基礎統計（小林）　期末試験　2009年7月24日金曜17:00-18:30　教科書、ノート、参考書、プリント、いずれも持込可、電卓、関数電卓持込可。仮説検定においては有意水準5%を用いること。

略解をつけます。実際の解答にはもっと丁寧に論理を書いてください。

1）確率変数 X は $P(X=1)=p,\ P(X=0)=1-p\quad(0\le p\le 1)$ を満たす。

期待値 E[X]，分散 V(X)，$E[(X-E(X))^3]$ を求めよ。（正答率 48%!!!!）

E[X]=1×p+0×(1-p)=p 分散 V(X)=$(1-p)^2$×p+ $(0-p)^2$(1-p)=p(1-p)，

$E[(X-E(X))^3]$=$(1-p)^3$×p+$(0-p)^3$(1-p)=p(1-p)=(1-p)p[$(1-p)^2$−$p)^2$]= (1-p)p[1-2p]

2）確率変数 X_1, X_2 は互いに無相関であり、ともに期待値　、分散 σ^2 を持つ。平均を $\bar{X}=\dfrac{X_1+X_2}{2}$ で表し、不偏分散を

$s^2=(X_1-\bar{X})^2+(X_2-\bar{X})^2$ と表すとき、次の問いに答えよ。正答率

(a) $E(\bar{X})=\quad,\ V(\bar{X})=\sigma^2/2,\ E[s^2]=\sigma^2$ を示せ。

解答：$E[X_1+X_2]/2=(\mu+\mu)/2=\mu$，$V([X_1+X_2]/2)=(1/4)V[X_1+X_2]=(1/4)(\sigma^2+\sigma^2)=(1/2)\sigma^2$

「X_1 と X_2 が無相関なら $Cov(X_1,X_2)=0$」を使う。

$$E[(X_1-\bar{X})^2+(X_2-\bar{X})^2]=E[(X_1-\quad-(\bar{X}-\quad))^2+(X_2-\quad-(\bar{X}-\quad))^2]$$
$$=E[(X_1-\quad)^2+(\bar{X}-\quad)^2-2(X_1-\quad)(\bar{X}-\quad)+(X_2-\quad)^2+(\bar{X}-\quad)^2-2(X_2-\quad)(\bar{X}-\quad)]$$
$$=V(X_1)+V(X_2)+2V(\bar{X})-2E[(X_1+X_2-2\quad)(\bar{X}-\quad)]$$
$$=5-4E[(\bar{X}-\quad)^2]=5-4V(\bar{X})=1$$

(b) 共分散 $Cov(\bar{X},X_1-\bar{X})$ を求めよ。

解答：$Cov(\bar{X},X_1-\bar{X})=(1/2)Cov((X_1+X_2,X_1)-V(\bar{X})=(1/2)V(X_1)-V(X_1)=(1/2)-(1/2)=0$

3）確率変数 $e_1,\ldots,e_n$ は期待値 0，分散 1 を持ち、互いに無相関である。定数 b と定数 $x_1,\ldots,x_n\quad(x_i\ne 0)$ を用いて、確率変数 $y_1,\ldots,y_n$ は、$y_1=bx_1+e_1,\ \ldots,\ y_n=bx_n+e_n$ により定義される。$\hat{b}=\sum_{i=1}^n y_ix_i\Big/\sum_{i=1}^n x_i^2$ と定義する。

(a) $\hat{b}=b+\sum_{i=1}^n e_ix_i\Big/\sum_{i=1}^n x_i^2$ を示し、それを用いて $E[\hat{b}]=b$，$V(\hat{b})=1\Big/\sum_{i=1}^n x_i^2$ を示せ。

解答：$\hat{b}=\sum_{i=1}^n y_ix_i\Big/\sum_{i=1}^n x_i^2$ に $y_t=bx_t+e_t$ を代入すると $\hat{b}=b+\sum_{i=1}^n e_ix_i\Big/\sum_{i=1}^n x_i^2$ は容易にえられる。b や x は確率変数ではないので、$c\,E[\sum_{i=1}^n e_ix_i]\Big/\sum_{i=1}^n x_i^2$ となる。$E[\sum_{i=1}^n e_ix_i]=\sum_{i=1}^n E[e_i]x_i=0$ なので、$E[\hat{b}]=b$

(b) $\mathrm{cov}(e_1,\ \hat{b})=x_1\Big/\sum_{i=1}^n x_i^2$，$\mathrm{cov}(y_1-\hat{b}x_1,\ \hat{b})=0$ を示せ。

解答：

$$\mathrm{cov}(e_1,\ \hat{b})=\mathrm{cov}(e_1,\ b+\sum_{i=1}^n e_ix_i\Big/\sum_{i=1}^n x_i^2)=\mathrm{cov}(e_1,\ \sum_{i=1}^n e_ix_i)/\sum_{i=1}^n x_i^2$$

$$=(1/\sum_{i=1}^n x_i^2)(cov(e_1,x_1e_1)+cov(e_1,x_2e_2)+\ldots+cov(e_1,x_ne_n))$$

である。$cov(e_1,x_ie_i)=\begin{cases}x_1\,\mathrm{cov}(e_1,e_1)=x_1V(e_1)\\ 0\end{cases}$ なので、$\mathrm{cov}(e_1,\ \hat{b})=x_1\Big/\sum_{i=1}^n x_i^2$ が導かれる。

$\mathrm{cov}(y_1-\hat{b}x_1,\ \hat{b})=\mathrm{cov}(bx_1+e-\hat{b}x_1,\ \hat{b})=\mathrm{cov}(bx_1,\ \hat{b})+\mathrm{cov}(e,\ \hat{b})-\mathrm{cov}(\hat{b}x_1,\ \hat{b})$ にまず注意。

ここで、$\mathrm{cov}(定数,\ \hat{b})=0$ より第一項は０、第二項は既に求めた。

第三項は $\mathrm{cov}(bx_1,\ \hat{b})$=$x_1\,\mathrm{cov}(\hat{b},\ \hat{b})$=$\mathrm{x}_1 V(\hat{b})=\mathrm{x}_1 / \sum_{i=1}^{n} x_i^2$ であることところから、$\mathrm{cov}(y_1-\hat{b}x_1,\ \hat{b})=0$ が得られる。

4) 確率変数 $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_n$ が $E(X_i)=\ \ _X$, $E(Y_i)=\ \ _Y$, $V(X_i)=\sigma_X^2, V(X_i)=\sigma_Y^2$, $Cov(X_i, Y_i)=c$ という一定の期待値、分散、共分散を持ち、$i \neq j$ のとき、$Cov(X_i, Y_j)=0, Cov(X_i, X_j)=0, Cov(Y_i, Y_j)=0$ とする。

$E\left(\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})/(n-1)\right)=c$ を示せ。

解答：$x_i = X_i -\ \ $ ，$y_i = Y_i -\ \ $ と表すと、$E\left(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right)=E\left(\sum_{i=1}^{n}x_i y_i - n\overline{xy}\right)=\sum_{i=1}^{n} E(x_i y_i)-nE(\overline{xy})$ m

において、第二項は $nE(\overline{xy})=(1/n)E(x_1+\ldots+x_n)(y_1+\ldots+y_n)=(1/n)(E[x_1y_1]+\ldots+E[x_ny_n]+\sum_{i\neq j}E[x_iy_j])$ であるが、

$i \neq j$ のとき $E[x_iy_j]=\mathrm{cov}(x_i, y_j)=0$ なので、$nE(\overline{xy})=\mathrm{cov}(X_i, Y_i)=c$ 。

第一項は $\sum_{i=1}^{n} E(x_iy_i)=\sum_{i=1}^{n} Cov(x_i, y_i)=nc$ 。よって $E\left(\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})\right)=c(n-1)$ が示される。

5) 男女 100 人ずつの学生について喫煙習慣の有無と飲酒習慣の有無の関係のデータをとり、2×2 の分割表によって独立性の検定をおこなった。男女別々に検定を行ったところ、いずれの場合も喫煙と飲酒の関係は有意水準 5%で有意であったが、これらのデータを合わせ 200 人のデータとして分析したろころ、喫煙と飲酒の関係は有意水準 5%で有意ではなかった。このような条件にあてはまる数値例を作り、検定を行え。このとき、統計量の分布のグラフを書き、その上に統計値、臨界値を示せ。

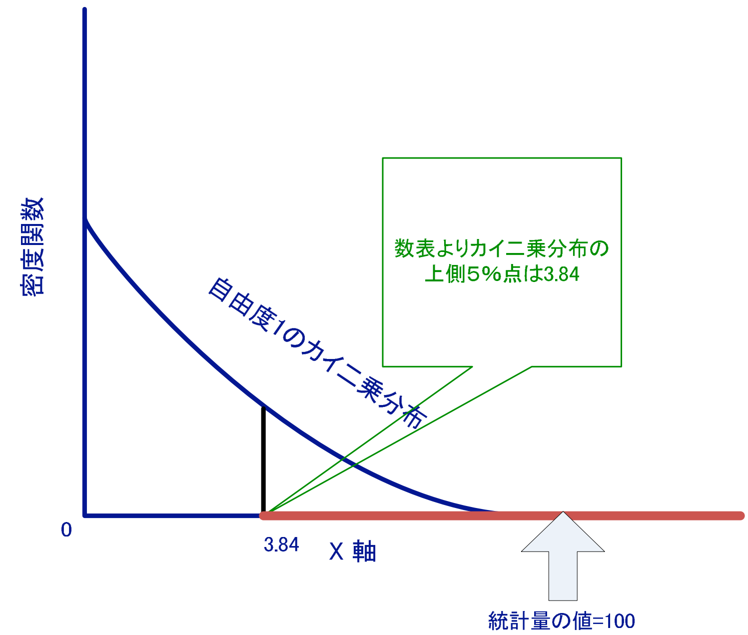
解答例：

男性	飲酒	非飲酒	計	女性	飲酒	非飲酒	計	全体	飲酒	非飲酒	計
喫煙	50	0	50	喫煙	0	50	50	喫煙	50	50	100
非喫煙	0	50	50	非喫煙	50	0	50	非喫煙	50	50	100
計	50	50	100	計	50	50	100	計	100	100	200

男性の場合の独立性の検定統計量の値は

$(50-25)^2/25+(50-25)^2/25+(50-25)^2/25+(50-25)^2/25=100$

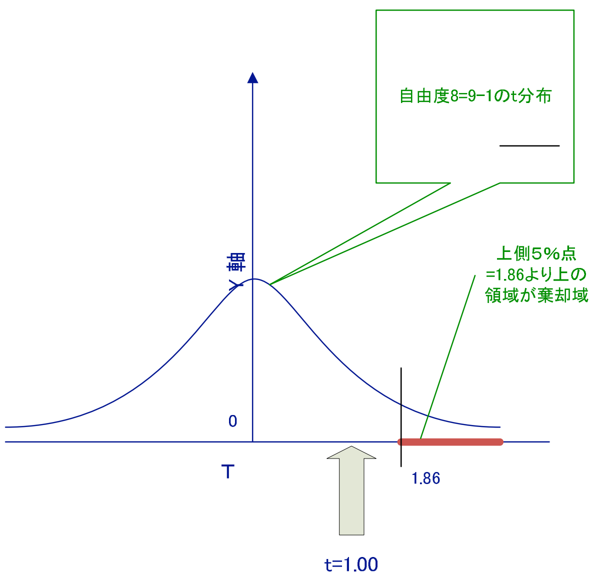
もしも仮説「喫煙と禁煙は独立」が正しければ、この統計量は自由度(2−1)(2−1)＝１のカイ二乗分布に従うので、100>3.84 より、有意水準 5%で仮説は棄却される。女性についても同様。しかし双方のデータを統合すると、検定統計量の値は０となり、仮説は棄却されない。



6） ある大学の新入生から 9 人をランダムに選んで尿酸の濃度を調査したところ、平均 7.5 ,標準偏差（不偏分散の平方根）1.5 であった。正規分布の仮定のもと、帰無仮説「この大学の新入生の尿酸の濃度分布の期待値は 7.0 以下（正常の範囲）である」にたいし t 検定を行い、結果を解釈せよ。このとき、必ず、統計量の分布のグラフを書き、その上に統計値、臨界値を示せ。

解答:t 検定を行う。濃度の期待値 とすると、統計量$t=\frac{(\bar{X}-\ )}{\sqrt{s^2/n}}$は自由度は 8=9-1 の t 分布にしたがう。S=1.5, n=9 を入れると、t=1.0 となる。仮説「正常範囲内」を棄却するのは、測定された濃度が十分高い時だから、片側検定となり、t>t0 の形の棄却域になる。 =7 のとき、統計量の値は(7.5-7)/(1.5/3)=1 であり、これは自由度 8 の t 分布の上側 5%点 1.86 を下回るので、水準 5%では仮説は棄却されない。すなわち、新入生の濃度の期待値が 7.0 を超えているとはいえない。

（ =7 のとき t>1.86 である確率は 5%であり、t>1.86 は稀な現象といえ、 <7 のとき t>1.86 はさらに稀になる）



7） ランダムに選んだ 2500 人を調査したら、50 パーセントの人がある政策を支持していた。観察された支持率が正規分布に従うと仮定し、標本は十分大きいと考え、真の政策支持率の 95 パーセント信頼区間を求めよ。

サンプルサイズ n,真の支持率を p とすると、支持率 r の分布は期待値 p, 分散 p(1-p)/n の正規分布に近似される。したがって、$P(-1.96<(r-p)/(p(1-p)/n))^{1/2}<1.96)=0.95$ が成立。これを変形すると、

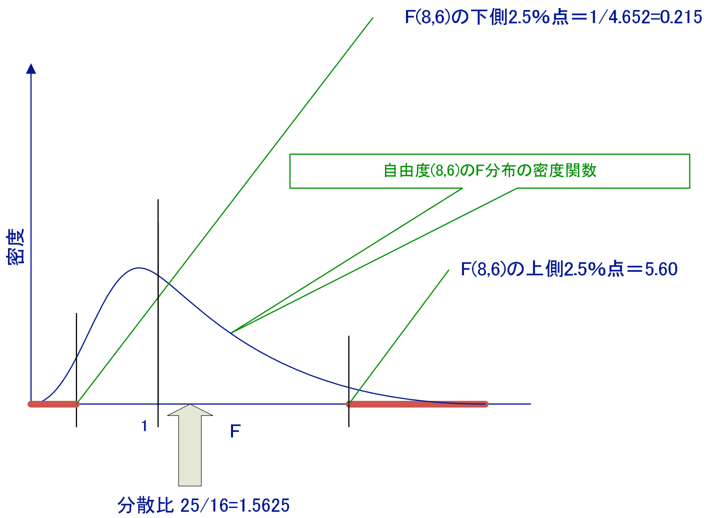
$P(r-1.96(p(1-p)/n))^{1/2}<p<r+(p(1-p)/n))^{1/2}1.96)=0.95$

区間推定の上限と下限の P=0.5 については推定値の 0.5 を代入すると、信頼区間は $0.5\pm 0.5/50*1.96=0.5\pm 0.0196$

8） ある動物について、オス 9 頭の体重の平均 25kg、標準偏差（不偏分散の平方根）5kg であり、メス 7 頭の体重の平均 17kg、標準偏差（不偏分散の平方根）4kg であった。ただし体重の分布は正規分布を仮定する。

(a) このデータから、オスとメスで体重のばらつき（母分散）が異なるといえるか検定を行い、結果を解釈せよ。必ず、統計量の分布のグラフを書き、その上に統計値、臨界値を示せ。

解答：オスメスの分散比は F=25/16=1.5625. これが 1 から離れているとき、同一分散の仮説が棄却されるので、両側検定となる。分散が等しい時、分散比は自由度 8,6 の F 分布に従うので、数表より上側 2.5%点は 5.6 である。下側 2. 5%点は数表にはないが、自由度 6,8 の F 分布の上側 2.5%点の逆数 1/4.652=0.215 としてもとめられる。分散比は棄却域に入らないので、同分散の仮説は棄却されず、このデータでは分散が異なるとは認められない。



(b) オスの母分散について、95 パーセント信頼区間を求めよ。

$(n-1)s^2/\sigma^2$ は自由度 n-1=8 のカイ自乗分布にしたがうので、自由度 8 のカイ自乗分布の下側 2.5%点、上側 2.5%点は数表より、2.18 と 17.53 なので、$P(2.18<(n-1)s^2/\sigma^2<17.53)=0.95$ を変形して、$P(8\times s^2/2.18>\sigma^2>8\times s^2/17.53)=0.95$ である。したがって s=5 のとき、信頼係数 95%の区間推定は 11.41 と 91.74 である。

9） 標本の大きさ 100 の実験をおこない、特性 A,B,C の出現回数が下記のように観察された。特性 A,B,C の出現確率が理論的には Pr(A):Pr(B):Pr(C)=1:2:1 であるとする。帰無仮説を「データは理論に従って発生している」とし、対立仮説（帰無仮説が否定されたときの結論）を「データは理論にしたがわない」として、適合度検定を行え。必ず、統計量の分布のグラフを書き、その上に統計値、臨界値を示せ。

特性	A	B	C	合計
観測回数	20	45	35	100

理論に従う平均頻度は A が 25、B が 50、C が 25 なので、観測値との差を示す統計量を次のように計算する。

カイ二乗統計量＝$(20-25)^2/25+(45-50)^2/50+(35-25)^2/25=5.5$。この統計量が 0 から大きくかい離していれば、理論は棄却されるので、片側検定を行う。