基礎統計 (安藤)2011 年度夏学期シケプリ Ver.2.16

小松憲人

平成 23 年 7 月 27 日

はじめに

過去のシケプリはたくさんありますが、このシケプリは次のような特徴を持っています。

- 用語集・公式集形式であること
- 2011 年度夏学期の基礎統計 (奇数)(安藤雅和) の授業に準拠し、重要度別に分類している
- 著者の TEX の練習
- シケプリ史上初 (著者調べ) のミニゲーム付き。このページのどこかに説明が書かれています。

次のように重要度を分けています。

- ★★：少なくとも覚えておきたい用語、公式
- ★：覚えておくとよい用語、公式
- ☆：授業でちょっと触れた程度の用語、重要な公式を導くのに使ったが直接はあまり意味のない公式、補足的な言葉の説明など
- ・：授業では触れたが、試験範囲ではないと明言されたもの

☆まで一応分かっておくとよいと思います。個人が勝手につけたので重要度が低いのが試験に出ても怒らないでくださいね。
このシケプリには定理の証明は全くついていません (打ち込むの大変だし)。証明を知りたい人は教科書を見てください。代わりに直感的な説明らしいものがありますが、全然厳密ではないので覚えるためのヒントとでも思ってください。

第 I 部
記述統計

記述統計とは、集団の特徴を記述するために、各個体から得られたデータを整理、要約する方法。平均をとるなど、一般人から見て一番普通な統計。

★★度数分布表

データを大きさに応じて範囲 (**階級**) に分け、分類したもの。階級は値の小さいものから第 n 階級という。
階級上限と階級下限の中間値を**階級値**、階級に属するデータの数を**度数**、第 n 階級までの度数の和を**累積度数**、(度数)/全データ数 を**相対度数**、第 n 階級までの相対度数の和を**累積相対度数**という。

★モード (最頻値)　Mo

度数が最大となる階級の階級値

★★ヒストグラム

度数分布表を柱状グラフにしたもので、横軸には階級値、縦軸には度数をとる。
データ分布の峰が 1 つのとき単峰、複数のとき多峰という。
また、左側に度数が高い階級が集中した分布を**右に歪んだ分布**、右側に集中した分布を**左に歪んだ分布**という。これは直感と逆になるので注意。右から押された形が右に歪んだ分布だと考えるとよい。

★★平均　$\bar{x}$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

まあこれは大丈夫でしょう、と思うかもしれないが次に示すようないろいろな性質がある。単位は元のデータと同じ。

★線形性

定義は数学 II を参照。基礎統計では、線形性のある関数 $f(x)$ が次の条件を満たすことが使われる。

$$f(ax+b) = af(x) + b$$

平均や**期待値**は線形性を持つ。平均の 1 つめの性質。

☆偏差　$x_i - \bar{x}$

データと平均の差。偏差の和は 0 になる。平均の 2 つめの性質。

☆平均の 3 つめの性質

$$\sum_{i=1}^{n}(x_i - c)^2 \text{は } c = \bar{x} \text{で最小}$$

このことを、$\bar{x}$ は最小 2 乗値であるという。

★メディアン (メジアン・中央値・中位値)　Md

データを大きい順に並べたとき、ちょうど真ん中のデータの値。全データ数が偶数のときは、真ん中の前後にある 2 つの値の平均となる。

☆モード・メディアン・平均の関係

分布が左右対称のとき、3 つの値は一致する。分布が左右非対称のとき、峰がモードに当たり、なだらかな方に下っていくと順にメディアン、平均がある。右に歪んだ分布だったら モード < メディアン < 平均、左に歪んだ分布だったら反対。

★★分散　S^2

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \text{偏差 2 乗の平均} = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$$

データの散らばりを表す指標。等式の最後の辺は計算を簡単にするのに役立つ。ただし単位は元のデータの 2 乗。
次の関係がある

$$y_i = ax_i + b \text{ のとき } S_y = a^2 S_x$$

散らばりだから全体にある数を足しても変わらず、全体が 2 乗されているから係数は 2 乗になる。当然だが分散は負にはならない。負だったら計算ミス。

★★標準偏差　S

$$\sqrt{S^2} = S$$

記号そのまんまの量。単位が元のデータに戻っているから、平均と足し引きしたりできるようになった。しかし直接標準偏差同士を散らばりの度合いとして比較してはならない (変動係数参照)
次の関係がある。

$$y_i = ax_i + b \text{ のとき } S_y = aS_x$$

分散と同じように全体に何かを足しても標準偏差は変わらないが、係数はそのまま残る。平方根とったから。

☆平均偏差

$$\frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

これから一度も現れない、意味があるんだか無いんだかよく分からない偏差。標準偏差以上になる性質があるらしい。

★k シグマ区間

区間 $[\bar{x} - kS, \bar{x} + kS] = [(\text{平均}) \pm k \times (\text{標準偏差})]$ のこと。k シグマ区間にどのくらいの割合のデータが含まれているかで、散らばりの度合いが分かる。

☆チェビシェフの不等式

$$k \text{ シグマ区間に含まれるデータの割合} \geq 1 - \frac{1}{k^2}$$

☆範囲 (レンジ)

データの最大値と最小値の差。そのまんま。

☆第 k 四分位点

n 個のデータを小さい順に並べたとき $kn/4$(に最も近い整数) 番目のデータの値。

$$\text{第 } 2 \text{ 四分位点} = Md$$

☆四分位範囲

$$\frac{\text{第 3 四分位点} - \text{第 1 四分位点}}{2}$$

★変動係数　CV

$$CV = \frac{S}{\bar{x}}$$

標準偏差は平均に比例する値より、それを平均で割って無名数 (単位のない値) としたもの。データ同士の散らばりの度合いを比較するときは変動係数で比較する。

☆基準化変量 (標準化変量)　z_i

$$z_i = \frac{x_i - \bar{x}}{S} = \frac{\text{データ} - \text{平均}}{\text{標準偏差}}$$

あるデータが全体の中でどのような位置にあるかを示す。このように平均を引いて標準偏差で割ることを基準化 (標準化する) といい、確率分布において重要になる。
ちなみに $10z_i + 50$ が偏差値である。偏差値はマイナスも 100 以上もある。100 人中 1 人が 100 点で残りが 0 点だったらその人の偏差値は 149.5、逆だったら -49.5 になる。50 万人が受けるセンター試験で 1 人が 900 点、残りが 0 点だったら偏差値は 7121。

★散布図

2 つの変数の関係を示すために、データを平面上にプロットしたもの。

★★相関

2 つの変数の関係を表すもので、2 つの変数が独立しており、一方が他方に影響を与えることがないときに用いる。
散布図が右肩上がりの傾向にあるとき**正の相関**、右肩下がりの傾向にあるとき**負の相関**を持つといい、関係が見られないとき**無相関**であるという。

★共分散　S_{xy}

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 2\text{変数の偏差の積の平均}$$

相関係数の重要な部分だが、回帰分析でも重要。
正の相関があると $(x_i - \bar{x})(y_i - \bar{y})$ の 2 つの因数が同符号になりやすく $S_{xy} > 0$、負の相関があると異符号になりやすく $S_{xy} < 0$ となる。単位は元のデータの 2 乗。

★★相関係数　r_{xy}

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\text{共分散}}{2\text{変数の標準偏差の積}}$$

標準偏差で割るのは標準化しているからと考えられる。常に $-1 \leq r_{xy} \leq 1$ で、$0 < r_{xy} \leq 1$ のとき正の相関、$-1 \leq r_{xy} < 0$ のとき負の相関。
特に $r_{xy} = \pm 1$ のとき**正 (負) の完全相関**といい、散布図ですべての点が 1 つの直線上に乗っている。無名数である。

・偏相関係数　$r_{xy\cdot z}$

$$r_{xy\cdot z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}}$$

小学生の足の大きさと学年には正の相関関係がある。また、学年と覚えている漢字の数にも正の相関関係がある。従って足の大きさと覚えている漢字の数には相関関係があるように見えるが、相関関係があると考えるのはおかしい。これを**見かけ上の相関**という。
この係数は、中間の変数 (ここでは学年) の影響を取り除いて本当の相関を調べるために用いられる。x, y が相関関係を知りたい変数 (足の大きさと漢字の数)、z が中間の変数である。

★★回帰

相関と違い、一方の変数がもう一方の変数に影響を与える場合、その背後には具体的な関数で表される関係があると予想される。そこで、2 変数 x, y の間に関数 $y = f(x)$ で表される関係があるとし、データから $f(x)$ を分析する。これを**回帰分析**という。
真の関係式 (これを求めることはできない) を $y = \beta_0 + \beta_1 x$ とする (β_1 が x の係数となることに注意) と、データは誤差を含んでいるので、

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

と表される。これを**回帰モデル**といい、β_0, β_1 を**回帰係数**、ϵ_i を**誤差項**という。

★★最小２乗法

回帰分析の一つの方法。

$$f(b_0, b_1) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

が最小となる $(b_0, b_1) = (\hat{\beta}_0, \hat{\beta}_1)$ が最もよくデータを要約すると考える方法。このとき得られる関数 $y = \hat{\beta}_0 + \hat{\beta}_1 x$ を**回帰直線 (推定回帰式)** といい、$\hat{\beta}_0, \hat{\beta}_1$ を**最小２乗推定値 (最小２乗値)** という。$\hat{}$ は推定値であることを示す。
あくまでも予測であるから $\hat{\beta}_0, \hat{\beta}_1$ は β_0, β_1 とは異なり、回帰直線は２変数の平均の点を通る。

★★最小２乗推定値　$\hat{\beta}_0, \hat{\beta}_1$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2} = \frac{\text{共分散}}{x\text{の分散}},\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$\hat{\beta}_0$ は、$\hat{\beta}_1$ と回帰直線が２変数の平均の点を通ることから導かれる。

★残差　$\hat{\epsilon}_i$

回帰直線に x_i を代入することで**推定値** $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1$ を求められるが、これも予測値であるから実際のデータ y_i とは異なる。この差を残差といい、$\hat{\epsilon}_i = y_i - \hat{y}_i$ となる。記号から分かるように誤差項の推定値版である。
残差について次の式が成り立つ。

$$\sum_{i=1}^{n} \hat{\epsilon}_i = 0,\ \sum_{i=1}^{n} x_i \hat{\epsilon}_i = 0$$

左の式は誤差をできるだけ小さくしているのだから当然。0 でなかったら最小２乗推定値を変えろよってことになる。

☆決定係数　R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\text{残差２乗和}}{y\text{の偏差２乗和}} = \frac{S_{xy}^2}{S_x^2 S_y^2}$$

実際のデータがどのくらい回帰直線にあっているのかを表す係数で、$0 \leq R^2 \leq 1$、大きいほど当てはまっている。最後の辺は計算に使える。
実は相関係数の 2 乗 r_{xy}^2 に等しい。

第 II 部
確率論

初めは普通の確率だが、期待値や分散と記述統計と対になる概念が登場する。後半は、あるデータをさいころの出た目のように確率変数の実現値として考え、次項の統計的推測の準備となる定理をまとめていく。
後半は今までの確率と同じものと考えて進めていくと混乱しやすいが、統計的推測で実際の問題が解ければいいのであまり重要ではない分野かもしれない。

☆統計解析

母集団から抽出した標本 (データ) から、母集団の性質を推論すること。記述統計に対し、確率論と統計的推測はこの分野に入る。

☆標本空間　Ω

試行の結果に起こりうる全体の集合、つまり全事象。すべての事象は Ω の部分集合になる。

☆余事象　A^c

記号だけ。試行とか和事象 $(A \cup B)$ とか積事象 $(A \cap B)$ とか互いに排反とかは略。

★★条件付き確率　$P(A|B)$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

B が起きたことが分かっている条件下で A が起こる確率。ちょっとわかりにくいが、今までの確率で散々使っている。
10 本中当たりが 3 本あるくじを B くんと A くんが順に引く。このとき 2 人とも当たる確率は、

$$\frac{3}{10} \times \frac{2}{9} = \frac{1}{15}$$

ここでの 2/9 が B くんが当たったという条件下で A くんが当たる確率、つまり $P(A|B)$。従って $P(B)P(A|B) = P(A \cap B)$ と定義式が導かれる。

☆全確率公式

標本空間 Ω が互いに排反な n この事象の和で書けるとき、

$$P(A) = \sum_{i=1}^{n} P(A|H_i)P(H_i)$$

条件付き確率の定義から $\sum$ 内は $P(A \cap H_i)$ となり、明らか。これも今まで当たり前にやってきたことを式にしただけ。
前項のくじで A くんが当たる確率は、B くんが当たったときと外れたときで場合分けして求める。B くんが当たる、外れる確率を H_1, H_2 とすればそのまま公式になる。

★ベイズの定理

標本空間 Ω が互いに排反な n 個の事象の和で書けるとき、

$$P(H_k|E) = \frac{P(E|H_k)P(H_k)}{\sum_{i=1}^{n} P(E|H_i)P(H_i)}$$

右辺を条件付き確率の定義と全確率公式で書き換えると $P(E \cap H_k)/P(E)$ となり定義そのままになるから覚える必要もないかも。
条件付き確率 $P(A|B)$ の前後 (A と B) を交換できるものと考えると応用に使いやすい。

☆事象の独立性

事象 A と B が独立 $\iff P(A \cap B) = P(A)P(B) \iff P(A|B) = P(A|B^c)$

左半分はおなじみ。さいころが 2 つのとき 2 つの確率をかけてよいことの根拠。右側も独立っていってるんだからその前の条件に左右されるわけがない、というところから明らか。

★★確率変数　X などの大文字

ある変数 X があって、X の取り得る値の全体 Ω が分かっており、Ω の各値に確率が与えられているとき X を確率変数という。
といってもよく分からないから、難しいことは考えずとりあえず変数だと思っておけばよい。宝くじやさいころそのものとも考えられるが、いろいろな演算が行われるので特別なものと考えると後の方でわかりにくくなる。
X が離散値しかとらないものを**離散型確率変数**、連続値をとるものを**連続型確率変数**といい、確率変数 X に対して実際に出た値は X の**実現値**という。関数 $P(X)$ を考えると、X に実現値を代入するとその確率が得られる。

★確率分布

確率変数 X の性質を定めるもの。つまり関数 $P(X)$ の中身。離散型確率変数の場合は次のように X の各値ごと定められる。

$$P(X = x_k) = p_k \quad (k = 1, 2, \cdots, N)$$

連続的確率変数の場合はある点ではなく X の範囲で確率が定められる (ぴったりある値である確率は限りなく小さい) から、次のように積分で表される。

$$P(A \leq X \leq B) = \int_a^b f(x)\,dx$$

ここで $f(x)$ は**確率密度関数**という。

★分布関数　$F(a)$

$$F(a) = P(X \leq a)$$

連続型確率変数で、X が a 以下の確率。正規分布の確率表などはこの形で書かれている。

★★期待値 (平均)　$E(X), \mu$

$$E(X) = \sum_{k=1}^{N} x_k P(X = x_k) = \sum_{k=1}^{n} x_k p_k \quad \text{(離散型確率変数)}$$

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,dx \quad \text{(連続型確率変数)}$$

記述統計の平均と対になるものだが、確率論ではそれぞれの値で重みが異なるのでデータ数で割る代わりに確率をかけている。
後に $E(X^2)$ などが出てくるが、この場合は $x_k^2 P(X = x_k), x^2 f(x)$ というように、$X^2 \times (X$ の確率$)$ と計算される。$E(g(X))$ は X のそれぞれの確率を $g(X)$ にかけ、和をとるという操作だと考えると分かりやすい。従って、$E(X^2) \neq E(X)^2$ であるから注意。
平均と同じように期待値も線形性など 3 つの性質を持つ。

★★分散　$D(X), \sigma^2$

$$V(X) = \sum_{k=1}^{N} (x_k - \mu)^2 P(X = x_k) = \sum_{k=1}^{N} (x_k - \mu)^2 p_k \quad \text{(離散型確率変数)}$$

$$V(X) = \int_{-\infty}^{\infty} (x - \mu) f(x)\, dx \quad \text{(連続型確率変数)}$$

記述統計の分散と基本的に同じ。記述統計の分散が偏差 2 乗の平均となるように、こちらでも期待値との差 $(=$ 偏差$)$ の 2 乗の期待値、つまり $V(X) = E\left((X - \mu)^2\right)$ と表せる。
また、計算を簡略化する公式

$$V(X) = E(X)^2 - \mu$$

も使える。$V(aX + b) = a^2 V(X)$ が成り立つことも記述統計と同じ。

★標準偏差　$V(X), \sigma$

$$D(X) = \sqrt{V(X)}$$

記述統計のときと全く同じ。$D(aX + b) = aD(X)$ が成り立つところも同じ。基準化変量 Z も同じように定義され、チェビシェフの不等式も成り立つ。

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

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

☆ X は (確率分布) に従う　$X \sim$ (確率分布)

X が具体的な確率分布を持つ、つまり P(X) が具体的な関数で与えられるとき、この表現を使って表す。確率分布は $B(n, p)$ のように分布の種類とパラメータで書かれる。
ここから様々な分布が出てくるが、その関数自体はあまり重要ではなく、平均と分散が重要である。

★ベルヌーイ試行

コイン投げを n 回行うと、表が出る確率はどの回でも一定で、しかも互いに独立である。このように、確率 p で成功 (表が出るなど) する試行を独立に n 回行うとき、これ全体を**成功確率** p、**長さ** n のベルヌーイ試行という。

★★2 項分布　$B(n,p)$

$$P(X=x) = {}_nC_x p^x (1-p)^{n-x}$$

n 回コインを投げて、X 回表が出るといった確率の分布。X を成功回数とすれば、成功確率 p、長さ n のベルヌーイ試行は 2 項分布 $B(n,p)$ に従う。
期待値、分散は次のようになる。

$$E(X) = np, \quad V(X) = np(1-p)$$

期待値は「コインは表が $1/2$ で出るから 2 回投げれば表が出るだろう、さいころは 1 が $1/6$ で出るから 6 回投げれば 1 が出るだろう」と考えるのが反映されている。
$p = 0.5$ のとき分布は左右対称になる。

☆ベルヌーイ分布　$Ber(p)$

$n = 1$ の 2 項分布 $B(1,p)$ で、連続型確率分布として扱われる。期待値は p、分散は $p(1-p)$。

★ポアソン分布　$Po(\lambda)$

$$P(X=x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

2 項分布を $\lambda = np, n \to \infty$ として得られる分布。2 項分布で、n が非常に大きく、p が非常に小さい場合、確率を計算するのは大変になる。このときにポアソン分布を近似式として使える。
期待値、分散は $\lambda = np, n \to \infty$ から次のようになる。

$$E(X) = V(X) = \lambda$$

★幾何分布　$Ge(p)$

$$P(X=x)=p(1-p)^{x-1}$$

さいころで1の目が出るまでに振る回数Xなど、ベルヌーイ試行でXを初めて成功する回とするとXは幾何分布に従う。期待値、分散は次のようになる。

$$E(X)=\frac{1}{p},\quad V(X)=\frac{1-p}{p^2}$$

さいころで1が出るまでの振る回数の期待値が$(1/6)^{-1}=6$というのは多すぎるような気がするが、そういうものだと考えよう。

☆無記憶性

$$P(X=a+b|X>b)=P(X=a)$$

を満たすこと。幾何分布には無記憶性がある。bまでに何も起こらなかったからといって、それから先で起こりやすいという訳ではないということ。
じゃんけんで負け続けているから今度は勝てるだろうと考えるのは間違いだということらしい。宝くじやパチンコもそう。

★★正規分布　$N(\mu,\sigma^2)$

$$P(A\leq X\leq B)=\int_a^b\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma}\right)dx$$

最も重要な確率分布。どんな分布もやがてこれに収束する(中心極限定理参照)。期待値と分散をそのままパラメータとしてとる。まさに王者の風格。
形は左右対称で、次の関係が成り立つ。

$$X\sim N(\mu,\sigma^2)\text{ のとき、}aX+b\sim N(a\mu+b,a^2\sigma^2)$$

期待値と分散がそのまま変換された形になる。

★★標準正規分布　$N(0,1)$

平均0、分散1の正規分布。$x=0$に関して対称。確率変数はZで表され、正規分布に従う確率変数Xは次のように標準化できる。

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

正規分布の確率を求める問題では、だいたい標準正規分布の分布関数 $\Phi(z)$ が表で与えられており、標準化して求める。具体的には次のようにする。

$$\begin{aligned} P(a \leq X \leq b) &= P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \\ &= P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$

$\Phi(z)$ の表は正の値しか与えられていない事が多いので、負の値は対称性を利用して $\Phi(-c) = 1 - \Phi(c)$ と求める。

★指数分布　$Ex(\lambda)$

$$P(A \leq X \leq B) = \int_a^b \lambda e^{-\lambda x}\, dx$$

幾何分布の連続版で、確率密度関数は $x > 0$ のときのみを考えて $x \leq 0$ では 0 になる。積分すると簡単に分布関数が求められる。

$$F(x) = 1 - e^{-\lambda x} \quad (x > 0)$$

λ は単位時間当たりに事象が起こる回数、$F(x)$ は時刻 x までに事象が起こる回数と考えられ、幾何分布と同じように無記憶性を持つ。
期待値、分散は次のようになる。

$$E(X) = \frac{1}{\lambda}, \quad V(X) = \frac{1}{\lambda^2}$$

期待値は幾何分布と同じ形だが分散は異なるので注意。

★★同時確率分布

確率変数が X, Y などと 2 つあるとき、(X, Y) を指定 (離散型だったら点、連続型だったら範囲) したときに確率がどうなるのかを表す分布。式では次のように表される。

$$\begin{aligned} &P(X = x, Y = y) = f(x, y) \quad \text{(離散型)} \\ &P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x, y)\, dxdy \quad \text{(連続型)} \end{aligned}$$

$f(x, y)$ は離散型では**同時確率関数**、連続型では**同時確率密度関数**という。
さいころ X, Y を振るとき、$(X, Y) = (1, 2)$ となる確率は同時確率分布。離散型ではさいころの確率表のように表で表すことも多い。

★★周辺確率分布

同時確率分布に対し、どちらか片方の変数だけを指定したときに確率がどうなるのかを表す分布。式では次のように表される。

$$P(X = x_i) = f_X(x_i) = \sum_{j=1}^{N} f(x_i, y_j) \quad (\text{離散型})$$

$$P(a \leq X \leq b) = \int_a^b f_X(x)\,dx = \int_a^b \left(\int_{-\infty}^{\infty} f(x, y)\,dy \right) dx \quad (\text{連続型})$$

$f_X(x)$ は離散型では**周辺確率関数**、連続型では**周辺確率密度関数**という。
さいころ X, Y を振るとき、$X = 1$ となる確率は周辺確率分布。同時確率分布の行、列の和で表される。

★多次元確率分布の期待値

前の期待値の項の補足として、$E(h(X,Y))$ (ただし $h(x,y)$ は x, y の関数) を考える。

$$E(h(X,Y)) = \sum_{i=1}^{M} \sum_{j=1}^{N} h(x_i, y_j) f(x_i, y_j)$$

簡単に言えば、$h(X,Y)$ にそれぞれの値を代入し、その (X,Y) の確率をかけて総和をとればよい。
$h(X,Y)$ が X, Y の片方の関数の場合は、周辺確率分布について今まで通り期待値をとればよい。

$$E(X) = \sum_{i=1}^{N} x_i f_X(x_i)$$

このとき $E(X) = \mu_x$ と書く。

☆条件付期待値

$$E(h(Y)|X = x) = \sum_{i=1}^{N} h(y_i) P(Y = y_i | X = x)$$

そのまま期待値の概念が拡張できる。

☆**共分散　$C(X,Y), \sigma_{XY}$**

$$\sigma_{XY} = E((X-\mu_X)(Y-\mu_Y))$$

記述統計において、平均を期待値としたもの。相関係数 $\rho_{XY} = \rho(X,Y)$ は

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

★★独立性

X と Y が独立 $\iff$

任意の x_i, y_jについて $P(X=x_i, Y=y_j) = P(X=x_i)P(Y=y_j)$

つまり、独立であるとは、任意の同時確率分布が周辺確率分布の積で表せるということ。X と Y が独立なとき、次の式が成り立つ。

$$E(X+Y) = E(X) + E(Y), \quad V(X+Y) = V(X) + V(Y)$$

実は期待値は独立でなくても成り立っている。
分散は一般の場合共分散の項が現れるが、独立の時共分散は 0 になる (よって独立のとき無相関) ことから導かれる。

ここからいよいよ確率変数 X は「測定値を文字でおいたもの」 と扱われます。さいころで考えるのはやめましょう。

★★標本平均　$\bar{X}$

$\bar{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$　ただし $X_i (1 \leq i \leq n)$ は互いに独立で同一の分布 F に従う

具体的に言えば同じ条件の実験 (互いに独立で F に従うことに対応) の測定値の平均にあたる。F が平均 μ、分散 σ^2 だとすると、$\bar{X}$ を確率変数としたときの期待値、分散は次のようになる。

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

期待値は、前項の式から X_i の和は $n\mu$ となり、期待値の線形性から n で割られてこのようになる。分散は同様に前項の式から X_i の和は $n\sigma$ となり、元の値を k 倍すると分散は k^2 倍になることから得られる。
次の独立性と上の関係から、次のことがいえる。これは統計的推測で最も重要な関係。

$X_i (1 \leq i \leq n)$ が互いに独立に同一正規分布 $N(\mu, \sigma^2)$ に従うとき

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

☆再生性

独立に同じ種類の分布に従う 2 つの変数の和が再び元の種類の分布に従うという性質。正規分布、2 項分布、ポアソン分布は再生性を持つ。

$$X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2) \Longrightarrow X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \quad \text{正規分布}$$
$$X \sim B(m, p), Y \sim B(n, p) \Longrightarrow X + Y \sim B(m + n, p) \quad 2 \text{ 項分布}$$
$$X \sim Po(\mu_1), Y \sim Po(\mu_2) \Longrightarrow X + Y \sim Po(\mu_1 + \mu_2) \quad \text{ポアソン分布}$$

★標本平均と正規分布

$$X_i \sim N(\mu, \sigma^2) \quad (1 \leq i \leq n) \Longrightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

再生性により、標本が正規分布に従うとき標本平均も正規分布に従う。従って、標本平均を標準化したものは標準正規分布に従う。

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1)$$

これにより、標準正規分布の確率表を用いて標本平均がどの辺りにあるのかを求められる。

★中心極限定理

$X_i (1 \leq i \leq n)$ が互いに独立に平均 μ、分散 σ^2 の確率分布に従うとき、n が大きくなると標本平均 $\bar{X}$ は正規分布 $N(\mu, \sigma^2/n)$ にいくらでも近づく。
従って n が大きいときはどんな分布でも正規分布に近似できる。

☆大数法則

$X_i (1 \leq i \leq n)$ が互いに独立に平均 μ の確率分布に従うとき、n が大きくなると標本平均 $\bar{X}$ は μ に近づく。
厳密には極限を使って書かれるが、要するにこういうこと。n が大きくなると近似される正規分布の分散は 0 に近づく (つまりばらつきが小さくなる) から。

第 III 部
統計的推測

確率論の後半で定義した考えを利用し、標本 (測定値) から母集団の性質、具体的には平均と分散を求めていく。ここからは具体的な目的があるので、様々な分布や公式が何のために使われるのかを把握していくと分かりやすい (と思う)。

★無作為標本

互いに独立に同一の分布に従っている標本のこと。つまり、無作為標本だと書かれていたら確率論後半の様々な定理が使える。無作為でないことはまずないので、一応の断り書きだと思っておけばよい。
標本の数を**標本の大きさ**という。

☆母集団

標本が小学生 50 人の身長だったら、母集団は全小学生の身長となる。母集団の分布、つまり標本が従う分布を**母集団分布**といい、分布の種類によって母集団は**正規母集団**、**ベルヌーイ母集団**などと呼ばれる。
母集団の平均を**母平均**、分散を**母分散**という。

ここから、断りがない場合は $X_i(1 \leq i \leq n)$ が正規母集団 $N(\mu, \sigma^2)$ からの無作為標本のときの時を考える。

★★不偏標本分散　s^2

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{\text{偏差 2 乗和}}{n-1} = \frac{n}{n-1}S^2$$

統計推測では標本分散の代わりに不偏標本分散が使われる。なぜなら、次の式が成り立つからである。

$$E(s^2) = \sigma^2 \quad (\text{不偏標本分散の期待値は母分散})$$
$$E(S^2) = \frac{n-1}{n}\sigma^2$$

これから母分散を求めるのだからそれに近い方がいいじゃない、ということ。でも何か不思議だねえ。

★カイ２乗分布　$\chi^2(k)$

$$Y = \sum_{i=1}^{k} Z_i^2 \sim \chi^2(k) \quad (Z_i \sim N(0,1)、互いに独立)$$

式はどうでもよい。大切なのは次項の公式。パラメータ k は自由度といい、独立変数の数に当たる。
グラフは左右非対称で、$X = k$、つまり自由度の付近で極大となる。

★★カイ２乗分布と不偏標本分散

$$\frac{(n-1)s^2}{\sigma^2} = \frac{(個数-1) \times 不偏標本分散}{母分散} \sim \chi^2(n-1)$$

この式では不偏標本分散 s^2 が確率変数として扱われている。これを使うと、不偏標本分散から母分散 σ^2 がどの辺りにあるか求めることができる。しかも母平均が式に出てこないので、母平均が分からなくても使える。
自由度が確率変数の数から 1 減っているのは、不偏標本分散の部分で標本平均 $\bar{X}$ を引いているから。

☆ステューデント比 (t 比)

$$\frac{\bar{X} - \mu}{\sqrt{s^2/n}}$$

$\bar{X}$ を標準化したもので、σ^2 を s^2 に置き換えたもの。
標準化して母平均を求めようとするとき、式中に σ^2 が含まれていて母分散を知らなくてはいけなかった。そこで、母分散を知らなくてもよいように置き換えてみようという考え。

★t 分布 $t(k)$

$$t = \frac{X}{\sqrt{Y/k}} \sim t(k) \quad (X \sim N(0,1), Y \sim \chi^2(k))$$

これも式より次項の関係が重要、というより次項の関係を満たすようにこの分布が作られた。パラメータ k は自由度。グラフは左右対称で、標準正規分布に似た形をしている。

★★t分布とステューデント比

$$\text{ステューデント比} = \frac{\bar{X}-\mu}{\sqrt{s^2/n}} \sim t(n-1)$$

標準正規分布に従った$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}$は$\sigma^2$を$s^2$に置き換えると今度は自由度$n-1$の$t$分布に従う。カイ2乗分布と同じように不偏標本分散中の$-\bar{X}$で自由度が1減っている。
これによって、母分散が分かっていなくても母平均μがどの辺りにあるのかを求めることができる。

★上側$100\alpha\%$点　z_α：正規分布，$t_\alpha(k)$：t分布，$\chi^2_\alpha(k)$：χ^2分布

$P(Z \geq z_\alpha) = \alpha$(正規分布のとき)を満たす点のこと。$Z$がこの点より大きいところにある確率は$\alpha$になる。信頼区間を表すのに使う。
$t_\alpha(k)$と$\chi^2_\alpha(k)$のkは自由度で、それぞれは1つの文字である。つまり$t_\alpha(k) \neq t_\alpha \times k$ということ。間違えないように。

★★信頼係数$100(1-\alpha)\%$の信頼区間

$$P(a \leq \mu \leq b) = 1-\alpha\text{を満たす区間}\,[a,b]$$
$$\text{ただし}\ P(\mu \leq a) = P(\mu \geq b) = \frac{\alpha}{2}$$

上の式の場合は母平均μに関する信頼区間で、μをσ^2に置き換えたものは母分散σ^2に関する信頼区間。$100(1-\alpha)\%$信頼区間ともいわれ、実際にはαに具体的な数値を入れて95%信頼区間などと使う。
$100(1-\alpha)\%$の確率で母平均や母分散がこの区間に入ることを表しているが、あくまでも変動するのは確率変数を含む区間の方で、母平均、母分散が動くわけではない。
母分散が分かっているときの母平均は具体的に次のように求める。
前項の式$P(Z \geq z_\alpha) = 1-\alpha$と$Z$の対称性から

$$P(Z \leq -z_\alpha \cup Z \geq z_\alpha) = 2\alpha$$
$$P(-z_\alpha \leq Z \leq z_\alpha) = 1-2\alpha$$

さらに、α を $\alpha/2$ として、$Z = \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}$(標本平均と正規分布の項目より標準正規分布に従う) を代入すると、

$$P\left(\bar{X} - z_{\alpha/2}\sqrt{\sigma^2/n} \leq \mu \leq \bar{X} + z_{\alpha/2}\sqrt{\sigma^2/n}\right) = 1 - \alpha$$

$$\text{信頼区間は} \left[\bar{X} \pm z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right]$$

となる。
同様にして、母分散が分かっていないときの母平均は

$$\left[\bar{X} \pm t_{\alpha/2}(n-1)\sqrt{\frac{s^2}{n}}\right]$$

母分散は

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

となる。母分散だけ $\chi^2_{\alpha/2}(n-1), \chi^2_{1-\alpha/2}(n-1)$ と特異な形になっているが、これは χ^2 分布が左右対称ではないからである。
といっても、実際は上側 $100\alpha\%$ 点は実際の数値で計算していくので、分布の確率表を見ながらやっていけばよい。また、中心極限定理より、正規母集団でなくても近似的に信頼区間を求められる。

☆推定量　T

前項では母平均や母分散を区間で推定したが、推定量は点で推定したもの。といっても、今まで扱ってきた標本平均 $\bar{X}$ や不偏標本分散 s^2 がそのまま推定量になる。
推定量に実際に値を入れたものを**推定値**という。

★不偏推定量

$E(T) = \theta$(θ は母平均や母分散など母集団のパラメータ) を満たす推定量 T のこと。
$\bar{X}$ や s^2 はもちろん、X_i(ある 1 つの測定値) も不偏推定量になる。

☆最小分散不偏推定量

不偏推定量ではだいたい何でもよいことになってしまうが、その中では推定値のばらつきが小さい方がよいといえる。よって、不偏推定量のうち分散の最も小さいものを最小分散不偏推定量として、よりよい推定量と決める。
正規母集団、ベルヌーイ母集団、ポアソン母集団では、標本分散 $\bar{X}$ はそれぞれ μ、p、λ の最小分散不偏推定量となる。

★統計的仮説検定 (検定)

今までは母平均や母分散を直接求めていたが、今度は仮説が存在してそれが正しいかを調べる。
仮説は**帰無仮説 H_0** と**対立仮説 H_1** からなり、「H_0 を棄却する」か「H_0 を採択する」を選択する。
この選択方法を**検定方式**という。一般に対立仮説が研究でそうなってほしい仮説であり、帰無仮説は採択されても特別なことが導かれない仮説 (効果がない、変化がないなど) となる。

★両側検定、片側検定

$$H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0 \quad \text{(両側検定)}$$
$$H_0 : \theta = \theta_0 \quad H_1 : \theta (> or <) \theta_0 \quad \text{(片側検定)}$$

上の形の仮説を持つ検定をそれぞれ両側検定、片側検定という。ただし、θ は母平均や母分散、θ_0 は仮説として設けられた推定量である。
実際の値が θ_0 より大きくなることも小さくなることもある場合は両側検定、大きくなっても小さくはならない、小さくなっても大きくはならない場合は片側検定を行う。片側検定は、H_1 が $\theta > \theta_0$(大きくなることがあっても小さくはならない) の**右側検定**と $\theta < \theta_0$ の**左側検定**に分けられる。
右側、左側は棄却域 (後述) のある側を示す。両側検定、右側検定、左側検定でそれぞれ検定方式が異なる。

ここから、まず初めに母分散が分かっているときの母平均の両側検定を考える。

★★検定統計量　T

$$T = \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}}$$

検定では、推定量 μ_0 が標本平均 $\bar{X}$ からどのくらい離れているかを調べればよい。そこで、推定値と標本平均の差が検定の指標となる。これに標準化を行ったのが検定推定量である。
実は信頼区間を求めるときに使った式の μ を推定量に置き換えたものであり、推定量が母平均に等しいと仮定したとき (つまり H_0 が採択されたとき) の標準化された標本平均の分布にあたる。従って、H_0 が採択されたという仮定の下、次の関係が成り立つ。

$$T \sim N(0, 1)$$

検定方式は次のようになる。

$$\begin{cases} |T| > c & \Longrightarrow \quad H_0\text{を棄却} \\ |T| \leq c & \Longrightarrow \quad H_0\text{を採択} \end{cases}$$

ここで c を**臨界値**という。

★第１種、第２種の誤り

第１種の誤りとは、帰無仮説 H_0 が正しいときに H_0 を棄却してしまうことで、第２種の誤りとは対立仮説 H_1 が正しいときに H_0 を採択してしまうことである。
例えば、H_0 : 新薬に効果がない、H_1 : 新薬に効果があるとすると、第１種の誤りは効果がないのに効果がある判断すること、第２種の誤りは効果があるのに効果がないと判断することである。
この例から分かるように、一般に第１種の誤りの方が重大である。従って、第１種の誤りの確率をコントロールして検定を行う。

★有意水準　α

第１種の誤りを起こす確率。統計推定量の項目より、統計推定量は H_0 が採択されたときの標本平均の分布であるから、有意水準とは統計推定量の分布のうち H_0 が棄却される確率にあたる。つまり、有意水準は次のように表され、対称性から臨界値も決まる。

$$\alpha = P(|T| > c) \iff c = z_{\alpha/2}$$

★★母平均の検定

$$\begin{cases} |T| > z_{\alpha/2} & \Longrightarrow \quad H_0\text{を棄却} \\ |T| \leq z_{\alpha/2} & \Longrightarrow \quad H_0\text{を採択} \end{cases} \quad \text{(両側検定)}$$

$$\begin{cases} T > z_{\alpha} & \Longrightarrow \quad H_0\text{を棄却} \\ T \leq z_{\alpha} & \Longrightarrow \quad H_0\text{を採択} \end{cases} \quad \text{(右側検定)}$$

$$\begin{cases} T < -z_{\alpha} & \Longrightarrow \quad H_0\text{を棄却} \\ T \geq -z_{\alpha} & \Longrightarrow \quad H_0\text{を採択} \end{cases} \quad \text{(左側検定)}$$

片側検定の時に式の形が変わるのは、母平均が推定量より大きく (小さく) なることはなく、有意水準も片方だけ考えればよいからである。

★棄却域、採択域

検定の棄却と採択の条件は、検定統計量Tがどこにあるのかということである。したがって、H_0が棄却されるTの範囲を棄却域、採択される範囲を採択域という。両側検定では棄却域は左右にあるが、片側検定では母平均があり得る方向だけに棄却域がある。
有意水準の概念が分かっていれば、前項の検定の公式は覚えていなくてもグラフを書いて棄却域、採択域を定めて作ることができる。

☆検出力

$$\text{検出力} = 1 - \text{第2種誤りの確率}$$

第2種誤りの確率は、H_1が点で定められているときだけに求められる。H_1が正しいと仮定したときの標本平均の分布のうち、H_0の採択域に入る部分の面積にあたる。
検出力が大きいほど検定の正当さが増すが、有意水準を小さくすると第2種誤りの確率は大きくなるというトレードオフの関係がある。

★ステューデントのt検定(t検定)

母分散が分かっていないときの母平均の検定。検定推定量はt検定推定量といい、次のようになる。

$$t = \frac{\bar{X} - \mu_0}{\sqrt{s^2/n}}$$

これはt分布に従い、検定方式は次のようになる。

$$\begin{cases} |t| > t_{\alpha/2}(n-1) \implies H_0\text{を棄却} \\ |t| \leq t_{\alpha/2}(n-1) \implies H_0\text{を採択} \end{cases} \quad \text{(両側検定)}$$

$$\begin{cases} t > t_{\alpha}(n-1) \implies H_0\text{を棄却} \\ t \leq t_{\alpha}(n-1) \implies H_0\text{を採択} \end{cases} \quad \text{(右側検定)}$$

$$\begin{cases} t < -t_{\alpha}(n-1) \implies H_0\text{を棄却} \\ t \geq -t_{\alpha}(n-1) \implies H_0\text{を採択} \end{cases} \quad \text{(左側検定)}$$

母平均が分かっているときと全く同じ形だ。

☆母分散の検定

母分散についての検定。検定推定量は今までと同様に、

$$Y = \frac{(n-1)s^2}{\sigma^2}$$

カイ 2 乗分布は左右非対称であるから、検定方式は次のようになる。

$$\begin{cases} Y < \chi^2_{1-\alpha/2}(n-1) \text{ または } Y > \chi^2_{\alpha/2}(n-1) \implies H_0\text{を棄却} \\ \chi^2_{1-\alpha/2}(n-1) \le Y \le \chi^2_{\alpha/2}(n-1) \implies H_0\text{を採択} \end{cases} \text{(両側検定)}$$

$$\begin{cases} Y > \chi^2_{\alpha}(n-1) \implies H_0\text{を棄却} \\ Y \le \chi^2_{\alpha}(n-1) \implies H_0\text{を採択} \end{cases} \text{(右側検定)}$$

$$\begin{cases} Y < \chi^2_{1-\alpha}(n-1) \implies H_0\text{を棄却} \\ T \ge \chi^2_{1-\alpha}(n-1) \implies H_0\text{を採択} \end{cases} \text{(左側検定)}$$

★ 2 標本問題

2 種類の異なる環境でデータをとり、そのデータから 2 つの母集団に差があるのかを調べる問題。母平均の差について考える。

ここからは $X_i \sim N(\mu_1, \sigma_1^2), Y_j \sim N(\mu_2, \sigma_2^2) \quad (1 \le i \le m, 1 \le j \le n)$ とする。

★標本平均の差の分布

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{m}\right)$$

正規分布の再生性から簡単に導かれる。分散が差にならないのは、Y も $-Y$ も分散、つまり散らばりは等しいから。
これを標準化すると次のようになる。

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0, 1)$$

2 つの母分散が分かっているときには、これを使えば $\bar{X} - \bar{Y}$ の信頼区間を求めたり検定をしたりできる。検定では一般に帰無仮説は $\mu_1 = \mu_2$ であり、$\mu_1 - \mu_2 = 0$ から検定統計量は次のようになる。

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

★プールされた分散　s^2

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

2 つの不偏標本分散をまとめて 1 つにしたもの。加重平均をとっている。これが全体の不偏標本分散として扱われる。

★2つの母分散が等しいと分かっているとき

2つの母分散が σ^2 で等しいと分かっているときには、母分散をプールされた分散に置き換えたものは自由度 $n+m-2$ の t 分布に従う。

$$Z = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{s^2\left(\frac{1}{m}+\frac{1}{n}\right)}} \sim t(m+n-2)$$

これを使えば母分散が等しいと分かっているだけで $\bar{X}-\bar{Y}$ の信頼区間を求めたり検定をしたりできる。
また、このとき次の式が成り立つ。

$$\frac{(m+n-2)s^2}{\sigma^2} \sim \chi^2(m+n-2)$$

これによって母分散についても信頼区間を求めたり検定をしたりできる。

・F 分布　$F(n_1, n_2)$

$$F = \frac{Y_1/n_1}{Y_2/n_2} \quad \big(Y_1 \sim \chi^2(n_1), Y_2 \sim \chi^2(n_2)\big)$$

自由度 (n_1, n_2) の F 分布といい、次の標本分散、母分散について次の関係がある。

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(m-1, n-1)$$

ここで確率変数は s_1^2, s_2^2 である。
この関係を用いて、$H_0 : \sigma_1^2 = \sigma_2^2$ の検定 (**F 検定**) を行うことができ、H_0 が採択されれば前項の母分散が等しいと分かっているときに帰着できる。
検定統計量は次のようになる。

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

対立仮説が $\sigma_1^2 < \sigma_2^2$ のとき、検定方式は次のようになる。

$$\begin{cases} F < F_{1-\alpha}(m-1, n-1) & \longrightarrow \quad H_0\text{を棄却} \\ F > F_{1-\alpha}(m-1, n-1) & \longrightarrow \quad H_0\text{を採択} \end{cases}$$

★統計解析における回帰分析

記述統計ではデータが母集団全体としていたが、今回は母集団が正規母集団に従うとしてモデルを考える。
回帰モデルは $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ となるが、このうち誤差項 ϵ_i を確率変数とする。β_0, β_1 は真の切片と傾き (未知)、x_i は具体的な値であり、y_i は ϵ_i によって決まる従属的な確率変数と見る。

★標準的仮定

x_iは確率変数でない
$E(\epsilon_i) = 0$　(誤差項の期待値は 0)
$V(\epsilon_i) = E(\epsilon_i^2) = \sigma^2$　(誤差項の分散は一定)
$C(\epsilon_i, \epsilon_j) = E(\epsilon_i \epsilon_j) = 0 \quad (i \neq j)$　(誤差項の共分散は 0)

分散の左二辺は $V(\epsilon_i) = E\left((\epsilon_i - E(\epsilon_i))^2\right)$ に期待値の条件を代入して、共分散の左二辺は $C(\epsilon_i, \epsilon_j) = E\left[(\epsilon_i - E(\epsilon_i))\,(\epsilon_j - E(\epsilon_j))\right]$ に期待値の条件を代入して得られる。
共分散の条件は誤差項が互いに独立であることを示す。
しかし実際には標準的仮定を満たす、$\epsilon_i \sim N(0, \sigma^2)$ の場合しか考えないのであまり問題にならない。

☆従属変数 y_i の平均、分散

$E(y_i) = \beta_0 + \beta_1 x_i$　(期待値は直線上にある)
$V(y_i) = V(\epsilon_i)\sigma^2$
$C(y_i, y_j) = 0 \quad (i \neq j)$　(互いに独立)

いずれも ϵ_i の条件に回帰モデルの式を代入して期待値、分散の性質を利用すれば簡単に出てくる。

★ B, w_i

$$B = \sum_{j=1}^{n} (x_j - \bar{x})^2 = x \text{ の偏差 2 乗和}$$

$$w_i = \frac{x_i - \bar{x}}{B} = \frac{x \text{ の偏差}}{x \text{ の偏差 2 乗和}}$$

それぞれの正式名称は不詳。記号だけ。だがこれから使われる。

★★最小 2 乗推定量　$\hat{\beta}_0, \hat{\beta}_1$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2},\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

記述統計では、「最小 2 乗推定値」だったが、こちらは y_i が確率変数のため確率変数になっているため名前も推定量になる。

$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i \epsilon_i$ という関係も成り立ち、他に回帰直線が平均の点を通ることなどを利用すると次のことが分かる。

$$E(\hat{\beta}_0) = \beta_0,\ V(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{B} \right)$$
$$E(\hat{\beta}_1) = \beta_1,\ V(\hat{\beta}_1) = \frac{\sigma^2}{B}$$

これより、最小 2 乗推定量は不偏推定量であることが分かる。

★残差　$\hat{\epsilon}_i$

$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

残差も確率変数になる。2 つの関係

$$\sum_{i=1}^{n} \hat{\epsilon}_i = 0, \sum_{i=1}^{n} x_i \hat{\epsilon}_i = 0$$

が成り立っているため、自由度は $n-2$ となる。
ここで、

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\epsilon}^2}{n-2}$$

となる値 $\hat{\sigma}^2$ を定めると、不偏標本分散と同じように次の式が成り立つ。

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

このことは $\hat{\sigma}^2$ が不偏標本分散と同じように偏差 2 乗和を自由度で割ったものであることからも分かる。この関係から誤差項の分散 σ を推定することも可能である。

☆回帰係数の t 検定

もし $\beta_1 = 0$ だったら、y_i は x_i に依らないということになり、回帰分析の意味が無くなる。従って、$H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$ として検定を行う。
母分散は分かっていないから次のような検定推定量となる。

$$t = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{B}}}$$

☆重回帰モデル

$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ というように独立変数を増やしたもの。