NO. 1

DATE

基礎統計シケプリ

- 教科書に準じて、問題を解くときに必要なことだけでまとめていきます。
  テストは<u>持込不可</u>なので、公式等も覚記が必要です。

- 次のページから本文ですが、第一章はなにも内容のないただのイントロなのでやりません。
  第二章からやっていきます。

[テスト情報]

- 重要語句（教科書太字で、かつ、講義資料で強調されているもの）の穴埋め問題がある可能性㊇

- 計算量は少なめにする（データの個数を減らす）予定。電卓も持込不可。

必ず出る
- 母分散既知のときの母平均 μ の推定（100α% 信頼区間）→ $\overline{x} - Z_{1-\alpha} \cdot \left(\frac{\sigma}{\sqrt{n}}\right) \leqq \mu \leqq \overline{x} + Z_{1-\alpha}\left(\frac{\sigma}{\sqrt{n}}\right)$
- 〃 未知 〃 （ 〃 ）→ $\overline{x} - t_{1-\alpha(n-1)} \cdot \left(\frac{s}{\sqrt{n}}\right) \leqq \mu \leqq \overline{x} - t_{1-\alpha(n-1)} \cdot \left(\frac{s}{\sqrt{n}}\right)$

[注意]

最終的に重要なテーマは
- （正規分布に関する）推定 → その中で、t 分布、χ^2 分布を利用
- （正規分布に関する）仮説検定 → 〃

です。

また、テストは上記の2つに加え
- 確率（ベイズの定理 or 独立と無相関）

は出題されると予想されます。（あくまで予想です。）

他は、わかりません。

ですから、最初の方は、ある程度流して、9〜12章を理解するようにして下さい。
9〜12 をやりつつ、前を見るかんじでいい と思います。

ただし、9〜12章は結構難しくて、私もあまりわからなかったので、このシケプリもちょっと雑でわかりにくいと思います。この辺は、適宜もう一方の素晴らしいシケプリを参照して下さい。

一応断っておきますが、本シケプリの内容に対して、まちがってても一切責任は負いません。

（by. 2010 17組 会計）

NO. 2.

DATE

第2章. 1次元のデータ (P17~P40)

※"1次元"とは、データが1種類ということです。例えば、A~Z氏について
身長のデータのみ ⇒ 1次元
身長、体重 ⇒ 2 〃 etc.

度数分布表とヒストグラム (P18~P20)

集められた膨大なデータをある程度のまとまりに分けて表やグラフにすることでみやすくする。

Ex) P18の表より。

表 2.1† 試験得点の度数分布表(某大学の統計学)

階級		階級値	度数	相対度数	累積度数	累積相対度数
0点以上	10点未満	5	12	0.032	12	0.032
10 〃	20 〃	15	10	0.027	22	0.059
20 〃	30 〃	25	19	0.051	41	0.110
30 〃	40 〃	35	42	0.113	83	0.223
40 〃	50 〃	45	40	0.193	155	0.416
50 〃	60 〃	55	82	0.220	237	0.635
60 〃	70 〃	65	54	0.145	291	0.780
70 〃	80 〃	75	38	0.102	329	0.882
80 〃	90 〃	85	25	0.067	354	0.949
90 〃	100点以下	95	19	0.051	373	1.000
合計			373	1.000		

試験得点の分布において、つねにこのような整った結果が得られるとはかぎらない。

階級: 観測値をある程度のまとまりにわけたもの ここでは10点ごとにわけている。

階級値: 階級の代表値。基本的には階級のまん中の値

度数: その階級にあるデータの個数。12人が0~10点をとったということ。

相対度数: 全体(ここでは373人)を1として、どれくらいの割合のデータがその階級にあるか。 $\frac{(度数)}{(全体)}$

累積度数: そこまでの度数の合計

累積相対度数: 累積度数 ÷ 全体

P19のグラフより

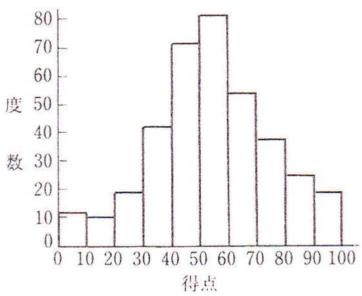


図 2.1 試験得点のヒストグラム
度数分布がこのように整った形となることはむしろめずらしい。階級数、階級幅などをうまくとらねばならない。

- 「階級」「度数」「相対度数」などの意味は、ほぼあきらかだろう。

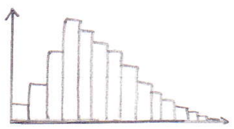
- "階級の幅をどうするか?"は案外悩ましい問題である。(広すぎるとヒストグラムが変わったりする)
  → 1つの解決策は スタージェスの公式。
  観測値が n コのとき、階級の数 k は
  $k \fallingdotseq 1 + \log_2 n$ くらいがよい。 ← 参考程度。

- 階級の数を決める際には、階級値がキレイな値になるようにするのもよい

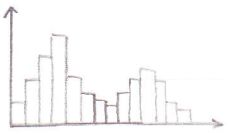
※ 度数分布表はたぶん読めれば十分です。このページは適当に飛ばして下さい。

● ヒストグラムの読みとり (P20~26)

○ 単峰型と双峰型

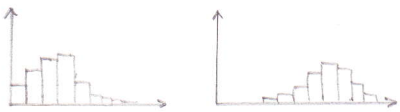


単峰型



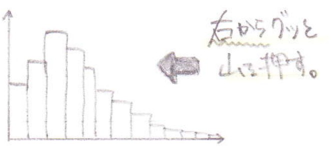
双峰型

↳ 2種類の異なる条件のグループがまざっているのでは？
(例えば、男性と女性を一緒に考えている など)
⇓ グループ分け [=層別 という] する

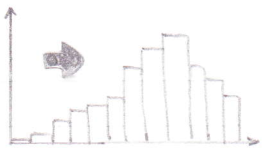


2つの単峰型に分けられる。

○ 歪んだ分布



右に歪んだ分布



左に歪んだ分布

● P26~28 までの内容「ローレンツ曲線」「測定の尺度」は省略します。
余裕のある方は、ご自分で。(範囲外とかではなく、面倒なので パスするということ)

● "歪んだ分布"に関しては、次の「代表値」のところと多少絡みます。

● 最初や最後の階級が開いている (テストの点数：0~100 ← とじている / 所得 ：0~たくさん ← ひらいている)
ときは、その部分は点線でてきとうにかく

※ ヒストグラムをかかせることはあまり考えられないので、このページはだいたいでいいでしょう

NO. 4.

DATE . .

ここからが最初の山場です。難しくはないので、しっかり理解して覚えましょう。

代表値 (P28～34)

◦ 平均 (算術平均) : mean

ただの平均です。$\bar{x}$ で表す。観測値のとる値を $v_1, v_2 \cdots$, その度数を $f_1, f_2, \cdots$ とすると

$$\bar{x} = \frac{f_1 \cdot v_1 + f_2 v_2 + \cdots + f_k \cdot v_k}{f_1 + f_2 + \cdots + f_k}$$ (←常識)

※個々のデータがわからないとき (つまり、度数分布表のとき) は、v_1, v_2 を階級値にすればよい (←常識)

※開いているデータは、ムシする / だいたいの階級値をとる etc. かなり適当に対処する。

※他にも、幾何平均 : $x_G = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n}$

調和 〃 : $\frac{1}{x_H} = \frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}\right)$ などがある。

◦ メディアン (中央値) : median

要は、nコのデータを小さい順に並べたときの、$\frac{n+1}{2}$ 番目のデータの値。

nが偶なら両側の値の平均

ex. 右の10コの数字 : 1. 1. 1. 1. 2. 3. 4. 5. 16. 20 で、平均は5.4、メディアンは 2.5

→ 平均は 5.4 だが、感覚的には、1～2とか、それくらいが代表的な値。

つまり、この場合、平均より メディアン がわかりやすい。

拡張⇓

◦ パーセンタイル (分位点)

観測値 (←上でいう数字) を、小さい順に並べて、

下から 25% のところ ⇒ 第1分位点 Q_1

〃 50% 〃 ⇒ 第2 〃 Q_2 (= メディアン)

〃 75% 〃 ⇒ 第3 〃 Q_3

◦ モード (最頻値) : mode

要は、いちばん度数の大きいところの値。上の ex では もちろん 1。

• では、右に歪んだ分布では、上の3つの関係はどうなるか。

→ 平均 > メディアン > モード。 になる (アルファベット順)

↙これも難しくないが重要。後々使うことになります。

散らばりの尺度（P35〜39）

以下のデータ A.B.C を考える

A: 0. 3. 3. 5. 5. 5. 5. 7. 7. 10

B: 0. 1. 2. 3. 5. 5. 7. 8. 9. 10

C: 3. 4. 4. 5. 5. 5. 5. 6. 6. 7

これらのデータはすべて　平均，メディアン，モードは5である。しかし分布は異なる。
→「散らばりぐあい」の尺度は，次のようなものがある。

(i) レンジ (range)

分布の存在する範囲をいう。Aなら0〜10，Bも0〜10，Cは3〜7
こんなのだれもつかわない。ムシしてよい。

(ii) 四分位偏差

$Q = \frac{1}{2}(Q_3 - Q_1)$　つまり　25%ラインと75%ラインの差を2で割ったもの。
0〜25%，75%〜100% という両端を除くことで，外れ値（←異質すぎるデータ。上の例でもし100とかが入っていればそれは外れ値）の影響を消した。
これもあまりよろしくない。

(iii) 平均偏差 → 省略。たぶん出ないでしょう。式は $\underset{\text{平均偏差}}{d} = \frac{1}{n}\{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \cdots + |x_n - \bar{x}|\}$

(iv) 標準偏差と分散

各観測値が平均からどれだけ離れているかを偏差という。
分散は S^2 という記号で表され

$$\boxed{S^2 = \frac{1}{n}\cdot\{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2\} = \frac{1}{n}\sum(x_i-\bar{x})^2}$$

$$\left(= \frac{1}{n}\left(x_1^2 + x_2^2 + \cdots + x_n^2 - 2\bar{x}\cdot\sum_{k=1}^{n} x_k + \bar{x}^2\right)\right)$$

$$\therefore \boxed{S^2 = \frac{1}{n}(x_1^2 + x_2^2 + \cdots + x_n^2) - \bar{x}^2 = (\text{2乗した値の平均}) - (\text{平均の2乗})}$$

もとのデータと単位をそろえるためルートをとって，標準偏差 S を得る

$$\boxed{S = \sqrt{S^2}}$$

※度数分布表から求めるとき（個々のデータの値がわからないとき）は，平均を求めるときと同様に，1つの階級ではすべて同じ値として求めればよい。

NO. 6

DATE

- 変動係数

分布の中心の位置が異なるときは、標準偏差でもって分布の散らばりぐあいを比較できない。
(∵ 各値が大きくなれば、標準偏差も大きくなり、たとえ散らばりが小さくても、大きな値になる)
よって 変動係数 (coefficient of variation) を考える

$$\boxed{C.V. = \frac{S_x}{\bar{x}}}$$

cf. これは無次元量

Ex. 1965年 ⇒ 所得平均 $\bar{x} = 26.6$ 万円, 標準偏差 $S = 7.5$ 万円
1975年 ⇒ 〃 $\bar{x} = 117.5$ 万円, 〃 $S = 23.8$ 万円
しかし、所得格差はひろがっているわけではない。なぜなら
$C.V. = 0.28 \rightarrow 0.20$ である。

- 標準化

データを一次変換して、互いに比べやすくすることを考える。

$$\boxed{z_i = \frac{x_i - \bar{x}}{S_x}}$$

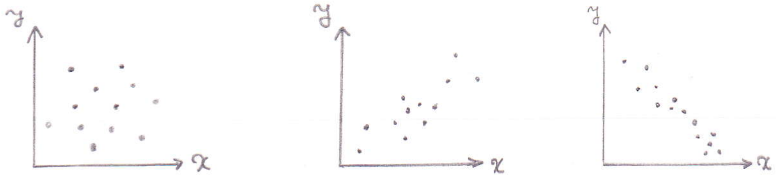
このように一次変換すると、$\bar{z_i} = 0$, $S_z = 1$ となる。
平均と標準偏差が揃うので、あとは z の値を比較してよい。
この変換を標準化という。z を標準得点という。

ちなみに、z に対してさらに $T_i = 10z_i + 50$ とした T_i は
いわゆる偏差値である。

第3章. 2次元のデータ (P41~66)

● 散布図と分割表 (P41~P47)

2次元のデータでは、次のような、xとyを軸にとった散布図を書くのがふつう。てか、必ず書く。



(i) (ii) (iii)

(i) のように x.yに 直線関係がない → 相関関係なし。
(ii) 〃 x.yに、一方の増加に伴いもう一方も増加する関係がある → 正の相関関係
(iii) 〃 x.yに 〃 ,もう一方は減少する 〃 → 負の 〃

しかし、一方が質的データ(性別、国籍など) の場合、上のような表がかけない。
よって、下のような分割表でかく。[これが問題になるのは たぶん稀なので、ムシしてもよい]

	日本人	留学生	合計
修士過程	2415	274	2,689
博士過程	2002	620	2,622
合計	4,417	894	5,311

← ヨコの列を表頭という。

↑ タテの列を表側という。

※ここは、だいたいわかるでしょう。それ以上でも以下でもありません。

NO. 8

DATE

ここは大切なポイントです。

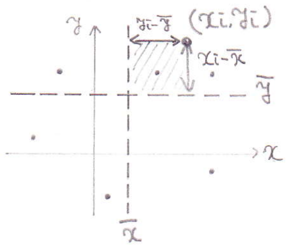
● 相関係数（P48〜57）

右ページの (i)〜(iii) などのようないろいろな相関について、
その相関の強さを数値化するのが下の相関係数。（$-1 \leq r \leq 1$ である。$|r|$ が大きい程、相関大）

○ 相関係数

$$r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})/n}{\sqrt{\sum(x_i-\bar{x})^2/n} \cdot \sqrt{\sum(y_i-\bar{y})^2/n}} = \frac{\text{共分散}}{\sqrt{x\text{の分散}} \cdot \sqrt{y\text{の分散}}}$$

○「共分散」の説明



ある点 (x_i, y_i) について、
$(x_i-\bar{x})(y_i-\bar{y})$ は左のような符号つきの面積。
$\bar{x}, \bar{y}$ によってわかれる4つの部分 のうち、対角の2つに
多くの点があつまれば、その値は大きくなり、4つの部分に
バラバラに分布すると"符号つき"面積なのでうちけしあって小さくなる。

◎ $r = \dfrac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2}\sqrt{\sum y_i^2 - n\bar{y}^2}}$ と表せる。これの方が重要

- ※ 0.7 とかあれば、十分な相関がある。
- ※ r の値のみに頼るのではなく、散布図をみるべき。例えば $y=x^2 \Rightarrow r=0$。しかし、相関は明らかにある。
- ※ r は外れ値の影響をうけやすいので、やはり散布図によって異常値がないか checkすべき。
  外れ値があるときは、適当な層別が可能かを考える。

○ 偏相関係数

ex. 収入を y、世帯人数を m、消費を c とする

- y と m の相関係数 $r_1 = 0.823$ → 強い正の相関！
  しかし、消費 c の影響をとりのぞく（＝同一の消費量とする）と…
- 偏相関係数 → $r' = -0.714$ → 負の相関！（同じ消費の世帯では 収入多 → 人数少）

⇓

偏相関係数［変数1と2の相関を、3を除いて考えたもの］

$$r_{12\cdot3} = \frac{r_{12} - r_{13}\,r_{23}}{\sqrt{1-r_{13}^2}\cdot\sqrt{1-r_{23}^2}}$$

（$r_{◎,□}$ … ◎と□の相関係数）

※ そこそこ重要な気もしますが、式を覚えるのもだるいので、
試験では諦めるのも一つの手かと。余力のある人向け。

○順位相関係数 ～ これも出る可能性はゼロではないが、どちらかというと余力のある人向けかも。1つ目はカンタンなのでそれだけさらっと押さえて次に進んでもいいと思います。

1. スピアマンの順位相関係数 r_s

nコのデータとすると、

あるデータを、2つの指標に基づいて順位づけしたときの順位を R_i、R_i' とする。($R_i = 1 \sim n$)

(人気調査で、i が、男性からは1番好まれ、女性からは3番目に好まれるとき、$R_i = 1$、$R_i' = 3$ など)

このとき

$$r_s = \frac{\sum (R_i - \frac{n}{2})(R_i' - \frac{n}{2})}{\sqrt{\sum (R_i - \frac{n}{2})^2}\sqrt{\sum (R_i' - \frac{n}{2})^2}} = \cdots = 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}(R_i - R_i')^2$$

2. ケンドールの順位相関係数 r_k

$$r_k = \frac{G-H}{n(n-1)/2}$$

G:「$R_i > R_j$ かつ $R_i' > R_j'$」or「$R_i < R_j$ かつ $R_i' < R_j'$」の数。

つまり、2つの指標で順位の上下が一致する組の数

H: その逆となる組の数

※ P55～P57：時系列と自己相関 の内容は省略します。

たぶん出ないと思います。出たら諦めましょう。やる気のある人はどうぞ。

NO. 10

DATE

● 回帰直線 (P58〜64)　重要です。

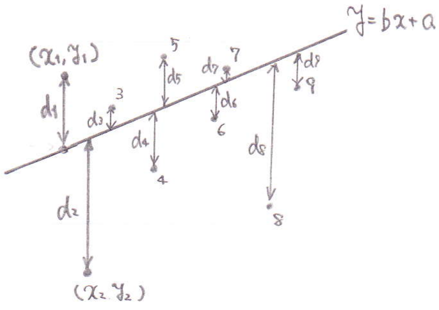
独立な変数 x と従属な変数 y の間に $y = bx + a$ という直線関係をあてはめることを考える。(= 回帰)

ここでは、その直線を(つまり係数 a と b を)どう決めるかを学ぶ。

↓

決め方は「最小二乗法」による。

重要!!　○最小二乗法 … 右図の $d_1, d_2 \cdots d_n$ の2乗の和が最小となるように直線を引く



これにのっとって考えると、求める a, b は、

$$L = \sum_{i=1}^{n} \{ \underbrace{y_i}_{\text{データのy}} - \underbrace{(bx_i + a)}_{\text{直線上のy}} \}^2 \quad \text{—①}$$

2乗は符号を外すためにやる。

の L を最小にする a, b である。

①式を a で偏微分して 0 とおくと、

$$na + (\Sigma x_i) b = \Sigma y_i \quad \cdots ②$$

①式を b で偏微分して 0 とおくと

$$(\Sigma x_i) \cdot a + (\Sigma x_i^2) b = \Sigma x_i y_i \quad \cdots ③$$

②から③をとくと、

$$\boxed{b = \frac{\Sigma x_i y_i - n \cdot \bar{x}\bar{y}}{\Sigma x_i^2 - n \cdot \bar{x}^2}, \quad a = \bar{y} - b\bar{x}}$$

$$b = \frac{(\text{共分散})}{(x\text{の2乗和}) - n \times (x\text{の平均の2乗})}$$

$$a = (y\text{の平均}) - b(x\text{の平均})$$

こうして得られた直線を回帰直線という。

※ 上式からわかるように、各々のデータがわからずとも、$\Sigma x_i y_i$, Σx_i^2, $\bar{x}$, $\bar{y}$ がわかれば直線はひける。

※ このページはかなりのポイントです。最小二乗法から $\boxed{a \cdot b\text{の公式}}$ を導く偏微分や②③の連立方程式は、自分で手をうごかしてみて下さい。

○ 決定係数

相関と回帰のつながりについて考える。

先程求めた b は

$$b = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}}$$

[← $\sum(x_i-\bar{x})(y_i-\bar{y}) = \sum x_i y_i - n\cdot\bar{x}\bar{y}$

$\sum(x_i-\bar{x})^2 = \sum x_i^2 - n\cdot\bar{x}^2$ を用いた]

とも表せる。

この2つの公式は重要です。(公式というほどのものでもないですが。) P12〜の練習問題

ここで、相関係数 $r = \dfrac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$ であるので

$$b = r\cdot\frac{\sqrt{\sum(y_i-\bar{y})^2}}{\sqrt{\sum(x_i-\bar{x})^2}} \qquad \therefore\ b = r\cdot\frac{S_y}{S_x}$$

さて、ここで、前ページの d_i について

$$L = \sum d_i = \sum\{y_i-(bx_i-a)\}$$
$$= (1-r^2)\cdot\sum(y_i-\bar{y})^2$$

が成り立つ。(計算はよくわかりません。たぶん上手く式変形すればできるんじゃないでしょうか)

つまり、相関係数 r について r^2 が大きい程 L は小さくなる(つまり d の和(誤差)が小さい)

この r^2 を決定係数といい、直線のあてはまりぐあいを表す。

※ P63〜64 の内容:平面のあてはめ、多項式回帰 は省略します。
難しいし、面倒なので。出ないと思います。

○ 次のページに、2.3章の復習用の超簡単な練習問題をのっけてみました。
いままでの説明はぶっちゃけ適当でいいので、それをやりつつ解き方を覚えるようにして下さい。
てか、第1回レポを自力でやった人はやんないでいいです。それをカンタンにしただけなので。

NO. 12

DATE

◎第2章 練習問題

1 10点満点のテストを5人が受けた。結果は下の通り。

2, 2, 5, 10, 3

(1) 平均は?

(2) メディアンは?

(3) モードは?

(4) レンジは?

(5) 分散は?

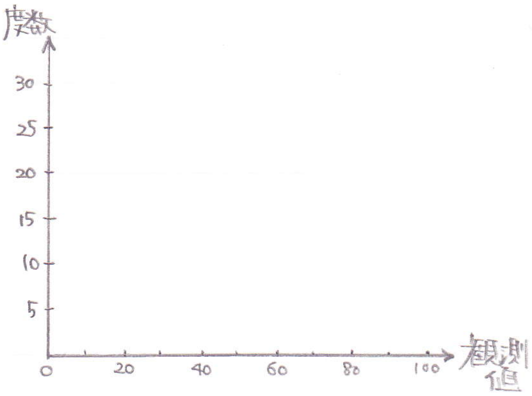
(6) 標準偏差は?

(5),(6)は有効4ケタでよい。

(7) (イミある作業ではないけど)
データをすべて標準化して同じ順に書き並べ、平均と標準偏差が(ほぼ)0と1になることを確かめよ。

2 次の度数分布表を得た(あるテスト)

階級	階級値	度数	累積度数
0以上~20未満		25	
20以上~40 〃		15	
40 ~ 60 〃		5	
60 ~ 80 〃		10	
80 ~ 100 〃		5	
合計	—	60	—



(1) 表の空欄を埋めよ。

(2) 平均は?

(3) ヒストグラムを右上に書け。

3 次式を示せ。(式変形を確認せよ)

(1) $\sum(x_i-\bar{x})^2 = \sum x_i^2 - n\cdot\bar{x}^2$ [分散]

(2) $\sum(x_i-\bar{x})(y_i-\bar{y}) = \sum x_i y_i - n\cdot\bar{x}\bar{y}$ [共分散]

↑

この2式はたいへん有用です。各々のデータの値がわからずとも
$\sum x_i^2$ と $n, \bar{x}$ さえもとまれば分散は求められることを意味するからです。

第3章 練習問題

1 ある10点満点の2つの試験P,Qを、10人の生徒が受験したところ、下の結果を得た。

生徒No.	1	2	3	4	5	6	7	8	9	10
P試験の点数 x	10	8	4	3	6	1	0	2	5	7
Q 〃 y	10	9	4	5	8	3	3	4	7	8

次の設問に答えよ

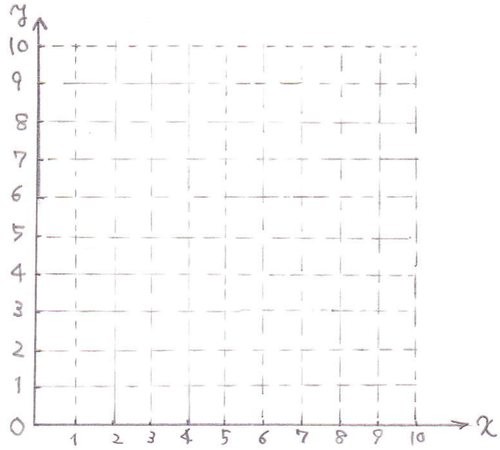
(1) 右表に点をプロットして散布図をつくれ。

(2) 相関係数を求めよ

(3) 決定係数を求めよ

(4) 回帰直線を求めてグラフに引け。

☆ (2)や(4)は、公式をそのままつかう方法も、第2章練習問題3を利用して考える方法もあります。後者に慣れることをオススメします。



やや難. でも面倒。

2 最小二乗法から、回帰直線の係数a.bをxやyの式で表現せよ。

3 相関と回帰は、どう違うか説明せよ。

第4章. 確率 (P67〜P86)

※基本的な内容については省略し、「条件付確率」のみ扱う。

条件付確率 (P81〜P83)

ある事象Bが起こったときに、事象Aが起こる確率 を、Aの条件つき確率 といい、P(A|B) と表す

$$P(A|B) = \frac{P(A\cap B)}{P(B)} \left(= \frac{\text{AもBもおきる確率}}{\text{Bが}} \right) \quad \leftarrow \text{あたり前の式。}$$

$$\left(\Leftrightarrow P(A|B)\cdot P(B) = P(A\cap B) \right)$$

ベイズの定理 (P84〜P85)

Aという結果に、$H_1, H_2 \cdots H_k$ という原因が考えられるとき、
当然「Aのとき、H_iである確率」を求めたい。しかしふつうは「H_iのとき、Aである確率」くらいしか求まらない。
$P(H_i|A)$　　$P(A|H_i)$

↓

以下の ベイズの定理 により、計算可能

$$\left(\text{結果の原因が }H_i\text{ である確率}=\right) P(H_i|A) = \frac{P(H_i)\cdot P(A|H_i)}{\sum_j P(H_j)\cdot P(A|H_j)} = \left(\frac{(\text{原因}H_i\text{が起こる確率})\cdot(\text{原因}H_i\text{の下でAになる確率})}{\text{上の積を各原因 }H_1\sim H_k\text{まで加えたもの}} \right)$$

$$= \frac{P(H_i\cap A)}{\sum_i P(H_i\cap A)} \quad (\leftarrow \text{この方が見やすい})$$

※例題1つとけば、わかると思います。
(問) ← 配布資料そのまま。

ある製品がA社,B社,C社で製造されていて、どの社の製品か見分けがつかないとする。

	A社	B社	C社
市場シェア	50%	30%	20%
不良率	10%	20%	25%

購入した製品が不良品であった。 それがA社、B社、C社である確率は？

↓

こんなん楽勝ですよね？
ふつうに考えると、"市場全体に不良品がどれくらいあるか"Ⓐ を求めて、
"市場全体に A社, B社, C社の不良品がどれだけあるか"Ⓑ を求めると、いいわけです。

さうすれば、$\frac{Ⓑ}{Ⓐ}$ が、求める確率ですから。これをいってるのが、ベイズの定理です。

一応、例題を定理にのっとって解説します。（わかってる人は、無視して次へ）

［答え］

まず、市場全体に、どれだけの割合で不良品があるか考える。（下図参照）

$$P(\text{不良品}) = \underbrace{0.5\times0.1}_{\text{Aの不良品がある割合}} + \underbrace{0.3\times0.2}_{\text{Bの不良品がある割合}} + \underbrace{0.2\times0.25}_{\text{Cの不良品がある割合}} = 0.16 \cdots ①$$

求めたいのは、不良品がA社である確率

$P(\text{A社}\mid\text{不良品})$

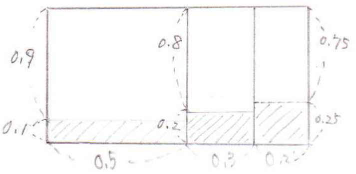
わかる（求まる）のは

$P(\text{不良品})$

$P(\text{不良品}\mid\text{A社})$

$P(\text{不良品}\cap\text{A社})$

など。



では、

A社かつ不良品である確率は

$P(\text{不良品}\cap\text{A社}) = 0.5\times0.1 = 0.05 \cdots ②$

①②より、ベイズの定理から、不良品がA社のものである確率は

$$P(\text{A社}\mid\text{不良品}) = \frac{0.05}{0.16} = \frac{5}{16}$$

同様に

・$P(\text{不良品かつB社}) = 0.3\times0.2 = 0.06$　よって $P(\text{B社}\mid\text{不良品}) = \frac{0.06}{0.16} = \frac{6}{16}$

・$P(\text{不良品かつC社}) = 0.2\times0.25 = 0.05$　よって $P(\text{C社}\mid\text{不良品}) = \frac{0.05}{0.16} = \frac{5}{16}$

以上より、不良品を見たら、$\frac{5}{16}$でA社、$\frac{6}{16}$でB社、$\frac{5}{16}$でC社のもの。//

※ここから先、example を用いることが増えるかと思います。それは、理論がやや面倒だったりわかりにくくても、実際には特殊な例が多く、それほど難しくないことが多いからです。説明がわからなくても、具体例がわかればいいです。そんなもんです。

※ぶっちゃけ、第4章は、このくらいの例題ができればだいたいOKです。所詮は統計で使うための確率ですし。（てことで、この章の練習問題はありません）

NO. 16.

DATE

第5章. 確率変数 (187~108)

この章のプリントは、ちょっと教科書とはスタイルを変えてまとめてみます。
いままでのシリーズとも ちょっと雰囲気がちがいますが、あまり気にしないで下さい。

ちなみに、数Cの確率に関して、完璧に理解してるって人は、連続型だけさらっと確認すればOKです。

・本章のポイント

…確率変数には、離散型と連続型がある。離散型は今まで扱ってきたようなもの
では、連続型では、期待値や分散などはどうなるか
結論から言うと、ほぼ同じになる。

○まず、「離散」と「連続」について説明する

。離散型の確率変数

例えば、サイコロの出目や人数など、値がとびとびの(自然数の)値しかとれないものをいう。(棒グラフで表すようなもの)

。連続型の確率変数

例えば、電池の寿命や地雷のある場所など、実数の値をとるもの(時間とか、距離とか、数直線で表したり、線グラフで表せるもの)

(注)
「確率変数」とは、「～である確率」と表現したときの「～」にくる値のこと。
ちなみに、「A社である」など、数以外のときは、例えば、A社⇒1, B⇒2…と対応づけて、変数といえる。

少し、連続型については具体例も使って説明しておく。
たとえば、全くRANDOMに地雷の埋め込まれた道を歩くとする。歩いた距離 x が変数である。
このような、randomな場合、x 歩いたときに地雷をふむ確率 $P(x)$ は、図1のような指数減少のグラフになるといわれているらしい。

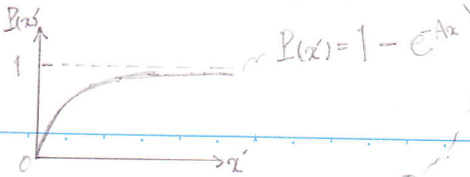
(図1)

y は、x あるいたときに地雷をふむ確率。このとき、$y=f(x)$ なる $f(x)$ を、確率密度関数という。

では、この指数関数はどのように表せるだろうか。$y=Ae^{-Bx}$ とおき、$(B>0)$
AとBの関係を考えよう。
全確率は1であるから、$\int_0^{\infty} P(x)\,dx = 1$ である。
これより
$$\int_0^{\infty} A\cdot e^{-Bx}dx = 1 \iff A\left[\frac{-1}{B}e^{-Bx}\right]_0^{\infty} = 1$$
$$\iff \frac{A}{B} = 1$$

ゆえに、$A=B$ である。つまり、このとき、$y=A\cdot e^{-Ax}$ と表せる。

※「歩くほど確率は上がるやろ!」とツッコミたい人は、勘違いしてます。上の文の"確率"は、そこまで歩く間に踏む確率で、累積確率になります。
累積確率 $P(x) = \int_0^{x} P(x)\,dx$ より、右図。
(x歩くまでにふむ)



○ では、このそれぞれについて、期待値と分散を考えます。（期待値 $E(x)$、分散 $V(x)$）

※ 期待値と分散の細かな定義や意味をやることは省略します。

・まず、離散型について

$$\boxed{E(x) = \sum_{x} x \cdot P(x)}$$

← 高校でやった通り

が、期待値です。これは全く問題ないでしょう。

では、分散はどうするか。今までの通りです。プリントNo.5にあるとおり、平均からのズレを考えていく。

$$(\text{分散}) = \left(\sum \{(\text{各値}) - (\text{平均})\}^2 \right) / n$$

であった。今回の分散の定義は下で

$$\boxed{\begin{aligned} V(x) &= \sum (X - E(x))^2 \cdot P(x) \\ &= E\{(X - E(x))\} \end{aligned}}$$

← $P(x)$ というのが それのようなものというか それを拡張したものというか…
うまく説明できないんですけど かなりの人は納得できると思います。

具体例1つやれば、わかってしまいます。

Ex. 次のような分布表があるとき、期待値 $E(x)$ と分散 $V(x)$ を求めよ

X	1	2	3	4	5	6
$P(x)$	$4/10$	$1/10$	$1/10$	$1/10$	$1/10$	$2/10$

$$E(x) = 1 \times \frac{4}{10} + 2 \times \frac{1}{10} + 3 \times \frac{1}{10} + 4 \times \frac{1}{10} + 5 \times \frac{1}{10} + 6 \times \frac{2}{10} = \underline{3}$$

$$V(x) = (3-1)^2 \times \frac{4}{10} + (3-2)^2 \times \frac{1}{10} + (3-3)^2 \times \frac{1}{10} + (4-3)^2 \times \frac{1}{10} + (5-3)^2 \times \frac{1}{10} + (6-3)^2 \times \frac{2}{10} = \underline{4}$$

※ 比較として… サイコロ振りを考えると（$P = \frac{1}{6}$ でぜんぶいっしょ）

$$\begin{cases} E(x) = 3.5 \ (\text{当然}) \\ V(x) = \dfrac{35}{12} \end{cases}$$

… 上より、ばらつきが小さい
（平均からはなれた値が少ない）

注．教科書では、$P(x)$ を、関数として $f(x)$ と表してる。

・では、連続型では、どうなるか。

→ Σ はとびとびに加える記号。連続的に加える記号は… $\int$

つまり、

連続型の期待値

$$\boxed{E(x) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx}$$

※ $f(x)$ ってのは、前のページの、確率密度関数です。要は、その値 x での確率です。

連続型の分散

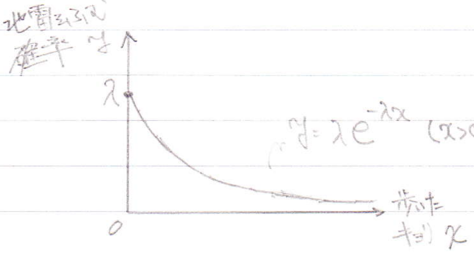
$$\boxed{V(x) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\, dx}$$

$E(x)$ は定数ですから、これを μ と表します。これ以後、断らず μ を使います。

Ex. 具体例として、先程の地雷を踏むやつを考えます。

x あるくと、$f(x) = \lambda \cdot e^{-\lambda x}$ の確率で地雷を踏むっていうやつです。

※ Aが λ になってますが、とくに意味はないです。教が λ なので、あわせてみました。



積分区間は、上の $\int_{-\infty}^{\infty}$ ではなく、正しくは"定義域内"です。これがあってます。

$$\text{期待値は}\quad E(x) = \int_0^{\infty} x \cdot \lambda e^{-\lambda x} dx = \lim_{x\to\infty} \left[-\left(\frac{1}{\lambda}+x\right) e^{-\lambda x} \right]_0^{x} = \frac{1}{\lambda}$$

$$\text{分散は}\quad V(x) = \int_0^{\infty} \left(x-\frac{1}{\lambda}\right)^2 \cdot \lambda e^{-\lambda x} dx = \cdots = \frac{1}{\lambda^2}$$

です。部分積分は、できますよね。

ついでに、とうぜん $\sqrt{V(x)}$ を標準偏差といいます。記号 σ で表します。(分散を σ^2 と書く)

○期待値と分散の性質

まず、最も重要な公式を。分散を計算するとき、いちいち「各値と期待値の差をとって二乗して…」とかするやつはほとんどいなくて、実際には、以前プリントNo.5でやったように、(2乗の平均)−(平均の2乗)のような式をつかいます。今回も同様に、

$$\boxed{V(x) = E(x^2) - (E(x))^2 \quad \left(= (2\text{乗の期待値}) - (\text{期待値の}2\text{乗}) \right)}$$

を使います。[証明省略。プリントNo.5と同じようにしてできます。教科書P97にものってます。]

ex. として、プリントNo.17のexを、この式でやると…

$$V(x) = \left\{ 1^2 \times \frac{4}{10} + 2^2 \times \frac{1}{10} + 3^2 \times \frac{1}{10} + 4^2 \times \frac{1}{10} + 5^2 \times \frac{1}{10} + 6^2 \times \frac{2}{10} \right\} - 3^2 = \underline{4}$$

つづいて、期待値の性質について。(ともに、証明略。難しい証明じゃないです。してみてはぜひ。)

$$\begin{cases} 1.\quad E(ax+b) = a \cdot E(x) + b & (a, b\text{は定数}) \\ 2.\quad E(x+Y) = E(x) + E(Y) & \leftarrow \text{独立性。} \end{cases}$$

つづいて、分散の性質について、

1. $V(ax+b) = a^2 \cdot V(x)$ (a, bは定数)

   この2つから、導かれます。
   $V(a) = 0$, $V(a \cdot x) = a^2 \cdot V(x)$
   (1つめは、あたりまえですよね。データがすべて定数なら、バラつきは0です。)

2. $V(x+Y) \neq V(x) + V(Y)$ ← こんなミスを犯してほしくないので、敢えて記しておきます。

また、例によって、下のように、標準化ができます。

$$\boxed{Z = \frac{X - E(x)}{\sqrt{V(x)}} \quad \text{このとき、必ず、} E(Z) = 0,\ V(Z) = 1}$$

※ モーメント、歪度、尖度、チェビシフの不等式は省略します。(できれば、チェビシフの不等式は、やってみて下さい。モーメントはやんないでいいです。)

第6章 確率分布 (P109～132)

先の章で学んだ観点から、実際に様々な確率分布モデルについて、考える。
(本章では多くの確率分布について取り扱っているが、そのうち大切なもののみここには記す。)

◎ 離散型の確率分布① ～二項分布～ ⇐ 離散型は、とにかくコレを押さえること!!

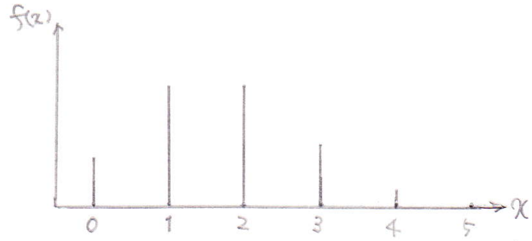
成功の確率を p とし、失敗の確率を $q(=1-p)$ とする1回の試行を n 回行う。
(これはベルヌーイ試行とよぶ)
このとき、成功が x 回(つまり失敗は $n-x$ 回)起こる確率 $f(x)$ は

$$\boxed{f(x) = {}_nC_x \cdot p^x \cdot (1-p)^{n-x}}$$

← これがわかんない人は、いませんよね? 高校レベルです。

この式による確率分布を、二項分布という。
二項分布の分布のようすは、n, p の2つのパラメタで決まるので、その関数というくらいの意味で、
二項分布を $Bi(n, p)$ と表す。
(※ ちなみに、$n=1$ をとくに、ベルヌーイ分布とよぶ)

右に例として、2人でジャンケン(勝つ確率 $\frac{1}{3}$)を5回やったときの二項分布 $Bi(5, \frac{1}{3})$ を示す。
(各自自分で、やってみて下さい。)



さて、この二項分布の期待値と分散を求めよう。
⟶ 証明はやや面倒なので、まず結果を示すと

$$\boxed{E(x) = np \quad , \quad V(x) = npq}$$

← 重要!

・直観的理解: そりゃ、1回につき p の確率で成功するなら、n 回では np 回成功するだろう。
分散は、確率のばらつきを示すが、$p \cdot q$ の値は $p = q = \frac{1}{2}$ で Max. つまり、
成功も失敗も等確率で起こるなら、そりゃ結果もバラバラになりやすい。

(一応証明書きますが、あんまり気にしないで下さい。)

$$E(x) = \sum_x x \cdot {}_nC_x \cdot p^x (1-p)^{n-x}$$
$$= np \cdot \sum_x {}_{n-1}C_{x-1} \cdot p^{x-1} (1-p)^{n-x}$$
$$= np \cdot \sum_y {}_{n-1}C_y \cdot p^y (1-p)^{(n-1)-y} \quad (y = x-1)$$

↳ 成功 p, y 回のベルヌーイ試行の確率の総和 ⇒ 全確率1

$$= np \quad ▨$$

$V(x) = E(x^2) - (E(x))^2$ より、$E(x^2)$ を求めると

$$E(x^2) = \sum x^2 \cdot {}_nC_x \cdot p^x \cdot q^{n-x}$$
$$= np \sum_y (y+1) \cdot {}_{n-1}C_y \cdot p^y \cdot q^{(n-1)-y}$$

(← 左と同じ、$y = x-1$)

$$= np \left\{ \sum_y y \, {}_{n-1}C_y \cdot p^y \cdot q^{(n-1)-y} + \sum_y {}_{n-1}C_y \cdot p^y \cdot q^{(n-1)-y} \right\}$$
$$= np \left\{ (n-1) \cdot p \cdot \sum_z {}_{n-2}C_z \cdot p^z \cdot q^{(n-2)-z} + 1 \right\}$$

(← $z = y-1$)

$$= np \{ (n-1)p + 1 \}$$

よって $V(x) = np\{(n-1)p+1\} - n^2p^2 = np(1-p) = npq$ ▨

● 離散型の確率分布② 〜ポアソン分布〜 ⇐ 重要度はやや低い。しかし、一読を薦めます。

二項分布において、nが大きく、pが小さい。つまり、大量の観察により、稀少現象がある程度確認されるようなとき、二項分布は次のように上手く近似される。

$n \to \infty$, $p \to 0$, かつ $E(x) = np \to \lambda$ （←つまり、期待値npという性質を保って$n \to \infty$, $p \to 0$にする）

の極限をとると、

$$\boxed{{}_nC_x \cdot p^x (1-p)^{n-x} \longrightarrow \frac{e^{-\lambda} \cdot \lambda^x}{x!}}$$

proof.

$${}_nC_x = \frac{n!}{x!(n-x)!}, \quad p^x \fallingdotseq \left(\frac{\lambda}{n}\right)^x$$

$$(1-p)^n \fallingdotseq \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}$$

$$(1-p)^{-x} \fallingdotseq 1$$

より、従う。（厳密さはビミョーだけど）

このように、

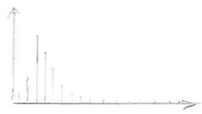
$$f(x) = \sum_x \frac{e^{-\lambda} \lambda^x}{x!}$$

に従う分布を、ポアソン分布という。

ポアソン分布のパラメタはλのみであり $P_o(\lambda)$ と表す。

ポアソン分布の形は、右に歪んだような形になる（∵ nが大でpが小）

《イメージ図》



ポアソン分布の期待値と分散は次のようになる。

$$\boxed{E(x) = \lambda \quad , \quad V(x) = \lambda}$$

証明は、二項分布のときと同じようにやるっぽい。

$E(x)$については教科書P115上部にのってます。

$V(x)$は、モーメント母関数というよくわかんない手法が付録にのってます。

これは、覚えるか、捨てて下さい。二項分布を、しっかりやるべきです。

◎ 連続型の確率分布① ～正規分布～ ⇐最重要事項．絶対にマスターすること!!

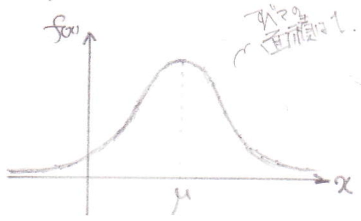
下のような期待値 μ，分散 σ^2（標準偏差 σ）を用いた，
確率密度関数による分布を，正規分布という。

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \cdot \exp\left\{-(x-\mu)^2/2\sigma^2\right\} \quad (-\infty < x < \infty)$$

※ 周知のこととおもいますが $\exp(f(x))$ は $e^{f(x)}$ のことです。

正規分布のパラメタは，期待値 μ（平均），分散 σ^2 であり，正規分布を $N(\mu, \sigma^2)$ と表す

正規分布の $\boxed{\text{期待値は } \mu \text{，分散は } \sigma^2}$ である



※ μを中心とする左右対称な分布

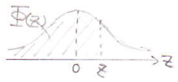
正規分布には以下の性質がある。

(a) X が正規分布 $N(\mu, \sigma^2)$ に従うとき，$Y = aX + b$ は
$N(a\mu + b, a^2\sigma^2)$ に従う。

(b) 標準化変数 $Z = \frac{X-\mu}{\sigma}$（このとき Z について期待値0，分散1）は，
正規分布 $N(0,1)$ に従う。この $N(0,1)$ を標準正規分布という

※ (a)，(b) より，すべての正規分布は，標準正規分布に帰着できる
それゆえ，標準正規分布の密度関数 $f(z)$ の，累積分布関数 $\Phi(z)$

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)\, dx$$



の値が与えられる。（←教科書 P280，付表1は，$1-\Phi(z)$ を与えている）

（注1）正規分布の式は複雑なので，別に自分で積分したりとかはしません。そのかわりに，
"$N(\mu, \sigma^2)$ が与えられたら，標準化を行い，$N(0,1)$ に帰着させ，付表を用いる" ことになります。

（注2）あとで例題1つ出すので，そこでなんとなくわかると思います。

（注3）．正規分布に従う例が多いこと。また，中心極限定理っていうやつによると，ランダムな試行については何か上手くやると必ず試行の数が多くなるとき正規分布になっていくことが，正規分布の大切さの一因。他にも，理論的な良さもあるっぽい。

（注4）「3シグマ範囲」… $N(\mu, \sigma^2)$ について，$\mu - 3\sigma \leqq x \leqq \mu + 3\sigma$ の範囲には，0.9973 の確率である事象が入ることから，「事実上すべて」といった意味で用いる。

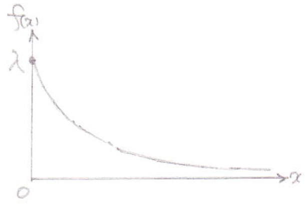
● 連続型の確率分布② ～指数分布～ ← 重要はそれほどでもない。でもカンタン

これは、もうやりました。最初に連続型を説明したときの例です。

確率密度関数

定義域を $-\infty \sim \infty$ にするための表現

$$f(x) = \lambda \cdot e^{-\lambda x} \quad (x \geqq 0), \quad 0 \quad (x < 0)$$



で定義される確率分布です。

プリント No.18 でやったように $\boxed{E(x) = \frac{1}{\lambda}, \quad V(x) = \frac{1}{\lambda^2}}$

※ 他の分布は省略。余裕のある人は
- 超幾何分布（離散型）
- ガンマ分布（連続型）

あたりをやってもいいかも。

※ 各分布の E, V は、出せるor覚える方がいいかも。
正規分布は、問題演習で理解すること。

第7章. 多次元の確率分布 (P133～154)

本章は、第5章の続きとして、2つの変数 X・Y があるときの
期待値や分散について考え、独立、無相関について学ぶ。

で、この章は、本当に適当にやります。ただし、それはこの章が大切でないとか、
テストに出ないとかじゃなくて、説明が面倒なわりに、結果は単純
であったりするからです。
では、始めましょう。

● 2次元の確率分布

2つの四面体 A.B があって、各面に 1～4 が書かれている。2つは正四面体ではなく、

$$\begin{cases} A: 1\to \frac{3}{10},\ 2\to \frac{5}{10},\ 3\to \frac{1}{10},\ 4\to \frac{1}{10} \\ B: 1\to \frac{1}{10},\ 2\to \frac{2}{10},\ 3\to \frac{3}{10},\ 4\to \frac{4}{10} \end{cases}$$

の確率で出るとき、A→x、B→y が出る確率 $f(x,y)$ は、下表のようになる。

B＼A	1	2	3	4
1	0.03	0.05	0.01	0.01
2	0.06	0.10	0.02	0.02
3	0.09	0.15	0.03	0.03
4	0.12	0.20	0.04	0.04

このような $f(x,y)$ の分布を、同時確率分布という

(とりあえず、こういうふうに2次元の分布があるとわかればよい)

● 分散と相関係数

第5章で学んだように、分散には加法が成立しない。($V(X+Y) \neq V(X) + V(Y)$)
では、$V(X+Y)$ はどうなるかというと、

$$\boxed{\begin{aligned} &V(X\pm Y) = V(X) + V(Y) \pm 2\cdot Cov(X,Y) \\ &\text{ただし}\quad Cov(X,Y) = E\{(X-E(X))(Y-E(Y))\} \quad :\text{共分散} \end{aligned}}$$

共分散は、昔やったのと、そんなにかわらないです。期待値は平均と同じようなものなので。
要は、「(xの平均からのズレ)×(yの平均からのズレ)」の平均です。

[補足] $V(X+Y) = V(X) + V(Y) + 2\,Cov(X,Y)$ の証明 (「−」も同じ)

$$V(X+Y) = E\{(X+Y)^2\} - \left(E(X+Y)\right)^2$$
$$= E(X^2+2XY+Y^2) - \left(E(X)+E(Y)\right)^2$$
$$= E(X^2) - E(X)^2 + E(Y^2) - E(Y)^2 + 2\left(E(XY) - E(X)E(Y)\right)$$

ここで、

$$E\{(X-E(X))(Y-E(Y))\} = E(XY - X\cdot E(Y) - Y\cdot E(X) + E(X)E(Y))$$
$$= E(XY) - E(Y)\cdot E(X) - E(X)\cdot E(Y) + E(X)E(Y)$$
$$= E(XY) - E(X)E(Y)$$

(← $E(X), E(Y)$ は定数、$E(\text{定数}) = (\text{その定数})$)

ゆえに、

$$V(X+Y) = V(X) + V(Y) + 2\,Cov(X,Y) \quad \blacksquare$$

この証明中の式 $\boxed{Cov(X,Y) = E(XY) - E(X)E(Y)}$ は、計算上重要。

- 相関係数と無相関、及び独立。

相関係数を、以前(第3章)定義したのと同様、次のように定義する

$$\boxed{\text{相関係数}\ \rho_{XY} = \frac{Cov(X,Y)}{\sqrt{V(X)}\cdot\sqrt{V(Y)}}}$$

(このとき $-1 \leqq \rho_{XY} \leqq 1$. $|\rho_{XY}|$ が大きいと直線関係)

$\rho_{XY} = 0$ となるとき、XとYが無相関であるという。ρ は分子 >0 より

$\rho = 0 \Leftrightarrow Cov(X,Y) = 0 \Leftrightarrow E(XY) = E(X)E(Y)$

すなわち $\boxed{\text{XとYが無相関} \Leftrightarrow E(XY) = E(X)E(Y)\ (\Leftrightarrow \rho_{XY} = 0)}$

一方、独立とは、Xの出方がYによらず、Yの出方がXに依らないことをいう。
また、確率を勉強するときにやったと思います。たとえば、積の法則 $P(X\cap Y) = P(X)\cdot P(Y)$ が成り立つことと同値です。

ここで、大切なのは $\boxed{\text{XとYが独立} \Rightarrow E(XY) = E(X)E(Y)\ (\text{逆は不成立})}$

つまり、$\boxed{\text{XとYが独立} \Rightarrow \text{XとYが無相関}\ (\text{逆は不成立!})}$

→ 例題に出してますので、checkして下さい。

※独立の数学的定義を飛ばしたのは、僕自身よくわかんなかったからです。自分で教P143とP137をやってみて下さい。

※第8章は、テスト範囲外(たぶん)なので、やりません。

第5,6章 練習問題

1. 標準化によって期待値(平均)が0、分散が1になることを示せ。

2. 赤玉が2コ、白玉が8コ入っているつぼから、n回ボールを引いてはもとに戻す試行を行う。このときx回赤玉が出る確率を$f(x)$とする。
   (1) この試行は何と呼ばれるか。また、確率密度関数$f(x)$をnとxの式で表せ。分布の名称も答えよ
   (2) 期待値$E(x)$と分散$V(x)$を求めよ。(公式を用いてよい。)

3. 「3シグマ範囲には0.9973の割合が当てはまる」ことを示せ。
   すなわち、$N(\mu, \sigma^2)$なる正規分布の$\mu-3\sigma \leqq x \leqq \mu+3\sigma$の範囲内の累積確率が約0.9973であることを示せ。教科書のP280.付表1を用いよ。
   <hint> まず、標準化しないと、付表を使えませんよね。標準化したとき、3シグマ範囲はどこからどこまででしょうか?

4. ←教科書P122の(例)と同じです。
   偏差値得点Tは、テストの実際の得点xを標準化し、(それをzとする)
   さらに$T=10z+50$としたものである。
   (1) Tの平均(期待値)と分散を求めよ。
   (2) Tは正規分布に従うとする。このとき、偏差値75以上の生徒の割合と、偏差値50以上51未満の生徒の割合を求めよ。付表1を用いよ。

第7章 練習問題

1. 2つのつぼA,Bの中に3個のボールを投げ入れる。(入る確率は$\frac{1}{2}$ずつ。)
   このとき、Aに入るボールをXコ、ボールの入っている箱の数をYコとするとき、XとYの同時確率分布を求めて(←表にする)
   XとYが無相関だが独立でないことを示せ。
   (教科書P153の問題7.3と同じ)

● 第9章．標本分布

・まず初めに、母集団と標本の違いを説明します。

母集団…その統計が本来対象とする集団全体
標本…母集団からランダムにとりだしたサンプル

私達は、母集団からとり出された標本から、
母集団全体を推定することになる。
本章では、母集団と標本の関係について述べられている。
しかし、結局はあまりテストに出る部分に限られている（ような気がする）ので、
それらについてのみ解説する。

● 母集団と標本の平均と分散

※平均と分散が重要なのは、この2つが、正規分布のパラメタであるから。

母集団の平均，分散は，今まで通り

$\mu = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$ などで定義される

一方，標本が n個 $X_1\ X_2 \cdots X_n$ とあったとき
平均は $\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n)$ となり，今まで通りなのだが

● 以後，表記の違いに注意!!

- μ, σ^2, σ：母集団の値
- $\bar{x}$, s^2, $\sqrt{s^2}$：標本の値

基本的に，$\bar{x}$, s^2 はわかっている。
μ, σ^2 などが推定したい値

標本分散（s^2 で表す）については

$$\boxed{s^2 = \frac{1}{n-1}\left\{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2\right\}}$$

となる。

これは、次の「不偏性」を保つための結果である（練習問題参照）

※不偏性

$$E(\bar{x}) = \mu \quad , \quad E(s^2) = \sigma^2$$

すなわち，標本の平均の平均は母集団の平均に、標本の分散の平均は母集団の分散になること

No. 28

Date

◎ パラメトリックとノン・パラメトリック ～次章以降のイントロとしての話～

「母集団が〇〇分布に従う」とわかっているとき、その分布のパラメタ(例えば正規分布なら μ と σ、二項分布なら n と p)がわかれば、全体がわかる。このような状況をパラメトリックという。

そうじゃなくって、母集団の従う分布がわかっていないとき、ノン・パラメトリックという。

今後、多くの問題で、正規分布を仮定する。つまり、パラメトリックを扱う。

その時、標本から母集団について推定したり、母集団に関する仮説を検定したりする。

※この章はあんまりよくわからないです。たぶん大丈夫です。
ただし、下の例題はテストに出そうかと言っていたので、それだけはやっとくといいでしょう。
(↑あくまで可能性で、絶対出すとは言ってませんが)

◎ 第9章練習問題

シケプリ No.27. の、不偏分散の不偏性 $E(s^2) = \sigma^2$ を示せ。

この章から先は、
「得られた標本から母集団について推定する」
ことを意識して読み進めて下さい。

第10章. 正規分布からの標本

得られた標本が、正規分布からのものであるとすると、比較的簡単に
母集団の平均や分散を考えることができる。(→平均と分散がわかれば、正規分布は一意に定まる)

詳しくは次章以降で学ぶので、本章ではいくつかの新たな統計量を定義したり、
その分布を考えたりする。よくわからなくても、すぐに進んでほしい。

・標本平均 $\overline{X}$ はどのような分布に従うか? ～母分散が既知のとき～

標本 $X_1, X_2 \cdots X_n$ が $N(\mu, \sigma^2)$ に従うとき、$\overline{X}$ はどうなるか。

$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ であるので、これも正規分布に従う。

このとき

$$E(\overline{X}) = \frac{1}{n} E(X_1 + X_2 + \cdots + X_n) = \frac{1}{n} \times (n \cdot \mu) = \mu$$

$$V(\overline{X}) = \frac{1}{n^2} V(X_1 + X_2 + \cdots + X_n) = \frac{1}{n^2} \times (n \cdot \sigma^2) = \frac{\sigma^2}{n}$$

である. つまり

$\boxed{\overline{X} \text{ は 平均 } \mu, \text{ 分散 } \frac{\sigma^2}{n} \text{ の 正規分布に従う。}}$

そこで、これを標準化した $Z = \dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}}$ は、$N(0, 1)$ に従う。

※「だから何!?」といいたいでしょうが、グッとこらえて下さい。
これをもとに、「μ」を推定することになるので。

Date

・標本分散 s^2 とカイ2乗分布 … この部分の説明は雑にやります。面倒なわりに、理解してもムダっぽい。覚えた方が早いので。

定義（χ^2分布）

$z_1, z_2 \cdots z_n$ を独立かつ、$N(0,1)$ に従う変数とすると

$$\chi^2 = z_1^2 + z_2^2 + \cdots + z_n^2$$

と定義した χ^2 が従う分布を、自由度 n の χ^2分布と呼ぶ

そして、実は、$s^2 = \frac{1}{n-1}\left\{(x_1-\overline{x})^2 + \cdots + (x_n-\overline{x})^2\right\}$

を次のように変形したとき

標準化っぽいね

$$\boxed{\frac{(n-1)s^2}{\sigma^2}} = \left(\frac{x_1-\overline{x}}{\sigma}\right)^2 + \cdots + \left(\frac{x_n-\overline{x}}{\sigma}\right)^2$$

この左辺の値は 自由度 $n-1$ の χ^2分布に従う。

※ これは母分散 σ^2 の推定につかいます。重要ですよ。

※表記法

$z = \frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$ の従う $N(0,1)$ において、その点より上側の確率が α となる点を z_α と表す。

たとえば、その上側に 0.25% しかふくまれない点は $z_{0.025}$ で表す。

同様に、χ^2分布については、$\chi^2_\alpha(n)$ と表す。nは自由度である。

この点が教科書の付表には記されている。

次のページで出す $t_\alpha(n)$ も同じである

・標本平均 $\bar{X}$ はどのような分布に従うか？ 〜母分散未知のとき〜

母分散既知のとき $Z=\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ が $N(0,1)$ に従うことがわかっていたが、
σ がわからない状況では、どうするか
↓
σ を $s^{(2)}$ で代用する！

定義（t分布）

$$t=\frac{\bar{X}-\mu}{s/\sqrt{n}}$$ で定義される t は、自由度 $n-1$ の t 分布に従う。

※ やはりこれも、μ の推定につかいます。

◎ 補足（重要！）標本平均の差の標本分布（教科書 P204〜）

2つの標本平均 $\bar{X}$, $\bar{Y}$ から、2つの母平均の差 $\mu_1-\mu_2$ を考えるとき、
以下の2つの場合を考える
(a) 母分散 σ_1^2, σ_2^2 が既知のとき
(b) 母分散 σ_1^2, σ_2^2 が未知であるが $\sigma_1^2=\sigma_2^2$ とわかっているとき

※教科書にはもう1つ、「母分散が全く未知のとき」がのっていますが、テストには出ないらしいです。

(a)のとき

$$Z=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{(\sigma_1^2/m)+(\sigma_2^2/n)}}$$ は $N(0,1)$ に従う。

(b)のとき

$$Z=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\left(\frac{1}{m}+\frac{1}{n}\right)\sigma^2}}$$ が 自由度 $n+m-2$ の t 分布に従う

この式を覚える必要はないです。テストに出すなら、どちらを使うかの判断が下せればよい、という形式にするそうです。

◎ 第10〜12章は、難しいです。本当はとても内容のあるところです。
でも、シケプリという性質上、本当に最低限のことをやることにしました。
ですから、ここに書いてあることは絶対にマスターして下さい。

◎ P207, 208 の 標本分散の比の標本分布 も、テストに出るかもしれませんが、
もう大変なのでやめます。やった方がいいとは思います。

第11章 推定 (区間推定のみ扱う。点推定は、よくわかんないので)

第10章で説明した通り、標本から母集団について推定する。
本章では、3つの区間推定について、マスターしてほしい

① 母分散 σ^2 が既知のとき、母平均 μ を推定する
② 〃 未知 〃 、 〃
③ (そもそも) 母分散を推定する。

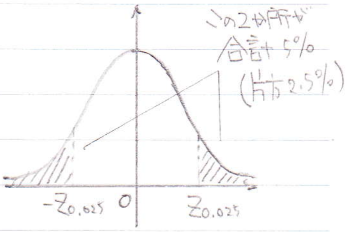
※ 言葉の定義
例えば、母平均 μ を推定するとき、95%の確率で μ がふくまれる区間を μ の95%信頼区間とよぶ。

① 母分散が既知のとき、母平均 μ を推定する。(95%信頼区間を求める)

標本平均 $\overline{X}$ を標準化した Z は、μ を用いて次のように表せる

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

そして、これは $N(0,1)$ に従う。



$N(0,1)$ で、95%の範囲になるのは、対称性より

$$-z_{0.025} \leq \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \leq z_{0.025}$$

μ について、解くと

$$\overline{X} - z_{0.025} \cdot \left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \overline{X} + z_{0.025} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$$

「必ずテストに出題するの」とのことです。
絶対に覚えましょう
過程が導けるとよいと思います

② 母分散が未知のとき、母平均 μ を推定する

Z のかわりの t を用いる。

$$t = \frac{\overline{X} - \mu}{s / \sqrt{n}}$$

は、自由度 $n-1$ の t 分布に従う

t 分布も対称性があるので 95%の範囲は、

$$-t_{0.025}(n-1) \leq \frac{\overline{X} - \mu}{s / \sqrt{n}} \leq t_{0.025}(n-1)$$

μ についてとくと

$$\overline{X} - t_{0.025}(n-1) \cdot \left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \overline{X} + t_{0.025}(n-1) \cdot \left(\frac{s}{\sqrt{n}}\right)$$

上の2つは、似ているということがわかって頂いたです。

③ 母分散を推定する

$\dfrac{(n-1)s^2}{\sigma^2}$ が χ^2分布（自由度 $n-1$）に従うことを利用。

χ^2分布には対称性がないので、ふつうに 0.025〜0.975 の範囲で考えると

$$\chi^2_{0.025}(n-1) \leqq \frac{(n-1)s^2}{\sigma^2} \leqq \chi^2_{0.975}(n-1)$$

σ^2についてとくと

$$\frac{(n-1)s^2}{\chi^2_{0.025}(n-1)} \leqq \sigma^2 \leqq \frac{(n-1)s^2}{\chi^2_{0.975}(n-1)}$$

※以上の3つを具体例を通して確認します。これをやれば、イメージがつかめるハズ。

(Ex.1)

座金の製造工場で、ある日に作られた座金の中から100コをとり出してその厚さを測定したところ、平均 $\overline{X} = 2.346$ (mm) であった。また、この工場で作られる座金の厚さの分散は既に $\sigma^2 = (0.047)^2$ (mm) とわかっている。

母平均 μ の信頼係数 90% の信頼区間を求めよ。

[Ans]

標本平均 $\overline{X}$ を標準化したら、μを用いて次のように表せる

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$ これが $N(0,1)$ に従う。

$N(0,1)$ で 90% の範囲をとるには、対称性より上下 5% を捨てて

$$-Z_{0.05} \leqq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leqq Z_{0.05}$$

付表1 より $Z_{0.05} = 1.645$，さらに各値を代入して

$$-1.645 \leqq \frac{2.346 - \mu}{0.047/\sqrt{100}} \leqq 1.645$$

$$\therefore \quad 2.338 \leqq \mu \leqq 2.354$$

よって、90%信頼区間は $[2.338,\ 2.354]$

（※直接、公式に代入してもOK！）

No. 34

Date

(ex. 2)

先程の ex.1 と似た状況を考える。

今回は、母分散はわかっておらず、標本分散を 100コのサンプルから求めると、$s^2 = (0.047)^2$ (mm) であったとする。

母平均 μ を推定せよ。(90%信頼区間で)

[Ans.]

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$ (自由度は n-1) を考える。

t分布にも対称性があるので(教科書でみて下さい)、上下5%ずつとすると、

$$-t_{0.05}(n-1) \leqq \frac{\overline{x} - \mu}{s/\sqrt{n}} \leqq t_{0.05}(n-1)$$

$$\therefore -t_{0.05}(99) \leqq \frac{2.346 - \mu}{0.047/\sqrt{100}} \leqq t_{0.05}(99)$$

ここで、 $t_{0.05}(99) = 1.660$ とすると (← 付表2とかにのってるのがふつう。)

$$-1.660 \leqq \frac{2.346 - \mu}{0.047/\sqrt{100}} \leqq 1.660$$

$$\therefore \left[2.346 - 1.660 \times \frac{0.047}{\sqrt{100}} ,\ 2.346 + 1.660 \times \frac{0.047}{\sqrt{100}} \right]_{//}$$

レポ問4

↓

(ex. 3) ある高校の1年生40名がモシをうけると、

$\overline{x} = 55$, $s^2 = 146.4$ だった。モシをうけた全受験生の分散 σ^2 の95% 信頼区間を求めよ。

→ 解答は RepAn 参照のこと

第12章 仮説検定

ある仮説（帰無仮説という）H_0 があるとき、それが正しい（妥当か）か否かを決めるにはどうすればよいだろうか。

↓

次のようにする。

1. 何らかの統計量を用いて、自分で決めた妥当性の基準（＝有意水準という）にあてはまる区间（採択域という）を求める。

2. 仮説の下でその統計量を計算し、採択域にあてはまるかをcheck.

早速、例題やります。慣れが肝心！

例5　仮説検定①（平均値の検定・母分散既知の場合）

【問題】　あるビール製造工場では、容器が 350ml の缶ビールが最も多く製造されている。ところが最近、どうも容量が設定値の 350ml とは異なることが多いとの報告が何回かあった。そこで、本当に要領が平均的に 350ml かどうかを検討するために、10 本の缶ビールをランダムに取って容量を測ったところ、次のようなデータが得られた。（単位：ml）

349.1　348.2　348.1　350.5　350.3　350.1　349.2　348.6　349.5　348.2

このデータから標本平均を計算すると、$\overline{X} = 349.18$ となる。また、過去の経験から、缶ビールの容量の（母）分散は $\sigma^2 = (0.9)^2$ であることが分かっている。この結果から、「やはり平均的に 350ml ではない」と結論付けてよいだろうか？

確かにそう結論付けても良いように見える。しかし、データのばらつきを考えれば、たまたま平均が 350 を下回っただけではないか、と反論することも可能だろう。

仮説は当然 $H_0 : \mu = 350\,\mathrm{ml}$

使う統計量は、$\overline{X}$, σ^2 から μ を考えるので、

$Z = \dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}}$ がよいだろう。

有意水準を 0.05 とすることにすると、

Z の 95% がふくまれる範囲は

$$-Z_{0.025} \leqq Z \leqq Z_{0.025}$$

付表1より、これは $-1.96 \leqq Z \leqq 1.96$ …①

↑
今回の採択域

一方、仮説 $H_0 : \mu = 350\,\mathrm{ml}$ として Z を計算すると

$$Z = \frac{349.18 - 350}{0.9 / \sqrt{10}} = -2.88$$

この値は①にふくまれない。つまり、H_0 が正しい確率は5%以下であり、仮説が正しいとはいえない。

（H_0 は有意水準 5% で棄却されたという。）

［※つまり、たった 0.82ml のズレとはいえ、10本の平均でこれだけずれれば有意な差があるといえるのです。］

次は、母分散が未知の場合です。
さらに、少しひねってあるのは、「〜より速いか否か」を判断するので、先の問題ではズレは大小どちらでもよかったけど、今回は、遅い方のズレは無視して、片側のみしかズレを棄却しないというところです。

例6 仮説検定②（平均値の検定・母分散未知の場合）

【問題】 陸上競技の選手であるBさんは、100m走で、平均で12秒より速く走れるといって自慢している。Bさんの友達は、このことを確認したかったので、Bさんに時間をおいて9回走ってもらったところ、以下の記録を得た（単位：秒）。

11.98　11.99　12.00　12.05　12.05　12.00　11.97　11.98　11.96

このデータの標本平均と標本分散は、それぞれ $\overline{X} = 11.997778$、$s^2 = 0.00104438$ と計算できる。Bさんは100mを平均で12秒より速く走ることができると結論できるだろうか？有意水準を 0.05 として仮説検定を行いなさい。

Bさんのタイムは正規分布に従うと仮定する。
帰無仮説を $H_0 : \mu = 12$
対立仮説を $H_1 : \mu < 12$
とする。（棄却されれば、12秒より速いといえる）

母分散がわかっていないので、統計量

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} \quad (\text{自由度 } n-1)$$

を用いる。

今回の採択域は、

$$-t_{0.05}(8) < t < \infty$$

である。（5%で下側だけをすてた）
↑速いほう

$t_{0.05}(8) = 1.86$ （∵付表2）
より

$$-1.86 < t \quad \cdots ①$$

が採択域である。

このとき、H_0 のもとで t を計算すると

$$t = \frac{11.99778 - 12}{\sqrt{0.00104438}/\sqrt{9}} \fallingdotseq -0.206$$

これは①にふくまれない。
つまり、
Bさんが平均で12秒より速く走れるとは限らない。

これで解説はすべておわりです。

第10章～12章 練習問題

1. 母平均 $\mu_1=2$, 母分散 $\sigma_1^2=3$ の正規母集団から大きさ $m=10$ の標本を、
母平均 $\mu_2=5$, 母分散 $\sigma_2^2=4$ の 〃 から大きさ $n=8$ の標本を
抽出する二つの標本平均の差 $\Delta=\overline{X}-\overline{Y}$ は、どのような分布をとるか？

2. 母分散 $\sigma=3$ である正規母集団から9個の標本をとりだしたとき、
この標本の平均は $\overline{X}=2$ であった。母平均 μ の95%信頼区間を求めよ。
（また、$\overline{X}=2$ が、6^4 個の標本による値だったとして、同じく μ の95%信頼区間を求めよ）

3. ある正規母集団から9個のサンプルをとり出した。
このサンプルについて、標本分布 $\overline{X}=30$, 標本分散 $s^2=4$ を計算により得たとするとき
母平均の95%信頼区間を求めよ

4. ある正規母集団からの10個のサンプルについて、$s^2=6$ であった。
母分散の95%信頼区間を求めよ

5. もう一度、ミケプリ No.33・34 の Ex を復習して下さい。

6. ある教授が教室に来る時間について、開始時刻からのズレを遅れは＋, はやく来たら − で
表したのが下の表である。
$\{-3, 0, -1, +2, 0, +4, +2, +6, 0, -1, +2, -1, +3\}$
この教授が定刻に来るという帰無仮説 $H_0: \mu=0$ を有意水準5%で
検定せよ。（ただし、対立仮説は $H_1: \mu\neq 0$ とし、両側検定にせよ）

7. 正規母集団について母分散 $\sigma^2=4$ がわかっている。ここから
25個のサンプルをとったとき、標本平均 $\overline{X}=14.6$ を得た。
このとき母平均 $\mu=15.0$ といえるか？ すなわち
$H_0: \mu=15.0$ を有意水準5%で検定せよ。（両側検定でよい）

NO. 12′-①

DATE

● 第2章練習問題の答え

[1]

(1) $\bar{x} = \dfrac{2+2+5+10+3}{5} = \underline{4.4}$ //

(2) 中央値は $\underline{3}$ //

(3) 最頻値は $\underline{2}$ //

(4) レンジは $10-2=\underline{8}$ //

(5) [解1] P5の上の式に代入

$$S^2 = \frac{1}{5} \times \left\{ (2-4.4)^2 + (2-4.4)^2 + (5-4.4)^2 + (10-4.4)^2 + (3-4.4)^2 \right\}$$

$$= \frac{1}{5} \times (5.76 + 5.76 + 0.36 + 31.36 + 1.96)$$

$$= 9.04$$

[解2] P5の下の式

$$S^2 = \frac{2^2+2^2+5^2+10^2+3^2}{5} - (4.4)^2$$

$$= 28.4 - 19.36$$

$$= 9.04 \qquad \therefore \underline{S^2 \fallingdotseq 9} \text{ //}$$

⇐ 断然ラク。

(6) ゆえに

$$S = \sqrt{9.04} \fallingdotseq \underline{3} \text{ //}$$

（本来なら関数電卓や開平法をやるべきかな？）

(7) 標準化の公式

$$z_i = \frac{x_i - \bar{x}}{S}$$

を利用。

与えられたデータは

$\underline{-0.8,\ -0.8,\ 0.2,\ 1.8\dot{6},\ -0.4\dot{6}}$ //

になる。このとき

$$\bar{z} = \frac{-0.8-0.8+0.2+1.8\dot{6}-0.4\dot{6}}{5} = \underline{0} \text{ //}$$

$$S_z = \sqrt{S_z^2}$$

$$= \sqrt{\frac{(-0.8)^2 + \cdots + (-0.46)^2}{5} - 0^2}$$

$$\fallingdotseq \sqrt{\frac{0.64+0.64+0.4+3.46+0.21}{5}}$$

$$= \sqrt{1.07} \fallingdotseq \underline{1} \text{ //}$$

[2]

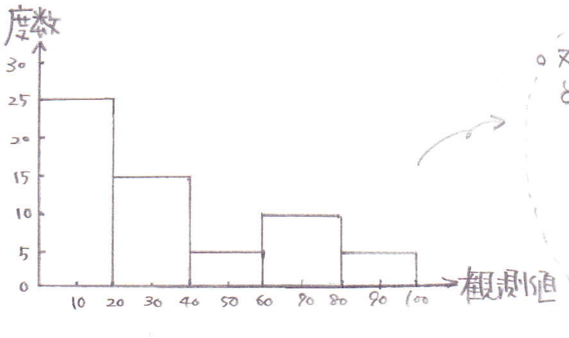
(1) 階級値はその階級の平均で、上から順に $\underline{10,\ 30,\ 50,\ 70,\ 90}$ //

累積度数は、そこまでの度数の和。上から順に $\underline{25,\ 40,\ 45,\ 55,\ 60}$ //

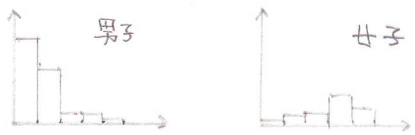
(2)
$$\bar{x} = \frac{10\times25 + 30\times15 + 50\times5 + 70\times10 + 90\times5}{60}$$

$$= \underline{35} \text{ //}$$

(3)



○ 双峰型なので、層別を考えたいところですね。たとえば男女で層別して、



みたいになると、それが適切な分け方だといえます。

（他にもいろいろ、わけ方はあるでしょうが）

NO. 12'-②

DATE

3

定数

(1) $$\sum (x_i - \bar{x})^2 = \sum (x_i^2 - 2\bar{x} \cdot x_i + \bar{x}^2)$$

$$= \sum x_i^2 - 2\bar{x}\left(\sum x_i\right) + \bar{x}^2\left(\sum 1\right)$$

$$= \sum x_i^2 - 2 \cdot \bar{x} \cdot n\bar{x} + \bar{x}^2 \cdot n$$

（∵ $\sum x_i$ はすべての要素の和. 平均の定義より、それは平均×個数）

$$= \sum x_i^2 - n \cdot \bar{x}^2 \quad ∎$$

(2) $$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i y_i - \bar{x} \cdot y_i - \bar{y} \cdot x_i + \bar{x} \cdot \bar{y})$$

$$= \sum x_i y_i - \bar{x} \cdot \sum y_i - \bar{y} \cdot \sum x_i + \bar{x}\bar{y} \cdot \sum 1$$

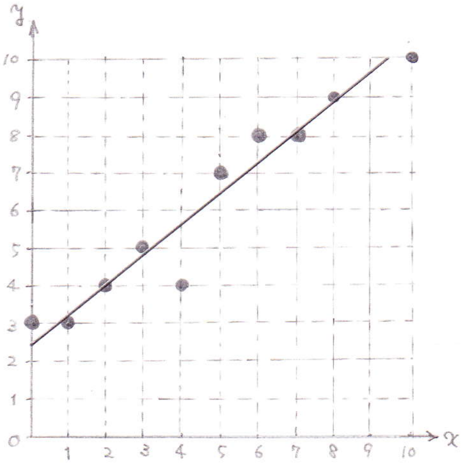
$$= \sum x_i y_i - \bar{x} \cdot n\bar{y} - \bar{y} \cdot n\bar{x} + \bar{x} \cdot \bar{y} \cdot n$$

$$= \sum x_i y_i - n \cdot \bar{x}\bar{y} \quad ∎$$

第3章練習問題の答え

1

(1), (4)のグラフ



(2)

$$r=\frac{\sum xy - n\cdot\bar{x}\bar{y}}{\sqrt{\sum x^2-n\bar{x}^2}\cdot\sqrt{\sum y^2-n\bar{y}^2}}$$

ここで

$\sum xy = 100+72+16+15+48+3+0+8+35+56$

$\therefore \sum xy = 353$

$\bar{x}=(10+8+4+3+6+1+0+2+5+7)\div 10$

$\therefore \bar{x}=4.6$

$\bar{y}=(10+9+4+5+8+3+3+4+7+8)\div 10$

$\therefore \bar{y}=6.1$

$\sum x^2 = 100+64+16+9+36+1+0+4+25+49$

$\therefore \sum x^2 = 304$

$\sum y^2 = 100+81+16+25+64+9+9+16+49+64$

$\therefore \sum y^2 = 433$

であるので

$$r=\frac{353-10\times 4.6\times 6.1}{\sqrt{304-10\times 4.6^2}\cdot\sqrt{433-10\times 6.1^2}} \fallingdotseq 0.9651\ldots$$

(3) よって、決定係数は

$r^2 \fallingdotseq 0.931\ldots$

(4)

回帰直線を $y=a+bx$ とすると、

$$b=\frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\cdot\bar{x}^2}, \quad a=\bar{y}-b\cdot\bar{x}$$

よって

$$b=\frac{353-10\times 4.6\times 6.1}{304-10\times(4.6)^2}=0.783\ldots$$

$\therefore a = 6.1 - 0.78 \times 4.6 = 2.512$

ゆえに、回帰直線は $y=0.78x+2.51$

2 解答省略。プリントNo.10も参照。

(②かつ③は、a.bの連立方程式でなので中2レベル)

3

(たぶん) 回帰は、一方が他方の原因であるという関係を指し、
相関は、因果がどちらからどちらにあるかをあまり気にしない(また、ともに
影響しあうような関係もふくむ)ような2つの変数の関係。

NO. 26′−①

DATE

第5.6章練習問題の答え

1 もとのデータの期待値をμ、分散をσ^2とする。このデータxを標準化すると、

$$z = \frac{x-\mu}{\sigma}$$

この新しいzについて、期待値と分散を定義に従って求める。
(もとの、xの期待値μや分散σ^2が定数であることに注意する。)

$$E(z) = E\left(\frac{x-\mu}{\sigma}\right)$$

$$= \frac{E(x)-\mu}{\sigma} \quad (\because E(ax+b) = aE(x)+b)$$

$$= 0 \quad (\because E(x) = \mu)$$

この2つも、証明可能。

また、$V(z) = V\left(\frac{x-\mu}{\sigma}\right)$

$$= \frac{V(x)}{\sigma^2} \quad (\because V(ax+b) = a^2V(x))$$

$$= 1 \quad (\because V(x) = \sigma^2)$$

[補足]

(1) $E(ax+b) = a\cdot E(x) + b$ の証明

$$E(ax+b) = \frac{1}{n}\sum^{n}(ax+b)$$

$$= \frac{1}{n}\left(a\sum^{n}x + nb\right)$$

$$= a\cdot E(x) + b$$

(2) $V(ax+b) = a^2\cdot V(x)$ の証明

$$V(ax+b) = E\{(ax+b)^2\} - (E(ax+b))^2$$

$$= E(a^2x^2+2abx+b^2) - (aE(x)+b)^2$$

$$= a^2E(x^2) + \cancel{2abE(x)} + \cancel{b^2} - a^2(E(x))^2 - \cancel{2abE(x)} - \cancel{b^2}$$

$$= a^2\left[E(x^2) - \{E(x)\}^2\right]$$

$$= a^2\cdot V(x)$$

2 (1) このような試行をベルヌーイ試行という。
n回やってx回赤が出る確率は

$$f(x) = {}_nC_x \cdot \left(\frac{2}{10}\right)^x \cdot \left(\frac{8}{10}\right)^{n-x}$$

このような分布を、二項分布という

(2) 公式より

$$E(x) = \frac{1}{5}n$$

$$V(x) = \frac{4}{25}n$$

3 $N(\mu, \sigma^2)$ なる正規分布を標準化すると、$N(0, 1)$ となり、このとき $z = \frac{x-\mu}{\sigma}$ であるので、

$$\mu - 3\sigma \leqq x \leqq \mu + 3\sigma$$

$$\therefore \mu - 3\sigma \leqq \sigma z + \mu \leqq \mu + 3\sigma$$

$$\therefore -3 \leqq z \leqq 3$$

ゆえに、標準正規分布で $-3 \leqq z \leqq 3$ になる部分を求めると、
教科書の付表1より $1 - 2 \times 0.0013499 \fallingdotseq 0.9973$

4 (1) zの平均(期待値)は0，分散は1ゆえ

$$\cdot\ E(10z+50) = 10 \cdot E(z) + 50$$

$$= 50$$

$$\cdot\ V(10z+50) = 10^2 \cdot V(z)$$

$$= 100$$

(2) ◦偏差値75以上

$$T \geqq 75 \Leftrightarrow z \geqq 2.5$$

$N(0,1)$で $z = 2.5$ となる確率は 0.0062097 (0.6%)

◦偏差値50以上51未満

$$\Leftrightarrow 0 \leqq z \leqq 0.1$$

$N(0,1)$で $z \geqq 0.1$ となる確率は 0.46017
$z \geqq 0$ 〃 0.5
よって $0.5 - 0.46017 = 0.03983$ (4%)

NO. 26-③

DATE

第7章 練習問題の答え

1

同時確率分布 $f(x,y)$ は

	Y \ X	0	1	2	3
$f(x,y)$	1	$\frac{1}{8}$	0	0	$\frac{1}{8}$
	2	0	$\frac{3}{8}$	$\frac{3}{8}$	0

上の表のようになる。

まず、無相関であるかだが

$$E(XY) = 0\times\frac{1}{8} + 1\times 0 + 2\times\frac{3}{8} + 3\times\frac{1}{8} + 4\times\frac{3}{8} + 6\times 0 = \frac{21}{8}$$

であり、$E(X) = 0\times\frac{1}{8} + 1\times\frac{3}{8} + 2\times\frac{3}{8} + 3\times\frac{1}{8} = \frac{12}{8}$

$E(Y) = 1\times\frac{2}{8} + 2\times\frac{6}{8} = \frac{14}{8}$

よって $Cov(X,Y) = \frac{21}{8} - \frac{12}{8}\times\frac{14}{8} = 0$ ∴無相関である

一方、独立性は、

たとえば

$P(X=3, Y=1) = \frac{1}{8}$

であり

$P(X=3) = \frac{1}{8}$, $P(Y=1) = \frac{2}{8}$

であるので

$P(X=3, Y=1) \neq P(X=3)\cdot P(Y=1)$

すなわち、

X, Y は独立でない。

第8章 練習問題の答え

まえおき

何故、標本分散は n ではなく $n-1$ で割るのか、それは、標本分散の期待値が、母集団の分散になるためです。

つまり、「$S^2 = \frac{1}{n-1}\underbrace{\{(X_1-\overline{X})^2 + \cdots + (X_n-\overline{X})^2\}}_{\sum_{i=1}^{n}(X_i-\overline{X})^2}$ としたとき、$E(S^2)=\sigma^2$」であることです。

(proof)

$$\sum_{i=1}^{n}(X_i-\overline{X})^2 = \sum_{i=1}^{n}\{(X_i-\mu)-(\overline{X}-\mu)\}^2$$

$$= \sum_{i=1}^{n}(X_i-\mu)^2 - 2(\overline{X}-\mu)\underbrace{\sum_{i=1}^{n}(X_i-\mu)}_{=n(\overline{X}-\mu)} + n(\overline{X}-\mu)^2$$

$$= \sum_{i=1}^{n}(X_i-\mu)^2 - n(\overline{X}-\mu)^2$$

$(\overline{X}-\mu)^2 = \left(\frac{X_1+\cdots+X_n}{n}-\mu\right)^2 = \frac{1}{n^2}\sum_{i=1}^{n}(X_i-\mu)^2$

$$\therefore E(S^2) = \frac{1}{n-1}\left\{E\{\sum_{i=1}^{n}(X_i-\mu)^2\} - n\,E\{(\overline{X}-\mu)^2\}\right\}$$

$$= \frac{1}{n-1}\left\{\sum_{i=1}^{n}E\{(X_i-\mu)^2\} - n\cdot\frac{1}{n^2}\cdot E\{(X_i-\mu)^2\}\right\}$$

$$= \frac{1}{n-1}(n\sigma^2 - \sigma^2)$$

$E\{(X_i-\mu)^2\} = n\cdot\sigma^2$

$$= \sigma^2$$ ∎

◎ 第10章～12章 練習問題の答え

1 ミケプリ No.31 より.

$$Z=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{(\sigma_1^2/m)^2+(\sigma_2^2/n)^2}}=\frac{(\overline{X}-Y)+3}{\sqrt{\left(\frac{3}{10}\right)^2+\left(\frac{4}{8}\right)^2}}$$ が 正規分布 $N(0.1)$ に従うので

$\Delta=\overline{X}-Y=\frac{\sqrt{34}}{10}Z-3$ は、正規分布 $N\left(-3,\frac{34}{100}\right)$ に従う。//

（∵ ある変数 X が $N(\mu,\sigma^2)$ に従うならば、$aX+b$ は $N(a\mu+b,a^2\sigma^2)$ に従う。）

2 母平均を μ とすると

$Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}=\frac{2-\mu}{3/\sqrt{9}}=2-\mu$ は $N(0.1)$ に従う。95%信頼区間は

$\therefore\ -Z_{0.025}\leqq 2-\mu\leqq Z_{0.025}$

$\therefore\ 2-Z_{0.025}\leqq\mu\leqq 2+Z_{0.025}$

付表より $Z_{0.025}=1.96$ ゆえ 95%信頼区間は $0.04\leqq\mu\leqq 3.96$ //

3 母平均を μ とすると

「$t=\frac{\overline{X}-\mu}{s/\sqrt{n}}$ が 自由度 $n-1$ の t 分布に従う」つまり

$\therefore\ t=\frac{30-\mu}{\sqrt{4}/\sqrt{9}}$ が 自由度 8 の t 分布に従う

よって、求める 95% 信頼区間は

$-t_{0.025}(8)\leqq\frac{30-\mu}{2/3}\leqq t_{0.025}(8)$

$\therefore\ 30-\frac{2}{3}\cdot t_{0.025}(8)\leqq\mu\leqq 30+\frac{2}{3}t_{0.025}(8)$

付表より $t_{(8)\,0.025}=2.306$ ゆえに $28.46\leqq\mu\leqq 31.54$

[注] 2, 3 は、教 P226 の (11.39)式，(11.43)式に代入するだけでもよい.

（2 別解）

(11.39)式に代入して $2-Z_{0.025}\cdot\frac{3}{\sqrt{9}}\leqq\mu\leqq 2+Z_{0.025}\cdot\frac{3}{\sqrt{9}}$ 以下上と同じ

（3 別解）

(11.43)式に代入して $30-\frac{\sqrt{4}}{\sqrt{9}}\cdot t_{0.025}(8)\leqq\mu\leqq 30+\frac{\sqrt{4}}{\sqrt{9}}\cdot t_{0.025}(8)$ 以下上と同じ

6 まず、与えられた標本について

個数：$n = 13$

標本平均：$\overline{x} = \frac{1}{13}\{(-3)+0+(-1)+2+\cdots+(-1)+3\} = 1$

標本分散：$s^2 = \frac{1}{12}\left[\{(-3)-1\}^2+(0-1)^2+\{(-1)-1\}^2+\cdots+(3-1)^2\right] = 6$

である。

母分散未知の場合であるので t検定となる。

$s^2 = \frac{1}{n-1}\cdot\sum(x_i-\overline{x})^2$ より

$s^2 = \frac{n}{n-1}V(x)$

$\therefore\ s^2 = \frac{n}{n-1}\left(E(x_i^2) - \{E(x)\}^2\right)$

ともできる。この方が使いやすい

本問では

$s^2 = \frac{13}{12}\left\{\frac{(-3)^2+0+(-1)^2+\cdots+3^2}{13} - 1^2\right\} = 6$

有意水準5%での採択域は $-t_{0.025}(12) \leqq t \leqq t_{0.025}(12)$ である。

付表より、これは、$-2.179 \leqq t \leqq 2.179$ …①

一方、帰無仮説 $H_0 : \mu = 0$ のもとで求めた t の値は

$$t = \frac{\overline{x}-\mu}{s/\sqrt{n}} = \frac{1-0}{\sqrt{6}/\sqrt{13}} = \sqrt{2.166\ldots} \fallingdotseq 1.4 \text{ よりちょっと大きい値}$$

これは①にあてはまる。よって 帰無仮説は棄却されない。//

7 母分散がわかってるので、$Z = \frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$ で検定する。

有意水準5%の採択域は $-Z_{0.025} \leqq Z \leqq Z_{0.025}$

付表より、これは、$-1.96 \leqq Z \leqq 1.96$

一方、帰無仮説 $H_0 : \mu = 15.0$ から求めた Z の値は

$$Z = \frac{14.6-15.0}{4/\sqrt{25}} = -0.5$$

これは採択域にあてはまる よって 帰無仮説は棄却されない //