Mathematics for Policy and Planning Science

Stephen Turnbull

Graduate School of Systems and Information

Homework: updated June 24, 2019

Abstract

All homework is collected here. The first few slides explain the submission process and format. The rest present the homework problems, *most recent first*.

Submitting Homework

Read and understand the following instructions on submission of homework. **If you do not follow them, you will not receive credit.**

Submit this assignment by *email*. Give the mail the subject "01CN101 Homework #<number> by <your name>" in *hankaku romaji* and send it to `turnbull@sk.tsukuba.ac.jp`. (This subject is necessary for automatically sorting incoming mail.)

Make sure that the body of the email contains your *name* and *student ID number*.

If you are late, submit the assignment for partial credit. The later, the less credit you will receive. If you believe that the late submission is in part due to lack of care by the instructor, or some event (such as hospitalization) required your full attention for two full days or more, you may explain for additional credit. Otherwise, I don't care why it was late.

Submit this assignment as a *plain text* e-mail. This homework requires no special symbols or wordprocessor features. (*Do not* attach a wordprocessor file such as Microsoft Word document or a Excel spreadsheet. *You will lose some credit!*)

Homework 4, due July 8, 12:00 noon

In these problems, unless otherwise specified, you may use the cumulative distribution, the mass function, or the density function, as is convenient.

1. An airplane regularly flies from Tokyo to Sapporo. It takes off at the same time every day, but it faces varying weather conditions. The airline checks its records and computes that the equation for the time to arrive is $t = 92 +$, where is distributed approximately normally with mean 0 and standard deviation 4.5. Time units are minutes.

   (a) What is the *domain model* being used?

   (b) What is the *statistical model* being used?

   Explain each model in some detail.

2. Here is some data about 10 college students:

Person	1	2	3	4	5	6	7	8	9	10
Gender	F	F	F	F	F	M	F	F	F	F
Age	23	20	21	20	19	18	20	22	18	22
Height	161.7	169.5	159.2	159.4	165.1	166.4	153.8	163.3	164.0	168.3
Weight	62.0	65.1	57.2	62.9	66.1	64.9	57.4	62.7	63.6	64.0

(a) Compute the joint frequency distribution of Gender and Age.

(b) Compute the cumulative joint distribution of Height and Weight.

(c) Compute a histogram of heights for 5cm ranges from 150 to 170 cm. Draw it as a bar graph.

(d) What is the percentile rank of the man's height?

(e) The BMI of a person is that person's weight in kilograms divided by the square of their height in meters: $BMI = W/H^2$. Describe the correct way to compute the average BMI of this group. Why is it correct?

3. Outliers For the following parts, refer to the distribution depicted below::

```
|_xxxoxoxxxoooxoooxo___...
a                 b
```

The left-to-right scale is quantitative and linear, increasing toward the right. "x" and "o" denote the two types of qualitative outcome. The underscore "_" indicates that there was no observation at that level of the quantitive variable. The bar at the extreme left labelled "a" indicates a theoretical lower limit on the quantitive variable (values lower than a are impossible). The observation labelled "b" will be referred to later. The ellipsis at the right indicates that (1) there is no theoretical upper limit, and (2) the absence of observed values continues forever to the right.

(a) Define *outlier* as used in statistics.

(b) Identify an outlier in the distribution above. (Choose the "worst" outlier.)

(c) Based on "looking at the picture," give a procedure to compute the outcome to be expected given a position on the line, based on the observed distribution above. You have to predict either "x" or "o" based on the position on the line. It does not have to be based on any particular statistical concept, but do the best you can.

(d) What does your procedure predict in the case of a new observation at the level "b"?

(e) Explain what *over-fitting* (also called "over-training") is.

(f) The ”1-nearest neighbor” classifier predicts that if an individual is observed at ”b”, it will be an ”x”. Explain how this is an example of ”over-fitting”.

(g) Explain how ”over-fitting” can occur in calculating the mean. (Note: This question was not discussed in class, and is intended to challenge those who consider themselves adept at mathematics. Questions like this that are ”difficult extensions” of class discussion will not be asked on the exam.)

4. Consider collecting data about people.

   (a) Give two examples of variables about people that are *qualitative* (also called *categorical*) but not ordered or quantitive (cardinal).

   (b) Give two examples of variables about people that are *ordered* (also called *ordinal*) but not quantitive (cardinal).

   (c) Give two examples of variables about people that are *quantitative* (also called *cardinal*).

   (d) Give the values for yourself of each variable you mentioned in parts a, b, and c.

   (e) For each type of variable (qualitative, ordered, quantitative), give one example of a statistical operation that *may* be performed on that kind of data, and one example of a statistical operation that *should not* be performed on that kind of data. (Your answer may be "none" when you believe any statistical operation is valid, respectively invalid, for that kind of data.)

5. Descriptive and inferential statistics

   (a) Briefly define *descriptive statistics* and *inferential statistics*.

   (b) The purely mathematical calculations for descriptive statistics and inferential statistics are mostly the same. How can you tell when someone is doing descriptive statistics and when it is inferential statistics?

   (c) There are two kinds of empirical variance, the population variance and the sample variance. The difference is that the sample variance uses the number of observations less one to correct for bias. Is the sample variance a descriptive statistic or an inferential statistic? How do you know? (This question requires some knowledge of statistics not discussed in class, although a native speaker of English can probably get the right answer without knowing about statistics. It will not be asked on the test.)

Homework 3, due June 17, 12:00 noon

Consider the "chair selection problem" of *Example 1.*

1. What are the probabilities of each pair of *numbers*?
2. What events (as sets of number pairs) correspond to the events that each committee member is selected as chair?
3. Is it possible for all members to become chair? If not, which members are possible, and which impossible?
4. Explain why the definition says "all pairs remaining."
5. What are the probabilities of each member becoming chair?

Homework 2, due June 17, 12:00 noon

As usual, answer questions in plain text in the email. Exception: You may attach a spreadsheet for the tables, including those used to construct distributions as well as the distributions themselves.

In class we showed that the distribution of sums of dots on a pair of dice is:

sum	2	3	4	5	6	7	8	9	10	11	12
frequency	1	2	3	4	5	6	5	4	3	2	1

Table 1: Sum of two dice

The distribution of ordered pairs is said to be *uniform*. For each pair, the frequency seen in a series of throws of the dice should be about the same. Theoretically, this occurs because each (ordered) pair occurs once in a list of pairs (not shown here – usually displayed as a square table with six rows and six columns). The frequency of sums is *nonuniform* (different frequencies for different values). Theoretically this difference occurs because there are often multiple pairs of dice that result in the same sum and the number of appropriate pairs varies according to the sum specified.

Homework 2(ii): Problems 1-2

1. Count the number of combinations of the number of dots on a pair of dice that achieves a particular:

   (a) product

   (b) difference

   (c) quotient

   of the pair and make a frequency table for each arithmetic operation (times, minus, and divide).

2. Did you notice any interesting similarities or differences between the frequency distributions? Did you notice any interesting similarities or differences between the random variables (*i.e.*, the function from the pairs of dice to the mathematical result)?

Homework 2(iii): Problems 3-4

3. Describe a model of ”what is the gender of the first person to arrive in the classroom of ’Mathematics for Policy and Planning Science’” with an underlying set whose probabilities are *uniform*. What ”model” means here is ”What things do you count to determine the probability that the first person to arrive is female?”

   Do you think this model actually describes the probability that the first person to arrive is female accurately? If so, why? If not, why not?

4. Think of a practical or daily life application where you can explain observed rates of occurance with a model with an underlying uniform distribution, but a nonuniform distribution for the observed (or practically relevant) outcomes. What is the underlying set composed of? Why should its elements be uniformly distributed? What is the function relating the underlying things to observed outcomes?

Homework 2(iv): Problem 5

5. **(Optional)** Here's a brain-breaker for those of you who think you're good at math. The numbers on dice are all positive, and therefore it makes sense to take logarithms. There's a one-to-one relationship between numbers and their logarithms, so given a number we can find its logarithm, and it is the unique number with that logarithm. And given a number that is a logarithm there's a unique number it is the logarithm for.

   Now, after taking logarithms, multiplication becomes addition. Therefore we might expect that there should be a one-to-one relationship between the distribution of *sums* of two dice and the distribution of *products* of two dice. That is wrong.

   Explain why.

Homework 1, due June 10, 12:00 noon

Here are the first three Peano Postulates describing what we call the *natural numbers*, $\mathcal{N} = \{0, 1, 2, \ldots\}$:

1. 0 is a number.
2. Every number n has exactly one successor, n'.
3. Two different numbers $m \neq n$ have different successors, $m' \neq n'$, and *vice versa*.

There are also finite number systems, such as the Boolean algebra on $\{0, 1\}$. These number systems are characterized by the number of elements, n. (The Boolean algebra has $n = 2$.)

Homework 1(ii) Problems

1. How do we use the following finite number systems in daily life? You may want to rename numbers to something more familiar. Some systems have more than one use!

$$n = 7, \qquad n = 12, \qquad n = 24, \qquad n = 60, \qquad n = 360.$$

2. Can you think of any other examples?