E3 確率統計

Mathematica では事前に、<<Statistics`Master` とパッケージを読みこんでおく。

【乱数】

確率の理論を実証するためには、極めて多くの乱数を必要とする。コインを繰り返し投げるとか、サイコロを多数回振るのは大変なので、パソコンにより擬似乱数を発生させて、それを利用することが多い。

１から６までの数をランダムに 30 個発生させる。
０から９９までのランダム整数を 20 個ずつ５組発生させる。

コインを１００回投げて表と裏が出た様子を記録する。このとき、表または裏が連続して６回続くことが珍しいことであるかどうか調べる。

【確率の計算】

事象 A の起こる確率を P(A) で表すとき、

加法定理 P(A∪B)＝P(A)+P(B)－P(A∩B)

乗法定理（条件付き確率） P(A∩B)=P(A)P(B | A) (P(A)＞０とする)

余事象の確率 P(Ā)＝１－P(A)

袋の中に赤玉１０個、白玉１０個があり、それぞれ１から１０まで番号が付けられている。

この袋の中から１球を取出すとき、赤玉であるか, 番号が１, ２, ３のいずれかである確率を求めよ。

この袋の中から、１個ずつ３球を取出し、赤玉、白玉、赤玉の順に出る確率を、復元抽出の場合と、非復元抽出の場合について求めよ。

２個のサイコロを振って, 何回のうちに少なくとも１回は６のゾロ目（２個とも６の目） が出るかという回数を当てる遊びがある。勝つ確率が 1/2 を越えるためには何回以上にかける必要があるか。（メレの問題）

ベイズの定理 互いに排反な事象 $A_1, A_2, ..., A_n$ のどれかが必ず起こり、すべての k について P(A_k)＞０とする。P(B)＞０である事象 B について

$$P(A_k \mid B) = \frac{P(A_k)P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)} \quad (k = 1, 2, ..., n)$$

A,B,C 社から同一種の製品を２：３：５の比で購入している工場がある。A,B,C 社の製品にはそれぞれ 2.5%, 1.5%, 1% の割合で不良品が含まれていることが分かっている。いま、これらの製品の中から任意に１個を抽出したところ、それが不良品である確率を求めよ。抽出した不良品が A,B,C 社の製品である確率をそれぞれ求めよ。

【離散確率分布】

２個のサイコロを振ったとき、出る目の和を X とする。X の確率分布 $P(X=k)=p_k \quad (k=2,3,...,12)$ の表を作り、そのグラフを描け。

平均あるいは期待値 E[X], 分散 V[X], 標準偏差 σ[X] を計算せよ。

次の各分布関数について、そのグラフを描くこと。このとき、定義式を使って描く方法と、組み込み関数を使う方法を試みよ。 サンプルのデータの他に、いろいろ替えてみるとよい。

また、各分布関数について、平均と分散を計算し、記述のようになることを確認すること。

このときも、定義式を使って計算する方法と、組み込み関数を使う方法を試みよ。

２項分布 binary distribution のグラフを描くこと。

ある事象Ａの起こる確率ｐが与えられているとき、ｎ回独立試行を行ってＡがｘ回起こる確率

$$f(x)={}_nC_x\,p^x(1-p)^{n-x} \quad (x=0,1,2,...,n)$$

平均 $\mu=np$　　分散 $\sigma^2=np(1-p)$

例えば　p = 1/6 とし、n = 10, 20, 30, 40, 50

幾何分布

確率を p とした「当たり」が始めて起こるまで何回試行が繰り返されたか、その回数の分布を表す

$$f(x)=p(1-p)^x \quad (x=0,1,2,...)$$

平均　$\dfrac{1-p}{p}$　　分散　$\dfrac{1-p}{p^2}$

<<Statistics`Master` のとき GeometricDistribution[p] が使える。

ポアソン（Poisson）分布のグラフを描くこと。

２項分布でｐが小さい（滅多に起こらない事象）のとき

$$f(x)=\frac{\mu^x}{x!}e^{-\mu} \quad (\mu=np)$$

平均 μ　　分散　$\sigma^2 = \mu$

例えば、μ = 2, 3 とする。

【チェビシェフの不等式】

任意の確率変数 X について、その平均を μ、標準偏差を σ とすると、任意の正数 k について $P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$ が成り立つ。

【大数の法則】

１回の試行で事象Ａの起こる確率が p である試行を、独立に n 回行うとき、Ａの起こる回数を X とするとき、X / n はＡの起こる相対度数である。n を大きくすると | X / n – p | は 0 に近づく。

このことを、プレゼンテーションで示すプログラムを作れ。

【連続確率分布】

X の確率密度関数が $f(x) = \begin{cases} k(1-x^2) & (|x| \leq 1) \\ 0 & (|x| > 1) \end{cases}$ で与えられるとき、k の値を定め、確率 P(-0.5<x<0.5) を求めよ。

また、平均 E[X], 分散 V[X], 標準偏差 σ[X] を計算せよ。

正規分布 normal distribution $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ のグラフを描くこと。

E[X]=μ　V[X]=σ^2 を確認すること。この確率分布を $N(\mu, \sigma^2)$ で表わす。

N(0, 1) は**標準正規分布**という。

２項分布　$f(x) = {}_nC_x p^x (1-p)^{n-x}$ で、ｎを大きくすると、正規分布に近づくことをグラフを描いて説明すること。

例えば、p = 0.2 の場合で、n = 10,20,30 の場合で、N(np, npq) (q = 1-p) と比較する。

逆に、n = 20 を固定し、p を 0.01 から 0.99 まで 0.02 づつ増やしながら、二項分布 B(n, p) の折れ線グラフと、同じ平均 np, 分散 np(1-p) を持つ正規分布のグラフを表示させ、そのときの p の値をｙ軸上の点として、また、そのときの平均と標準偏差をｘ軸上の区間として同時に表示させる。これらのグラフをアニメーションとして見ること。

$N(\mu,\sigma^2)$で$Z=\dfrac{X-\mu}{\sigma}$ とおくとき、Zの分布関数は $g(z)=\dfrac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$ で あり、この操作を**正規化**という。

正規乱数　正規分布に従う乱数

一様乱数から正規乱数を発生させる Box and Muller の方法がある。

$z=(-2\ln u_1)^{1/2}\cos 2\pi u_2$ （u_1,u_2は[0,1]の一様乱数）によって変換すればよい。

これを用いて１００個の正規乱数を求めよ。

Mathematica では <<Statistics`NormalDistribution` と読み込み
Table[Random[NormalDistribution[0,1]],{100}] により得られる。

誤差関数 error function と**逆誤差関数**

標準正規分布（ガウス分布）の密度関数を積分した $Erf(z)=\dfrac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt$

が誤差関数 Erf[z] として、Mathematica に組み込まれている。またその逆関数として、逆誤差関数

InverseErf[s]

があり、信頼区間の解析などに利用される。

$\phi(z)=\dfrac{1}{\sqrt{2\pi}}\int_z^\infty e^{-\frac{x^2}{2}}dx$ を Erf を用いて計算すること。

φ(z)=(1 - Erf[z/Sqrt[2]])/2

$s=\dfrac{1}{\sqrt{2\pi}}\int_z^\infty e^{-\frac{x^2}{2}}dx$ で s=0.3085, s=0.0668, s=0.00621 となる z を InverseErf を用いて計算すること。

z=Sqrt[2] InverseErf[1-2s]

計算例　φ(0.0)=0.5000, φ(1.0)=0.1587, φ(2.0)=0.0228
φ(0.5)=0.3085, φ(1.5)=0.0668, φ(2.5)=0.00621

【２変数の確率分布】

さいころＡ，Ｂを同時に振り、Ａの目の数が１または６のときＸ＝１，その他のときＸ＝２とし、Ｂの目の数が奇数のときＹ＝１，偶数のときＹ＝２とする。このとき（Ｘ，Ｙ）の確率分布と周辺分布を表に示せ。

$f(x,y)=\frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}}$ は確率密度関数であることを示し、X，Yの周辺密度関数は、共に、標準正規分布であることを示せ。

【中心極限定理】

$a_1,a_2,...,a_n$が定数で、$X_1,X_2,...,X_n$が互いに独立な確率変数で、それぞれ正規分布 $N(\mu_1,{\sigma_1}^2),N(\mu_2,{\sigma_2}^2),...,N(\mu_n,{\sigma_n}^2)$ に従うとき、確率変数 $\sum_{i=1}^{n}a_1X_i$ は正規分布 $N(\sum_{i=1}^{n}a_i\mu_i,\sum_{i=1}^{n}{a_i}^2{\sigma_i}^2)$ に従う。

確率変数$X_1,X_2,...,X_n$が互いに独立で、すべて有限な平均 μ，分散 σ^2 をもつ同一の確率分布に従うとき、$\overline{X}=\frac{X_1+X_2+\cdots+X_n}{n}$ とおくと、nが大きいならば、$\overline{X}$ は正規分布 $N(\mu,,\frac{\sigma^2}{n})$ に従う。

確率変数 X，Yが互いに独立で、それぞれ正規分布 N(5，1)，N(6,1) に従うとき、確率 P(10＜X＋Y＜12) を求めよ。また確率P(2X＞3Y) を求めよ。

１冊の重さが20ｇ，標準偏差 1ｇの確率分布に従うパンフレットがある。これを３０冊束ねるとき、その重さが590ｇ以上610ｇ以下である確率を求めよ。

【１変数データの整理】

<u>度数分布表</u>

サイコロを５０００回振り、１の目が出た度数を５００回毎に集計し、相対度数の表を作成すること。

２１から３０までのランダム整数を１０００個発生させる。このとき同じ数が何回現れたかを２１が９８，２２が１０３，．．．のように表示させること。

Box and Muller の方法で求めた６０００個の正規乱数を、-3 ～ 3 の範囲で 0.1 刻みでカウントすること。

あるスポーツチームの選手の身長を測定した次のデータ

```
spo={178.5 ,177.3 ,176.6 ,179.3 ,181.8 ,175.4 ,173.7 ,172.3, 184.9 ,175.2 ,
     177.2 ,179.7 ,186.7 ,185.7 ,175.6 ,178.3, 185.6 ,179.0 ,177.0 ,172.9 ,
```

176.9 ,172.6 ,185.9 ,183.7, 176.5 ,174.7 ,178.9 ,175.9 ,179.6 ,175.9 ,
172.1 ,173.3, 176.9 ,172.8 ,179.8 ,174.2 ,181.5 ,175.1 ,177.0 ,180.0}

において、データの数Ｎのとき、範囲Ｒ＝ＭＡＸ－ＭＩＮとし、階数の数ｎを

$$n \approx 1 + \frac{\log_{10} N}{\log_{10} 2}$$ （スタージェスの公式）

で求めよ。（通常７～１５程度）

このデータの度数分布、累積度数を求めよ。

Excel では Frequency

Mathematica では、<<Statistics`Master` をロードし、BinCounts

ヒストグラム 度数分布を柱状グラフで表したもの

度数折れ線 ヒストグラムにおいて、各階級の中央上の高さを表す点を順次線分で結び、両端に高さ０の点を加えて結ぶ折れ線

Excel では ｸﾞﾗﾌｨｸｽｳｨｻﾞｰﾄﾞ

Mathematica では <<Graphics`Master` を読み込み BarChart, ListPlot

標本平均（平均値）・中央値・最頻値（モード）

データ $x_1, x_2, \ldots, x_n$ について $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$

Excel では Average, Median, Mode,

Mathematica では <<Statistics'Master' を読み込み Mean, Median

標本分散と標本標準偏差・不偏分散

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Excel では Var, Stdev, Varp, Stdevp

Mathematica では <<Statistics`Master` を読み込み

VarianceMLE, StandardDeviationMLE, Variance, StandardDeviation

データの標準化

平均０、標準偏差１に変換する

歪度（わいど）・尖度（せんど）

Ｎ個のデータ $x_1, x_2, \ldots, x_N$ に対して、

歪度 $a_3 = \sum_{i=1}^{N}\left(\dfrac{x_i - \bar{x}}{s}\right)^3$ Excel では Skew, Mathematica では Skewness

尖度　$a_4 = \sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{s}\right)^4$　　Excel では Kurt, Mathematica では Kurtosis

ただし、$\bar{x}$ は平均、s は標準偏差。　<<Statistics`Master` を読み込み利用する。

【２変数データの整理】

散布図（相関図）

あるラグビーチーム 20 名の身長(cm), 体重(kg) のデータ

```
rug={{183.2,76.8},{182.4,79.5},{179.4,87.2},{177.4,74.6},{180.4,77.5},
     {173.2,69.0},{182.1,81.8},{174.7,80.5},{180.7,80.2},{185.8,86.5},
     {181.3,77.7},{184.0,85.4},{179.3,76.4},{169.5,74.5},{176.5,83.7},
     {184.2,85.9},{187.7,89.5},{186.4,99.8},{177.8,81.7},{186.5,75.3}}
```

の (x, y) の点を描くこと。

Excel ではデータ範囲を指定して グラフウィザード

Mathematica では ListPlot[rug, AxesLabel->{“身長”,”体重”}]

標本共分散・標本相関係数

データ $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ について

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \quad C_{xy} = \frac{s_{xy}}{s_x s_y}$$

但し x の平均を $\bar{x}$, 標本標準偏差を s_x, y の平均を $\bar{y}$, 標本標準偏差を s_y とする。

Excel では Cov、Correl

Mathematica では <<Statistics`MultivariateDescriptiveStatistics` を使用

CovarianceMLE[{{x1,y1},...,{xn,yn}}]

Correlation[{{x1,y1},...,{xn,yn}}]

回帰直線

y の x に対する回帰直線　$y - \bar{y} = \frac{C_{xy}}{s_x^{\,2}}(x - \bar{x})$

x の y に対する回帰直線　$x - \bar{x} = \frac{C_{xy}}{s_y^{\,2}}(y - \bar{y})$

Excel では Linest

Mathematica では <<Statistics`LinearRegression` を使用

Fit[{{x1,y1},...,{xn,yn}},{1,x},x]　y の x に対する回帰直線

【正規母集団に対する分布関数】

自由度ｎの χ^2 分布

ｎ個の変数 $X_1, X_2, \ldots X_n$ が互いに独立で、それぞれ $N(0,1)$ に従うとき

$Z = X_1^2 + Z_2^2 + \ldots + Z_n^2$ が従う分布

$$T_n(z) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\dfrac{n}{2}\right)} z^{\frac{n-2}{2}} e^{-\frac{z}{2}} & (z > 0) \\ 0 & (z \le 0) \end{cases}$$

$\alpha = \int_t^{\infty} T_n(x)dx$　およびその逆関数

αとｎを与えてｔの値を求める χ^2 分布表を作成すること。

平均 n　分散 2n

自由度ｎ(n=1,2,...10)の χ^2 分布のグラフと、標準正規分布 N(0,1)のグラフを描き、アニメーションで表示させること。

n=2,3,4,5,6 のときの ChiSquareDistribution[n] に従う乱数をそれぞれ 1000 個発生させ、[0, 10]の区間を 20 等分した小区間にデータがそれぞれ何個あるか集計し、その様子を ListPlot で５本の折れ線で表示すること。

３組の NormalDistribution[0,1] に従う乱数 1000 個を使って、
$Z = X_1^2 + Z_2^2 + Z_3^2$ の分布のグラフと、ChiSquareDistribution[3] のグラフを比較せよ。

Ｆ分布　自由度(m,n)の（スネデッカーの）Ｆ分布

X_1, X_2 が互いに独立で、それぞれ自由度m, ｎの χ^2 分布に従っているときの、

$X = \dfrac{\dfrac{X_1}{m}}{\dfrac{X_2}{n}}$ が従う分布

$$f_{m,n}(x) = \begin{cases} \dfrac{m^{\frac{m}{2}} n^{\frac{n}{2}} x^{\frac{m-2}{2}}}{\mathrm{B}\left(\dfrac{m}{2}, \dfrac{n}{2}\right)(mx+n)^{\frac{m+n}{2}}} & (x > 0) \\ 0 & (x \le 0) \end{cases}$$

$\int_t^{\infty} f_{m,n}(x)dx = \alpha$ でα=0.05, 0.01 のときのｔの値を求めるＦ分布表を作成すること。

平均　$\dfrac{n}{n-2}$　　分散　$\dfrac{2(m+n-2)n^2}{m(n-2)^2(n-4)}$

Mathematica では FRatioDistributon[m,n] が利用できる

ｔ分布　自由度ｎの（スチューデントの）ｔ分布

自由度(1,n)のＦ分布 $f_{1,n}(x)$ において、 $X=T^2$ と変換したときの $T=t$ の従う分布

$$f_n(t)=\frac{1}{\sqrt{n}\mathrm{B}\left(\frac{1}{2},\frac{n}{2}\right)\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}}$$

$\int_t^{\infty} f_n(x)dx=\frac{\alpha}{2}$ において、αとｎからｔを求めるｔ分布表を作成せよ。

平均　0　　分散　$\frac{n}{n-2}$

Mathematica では StudentTDistribution[n] が利用できる。

【推定】

母平均μの区間推定（母分散σ^2が既知のとき）

正規母集団から大きさｎの標本を無作為抽出して標本平均 $\overline{X}$ をつくったとき、母平均μを信頼水準γ (=0.95 or 0.99)で推定する。

$$\overline{X}-\frac{\sigma}{\sqrt{n}}z_1<\mu<\overline{X}+\frac{\sigma}{\sqrt{n}}z_1$$

ただし、z_1は　$\int_{-z_1}^{z_1} g(z)dz=\gamma$　となるように選ぶ。

γ=95% のときz_1=1.960　γ=99%のときz_1=2.576 である。

母平均の区間推定（母分散が未知のとき）

大きさｎの標本を無作為抽出して標本平均 $\overline{X}$、標本分散S^2から母平均μを信頼水準γ (=0.95 or 0.99)で推定する。

$$\overline{X}-\sqrt{\frac{x_1}{n-1}}S<\mu<\overline{X}+\sqrt{\frac{x_1}{n-1}}S$$

ただし、x_1は$\int_0^{x_1} f_{1,n}(x)dx=\gamma$となるように選ぶ。

母比率の区間推定

標本比率をＲとすると、標本の大きさｎが大きいとき、母比率ｐに対する信頼区間は、信頼水準γに対するz_1を$\int_{-z_1}^{z_1} g(z)dz=\gamma$　となるように選ぶと

$$\left[R - z_1\sqrt{\frac{R(1-R)}{n}},\quad R + z_1\sqrt{\frac{R(1-R)}{n}} \right]$$

とくに、γ=95% のときz_1=1.960　γ=99%のときz_1=2.576 である。

母分散の推定

正規母集団から大きさｎの標本を無作為抽出したときの標本分散S^2から母分散σ^2を信頼水準γ (=0.95 or 0.99)で推定する。

$$\frac{nS^2}{x_2} < \sigma^2 < \frac{nS^2}{x_1}$$

ただし、x_1, x_2は　$\int_0^{x_1} T_{n-1}(x)dx = \frac{1-\gamma}{2}$, $\int_{x_2}^{\infty} T_{n-1}(x)dx = \frac{1-\gamma}{2}$ となるように選ぶ。

母相関係数の区間推定

母集団から大きさｎの標本を抽出し、その標本相関係数がC_{xy}であるとする。

$Z = \tanh^{-1} C_{xy} = \frac{1}{2}\log\frac{1+C_{xy}}{1-C_{xy}}$ とおく。母相関係数 $\rho_{xy} = \tanh\zeta$　は$T = \sqrt{n-3}(Z-\zeta)$

が N(0,1) に従うことから、信頼水準９５%で　-1.960 ＜ $\sqrt{n-3}(Z-\zeta)$ ＜ 1.960

即ち、 $\tanh\left(Z - \frac{1.960}{\sqrt{n-3}}\right) < \rho_{xy} < \tanh\left(Z + \frac{1.960}{\sqrt{n-3}}\right)$

ある店で買った１０個の LL サイズの卵の重(単位ｇ）を計ったところ

65.1, 67.5, 71.5, 68.4, 70.1, 72.2, 68.7, 69.3, 70.6, 67.1

であった。 LL サイズの卵全体は正規母集団と考え、母分散は 4.0 とする。
母平均の信頼区間を信頼水準\ 95% で推定せよ。

全国から無作為抽出した２５００世帯について、年間の米購入量を調査したところ、平均値１１８Ｋｇ，標準偏差３８．０Ｋｇであった。全国の１世帯あたりの平均購入量を信頼度９９%で推定せよ。

ある地域で有権者５０００人を無作為に抽出して、A政党の支持者を調べたところ、２３５０人であった。この地域のA政党支持率ｐを、信頼度９５%で推定せよ。

ある溶液のｐＨを８回測定し、次の値を得た。

7.86, 7.90, 7.81, 7.94, 7.84, 7.92, 7.91, 7.93

この溶液のｐＨの測定値は正規母集団をつくり、上の値はこれから抽出した標本の

実現値とみて、母分散の９５%信頼区間を求めよ。

ある大学の学生の中から１００人を無作為に選び、体重と胸囲を測ったところ、標本相関係数が ０．８７であった。母相関係数を信頼水準９５%で推定せよ。

【検定】

<u>母平均の検定</u>

大きさNの母集団で、母平均をμ、分散をσ^2とする。ただし、これらは実際は分かっていない。この母集団から大きさnの標本を抽出し、標本平均$\overline{X}$、標本分散S^2、不偏分散U^2の実現値をそれぞれ$\bar{x}, s^2, u^2$とする。このとき母平均μに関する帰無仮説「$H_0 : \mu = \mu_0$」を検定する。H_0が真であると仮定するとき、

母集団が正規母集団$N(\mu, \sigma^2)$の場合

σ^2が既知であるか、σ^2が未知でnが大きいとき

$Z = \dfrac{\overline{X} - \mu_0}{\sqrt{\dfrac{\sigma^2}{n}}}$ は標準正規分布に従うので検定統計量としてZを用いる。

危険率αに対して、Zの値が棄却域にあれば、H_0を棄却する。

σ^2が未知でnが小さいとき

$T = \dfrac{\overline{X} - \mu_0}{\sqrt{\dfrac{U^2}{n}}}$ は自由度(n－１)のｔ分布に従うので、Tを検定統計量とする。

一般の母集団の場合

N，nが大きいとき

$Z = \dfrac{\overline{X} - \mu_0}{\sqrt{\dfrac{N-n}{N-1}\dfrac{\sigma^2}{n}}}$ は近似的に標準正規分布に従うので、Zを検定統計量とする。

なお、σ^2が未知のときは、その一致推定値であるS^2を代用する。

N が非常に大きく$\dfrac{N-n}{N-1}$が１に近いとき

$Z = \dfrac{\overline{X} - \mu_0}{\sqrt{\dfrac{\sigma^2}{n}}}$ を検定統計量とする。

<u>母分散の検定</u>

標本分散S^2をもとにして、仮説「$H_0 : \sigma^2 = {\sigma_0}^2$」を検定する。対立仮説として「$H_1 : \sigma^2 \neq {\sigma_0}^2$」や「$H_1 : \sigma^2 > {\sigma_0}^2$」や「$H_1 : \sigma^2 < {\sigma_0}^2$」をおき、両側検定・片側検定

に入る。H_0が真であると仮定すると、$T=\frac{nS^2}{\sigma_0^{\ 2}}$が自由度（n－1）の$\chi^2$分布に従うので、Tを検定統計量として用いる。

母比率の検定

大きさN、母比率pの二項母集団があるとき、仮説「$H_0 : p = p_0$」を検定する。

標本の大きさをn、標本比率を$\hat{P}$で表すと、N，nが大きいとき

$$Z=\frac{\hat{P}-p_0}{\sqrt{\frac{N-n}{N-1}\frac{p_0 q_0}{n}}} \quad (q_0=1-p_0)$$ が近似的に標準正規分布に従うことを使う。

無相関の検定

N個の対応する2変量のデータの相関係数rから検定統計量

$$T(r)=\frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

を求め、有意水準α（0.05または0.01）に対して

$$T(r)\le -f_{N-2}(\frac{\alpha}{2}) \quad \text{または} \quad f_{N-2}(\frac{\alpha}{2})\le T(r)$$

ならば、「仮説H_0：相関がない」を棄却する。すなわち、「相関がある」

母集団の母相関係数の検定

サンプルの相関係数Rから、母相関係数 $\rho=\rho_0$ を検定するには、統計量

$$T=\frac{\sqrt{n-3}}{2}\left(\log\frac{1+R}{1-R}-\log\frac{1+\rho}{1-\rho}\right)$$

を用いる。ただし、標本の大きさnは相当大きいとする。T はほぼ標準正規分布 $N(0,1)$ に従う。

２つの母集団の母相関係数の比較検定

２つの母集団の標本相関係数を、それぞれR_1, R_2 とする。２つの標本の大きさは相当大きいとする。このとき、２つの母集団の母相関係数が等しいという検定をする。そのための統計量は

$$T=\frac{\frac{1}{2}\log\frac{1+R_1}{1-R_1}-\frac{1}{2}\log\frac{1+R_2}{1-R_2}}{\sqrt{\frac{1}{n-3}+\frac{1}{m-3}}}$$

である。これはほぼ正規分布 $N(0,1)$ に従う。

2つの正規母集団の等分散の検定

2つの互いに独立な正規母集団 $N(\mu_1,\sigma_1^{\,2}), N(\mu_2,\sigma_2^{\,2})$ から大きさ n_1, n_2 の標本を抽出し、それらの不偏分散をそれぞれ $U_1^{\,2}, U_2^{\,2}$ とおくと、 $F=\dfrac{U_1^{\,2}}{\sigma_1^{\,2}}\Big/\dfrac{U_2^{\,2}}{\sigma_2^{\,2}}$ は自由度 (n_1-1, n_2-1) のF分布に従う。統計量 F を用いて仮説「$H_0:\sigma_1^{\,2}=\sigma_2^{\,2}$」が検定できる。対立仮説を「$H_1:\sigma_1^{\,2}>\sigma_2^{\,2}$」にとる場合、$H_0$ が真ならば、F の実現値 $f=u_1^{\,2}/u_2^{\,2}>1$ と予想されるので、f を検定統計量として右側検定を行う。

2つの正規母集団の母平均の差の検定

2つの互いに独立な正規母集団 $N(\mu_1,\sigma_1^{\,2}), N(\mu_2,\sigma_2^{\,2})$ から大きさ n_1, n_2 の標本を抽出し、それらの標本平均を $\overline{X_1}, \overline{X_2}$ とし、標本分散を $S_1^{\,2}, S_2^{\,2}$ とする。 $\overline{X_1}-\overline{X_2}$ は $N(\mu_1-\mu_2, \dfrac{\sigma_1^{\,2}}{n_1}+\dfrac{\sigma_1^{\,2}}{n_2})$ に従う。従って、 $Z=\dfrac{\overline{X_1}-\overline{X_2}-(\mu_1-\mu_2)}{\dfrac{\sigma_1^{\,2}}{n_1}+\dfrac{\sigma_1^{\,2}}{n_2}}$ は $N(0,1)$ に従うので、$\sigma_1^{\,2}, \sigma_2^{\,2}$ が既知であれば、仮説「$H_0:\mu_1-\mu_2=0$」の検定ができる。

$\sigma_1^{\,2}, \sigma_2^{\,2}$ が既知でなくても、$\sigma_1^{\,2}=\sigma_2^{\,2}$ であることが分かっていれば、

$X=\dfrac{(n_1-n_2-2)\{\overline{X_1}-\overline{X_2}-(\mu_1-\mu_2)\}}{\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)\left(n_1S_1^{\,2}+n_2S_2^{\,2}\right)}$ が自由度 $(1, n_1+n_2-2)$ のF分布に従うこと

を用いる。

適合度の検定

母集団が互いに排反なn個のクラス $A_1, A_2, \ldots, A_n$ に分けられていて、ある固体が各クラスに属する確率は $p_1, p_2, \ldots, p_n$ $(p_1+p_2+\cdots+p_n=1)$ とする。この母集団から大きさNの標本を抽出したとき、各クラスに属する固体の数は $p_1N, p_2N, \ldots, p_nN$ と期待される。これを**期待度数**という。一方、抽出した標本で実際に各クラスに入っている固体の数 $x_1, x_2, \ldots, x_n$ $(x_1+x_2+\cdots+x_n=N)$ は**観測度数**という。この両者を比較する。Nが大きいとき

$X=\dfrac{(x_1-p_1N)^2}{p_1N}+\dfrac{(x_2-p_2N)^2}{p_2N}+\cdots+\dfrac{(x_n-p_nN)^2}{p_nN}$ ， $p_iN\geq 5\ (i=1, 2, \ldots, n)$

は自由度n－1の χ^2 分布をなす。

独立性の検定

2つの属性 A, B についてのm×ｒ分割表（クロス集計表・連関表）

$B \cdot \cdot A$	A_1	A_2	…	A_r	sum
B_1	f_{11}	f_{12}	…	f_{1r}	$f_{1\circ}$
B_2	f_{21}	f_{22}	…	f_{2r}	$f_{2\circ}$
⋮	⋮	⋮	⋱	⋮	⋮
B_m	f_{m1}	f_{m2}	…	f_{mr}	$f_{m\circ}$
sum	$f_{\circ 1}$	$f_{\circ 2}$	…	$f_{\circ r}$	N

（$f_{ij} \geq 5$）

に対して、次の表の統計量 $f_{i\circ} f_{\circ j}$ を計算する。

$B \cdot \cdot A$	A_1	A_2	…	A_r
B_1	$f_{1\circ} f_{\circ 1}$	$f_{1\circ} f_{\circ 2}$	…	$f_{1\circ} f_{\circ r}$
B_2	$f_{2\circ} f_{\circ 1}$	$f_{2\circ} f_{\circ 2}$	…	$f_{2\circ} f_{\circ r}$
⋮	⋮	⋮	⋱	⋮
B_m	$f_{m\circ} f_{\circ 1}$	$f_{m\circ} f_{\circ 2}$	…	$f_{m\circ} f_{\circ r}$

（$\dfrac{f_{i\circ} f_{\circ j}}{N} \geq 5$）

ここで、検定統計量 $T(f_{ij}) = \sum_{i=1}^{m} \sum_{j=1}^{r} \dfrac{(Nf_{ij} - f_{i\circ} f_{\circ j})^2}{Nf_{i\circ} f_{\circ j}}$ を求める。

有意水準 α（0.05 または 0.01）に対して

$T(f_{ij}) \geq \chi^2_{(m-1)(r-1)}(\alpha)$

ならば、「仮説 H_0 ：２つの属性 A, B は独立である」を棄却する。

ある植物の種子の発芽率は７５％といわれている。この種子の中から３００個の無作為標本を抽出して発芽実験をしたところ、２０５個が発芽した。この種子の発芽率は７５％と考えてよいか。危険率５％で検定せよ。

２５０ｇと表示されている缶詰の中から、無作為に１００個取り出し、その重さを調べたところ、平均値２４７ｇ，標準偏差６．５ｇを得た。このことから、この表示に誤りがあるといえるか。危険率５％で検定せよ。

ある県の中学校２９３校について、各学校の生徒数と学力テストの平均点との相関係数を求めたところ、R=0.63 であった。学校の生徒数と学力テストの成績に相関関係があるか、危険率５％で検定せよ。

相関関係あり

ある県の公立中学校について、各中学校の生徒数と学力テストの成績の相関係数を調べた。都市部の中学１０１校では R_1=0.51，郡部の中学１９２校では R_2=0.63 であった。都市部と郡部の間で相関係数に差があるか。危険率５％で検定せよ。

ある中学校で、１５人を無作為に１５人選び、数学と英語の成績を調べた。

数学 16 13 18 8 13 17 19 7 15 16 20 14 10 16 18
英語 18 17 15 10 14 14 12 12 12 19 18 13 12 15 17

数学と英語の成績の間に相関関係があるか。危険率５％で検定せよ。

相関関係あり

ある大学で、学生１０人を無作為に選び、身長と足の裏の長さを調べた。

身長 173 171 165 166 171 175 162 164 168 168 ㎝
足裏 24.7 25.8 24.0 24.6 24.4 25.5 23.4 24.6 25.1 24.6 ㎝

身長と足の裏の長さに相関関係があるか。危険率１％で検定せよ。

ある大学の２年生以上の学生４１０名について、喫煙と留年の関係を調べたところ、次の表を得た。

留年の有無	有り	無し	計
煙草を吸う	75	122	197
吸わない	44	169	213
計	119	291	410

この結果から煙草を吸う学生の方が留年の経験が多いかどうか、危険率１％で検定せよ。

独立であるとはいえない。

参考書

例題中心 確率･統計入門 坂光一･水原昂廣 学術図書
確率統計(理工系の数学入門コース7) 薩摩順吉 岩波書店 1989
確率統計 田河生長他 大日本図書 1995
精解 確率と統計演習 鈴木七緒他 共立出版 1979
Mathematica確率統計入門 小林道正 トッパン 1994
すぐわかる統計解析 すぐわかる多変量解析 石村貞夫 東京図書 1993
すぐわかるＥｘｃｅｌによる統計解析 内田 治 東京図書
Ｅｘｃｅｌ統計解析フォーム集 シンクスタット 共立出版 1997