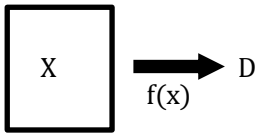
基礎統計（2014 火 5 佐藤俊樹） 重要ポイントまとめ

凡例［赤字：専門用語　紫字：数式　緑字：重要ポイント］　講義ノートをまとめ直したものです。ノートと併用して下さい。（14 年度 L1-9 河野）

☆超重要☆



統計には 2 つの世界がある！「X の世界」と「D の世界」

・X（確率変数）の世界：あるデータの集まりが出てくるだろうと推測される世界。理論上の世界。仮定の世界。
　確率的なことを取り扱う。この世界に属する数値を理論値と言い、基本的に「母〇〇」と呼ぶ
　この世界に属するもの：母平均・母分散・モーメントなど。ギリシャ文字で表す（母比率のみラテン文字 p を使う）

・D（データ）の世界：実際に測定された観測値（データ）を取り扱う。
　この世界に属するもの：平均・分散など。ローマ字で表す

・この 2 つの世界を結ぶものが確率関数 f(x)

・2 つの世界を混同しないこと！

例えば、「サイコロは 1~6 の目が等確率で出る」は X の世界

「実際にサイコロを投げると 1 が 10 回連続で出た」は D の世界。D の世界ではそういうこともあり得る。

ただ、試行回数を多くしていくと、D は X の世界の分布に近づいていく。つまり、「十分に多い観測」を行うと、1~6 の出目は等確率になる。

★D の世界で、実際に観測されたデータ（観測値）の集まりの傾向を測る指標をいくつか紹介する。

※　このデータの集まりのことを「分布」と言う

データの個数＝N、それぞれのデータ＝$\{x_1,\ x_2, x_3, \dots x_N\}$とする

①（標本）平均(mean)：分布の位置がどの辺りなのか

$$m = \frac{1}{N}\sum_i x_i$$

※$\sum_i$とは、$\sum_{i=1}^{N}$のこと

②（標本）分散(variance)：分布がどの程度ばらついているか

$$V = \frac{1}{N}\sum_i (x_i - m)^2$$

※分散のことをs^2と書くことも多い。理由は本シケプリ p. 2 の「標準偏差」参照

この式の意味

例えば、{1,5,2}というデータがあるとする。平均は$\frac{8}{3}$

ばらつき、つまり平均からどれぐらい離れているかを調べたいので、それぞれのデータの平均との差を取ると、$-\frac{5}{3}, \frac{7}{3}, -\frac{2}{3}$

それらを全部合わせた時ばらつきがどれぐらいかを知りたいので、基本は足せばいいのだが、単純に足すと $-\frac{5}{3}+\frac{7}{3}-\frac{2}{3}=0$ となりダメ

そこで、正負の影響を消すためにそれぞれ 2 乗した上で足す。（※ ）

最後に、データの個数によって分散の大きさが変わってこないように、個数 N で割る。

※　正負の影響を消すには、2 乗ではなく絶対値を付けてもいいのでは？ → 平均偏差というものがある

平均偏差＝$\frac{1}{N}\sum_i |x_i - m|$。だが、絶対値は扱いにくい・分布関数（本シケプリ p.2 参照）につながらないなどの理由であまり使われない。

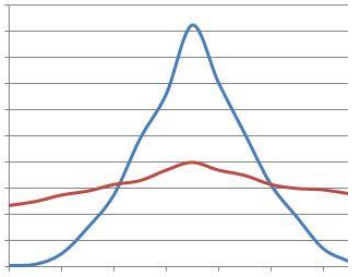
※実は、N ではなく N-1 で割った方が母分散（本シケプリ p.3）に近づき、よい推定・検定（p.8, p.11）ができる

不偏分散：$\frac{1}{N-1}\sum_i (x_i - m)^2$

その理由：自由度（p.6）が関係してくる（詳しくは http://homepage2.nifty.com/nandemoarchive/toukei_hosoku/jiyudo.htm 参照）

統計では、標本分散ではなく不偏分散を使うことの方が多い

だが、本講義では標本分散のみを扱う（テストでは必ず N で割ること！教科書に不偏分散が載っていても騙されないこと！）



← 青は分散小、赤は分散大。どちらも平均は同じ

②’ 分散は 2 乗されているので、データ・平均と単位が異なってしまう。(例. データがメートルで表されていれば、分散は m^2 で表されてしまう) そこで平方根を取り比較を可能にする。

標準偏差(standard deviation) : 分布がどの程度ばらついているか(単位をデータ・平均と同じにしたバージョン)

$$s = \sqrt{V}$$

さらに、分布が全体的に大きい値を取れば（＝平均が大きければ）標準偏差も大きくなる

平均の大きさの影響を取り除くために標準偏差 s を平均 m で割る

→ 分布の大きさが異なるもののばらつき方を比較可能になる：変動係数(coefficient of variation) $\frac{s}{m}$

③ ①②と同様に、x_i-m を 3 乗してみると…

この値が大きい＝平均から正に外れたデータが多い
この値が小さい＝平均から負に外れたデータが多い
} ということが分かる

歪度(skewness)：分布が正負どちらにどれぐらい偏っているか。$\frac{1}{N}\sum_i(x_i - m)^3$

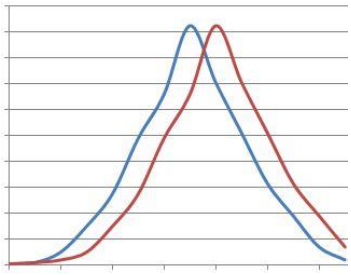
④ x_i-m を 4 乗してみると…

この値が大きい＝尖っている
この値が小さい＝丸みがかっている
} ということが分かる

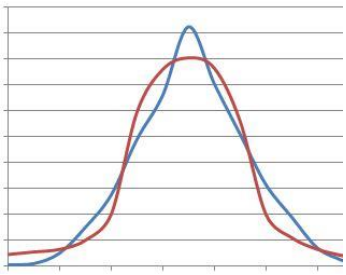
尖度(kurtosis)：分布がどれぐらい尖っているか。$\frac{1}{N}\sum_i(x_i - m)^4$ …(ア)

※最も基本的な分布である正規分布(後述)の尖度を 0 にして基準とするため、3 を引くこともある。

$\frac{1}{N}\sum_i(x_i - m)^4 - 3$ …(イ)。正規分布の尖度は、(ア)では 3、(イ)では 0。



↑赤の方が歪度大



↑青の方が尖度大

①～④のように、$\frac{1}{N}\sum_i(x_i - m)$を 1 乗、2 乗、3 乗…するとその分布の特徴が表れてくる

→ そこで、それを全部足し合わせてみると、その分布の全体像が見え、その分布と 1 対 1 対応する

すなわち、ある分布の分布関数$= \frac{1}{N}\sum_i(x_i - m) + \frac{1}{N}\sum_i(x_i - m)^2 + \frac{1}{N}\sum_i(x_i - m)^3 + \cdots$ と表せる

こういうことができるのが統計学（数理統計学）の面白さ

★X の世界の話

X：確率変数。$\{x_1, x_2, x_3, \cdots\}$という値をそれぞれ $f(x_1), f(x_2), f(x_3)$ … という確率で取るもの

例．サイコロ

「出てくる値は{1,2,3,4,5,6}、それぞれの確率は $f(1)=f(2)=f(3)=f(4)=f(5)=f(6)=\frac{1}{6}$」

こういう挙動すべてを含めて「確率変数」と言う

X のとる値とその確率の間の関係式を確率関数／確率分布 f(x)という

サイコロの場合、$f(x)=\frac{1}{6}$という確率関数（確率分布）が成り立つ

X が出し得る値$\{x_1, x_2, x_3, \cdots\}$の特徴について

① 母平均＝期待値：X により出てくる値が大体どの辺りか

$$E(X) = \sum_i x_i f(x_i)$$ ※母平均のことは E(X)もしくは μ と書く：平均 m のギリシャ文字

母平均の性質（証明は講義ノート p.16 参照）：母平均は普通に加減乗除して良い

(i) $E(c)=c$（c：定数）

(ii) $E(X+c)=E(X)+c$

(iii) $E(cX)=cE(X)$

(iv) $E(X+Y)=E(X)+E(Y)$

※ X+Y の意味

X の取り得る値 x_i と Y の取り得る値 y_j のすべての組み合わせ(x_i, y_j)について、

x_i+y_j を値とし、x_i と y_j が同時に成立する確率 $f(x_i+y_j)$を値 x_i+y_j の出現確率とする確率変数

例．サイコロ2つを投げたとき、X+Y=Zとすると、X+Yの取りうる値は2, 3, 4⋯, 12

それぞれの確率は$f(1,1)=f(1,2)=\cdots=f(6,6)=\frac{1}{36}$。$E(X+Y)=(1+1)\times f(1,1)+(1+2)f(1,2)+\cdots=7$ だが、$E(X+Y)=E(X)+E(Y)=7$ としても良い

② 母分散：X により出てくる値がどれぐらいばらついているか

$$V(X) = \sum_i \big(x_i - E(X)\big)^2 f(x_i) = E\big(X - E(X)\big)^2 = E(X^2) - \big(E(X)\big)^2$$

（$E(X^2) = \sum_i x_i^2 f(x_i^2)$）

※母分散のことは V(X)もしくは σ^2 と書く（σ は母標準偏差。標準偏差 s のギリシャ文字）

※ $E\big(X - E(X)\big)^2 = E(X^2) - \big(E(X)\big)^2$の証明：講義ノート p.18 参照（※E(X)を定数として扱っていることに注意）

母分散の性質（証明は講義ノート p.16 参照）

(i) $V(c)=0$

(ii) $V(X+c)=V(X)$

(iii) $V(cX)=c^2V(X)$ ←V は 2 乗の話なので c が 2 乗されていることに注意

(iv) $V(X+Y)=V(X)+V(Y)+2Cov(X,Y)$ ←V(X)+V(Y)に、2×Cov(X,Y)を足さなければいけない

※ Cov(X,Y)＝共分散(covarience)

$$Cov(X,Y)=E\{(X-E(X))(Y-E(Y))\}$$

共分散の意味：X と Y の間の関係が分かる

共分散が正：X の取る値 x が大きくなると、Y の取る値 y も大きくなる

共分散が負：X の取る値 x が大きくなると、Y の取る値 y は小さくなる

共分散が 0 に近い：X と Y にはあまり関係がない

例．株を買う時、安全策を取る人は共分散が負となるように、一攫千金を狙う人は共分散が正となるようにする

母平均・母分散は、平均・分散と発想は同じだが、X の世界か D の世界かという点で根本的に異なる

同様に、母歪度 $E\{(X-E(X))^3\}$・母尖度 $E\{(X-E(X))^4\}$も考えることができる。

母分散$=E(X^2)-(E(X))^2$、母歪度$=E(X^3)-3E(X)E(X^2)+2(E(X))^3$ のように、これらは $E(X^n)$の足し算の形に変形できる

$E(X^n)$を、X の(原点回りの)n 次のモーメントと呼ぶことにする

以下は試験範囲外

$1+E(X)t+\left\{\frac{E(X^2)}{2!}\right\}t^2+\left\{\frac{E(X^3)}{3!}\right\}t^3+\cdots$ という関数（モーメント母関数）を定めると、

テイラー展開によりモーメント母関数$=E(e^{tX})$という簡単な関数になるので、これを利用して

n 次のモーメント、さらには母平均・母分散…が求められて、X の特性がすべて分かる

さらに D の世界の分布関数と関連付けることで、データから X を推定できる：モーメント法

★数理統計学の目標＝①推定　②検定

①推定：データに対して特定の確率変数をモデルとしてあてはめることで、データの出てくるしくみを推測

②検定：ある仮説を特定の確率変数によってモデル化することで、その仮説がデータにどのくらいあてはまるかを(数量的に)判定

つまり、推定は「D→Xを推測する」、検定は「Xの仮説→Dに当てはめる」

★色々な確率分布

確率分布＝確率関数＝X(確率変数)と確率の間の関係

データの多くは、いくつかの有名な確率分布に従って分布する → それらを用いて推定・検定を行う

確率分布は X の世界の話だということに注意（取りうる値から外れることもある）

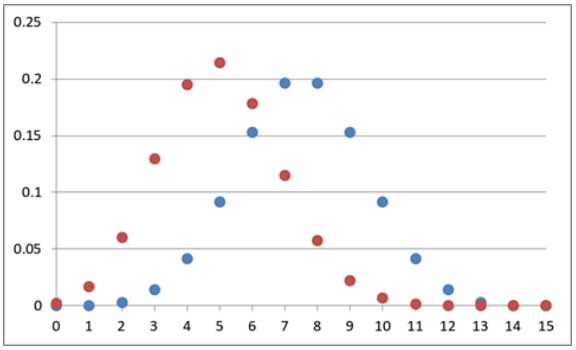
①二項分布

表/裏や赤/白など、2 つの値のどちらかを取るものが n 個ある、という話

一方の値を 1、もう一方の値を 0 とし、1 を取る確率を p、0 を取る確率を 1-p とする

n 個のうちちょうど x 個が 1 (n-x 個が 0)となる確率$={}_nC_x\,p^x(1-p)^{(n-x)}$　←これが二項分布の確率関数

横軸を x、縦軸を確率としてグラフを書くと、下のようになる（n=15、青は $p=\frac{1}{2}$、赤は $p=\frac{1}{3}$）



x は整数値しか取らないので連続でない：離散型確率変数

母平均$=np$、母分散$=np(1-p)$

（証明は意外と大変。気になる人は
http://www012.upp.so-net.ne.jp/doi/biostat/CT39/bin_and_poi_E_and_V.pdf 参照）

②正規分布：最も基本的な分布

その理由：中心極限定理によって、どんな複雑な分布でも正規分布に近似できるから

中心極限定理：母平均μ、母分散σ^2 の確率変数Z から出現する N 個の観測値の平均は、

漸近的に母平均μ、母分散$\frac{\sigma^2}{N}$の正規分布に従う。

（もう少し分かりやすく：

母平均 E(Z)= μ、母分散 V(Z)= σ^2 の確率変数 Z がある。　←X の世界

「Z に従う N 個のデータを観測し、その平均$\frac{1}{N}\sum_i z_i$ を求める。」…(＊)　←D の世界

(＊)を繰り返し行い、k 回目の試行で求められた平均を y_k とする。

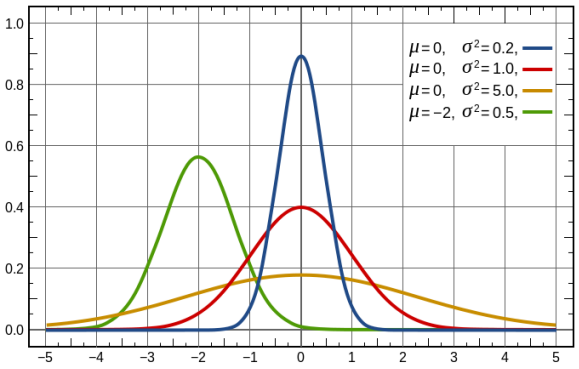
$y_1, y_2, y_3\cdots$ はそれぞれが確率変数 Y の観測値であると言える。　←X の世界

このとき、Z がどんな確率変数であっても、Y は正規分布に従って分布する。(証明はとても難しい)

Y の母平均 $E(Y) = E\left(\frac{1}{N}\sum_i z_i\right) = \mu$　　Y の母分散 $V(Y) = V\left(\frac{1}{N}\sum_i z_i\right) = \frac{\sigma^2}{N}$ ）

※ http://www.kwansei.ac.jp/hs/z90010/sugakuc/toukei/tyuusin2/chuusin.htm で実際にシミュレーションができます。

形：つりがね型



※ 連続型

正規分布は値が連続的に分布する連続型確率変数（⇔ 二項分布は値が飛び飛びの離散型確率変数）

連続型確率変数：-∞<x<∞の間で値を取り、幅 a<x<b になる確率 P(a<x<b) が決まっているもの

$P(a<x<b) = \int_a^b f(x)\,dx$ となるような f(x)を確率密度関数という

これは離散型で言う確率関数と同じ（離散型では、$P(a<x<b)=\sum_{i=a}^{b} f(x)$ ）※<か≦かは気にしなくて OK

連続型では、離散型の Σ を全部 ∫ に変えれば良い

母平均 $E(X) = \int x f(x) dx$　　※∫とは、$\int_{-\infty}^{\infty}$のこと

母分散 $V(X) = E\left(X - E(X)\right)^2 = \int \left(x - E(X)\right)^2 f(x)\,dx$

母平均 μ、母分散 σ^2 の正規分布の確率密度関数 $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

※exp(x)は e(自然対数の底)の x 乗という意味。つまり $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

中でも、母平均 μ=0、母分散 σ^2=1 である正規分布を標準正規分布という（上図の赤いグラフ）

標準正規分布において、値がある区間内に入る確率は、標準正規分布表にまとめられている

『統計学入門』であれば、標準正規分布で u 以上の値を取る確率（$\int_u^{\infty} f(x)dx$）が p.280 にまとめられている

すべての正規分布は、標準正規分布に変換することで確率が求められる

母平均 μ を 0 にし、母分散 σ^2 を 1 にすればいいので、$\frac{値-\mu}{\sigma}$ となる確率を求めれば良い

(逆に言えば、標準正規分布の値を u とすると、元の正規分布の値＝μ+uσ → u は値が母平均から σ 何個分離れているかを表す)

なお正規分布は左右対称なので、値-μ が負の場合は$\frac{|値-\mu|}{\sigma}$の確率を求めた後 1 から引けば良い

例．母平均 0.5、母分散 0.4 の正規分布に従う確率変数が 0.3～0.6 に入る確率は、

$\frac{0.3-0.5}{0.4} = -0.5$ → 確率 $1 - 0.309 = 0.691$　　$\frac{0.6-0.5}{0.4} = 0.25$ → 確率 0.401　　$0.691 - 0.401 = 0.290$

③ポアソン分布

ある観測単位（例．1 時間あたり、1km あたり…）の中で n 回の試行が行われたとき、ある事象が起こる確率が p であるとき、その観測単位の中でその事象が x 回起きる確率は、

確率関数 $f(x)=\frac{e^{-\lambda}\lambda^{x}}{x!}$ $(\text{ただし}\ \lambda=np)$ に近似的に従って分布する（ポアソン分布は離散型）

つまり、観測単位の中で事象が平均 λ 回発生するとき、ちょうど x 回(x=0,1,2…)発生する確率の分布

これは二項分布の $f(x)={}_nC_x\,p^x(1-p)^{(n-x)}$でも求めることができるが、n が大きいときは ${}_nC_x$ や$(1-p)^{n-x}$ の計算が大変

→ よって、p が小さい時はポアソン分布で近似する（p が大きい時は正規分布で近似する：本シケプリ p.8, p.10 参照）

また、n や p が分かっていなくても、λ=np（=平均の発生回数）が分かっていれば分布が分かる

例．30 年に 1 回巨大地震が起きることが分かっていれば、すべての地震が起こる回数 n が何回か分からなくても良い

※ ポアソン分布については http://www.ntrand.com/jp/poisson-distribution/の説明が分かりやすい

④ t 分布

正規分布に従う確率変数 X の観測値が n 個ある。n が十分大きいと観測値は正規分布に従うが、

n が小さいと、正規分布と少し異なる「t 分布」に従う。

（具体的には、n が小さいと、母平均から外れた値が出る確率が正規分布より少し高くなる）

確率変数 X の母平均を μ とし、n 個の観測値の標本平均を m、標本分散を s^2 とすると、

$\frac{m-\mu}{\frac{s}{\sqrt{n-1}}}$ が自由度 n－1 の t 分布に従う（＝t 統計量こそが確率変数であり、これは t 分布に従う）

これを t 統計量と呼ぶ。こういう計算をするのは、正規分布を $\frac{値-\mu}{\sigma}$ によって標準正規分布に変換するのと同じ理由

※ 自由度：自由に動ける変数の数

今回の場合、n 個の観測値により μ を決める。n-1 個の観測値を決めると残り 1 つは自動的に決まってしまうので、自由度は n-1 となる。

（自由度の分かりやすい説明、と言ってもこれは不偏分散の説明だけど：

http://homepage2.nifty.com/nandemoarchive/toukei_hosoku/jiyudo.htm）

t 分布は、正規分布に従うとされるデータの、母平均 μ を推定・検定したい時に使われる

ただし、実際には正規分布と少しずれていてもきちんと推定・検定できる（頑健性がある）

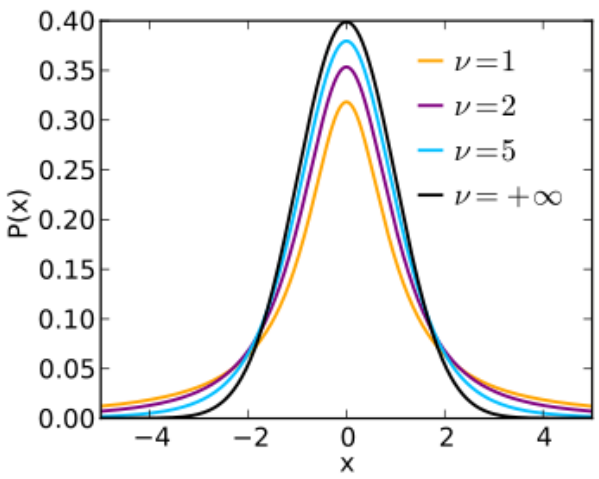
※ 頑健性：仮定されたある条件を満たしていなくても、ほぼ妥当な結果が得られること

⑤カイ二乗(χ^2)分布

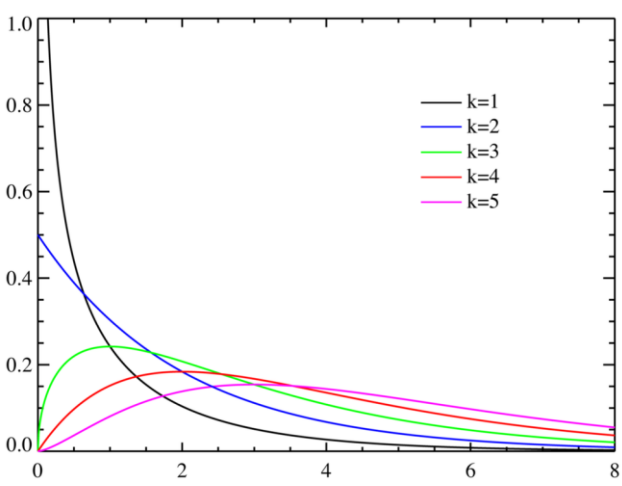
標準正規分布に従う k 個の確率変数 $X_1, X_2, X_3 \cdots X_k$ があるとき、$Y=\sum_i {X_i}^2$ を求めると、Y は新たな確率変数である。Y の確率分布を「自由度 k のカイ二乗分布」という（これは習ってないので覚えなくて良い）

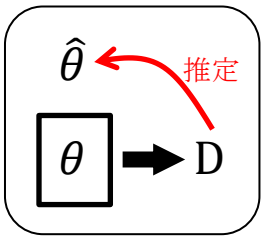
カイ二乗分布は、母分布の検定（適合度検定・独立性検定）や母分散の推定・検定などに使われる

自由度 ν の t 分布（自由度∞は正規分布と一致）



自由度 k のカイ二乗分布





★推定

(統計的)推定：観察されたデータから確率変数の理論値を推測すること

その際、サンプル調査(標本調査)を行う：標本から母集団の特徴・性質を推定する方法

＝D の世界を元に X の世界を推定する

母集団：本来調査したい対象　　標本：母集団から無作為に抽出された集合　サンプル規模：標本の大きさ

推定値は $\hat{\theta}$ のように、理論値 θ の上に^(ハッシュ)を付ける

推定では、推定値が理論値と一致する保証はないが、どの程度の精度なのかを数値的に評価できるのが特徴

例として、視聴率調査を挙げる

母視聴率 p を求めたい。日本全国の全世帯(=母集団)の調査はできないので、n 世帯(=サンプル規模 n の標本)のみ調査をし、そのうち x 世帯が番組を視聴しているとき、p を推定する(=$\hat{p}$を求める)。

このとき、直感的に考えれば視聴率は $\frac{x}{n}$ と考えられる…(1)。これが推定値。が、本当にそれでよいのか？

推定の方法として、点推定と区間推定がある

点推定：推定値として 1 つの特定の値を求める

区間推定：推定値としてある区間(幅)を求め、その区間に入る確率も求める

点推定

視聴率は「見る/見ない」の 2 択なので、二項分布に従う

n 世帯中 x 世帯が見ている確率は、母視聴率 p とすると $f(x)={}_nC_x\, p^x(1-p)^{n-x}$ …(2)

最尤推定法を使用する

最尤推定法：x を定数、p を変数と見たとき、f(x)が最大になるような(=確率が最大となるような)

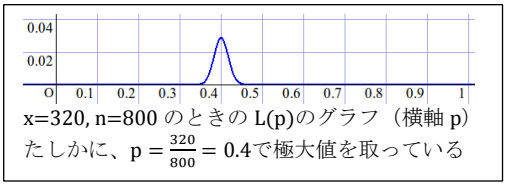
p を求めれば、現実の世界に最も近い(＝最も尤(もっと)もらしい)推定値になるはず！

そこで②の式を、x を定数、p を変数として見た尤度関数 L(p)を導入する：$L(p)={}_nC_x\, p^x(1-p)^{n-x}$

L(p)の最大値は、p を微分して L(p)が極大値を取るような p を求めれば良い。

最尤推定法により求められた結果は $p=\frac{x}{n}$ となる。これを最尤推定値という。

ゆえに、数学的に(1)が正しいことが証明された。



x=320, n=800 のときの L(p)のグラフ（横軸 p）
たしかに、$p=\frac{320}{800}=0.4$で極大値を取っている

なお、推定値として考えられるものは最尤推定値以外にもある

良い推定値が持つべき様々な性質

・不偏性：$E(\hat{\theta})=\theta$（推定値 $\hat{\theta}$ の母平均が理論値 θ と同じ）

例.　推定値：標本平均 m は、理論値：母平均 μ の不偏推定値

推定値：$\frac{1}{n-1}\sum_i(x_i-m)^2$は、理論値：母分散 σ^2 の不偏推定値

・漸近不偏性：$\lim_{n\to\infty} E(\hat{\theta})=\theta$（サンプル規模 n が大きくなるにつれて$\hat{\theta}$ の母平均が θ に近づく）

例. 最尤推定値は漸近不偏性を持つ

・一致性：$\lim_{n\to\infty} P\left|\hat{\theta}-\theta\right|=0$（n が大きくなるにつれて、$\hat{\theta}$=θ となる確率が 1 に近づく・収束する）

例. 最尤推定値は一致性を持つ

推定値：$\frac{1}{n}\sum_i(x_i-m)^2$は、理論値：母分散 σ^2 の不偏推定値（正規分布などの場合）

・有効性：不偏推定値のうち、分散 $V(\hat{\theta})=E\left(\hat{\theta}-E(\hat{\theta})\right)^2$ が最小。$E(\hat{\theta})=\theta$ より、$E(\hat{\theta}-\theta)^2$が最小

・平均平方誤差が最小：$E(\hat{\theta}-\theta)^2$が最小（不偏推定値ではないがこれが成り立つ場合もある）

例. 推定値：$\frac{1}{n+1}\sum_i(x_i-m)^2$は、理論値：母分散 σ^2 の平均平方誤差最小推定値（正規分布の場合）

推定値：$\frac{x+1}{n+2}$ は理論値：母比率（視聴率など）p の平均平方誤差最小推定値（p≒0.5 のとき）

例えば内閣を支持するか 3 人に聞いて 3 人とも no と言ったとき、支持率 0%より $\frac{0+1}{3+2}$=20%の方がもっともらしい

最尤推定値は、漸近不偏性・漸近一致性・有効性を持つ（長所）が、頑健性がない（短所）

※頑健性がない ＝ 推定値の導出に使用した確率分布(=確率関数=尤度関数)と実際の確率分布が、少しずれていれば推定値が成り立たない

区間推定

a≦θ≦b に入る確率は α、という風に求まる（長所：確率＝精度を数値的に評価できること）

このとき、a～b を信頼区間、α を信頼係数という

・母比率 p を推定 → 二項分布に従うときは、正規分布で近似する

問. n=800 世帯で視聴率調査を行うと、x=320 世帯が番組を見ていた。信頼係数 0.68 の信頼区間を求めよ。

（ノートの p.43 で扱った問題のやり方 2-1 を少し言い換えた）

標本による比率 $\frac{x}{n}=\frac{320}{800}=0.4$。これが a~b に入る確率は 0.68

中心極限定理を使う場合は繰り返し観測しなければいけない→「ある 1 世帯を観測する」というのを n 回繰り返す（本シケプリ p.10 も参照）

ある 1 世帯が番組を見ていたかどうかは二項分布に従う。母平均 p、母分散 p(1-p)

→ 中心極限定理より、これは母平均 μ=p、母分散 $\sigma^2=\frac{p(1-p)}{n}$ の正規分布に近似できる

※『統計学入門』の標準正規分布表には、u の欄に値が u 以上となる確率が書かれている

u 以上となる確率は $\frac{1-0.68}{2}=0.16$ より、対応する u を見ると u=1.0

標準正規分布で値 u →元の正規分布で値 μ+uσ

つまり標準正規分布では信頼区間-1.0~1.0 となる→元の正規分布では信頼区間 p-σ ~ p+σ となる

$p-\sigma\leqq\frac{x}{n}=0.4\leqq p+\sigma \quad\Leftrightarrow\quad \frac{x}{n}-\sigma\leqq p\leqq\frac{x}{n}+\sigma\cdots①$ ←求めたいのは μ の範囲なので移項

ここで、母標準偏差 $\sigma=\sqrt{\frac{p(1-p)}{n}}$ の値は分からないが、$p=0.4$ と仮定すると $\sigma=\sqrt{\frac{0.4(1-0.4)}{800}}=0.017$ と求まる

①に代入すると、0.383≦p≦0.417 …(答)

問題点：勝手に p=0.4 と仮定してしまっている（推定の範囲を出す前に p を推定してしまっている）：Wald 型

↓これを改善した方法：ベイズ推定（ノートのやり方 2-2）

$p-\sigma\leqq\frac{x}{n}=0.4\leqq p+\sigma \quad\Leftrightarrow\quad -\sigma\leqq\frac{x}{n}-p\leqq\sigma \quad\therefore\quad \left(\frac{x}{n}-p\right)^2\leqq\sigma^2=\frac{p(1-p)}{n}$

あとは、p についての 2 次不等式を解けば良い → 0.383≦p≦0.417 …(答) ←答えは Wald 型にほぼ一致

・母平均 μ を推定 → 正規分布に従う時は、t 分布で近似する（頑健性があるので、少し正規分布とずれていても問題ない）

問. 正規分布に従う確率変数による 10 個の観測値「1.4 1.8 0.9 2.1 1.7 1.8 1.5 1.7 1.8 1.9」がある。この母平均の信頼係数 95%の信頼区間は？

標本平均 $m=\frac{1}{n}\sum_i x_i=1.66$ 標本分散 $s^2=\frac{1}{n}\sum_i(x_i-m)^2=0.0984$ 標準偏差 $s=\sqrt{0.0984}=0.3137$

t 統計量 $\frac{m-\mu}{\frac{s}{\sqrt{n-1}}}$ が自由度 n-1＝9 の t 分布に従う

※『統計学入門』の t 分布表の(ν,α)の欄に書かれている値は、自由度 ν でその値以上になる確率が α ということを表す

t 分布は左右対称 → 上下を切り捨てて 95%になれば良いので、$\alpha=\frac{1-0.95}{2}=0.025$ の欄を見れば良い → 2.262

t 統計量が-2.262~2.262 の範囲に入る→ $-2.262\leqq\frac{m-\mu}{s/\sqrt{9}}\leqq 2.262$ → μ について解くと、1.24≦μ≦2.08…(答)

・母分散 σ² を推定 → 正規分布に従うときは、カイ二乗分布で近似する

母分散をカイ二乗分布で推定・検定するときの検定統計量は $\frac{ns^2}{\sigma^2}$（検定統計量の使い方は本シケプリ p.10 で詳述）

※ 検定統計量：分布についての色々な性質（m, s, n, μ, σ など）から計算される量で、
それがある確率分布（t 分布とかカイ二乗分布とか）に従うというもの

例. 母平均を推定・検定するとき… t 分布の検定統計量：t 統計量 $\frac{m-\mu}{\frac{s}{\sqrt{n-1}}}$：これが t 分布に従う

問. 上の 10 個の観測値の母分散を信頼係数 90%で区間推定せよ。

※『統計学入門』のカイ二乗分布表(ν,α)の欄に書かれている値は、自由度 ν でその値以上になる確率が α ということを表す

カイ二乗分布は左右対称ではない → 上下を切り捨てて 90%になれば良いので、α=0.95 の欄と α=0.05 の欄を見れば良い

カイ二乗分布の母分散についての検定統計量 $\frac{ns^2}{\sigma^2}$ が 3.325~16.919 の範囲に入る

→ $3.325\leqq\frac{ns^2}{\sigma^2}\leqq 16.919$ → n=10, s²=0.0984 を代入して σ² について解くと、0.0582≦σ²≦0.296…(答)

★(仮説)検定：仮説(H_0)を特定の確率変数に当てはめることで、データ(D)とどの程度適合しているかを調べること

検定を行う際には、危険率・検出力を考慮して区間を定める必要がある

例．H_0：母視聴率 40%　D：200 世帯中△△世帯が見ていた（サンプル視聴率▽▽%）→H_0 はあり得るか？

※母視聴率＝視聴率の理論値＝本当の視聴率

まず、H_0 が合っていると仮定したときに、D として出てきそうな値の区間を設定する

母視聴率 40%なら D のサンプル視聴率は 36%~44%（世帯数 72~88 世帯）ぐらいなら出てくると考えてみよう

D のサンプル視聴率が 36%~44%の外なら仮説は×、区間内なら仮説は×ではない（○の可能性がある）、と判定する

※検定は、仮説が×か×でないかを判定するもの　→「棄却される」という

H_0 が合っているときに、サンプル視聴率 36%~44%に入る確率は$\sum_{72}^{88} {}_{200}C_x(0.4)^x(0.6)^{1-x} = 0.75$

＝確率 0.25 で、H_0 が合っているのに×と判定されてしまう：この確率を危険率(有意水準)α と言う（α=0.25）

区間を広く取ると危険率は下がる（例．34%~46%と取ると危険率 α=0.1）

検出力 1-β：講義ノートの p.53~p.54 が分かりやすいのでそこを見て下さい。ここでは省略（書くのめんどい）

まとめると、

	本当は○	本当は×
判定は×ではない	検定は合っている	第二種の誤り β
判定は×	第一種の誤り α	検定は合っている

危険率 α は低い方が良く、検出力 1-β は高い方が良い

検定では最初に危険率 α の最大値を定めておく。（普通は、出題のされ方が「この仮説を危険率 0.1 で検定せよ。」という形）

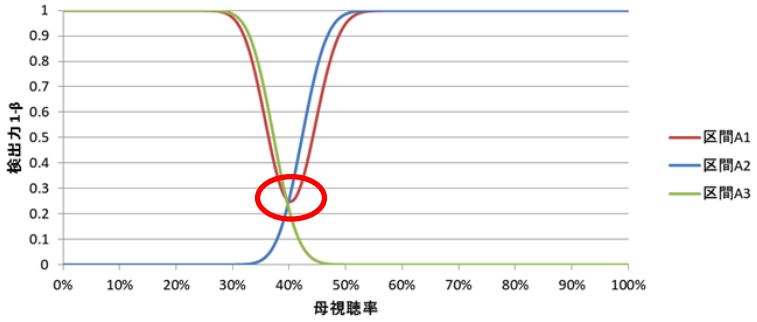
→ 調べる範囲全てで危険率が α 以下で、検出力が最大になるような区間を設定すれば良い

※調べる範囲＝母視聴率として取りうる値。絶対に母視聴率が 50%以下だと分かっていたら、調べる範囲は 0%~50%

理論値(本当の値、今回であれば母視聴率)によって、どの区間を使うと一番検出力が高いかは変わってくる

最強力：ある理論値で一番検出力が高い。一様最強力：調べる範囲の値全てで一番検出力が高い

→ 検出力関数：母視聴率を定義域とする関数（グラフは下の通り）



母視聴率<40%なら区間 A3 (37%~100%)が、母視聴率>40%なら区間 A2(0%~42%)が最強力

区間 A1(36%~44%)はバランス型（全体的に悪くないが局所的に見ると A2, A3 の方が優れている）

→ 一様最強力な検定は A1～A3 には存在しない

不偏検定：調べる範囲全てで検出力≧危険率（1-β≧α）

※上の図を見ると、A1 は不偏検定だが、A2 は 0%~40%、A3 は 40%~100%で検出力<0.25 となっている

一様最強力不偏検定：不偏検定の中では一様最強力（今回は、調べる範囲が 0%~100%だと A1、0%~40%だと A3、など）

調べる範囲が決まっていない場合（母視聴率の見当が付かない場合）：両側検定 → A1 を使う

調べる範囲が仮説以上か以下か決まっている場合：片側検定 → A2 や A3 を使う

※ 片側検定を使う例．汚染物質が基準値より多いか：少ない場合はどうでもいいので範囲から外す。多い場合は高精度で検定したい

実は検出力関数は、検出力と同時に危険率も表している

理論値が仮説に近い場合（今回なら、本当の母視聴率が仮説：母視聴率 40%に近い）、検出力≒危険率（上グラフの○部分。説明はノート p.70）

→検出力関数の形状を見れば、検出力と危険率の関係が分かる → その区間の決め方・検定方法だとどういう判断をして良いのかが分かる

また、理論値＝仮説（母視聴率＝40%）のとき検出力＝危険率となるので、どんなに良い検定方法でも検出力は理論値付近で落ち込む

逆に言えば、良い検定方法：検出力関数が仮説付近で急激に落ち込むもの

仮説付近で大きく凹む＝今回ならば、母視聴率 39%でも仮説が棄却されない → だが、仮説に近いので、仮説は○と判断して問題ない

仮説付近であまり凹まない時は、仮説が棄却されなくても安易に「仮説は○」と判断してはいけない

仮説付近で大きく凹むための条件：危険率が低い・サンプル規模が大きい・(適合度の検定では)カテゴリ k の数が少ない、など

検定の方法：　①仮説を立てる　②区間を定める　③特定の確率分布に従う検定統計量を求める
　　④検定統計量が②の区間に入っている否かで検定する

ある種類の検定をするときに、どの検定統計量（本シケプリ p.8 参照）・確率分布を使うかは決まっている

・母比率 p の検定 …（ランダムに選ぶ場合）検定統計量 $\frac{x}{n}$、確率分布：二項分布 or 正規分布/ポアソン分布に近似

問. A さんはトランプのマークが黒 or 赤を当てる能力があるという。52 枚のトランプを当てさせたところ 40 枚が当たった。A さんの能力はあると言えるか、危険率 0.05 で検定せよ。

H_0：正解する確率は $\frac{1}{2}$（①）←能力があった場合棄却される仮説を立てる（検定は×か×でないかを判定するから）

D：52 枚中 40 枚正解

正解 or 不正解の 2 択なので、二項分布に従う。H_0 が正しいとすると、$p=\frac{1}{2}$

ある 1 枚を取り出すとき、それが正解 or 不正解（正解の枚数=0 枚か 1 枚か）は二項分布に従う（n=1, x=0or1）

※ n=52 としない理由：中心極限定理は「繰り返し観測」しなければいけない →「ある 1 枚を観測する」ことを 52 回繰り返す

$f(x)={}_1C_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{1-x}$　母平均 $\frac{1}{2}$　母分散 $\frac{1}{2}\left(1-\frac{1}{2}\right)=\frac{1}{4}$　の二項分布

中心極限定理より、これは母平均 $\mu=\frac{1}{2}$、母分散 $\sigma^2=\frac{1}{4}\cdot\frac{1}{52}$ の正規分布に近似できる（$\sigma=0.0693$）

特に条件がないので両側検定。危険率 0.05 より、標準正規分布での区間は-1.96~1.96（②）

（『統計学入門』では、確率 $\frac{0.05}{2}=0.025$、正確には 0.24998 に対応する値を見ると 1.96 と書いてある）

40 枚/52 枚正解したので、正規分布で取る値は $\frac{40}{52}=0.769$。標準正規分布に直すと、$\frac{0.769-0.5}{0.0693}=3.882$（③）

↘ ある 1 枚について見ているので 52 で割る。これが検定統計量

3.882 は②の区間に入らないので、H_0 は棄却される。（④）よって、A さんは能力があると言える。

ちょっと変わった例. これは D が最初から区間になってるし、82 年繰り返したりするから、「いわゆる検定」とはちょっと違う

ロンドンの年間出生数を 82 年間観察すると、年間平均 11429 人、うち男子の比率は 0.5027~0.5362 だった（男女半々より男子が多い）。このとき、「男子と女子が生まれる確率は等しい」という仮説は成り立つか？

H_0：男子が多い確率＝女子が多い確率＝0.5　　D：男子の比率 0.5027~0.5362

「男子が多い」or「女子が多い」の 2 択なので、二項分布に従う。H_0 が正しいとすると、$p=\frac{1}{2}$

まずは、ある 1 年だけを見て D になる確率を求める → 82 年連続でそうなる＝確率を 82 乗すれば良い

（※ 一つの方法は、二項分布を直接計算する方法

男子の年間出生数の平均は(11429×0.5027=)5745 人～(11429×0.5362=)6128 人

D が出現する確率は、n=11429 のうちちょうど 5745, 5746,…6128 回「男子」が出る確率

$=\sum_{x=5745}^{6128}{}_{11429}C_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{11429-x}$：この計算はできなくもないが計算量がすごい（計算すると 0.287）

↙ 講義では計算間違ってた）

計算量を減らすために、中心極限定理を使って正規分布で近似する

ある 1 人だけを見て、その人が男子か女子か（=男子が 0 人か 1 人か）は二項分布に従う（n=1, x=0 or 1）

$f(x)={}_1C_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{1-x}$　母平均 $\frac{1}{2}$　母分散 $\frac{1}{2}\left(1-\frac{1}{2}\right)=\frac{1}{4}$　の二項分布

中心極限定理より、これは母平均 $\mu=\frac{1}{2}$、母分散 $\sigma^2=\frac{1}{4}\cdot\frac{1}{11429}$ の正規分布に近似できる（$\sigma=0.004677$）

これが 0.5027~0.5362 の値を取る → 標準正規分布に直すと、$\frac{0.5027-0.5}{0.004677}$～$\frac{0.5362-0.5}{0.004677}$、つまり 0.577～7.74

標準正規分布表より、$0.282-0=0.282$：二項分布での計算と割と一致　　講義では計算間違ってた ↗

1 年だけ見て D になる確率が 0.282 → 82 年連続で D になるのは $(0.282)^{82}=8.32\times10^{-46}$：確率ほぼゼロ

→ よって仮説は棄却される…（答）

・母平均 μ の検定 …（観測値が正規分布に従って分布する場合）検定統計量 $\frac{m-\mu}{\frac{s}{\sqrt{n-1}}}$、確率分布：自由度 n-1 の t 分布

例は挙げないが、やり方は母比率 p の検定と同じ。

たとえば危険率 0.1 で両側検定せよ、と言われたら、t 分布表で 2α=0.1 の列を見れば良い

・母分散 σ^2 の検定 …（観測値が正規分布に従って分布する場合）検定統計量 $\frac{ns^2}{\sigma^2}$、確率分布：カイ二乗分布

例は挙げないが、やり方は母比率 p の検定と同じ。たとえば危険率 0.1 で両側検定せよ、と言われたら、α=0.95 と α=0.05 の欄を見れば良い（カイ二乗分布は左右対称ではないため）

・母分布の検定 …（f_i が多次元正規分布に従って分布する場合）検定統計量 $\sum_{i=1}^{k}\frac{(f_i-np_i)^2}{np_i}$ 、確率分布：カイ二乗分布

※多次元正規分布（試験範囲外）：正規分布の拡張版。例えばサンプルを無作為に抽出すると多項分布（二項分布の拡張版）に従う
→中心極限定理の拡張版により、漸近的に多次元正規分布に従う

k 個のカテゴリに分かれている（カテゴリの例．サイコロの出目が 1 のカテゴリ、2 のカテゴリ… → k=6）

理論上（X の世界）では、それぞれの値の出現確率が $p_1, p_2, p_3 \cdots p_k$

実際（D の世界）に n 個の値を観測したところ、それぞれの値の出現個数が $f_1, f_2, f_3 \cdots f_k$ だった

検定統計量は、観測値の個数（観測度数）=O、理論値の個数（期待度数）=E とすると、$\sum_i \frac{(O_i-E_i)^2}{E_i}$ と言い換えられる

1. このとき、D の世界の分布が X の世界の分布に適合しているかを検定できる（適合度の検定）

検定統計量は、自由度 k-1 のカイ二乗分布に従う。

問．メンデルの仮説「豆の黄：緑と丸い：皺々はそれぞれ 3:1 で出現し、相互が独立」

実際にとれた豆を見ると、全部で n=556 個中、黄丸 315、黄皺 101 個、緑丸 108 個、緑皺 32 個だった。

とれた豆は無作為に出現したとして、メンデルの仮説を検定せよ。

H_0：D の世界の分布は X の世界の分布に適合する　　D：D の世界の分布

カテゴリの個数 k=4 → 検定統計量は自由度 3 のカイ二乗分布に従う

$p_i = \left\{\frac{3}{4}\cdot\frac{3}{4}, \frac{3}{4}\cdot\frac{1}{4}, \frac{1}{4}\cdot\frac{3}{4}, \frac{1}{4}\cdot\frac{1}{4}\right\} = \left\{\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right\}$　　$f_i = \{315, 101, 108, 32\}$　→ 検定統計量＝0.470

母分布に関する検定は、検定統計量が正の値しか取らないので片側検定（例外はあるが本講義では扱わない）

自由度 3 の欄を見ていくと、危険率 0.9 以下では（0.2 でも 0.5 でも 0.9 でも）棄却されない

※危険率 0.9 とは「実際は正しいのに 90%の確率で棄却する」という厳しい基準だが、それでも棄却されない
だが、今回の H_0 は肯定したい仮定なので、高い危険率でも棄却されなくてようやく厳しい検定を通過したと言える
通常の検定の H_0 は否定したい仮定なので、危険率が低い＝検出力が低くても棄却されないと厳しい検定を通過したと言える

高い危険率でも仮説は棄却されないので、適合していると言える。…（答）（※適合しすぎなので捏造説もある）

2. D の世界で観測された 2 つの確率変数 A, B が独立かどうかを検定できる（独立性の検定）

問．男女に死後の世界を信じるか否かのアンケートを取ると、下表のようになった。性別と信じるか否かは独立か？危険率 0.05 で検定せよ。

A ＼ B	信じる(j=1)	信じない(j=2)	合計
男性(i=1)	350	100	450
女性(i=2)	400	150	550
合計	750	250	n=1000

f_{ij}＝クロス集計（i 行 j 列の値）とする
また、 $f_{1.}$=1 行目の合計=$f_{11}+f_{12}$=450
$f_{.2}$=2 列目の合計=$f_{12}+f_{22}$=250
などとする

独立であれば、X の世界の分布について、$p_{ij} = p_{i.}p_{.j}$ が成り立つ（数学 A でやった）

ここで、$p_{i.}$ の最尤推定値は $\frac{f_{i.}}{n}$、$p_{.j}$ の最尤推定値は $\frac{f_{.j}}{n}$ なので（母比率の推定）、$p_{ij} = \frac{f_{i.}}{n}\frac{f_{.j}}{n} = \frac{f_{i.}f_{.j}}{n^2}$ より $np_{ij} = \frac{f_{i.}f_{.j}}{n}$

よって、検定統計量は $\sum_i\sum_j \frac{\left(f_{ij}-\frac{f_{i.}f_{.j}}{n}\right)^2}{\frac{f_{i.}f_{.j}}{n}}$ となる。これは、自由度$(i-1)(j-1)$のカイ二乗分布に従う

※なぜ自由度(n-1)(j-1)なのか：各行・列の合計（$f_{i.}, f_{.j}$）は始めから決まっている。例えば今回の問なら、f_{11}=350 と決めた時点で、f_{12}=100, f_{21}=400 が合計から自動的に求められ、ということは f_{22}=150 と決まるので、自由に決められる変数は f_{11} のみ、つまり自由度は 1。

H_0：A と B は独立（すなわち $np_{ij} = \frac{f_{i.}f_{.j}}{n}$ が成立）　　D：確率変数 A, B の分布

検定統計量＝$\frac{\left(350-\frac{450\cdot750}{1000}\right)^2}{\frac{450\cdot750}{1000}} + \frac{\left(100-\frac{450\cdot250}{1000}\right)^2}{\frac{450\cdot250}{1000}} + \frac{\left(400-\frac{550\cdot750}{1000}\right)^2}{\frac{550\cdot750}{1000}} + \frac{\left(150-\frac{550\cdot250}{1000}\right)^2}{\frac{550\cdot250}{1000}} = 3.37$

これが自由度(2-1)(2-1)=1 のカイ二乗分布に従う。片側検定なので α=0.05 の欄を見ると 3.84

よって仮説は棄却されないので、独立でないとはいえない。…（答）