基礎統計

furaga

はじめに

すでにすばらしい基礎統計のシケプリ[*1]がネット上に存在する今日、絶対需要ないよなーとか思いながらも自己満で作りました。しかしどのシケプリにも他にはない特徴が一つや二つはあるものです。このシケプリの特徴は以下の二点。

1. 統計学入門の解答例がある。たとえ間違いだらけだったとしても、考え方の指針は大体あってると思うので、少しは参考になると思います。
2. このシケプリ、TEX[*2]というもので書かれているのですが、そのソースコードを公開している。
   来年以降、下級生の方々がシケプリを作るのが少しだけ楽になりましょう。（シケ対が TeX を使うなら、ですが）

なので、この用語解説はあんまり見る価値ないかもしれませんが、問題の解答例は一瞬だけ斜め読みして一笑に付すだけの価値はあろうかと思います。
あと、TEX が使える人は、間違いなどがあれば積極的に修正してアップしなおしましょう。（シケプリはクラス共有のものだと思っています）

お断りなど

- 自分の担当は 7 章から 12 章であり、それ以前の章のシケプリ制作に関しては微塵も責務を有しておりません。
- 後になるにつれて説明が雑＆適当になっている気がしますが、それは仕様です。
- 誤字・脱字・説明不足などが横行していると思われますが、そうしたものを見つけた場合の対処法を示しておきます。

  1. 自分の心のうちにそっとしまいこむ。（推奨）
  2. 一緒にアップされているソースコードを編集・修正する→コンパイル→ dvi ファイルを pdf ファイルに変換→クラスのホームページにアップしなおす。（TeX 経験者のみに推奨）
  3. 掲示板・メーリングリストなどで僕に知らせる。（多分やらないと思うので、あまり推奨しない）

[*1] なんで今でも「プリント」っていうんでしょうね。
[*2] 数式などをとてもキレイに書けるソフト。

7 多次元の確率分布

7.1 同時確率分布と周辺確率分布

- 同時確率分布

  二つの離散型の確率変数 X,Y をベクトル (X,Y)[*3]と表したとき、X=x かつ Y=y である確率。

$$P(X = x, Y = y) = f(x, y)$$

  と定義する。また、事象 A が起こる確率 P(A) は

$$P((X, Y) \in A) = \sum \sum_{A} f(x, y)$$

  で求められます。

- 同時確率密度関数

  上の同時確率分布の積分バージョン。確率変数 X,Y が連続型のときのもの。

$$P((X, Y) \in A) = \int \int_{A} f(x, y) dx dy$$

- 周辺確率分布

  離散型確率変数 X、Y 単独の確率分布。要する (?) に、条件が「X=x」(または「Y=y」) だけしかなくて、Y（または X）の値は別に何でもいいよーってときの確率。同時確率分布の式から、

$$g(x) = P(X = x, Y : \text{すべての実数}) = \sum_{y} f(x, y)$$

$$h(y) = P(X : \text{すべての実数}, Y = y) = \sum_{x} f(x, y)$$

- 周辺確率分布関数

  上の周辺確率分布の積分バージョン。（X,Y が連続型）

$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

$$h(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

- 共分散

  確率変数 X、Y が関連しながらばらつく程度を表したもの。X,Y の関係の方向[*4]を表す。

$$\begin{aligned} Cov(X, Y) &= E\{(X - \mu_X)(Y - \mu_Y)\} \qquad (\mu_X = E(X), \mu_Y = E(Y) \\ &= E(XY) - E(X)E(Y) \qquad (\text{実際の計算ではこちらを使う}) \end{aligned}$$

[*3] つまり、X、Y はそれぞれ xy 座標系の x 座標、y 座標の値

[*4] 正の相関関係があるか、負の相関関係があるか、など

と定義されます。

- 相関係数

  共分散を X,Y の標準偏差で割って、X,Y の関係の強さの程度を判断できるようにしたもの。

$$\begin{aligned}\rho_{XY} &= \frac{Cov(X,Y)}{D(X)D(Y)} \\ &= \frac{Cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}} \qquad (V(X), V(Y)：それぞれ\ X, Y\ の分散)\end{aligned}$$

  なお、$-1 \leq \rho \leq 1$。

  $0 \leq \rho$ なら、X, Y と同じ大小の向きに変化し、$\rho \leq 0$ ならその逆になる傾向があります。ここでいう傾向とは平均的・確率的なもので、ρ の絶対値が大きいほど、その傾向は確定的になります。

  特に $\rho = \pm 1$ のとき Y=aX+b という一次式が成立します。(ただし $\rho = 1$ なら $a > 0$、$\rho = -1$ なら $a < 0$)

- 無相関

  $\rho_{XY} = 0$ であるとき、「X と Y は無相関である」といいます。独立とは似て非なるもの。

- 独立

  任意の x,y について、条件

$$f(x) = g(x)h(y)$$

  が成り立つとき、X,Y は互いに独立であるといいます。

  独立ならば無相関ですが、無相関でも独立だとは限りません (独立⇒無相関)

7.2 条件付確率分布と独立な確率変数

- 条件付確率

  あらかじめ、何かしらの条件が与えられた後に、ある事象が起こる確率。

- 条件付確率密度関数

  Y=y という条件が与えられたときの X の条件付確率密度関数を

$$g(x|y) = \frac{f(x,y)}{h(y)}$$

  X=x という条件が与えられたときの Y の条件付確率密度関数を

$$h(y|x) = \frac{f(x,y)}{g(x)}$$

  と定義します。

- 条件付期待値・条件付分散

  条件付確率における、期待値と分散。

  Y=y と与えられたとき、X の条件付期待値、条件付分散は、

  X，Y が離散型のとき、

$$E(X|y) = \mu_{X|Y} = \sum_x x \cdot g(x|y) \tag{1}$$

$$V(X|y) = \sum_x (x - \mu_{X|Y})^2 g(x|y) \tag{2}$$

  連続型のとき、

$$E(X|y) = \mu_{X|Y} = \int_{-\infty}^{\infty} x \cdot g(x|y) dx$$

$$V(X|y) = \int_{-\infty}^{\infty} (x - \mu_{X|Y})^2 g(x|y) dx$$

- 独立[*5]

  任意の x,y について、条件

$$f(x) = g(x)h(y)$$

  が成り立つとき、X,Y は互いに独立であるといいます。

  このとき、

$$g(x|y) \equiv {}^{*6}g(x), \qquad h(y|x) \equiv h(y)$$

  が成り立ちます。（条件付確率の式 (1)(2) に $f(x) = g(x)h(y)$ を代入すれば出てきます）

- 積の期待値

  X,Y が独立のとき、積 XY の期待値 E(XY) について、

$$E(XY) = E(X)E(Y)$$

  が成り立ちます。

- 独立と無相関の関係

  上の式から、X，Y が独立のとき、

$$Cov(X,Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0 \quad \text{より、}$$

$$\rho_{XY} = \frac{Cov(X,Y)}{D(X)D(Y)} = 0$$

  よって、独立　⇒　無相関。逆は一般に成り立ちません。

[*5] 大事そうなので、あえて二回書きました。

[*6] 両辺は同値ですよって意味

- 独立のときのモーメント母関数 X,Y が独立ならモーメント母関数について、

$$M_{X+Y}(t) = X_X(t)M_Y(t)$$

7.3 独立な確率変数の和

- 期待値、分散の加法性確率変数 $X_1,\ X_2,\ \cdots,\ X_n$ について、
  期待値は独立の如何にかかわらず、つねに

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

  が成立します。一方分散は $X_1,\ X_2,\ \cdots,\ X_n$ が独立の時に限り、

$$V(X_1 \pm X_2 \pm \cdots \pm X_n) = V(X_1) + V(X_2) + \cdots + V(X_n)$$

  特に $X_1,\ X_2,\ \cdots,\ X_n$ が同一の確率分布に従うとき、その期待値・分散を μ、σ^2 とすれば、

$$E(X_1 + X_2 + \cdots + X_n) = n\mu, \qquad V(X_1 + X_2 + \cdots + X_n)n\sigma^2$$

  であり、標準偏差は、$D(X_1 + X_2 + \cdots + X_n) = \sqrt{n}\sigma^2$。したがって、**標準偏差は $\sqrt{n}$ に比例**する。
- 相加平均 $X_1,\ X_2,\ \cdots,\ X_n$ の平均 $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$ の期待値・分散はそれぞれ、

$$E(\overline{X}) = \mu, \qquad V(\overline{X}) = \frac{\sigma^2}{n}$$

  したがって、変数の数 n が大きくなるにつれて、分散は小さくなり、0 に収束します。(→大数の法則)
- たたみこみ
  独立な二つの確率変数 X,Y(それぞれの確率分布を g(x),h(y) とする) において、$X + Y(= z)$ の確率分布 k(z) は、

$$k(z) = \int_{-\infty}^{\infty} g(x)h(z - x)$$

- 再生性確率変数 X,Y が同一種類の確率分布にしたがっているとき、
  g,h のたたみこみの結果、ふたたび同一種類の確率分布 k がえられるとき、その確率分布は再生的であるという。再生的な確率分布の例：二項分布・ポアソン分布・正規分布など

8 大数の法則と中心極限定理

- 大数の法則
  試行回数をを多くすると、観測結果が真の値に近づく、という法則。
  → 大標本では、観測された標本平均を真の平均値（母平均）とみなせます。

- 中心極限定理
  母集団分布が何であれ、確率変数の和の確率分布は、n が大きくなるにつれて正規分布に近づきます。
  つまり、母集団分布の平均、分散（母平均、母分散）を μ, σ^2 とすると、確率変数の和
  $S_n = X_1 + X_2 + \cdots + X_n$ は、$N(n\mu, n\sigma^2)$ に従い、ゆえに、
  $\overline{X} = \frac{X_1+X_2+\cdots+X_n}{n}$ は、$N(\mu, \frac{\sigma^2}{n})$ に従います。つまり、n を十分大きくすると、標本平均 $\overline{X}$ は母平均 μ に限りなく近づきます。 → 大数の法則

- コンピュータ・シミュレーション
  コンピュータでシミュレーションすること。
  しばしば乱数が使われる。Excel なら「= RAND()」と書けば乱数を発生させられます。

- ベルヌーイ試行
  一回の実験で 2 種類の事象のいずれかが生じ、しかもそのような事象が常に一定の確率で起こるような試行のこと。
  例：コインを投げて裏表を見る試行[*7]

[*7] たまにコインが立つこともあるけど無視しましょう。

9 標本分布

- 母集団
  自分が持っている標本のデータから知りたいと思う集団全体のこと。
  (例) 日本人の意識調査を行う場合　→　日本人全体が母集団。

- 標本
  母集団から分析のために選び出された要素・属性値のこと。

- 標本抽出
  母集団から標本を選び出すこと。

- 統計的推論
  母集団について何か知りたくても、現実には不可能なことがあります。例えば、
  1. 母集団が非常に多く (無限大の場合もある) の要素からなる場合。
  2. 全体の調査が意味を持たなかったり、予算上の問題から全数の調査が無理な場合。
  3. 将来に起こるため、現在は測定が不可能な要素を含む場合。

  そんなとき、
  1. 母集団からその一部を選び出し（標本抽出をして)、
  2. 標本をを分析して、
  3. 母集団について推測する。

  ということが行われ、これを統計的推論といいます。

- 母集団分布
  母集団の確率分布。得られた標本は、母集団分布に従う確率変数だと考えます。統計的推論の最終目標は、これらの標本をほげほげして母集団分布を求めること。

- 標本の大きさ
  標本の数。同一の母集団分布 $f(x)$ に従う独立な確率変数の数とも。

- パラメトリックの場合
  いくつかの定数さえわかれば、母集団分布についてすべて知ることができる場合。
  つまり母集団分布が、既知の確率分布（正規分布・ポアソン分布・一様分布などなど）であるとわかっている場合。

- 母数
  パラメトリックの場合の、求めるべき定数（パラメータ）のこと。
  (例)
  1. 正規分布 $N(\mu, \sigma^2)$ における、平均 (期待値)μ と、分散 σ^2。

2. ポアソン分布 P(λ) における λ。

- ノン・パラメトリック
  パラメトリックじゃない場合。もとい、いくつかのパラメータだけでは母集団分布を決定できない場合。この場合、平均・メディアン・モード・分散・レンジ・歪度・尖度などを調べて、母集団分布の形状を考えていきます。

- 復元抽出と非復元抽出
  母集団から標本を抽出する際、一度抽出した要素を再び母集団に戻すかどうかという話。もとに戻す抽出方法を復元抽出、戻さない方法を非復元抽出といいます。
  前者と後者では、組合せの数などが微妙に違ってくるため得られる数値も変わります。しかし、取りだす母集団の要素の数 N が標本の大きさ n と比べて十分大きいなら、どっちの方法でもほとんど差はないので、手間のかからない非復元抽出がよく行われるようです。

- 単純ランダム・サンプリング
  母集団から標本を選び出す方法のひとつ。要素数 N の母集団から、n 個の標本を抽出するとき、母集団の各要素が標本として選ばれる確率が等しく n/N になるように選ぶ方法。乱数がしばしば使われます。
- 母平均
  母集団分布 $f(x)$ の平均。

$$\mu = \int_{-\infty}^{\infty} x f(x) dx \qquad \text{あるいは} \qquad \mu = \sum_{x} x f(x)$$

  とかけます。母数の一つ。

- 母分散
  母集団分布の分散。

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \qquad \text{あるいは} \qquad \sigma^2 = \sum_{x} (x-\mu)^2$$

  とかけます。母数の一つ。

- 統計量
  標本を要約し、母集団の母数のいろいろな推測に使われる数値のこと。
  (例)　標本の平均、分散、標準偏差、メディアン、最小値、最大値、相関係数などなど。

- 統計分布
  統計量の確率分布。

- 標本平均
  標本の平均　$\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$　のこと。
  母平均 μ を知るのが現実的に難しい場合、標本平均を代用する。その心は、標本平均の期待値 $E(\overline{X}) = \mu$

であり、標本の数が大きくなるほど、$\overline{X}$ は μ に確率収縮するから（大数の法則)。統計量の一つ。

- 標本分散
  標本の分散。

$$s^2 = \frac{1}{n-1}(X_1 - \overline{X}) + (X_2 - \overline{X}) + \cdots + (X_n - \overline{X})$$

  注意すべきは、分母が n ではなく n-1 であること。その心は、計算すると、s^2 の期待値　$E(s^2) = \sigma^2$ だから。
  母分散 σ^2 の不偏推定量、または不偏分散という。統計量の一つ。

- 偏りのある標本分散

$$S^2 = \frac{1}{n}(X_1 - \overline{X}) + (X_2 - \overline{X}) + \cdots + (X_n - \overline{X})$$

  のこと。このとき、期待値は $\mathrm{E}(S^2)$ は、

$$E(S^2) = \frac{n-1}{n}\sigma^2$$

  となり、実際の母分散の値より少し小さい値が出る（σ^2 の過小評価が起こる)。統計量の一つ。

- 標本和の標本分布
  パラメトリックの場合で、母集団分布が再生性を持つ場合、標本和の確率分布は結構簡単に求まります。
  (例)
  1. 二項母集団母集団分布が母数 (片方の事象が起こる確率)p のベルヌーイ分布なら、標本分布は二項分布 Bi(1,p) に、和 $X_1 + X_2 + \cdots + X_n$ は二項分布 Bi(n,p) に従います。
     (例)　製品に含まれる不良品の数・社会調査法におけるある事項に関する賛否

  2. ポアソン母集団母集団分布がポアソン分布 $P_0(\lambda)$ のとき、$X_1 + X_2 + \cdots + X_n$ はポアソン分布 $P_0(n\lambda)$ に従います。
     (例)　交通事故死亡者数

  3. 正規母集団母集団分布が正規母集団 $N(\mu, \sigma^2)$ のとき、$X_1 + X_2 + \cdots + X_n$ は正規分布 $N(n\mu, n\sigma^2)$ に従います。($\overline{X}$ は正規分布 $N(\mu, \sigma^2/n)$ に従う)
     (例)　測定誤差

- 漸近的正規性
  標本平均の分布は、n が十分の大きければ正規分布で近似でききます（中心極限定理）

- 有限母集団修正以上はすべて母集団の大きさが無限大の無限母集団についての話でしたが、母集団の大きさ N があまり大きくないときや、n/N が大きい場合、以上の内容をそのまま適用するのは無理があ

ります。そこで、母集団の大きさが有限であることを考慮して、修正を行う必要があります。
有限母集団における、標本平均の期待値・分散をそれぞれ $E(\overline{X})$、$V(\overline{X})$ とすると、

$$E(\overline{X}) = \mu \qquad V(\overline{X}) = C_N \frac{\sigma^2}{n} = \frac{N-n}{N-1}\frac{\sigma^2}{n}$$

このとき、　$C_N(=\frac{N-n}{N-1})$　を有限母集団修正といいます。

10 正規分布からの標本

- 正規標本論
  正規分布に従う確率変数 $X_1, X_2, \cdots, X_n$ から得られる統計量の標本分布を計算するための理論。

- 正規分布

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

  のこと。ガウスさんが体系付けたんだって。[*8]
  また、平均 $\mu = 0$、分散 $\sigma^2 = 1$ のとき、とくに**標準正規分布**といい、N(0,1) とかきます。
  さらに、X が正規分布 $N(\mu, \sigma)$ に従っているとき、

$$Z = \frac{X - \mu}{\sigma}$$

  は、標準正規分布 N(1,0) に従います。
  ちなみに、正規分布においては、平均　＝　メディアン　＝　モード

- X_kが $N(\mu, \sigma^2)$ に従うとき、$\overline{X}$ は $N(\mu, \sigma^2/n)$ に従い、$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ は N(0,1) に従います。単独の X より、標本平均 $\bar{X}$ の方が分散が小さくなるのですぐれた測定値となります。[*9]

- χ^2 分布[*10]
  ときに、標本分散というのは、$s^2 = \frac{1}{n-1}\sum_n (X_n - \bar{X})^2$ で表せるわけだけど、さらに母集団が正規分布であると仮定すれば、標本分散 s^2 の標本分布を求めて少し幸せな気分になれるらしい。
  そこで、Z_k を標準正規分布 N(0,1) に従う確率変数として、

$$\chi^2 = {Z_1}^2 + {Z_2}^2 + \cdots + {Z_k}^2$$

  を定義する。このとき、確率変数 x^2 は、自由度 k の x^2 分布 $x^2(k)$ [*11]に従うという。

- χ^2 分布と標本分散との関係。
  統計量

$$\chi^2 \equiv \frac{(n-1)s^2}{\sigma^2} (\text{標準化})$$

  は自由度 n-1 のカイ二乗分布に従う。
  (例)　$\mu = 50, \sigma^2 = 25, n = 10$

$$P(s^2 > 50) = \quad = 0.035$$

[*8] ド・モアブル、ラプラスがその前に発見してたけど、ちゃんと体系づけたのはガウスさんが初めてだったんだって。
[*9] これを反覆数の原理というらしい。
[*10] エックスではなくカイ
[*11] 確率密度関数は、ガンマ分布 $G(\frac{k}{2}, \frac{1}{2})$ と同じ

- 分散が未知　⇒　何か値を与えなければならないので、σ^2 のかわりに s^2 を用います。

- スチューデントの t 統計量
  標準化変数 Z の代わりに
  $$t = \frac{\bar{X} - \mu}{\sqrt{s^2/n}}$$
  を使い、これは自由度 n-1 の t 分布 t(n-1) に従う。
  t(n-1) の (n-1) は、正規分布 N(0,1) における (0,1) と同じノリ。断じて、t に n-1 をかけるという意味ではありません。

- 1. Z は標準正規分布 N(0,1) に従う。
  2. Y は自由度 k の χ^2 分布 $\chi^2(k)$ に従う。
  3. Z と Y は独立である。
  以上の条件を満たすとき、
  $$t = \frac{Z}{\sqrt{Y/k}}$$
  は自由度 k の t 分布 t(k) に従います。

- 自由度
  値がどうとでもとれる変数の数らしい。t(k) などの k。自由度が大きくなるにつれ t 分布は標準正規分布に近づきます。

- 二標本問題
  異なる二種の標本による母集団の比較を扱う問題。二つの標本の平均を $\bar{X}, \bar{Y}$ として、$\bar{X} - \bar{Y}$ を調べたり。

- 標本平均の差
  母平均の差 $\mu_1 - \mu_2$ の確率分布を知るには、標本平均の差 $\bar{X} - \bar{Y}$ を見ればよい。
  母分散に依存するので、場合別に考える必要があります。

  $\overline{X}, \overline{Y}$ の平均、母分散をそれぞれ、$(\mu_1, {\sigma_1}^2/n), (\mu_2, {\sigma_2}^2/n)$ とおくと、

  1. 母分散が両方ともわかっているとき、
     $\overline{X} - \overline{Y}$ は、$(\mu_1 - \mu_2, ({\sigma_1}^2/n) + ({\sigma_2}^2/n))$ に従う。
     このとき、標準化変数 Z は
     $$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{({\sigma_1}^2/m) + ({\sigma_2}^2/n)}}$$
     となります。[*12]

[*12] 基本的に、推定とか検定とかは標準化変数に対して行うので、暇があれば標準化変数に直すようにしてるらしい。

2. 母分散が未知だけど等しいと分かっているとき
   母分散が分からないので、標本分散を使うわけですが、このとき次の**合併した分散**を使う。

$$s^2 = \frac{\sum_{i=1}^{m}(X_i - \overline{X})^2 + \sum_{j=1} n(Y_j - \overline{Y})^2}{m+n-2} = \frac{(m-1){s_1}^2 + (n-1){s_2}^2}{m+n-2}$$

   ${s_1}^2, {s_2}^2$ は X,Y の標本分散。m、n は X,Y の標本の大きさ。
   ところで、$\overline{X} - \overline{Y}$ の標準化変数は、($\sigma_1^2 = \sigma_2^2 = \sigma^2$ とすると)

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right)\sigma^2}}$$

   しかし、σ が分からないので、代わりに上の s を使うと、

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{s\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

   これを推定や検定に使います。

3. 母分散が未知で等しいとは限らない（等しいかどうか分からない）とき
   この場合、正確に母分散の差の分布は求められないそうですが、ウェルチの近似法なるものを使って、近似的に求められます。
   ウェルチの近似法によると、統計量

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

   は、自由度について

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}}$$

   に最も近い整数 ν^* の t 分布 $t(\nu^*)$ に近似的に従います。

- F 分布
  上記のように、二つの母集団分布の分散が等しいかどうかが結構大事です。
  そこで標本分散の比 $\frac{{s_1}^2}{{s_2}^2}$ を調べると、そういったことが（確率的に）分かってちょっと幸せになれます。その標本分散の比を調べるのに使われるのが F 分布です。

  ある 2 つの標本分散 ${s_1}^2$, ${s_2}^2$ があり、それぞれの標本の大きさがm、nであるとする。このとき、確率変数 F を

$$F = \frac{{\sigma_1}^2}{{\sigma_2}^2}\frac{{s_1}^2}{{s_2}^2} \qquad (\sigma_1{}^2、\sigma_2{}^2は各標本の母分散)$$

  と定義すると、F は自由度 (m-1,n-1) の F 分布 $F(m-1, n-1)$ に従います。
  特に、${\sigma_1}^2 = {\sigma_2}^2$ のとき、F 分布は標本の分散比

$$F = \frac{{s_1}^2}{{s_2}^2}$$

の確率分布になります。

- 標本相関係数
  二種類の観測値 X,Y があって、その相関係数を調べたいことがあります。それぞれの標準偏差と共分散が分かっているときは、

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

  で求められますが、分からないときは以下のように、

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

  で代用します。ただし、$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$（標本共分散)。

- フィッシャーの z 変換
  ところで、母相関係数 ρ_{XY} に対しての標本共分散 r_{XY} の標本分布を求めるのは大変難しいそうです。
  (これができないと、r_{XY} の扱いがむずかしくなります)
  しかし、フィッシャーの z 変換を使うと近似的にその標本分布が求まります。
  これは、z, η を、

$$z = \frac{1}{2} log \frac{1+r}{1-r}$$

$$\eta = \frac{1}{2} log \frac{1+\rho}{1-\rho}$$

  とすると、z は n が大きいときに正規分布 $N(\eta, \frac{1}{n-3})$ に従うというもの。
  ただし、$r = r_{XY}, \rho = \rho_{XY}$。

  このことから、$\sqrt{n-3}(z-h)$ は n が大きいとき標準正規分布 N(0,1) に従うことがわかるので、そこからいろいろ分析を進めることができます。

11 推定

- 点推定
  母集団の母数の推定法のひとつ。
  標本から得られた、なにかしらの一つの値で母数を推定する。
  （例）標本平均 $\overline{X}$ で μ を推定する。

- 区間推定
  真のパラメータの値が入る確率が $1-\alpha$ 以上となる区間を求める、つまり、

$$P(L \leq \mu \leq U) \geq 1-\alpha$$

  なる L、U を求めるものです。

- 推定量
  母集団の母数を推定するために標本から求められる統計量。
  （例）母平均 μ に対する標本平均 $\overline{X}$。母分散 σ^2 に対する標本分散 s^2。

- 推定値
  標本として具体的に n 個の観測値が与えられたとき、これを代入して計算される値（具体的な数値）。
  つまり、推定量に値を入れて電卓で計算した結果のこと。

- モーメント法
  k 次のモーメント μ_k の推定量を $\hat{\mu}_k = \sum {X_i}^k/n$ で推定することを考えます。
  これを k=1,2,3,$\cdots$ について行って、以下のような連立方程式を作ります。

$$\begin{aligned}
\hat{\mu}_1 &= \sum {X_i}^1/n \\
\hat{\mu}_2 &= \sum {X_i}^2/n \\
\hat{\mu}_3 &= \sum {X_i}^3/n \\
&\cdots
\end{aligned}$$

  これらから、各母数を求めていくというのが一般的なモーメント法です。[*13]
  (例) 分散 σ^2 は、$\sigma^2 = \mu_2 - {\mu_1}^2$（定義）で求められる。

[*13] 一般的じゃない方法はしらない。教科書に載っていない。

- 最尤法[*14]

  最尤原理「現実の標本は確率最大のものが実現した。」[*15]という原理を使います。

  たとえば、表になる確率 p、裏になる確率 1-p のコインを 5 個投げた時、表が 4 回出たとします。高校数学チックに考えると、表が 4 回出る確率を L(p) とすると、

$$L(p) = 5p^4(1-p)$$

  と書けます。このとき、L(p) は、p のいろいろな値におけるもっともらしさを表す関数と見ることができ、尤度関数（ゆうどかんすう）と呼ばれます。この尤度関数を最大にする p の値（上の例では p=0.8）を最尤推定値 (場合によっては、最尤推定量) といいます。この最尤推定値こそが求める推定値だとほざいているのが最尤原理です。

- 点推定の基準

  点推定は上のような方法を使ったりして行うそうですが、こうした推定を行う際に「どの統計量を推定量として使うか」というのが大事になります。

  実用上、推定量になるためには以下のような条件を満たしていることが望ましいそうです[*16]。

  1. 不偏性

     推定量の期待値が、真の母数の値となること。($E(\hat{\theta}) = \theta$)

     （例）$\overline{X}$、s^2 は不偏推定量。なぜなら、

$$E(\overline{X}) = \mu, \quad E(s^2) = \sigma^2$$

     また、S^2 は不偏推定量ではないので、推定量として使うのは望ましくない。

  2. 一致性

     標本の大きさ n が大きくなるに従い、真の母数の値に近づくこと。これを満たす推定量を一致推定量と言います。

     （例）$\overline{X}$ は μ の一致推定量。[*17]

  3. 漸近正規性

     n が大きいときに、正規分布に従うとみなせること。これを満たす推定量を漸近正規推定量といいます。

  4. 有効性

     ほかのどんな不偏推定量よりも分散が小さいこと。そのような推定量を有効推定量（または最小分散不偏推定量）という。

[*14] 「さいゆうほう」と読む。

[*15] 「意味がわからないですね☆」by 安藤先生

[*16] 教科書の書き方を見ると、絶対に満たしてないといけないってわけでもないようですが。

[*17] 私は真に驚くべき証明を見つけたが、この余白はそれを書くには狭すぎる (嘘。チェビシェフの不等式と大数の法則から自明ということにしてください。

（例）母集団分布が $N(\mu, \sigma^2)$ のとき、標本平均 $\overline{X}$ は μ の有効推定量。

しかし、実際には、有効推定量を見つけるのは相当うざい作業らしいので、かわりに漸近的有効推定量というものが使われます。これは、n が十分大きいときに分散が最小になる推定量のことです。（この性質を漸近有効性といいます）最尤法で求めた推定量（最尤推定量）は、一般に漸近的有効推定量。

- 点推定の例

最尤法を使って、いくつかの分布に関する推定量を求めます。

- 正規分布 $N(\mu, \sigma^2)$

  正規分布の定義より、観測値 $X_1, X_2, \cdots, X_n$ が得られた場合、尤度関数は、

$$L(\mu, \sigma^2) = \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp(-(X_i - \mu)^2/2\sigma^2)$$

  両辺をログって、

$$logL(\mu, \sigma^2) = -nlog(\sqrt{2\pi}\sigma) - \sum_{i=1}^{n}(X_i - \mu)^2/2\sigma^2$$

  尤度関数の最大値を求めたいので、このログった式を μ, σ^2 でそれぞれ偏微分してイコール 0 とおきます。それを解くと、

$$\hat{\mu} = \sum_{i=1}^{n} X_i/n = \overline{X}, \qquad \hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/n$$

  しかし、分散については、得られた答えは不偏なので、実際には、標本分散が使われます。
  なお、モーメント法でやっても結果は同じです。

- 二項分布

  母集団分布が、標本 $X_i = 0, 1$ を出す母数 p のベルヌーイ分布 Bi(1,p) の二項分布のとき同様に解くと、p の推定量は

$$\hat{p} = \overline{X}$$

  モーメント法でも結果は同じ。

- ポアソン分布 $Po(\lambda)$

$$\hat{\lambda} = \overline{X}$$

  モーメント法でも結果は同じ。

- 一様分布（区間 (a,b)[*18]）

  最尤法では、

[*18] このとき、a,b が母数

$$a = MinX_1, X_2, \cdots, X_n, \qquad b = MaxX_1, X_2, \cdots, X_n$$

モーメント法では、

$$a = \overline{X} - \sqrt{3}S, \qquad b = \overline{X} + \sqrt{3}S \quad （S \text{ は標本標準偏差}）$$

– ノンパラメトリックの場合

母集団分布の式が分からないので、最尤法は使えない。

モーメント法で求めた結果は、

$$\hat{\mu} = \sum_{i=1}^{n} X_i/n = \overline{X}, \qquad \hat{\sigma}^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2/n$$

しかし、分散については、得られた答えは不偏なので、実際には、標本分散が使われます。

- 区間推定

真のパラメータの値が入る確率が $1-\alpha$ 以上となる区間を求める、つまり、

$$P(L \leq \theta \leq U) \geq 1 - \alpha$$

なる L、U を求めるものです。[*19]

このとき、L,U はそれぞれ、下側信頼限界、上側信頼限界といい、$1-\alpha$ は信頼係数と呼ばれ、区間 $[L, U]$ を信頼区間と呼びます。$1-\alpha$ の値はたいてい、0.99 か 0.95 に設定されることが多いようです。

- いろいろな母集団分布に関する区間推定

信頼係数 $1-\alpha$ としたときの、信頼区間を求めます[*20]。

– 正規分布

平均 μ、分散 σ^2 の信頼区間はそれぞれ、

$$\left[\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

（分散が分からないときは $Z_{\alpha/2}$を $t_{\alpha/2}(n-1)$ に置き換える）

$$\left[\frac{(n-1)s^2}{\chi_{\alpha/2}(n-1)}, \quad \frac{(n-1)s^2}{\chi_{1-\alpha/2}(n-1)}\right]$$

[*19] 大事なことなので 2 回書きました

[*20] 心の余白が足りないので証明は略。申し訳ないですけど教科書見て下さい（226 ページくらい）

また、二つの母集団分布がそれぞれ $N(\mu_1, {\sigma_1}^2), N(\mu_2, {\sigma_2}^2)$ であり、

・${\sigma_1}^2 = {\sigma_2}^2$ のとき、母平均の差 $\mu_1 - \mu_2$ の信頼区間は、

$$\left[\overline{X} - \overline{Y} - t_{\alpha/2}(m+n-2)s\sqrt{\frac{1}{m}+\frac{1}{n}}, \quad \overline{X} - \overline{Y} + t_{\alpha/2}(m+n-2)s\sqrt{\frac{1}{m}+\frac{1}{n}}\right]$$

・${\sigma_1}^2 \neq {\sigma_2}^2$ のときは、

$$\left[\overline{X} - \overline{Y} - t_{\alpha/2}(\nu^*)\sqrt{\frac{{s_1}^2}{m}+\frac{{s_2}^2}{n}}, \quad \overline{X} - \overline{Y} + t_{\alpha/2}(\nu^*)\sqrt{\frac{{s_1}^2}{m}+\frac{{s_2}^2}{n}}\right]$$

さらにこのとき、母分散の比 ${\sigma_1}^2/{\sigma_2}^2$ の信頼区間はは、

$$\left[F_{1-\alpha/2}(m-1, n-1){s_2}^2/{s_1}^2, \quad F_{\alpha/2}(m-1, n-1){s_2}^2/{s_1}^2\right]$$

– 二項分布 Bi(1,p)

p の信頼区間を定義から求めるのはしんどいので、nが大きいときは中心極限定理から正規分布で近似して考える。

すると、pの信頼区間は近似的に、

$$\left[\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

と求められます。

– ポアソン分布 $Po(\lambda)$

同様に、近似的に λ の信頼区間は、

$$\left[\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

と求められます。

12 仮説検定

12.1 検定の考え方

- 仮説検定
  母集団について仮定された命題を、標本に基づいて検証すること。

- 有意性検定
  標本から得られた数値と、理論的に求めた数値との差に、なにか意味が有るのか調べること。

- 採択
  仮説が正しいと判断すること。

- 棄却
  仮説が間違っていると判断すること。

- 有意水準
  もし、仮説が正しいときにその事象が起こる確率がどの程度以下になったら希少だと考えるか、という基準。α で表します。(α =1,5,10 %とおくことが多い)

  (例) コイン投げ
  「コインに歪みがない」と仮定すると、表が１４回出る確率は 5.8 %くらいになる。
  もし、有意水準を１%または５%とおくと仮説は棄却されないが、１０%とおくと仮説は (起こった事象があまりに希少なため) 棄却される。

- 帰無仮説
  はじめに立てる仮説。上の例で言うと、「コインに歪みがない (表ができる確率 $p = 1/2$)」という仮説。H_0 で表す。

- 対立仮説
  帰無仮説に対立する仮説。帰無仮説が棄却されると、これが代わりに採択される。上の例で言うと、「コインに歪みがある ($p \neq 1/2$)」という仮説。H_1 で表す。

- 第一種の誤り
  本当は帰無仮説が正しいのに、誤って棄却してしまうこと。

- 第二種の誤り
  本当は帰無仮説が間違っているのに、誤って採択 (正しいと判断) してしまうこと。

- 検定統計量
  検定に用いる統計量。平均・分散・標準偏差など。

- 棄却域
  検定統計量の値がこの領域内だったら、帰無仮説を棄却しますよ、という領域。

- 受容域
  検定統計量の値がこの領域内だったら、帰無仮説を採択 (受容) しますよ、という領域。

- 両側検定
  ある範囲より大きくても小さくても棄却する検定方法。

- 片側検定
  ある範囲より大きければ OK で、小さいときは NG（あるいは、ある範囲より小さければ OK で、大きいときは NG）とする検定方法。

12.2 正規母集団に関する仮説検定

- 母平均に関する仮説検定
  母平均の仮説 (例：「$\mu = \mu_0$である」) の検定をするとき、以下のように条件を設定する。
  1. 両側検定・母分散が既知のとき
     - 帰無仮説　$H_0 : \mu = \mu_0$
     - 対立仮説　$H_1 : \mu \neq \mu_0$
     - 有意水準　α
     - 棄却域　　$R = \{Z||Z| > Z_{\alpha/2}\}$

     と設定する。ただし、Z は検定統計量で、

     $$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

     と定義される。[*21]

  2. 両側検定・母分散が未知のとき
     - 帰無仮説　$H_0 : \mu = \mu_0$
     - 対立仮説　$H_1 : \mu \neq \mu_0$
     - 有意水準　α
     - 棄却域　　$R = \{t||t| > t_{\alpha/2}\}$

     と設定する。ただし、t は検定統計量で、

     $$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

[*21] 標準化変数と微妙に違うので気をつけた方がいいと思いました。

と定義される。

3. 片側検定・母分散が既知のとき
   - 帰無仮説　$H_0 : \mu = \mu_0$
   - 対立仮説　$H_1 : \mu > \mu_0$（右片側検定）
   - 有意水準　α
   - 棄却域　　$R = \{Z | |Z| > Z_\alpha\}$

   と設定する。ただし、Z は検定統計量で、

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

   と定義される。

4. 片側検定・母分散が未知のとき
   - 帰無仮説　$H_0 : \mu = \mu_0$
   - 対立仮説　$H_1 : \mu > \mu_0$（右片側検定）
   - 有意水準　α
   - 棄却域　　$R = \{t | |t| > t_\alpha\}$

   と設定する。ただし、Z は検定統計量で、

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

   と定義される。

- 母分散の検定
  1. 母分散の検定（両側検定）
     - 帰無仮説　$H_0 : \sigma^2 = {\sigma_0}^2$
     - 対立仮説　$H_1 : \sigma^2 \neq {\sigma_0}^2$
     - 有意水準　α
     - 棄却域　　$R = \{\chi^2 | \chi^2 < \chi^2_{1-\alpha/2}(n-1), \chi^2 > \chi^2_{\alpha/2}(n-1)\}$

     と設定する。ただし、χ^2 は検定統計量で、

$$\chi^2 = (n-1)\frac{s^2}{{\sigma_0}^2}$$

     と定義される。

2. 母分散の検定（片側検定）

・右片側検定のとき、

– 帰無仮説　$H_0 : \sigma^2 = {\sigma_0}^2$

– 対立仮説　$H_1 : \sigma^2 > {\sigma_0}^2$

– 有意水準　α

– 棄却域　　$R = \{\chi^2 | \chi^2 > \chi^2_{\alpha/2}(n-1)\}$

と設定する。ただし、χ^2 は検定統計量で、

$$\chi^2 = (n-1)\frac{s^2}{{\sigma_0}^2}$$

と定義される。

・左片側検定のとき、

– 帰無仮説　$H_0 : \sigma^2 = {\sigma_0}^2$

– 対立仮説　$H_1 : \sigma^2 < {\sigma_0}^2$

– 有意水準　α

– 棄却域　　$R = \{\chi^2 | \chi^2 < \chi^2_{1-\alpha/2}(n-1)\}$

と設定する。ただし、χ^2 は検定統計量で、

$$\chi^2 = (n-1)\frac{s^2}{{\sigma_0}^2}$$

と定義される。

- 母平均の差の検定

基本的に、μをμ_1に、μ_0をμ_2 に置き換えて考えます。ただし、分散について、

・分散が等しいとき、${\sigma_1}^2 = {\sigma_2}^2 = \sigma^2$ とおいて計算。

・分散が等しくないとき、ウェルチの近似法を使って考える。

- 母分散の比の検定（両側検定のときのみを示す）

– 帰無仮説　$H_0 : {\sigma_1}^2 = {\sigma_2}^2$

– 対立仮説　$H_1 : {\sigma_1}^2 \neq {\sigma_2}^2$

– 有意水準　α

– 棄却域　　$R = \{F | F < F_{1-\alpha/2}(m-1, n-1), F > F_{\alpha/2}(m-1, n-1)\}$

ただし、

$$F = \frac{{s_1}^2}{{s_2}^2}$$

12.3　いろいろな χ^2 検定

検定統計量は χ^2 分布に従うので、そこから適合度や独立性を調べることができる。

- 適合度検定
  観測値が、理論値と近ければ適合しているとする検定方法。

- 独立性検定
  観測度数が、独立だと仮定したときの理論値に近ければ、独立だと考えましょう、という検定方法。

12.4 中心極限定理を用いる検定

検定統計量が近似的に正規分布に従うなら、正規分布のときとまったく同じ方法で検定を行うことができます。
（例）母比率の検定
確率変数 $X_i(=0,1)$ が Bi(1,p) に従うとき、検定統計量

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

は、nが大きいとき、標準正規分布 N(0,1) に従う。

12.5 検出力

- 検出力
  $1-\beta$:帰無仮説が真でないときにちゃんと棄却する確率。（第二種の誤りを犯さなさい確率）
  検出力が大きいほど、検定方法は良いものであると評価されます。