2nd-Half Examination

Mathematics for Policy and Planning Science

July 1, 2019

General instructions / 一般説明

Do not forget to write your name and student ID number on each page.

Several problems in basic mathematics are presented below. **You may answer in Japanese or English.** In Japanese, please take great care in writing kanji. Avoid abbreviated kanji; the only one I know is the 3-stroke mongamae.

Use of notes, textbooks, and so on is prohibited. All calculations are simple, and allowing use of electronic devices is unfortunately far to great a temptation to some students, so the use of calculators and electronic dictionaries is also prohibited. Some dictionaries will be provided.

On the desk you may place pens, pencils, erasers, pencil sharpeners, and tissues or handkerchief of reasonable size. Take out your cellphone, turn it off, and place it on the desk in front of you. All other items must be put in your bag and placed at your feet.

Except for calculations, many problems can be completely answered within 3 lines. A few questions can be answered within 2 or 3 words. Below each problem ample space is provided. Please write your answers there. Special space will be provided for graph problems. Please use it. In calculations, in addition to the result itself, please also write any equations used.

名前と学籍番号を忘れずに各ページに記入してください。

以下の設問のすべてに解答せよ。**解答の言語は日本語でも英語でも構わない。**もし日本語で書けば漢字などの書き方に十分注意してください。たとえ、省略した漢字などを使わないで。（三角門構えの他は分からなく、そして私が読めない場合には省略した文字を「間違え」と採点します。）

ノート・教科書・電子辞書・電卓・携帯電話・その他のメモリ付き電子製品が誘惑ものとなるので使用は禁止である。全ての計算は簡単で電卓などは不要。

机の上にペン・鉛筆・消しゴム・鉛筆削り・時計・この試験用紙の他の物を置かないこと。携帯をお持ちなら電源を切って机の上に置くこと。その他のものを鞄などに締めて足元に置くこと。

裏面の空白を使ってもよい。ただ、その場合に表にメモを書くこと。メモがない場合、別所で書いた文はカウントされない場合があります。

Problems / 問題

The point value of each question is marked with the question.

それぞれの問題は基本的点数が１０点。場合により、５点、２０点などの問題し、確認してください。

1. `[Problem ID #50] Anscombe`
   [5 points] Recall the *Anscombe data sets*. Mark each of the following statements as *True* or *False*.

   (a) *True* All of the 4 data sets are extremely similar in terms of one-variable descriptive statistics.

   (b) *False* We can easily distinguish each of the data sets by looking at regression coefficients, standard errors, or R^2.

   (c) *False* All of the 4 data sets seem to be based on a trend plus a substantial amount of randomness.

   (d) *True* We can easily distinguish each of the data sets by looking at X-Y plots of the data sets themselves.

   (e) *True* We can easily distinguish each of the data sets by looking at X-Y plots of the regression residuals for the data sets.

2. `[Problem ID #46] fractile: vs. fractile rank`
   [5 points] Explain the difference between *fractile* (or *percentile* if you prefer) and *fractile rank* (respectively *percentile rank*).
   A fractile is computed by solving the equation $F(c) = r$, where r is the fractile rank *and c is the r*-fractile. *The fractile is a number given in units of the random variable, while the rank is a number between 0 and 1, or a percentage between 0% and 100%.*

3. `[Problem ID #52] big data: definition`
   [10 points] What are the three characteristics of "big data" that make it appropriate to use modern statistical methods such as machine learning instead of the more traditional regression or analysis of variance methods?
   From the notes: The characteristics that tend to define big data are:

   - *measurements designed for some purpose other than proving a theory by statistics*
   - *relatively low information density*
   - *frequent updates with new measurements*

   You were not asked to explain why *modern methods are needed but if you did so, it would likely be something like "the low information density is often due to a large number of variables measured, which leads to the 'curse of dimensionality', and conventional regression and correlation analysis is not good at dealing with that."*

4. `[Problem ID #39] Peano system: 0 not a successor`
   [5 points] Recall the Peano Postulates describing the *natural numbers*, $\mathcal{N} = \{0, 1, 2, \ldots\}$:

   (a) 0 is a number.

   (b) Every number n has exactly one successor, n'.

   (c) Two different numbers $m \neq n$ have different successors, $m' \neq n'$, and *vice versa*.

   (d) 0 is not the successor of any (natural) number.

   What is the role of the fourth postulate ("0 is not a successor")? Hint: what can happen without it?
   The fourth postulate ensures that $\mathcal{N}$ is infinite. Otherwise, taking the successor will cycle back to 0 with a finite number of "numbers." Sometimes that's useful ("clock arithmetic") but we also need a way to "count as high as needed."

5. `[Problem ID #IDNO] set computation`
   [5 points] Let $A = \{math, physics, chemistry, biology\}$ and $B = \{math, English, planning, finance\}$. Compute the following sets:

   (a) $A \cup B$
   $A \cup B = \{math, physics, chemistry, biology, English, planning, finance\}$.

   (b) $A \cap B$
   $A \cap B = \{math\}$.

   (c) $A \backslash B$
   $A \backslash B = \{physics, chemistry, biology\}$.

   (d) $B \backslash A$
   $B \backslash A = \{English, planning, finance\}$.

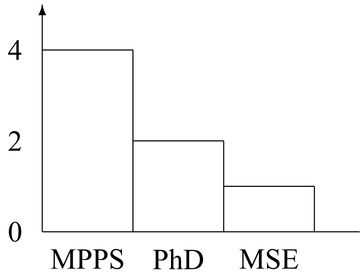
6. `[Problem ID #35] probability algebra independence`
   [5 points] The probability of the event A is a, and the probability the event $A \cap B$ is b. What is the probability of event B?
   This cannot be computed without knowing whether A and B are independent. If you assume explicitly *that A and B are independent events, then the probability is ab.* ***This is a trick question.*** *(Compare problem* **??***.)*

7. `[Problem ID #42] data generating process`
   [5 points] What is a *data generating process*? Explain why it is important in practical use of statistics.
   A data generating process *is a combination of a domain (scientific) model and a statistical model. The data generating process determines what the biases of any given statistical procedure are. For example, if the disturbances in a regression are correlated, ordinary least squares may be biased and inefficient.*

8. `[Problem ID #45] distribution: absolute vs. relative`
   [5 points] What is the difference between an *absolute distribution* and a *relative distribution* in statistics? What is the relation of these distributions to probability distributions?
   An absolute distribution *simply counts the number of times a particular value (or vector of values) occurs in a data set. A* relative distribution *gives the fraction of the entire data set in which the value (vector) occurs. The arithmetic of relative distributions is very similar to that of probability distributions, and in fact relative distributions are often used to estimate probability distributions.*

9. `[Problem ID #44] histogram`
   [5 points] In Prof. A's advanced statistics class, there are 2 PhD students, 1 MSE student, and 4 MPPS students.

   (a) Draw a histogram of this distribution.



   (b) Is there any particular reason for the way you draw this histogram?
   Yes. It is often useful to use a "Pareto transformation" on a distribution, that is, sort the values from highest frequency to lowest when displaying in a table of histogram. If you didn't have a special reason, "no" is a perfectly good answer.

10. `[Problem ID #49] descriptive vs. inferential`
    [5 points] What is the difference between *descriptive* and *inferential* statistics? How can you tell the difference in a particular use case?
    Descriptive statistics *refers only to the data in the given data set.* Inferential statistics *estimates or predicts data (including parameters or unobserved values) outside of the data set.*

    Sometimes it is difficult to determine the intent, but usually discussion of inferential statistics does explicitly refer to data outside the data set.

11. `[Problem ID #IDNO]`
    [10 points] Why do we use mathematics in Policy and Planning Sciences?

    (a) Mathematics is useful because mathematical models describe real world behavior more accurately and realistically than theories expressed in words.

(b) Mathematics is useful because equations and statistics can be analyzed by standard methods and algorithms, and the researcher can focus on matching the model to real behavior rather than analysis techniques.

(c) Mathematics is useful because we can use computers to calculate and analyze statistics and equations, and so get numerical answers.

Choose one of the above statements which you think is *most accurate*, *copy it* to your answer, and *answer both* of the following questions

(a) Why do you think it is true?

(b) What does it say about what part of the work of researchers and analysts in Policy and Planning Science is most important?

I hope you chose statement (11b). If most of you do, I will consider my most important goal for this course fulfilled. You know when *and* why *to use mathematics, and I hope that you will choose well which mathematics and how much to study. You can spend your whole life studying mathematics. (I think that's heaven!) But for most people, mathematics is just a tool. Optimize your use!*

If you picked (11a), one plausible rationale is something like the following. People talk about things in vague and ambiguous ways, and when problems are complex and have underlying quantitive processes, others may incorrectly understand their words. In converting verbal descriptions to variables and equations, we need to make precise definitions, and explain how to measure the variables. This process makes mathematical models more realistic and accurate than verbal discussion. The costs of mistaken policies and negotiation of ambiguous goals and processes are very high, and minimizing them is the most important function of mathematics.

The most important task of people trained in PPS is to turn statements of policy-makers about goals and processes into mathematics.

If you picked (11b), one plausible rationale is something like the following. Although policy-makers will take the final decisions, when their statements are vague and ambiguous, staff have an important role to play in clarifying them. Whenever possible, this role needs to take precedence over the relatively mechanical tasks of measurement, formulation, and calculation. The more you know about mathematics, the more of the mechanical tasks you can delegate to computers or to junior colleagues, because they can reliably follow algorithms even though you don't trust their judgment as much as your own.

The most important tasks for those trained in PPS are helping policy-makers to specify the goals and processes involved, to specify the requirements for a solution, and to convert these specifications into computable models.

If you picked (11c), one plausible rationale is something like the following. Modern social engineering is a quantitive activity. We can get very precise *financial information, and plan construction and manufacturing activities very accurately. Risk and various social indicies are harder, but we're getting better at those as well. With all this information, and computation so cheap, it makes sense to use these abundant resources as much as possible.*

The most important tasks for those trained in PPS are collecting and analyzing data, optimizing policy, and simulating outcomes of different policies.

I hope you already noticed, but the modeling (11b) motivation incorporates both the specification (11a) and computational (11c) motivations. What an expert in PPS should provide to their client is the ability to balance the three aspects appropriately, using their knowledge of what mathematics is and is not useful for.

12. `[Problem ID #59] DESCRIPTION`
    [5 points] Define (explain what is special about) *social media platform*, compared to social media in general. How are platforms related to "big data"?
    A social media platform *is an automated system, usually on the Internet, for managing a social network.*

    Platforms frequently collect a wide variety of information, including activity history, about individual users. This means storing a large number of variables, which is a characteristic of "big data."

13. `[Problem ID #53] big data: not size`
    [5 points] Why isn't the size of a data set particularly important in defining "big data"?
    Modern computer systems can handle very large data sets with traditional tools such as regression analysis. In fact, the computational burden is typically smaller for the traditional tools. Thus the reason "big data" indicates use of newer tools must be something other than size.

14. `[Problem ID #55] big data: cross-validation`
    [5 points] What is *cross-validation*, and why is it necessary?
    Cross-validation *is the practice of dividing a data set into "training" and "testing" subsets. The model is estimated using the training subset, and then that model is used to "predict" the values in the test subset. This helps recognize* over-fitting.

15. `[Problem ID #IDNO] probability model: construct from hypothesis`

    [10 points] Consider the problem of preparing for a test. Suppose that the same instructor has taught a course in each of the past ten years. She has posted all of the past exams on her home page, as well as a testbank containing all of the exam questions, as well as some questions never used on an exam. You want to model the instructor's question selection process using probability theory.

Consider **Hypothesis X**: "The instructor writes questions after each lecture, and adds them to the testbank. In the week before the final, she selects questions from the testbank to create the test." Can this hypothesis be tested with a statistical hypothesis test, or not? **Explain** why you think so.
In my opinion, this hypothesis can't be tested statistically. Several aspects of the data generating process are unclear, but are necessary to test the hypothesis. Most important, it seems unlikely that it is possible to determine when *questions were written and added to the testbank. The best we can do is check whether the questions were on the testbank by the last week of the term, but this isn't really a statistical test.*

Other answers may be possible but that one seems most likely to me.

16. `[Problem ID #48] data types: ordinal vs. cardinal`
[5 points] What is the difference between *ordinal* and *cardinal* (also, *quantitative)* types of data? Given an example of each, and explain how your answer applies.
Both ordinal and cardinal values can be sorted into a "natural" order, but only in the case of cardinal values is there a natural unit of measure so that you can say "A is twice as far from B as C is," or even "A is twice as big as B" (for the latter you need a "natural zero" or origin.

For example, you probably have a pretty good sense when studying for a class whether you learned much or little, but it's just a figure of speech to say "I learned twice as much in the morning as I did in the afternoon!" On the other hand, "I got twice as many points on the final as on the midterm" is an objective statement.

17. `[Problem ID #IDNO] probability model: construct from hypothesis`

[10 points] What are the **three generic ways** (or interpretations of "probability") you can use to determine the probabilities to use for the events you are modeling? **Briefly explain why** (or when, for what kind of situation) you might use each one.

Frequency *In some cases, the exact event occurs repeatedly (such as balls bouncing off pins in a pachinko machine), and you can count how often it occurs, and use the relative frequency of each event as its probability.*

Symmetry *Sometimes there's a fundamental symmetry, such as the cubical die or the sides of a coin, such that it seems reasonable to decide that symmetrical events have the same probability.*

Subjectivity *Sometimes you simply estimate "in your head." Humans aren't very good at this, but it's a fundamental behavior in an uncertain world.*

18. `[Problem ID #21] statistics model`
[10 points] 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 $t = 92 + \epsilon$, where ϵ is distributed approximately normally with mean 0 and standard deviation 4.5. Time units are minutes.

Explain each model in some detail.

(a) What is the *domain model* being used?
The domain model *is that it takes a certain amount of time to fly the route. This time is 92 minutes. It doesn't say why a constant time is an appropriate model. The model I had in mind is that the plane needs a certain time to take off and land, and then travels using a specific amount of power to the engines over the same route every time. There are other models that would give the same equation.*

(b) What is the *statistical model* being used?
The statistical model is that there is a random variation in flight time. The distribution was measured empirically, and the use of a statistical model is based on the weather (presumably wind direction and speed). This last statement matters because it means that