データ解析基礎- Basic Data Analysis

問題集（例） - Sample Examination Questions

These questions may or may not appear on any given examination.

Problems

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. That is $\frac{7}{5}$ and $1\frac{2}{5}$ are OK, please try to avoid $\frac{28}{20}$. $\sqrt{5}$ is OK.

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。例：$\frac{7}{5}$も $1\frac{2}{5}$もよいが、$\frac{28}{20}$を避けてもらいたい。$\sqrt{5}$はよ い。

For Problems 1 to 5, use **Data Set A**. (Each student receives a different data set.) Note that the data set has the *true* standard deviation reported; you do not need to calculate it. Data Set A is a data set of the *monthly salary in 1000 yen* of a sample of new graduates of a university.

問題１～５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はある大学の新卒学生の月間給料（千円単位）。

Copy your data set in the space below: ここにデータを写ってください：

1. `[Problem ID #IDNO] applying statistics: two ways`

   What are the two ways that statistics can be used in research and business applications? Explain each briefly.

   研究やビジネスでは統計がよく使われている。主に２つの使い方がある。 その２つを書け。それぞれを説明すること。

   *Statistics may be used for **description** and for **inference**. **Descriptive statistics** are used to reduce large amounts of data to a small number of facts (usually numerical) that represent important information about the whole data set, such as averages. The basic tool is the (empirical) distribution, which summarizes data about large numbers of individual observations by counting how often each configuration of values occurs. **Statistical inference** is used to discover or confirm causal relationships based on the data, and in generating predictions about events that have not yet been observed. The basic tool is probability theory, which is used to calculate unbiased predictions and to estimate the accuracy of those predications based on the data.*

2. `[Problem ID #IDNO] data mining: definition`

   What is *data mining*, and why is it needed?

   □□□□□□□□□□□□□□□□□□□□□□□□□□□□ □□□□□□□□

   ***Data mining** is used to describe a variety of statistical techniques used to discover structure in (large) data sets. It can be used to efficiently update descriptive statistics or inferences based on large data sets as new data is*

*observed. It also can be used to group explanatory variables that have similar effect into **factors**, to infer the meaning of previously unobserved data values from context, and to accurately **classify** data according to imprecise criteria using easily measureable variables.*

3. `[Problem ID #IDNO] sampling: random`
   Is a random sample perfectly representative? Explain your answer.
   ランダムサンプルは完全に代表的サンプルになるか。説明せよ。

   A random sample (properly designed and implemented) is representative, but not perfectly so. Because it is random, the frequency of each value will vary if samples are taken repeatedly. Not all of these frequencies can perfectly match the underlying distribution, so the sample cannot be perfectly representative. On the other hand, a large random sample will tend to have a relative frequency distribution very close to that of the underlying distribution, including correlations with unobserved variables. This makes it suitable (representative) for statistical inference, because it eliminates the possibility of selection bias in the sample.

4. `[Problem ID #IDNO] independence: matrix game`
   Ann and Bob repeatedly play a simple game where Ann chooses Up or Down, and at the same time, Bob chooses Left or Right. The relative frequencies of each combination of choices observed is given in Table 1.
   アンちゃんとボッブ君がゲームで遊ぶ。アンは上(Up) か下(Down) を 選択し、ボッブは左(Left) と右(Right) を同時に選択する。観察した各 選択組の頻繁さを表 1 でそろえてある。

		Bob: Left	Bob: Right
Ann	Up	$\frac{1}{12}$	$\frac{1}{6}$
	Down	$\frac{1}{4}$	$\frac{1}{2}$

Table 1: Game of Ann and Bob

   (a) What is the relative frequency of Ann playing Up?
   アンが Up を選択する割合を計算せよ。
   $f(Up) = \frac{1}{12} + \frac{1}{6} = \frac{1}{4}$

(b) What is the relative frequency of Bob playing Left?
ボッブが Left を選択する割合を計算せよ。
$f(\textit{Left}) = \frac{1}{12} + \frac{1}{4} = \frac{1}{3}$

(c) Interpreting the relative frequencies as probabilities, are the Up and Left independent events?
割合を確立として解釈した場合には Up と Left は独立した事象になるか について答えて説明せよ。
Yes: $f(\textit{Up}) \times f(\textit{Left}) = \frac{1}{4} \times \frac{1}{3} = \frac{1}{12} = f(\textit{Up and Left})$

5. `[Problem ID #IDNO] hypothesis test: P-value and type II error`
Explain the relationship between the P-value of an hypothesis test and the Type II error of the test.
仮説検定におけるP値とタイプIIエラーの関係を説明せよ。

The P-value of a test is the probability of a Type II error, calculated by assuming the values of the parameters of the distribution of the test statistic are given by the null hypothesis H_0.

6. `[Problem ID #IDNO] hypothesis test: type I, II error`
Define the Type I error and Type II error of an hypothesis test.
仮説検定におけるタイプIエラーとタイプIIエラーの定義を書け。

A Type I error occurs when the null hypothesis H_0 of the test is false, but the test accepts H_0. A Type II error occurs when the null hypothesis H_0 of the test is true, but the test rejects H_0.

7. `[Problem ID #IDNO] hypothesis test: power computation`
Why is computation of the *power* of an hypothesis test (probability of accepting a false null hypothesis) more difficult than computation of the test's *significance* (probability of rejecting a true null hypothesis)? なぜ仮説検定のパワー（仮説が不実の場合に認められる確立）の計算が 有意性（仮説が事実の場合に認められない確立）の計算より難しいことを説明せよ。

The null hypothesis H_0 is defined by some scientific theory characterized by a given value for the parameter of the distribution of the test statistic. Computation of the significance *is therefore straightforward: simply compute the probability using that value of the parameter. Computation of the* power *is usually no more difficult as a numerical calculation, but theory doesn't give a value for the parameter, making choice of that parameter a challenge to the judgment of the statistician.*