School of Public Policy

ANSWER KEY TO PRACTICE QUESTIONS IN PRE-ARRIVAL MATHS AND STATISTICS MODULES 1 – 7

School of Public Policy

Table of Contents

MODULE 1

1. What is the difference between a theory and a hypothesis?

A hypothesis is a tentative assumption created before the study or data collection begins, which is tested during the course of the research.

A theory is supported by preexisting studies and evidence, and helps to explain findings already seen in the data. As a result, a theory is a much stronger claim than a hypothesis.

2. Imagine you are a policymaker, and you have been given additional funding to improve education outcomes in your district. You could spend these funds in many different ways, but the head of your department suggests that hiring more teachers to reduce the size of the class will improve test scores.

Using this idea, try to develop a research question, theory, and set of hypotheses.

Here is one way to complete this table:

RESEARCH QUESTION	What is the relationship between small class size and education outcomes such as test scores?
THEORY	Smaller classes means more attention is given to each student, which will improve their knowledge and academic performance.
HYPOTHESIS (H_A)	Children in smaller classes will have higher test scores.
NULL NYPOTHESIS (H_0)	There is no effect of class size on test scores.

3. What is the difference between qualitative and quantitative data?

Qualitative data is descriptive and non numerical, features a small and unrepresentative sample, and cannot be used with statistical analysis. Qualitative data aims for in depth understanding of the context or subjects in a study.

Quantitative data is numerical, features a large and representative sample, and can be used in statistical analysis. Quantitative data aims to test the hypotheses of a study, in order measure or quantify the causal effect in the research study.

4. Given this list of data, identify which are "qualitative" and which are "quantitative."

a. Number or measurements - QUANTITATIVE
b. Open ended interviews - QUALITATIVE
c. First hand accounts (such as diaries, or other primary sources) - QUALITATIVE
d. Survey data - QUANTITATIVE
e. Newspaper reports - QUALITATIVE

Note: interviews are typically considered to be qualitative because they are unstructured and typically feature a small sample, but a large scale survey of respondents (a survey is a type of interview) that asks open ended questions could be coded in a numerically meaningful way. It's also worth noting that, depending on the context, some qualitative data can in fact be quantified (or coded to be numerically meaningful).

MODULE 2

1. What do we mean when we say "correlation does not imply causation"?

This means that just because two events occur together, or correlate, does not mean that one has a causal effect on the other.

2. Suppose someone supplied you with data that showed a positive correlation between the number of nurses in a hospital (X) and the number of patient deaths (Y). They then tried to argue that increasing the number of nurses would cause an increase in deaths, and therefore nurses are bad for patient health. Using the concept of omitted variables, or any other logical reasoning, and explain why this conclusion might be wrong.

It's unlikely that more staff leads to worse outcomes for patients in a hospital, so there is most likely an omitted variable.

One such omitted variable might be the type of area in which the hospital is located — for example, maybe this area has an abnormally high rate of accidents. If this is systematically true, it also means that the hospital increased its staff in expectation of higher numbers of patients (and, the higher number of severe accidents, naturally the mortality rates would increase as well).

Another omitted variable might have to do with the type of hospital. Perhaps the hospital is responsible for caring for all the severely ill patients in the region, and these are the types of patients with high mortality rates.

Either omitted variable could both cause i) the number of nursing staff to increase, and ii) higher number of patient deaths.

For a real world example of this, see http://www.manchester.ac.uk/discover/news/national-study-casts-doubt-on-higher-weekend-death-rate-and-proposals-for-seven-day-hospital-services.

3. Imagine you are a policymaker, and you are focusing on issues with low turnout in your district (as in, the number of people voting in each election has been steadily decreasing). Here, turnout is your dependent variable (Y). List as many independent variables (X) as you can that could plausibly affect turnout on the day of the election.

Y= Voter turnout

X=

1. Adverse weather conditions
2. Dissatisfaction with candidates or parties
3. Uncompetitive elections
4. Illegal voter intimidation
5. Polling place too far away
6. Age or income of the voting population
7. And there are many more….

MODULE 3

1. Using Tables 3.1 and 3.2, write out this expression in words:

$$\forall x \in X \text{ and } \forall y \in Y, \exists z \in Z \text{ s.t. } x + y = z$$

For every x in the set X and every y in the set Y, there is an element z in the set Z such that x plus y equals z.

2. Evaluate: $2x^2 + 3x$ for $x = 3$

$2(3^2) + 3(3) = 0$

$2(9) + 9 = 0$

$18 + 9 = 0$

27

3. Solve for x:

(x+4) -(x+2)=6x

(x+4) - (x+2)=6x	Carry the negative sign through
x+4 - x - 2 = 6x	Carry the negative sign through
x + 4 - x -2 -6x = 0	Subtract 6x from both sides
x - x - 6x = -4 +2	Keep "x" terms to the left, subtract 4 and add 2 to both sides
-6x = -2	Combine like terms
x= -2/6	Simplify
x=1/3	

4. Of all the countries in the world, the country of Rwanda currently has the highest proportion of female politicians serving in the national legislature.

Term	Number of Female Legislators	Total Number of Legislators
2013	49	80
2008	45	80

Using this information, answer the following questions (you may round to one decimal point):

a. What percentage of legislators were women in 2008?

56.3%

b. What percentage of legislators were women in 2013?

61.3%

c. The number of female legislators change by how many percentage points from 2008 to 2013?

5 percentage points

d. What was the percentage increase in female legislators from 2008 to 2013?

8.9%

MODULE 4

1. Levels of measurement are usually placed in a sequence, from weakest measurement to strongest measurement (or said in another way, from the least informative to the most informative).

Place these in order, from the least informative to most informative: ratio, nominal, interval, ordinal.

Answer:

Nomin
al
Ordin
al
Interv
al
Ratio

2. For each variable, provide the type of measurement:

a. The number of citizens in a town
RATIO
b. A set of categories measuring the respondent's country (1=UK, 2=Singapore, 3=India, 4=Mexico)

NOMINAL

c. A "feeling thermometer", or a survey question that asks where a respondent likes a policy idea (with 1 being don't like it at all, and 10 being like it very much)

ORDINAL

3. What does the abbreviation SD stand for?

Standard Deviation

4. For each of these terms, list whether they are a measure of central tendency or dispersion.

Central tendency: mean, mode
Dispersion: SD, variance, range

5. Fill in the blank, with "large" or "small":

If the data are spread out far from the mean, the standard deviation will be large.

If the data are bunched tightly together around the mean, the standard deviation will be small.

MODULE 5

1. Please fill in the blanks in the following statement:

   Statistical inference is a procedure, in which we use what we know about the data from a **sample** to infer what is likely to be true about the **population**.

2. For sufficiently large samples, the sample means of a variable that is not normally distributed will be normally distributed.

   True

3. The population has a mean of 14 and a standard deviation of 3. The sample size is n=1,000. What is the mean of the distribution of the sample means from that population?

   Due to the central limit theorem, the mean of the distribution of the sample means will be equal to the population mean, which is 14.

4. The population has a mean of 120 and a standard deviation of 12. The sample size is n=16. What is the standard error of the mean?

   Using the formula for the standard error of the mean, we get $\frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{16}} = 3$.

MODULE 6

1. The higher the p-value of the test for a relationship between two variables, the more confident we are that there is a significant relationship. **True**

2. Please fill in the blanks in the following statement:

   Type-I error occurs when we **reject** a **true** null hypothesis. Type-II error occurs when **we fail to reject** a **false** null hypothesis.

3. Suppose you have a sample of 25 students from the MPA programme, and their mean exam score is 67. The standard deviation of their scores is 10. Using this sample ofdata, please test the hypothesis that the mean test score of all students in the programmeis 60. Use the significance level of 0.05.

   Our null hypothesis is $H_0 : \mu = \mu_0 = 60$. Our alternative hypothesis is that $H_1 : \mu \neq 60$.

   We don't know the variance of scores of all students, so we need to calculate the t-statistic: $t = \frac{\bar{X} - \mu_0}{s_X/\sqrt{n}} = \frac{67 - 60}{10/\sqrt{25}} = 3.5$. This number is larger than 3.467, which is the critical value for 24 degrees of freedom and the p-value of .002. So we can clearly reject the null hypothesis that the mean score of all students is 60 at 5 % significance level.

We don't know the variance of scores of all students, so we need to calculate the t-

$- \mu_0$ 67 – 60

$\bar{x}$

statistic $/\sqrt{n}$ = 10/ $\sqrt{25}$ = 3.5. This number is larger than 3.467, which is the critical : t =

s_X

value for 24 degrees of freedom and the p-value of .002. So we can clearly reject the null hypothesis that the mean score of all students is 60 at 5 % significance level.

4. In the latest YouGov survey in UK, 811 out of 1763 British adults say that their most favorite outcome of Brexit negotiations is to remain in the European Union. You can access the survey results at the following link: https://d25d2506sfb94s.cloudfront. net/cumulus uploads/document/ete4gzwp4j/UCL_Brexit_190529_w.pdf. Please calculate the 95% confidence interval for the percentage of British adults in UK, whose most favorite outcome is to remain in the EU.

   If we think of the group that wants to remain in EU as 1's and the rest as 0's, the sample mean is $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{(1 \times 811) + (0 \times 952)}{1763} = 0.46$.

   The standard deviation in the sample is: $s_X = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}}$

   $= \sqrt{\frac{811 \times (1 - 0.46)^2 + 952 \times (0 - 0.46)^2}{1763 - 1}} = 0.499$. So the standard error of the mean $= \frac{0.499}{\sqrt{n}}$

$\sigma_{\bar{x}} = \sqrt{\frac{}{1763}} = 0.012.$

Therefore, if we subtract and add two times the standard error from and to the sample

mean, we get the following 95 % confidence interval: $\bar{X} \pm 2 \times \sigma_{\bar{x}} = 0.46 \pm (2 \times 0.012) =$

0.46 ± 0.024 ([.436; 0.484])

We can say with 95% confidence that the percentage of British adult population thatprefer to remain in the EU is between 43.6% and 48.4%.

MODULE 7

1. Please fill in the blanks in the following statement:

   In an experiment to determine the effectiveness of a new drug, the treatment group would be those who receive the new drug, and the control group would be those whodon't.

2. Why is it important that subjects are randomly assigned to the treatment and control groups in an experiment? Briefly explain.

   Thanks to random assignment, the treatment and the control groups are not systematically different from each other except the fact that those in the treatment group have re- received the treatment and those in the control group have not. Therefore, if the researcher observes a difference between the two groups' responses or values of the dependent variable, that difference must be because of the treatment assigned by the researcher.

3. What types of data are used in natural experiments?
   The data used in natural experiments are observational.

4. The legal drinking age in US is 21. Suppose we are interested in establishing the causal effect of legal access to alcohol on death rates in US. How can we use this information to construct a research design to establish this causal effect? Briefly describe.

   We can use this information to construct a regression discontinuity design and compare the death rates among American youth right above the cutoff of 21 years of age, andbelow 21 years of age. The assumption that this design relies on is that these two groupsdo not systematically differ from each other on factors other than legal access to alcohol and age.