(金曜 2 講時, 尾畑)

数理統計学概要・レポート課題１ (2020.06)

- レポート原稿をどのような形式で準備するかは自由. WORD 等を利用して準備してもよいし, 手書きでも構わない. また, それらの混合でもよい.
- 提出にあたっては, 作成したレポートを 1 個の PDF ファイルにまとめること. したがって, 手書き原稿をスキャンする場合は最終的に PDF に変換すること.
- WORD, EXCEL, PowerPoint のファイルや JPEG などの画像ファイルは受理しません.
- 提出先：ISTU にアップロード（投稿）する.
- 提出期限：2020 年 6 月 18 日 23 時 59 分
- 提出方法に困難がある場合は別途申し出ること. 期限を延長することはしない.

[1] 「オリンピック代表選手.xlsx」にはロンドン五輪 (2012) へ派遣された日本, 中国, 米国選手の身長や体重等のデータが収められている. このデータを統計処理して次のことを考察せよ. ただし, 収録されているデータには欠損（エクセルのセルが空欄）があるので, そのようなデータは除外して扱うこと.

(1) 選手の身長には男女の違いや国別の違いはあるか? 統計量を計算して比較せよ.

(2) 身長と体重には相関があるか? 散布図からわかることを述べよ.

(3) 身長と体重の相関係数について男女の違いや国別の違いはあるか? 計算して比較せよ.

[2] サイコロを 2 個投げるとき, 出る目の和を X, 積を Y とする. 答は既約分数で記せ.

(1) X の確率分布, 平均, 分散, 標準偏差を求めよ.

(2) Y の確率分布, 平均, 分散, 標準偏差を求めよ.

[3] ある国では, 病気 A の感染者が $100q$ % あるという $(0 < q < 1)$. 検査 B は, 感染者の 90%に陽性反応を示すが, 非感染者の 5% にも陽性反応が出てしまう.

(1) $q = 0.04$ とする. この検査を受けて陽性反応が出た人が感染者である確率を求めよ. 答は適当な桁で四捨五入した小数で記せ.

(2) $q = 0.1$ とする. この検査を受けて陰性反応が出た人が非感染者である確率を求めよ. 答は適当な桁で四捨五入した小数で記せ.

(3) $0 < q < 1$ として, この検査を受けて陽性反応が出た人が感染者である確率 P を求めよ. この P が q とともにどのように変化するかグラフを示せ. その変化の特徴からこの確率 P を現実問題に適用する際の注意を述べよ.

(金曜 2 講時, 尾畑)

数理統計学概要・レポート課題２ (2020.07.07)

- [1]–[3] のうち 2 題を選択解答せよ. (3 題解答しても評価しません.)
- レポート原稿をどのような形式で準備するかは自由. WORD 等を利用して準備してもよいし, 手書きでも構わない. また, それらの混合でもよい.
- 提出にあたっては, 作成したレポートを 1 個の PDF ファイルにまとめること. したがって, 手書き原稿をスキャンする場合は最終的に PDF に変換すること. WORD, EXCEL, PowerPoint のファイルや JPEG などの画像ファイルは受理しません.
- 提出先：ISTU にアップロード（投稿）する.
- 提出期限：2020 年 7 月 22 日 23 時 59 分
- 提出方法に困難がある場合は別途申し出ること. 期限を延長することはしない.

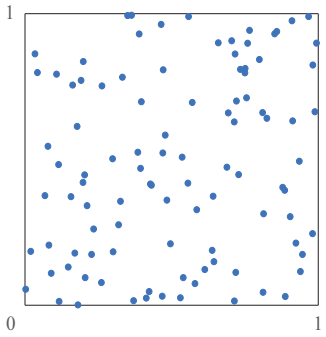
[1] エクセルには $0 < x < 1$ を満たす一様乱数を発生するコマンド rand() がある. これを利用することで, 正方形 $\{(x, y) \,|\, 0 \leq x \leq 1, 0 \leq y \leq 1\}$ 内にランダムに点を打つことができる. たとえば, 100 個のランダム点を打ち出したものを例示してある (左図). 座標平面において, 放物線 $y = x^2$, 直線 $x = 1$, および x 軸によって囲まれる図形を D とする (右図). 正方形に打ったランダム点のうち, D 内に落ちる点の個数を数えることで, D の面積を求めたい.

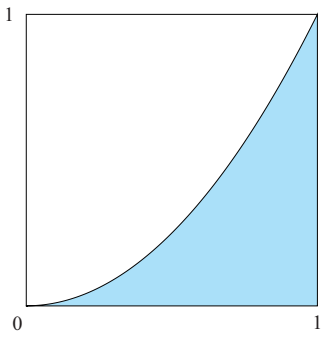
(1) D の面積を積分によって小数第 2 位まで求めよ.

(2) 正方形に多数のランダム点を打ち, D に落ちる点の個数の割合をもって D の面積の近似値とできることを数学的根拠をもって説明せよ.

(3) ランダム点の個数を 100 個として, D の面積をどの程度近似できるか. 多数の実験を繰り返して考察せよ. また, 理論的にも説明せよ.

(4) D の面積を小数第 2 位まで厳密値に一致させるためにはランダム点を何個取ればよいか考察せよ.





[2] 双六に動機づけられて, サイコロを繰り返し投げて 100 マスをできるだけ早く (少ない回数で) 塗りつぶすゲームを考案した. このとき, 次の戦略 A, B のうちどちらが有利かを考察せよ. 様々な観点がありうるので, そこも評価する.

(A) サイコロを 1 個投げて出目の 2 倍の個数のマス目を 1 回で塗りつぶすことができる.

(B) サイコロを 2 個投げて出目の和の個数のマス目を 1 回で塗りつぶすことができる.

[3] 次の数値列は正規母集団 (分布が正規分布であるような母集団) から取り出された無作為標本である.

39.2 12.7 41.8 43.9 2.6 41.4 22.0 7.3 43.9 27.4 45.2 37.3

(1) 母分散が $\sigma^2 = 10^2$ であるとき, 母平均の 95% 信頼区間を求めよ.

(2) 母分散が未知であるとき, 母平均の 95% 信頼区間を求めよ.

(金曜 2 講時, 尾畑)

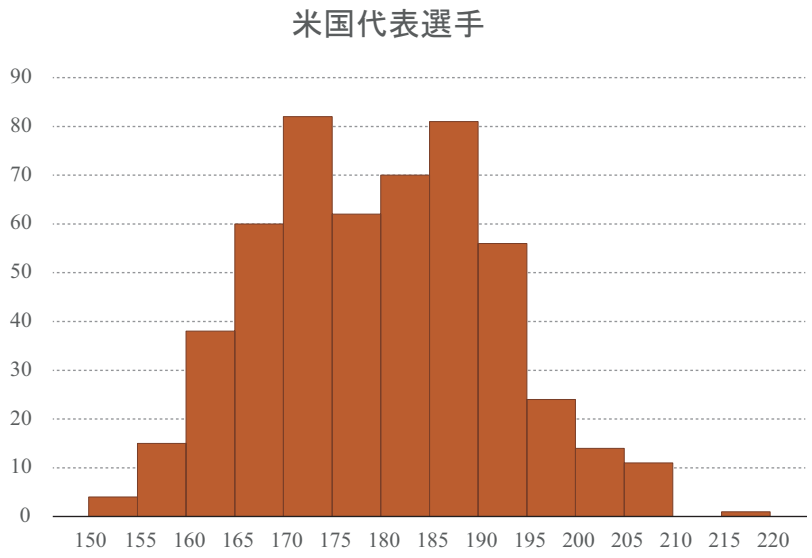
数理統計学概要・レポート課題３ (2020)

- レポート原稿をどのような形式で準備するかは自由. WORD 等を利用して準備してもよいし, 手書きでも構わない. また, それらの混合でもよい.
- 提出にあたっては, 作成したレポートを 1 個の PDF ファイルにまとめること. したがって, 手書き原稿をスキャンする場合は最終的に PDF に変換すること.
- WORD, EXCEL, PowerPoint のファイルや JPEG などの画像ファイルは受理しません.
- 提出先および期限：2020 年 8 月 16 日 23 時 59 分 までに ISTU にアップロードされていること. これ以降は理由の如何によらず受け取りません.
- 解答にあたり, ロンドン五輪 (2012) へ派遣された日本, 中国, 米国選手の身長や体重等のデータ「オリンピック代表選手.xlsx」を用いよ.
- 当然ですが, 解答には統計学的 (より一般には科学的) な論拠が必要であり, 視覚的な説明や単なる数値の比較は評価しない.
- 参考文献を明記せよ.

[1]] 教科書には「ヒトの身長や体重をはじめ, 動物の身体的特徴の統計を取ると正規分布になる」というような記述がしばしば見られる. 米国選手の身長のデータ (518 個) から

平均値：179.0(cm), メディアン：178(cm), 標準偏差 12.1(cm)

のように計算され, ヒストグラムは下図のようになった.

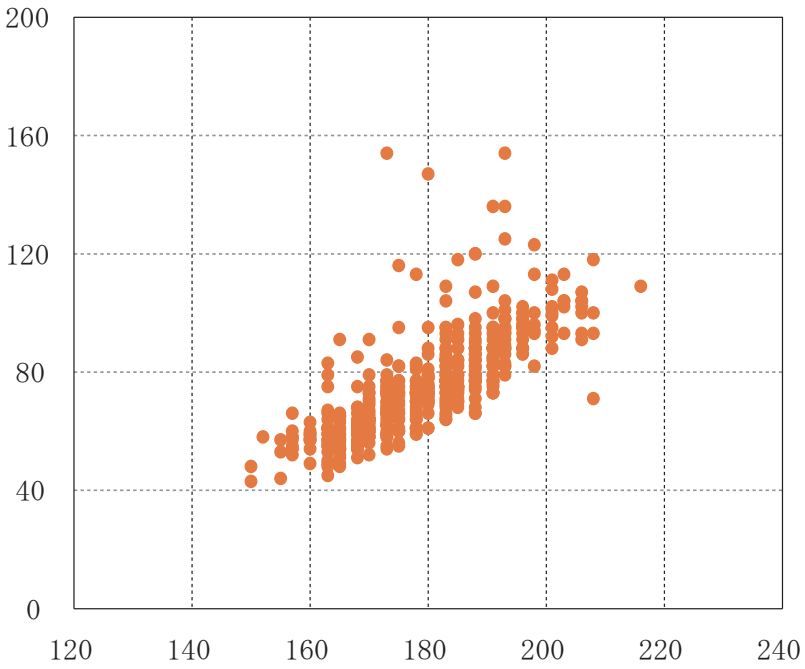


(1) 上のヒストグラムから米国選手の身長は正規分布になっているとは思われない. 正規分布ではないという論拠を示せ.

(2) 教科書の記述に反しているのは, 「米国選手のデータ」を全部用いたことが適切でなかったと考えられる. 調査対象を区分けすることで教科書の記述を支持することができるか考察せよ.

(3) そもそも「動物の身体的特徴の統計を取ると正規分布になる」という記述は標語としてはよいが, 一般的に適用するのは危険である. より適切と思われる記述に書き直して, その論拠を示せ.

[2] 米国選手の身長と体重のデータ (488 個) をもとに散布図を作った. ただし, 横軸が身長 (cm), 縦軸が体重 (kg) である.



(1) 相関関係について相関係数も合わせて述べよ.

(2) 身長が高いほど体重のばらつきが大きくなるといえるか? 数値的, 統計的に述べよ.

[3] 次のデータは, ロンドン五輪 (2012) へ派遣された男子選手 10 人の身長を測定したものである. この 10 人は日本, 中国, 米国いずれかの代表団からランダムに選ばれたものである. どの国から選ばれたものであるか仮説検定などを用いて考察せよ.

164 180 160 190 186 178 185 212 186 200

[4] 与えられたデータ「オリンピック代表選手.xlsx」を用いて, 国別の統計的特徴を考察せよ. 単なる数値の比較は評価しない.

レポート課題 1　講評

[1] 国別性別に身長と体重に関する基本的な統計量を計算して, ヒストグラムや散布図を図示することが主な課題である. 算出した統計量の大小を単に比較して, どちらが大きいというだけでは統計的な分析としては不十分であるが, そのような話題 (母集団の比較) は講義の後半で扱う.

(1) 小数点以下の桁数に無頓着なのはいけない. 元のデータに対して 1-2 桁多くとる程度として, 全体を通して統一する.

(2) データの個数を明記する.

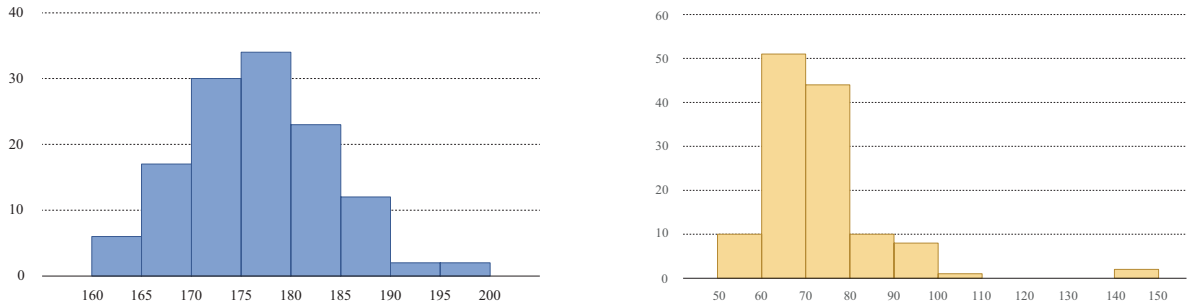
(3) 平均値の計算に必要だからといってデータ値の総和を記載する意味はない. むしろ記載してはいけない.

(4) 散布図は縦軸と横軸を調整して, 表示される点が団子状態ではなくある程度ばらつくようにする.

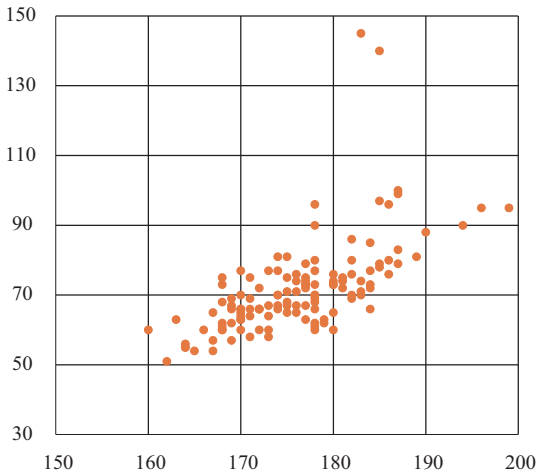
(5) 相関係数に関連して回帰直線による比較をするのもよい.

統計量のまとめ方を「日本代表の男子」を例にとって示しておく. ただし, ここに示したのは最も基本的な統計量に限っており, 別の代表値や散布度を用いて様々に検討するのがなお良い.

London 2012 日本代表男子	
標本数	126
身長の平均値 (cm)	176.2
身長の標準偏差 (cm)	7.2
体重の平均値 (kg)	71.9
体重の標準偏差 (kg)	13.4
身長と体重の相関係数	0.63



身長（左図）と体重（右図）



なお, 体重が著しく重い (外れ値といってよい) 標本が 2 個含まれており, 体重の標準偏差を大きくずらし, 身長と体重の相関を弱める効果が出ている. 実際, 2 個の外れ値を除いて計算した結果と比較するとよい. 外れ値に意味を見出すのか, 外れ値を除いた統計量に意味を見出すかは目的 (文脈) による.

London 2012 日本代表男子	
標本数	124
身長の平均値 (cm)	176.1
身長の標準偏差 (cm)	7.2
体重の平均値 (kg)	70.8
体重の標準偏差 (kg)	10.1
身長と体重の相関係数	0.73

[2] (1) X の確率分布

k	2	3	4	5	6	7	8	9	10	11	12
$P(X=k)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

$$\mathbf{E}[X] = \sum_{k=2}^{12} kP(X=k) = 7,$$
$$\mathbf{E}[X^2] = \sum_{k=2}^{12} k^2P(X=k) = \frac{1974}{36},$$
$$\mathbf{V}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2 = \frac{1974}{36} - 7^2 = \frac{210}{36} = \frac{35}{6},$$
$$\sqrt{\mathbf{V}[X]} = \frac{\sqrt{210}}{6}.$$

(2) Y の確率分布

k	1	2	3	4	5	6	8	9	10
$P(X=k)$	1/36	2/36	2/36	3/36	2/36	4/36	2/36	1/36	2/36

k	12	15	16	18	20	24	25	30	36
$P(X=k)$	4/36	1/36	2/36	2/36	2/36	2/36	1/36	2/36	1/36

$$\mathbf{E}[Y] = \sum_{k} kP(Y=k) = \frac{441}{36} = \frac{49}{4},$$
$$\mathbf{E}[Y^2] = \sum_{k} k^2P(Y=k) = \frac{8281}{36},$$
$$\mathbf{V}[Y] = \mathbf{E}[Y^2] - \mathbf{E}[Y]^2 = \frac{8281}{36} - \Big(\frac{49}{4}\Big)^2 = \frac{11515}{144},$$
$$\sqrt{\mathbf{V}[Y]} = \frac{\sqrt{11515}}{12}.$$

(注意) 平均値や分散の計算では, 実は, 1 個目のサイコロを A, 2 個目のサイコロを B として,

$$X = A + B, \qquad Y = AB$$

を考える方が簡単. このとき, A と B が独立である事にも注意する. まず, サイコロ 1 個の場合に,

$$\mathbf{E}[A] = \sum_{k=1}^{6} kP(A=k) = \frac{7}{2},$$
$$\mathbf{E}[A^2] = \sum_{1=2}^{6} k^2 P(A=k) = \frac{91}{6},$$
$$\mathbf{V}[A] = \mathbf{E}[A^2] - \mathbf{E}[A]^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

を確認しておく. これらは, $\mathbf{E}[B], \mathbf{E}[B^2], \mathbf{V}[B]$ と一致する. そうすると, $X = A + B$ に対しては,

$$\mathbf{E}[X] = \mathbf{E}[A+B] = \mathbf{E}[A] + \mathbf{E}[B] = 2 \times \frac{7}{2} = 7,$$
$$\mathbf{V}[X] = \mathbf{V}[A+B] = \mathbf{V}[A] + \mathbf{V}[B] = 2 \times \frac{35}{12} = \frac{35}{6}.$$

$Y = AB$ に対しては,

$$\mathbf{E}[Y] = \mathbf{E}[AB] = \mathbf{E}[A]\mathbf{E}[B] = \left(\frac{7}{2}\right)^2 = \frac{49}{4}.$$

また,

$$\mathbf{E}[Y^2] = \mathbf{E}[A^2B^2] = \mathbf{E}[A^2]\mathbf{E}[B^2] = \left(\frac{91}{6}\right)^2 = \frac{8281}{36}.$$

したがって,

$$\mathbf{V}[Y] = \mathbf{V}[Y^2] - \mathbf{V}[Y]^2 = \frac{8281}{36} - \left(\frac{49}{4}\right)^2 = \frac{11515}{144}.$$

[3] (1) ベイズの公式によって,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

である. 与えられた条件から,

$$P(A|B) = \frac{0.9q}{0.9q + 0.05(1-q)} = \frac{18q}{1+17q}$$

となる. $q = 0.04$ を代入して, $P(A|B) = 0.429$ となる.

(2) も同様である.

$$P(A^c|B^c) = \frac{P(B^c|A^c)P(A^c)}{P(B^c|A^c)P(A^c) + P(B^c|A)P(A)}$$

である. 与えられた条件から,

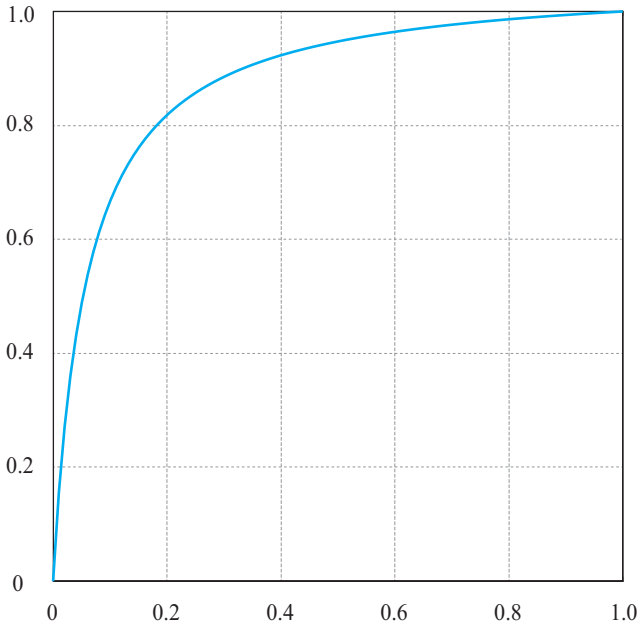
$$P(A^c|B^c) = \frac{0.95(1-q)}{0.95(1-q) + 0.1q} = \frac{19-19q}{19-17q}$$

となる. $q = 0.1$ を代入して, $P(A^c|B^c) = 0.988$ となる.

(3) では (1) で求めた

$$P(A|B) = \frac{18q}{1+17q}$$

のグラフを描いて考察すればよい.



ポイントは, q によって, $P(A|B)$ は 0 から 1 まで変化すること. したがって, q がわからない状況で, 適当に設定する (たとえば, わからないから半々でのように考えて $q = 1/2$ とおく) ことは危険である. $P(A|B)$ が大きいほど効率の良い検査ということができるが, 感染者の割合 q が小さいと効率は低い.

レポート課題２ 講評

[1] 一様乱数を 2 個発生させて, それらを x 座標, y 座標とする点を座標平面に打つことを考える. その点の両座標を X, Y とすると, それらは独立な確率変数であり, $[0,1]$ 上の一様分布に従う. その点が, 正方形内の小さな長方形の領域 $A = \{a \le x \le b, c \le y \le d\}$ に落ちる確率は,

$$P(a \le X \le b, c \le Y \le d) = P(a \le X \le b)P(c \le Y \le d) = (b-a)(d-c)$$

となり, その長方形の面積に一致する. 問題の領域 D は小長方形の和で近似できるので, ランダム点が D 内に落ちる確率は D の面積 $1/3$ に一致する. こうして, 1 個 1 個のランダム点が D に落ちるかどうかは, 成功確率 $1/3$ のベルヌイ試行 (コイン投げ) と同じである.

ランダム点を n 個打った時, D に落ちる点の個数を S とする. 大数の法則から

$$P\Big(\lim_{n\to\infty} \frac{S}{n} = \frac{1}{3}\Big) = 1$$

であるから, n を大きくとれば, D に落ちたランダム点の相対度数が $1/3$, つまり D の面積に近づく. どのくらいの精度で近づくかは中心極限定理 (二項分布の正規分布近似) によって見積もることができる. つまり,

$$S \sim B\Big(n, \frac{1}{3}\Big) \approx N\Big(\frac{n}{3}, \frac{2n}{9}\Big)$$

であることから, 相対度数は,

$$\frac{S}{n} \sim N\Big(\frac{1}{3}, \frac{2}{9n}\Big)$$

に従うとしてよい.

厳密値 $1/3 = 0.333\ldots$ は既知であるとして, 小数第 3 位まで厳密値を再現するのに要するランダム点の個数 n を求めよう. もちろん, 確率的な話になるので, たとえば, 99% の確率で再現するならば,

$$P\Big(0.3325 \le \frac{S}{n} < 0.3335\Big) = 0.99$$

を満たす n を求めればよい. いつも通り, 標準化して, $Z \sim N(0,1)$ とすれば,

$$P\Big(\frac{0.3325 - 1/3}{\sqrt{2/9n}} \le Z < \frac{0.3335 - 1/3}{\sqrt{2/9n}}\Big) = 0.99$$

ここで,

$$a = \frac{0.3335 - 1/3}{\sqrt{2/9n}}$$

とおけば,

$$\frac{0.3325 - 1/3}{\sqrt{2/9n}} = -5a$$

であるから,

$$P(-5a \le Z < a) = 0.99$$

となる a をまず求めればよい. $0.99 = P(-5a \le Z < a) \approx P(Z < a)$ から a は両側 2%点となり, $a = 2.33$ となる. そうすると,

$$\frac{0.3335 - 1/3}{\sqrt{2/9n}} = 2.33$$

から $n = 43431200$. 小数第 2 位まで厳密値を再現するのなら, サンプル数は $1/100$ でよい.

より一般の図形の面積を推定する文脈では面積 $0 < p < 1$ は未知である. このとき,

$$\frac{S}{n} \sim N\Big(p, \frac{p(1-p)}{n}\Big)$$

から始める. 小数第 3 位まで厳密値を再現するというのを誤差を ± 0.0005 に抑えるという意味に解釈して, それが 99% の確率で保証されるということは, ランダム点の個数を n とすれば,

$$P\Big(\Big|\frac{S}{n} - p\Big| < 0.0005\Big) = 0.99.$$

標準化して, $Z \sim N(0,1)$ とすれば,

$$P\Big(|Z| < \frac{0.0005}{\sqrt{p(1-p)/n}}\Big) = 0.99$$

ここで, $P(|Z| < 2.58) = 0.99$ と比較して,

$$2.58 = \frac{0.0005}{\sqrt{p(1-p)/n}} \quad \Leftrightarrow \quad \sqrt{n} = \frac{2.58}{0.0005}\sqrt{p(1-p)}$$

また $p(1-p) \le 1/4$ を用いて,

$$n = \left(\frac{2.58}{0.0005}\frac{1}{2}\right)^2 = 6656400$$

小数第 2 位まで厳密値を再現するのなら, サンプル数は $1/100$ でよい.

[2] 戦略 A で 1 回に塗りつぶすことのできるマス目の個数を X とすれば, X の確率分布は

k	2	4	6	8	10	12
$P(X=k)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

戦略 B で 1 回に塗りつぶすことのできるマス目の個数を Y とすれば, Y の確率分布は

k	2	3	4	5	6	7	8	9	10	11	12
$P(Y=k)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

簡単な計算で $\mathbf{E}[X] = \mathbf{E}[Y] = 7$ がわかる. したがって, 1 回あたり塗りつぶすことのできるマス目の個数の平均値, つまり平均速度はともに 7 となる. ゴールのない競争であり, ある時点でどちらが先行しているかだけを見るのであれば, 戦略 A,B で平均速度は同じなので優劣はない. さらに, X, Y の分散に注目するのは自然な考え方である. 実際,

$$\mathbf{V}[X] = \frac{70}{6} = 3.416^2, \qquad \mathbf{V}[Y] = \frac{35}{6} = 2.415^2$$

である. このことから戦略 A の方がギャンブル性が大きい, つまり, 大きくリードしたり, 大きく出遅れたりしやすいことがわかる.

しかしながら, 双六にはゴールがあり, どちらが早くゴールするかという問題ははるかに複雑である. 100 マスを塗りつぶすのに要する回数（時間と言ってもよい）が重要であるが, 平均速度 7 をもとに,

$$\frac{100}{7} = 14.2857$$

をゴールに要する平均時間と言いたいところであるが, これは間違い. たとえば, 100 マスではなく 6 マスの双六を考えてみよう. 平均速度が 7 だからと言って,

$$\frac{6}{7} = 0.8571$$

がゴールするまでの平均時間であろうか? 明らかに違う. なぜなら, ゴールするためには 1 回以上サイコロを振らなければならないから, 平均時間は明らかに > 1 のはずだ. 実際, 6 マス競争において, 戦略 A,B によってゴールに要する時間をそれぞれ T_A, T_B とすると, 組合せ数を数えることで,

$$\mathbf{E}[T_A] = \frac{49}{36} = 1.3611, \qquad \mathbf{E}[T_B] = \frac{1661}{1296} = 1.2816,$$

がわかる. したがって, 戦略 B の方がゴールするまでの平均時間は短い.

100 マスの場合を厳密値を与える公式を導出することは困難であり, ある種の漸化式 (下記) を用いて数値的に求めるか, サイコロ振りのシミュレーションを多数回繰り返してゴールに要する時間の分布を求めることになる. 漸化式について, 戦略 A によって x マス進んだ時点でゴールの 100 マスまでに要する平均時間を A_x とすると,

$$A_x = \frac{1}{6}A_{x+2} + \frac{1}{6}A_{x+4} + \cdots + \frac{1}{6}A_{x+12} + 1$$

がわかる (正しくは, マルコフ連鎖の考え方が必要だが, 直感的にも理解できるであろう). この漸化式を

$$A_{100} = A_{101} = A_{102} = \cdots = 0$$

を条件として逆向きに解いて, A_0 を求める. これが戦略 A でゴールに要する平均時間を与える:

$$\mathbf{E}[T_A] = A_0 = 14.7619.$$

同様に, 戦略 B によって x マス進んだ時点でゴールの 100 マスまでに要する平均時間を B_x とすると,

$$B_x = \frac{1}{36}\,B_{x+2} + \frac{2}{36}\,B_{x+3} + \frac{3}{36}\,B_{x+4} + \cdots + \frac{1}{36}\,B_{x+12} + 1$$

であり,

$$\mathbf{E}[T_B] = B_0 = 14.7738$$

が得られる. したがって, 戦略 A の方がわずかに有利である. シミュレーションによる方法では, Python や C++ でコードを書いて, 1000 回程度のシミュレーションで優劣を見出したり, シミュレーションを 100 万回程度まで行って上記の厳密値を再現している強者もあり大変感心した.

※ 実は, ゴールまでの距離 G によって戦略 A,B の優劣 (ゴールまでの平均時間) が変化する. 今回 $G = 100$ を問題にして, 戦略 A が有利としたが, $G = 99$ なら戦略 B が有利である. 数値計算で $G \leq 100$ まで確認したところ, ゴールまでの距離 $G = 1, 2$ なら両戦略に違いはなく, G が偶数で $G \neq 4, 6, 14, 16$ ならば戦略 A の方が有利, それ以外は戦略 B の方が有利である. $G \geq 17$ では G が奇数なら戦略 B, G が偶数なら戦略 A の方が有利であると一般に言えそうであるが, 証明にはちょっと時間を要しそうなので中断している.

[3] 信頼区間の公式を適用するだけである. まず, 標本の大きさ, 平均値と不偏分散

$$n = 12, \qquad \mu = 30.39, \qquad u^2 = 15.547^2$$

を求めておく.

(1) 標準正規分布 $N(0,1)$ の両側 5%点は $z(0.05) = 1.96$. したがって,

$$30.39 \pm z(0.05)\frac{\sigma}{\sqrt{n}} = 30.39 \pm 5.66$$

(2) 自由度 11 の t 分布の両側 5%点は $t_{11}(0.05) = 2.201$. したがって,

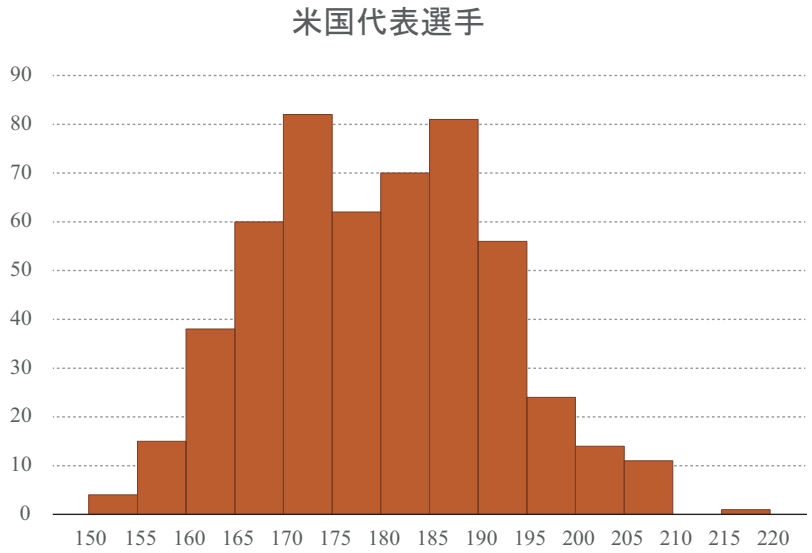
$$30.39 \pm t_{11}(0.05)\frac{u}{\sqrt{n}} = 30.39 \pm 9.88$$

レポート課題３ 講評

[1] (1) 正規分布の基本的な特徴は, 平均値, メディアン, モードが一致することと密度関数（ヒストグラム）が平均値に関して線対称になることがあげられる. 問題文に示したように, 米国代表選手の平均値とメディアンはおおむね一致しているが, ヒストグラムを見るとモードが大きくずれている（階級を超えている）. さらに, モード（正確には分布のピーク）は明確に 2 つある. これらから, 米国代表選手の身長は正規分布に従っているとは言えない.

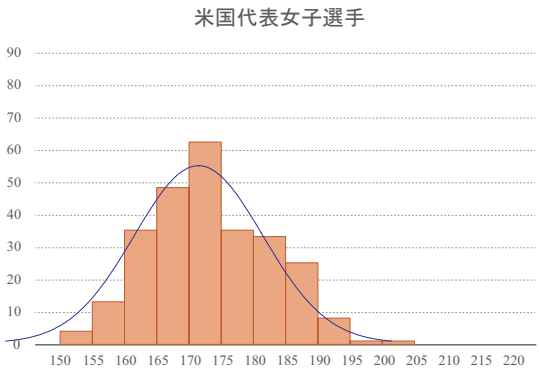
※ 2 つのピークに言及せずとも, 平均値の前後で分布が単調増加から単調減少に変化するかに注目してもよい. 分布が平均値に関して線対称になるというだけではダメ.

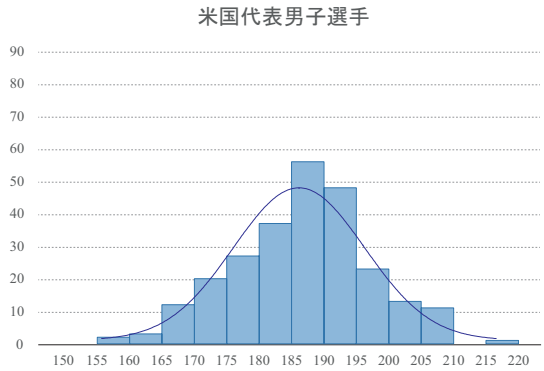
※ 実データに対する考察であるから, 数学のような厳密さを追求するのは筋違いである. たとえば, 横軸が正規分布の密度関数のグラフの漸近線であるというのは数学的事実であるが, データが有限個しかない実データに対して漸近線になっていないと主張するのは意味がない.



(2) 米国代表選手ヒストグラムを見るとピークが 2 つあり, うっすらと 2 つの正規分布が重なっているように見えるところが出発点である. そこで, 調査対象が男女混合であって, 男女には生来の体格差があるので, 男女に分けてみようと思う（数理統計学以外の知識を活用）.

	女子	男子
標本数	265	253
平均値	172.47	185.94
標準偏差	9.56	10.53





男女別にヒストグラムを見れば, 対称性が崩れているようには見えるが, ともに正規分布により近い形になっているだろう. 同じ平均値と分散をもつ正規分布の密度関数を重ねてみると, だいたい合っているように見える. さらに, 米国代表選手ヒストグラムに見られた 2 つのピークは, 男女それぞれの平均値に相当していることがわかるので, 米国代表選手が正規分布になっていないのは, 異なる平均値をもつ 2 つの正規分布の混合であると結論することに合理性はある. (ここでは目視レベルの考察で可とする.)

※ 数理統計学には正規分布の適合度を判定する手法がさまざま用意されている. 授業の最後の課題であったカイ 2 乗検定はその一つであり, それを試みた解答も多数あった (自由度の決め方に注意を要する). なお, 有意水準 5%が絶対的な基準ではないので注意! 有意水準 5%は練習のために採用した基準であると認識せよ.

一例を記しておく. 米国選手に対して 150cm から 5cm 刻みとし, 205cm 以上は 1 つにまとめた 12 階級としてカイ 2 乗値を求めると $\chi^2 = 22.74$ となる. 自由度 $9 = 12 - 3$ のカイ 2 乗分布を用いて P 値を求めると 0.0068 となる. この P 値は, 米国選手が正規母集団 $N(179.05, 12.09^2)$ に従うと仮定して, そこから 518 個の標本を取り出したとき, 問題にあるようなヒストグラム（および正規分布からもっとずれているもの）が得られる確率を意味する. 通常, この小さな P 値によって, 米国選手の身長が正規分布に従っていないと結論付ける. 同様に, 男子選手については自由度 7 のカイ 2 乗分布を用いて $P = 0.254$, 女子選手については自由度 6 のカイ 2 乗分布を用いて $P = 0.044$ が得られる. 男子は正規分布への適合性が認められるが, 女子はなお認めがたいというのがカイ 2 乗検定の結論となる.

※ 男女別のヒストグラムに同じ平均値と分散をもつ正規分布の密度関数を重ねた図から, 適合度検定を用いないで正規分布に適合していると結論付けるところは各研究分野における知識や経験が必要である. 知識のない者が図から適合性を判断したとすれば, それは主観であり科学にならないが, 本題では適合度検定までは要求せず, 観察のレベルまでで良しとした.

※ さらに, 自学自習によって正規分布への適合度を別の方法 (正規確率紙 = Q-Q プロット, コルモゴロフ-スミルノフ検定, シャピロ-ウィルク検定など) で試みている意欲的なものもあり感心した. 正規確率紙は計算機のない時代から使われている伝統的手法であり, エクセルで割と簡単に作成できるので試みるとよい. 一種の散布図を作って, データが直線に乗っているかどうかを目視で判定するというものである. 目視なので, 職人芸が必要となり客観性に難がある. 検定を用いれば, 確率的な指標によって適合性が判定されるのでより合理的である. どの検定が優れているかは一概には言えない.

(3) 考えている量が, 小さな揺らぎが独立に積み重なっているとみなされる状況では, その測定値が平均値の周りに正規分布で揺らぐ. 数学的には中心極限定理として証明される. たとえば, 細胞の長さが多数の遺伝子（のオン・オフ）で決まっているとすれば, 実際に測定される細胞の長さは二項分布のようになり, それは正規分布で近似される. さらに, 成長の過程ではさまざまな環境の影響があるだろう. 多数の独立な環境要因の積み重ねで細胞の長さが決まるのであれば, 中心極限定理が（数学で言う厳密性は欠けるが）適用できる. 動物の身体的特徴を決める要因は複雑であるが, 遺伝子と成長する環境に類似性がある個体を多数集めれば, 正規分布に近くなることが予想される. 本題の米国代表選手の身長については, まず男女は遺伝的に大きく異なるので, 分離することで正規分布に近い分布が得られたと考えられる.

※ 男女を区別した後に, さらに種目別に考えるのも面白い. オリンピック選手であっても身長は自分で調整できないので自然の揺らぎに任せるしかないと思われるが, 一方, バスケットボールなどでは高身長の選手ばかりを集めているだろう. そのような効果が正規分布に反映しているかどうか調べてもよい. ただし, 標本数が十分ないと結論を得るのは難しい.

※ 一方, オリンピック選手ともなれば, 体重制限のある種目でなくとも, 体重は自分で調整していると推察されるので, 身長とは違った統計的特徴が見えるかもしれない.

※ 数学の主張である中心極限定理の前提を満たすような測定環境は現実には整備しがたいし, 無限回の実験や観察を繰り返すことは不可能である. したがって, あくまで「近似的に」正規分布になっていることを議論するのである. どの程度のずれを許容するかは, 物理学, 生命科学, 社会科学などの分野によって大きく異なる.

[2] 相関係数は 0.745 となり, 強めの正の相関と言える. 散布図を見ると, 身長が高いほど体重のばらつきが大きいように見えるが, これを数値的に確認するには, 身長で区分けして, 各グループ内で体重のばらつきを数値化して比較すればよい. 一例を示す.

区分	165 以下	165〜175	175〜185	185〜195	195 超	計
標本数	74	130	136	99	49	488
標準偏差	7.88	11.91	11.78	14.23	9.07	
相関係数	0.285	0.321	0.282	0.328	0.226	

概ね, 身長が高いほど体重の標準偏差が大きくなると言える. 195 超のグループでは標本数が少ないことを考慮して除外してもよいし, 185 超としてまとめてもよい. あるいは, 195cm までは体重の標準偏差が大きくなるが, それを越すと小さくなるという観察でもよい.

※ 各グループに対して相関係数を計算しても, 体重のばらつきをうまく反映する量とは言えない. 今の問題では, 相関係数は身長の変化に伴う体重の変化を表すものになるが, 身長の変域を限定しているため, その範囲だけの散布図は団子状態になりやすい. 実際, 相関係数が小さくなっている. 団子が右上がりに連なることで, 全体として大きな相関係数 0.745 が得られている.

※ グループ分けの仕方は自由であるが, 同身長のデータを別のグループに配分するのはまずい.

※ 本題の結論「身長が高いほど体重のばらつきが大きくなる」に対して肯定・否定のいずれでもよい. そもそも「ばらつき」をどのように計量するかは一意的ではなく, それを表す統計的なきちんとした指標（たとえば分散）を設定して結果を主張しているかがポイント.

[3] 10 個のデータの平均値は $\bar{x} = 184.1$ である. これをもとに, 日本, 中国, 米国の男子選手からなる 3 つの母集団のいずれから抽出されたものであるかを仮説検定で論じる.

その準備として, 各母集団の平均値と標準偏差を計算しておく. （各国の男子代表選手を母集団として扱うので, その大きさは以下の議論では不要であるが, 計算根拠として記してある.）

	日本	中国	米国
母集団の大きさ	135	145	253
平均値	175.79	182.54	185.94
標準偏差	7.48	11.12	10. 55

(1) 日本の男子選手からの無作為標本とみなせるかを仮説検定で判断する. 仮説を

$$H_0 : \mu = 175.79, \qquad H_1 : \mu \neq 175.79$$

として, 有意水準を $\alpha = 0.05$ とする.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{10}} = \frac{\bar{X} - 175.79}{7.48/\sqrt{10}} \sim N(0, 1)$$

を用いる. 実現値は

$$z = \frac{184.1 - 175.79}{7.48/\sqrt{10}} = 3.51 > 1.96$$

となり, 5%棄却域に落ちる. もっと言えば, > 2.58 なので 1%棄却域に落ちる. したがって, 有意水準 1% の両側検定によって, 得られた 10 個の標本は日本男子の母集団から選ばれたものではないと言える.

※ 題意から H_0 が棄却された場合, 平均値の大きい米国または中国から選ばれたことになるので, 実際は片側検定がより適していると言える. 結果は, たとえば「有意水準 1% の片側検定によって, 得られた 10 個の標本は日本男子の母集団から選ばれたものではない」となる. 実際, 片側の P 値は 0.00022 という僅少な確率である. 自信をもって棄却できる.

(2) 中国の男子選手からの無作為標本とみなせるかを仮説検定で判断する. 仮説を

$$H_0 : \mu = 182.54, \qquad H_1 : \mu \neq 182.54$$

として, あとの議論は (1) と同様である. 実現値は

$$z = \frac{184.1 - 182.54}{11.12/\sqrt{10}} = 0.44 < 1.96$$

となり, 有意水準 5%の両側検定で棄却されない.

(3) 米国の男子選手からの無作為標本とみなせるかを仮説検定で判断する. 仮説を

$$H_0 : \mu = 185.94, \qquad H_1 : \mu \neq 185.94$$

として, あとの議論は (1) と同様である. 実現値は

$$z = \frac{184.1 - 185.94}{10.55/\sqrt{10}} = -0.55 > -1.96$$

となり, 有意水準 5%の両側検定で棄却されない.

仮説検定によって, 日本の選手でないことは結論される. 中国か米国の判断は, 実現値 z を比較して, 中国である確率が大きいことがわかる.

※ 仮説検定（母分散既知の場合）を主眼とした課題である. 日本の男子選手には身長 212cm の者はいないことなどを観察して, 10 個のデータの由来を特定するという（決定論的な）アイデアもあるが, ちょっと趣旨が違う. 出題の仕方が良くないと言われればその通りであるが. 実際, 10 個のデータは中国から乱数を用いて抽出したものであった.

※ 問題文にあるように, 男子選手を扱っているので, 日本, 中国, 米国とも男子選手だけを取り出して母集団とするべきで男女混合で扱ってはいけない. さらに, そうすることで [1] で考察したように母集団が正規分布に従うとしてよい.

※ 母分散が既知である場合は, t 検定より正規分布による検定 (Z 検定) のほうが優れている. なお, Z 検定では標本数が大きければ, 任意の母集団分布を扱うことができる（中心極限定理）.

[4] 自由に見解を述べて, 統計的な議論をすればよい. BMI を比較するなど医学部学生ならではの視点もあり感心した。本来ならグループディスカッションをしてもらうところであるが, 今学期は断念せざるを得ない.