一番優しい、医薬品開発に必要な統計学の教本

統計検定２級の問題解説

2018 年 11 月の問題

2018年11月に実施された問題の解説です。
問題は、統計検定のHPにありますので、そちらからダウンロードしてください。
http://www.toukei-kentei.jp/about/grade2/

問１（１）

この問題で重要なことはただ一つ。
相対度数をすべて足すと、100%（割合の場合は１）になる
ということ。
これさえわかればOKです。

（ア）：100-(85.1+2.1) = 12.8%
（イ）：100-(76.6+17.0+2.1) = 4.3%

問１（２）

箱ひげ図の問題ですが、実は最大値に着目すれば解ける問題。
各年代の最大値は以下の通りですよね。
1952年の最大値：60校以上80校未満
1985年の最大値：100校以上120校未満
2017年の最大値：120校以上140校未満

ということで、箱ひげ図の縦軸のスケールに着目しましょう。
縦軸の一番大きい数字が70であるAが1952年
縦軸の一番大きい数字が100であるBが1985年
縦軸の一番大きい数字が140であるCが2017年
となります。

問１（３）

四分位範囲とは、全データを小さい（大きい）順に並べて、下（上）から25%の点と75%の点の間のことをいいます。
箱ひげ図では、箱の部分が四分位範囲。
そのため、A,B,Cの順で四分位範囲が大きくなっていることが分かります。

そのため、Iは×。

1952年の最大値は60校以上80校未満、1985年の最大値は100校以上120校未満ですから、1952年の最大値は1985年の最大値の半分以下、というのは間違いです。
そのため、IIは×

箱ひげ図において、中央値は箱の中の横線です。
そのため、A,B,Cの順に中央値が大きくなっています。
そのため、IIIは○。

問 2

この問題で重要な知識は、相関係数は直線関係の指標である、ということ。
つまり、二次曲線の関係があったとしても、相関係数の絶対値は小さくなります。
それを念頭に踏まえて。

“男性・正社員”では、50-54歳を頂点とした、放物線（二次曲線）の形をしています。前述の通り、相関係数は「直線関係」を表している指標です。そのため、このグラフは相関係数のみで判断してはいけません。
そのため、Iは〇。

“女性・正社員”についても、“男性・正社員”と同様に50-54歳を頂点とした放物線になっています。ですが、20-24歳から50-54歳までを見ると、比較的直線関係が見えています。ということであれば、69歳までの全体の相関係数よりも絶対値は大きくなるはず。
そのため、IIは×。

相関係数は、ｘが大きくなるとｙがどれだけ上がるか、ということを反映した指標ではありません。繰り返しになりますが、直線関係がどれほどあるか、という指標です。
そのため、IIIは×。
ちなみに、ｘが大きくなるとｙがどれだけ上がるか、ということを調べたい場合には、回帰分析を行います。

問３（１）

変化率は、以下の式で求めることができます。
前月からの変化率＝（求めたい月のデータ　-　前月のデータ）/前月のデータ×100
そのため、（ア）をxとすると、4.98=(111.7-x)/ x *100を求めればOKです。
ということで、106.4になります。

問３（２）

３項移動平均とは、軸となるデータ（今回の問題では2017年10月）と、その前後のデータの平均値です。
そのため、2017年9月、10月、11月の3つのデータの平均値になります。
つまり、(110.3+107.9+109.5）/3が正解です。

問 4

ラスパイレス指数の計算式を知っていれば解ける問題です。

$$\text{ラスバイレス指数} = \frac{\text{(比較年の個数×基準年の価格)を全て足したもの}}{\text{(基準年の個数×基準年の価格)を全て足したもの}} \times 100$$

つまり、今回の問題だとこのような式になります。

$$\text{ラスバイレス指数} = \frac{49.30 \times 3827 + 115.36 \times 2422}{48.86 \times 3827 + 107.09 \times 2422} \times 100$$

問 5

抽出に関する基本問題です。
I：◯
これは正しいです。
II：×
層の標本サイズが同じなら、層別しない抽出方法と同程度の分散を期待することができます。
ですが、標本サイズが違う場合には、平均値の分散は大きくなります。
III：◯
これは正しいです。

問 6

このような、抽出を何段階かに分けて実施することを多段抽出と呼びます。
この問題では、２段階（市区町村と世帯）の抽出をしているため、２段抽出と呼びます。

問７（１）

抽出した箱が工場Aからのものである確率は、0.7である。
更に、この箱にカモノハシの絵がプリントされているのは0.02である。
ということは、抽出した箱がAであり、更にカモノハシの絵がプリントされている確率は、0.7×0.02=0.014である。
同様に、抽出した箱がBであり、更にカモノハシの絵がプリントされている確率は、0.3×0.08=0.024である。
よって、抽出した箱にカモノハシの絵がプリントされている確率は
0.014+0.024=0.038となる。

問７（２）

条件付確率の定義は以下の通りです。

$$\mathrm{P}\left(\mathrm{A}\middle|\text{カモノハシ}\right)=\frac{P(A\cap\text{カモノハシ})}{P(\text{カモノハシ})}$$

そのため、求める確率は次の通り。

$$\mathrm{P}\left(\mathrm{A}\middle|\text{カモノハシ}\right)=\frac{0.0014}{0.038}=0.368$$

問８（１）

ちょっとだけ難しい問題です。
ですが、まずは与えられた問題の通りに計算していきます。

$$P(Y \geq 0) = 0.95$$
$$P(0.3 + 2x + U \geq 0) = 0.95$$
$$P(U \geq -2x - 0.3) = 0.95$$

で、ここからが頭を使うのですが、Uは平均0、分散1の正規分布に従います。
つまり、正規分布表を確認して、95%よりも大きくなるUの点を探します。
すると、Uが-1.64以上となる確率は95%と確認できます。
そのため、以下の等式を解けばよいということが分かります。

$$-2x - 0.3 = -1.64$$

よって、$x = 0.67$となります。

問８（２）

問題文より、$Y = 0.3 + 2x + U$です。
95%点はUが1.64の時なので、$Y = 0.3 + 2x + 1.64 = 2x + 1.94$となります。
つまり、一次方程式であることがわかりました。
直線関係のグラフは1なので、答えは１です。

問９（１）

計算が面倒なので、ミスに注意です。
まず、2以下の目が出る確率は1/3であり、それ以外の目が出る確率は2/3です。
7回サイコロを振って$x+1$回だけ2以下の目が出る確率は、以下の通りです。

$$ {}_7C_{x+1}\frac{1}{3}^{x+1}\frac{2}{3}^{7-(x+1)} $$

同様にして、x回だけ2以下の目が出る確率は、以下の通りです。

$$ {}_7C_x\frac{1}{3}^{x}\frac{2}{3}^{7-x} $$

これを式展開すると、以下の通りになります。

$$ \frac{P(X=x+1)}{P(X=x)}=\frac{-x+7}{2x+2} $$

問９（２）

（１）で求めた式に0～7を代入してみます。
すると、ｘ＝２の時に最大になります。

問 10

標本平均の期待値と、標準誤差の知識があれば一発で回答できます。
標本平均の期待値は、確率変数の期待値と一緒であるため、（ア）はμ。
標準誤差（標本平均の標準偏差）はσ/sprt(n)となるため、分散はその2乗。
つまり、σ2/nとなる。

問 11（１）

歪度と尖度の問題です。
どちらも、正規分布の場合には 0 になります。

問 11（２）

結構な難問でした。
確率変数Xが一様分布$U(a,b)$に従う時、平均値と分散は以下の定義です。

$$\text{平均値} = \frac{a+b}{2}$$

$$\text{分散} = \frac{(b-a)^2}{12}$$

よって、一様分布$U(-1,1)$では、平均値が0,分散が1/3となります。
よって、問題文の歪度の定義から以下のように求めることができます。

$$\frac{E[(X-\mu)^3]}{\sigma^3} = \frac{E[(X-0)^3]}{\sigma^3} = \frac{E[(X)^3]}{\sigma^3} = \frac{\int_{-1}^{1} x^3 f(x)dx}{\sigma^3} = \frac{0}{\sigma^3} = 0$$

また、尖度も同様に計算すると、-1.2となります。（計算式は割愛します）

問 11（３）

I：×
逆です。
右に裾が長い分布では、歪度は正の値になり、左に裾が長い分布では、歪度は負の値になります。
II：×
こちらも逆です。
正規分布よりも尖っている分布では尖度は正の値に、正規分布よりも中心部が平坦な分布では尖度は負の値になります。
III：×

自由度が大きくなるにつれて、t分布は正規分布に近づきます。すなわち、尖度は0に近づくので、絶対値は小さくなります。

問 12

比率の信頼区間。
統計検定２級では、かなりの頻度で出題されます。
割合をpとして、信頼区間は以下の式で計算できます。(暗記しておいていいレベルです)

$$P \pm 1.96\times\sqrt{\frac{p(1-p)}{n}}$$

今回のデータを当てはめると、以下のようになります。

$$0.02 \pm 1.96\times\sqrt{\frac{0.02(1-0.02)}{1338}} = 0.02 \pm 0.008$$

問 13

T 統計量の基礎問題です。
サンプルサイズを n、母平均を μ、標本平均を$\bar{x}$、不偏分散をs^2とすると、t 統計量は次の式から求められます。

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}}$$

よって、t 統計量は値を代入して以下のようになります。

$$t = \frac{85.6 - 90.0}{\sqrt{\frac{121.9}{20}}} = 1.78$$

両側 5%ということは、片側 2.5%であるため、T 分布表から自由度 20-1=19 の棄却域は 2.093 です。
そのため、棄却域＞T 統計量なので棄却されません。

問 14（１）

「分散が等しいかどうか」の検定は、F 検定です。
覚えていただきたいのは 2 点。
「F 統計量の式」と「自由度」です。
2 群の不偏分散をそれぞれs_1^2, s_2^2とすると、F 統計量は以下の式で算出できます。

$$\mathrm{F} = \frac{s_1^2}{s_2^2}$$

また、このとき 2 群のサンプル数をn_1,n_2とすると、自由度は以下の通りです。

$$(m_1, m_2) = (n_1 - 1, n_2 - 1)$$

そのため、$(m_1, m_2) = (29,30)$となります。

問 14（２）

検定の多重性の問題です。
３つの検定をして「１つでも有意であればよい」ときの確率を求めます。
１つでも有意であれば良い、ということを直接計算するとややこしいので、逆転の発想をします。
それは、1-（１つも有意にならない確率）を求めるということです。
１つの検定のαエラーを 5%とした時、有意にならない確率は 1-0.05 です。
つまり、以下の式を解けば良いです。

$$1 - (1 - 0.05) \times (1 - 0.05) \times (1 - 0.05) = 0.14$$

問 15（１）

「不良品、不良品でない」という２値のデータですので、二項分布が当てはまります。
二項分布は統計検定２級では必ず出題されますので、必ず理解しておきましょう。
二項分布の平均値はnpで求めることができ、分散は$np(1-p)$で求めることができます。
つまり、平均値は$np = 200\times0.05 = 10$であり、分散は$np(1-p) = 200\times0.05\times 0.95 = 9.5$となります。

問 15（２）

標本平均を$\hat{p}$、母比率をp_0、サンプル数をnとすると、以下の式から求めることができる z 統計量は、標準正規分布に従います。

$$\mathrm{z} = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$

200 個のうち、16 個が不良品なので、標本平均は 16/200 となります。
そのため、z 統計量は以下の通りです。

$$\mathrm{z} = \frac{\dfrac{16}{200} - 0.05}{\sqrt{\dfrac{0.05(1-0.05)}{200}}} = 1.95$$

標準正規分布表を見ると、1.95 より大きくなる確率は 0.026 であることがわかる。

問 15（3）

２つの群の標本比率をそれぞれ$\widehat{p_1}$,$\widehat{p_2}$とし、サンプルサイズをn_1,n_2とすると、次の式から求められる z 統計量は、標準正規分布に従う。

$$z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$$

ただし、$\hat{p}$はプールした標本比率のこと。

$$\hat{p} = \frac{n_1\widehat{p_1} + n_2\widehat{p_2}}{n_1 + n_2}$$

問題文より、A 社の標本比率は 16/200=0.08 であり、B 社の標本比率は 17/200=0.085 です。そのため、プールした標本比率は、以下の通り。

$$\hat{p} = \frac{200\times0.08 + 200\times0.085}{200 + 200} = 0.0825$$

よって、z は以下の通り。

$$z = \frac{0.08 - 0.085}{\sqrt{0.0825(1-0.0825)(\frac{1}{200}+\frac{1}{200})}} = -0.18$$

標準正規分布より、-0.18 より小さくなる確率は 0.43 であるため、両側検定ではその 2 倍である、0.86 が P 値になります。

問 16（1）

適合度検定において、 カイ 2 乗統計量は以下の通り。

$$\chi^2 = \frac{(\text{データ} - \text{理論値})^2 \text{の総和}}{\text{理論値}}$$

よって、これを満たしているのは 1 のみ。

問 16（2）

カイ二乗検定の自由度は、カテゴリの数-1 であるため、5 です。
自由度 5 のカイ二乗分布における上側 5%点は 11.07 となります。
１つ前の問題で 1 がカイ二乗統計量でしたので、2.59<11.07 より、棄却されない、という結果になります。

問 17（１）

回帰分析の読み取り方です。
国の数（データの数）を読み取るには、自由度を読み取る必要があります。
自由度は degree of freedom もしくは、DF と表記されます。
２通りの読み取り方があります。

１つ目は、Residual standard error の degree of freedom に着目する方法。
Residual standard error とは、残差の標準誤差のことです。
この時の 52 degree of freedom は、（データの数-説明変数の数-1）から算出されますので、52=X-2-1 が成り立ちます。
そのため、X＝55 となります。

２つ目は、F-statistic の DF に着目する方法です。
「2 and 52 DF」は、「説明変数の数 and 残差 DF」を示しています。
そのため同様に、52=X-2-1 の関係式が成り立ちますので、X＝55 となります。

問 17（２）

I：×
α の推定値は Intercept の行です。
説明変数のないパラメータは日本語では「切片」と呼び、統計ソフトでは Intercept と表示されますので、覚えましょう。
そのため、α の標準誤差は 113.7 です。
II：◯
ここで重要なのが e の読み取り方です。
統計ソフトの出力で e+は 10 倍を示していますので、例えば e+02 であれば、10 倍の 10 倍、つまり 100 倍するということです。
一方で、e-は 1/10 するということを示していますので、例えば e-02 であれば 1/100 するということです。
それを念頭におくと、全てのパラメータで 0.05 を下回ることがわかりますので、

有意水準 5%で 0 ではないという結果になります。
III：×
自由度調整済み決定係数は、Adjusted R-squared を見ます。
すると、0.8141 ですので、間違いです。
自由度を調整しない決定係数は、Multiple R-squared になります。

問 17（３）

I：◯
人口密度（Population）の点推定値（Estimate）を見ると、マイナスの値になっています。
そのため、人口密度が高い国では、自動車普及率は低い傾向がある。
II：◯
１人当たり GDP の点推定値（Estimate）を見ると、プラスの値になっています。
そのため、１人当たり GDP が高い国では、自動車普及率は高い傾向がある。
III：◯
得られた推定値を回帰式に代入すると、450 となる。

問 18（1）

I：◯
残差の標準誤差 ^{2}=残差平方和/(サンプル数-説明変数の数-1)という関係を知っているかどうか、という問題でした。
この方程式を解くと、$0.608^2 \times (5-1-1)=1.1$ となります。
II：×
単位を変えても、t 値は変わりません。
III：◯
単位を変えると、推定値は変わります。

問 18（2）

I：×
上記の問題と同様に、推定値は単位を変えれば変わります。
そのため、説明変数が不要かどうかは、推定値が 0 に近いかどうは関係ありません。
一般的には、P 値を確認して判断します。
II：×
説明変数間の相関が高い場合には、多重共線性の問題が発生します。
標本サイズは関係ありません。
III：×
得られた P 値（0.559）>有意水準（0.05）であるため、棄却されません。

問 18（3）

I：×
説明変数の数が異なれば、得られる推定値は異なります。
II：◯
これはその通りです。
III：×
有意ではないため、x が 1 万円大きい時 y が 6.462 万円小さくなる、ということ

はできません。