データ解析基礎- Basic Data Analysis

中間試験- Midterm Test

May 26, 2011

Problems

You received data sets on a separate sheet of paper, that look something like this: A typical dataset:
1. Your data sets have data set ID #1.
Be sure to enter your data set ID in the space provided.
Data Set A: 21 21 22 16 22
Data Set B: 100 63 100 92 90 100 82 54 81 100

When you are asked to do a calculation, you do not need to compute the decimal equivalent of a fraction or radical (square root). Fractions should be reduced to lowest terms for convenience in grading. Radicals do not need to be reduced.

計算を行うときには分数または根数のままを書いてもよい。少数にする必要は ない。ただし、分数の分母と分子は互いに素にすること。

For Problems 1 to 6, use **Data Set A**. (Each student receives a different data set. Make sure your Data Set ID is correctly entered in the space at the top of the page.) Data Set A is a data set of defects per 1000 cars recorded at workstations of a large manufacturer. Classify the sizes according to the following *subjective size* scale: ≤ 5 is "very small," 6-10 is "small," 11-20 is "large," and ≥ 21 is "very large."

問題1～6に**データセットA**を利用してください。(注意: 皆に別のデータ を用意する。必ずデータセットIDを確認すること。) データセットAは工場の ワークステーションで見つかった1000台あたりの欠陥数である。 サイズを「主観的サイズ」に以下の表によって区別する: 5 以下は「非常に小さい」、6～10は「小さい」、11～20は「大きい」、そして 21以上は「非常に大きい」。

Copy your data set here: ここにデータをここに写ること:
Data Set A: 21 21 22 16 22

1. `[Problem ID #IDNO] DESCRIPTION`
   Sort your raw data, and write the sorted data here.
   データを順序に並べてここに書くこと。

   16 21 21 22 22

2. `[Problem ID #IDNO] DESCRIPTION`
   Convert your data to "subjective sizes," and enter the absolute, relative, and cumulative relative frequency distributions here. What is the *median* of your data?
   データを「主観的サイズ」に変換して絶対頻度分布、相対頻度分布、または（相対）累積分布を書け。中央値（メディアン）はどれですか？

size	*value*	*absolute*	*relative*	*cumulative*
very small	*2.5*	*0*	*0%*	*0%*
small	*8.0*	*0*	*0%*	*0%*
large	*15.0*	*1*	*20%*	*20%*
very large	*23.5*	*4*	*80%*	*100%*

   The median is ``very large''.

3. `[Problem ID #IDNO] DESCRIPTION`
   For each "subjective size," choose a *representative numerical size*. Enter the correspondence here. **Explain why you chose *each* numerical value.**
   各「主観的サイズ」に代表的量的サイズを選んで、ここに記入。**それぞれの値を選んだ理由を説明せよ。**

   I chose the values in the ``value'' column in the table above.
   This problem is relatively difficult because of the uneven sizes of the cells (5, 5, 10, and infinite). One reasonable procedure is to take the midpoint of each of the lower cells, and skipping a somewhat larger amount from the value for the second largest cell to the value for the largest.
   Why choose cells of uneven size like this? Perhaps most of the time the distribution has a small median, but a ``fat'' upper tail. These cells may express this.
   Note that the example data (which were generated randomly) does not *satisfy the description of the ``typical'' distribution for this kind of data. Sometimes that happens! But the cells will be chosen to fit the typical data, not the (unusual) data you happen to get in a particular trial.*

4. `[Problem ID #IDNO] DESCRIPTION`
   Using the representative sizes you chose in the previous problem, compute the *mean*, *variance*, and *standard deviation* of the *distribution* of sizes. Show your work (*e.g.*, using a table like that in Homework 2's spreadsheet).

   前の問題で選んだ代表的サイズと分布を用いてサイズ分布の平均値、分散、 と標準偏差を計算せよ。計算方法を表すテーブルなどを含むこと。（たとえ、第２宿題のシートのようなもの。）

size	x	$f(x)$	$xf(x)$	$x-\mu$	$(x-\mu)^2$	$(x-\mu)^2 f(x)$
very small	*2.5*	*0.00*	*0.00*	*-19.3*	*372.49*	*0.000*
small	*8.0*	*0.00*	*0.00*	*-13.8*	*190.44*	*0.000*
large	*15.0*	*0.20*	*3.0*	*-6.8*	*46.24*	*9.248*
very large	*23.5*	*0.80*	*18.8*	*1.7*	*2.89*	*2.312*
		1.00	*21.8*			*11.56*

The mean is $\mu = 21.8$*, the variance is* $\sigma^2 = 11.56$*, and the standard deviation is* $\sigma = \sqrt{11.56} = 3.4$.

5. `[Problem ID #IDNO] DESCRIPTION`
   Pick **one** of the following three cases, and answer the question in the space provided below.

   以下の状況説明から一つを選んで、下記の問aと問bを答えろ。

   (a) The data set of defects was derived from the monthly count for each month *over a 5-month period in 2010* at one workstation in a particular factory.
   欠陥数データは２０１０年の5ヵ月分のある１つのワークステーション の月次欠陥数(1000台あたり）データである。
   Note that your answer to this kind of question does not depend on the data!
   Whether the workers are the same in each month is not stated. If the workers are different, one would expect different error distributions depending on experience and native ability.

   (b) The data set of defects was derived from the monthly count for each month over a 5-month period in 2010 for a particular *new worker* in a particular factory.
   欠陥数データは２０１０年の5ヵ月分のある１人の新社員 の月次欠陥数(1000台あたり）データである。
   One would expect the worker's (average) error rate to fall over time as he acquires training and experience.

(c) The data set of defects was taken in *May 2010 at 5 different workstations* in one factory.
欠陥数データは２０１０年５月の１つの工場の５ヵ所のデータである。
Since the work differs, one would expect the error distributions to be different at each station.

For your chosen case, give *one* example of a "hidden factor" relating different observations in the data set that could affect the way you interpret your statistics. Explain why this matters.

選択状況には「隠された要因」により観察間関係が現れ、統計量の解釈に影 響を及ぼすことがある。その要因・関係をひとつを選んで説明せよ。

6. The executives of the company are worried about large defect counts. The Board is considering reassignment, retraining, or workflow redesign in cases of *very large* counts, *i.e*., classes with 21 or more defects per workstation. Assume your data set A is based on the counts for a sample of 10 different workstations in one factory taken in May 2010.

会社の役員は厳しい市場状況を受け、大きすぎる(「非常に大きい」)ワー クステーションを改善する案を検討する。 欠陥数データセットAは２０１０年５月のある工場の１０ヶ所のワークステー ションのデータとする。

(a) Based on your Data Set A, do you think there is a problem related to defects? Explain why or why not, based on statistics you compute in Problem 4.
問題があるかどうかをデータセットAに基づいて検討せよ。問4で計算した 統計量に基づき理由を説明すること。

(b) Based on your Data Set A, do you think the proposed process improvements for workstations with excess defects will *change the distribution* of defect counts in factory? Explain why or why not. (You may assume that any process improvements are effective in reducing defects. This question is about the numbers, not about industrial engineering.)
データセットAを参照して改善案は工場の欠陥分布に大きい影響が与えら れるかについて自分の意見を述べてその理由を説明せ

よ。（改善案は効果 的であることの仮定上に答える。）

According to the descriptions from ``very small'' to ``very large,'' in this data set both mean and median are in the ``very large'' range. Presumably this is unexpectedly high, so yes, it seems reasonable to expect an improvement in results from an improvement in process.

For Problems 7 and 8, use **Data Set B**. (Each student receives a different data set.) Data Set B is a data set of examination scores on a 0-100 scale.

問題7・8に**データセットB**を利用してください。（注意：皆に別のデータ を用意する。必ずデータセットIDを確認すること。）データセットBはある試 験の点数データで、0〜100の範囲である。

Copy your data set here: ここにデータを写ってください：

Data Set B: 100 63 100 92 90 100 82 54 81 100

7. `[Problem ID #IDNO] DESCRIPTION`

Give the definition of *median*. Find the median of the *raw* data from Data Set B. Now, convert Data Set B to letter grades according to the usual scale, and enter a table containing the letter grade, the *scale□nterval*, the absolute frequency, the relative frequency, and the cumulative frequency distribution. What is the *median* of the distribution of letter grades? Compare it to the raw (point score) median.

中央値（メディアン）の定義を書け。データセットBの中央値を記入せよ。デー タセットBを普通のスケールでレターグレードに変換し、レターグレード、 スケール範囲、絶対同数、相対同数、と（相対）累積頻度分布を表に記入す ること。レターグレード分布の中央値を求めよ。点数のメディアンと比較せよ。

The median is the 50-percentile value, that is, the value such the 50% of the distribution is larger than this value and 50% is smaller.

Sorting the data set gives 54 63 81 82 90 92 100 100 100 100, so the median is some number between 90 and 92. Some statisticians report both (the median values are 90 and 92), others take the average (the median value is 91).

The letter grade distribution is

grade	interval	absolute	relative	cumulative
D	0-59	1	10%	10%
C	60-69	1	10%	20%
B	70-79	0	0%	20%
A	80-100	8	80%	100%

The median of the letter grades is A. This corresponds accurately to the letter grade corresponding to the median point score.

8. `[Problem ID #IDNO] DESCRIPTION`
Draw a histogram for the *raw* data set. Drawing a histogram involves a choice of division into cells of values. (Recall that a *cell* is a group of values that are close to each other.) *Explain why* you chose the cells you did.

点数データのヒストグラムを描け。ヒストグラムの作成には値の区間（セル、 仕切り）の選択が必要だ。（区間は値の範囲だ。）区間の選択の理由を説明 せよ。

There are two very reasonable answers. The first is to use cells that correspond to grade boundaries (0-59, 60-69, 70-69, and 80-100), and use a density histogram. The second is to use cells of width 10 (most of which have frequency zero).
Since the problem specifies the raw data set, a histogram for the distribution of letter grades is not acceptable.

9. `[Problem ID #IDNO] DESCRIPTION`
Which of *random variable* and *distribution* can you define and describe without using *primitive events*, but only using ordinary events? Explain your answer by referring to the definitions.

「乱数」と「分布」のどちらが「primitive event」を使わずに定義できる か。（普通の「event」（事象）はもちろん使う。）その理由を説明すること。

Defining a distribution does not require the concept of primitive events. Defining random variable does, because the primitive events are the domain (source set) for the random variable's function.

10. `[Problem ID #IDNO] DESCRIPTION`
Suppose you pick a child at random from an elementary school. Are the events ``the child is in 2d grade'' and ``the child will graduate this year'' *independent*? Are they *mutually exclusive*? Explain.

小学校の１人の生徒をランダムに選ぶ。「２年生である」と「来年卒業する」 という事象を定義する。２つの事象は「independent」ですか。「mutually exclusive」ですか。その理由を説明すること。

These events are (normally) mutually exclusive, since children don't finish elementary school until they are in 6th grade. Since knowing the child is going to graduate means you can deduce they are in the 6th grade, not the 2nd, your assessment of the likelihood of being in the second grade changes dramatically given that information.

11. `[Problem ID #IDNO] DESCRIPTION`
Class B of the 3rd grade in East Takezono Elementary School is supposed to be the most international class in Tsukuba City, in terms of nationality of its students. Take a student from Class B at random, and consider the events A = "the student is Chinese," and B = "the student is female." State as many facts as you can about $\Pr(\{\})$, $\Pr(A)$, $\Pr(B)$, $\Pr(A \cap B)$, $\Pr(A \cup B)$, and $\Pr(\Omega)$, including comparing the probabilities of two events (*e.g.*, $\Pr(A) < \Pr(\Omega)$).

つくば市のもっとも国際的な年度組は竹園東小学校の３年B組と言われてい る。つまり、生徒の国籍が一番多い。ランダムにB組のひとりの生徒を選出 し、A=「中国人だ」とB=「女性だ」という事象を考察しよう。$\Pr(\{\})$、 $\Pr(A)$、$\Pr(B)$、$\Pr(A \cap B)$、$\Pr(A \cup B)$、$\Pr(\Omega)$ についてできるだけ多くの事実を書け。事象の確率の比較を含む。（例：$\Pr(A) < \Pr(\Omega)$。）

$Pr(\{\} \leq \Pr(A) \leq \Pr(A \cup B) \leq \Pr(\Omega)$
$Pr(\{\} \leq \Pr(B) \leq \Pr(A \cup B) \leq \Pr(\Omega)$
$\Pr(A \cap B) \leq \Pr(A) \leq \Pr(\Omega)$
$\Pr(A \cap B) \leq \Pr(B) \leq \Pr(\Omega)$