基礎統計　（金５　倉田）試験対策プリント

by Takumi Kida

はじめに

基礎統計のシケプリです。内容は授業ノートとプリントの問題を整理して書いたものになっているので、授業に出席してノートをきちんととっている人にとっては必要のないものであると思います。しかも大事なことは全部教科書に載っているのでそちらも参照してください。

1　データの整理

1.1　1 次元のデータの整理

1 次元データとは ⟶1 つの側面に注目したデータのこと

例：　$x_1, x_2, x_3 \ldots, x_n$

それに対して 2 次元のデータはこのような形をとる

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

3,4,5,. . . ,n 次元ついても同様。

1.1.1　代表値の考察

集めたデータの性質を知るために代表値という様々な値について考察する必要がある。
1, 平均

$$\text{平均} \qquad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \cdots \text{元データと同じ単位を持つ。}$$

※データの取りうる値をいくつかの階級 (class) に分けてそれぞれの階級にデータがいくつあるか（度数）ヒストグラムを作って調べる。階級値とは階級を代表する値のこと（教科書 18 ページ）

2, モード/メジアン

- モード Mo ⋯ 度数が最大となる階級の階級値のこと
- メジアン Md(中位数)
  $x_1, x_2, \ldots, x_n$ を小さいほうから並べて $x_(1) \leq x_(2) \leq \ldots \leq x_(n)$ とする。
  このとき　Md= $x_(\frac{n+1}{2})$(n:奇数),　$x_(\frac{n}{2})$ (n:偶数)

3, $\overline{x}$ の持つ意味：最小二乗値について考察する。
例： {1,2,5} …与えられたデータ　数直線上の点 c と与えられたデータの点との距離を考える。そこで

$$f(c) = (1-c)^2 + (2-c)^2 + (5-c)^2 \tag{1}$$

とおく。これは c に関する 2 次式なので平方完成して

$$f(c) = (c - \frac{8}{3})^2 + \frac{26}{3} \tag{2}$$

$$c = \frac{8}{3} \quad \text{で最小} \tag{3}$$

これはデータ {1,2,5} の平均に一致している。

＜問題 1＞

(1)$f(c) = \sum_{i=1}^{n}(x_i^c)^2$ は $c = \overline{x}$ で最小値を取ることを示せ
(2)$\sum_{i=1}^{n}(x_i - \overline{x}) = 0$ を示せ

1.1.2　散らばりの尺度の考察

1, 分散
分散の定義：　$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$

※ S^2 とは分散を表す文字

2, 標準偏差：　$S = \sqrt{S^2}$　　… 元データと同じ単位を持つ

3, 変動係数：　$= \frac{S}{\overline{x}}$（単位なし）

※平均と標準偏差との比をとることでちらばりの大きさを測る。詳しくは教科書 38 ページ。

4, 標準化：標準偏差をメモリに取ったデータの表現
定義：$Z_i \equiv \frac{x_i - \overline{x}}{S}$　Z_iは x_iの標準化

定理　$x_1, x_2, \ldots, x_n$のデータに対して　$y_i = ax_i + b$ とおく $(i = 1, 2, 3, \ldots, n)$ このとき
(1)$\overline{y} = a\overline{x} + b$
(2)${S_y}^2 = a^2 {S_x}^2$
(3)$S_y = a\sqrt{S_x}$

計算すれば自明なので証明略

Z を $Z_i = \frac{1}{S}x_i + \frac{\overline{x}}{S}$ と変形し、上の定理を用いる。$a = \frac{1}{S}, b = \frac{\overline{x}}{S}$ に相当するから

- $\overline{Z} = 0$
- ${S_z}^2 = 1$

を導くことができる。この標準化の性質は重要！

1.2　二次元のデータ

1.2.1　相関

1, 散布図 2 次元のデータを扱う際は要素が二つあるため、データをｘｙ平面の点にプロットして表現する。これを散布図という。

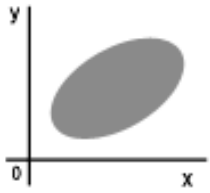


図 1: x が大きくなると y も大きくなる→正の相関

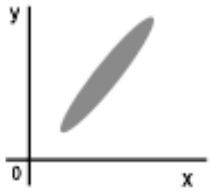


図 2: 直線に近いほうが相関性は強い

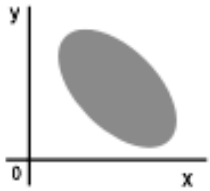


図 3: 負の相関

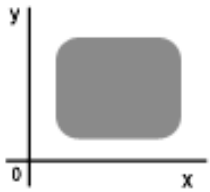


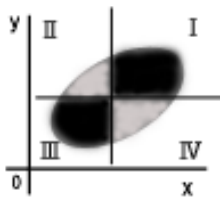
図 4: 全体にばらばら→相関なし

2, 共分散

$$C_{xy} \equiv \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

★共分散のもつ意味を考えてみる

性質：$C_{xy} = C_{yx}$
$C_xy > 0$ →正の相関
$C_xy < 0$ →負の相関



$(x_i - \overline{x})(y_i - \overline{y})$ の値は
，　→正
，　→負

を取る。1,2,3,. . . ,n の総和を考えたとき黒い部分の面積と相関、共分散との関係を考えれば納得できる。

3, 相関係数　　…相関の強さをあらわす係数

定義：　$r_{xy} \equiv \frac{C_{xy}}{S_x S_y}$

$$= \frac{1}{n}\sum_{i=1}^{n}(\frac{x_i - \overline{x}}{S_x})(\frac{y_i - \overline{y}}{S_y})$$

$$= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

定理

(1)$-1 \leq r_{xy} \leq 1$

(2)$r_{xy} = \pm 1 \Leftrightarrow$ 同一直線上

※証明: Cauchy-Schwarz の式 $\sum(a_i b_i)^2 \leq (\sum(a_i{}^2))(\sum(b_i{}^2))$ を用いる。$a_i = x_i - \overline{x}, b_i = y_i - \overline{y}$ とすると $C_x y^2 \leq S_x{}^2 S_y{}^2, C_x y^2 \leq 1$

★相関係数の表す事柄

$0 < r_{xy} < 1$ →正の相関

r_{xy}　0 →相関無し

$-1 < r_{xy} < 0$ →負の相関

1.2.2　回帰分析と決定係数

1, 回帰分析　x が y を説明しているという考えによってデータを見る

仮に x,y の間に $y = \alpha + \beta x_i + \epsilon_i$ という直線的関係があるとする

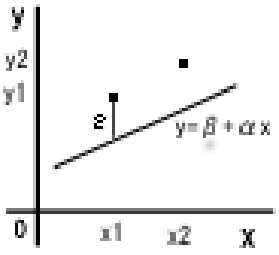
ここで

x:説明変数

y:被説明変数

α, β : 回帰係数（未知）

ϵ : 誤差項（未知）



α, β を決定する方法：最小２乗法（各点に最も近い直線にすればよい)

各点の誤差 ϵ の和を L とする。

$$L = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i)^2)$$

図 5: y 軸に平行なラインで距離を測る

定理　このとき $\beta = \frac{C_{xy}}{S_x{}^2}, \quad \alpha = \overline{y} - \beta\overline{x}$

証明:

2, 決定係数　回帰直線のあてはまり具合を表す係数

- $\hat{y}_i = \alpha + \beta x_i$ ⋯ 回帰直線のあてはめ値
- $d_i = y_i - \hat{y}_i$ ⋯ 残差

を定義して $\sum d_i$ を見る。しかし、d_iはy と同じ単位をもつため、判断がむずかしい。

定理

$$\begin{aligned} \sum(y_i - \bar{y})^2 &= \sum(\hat{y}_i - \bar{y})^2 + \sum d_i \\ \text{y の全変動} &= \text{回帰変動} + \text{残差変動} \\ \text{A} &= \text{B} + \text{C とする。} \end{aligned}$$

これに対して 決定係数：$R^2 \equiv \frac{B}{A} = 1 - \frac{C}{A}$ を定義する
性質：　$(1) 0 \leq R^2 \leq 1, \quad (2) R^2 = 1 \Longleftrightarrow$ 回帰直線上
★ R^2 が 1 に近いほどよく当てはまっている
定理

$$\sum_{i=1}^{n} {d_i}^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 (1 - {r_{xy}}^2)$$

証明は教科書 60〜61 ページ

結論：R^2　　=　　${r_{xy}}^2$
　　決定係数　　相関係数

＜問題２＞

(1)(相関係数は一次変換しても変わらない)　$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ を $z_i = ax_i + b, w_i = cy_i + d (i = 1, 2, \ldots, n)$ によって $(z_1, w_1), (z_2, w_2), \ldots, (z_n, w_n)$ と変換する。$a, c > 0$ とする。このとき“ $r_{xy} = r_{zw}$ ”を示せ。

$$(2) \quad C_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x}\bar{y}), \qquad {S_x}^2 = \frac{1}{n}\sum_{i=1}^{n}({x_i}^2 - \bar{x}^2) \quad \text{を示せ}$$

(3) 野生のイルカ５頭を捕獲してその年齢 x と血中の PCB の量 $y(\mu g)$ を測定したところ

$$(x, y) = (3, 1.5), (9, 9.1), (13, 5.8), (21, 17.6), (32, 15.1)$$

であった。このとき回帰直線 $y = a + bx$ を求めよ。なお、$\bar{x} = 15.6, \bar{y} = 9.8, S_x = 10.1, S_y = 5.9, C_x y = 49.7$
資料は柳川尭『環境と健康データ』（共立出版）らしい。

2　確率

2.1　基本的概念と公式

2.1.1　定義

- 試行：実験や観測などを総称して試行 (trial) という
- 標本空間Ω: 試行の結果の集合を標本空間 ($samplespace$) という
- 標本点：標本空間Ωの各要素を標本点 ($samplepoint$) という
- 標本空間Ωの部分集合を事象 ($event$) という
- 標本空間ΩはΩ の部分集合だから、Ω はひとつの事象である（全事象)
- 空集合∅ もΩの部分集合であるから、∅ もひとつの事象である（空事象)
- 要素がただひとつである事象 $\{\omega\}$ を根元事象という
- A^cを A の補事象といい、A に含まれない標本点の集合をあらわす。但し、$\Omega^c = \emptyset, \emptyset^c = \Omega$

P が Ω 上の確率である
$\Longleftrightarrow$

$$(\mathrm{a})0 \leq P(A) \leq 1$$
$$(\mathrm{b})P(\Omega) = 1, \quad P(\emptyset) = 0$$
$$(\mathrm{c})A_1, A_2, \ldots, \quad \text{が排反のとき } P(A_1 \cup A_2 \cup A_3 \ldots) = P(A_1) + P(A_2) + \ldots$$

2.1.2　公式

$A_1, A_2, \ldots, A_n$が排反事象であるとき

$$1, \quad P(A_1 \cup A_2 \cup \ldots \cup A_n) = \textstyle\sum_{i=1}^{n} P(A_i) \qquad (1)$$

証明：定義 (c) において A_{n+1}以降をすべて∅ と考えれば (b) から $P(\emptyset) = 0$ だから上の式は成り立つ。

$$2, \quad P(A^c) = 1 - P(A) \qquad (2)$$

証明：$\Omega = A \cup A^c$.A, A^cは排反。よって 1 より $P(\Omega) = P(A) + P(A^c)$.$P(\Omega) = 1$ であるから $P(A^c) = 1 - P(A)$

$$3, \quad P(A \cup B) = P(A)) + P(B) - P(A \cap B) \qquad (3)$$

証明：$P(A \cup B) = P((A \cap B^c) \cup (A^c \cap B) \cup (A \cap B)).$ $\cdots (*)$
$P(A) = P((A \cap B) \cup (A \cap B^c)).$
$P(B) = P((A \cap B) \cup (A^c \cap B)).$
事象 $A \cap B$ と $A \cap B^c$, $A^c \cap B$ は排反事象であるから、$P(A) + P(B) = P(A \cap B^c) + P(A^c \cap B) + 2P(A \cap B)$
$(*)$ から $P(A)+P(B) = P(A \cup B)+P(A \cap B)$.移項すれば (3) を得ることができる.

2.1.3 条件付確率

定義

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad (\text{但し } P(B) > 0 \text{ とする。})$$

定義式を変形することで　$P(A \cap B) = P(A|B)P(B)$（乗法公式）を得ることができる。

※　このとき | は given という意味で、| より右の条件が与えられているということをあらわしている。高校数学（数 B）の範囲なので説明はラフです。

2.2 事象の独立

・$P(B|A) = P(B)$ のとき事象 B の起こり方に A は無関係。(但し $P(A) \neq 0$)

$$\Longleftrightarrow \frac{P(A \cap B)}{P(A)} = P(B) \Leftrightarrow P(A \cap B) = P(A)P(B)$$

$\Longleftrightarrow A$ と B は独立。$(A \perp B)$

定理

$(1) A \perp B \Leftrightarrow A^c \perp B \Leftrightarrow A \perp B^c \Leftrightarrow A^c \perp B^c$

$(2) P(A|B = P(A|B^c) \Leftrightarrow A \perp B$

2.2.1 Bayes の定理（範囲外）

$\Omega = H_1 \cup H_2 \cup \ldots \cup H_n$.　$H_i (i = 1, 2, 3, \ldots, n)$ はΩ上の事象で互いに排反

このときある事象 A に対して $P(A) = P((A \cap H_1) \cup (A \cap H_2) \ldots \cup (A \cap H_n)$

$$= \sum_{i=1}^{n} P(A \cap H_i)$$

$$= \sum_{i=1}^{n} P(A|H_i)P(H_i) \qquad (\text{全確率の公式})$$

$$\text{一般に } P(H_j|A) = \frac{P(H_j) \cap A}{P(A)} = \frac{P(A|H_j)P(H_j)}{\sum_{i=1}^{n} P(A|H_i)P(H_i)} \qquad (Bayes \text{ の定理})$$

＜問題３＞

$(1) P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B)$ を示せ

$(2) A \perp B \Longleftrightarrow A^c \perp B^c$を示せ

2.3 確率変数 (r.v)

確率変数 (以下 r.v) とはそれが取る値に対してそれぞれ確率与えられている変数のことである。統計ではデータは確率変数の実現値として捉えられる。r.v には離散型と連続型の r.v がある。

2.3.1　離散型

定義　確率変数 X が $R(X) = \{x_1, x_2, \ldots\}$ と飛び飛びの値をとるとき、X を離散型の確率変数という。X の確率的な挙動は、

$$P(X = x_k) = f(x_k) \quad (k = 1, 2, \ldots) \quad \sum_{k=1}^{\infty} = 1$$

によって定まる。これを X の確率分布という。

2.3.2　連続型

定義
区間があり、負でない連続関数 $f(x)$ に対して $P(a \leq X \leq b) = \int_a^b f(x)dx (\forall a \leq b)$ のとき、X を連続型確率変数といい、$f(x)$ を確率密度関数という。(グラフの面積が確率を表している)
※性質：　$1, \int_{-\infty}^{+\infty} f(x)dx = 1$　$2, P(X = a(\text{定数})) = 0$　$3, P(a \leq X \leq b) = P(a < x < b)$

＜問題４＞

$(1) X$ は $P(X = i) = \frac{1}{n}(i = 1, 2, \ldots, n)$ なる確率分布に従う離散型確率変数とする。これを離散一様分布という。このとき $E(X), E(2X + 5), E(X^2), V(X)$ を求めよ

(2) は次の確率密度関数に $f(x)$ を持つ連続型確率変数とする。

$$f(x) = 1 - |x|(|x| \leq 1), = 0(|x| > 1)$$

$E(X), E(2X + 5), E(X^2), V(X)$ を求めよ

2.3.3　期待値・分散・標準化

データについて行ったのと同様に r.v に対しても期待値や分散などの値を定義することができる。1, 期待値

定義： $E(X) = \mu = \sum_{k=1}^{\infty} x_k f(x_k)$ (離散型)，　$\int_{-\infty}^{+\infty} xf(x)dex$ (連続型)
また、関数 $\phi(x)$ に対して $E\{\phi(x)\} = \sum_{k=1}^{\infty} \phi(x_k)f(x_k)$,　$\int_{-\infty}^{+\infty} \phi(x)f(x)dx$ は $\phi(x)$ の期待値
定理 a, b を定数とする
$(1) E\{aX + b\} = aE(X) + b$
$(2) E(a) = a$
$(3) f(c) = E\{(x - c)^2\}$ は $c = \mu$　で最小値を取る

(3) の証明:$f(c) = E(X^2 - 2cX + c^2)$
$= c^2 - 2cE(X) - E(X^2)$
$= (c - E(X))^2 + E(X^2) - \{E(X)\}^2$　よって $c = E(X) = \mu$で最小

展開し、定理 (1),(2) を使う

2, 分散

定義：$V(X) = \sigma^2 = E\{(x-\mu)^2\}$
標準偏差：　$\sigma = \sqrt{\sigma^2}$
定理：
$(1)V(X) = E(X^2) - \{E(X)\}^2$
∵期待値の定理 (3) の $f(c)$ において、$c = \mu$とすると $f(\mu) = V(X)$
$(2)V(aX+b) = a^2V(X)$ ※ (1) を使って展開すれば自明

3, 標準化
$X = \mu + Z\sigma$とおくと　$Z = \frac{X-\mu}{\sigma}$,　$E(Z) = 0, V(Z) = 1$…すべてデータの整理と同じ。

2.3.4 Chebyshev の不等式

Chebyshev 不等式
確率変数 X の平均をμ, 分散をσ^2とする。このとき任意の $k \geq 0$ に対して以下の不等式が成り立つ

$$\cdot P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

$$\cdot P(|X-\mu| \leq k\sigma) \geq 1 - \frac{1}{k^2}$$

※ Chebyshev 不等式はかなり粗い不等式である。後の正規分布を例に取れば正規分布では k=2 のときの確率は 0.9545 であるが Chebyshev 不等式では 0.75 である（いくらなんでも粗すぎだと思わない？）。しかし、この不等式はどんな分布に対しても成り立つという長所もある。

2.4 確率分布

この節では様々な確率分布の種類について扱います。
用語：Bernoulli 試行：
定義：$(1)A_1, A_2, \ldots A_n$:⊥(←独立)。このとき $A_1 \cdots A_n$から任意の個数 m 個の試行を選び、$B_1, B_2, \ldots B_n$　とすると
$P(B_1 \cap B_2 \cap \cdots \cap B_n) = P(B_1)P(B_2)\cdots P(B_n)$

$(2)\{0,1\}$ からなる試行において $P(1) = p$,　$P(0) = 1-p$ であるような試行を n 回独立に行う。これを長さ n の Bernoulli 試行という。簡単に言えばコイン投げにおいて表=1、裏=0 としたようなもの。

2.4.1 二項分布

定義：X は r.v。長さ n の Bernoulli 試行において X = "1"の回数とすると
$X \sim Bi(n,p)$ で $f(x) = {}_nC_k p^x(1-p)^{n-x}(x = 1,2,3,\ldots,n)$. このとき X は二項分布に従うという。傍注にもあるように簡単なので詳しい説明は省略。プリント

※ f(x) が定義のようになる理由について：表が出る確率が p のコインを n 回投げて表が x 回出る確率を求めるという問題を想像すれば納得が行くはず。高校数学です！

を持っている人は p17 に書いてあるので見てください。

定理： $X \sim bi(n,p)$
(1)E(X)=np
(2)V(X)=np(1-p)
※証明には二項定理を使います。

二項定理：$(A+B)^m = {}_mC_0B^m + {}_mC_1AB^{m-1} + \cdots + {}_mC_mA^m = \sum_{k=0}^{m} {}_mC_kA^kB^{m-k}$

証明：

$$
\begin{aligned}
E(X) &= \sum_{x=0}^{n} x\,{}_nC_k p^x(1-p)^{n-x} && (4)\\
&= \sum_{x=0}^{n} x\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \qquad (\because \text{二項係数の定義から}) && (5)\\
&= \sum_{x=1}^{n} x\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \qquad (\because x=0 \text{ の項は } 0 \text{ なので}) && (6)\\
&= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!}p^x(1-p)^{n-x} \quad (\because \frac{x}{x!} = \frac{1}{(x-1)!}) && (7)\\
&= np\sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!}p^{x-1}(1-p)^{n-x} \quad (np \text{ をΣの外に出した}) && (8)\\
&= np\sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!((n-1)-(x-1))!}p^{x-1}(1-p)^{((n-1)-(x-1))} && (9)
\end{aligned}
$$

ここで $m=n-1, k=x-1$ と文字を置き換えると

※ (6)(n-x=(n-1)-(x-1))

$$
\begin{aligned}
&= np\sum_{k=0}^{m} {}_mC_k p^k(1-p)^m - k && (10)\\
&= np(p+(1-p))^m && (11)\\
&= np && (12)
\end{aligned}
$$

よって $E(X) = np$ は示せた。　分散について、$V(X) = E(X^2) - \{E(X)\}^2$ より $V(X) = E(X(X-1)) + E(X) - \{E(X)\}^2$。上と同じ方法で (4) 式の $(x-1) \to (x-2)$ となっているものを得る。さらに、(4) → (5) → (6) の変形の代わりに

$$E(X(X-1)) = n(n-1)p^2\sum_{x=1}^{n} \frac{(n-2)!}{(x-2)!((n-2)-(x-2)}p^{x-2}(1-p)^{((n-2)-(x-2))}$$

となるので m=n-2,k=x-2 と置き換えれば $E(X(X-1) = n(n-1)p^2$ である。

$$V(X) = n(n-1)p^2 + np - (np)^2 = np(1-p) \qquad (13)$$

証明終わり

2.4.2 Poisson 分布

Poisson 分布は二項分布においてpが小さく、nが大きいときに用いる確率分布である。（交通事故の発生など）

定理 $E(X) = np = \lambda$（一定）として、$n \to \infty$ の極限を考える

このとき ${}_nC_x p^x (1-p) n - x \longrightarrow e^{-\lambda} \frac{\lambda^n}{x!}$ と近似される。

定義　$f(x) = e^{-\lambda} \frac{\lambda^n}{x!} \Longleftrightarrow X \sim Po(\lambda), \lambda = np$

2.4.3 幾何分布 Ge(p)

定義：長さを指定しなBernoulli試行に対して $X =$ (はじめて1をとる（表が出る）回) とおくと $X \sim Ge(p)$

$$X \sim Ge(P) \Longleftrightarrow f(x) = p(1-p)^{x-1} (x = 1, 2, 3 \cdots)$$

公式（等比数列の和）：

$$\sum_{i=0}^{n} r^i = \frac{1 - r^{n+1}}{1 - r}, \quad \sum_{i=0}^{\infty} r^i = \frac{1}{1 - r}$$

これを使って

- $(1) E(X) = \frac{1}{p}$
- $(2) V(X) = \frac{1-p}{p^2}$

がわかる。・幾何分布の無記憶性
幾何分布では　　$P(X = a + b | X > b) = P(X = a)$
要はある時点まで0を取っているということがわかってしまえば、その先で1が初めて起こる確率は最初から試行をしてはじめて出る確率と同じと考えられるということ。

＜問題5＞

(1)

1. 表の出る確率が0.6であるようなコインを5回投げる試行を考える。表の出る回数XがX=3となる確率を求めよ
2. 表の出る確率が0.0002であるようなコインを10000回投げる試行を考える。表の出る回数XがX=3となる確率を求めよ
3. 表の出る確率が0.6であるようなコインを表が出るまで投げ続ける試行を考える。X回目に初めて表が出るとするときX=3となる確率を求めよ。
4. 表が出る確率がpであるようなコインを5回投げる試行を考える。B={表は3回出る}という事象が起こったという条件の下での事象A={1回目に表が出る}条件付確率 $P(A|B)$ を求めよ。

(2) $X \sim Ge(p)$ のとき、$P(X \geq x + y | X > x) = P(X > y)$ を示せ

2.4.4　正規分布

定義 $X \sim N(\mu, \sigma^2)$

$$\Longleftrightarrow f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

定理 $X \sim N(\mu, \sigma^2)$ のとき
(1)$E(X) = \mu$
(2)$V(X) = \sigma^2$
(3) 正規分布の線形変換はやはり正規分布する。$aX + b \sim N(\mu, \sigma^2)$
このとき $a = \frac{1}{\sigma}, b = \frac{-\mu}{\sigma}$ とすると標準化 $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$ →標準正規分布
正規分布には数表があり、（教科書 80 ページに載っています）様々な区間の確率を知ることができる。以下は重要な数値で覚えたほうが良い。

- $P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.683$
- $(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.9545$
- $(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.9997$

＜問題６＞

ある工程で作られたボールベアリングの直径は平均 $0.5998cm$, 標準偏差 $0.0005cm$ の正規分布に近似して分布することがわかっている。規格によるとボールベアリングの直径が $0.6000 \pm 0.0007cm$ の範囲のものが要求されている。さて、この工程で作られた製品が不合格となる確率を求めよ。

2.4.5　指数分布

定義: $X \sim E_x(\lambda)$

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

定理$X \sim E_x(\lambda)$ のとき

$$P(X > a + b|X > b) = P(X > a)\text{（指数分布の無記憶性）}$$

連続分布で無記憶性を持つのは指数分布だけである。

2.5　多次元の r.v

確率変数が X,Y と二つある場合について考える。

2.5.1　同時確率分布と周辺確率分布

1, 同時確率分布**定義：**X,Y を離散型の r.v とする。これを (X,Y) と表し、2 次元確率変数と呼ぶ。式で表すと

$$(X, Y) \text{の取りうる値の集合} = \{(x_i, y_j)|i, j = 1, 2, 3, \cdots\}$$

このとき (X,Y) の確率的挙動は

$$P\left((X,Y)=(x_i,y_j)\right)=P\left(X=xi,Y=y_j\right)=f(x_i,y_j)(i,j=1,2,\cdots)$$

によって定まる。これを (X,Y) の同時確率分布という。
連続型分布の場合の定義式は

$$P\left(a\leq bX\leq b,c\leq X\leq d\right)=\int_c^d\int_a^b f(x,y)dxdy(\forall a\leq b,c\leq d)$$

となり、f(x,y) を同時確率密度関数という。

2, 周辺確率分布
定義：X,Y は r.v $(i,j=1,2,3,\ldots,n)$
$P(X=x_i)=P(x_i,y_1)\cup P(x_i,y_2)\cup\cdots\cup P(x_i,y_n)=\sum_{k=1}^n f(x_i,y_k)=g(x_i)$
この g(x) を x の周辺確率分布という。
定義： 期待値の定義
$\phi(X,Y)$ を関数とする。$E\{(\phi(X,Y))\}=\sum_{i=1}^\infty\sum_{j=1}^\infty f(x_i,y_j)=\sum_{i=1}^\infty x_i(\sum_{j=1}^\infty f(x_i,y_j))=\sum_{i=1}^\infty x_i(g(x_i))$ X,Y を aX+b などのように変換しても同様に成立する。
¥textbf 定理：
(1)E(X+Y)=E(X)+E(Y)
(2)V(X+Y)=V(X)+V(Y)+2Cov(X,Y)
証明：$\mu_x=E(X),\mu_y=E(Y)$ とする。

※忘れているかもしれないけど、Cov は共分散ですよ

$$E(X+Y)=\sum_{i=1}^\infty\sum_{j=1}^\infty(x_i+y_j)f(x_i,y_j)$$

$=\sum\sum x_if(x_i,y_j)+\sum\sum y_jf(x_i,y_j)=E(X)+E(Y)$

$$V(X+Y)=E\{((x+y)-(\mu_x+\mu_y))^2\}$$

$=E((x-\mu_x)^2+(y-\mu_y)^2+2(x-\mu_x)(y-\mu_y))=V(X)+V(Y)+2Cov(X,Y)$
証明終わり

＜問題７＞

袋の中に同じ大きさの球 5 個が入っていて、それぞれ 1,2,3,4,5 の数字が記されている。非復元抽出によって 1 球ずつ取り出し、第 1 球の数字を X, 2 球の数字を Y とする。
(1)(X,Y) の同時確率分布を求めよ
(2)X と Y の周辺確率分布を求め、$E(X),E(Y)$ を求めよ
(3)$X+Y$ の確率分布を求め、$E(X+Y)=E(X)+E(Y)$ となっているかを確かめよ

2.5.2 独立性

定義： X,Y を離散型 r.v とする。

$$X,Y:\perp\Longleftrightarrow P(X=x_i\cap Y=y_j)=P(X=x_i)P(Y=y_j)\quad(\forall i,j)$$

$$\Longleftrightarrow f(x_i, y_j) = g(x_i)h(y_j) \qquad g, h \text{ は } x, y \text{ の周辺確率分布}$$

積事象の確率が確率の積で表せる→独立。これは多次元の場合も同じ。

定理：

$(1) X, Y : \perp \Longrightarrow E(XY) = E(X)E(Y)$

$(2) X, Y : \perp \Longrightarrow Cov(X, Y) = 0$

$(3) X, Y : \perp \Longrightarrow V(X + Y) = V(X) + V(Y)$

証明：

$(1) E(XY) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} x_i y_j P(X = x_i, Y = y_j)$.独立性から $P(X = x_i, Y = y_j) = P(X = x_i)P(Yy_j)$

$\sum_{i=1}^{\infty} x_i P(X = x_i) \sum_{j=1}^{\infty} y_j P(Yy_j) = E(X)E(Y)$

$(2) Cov(X, Y) = E(XY) - E(X)E(Y)$ だから (1) から $Cov(X, Y) = 0$

$(3) V(X + Y) = V(X) + V(Y) + 92Cov(X, Y)$ だから (2) より明らか 終わり。

2.5.3 和の分布

今、以下の条件が成立していると仮定する。n 個 の確率変数 $X_1 \cdots X_n$について

- $X_1, X_2, X_3, \ldots, X_n : \perp$
- $\forall X_i (i = 1, 2, 3, \ldots, n)$ について X_iは同一の分布に従う
- $E(X_1) = E(X_2) = \cdots = E(X_n) = \mu$
- $V(X_1) = V(X_1) = \cdots = V(X_n) = \sigma^2$

定理

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \text{とおくと } E(\bar{X}) = \mu, \quad V(\bar{X}) = \frac{\sigma^2}{n}$$

この定理は基礎統計の定理の中で最も重要なので必ず覚える！

証明：

$$(1) E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n) = n\mu$$

$$E(\bar{X}) = E(\frac{1}{n} \sum_{i=1}^{n} X_i) = \frac{1}{n} E(\sum_{i=1}^{n} X_i)$$

$$= \frac{1}{n} \times n\mu = \mu$$

$$(2) V(X_1 + \cdots + X_n) = E((X_1 + \cdots + X_n) - n\mu)^2)$$

$$= E((x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (X_n - \mu)^2 + 2((X_1 - \mu)(x_2 - \mu) \cdots (X_n - \mu))$$

$$= \sum_{i=1}^{n} V(X_i) + 2(Cov(X_1, X_2) + \cdots + Cov(X_n - 1, X_n))$$

$$\text{ここで条件から } Cov(X_i, X_j) = 0, \sum_{i=1}^{n} V(X_i) = n\sigma^2 \text{だから}$$

$$V(\bar{X}) = V(\frac{1}{n} \sum_{i=1}^{n} V(X_i)) = \frac{1}{n^2} V(\sum_{i=1}^{n} V(X_i)) = \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}$$

証明終わり

＜問題８＞

$(1)X_1, X_2, \ldots, X_n : \perp \sim N(0,1)$ とする。次の確率を求めよ。

1. $P(-1 \leq X_1 \leq 1)$
2. $P(-1 \leq \frac{X_1+X_2}{2} \leq 1$
3. $P(-1 \leq \frac{X_1+\ldots+X_5}{5} \leq 1$

$(2)X \sim N(2,3^2), Y \sim N(3,4^4), X \perp Y$ とする。次の確率を求めよ

1. $P(X+Y \geq 3)$
2. $P(3 \leq X+Y \leq 6)$

3　標本分布

3.1　母集団と標本

統計ではある母集団から取ったデータを使ってその母集団性質を知労とする。

標本（データ）：$x_1, x_2, \ldots x_n \Longleftarrow^{\text{無作為抽出}}$　母集団

⇒確率モデル化　たとえば $X_1, X_2, \cdots, X_n : r.v, \perp \sim N(\mu\sigma^2)$ →無作為抽出の表現形式のひとつ。ここでは正規分布で確率モデル化している。今、$\mu\sigma^2$はこの分布を支配しているパラメータだが、データを取った時点ではこれらの値は未知である。分布の特徴を知るにはパラメータを知る必要がある。このパラメータをデータから推測する。用語：
・μ(母平均)　$\leftarrow^{\text{推測}} \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$(標本平均)
・σ^2(母分散) $\leftarrow^{\text{推測}} S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$(標本分散)

3.2　統計量

母集団を推測するのに必要なデータから得られた数値を統計量という。注意すべきことは、統計量はあくまでr.vであるということである。統計量の例としては上に上げた標本平均、標本分散のほかに $s^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$(標本不偏分散 .$E(s^2) = \sigma^2$という性質がある) やデータそのものも統計量である。

3.3　再生性

再生性とは、独立な二つ以上のr.vが、同じ種類の分布に属しているとき、それらの和もまたその分布に属するという性質である。いくつかの具体例を紹介する。二つの $r.vX, Y : \perp$とする。

- $X_i \sim Bi(1,p) \Longrightarrow \sum_{i=1}^{n} X_i \sim Bi(n,p)$
- $X_i \sim Po(\lambda_i)(i=1,2,3,\cdots) \Longrightarrow \sum_{i=1}^{n} X_i \sim Po(\lambda_1+\lambda_2+\cdots)$
- $X_i \sim N(\mu_i,{\sigma_i}^2) \Longrightarrow \sum_{i=1}^{n} X_i \sim N(\sum_{i=1}^{n}\mu_i,\sum_{i=1}^{n}{\sigma_i}^2)$
- $X_i \sim N(\mu,\sigma^2)$(同一分布) $\Longrightarrow \sum_{i=1}^{n} \sim N(n\mu,n\sigma^2$

＜問題９＞

$X_1,\ldots,X_n : \perp \sim Bi(1,p)$.このとき再生性によって$\sum_{i=1}^{n} X_i \sim Bi(n,np)$ である。

$E(X)=np, V(X)=np(1-p)$ を証明せよ

3.4 大数の法則（LLN) と中心極限定理（CLT)

3.4.1 大数の法則 (LLN)

大数の法則とはデータの数を増やせば増やすだけ標本平均は母平均に近づくという法則である。式で書くとこうなる

条件：$X_1.\cdots,X_n : \perp \sim$ 同一分布.$\forall i$ に対して $E(X_i)=\mu, V(X_i)=\sigma^2 \ \ (*)$

条件 $(*)$ が成り立つとき、　$\forall \epsilon > 0$ に対して $\lim_{n\to\infty} P(|\bar{X}-\mu|>\epsilon)=0$

※数列の収束条件に似てる。

3.4.2 中心極限定理 (CLT)

中心極限定理とは n が十分に大きければ標本平均は近似的に正規分布するという定理である。証明は難しいので省略。教科書 164 ページに略証がある。式で書くとこうなる：

$$(*) \text{の条件下で} n \text{が十分大きい} \Longrightarrow \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

この定理は標準化などにも適用できる。標準化の場合→標準正規分布

＜問題１０＞

ある生産工程では製品の１０%が不良品であるという。この製品の山から無作為に４００個を取り出すとき (1) ５０個以上不良品である確率
(2)不良品の数 X が $40-c \le X \le 40+c$ 個となる確率が 0.9 になるような c を中心極限定理を使って求めよ

3.5 正規母集団からの標本

正規母集団からの標本とは、母集団が正規分布しているとデータを確率モデル化した場合の標本である。
まずは 1 標本の問題から考える。$X_1\ldots,X_n : \perp \sim N(\mu,\sigma^2)$ とする。

3.5.1　$\bar{X}$の標本分布

定理： $\bar{X} \sim N(\mu, \sigma^2)$　　**…** 正規分布の再生性による σ^2を既知の値として、μを$\bar{X}$から推定する。
正規分布の再生性から標準化 $Z \sim N(0,1)$ 標準正規分布の数表から

$$P(|Z| \leq 1.96) = 0.95$$

これを Z の定義に従ってμについて整理すると

$$P(\bar{X} - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{X} + 1.96\sqrt{\frac{\sigma^2}{n}}) = 0.95$$

この区間は確率0.95で未知のμを含むことがわかる。これをμに関する信頼係数0.95の信頼区間という。

＜問題１１＞

ある県の公立高校生 n 人と私立高校生 m 人を無作為に選び、学力テストを行った。公立高校生の結果を $X_1, \ldots X_n$, 私立高校生の結果を $Y_1, \ldots Y_n$とスル。公立高校生の試験結果は互いに独立に同一の正規分布 $N(\mu_A, {\sigma_A}^2)$ に従う。私立高校生の試験結果は互いに独立に同一の正規分布 $N(\mu_B, {\sigma_B}^2)$ に従う。また、公立高校生と私立高校生の試験結果は独立とする。
(1) 公立高校生の標本平均$\bar{X}$を求めよ
(2) 私立高校生の標本平均$\bar{Y}$を求めよ
(3) 両者の差$\bar{X} - \bar{Y}$の分布を求めよ

3.5.2　S^2, s^2の標本分布

$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2, E(X) = \frac{n-1}{n}\sigma^2$
$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, E(X) = \sigma^2$

3.5.3　χ^2分布

定義： $Z_1, Z_2, \ldots, Z_n \sim N(0,1)$（標準正規分布）とする。
このとき $Y = \sum_{i=1}^{k} {Z_i}^2 \sim \chi^2(k)$　これを自由度 k のχ^2分布という
$X_1, \cdots X_n : \perp \sim N(\mu, \sigma^2)$ のとき標本不偏分散 $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$に対して

$$Y = \frac{(n-1)s^2}{\sigma^2} = \sum \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$$

3.5.4　t 分布

$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1), \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}} \sim t(n-1)$$

$$\text{定義：} X \sim N(0,1), Y \sim \chi^2(k), X \perp Y \Longrightarrow \frac{X}{Y/k} \equiv t \sim t(k)$$

t 分布の特徴

- 左右対称
- グラフはN(0,1)よりも裾野が厚い形をしている
- 自由度を大きくするとN(0,1)に近づく

3.6　2標本問題

$X_1 \cdots X_n : \perp \sim N(\mu_1, \sigma_1^2), Y_1, \cdots Y_n : \perp \sim N(\mu_2, \sigma_2^2)$ とする。今$\mu_1 - \mu_2$の値を知りたい。平均に着目すると

$$\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n})$$

これを標準化すれば、σが既知のときは$\mu_1 - \mu_2$信頼区間を作れる。今、$\sigma_1 = \sigma_2 = \sigma$（未知）とする。

※ s という形で二つの数を合併している。数学的には内分の概念。

$$s_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2$$

$$s_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$s^2 = \frac{1}{m+n-2}\{(n-1)s_1^2 + (m-1)s_2^2\}$$

$$s^2 = \frac{1}{m+n-2}(\sum_{i=1}^{m}(X_i - \bar{X})^2 - \sum_{i=1}^{n}(Y_i - \bar{Y})^2)$$

ちなみに$\frac{(m+n-2)s^2}{\sigma^2} \sim \chi^2(m+n-2), \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{s^2(\frac{1}{m} + \frac{1}{n})}} \sim t(m+n-2), E(s^2) = \sigma^2$

であるから、数表を使って信頼区間を作ることができる。

＜問題１２＞

(1) 母平均$\mu = 3$の正規母集団 (σ^2は未知) から、大きさ $n = 10$ の標本を取り出す。

1. 標本平均$\bar{X}$が３と６の間にある確率を求めよ
2. 標本不偏分散 s^2が a を超える確率が 0.05 となるに定数 a の値を定めよ

(2) 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$\mu_1 = 5$, 母分散${\sigma_2}^2 = 4$ の正規母集団から大きさ $n = 8$ の標本を抽出する。二つの標本平均の差の標本分布を求めよ

4　推定

$X_1, X_2, \ldots, X_n : \perp \sim$ 分布 (任意)(μ, σ^2) を前提とする。

4.1　推定量

推定量とは推定に用いる統計量のことである。その実現値を推定値という。推定には二種類あり、点推定と区間推定がある。

4.2　点推定

点推定では、ある一点について推定する。

4.2.1　不偏推定量

定義：$\hat{\theta}$はθの不偏推定量 $\Longleftrightarrow E(\hat{\theta}) = \theta$
$\hat{\theta}$はθの値を正確に推定していると考えられる。
例：$x_1, x_2, \ldots . x_n \sim N(\mu, \sigma^2) \rightarrow \bar{x}$は$\mu$の不偏推定量。$s^2$は$\sigma^2$の不偏推定量。でも s はσの不偏推定量にはなっていない。これは非線形の変換に対して不偏性が保存されないことを表している。

4.2.2　モーメント法

定義：X:r.v

- $\mu_r \equiv E(X^r)$: r 次母モーメント
- $\hat{mu_r} \equiv \frac{1}{n}\sum_{i=1}^{n} {X_i}^r$: r 次標本モーメント
- $\mu'_r \equiv E\{(X-\mu)^r\}$: 平均回りの r 次母モーメント
- $\hat{\mu'_r} \equiv \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^r$: 平均回りの r 次標本モーメント

モーメント法とは、r 次の母モーメントを標本モーメントで推定するという方法である。ちなみに $\mu_1 = \mu, \hat{\mu_1} = \bar{X}, \mu'_2 = \sigma^2$　である。

4.2.3　最尤法

試験版範囲外なので割愛

4.3　区間推定

区間推定では事前に$\alpha(0 < \alpha < 1)$ を定めて
$P(L \leq \theta \leq U) = 1 - \alpha$($\theta$は調べるべきパラメーター) となるような区間 (L, U) を求める（通常$\alpha = 0.1$ や 0.05) ここで、$1 - \alpha$　を信頼係数、区間 (L, U) を信頼区間という

4.3.1　CLT を使った区間推定

・二項分布の例
$X_1, X_2, \ldots, X_n : \perp \sim Bi(1, p)$ とする (P は未知)

nが大きい時、CLTによって$X \sim N(p, \frac{p(1-p)}{n})$と近似できる。
よって、標準化$Z = \frac{\bar{X}-p}{sqrt\frac{p(1-p)}{n}} \sim N(0,1)$したがって

$$P(|Z| \leq Z_{\frac{\alpha}{2}}) = 1-\alpha \qquad p\text{について整理すると}$$

$$P(\bar{X} - Z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \bar{X} + Z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}) = 1-\alpha$$

しかしこのままでは区間に未知数pが含まれるため意味がない。nが十分大きい状況では$p(1-p) \quad \bar{X}(1-\bar{X})$と近似できるので

$$[\bar{X} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}]$$

という近似的な区間を得る。この方法は他の分布に対しても使える。

＜問題１３＞

(1)$X_1, \ldots, X_n : \perp \sim Bi(1,p)$.このとき母平均$\mu = p$,母分散$\sigma^2 = p(1-p)$である。母平均$\mu = p$の推定量はモーメント法で考えれば$\bar{X}$となり、これは不偏である。一方母分散$\sigma^2 = p(1-p)$推定量はモーメント法で考えれば$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$となる。(あるいは$s^2$でもよい)。また、$\sigma^2 = p(1-p)$であるから$\hat{\sigma^2} = \bar{X}(1-\bar{X})$という推量も自然であろう。実は今の場合$S^2 = \hat{\sigma^2}$となる。このことを示せ。

(2)参議院選挙で、ある地方区から大きさ１００の無作為標本を取り出し、意見を聞いたところA候補者を支持する者が６０人いた。A候補者の支持率pに冠する信頼係数0.95の信頼区間を作れ。無作為標本の大きさを1000とし、支持者が600人であった場合の信頼区間も計算せよ。

4.4 検定

4.4.1 検定の考え方

検定ではある確率変数Xに対してそのパラメータが本当に正しいか（またはXに変化があった場合本当に変化があったか）を調べる手段ある。
1, 二つの仮説
今調べるべきパラメータを$r.v X$の平均μとする。今$\mu = \mu_0$であるかどうかを検定するために二つの仮説を立てる

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0$$

前者を帰無仮説、後者を対立仮説という。どちらの仮説が正しいかを調べるためにn個のデータを集める。

$$X_1, X_2, \ldots, X_n : \perp \sim N(\mu\sigma^2)(\sigma : \text{既知})$$

まずは$\bar{X}$に注目する。$|\bar{X} - \mu_0|$が
大きい時→H_0を棄却（H_1を選択）

小さいとき→ H_0を棄却しない
という方法を取りたいが、具体的な値を見ても大きいか小さいかはよくわからない。
→そこで標準化することで標準偏差のメモリで判断する必要が出てくる。標準化をZとすると $Z \sim N(0,1)$

$\frac{|\bar{X}-\mu_0|}{\sqrt{\frac{\sigma^2}{n}}} > c \Longrightarrow H_0$を棄却。$\frac{|\bar{X}-\mu_0|}{\sqrt{\frac{\sigma^2}{n}}} \leq c$ ここで c は判断の基準とする定数

2,cの決定にあたって起こりうる二つの誤り

- 第1種の誤り　→ H_0が正しいときに H_0を棄却してしまうという誤り
- 第2種の誤り　→ H_1が正しいときに H_0を棄却しないという誤り

第1種の誤りを極力犯さないようにcを決定する。
→そこで、P(第一種) $= \alpha(0.05, 0.1etc)$　などをあらかじめ設定する。これを有意水準という。
まず、H_0が正しいと仮定する。$\mu = \mu_0$で、$Z \sim N(0,1)$ がわかっているので

$$P(\frac{|\bar{X}-\mu_0|}{\sqrt{\frac{\sigma^2}{n}}} > c) = \alpha$$を満たす c を数表から探す。

あとは定めたcを基準にして検定をすればよい。この場合標準正規分布の両側から(絶対値で)区間を押さえているので有意水準 α の両側検定という。

4.4.2　t検定

考え方の節では分散を既知のものとして扱ったが、分散が未知の場合かわりに標本不偏分散を使って評価することになる。このとき標準正規分布の変わりにt分布を使う。

$$X_1, X_2, \ldots, X_n : \perp \sim N(\mu\sigma^2)(\sigma : \text{未知})$$

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0$$

$\frac{\bar{X}-\mu}{\sqrt{\frac{s^2}{n}}} \equiv t \sim t(n-1)$　この t が検定統計量である。

$|t| > c \rightarrow H_0$棄却, $|t| < c$　→棄却しない　　という条件で c を決定する。あとは有意水準αを設定して数表から数字を探して c を決定する。式で言うとこういうこと

$t \sim t(n-1$のとき $P(t > t_{\alpha(n-1)}) = \alpha$であるから $P(|t| > t_{\frac{\alpha}{2}(n-1)}) = \alpha$　　$\therefore c = t_{\frac{\alpha}{2}(n-1)}$

これを有意水準 αの両側 t 検定という

$$\text{もし } H_0 : \mu = \mu_0, \qquad H_1 : \mu > \mu_0$$

という仮説であったとすれば判断基準は
$t > c \rightarrow H_0$棄却, $t < c$　→棄却しない　　となるので
$c = t_{\alpha(n-1)}$となる。　これを有意水準αの片側 t 検定という。

4.4.3　母分散の検定

　つぎに注目するパラメータを分散σ^2　としてみる。
このとき注目すべき統計量はs^2である。

$$H_0 : \sigma^2 = {\sigma_0}^2 \qquad H_1 : \sigma^2 \neq {\sigma_0}^2$$

$$Y \equiv \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \quad Y \text{が検定統計量}$$

$$Y > c \rightarrow H_0\text{棄却} \quad Y < c \rightarrow\text{棄却しない} \quad \text{として } c \text{ を決定する}$$

あとは￥chi^2分布の数表を使うだけでやることは上の二つと同じなので説明は割愛。

4.4.4　2 標本問題

$$X_1, \ldots X_m : \perp \sim N(\mu_1, \sigma^2)$$

$$Y_1, \ldots Y_n : \perp \sim N(\mu_2, \sigma^2)(\text{分散は共通})$$

$$H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2$$

$\mu_1 - \mu_2$に注目する。　検定統計量：$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{s^2(\frac{1}{m}+\frac{1}{n})}}$　$\sim t(m+n-2)$ 後は同じなので割愛

4.4.5　CLT を使った検定

　今まで正規母集団に対する検定を行ってきたが、CLT を使うことで二項母集団の母比率 p などの検定も行うことができる

$$X_1.X_2, \ldots, X_n \sim Bi(1, p)$$

まずは$\bar{X}$に注目　nが大きい時CLTによる近似：　$\bar{X} \sim N(p, \frac{p(1-p)}{n})$ あとはセオリーどおり

$$\text{検定統計量：} Z \equiv \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

$$H_0 : p = p_0 \qquad H_1 : p > p_0$$

あとは有意水準 α を定めて c を求めるだけ。説明は割愛。CLT があれば様々な分布において検定が可能。しかも検定はワンパターン。

＜問題 14＞

ある薬品中に含まれるケイ酸マグネシウムの含有率は$\mu = 0.8$ %と成分表示されている。その薬品についてケイ酸マグネシウムの含有率を 4 回測定したところ、標本平均$\bar{X} = 0.868$ %、標本不偏部さん $s^2 = 0.056$ %を得た。この含有率は成分表示より著しく多いといえるか。正規母集団 $N(\mu, \sigma^2)$ を仮定し、有意水準 0.05 で検定せよ。有意水準 0.01 ではどうか。

　いじょー。詳しいことは教科書を読みましょう。