§1. 確率

統計学演習 (2004 年度) 渡辺

□□□ テーマ □□□

1. 飛行機事故

飛行機事故は続いて起こるとよく言われる. 下の図は, 1953 年から 1977 年までの間に起こった 145 件の世界の航空機事故の日数間隔分布である. このデータから, 航空機事故は互いに関連があると言えるだろうか. (右図は武者利光著「ゆらぎの世界」(講談社) より転載)

(確率分布の問題)

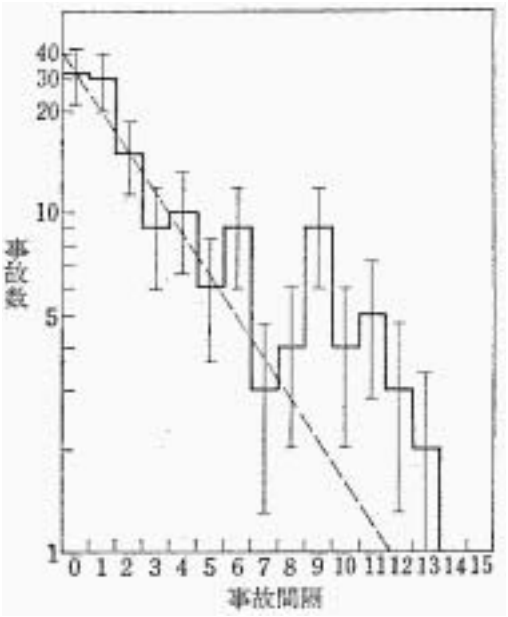


図 1. 飛行機事故の間隔
(1953 年から 1977 年まで 145 件の事故)

2. バスの待ち時間

自宅前の停留所に無作為に行くと, バスが来るまで平均して 10 分待たなければならないことが分かった. 営業所に聞いてみると, 交通渋滞でダイヤはかなり乱れているが, 平均して 10 分に 1 本の割合で運転していると言う. 本当だろうか.

(平均の問題)

3. 1 列並びの利点

ある銀行には 2 台の ATM がある. 2 台の前にそれぞれ列を作って待つ場合も, 1 列並びで待つ場合も, 待ち時間の平均は同じだろう. それでは, 1 列並びの利点とは何なのだろうか.

(揺らぎの問題)

■□■　演習問題　■□■

1

ある放射性物質が 1 回崩壊してから，次の崩壊が起きるまでの平均時間は 12 分 30 秒である．最後の崩壊を観測して 1 分後から 10 分間測定器から離れていたために，その間に何回崩壊したか分からなくなった．測定器のそばに戻ってから次の崩壊が起きるまでの時間は，どれほどであると期待されるか．

2

下図は，東名高速の横浜インター付近で，通過する自動車の車間時間分布を測定した結果である．比較のために，完全にランダムな自動車の流れをコンピューターで作った結果も示してある．この図から，自動車の走行状態は，互いにどのように影響し合っていると考えられるか．(右図は武者利光著「ゆらぎの世界」(講談社) より転載)

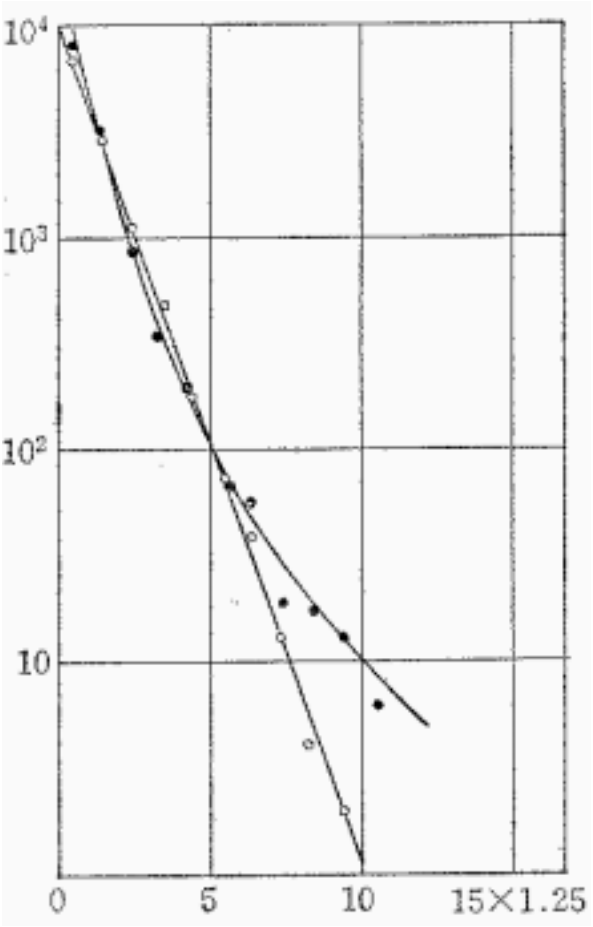


図 2.　車間時間 （秒）

●は実測，○はコンピューターシミュレーション

3

ある工場で生産される製品の中に稀に不良品が含まれることがあり，不良品が作られるという現象は，完全にランダムに起きることが分かっている．時刻 s から時刻 t までの間に不良品が作られない確率を $p(s,t)$ とすると，

$$p(t_1,t_2)p(t_2,t_3)=p(t_1,t_3)$$

が成り立つことを示せ．また，この式が成り立つ根拠は，結局どこにあると言えるか．

4

平均すると 1 時間に 10 本の割合で，完全にランダムに来るバスがある．

(1) 1 時間に $n(=0,1,2,\cdots)$ 本のバスが来る確率を求めよ．

(2) ランダムに停留所に行くとき，6 分以内にバスが来る確率を求めよ．

5

3 台の ATM の場合に，1 列並びの待ち時間を論ぜよ．

□□□ 確率 ・ノート □□□

1．飛行機事故

(1) 確率モデル：事故は完全にランダムに起こると仮定する．すなわち

(a) same chance whenever
時間 $[t, t+\Delta t]$ に起きる確率は t によらず λt である（Δt：小，$\lambda > 0$）

(b) independent events
互いに独立に起きる

(c) no more than on at a time
時間 $[t, t+\Delta t]$ に 2 度以上起きる確率は無視できる．

(2) 時間 $[0, t]$ に事故が起きない確率 $p(\lambda, t)$ を求める．

$[0, t]$ を N 等分し，$\Delta t = \dfrac{t}{N}$ とおく．

$[n\Delta t, (n+1)\Delta t]$ に事故が起きる確率 $= \lambda\Delta t$

$[n\Delta t, (n+1)\Delta t]$ に事故が起きない確率 $= (1 - \lambda\Delta t)$

よって， $p(\lambda, t) \fallingdotseq (1-\lambda\Delta t)^N = (1 - \lambda\dfrac{t}{N})^N \xrightarrow[N\to\infty]{} e^{-\lambda t}$

(3) $t = 0$ に事故が起きたとして，次の自己が $[t, t+\Delta t]$ に起きる条件付確率

$p(\lambda, t)\lambda\Delta t = e^{-\lambda t}\lambda\Delta t$

事故の時間間隔の分布密度は

$\lambda e^{-\lambda t}$ （指数分布）

2．バスの待ち時間

(1) 確率モデル：

バスは完全にランダムに来ると仮定する．時刻 0 から停留所で待つとして，

$[0, t]$ にバスが来ない確率 $p(\lambda, t) = e^{-\lambda t}$

$[t, t+\Delta t]$ にバスが来る確率 $= \lambda\Delta t$

待ち時間 T の分布密度は $\lambda e^{-\lambda T}$

・ 全確率 $= \displaystyle\int_0^\infty \lambda e^{-\lambda T} dT = 1$

(2) 待ち時間 T の平均値 μ を求める

$\mu = \displaystyle\int_0^\infty T\lambda e^{-\lambda T} dT = \frac{1}{\lambda}$

(3) バスが来る頻度

$[0, t]$ に n 回来る確率を p_n とする．

$p_0 = p(\lambda, t) = e^{-\lambda t}$

$p_1 = \int_0^t p(\lambda, t_1)p(\lambda, t-t_1)\lambda dt_1 = \int_0^t \lambda e^{-\lambda t} dt_1 = \lambda t e^{-\lambda t}$

$p_2 = \int_0^t \int_0^{t_2} p(\lambda, t_1)p(\lambda, t_1-t_2)p(\lambda, t-t_2)\lambda dt_1 \lambda dt_2 = \lambda^2 e^{-\lambda t} \int_0^t \int_0^{t_2} dt_1 dt_2 = \frac{1}{2}\lambda^2 t^2 e^{-\lambda t}$

$p_n = \frac{1}{n!}(\lambda t)^n e^{-\lambda t}$ （Poisson 分布）

・ 全確率 $= 1$

n の平均 $= \sum_{n=0}^{\infty} n p_n = \sum_{n=1}^{\infty} \frac{1}{(n-1)!}(\lambda t)^n e^{-\lambda t} = \lambda t$

よって，単位時間あたり λ 回来る．

(4) ２項分布を用いて p_n を求める．

$$
\begin{aligned}
p_n &= {}_N\mathrm{C}_n (\lambda \Delta t)^n (1-\lambda \Delta t)^{N-n} \\
&= \tfrac{N(N-1)\cdots(N-n+1)}{n!} \left(\tfrac{\lambda t}{N}\right)^n \left(1-\tfrac{\lambda t}{N}\right)^{N-n} \\
&= \tfrac{1}{n!}\, 1\, (1-\tfrac{1}{N})\, (1-\tfrac{2}{N}) \cdots (1-\tfrac{n-1}{N})\, (\lambda t)^n \frac{(1-\frac{\lambda t}{N})^N}{(1-\frac{\lambda t}{N})^n} \\
&\to \tfrac{1}{n!}\, (\lambda t)^n e^{-\lambda t}
\end{aligned}
$$

3．待ち行列

(1) 確率モデル

一人一人の処理がいつ終わるかは完全にランダムであるとする．

窓口が１つ

n 人が並んでいて，自分は $n+1$ 人目

一人目の処理が時刻 0 に始まった

(2) 自分の処理が $[T, T+\Delta T]$ に始まる確率 $f_n(T)\Delta T$ を求める．

時刻 T までに $n-1$ 人の処理が終わる確率 $p_{n-1}(T) = \frac{(\lambda T)^{n-1}}{(n-1)!} e^{-\lambda T}$

よって

$$f_n(T)\Delta T = p_{n-1}(T)\lambda \Delta T = \frac{\lambda^n T^{n-1}}{(n-1)!} e^{-\lambda T} \Delta T$$

待ち時間の分布密度 $= f_n(T) = \frac{\lambda^n T^{n-1}}{(n-1)!} e^{-\lambda T}$

(3) 待ち時間 T の平均

$$\mu = \int_0^{\infty} T f_n(T) dT = \frac{\lambda^n}{(n-1)!} \int_0^{\infty} T^n e^{-\lambda T} dT = \frac{n}{\lambda}$$

また，T^2 の平均は

$$E(T^2) = \int_0^{\infty} T^2 f_n(T) dT = \frac{\lambda^n}{(n-1)!} \int_0^{\infty} T^{n+1} e^{-\lambda T} dT = \frac{n(n+1)}{\lambda^2}$$

T の分散

$$\sigma^2 = E(T^2) - \mu^2 = \frac{n}{\lambda^2}$$

(4) 窓口２つ

n 人ずつ 2 列並び，自分は $n+1$ 人目

$\mu = \dfrac{n}{\lambda}$，$\sigma^2 = \dfrac{n}{\lambda - 2}$

4．1 列並びの利点

(1) model

一人一人の処理がいつ終わるかは完全にランダム

窓口が２つ

$2n+1$ 人が１列並び，自分は $2n+2$ 番目

１人目，２人目の処理が時刻 0 に始まった

(2) 自分の処理が $[T, T+\Delta T]$ に始まる確率 $g_{2n+1}(T)\Delta T$ を求める．

・窓口１で $[T, T+\Delta T]$ に m 人目の処理が終わり，自分の処理が始まった．

・窓口２で時刻 T までに $2n-m$ 人の処理が終わり，$2n+1-m$ 人目の処理をしているとすると，その確率は

$$\begin{aligned} f_m(T)\Delta T \cdot p_{2n-m}(T) &= \frac{\lambda^m T^{m-1}}{(m-1)!} e^{-\lambda T}\Delta T \cdot \frac{(\lambda T)^{2n-m}}{(2n-m)!} e^{-\lambda T} \\ &= \frac{\lambda^{2n} T^{2n-1}}{(m-1)!(2n-m)!} e^{-2\lambda T}\Delta T \\ &= \frac{{}_{2n-1}\mathrm{C}_{m-1}}{(2n-1)!}\lambda^{2n} T^{2n-1} e^{-2\lambda T}\Delta T \end{aligned}$$

$m = 1, 2, \cdots, 2n$ について加え，窓口１，２を交換した場合も考え，

$$\begin{aligned} g_{2n+1}(T)\Delta T &= 2\sum_{m=1}^{2n} f_m(T)\Delta T \cdot p_{2n-m}(T) \\ &= 2\sum_{m=1}^{2n} \frac{{}_{2n-1}\mathrm{C}_{m-1}}{(2n-1)!}\lambda^{2n} T^{2n-1} e^{-2\lambda T}\Delta T \\ &= 2\frac{2^{2n-1}}{(2n-1)!}\lambda^{2n} T^{2n-1} e^{-2\lambda T}\Delta T \\ &= 2\lambda\frac{(2\lambda T)^{2n-1}}{(2n-1)!} e^{-2\lambda T}\Delta T \end{aligned}$$

(3) 待ち時間 T の平均は

$$\mu_T = \int_0^\infty T g_{2n+1}(T) dT = \frac{(2\lambda)^{2n}}{(2n-1)!}\int_0^\infty T^{2n} e^{-2\lambda T} dT = \frac{(2\lambda)^{2n}}{(2n-1)!}\frac{(2n)!}{(2\lambda)^{2n+1}} = \frac{2n}{2\lambda} = \frac{n}{\lambda}$$

$$\begin{aligned} E(T^2) &= \int_0^\infty T^2 g_{2n+1}(T) dT = \frac{(2\lambda)^{2n}}{(2n-1)!}\int_0^\infty T^{2n+1} e^{-2\lambda T} dT \\ &= \frac{(2\lambda)^{2n}}{(2n-1)!}\frac{(2n+1)!}{(2\lambda)^{2n+2}} = \frac{2n(2n+1)}{(2\lambda)^2} = \frac{n(n+\frac{1}{2})}{\lambda^2} \end{aligned}$$

$$\sigma_T^2 = E(T^2) - E(T)^2 = \frac{n(n+\frac{1}{2})}{\lambda^2} - \frac{n^2}{\lambda^2} = \frac{n}{2\lambda^2}$$

§2. 推定

統計学演習 (2004 年度) 渡辺

□□□ テーマ □□□

1. 金星のネコ

A 氏と B 氏がつぎのような会話をした.

A B さん. 金星に生物の存在する可能性について，あなたの意見をお聞かせ下さい.
B よろしい. 私は金星についての知識が全くありませんので，生物の存在する可能性と存在しない可能性とは同様に確からしいと考えます. よって，どちらも確率 $\frac{1}{2}$ であると答えましょう.
A なるほど. ではこの問題を別の角度から考えてみます. 金星にネコが存在しない確率はいくつでしょうか.
B くどいようですが，私はその方面に無知ですので，確率は $\frac{1}{2}$ です.
A では，ドラエモンがいない確率は？
B やはり $\frac{1}{2}$ です.
A 分かりました. でもそうすると，ネコもドラエモンもいない確率は，$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ となりますね.
B (自分の立場に気づき始めて) さあ，どうでしょう，……
A そうすると，ネコかドラエモンのいる確率は $\frac{3}{4}$ ですから，金星に生物の存在する確率は少なくとも $\frac{3}{4}$ となり，最初にあなたがおっしゃった $\frac{1}{2}$ と矛盾しますね.

とうとう B 氏は何も言えなくなった. A 氏と B 氏のどちらが正しいのだろうか.

(確率とは何かという問題)

2. 視聴率

関東地区でランダムに 2000 世帯を選んで調べたところ，テレビ局 A の番組を 35 % の世帯で見ていた. このことから，この番組の真の視聴率について，どのようなことが言えるだろうか.

(出現率の推定の問題)

3. 出生性比

1950 年から 1996 年までの人口動態統計によると，1 年間に日本で生まれた女児 100 に対する男児の数は，次の表の通りである. このデータから，出生性比はどれほどであると推定できるか.

年	出生性比
1950	106.1
1955	105.8
1960	105.6
1965	105.3
1970	107.1
1975	106.2
1980	106.0
1985	105.6

年	出生性比
1990	105.4
1991	105.7
1992	106.0
1993	105.6
1994	105.6
1995	105.2
1996	105.6

(正規母集団の母平均を推定する問題)

□■□ ノート □■□

1. 確率とは何か

- 「金星に生物が存在する確率は $\frac{1}{2}$ である」という言明は何を意味しているのか?
  - ある事柄について何も知らないとき,「確率は $\frac{1}{2}$ である」と主張できるか?
    「確率は $\frac{1}{2}$ である」とはどういう意味か?
    「確率」とは何か?
  - 金星は一つしか存在しない.
    唯一の金星について確率という概念は意味をもたない.
    同様に,「明日」は 1 回しかないので,「明日晴れる確率」は意味をもたない.
  - 試行のたびに結果が random に変わる現象において，確率という概念が意味をもつ.
- 「金星に生物が存在する確率は $\frac{1}{2}$ である」という言明は何を意味すべきか?
  - 全宇宙の金星型天体を考える
  - 全宇宙の金星型天体が n 個，そのうち生物が存在する天体が k 個存在するとする．このとき，$p = \frac{k}{n}$ という量は意味がある.
  - この意味において,「金星型天体に生物が存在する確率は $\frac{1}{2}$ である」と主張するには,根拠となる data が必要である.

2. 確率の測定 (推定)

- コインが表を出す確率 p を知るには？
  - 繰り返しコインを投げる
  - 例えば，2000 回コインを投げて 700 回表が出たとして，$p = \frac{700}{2000}$ と推測する.

- 確率は物理量であり，測定によって定められる
  - 測定には誤差がつきまとう
  - 誤差はいくらでも小さくすることができる (はず)
- 測定 (推定) した確率はどれほどの誤差を含むか？
  - 確率をいくらでも精密に測定し得るか?

3．確率の測定誤差

- コインが表を出した回数は，2 項分布に従うと考える．
- **2 項分布 (Binomial distribution)**
  確率 p で表を出すコインを n 回投げる．各試行が **独立** であるとすると，n 回中 k 回表が出る確率は
  $$P_{n,k} = {}_n\mathrm{C}_k p^k (1-p)^{n-k} \ , \quad k = 0, 1, 2, \cdots, n$$
  である．この確率分布を 2 項分布といい，$B_{n,p}$ と書く．
  このとき，k の平均 μ と分散 σ^2 は，
  $$\mu = E(k) = np$$
  $$\sigma^2 = E((k-np)^2) = np(1-p)$$
  で与えられる．
- $\dfrac{k}{n}$ の平均 $= E(\dfrac{k}{n}) = p$,
  $\dfrac{k}{n}$ の分散 $= E((\dfrac{k}{n} - p)^2) = \dfrac{p(1-p)}{n}$
  確率の測定値 $\dfrac{k}{n}$ と真の確率 p の差は，大体 $\sqrt{\dfrac{p(1-p)}{n}}$ 程度だろう (n が大きいとき，非常に小さい)．
- 試行回数 n を大きくすれば，確率の測定値は真の確率にいくらでも近付く (と期待できる)．
  従って,「$p = \dfrac{k}{n}$ であろう」とする推定 (**点推定**) は妥当であると考えられる．

4．信頼区間 (confidence interval)

- 2000 回コインを投げて 700 回表が出たとき，
  コインが表を出す確率 p は，区間 $\left[\dfrac{700}{2000} - a, \dfrac{700}{2000} + a\right]$ に属する
  という形の推定 (**区間推定**) をすることを考える．このような区間を **信頼区間** という．
- 区間推定は，どの程度信頼できるか？
  区間の幅 $2a$ が小さいと，推定の信頼性は低下する．
  $2a$ を大きくとれば，推定の信頼性は増すが，推定としての意味を失う．

5．2 項分布と正規分布

- 信頼区間の信頼性を評価することを考える．計算を容易にするために，2 項分布を正規分布で近似する．
- **正規分布 (normal distribution)**
  確率変数 x が $a < x < a + \Delta a$ (Δa : 微小量) なる値をとる確率が
  $$\mathrm{Prob}(a < x < a + \Delta a) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(a-\mu)^2}{2\sigma^2})\Delta a$$
  で与えられるとき，即ち，確率密度関数が
  $$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$
  で与えられるとき，x は，平均 μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ に従うという．
  x が $N(\mu, \sigma^2)$ に従うとき，
  $$z = \frac{x-\mu}{\sigma}$$
  は基準正規分布 $N(0,1)$ に従う．
- **2 項分布と正規分布**
  2 項分布 $B_{n,p}$ は，n が大きいとき，平均 np, 分散 $np(1-p)$ の正規分布 $N(np, np(1-p))$ に近づく．

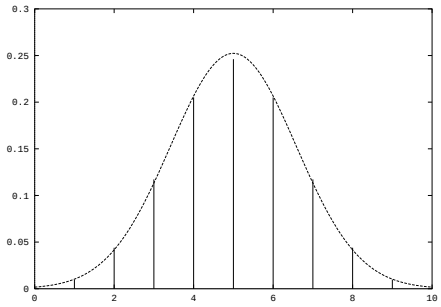


図 1.　$B_{10,1/2}$ と $N(5, \frac{5}{2})$

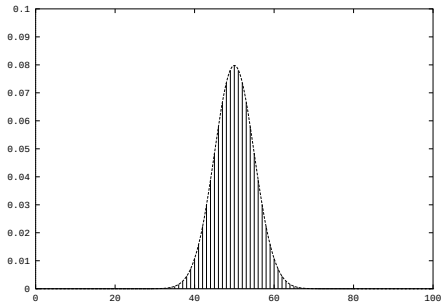


図 2.　$B_{100,1/2}$ と $N(50, 25)$

6．信頼水準 (confidence level)

- コインを n 回投げて k 回表が出たとする．
- k の平均 $= np$,
  k の標準偏差 $= \sqrt{np(1-p)}$
  n が大きいとき，k の確率分布は近似的に正規分布 $N(np, np(1-p))$ となる．
- $z = \dfrac{k-np}{\sqrt{np(1-p)}}$ とおくと，z は基準正規分布 $N(0,1)$ に従うと考えてよい．
  従って，たとえば
  $$P(|z| \leq 1.96) = 0.95$$

- $|z| \leq 1.96 \Leftrightarrow \dfrac{|k-np|}{\sqrt{np(1-p)}} \leq 1.96 \Leftrightarrow \left|\dfrac{k}{n} - p\right| \leq 1.96\sqrt{\dfrac{p(1-p)}{n}}$ $\quad (*)$
  確率 0.95 で $(*)$ が成立する.
- 実際に 2000 回中 700 回表が出たとする．このとき，$(*)$ に $n = 2000$，$k = 700$ を代入した式
  $$\left|\frac{700}{2000} - p\right| \leq 1.96\sqrt{\frac{p(1-p)}{2000}} \qquad (**)$$
  が成立すると期待したい．これを p について解くと，
  $$0.329402 \leq p \leq 0.371174$$
  この区間を，**信頼水準** 95% の **信頼区間** という.
- 「信頼水準 95%」の意味.

  1 回の調査「2000 回中 700 回表が出た」をもとにして，
  「$0.329402 \leq p \leq 0.371174$ だろう」と推定する

  この調査・推定を多数回繰り返すとき，推定が正しいこともあれば誤っていることもある.
  この方法で p を推定するとき，正しい推定をする確率は 95% である.
- $(**)$ において，$p \fallingdotseq \dfrac{700}{2000}$ であるから，$(**)$ の右辺の p を $\dfrac{700}{2000} = 0.35$ で置き換えると，
  $|0.35 - p| \leq 1.96\sqrt{\dfrac{0.35 \times 0.65}{2000}}$
  これを解くと $0.329096 \leq p \leq 0.370904$
  実用上はこれで十分.

7. 母集団と標本

- 関東地区で調査した 2000 世帯のうち 700 世帯でテレビ局 A の番組を見ていたとする.
  関東地区の全世帯のうちどれだけがこの番組を見ているか推定する場合,
  母集団 (population) = 関東地区の全世帯
  標本 (sample) = 調査した 2000 世帯
  統計学の仕事：標本を通して母集団の統計的性質を推定する.
- 母集団の性質を正しく推定するには，質の良い（random に選ばれている）標本が必要である．**random に選ぶ** とは,

  (1) 母集団の各成員が，同じ chance で標本に現れる,
  (2) それらがたまたま標本に入り込むのは互いに独立であるようにする,

  ということ.

- 母集団の統計的性質の例

  出現率 p (あるテレビ番組を見ている世帯の割合)

  母平均 (population mean) μ

  母分散 (population variance) σ^2

- （例）　日本人全体の身長 x_i (i は人の番号：$i = 1, 2, \cdots, N$，N = 日本人の全人口）

  $\mu = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} x_i$

  $\sigma^2 = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(x_i - \mu)^2 = \dfrac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$

- 標本の統計的性質

  標本の大きさ n

  標本 x_1, x_2, $\cdots$, x_n

  標本平均 (sample mean) $\bar{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$

  標本分散 (sample variance) $s^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2 = \dfrac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$

  不偏分散 (unbiased variance) $v^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2 = \dfrac{n}{n-1}s^2$

- **定理**

  $E(\bar{x}) = \mu$ （$\bar{x}$ は μ の不偏推定量 (unbiased estimator) である）

  $E(v^2) = \sigma^2$ （v^2 は σ^2 の不偏推定量 (unbiased estimator) である）

  （証明） $E(x_k) = \bar{x}$ から $E(\bar{x}) = E(\dfrac{1}{n}\displaystyle\sum_{k=1}^{n} x_k) = \bar{x}$

  また，$\sigma^2 = E(x_k^2) - E(x_k)^2 = E(x_k^2) - \mu^2$ から

  $E(s^2) = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} E(x_k^2) - E(\bar{x}^2) = \sigma^2 + \mu^2 - E(\bar{x}^2)$

  さらに独立性から $x_i \neq x_j$ のとき $E(x_i x_j) = E(x_i)E(x_j)$ を用いると

$$
\begin{aligned}
E(\bar{x}^2) &= \frac{1}{n^2}\sum_{i,j} E(x_i x_j) = \frac{1}{n^2}\sum_{i} E(x_i^2) + \frac{1}{n^2}\sum_{i\neq j} E(x_i x_j) \\
&= \frac{1}{n^2} n(\sigma^2 + \mu^2) + \frac{1}{n^2}\sum_{i\neq j} E(x_i)E(x_j) = \frac{1}{n}(\sigma^2 + \mu^2) + \frac{1}{n^2} n(n-1)\mu^2 \\
&= \frac{1}{n}\sigma^2 + \mu^2
\end{aligned}
$$

  よって $E(s^2) = (1 - \dfrac{1}{n})\sigma^2$

  故に $E(v^2) = \sigma^2$　　（証明終わり）

8. 正規母集団の母平均の推定

- 母集団 M から random に 1 つの個体を取り出すとき，その個体の値 x が正規分布 $N(\mu, \sigma^2)$ に従うとする．

  母集団 M から random に選ばれた sample を用いて，母平均を推定したい．

- **t 分布 (t-distribution)**

  x の確率密度関数が

  $$f(x) = c_n \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

  で与えられるとき，x は，**自由度 n の t 分布** (t_n) に従うという．ただし，c_n は，

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

  を満たすように選ばれた定数である．

- **正規母集団の母平均の推定**

  定理

  母平均 μ の正規母集団 M を考える．$x_1, x_2, \cdots, x_n$ を M の sample，$\bar{x}$ と s^2 をそれぞれ標本平均と標本分散とするとき，$t = \dfrac{\bar{x} - \mu}{\frac{1}{\sqrt{n-1}}s}$ は，自由度 $n-1$ の t 分布に従う．

- 毎年の出生性比は，母平均 μ の正規分布に従うとする．

  1950 年から 1996 年までの data に対して，$\bar{x} = 105.7867$, $s^2 = 0.199822$, $s = 0.447015$ が得られる．

  一方，自由度 14 の t 分布において $P(|t| \geq 2.145) = 0.025$ が知られている．

  よって，95% 信頼区間は $\dfrac{|105.7867 - \mu|}{\frac{0.447015}{\sqrt{14}}} \leq 2.145$,

  即ち， $105.53 \leq \mu \leq 106.04$

■□■ 演習問題 ■□■

1 3 人の人が，2 枚の硬貨を投げて表が出る枚数を当てるゲームをしている.

A さて，表を出している硬貨は 0 枚か 1 枚か 2 枚です. あなたはどれに賭けますか？

B 表裏の出方は，表表，表裏，裏表，裏裏の 4 通りの場合があって，表が 0 枚となるのはそのうち 1 通り，1 枚となるのは 2 通り，2 枚となるのは 1 通りですから,「1 枚」に賭けますね.

C しかし，2 枚の硬貨が寸分違わず同じで，全く見分けがつかないとしたらどうでしょう. 2 枚のうち 1 枚だけ表だと言っても，どちらの硬貨が表なのか分かりませんから，表裏の組み合わせは全部で 3 通りと考えるべきです. そして，どれも同じ確率で起きるはずです.

B と C のどちらが正しいのか，先験的 (a priori) に判断できるだろうか.

2 信頼区間を用いて確率を推定する立場から,「金星のネコ」に関する B 氏の主張を批判せよ.

3 統計的判断に完璧な正しさを期待することはできない. 統計的判断において誤りを避けることができない根本的理由は何か.

4 信頼区間を求める手続きに従って信頼水準 95% での信頼区間を求めたところ $[0.3,\ 0.4]$ となったとする. このとき信頼水準が 95% であるということの意味について，正しい文はどちらか.

(1) The probability that this estimation is correct is 95%.

(2) The probability that this procedure gives a correct estimation is 95%.

5 関東地区で視聴率調査をするとき 2000 件のサンプルで十分だとすれば，日本全国の視聴率調査をするときはどうか.

6 硬貨を 500 回投げたところ，250 回表が出た. 信頼水準 99% で，この硬貨が表を出す確率 p の信頼区間を求めよ. また，50000 回投げて 25000 回表が出たとしたらどうか.

7 2000 世帯のうち 700 世帯で，あるテレビ番組を見ていた．視聴率の信頼区間 $[0.345, 0.355]$ の信頼水準を求めよ．

8 ある県の中学 3 年の男子 121 名をランダムに選び，身長を測ったところ，平均は 161.5cm，標準偏差は 6.0cm だった．信頼水準 95% で，母平均の信頼区間を求めよ．また，信頼水準 99% ではどうか．

9 腎機能障害の患者 6 名の血清クレアチニン濃度 $(\mathrm{mg/d}\ell)$ のデータ

$$4.0 \quad 3.9 \quad 3.8 \quad 4.0 \quad 4.4 \quad 3.9$$

から，この疾患の血清クレアチニン濃度の母平均について，信頼水準 95% で，信頼区間を求めよ．

§3. 検定

統計学演習 (2004 年度) 渡辺

□□□ テーマ □□□

1. 領主の困惑

ヨハン 領主様。トーマスの言うことを信じてはいけません。
トーマス 私はただひよこを売っただけです。
ヨハン 買ったひよこを育ててみればおんどりばかり！
トーマス それはたまたま，…
ヨハン たまたま，ひよこが 10 羽全部牡だったなど，まずあり得ぬこと。
領　主 ほう。ひよこの雌雄を見るのは至難と聞いているが，…
ヨハン いえいえ，トーマスにはそれが分かるのです。
領　主 その証拠は？
ヨハン 証拠は，私のおんどり達です。ひよこの雌雄を見抜かない限り，このようなことが起きるはずはありません。
トーマス それはいいがかりと言うもの。
領　主 うーむ。何もしかけがないとすれば，まず起きそうもないことが起きた。それを理由に，しかけがあると断じ得るものであろうか …

(統計的仮説検定と過誤の問題)

2. メンデルの実験

メンデルは 1865 年，エンドウ豆の遺伝形質を調べた実験において，次のような結果を得た.

表現型	黄色・丸い	黄色・しわがある	緑色・丸い	緑色・しわがある	合計
観測度数	315	101	108	32	556
理論比	9	3	3	1	16

このデータは，“メンデルの法則”の正しさを証明していると言えるだろうか？

(適合度の χ^2 検定)

3. 天気予報

天気予報が当るものかどうか判断するには，どのようなデータを集めて，どのような処理をすればいいだろうか？

(独立性の χ^2 検定)

□■□ ノート □■□

(1) 仮説検定の考え方

『もしもある仮説 H_0 が正しければとうてい起きそうもないことが起きた.
このことを根拠に H_0 は誤りであると判断する.』

H_0 を **帰無仮説 (null hypothesis)** という.

ヨハンのひよこの例に即して考える.

トーマスが雄鶏を選ぶ確率を p とする. 10 羽中 k 羽が雄鶏である確率は 2 項分布 $B_{10.p}$ に従うと考える.

k	0	1	2	3	4	5	6	7	8	9	10
$p=\frac{1}{2}$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001
$p=\frac{3}{4}$	0.000	0.000	0.000	0.003	0.016	0.058	0.146	0.250	0.282	0.188	0.056

帰無仮説「H_0 : $p=\dfrac{1}{2}$」を検定する. (H_0 を否定したい)
次のルールを考える.

- $3 \leq k \leq 7$ となったら H_0 を **採択 (accept)** する.
- $k \leq 2$ または $k \geq 8$ となったら H_0 を **棄却 (reject)** する.

棄却する場合の k の範囲

$$k \leq 2 \text{ or } k \geq 8$$

を **棄却域 (critical region)** という.

もちろん, このルールは誤った結論に導くことがある.

第 1 種の過誤 : トーマスにぬれぎぬを着せる
第 2 種の過誤 : トーマスにだまされる

(2) 有意水準

- **第１種の過誤 (error of the first kind)**
  帰無仮説 H_0 が正しい場合にも, それを棄却してしまうことがある. その確率は

$$\alpha = 0.001 + 0.010 + 0.044 + 0.044 + 0.010 + 0.001 = 0.11$$

この値を **有意水準 (significance level)** という.

- 例

  ヨハンのひよこのうち，８羽が雄鶏だったとする．

  　→有意水準 0.11 で H_0 を棄却する．

  ヨハンのひよこのうち，３羽が雄鶏だったとする．

  　→有意水準 0.11 で H_0 を採択する．（むしろ，棄却できなかった，と言うべき．）

(3) 棄却域の取り方

- **両側検定 (both-sided test) と片側検定 (one-sided test)**

  棄却域は左右対称にしなくてもよい．トーマスのひよこの例では

  $k \leq 2, k \geq 8$　　有意水準 0.11 の両側検定

  $k \leq 3$　　有意水準 0.172 の下側検定 ｝片側検定

  $k \geq 7$　　有意水準 0.172 の上側検定 ｝片側検定

  一般に，有意水準が同じでも，棄却域の取り方によって検定法に優劣が生じる．棄却域はどのように取るべきか．

- **第２種の過誤 (error of the second kind)**

  帰無仮説が誤っているのに採択してしまう（棄却できない）ことがある．

  その確率 β を求めることができるだろうか．

  帰無仮説が成立しないというだけでは，それに代わる仮説が立たないので，仮に

  　トーマスが雄鶏を選ぶ確率 p が $\frac{1}{2}$ か $\frac{3}{4}$ のどちらかである

  としよう．

  H_0: $p = \dfrac{1}{2}$　（帰無仮説）

  H_1: $p = \dfrac{3}{4}$　（対立仮説，alternative hypothesis）

  H_0 が成立せず，H_1 が成立している場合に，誤って H_0 を採択する（H_1 を棄却する）確率 β は，

  棄却域が $k \leq 2$ または $k \geq 8$ $\Rightarrow \beta = 0.003 + 0.016 + 0.058 + 0.146 + 0.250 = 0.473$

  棄却域が $k \leq 3$　　$\Rightarrow \beta = 0.997$

  棄却域が $k \geq 7$　　$\Rightarrow \beta = 0.223$

  上の３つの棄却域に対し，第１種の過誤が起きる確率 α はあまり変わらない．しかし，第２種の過誤が起きる確率 β は異なる．

  上の例では，棄却域を $k \geq 7$ としたとき，第２種の過誤が最も起きにくい (**検出力 (power)** が高い)．しかし，片側検定は恣意的になりがちなので，多用すべきではない．

(4) 仮説検定の考え方 (まとめ)

- 仮説検定

  母集団についての仮説の真偽を，標本に基づいて検証すること．通常は，

  　もしも仮説 H_0 が正しければ到底起きそうもないことが起きた．この事を根拠に，
  　H_0 を棄却する．

  という形を取ることが多い．このときの仮説 H_0 を帰無仮説という．

- 第 1 種の過誤

  帰無仮説が正しいにも拘らず，たまたま棄却域の値が出たため，正しい帰無仮説を棄却する誤り．第 1 種の過誤が起きる確率を有意水準という．

- 第 2 種の過誤

  帰無仮説が偽であるにも拘らず，たまたま棄却域の値が出なかったため，偽である帰無仮説を採択する誤り．第 2 種の過誤が起きにくい検定法は検出力が高いという．

(5) χ^2 **分布 (chi-square distribution)**

基準正規分布 $N(0,1)$ に従う独立な確率変数 $X_1, X_2, \cdots, X_n$ の平方和

$$\chi^2 = \sum_{i=1}^{n} X_i^2$$

が従う分布を，自由度 (degree of freedom) n の χ^2 分布 (χ_n^2) という．分布 χ_n^2 の確率密度関数は，次式で与えられる．

$$f_n(x) = \begin{cases} c_n x^{n/2-1} e^{-x/2}\,, & x > 0 \\ 0\,, & x \le 0 \end{cases}$$

ただし，c_n は (全確率を 1 にするための) 定数である．

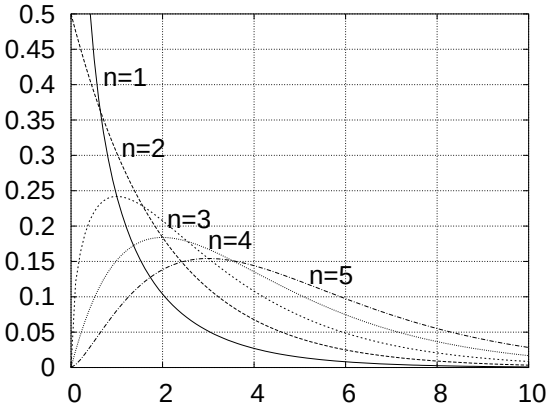


図 1.　自由度 n の χ^2 分布の確率密度関数 $f_n(x)$

(6) **適合度の** χ^2 **検定 (**χ^2**– test of goodness of fit) の原理**

母集団からランダムに選んだ N 個のものを，属性 A によって，階級 $A_1, A_2, \cdots, A_r$ に分割したとき，性質 A_i をもつものの数を x_i とする．

このとき，

仮定： 母集団全体において，各個体が性質 A_i をもつ確率は p_i である

のもとで，N 個のもののうち，性質 A_i をもつものの理論度数は

$$y_i = p_i N$$

となり，統計量

$$\chi^2 = \sum_{i=1}^{r} \frac{(x_i - y_i)^2}{y_i}$$

は，近似的に，自由度 $r-1$ の χ^2 分布に従う.

階級	A_1	A_2	$\cdots$	A_r	計
観測度数	x_1	x_2	$\cdots$	x_r	N
理論度数	y_1	y_2	$\cdots$	y_r	N

(7) **メンデルの実験**

帰無仮説 H_0：観測度数はメンデルの法則 (9:3:3:1) に適合している.

観測度数 x_i	315	101	108	32
理論度数 y_i	312.75	104.25	104.25	34.75

H_0 を有意水準 5% で検定する（上側検定）.

$$\chi^2 = \sum_{i=1}^{4} \frac{(x_i - y_i)^2}{y_i} = 0.470$$

H_0 の下で，χ^2 は近似的に自由度 3 の χ^2 分布に従い，$\mathrm{Prob}(\chi \geq 7.815) = 0.05$. 従って，$H_0$ を採択する.

余談：

$\mathrm{Prob}(\chi^2 \leq 0.58) = 0.1$ だから $\chi^2 = 0.47$ は小さすぎる. つまり，メンデルの実験結果は理論に合いすぎている.（有意水準 0.1 の下側検定で帰無仮説は棄却される.）メンデルは多数の実験を繰り返して，もっとも自説に合っているものだけを発表したのではないか？(Fisher)

自由度 n の χ^2 分布では，大まかに言って，自由度 $(= n)$ の付近の値を取ると考えてよい. 理論と実験データの差を χ^2 を使って測るとき，1 自由度あたり 1 程度の誤差が見えるのが自然.

(8) **独立性の χ^2 検定 (χ^2– test for independence) の原理**

母集団からランダムに選んだ N 個のものを，二つの属性 A, B によって，階級 $A_1, A_2, \cdots, A_r$, $B_1, B_2, \cdots, B_s$ に分割したとき，性質 A_i と性質 B_j の両方をもつものの数を x_{ij} として，次のような **分割表 (contingency table)** を作る.

	B_1	B_2	$\cdots$	B_s	計
A_1	x_{11}	x_{12}	$\cdots$	x_{1s}	$x_{1.}$
A_2	x_{21}	x_{22}	$\cdots$	x_{2s}	$x_{2.}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_r	x_{r1}	x_{r2}	$\cdots$	x_{rs}	$x_{r.}$
計	$x_{.1}$	$x_{.2}$	$\cdots$	$x_{.s}$	N

このとき，

仮定： 母集団全体において，性質 A, B は独立である

のもとで，N 個のもののうち，性質 A_i, B_j をもつものの理論度数は

$$y_{ij} = \frac{x_{i.}}{N}\frac{x_{.j}}{N}N = \frac{x_{i.}x_{.j}}{N}$$

となり，統計量

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(x_{ij}-y_{ij})^2}{y_{ij}}$$

は，近似的に，自由度 $(r-1)(s-1)$ の χ^2 分布に従う.

(9) **天気予報** 2002 年に関東甲信地方に出された予報（17 時発表）

(a) 今夜 17 時－ 24 時の降水予報と実況

	予報		
(x_{ij}) の表	降水あり	降水なし	計
実　降水あり	40	21	61
況　降水なし	14	290	304
計	54	311	365

気象庁発表の百分率から調整したデータ（365 回の予報の成績として計算）

帰無仮説 H_0：予報と実況は独立である.

H_0 の下での理論度数を (y_{ij}) とする.

	予報	
(y_{ij}) の表	降水あり	降水なし
実　降水あり	9.02	51.98
況　降水なし	44.98	259.02

（例）　$\frac{54}{365} \times 61 = 9.02$

$$\chi^2 = \sum_{i,j}\frac{(x_{ij}-y_{ij})^2}{y_{ij}} = \frac{(40-9.02)^2}{9.02} + \cdots = 149.91$$

H_0 の下で，χ^2 は（近似的に）自由度 1 の χ^2 分布に従い，$\mathrm{Prob}(\chi^2 \geq 7.88) = 0.005$．ほぼ確実に H_0 は棄却される.

(b) 明後日（0 時－ 24 時）の降水予報と実況

	予報		
(x_{ij}) の表	降水あり	降水なし	計
実　降水あり	57	55	112
況　降水なし	22	231	253
計	79	286	365

(y_{ij}) の表		予報 降水あり	予報 降水なし
実況	降水あり	24.2	87.8
	降水なし	54.8	198.2

$\chi^2 = 81.8$

やはり，ほぼ確実に H_0 は棄却される．

■□■ 演習問題 ■□■

1 メンデルの実験に関する χ^2 検定において上側検定を用いた．この場合，上側検定が妥当である理由は何か．

2 工場 F で製造された砲丸投げ用の砲丸について，重心の狂いが 1mm 以上のものは 5% しかなく，2mm 以上のものは 1% しかないという．ある砲丸は，重心の狂いが 1.5mm あった．

(1) 「この砲丸は工場 F で作られたものである」という仮説は，有意水準 5% で棄却される．しかし，

「この砲丸は工場 F で作られたものでない」という判断は確率 95% で正しい

と考えてはならない．「有意水準 5% で棄却される」とは，どういう意味か．

(2) 「この砲丸は工場 F で作られたものである」という仮説は，有意水準 1% で採択される．しかし，

「この砲丸は工場 F で作られたものである」と積極的に主張できる

と考えてはならない．「有意水準 1% で採択される」とは，どういう意味か．

3 IBM のスーパーコンピューター “Deep Blue” は，1996 年，13 代チェス世界チャンピオン，ガルリ・カスパロフと対戦し，1 勝 3 敗 2 分でチャンピオンに負けた．この成績を 2 勝 4 敗と見なすとき，実力の差は “決定的” と言えるだろうか．統計的仮説検定の立場から論ぜよ．

(1997 年に行われた 2 度目の対戦では 2 勝 1 敗 3 分で Deep Blue が勝った.)

4 Hayes は，1943 年から 1958 年までの 16 年間に，ノースカロライナのある病院における記録をもとに，急性白血病の発症数を月別に集計して，次のような表を得た．この結果において，急性白血病の発症に季節変動が認められるか．有意水準 5% で検定せよ．

月	1	2	3	4	5	6	7	8	9	10	11	12	計
頻度	23	21	15	20	14	8	11	11	14	17	10	20	184

5

下の表は，イギリスにおいて，60 – 64 歳の軍人恩給受給者に喫煙習慣をアンケートでたずね，6 年後の生存・死亡を調べたものである．喫煙習慣と生存・死亡の間に関連が認められるか．有意水準 5% で検定せよ．(年齢を制御したのは，死亡率が年齢に依存するからである.)

	非喫煙者	パイプ喫煙者	計
生存	117	54	171
死亡	950	348	1298
計	1067	402	1469

6

総務省統計局の「労働力調査報告」によると，2002 年の労働力人口は次の表の通りである．この資料から，労働者の男女構成は年齢層によって異なると言ってよいか．

	男	女
若年層 15–29 歳	816	671
中年層 30–54 歳	2159	1466
高年齢層 55 歳以上	981	594

(単位＝万人)

7

第 4 問の Hayes による急性白血病の発症数調査について再考する．一般に，全度数を N，m 月における度数を x_m $(m = 1, 2, \cdots, 12)$ とし，

$$C = \sum_{m=1}^{12} x_m \cos\frac{m\pi}{6}\,, \quad S = \sum_{m=1}^{12} x_m \sin\frac{m\pi}{6}$$

とおいて，統計量

$$R = \frac{2}{N}(C^2 + S^2)$$

を考える．このとき,「季節変動が存在しない」という仮定のもとで，R は自由度 2 の χ^2 分布に従う．この方法を用いると，有意水準 1% で，Hayes のデータに季節変動が認められることを示せ.

§4. 正規母集団の検定

統計学演習 (2004 年度) 渡辺

□□□ テーマ □□□

1. 再処理の効果

ある製鉄メーカーが製造している鋼板の鉄の純度は，平均 97.10% である. 試みに 1 工程増やして純度を高めるテストをした結果，試験的に製造された 10 枚の鋼板の鉄の純度は

97.06 97.07 97.93 97.03 97.50 97.60 97.67 97.51 97.06 97.01 (%)

となった. これらの値の平均は 97.544% であるが,「製品の純度は高くなった」と考えていいだろうか.

(母平均の検定)

2. 気温の比較

下の表は，東京と大阪の 1988 年 8 月前半の最高気温である.

日	東京	大阪
1	32.1	35.4
2	26.2	34.6
3	27.5	31.1
4	31.8	32.4
5	32.1	33.3

日	東京	大阪
6	31.2	34.7
7	30.1	35.3
8	32.4	34.3
9	32.3	32.1
10	29.9	28.3

日	東京	大阪
11	29.6	33.3
12	26.6	30.5
13	31.2	32.6
14	30.9	33.3
15	29.3	32.2

このデータによると，東京の最高気温の平均は 30.2°C，大阪の最高気温の平均は 32.9°C であるが,「大阪は東京より暑い」と言ってよいだろうか？

(母平均の差の検定)

3. 睡眠時間

次の表は，1996 年 4 月から 1997 年 1 月にかけて，日本学校保険会が日本の小学生，中学生，高校生の睡眠時間を調査した結果である.

	(人数)	平均値	標準偏差
小学 3・4 年生男子	(350)	9 時間 03 分	0 時間 40 分
小学 3・4 年生女子	(322)	9 時間 01 分	0 時間 40 分
小学 5・6 年生男子	(530)	8 時間 56 分	0 時間 42 分
小学 5・6 年生女子	(481)	8 時間 45 分	0 時間 42 分
中学生男子	(992)	7 時間 34 分	1 時間 07 分
中学生女子	(941)	7 時間 10 分	1 時間 02 分
高校生男子	(1073)	6 時間 55 分	1 時間 21 分
高校生女子	(1985)	6 時間 42 分	1 時間 15 分

(「児童生徒の健康状態サーベイランス事業報告書」1998 年)

A　大まかには，睡眠時間は小学 3・4 年生の男子が最も長く，高校生の女子が最も短いということだね.

B　睡眠時間を聞いて，年齢と性別を当てることもできそうだ.

C　そうかな．たとえば高校生の男子と女子の睡眠時間の分布はほとんど重なっているよ.

D　つまり高校生の男女の睡眠時間には統計的な差がないということさ.

さて，正しいことを言っているのは誰だろうか.

(Welch の検定)

□■□ ノート □■□

この節では，母集団は全て正規分布すると仮定する．

(1) χ^2 分布 (chi-square distribution) と t 分布 (t-distribution)

基準正規分布 $N(0,1)$ に従う独立な確率変数 $X_1, X_2, \cdots, X_n$ の平方和

$$\chi^2 = \sum_{i=1}^{n} X_i^2$$

が従う分布を，自由度 n の χ^2 分布 (χ_n^2) という．

Z が $N(0,1)$ に従い、χ^2 が χ_n^2 に従う確率変数で、Z と χ^2 が独立のとき、

$$t = \frac{Z}{\sqrt{\frac{1}{n}\chi^2}}$$

が従う分布を，自由度 n の t 分布 (t_n) という．分布 t_n の確率密度関数は

$$f_n(x) = c_n \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

で与えられる．ただし，c_n は，

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

を満たすように選ばれた定数である．

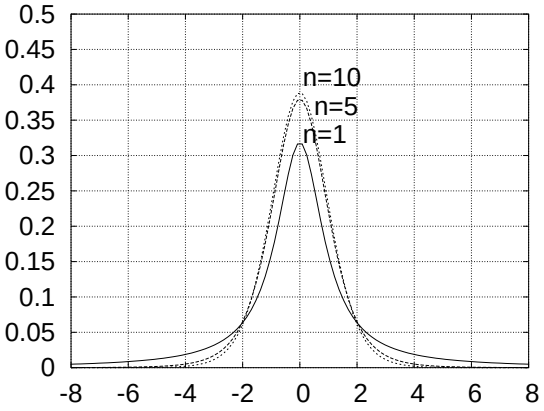


図 1.　自由度 n の t 分布の確率密度関数 $f_n(x)$

注意：n が大変大きいとき，分布 t_n は基準正規分布 $N(0,1)$ で近似される．

(2) 母平均の検定

母平均 μ の正規母集団 M から選んだ標本 $x_1, x_2, \cdots, x_n$ に対し，その標本平均を $\bar{x}$，標本分散を s^2 として

$$t = \frac{\bar{x} - \mu}{\frac{1}{\sqrt{n-1}}s}$$

とおくと，t は自由度 $n-1$ の t 分布に従う．

(3) **再処理の効果（テーマ 1) — 母平均の検定**

鋼板の純度

$$97.06 \;\; 97.07 \;\; 97.93 \;\; 97.03 \;\; 97.50 \;\; 97.60 \;\; 97.67 \;\; 97.51 \;\; 97.06 \;\; 97.01 \; (\%)$$

に対し，

帰無仮説 H_0： $\mu = 97.10$

を検定する．

$\bar{x} = 97.544$, $s^2 = 0.2598$, $n = 10$ から，

$$t = \frac{97.544 - 97.10}{\frac{1}{\sqrt{9}}\sqrt{0.2598}} = 2.613$$

t は，自由度 9 の t 分布に従う．

(a) $\mathrm{Prob}(|t| \geq 2.262) = 0.05$
よって，有意水準 0.05 の両側検定で H_0 を棄却する．

(b) $\mathrm{Prob}(|t| \geq 3.250) = 0.01$
よって，有意水準 0.01 の両側検定で H_0 を採択する．

(c) $\mathrm{Prob}(t \geq 2.821) = 0.01$
よって，有意水準 0.01 の上側検定で H_0 を採択する．

(4) **等平均の検定**

二つの正規母集団 M_x, M_y の母平均を μ_x, μ_y，母分散を σ_x^2, σ_y^2 とする．

$x_1, x_2, \cdots, x_m$ を M_x の標本，$y_1, y_2, \cdots, y_n$ を M_y の標本，それぞれの標本平均を $\bar{x}, \bar{y}$，不偏分散を v_x^2, v_y^2 とする．

このとき，

$$c = \frac{(m-1)v_x^2 + (n-1)v_y^2}{m+n-2}\left(\frac{1}{m} + \frac{1}{n}\right)$$
$$t = \frac{\bar{x} - \bar{y}}{\sqrt{c}}$$

とおくと，

仮定： $\sigma_x^2 = \sigma_y^2,\ \mu_x = \mu_y$

のもとで，t は，自由度 $m+n-2$ の t 分布に従う．

注意：この定理を用いると，

等分散の仮定： $\sigma_x^2 = \sigma_y^2$

の下で，等平均 $\mu_x = \mu_y$ の検定をすることができる．

(5) F **分布 (F-distribution)**

X が χ^2_m に従い、Y が χ^2_n に従う確率変数で、X と Y が独立のとき、$F = \dfrac{\frac{1}{m}X}{\frac{1}{n}Y}$ が従う分布を、自由度対 (m, n) の F 分布 (F_n^m) という.

分布 F_n^m の確率密度関数は

$$f_{mn}(x) = \begin{cases} c_{mn} x^{m/2-1} \left(1 + \dfrac{m}{n}x\right)^{-(m+n)/2} , & x > 0 \\ 0 , & x \leq 0 \end{cases}$$

で与えられる. ただし，c_{mn} は (全確率を 1 にするための) 定数である.

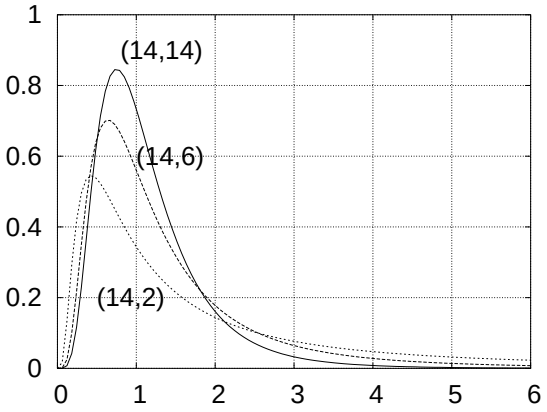


図 2.　自由度 (m, n) の F 分布の確率密度関数 $f_{mn}(x)$

(6) **等分散の検定**

二つの正規母集団 M_x, M_y の母分散を σ_x^2, σ_y^2 とする.

$x_1, x_2, \cdots, x_m$ を M_x の標本，$y_1, y_2, \cdots, y_n$ を M_y の標本，それぞれの不偏分散を v_x^2, v_y^2 とする. このとき,

$$\textbf{仮定：} \quad \sigma_x^2 = \sigma_y^2$$

のもとで,

$$F = \frac{v_x^2}{v_y^2}$$

は，自由度対 $(m-1, n-1)$ の F 分布 F_{n-1}^{m-1} に従う.

注意： $\dfrac{1}{F} = \dfrac{v_y^2}{v_x^2}$ は自由度対 $(n-1, m-1)$ の F 分布 F_{m-1}^{n-1} に従う. F 分布表では $F > 1$ の範囲の数値が与えられているので，$F < 1$ の場合は $\dfrac{1}{F}$ の分布を利用する.

(7) **気温データ（テーマ 2）** — **等分散の検定**

	標本平均	標本分散	不偏分散
東京	$\bar{x} = 30.21$	$s_x^2 = 3.936$	$v_x^2 = \frac{15}{14}s_x^2 = 4.217$
大阪	$\bar{y} = 32.89$	$s_y^2 = 3.481$	$v_y^2 = \frac{15}{14}s_y^2 = 3.730$

帰無仮説 H_0： $\sigma_x^2 = \sigma_y^2$ （東京と大阪の最高気温の母分散は等しい）

を検定する.

$F = \dfrac{v_x^2}{v_y^2} = \dfrac{4.217}{3.730} = 1.131$

F は F_{14}^{14} 分布に従い,

$$\mathrm{Prob}(F \geq 2.99) = 0.025$$

また, $\dfrac{1}{F}$ は F_{14}^{14} 分布に従い,

$$\mathrm{Prob}(F \leq \frac{1}{2.99}) = \mathrm{Prob}(\frac{1}{F} \geq 2.99) = 0.025$$

よって，有意水準 5% の両側検定で H_0 を採択.

(8) **気温データ（テーマ 2） — 等平均の検定**

等分散の帰無仮説を採択した状況で，等平均の仮説検定をする.

帰無仮説 H_0： $\mu_x = \mu_y$ （東京と大阪の最高気温の母平均は等しい）

を検定する． $m = 15, n = 15$ だから,

$c = \dfrac{14 \times 4.217 + 14 \times 3.730}{28}(\dfrac{1}{15} + \dfrac{1}{15}) = 0.5298$

$t = \dfrac{30.21 - 32.89}{\sqrt{0.5298}} = -3.682$

自由度 28 の t 分布によると $\mathrm{Prob}(|t| \geq 3.055) = 0.005$．よって，有意水準 0.5% の両側検定で H_0 を棄却する.

注意：東京と大阪の気温のデータはペアをなして対応している**対標本 (paired sample)** である．このような場合には，気温の差 $z = x - y$ が正規分布するという仮定の下で，z の母平均が 0 であるかどうか検定する方法もある (演習問題 3 参照).

(9) **Welch の検定**

等分散の仮定が成立しない場合にも用いることができる「等平均の検定法」の 1 つ.

二つの正規母集団 M_x, M_y の母平均を μ_x, μ_y とする.

$x_1, x_2, \cdots, x_m$ を M_x の標本，$y_1, y_2, \cdots, y_n$ を M_y の標本，それぞれの標本平均を $\bar{x}, \bar{y}$，不偏分散を v_x^2, v_y^2 とする.

このとき,

$$r_x = \frac{v_x^2}{m}\ , \quad r_y = \frac{v_y^2}{n}$$

$$\nu = \frac{(r_x + r_y)^2}{\dfrac{r_x^2}{m-1} + \dfrac{r_y^2}{n-1}}$$

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{r_x + r_y}}$$

とおくと，

仮定： $\mu_x = \mu_y$

のもとで，t は近似的に，自由度 ν^* の t 分布に従う．ただし，ν^* は ν に最も近い整数である．

(10) **気温データ（テーマ 2）— Welch の検定**

$m = n = 15$,

$\bar{x} = 30.21$, $v_x^2 = 4.217$, $r_x = \dfrac{4.217}{15} = 0.2811$

$\bar{y} = 32.89$, $v_y^2 = 3.730$, $r_y = \dfrac{3.730}{15} = 0.2487$

$\nu = \dfrac{(0.2811 + 0.2487)^2}{0.2811^2 + 0.2487^2} \times 14 = 27.896$ から $\nu^* = 28$ として，

$$t = \frac{30.21 - 32.89}{\sqrt{0.2811 + 0.2487}} = -3.682$$

t は自由度 28 の t 分布に従い，

$\mathrm{Prob}(|t| \geq 3.055) = 0.005$

よって有意水準 0.5% の両側検定で等平均の帰無仮説を棄却．

(11) **高校生男子，女子の睡眠時間の場合（テーマ３）— Welch の検定**

男子	$m = 1073$	$\bar{x} = 6:55$	$s_x^2 = 81^2$	$v_x^2 = \frac{1073}{1072} s_x^2 = 6567$	$r_x = 6.120$
女子	$n = 1985$	$\bar{y} = 6:42$	$s_y^2 = 75^2$	$v_y^2 = \frac{1985}{1984} s_y^2 = 5628$	$r_y = 2.835$

$\nu = 2056.7$ から $\nu^* = 2057$ として，

$t = \dfrac{13}{\sqrt{8.955}} = 4.344$ に対して t_{2057} 分布を用いればよい．

自由度が大きいとき t 分布はほぼ標準正規分布 $N(0,1)$ に等しい．

正規分布を用いると， $\mathrm{Prob}(|t| \geq 3.89) = 0.00005 \times 2 = 0.0001$

よって，有意水準 0.01% で等平均の帰無仮説を棄却．

■□■ 演習問題 ■□■

1 一つの問題に対して，複数の検定法が存在する場合，ある検定法では帰無仮説を棄却する結果になり，他の検定法では採択する結果になることがある．これは，統計学が矛盾を含むことを意味しないだろうか．

2 あるデータを統計分析するために，有意水準を 5% として複数の検定法を試みたとする．このとき，

> 少なくとも一つの方法において帰無仮説を棄却することができたら，期待される結論は統計的に正しいと主張してよい

と言えるか．

3 テーマ 2 のデータを再考する．東京の気温 x と大阪の気温 y の差 $z = x - y$ が正規分布すると仮定して，z の母平均が 0 であるかどうか，有意水準 0.1% で両側検定せよ．

4 透析患者の免疫グロブリンの一つである IgG 値 (mg/100mℓ) が健常者に比べて高いかどうか調べるために，40 歳代男性の透析患者 9 名，同年代の病院職員の健常者 7 名の IgG 値を測定し，次の結果を得た．

透析患者	1326	1418	1820	1516	1635	1720	1580	1452	1600
健常者	1220	1080	980	1420	1170	1290	1116		

このデータについて，透析患者の IgG 値が健常者に比べて高いかどうか，

(1) 有意水準 5% の両側 F 検定により，等分散の仮説を検定せよ．

(2) 有意水準 1% の両側 t 検定により，等平均の仮説を検定せよ．

(3) 有意水準 1% の両側 Welch 検定により，等平均の仮説を検定せよ．

5 Welch の検定を用いて，小学 $3 \cdot 4$ 年生の男女の睡眠時間の差を検定せよ．

§5. 相関と回帰

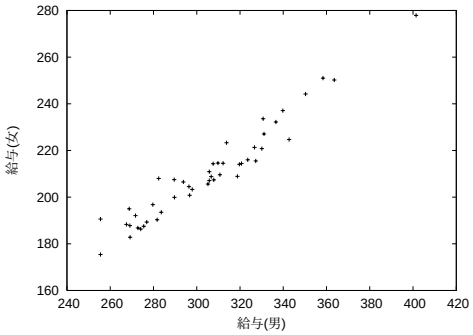
統計学演習 (2004 年度) 渡辺

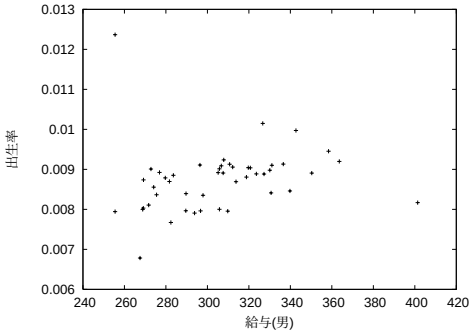
□□□ テーマ □□□

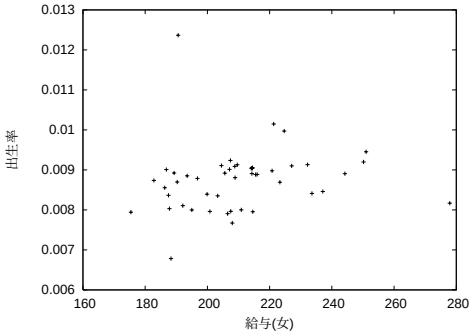
1. 出生率と給与額

次の表は，2003 年の出生率 (出生数/人口) [1] と，2004 年の給与額[2] （単位：千円）の調査結果である．出生数と男女の給与額は互いに関連があると言えるか． **(相関の問題)**

	出生率	給与額 (男)	給与額 (女)
北海道	0.007907	293.8	206.5
青森	0.007943	255.5	175.4
岩手	0.008032	269.1	187.8
宮城	0.008807	318.8	208.9
秋田	0.006784	267.5	188.3
山形	0.008106	271.7	192.1
福島	0.008852	283.6	193.5
茨城	0.008883	327.3	215.5
栃木	0.009039	320.7	214.4
群馬	0.009056	312.2	214.5
埼玉	0.009131	336.6	232.2
千葉	0.008907	350.3	244.2
東京	0.008160	401.4	277.9
神奈川	0.009453	358.4	251.0
新潟	0.007965	289.6	207.5
富山	0.008352	297.9	203.3
石川	0.009237	307.9	207.4
福井	0.009013	305.8	207.1
山梨	0.008692	313.8	223.3
長野	0.008909	307.6	214.3
岐阜	0.009088	306.7	208.8
静岡	0.009041	319.7	214.1
愛知	0.009972	342.7	224.7
三重	0.008886	323.6	216.0
滋賀	0.010149	326.7	221.3
京都	0.008460	339.8	237.1
大阪	0.009199	363.6	250.2
兵庫	0.009101	331.1	227.1
奈良	0.008412	330.7	233.6
和歌山	0.008001	305.8	210.9
鳥取	0.008924	276.9	189.3
島根	0.007999	268.8	195.0
岡山	0.009108	296.4	204.5
広島	0.009129	310.7	209.6
山口	0.007962	296.7	200.8
徳島	0.007955	309.8	214.6
香川	0.008918	305.2	205.6
愛媛	0.008394	289.7	199.9
高知	0.007671	282.4	208.0
福岡	0.008979	330.1	220.8
佐賀	0.009009	272.8	186.8
長崎	0.008555	274.1	186.3
熊本	0.008787	279.7	196.8
大分	0.008365	275.5	187.5
宮崎	0.008737	269.2	182.8
鹿児島	0.008696	281.7	190.3
沖縄	0.012367	255.5	190.6







[1] 出生数は人口動態総覧 (厚生労働省) による．
http://www.mhlw.go.jp/toukei/saikin/hw/jinkou/geppo/nengai03/toukei7.html
人口は平成 12 年国勢調査　第 1 次基本集計結果（全国結果）統計表第 2 表「男女別人口及び世帯の種類（2 区分）別世帯数」による．
http://www.stat.go.jp/data/kokusei/2000/kihon1/00/hyodai.htm

[2] 平成１６年賃金構造基本統計調査結果 (厚生労働省）による．
http://www.mhlw.go.jp/toukei/itiran/roudou/chingin/kouzou/t04/index.html

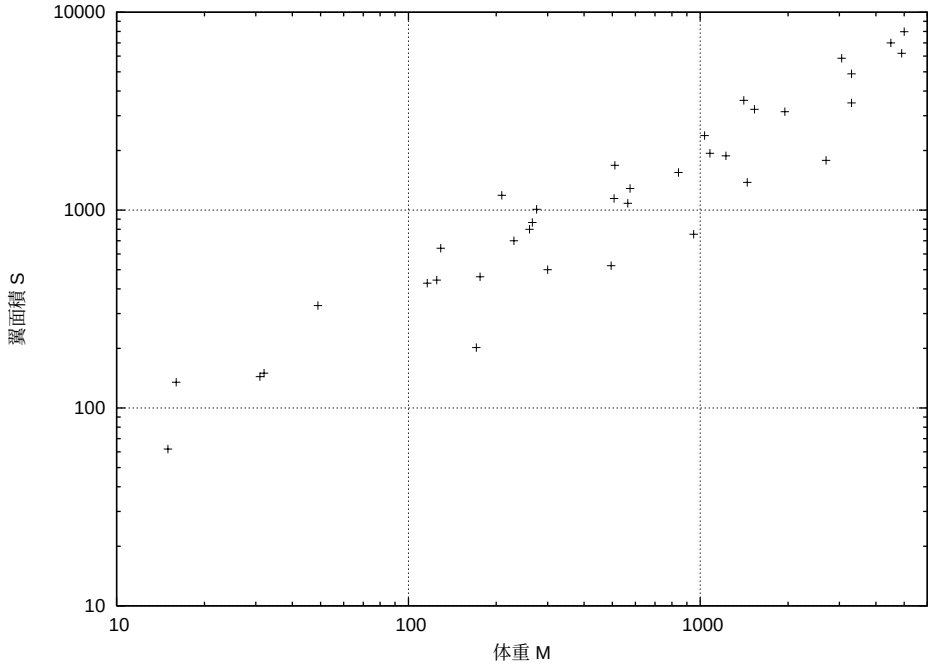
2. 鳥と翼

次の表は，鳥の体重と翼面積 (両翼の面積の和) のデータである．翼面積 S と体重 M の間に近似的な関係式が成立するように見える．鳥の身体構造が種類によらず相似であるとすると，S は $M^{2/3}$ に比例するはずだが，このデータは「相似仮説」を支持しているだろうか．

(回帰の問題)

鳥の種類	体重 M (g)	翼面積 S (cm^2)
オオモズ	31	144
シジュウカラ	15	62
ヒバリ	32	150
ミヤマガラス	575	1285
ハシボソガラス	507	1144
コクマルガラス	230	700
ホシガラス	176	460
カケス	125	443
ヤツガシラ	49	329
ツバメ	16	135
セグロカモメ	565	1082
	842	1550
	1035	2380
	1225	1880
	1080	1936
コウノトリ	3300	4880
オジロワシ	5000	7973
	4500	7000
	4900	6200

鳥の種類	体重 M (g)	翼面積 S (cm^2)
ハゲワシ	1535	3233
チョウゲンボウ	129	642
ミサゴ	3055	5852
	1950	3142
チゴハヤブサ	510	1684
ハイタカ	260	800
	266	866
チョウヒ	209	1188
トラフズク	275	1010
キジ	950	755
ヨーロッパオオライチョウ	2700	1785
	1450	1380
インドクジャク	3300	3480
クイナ	171	202
オオバン	495	524
ヤマシギ	300	500
アオサギ	1410	3584
アジサシ	116	427



3. ことわざと処世観

1. 雨だれ石をもうがつ
2. 君子危うきに近寄らず
3. 当たって砕けよ
4. 石の上にも三年
5. 後ろ足で砂をかける
6. 好機逃すべからず
7. 長い物に巻かれよ
8. 虎穴に入らずんば虎児を得ず
9. 触らぬ神にたたりなし
10. 塵も積もれば山となる
11. 恩を仇で返す
12, 陰にいて枝を折る
13. 親しき仲にも礼儀あり
14. 細き流れも大河となる
15. 人間至る処青山あり
16. 危うい橋も一度は渡れ

上記の 16 個のことわざについて，自分の処世観とどの程度合致するかを約 300 名の学生に標定させた．そのデータに基づいて，各ことわざの評点間の相関係数を調べたところ，以下のような結果が得られた．この結果から，学生によることわざの評価の背景にどのような処世観があると推測されるか．**(主因子分析の問題)**

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1.00	0.09	0.16	0.52	-0.11	0.15	0.07	0.05	-0.02	0.50	-0.23	-0.05	0.22	0.59	0.33	0.11
2	0.09	1.00	-0.17	0.07	0.12	-0.10	0.29	-0.03	0.38	0.01	-0.07	-0.09	0.06	-0.05	-0.17	-0.17
3	0.16	0.17	1.00	0.32	-0.01	0.31	-0.20	0.40	-0.40	0.26	-0.05	0.00	0.25	0.23	0.43	0.45
4	0.52	0.07	0.32	1.00	-0.15	0.23	0.03	0.28	0.08	0.47	-0.34	-0.16	0.37	0.51	0.32	0.16
5	-0.11	0.12	-0.01	-0.15	1.00	0.04	0.12	0.02	0.05	-0.06	0.45	0.51	-0.28	-0.10	-0.07	0.03
6	0.15	0.10	0.31	0.23	0.04	1.00	-0.11	0.35	0.03	0.35	-0.21	-0.08	0.28	0.29	0.44	0.38
7	0.07	0.29	-0.20	0.03	0.12	-0.11	1.00	-0.14	0.34	0.01	-0.03	0.15	0.00	0.18	-0.03	-0.17
8	0.05	0.03	0.40	0.28	0.02	0.35	-0.14	1.00	0.07	0.12	0.03	0.04	0.02	0.10	0.34	0.48
9	-0.02	0.38	-0.40	0.08	0.05	0.03	0.34	0.07	1.00	0.06	-0.05	0.02	0.00	0.09	-0.02	-0.06
10	0.50	0.01	0.26	0.47	-0.06	0.35	0.01	0.12	0.06	1.00	-0.27	-0.08	0.24	0.68	0.37	0.20
11	-0.23	0.07	-0.05	-0.34	0.45	-0.21	-0.03	0.03	-0.05	-0.27	1.00	0.54	-0.43	-0.20	-0.17	0.06
12	-0.05	0.09	0.00	-0.16	0.51	-0.08	0.15	0.04	0.02	-0.08	0.54	1.00	-0.40	0.01	-0.04	0.05
13	0.22	0.06	0.25	0.37	-0.28	0.28	0.00	0.02	0.00	0.24	-0.43	-0.40	1.00	0.37	0.18	0.06
14	0.59	0.05	0.23	0.51	-0.10	0.29	0.18	0.10	0.09	0.68	-0.20	0.01	0.37	1.00	0.42	0.20
15	0.33	0.17	0.43	0.32	-0.07	0.44	-0.03	0.34	-0.02	0.37	-0.17	-0.04	0.18	0.42	1.00	0.54
16	0.11	0.17	0.45	0.16	0.03	0.38	-0.17	0.48	-0.06	0.20	0.06	0.05	0.06	0.20	0.54	1.00

芝祐順著「相関分析法」東京大学出版会 (1975) による

□■□ ノート □■□

1. 相関係数

2 つの確率変数 X, Y に対し，その母平均を

$$\mu_X = E(X)$$
$$\mu_Y = E(Y)$$

母分散を

$$\sigma_X^2 = E((X-\mu_X)^2)$$
$$\sigma_Y^2 = E((Y-\mu_Y)^2)$$

とする．そして X, Y の (母) **共分散 (covariance)** を

$$\mathrm{cov}(X,Y) = E((X-\mu_X)(Y-\mu_Y))$$

で定める．このとき,

$$r_{XY} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

で定義される数 r_{XY} を (母) **相関係数 (correlation coefficient)** という.

特に，$X = Y$ のとき,

$$\mathrm{cov}(X,X) = E((X-\mu_X)^2) = \sigma_X^2$$

であるから $r_{XX} = 1$ となる．また，X と Y が独立であるとき,

$$\mathrm{cov}(X,Y) = E((X-\mu_X)(Y-\mu_Y)) = E(X-\mu_X)E(Y-\mu_Y) = 0$$

であるから $r_{XY} = 0$ となる．一般に

$$-1 \le r_{XY} \le 1$$

が成り立つ.

2. 標本相関係数

2 つの変量 X, Y のサンプル値

$$(x_i, y_i) \quad i = 1, 2, \cdots, n$$

に対し，$\bar{x}, \bar{y}$ を標本平均として，X, Y の (標本) **共分散** を

$$\mathrm{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

で定める．さらに，s_x^2, s_y^2 を標本分散として，(標本) **相関係数** を

$$r = \frac{\mathrm{cov}(x,y)}{s_x s_y}$$

で定める．一般に

$$-1 \leq r \leq 1$$

が成り立つ．標本相関係数は母相関係数の (点) 推定値として用いられる．

定理　母集団において，変量 X, Y が正規分布に従うとき，[3]

帰無仮説：X, Y は互いに独立 (無相関) である

のもとで,

$$t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

は，自由度 $n-2$ の t 分布に従う．

サンプルが与えられたとき，上記の定理を用いて，母集団における X, Y の相関の有無を仮説検定をすることができる．

3．出生率と給与額 — 相関

東京と沖縄のデータは飛び抜けているおり，特殊な事情があると推測されるので，除外して計算する．サンプルの大きさは $n = 45$ となる．

$X =$ 出生率，$Y =$ 給与額 (男)，$Z =$ 給与額 (女) とすると，

$$\bar{x} = 0.00867922\ , \quad \bar{y} = 304.404\ , \quad \bar{z} = 209.102$$

また,

$$\begin{pmatrix} s_x^2 & \mathrm{cov}(x,y) & \mathrm{cov}(x,z) \\ \mathrm{cov}(y,x) & s_y^2 & \mathrm{cov}(y,z) \\ \mathrm{cov}(z,x) & \mathrm{cov}(z,y) & s_z^2 \end{pmatrix} = \begin{pmatrix} 3.64768 \times 10^{-7} & 0.00931848 & 0.00472633 \\ 0.00931848 & 700.779 & 446.906 \\ 0.00472633 & 446.906 & 309.878 \end{pmatrix}$$

よって，X, Y の (標本) 相関係数 r_{xy} は

$$r_{xy} = \frac{\mathrm{cov}(x,y)}{s_x s_y} = \frac{0.00931848}{\sqrt{3.64768 \times 10^{-7}}\sqrt{700.779}} = 0.582836$$

	出生率	給与額 (男)	給与額 (女)
出生率	1.	0.582836	0.44455
給与額 (男)	0.582836	1.	0.959028
給与額 (女)	0.44455	0.959028	1.

[3] ここで X と Y の同時確率分布を考えている．即ち．X と Y の組 (X, Y) を考え，点 (X, Y) の分布が 2 次元正規分布であることを仮定する．

4. 出生率と給与額 — 相関の仮説検定

さらに，出生率 X と男性の給与額 Y が正規分布に従うとすると，

帰無仮説：X, Y は互いに独立 (無相関) である

のもとで，

$$t = \sqrt{43}\frac{r_{xy}}{\sqrt{1 - r_{xy}^2}}$$

は，自由度 43 の t 分布に従い，

$$\mathrm{Prob}(|t| \geq 2.695) = 0.01$$

が成り立つ．

与えられたサンプルに対して，

$$t = \sqrt{43}\frac{0.582836}{\sqrt{1 - 0.582836^2}} = 4.70337$$

であるから，有意水準 1% の両側検定で帰無仮説を棄却する．この意味において，出生率と男性の給与額の間に相関があると考える．

注意　相関があるといっても，因果関係があると主張しているわけではない．相関が生じるメカニズムを探究するのは，また別の問題である．

5. 回帰

2 つの変量 X, Y の間に何らかの近似的な関係があり，

$$Y \fallingdotseq f(X)$$

のような関係式が近似的に成立することがある．X, Y のサンプル値

$$(x_i, y_i) \quad i = 1, 2, \cdots, n$$

を用いて，上のような近似式を求めることを **回帰 (regression)** と言う．このとき，近似の誤差

$$E = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

がなるべく小さくなるように関数 $f(X)$ を選ぶ．この方法を **最小 2 乗法 (method of least squares)** と言う．

たとえば 1 次式

$$Y \fallingdotseq \alpha + \beta X$$

によって回帰するには，

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

が最小となるように α, β を選ぶ.

定理　$E(\alpha, \beta)$ を最小にする α, β は次式で与えられる.

$$\beta = r\frac{s_y}{s_x}$$
$$\alpha = \bar{y} - \beta\bar{x}$$

6. 鳥と翼 — 回帰直線

鳥の体重を M，翼の面積を S とする．MS 平面で両対数プロットして見ると，X, Y の対数

$$X = \log_{10} M$$
$$Y = \log_{10} S$$

の間に近似的な関係

$$Y = \alpha + \beta X$$

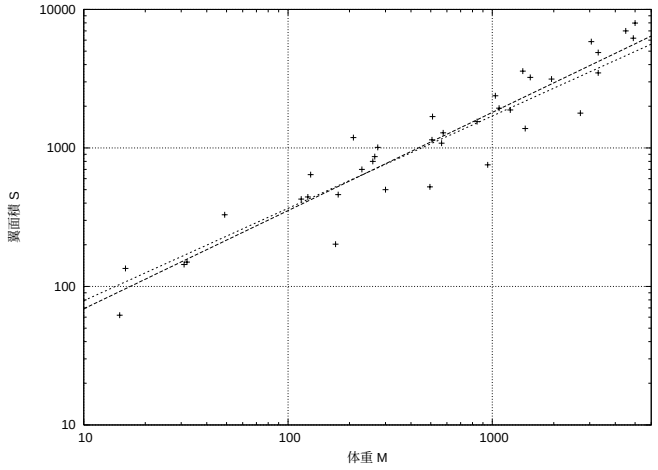
が成立するように見える.

α, β の最適な値を，最小 2 乗法を用いて求める.

$$n = 37$$
$$\bar{x} = 2.66747$$
$$\bar{y} = 3.02521$$
$$s_x^2 = 0.469579$$
$$s_y^2 = 0.263373$$
$$\mathrm{cov}(x, y) = 0.332836$$
$$r = \frac{\mathrm{cov}(x, y)}{s_x s_y} = 0.946432$$
$$\beta = \frac{r s_y}{s_x} = 0.708795$$
$$\alpha = \bar{y} - \beta\bar{x} = 1.13452$$

よって次の回帰式が得られる.

$$Y = 1.134 + 0.7088X$$
$$S = 13.49 \times M^{0.7088}$$



← $S = 13.49 \times M^{0.7088}$ (最小 2 乗法)
↖ $S = 17.00 \times M^{2/3}$ (相似仮説)

7. 回帰係数の推定と検定

母集団において，変量 X, Y が正規分布に従うとする．このとき，確率変数 X の値が x に固定されたとして (母集団全体での) Y の平均値 (期待値) を $f(x)$ とすると，

$$f(x) = \alpha^* + \beta^* x$$

が成り立つ．

この α^*, β^* は，(サンプルを用いるのではなく) 母集団全体で計算した回帰係数 α, β に等しいことが知られている．α^*, β^* を**母回帰係数** と言う．そして，(サンプルを用いて計算した) 標本回帰係数 α, β は，母回帰係数 α^*, β^* の (点) 推定値として用いられる．

定理　母集団において，変量 X, Y が正規分布に従うとき，

$$t = (\beta - \beta^*)\sqrt{\frac{(n-2)s_x^2}{(1-r^2)s_y^2}}$$

は，自由度 $n-2$ の t 分布に従う．

サンプルが与えられたとき，上記の定理を用いて，β^* の信頼区間を作ったり (区間推定)，β^* についての仮説検定をすることができる．

8. 鳥と翼 — 回帰係数の推定

母集団において，変量 X, Y が正規分布に従うとき，

$$t = (\beta - \beta^*)\sqrt{\frac{(n-2)s_x^2}{(1-r^2)s_y^2}}$$

は，自由度 $n-2$ の t 分布に従い，$n=37$ のとき

$$\mathrm{Prob}(|t|>2.03)=0.05$$

が成り立つ．また

$$\begin{aligned} n &= 37 \\ \beta &= 0.708795 \\ \sqrt{\frac{(n-2)s_x^2}{(1-r^2)s_y^2}} &= 24.4641 \end{aligned}$$

であるから，信頼水準 95% の信頼区間は

$$\begin{aligned} & |0.708795-\beta^*|\times 24.4641 < 2.03 \\ \therefore\quad & 0.6258 < \beta^* < 0.7918 \end{aligned}$$

となる．相似仮説に対応する値 $\beta^*=\dfrac{2}{3}$ はこの信頼区間に属している．

注意　相似仮説の妥当性を検定する問題については，演習問題３参照.

9．ことわざと処世観 — 主成分分析

ことわざ i とことわざ j に対する学生の評点の相関係数を r_{ij} とし，r_{ij} を (i,j) 成分とする 16×16 行列 を R とする．

行列 R の固有値 λ と (長さが 1 の) 固有ベクトル $\mathbf{u}$ を計算する．

$$\begin{aligned} R\mathbf{u} &= \lambda\mathbf{u} \\ |\mathbf{u}| &= 1 \end{aligned}$$

固有値と固有ベクトルは 16 組あり，固有値を大きい順に並べると

$$4.08, 2.38, 1.97, 1.33, 0.91, 0.84, 0.77, 0.64, 0.57, 0.5, 0.44, 0.39, 0.37, 0.33, 0.27, 0.21$$

(全部で 16 個)

となる．また固有ベクトルは次の表のようになる．

		4.08	2.38	1.97	1.33	0.91	···
1	雨だれ石をもうがつ	**-0.30**	-0.13	-0.20	-0.34	-0.12	···
2	君子危うきに近寄らず	0.04	-0.27	**-0.30**	0.34	-0.54	···
3	当たって砕けよ	-0.28	**0.30**	0.15	-0.05	-0.43	···
4	石の上にも三年	**-0.35**	-0.12	-0.11	-0.05	-0.31	···
5	後ろ足で砂をかける	0.13	**0.29**	**-0.39**	-0.03	-0.23	···
6	好機逃すべからず	**-0.29**	0.15	0.00	0.27	0.23	···
7	長い物に巻かれよ	0.04	-0.22	**-0.43**	0.06	0.10	···
8	虎穴に入らずんば虎児を得ず	-0.20	**0.30**	-0.03	0.44	-0.19	···
9	触らぬ神にたたりなし	0.02	-0.23	**-0.39**	0.47	0.29	···
10	塵も積もれば山となる	**-0.35**	-0.07	-0.19	-0.22	0.11	···
11	恩を仇で返す	0.23	**0.37**	-0.23	-0.12	-0.08	···
12	陰にいて枝を折る	0.13	**0.34**	**-0.40**	-0.23	0.03	···
13	親しき仲にも礼儀あり	-0.27	-0.24	0.16	0.07	-0.19	···
14	細き流れも大河となる	**-0.36**	-0.09	-0.26	-0.29	0.17	···
15	人間至る処青山あり	**-0.34**	0.20	-0.04	0.09	0.29	···
16	危うい橋も一度は渡れ	-0.24	**0.37**	0.01	0.25	0.13	···

大きい固有値とその固有ベクトルが，評点を左右する何らかの処世観に対応していると考え，上位 3 固有値 4.08, 2.38, 1.97 を選ぶ．これらの 3 つの固有値に対する固有ベクトルの成分 (全部で 16×3 個ある) のうち，絶対値が 0.29 以上のものを見ると (太字の部分)，16 個のことわざがほぼ 3 つのグループに分かれる．

たとえば，第 1 固有値 4.08 の固有ベクトルに深く関係することわざは，

1. 雨だれ石をもうがつ
4. 石の上にも三年
6. 好機逃すべからず
10. 塵も積もれば山となる
14. 細き流れも大河となる
15. 人間至る処青山あり

であり，忍耐を尊びながら楽観的に構えようという処世観の存在を示唆していると言えそうだ．第 2，第 3 の固有値に対応する処世観も同様に解釈できるだろう．そして与えられた統計データから，次のような描像が得られる．

- 3 つの処世観が，各人の中で **独立な重み** を持って配合されている
- このように配合された処世観が，ことわざへの評点として現わている

上記のように，多数のことわざの間の相関を支配している少数の処世観 (固有値と固有ベクトル) を**主成分** と呼び，このような分析法を **主成分分析** という．

■□■ 演習問題 ■□■

1 出生率と女性の給与額 (テーマ 1) の相関の有無を適当な有意水準を設定して検定せよ.

2 男女の給与額 (テーマ 1) の回帰式を求めよ.

3 鳥の体重 M と翼面積 S (テーマ 2) の回帰式

$$\log S = \alpha + \beta \log M$$

において,

相似仮説 : 母回帰係数 β^* は $\dfrac{2}{3}$ に等しい

を適当な有意水準を設定して検定せよ.

4 入学試験の成績と入学後の成績があまり相関しないという現象が知られている. しかしこれは必ずしも入学後の学生の変化を意味するとは限らず, 入学試験の成績が合格点より低い学生のデータを欠いていることに起因する見かけの現象に過ぎないという見方がある. 一般に, 良く相関している 2 変量 X, Y のサンプル (x_i, y_i), $i = 1, 2, \cdots, n$, に対し, x_i がある値より大きいものだけを選ぶと相関係数が小さくなることがある. 男性の給与額と出生率のデータ (テーマ 1) のうち, 男性の給与額が 336 を超える上位 6 件 (東京を除く) に限定して, 相関係数を計算せよ.

	出生率	給与額 (男)
大阪	0.009199	363.6
神奈川	0.009453	358.4
千葉	0.008907	350.3
愛知	0.009972	342.7
京都	0.008460	339.8
埼玉	0.009131	336.6

§6. まとめ

(1) **2 項分布**

確率 p で表を出すコインを n 回投げたとき，k 回表が出る確率は

$$P_{n,k} = {}_n\mathrm{C}_k p^k (1-p)^{n-k} \ , \quad k = 0, 1, 2, \cdots, n$$

である．この確率分布を 2 項分布といい，$B_{n,p}$ と書く．$B_{n,p}$ の平均 μ と分散 σ^2 は，

$$\mu = np$$
$$\sigma^2 = np(1-p)$$

で与えられる．2 項分布 $B_{n,p}$ は，n が大きいとき，平均 np, 分散 $np(1-p)$ の正規分布に近づく．

(2) **正規分布**

確率変数 x が $a < x < a + \Delta a$ (Δa : 微小量) なる値をとる確率が

$$\mathrm{Prob}(a < x < a + \Delta a) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(a-\mu)^2}{2\sigma^2})\Delta a$$

で与えられるとき，即ち，確率密度関数が

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

で与えられるとき，x は，平均 μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ に従うという．

x が $N(\mu, \sigma^2)$ に従うとき，

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

は基準正規分布 $N(0,1)$ に従う．

(3) **χ^2 分布**

基準正規分布 $N(0,1)$ に従う独立な確率変数 $X_1, X_2, \cdots, X_n$ の平方和

$$\chi^2 = \sum_{i=1}^{n} X_i^2$$

が従う分布を，自由度 (degree of freedom) n の χ^2 分布 (χ_n^2) という．分布 χ_n^2 の確率密度関数は，次式で与えられる．

$$f_n(x) = \begin{cases} c_n x^{n/2-1} e^{-x/2} \ , & x > 0 \\ 0 \ , & x \leq 0 \end{cases}$$

ただし，c_n は，

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

を満たすように選ばれた定数である．

(4) t **分布**

Z が $N(0,1)$ に従い、χ^2 が χ_n^2 に従う確率変数で、Z と χ^2 が独立のとき、

$$t = \frac{Z}{\sqrt{\frac{1}{n}\chi^2}}$$

が従う分布を，自由度 n の t 分布 (t_n) という．分布 t_n の確率密度関数は

$$f_n(x) = c_n \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

で与えられる．ただし，c_n は (全確率を 1 にするための) 定数である．特に，n が大変大きいとき，分布 t_n は基準正規分布 $N(0,1)$ で近似される．

(5) F **分布**

X が χ_m^2 に従い、Y が χ_n^2 に従う確率変数で、X と Y が独立のとき、$F = \frac{\frac{1}{m}X}{\frac{1}{n}Y}$ が従う分布を、自由度対 (m,n) の F 分布 (F_n^m) という．

分布 F_n^m の確率密度関数は

$$f_{mn}(x) = \begin{cases} c_{mn} x^{m/2-1} \left(1 + \frac{m}{n}x\right)^{-(m+n)/2}, & x > 0 \\ 0\,, & x \leq 0 \end{cases}$$

で与えられる．ただし，c_{mn} は (全確率を 1 にするための) 定数である．

(6) **適合度の** χ^2 **検定**

定理　母集団からランダムに選んだ N 個のものを，属性 A によって，階級 $A_1, A_2, \cdots, A_r$ に分割したとき，性質 A_i をもつものの数を x_i とする．

このとき，

仮定： 母集団全体において，各個体が性質 A_i をもつ確率は p_i である

のもとで，N 個のもののうち，性質 A_i をもつものの理論度数は

$$y_i = p_i N$$

となり，統計量

$$\chi^2 = \sum_{i=1}^{r} \frac{(x_i - y_i)^2}{y_i}$$

は，近似的に，自由度 $r-1$ の χ^2 分布に従う．

階級	A_1	A_2	$\cdots$	A_r	計
観測度数	x_1	x_2	$\cdots$	x_s	N
理論度数	y_1	y_2	$\cdots$	y_s	N

(7) **独立性の χ^2 検定**

定理　母集団からランダムに選んだ N 個のものを，二つの属性 A, B によって，階級 $A_1, A_2, \cdots, A_r$，$B_1, B_2, \cdots, B_s$ に分割したとき，性質 A_i, B_j をもつものの数を x_{ij} として，次のような分割表 (contingency table) を作る．

	B_1	B_2	$\cdots$	B_s	計
A_1	x_{11}	x_{12}	$\cdots$	x_{1s}	$x_{1.}$
A_2	x_{21}	x_{22}	$\cdots$	x_{2s}	$x_{2.}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_r	x_{r1}	x_{r2}	$\cdots$	x_{rs}	$x_{r.}$
計	$x_{.1}$	$x_{.2}$	$\cdots$	$x_{.s}$	N

このとき，

仮定： 母集団全体において，性質 A, B は独立である

のもとで，N 個のもののうち，性質 A_i, B_j をもつものの理論度数は

$$y_{ij} = \frac{x_{i.}}{N}\frac{x_{.j}}{N}N = \frac{x_{i.}x_{.j}}{N}$$

となり，統計量

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(x_{ij}-y_{ij})^2}{y_{ij}}$$

は，近似的に，自由度 $(r-1)(s-1)$ の χ^2 分布に従う．

(8) **母平均の推定，検定**

定理　母平均 μ の正規母集団 M から選んだ標本 $x_1, x_2, \cdots, x_n$ に対し，その標本平均を $\bar{x}$，標本分散を s^2 として

$$t = \frac{\bar{x}-\mu}{\frac{1}{\sqrt{n-1}}s}$$

とおくと，t は自由度 $n-1$ の t 分布に従う．

(9) **等分散の検定**

定理　二つの正規母集団 M_x, M_y の母分散を σ_x^2, σ_y^2 とする．

$x_1, x_2, \cdots, x_m$ を M_x の標本，$y_1, y_2, \cdots, y_n$ を M_y の標本，それぞれの不偏分散を v_x^2, v_y^2 とする．このとき，

仮定：　$\sigma_x^2 = \sigma_y^2$

のもとで，

$$F = \frac{v_x^2}{v_y^2}$$

は，自由度対 $(m-1, n-1)$ の F 分布 F_{n-1}^{m-1} に従う．

(10) **等平均の検定**

定理　二つの正規母集団 M_x, M_y の母平均を μ_x, μ_y，母分散を σ_x^2, σ_y^2 とする．

$x_1, x_2, \cdots, x_m$ を M_x の標本，$y_1, y_2, \cdots, y_n$ を M_y の標本，それぞれの標本平均を $\bar{x}, \bar{y}$，不偏分散を v_x^2, v_y^2 とする．

このとき，

$$c = \frac{(m-1)v_x^2 + (n-1)v_y^2}{m+n-2}\left(\frac{1}{m} + \frac{1}{n}\right)$$
$$t = \frac{\bar{x} - \bar{y}}{\sqrt{c}}$$

とおくと，

$$\textbf{仮定：} \quad \sigma_x^2 = \sigma_y^2,\ \mu_x = \mu_y$$

のもとで，t は，自由度 $m+n-2$ の t 分布に従う．

(11) **Welch の検定**

定理　二つの正規母集団 M_x, M_y の母平均を μ_x, μ_y とする．

$x_1, x_2, \cdots, x_m$ を M_x の標本，$y_1, y_2, \cdots, y_n$ を M_y の標本，それぞれの標本平均を $\bar{x}, \bar{y}$，不偏分散を v_x^2, v_y^2 とする．

このとき，

$$r_x = \frac{v_x^2}{m}, \quad r_y = \frac{v_y^2}{n}$$
$$\nu = \frac{(r_x + r_y)^2}{\dfrac{r_x^2}{m-1} + \dfrac{r_y^2}{n-1}}$$
$$t = \frac{\bar{x} - \bar{y}}{\sqrt{r_x + r_y}}$$

とおくと，

$$\textbf{仮定：} \quad \mu_x = \mu_y$$

のもとで，t は近似的に，自由度 ν^* の t 分布に従う．ただし，ν^* は ν に最も近い整数である．

(12) **相関係数**

2 つの変量 X, Y のサンプル

$$(x_i, y_i) \quad i = 1, 2, \cdots, n$$

に対し，$\bar{x}, \bar{y}$ を標本平均，s_x^2, s_y^2 を標本分散とするとき，

$$\mathrm{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

を X, Y の (標本) 共分散，

$$r = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$

を (標本) 相関係数という．

(13) **相関関係の検定**

定理　母集団において，変量 X, Y が正規分布に従うとき，

帰無仮説：X, Y は互いに独立 (無相関) である

のもとで，

$$t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

は，自由度 $n-2$ の t 分布に従う．ただし，r は (標本) 相関係数である．

(14) **回帰係数**

2 つの変量 X, Y に対し，1 次式による回帰

$$Y \fallingdotseq \alpha + \beta X$$

を考える．

定理　2 つの変量 X, Y のサンプル

$$(x_i, y_i) \quad i = 1, 2, \cdots, n$$

に対し，

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

を最小にする α, β (標本回帰係数) は次式で与えられる．

$$\beta = r\frac{s_y}{s_x}$$
$$\alpha = \bar{y} - \beta\bar{x}$$

ただし，$\bar{x}, \bar{y}$ は標本平均，s_x^2, s_y^2 は標本分散，r は (標本) 相関係数である．

(15) **回帰係数の推定と検定**

定理　母集団において，変量 X, Y が正規分布に従うとき，

$$t = (\beta - \beta^*)\sqrt{\frac{(n-2)s_x^2}{(1-r^2)s_y^2}}$$

は自由度 $n-2$ の t 分布に従う．ただし，s_x^2, s_y^2 は標本分散，r は標本相関係数，α, β は標本回帰係数である．