2011年度前期　統計データ解析Ａ　第１４回
演習（４）　問題

１。下の記述は，統計学の観点からみて正しいかどうかを答えよ。正しいときは理由を説明せよ。正しくないときは，どういう点がどのように正しくないかを説明せよ。

> 母集団の平均の区間推定を行なう問題で，取り出した標本を使って計算した結果，「母平均の 95 パーセント信頼区間は 20 から 30 である」という結論を得た。このとき，母平均が 20 以上 30 以下である確率は 95 パーセントである。

２。X薬品の「Y」という薬は，１つ１グラムの錠剤となっている。いま，10 個の錠剤を無作為抽出し，各々の錠剤に含まれる物質Ｐの量を調べた。その結果，各錠剤の物質Ｐの量（ミリグラム）は次の通りであった。

1.2　0.8　0.9　0.9　1.0　1.3　1.2　1.0　0.8　0.9

1. 講義で説明した知識を使って，「薬Ｙに含まれる物質Ｐの割合」を区間推定するには，この測定がどのようなものであると仮定できる必要があるか。
2. 1. で答えた仮定が正しいとして，信頼係数 95% で 1. の区間推定を行なえ。

3。

1. ある工場では，「製品の強さが正規分布にしたがって分布するものとして，製品の強さの平均値の 95%信頼区間を求めたとき，その下限が 100kg 以上であること」という条件の製品の注文を受けた。そこで５個の試作品を作って強さを測定した結果，100, 105, 105, 110, 135(kg) という結果を得た。ところが，工場長は，データを 100, 105, 105, 105, 105(kg) のようにわざと低い数値に書き直して注文主に提出した。工場長は，なぜこのように「良いデータをわざわざ悪く書き換える」ような不正を行なったのか，その動機を工場長の立場になって説明せよ。
2. この例は問題のために作成したもので，実際には，このような不正な改ざんはもちろんしてはならない。では，工場長はどのような行動をとるべきだったか，簡単に述べよ。

解答例

１。母平均は，調査した人が知らないだけで，標本抽出には関係なく既に決まっている。だから，「20 以上 30 以下」という具体的区間を求めた段階で，母平均が「20 以上 30 以下」であるかどうかはすでに決まっている。「95% 信頼区間」とは，この方法で信頼区間を何度も求めると，そのうち 95% は母平均を本当に含んでいる区間である，という意味であり，「20 以上 30 以下」はそれらの信頼区間のうちのひとつにすぎない。

２。

1. 薬Ｙの各錠に含まれる物質Ｐの割合が，正規分布にしたがっていること。
2. 薬Ｙ全体での１グラムあたりの物質Ｐの量を μ，抽出された各錠剤での物質Ｐの量の平均を $\bar{X}$，その不偏分散を s^2，抽出された錠剤の数（標本サイズ）を n とする。このとき，問題文と 1. の仮定から，各錠剤が含む物質Ｐの量は，平均 μ の正規分布にしたがうと考えられる。したがって，

$$t = \frac{\bar{X} - \mu}{\frac{s^2}{n}} \tag{1}$$

という値（t 統計量）は自由度 $n-1$ の t 分布 $t(n-1)$ にしたがうので，$t_{0.025}(n-1)$ を $t(n-1)$ の上側 2.5 パーセント点とすると

$$P\left(\bar{X} - t_{0.025}(n-1)\sqrt{\frac{s^2}{n}} \leq \mu \leq \bar{X} + t_{0.025}(n-1)\sqrt{\frac{s^2}{n}}\right) = 0.95 \tag{2}$$

となるので，μ の 95%信頼区間の上限・下限は，(2) 式のかっこ内の不等式の上限・下限となる。
$\bar{X} = (1.2+0.8+\cdots+0.8+0.9)/10 = 1.0$，$s^2 = \{(1.2-1.0)^2+\cdots+(0.9-1.0)^2\}/(10-1) = 0.031$，$n = 10$，$t_{0.025}(9) = 2.262$ を (2) 式に代入すると，μ の 95%信頼区間は $[0.87, 1.13]$（ミリグラム）となる。したがって，物質Ｐの割合は $[0.087, 0.113](\%)$ となる。

３。

1. 製品の強さは正規分布にしたがうので，標本から製品の強さの平均値の信頼区間を求めるには t 分布にもとづく区間推定を用いる。書き換え前の標本については，

標本平均 $\bar{X} = (100 + 105 + 105 + 110 + 135)/5 = 111$,
不偏分散 $s^2 = ((100-111)^2 + (105-111)^2 + (105-111)^2 + (110-111)^2 + (135-111)^2)/(5-1) = 192.5$

である。また，標本サイズ $n = 5$ で，上側 2.5 パーセント点 $t_{0.025}(5-1) = 2.776$ である。このとき，製品の強さの平均値 μ について

$$P\left(\bar{X} - t_{0.025}(n-1)\cdot\sqrt{\frac{s^2}{n}} \geq \mu \geq \bar{X} + t_{0.025}(n-1)\cdot\sqrt{\frac{s^2}{n}}\right) = 0.95$$

がなりたつので，上の数値を代入すると，製品の強さの平均値の 95%信頼区間は $[93.775, 128.225]$ となる。一方，書き換えた後の標本について同様に計算すると，95%信頼区間は[101.224, 106.776] となる。

このように，データを書き換える前のほうが，「よい」値が入っているにもかかわらず値のばらつきが大きいために，不偏分散が大きくなり，信頼区間の下限が注文主の指定する 100kg を下回ってしまった。そこで，工場長はデータを書き換えてばらつきを小さくし，注文主の指定した範囲に信頼区間が入るようにみせかけたと思われる。

（正規分布の仮定がなりたっているということは，ヒストグラムが左右対称なわけですから，135kg という飛び抜けて大きな値が現れている以上，極端に小さな値も現れる可能性がある，ということになります）

2. 試作品が無作為抽出された標本であるならば，各試作品の強さも正規分布にしたがっているので，「135kg」という飛び抜けて大きな値が得られる確率は小さい。したがって，この場合「製品の強さが正規分布している」という仮定が成り立っていないと疑われる。製造工程になんらかの異常が発生している可能性があるので，製造工程を点検すべきである。