基礎統計（2010 年度冬学期）授業ノート

1. 集めて測る

1.1　統計学の発見

統計は何をやっているのか？⇒「集めて測る」

集める事で何か見えてくる！

18 世紀頃から形成、「死亡秩序 mortal order」の発見

(1)イングランドの 1811 年の平均寿命推計

農村部 41、10 万未満の都市 32、10 万以上の都市 30

(2)16～17 世紀、イングランドの乳児死亡率

都市：ロンドン 4 教区 0.228、ヨーク教区 0.237

市場町：ゲインズバラ 0.222、バンベリィ 0.165

農村：6 教区平均 0.121(0.090～0.149)

(3)18 世紀イングランドの乳児死亡率

都市：ロンドン 0.319

市場町：ゲインズバラ 0.236、バンベリィ 0.220

農村：11 教区平均 0.165

John Graunt『死亡表 Bills of Mortality』

(1)「死には二種類ある」

「総埋葬数に対して恒常的な比率を保つもの/そうでないもの」

第一種の死：脳卒中、心臓発作、事故、自殺

第二種の死：伝染病（ペスト、猩紅熱など）

病気による死も集めて測ると違った様相を見せる

事故や自殺と同じものと違うもの

事故は偶然によるが、集めて測れば法則性を持つ

→偶然を支配する法則が存在する

⇒「極限定理 limit theorem」：個数 N が大きくなると特定の法則に従う

(2)「生を超過する死」

都市では死亡が出生を常に超過する（ex.ロンドン）

「都市＝蟻地獄」仮説

(3)「変わる数と普遍法則」

男女の出生数が違う。「出生数の男女比は、1:1 ではなく、約 14:13」を発見。

「アダムとイブ」なら神は 1:1 で作ったはず！なのに何故？

→誤差と法則性をどう識別するか

(1)の問題でもあるが、出生比 not1（真のなるべき値）は聖書がらみなので特に衝撃的

1.2　分布の形状を調べる

Graunt の発見「変わる数と普遍法則」

＝「出生数は大きく変わるのに、出生比は恒常的」

1)集まりが全体としてどの辺りにあるのか

：集まっている位置を表す指標『平均 mean』$(1/N)\Sigma i(xi)$

※以下、$(1/N)\Sigma i(xi)$で表記する

2)集まり方がどの程度ばらついているのか

：散らばり方（集まり方）の程度を示す指標『分散 variance』

$(1/N)\Sigma i(xi-1/N\Sigma i(xi))^2=1/N\Sigma i(xi-m)^2$

慣習的に、平均は m、分散は V または $s^2$ で表す

定義式の意味

①何故 N で割るのか？

そうしないと、具体的な値の数によって大小が違う。

←$\Sigma i(xi)$は xi の個数に影響されるから N で割って…

②なぜ$(xi-m)^2$と 2 乗するのか？

そうしないと、差の正負が打ち消し合ってしまう。

集まりは、位置と幅でまず押さえる。専門的な言い方では「平均と分散は 1 変数の分布をみる最も基本的な指標」

『分布 distribution』: 集まり（散らばり）のあり方

『変数 variable』: 統計学が捉えたいと思っている対象。変動する値を持つもの。

ロンドンの出生数と出生比に話を戻すと…

82 年間の

男性の出生数：平均 5907、分散 2732354

女性の出生数：平均 5533、分散 2531195

男女の出生比：平均 1.07、分散 0.000986

出生数の分散は出生比の分散に比べて明らかに大きい

「出生数は大きく動くが、出生比は恒常的」と言えるか？

―まだダメ

元の数の絶対的な大きさが影響する
例えば、単位を変えるだけで大きさが変わる
その影響をどうすれば外せるか？
　→測りたい特性だけをいかに取り出すか？
　→直観的な感覚をいかにうまく再現するか？
注目点⇒平均が大きくなるほど分散は大きい
　　⇒相殺させればよい！

(1)分散/平均
　ちょっと変…
　分散は定義式をみると 2 乗している。だから単位の影響も 2 乗。（次元解析の考え方）
　ならば…
(2)(分散)$^{0.5}$/平均
　つまり、分散の平方根を平均で割ればよい。
　分散の平方根を『標準偏差 standard deviation』と呼ぶ
　『変動係数 coefficient of variation』：標準偏差/平均＝s/m
82 年間での
　男性出生数：変動係数 0.278
　女性出生数：変動係数 0.286
　男女の出生比：変動係数 0.029
たしかに出生比の変動係数は小さい

●補足
　散らばり方を示すもう一つの指標：1/N Σi|xi－m|『平均偏差 mean deviation』
　→各観測値が平均からどれだけ離れているかについての平均を求めたもの
　男性の出生数：平均 5907、平均偏差 1424
　女性の出生数：平均 5533、平均偏差 1372
　男女の出生比：平均 1.07、平均偏差 0.0235
　男性の平均偏差/平均 0.2411、変動係数 0.278
　女性の平均偏差/平均 0.2480、変動係数 0.286
　男女の出生比の平均偏差/平均 0.021、変動係数 0.029
　…

1.3　指標から分布をみる
　平均と分散以外にも測る指標はありうる

例えば、$1/N\Sigma i(xi-m)^3$

これは、正負どちらにどの程度振れているか＝歪み具合を示す

→『歪度 skewness』: $(1/N\Sigma i(xi-m)^3)/s^3$

$1/N\Sigma i(xi-m)^4$、これはどの程度真ん中に集まっているか＝尖り具合（正確には尖らなさ具合）を示す

→『尖度 kurtosis』: $(1/N\Sigma i(xi-m)^4)/s^4$、または、$(1/N\Sigma i(xi-m)^4)/s^4-3$

「－3」の 3 は正規分布の$(1/N\Sigma i(xi-m)^4)/s^4$ だから、$(1/N\Sigma i(xi-m)^4)/s^4-3$ は正規分布と比べて尖っているかいないかを＋－で示す

『平均』全体的な位置：xi の 1 乗に関わる

『分散』散らばり方：xi の 2 乗に関わる

『歪度』歪み方：xi の 3 乗に関わる

『尖度』尖り方：xi の 4 乗に関わる

5 乗項や 6 乗項や…に関わる指標も有り得て、それぞれ数の集まり＝『分布』の何らかの特徴を表す。経験的に、4 乗項ぐらいで大体わかるので、特別な用語はない。

しかし、理論的に考えると…

1 乗項、2 乗項、3 乗項、4 乗項、…は分布のあり方の特徴をそれぞれ表現する

だとすれば、それら全部を集める…

分布 → 平均／分散／歪み／尖り／… ⇔「分布＝平均＋分散＋…」

実は、少し工夫すれば本当に数式で成り立つ！

分布関数 → 1 乗項 $1/N\Sigma xi$／2 乗項 $1/N\Sigma xi^2$／3 乗項 $1/N\Sigma xi^3$／…

⇔「分布関数は $1/N\Sigma xi+1/N\Sigma xi^2+1/N\Sigma xi^3+\cdots$ で捉えられる」

→『母関数 generating function』という発想へ

大体の位置は 1 乗項、散らばりは 2 乗項でうまく捉えられる

⇒3 乗項や 4 乗項にも対応する特性がある

⇒各 n 乗項にもそれぞれ対応する特性がある

⇒各 n 乗項を「合わせる」と分布の全体像に

各 step はそれぞれ集めて測ることだが、step ごとにより一般化が進んでいる
　step3、step4 が成立するところが統計学の面白さ
　step4 が厳密に数式でも成立するところは特にそう
　＝数式を通して考える数理科学

1.4　幅のある判断、確率の導入
　Graunt よりもっと厳密に考えてみた人がいた
　John Arbuthnot　1712 年の論文
　Newton や Swift と同時代人でイギリス人の仇名 “John Bull”の名づけ親
　本当に「1:1」ではないのか
　→統計的検定の第一歩

H0「女性が多い年と男性が多い年は 1/2 の確率で起こる」
D「男性が多い年が 82 年続いている」
⇒H0 の下で D が出現する確率は$(0.5)^{82}$で極めて小さいから H0 であるとは言えない
偶然の誤差を考えた上での判断！
集めて測ることのもう一つの面
もともと集まり＝『分布』は幅のあるもの
Arbuthnot の判定は「めったに起こらないから 1:1 であるとは言えない」≠「1:1 ではない」
しかし、ただ「言えない」というだけではなく、どの程度言えないのかを数値化してみせる
言えない程度を数値化して考える
程度を測るという発想⇒幅とか集まりで判断する良さの一つ！

2. 確率変数として見る
2.1 確率的なモデル
Arbuthnot の判定をもう一度振り返ってみると
H0「女性が多い年と男性が多い年はそれぞれ 0.5 の確率でおこる」
D「男性が多い年が 82 年続いている」
⇒H0 の下で D が出現する確率は$(0.5)^{82}$で極めて小さいから H0 であるとは言えない

H0 という性質
　二つの値｛女性が多い、男性が多い｝をとり
　各種の確率は P（女性が多い）=0.5、P（男性が多い）=0.5
をもつ変数がある
観察されたデータ D＝｛男性が多い、男性が多い、男性が多い、…｝

Arbuthnot の判定：
H0 の性質をもつ変数から D｛男性が多い、男性が多い、…｝が出現したといえるか？
⇒出現する確率が$(0.5)^{82}$できわめて小さいから、そうであるとはいえない

『確率変数 variable』＝「変動する値を確率的にとるもの」
例えば、値は｛女性が多い、男性が多い｝、各値をとる確率は P（女性が多い）＝0.5、P（男性が多い）＝0.5

それぞれ決まった呼び方（定義）があって
『確率変数』X とは、{x1、x2、x3…}という値をそれぞれ確率 P(x=xi)でとるもの
{x1、x2、x3…}は『観測値』または『値』
P(x=xi)をあたえる関数 f(xi)すなわち f(xi)=P(x=xi)となる関数を『確率関数』という

Arbuthnot の判定とは、
　値域{x1、x2}={男性が多い、女性が多い}
　確率関数 f(x1)=0.5、f(x2)=0.5
という確率変数 X があって、観測データ D はその結果であると言えない
⇒出現確率は$(0.5)^{82}$だから、言えない

{男性が多い、女性が多い}も数値化すれば、
確率変数 X
　値域{x1、x2}={1,0}
　確率 f(x1)=0.5、f(x2)=0.5
観測されたデータは{1,1,1,1,1,…}

確率変数はふつう大文字で表し　その値をその小文字で表す

なんで値もデジタル化するの？
→確率関数の特徴をうまく指標化できる

確率変数 X（の確率関数 f(x)）はどのような特徴をもつか？
①確率変数 X が出す値は大体どのくらいか『X の期待値』（略して「期待値」）という
$E(X)=\Sigma x_i \cdot f(x_i)$　→ E は期待 expectation の頭文字
Arbuthnot の判定の例では E(x)＝0.5
②確率変数 X が出す値はどれだけばらつきがあるか、これを「X の期待値周りのばらつきの期待値」とよぶと　面倒なので以下「ばらつき期待値」とよぶと

$E(\text{ばらつき})=\Sigma(xi-E(X))^2\cdot f(xi)$

Arbuthnot の判定の例では

$E(\text{ばらつき})=(1-0.5)^2\times0.5+(0-0.5)^2\times0.5=0.25\times0.5+0.25\times0.5=0.25$

同じように

③歪み期待値

$E(\text{歪み})=\Sigma(xi-E(X))^3\cdot f(xi)$

④尖り期待値

$E(\text{尖り})=\Sigma(xi-E(X))^4\cdot f(xi)$

『分布』の捉え方と発想は同じなので、①は平均②は分散ともいえる
ただし確率変数の期待値やばらつき期待値はその挙動を確率的に示すもので、実際の観測される値（データ）の平均や分散とは別である

期待値やばらつき期待値は、確率関数 f(x)即ち、観測値{x1,x2,x3,…}に確率値{P(x1),P(x2),P(x3),…}を対応させる関係の特徴を指標化したもの
（→指標にするために値をデジタル化した）
期待値やばらつき期待値をみれば、どんな範囲で観測値が出てきやすいかはわかるが　実際に観測されるデータの特徴ではない
なので　平均や分散という表現を使う場合でも、正式には期待値は『母平均』、ばらつき期待値は『母分散』と、『母』をつけて表記する

例えば Arbuthnot の判定であれば
仮定した確率関数の期待値は 0.5 でばらつき期待値 0.25 だから
観測値は平均が 0.5 くらいで、分散は 0.25 くらいになりやすい（＝と期待できる）が、実際には{1,1,1,…}というデータが出ることもあり得る

このデータは平均 1 分散 0、つまり、期待値 0.5、ばらつき期待値 0.25

○期待値と実際の観測値の関係
これも数値例で考えるとイメージしやすい
例えば
ロンドンの性別出生という確率変数 X が
値域｛x1,x2｝＝｛1,0｝
確率関数 f(x1)=0.5、 f(x2)=0.5
にしたがうとしよう

すでに計算したように　この X の期待値は 0.5、ばらつき期待値は 0.25

{1,1,1,1}なら、データの平均 1、分散 0
{1,1,1,0}…{0,1,1,1}なら、データの平均 0.75、…

{1,1,1,1}が出る確率は$(0.5)^4 \times 1$

{0,0,0,0}が出る確率は$(0.5)^4 \times 1$

つまり、データの平均や分散は必ずしも期待値やばらつき期待値と同じ値にはならないが、…

2.2 期待値の意味と意義
確率変数とその確率関数にはいろいろある
一番簡単で一番古く、今もよく使われるものは高校の数学でおなじみ！
「白い球が p、赤い球が q の比率（ただし p＋q=1）で入っている袋から、n 回球をとりだしてそのうち x 個が白い球である確率を求めよ」

これも確率変数とみなす　図に描くと

| | 「x の値」 |
|---|---|
| | →0 |
| 確率変数 X | →1 |
| =「白い球の数」 | →2 |
| | →… |
| | →n |

この確率変数 x は
　値域={x ｜0,1,2,3,4,… n}
　確率変数 f(x)＝

この形の確率関数を「二項分布」、「ベルヌーイ分布」とよび、確率変数を「二項分布にしたがう確率変数」とよぶ
二項分布の特性
　期待値 np、ばらつき期待値 np(1－p)

確率変数の期待値やばらつき期待値は「その変数の基本性能」⇔観測（観察）された値の集まりとは必ずしも一致しない

例えばDもありうる
「十分多い観察」という条件の下で、観測された値の平均や分散は、期待値やばらつき期待値に一致する

期待値やばらつき期待値は独自の演算規則をもつ
期待値 E(・)では
1)E(c)＝c（c は定数 constant を表す）
2)E(X＋c)＝E(X)＋c
3)E(cX)＝cE(X)
4)E(X＋Y)＝E(X)＋E(Y)

ばらつき期待値は通常 V(・)で表すが
V(・)では
1)V(c)＝0
2)V(X＋c)＝V(X)
3)V(cX)＝$c^2$V(X)
4)V(X＋Y)＝V(X)＋V(Y)＋2E{(X－E(X))(Y－E(Y))}
つまり加法則は(X－E(X)),(Y－E(Y))の期待値が 0 になるとき以外は成立しない
E{(X－E(X))(Y－E(Y))}を Cov(X,Y)で表す

●確率変数同士の演算の考え方
確率変数 X と Y の和 X＋Y とは、（X と Y 各々の値 xi と yi とすると）「ありうる全ての値の組み合わせ(xi,yi)に関して、xi＋yi を値とし、xi かつ yi となる（＝xi と yi が同時に成立する）確率 P(xi,yi)を値 xi＋yi の出現確率とする確率変数」
※値域が xi＋yi で定義され、各「xi＋yi」に対してその確率が定義できている（＝確率関数が与えられている）ので…

期待値 E(・)の加法測を…

X は xi,x2、Y は y1,y2 の 2 つの値をもつとする。E(・)は確率変数の値にその出現確率をかけたものだから
E(X＋Y)
=(x1+y1)P(x1,y1)+(x2+y2)P(x2,y2)+(x2+y1)P(x2,y1)+(x1+y2)P(x1,y2)
=…
=x1P(x1)+x2P(x2)+y1P(y1)+y2P(y2)
=E(X)＋E(Y)

で E(・)では加法則が成り立つ

V(・)についてみると
V(X)は確率変数 X の各値 xi から期待値 E(X)を引いたものの 2 乗、即ち(xi－E(X))$^2$ に xi の出現確率 P(xi)をかけたものだから、(xi－E(X))$^2$ を値にする確率変数「(X－E(X))$^2$」の期待値に等しい
同様に、V(X＋Y)は(xi＋yi－E(X＋Y))$^2$ に xi＋yi の出現確率 P(xi,yi)をかけたものだから、
…
※X と Y は確率変数、E(X＋Y)は E(・)の加法測より、E(X)＋E(Y)
…
Cov(X,Y)が正になる場合とは
→x が期待値 E(X)より大きく（小さく）なるときは、y もその期待値 E(Y)より大きく（小さく）なる関係が
Cov(X,Y)が負になる場合とは
→x が期待値 E(X)より小さく（大きく）なるときは、y はその期待値 E(Y)より大きく（小さく）なる関係が
それぞれある場合にあたる

株式でいえば
株 X が暴騰する時は株 Y も暴騰し、X が暴落するときは Y も暴落するような関係にあれば
Cov(X,Y)は正になる　⇒　V(X+Y)>V(X)+V(Y)
株 X が暴騰する時は株 Y は暴落し、X が暴騰するときは Y は暴落するような関係にあれば
Cov(X,Y)は負になる　⇒　V(X+Y)<V(X)+V(Y)

2.2　モーメントと分布指導
もうちょっと一般化してみよう…
確率変数（確率関数）の挙動を示す値を『理論値（母数）parameter』と呼ぶ
E(・)や V(・)はその一つ
※確率変数 X に対して E(X)や V(X)は特定の値をとる。E(・)や V(・)は確率変数を特定の値に変換する演算でもある

E(・)と V(・)だけがわかればその確率関数の挙動は完全にわかるのか？
　⇒一般にはダメ　全く違う確率関数がたまたま同じ E(・)や V(・)をもちうるから

では、ある確率変数の確率関数の挙動を完全に特定できる指標はあるのか？
ある⇒モーメント

確率変数 X における、その値の n 乗 $X^n$ を値とする確率変数 $X^n$ の期待値＝$E(X^n)$を『X の n 次のモーメント』と呼ぶ

各次のモーメントは次のような性質をもつ
①E(X)つまり期待値は 1 次のモーメント
②ばらつき期待値 $V(X)=E\{(X-E(X))^2\}$は 2 次のモーメント
③歪み期待値 $E\{(X-E(X))^3\}$は 3 次のモーメント
④尖り期待値 $E\{(X-E(X))^4\}$は 4 次のモーメント

モーメントは特に『モーメント母関数 moment generating function』で重要になる
モーメント母関数とは各次のモーメントを係数にもつ多項式
$1+E(X)t+\{E(X)^2/(2!)\}t^2+\{E(X)^3/(3!)\}t^3+\cdots$のこと
この t(と X)の関数を n 回微分して t＝0 とすれば n 次のモーメントとなる…

…という簡単な確率関数 $e^{tX}$（e は自然対数の底）の期待値の形に変形できるので、元の確率関数 f(x)よりあつかいやすいことが多い
例えば二項分布…

他にもいろいろ便利な特性があるので　確率変数の特徴を捉える上で強力な道具になる
例えば
1)元の確率関数 f(x)と一対一に対応する
　モーメント母関数が
2)2 つの確率変数 X と Y が独立なら、その和 X＋Y は確率変数で、そのモーメント母関数は X と Y 各々のモーメント母関数の積になる

| | | |
|---|---|---|
| | → | 平均 |
| | → | 分散 |
| 分布 | → | 歪み |
| | → | 尖り |
| | → | … |

という考え方は少し工夫すれば

| | |
|---|---|
| | →1 乗項 |
| | →2 乗項 |
| 分布関数 | →3 乗項 |
| | →4 乗項 |
| | →… |

つまり各次のモーメントがわかれば
　→モーメント母関数がわかり
　→元の確率関数がわかる

2.4　モーメント母関数
n 次のモーメントが特定されればその確率変数の挙動は完全にわかる。つまり確率関数は一意に決まる。これを数式的に表現するのが『モーメント関数』すなわち
　$1+E(X)t+\{E(X^2)/(2!)\}t^2+\{E(X^3)/(3!)\}t^3+\cdots$
という x と t の多項式(t は新たに導入した変数)

2.5 数理統計学のコンセプト
この確率変数という考え方を使った判断の方法が『数理統計学（推測統計学)』
天下り的&おおざっぱに言えば、数理統計学とは
a)観察された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する
b)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、その仮説が観察されたデータにどのくらいあてはまるかを数量的に判定する
a)を『推定』、b)を『検定』とよぶ
例えば Arbuthnot の判定は歴史上最初の検定の例にあたる

もう一つ重要なポイント！
数理統計学では二種類の数値が出てくる
1)観察されたデータ上での数値
2)確率変数の挙動を示す数値
1)を『観測値』、2)を『理論値（母数）parameter』とよぶ

2.1 で述べたように、この二つをきちんと区別できることが大切
二つは違った種類の概念だが、十分多い観察の下では、それぞれに対応する値がある

例えば、E(・):「理論値の一つで、十分多い観察の下では観測値の平均mがとるであろう値」
V(・):「理論値の一つで、十分多い観察の下では観測値の分散 $s^2$ がとるであろう値」

理論値（母数）はふつうギリシア文字で表す
日本語では「母」をつける

例えば
E(・)は$\mu$（ミュー）←平均 mean の頭文字
「m」にあたるギリシア文字 ＝「母平均」
V(・)は$\sigma^2$（シグマ 2 乗）←分散 $s^2$ の平方根
『標準偏差 standard deviation』の頭文字
「s」にあたるギリシア文字 ＝「母分散」

モーメント $E(X^n)$は$\mu 1$、$\mu 2$、…$\mu n$
…

混乱しやすい有名な例
中心極限定理：「平均$\mu$、分散$\sigma^2$の確率変数 Z からの N 個のサンプルの平均 $1/N\Sigma zi$ は、漸近的に平均$\mu$、分散$\sigma^2/N$の正規分布に従う」
…なにがなんやらだが　丁寧に書き直すと
中心極限定理：「期待値 $E(Z)=\mu$、ばらつき期待値 $V(Z)=\sigma^2$の確率変数 Z から出現する N 個の観測値の平均 $1/N\Sigma zi$ は、それ自体が確率変数になり、N が無限大に近づくにつれて、期待値 $E(1/N\Sigma zi)=\mu$、ばらつき期待値 $V(1/N\Sigma zi)=\sigma$ ^2/N の正規分布にしたがう」

3.　確率変数の応用
3.1 連続型確率変数
2.1 で定式化した確率変数は正確には確率変数の一つの型
　⇒『離散型確率変数』とよばれる
確率変数には二つの種類がある
『離散型 discrete』と『連続型 continuous』
離散型は「ある確率で特定の値をとる確率変数」
連続型は「ある確率である幅に値が入る確率変数」いわば幅でおさえる捉え方

ロンドンの出生率でいえば
「ロンドンの女性出生力」という確率変数として
a)離散型で
{x1,x2,x3…}={4688,4457,4102,4590,…}
確率関数 f=…
という変数も想定できるが、
b)連続型で
~3000 が 0.073,~3500 が 0.098
…

例えば
1)区間でみていくとわかることも
ロンドンの女性出生数と男性出生数
…
区間で切る事で規則性が見えることもある

2)連続型の方が数学的にあつかいやすい
実際にはこれが大きい　最初は離散型の近似計算法だった
←コンピュータができるまでは積分で計算する方がずっと簡単だった

連続型確率変数 X とは
$-\infty \leqq x \leqq +\infty$のなかで値をとり
幅 $a \leqq x \leqq b$ になる確率 P（$a \leqq x \leqq b$）が決まっている「何か」＝変数である

このとき P(a≦x≦b)を与える関数 f(x)…

$P(a \leqq x \leqq b) = \int_a^b f(x)dx$ ただし $\int f(x)dx = 1$
（特に限定しない場合∫は$-\infty \sim +\infty$）
を『確率密度関数』とよぶ
これは離散型の確率関数にあたるもの
　和記号Σ‥⇔積分記号∫‥dx
離散型なら
$P(a \leqq x \leqq b) = \Sigma xi_a^b f(xi)$ただし$\Sigma if(xi)=1$

つまりΣ～を∫～dx に置き換えたもの

連続型でも離散型と同じく、期待値やばらつき期待値などの理論値が定義されている
①期待値 $E(X)=\int xf(x)$(特に指定がなければ∫は$-\infty \sim +\infty$まで)
②ばらつき期待値 V(X)
　$E(X-E(X))^2=\int (x-E(X))^2 f(x)dx$
これで数学的にも自然な定義になる
連続型の確率密度関数は指数関数などの形をとる事が多い

連続型で最も有名なのが『正規分布 normal distribution』
正規分布とは　その確率密度関数が
$f(x)=1/(2\pi\ \beta^2)^{0.5} \cdot \exp_e\{-(x-\alpha)^2/(2\beta^2)\}$

e は自然対数の底　$\exp_e$〜は e の〜乗
という形になるもの

期待値は $E(\cdot)=\int xf(x)dx=\alpha$
ばらつき期待値は $V(\cdot)=\int (x-\mu)^2 f(x)dx=\beta^2$

正規分布ではその挙動を示す 2 つの母数（理論値） $\alpha$ と $\beta$ の値とばらつき期待値にちょうどひとしいので、慣例的に
f(x)=1/…

二つの母数がたまたま期待値とばらつき期待値に一致すると考えた方がよい

期待値とちらばり期待値は直接計算もできるが　正規分布のモーメント母関数が
$\exp_e\{\alpha t+(\beta^2 t^2)/2\}$
になることから簡単に求まる
正規分布も確率密度関数よりもモーメント母関数が簡単であるものの一つ

3.2　二項分布を使ったモデル
確率変数というモデルを使った判断を実例で説明しよう
まずモデルに使う確率（密度）関数から
いろいろある　例えば正規分布　連続型の確率密度関数で最も旧くまたよくつかわれる

離散型の確率関数で最も旧くまたよく使われるのが『二項分布』
$f(x)={}_nC_x\cdot(p)^x\cdot(1-p)^{n-x}$ (x は 0 から n までの整数)
二項分布は「あることが生起する確率が p の場合に、n 回やってみてそのうち x 回それが実際に起こる確率」にあたる
母比率 p は理論値なのでふつうギリシア文字を使うが、円周率 $\pi$ と区別するため p を使う

ロンドンの出生比の問題
これも二項分布をモデルに使える
Arbuthnot の判定
1)H0「男性が多い年と女性が多い年はそれぞれ 0.5 の確率で起こる」
2)実際のデータ D は「男性が多い年が 82 年間続いている」
3)H0 が正しければ D が出る確率は $(0.5)^{82}$ これは極めて小さいから H0 だとは言えない

これも二項分布

「男性が多い年が生起する確率が 0.5 の場合に、82 回やってみてそのうち 82 回男性が多い年になる確率」

⇒「白い球と赤い球が 0.5 ずつの袋から、82 回球を取り出して 82 個白い球である確率」と同じ

実際、二項分布の確率関数に n=82、x=82 を代入すれば

$f(x)={}_{82}C_{82}$・…

このモデルはどれだけ適切であろうか、疑問点が二つある

a)出生は本来個体レベルで起こる

→「1:1」は男性が多い年の出現確率というより、出生児一人一人の男女の出現確率

実際、男性の多い年は 14/27 で出現するわけではない

b)男性がどのくらい多いかがわかっている

→男性の比率は 0.5027~0.5362　つまり「男性の比率が 0.5027~0.5362 である」年が 82 年間つづいたという情報を取り込んだ方がより良い

ここで「出生児が男性である確率が 0.5 である場合に、1 年の出生児全体での男性の比率が 0.5027～0.5362 である」確率が求まれば、それが 82 年間続いた確率はその 82 乗

従って、「出生児が男性である確率が 0.5 である場合に、1 年の出生児全体での男性の比率が 0.5027～0.5362 である」確率が計算できればよい

☆やり方 1

二項分布で直接計算する

82 年間のロンドンの平均出生児数を 11429 とすると、「出生児が男性である確率が 0.5 である場合に、1 年の出生児全体での男性の比率が 0.5027～0.5362 である」＝「出生児が男性である確率が 0.5 である場合に、11429 人の出生児の中で男性が 5745～6128 である確率」

⇒「白い球と赤い球が 0.5 ずつの袋から、11429 回球を取り出して、白い球が 5745 個から 6128 個である確率」

これは一つずつのケースを足していけばよい

例えば、5745 個である確率は

${}_{11429}C_{5745}\times(0.5)^{5745}\times(0.5)^{11429-5745}$

同様に、5746 個である確率は…

5747 個である確率は…

6128 個である確率は、${}_{11429}C_{6128}\times(0.5)^{6128}\times(0.5)^{11429-6128}$

まとめて書くと、…

実際に計算した人がいる

結果のみ紹介すると、平均出生児数を 11429 として

$\Sigma_{x=5745}^{6128}{}_{11429}C_x \times (0.5)^x \times (0.5)^{11429-x} = 0.292$

その 82 乗は$(0.292)^{82}$

Arbuthnot の判定では、$(0.5)^{82}$だからそれよりさらに小さい値になる

←より正確にモデル化することでもっと有り得ないと言える

☆やり方 2

近似計算する

「出生児が男性である確率が 0.5 である場合に、1 年の出生児全体での男性の比率が 0.5027～0.5362 である」

一回一回の出生は…

男性を「1」、女性を「0」の値に置き換えると

「袋」　→　1or0

これも二項分布に従う確率変数

変数 Z：値域{男性 or 女性}={1,0}

確率関数 $f(z)={}_1C_x \times (0.5)^x \times (0.5)^{1-x}$

p=0.5 の袋の中から、n=1 回球を取り出して、それが白（＝男性＝1）か赤（＝女性＝0）

Z は二項分布に従うので、その期待値とばらつき期待値は

E(Z)=np=0.5

V(Z)=np(1－p)=0.25

この Z の値を 11429 回観察して、平均 1/11429 Σzi をとったのが「1 年の出生児数 11429 での男性の比率」

1/11429 Σzi は確率変数の観測値の平均

中心極限定理で説明したように、これも確率変数で、観察回数即ち zi の数が大きい場合には近似的に正規分布に従う

中心極限定理「期待値 E(・)=$\mu$、ばらつき期待値 V(・)=$\sigma^2$の確率変数 Z から N 個のサンプルの平均 1/N Σzi は、漸近的に、E(・)=$\mu$、V(・)=$\sigma^2$/N の正規分布に従う」

※この N と二項分布の n を混同しないように

したがって、1/11429 Σzi は

E(・)=0.5

V(・)=0.25/11429=$(0.004679)^2$の正規分布に近似的に従う

だからこういう E と V をもつ正規分布の値が 0.5027～0.5362 の区間に入る確率を求めてやれば、それが 1/11429ΣziがⅠ0.5027～0.5362 の値をとる確率にほぼ等しい

正規分布に関しては値がどの区間にどのくらいの確率で入るかはすでに計算されている
⇒『正規分布表』
→これは、期待値より正または負の方向に u 分
u とは、(値－E(・))/(V(・))$^{0.5}$=(値－$\mu$)/$\sigma$
つまり、「期待値$\mu$からその値がどれだけ離れているかを、標準偏差の期待値$\sigma$を 1 単位として表したもの」
「0.5027」であれば
u=(0.5027－0.5)/0.004677=0.577 だから、表の 0.5 の行の 0.07 と 0.08 の間
ほぼ 0.282
→つまり、0.5027 以上の値が出る確率は、0.282
「0.5362」であれば、
u=(0.5362－0.5)/0.004677=6.97 だから、表に出てこないくらい小さい
ほぼ 0
→つまり、0.5362 以上の値が出る確率は、0

したがって、「出生児が男性である確率が 0.5 である場合に、1 年の出生児全体での男性の比率が 0.5027～0.5362 である」確率は、0.282－0=0.282
やり方その 1 とかなり近い値になる

●出生比 14/27 という仮説の妥当性
同じように、「出生児が男性である確率が 14/27≒0.52 である場合に、17 世紀後半のロンドンぐらいの大きさの集団で、1 年の出生児中の男性の比率が 0.5027~0.5362 である」確率も計算できる

やり方 2 でやると、
まず出生比の確率変数の E(・)と V(・)は
E(Z)=np=0.52
V(Z)=np(1-p)=0.2496
になる
したがって、1/NΣzi は N が十分大きい場合には中心極限定理より
E(1/NΣzi)=0.52
V(1/NΣzi)=0.2496/11429=(0.004637)$^2$
の正規分布にしたがう

あとは正規分布表を見ればよい
「0.5027 以下」の値が出てくる確率は
u=(0.5027－0.52)/0.004673=－3.7 だから　ほぼ 0
「0.5362 以上」の値が出てくる確率は
u=(0.5362－0.52)/0.004673=3.47 だから　ほぼ 0
したがって、「0.5027~0.5362」が出る確率は 0.99~1
その 82 乗は 0.77~1
だから p=14/27≒0.52 という仮説はデータにあてはまらないとはいえない

モデルの使い方で重要な事
(1)あてはめられるモデルは複数ある
例えばロンドンの出生比でいえば
①Arbuthnot のモデルもあれば
②s'Gravesande のモデルもあれば
③s'Gravesande のモデルを連続型でさらに近似したモデルもある
連続型でも離散型でもいい
(2)そのなかで
a)仮説の理論によりあったもの
b)データの情報をより活かせるもの
c)数値計算が楽なもの
を選べば仮説の妥当性についてより適切な評価ができる　あとは
d)汎用性が高いもの　も応用では重要
⇒現在の統計的検定の原型　つまり
ある仮説：「出生児が男性の確率 p が 0.5」あるいは「p が 0.52」
が正しければ
観測値から求められる量：$1/N\,\Sigma\,zi$　が特定の確率分布:E(・)=np、V(・)={np(1-p)}/N の正規分布にしたがう確率変数になることを使ってその量 $1/N\,\Sigma\,zi$ の出現確率の大きさで仮説「p が 0.5」や「p が 0.52」の正しさを評価する手続き
この「観測値から求められる量」を『検定統計量 test statistic』とよぶ

4　統計的推定
4.1　視聴率のしくみ

数理統計学（推測統計学）:『確率変数 Variable』＝「変動する値を確率的にとるもの」を使ったモデルで考えていく営み

いろいろ例があるが　身近な一つが「視聴率」

視聴率にもいろいろある
a)個人視聴率と世帯視聴率
b)瞬間視聴率と番組視聴率

測定法はいくつかありそれぞれの妥当性も重要な問題だが今は無視する
以下では 1 世帯に 1 台 TV があると仮定する

いわゆる「視聴率」とはサンプル調査の視聴率＝サンプル集団での「視ている」比率

『サンプル調査』: 本来調査したい対象から一定のやり方で一部を抜き出して調べる
抜き出したものが『サンプル（標本)』
本来調査したい対象が『母集団』
⇒母平均とか母分散の「母」と同じ意味

視聴率調査でのサンプルの大きさ（『サンプル規模 sample size』）
関東・関西・名古屋で 600 世帯
札幌・仙台・広島・福岡などが 200 世帯
母集団の大きさは例えば関東地区全体の世帯数は 1455 万世帯

少し小さすぎる？サンプル調査で視聴率 30%とか 20%が出てもどのくらいあてになる？
サンプル視聴率は母集団の視聴率をある程度反映するだろうが、本当の視聴率（＝関東地区全体の視聴率）と同じになる保証はない　むしろ違うのがあたりまえ
重要なのは、その精度が計算できる＝数値的に評価できる！

4.2　サンプル調査の精度
注目する点⇒サンプル視聴率の値は二項分布にしたがう確率変数になる
つまり、「関東地区で母視聴率（母比率）が p である場合に、n 個の世帯取り出してそのうち x 個が視ている確率は？」

→番組をみている or 視ていない
『袋』　→番組を視ている or 視ていない
→　…

で n 回取り出してそのうち x 回「番組を視ている」が出る確率を求めるのと同じ

ロンドンの出生比(Arbuthnot の判定)の話では

→男の子 or 女の子

『袋』　→男の子 or 女の子

→　…

で、「出生時は男」の確率が 0.5 や 0.52 の場合に

二項分布の確率関数

$f(x)={}_nC_x \cdot p^x \cdot (1-p)^{n-x}$

を使えば、サンプル規模（n の大きさ）ごとに　サンプル視聴率 x/n の値が出現する確率が計算できる

n=1 すなわち 1 人だけをサンプル集団として調査する場合

サンプル視聴率 x/n は「0%」「100%」どちらかで

「0%」　${}_1C_0 \cdot p^0 \cdot (1-p)^{1-0}$

「100%」　${}_1C_1 \cdot p^1 \cdot (1-p)^{1-1}$

n=2 だと x/n は「0%」「50%」「100%」のどれかで

「0%」　${}_2C_0 \cdot p^0 \cdot (1-p)^2$

「50%」　${}_2C_1 \cdot p^1 \cdot (1-p)^1$

⇒p が与えられればサンプル視聴率 x/n の値の出方は求められる

＝母視聴率が p のときにサンプル視聴率 x/n としてどの値が出やすいかを計算できる

サンプル調査の特性

1)サンプル視聴率 x/n はどんな人がサンプルになるかによって変わる確率変数

2)サンプル視聴率としてどの値がどのくらい出やすいかはサンプル規模 n と母視聴率 p だけで決まる（二項分布だから当たり前だが）

→母集団の大きさは関係ない！

3)サンプル規模が大きくなるにつれて、母視聴率に近い値がサンプル視聴率として出てくる確率が高くなる

4.3 点推定

サンプル調査の精度では、p がわかっているものとしてサンプル視聴率 x/n のどの値がどんな確率で出現するかを求めた

しかし実際のサンプル調査では p は未知数で n と x が既知

知りたいのは母視聴率 p の方！

これを統計学では『推定』という
『推定』:「観測されたデータから確率変数の理論値を推定する」

視聴率調査での理論値は母視聴率（母比率）

サンプル視聴率 x/n からどうやって母視聴率 p を推測できるか？

やり方はいくつかあるが　よく使われるのは
『最尤推定法による点推定』
『点推定』: 特定の値の形で推定する
『区間推定』: 幅（区間）の形で推定する
では最尤推定とは？
―具体的な例で説明しよう
今サンプル集団が 800 世帯、その視聴率が 40%、つまり 320 世帯が視ていたとする
ここで知りたいのは関東地区全体の視聴率つまり母視聴率 p
→あてずっぽうではなく　ちゃんとした手がかりで推論するには？
既知のものから未知のものを知るうえで、一番の手がかりは両者の関係　つまり観測値と理論値の間の関係
→これは二項分布ですでにわかっている
$f(x)={}_nC_x\cdot\cdots$
この場合は
$f(320)={}_{800}C_{320}\cdot p^{320}\cdot(1-p)^{480}$
ではこの式から p の値を測定するには？
いろいろ考え方があるが、一つのやり方は
f(x)が確率を示す点に注目
現実におきたことは「サンプル集団で 320 世帯がみていた」とすればこれに最も近い状態になる p の値を p の推定値にすればよい
「もっともなりやすい＝最も確率が高い」だから f(320)を最大にする p を求めればよい

直感的に言えば
現実に生起した事態は確率 1 だと考えられる
⇒それに一番近いのは f(x)が最大になる状態

これを『最尤推定法』という
具体的な計算は　最大値は極値だから

$f(320)= {}_{800}C_{320} \cdot p^{320} \cdot (1-p)^{480}$
を p の関数として p で微分して=0 とおく

この関数を『尤度関数　likelihood function』というふつう L(p)と表記する
確率関数 f(x)→p が定数で x が変数
尤度関数 L(p)→x が定数で p が変数

計算結果だけ紹介すると、この場合は p＝0.4
つまりサンプル視聴率と同じ値が…

最尤推定法の長所（ゆるい条件付きだが）
1）漸近不偏性：サンプル規模が大きくなるにつれて（n→∞）、$\tilde{a}$期待値が真の値になる　すなわち $E(\tilde{a})=a$
2)漸近有効性：サンプル規模が大きくなるにつれてばらつき期待値 $E(\tilde{a}-a)^2$ が不偏推定値のなかで最小になる
3)一致性：サンプル規模が大きくなるにつれて、推定値$\tilde{a}$が真の値 a に一致する確率が 1 に近づく

短所
1)上の長所が『頑健 robust』ではない：想定している確率関数（尤度関数）が少し違うと成り立たない

●推定値の良さの基準
不偏性や有効性以外にも「良さの基準」はある
例えば『平均平方誤差最小』: ばらつき期待値 $E(\tilde{a}-a)^2$ が最小（有効推定値は不偏推定値のなかで平均平方誤差が最小）

一つの推定量が多くの良さを兼ね備えるとは限らない
例えば $\sigma^2$ の点推定では
1)不偏推定値は $1/(n-1) \cdot \Sigma i(xi-m)^2$
2)最尤推定値は $1/n \cdot \Sigma i(xi-m)^2$　（正規分布などの場合）
3)平均平方誤差最小推定値は $1/(n+1) \cdot \Sigma i(xi-m)^2$

4.4 区間推定
点推定：特定の値すなわち点で推定
区間推定：「この幅（区間）に」の形で推定

大まかだが確からしさを数値的に評価できる

例えば
●やり方その 1
母視聴率 40%でサンプル集団 800 世帯なら「サンプル集団では 38%から 42%の視聴率が出てくる」確率は 0.8
いいかえると
＝「サンプル視聴率が母視聴率±2%の間にある」確率が 0.8
=「母視聴率－0.02≦サンプル視聴率≦母視聴率+0.02 である」確率が 0.8
「」内に注目すると
「母視聴率－0.02≦サンプル視聴率≦母視聴率+0.02 である」は不等式の両辺を移行すると
「サンプル視聴率－0.02≦母視聴率≦サンプル視聴率+0.02」になる

したがって「サンプル視聴率－0.02≦母視聴率≦サンプル視聴率+0.02 である」確率が 0.8

サンプル視聴率 40%=0.4 を代入すれば
「0.4－0.02≦母視聴率≦0.4+0.02 である」確率が 0.8
=「38%≦母視聴率≦42%である」確率が 0.8
だといえる

上の区間推定にはずっと「母視聴率 0.4」という仮定が使われている

この仮定を置かずに求めるには？
●やり方その 2
母比率 p の母集団から標本 1 個を抜き出して
該当=1、非該当=0 とすると、これは
E(・)=p
V(・)=p(1－p)
の二項分布にしたがう

この確率変数からとった n 個のサンプルの値の平均、すなわち標本比率 x/n は
中心極限定理によって
E(・)＝p
V(・)=p(1－p)/n
の正規分布に漸近的にしたがう

正規分布である事より
「x/n が E(・)±V(・)$^{0.5}$ 内」(=「標準偏差の期待値を 1 単位として、期待値± 1 単位分の内」)
にある確率は約 0.70（正規分布表で u=1.0 をみると 0.158）

あとは　やり方その 1 と同じ考え方で
「」内だけ注目してこれを変形する　ただし
その 1 では一次不等式の変形だが　その 2 では二次不等式を解く形になる

「x/n が E(・)±V(・)$^{0.5}$ 内」
→「$|x/n-E(\cdot)| \leqq V(\cdot)^{0.5}$」
→「$|x/n-p| \leqq (p(1-p)/n)^{0.5}$」
→…
→「$(1+1/n)p^2-(2x/n+1/n)p+(x/n)^2 \leqq 0$

…

だから、「$a \leqq p \leqq b$」の確率は 0.70 といえる

a と b はサンプル数 n とサンプル集団の視聴世帯数 x で表されるから、p の区間を n と x だけで（＝母視聴率の値を仮定せずに）推定できる

区間推定における
推定された値のありうる範囲を『信頼区間』
信頼区間にある確からしさを『信頼係数』
とよぶ

例えば
「サンプル視聴率 40%の場合、母視聴率の信頼係数 0.7 の信頼区間は 38%～42%」（やり方 1）
「サンプル視聴率 40%の場合、母視聴率…

区間推定のやり方も複数ある　仮定の置き方がちがう　ロンドンの出生比のときと同じ

やり方その 1 :「母視聴率の値」を仮定
やり方その 2 : 正規分布に従うと仮定（中心極限定理では漸近的に正規分布にしたがうといえるだけ）

他にもやり方はある

統計的推定とは　一定の仮定をおくことで数量的にその良さが評価できる形で推定する
それによって
①どんな仮定を置いたかが明確になる
②仮定の負荷を減らす方向に改良できる

4.5　t 分布をつかった区間推定の例

仮定の負荷を減らせる場合は他にもある
例えば、観測値の出方が正規分布にしたがう、つまり正規分布にしたがう確率変数の観測値から理論値α（母平均μ）を推定する場合
t 分布をつかって推定するのが一般的

この推定法の長所は
１）他の仮定がいらない
２）頑健性が高い

詳しくいえば――
正規分布に従う確率変数の観測値が n 個ある場合、母平均をμ、標本平均をm、標本分散を $s^2$ とすると、

$(m-\mu)/(s/(n-1)^{0.5})$が『自由度 n－1 の t 分布』にしたがう

この$(m-\mu)/(s/(n-1)^{0.5})$を『t 統計量』とよぶ（t 統計量は t 分布にしたがう確率変数）

自由度 n－1 の t 分布がどの幅の値をどんな確率でとるかは t 分布表からわかる
例えば
n が 10 個ならば自由度 9 の t 分布
t 分布表を見ると自由度ν（ニュー）＝9 の行でα＝0.15 の値は 1.1
→$(m-\mu)/(s/(n-1)^{0.5})$が+1.1 より大きくなる確率は 0.15

したがって「$-1.1 \leqq (m-\mu)/(s/(n-1)^{0.5}) \leqq +1.1$」になる確率は 1－0.15×２=0.7 になる

あとは「」のなかを変形していけばよい
「$-1.1 \leqq (m-\mu)/(s/(n-1)^{0.5}) \leqq +1.1$」

→…
→「$m-1.1(s/(n-1)^{0.5}) \leqq \mu \leqq m+1.1(s/(n-1)^{0.5})$」

つまり
母平均$\mu$の、信頼性係数 0.7 の信頼区間は
$m-1.1(s/(n-1)^{0.5})$～$m+1.1(s/(n-1)^{0.5})$
である

数値例でいうと　例えば
観測値が 10 個で
1.4,1.8,0.9,2.1,1.7,1.8,1.5,1.7,1.8,1.9 だとすると

$m=1.66$
$s^2=0.0984$ したがって $s=0.3137$
$1.1(s/(n-1)^{0.5})=1.1\times 0.3137/9^{0.5}=0.115$

つまり
母平均$\mu$の信頼係数 0.7 の信頼区間は $1.66\pm 0.115$ すなわち 1.545～1.775

5. 仮説の検定
5.1　基本的な考え方
推定：データから母集団の値を測定
検定：母集団に関する予想をデータで検証

原型はやはり Arbuthnot の判定
H0：男性が多い年と女性が多い年は 1/2 の確率で生起する
D：男性が多い年が 82 年続いている

H0 の下で D が出現する確率は$(1/2)^{82}$　この値はきわめて小さいから　予想 H0 は外れていると判断する
⇒偶然の誤差を考えた上での判定

具体的にどうするかというと
H0：母視聴率が 0.4
D：サンプル数 200 でサンプル視聴率 0.37
このとき仮説 H0 は外れと判断できるか？

考え方：一定の基準で見て　H0 の下では D は出にくいならば「母視聴率が 0.4」という仮説は×であると判定する

●やり方その 1
D「サンプル数 200 でサンプル視聴率 0.37」が出る確率を直接計算する
=0.4
この形だと出しにくさは評価しにくい
例えば
H0「0.41」なら D の出る確率は 0.03
H0「0.39」なら D の出る確率は 0.05
●やり方その 2
区間推定のように考えて
1)事前に、「母視聴率が 0.4」という仮説が正しければ、サンプル視聴率の値として出てもおかしくない区間を置いて
2)実際に調べたサンプル視聴率がその区間の外ならば「仮説は×」区間の内なら「仮説は×でない」と判断する
例えば
区間として「36～44%」を置けば
もしサンプル視聴率が 46%なら区間外なので、仮説は×
もしサンプル視聴率が 43%なら区間内なので、仮説は×でない
とする　ここで注意すべき点！
「×か×でないか」を判定するもので、「○か○でないか」ではない

区間をどうおけばいいか？
素朴に考えて　区間が狭ければ「×」と判定しやすい

例えばもし「母視聴率が 0.4」という仮説が正しければ　サンプル数 200 だと
区間 A「36～44%」なら　区間内にサンプル視聴率が入る確率は 0.75
⇒実際のサンプル視聴率が区間内でなかったら仮説は×だと判定して、それがまちがいである確率は 0.25
区間 B「34～46%」なら　区間内にサンプル視聴率が入る確率は 0.90
⇒実際のサンプル視聴率が区間内でなかったら仮説は×だと判定して、それがまちがいである確率は 0.1

A か B かはその判定にどれだけのリスクを許すかで決めればよい（決めるしかない）

例えば新薬の効果を調べるなら
H0「差はない」が本当は○なのに「×である」とする確率はかなり小さい方がよいだろう（副作用や経済的コストを考えると）

仮説が本当は○なのに「×である」と判定する確率：『有意水準』あるいは『危険率』
例えば　区間Aは危険率 0.25　区間Bは危険率 0.1 になる

以上まとめると
1)仮説を立てる
2)事前に、一定の危険率を見込んで区間を決める
3)実際に調べた値が 2)の区間に入らなければ「仮説は×である」と判定する

5.2 有意水準と検出力
5.1 で紹介した形ではまだ不十分
⇒危険率 0.25（仮説がただしければその区間内の値をとる確率 0.75）の区間は複数ある！
例えば
「サンプル視聴率が 36～44%」
「サンプル視聴率が 0～42%」
どちらも区間にも確率 0.75 でおちる

☆区間推定の信頼区間も本当は複数ある！

危険率だけならどちらも同じ
でも直感的には「サンプル視聴率が 36～44%」の方が「サンプル視聴率が 0～42%」よりも良い幅のとり方だと思える

なぜ良いと思えるのだろうか？
サンプル数 200 で「母視聴率 0.4」という仮説を危険率 0.25 で検定するとしよう

区間 A の置き方は複数ある　例えば
区間 A1：サンプル視聴率が 36～44%
区間 A2：サンプル視聴率が 0～42%
区間 A3：サンプル視聴率が 37～100%

では三つの区間の置き方は何が違う？
⇒「母視聴率 0.4」という仮説が本当は間違っている場合に違いが出てくる

例えば
本当は母視聴率 0.34 だったりすると
区間 A1 に入る確率は 0.25
⇒「母視聴率 0.4」という仮説は本当は×なのに「×でない」とする確率は 0.25
区間 A2 に入る確率は 0.99
⇒「母視聴率 0.4」という仮説は本当は×なのに「×でない」とする確率は 0.99
区間 A3 に入る確率は 0.17
⇒「母視聴率 0.4」という仮説は本当は×なのに「×でない」とする確率は 0.17
→区間 A3 が一番間違う確率が低い　それゆえ best な置き方

母視聴率 0.46 だったりすると

区間 A1 に入る確率は 0.31
⇒「母視聴率 0.4」という仮説は本当は×なのに「×でない」とする確率は 0.31
区間 A2 に入る確率は 0.14
⇒「母視聴率 0.4」という仮説は本当は×なのに「×でない」とする　確率は 0.14
区間 A3 に入る確率は 0.99
⇒「母視聴率 0.4」という仮説は本当は×なのに「×でない」とする確率は 0.99
→区間 A2 が一番間違う確率が低い　それゆえ best な置き方

区間をどこに置くかで、仮説が本当は×だった場合、×なのに「×でない」とする確率がちがってくる
この「仮説が本当は×なのに「×でない」とする」ことを「第二種の誤り」という

『第一種の誤り』: 仮説は本当は○なのに「×である」とする誤り
『第二種の誤り』: 仮説が本当は×なのに「×でない」とする誤り

危険率は第一種の誤りの確率にあたる　これをふつう $\alpha$ で表す
第二種の誤りの確率はふつう $\beta$ で表す　特に $1-\beta$ を『検出力 power』という

5.3 最強力検定の考え方
検定では事前に（つまり調べる前に）区間の大きさと置き方を決める
⇒第一種と第二種両方の誤りを考慮する必要

一番良い区間とは？
「第一種の誤りの確率（＝危険率）が一定の値 $\alpha$ 以下で、第二種の誤りの確率 $\beta$ が最小（＝

検出力 1－β が最大）の区間」と考えればよい

検出力の大きさは母視聴率の値でかわる
母視聴率がある値の下で検出力が最大
＝『最強力　most powerful』

例えば母視聴率 0.34 なら区間 A3 が、母視聴率 0.46 なら区間Ａ２が最強力
検出力の大きさは母視聴率の値で変わる

⇒検出力は母視聴率を定義域とする関数　になっている
これを『検出力関数』という

母視聴率がどの値でも検出力が最大
＝『一様最強力　uniformaly most powerful』

一様最強力な検定はつねにあるわけではない
例えば A1、A2、A3 の場合にもない
ない場合によく使われるのが『一様最強力不偏　uniformaly most powerful unbiased』検定＝『不偏検定』: 検出力 1－β が危険率 α を下回らない検定法　のなかで一様最強力
A1 は一様最強力不偏検定になっている

先の例で考えてみると　区間 A1 の良さ
①A2 や A3 は母視聴率の値によっては A1 より強力だが、検出力が 0.25 を下回ることも
②（①の言い換え）A1 は仮説の値 0.4 付近で検出力が低い
→本当の母視聴率が仮説の 0.4 に近い場合は仮説を「×ではない」と判定しやすく　遠い場合には「×ではない」と判定しにくい
いわば「当たらずとも遠からず」！

A2 や A3 ではそうはかぎらない
だから本当の母視聴率が検討がつかない場合には区間 A1 型がよい

まとめると　統計的検定とは
1)仮説を立てる
2)事前に、危険率 α を設定して区間の大きさを、検出力 1 － β を考慮して区間の位置を決める

A1 型の置き方を『両側検定』とよぶ
A2 や A3 の置き方を『片側検定』とよぶ
実際には特に理由のない限り『両側検定』、仮説の値以上か以下かどちらかにしかならないなどの特定の場合に『片側検定』
3)実際に調べたデータでの値が 2)での区間の外なら「仮説は×」と判定する

5.4　いろいろな検定法
検定とは何かをおさらいしておくと、ある仮説が正しければ観測値から求められる量が特定の確率分布にしたがう確率変数になることを使ってその量の出現確率の大きさで仮説の正しさを評価する手続き
この「観測値から求められる量」を検定統計量とよぶ

例えばロンドンの出生比の話でいえば
ある仮説：「出生児が男性の確率 p が 0.5」あるいは「p が 0.5 ある」
…

観測値から求められる量：$1/N\,\Sigma\,zi$ が特定の確率分布：$E(\cdot)=np$、$V(\cdot)=(np(1-p))/N$ の正規分布にしたがう確率変数になることを使って、その量 $1/N\,\Sigma\,zi$ の出現確率の大きさで仮説「p が 0.5」や「p が 0.52」の正しさを評価する手続き

母視聴率の検定の例では
仮説：母視聴率つまり母集団での比率
検定統計量：サンプル視聴率
で、特定の確率分布：サンプル視聴率が二項分布、または近似的に正規分布に従うことを使って、…

仮説の種類に応じてどの量を使うか（＝何が検定統計量になるか）は変わってくる
だから、実際に使う上では仮説の種類と検定統計量（＋それがどんな場合にどんな確率分布にしたがうのか）の組み合わせを知っておく必要がある（対応する検定統計量がない場合には検定できない

例えば――
(1)母比率の検定
これは説明済み
仮説の種類は「母集団のある変数の比率 p 」
検定統計量は「サンプル集団でのその変数の比率 x/n」

サンプル集団が無作為抽出でつくられた場合、この検定統計量は二項分布　もしくは中心極限定理より近似的に正規分布に従う

(2)母平均の検定
仮説の種類は「母集団のある変数がとる平均値$\mu$」
検定統計量は「$(m-\mu)/(s/(n-1)^{0.5})$」（仮説により$\mu$の値があたえられるのであとはデータだけから計算できる）
n 個の観測値が正規分布にしたがって出現する場合、この検定統計量は自由度 n－1 の t 分布にしたがう

詳しくいうと――
母平均$\mu$は、標本平均をm、標本分散を $s^2$ で表すと、$(m-\mu)/(s/(n-1)^{0.5})$が自由度 n－1 の t 分布に従う

『t 検定』の一つで単に「t 検定」というとこれを指すことが多い
$(m-\mu)/(s/(n-1)^{0.5})$を『t 検定料』という

●実際の使い方
4.5 で扱った数値例をそのまま使えば、
観測値が 1.4,1.8,0.9,2.1,1.7,1.8,1.5,1.7,1.8,1.9 の 10 個だとする
このデータの下で「母平均$\mu$＝1.4」という仮説を危険率 10%で検定してみよう

観測値が 10 個なので、$(m-\mu)/(s/(n-1)^{0.5})$は自由度 9 の t 分布にしたがう
特に理由がないので両側検定（区間 A1 型）
⇒t 分布表での自由度 9 の危険率 10%の値を見ると 1.833（$\nu$＝9、$\alpha$＝0.10 のところ）
したがって棄却域（そこに入ると仮説は×だと…

あとは計算するだけ
m＝1.66、$s^2$＝0.0984 したがって s=0.3137
$s/(n-1)^{0.5}=0.337/(9^{0.5})=0.115$
$\mu$＝1.4 という仮説の下では
$(m-\mu)/(s/(n-1)^{0.5})=(1.66-1.4)/0.115=2.26$ で棄却域に入る
⇒$\mu$＝1.4 という仮説は棄却される

(3)母分布に関する検定
仮説の種類は「母集団のある変数の k 個の値の分布 pi(p1,p2,・・・pk)」

検定統計量は「Σi(fi－npi)² /npi」（仮説から pi があたえられるのであとはデータでのその変数の値の分布 fi(f1,f2,・・・fk)から計算できる）
観測される分布が正規分布にしたがって出現する場合、この検定統計量は『自由度 k－1 のカイ 2 乗分布』にしたがう

詳しくいうと――
母分布を p1,p2,…pk(Σipi=1)、観測された分布を f1,f2,…fk(Σifi=n)で表すと、Σi(fi－npi)² /npi は自由度 k-1 のカイ 2 乗分布に従う

『カイ 2 乗検定』とよばれるものの一つΣi(fi－npi)²/npi を『カイ 2 乗統計量』とよぶ

サンプル集団が無作為抽出でつくられた場合の母集団でのある変数の値の分布 pi とサンプル集団でのその変数の値の分布 fi でも
Σi(fi－npi)² /npi は近似的にカイ 2 乗分布に従う（←中心極限定理によりサンプル比率は近似的に正規分布にしたがって出現するので）

●実際の使い方その１：適合度の検定
メンデルの仮説（メンデルの法則）
「黄/緑と丸い/皺皺はそれぞれ 3:1 の比率で出現して相互がどくりつである」具体的に言えば「{黄・丸、黄・皺、緑・丸、緑・皺} の母分布は {9/16、3/16、3/16、1/16}」
実際のデータは {315、101、108、32}
このデータの下でメンデルの仮説を危険率 0.1(=10%)で検定してみると
分布の個数が 4 なのでΣi(fi－npi)² /npi は自由度 3 のカイ 2 乗分布にしたがう
Σi(fi－npi)² /npi は正の値しかとらないので片検定

したがって棄却値は 6.25 以上

あとは計算するだけ 母分布 pi は {9/16,3/16,3/16,1/16}、観測された分布 fi は {315、101、108、32} だから
Σi(fi－npi)² /npi =0.470
だから棄却域には入らない
⇒メンデルの仮説は棄却されない

実は危険率 0.9 でも（つまり×でないのに×とする可能性が 90%という極めて棄却しやすい基準でも）メンデルの仮説は棄却されない
データにきわめてよくあてはまる仮説だった

…教科書的にはそういうことだが　このケースはなにか違和感がある

（理由その１）出来すぎのデータ
かえって怪しげ　→　捏造の疑いあり　真実は神のみぞ知るだが

（理由その２）仮説検定の論理的特性
検定とは本来、ある仮説を、それが正しいと考えた場合「大目にみても」あり得ない結果がデータで出ていることを示すことで否定する手続き
Arbuthnot の判定もそうだった
⇔メンデルの仮説のケースは肯定したい仮説

こういう場合どうするかはいろいろな立場があって、共通見解があるとはいえない
→仮説検定の適用限界

●実際の使い方その２：独立性の検定
サンプル集団で変数 X(xi)と変数 Y(yi)のクロス集計の値が fij だった場合、母集団において変数 X と変数Ｙが独立かどうかを検定できる

「独立である」という仮説が正しければ、母集団における値の分布 pij について
$pij=(fi./n)\times(f.j/n)=(fi.\times f.j)/n^2$
が成り立つ（後述）
それゆえ、$\Sigma ij(fij-npij)^2/npij$
$=\Sigma ij(fij-fi.f.j/n)^2/(fi.f.j/n)$
はカイ 2 乗分布にしたがう
ただし、この時の自由度は ij－1 ではなく…

例えば、男性か女性かと死後の世界を信じるかどうかは関係あるかを調べたところ、サンプル調査では次のような結果になった

| | 信じる | 信じない |
|---|---|---|
| 男性 | 350 | 100 |
| 女性 | 400 | 150 |

このとき「性別と死後の世界を信じる信じないは独立である」という仮説を危険率（有意水準）0.05 で検定してみよう

性別を変数 X（1＝男性、2＝女性)、信じるかどうかを変数 Y（1＝信じる、2＝信じない）とすると
上の表は
f11=350,f…
このとき
f1.=f11+f12=460,f2.=f21+f22=550
f.1=f11+f21=750,f.2=f12+f22=250
だから
$\Sigma_{ij}(fij-npij)^2/npij$
$=\Sigma_{ij}\{(fij-fi.f.j/n)^2/(fi.f.j/n)\}$
$=(f11-f1.f.j/n)^2/(fi.f.1/n)+\cdots$
＝3.37

自由度 1 のカイ 2 乗分布の危険率（有意水準）0.05 の値は 3.84、したがって独立であるとは言えない

●独立性の検定と適合度の検定のちがい
独立性の検定は適合度の検定の特殊ケースとどこがちがうか？
通常の適合度の検定は理論値が既知
例えば、メンデルの仮説が正しければ色と形の比率はこうなる、など
独立性の検定では理論値は未知
独立という仮説の下で最尤推定法によって母集団の値の分布を求めると
pij=(fi./n)×(f.j/n)
がなりたつ
ただしこの推定の際に pij の値のとり方にある条件を与えるため自由度が(i－1)(j－1)になる

5.5 検出力関数の意味
検出力関数とは？
―この関数は検出力とともに危険率を表す
…以外に思うかもしれないが
理論値（例えば母比率）がどこにあろうと判定手続き自体は同じ　つまりサンプル視聴率が特定の区間外なら「仮説は×」だと判定する
それゆえ、「本当の母視聴率が 0.4001 である場合のサンプル視聴率の出方」≒「本当の母視聴率が 0.4 である場合のサンプル視聴率の出方」である以上、「本当の母視聴率が 0.4001 である場合に仮説を×とする確率」≒「本当の母視聴率が 0.4 である場合に仮説を×とする確率」であるはず

より一般的な形で言えば
「本当の母視聴率θが仮説の値θ0（例えば 0.4）にごく近い場合に仮説を×とする確率」
≒「本当の母視聴率θが仮説の値θ0である場合に仮説を×とする確率」

したがって、本当の母視聴率θがθ0にごく近い場合には検出力1－β≒危険率α（母視聴率がθ0である場合に「仮説は×」とする確率）になる

検出力関数の形状をみれば　危険率と検出力がどんな関係にあるかがわかる
例えば
(1)最強力検定「危険率がα以下で、検出力1－βが最大」とは
＝「母視聴率がθ0ならB(θ)はα以下（すなわちB(θ0)≦α）をみたす検出力関数で、母視聴率がある値の時B(θ)が最大のもの」
(2)一様最強力検定「危険率がα以下で、母視聴率がどの値でも検出力1－βが最大」

検定結果からどんな判断ができるか　も検出力関数からある程度いえる
→検出力関数の形状を頭にいれておくと　いろんなことが明確に整理できる

検出力関数の形状の基本的な特徴
1)B(θ)は(θ0、α)を必ず通る
θ0とαは検定を使う人が決める
2)αは通常低い値をとる（例えば0.1や0.05）
3)検出力1－βはできるだけ高い方がよい

したがって、望ましい形状はθ0の近くでB（θ）が急激に凹む形　凹み方は急激であるほうがよい

凹み方を決める要因は三つ
a）仮説が本当は×の場合に統計量が従う確率分布＝「非心分布」の種類
b）危険率α
c）サンプル規模n
例えば　仮説の母比率＝0.4の場合　区間Ａ１型＝両側検定の検出力は…

もしθ0の近くでB(θ)が急激に凹であれば
仮説検定で「×でない」という結果から「仮説は○」と判断しても大外れはない
なぜならば前に説明したように

$\theta 0$ 近く（$\theta \fallingdotseq \theta 0$）では検出力は$\alpha$程度しかない、つまり本当は仮説は×なのに「×でない」と判断する確率が高いが、そもそも$\theta \fallingdotseq \theta 0$だから「仮説は○（＝真の値は$\theta 0$)」だと判断しても大きな誤りにはならない　「当たらずとも遠からず」！

逆に言えば、B($\theta$)は$\theta 0$の近くで急激に凹にならなければ
⇒「仮説が×でない」という結果から…

●片側検定と両側検定
「当たらずとも遠からず」
⇒使う目的によっては　仮説「$\theta = \theta 0$」は本当は×だけれど「×でない」と判断してかまわない場合　もあるということ
検出力関数の形状で言えば
$\theta$が特定の区間ではB($\theta$)＝0でよい場合

いくつかの可能性
①$\theta \fallingdotseq \theta 0$なら「$\theta = \theta 0$」と等価である場合
いいかえると
$\theta$は大体$\theta 0$辺りだとわかれば問題ない場合
例えば多くの経験的な研究ではそう　知りたい値が大体$\theta 0$辺りだとわかればよい
②$\theta > \theta 0$なら「$\theta = \theta 0$」と等価である場合
いいかえると
$\theta$が$\theta 0$より低ければ問題だが、$\theta 0$より高ければ特に問題ない場合
例えば一定の学力未満なら不合格にすべきケースなど
③$\theta < \theta 0$なら「$\theta = \theta 0$」と等価である場合
いいかえると
$\theta$が$\theta 0$より高ければ問題だが$\theta 0$より低ければ特に問題ない場合
例えば一定の濃度以上だと人体に害がある排出物の出方を監視するケース

①なら$\theta \fallingdotseq \theta 0$の区間はB($\theta$)＝0でよい
⇒両側検定を使う
サンプル規模が小さいときは危険率$\alpha$を小さくしすぎない方がいい
②なら$\theta > \theta 0$の区間　③なら$\theta < \theta 0$の区間はB($\theta$)＝0でよい
⇒0でよい区間の反対側に棄却域をおくやり方
＝片側検定（区間A2やA3型）を使う
なぜならば0でよい区間以外では一様最強力になるから