第１章　標本と母集団

問１　次の文章の中から誤っているものを１つ選びその番号を答えなさい
① 標本は母集団から抽出される。
② 系統抽出は最も代表性の高い標本を抽出する方法である。
③ 母集団の特性を知るための調査で抽出される標本の背後にある母集団の大きさはかならず有限である。
④ 繰り返し可能な実験の背後にある母集団は仮説的なものである。

正解　②
　系統抽出は代表性の高い標本を得る方法ではあるが、完全無作為抽出と比較してどちらが良いかは必ずしも言えない。

問２　正の値をとる確率変数 X,Y の期待値に関する次の式の中で、必ずしも成立しない場合があるものを選び、その番号を答えよ。
① $E[2X+3Y] = 2\,E[X]+3E[Y]$
② $E[X^2 +Y^2] \sim E[X^2] + E[Y^2]$
③ $E[\,2XY] = 2E[X]E[Y]$
④ $E[\log 2XY] = \log 2 + E[\log X] +E[\log Y]$

正解　③
　確率変数 X と Y とが独立でないと成立しない。
　X, Y が正の値しかとらなければ、$\log XY = \log X +\log Y$ と書けることに注意。

問３　A 部長と B 課長との財布の中に現金が幾ら入っているかを推定することにした。ただし、実際に幾ら持っているかはどちらか一人にしか聴けない状況がある。次の 4 つの方法の中で、両者の手持ち金額の合計の偏りの無い推定となっているのはどれか。番号を答えよ。

① B 課長の手持ち金額を調べ 2 倍する。
② A 部長は 1 万円、B 課長は 5000 円持っていると先ず事前に想定する。2 名を無作為に抽出し、A 部長を調べることになったら、その手持ち金額に 5000 円を足す。B 課長を調べることになったら、その金額に 1 万円を足す。
③ A 部長も B 課長もどちらも 5000 円持っていると先ず事前に想定する。2 名を無作為に抽出し、A 部長を調べることになったら、その手持ち金額に 5000 円を足す。B 課長を調べることになったら、その金額に 5000 円を足す。
④ 2 名を無作為に抽出し、どちらが抽出されてもその 2 倍を合計手持ち金額として推定する。

正解　④

A 部長の手持ち金額を X、B 課長の手持ち金額を Y とする。

① は２Y の不偏推定である。

② は、1/2 の確率で X+5000,1/2 の確率で Y+10000 となるので、その期待値は（X+Y）/2+7500

③ は（X+Y）/2+5000 が期待値である。

④ は、1/2 の確率で 2X,1/2 の確率で２Y となり、その期待値は両者の手持ち金額の合計（X+Y）と一致する。

第２章　正規分布とは

問１　【標準正規分布表が必要】

標準正規分布について説明している以下の文章中の空欄(1)～(4)に当てはまる正しい用語・数値について適切なものを、次の中から一つ選びなさい。

・標準正規分布に従うデータについて、平均 $\boxed{(1)}$ から±２の範囲に含まれるデータの割合は、$\boxed{(2)}$ ％である。
・標準正規分布表によると、$\boxed{(3)}$ 以上の値が生じる確率は、約 5%であることがわかる。また、$\boxed{(4)}$ 以下の値が生じる確率は、約 97.5%であることがわかる。

① (1)0　(2)68.3%　(3)1.96 (4)1.65
② (1)0　(2)95.4%　(3)1.65 (4)1.96
③ (1)1　(2)68.3%　(3)1.65 (4)1.96
④ (1)1　(2)95.4%　(3)1.96 (4)1.65

正解　②
・標準正規分布の 平均は０、分散は１である。
・標準正規分布では、以下が成立する。
　○平均０から±１の範囲に含まれるデータは 68.3%。
　○平均０から±２の範囲に含まれるデータは 95.4%。
　○平均０から±３の範囲に含まれるデータは 99.7%。
・標準正規分布表から、問題文に示されている上側確率（5%）に対応する点を探すと、1.65以上の値の生じる確率が約 5%であることがわかる。また、1.96 以上の値の生じる確率は 2.5%であることから、これを１から引くことで、1.96以下の値の生じる確率は 97.5%であることがわかる。

問２

２つの独立な確率変数 X_1、X_2 があり、それぞれ、標準正規分布に従っているとする。このとき、以下の文章中の空欄(1)～(3)に当てはまる正しい用語及び数値について、適切なものを、次の中から一つ選びなさい。

・X_1 と X_2 が両方とも、0 から±1 の範囲に生じる確率は、$\boxed{(1)}$ ％である。
・X_1 と X_2 のうち、一つだけが 0 から±1 の範囲に生じる確率は、$\boxed{(2)}$ ％である。
・X_1 と X_2 のうちの最大値が 1 以下である確率は、$\boxed{(3)}$ ％である。

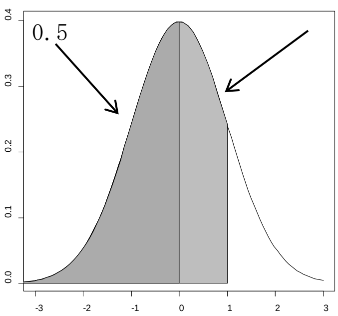
① (1)47%　(2)43%　(3)70%
② (1)47%　(2)46%　(3)60%
③ (1)33%　(2)49%　(3)70%
④ (1)33%　(2)52%　(3)60%

正解　①
・X_i（$i = 1,2$）が 0 から±１の範囲に生じる確率は、68.3%(0.683)である。
・このことから、X_1 と X_2 が両方とも、0 から±1 の範囲に生じる確率は、
　0.683×0.683 ＝ 0.466（≒ 47%）となる。
・X_1 と X_2 のうち、どちらか一つだけが 0 から±1 の範囲に生じるという事象に関しては、以下の２つの場合が考えられる。

（a）X_1 が±１の範囲に生じて、X_2 が±１の範囲の外に生じる。
（b）X_2 が±１の範囲に生じて、X_1 が±１の範囲の外に生じる

このとき、事象(a)と事象(b)の生じる確率は、どちらも
0.683×(1−0.683) = 0.217
となる。
ここで、事象(a)と事象(b)は、排反な事象であることから、２つの場合の確率を足すことで、以下のようにして求めることができる。
0.217 + 0.217 = 0.434（≒ 43%）

・X_1 と X_2 のうちの最大値が１以下になるということは、X_1 が１以下であり、かつ X_2 も１以下ということである。したがって、「X_1 と X_2 がともに１以下である確率」を求めると、それは、「最大値が１以下である確率」と等しいことになる。
・ここで、それぞれの X_i（$i = 1,2$）が１以下である確率は、
0.683 / 2 + 0.5 = 0.8415（84%）
となる（以下の図を参照。）。
※-1 から 1 までの面積が 0.683 であることから、図の右側（正）の部分の面積は、0.683 / 2 となる。



・したがって、X_1 と X_2 がともに１以下となる確率は、これらの積により、
0.8415 × 0.8415 = 0.708（≒ 70%）
となる。

問３　【標準正規分布表が必要】
ある野球選手の打率が３割であるとする。この選手が今シーズンの残り 20 打席で５本以上のヒットを打つ確率を、正規分布による近似（中心極限定理）を用いて求めた場合、適切なものを、次の中から一つ選びなさい。
なお、正規近似による確率の計算に当たっては、連続補正は行わなくてよいものとする。また、以下の計算結果を用いてもよいものとする。

X	4.0	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9
$\sqrt{X}$	2.000	2.025	2.049	2.074	2.098	2.121	2.145	2.168	2.191	2.214

X	5.0	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9
$\sqrt{X}$	2.236	2.258	2.280	2.302	2.324	2.345	2.366	2.387	2.408	2.429

① 27%
② 48%
③ 69%
④ 90%

正解　③
・提示された問題は、
「表の出る確率が 0.3 であるコインを 20 回投げたときに、表が５回以上出る確率を求める問題」
に置き換えることができる。
・そこで、このようなコイン投げの問題を考えるために、確率パラメータ $p = 0.3$、試行回数 $n = 20$ の２項分布に従う確率変数 Y を考えることとする。
・このとき、Y の平均 $E[Y]$ と分散 $V[Y]$（標準偏差 $STD[Y]$）を求めると、以下のようになる。

$$E[Y] = np = 20 \times 0.3 = 6$$
$$V[Y] = np(1-p) = 20 \times 0.3 \times 0.7 = 4.2$$
$$STD[Y] = \sqrt{V[Y]} = \sqrt{np(1-p)} = \sqrt{4.2} = 2.049$$

※問題文中にある表を用いた。

・このとき、以下のように Y を標準化した確率変数 Z は、中心極限定理により、近似的に、標準正規分布に従うとみなすことができる。

$$Z = \frac{Y - E[Y]}{STD[Y]} = \frac{Y-6}{2.049}$$

・以上を基に、Y が 5 以上の値をとる確率（$P(Y \geq 5)$）を、以下のように計算する。

$$P(Y \geq 5) = P\left(\frac{Y-6}{2.049} \geq \frac{5-6}{2.049}\right)$$
$$= P(Z \geq -0.49)$$
$$= 0.6879$$

第３章　様々な分布とその応用

問１

次の３つの現象に確率モデルをあてはめる際に最も適切であると思われる分布モデルの組み合わせはどれか。正しい組み合わせを選択肢の中から選べ。

現象(A)：支持率が 50%と想定される政策に対して、ランダムに選んだ 30 人の住民の中にいる反対者の数
現象(B)：ある交差点における１か月間の交通事故の数
現象(C)：サイコロを振ってでる六の目の数

理論分布（い）：二項分布
理論分布（ろ）：ポアソン分布
理論分布（は）：一様分布

①　(A) - (い)、(B) - (ろ)、(C) - (は)
②　(A) - (い)、(B) - (ろ)、(C) - (い)
③　(A) - (は)、(B) - (ろ)、(C) - (は)
④　(A) - (は)、(B) - (い)、(C) - (は)

正解　②

問２

古い住宅では火災報知器が未設置であるところもある。そこで、ある地域で築 10 年以上の住宅から 25 の住宅を無作為に抽出し、調査を行った。このとき、その地域の築 10 年以上の住宅における未設置率を 20%と想定した場合に、10 件以上の住宅が未設置である確率を求める式として適切なものを選択せよ。ここで、${}_{n}C_{k}$ は、nからkを選ぶ組み合わせの数を表す。

①　${}_{25}C_{10}0.8^{25}0.2^{10}$
②　$1-({}_{25}C_{0}0.8^{25}0.2^{0}+{}_{25}C_{1}0.8^{24}0.2^{1}+\cdots+{}_{25}C_{9}0.8^{16}0.2^{9})$
③　${}_{10}C_{25}0.8^{10}0.2^{15}$
④　${}_{25}C_{10}0.8^{25}0.2^{10}+{}_{25}C_{11}0.8^{25}0.2^{11}+\cdots+{}_{25}C_{10}0.8^{25}0.2^{25}$

正解　②

問３

　ある大学の学生の平均年間自殺率は0.01％である。現在、25,000人の学生が在学しているが、今年は既に４名の自殺者が出た。年間自殺率 0.01％とした場合に、この大学で４名以上の学生が自殺する確率を求めるエクセル関数の書式として、最も適切なものを選べ。

① BINOM.DIST（4,25000,0.01,0）
② POISSON.DIST(4,0.01,0)
③ POISSON.DIST(4,2.5,1)
④ 1－POISSON.DIST(3,2.5,1)

正解　④

第４章　標本調査と母集団推定

問１

標本抽出方法に関する以下の記述のうち、最も適切なものを一つ選べ。

① 非復元抽出法では、同じ対象が重複して選ばれることは決してない。
② 系統抽出法では、抽出間隔は必ず２以上としなければならない。
③ 層化抽出をするときには、層の大きさが等しくなるよう層化しなければならない。
④ 単純無作為抽出法で選んだ大きさ1,000 の標本よりも、集落抽出法で選んだ大きさ2,000 の標本の方が標本誤差は必ず小さい。

正解　①
① 復元抽出法では同じ対象が重複して選ばれる可能性があるが、非復元抽出法では同じ対象が重複して選ばれることはない。
② 最適な方法か否かはともかく、抽出間隔を1 として、連続したn の対象を標本としてもよい。
③ 例えば都道府県ごとの推定を行いたいときには都道府県を層とし、必ずしも層の大きさは等しい必要はない。
④ 集落抽出法は単純無作為抽出法よりも一般に標本誤差が大きくなりやすい。標本の大きさが大きいからといって、必ずしも標本誤差が小さいとは限らない。

問２

95 %信頼区間に関する以下の記述のうち、最も適切なものを一つ選べ。

① 95 %信頼区間とは、母数のちょうど95 %を含む区間のことである。
② 95 %信頼区間は、推定値の標準誤差に1.96 を乗じた値を、真の母数から引いた値と、真の母数に加えた値でできる区間のことである。
③ 95 %信頼区間は、データの標準偏差に1.96 を乗じた値を、推定値から引いた値と、推定値に加えた値でできる区間のことである。
④ 95 %信頼区間の中に、真の母数が含まれるとは限らない。

正解　④
① 「母数の95 %」は意味をなさない。
② 信頼区間は「真の母数」の周りに構成するのではなく、推定値の周りに構成する。
③ 区間の幅を求めるには「データの標準偏差」ではなく「推定値の標準誤差」を用いる。
④ 信頼区間の中に、真の母数が含まれるとは限らない。

問３

　ある県では、保育園に関する施策の充実のため、昨年度生まれた子供がいる10,000 世帯を母集団として、保育園に関するアンケート調査を行うことにした。予算の都合から標本調査とすることとしたが、入園希望の有無などといった割合に関する結果の標準誤差は、１パーセントポイントとしたい。何世帯程度を抽出すればよいか。以下の中から最も適切なものを一つ選べ。

① 500世帯程度
② 1,000世帯程度
③ 2,000世帯程度
④ 5,000世帯程度

正解　③

標準誤差は

$$SE(\hat{\pi}_y) = \sqrt{\frac{N-n}{N-1} \times \frac{\pi_y(1-\pi_y)}{n}}$$

であるが、母集団割合 π_y が分からないため、標準誤差が最大となる $\pi_y = 0.5 = 1/2$ とする。

$$SE(\hat{\pi}_y) = \sqrt{\frac{N-n}{N-1} \times \frac{0.5(1-0.5)}{n}} = \sqrt{\frac{N-n}{N-1} \times \frac{1}{4n}} = \sqrt{\frac{1}{4(N-1)} \times \left(\frac{N}{n} - 1\right)}$$

n に関して整理すると、次式が得られる。

$$n = \frac{N}{4(N-1)\{SE(\hat{\pi}_y)\}^2 + 1}$$

この式に $N = 10,000$ と $SE(\hat{\pi}_y) = 0.01$ を代入すると

$$n = \frac{10,000}{4 \times (10,000-1) \times 0.01^2 + 1} = \frac{10,000}{4.9996} = 2,000.16$$

第５章　公的統計の実際

問１

国勢調査による就業者数が経済センサスによる従業者数と一致しない理由のうち適切でない記述を、次のうちから一つ選びなさい。

① 調査日が異なる。
② 経済センサスからは農林漁業に属する個人経営の事業所が除外されている。
③ 経済センサスからは集合住宅の中にある事業所が除外されている。
④ 経済センサスでは複数の事業所で働いている人はその分計上される。

正解　③

国勢調査と経済センサスでは、いずれも就業状況を調査しているものの、就業者に関する両調査の結果に差があることから、その理由を問う問題である。

① 適切である。国勢調査は西暦末尾が 0 又は 5 の年 10 月 1 日において実施されるのに対し、経済センサスは他の年に行われる。
② 適切である。経済センサスの調査対象に農林漁業に属する個人経営の事業所は含まれていない。
③ 適切でない。外見からは住宅であっても事業所であれば調査対象である。
④ 適切である。同一人物が複数の事業所で働いている場合、国勢調査では一人として世帯で調査されるのに対し、経済センサスでは、それぞれの事業所で計上される。

問２

労働力調査のデータのチェックの仕方として適切でないものを次のうちから一つ選びなさい。

① 世帯主の男女の別に記入がなかったが、配偶者が女性の欄に記入をしているので世帯主は男性とみなした。
② おもに仕事をしている欄と仕事を探していた欄の両方に記入をしているが、国際基準に従えば、仕事をしている人は失業者とみなさないので、仕事を探していた欄の記入を削除した。
③ 月末１週間に仕事をしたかどうかの別に記入がなかったが、月末１週間に仕事をした日数に５と記入していたので、おもに仕事をしているとみなした。
④ ２年目の調査票の出生の年月に記入がなかったので１年目の調査票に書かれた内容と同じとみなした。

正解　②

① 適切である。夫婦の性別は逆であるという事実を適用した判断である。

② 適切でない。いずれの記入が正しいか判断がつけられる状況ではない。

③ 適切である。記入がなかった月末１週間の状況を他の情報で推論するものであり、仕事の日数からして適切な判断であるとみなすことができる。

④ 適切である。同一人物であれば生年月に違いはないという事実を適用した判断である。

問３

ある地域を 30 の調査区に分けて、1 つの調査区を抽出し、就業者数を調査したところ 60 人であった。地域の人口が 10,000 人、調査をした調査区の人口が 100 人のとき、地域の就業者数の人口による比推定の結果を次のうちから一つ選びなさい。

① 2,000 人

② 3,000 人

③ 3,333 人

④ 6,000 人

正解　④

人口に比例して就業者数を推定することになる。調査をした調査区の人口は 100 人、地域の人口は 10,000 人なので、地域の人口は調査区の人口の$\frac{10,000}{100}=100$倍である。したがって、地域の就業者数は、60×100＝6,000 人と推定され、④が正解である。