standard deviations

e) An investigator measuring various characteristics of a large group of athletes found that the correlation coefficient between the weight of the athlete and the weight that the athlete could lift was r = 0.60. Determine whether each statement is true or false.

(i) If an athlete gains 5 kg, he/she will be able to lift an additional 3 kg.
(ii) The more an athlete can lift, on the average the more that athlete weighs.
(iii) 36 per cent of the athlete's lifting ability can be attributed to his or her weight alone.
(iv) 60 per cent of the athlete's lifting ability can be attributed to his or her weight alone.

Q14. An expert consultant in hospital resource planning states that the number of open beds that a hospital can use effectively should be estimated by the number of FTEs (full-time equivalent employees) on staff. The consultant collected data on the number of open beds and number of FTEs for 12 hospitals, and computed the means and SDs as follows:

Number of open beds:	Mean = 50	SD = 20
Number of FTEs:	Mean = 140	SD = 40

She computed the least squares regression equation and found that for a hospital with 100 FTEs, the estimated number of open beds was 32.

a) Use this information to compute the value of the correlation coefficient.
b) What is the least squares regression equation she found?
c) From the available data, what would you predict the number of open beds to be for a hospital with an unknown number of FTEs?
d) What fraction of the variation in number of open beds is explained by the number of FTEs?
e) Another expert consultant, this one in hospital administration, claims that the regression was done the wrong way around, and that the number of FTEs required in a hospital should be estimated from the number of open beds in the hospital. What would the value of the correlation coefficient be if the analysis were done this way?

Q15. The human resources department of a pulp and paper company is concerned about growing absenteeism, measured by number of sick days, among shift workers at the mill. The HR Analyst randomly selects 15 long-time employees and 15 relatively new employees (one from each of the 15 union job types at the mill), for a total of 30 employees, and finds the presented below.

	Long-time Employees	**New Employees**
Mean	5.1	3.7
Standard Deviation	2.9	2.2

Is there evidence of a difference between long-time employees and new employees with respect to their absenteeism in 2007? Carry out a hypothesis test at 5% level of significance.

Q16. Which of the following situations would probably best be handled by a paired *t*-procedure?
A. Comparing average salaries of accountants versus real estate agents.
B. Exploring the differences between average vocabularies of French vs. Spanish speakers.
C. Comparing average grades in an accounting class and a real estate class for a sample of students who took both classes.
D. Exploring the difference in language proficiency based on a group of executives before and after a three-week intensive French class.
E. Comparing experimental results when the Central Limit Theorem can be approximated by a *t*-distribution.

Q17. The following data come from a recent study of the punctuality (i.e. on-time arrival) of public transportation in a European city. The study's aim was to test whether the type of transportation was related to punctuality.

	Exactly on time	<1 minute late	1-5 minutes late	>5 minutes late
Bus	27	35	33	25
Train	13	15	27	25

a) What proportion of all public transportation was on time or less than 1 minute late?
b) What proportion of the buses were more than 5 minutes late?
c) Test the null hypothesis that there is no association between punctuality and type of public transportation.

Q18. Consider the following 2x2 table of frequencies with one cell frequency missing. What should that missing frequency be in order to get a chi-square statistic equal to zero?

4	12
10	

Q19. As more women become owners of small businesses, there is some question about how they are treated by banks. To answer questions about possible bank discrimination against women business owners, researchers surveyed over one thousand male and female business owners. The survey asked questions of both men and women business owners who applied for loans during the previous month. It determined the nature of the business, its size, and its age, and asked about owners' experiences in dealing with banks. Here are some of the variables for which data were collected (they are labeled V1 through V6, for convenience). Assume V3, V4 and V6 are normally distributed.

- V1 Gender of the business-owner (1=female, 2=male)
- V2 Form of business (1=proprietorship, 2=partnership, 3=corporation)
- V3 Annual gross sales (thousands of dollars)
- V4 Age of the business (years)
- V5 Was the loan approved? (1=no, 2=yes)
- V6 If the loan was approved, what interest rate did you get (percentage scale)

For each of the following research questions, which of the standard techniques of statistical inference is the most appropriate to address the question. Here is a list of techniques to choose from. Note that some may be used more than once, others not at all. Write the number of the <u>most appropriate</u> technique in the space to the left of the research question.

1 = One-sample z-test of a proportion
2 = Two-sample z-test of two proportions
3 = One-sample t-test of a mean
4 = Two-sample t-test of two independent means
5 = Matched pairs t-test of two dependent means
6 = Chi-square test of independence
7 = Simple linear regression
8 = Multiple regression
9 = One-way analysis of variance

__ a) Do male-owned and female-owned (V1) businesses have different annual gross sales (V3)?
__ b) Is there a difference in age of the business (V4) for male-owned versus female-owned (V1) businesses?
__ c) Is the rate of loan approval (V5) the same for male business owners as for female business owners (V1)? (Two answers are possible here.)
__ d) Do all three forms of business (V2) who have loans approved receive the same mean interest rate (V6)?
__ e) Is the percentage of women-owned businesses (V1) in this survey different from the 10% rate found in nation-wide censuses?
__ f) Are the forms of business (V2) owned by women different from the forms owned by men (V1)?
__ g) Is interest rate (V6) related to gross annual sales (V3)?

Q20. Write a clear, grammatically correct, statistically accurate, and jargon-free explanation of a P-value. Use one or two sentences and language that a layperson could understand.

END OF QUESTIONS
SOLUTIONS FOLLOW

SAUDER School of Business | ROBERT H. LEE GRADUATE SCHOOL

STATISTICS Preparation Session

Self-Assessment Test (or Let's Play "Twenty Questions")

Prepared by Jonathan Berkowitz, 2013

SOLUTIONS

Do the self-assessment questions BEFORE reviewing the solutions.
A grading scheme is given at the end.

SOLUTION Q1.
Categorical, Neither, Measurement, Categorical, Measurement, Measurement

Comment: Identifying the type of data/variable is the first, and perhaps most important, principle of data analysis. Note that date of birth can be used to compute age, but cannot be usefully analyzed by itself.

SOLUTION Q2.
a) Mean; Median; Proportion
b) Histogram or stem-and-leaf display (or boxplot); histogram or stem-and-leaf display (or boxplot); bar chart. (Other choices of graphs are possible.)

Comment: The choice numerical and graphical summary depends on the type of data and the distribution.

SOLUTION Q3.
a) B. 1 year. The typical distance from the mean age is likely to be about one year.
OR: Most students will be about the same age, give or take a year.
b) C. 25 years (Reason: [5(30)–50]/4 = 25)
c) D. Cannot be determined
d) C. 100 (Reason: [750-150]/6 = 100)
e) B. Third quartile.
f) 3 kph (It takes 30 min. at 20 kph and 10 min. at 60 kph. Total distance = 20 km; total time = 40 min. =2/3 hr. Therefore: Average speed = 20 km divided by 2/3 hr = 30 kph)

Comment: Basic concepts of numerical summaries, especially the standard deviation.

SOLUTION Q4.
a) Range = $26,000 (i.e. 57,200 – 31,200)
Interquartile range = $11,000 (i.e. 48,400 – 37,400)
b) No: Q1–1.5×IQR = 20,900; Q3+1.5×IQR = 64,900
There are no values outside this range, so there are no outliers.
c) The distribution is skewed to the right (i.e. long right-hand tail) since the distance between the median and Q3, and between Q3 and the maximum, are greater than the distance between the median and Q1, and between Q1 and the minimum.

Comment: Computation of numerical summaries for all distributions, not just symmetric ones.

SOLUTION Q5.
Median: 2, IQR: 1, Range: 1, SD: 1

Comment: Effect of coding/changing units on numerical summaries

SOLUTION Q6.
a) 18.1% (34/188)
b) 33.3% (34/102)
c) 53.1% (34/64)
d) 54.3% (102/188)
e) Yes: The age distribution (ratio of younger to older) is about the same for each mode (i.e. type) of investment.

Comment: Marginal and conditional distributions of a two-way table for categorical data.

SOLUTION Q7.
a) Pr (X < 2.00) = Pr (Z < [2.00–2.84]/0.40) = Pr (Z < –2.10) = 0.0179 or 17.9%.
b) Z = 0.84; X = 2.84 + 0.84(0.40) = 3.18 (or 3.176)
c) Q1 = 2.57; Q3 = 3.11; IQR = 0.54
Q1 for Z = –0.675; X = 2.84 + (–0.675)(0.40) = 2.57
Q3 for Z = 0.675; X = 2.84 + (0.675)(0.40) = 3.11
IQR = 3.11 – 2.57 = 0.54

Comment: Calculations using the normal distribution.

SOLUTION Q8.
Low-tech z-score = (10–15)/3 = -1.67
High-tech z-score = (8–10)/1 = -2.00
The high-tech plant is more unlikely since it has a more negative z-score.

Comment: Use of Z-scores for comparing normal distributions.

SOLUTION Q9.
a) Normal with Mean = 520 [= 400x1.3], Variance = 100 [= $400(0.5)^2$], SD = 10 [=√100]
b) 544 total slices [z=2.33 has a right-tail prob. of 0.01; x = 520 + 2.33(10) = 543.3; round up]

Comment: Combining random variables, sampling distribution of a sum, normal calculation.

SOLUTION Q10.
a) Pr ($\hat{p}$ > 0.80) = Pr (Z > [0.80–0.70] / √[(0.70)(0.30)/100] = Pr (Z > 2.18) = <u>0.0145</u> or 1.45%
b) Pr ($\bar{x}$ < 500) = Pr (Z < [500–480] / [250/√400]) = Pr (Z < 1.60) = <u>0.945</u> or 94.5%

Comment: Central Limit Theorem and probability calculations using the sampling distributions of $\hat{p}$ and $\bar{x}$.

SOLUTION Q11.
a) $\hat{p}$=55/100 = 0.55; 95% CI: 0.55 ± 1.96√[(0.55)(0.45)/100] = 0.55 ± 1.96(0.050) = 0.55 ± 0.098
It can also be written in any of these forms: (0.452,0.648) or (45.2%,64.8%) or (45%,65%).
b) $n = (2.576^2)(0.5)(0.5)/(0.05^2) = 664$

Comment: Confidence interval and sample size calculation for one proportion.

SOLUTION Q12.

a) t^*_{120} = 1.980: CI = 6.4 ± 1.980(1.8/√121) = 6.4 ± 0.324

6.4 $\pm$ 0.324

b) H_0: μ = 6.8; H_a: μ < 6.8

t = $\frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{6.4-6.8}{1.8/\sqrt{121}}$ = -2.44

0.005 < P-value < 0.01 (from t-distribution table); reject H_0 at the 1% significance level

There is strong evidence to say that the system is improving (or that mean lateness has decreased).

Comment: One-sample confidence interval and hypothesis test for a mean.

SOLUTION Q13.

a) Perfect negative correlation: r = –1 (Plot the data points; they fall on a straight line.)

b) r = 0.46, unchanged (Correlation is invariant to the measurement scales.)

c) No – prediction at 10 years requires extrapolation beyond the range of data

d) D. The square of the correlation coefficient.

e) False, True, True, False

Comment: Basic concepts of correlation and least squares regression.

SOLUTION Q14.

a) There are multiple ways to solve this:

Method 1: 100 is 1 SD below average in the X-variable; this equals 32, which is r SDs below average in the Y-variable. So 32 = 50–r(20). Hence r = 0.9.

Method 2: Since $\bar{y} = b_0 + b_1\bar{x}$, then $50 = b_0 + b_1(140)$; and since $\hat{y} = b_0 + b_1 x$, then $32 = b_0 + b_1(100)$

Solve for b_1 by taking the first equation minus the second equation: b_1 = 18/40 = 0.45

Since b_1 = r (S_y/S_x) then 0.45 = r (20/40). Hence r = 0.9.

b) $\hat{y} = -13 + 0.45x$ [b = r (S_y/S_x) = 0.90(20/40) = 0.45; a = $\bar{y} - b\bar{x}$ = 50 – 0.45(140) = -13]

c) If X is unknown, predict Y to be the mean of y, which is 50.

d) r^2 = 0.81

e) 0.9, unchanged.

Comment: Use of least squares regression formulas for slope and intercept.

SOLUTION Q15.

Use a two-sample t-test

H_0: $\mu_1 - \mu_2 = 0$; H_a: $\mu_1 - \mu_2 \neq 0$ (or H_0: $\mu_1 = \mu_2$; H_a: $\mu_1 \neq \mu_2$)

Pooled variance version: $s_p^2 = [14(2.9)^2+14(2.2)^2]/28 = 6.625$; $s_p = 2.574$

Test stat: t = (5.1–3.7) / (2.574√[1/15 +1/15]) = 1.49

OR

Unpooled variance version:

Test stat: t = (5.1–3.7) / √[2.9^2/15+2.2^2/15] = 1.49

If pooled t: P-value = 2×Pr(t(28 df) > 0.950) > 0.20 (see Table T, two-tail prob.)

If unpooled t: P-value = 2×Pr(t(14 df) > 0.950) > 0.20 (see Table T, two-tail prob.)

For either version, there is no evidence of a difference between groups of employees with respect to mean number of sick days.

Comment: Carry out a two-sample t-test of two independent means from summary statistics.

SOLUTION Q16.
C and D

Comment: Identify situations requiring a matched pairs t-test.

SOLUTION Q17.
a) 90/200 =0.45
b) 25/120 = 0.208
c) H_0: Punctuality and type of public transportation are independent (i.e. unrelated)
H_a: Punctuality and type of public transportation are dependent.
Chi-square test statistic = 5.729, with 3 degrees of freedom
Since Chi-square(0.10,3 df) = 6.25, P-value > 0.10
Therefore, do not reject H_0 at the 0.10 significance level. There is no evidence that punctuality is related to type of public transportation.

Comment: Analysis of a cross-tabulation using the chi-square test of independence.

SOLUTION Q18. Missing cell = 30. The proportions must be the same in both rows: 4 is to12 as 10 is to 30.

Comment: Understanding the structure of a chi-square statistic.

SOLUTION Q19.
a) 4; b) 4; c) 2 or 6; d) 9; e) 1; f) 6; g) 7

Comment: Identifying the right technique of standard statistical inference.

SOLUTION Q20.
Here are four possibilities; other answers may also be correct.

- A P-value is the likelihood of getting the data you observed assuming the claim you started with is true.
- A P-value is the probability that sampling variability can explain the difference between what you observed and what you hypothesized.
- A P-value assesses how compatible the data are with the starting hypothesis; the lower the P-value, the less the compatibility.
- A low enough P-value means that the results are not likely to have happened by chance; rather, some real effect has happened.

Comment: This is probably the hardest question on the Self-Test. P-values have bedevilled students since P-values were invented!

How to score this test

There are 20 questions of varying lengths and complexity in this self assessment test. All questions address material covered in an introductory business statistics course. Question 1 to 8 concern descriptive statistics; Questions 9 to 20 are about inferential statistics.
A score of 80% or higher (16 out of 20) might indicate that your basic statistics knowledge is sufficient and you might not need the preparation course in Statistics. In particular if you missed 2 or more questions about descriptive statistics, and 3 or more questions about inferential statistics, it indicates that taking the Statistics preparation course would be advised to optimize your MBA or ECM learning experience. Remember that Statistics crosses all other areas of study! As the great Canadian humorist, Stephen Leacock, wrote: "Ah, statistics! Wonderful things, statistics; very fond of them myself."