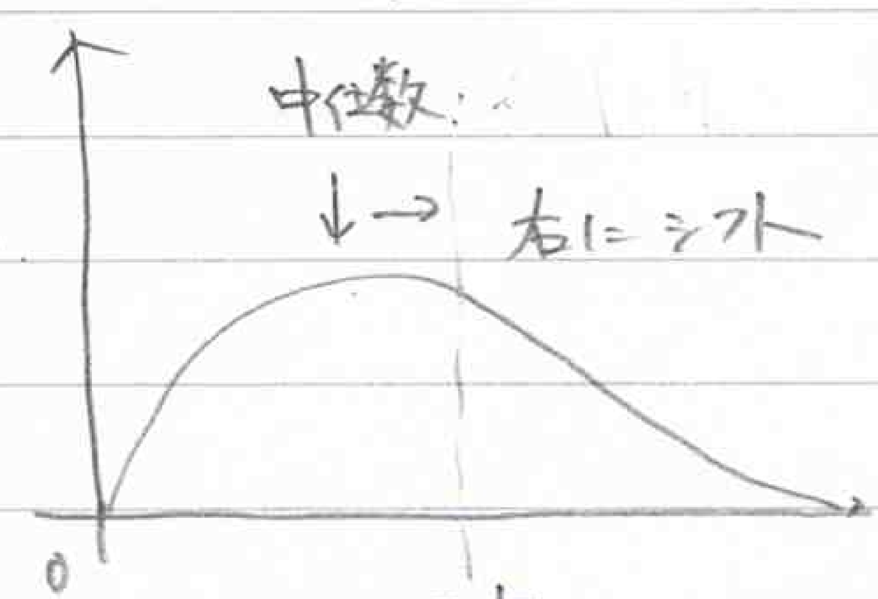
基礎統計

データ：「東大生の親は金持ちだ」

⇒ 親の平均所得が1000万をこえている。

一般家庭の所得資産



世帯数

平均月収 50万2千円

平均以下の世帯が6割

中位数：(上から50%)

= median

・中位数と平均に平均がある。

→ 総資産になると差は広がる。

median : 1006万円

平均 : 1772万円

・散らばりの尺度　分散　$\sqrt{\text{分散}}$ = 標準偏差

バラツキの尺度　標準偏差

x_1

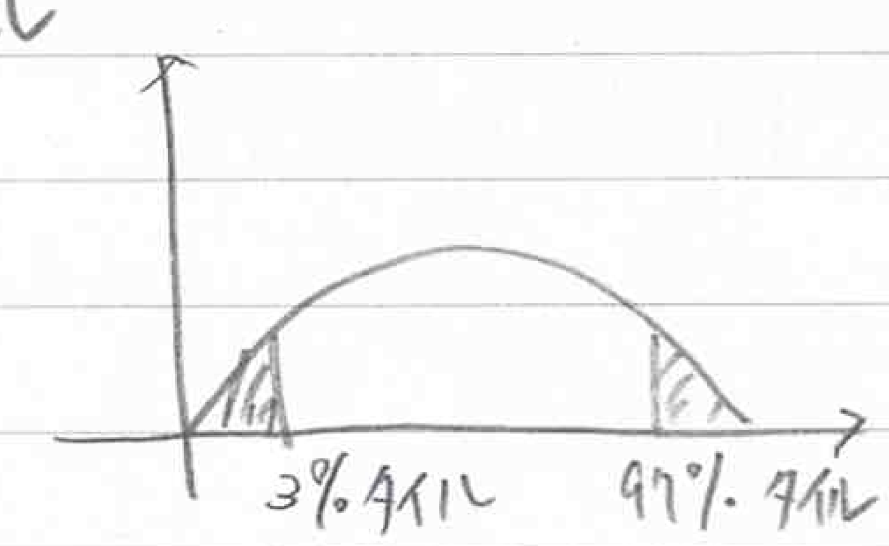
$\vdots$

x_n

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

なぜバラツキの尺度が必要か？

・タイル



7.6kg　11.0kg

公式を使う

$\boxed{\text{分散}}$ = (中心からの距離)^2 の期待値

$$\text{分散} = \frac{1}{3-1}\left((9000-8000)^2 + (8000-8000)^2 + (7000-8000)^2\right)$$

N=3

$= 1 \times 10^6$

7000g

8000g

9000g

標準偏差 = $\sqrt{\text{分散}}$ = 10^3 ≒ 中心からの平均距離

散らばり

1. 不平等，製品の精度
(2). 推定の精度

ジニ係数 ⇒ 不平等

1989 USA ジニ係数 0.4
日本 0.43

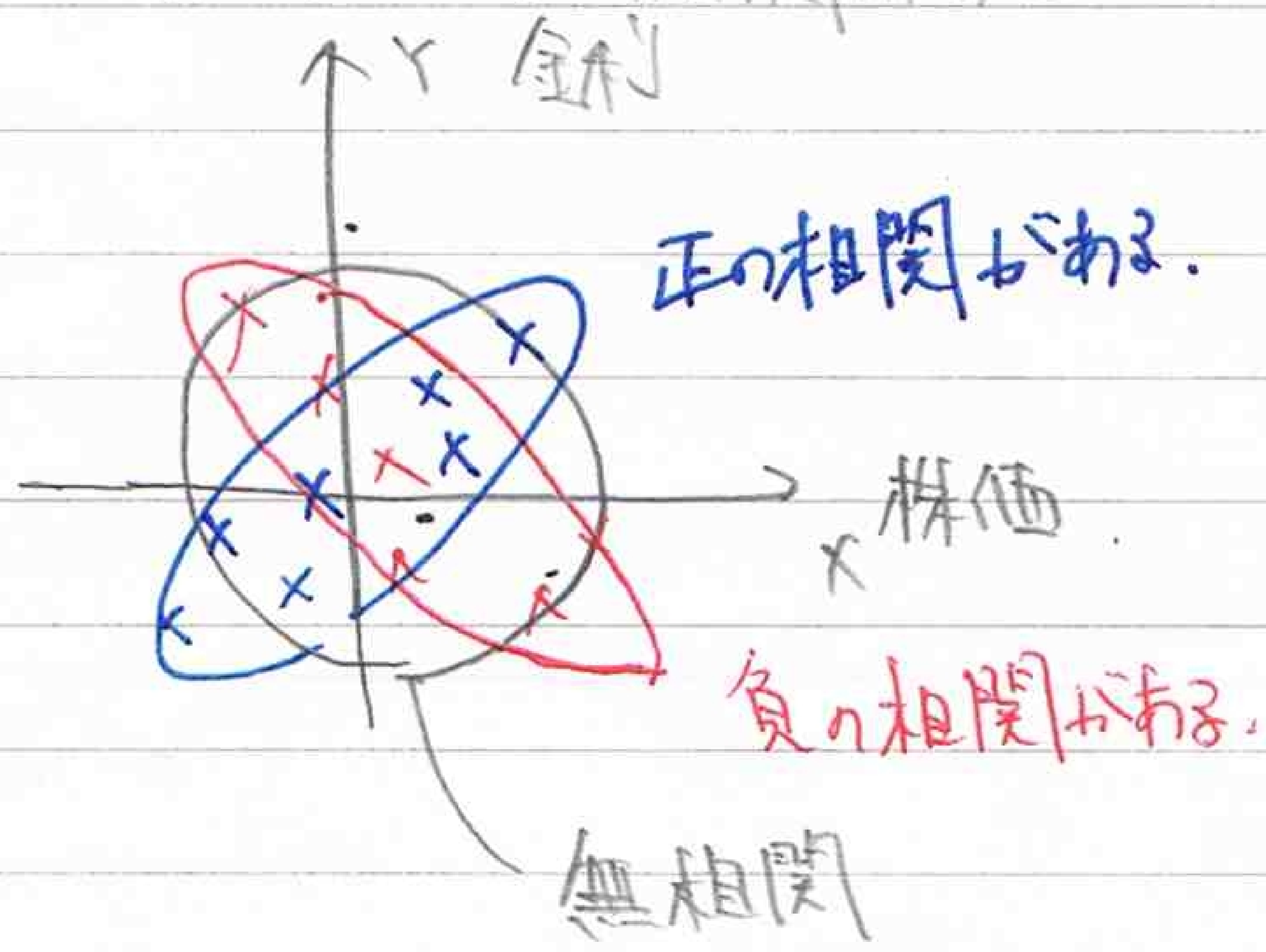
日本は平等社会(?)という考えに疑い。

0≦ジニ≦1 ← 何を基にはかっているのか?
完全平等 完全不平等

→ ・不平等度上昇 ・何故か?

2つ以上の変数の関係

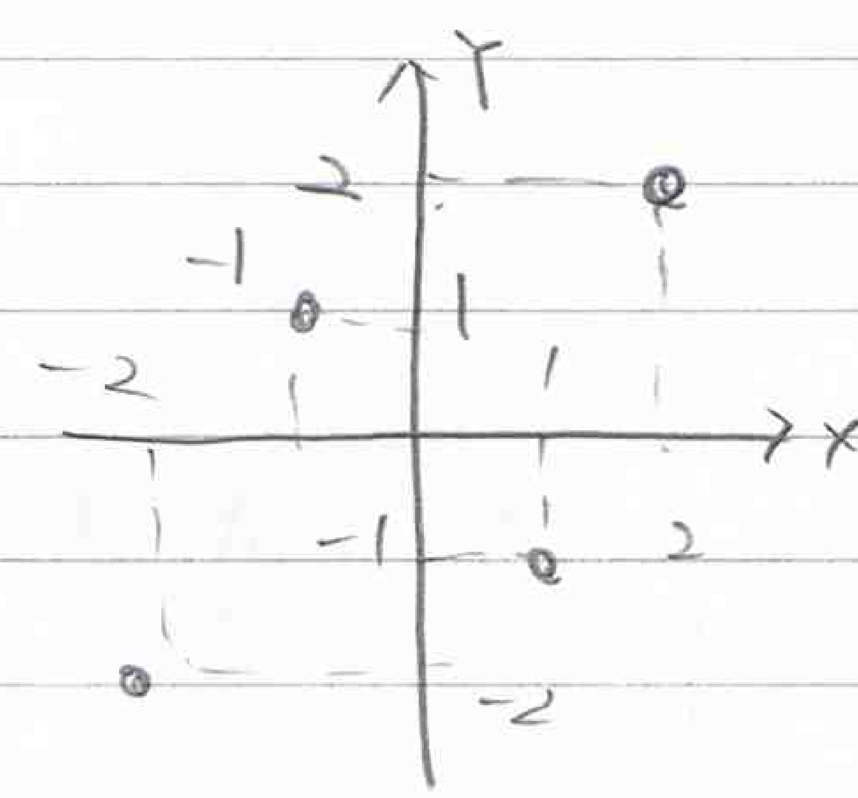
X, Y ⇒ 相関係数 正の相関，負の相関，無相関



相関を数値化

	X	Y
1	2	2
2	1	-1
3	-1	1
4	-2	-2

⇒



$$\bar{X} = \frac{1}{4}(2+1-1-2) = 0$$
$$\bar{Y} = \frac{1}{4}(2-1+1-2) = 0$$

	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$
1	+	+	+
2	+	−	−
3	−	+	−
4	−	−	+

株価　　　金利

$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \Rightarrow$ 共分散　convariance

$= \frac{1}{N}\left\{(x_1-\bar{x})(y_1-\bar{y})+\cdots+(x_N-\bar{x})(y_N-\bar{y})\right\}$

正 or 負 → 相関

共分散 = 200

X ドル → セント　　X は 100倍

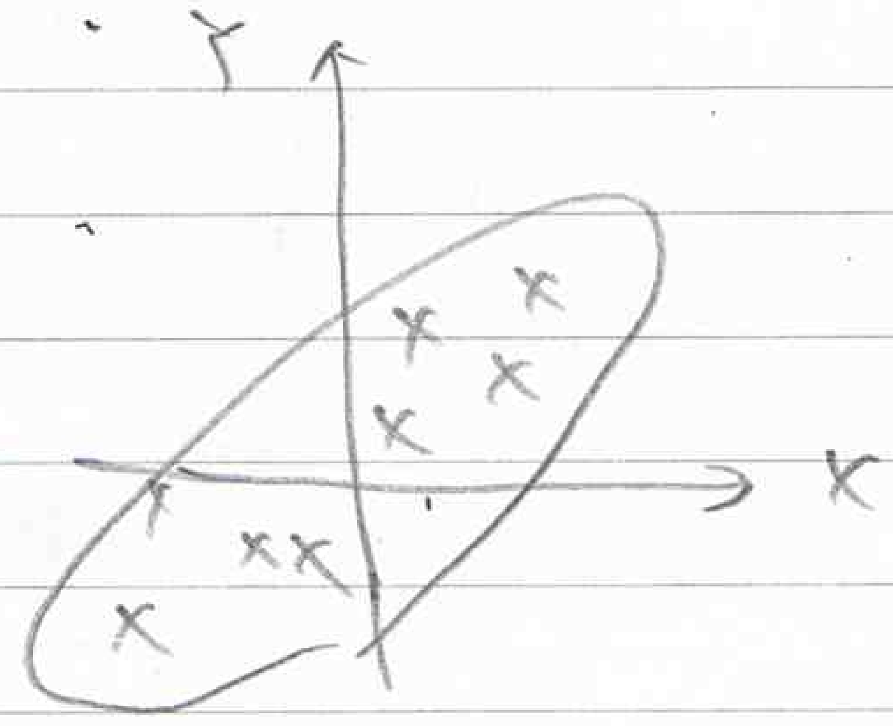
Y ドル → セント　　Y 〃

共分散は 100^2倍 になってしまう

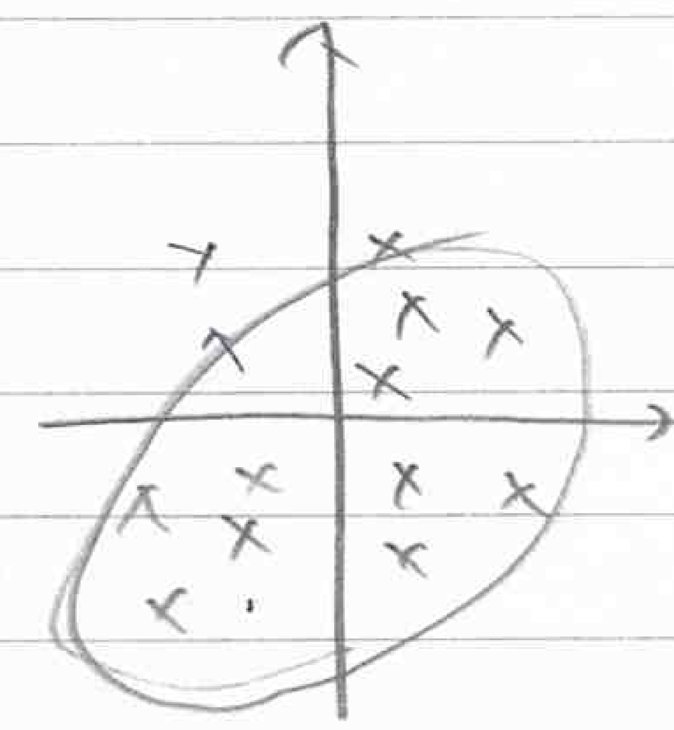
単位に無関係な相関の尺度

(単位によらず)

強い正の相関



弱い正の相関



⇒ 単位に無関係な値を出すにはどうするか？　教科書49ページ

$$\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{N}\sum(x_i-\bar{x})^2 \cdot \frac{1}{N}\sum(y_i-\bar{y})^2}}$$

$$= \frac{\text{共分散}}{\text{Xの標準偏差} \times \text{Yの標準偏差}}$$

相関係数 1. 単位と無関係

直線的な関係 2. 原点とも無関係

3. $-1 \leq$ 相関係数 ≤ 1　(シュワルツ不等式)

教科書 p49へ50

正の傾きの直線上に存在する

$\rho = 1$

$\rho = 1$

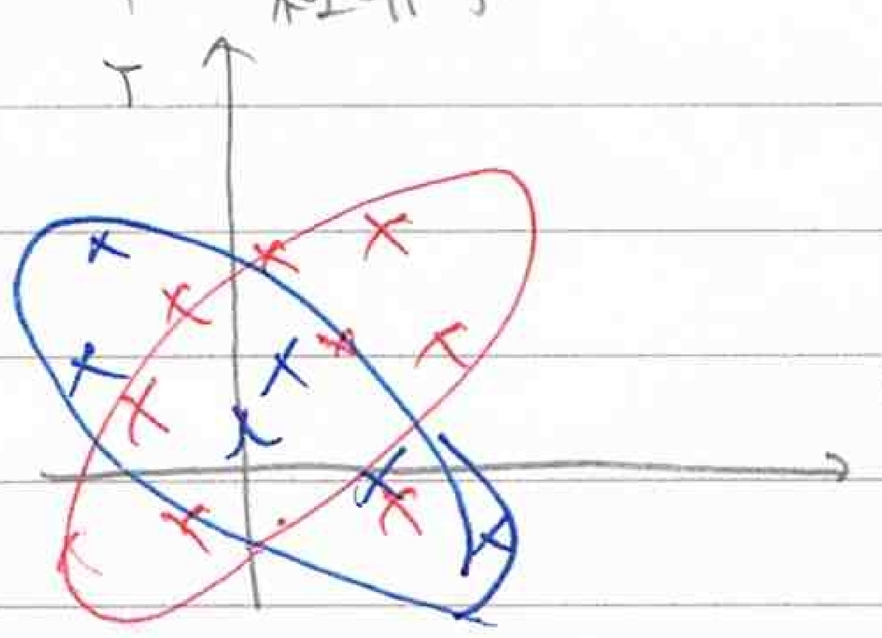
No.　　Date

経験的に相関係数の絶対値0.5が肉眼で分かる「相関」

相関と因果

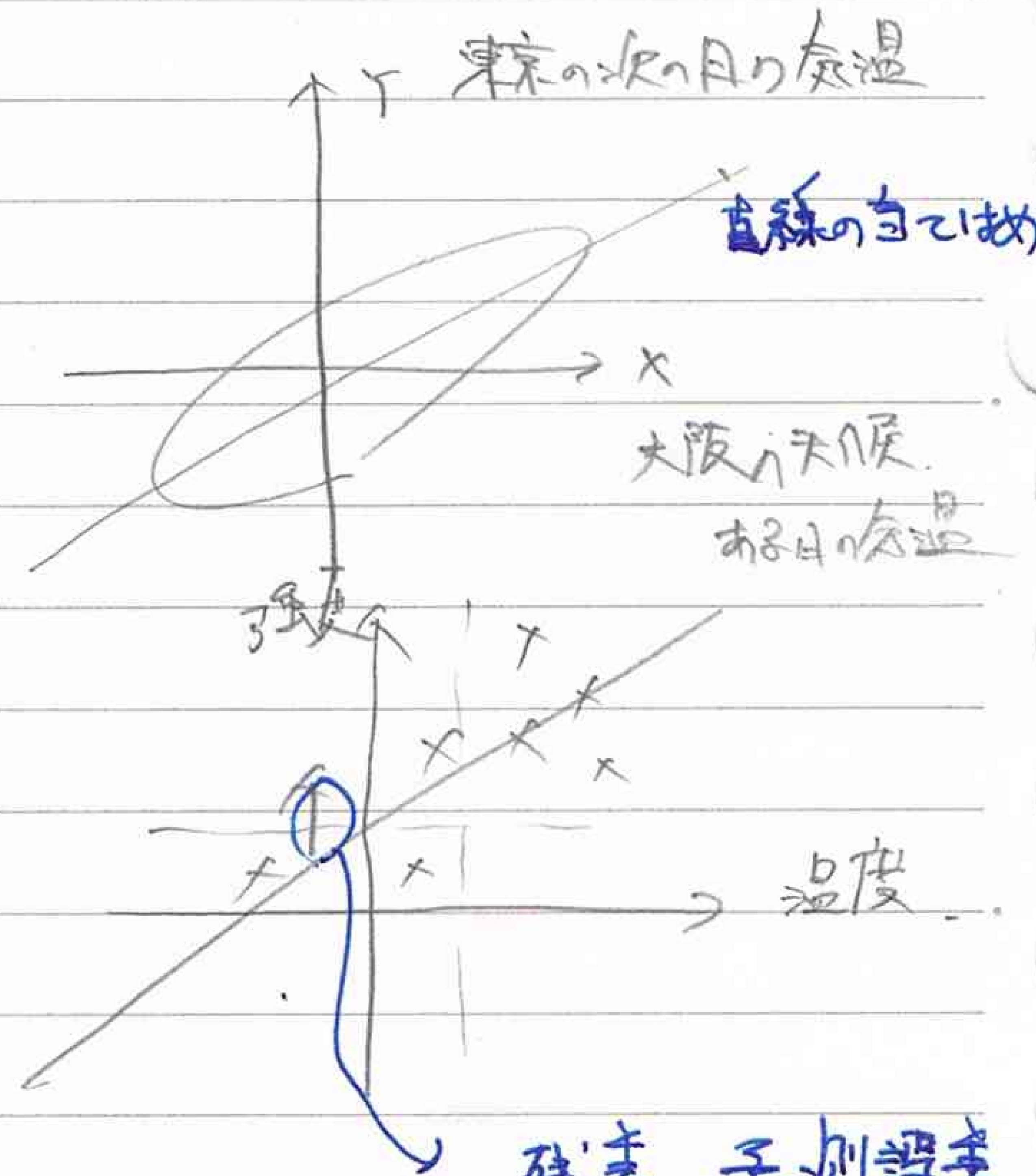
$X \Leftrightarrow Y$　　$X \to Y$
　　　　原因　結果

▷ 回帰分析



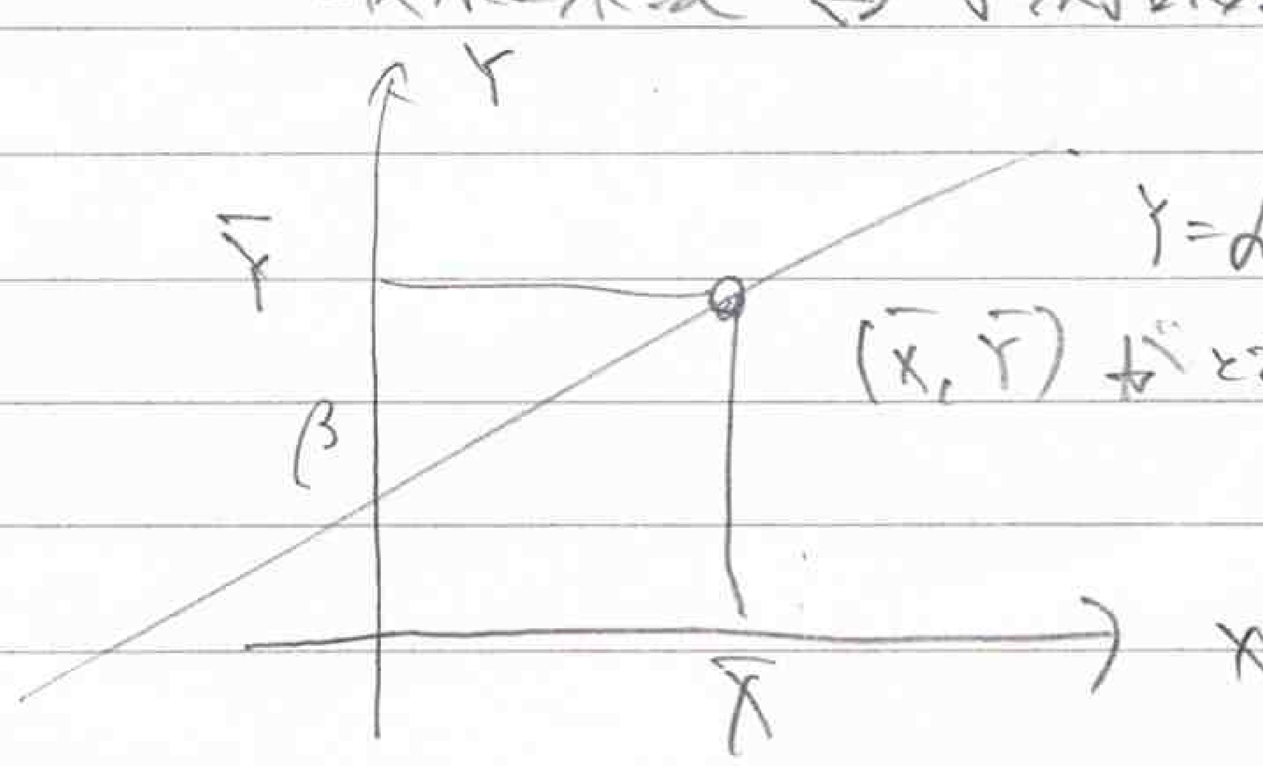
10人の学は
予測 →

コントロール →



$\hat{e}_1$ | … 予測誤差
$\hat{e}_2$ | の二乗の和を最小にする
$\hat{e}_3$ | ような直線を当てはめる．
$\hat{e}_4$ | ⇒ 公式計算機　Excel

最小二乗法 ⇔ 予測誤差の絶対値の和を最小．



$Y = \alpha + \beta X$
$(\bar{X}, \bar{Y})$ がとおることもね．

X：説明変数
　独立変数
Y：被説明変数

決定係数

残差の二乗和

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

α：係数
$\hat{\alpha}$：推定値

R^2が大 ⇔ あてはまりやすい　　$1 \geq R^2 \geq 0$．

Date . . No.

相関係数：ρ . $-1 \leq \rho \leq 1$ $R^2 = \rho^2$

一般的

回帰分析で説明変数が2つ以上重回帰分析

確率

。確率変数 独立性．条件付き確率

	G	B	
鬼	200	100	300
仏	100	200	300
	300	300	

A 講義がうまい．

B 仏 $P(A) = \frac{1}{2}$ $P(B) = \frac{1}{2}$

A^C：Aではない $P(A^C) = \frac{1}{2}$ $P(B^C) = \frac{1}{2}$

うまい先生に当たったら $P(鬼) = 2/3$

$P(B) = 1/2 \to P(B/A) = 1/3 \Rightarrow P(B^C/A) = 2/3$

条件付き確率

	G	B	
鬼	1/3	1/6	1/2
仏	1/6	1/3	1/2
	1/2	1/2	

周辺確率

確率変数 $P(B/A)$

$= \dfrac{1/3}{1/2}$

$= \dfrac{P(A \cap B)}{P(A)}$ 同時確率 周辺確率

$P(B/A) = P(B)$ Aは情報として役に立たない．

	G	B	
鬼	150	150	300
仏	150	150	300
	300	300	600

$P(仏) = \frac{1}{2}$

$P(仏|G) = \frac{1}{2}$

$P(仏/B) = \frac{1}{2}$

$P(鬼/B) = \frac{1}{2}$

$= P_r(鬼) = \frac{1}{2}$

すべてのX，Yについて

$P_r(X/Y) = P(Y)$

$\Leftrightarrow$ このときXとYは独立である $\Leftrightarrow$ $P(B|A) = P(B)$

No. Date . .

$$P(B|A) = \frac{P(A \wedge B)}{P(A)} = P(B) \Leftrightarrow P(A \wedge B) = P(A) \times P(B)$$

$$\Leftrightarrow \frac{P(A \text{ and } B)}{P(B)} = Pr(A|B)$$

$$= P(A)$$

Q ベイズの定理 ・診断 ・ジャンク（スパム）の検出
・ウイルスの検出

スパム → Sとおく。

ex. 件名がない・短い・画像付き。

Pr（病）$= \frac{1}{1000}$ Pr（健）$= \frac{999}{1000}$

Pr（陽｜病）$= 0.95$ Pr（陽｜健）$= 0.01$

$\Rightarrow$ Pr（病｜陽）

$= \dfrac{Pr(\text{病かつ陽})}{Pr(\text{陽})}$

$= \dfrac{\text{病} \times \text{陽}}{\text{健} \times \text{陽} + \text{病} \times \text{陽}}$

0.001×0.95 0.095

$\longrightarrow 0.01$

0.999×0.01 0.001×0.95

P（病｜陽）$= \dfrac{0.0095}{0.01}$

$= 0.086$ 1/1. 以下

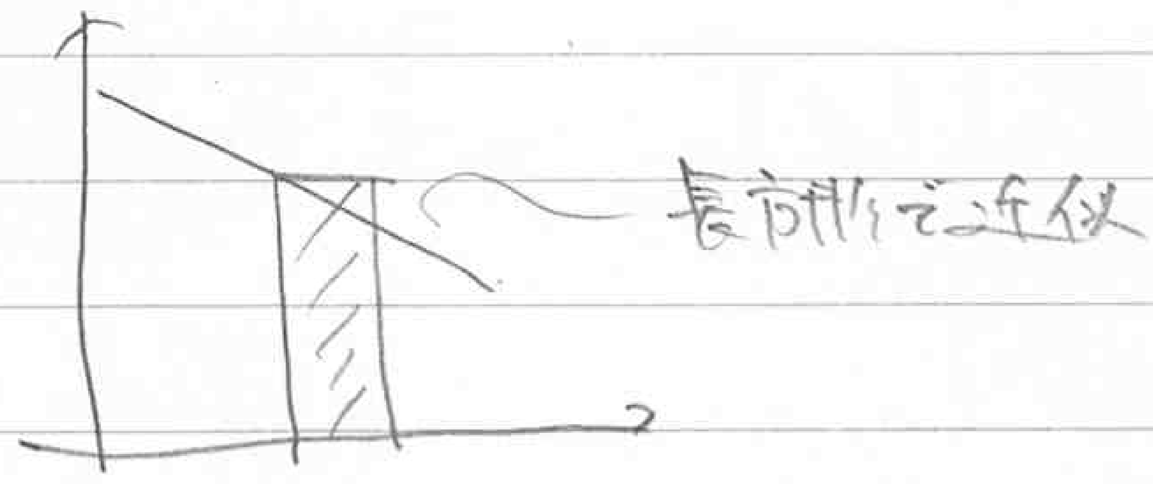
P(A) Spam

$\xrightarrow{0.1}$ 短い P(B|A)

$\longrightarrow$ 件名がない

$\xrightarrow{0.2}$ 画像

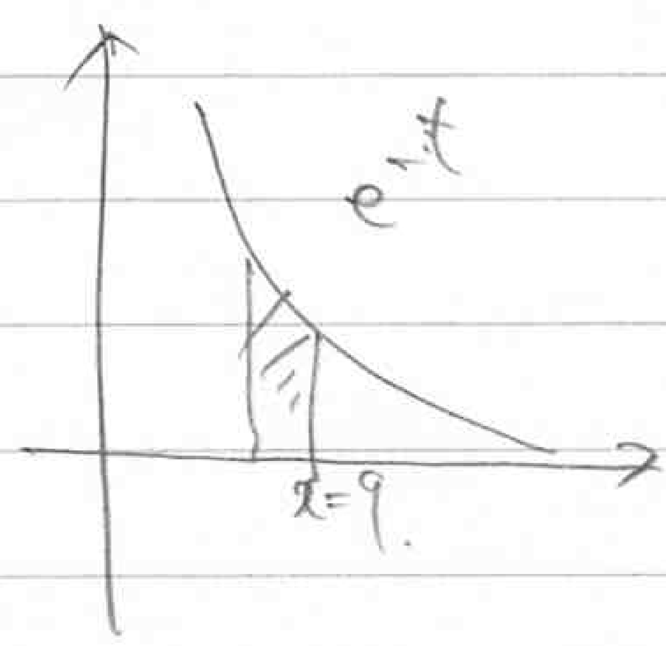
連続型確率変数 : 密度関数 (Density)



$p(9 < x < 9.1) = f(9) \times 0.1$

密度 × 区間幅

= 確率　　$f(x) = e^{-t}$



$p(x = 9.00) = 0.$

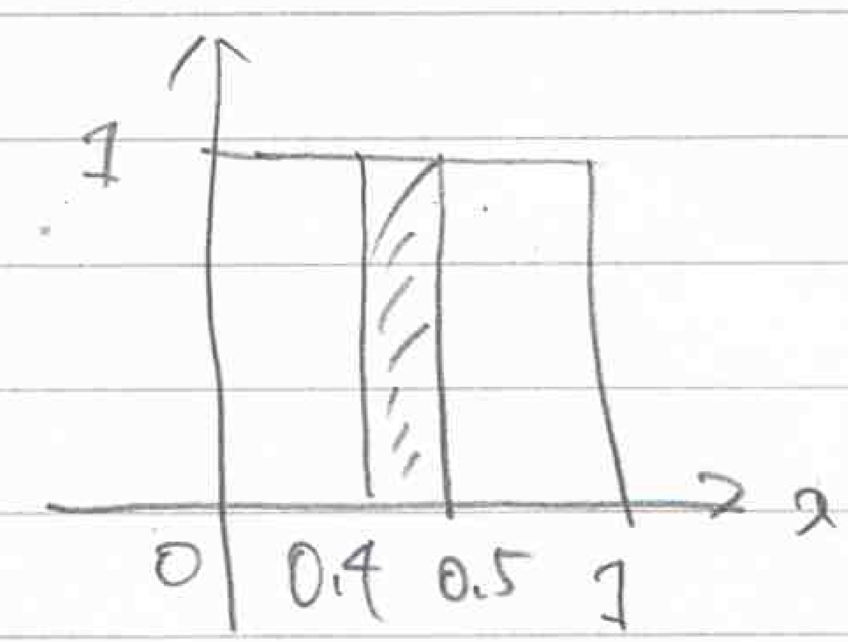
$p(x = 9) = 0$

幅がないため.

一様分布　$0 < U < 1$

$p(x = 0.5) = 0$

$p(0.4 < x < 0.5) = 0.1 \times 1 = 0.1$



密度 : 高さ

*** 分布に従う = 自然、社会現象 ⇒ 正規分布 (誤差) $-\infty < x < \infty$

→ 中心極限定理によって理由付けされる。

$\int_{-\infty}^{\infty} f(x)\,dx = 1$

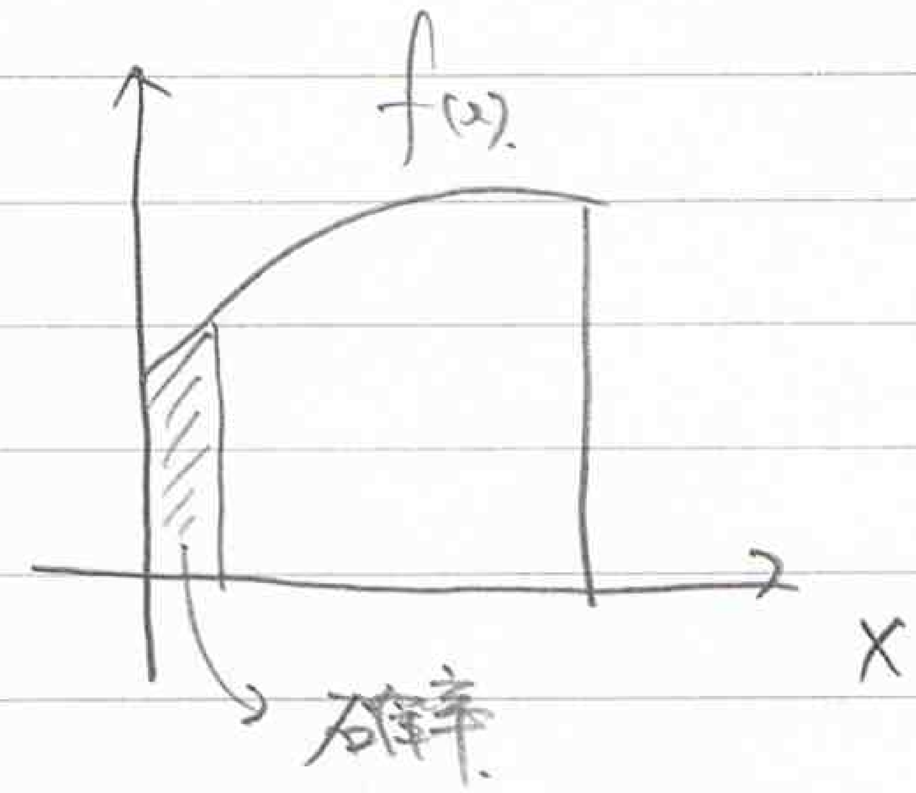
連続分布の期待値

$E(x) = x_1 \times \Delta f(x_1)$

$+ \cdots$

$+ \underline{x_N} \times \Delta f(x_N)$

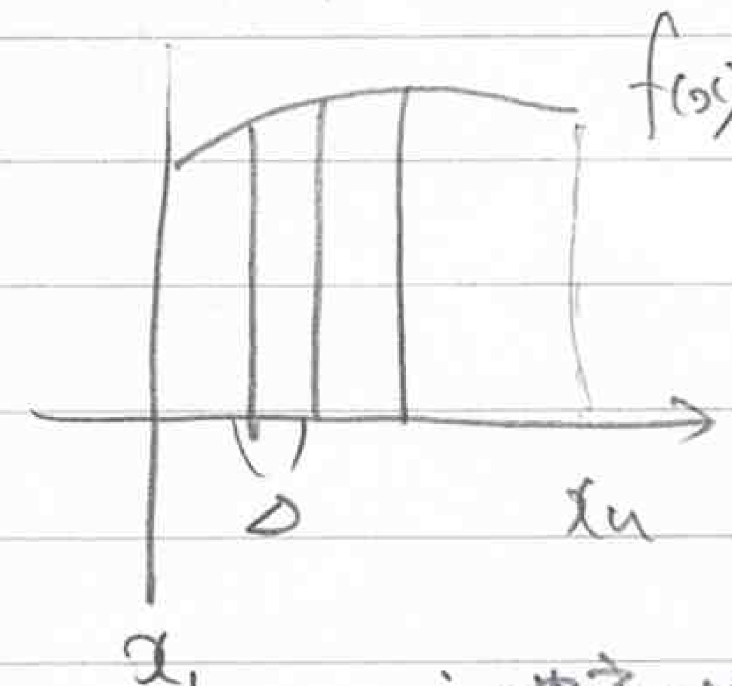
値



$\int_{-\infty}^{\infty} f(x)\,dx = \Delta f(x_1) + \cdots + \Delta f(x_N) \quad \Delta \to 0$

$\int_{-\infty}^{\infty} x f(x)\,dx = \Delta x_1 f(x_1) + \cdots \Delta x_N f(x_N) \quad \Delta \to 0.$

$E(x^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx.$



$\int_{x_1}^{x_n} f(x)\,dx$

$- P(x_1 < x < x_n)$

※確率は積分で表せる!

$$f(x) = \begin{cases} 0 & \text{それ以外} \\ 1 & (0,1) \end{cases}$$

$$E[x] = \int_0^1 x \cdot 1\, dx = \left[\frac{1}{2}x^2\right]_0^1 = \frac{1}{2}$$

$$V(x) = E\left((X - E(X))^2\right)$$

$$= \int_0^1 \left(x - \frac{1}{2}\right)^2 dx$$

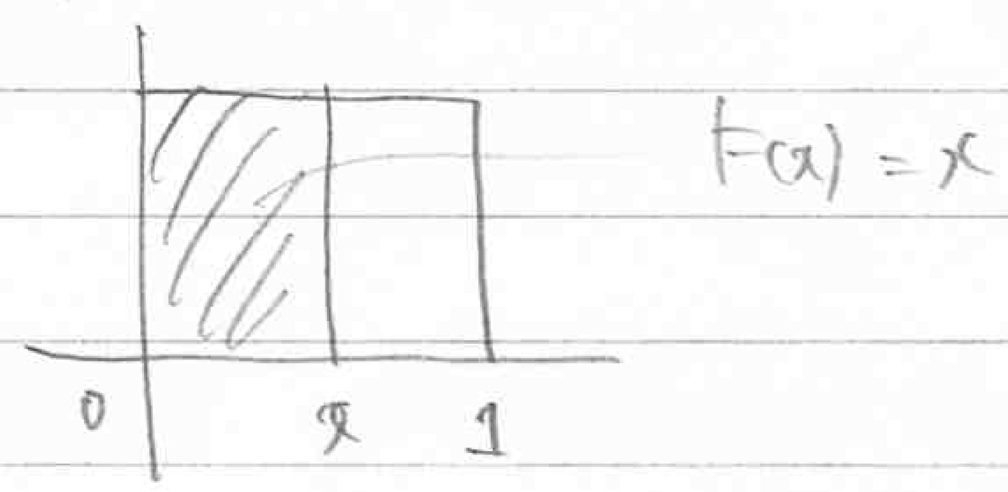
$$= \left[\frac{1}{3}\left(x - \frac{1}{2}\right)^3\right]_0^1$$

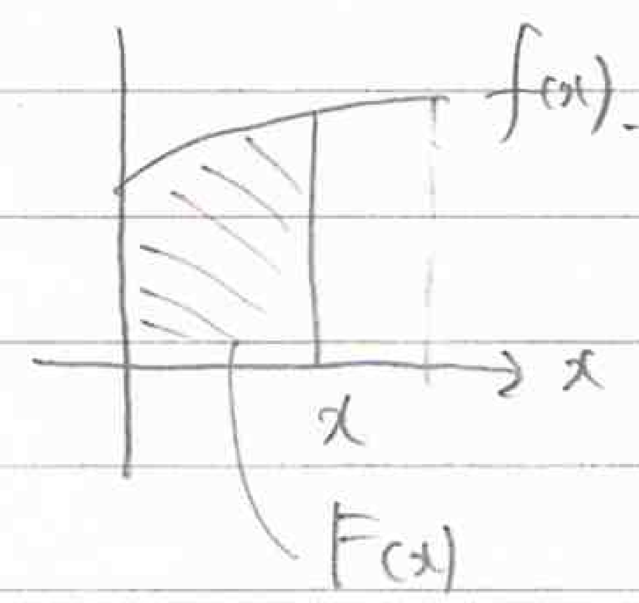
$$= \frac{1}{12}$$

$$\therefore \sqrt{V(x)} = \frac{1}{2\sqrt{3}}$$

p97. 累積分布関数

$Pr(X<x) = F(x)$　　Xが$(0,1)$の一様

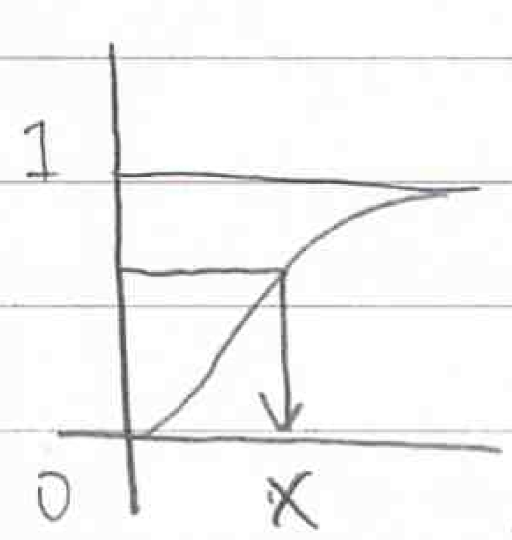




$$F(x) = \int_{-\infty}^{x} f(x)dx \qquad F'(x) = f(x) \qquad (\text{分布関数})' = \text{密度}$$

U… 一様乱数　　(0,1)の一様分布.

シミュレーション　　$X = F^{-1}(U)$



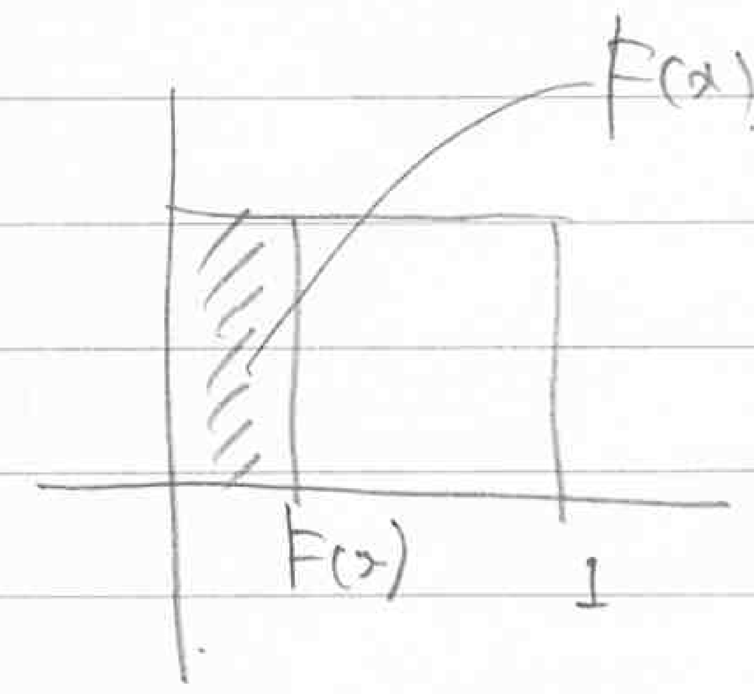
$$F(x) = \int_{-\infty}^{x} f(x)dx$$

≥0

単調増加（広義）

Xは分布関数F(x)を持つ.

$p(F^{-1}(U) < x)$ 　 $F^{-1}(U)$の分布関数



$p(F^{-1}(U) < x)$

$= P(U < F(x)) = \begin{cases} 0 \\ F(x) \\ 1 \end{cases}$ 　(∵ $F(x)$が単調増加)

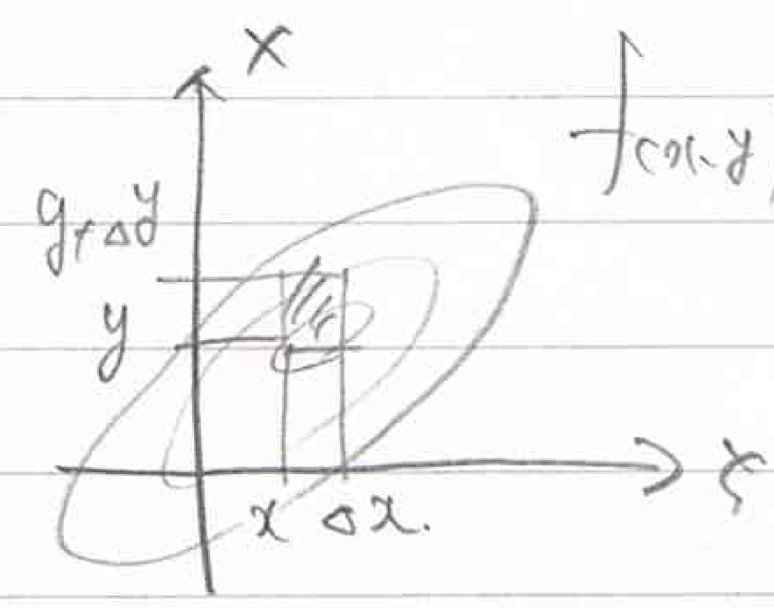
$P(U<z) = \begin{cases} 0 \\ z \\ 1 \end{cases}$ 　 $0<z<1$

$X = F^{-1}(U)$ 　 Uは一様乱数は分布関数$F(x)$を持つ.

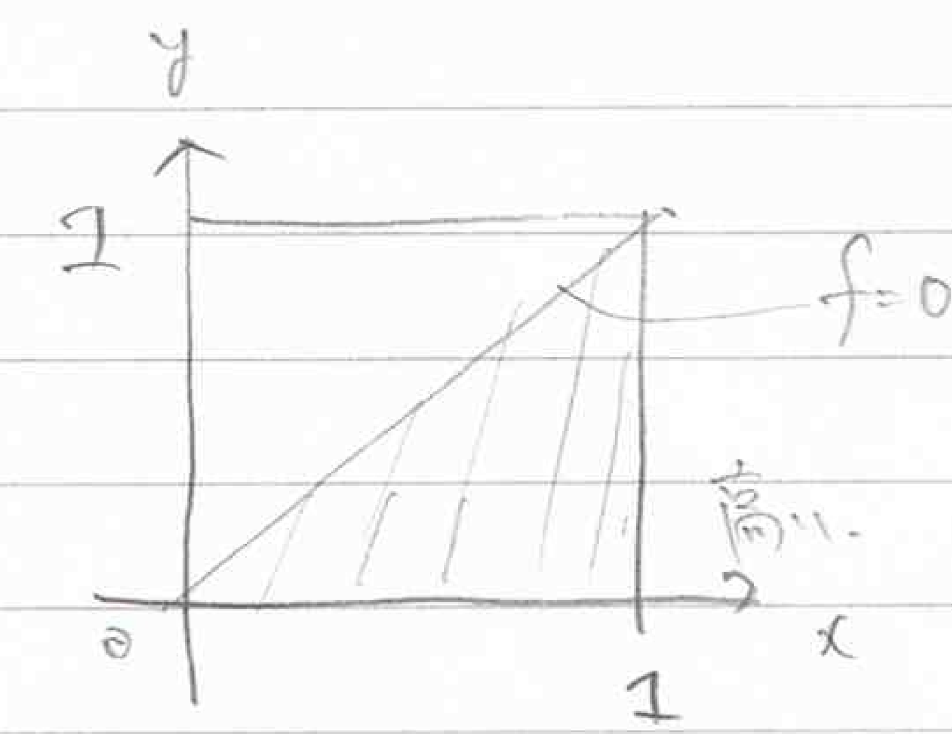
モンテカルロ実験(乱数を発生) 架空の実験

$P(x<X<x+\Delta \quad y<Y<y+\Delta) = f(x,y)\,\Delta x \Delta y$

同時密度



$f(x,y) = 6(x-y)$ 　 $0 \le y < x \le 1$



Xはオプション:保険

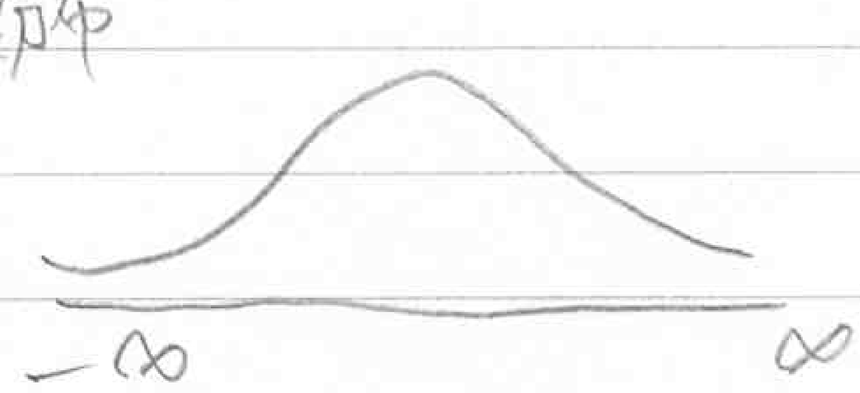
Xがある値を超えたらY円, それ以外0

⇒ Yの(周辺密度)

$f(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$

$= f(x_1,y)\Delta x \cdots + f(x_N,y)\Delta x$

正規分布



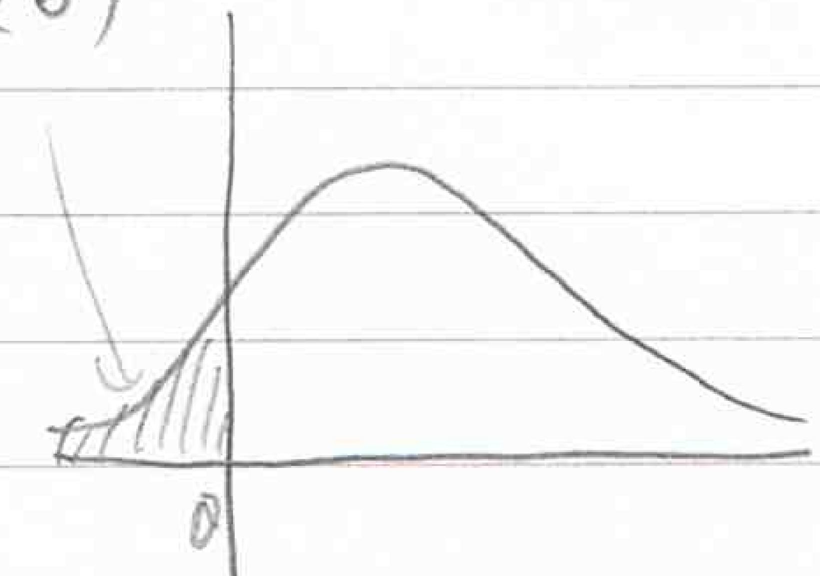
左右対称, ベル型

Xが正規
Yも正規 ⇒ X+Yも正規

ポートフォリオ

付表1. p280 上側確率

$P(X+Y<0)$



$E(X)=0$ 　 $V(X)=1$ 　 X,Yは独立.

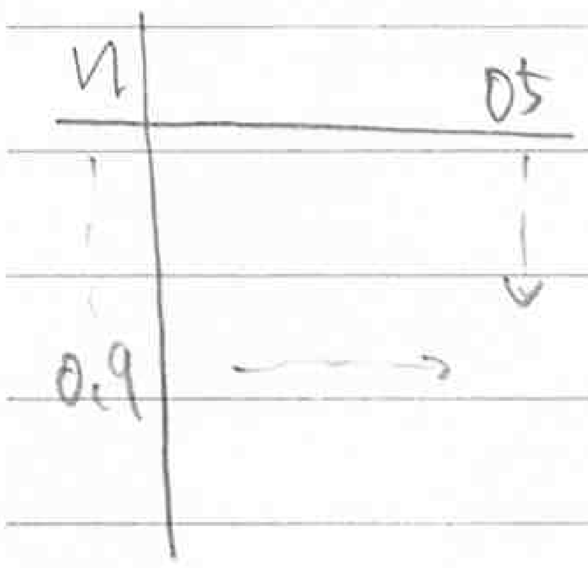
$E(Y)=0$ 　 $V(Y)=1$

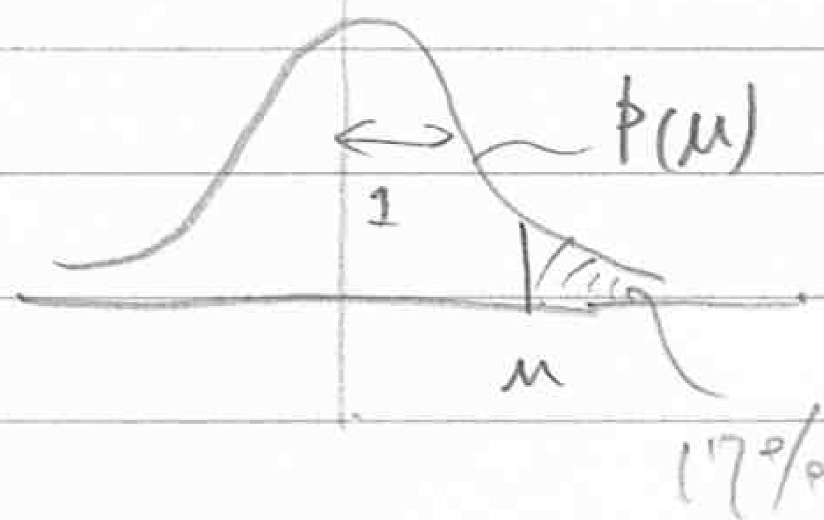
No. Date . .

$V(X+Y)=2$, $E(X+Y)=0$ X,Yの分布が正規(仮定)
⇒ X+Yも正規

期待値 0、分散1の正規分布の上側確率

$P_1(X>0.95)$
$= .1761$





$P(X<-1.95) = P(X>1.95)$
(∵左右対称)

$P(X+Y<0) = P\left(\frac{X+Y}{\sqrt{2}}<0\right) = 0.5$

$E(X+Y)=0.5$ $V(X+Y)=2$

$P(X+Y<0) = P(X+Y-0.5<-0.5)$ ⟶ 期待値を0にした。

$= P\left(\frac{X+Y-0.5}{\sqrt{2}}<-\frac{0.5}{\sqrt{2}}\right)$ 標準偏差を1にした。

0.35

$=0.363$ 36% で破算。

Value at Risk.

P122の例題

覚えていく $P(|X|>1.96)=0.05$
$P(|X|>2) = \frac{2}{1000}$

精度を上げる

N=600 視聴率調査. → Nを増やす

→ N=600であきらめる。精度を明らかにする。

大数の法則

$X_1, \cdots, X_N$ 独立 $E(X_i)=\mu$ $V(X_i)=\sigma^2$ (一定)

$\bar{X} = \frac{1}{N}(X_1+\cdots+X_N)$

$\to \mu$ (確率収束) $(N\to\infty)$

$V(\bar{X}) = V\left[\frac{1}{N}(X_1+X_2+\cdots+X_N)\right]$

$= \frac{1}{N^2}V(X_1+X_2+\cdots+X_N)$

$= \frac{1}{N^2}\sum_{i=1}^{N}V(X_i)$ ($\because$ X_iは互いに独立)

$= \frac{\sigma^2}{N}$

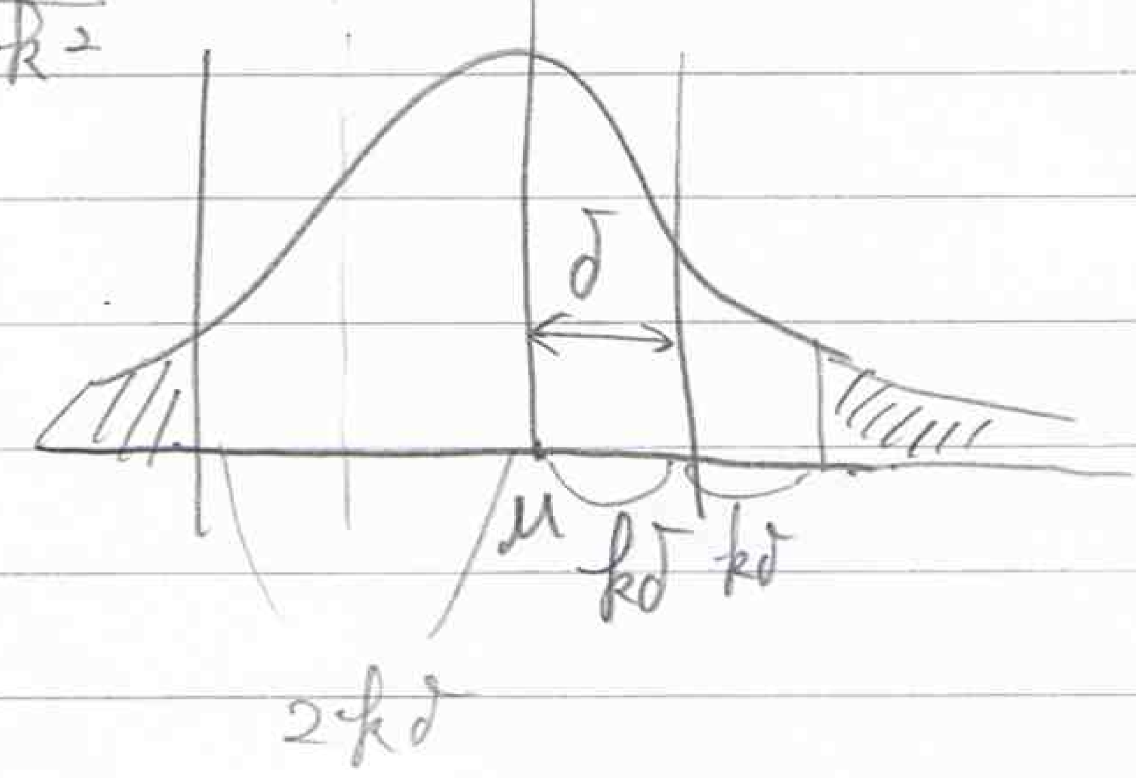
$\longrightarrow 0$ $(N\to 0)$

チェビシェフの不等式

普通：分布 → 分散. チェビシェフ：分散 → 分布.

$P(|X-\mu|>k\sigma) \le \frac{1}{k^2}$

σ：標準偏差



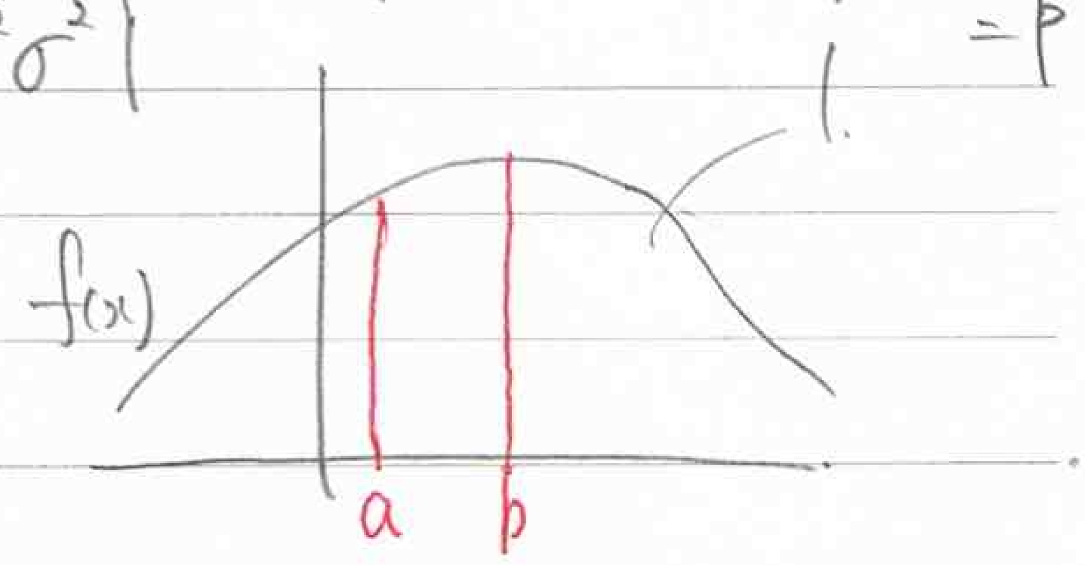
$\sigma^2 = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)\,dx$

$I\ \{x \mid |x-\mu|^2 \ge k^2\sigma^2\}$

$\ge \int_I (x-\mu)^2 f(x)\,dx$

$\ge k^2\sigma^2 \int_I f(x)\,dx = k^2\sigma^2 Pr(|x-\mu|>k\sigma)$

$Pr(a<x<b) = \int_a^b f(x)\,dx = P$



No. Date

$\sigma^2 \geq \sigma^2 k^2 \cdot P(|x-\mu| > k\sigma)$ $\frac{1}{k^2} \geq P(|x-\mu| > k\sigma)$ □

大数の法則

$V(\bar{X}) = \frac{1}{N}\sigma^2$ $N \to \infty$

$P(|\bar{X}-\mu| > \underbrace{\varepsilon}_{固定}) \to 0$ ： 確率収束

proof $\varepsilon = k\sigma_N$ とおると. $\left(k = \frac{\varepsilon}{\sigma_N}\right)$

$P(|\bar{X}-\mu| > \varepsilon) \leq \frac{1}{k^2} = \frac{\sigma_N^2}{\varepsilon^2}$ $(\sigma_N \to 0 \quad (N \to \infty))$

中心極限定理

$\bar{X} = \frac{1}{N}\{X_1 + \cdots + X_N\}$

$N \to \infty$ $X_1, \cdots, X_N$ は独立. $E(X_1) = \mu$ $V(X) = \sigma^2$

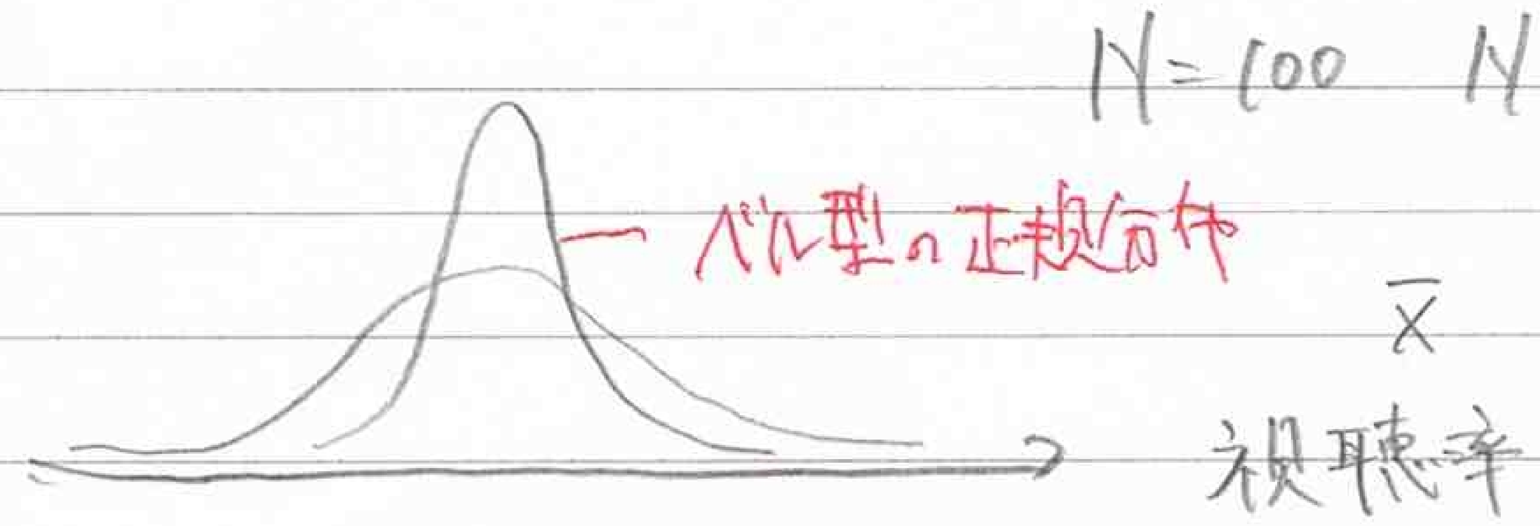
$V(\bar{x}) = \frac{N}{\sigma^2}$ $\frac{\bar{X}-\mu}{\sqrt{\sigma^2/N}} = \sqrt{N}\left(\frac{\bar{X}-\mu}{\sigma}\right)$

標準化 (期→0 分→1)

$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/N}}$ は $N \to \infty$ は 標準正規分布に近づく。(分布収束)

期待値0, 分散1

$\bar{X}$の分布 $N=100$ $N=200$

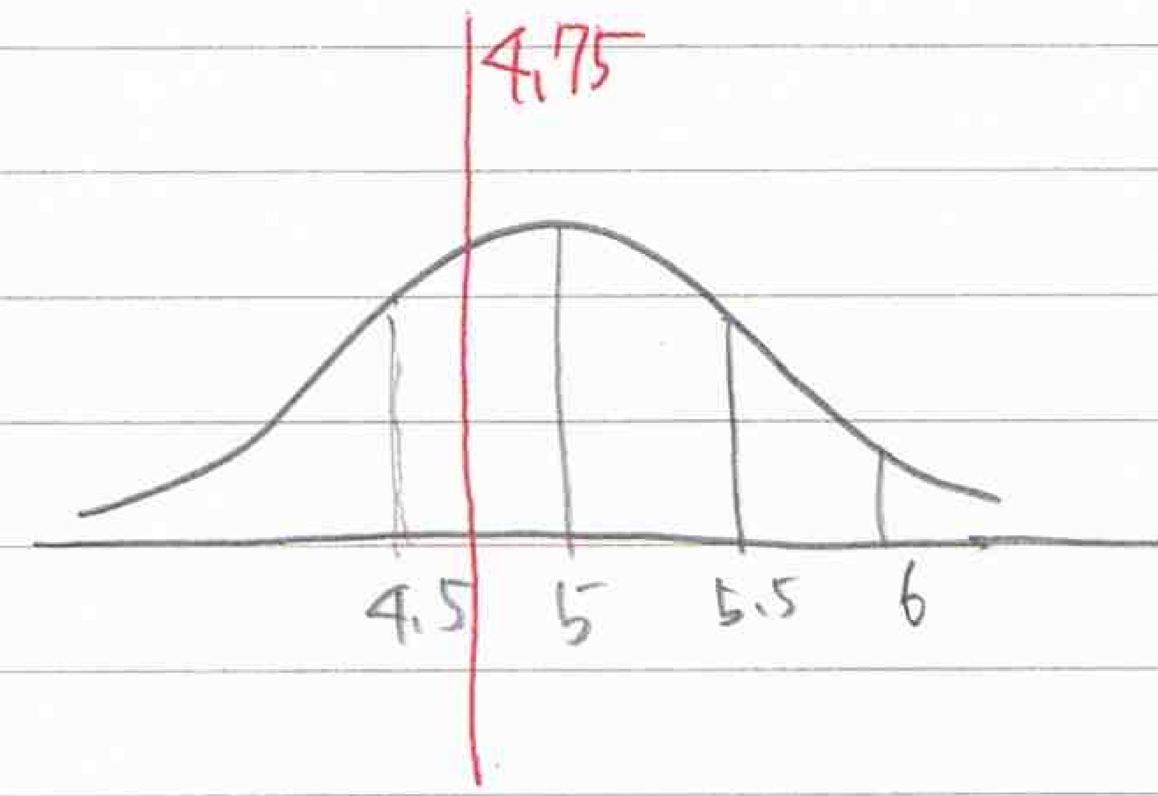


Date . . No.

サイコロの目 X　　$E(X) = \frac{7}{2}$　　$V(X) = \frac{35}{12}$　　$D(X) = 1.7$

サイコロを2つ投げ　$P(\bar{X} \geq 5)$

$$(X_1, X_2) = \begin{cases} (6,6) \\ (6,5) \\ (5,6) \\ (5,5) \\ (4,6) \\ (6,4) \end{cases} \qquad \frac{6}{36} = \frac{1}{6} = 0.166$$



$$P(\bar{X} > 4.75)$$

$$= E(\bar{X}) = \frac{7}{2}$$

$$V(X) = \frac{35}{12}$$

$$V(\bar{X}) = \frac{35}{2 \times 12} = \frac{35}{24} \fallingdotseq 1.46$$

$$P(\bar{X} > 4.75)$$

$$= P(\bar{X} - 3.5 > 4.75 - 3.5)$$　期待値を0にする

$$= P\left(\frac{\bar{X} - 3.5}{\sqrt{1.46}} > \frac{4.75 - 3.5}{\sqrt{1.46}}\right)$$

$$= P\left(Z > \frac{1.25}{1.208}\right)$$

$$= P(Z > 1.03)$$

$$= 0.1515$$

$N=2$　$\bar{X} \geq 5$　厳密：0.166　正規分布 0.1515

$N=4$　$\bar{X} \geq 5$　厳密：0.054　〃　0.0537

左右対称だと正規分布に良く近似できる。

No. Date . .

視聴率 $p=0.2$ $X=0\sim5$ $N=5$

$P(X=0)=(0.8)^5=0.32$
$P(X=1)=0.4$
$P(X=2)=0.20$
$P(X=3)=0.05$
$P(X=4)=0.0064$
$P(X=5)=0.0003$

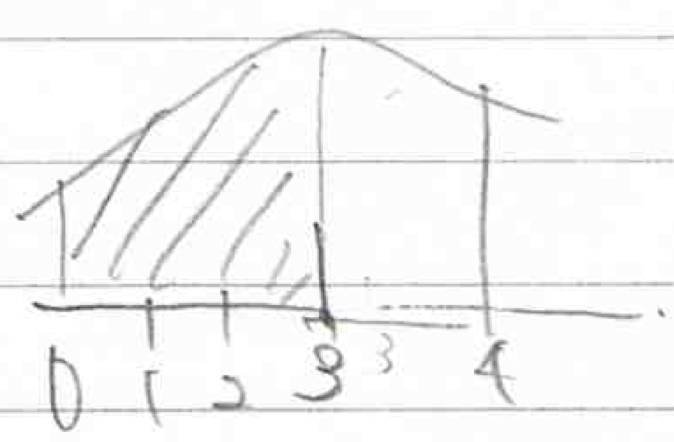
P3. $p=0.5$ $N=10$ X:表の枚数 $(\overline{X})$

$E(X)=5$ 1枚 $p(1-p)$ 分散.
$V(X)=10\times0.5(1-0.5)$
$=2.5$

$D(X)=1.581$

$P(X\leqq3)=P_r(X\leqq3.5)$

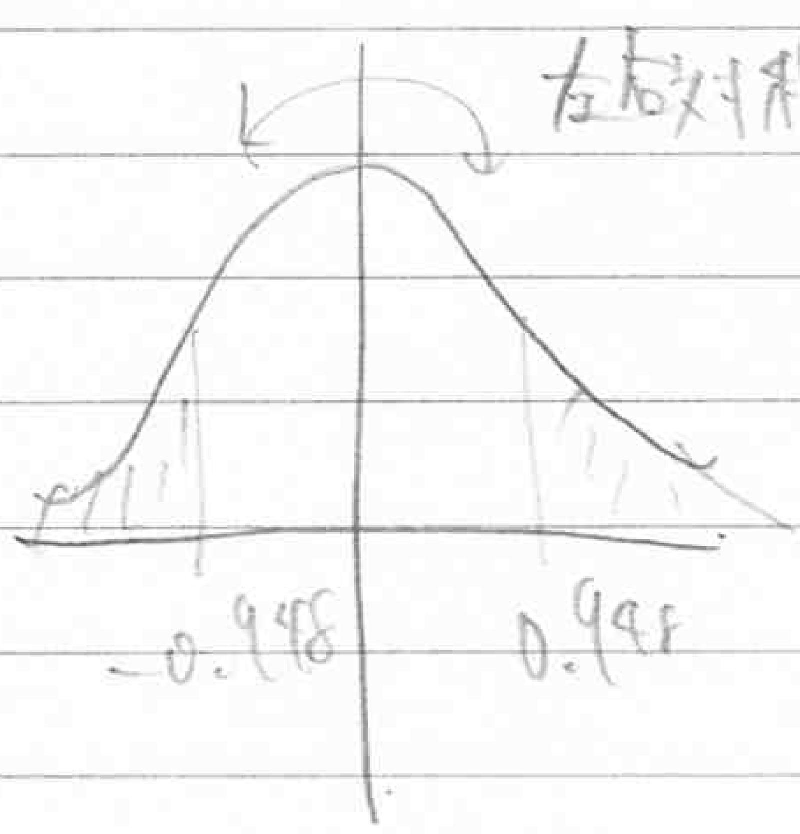
標準化.



$=P\left(\frac{X-5}{1.581}\leqq\frac{3.5-5}{1.581}\right)$

$=-0.948\fallingdotseq-0.95$

Z $N(0,1)$ Normal



$=P(Z\leqq-0.948)$

$=P(Z\geqq0.948)$

$\fallingdotseq P(Z\geqq0.95)$

$=0.17106$

$\fallingdotseq17\%$

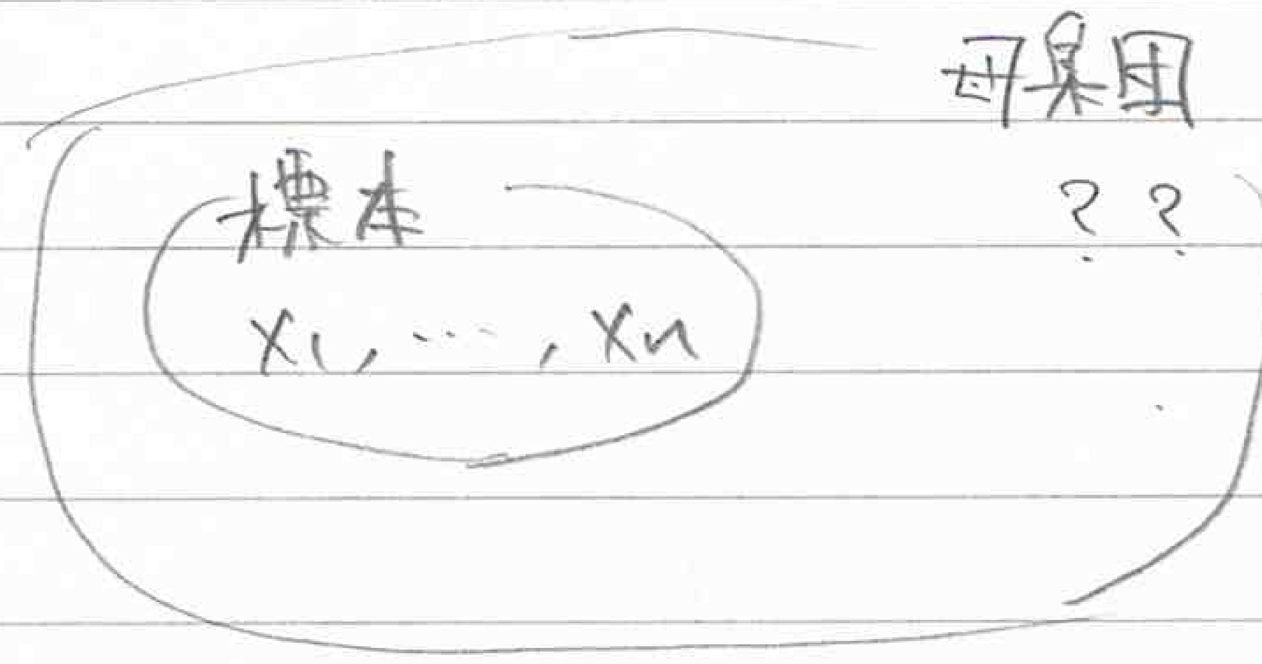
Date . . No.

p4 (p185 統計学入門)

トマトジュースのビタミンC $N=17$ $s^2=15.875$ $s=3.984$ $\bar{X}=20$

「$x_1, \cdots, x_{17}$」が正規分布に従っている.

1. 推定 (Estimation).
2. 検定 (test)



→ 日本で生産されている トマトジュース (今年)

→ ・海外との比較
・昨年
・検定

仮定

$X_i \sim N(\mu, \sigma^2)$ 期待値 μ, 分散σ^2

$X_1 \cdots X_n$ は独立 $\Rightarrow E(\bar{X}) = \mu$

$= \frac{1}{n^2}\{V(x_1) + \cdots + V(x_n)\}$

$= \frac{\sigma^2}{n}$

$\bar{X}$も正規

$$\bar{X} - \mu \sim N\left(0, \frac{\sigma^2}{n}\right)$$

標準化

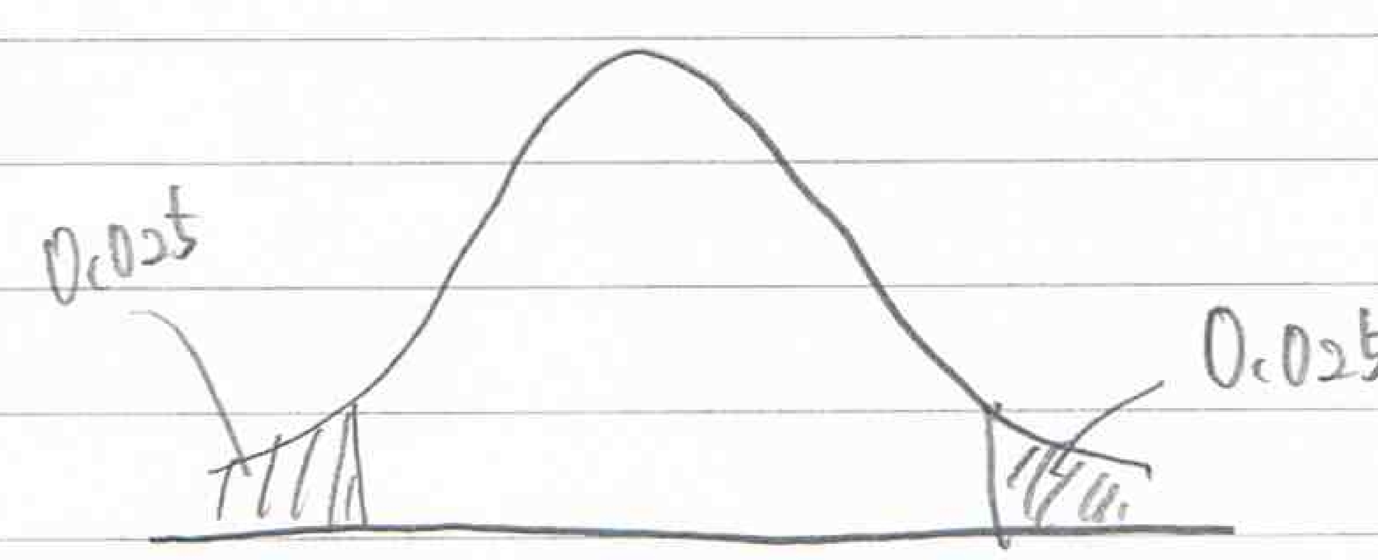
$$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \sim N(0.1)$$

標準正規

$$P\left(-1.96 < \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} < 1.96\right) = 95\%$$

$\Rightarrow$ p.280

1%〜5% : rare
(めったにない)



$$P\left(-1.96 < \frac{\mu-\bar{X}}{\sqrt{\sigma^2/n}} < 1.96\right)$$

No.　　　Date　　.　　.

$$= P\left(\bar{X} - 1.96 \times \sqrt{\sigma^2/n} < \underbrace{\mu}_{\text{真の期待値}} < \bar{X} + 1.96 \times \sqrt{\sigma^2/n}\right)$$

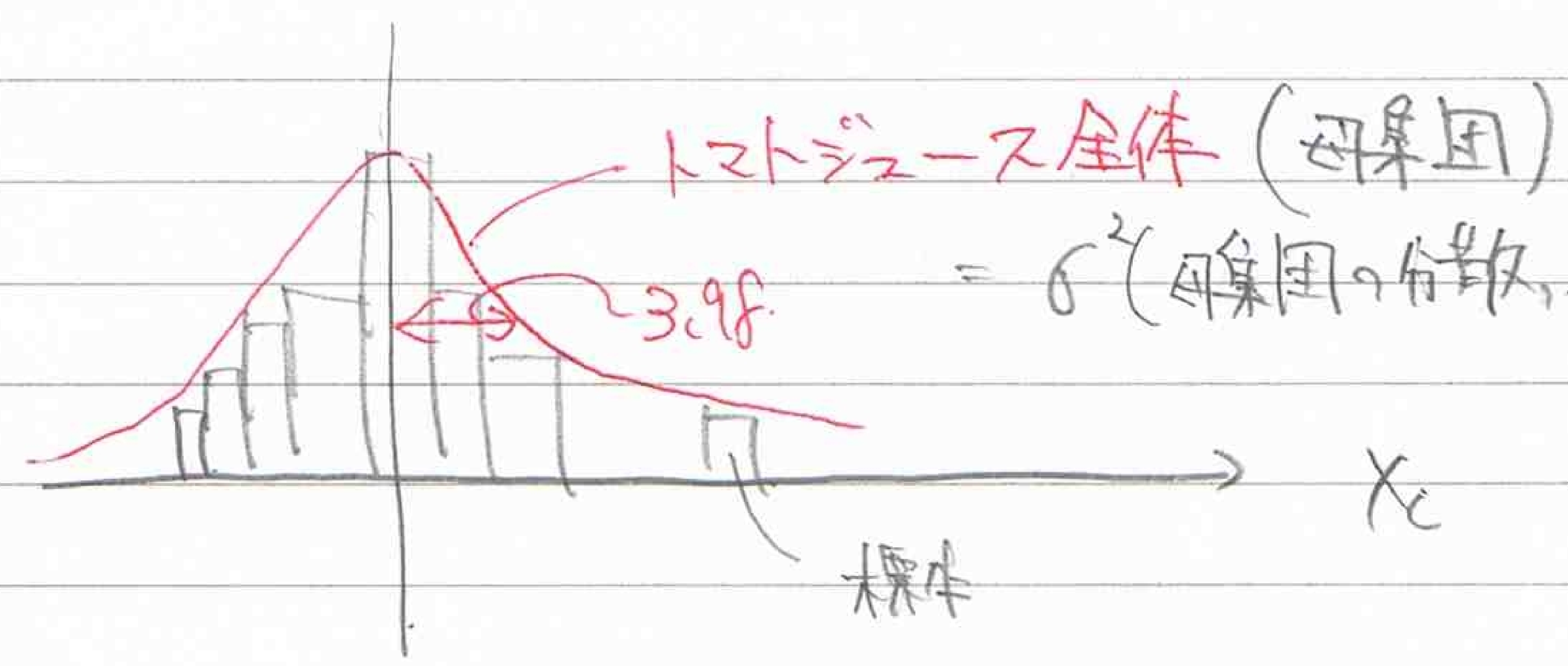
大数の法則

$$\bar{X} \to \mu \quad (N \to \infty) \qquad E(X) = \mu \qquad E\left((X-\mu)^2\right) = \sigma^2$$

$$\frac{1}{N}\left[(x_1-\mu)^2 + \cdots (x_N-\mu)^2\right] \to \sigma^2$$

$$\frac{1}{N-1}\left\{(x_1-\bar{x})^2 + \cdots + (x_N-\bar{x})\right\} \to \sigma^2$$

Nが十分大きければ $N \fallingdotseq N-1$



$= \sigma^2$（母集団の分散、母分散）

$$\underset{20}{\bar{X}} \pm 1.96 \times \frac{3.984}{\sqrt{17}} \fallingdotseq 18 \sim 22$$ にμが収まる確率は95%

→ 区間推定　　昨年のトマトジュースの平均25mg
⇒ 昨年よりもビタミンCは少ない

$$95\% = P\left(\bar{X} - 1.96 \times \sqrt{\sigma^2/n} < \mu < \bar{X} + 1.96 \times \sqrt{\sigma^2/n}\right)$$

$$99\% = P\left(\bar{X} - 2.58 \times \sqrt{\sigma^2/n} < \mu < \bar{X} + 2.58 \times \sqrt{\sigma^2/n}\right)$$

5%：危険率有意水準

95%：信頼係数

Date . . No.

$$\mu = 20 \pm 2.58 \times \frac{3.9}{\sqrt{17}} = (17.5,\ 22.5) \quad : 99\%$$

仮説検定

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

従来の薬 $\mu = 25$　　$\bar{X} = 20$

仮説 $\mu = 25$ と考えと…　$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \sim N(0,1) = t$ とする

$P(|t| > 1.96) = 5\%$　　$P[|t| > 2.58] = 1\%$

$$t = \frac{20-25}{\sqrt{15/17}} \fallingdotseq -5$$

$\Rightarrow$ 前提が否定
$\Leftrightarrow$ $\mu = 25$ が棄却（否定）

1.96　有意水準　5%
2.58　〃　1%

$P = 226$　　$N = 100$　　$X_i \sim N(\mu, \sigma^2)$

$\bar{X} = 2.346$　　$S = 0.047$

$$P\left(\left| \frac{\bar{X}-\mu}{\sqrt{\sigma^2/N}} \right| < 1.65 \right) = 90\%$$

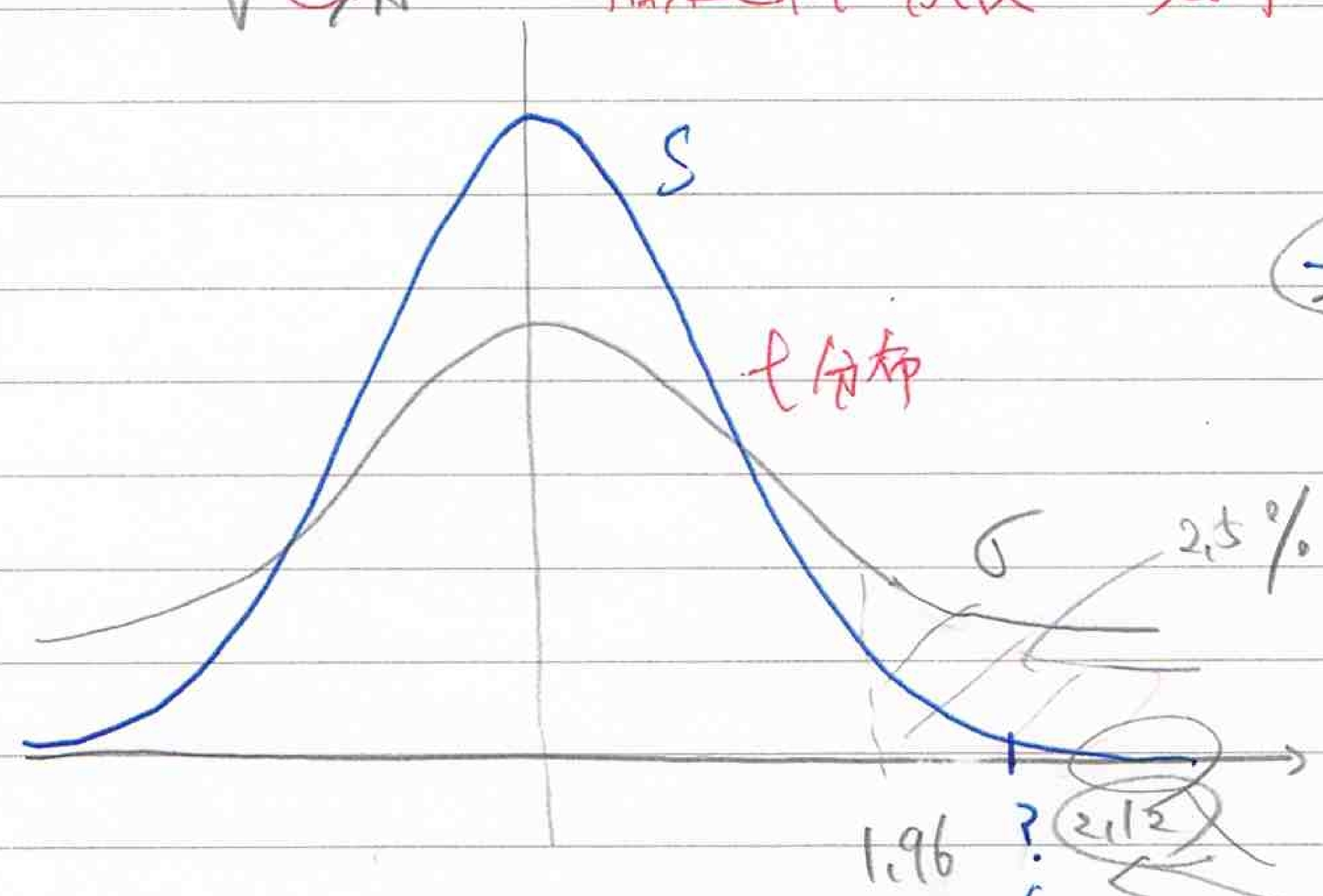
$$\bar{X} - 1.65 \times \sqrt{\sigma^2/n} < \mu < \bar{X} + 1.65 \times \sqrt{\sigma^2/n}$$

$$\sqrt{S^2/n} = \frac{0.047}{10} = 0.0047 \quad \therefore \mu \in (2.34,\ 2.35)$$

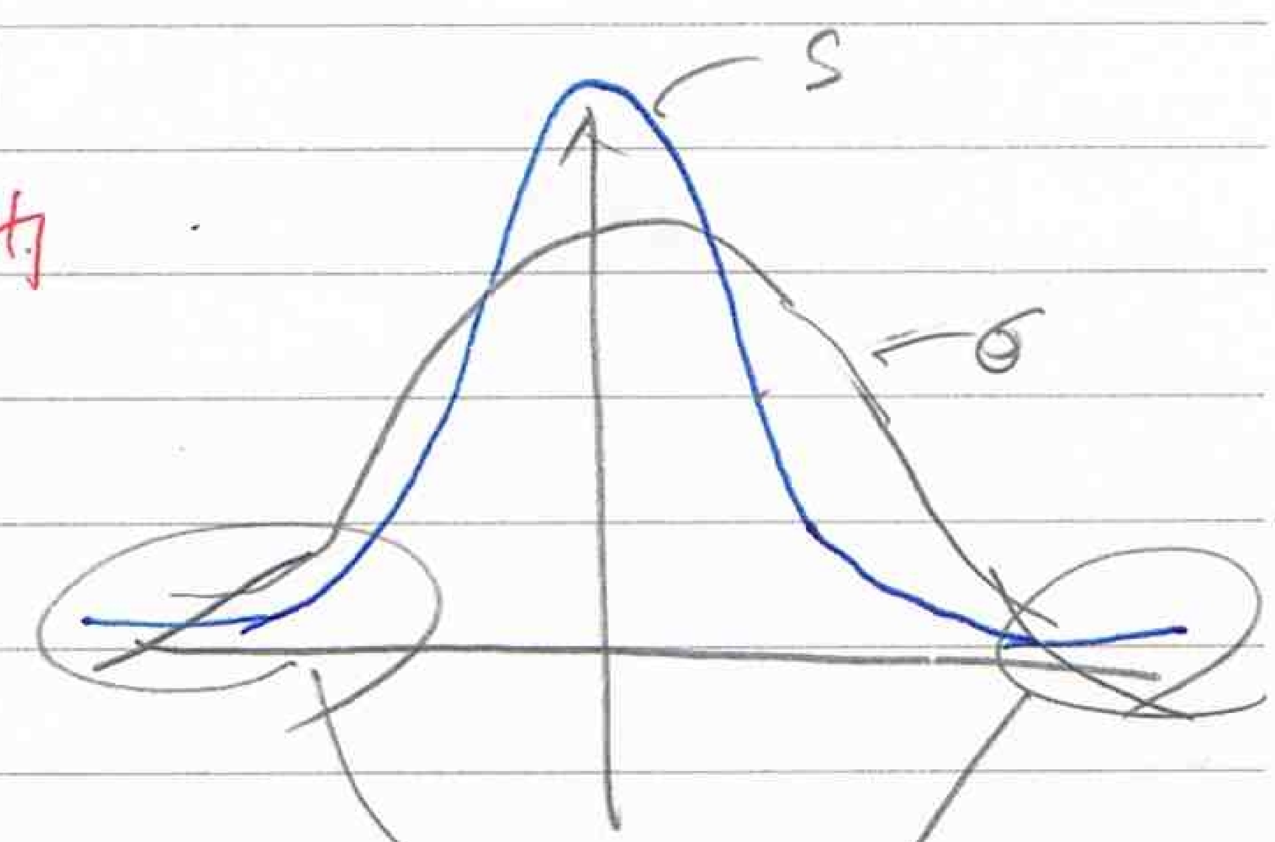
1) Nが大 or σ^2が既知
$S^2 \fallingdotseq \sigma^2$

2) Nが小さい $S^2 \overset{?}{\fallingdotseq} \sigma^2$

$t = \dfrac{\bar{X}-\mu}{\sqrt{S^2/N}} \sim N(0,1)$ 推定された分散 ⇒ 変動

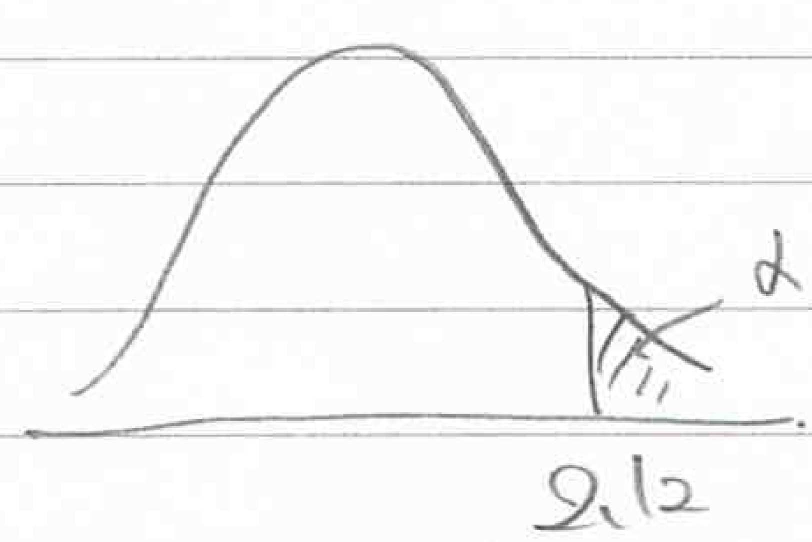


1.96より大きい値で2.5%を超える.



裾が広がる.
Fat Teil

P.281 付表2.



自由度	α 0.025
N−1 16	2.12 > 1.96
	1.96

$s^2 = \dfrac{\Sigma(x_i-\bar{x})^2}{N-1}$

※ ν(自由度) ≧ 20
正規分布と ~~数布~~ t分布 (σ^2の代わりに s^2) を使ったもの
はほとんど一致

区間推定 N=17

$\bar{X} \pm \cancel{1.96} \times \sqrt{S^2/N}$
↓
2.12 (∵ Sを使っているから, t分布)

⇒ t分布を使った区間推定

Date . . No.

4.

$$S^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$

☆何故N−1で割ったかの説明.

$$(分子) = \sum_{i=1}^{N}(x_i - \mu + \mu - \bar{x})^2 = ☆$$

$$E(x) = \mu \qquad V(x) = \sigma^2$$

$$☆ = \sum_{i=1}^{N}(x_i - \mu)^2 + \sum_{i=1}^{N}(\bar{x} - \mu)^2 - 2\sum_{i=1}^{N}(\bar{x} - \mu)(x_i - \mu)$$

$$= \sum_{i=1}^{N}(x_i - \mu)^2 + N(\bar{x} - \mu)^2 - 2(\bar{x} - \mu)\underbrace{\sum_{i=1}^{N}(x_i - \mu)}_{= N(\bar{x} - \mu)}$$

$$\left(\because\ \bar{x} = \frac{\sum x_i}{N}\right)$$

$$= \sum_{i=1}^{N}(x_i - \mu)^2 - N(\bar{x} - \mu)^2.$$

分子の期待値 $= N \times \sigma^2$

$$V(x) = E(x - \mu)^2 = \sigma^2 \qquad E(\bar{x}) = \mu \qquad V(\bar{x}) = \frac{\sigma^2}{N}$$

$$E\left((\bar{x} - \mu)^2\right) = V(\bar{x}) = \frac{\sigma^2}{N}$$

$$E(S^2) = \frac{1}{N-1}\left\{N \times \sigma^2 - N \cdot \frac{\sigma^2}{N}\right)$$

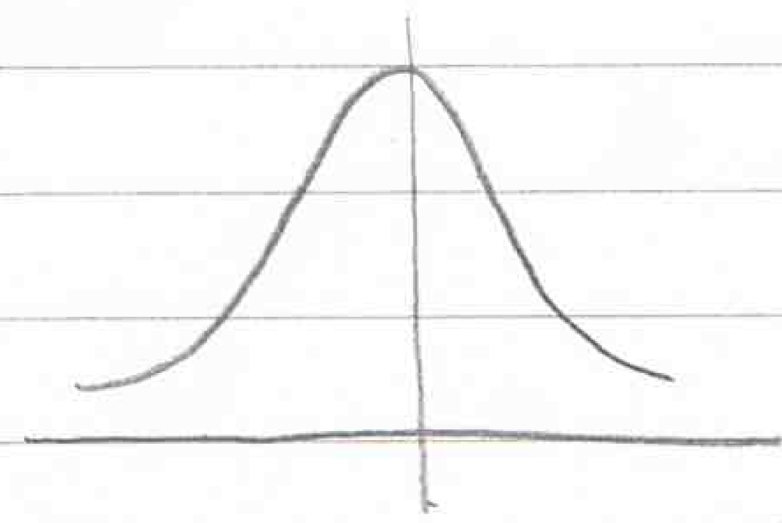
$$= \frac{1}{N-1}(N-1)\sigma^2$$

$$= \sigma^2$$

No.　Date　.　.

t検定　前提 $X_1, \cdots, X_N$

正規分布.



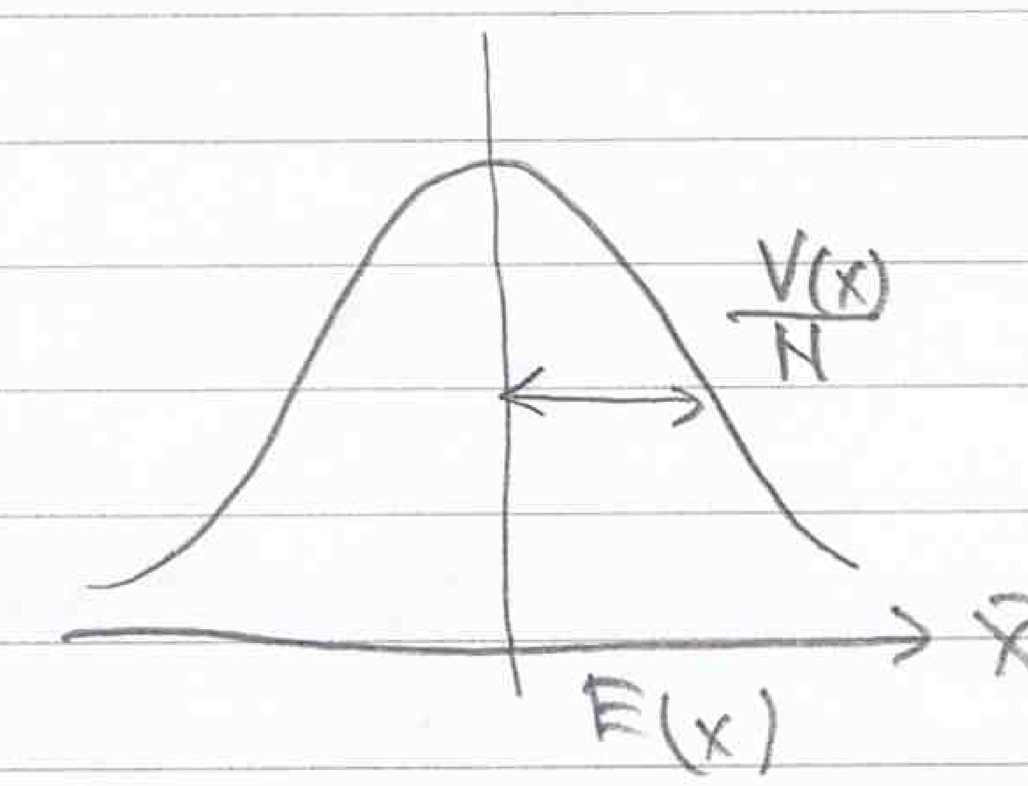
$$X_i = \begin{cases} 1 & \text{視聴} \\ 0 & \text{No} \end{cases}$$

$$\bar{X} = \frac{X_1 + \cdots + X_N}{N}$$

$N \to \infty \Rightarrow \bar{X}$は正規分布に近似.

中心極限定理

$E(X) = p$　　$V(X) = \dfrac{p(1-p)}{N}$



$$X_i = \begin{cases} 1 & p : \text{確率} \\ 0 & 1-p : \text{否確率} \end{cases}$$

$$E(X_i) = p \times 1 + (1-p) \times 0 \quad : \text{加重平均}$$
$$= p$$

$$V(X_i) = E\left((x_i - p)^2\right)$$
$$= p \times (1-p)^2 + (1-p) \times (0-p)^2$$
$$= p(1-p)$$

信頼係数 95%

$$\bar{X} \sim N\left(p, \frac{p(1-p)}{N}\right)$$

$$\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{N}}} \sim N(0, 1)$$

仮説検定と区間推定

・ t検定, t分布 ⇒ 100%

$N=24$ $\bar{X} = -0.154$ X_i : 薬効, 訓練, cost

分散 0.809 ほとんどの場合には $N-1$ で割ったもの (不偏分散)

→ 標準偏差 (SD) $\sqrt{0.809} = 0.9$

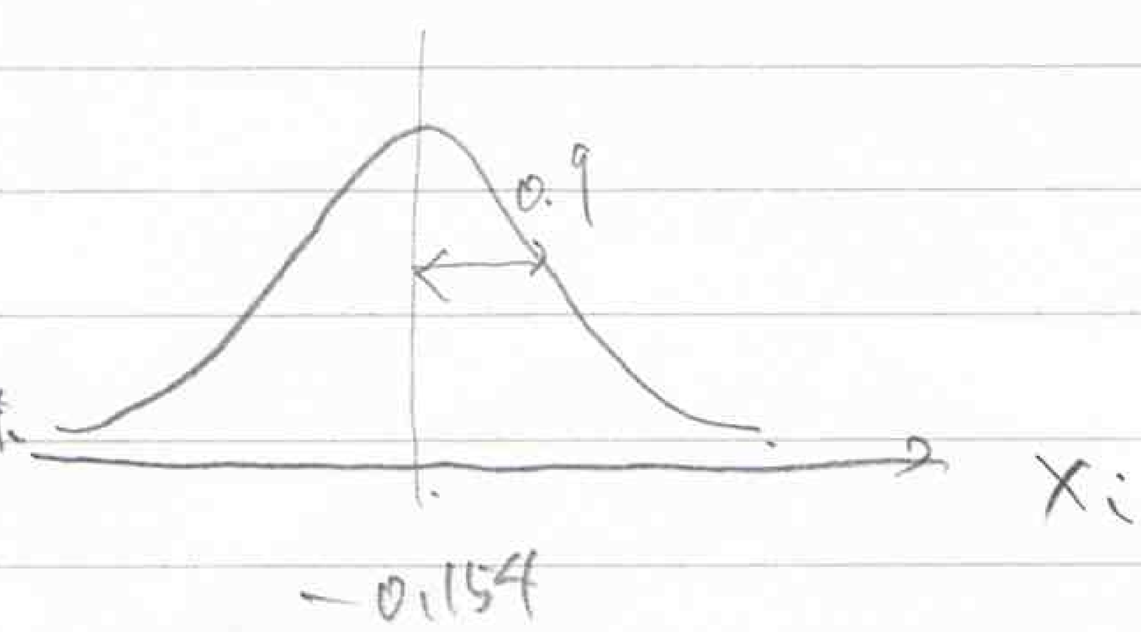
⇒ X_i は独立に正規分布

$$V(\bar{X}) = \frac{1}{N}\sigma^2 = \frac{1}{N}V(X_i)$$

$$\sqrt{V(\bar{X})} = \frac{1}{\sqrt{N}}\sigma$$: 標準誤差 → $\bar{X}$ の標準偏差のことを

⇐ 推定量のSDのことをいう。

$$\frac{0.9}{\sqrt{24}} = \frac{0.1837}{0.154}$$



"〜" : は従うという意味

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{N}} \sim N(0,1)$$

$$\frac{\bar{X}-\mu}{s/\sqrt{N}} \sim t(24-1)$$ ← $N-1$

※ $$s^2 = \frac{(x_1-\bar{x})^2 + \cdots + (x_N-\bar{x})^2}{N-1}$$

区間推定 $$p\left(\left|\frac{\bar{X}-\mu}{s/\sqrt{N}}\right| < 2.069\right) = 95\%$$

$$\Leftrightarrow p\left(\bar{X} - 2.069 \times \frac{s}{\sqrt{N}} < \mu < \bar{X} + 2.069 \times \frac{s}{\sqrt{N}}\right) = 95\%$$

$$p\left(\bar{X} - 2.8 \times \frac{s}{\sqrt{N}} < \mu < \bar{X} + 2.8 \times \frac{s}{\sqrt{N}}\right) = 99\%$$

→ μ の範囲が絞れる。

No. Date

$$t = \frac{\bar{x} - \mu}{s/\sqrt{N}} \sim t(N-1) \qquad s^2 : 不偏分散$$

t値

仮説検定

t: 仮説の μ (期待値) から $\bar{x}$ が標準誤差で
いくつ離れているか.

$t = 3, 4, 5$ となると μ の仮説が間違っていたことになる

$\mu = 0$ で検定することで効果の有無が分かる

$|t| < 2.086$

$\Leftrightarrow$ $\mu = 0$ の可能性有り

$\Leftrightarrow$ $\mu = 0$ は区間推定の中. $\mu = 0$ は棄却されない.

比較.

class A X 教授法 20人
B Y 教授法 20人

※ $\bar{x}$, $\bar{y}$ も誤差を持つ.

$\bar{x} - \bar{y}$ の分析 (=標本) 問題 $N = 2$ sample size

x_1 y_1 x_1
⋮ ⋮ x_2.
x_N y_M

$x_1 \cdots x_N$ | すべて独立.
$y_1 \cdots y_N$ | $x_i \sim N(\mu_1, \sigma_1^2)$ $y_i \sim N(\mu_2, \sigma_2^2)$

$$V(\bar{x} - \bar{y}) = V(\bar{x}) + V(\bar{y}) + 2\,\mathrm{cov}(\bar{x}, \bar{y})$$

$$= \frac{1}{N}\sigma_1^2 + \frac{1}{M}\sigma_2^2 \qquad x, y \text{ が独立} \rightarrow \text{無相関}$$

case 1. σ_1^2 σ_2^2: 既知

$$t = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{N}\sigma_1^2 + \frac{1}{M}\sigma_2^2}} \sim N(0, 1)$$

Date . . No.

case 2. $\sigma_1^2 = \sigma_2^2 \ (= \sigma^2)$

$$t = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S\sqrt{1/N + 1/M}} \sim t(M+N-2)$$

(σをSに置き換えた)

$$S^2 = \frac{\sum(X_i - \bar{X})^2 + \sum(Y_i - \bar{Y})^2}{M+N-2}.$$

不偏分散

覚えろ!!

$$E\left(\sum(X_i - \bar{X})^2\right) = \sigma^2(N-1)$$
$$E\left(\sum(Y_i - \bar{Y})^2\right) = \sigma^2(M-1)$$

仮定 $\mu_1 = \mu_2$.

$$t = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{N} + \frac{1}{M}}}.$$

$$\begin{cases} 0.906 = \frac{1}{N-1} \sum (X_i - \bar{X})^2 \\ 1.988 = \frac{1}{M-1} \sum (Y_i - \bar{Y})^2 \end{cases}$$

$$S^2\text{の分子} = 0.906 \times \overbrace{(N-1)}^{19} + 1.988 \times \overbrace{(M-1)}^{19}$$

$$= 54.986 \qquad (\because N = M = 20)$$

$$S^2 = \frac{54.986}{38} = \underline{1.477}.$$ P2のプールされた分散.

$$\therefore S = \sqrt{1.447} \fallingdotseq 1.202.$$

$$t = \frac{0.022 - (-0.118)}{1.202 \times \sqrt{\frac{1}{20} + \frac{1}{20}}} \fallingdotseq 0.368$$
$$\fallingdotseq 0.37$$

No. Date . .

結論

仮説：「$\mu_1=\mu_2$」は有意水準 5% で棄却されてない

$\Leftrightarrow$ $\mu_1=\mu_2$ の可能性がある。（が否定できない

$$P\left(-1.96 < \frac{\bar{X}-p}{\sqrt{\frac{p(1-p)}{N}}} < 1.96\right) = 95\%$$

$$\Leftrightarrow P\left(\bar{X} - 1.96 \times \sqrt{\frac{p(1-p)}{N}} < p < \bar{X} + 1.96 \times \sqrt{\frac{p(1-p)}{N}}\right) = 95\%$$

pに$\hat{p}$（推定値）を代入（両側の範囲は pが定まらないと決められない）

中心極限定理　N大 $\Rightarrow$ $\hat{p} \fallingdotseq p$にある程度近い

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{N}\right)$$

$$\frac{0.769-p}{\sqrt{\frac{p(1-p)}{52}}} \leq 1.96$$

を解けば良いのでは？

$$t = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{N}}} \sim N(0,1)$$

もしも確率pで$X=1$

$$\hat{p} = \bar{X}$$

$$t = \frac{\bar{X}-0.5}{\sqrt{\frac{0.25}{52}}} \sim N(0,1) \qquad \bar{X} = \frac{40}{52} = 0.769$$

$t = \frac{0.269}{0.069} = 3.87 \geq |\pm 2|$ より有意ではない

$p = 0.5 \Rightarrow \bar{X} \geq 0.769$

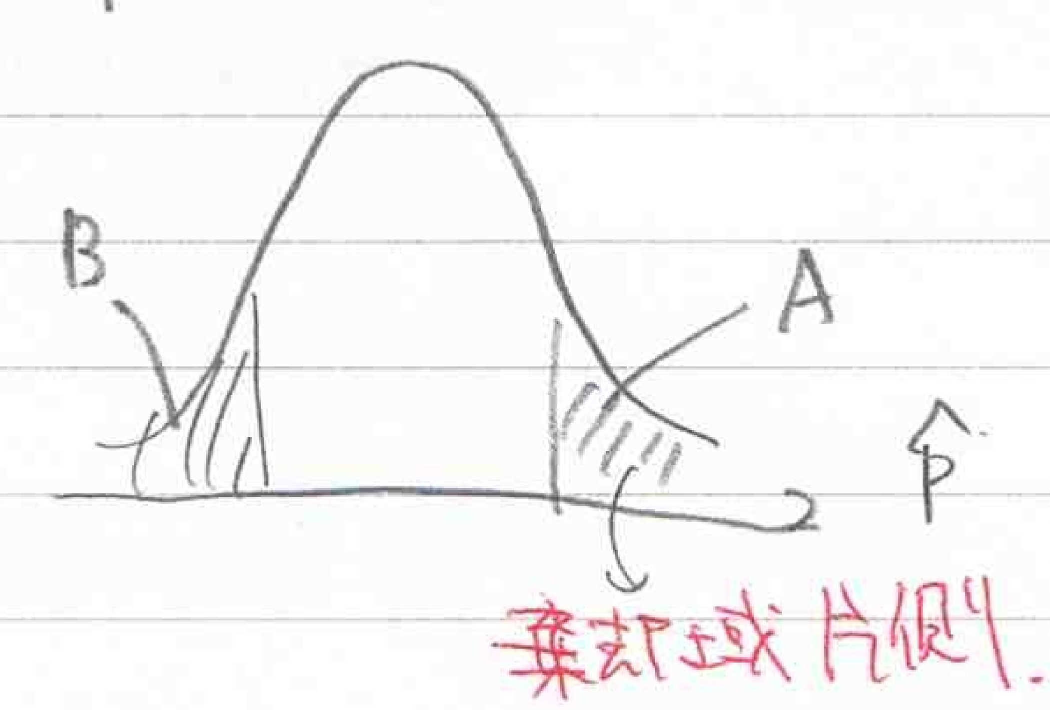
$\Rightarrow$ $p=0.5$は有意水準 ~~95%~~ 5%で棄却される。

片側検定・両側検定

「支持率は過半数か？」　$p > 0.5$か $p < 0.5$

A : X党　$p > 0.5$

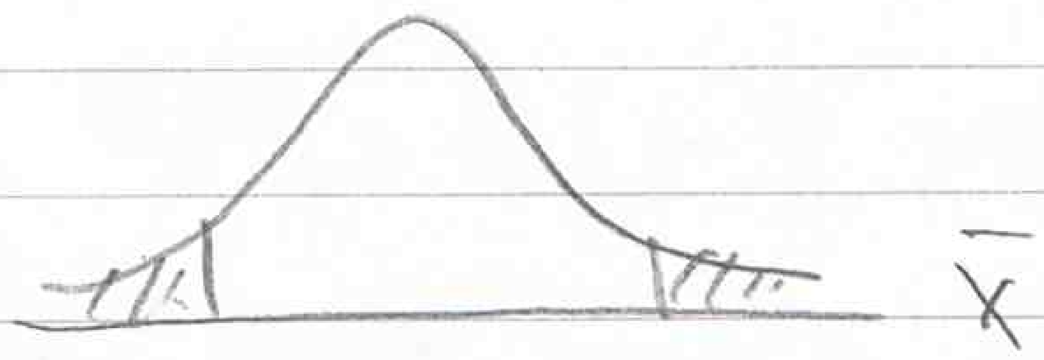
B : X党　$p < 0.5$



No. Date

ねじの大きさ = 5mm

$\bar{X} = 4.9$ $S = 0.1$ $N = 20$



大きすぎても小さすぎても×

⇒ 両側検定を使う

→ どちらを使うか分からない時は両側検定を用いる.

$X_1, \dots, X_N$: X社の株の収益率

$X_i \sim N(\mu, \sigma^2)$ 期待値

↳ 分散 = Risk = 不確実性

具体例 →

$\sigma^2 = \sigma_0^2$ = 他社のRisk. σ^2 : 新規参入のRiskがσ_0^2に等しいかどうかを調べる

品質管理 平均的 100 km

50 km しか出ない… ではX.

120 km まで出る…

→ 分散の小ささが重要となる

正規分布 t分布 正規よりFat-tail

χ^2 カイ自乗分布 : F分布

$z_1 \dots z_n \sim N(0,1)$ z_i $(i = 1, 2 \dots n)$ は独立.

$\sum_{i=1}^{n} z_i^2 \sim \chi^2(n)$ n = 自由度.

$E(z^2) = 1$

$E[(z-a)^2] = \sigma^2 = 1$

$x_i \sim N(\mu, \sigma^2)$

$\sum_{i=1}^{n} (x_i - \mu)^2 \sim \chi^2(n) \times \sigma^2$

$\sum_{i=1}^{n} (x_i - \bar{x})^2 \sim \chi^2(n-1) \times \sigma^2$

← 証明は今の段階ではできない.