標本分布について

- 黒木玄

- 2022-04-11〜2022-04-29

このノートではJulia言語 (https://julialang.org/)を使用している:

- Julia言語のインストールの仕方の一例 (https://nbviewer.org/github/genkuroki/msfd28/blob/master/install.ipynb)

自明な誤りを見つけたら, 自分で訂正して読んで欲しい. 大文字と小文字の混同や書き直しが不完全な場合や符号のミスは非常によくある.

このノートに書いてある式を文字通りにそのまま読んで正しいと思ってしまうとひどい目に会う可能性が高い. しかし, 数が使われている文献には大抵の場合に文字通りに読むと間違っている式や主張が書いてあるので, 内容を理解した上で訂正しながら読んで利用しなければいけない. 実践的に数学を使う状況では他人が書いた式をそのまま信じていけない.

このノートの内容よりもさらに詳しいノートを自分で作ると勉強になるだろう. 膨大な時間を取られることになるが, このノートの内容に関係することで飯を食っていく可能性がある人にはそのためにかけた時間は無駄にならないと思われる.

目次

- 1 標本分布とその使い方
  - 1.1 同時確率質量函数と同時確率密度函数
  - 1.2 確率変数の独立性の定義
    - 1.2.1 独立な確率変数達の同時確率質量函数
    - 1.2.2 独立な確率変数達の同時確率密度函数
    - 1.2.3 独立性に関する大雑把なまとめ
    - 1.2.4 分散を0に近付けたときの正規分布について
  - 1.3 独立同分布(i.i.d., iid)の定義
  - 1.4 標本分布の定義
  - 1.5 確率分布の積と n 乗の定義
  - 1.6 分布 D の標本分布の主な使用用途
  - 1.7 試行回数 n のBernoulli試行の分布はBernoulli分布の標本分布
  - 1.8 二項分布による推定の確率的揺らぎの記述
  - 1.9 問題: 大阪都構想に関する住民投票の結果について
    - 1.9.1 Julia言語による計算の例
    - 1.9.2 WolframAlphaによる計算の例:
    - 1.9.3 Clopper-Pearsonの信頼区間とそれを与えるP値
    - 1.9.4 信頼区間よりも情報量が大きなP値函数のプロット
    - 1.9.5 Sternの信頼区間とそれを与えるP値函数
    - 1.9.6 Sternの信頼区間を与えるP値函数の実装例
  - 1.10 対ごとに独立であっても全体が独立であるとは限らない
  - 1.11 確率変数の独立性の現実における解釈に関する重大な注意
- 2 確率変数達の共分散と相関係数と無相関性
  - 2.1 確率変数達の共分散と相関係数の定義
  - 2.2 確率変数達の無相関の定義
  - 2.3 問題: 確率変数の相関係数の計算例
  - 2.4 問題: Cauchy-Schwarzの不等式
  - 2.5 問題: 等確率有限離散分布の相関係数と $\cos\theta$ の関係
  - 2.6 問題: 相関係数の計算
  - 2.7 問題: 共分散が 0 に近くても相関係数が 0 から遠い場合がある
  - 2.8 問題: 独立ならば無相関である (実質1行で解ける)
  - 2.9 問題: 無相関でも独立とは限らない
  - 2.10 問題: 無相関な確率変数達の和の分散はそれぞれの分散の和になる
  - 2.11 問題(メタアナリシスの出発点): 共通の期待値と異なる分散を持つ確率変数の荷重平均
  - 2.12 問題: 二項分布と負の二項分布の平均と分散のBernoulli分布と幾何分布の場合への帰着
  - 2.13 問題: 番号が異なる確率変数達が無相関なときの確率変数の和の共分散
  - 2.14 分散共分散行列とその半正定値性
- 3 標本(サンプル, データ)の平均と分散と共分散と相関係数
  - 3.1 標本平均の定義
  - 3.2 問題: 無相関な確率変数達の標本平均の分散
  - 3.3 標本分散と不偏分散の定義
  - 3.4 不偏推定量について: 不偏分散の定義ではどうして n ではなく $n-1$ で割るか
  - 3.5 データの共分散の定義

3.6 問題: 標本平均達の共分散
3.7 問題: 不偏共分散の定義ではどうして n ではなく $n-1$ で割るか
3.8 データの相関係数の定義 (以上の定義のまとめにもなっている)
4 最小二乗法による線形回帰
4.1 問題: 最小二乗法による線形回帰の公式の導出
4.1.1 解答例1: 単回帰の公式の素朴な導出
4.1.2 解答例2: データの5つの要約値だけで平均二乗残差を表示
4.1.3 コンピュータによる最小二乗法の計算例
4.2 必修問題: 最小二乗法の計算例 (Anscombe's quartet)
4.2.1 WaolframAlphaでAnscombeの例1を扱う方法
4.2.2 Julia言語でAnscombeの例を扱う方法
4.3 Galton (1886)が描いた2つの回帰直線
4.3.1 Galtonのグラフの再現
4.3.2 グラフ中の楕円と2変量正規分布モデルや回帰直線の関係
5 モーメントとキュムラントと歪度と尖度
5.1 モーメントとその母函数と特性函数とキュムラント母函数の定義
5.2 特性函数による期待値の表示
5.3 問題: 分布のアフィン変換のキュムラント
5.4 問題: 標準正規分布のモーメント母函数と特性函数とキュムラント母函数
5.5 確率変数の標準化と標準化キュムラントと歪度と尖度
5.6 問題: 正規分布の歪度と尖度
5.7 問題: 一様分布のキュムラント母函数と歪度と尖度
5.8 問題: 独立な確率変数達の和のモーメント母函数と特性函数とキュムラント母函数
5.9 問題: ガンマ分布のキュムラント母函数と歪度と尖度
5.10 問題: 二項分布のキュムラント母函数と歪度と尖度
6 独立同分布な確率変数達の不偏分散の分散
6.1 期待値 0, 分散 1 の場合への帰着
6.2 標本平均と不偏分散の共分散の計算
6.3 不偏分散の分散の計算
6.4 歪度と尖度に関する不等式
6.5 問題: 不偏分散と不偏補正されていない標本分散の平均二乗誤差の比較
7 標本分布における標本平均と不偏分散の同時分布の視覚化
7.1 正規分布の標本分布における $\bar{X}, S^2$ の同時分布
7.2 一様分布の標本分布における $\bar{X}, S^2$ の同時分布
7.3 単峰型のガンマ分布の標本分布における $\bar{X}, S^2$ の同時分布
7.4 指数分布の標本分布における $\bar{X}, S^2$ の同時分布
7.5 非対称なベータ分布の標本分布における $\bar{X}, S^2$ の同時分布
7.6 非対称な2つ山の混合正規分布の標本分布における $\bar{X}, S^2$ の同時分布
7.7 非対称なBernoulli分布の標本分布における $\bar{X}, S^2$ の同時分布
7.8 対数正規分布の標本分布における $\bar{X}, S^2$ の同時分布
8 正規分布の標本分布の場合
8.1 準備: $\sqrt{n}\,\bar{x}$ を含む正規直交座標系
8.2 準備 $n-2$ 次元単位球面上の微小面積要素
8.3 正規分布の標本分布における $\bar{X}, S^2$ の同時分布の計算
8.4 正規分布の標本分布から自然に t 分布に従う確率変数が得られること
8.5 t 分布の使いどころ
8.6 問題: T の計算例

In [1]:

```
ENV["LINES"], ENV["COLUMNS"] = 100, 100
using BenchmarkTools
using Distributions
using LinearAlgebra
using Optim
using Printf
using QuadGK
using Random
Random.seed!(4649373)
using Roots
using SpecialFunctions
using StaticArrays
using StatsBase
using StatsFuns
using StatsPlots
default(fmt = :png, titlefontsize = 10, size = (400, 250))
using SymPy
```

1 標本分布とその使い方

1.1 同時確率質量函数と同時確率密度函数

確率変数達 $X_1, \dots, X_n$ が同時確率質量函数 $P(x_1, \dots, x_n)$ を持つとは, $P(x_1, \dots, x_n) \geq 0$ でかつ, $P(x_1, \dots, x_n)$ の総和が 1 であり,

$$E[f(X_1, \dots, X_n)] = \sum_{x_1} \cdots \sum_{x_n} f(x_1, \dots, x_n) P(x_1, \dots, x_n)$$

が成立することである. このとき, X_i 単独の確率質量函数 $P(x_i)$ は

$$P(x_i) = \sum_{x_1} \cdots \widehat{\sum_{x_i}} \cdots \sum_{x_n} P(x_1, \cdots, x_i, \dots, x_n)$$

になる. ここで, $\widehat{\sum_{x_i}}$ は $\sum_{x_i}$ を除くという意味である. なぜならば,

$$\begin{aligned} E[f(X_i)] &= \sum_{x_1} \cdots \sum_{x_i} \cdots \sum_{x_n} f(x_i) P(x_1, \cdots, x_i, \dots, x_n) \\ &= \sum_{x_i} f(x_1) \left(\sum_{x_1} \cdots \widehat{\sum_{x_i}} \cdots \sum_{x_n} P(x_1, \cdots, x_i, \dots, x_n) \right). \end{aligned}$$

注意: $i \neq j$ のとき $P(x_i)$ と $P(x_j)$ は一般に異なる確率質量函数になることに注意せよ. 区別を明瞭にするためには $P(x_i)$ を $P_i(x_i)$ のように書くべきであるが, 統計学の世界での慣習に従ってこのような記号法上の省略を行った. これが気に入らない人は自分でノートを作るときに, 区別されるべきものに異なる記号を割り振るようにすればよいだろう. 以下についても同様である.

確率変数達 $X_1, \dots, X_n$ が同時確率密度函数 $p(x_1, \dots, x_n)$ を持つとは, $p(x_1, \dots, x_n) \geq 0$ でかつ, $p(x_1, \dots, x_n)$ の積分が 1 になり,

$$E[f(X_1, \dots, X_n)] = \int \cdots \int f(x_1, \dots, x_n) p(x_1, \dots, x_n)\, dx_1 \cdots dx_n$$

が成立することである. ここで $\int$ は定積分を表す. このとき, X_i 単独の確率密度函数 $p(x_i)$ は

$$p(x_i) = \int \cdots \widehat{\int} \cdots \int p(x_1, \cdots, x_i, \dots, x_n)\, dx_1 \cdots \widehat{dx_i} \cdots dx_n$$

になる. ここで, $\widehat{\int}$, $\widehat{dx_i}$ はそれらを除くことを意味する. なぜならば,

$$\begin{aligned} E[f(X_i)] &= \int \cdots \int \cdots \int f(x_i) p(x_1, \cdots, x_i, \dots, x_n)\, dx_1 \cdots dx_i \cdots dx_n \\ &= \int f(x_i) \left(\int \cdots \widehat{\int} \cdots \int p(x_1, \cdots, x_i, \dots, x_n)\, dx_1 \cdots \widehat{dx_i} \cdots dx_n \right) dx_i. \end{aligned}$$

1.2 確率変数の独立性の定義

確率変数達 $X_1, \dots, X_n$ が与えられており, それらの函数の期待値 $E[f(X_1, \dots, X_n)]$ が定義されているとする. このとき, $X_1, \dots, X_n$ が **独立** (independent)であるとは,

$$E[f_1(X_1) \cdots f_n(X_n)] = E[f_1(X_1)] \cdots E[f_n(X_n)] \tag{1}$$

のように X_i ごと函数達 $f_i(X_i)$ の積の期待値が各々の $f_i(X_i)$ の期待値の積になることだと定める.

注意: 厳密には函数 f_i 達を動かす範囲を確定させる必要があるが, このノートではそのようなことにこだわらずに解説することにする. 厳密な理論の展開のためには測度論的確率論の知識が必要になるが, そのような方向はこのノートの目標とは異なる. 測度論的確率論に興味がある人は別の文献を参照して欲しい. ただし, 測度論的確率論の理解と統計学の理解は別の話題になってしまうことには注意して欲しい. 測度論的確率論と統計学では興味の方向が異なる.

1.2.1 独立な確率変数達の同時確率質量函数

確率変数達 $X_1, \dots, X_n$ が同時確率質量函数 $P(x_1, \dots, x_n)$ を持つとき, X_i 単独の確率質量函数を $P_i(x_i)$ と書くならば, 確率変数達 $X_1, \dots, X_n$ が独立であることと,

$$P(x_1, \dots, x_n) = P_1(x_1) \cdots P_n(x_n) \tag{2}$$

が成立することは同値である. すなわち, $(X_1, \dots, X_n)$ の函数の期待値が次のように表されることと同値である:

$$E[f(X_1,\dots,X_n)] = \sum_{x_1}\cdots\sum_{x_n} f(x_1,\dots,x_n)P_1(x_1)\cdots P_n(x_n)$$

注意: ここでは気分を変えて, $P_i(x_i)$ をずぼらに $P(x_i)$ と書く流儀をやめてみた.

証明: (2)が成立しているならば,

$$\begin{aligned}
E[f_1(X_1)\cdots f_n(X_n)] &= \sum_{x_1}\cdots\sum_{x_n} f_1(x_1)\cdots f_n(x_n)P_1(x_1)\cdots P_n(x_n)\\
&= \sum_{x_1} f_1(x_1)P_1(x_1)\cdots\sum_{x_n} f_n(x_n)P_n(x_1)\\
&= E[f_1(X_1)]\cdots E[f_n(X_n)]
\end{aligned}$$

と(1)が成立する.

逆に(1)が成立しているならば, $f_i(x_i)$ として $x_i = a_i$ のときにのみ 1 でそれ以外のときに 0 になる函数を取ると,

$$P(a_1,\dots,a_n) = E[f_1(X_1)\cdots f_n(X_n)] = E[f_1(X_1)]\cdots E[f_n(X_n)] = P_1(a_1)\cdots P_n(a_n)$$

なので, (2)が成立する.

1.2.2 独立な確率変数達の同時確率密度函数

確率変数達 $X_1,\dots,X_n$ が同時確率密度函数 $p(x_1,\dots,x_n)$ を持つとき, X_i 単独の確率密度函数を $p_i(x_i)$ と書くならば, 確率変数達 $X_1,\dots,X_n$ が独立であることと,

$$p(x_1,\dots,x_n) = p_1(x_1)\cdots p_n(x_n) \tag{3}$$

が成立することは同値である. すなわち, $(X_1,\dots,X_n)$ の函数の期待値が次のように表されることと同値である:

$$E[f(X_1,\dots,X_n)] = \int\cdots\int f(x_1,\dots,x_n)p_1(x_1)\cdots p_n(x_n)\,dx_1\cdots dx_n.$$

ここで $\int\cdots dx_i$ は適切な範囲での定積分を表す.

注意: ここでは気分を変えて, $p_i(x_i)$ をずぼらに $p(x_i)$ と書く流儀をやめてみた.

(雑な)証明: (3)が成立しているならば,

$$\begin{aligned}
E[f_1(X_1)\cdots f_n(X_n)] &= \int\cdots\int f_1(x_1)\cdots f_n(x_n)p_1(x_1)\cdots p_n(x_n)\,dx_1\cdots dx_n\\
&= \int f_1(x_1)p_1(x_1)\,dx_1\cdots\int f_n(x_n)p_n(x_n)\,dx_n\\
&= E[f_1(X_1)]\cdots E[f_n(X_n)]
\end{aligned}$$

と(1)が成立する.

簡単のため, 密度函数達 $p(x_1,\dots,p_n)$, $p_i(x_i)$ は連続函数になっていると仮定する(応用上は概ねこれで十分). このとき,

$$f_i(x_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{(x_i-a_i)^2}{2\sigma_i^2}\right)$$

とおくと, $\sigma_i \searrow 0$ のとき

$$\begin{aligned}
E[f_i(X_i)] &= \int f_i(x_i)p_i(x_i)\,dx_i \to p_i(a_i),\\
E[f_1(X_1)\cdots f_n(X_n)] &= \int\cdots\int f_1(x_1)\cdots f_n(x_n)p(x_1,\dots,x_n)\,dx_1\cdots dx_n \to p(a_1,\dots,a_n)
\end{aligned}$$

となるので, 逆に(1)が成立しているならば

$$p(a_1,\dots,a_n) = p_1(a_1)\cdots p_n(a_n)$$

と(3)が成立する.

確率密度函数と確率質量函数が混ざっている場合も同様である.

1.2.3 独立性に関する大雑把なまとめ

大雑把に誤解を恐れずに言うと, 確率が積になるときに独立だという.

実際の計算では, 積の期待値が期待値の積になるという形式で確率変数の独立性を使う.

In [2]:
```
# 以下で使う図の準備

var"正規分布は分散が小さくなると一点に集中して行く." = plot()
for σ in (1.0, 0.5, 0.25, 0.1)
    plot!(Normal(0, σ); label="σ = $σ")
end
title!("Normal(μ=0, σ)");
```

1.2.4 分散を0に近付けたときの正規分布について

X_σ を平均 μ, 分散 σ^2 の正規分布に従う確率変数とすると, 有界な連続函数 $f(x)$ について

$$\lim_{\sigma \searrow 0} E[f(X_\sigma)] = f(\mu)$$

となる. 証明が知りたい人は次のリンク先の「総和核」に関する解説(特にGauss核の説明)を参照せよ:

- [12 Fourier解析 - 総和核](https://nbviewer.org/github/genkuroki/Calculus/blob/master/12%20Fourier%20analysis.ipynb#%E7%B7%8F%E5%92%8C%E6%

大雑把に言えば, これは $\sigma \searrow 0$ の極限で正規分布が一点 $x = \mu$ に集中することを意味している.

In [3]:
```
var"正規分布は分散が小さくなると一点に集中して行く."
```

Out[3]:

$E[f(X_0)] = f(\mu)$ を満たす確率変数 X_0 は実質定数 μ に等しく, 実質定数に等しい確率変数が従う分布を **Dirac分布** と呼ぶことがある.

1.3 独立同分布(i.i.d., iid)の定義

確率変数達 $X_1, \dots, X_n$ が **独立同分布** (independent and identically distributed, **i.i.d.**, **iid**) であるとは, それらが独立でかつ X_i 達が従う分布が互いにすべて等しいことであると定める.

1.4 標本分布の定義

n 個の確率変数達 $X_1, \dots, X_n$ は独立同分布であり, 各 X_i は共通の確率分布 D に従うと仮定する:

$$X_1, \dots, X_n \sim D \quad \text{(independent)}$$

このとき, $X_1, \dots, X_n$ を分布 D の **標本** または **サンプル** (sample)と呼び, $X_1, \dots, X_n$ の同時確率分布を **分布 D のサイズ n の標本分布** (distribution of samples of size n from distribution D)と呼ぶことにする.

注意: 独立同分布な確率変数達 $X_1, \dots, X_n$ (確率変数は数ではなく函数であった)そのものではなく, それらの実現値 $x_1, \dots, x_n$ (函数としての確率変数の値達)を標本もしくはサンプルと呼ぶことがある. 1つひとつの値 x_i をサンプルと呼ぶのではなく, 数の列 $x_1, \dots, x_n$ をサンプルと呼ぶことに注意せよ. その辺が紛らわしい場合には数の列 $x_1, \dots, x_n$ を **データ** (data)と呼ぶことがある. $X_1, \dots, X_n$ もデータと呼ぶことがある. この辺の用語の使い方はかなりイーカゲンになり易いので注意して欲しい.

確率変数 X_i 達が同一の確率密度函数 $p(x_i)$ を持つとき, $X_1, \dots, X_n$ の同時確率密度函数が

$$p(x_1, \dots, x_n) = p(x_1) \cdots p(x_n)$$

であることと, $X_1, \dots, X_n$ の同時確率分布が確率密度函数 $p(x)$ が定める連続分布の標本分布であることは同値である.

確率変数 X_i 達が同一の確率質量函数 $P(x_i)$ を持つとき, $X_1, \dots, X_n$ の同時確率質量函数が

$$P(x_1, \dots, x_n) = P(x_1) \cdots P(x_n)$$

であることと, $X_1, \dots, X_n$ の同時確率分布が確率質量函数 $P(x)$ が定める離散分布の標本分布であることは同値である.

1.5 確率分布の積と n 乗の定義

$X_1, \dots, X_n$ が独立な確率変数達であり, X_i が従う分布が D_i であるとき, $(X_1, \dots, X_n)$ が従う分布を

$$D_1 \times \cdots \times D_n$$

と書き, 分布 $D_1, \dots, D_n$ の **積** (product, 直積, direct product)と呼ぶ:

$$(X_1, \dots, X_n) \sim D_1 \times \cdots \times D_n \iff X_i \sim D_i \text{ for } i = 1, \dots, n \text{ and } X_1, \dots, X_n \text{ are independent.}$$

例えば, 分布 D_i が確率密度函数 $p_i(x_i)$ を持つとき, 積分布 $D_1 \times \cdots \times D_n$ は確率密度函数

$$p(x_1, \dots, x_n) = p_1(x_1) \cdots p(x_n)$$

によって定義される多変量連続分布になる. 確率密度函数が積の形になるので, 分布の積と呼ぶ.

D_i がすべて同一の分布 D であるとき, 積 $D_1 \times \cdots \times D_n$ を D^n と書き, 分布 D の **べき乗** (冪乗, power)もしくは **累乗** と呼ぶ. 前節で説明したように分布 D の n 乗 D^n は分布 D のサイズ n の標本分布と呼ばれる:

$$(X_1, \dots, X_n) \sim D^n \iff (X_1, \dots, X_n) \text{ is a sample of size } n \text{ from the distribution D.}$$

例: 試行回数 n, 成功確率 p のBernoulli試行の分布は成功確率 p のBernoulli分布の n 乗になる. 以下の節も参照せよ.

1.6 分布 D の標本分布の主な使用用途

分布 D のサイズ n の標本分布は, 現実に得られる長さ n の数値列 $x_1, \dots, x_n$ の形式のデータの生成法則のモデル化としてよく使われている. そのとき, 分布 D やその標本分布 D^n は **統計モデル** と呼ばれる.

ただし, 我々は, 分布 D のサイズ n の標本分布 D^n を各々が分布 D に従う独立同分布な n 個の確率変数達 $X_1, \dots, X_n$ の同時確率分布のことだと定義したので, データ $x_1, \dots, x_n$ の生成法則が独立同分布という強い条件を近似的に満たしていなければ, そのようなモデル化は妥当ではなくなる.

例: データ $x_1, \dots, x_n$ がコンピュータで rand() 函数を n 回繰り返して得られた乱数列とするとき, そのデータの生成法則は一様分布 $\mathrm{Uniform}(0, 1)$ のサイズ n の標本分布に近似的に従っているとみなされる.

例(基本!): S市の中学3年生男子の総人数は N 人であるとし, その身長の値全体を $a_1, a_2, \dots, a_N$ と書く. 等確率 $1/N$ で a_i 達のどれかの値が選ばれるという条件で定義される有限離散分布を Q と書き, S市の中3男子の身長の分布と呼ぶことにする. (全数調査前なので分布 Q は未知であると仮定する.) $1, 2, \dots, N$ から n 個の数 $i_1, \dots, i_n$ を無作為に選んで, n 人分の中3男子の身長を測定してデータ $(x_1, \dots, x_N) = (a_{i_1}, \dots, a_{i_N})$ が得られたとする. このとき, データ $x_1, \dots, x_N$ の生成法則はS市の中3男子の身長の分布 Q のサイズ n の標本分布になっていると考えられる. このとき, 未知の分布 Q についての推定・推測・推論をデータ $x_1, \dots, x_N$ を用いて行いたい. そのときに, Q のモデル化として平均 μ と分散 σ^2 をパラメータに持つ正規分布を設定し, データ $x_1, \dots, x_N$ の生成法則のモデル化として, 正規分布のサイズ n の標本分布 $\mathrm{Normal}(\mu, \sigma)^n$ を採用することが考えらえる. この場合には正規分布モデルを統計モデルとして採用したことになる. パラメータ μ, σ はデータ $x_1, \dots, x_N$ を使って推定されることになる. 未知であるS市の中3男子の身長の分布 Q (よく **真の分布** と呼ばれる)が正規分布からかけ離れた形をしていると, 正規分布モデルの採用は妥当ではなくなるが, 真の分布が十分に正規分布に近いことが十分に確からしい場合には正規分布モデルの採用は適切になる.

例: 当たりが常に同じ確率で出続けると考えられるルーレットを n 回まわして, 当たりなら 1 を記録し, 外れなら 0 を記録することによって得られる $1, 0$ で構成された長さ n の数列 $x_1, \dots, x_n$ の生成法則は, 未知の成功確率 p を持つ試行回数 n のBernoulli試行の確率分布 $\mathrm{Bernoulli}(p)^n$ (これはBernoulli分布の標本分布に等しい)に近似的に従うと考えられる.

1.7 試行回数 n のBernoulli試行の分布はBernoulli分布の標本分布

試行回数 n, 成功確率 p のBernoulli試行の確率質量函数は

$$P(x_1, \dots, x_n) = p^{x_1 + \cdots + x_n}(1-p)^{n-(x_1+\cdots+x_n)} = \prod_{i=1}^{n} (p^{x_i}(1-p)^{1-x_i}) \quad (x_i = 1, 0)$$

であった. これは成功確率 p のBernoulli分布の確率質量函数

$$P(x_i) = p^{x_i}(1-p)^{1-x_i} \quad (x_i = 1, 0)$$

の積になっているので, 試行回数 n, 成功確率 p のBernoulli試行の確率分布は, 成功確率 p のBernoulli分布のサイズ n の標本分布になっている.

未知の確率 p で当たりが出るルーレットを n 回まわしたときの, 長さ n の当たりと外れからなる列をそのルーレットの出目のサイズ n の **標本** (**サンプル**, sample)と呼ぶ.

その意味でのルーレットの出目のサンプルの確率的揺らぎは, Bernoulli分布の標本分布(すなわちBernoulli試行の分布)でモデル化される.

例えば, 未知の確率 p で当たりが出るルーレットを $n = 1000$ 回まわしてサンプル(データ)を取得したら, 1000 回中当たりが 308 回で外れが 692 回ならば, そのルーレットで当たりが出る確率は3割程度だろうと推定できる. 実際にはルーレットを 1000 回まわし直すたびに当たりの回数は確率的に揺らぐので, 推定結果も確率的に揺らぐことになる. そのような揺らぎを数学的にモデル化するために標本分布は使用される.

言葉使いに関する重要な注意: 「標本」「サンプル」は一般に複数の数値の集まりになる. 上のルーレットの場合には当たりには 1 を対応させ, 外れを 0 に対応させると, サンプルは 1 と 0 からなる長さ n の列になる. 一つひとつの数値をサンプルと呼ぶのではなく, 複数の数値の集まりをサンプルと呼ぶ. この専門用語的言葉遣いは日常用語的なサンプルという言葉の使い方からはずれているので注意が必要である. この辺の言葉遣いで誤解を防ぎたい場合には「データ」と呼ぶこともある.

1.8 二項分布による推定の確率的揺らぎの記述

前節の設定を引き継ぐ.

Bernoulli分布のサイズ n の標本分布における 1 の個数の分布は二項分布になるのであった. 仮に

> n 回中 k 回の当たりが出たときに, 未知である当たりが出る確率は k/n に近いだろうと推定

することにしたときの, k/n の確率的な揺らぎは二項分布によって計算できる.

K は二項分布 $\mathrm{Binomial}(n, p)$ に従う確率変数であるとし, 確率変数 $\hat{p}$ を $\hat{p} = K/n$ と定める. この確率変数 $\hat{p} = K/n$ は上のルールで定めた未知の確率 p の推定の仕方の数学的記述になっている. このとき, $\hat{p}$ はパラメータ p の **推定量** (estimator)であるという. 確率変数 $\hat{p}$ は「n 回中 k 回当たりが出た」というデータに対応する確率変数 K の函数になっている.

$\hat{p} = K/n$ の期待値と分散は, 二項分布の期待値と分散がそれぞれ np と $np(1-p)$ であることより, 以下のように計算される:

$$
\begin{aligned}
E[\hat{p}] &= E[K/n] = \frac{E[K]}{n} = \frac{np}{n} = p, \\
\mathrm{var}(\hat{p}) &= \mathrm{var}(K/n) = \frac{\mathrm{var}(K)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.
\end{aligned}
$$

例えば, $n = 1000$, $p = 0.3$ のとき, $\hat{p}$ の標準偏差は

$$
\mathrm{std}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.3 \cdot 0.7}{1000}} \approx 1.45\%.
$$

モデル内の設定では大雑把に推定結果はこの2倍の $\pm 3\%$ 程度揺らぐと考えらえる.

未知の p を使わずに $\mathrm{SE} = \mathrm{std}(\hat{p})$ の値を推定するためには, その式の中の p に p の推定量 $\hat{p}$ を代入して得られる公式

$$
\widehat{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
$$

を使えばよいだろう. ($\mathrm{SE} = \mathrm{std}(\hat{p})$ は **標準誤差** (standard error)と呼ばれる. $\widehat{SE}$ は標準誤差の推定量である. $\widehat{SE}$ のことも標準誤差と呼ぶことがある. この辺の言葉遣いもイーカゲンな場合が多いので注意が必要である. 何をどう呼ぶかよりも, それが正確には何を意味しているかが重要である.

このような統計分析の結果が現実において信頼できるかどうかは, Bernoulli分布の標本分布によるモデル化の現実における妥当性に依存する. モデルが現実において妥当である証拠が全然無ければ, このような推測結果も信頼できないことになる. モデルの現実における妥当性の証拠の提示は統計モデルのユーザー側が独自に行う必要がある.

In [4]:

```
n, p = 1000, 0.3
dist = Binomial(n, p)
L = 10^6
K = rand(dist, L)
p̂ = K/n
histogram(p̂; norm=true, alpha=0.3, label="p̂ = K/n", bin=100)
vline!([p]; label="p=$p", xlabel="p̂", xtick=0.30-0.12:0.03:0.30+0.12)
title!("n = $n, p = $p")
```

Out[4]:

確かにランダムに決まる $\hat{p} = K/n$ の値は, 推定先の値である $p = 0.30 = 30\%$ を中心に大雑把に $\pm 3\%$ 程度の範囲に分布している.

実際には小さな確率でもっと大きく外れることもあることに注意せよ.

現実の統計分析では, データを上のグラフの例のように100万回撮り直したりできないので, このようなグラフを描くことはできない. たった1つの「n 回中 k 回成功」(もしくは「n 人中 k 人」)の型のデータだけを使って判断を下さなければいけなくなる. 真っ暗闇の中を手探りで進むような感じになる.

1.9 問題: 大阪都構想に関する住民投票の結果について

2015年と2020年の大阪都構想に関する住民投票の結果は

- 2015年: 賛成: 694,844 (49.6%)　反対: 705,585 (50.4%)
- 2020年: 賛成: 675,829 (49.4%)　反対: 692,996 (50.6%)

であった([検索 (https://www.google.com/search?q=%E5%A4%A7%E9%98%AA%E9%83%BD%E6%A7%8B%E6%83%B3++%E4%BD%8F%E6%B0%91%E6%8A%95%E7%A5%A](https://www.google.com/search?q=%E5%A4%A7%E9%98%AA%E9%83%BD%E6%A7%8B%E6%83%B3++%E4%BD%8F%E6%B0%91%E6%8A%95%E7%A5%A)

どちらでも僅差で反対派が勝利した. パーセントの数値を見ると大変な僅差であったようにも見える. この数値に二項分布モデルを適用したらどうなるかを(それが妥当な適用であるかどうかを度外視して)計算するのがこの問題の内容である.

確率変数 K は二項分布 $\mathrm{Binomial}(n, p)$ に従うと仮定する.

(1) $n = 694844 + 705585 = 1400429$, $p = 0.5$ のとき, 確率 $P(K \leq 694844)$ の2倍の値を求めよ.

(2) $n = 675829 + 692996 = 1368825$, $p = 0.5$ のとき, 確率 $P(K \leq 675829)$ の2倍の値を求めよ.

(3) $n = 694844 + 705585 = 1400429$ のとき, 以下を求めよ:

- $P(K \geq 694844) = 2.5\%$ になるようなパラメータ p の値 p_L,
- $P(K \leq 694844) = 2.5\%$ になるようなパラメータ p の値 p_U.

(4) $n = 675829 + 692996 = 1368825$ のとき, 以下を求めよ:

- $P(K \geq 675829) = 2.5\%$ になるようなパラメータ p の値 p_L,
- $P(K \leq 675829) = 2.5\%$ になるようなパラメータ p の値 p_U.

確率やパラメータの数値は有効桁4桁まで求めよ. 0.00000000000000000001234 のように 0 を沢山含む表示は見難いので,

$$1.234 \times 10^{20} = 0.\underbrace{00000000000000000001}_{20}234$$

のように書かずに,

$$1.234\mathrm{E}{-20} = 1.234\mathrm{e}{-20}$$

のように書くこと.

注意: この問題の内容を一般化するだけで **検定** (統計的仮説検定, statistical hypothesis testing)や **信頼区間** (confidence interval)の一般論が得られる. (1), (2)はP値と求める問題になっており, (3), (4)は 95% 信頼区間 $[p_L, p_U]$ を求める問題になっている(ただし両方Clopper-Pearsonの信頼区間の場合).

解答例:

(1) $P(K \leq 694844) \approx 1.130\mathrm{e}{-19}$

(2) $P(K \leq 675829) \approx 9.687\mathrm{e}{-49}$

(3) $[p_L, p_U] \approx [0.4953, 0.4970]$

(4) $[p_L, p_U] \approx [0.4929, 0.4946]$

解答終

1.9.1 Julia言語による計算の例

In [5]:
```
# 確率計算を素朴に行うには対数を取った結果を主な対象にしないと失敗する.
# 次は二項分布における確率質量函数の対数である.
logP(n, p, k) = logabsbinomial(n, k)[1] + k*log(p) + (n-k)*log(1-p)
```

Out[5]:
```
logP (generic function with 1 method)
```

In [6]:
```
# (1)
@show 2exp(logsumexp(logP(694844 + 705585, 0.5, k) for k in 0:694844))
@show 2cdf(Binomial(694844 + 705585, 0.5), 694844);
```

```
2 * exp(logsumexp((logP(694844 + 705585, 0.5, k) for k = 0:694844))) = 1.130422573100686e-19
2 * cdf(Binomial(694844 + 705585, 0.5), 694844) = 1.1304225734350622e-19
```

In [7]:
```
# (2)
@show 2exp(logsumexp(logP(675829 + 692996, 0.5, k) for k in 0:675829))
@show 2cdf(Binomial(675829 + 692996, 0.5), 675829);
```

```
2 * exp(logsumexp((logP(675829 + 692996, 0.5, k) for k = 0:675829))) = 9.687442124914083e-49
2 * cdf(Binomial(675829 + 692996, 0.5), 675829) = 9.687442131513807e-49
```

In [8]:
```
# (3)
f(t) = ccdf(Binomial(694844 + 705585, t), 694844-1) - 0.025
g(t) =  cdf(Binomial(694844 + 705585, t), 694844)   - 0.025
@show p_L = find_zero(f, (0, 1))
@show p_U = find_zero(g, (0, 1));
```

```
p_L = find_zero(f, (0, 1)) = 0.4953366705819138
p_U = find_zero(g, (0, 1)) = 0.4969935526649652
```

In [9]:
```
# (3)
n = 694844 + 705585
k = 694844
α = 0.05
@show p_L = quantile(Beta(k, n-k+1), α/2)
@show p_U = quantile(Beta(k+1, n-k), 1 - α/2);
```

```
p_L = quantile(Beta(k, (n - k) + 1), α / 2) = 0.4953366705819138
p_U = quantile(Beta(k + 1, n - k), 1 - α / 2) = 0.49699355266496525
```

In [10]:
```
# (3)
n = 694844 + 705585
k = 694844
α = 0.05
@show p_L = beta_inc_inv(k, n-k+1, α/2)[1]
@show p_U = beta_inc_inv(k+1, n-k, 1 - α/2)[1];
```

```
p_L = (beta_inc_inv(k, (n - k) + 1, α / 2))[1] = 0.4953366705819138
p_U = (beta_inc_inv(k + 1, n - k, 1 - α / 2))[1] = 0.49699355266496514
```

In [11]:
```
# (4)
f(t) = ccdf(Binomial(675829 + 692996, t), 675829-1) - 0.025
g(t) =  cdf(Binomial(675829 + 692996, t), 675829)   - 0.025
@show p_L = find_zero(f, (0, 1))
@show p_U = find_zero(g, (0, 1));
```

```
p_L = find_zero(f, (0, 1)) = 0.49289139358630474
p_U = find_zero(g, (0, 1)) = 0.4945672196912542
```

In [12]:
```
# (4)
n = 675829 + 692996
k = 675829
α = 0.05
@show p_L = quantile(Beta(k, n-k+1), α/2)
@show p_U = quantile(Beta(k+1, n-k), 1 - α/2);
```

```
p_L = quantile(Beta(k, (n - k) + 1), α / 2) = 0.49289139358630474
p_U = quantile(Beta(k + 1, n - k), 1 - α / 2) = 0.4945672196912543
```

In [13]:
```
# (4)
n = 675829 + 692996
k = 675829
α = 0.05
@show p_L = beta_inc_inv(k, n-k+1, α/2)[1]
@show p_U = beta_inc_inv(k+1, n-k, 1 - α/2)[1];
```

```
p_L = (beta_inc_inv(k, (n - k) + 1, α / 2))[1] = 0.49289139358630474
p_U = (beta_inc_inv(k + 1, n - k, 1 - α / 2))[1] = 0.4945672196912543
```

1.9.2 WolframAlphaによる計算の例:

(1) 2 cdf(BinomialDistribution(694844 + 705585, 0.5), 694844) (https://www.wolframalpha.com/input?i=2+cdf%28BinomialDistribution%28694844+%2B+705585%2C+0.5%29%2C+694844%29)

(2) 2 cdf(BinomialDistribution(675829 + 692996, 0.5), 675829) (https://www.wolframalpha.com/input?i=2+cdf%28BinomialDistribution%28675829+%2B+692996%2C+0.5%29%2C+675829%29)

(3) p_L: solve cdf(BinomialDistribution(694844 + 705585, q), 705585) = 0.025 (https://www.wolframalpha.com/input?i=solve+cdf%28BinomialDistribution%28694844+%2B+705585%2C+q%29%2C+705585%29+%3D+0.025) として, これを1から引いた値を求める: InverseBetaRegularized(1/40, 694844, 705586) (https://www.wolframalpha.com/input?i=InverseBetaRegularized%281%2F40%2C+694844%2C+705586%29). もしくは, InverseBetaRegularized(0.025, 694844, 705585 + 1) (https://www.wolframalpha.com/input?i=InverseBetaRegularized%281%2F40%2C+694844%2C+705586%29)

p_U: solve cdf(BinomialDistribution(694844 + 705585, p), 694844) = 0.025 (https://www.wolframalpha.com/input?i=InverseBetaRegularized%280.025%2C+694844%2C+705585+%2B+1%29). もしくは, InverseBetaRegularized(1 - 0.025, 694844 + 1, 705585) (https://www.wolframalpha.com/input?i=InverseBetaRegularized%281+-+0.025%2C+694844+%2B+1%2C+705585%29&lang=ja)

(4) p_L: solve cdf(BinomialDistribution(675829 + 692996, q), 692996) = 0.025 (https://www.wolframalpha.com/input?i=solve+cdf%28BinomialDistribution%28694844+%2B+705585%2C+q%29%2C+705585%29+%3D+0.025) として, これを1から引いた値を求める: InverseBetaRegularized(1/40, 694844, 705586) (https://www.wolframalpha.com/input?i=InverseBetaRegularized%281%2F40%2C+694844%2C+705586%29). もしくは, InverseBetaRegularized(0.025, 675829, 692996 + 1) (https://www.wolframalpha.com/input?i=InverseBetaRegularized%280.025%2C+675829%2C+692996+%2B+1%29)

p_U: solve cdf(BinomialDistribution(675829 + 692996, p), 675829) = 0.025 (https://www.wolframalpha.com/input?i=solve+cdf%28BinomialDistribution%28675829+%2B+692996%2C+p%29%2C+675829%29+%3D+0.025). もしくは, InverseBetaRegularized(1 - 0.025, 675829 + 1, 692996) (https://www.wolframalpha.com/input?i=InverseBetaRegularized%281+-+0.025%2C+675829+%2B+1%2C+692996%29)

ここで, (3), (4) の p_L の計算では $\mathrm{Binomial}(n, p)$ で k 以上の確率は $\mathrm{Binomial}(n, 1-p)$ で $n-k$ 以下になる確率に等しいことを使った.

1.9.3 Clopper-Pearsonの信頼区間とそれを与えるP値

(3)と(4)で計算した値から得られる区間 $[p_L, p_U]$ は **母比率に関する信頼度95%のClopper-Pearsonの信頼区間** として統計学ユーザーのあいだでよく知られている.

(1)と(2)での2倍する前の確率は **片側検定のP値** になっている. 2倍後の値は **両側検定のP値** (通常はこちらを使う)の一種になっており, Clopper-Pearsonの信頼区間を与える.

P値は採用した統計モデルとデータの整合性の指標である(P値が小さければ整合性が低い). (1)と(2)で求めたP値の値は極めて小さいということは, 成功確率 $p = 0.5$ の二項分布モデルと2015年と2020年の大阪都構想に関する住民投票の結果の整合性が極めて低いということを意味している. 2015年と2020年の大阪都構想に関する住民投票の結果については, 成功確率 $p = 0.5$ の二項分布モデルは捨て去る必要がある.

(3)と(4)で求めた区間 $[p_L, p_U]$ はデータから計算されるP値が 5% 以上になるパラメータ p の値全体の集合になっている. すなわち, P値に関する 5% の閾値(**有意水準** と呼ばれる)で整合性が低すぎるという理由で捨て去られずにすむ p の値全体が信頼度 $1 - 5\% = 95\%$ の信頼区間になっている. このパターンは一般の場合にもそのまま通用する.

Clopper-Pearsonの信頼区間の効率的計算には, 二項分布の累積分布函数はベータ分布の累積分布函数で書けることが使われる:

$$\begin{aligned} 1 - \mathrm{cdf}(\mathrm{Binomial}(n, p), k - 1) &= \mathrm{cdf}(\mathrm{Beta}(k, n - k + 1), p), \\ \mathrm{cdf}(\mathrm{Binomial}(n, p), k) &= 1 - \mathrm{cdf}(\mathrm{Beta}(k + 1, n - k), p). \end{aligned}$$

これらの公式から, n, k が与えられていて $K \sim \mathrm{Binomial}(n, p)$ のとき, $P(K \geq k) = \alpha/2$ の解 p_L と $P(K \leq k) = \alpha/2$ の解 p_U はそれぞれ次のように書ける:

$$p_L = \mathrm{quantile}(\mathrm{Beta}(k, n - k + 1), \alpha/2), \qquad p_U = \mathrm{quantile}(\mathrm{Beta}(k + 1, n - k), 1 - \alpha/2).$$

ベータ分布の累積分布函数が **正則化された不完全ベータ函数** (regularized incomplete Beta function)になっている:

$$P(T \leq \theta) = I_\theta(\alpha, \beta) = \frac{\int_0^\theta t^{\alpha-1}(1-t)^{\beta-1}}{B(\alpha, \beta)} \quad \text{if} \quad T \sim \mathrm{Beta}(\alpha, \beta)$$

このことから, `quantile` 函数の代わりに $\theta \mapsto p = I_\theta(\alpha, \beta)$ の逆函数を直接使ってもよい:

$$p_L = \mathrm{beta_inc_inv}(k, n - k + 1, \alpha/2)[1], \qquad p_U = \mathrm{beta_inc_inv}(k + 1, n - k, 1 - \alpha/2)[1].$$

Julia言語の `SpecialFunctions.jl` では正則化された不完全ベータ函数とその逆函数はそれぞれ `using SpecialFunctions` した後に `p = beta_inc(α, β, θ)[1]` と `θ = beta_inc_inv(α, β, p)` で使える.

Wolfram言語では `BetaRegularized` (https://reference.wolfram.com/language/ref/BetaRegularized.html) と InverseBetaRegularized (https://reference.wolfram.com/language/ref/InverseBetaRegularized.html) を使う.

Clopper-Pearsonの信頼区間を使うことのメリットは, 二項分布の累積分布函数を和で計算してからパラメータに関する方程式を解くという面倒な手続きを経由せずに, 基本特殊函数の1つである正則化された不完全ベータ函数の逆函数に帰着して効率的に計算できることである. 別の信頼区間の定義の仕方との比較でデメリットもある. Sternの信頼区間(Stern (1954) (https://www.jstor.org/stable/2333026))との相対比較では, Clopper-Pearsonの信頼区間の方が無駄に広くなってしまう場合が多い. np が大きな場合にはどちらを使っても実践的に意味のある差は出ない.

In [14]:

```
n, k = 16, 7

P1 = plot(p -> 1 - cdf(Binomial(n, p), k-1), 0, 1;
    label="1 - cdf(Binomial(n, p), k-1)", legend=:bottomright)
plot!(p -> cdf(Beta(k, n-k+1), p), 0, 1;
    label="cdf(Beta(k, n-k+1), p)", ls=:dash)
title!("n = $n,  k = $k"; xlabel="p")

P2 = plot(p -> cdf(Binomial(n, p), k), 0, 1;
    label="cdf(Binomial(n, p), k)", legend=:topright)
plot!(p -> 1 - cdf(Beta(k+1, n-k), p), 0, 1;
    label="1 - cdf(Beta(k+1, n-k), p)", ls=:dash)
title!("n = $n,  k = $k"; xlabel="p")

plot(P1, P2; size = (800, 250), bottommargin=4Plots.mm)
```

Out[14]:

上のグラフでは $1 - \mathrm{cdf}(\mathrm{Binomial}(n, p), k-1)$ と $\mathrm{cdf}(\mathrm{Beta}(k, n-k+1), p)$ がぴったり一致し, $\mathrm{cdf}(\mathrm{Binomial}(n, p), k)$ と $1 - \mathrm{cdf}(\mathrm{Beta}(k+1, n-k), p)$ がぴったり一致し, 区別がつかなくなっている. ここで $\mathrm{cdf}(D, x)$ は分布 D の累積分布函数を表す:

$$X \sim D \implies \mathrm{cdf}(D, x) = P(X \leq x).$$

1.9.4 信頼区間よりも情報量が大きなP値函数のプロット

二項分布モデルにおいて n, k が与えられたときに, パラメータ p に対してP値を対応させる函数を **P値函数** (p-value function)と呼ぶ. P値函数の値が有意水準 α 以上の p 全体の集合が信頼度 $1-\alpha$ の信頼区間になる. この意味でP値函数はすべての信頼度に関する信頼区間の情報をすべて持っており, 適当な条件の下ではすべての信頼度に関する信頼区間が与えられていればそこからP値函数を逆に作れる. この意味でP値函数と信頼区間達は表裏一体の関係になっている.

以下では上の問題の場合についてのP値函数をプロットしてみよう.

In [15]:

```julia
function plot_pvalue_function!(pvalue_func, n, k; label="", kwargs...)
    p̂ = k/n
    SE = √(p̂*(1 - p̂)/n)
    ps = range(p̂ - 4SE, p̂ + 4SE, 1000)
    plot!(ps, p -> pvalue_func(n, k, p); label, kwargs...)
end

function plot_pvalue_function(pvalue_func, n, k; label="", kwargs...)
    plot()
    plot_pvalue_function!(pvalue_func, n, k; label)
    title!("n = $n, k = $k")
    plot!(; ytick=0:0.1:1, xlabel="parameter p", ylabel="p-value")
    plot!(; kwargs...)
end

function pvalue_clopper_pearson(dist::DiscreteUnivariateDistribution, x)
    min(1, 2cdf(dist, x), 2ccdf(dist, x-1))
end
pvalue_clopper_pearson(n, k, p) = pvalue_clopper_pearson(Binomial(n, p), k)

# (3)
P1 = plot_pvalue_function(pvalue_clopper_pearson, 694844 + 705585, 694844;
    label="Clopper-Pearson", xtick=0:0.0005:1)

# (4)
P2 = plot_pvalue_function(pvalue_clopper_pearson, 675829 + 692996, 675829;
    label="Clopper-Pearson", xtick=0:0.0005:1)

plot(P1, P2; size=(800, 250),
    leftmargin=4Plots.mm, bottommargin=4Plots.mm, rightmargin=4Plots.mm)
```

Out[15]:

データ x を与えられているとき(上の場合には「n 人中 k 人が賛成」というデータを与えられているとき), 統計モデルのパラメータ θ に(上の場合には p に)対してP値を対応させる函数をP値函数と呼ぶ. 上のグラフは大阪都構想に関する住民投票の結果データとし, 統計モデルを二項分布にしたときのP値函数のプロットである.

P値が高いパラメータ値ほど与えられたデータに統計モデルがよくフィット(fit, 適合)している. P値が低い場合にはその逆になる. あまりにも低いP値を持つパラメータ値については, 統計モデルとそのパラメータ値の組が与えられたデータに全然適合していないことになる.

そのとき重要なことは, P値が低いパラメータ値について, そのパラメータ値のみについてデータに適合していないと考えるのではなく, 統計モデルとそのパラメータ値の組がデータにフィットしていないと考えなければいけないことである.

例えば, 上のグラフを見れば, グラフに描いた部分の横軸の目盛りをはみ出しているパラメータ値 $p = 0.5$ のP値はおそろしく低い. 二項分布モデル内で $p = 0.5$ は賛成派の比率がちょうど 50% であることを意味している. $p = 0.5$ のP値がおそろしく低いことを理由に「賛成派の比率が 50% であるという仮説は否定された」のように単純に考えてはいけない. そうではなく, 「賛成派の比率が 50% であることを意味するパラメータ値の二項分布モデルの現実における妥当性は疑わしい」と考えなければいけない. 二項分布モデルも現実における妥当性を疑う対象に入れる必要がある.

この手のことが大学学部生向けの教科書に書かれていないせいで, 大学で統計学の講義を受講した人達の多くが統計学の最も基本的な部分をひどく誤解したまま単位を取得しているように思われる.

誤解を修正するためには次の講義動画がお勧めである:

- 京都大学大学院医学研究科 聴講コース 臨床研究者のための生物統計学「仮説検定とP値の誤解」佐藤 俊哉 医学研究科教授. [YouTube (https://youtu.be/vz9cZnB1d1c)]

その講義で扱われているP値に関するASA声明の翻訳とその翻訳の経緯の解説を以下の場所で読める:

- 統計的有意性とP値に関するASA声明 [pdf (https://www.biometrics.gr.jp/news/all/ASA.pdf)]
- 佐藤俊哉, ASA声明と疫学研究におけるP値, 計量生物学, 2018年38巻2号, pp. 109-115. [link (https://www.jstage.jst.go.jp/article/jjb/38/2/38_109/_article/-char/ja/)]

1.9.5 Sternの信頼区間とそれを与えるP値函数

Sternの信頼区間を与えるP値函数の定義は, 二項分布の確率質量函数を

$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k} \quad (k = 0, 1, \ldots, n)$$

と書くとき,

$$\mathrm{pvalue}_{\mathrm{Stern}}(k|n,p) = \sum_{j \text{ with } P(j|n,p) \leq P(k|n,p)} P(j|n,p)$$

と $P(k|n,p)$ 以下となるような $P(j|n,p)$ 達の和として定義される. すなわち, 二項分布 $\mathrm{Binomial}(n,p)$ においてその値が生じる確率がデータの数値 k が生じる確率以下になる確率としてSternの信頼区間を与えるP値は定義される.

そして, 与えられた n, k について, Sternの信頼区間はこのP値が α 以上になるパラメータ p の範囲として定義される. (実はその定義だと区間になるとは限らない場合が稀にあるので, その場合には適当に定義を訂正することになる.)

1.9.6 Sternの信頼区間を与えるP値函数の実装例

以下はSternの信頼区間を与えるP値函数の実装例である.

Clopper-Pearsonの信頼区間を与えるP値函数の実装(実質1行!)と比較すると相当に複雑になっている.

そして, 実装の仕方によって計算効率に大きな違いが生じていることにも注目せよ.

In [16]:

```julia
x ≲ y = x < y || x ≈ y

# Naive implementation is terribly slow.
function pvalue_stern_naive(dist::DiscreteUnivariateDistribution, x; xmax = 10^6)
    Px = pdf(dist, x)
    Px == 0 && return Px
    ymin, maxdist = minimum(dist), maximum(dist)
    ymax = maxdist == Inf ? xmax : maxdist
    sum(pdf(dist, y) for y in ymin:ymax if 0 < pdf(dist, y) ≲ Px; init = 0.0)
end
pvalue_stern_naive(n, k, p) = pvalue_stern_naive(Binomial(n, p), k)

# Second implementation is very slow.
function pvalue_stern_old(dist::DiscreteUnivariateDistribution, x)
    Px = pdf(dist, x)
    Px == 0 && return Px
    distmin, distmax = extrema(dist)
    m = mode(dist)
    Px ≈ pdf(dist, m) && return one(Px)
    if x < m
        y = m + 1
        while !(pdf(dist, y) ≲ Px)
            y += 1
        end
        cdf(dist, x) + ccdf(dist, y-1)
    else # k > m
        y = m - 1
        while !(pdf(dist, y) ≲ Px)
            y -= 1
        end
        cdf(dist, y) + ccdf(dist, x-1)
    end
end
pvalue_stern_old(n, k, p) = pvalue_stern_old(Binomial(n, p), k)

### The following implementation efficient.

_pdf_le(x, (dist, y)) =  pdf(dist, x) ≲ y

function _search_boundary(f, x0, Δx, param)
    x = x0
    if f(x, param)
        while f(x - Δx, param) x -= Δx end
    else
        x += Δx
        while !f(x, param) x += Δx end
    end
    x
end

function pvalue_stern(dist::DiscreteUnivariateDistribution, x)
    Px = pdf(dist, x)
    Px == 0 && return Px
    m = mode(dist)
    Px ≈ pdf(dist, m) && return one(Px)
    if x < m
        y = _search_boundary(_pdf_le, 2m - x, 1, (dist, Px))
        cdf(dist, x) + ccdf(dist, y-1)
    else # x > m
        y = _search_boundary(_pdf_le, 2m - x, -1, (dist, Px))
        cdf(dist, y) + ccdf(dist, x-1)
    end
end
pvalue_stern(n, k, p) = pvalue_stern(Binomial(n, p), k)
```

Out[16]:

```
pvalue_stern (generic function with 2 methods)
```

In [17]:
```
n = 10
k = -1:11
p = 0.4
a = @time pvalue_stern_naive.(n, k, p)
b = @time pvalue_stern_old.(n, k, p)
c = @time pvalue_stern.(n, k, p)
d = @time pvalue_clopper_pearson.(n, k, p)
@show a ≈ b ≈ c
[a b c d]
```

```
  0.116795 seconds (440.94 k allocations: 24.335 MiB, 99.65% compilation time)
  0.076968 seconds (211.61 k allocations: 11.447 MiB, 99.31% compilation time)
  0.084257 seconds (190.24 k allocations: 10.236 MiB, 20.31% gc time, 99.48% compilation time)
  0.029716 seconds (125.57 k allocations: 6.830 MiB, 98.96% compilation time)
a ≈ b ≈ c = true
```

Out[17]:
```
13×4 Matrix{Float64}:
 0.0          0.0          0.0          0.0
 0.00772434   0.00772434   0.00772434   0.0120932
 0.058652     0.058652     0.058652     0.0927148
 0.333528     0.333528     0.333528     0.33458
 0.749177     0.749177     0.749177     0.764561
 1.0          1.0          1.0          1.0
 0.534186     0.534186     0.534186     0.733793
 0.212596     0.212596     0.212596     0.332477
 0.101119     0.101119     0.101119     0.109524
 0.0183412    0.0183412    0.0183412    0.0245891
 0.00167772   0.00167772   0.00167772   0.00335544
 0.000104858  0.000104858  0.000104858  0.000209715
 0.0          0.0          0.0          0.0
```

In [18]:
```
# (3)の場合に
# pvalue_stern_naive は pvalue_stern_old よりも数百倍遅く,
# pvalue_stern_old は pvalue_stern よりも数百倍遅く,
# pvalue_stern は pvalue_clopper_pearson よりも少し遅い.
n = 694844 + 705585
k = 694844
a = @btime pvalue_stern_naive($n, $k, 0.5)
b = @btime pvalue_stern_old($n, $k, 0.5)
c = @btime pvalue_stern($n, $k, 0.5)
d = @btime pvalue_clopper_pearson($n, $k, 0.5)
@show a ≈ b ≈ c ≈ d
a, b, c, d
```

```
  131.680 ms (0 allocations: 0 bytes)
  545.600 μs (0 allocations: 0 bytes)
  1.510 μs (0 allocations: 0 bytes)
  970.000 ns (0 allocations: 0 bytes)
a ≈ b ≈ c ≈ d = true
```

Out[18]:
```
(1.1304225731963725e-19, 1.1304225734350622e-19, 1.1304225734350622e-19, 1.1304225734350622e-19)
```

In [19]:
```
# 極端な場合
n = 694844 + 705585
k = 600000
b = @btime pvalue_stern_old($n, $k, 0.5)
c = @btime pvalue_stern($n, $k, 0.5)
d = @btime pvalue_clopper_pearson($n, $k, 0.5)
b, c, d
```

```
  94.168 ns (0 allocations: 0 bytes)
  96.215 ns (0 allocations: 0 bytes)
  299.608 ns (0 allocations: 0 bytes)
```

Out[19]:
```
(0.0, 0.0, 0.0)
```

In [20]:
```
# この場合には pvalue_stern_naive はさらに遅い.
n = 100000
k = 49500:50500
a = @time pvalue_stern_naive.(n, k, 0.5)
b = @time pvalue_stern_old.(n, k, 0.5)
c = @time pvalue_stern.(n, k, 0.5)
d = @time pvalue_clopper_pearson.(n, k, 0.5)
@show a ≈ b ≈ c ≈ d;
```

```
 10.735675 seconds (3 allocations: 8.094 KiB)
  0.026518 seconds (3 allocations: 8.094 KiB)
  0.001486 seconds (3 allocations: 8.094 KiB)
  0.001087 seconds (3 allocations: 8.094 KiB)
a ≈ b ≈ c ≈ d = true
```

In [21]:
```
# 以上の実装は超幾何分布でも使える.
dist = Hypergeometric(9, 9, 9)
k = -1:10
a = @time pvalue_stern_naive.(dist, k)
b = @time pvalue_stern_old.(dist, k)
c = @time pvalue_stern.(dist, k)
d = @time pvalue_clopper_pearson.(dist, k)
@show a ≈ b ≈ c ≈ d
[a b c d]
```

```
  0.093384 seconds (360.68 k allocations: 19.780 MiB, 13.67% gc time, 99.47% compilation time)
  0.047890 seconds (166.48 k allocations: 8.991 MiB, 99.12% compilation time)
  0.062885 seconds (150.23 k allocations: 8.066 MiB, 99.29% compilation time)
  0.046932 seconds (171.79 k allocations: 8.868 MiB, 99.11% compilation time)
a ≈ b ≈ c ≈ d = true
```

Out[21]:
```
12×4 Matrix{Float64}:
 0.0          0.0          0.0          0.0
 4.11353e-5   4.11353e-5   4.11353e-5   4.11353e-5
 0.0033731    0.0033731    0.0033731    0.0033731
 0.0566845    0.0566845    0.0566845    0.0566845
 0.346935     0.346935     0.346935     0.346935
 1.0          1.0          1.0          1.0
 1.0          1.0          1.0          1.0
 0.346935     0.346935     0.346935     0.346935
 0.0566845    0.0566845    0.0566845    0.0566845
 0.0033731    0.0033731    0.0033731    0.0033731
 4.11353e-5   4.11353e-5   4.11353e-5   4.11353e-5
 0.0          0.0          0.0          0.0
```

In [22]:
```
# 以上の実装は無限離散分布のPoisson分布でも使える.
dist = Poisson(4)
k = -1:10
a = @time pvalue_stern_naive.(dist, k)
b = @time pvalue_stern_old.(dist, k)
c = @time pvalue_stern.(dist, k)
d = @time pvalue_clopper_pearson.(dist, k)
@show a ≈ b ≈ c
[a b c d]
```

```
  0.624843 seconds (434.08 k allocations: 24.237 MiB, 2.56% gc time, 22.52% compilation time)
  0.047628 seconds (158.29 k allocations: 8.645 MiB, 99.02% compilation time)
  0.058490 seconds (145.67 k allocations: 7.890 MiB, 99.25% compilation time)
  0.046568 seconds (165.03 k allocations: 8.617 MiB, 99.10% compilation time)
a ≈ b ≈ c = true
```

Out[22]:
```
12×4 Matrix{Float64}:
 0.0          0.0          0.0          0.0
 0.0396791    0.0396791    0.0396791    0.0366313
 0.202252     0.202252     0.202252     0.183156
 0.452973     0.452973     0.452973     0.476207
 1.0          1.0          1.0          0.86694
 1.0          1.0          1.0          1.0
 0.609266     0.609266     0.609266     0.742326
 0.306448     0.306448     0.306448     0.429739
 0.12899      0.12899      0.12899      0.221348
 0.0694493    0.0694493    0.0694493    0.102267
 0.0213634    0.0213634    0.0213634    0.0427269
 0.00813224   0.00813224   0.00813224   0.0162645
```

In [23]:

```
# (3)
P1 = plot_pvalue_function(pvalue_stern, 694844 + 705585, 694844;
    label="Stern", xtick=0:0.0005:1)

# (4)
P2 = plot_pvalue_function(pvalue_stern, 675829 + 692996, 675829;
    label="Stern", xtick=0:0.0005:1)

plot(P1, P2; size=(800, 250), leftmargin=4Plots.mm, bottommargin=4Plots.mm)
```

Out[23]:

In [24]:

```
# Clopper-Pearsonの信頼区間を与えるP値函数とSternの信頼区間を与えるP値函数の比較
# (3)の場合にはほぼぴったり一致している.
n, k = 694844 + 705585, 694844
plot(title="p-value functions for n = $n, k = $k", ytick=0:0.1:1)
plot_pvalue_function!(pvalue_stern, n, k; label="Stern")
plot_pvalue_function!(pvalue_clopper_pearson, n, k; label="Clopper-Pearson", ls=:dash)
```

Out[24]:

このように数値的にぴったり一致する場合にはClopper-Pearsonの信頼区間を与えるP値函数とSternの信頼区間のどちらを使うか悩む必要はないだろう. どちらを使っても(ほぼ)同じ結果が得られる.

In [25]:

```
# Clopper-Pearsonの信頼区間を与えるP値函数とSternの信頼区間を与えるP値函数の比較
# n = 50, k = 15 の場合には違いが見える.
n, k = 50, 15
plot(title="p-value functions for n = $n, k = $k", ytick=0:0.1:1)
plot_pvalue_function!(pvalue_stern, n, k; label="Stern")
plot_pvalue_function!(pvalue_clopper_pearson, n, k; label="Clopper-Pearson", ls=:dash)
```

Out[25]:

Sternの信頼区間を与えるP値函数の値はClopper-Pearsonの信頼区間を与えるP値函数の値よりも小さくなりことが多い. (常にそうなるわけではない.) その結果, 対応する信頼区間もSternの側が狭くなってくれることが多い.

1.10 対ごとに独立であっても全体が独立であるとは限らない

確率変数達 X, Y, Z の互いに異なる任意の2つが独立であっても, X, Y, Z の全体が独立でない場合があることを具体的な例を作ることによって証明しよう.

確率変数 X, Y, Z の同時確率質量函数が $P(x, y, z)$ であるとき, X 単独の確率質量函数は $P(x) = \sum_{y,z} P(x, y, z)$ になり, (X, Y) の同時確率質量函数は $P(x, y) = \sum_z P(x, y, z)$ になる. 他の場合も同様である.

注意: 変数名で異なる函数を区別するスタイルを採用したので記号法の運用時に注意すること. このスタイルでは $P(x, y)$ と $P(x, z)$ は異な函数になる. 値を代入する場合には $P(x = 1, y = 1)$ や $P(x = 1, z = 1)$ のように書いて区別できるようにする.

$P(x, y, z)$ で $P(x, y) = P(x)P(y)$, $P(x, z) = P(x)P(z)$, $P(y, z) = P(y)P(z)$ を満たすが $P(x, y, z) \neq P(x)P(y)P(z)$ となるものを具体的に構成すればよい.

例えば以下のような確率質量函数 $P(x, y, z)$ の例を作ることができる:

	$z = 1$		$z = 0$	
	$y = 1$	$y = 0$	$y = 1$	$y = 0$
$x = 1$	$P(1, 1, 1) = 1/4$	$P(1, 0, 1) = 0$	$P(1, 1, 0) = 0$	$P(1, 0, 0) = 1/4$
$x = 0$	$P(0, 1, 1) = 0$	$P(0, 0, 1) = 1/4$	$P(0, 1, 0) = 1/4$	$P(0, 0, 0) = 0$

x, y, z はそれぞれ $1, 0$ を動くとする. このとき, $P(x, y)$, $P(x, z)$, $P(y, z)$ の値はすべて $1/4$ になり, $P(x)$, $P(y)$, $P(z)$ の値はすべて $1/2$ になることがわかる. たとえば, $P(x = 1, y = 1) = P(1, 1, 1) + P(1, 1, 0) = 1/4 + 0 = 1/4$ であり, $P(x = 1) = \sum_{y,z} P(1, y, z) = 1/4 + 0 + 0 + 1/4 = 1/2$.

だから, $P(x, y) = P(x)P(y)$, $P(x, z) = P(x)P(z)$, $P(y, z) = P(y)P(z)$ が成立するが, $P(x, y, z) \neq 1/8 = P(x)P(y)P(z)$ となっている.

このとき, X, Y, Z が同時確率質量函数 $P(x, y, z)$ を持つとすると, そのうちの任意の異なる2つは独立だが, X, Y, Z の全体は独立ではない.

1.11 確率変数の独立性の現実における解釈に関する重大な注意

上の例は具体的には次のような状況だと解釈可能である.

(1) $P(x) = 1/2$ の解釈: $X = 1$ は薬Aを与えたことを, $X = 0$ は薬Bを与えたことを意味する. 全員に確率 $1/2$ で薬Aまたは薬Bを与えた.

$x = 1 =$ 薬A	$P(x = 1) = 1/2$
$x = 0 =$ 薬B	$P(x = 0) = 1/2$

(2) $P(y) = 1/2$ の解釈: $Y = 1$ は薬に効果があったことを, $Y = 0$ は薬に効果が無かったことを意味する. 全体で見たら, $1/2$ の確率で薬には効果があった.

$y=1=$ 効果有	$y=0=$ 効果無
$P(y=1)=1/2$	$P(y=0)=1/2$

(3) $P(z)=1/2$ の解釈: $Z=1$ は女性であることをを, $Z=0$ は男性であることを意味する. 女性である確率と男性である確率は $1/2$ だった.

$z=1=$ 女性	$z=0=$ 男性
$P(z=1)=1/2$	$P(z=0)=1/2$

(4) 男女の区別をやめると, 薬Aも薬Bも効果がある確率は半々であり, 薬Aと薬Bのどちらを与えたかと効果があったかどうかは独立である. 男女を合わせた($z=1,0$ の場合の和を取って得られる)確率質量函数 $P(x,y)$ の表

	$y=1=$ 効果有	$y=0=$ 効果無
$x=1=$ 薬A	$P(x=1,y=1)=1/4$	$P(x=1,y=0)=1/4$
$x=0=$ 薬B	$P(x=0,y=1)=1/4$	$P(x=0,y=0)=1/4$

は$x=1$ の薬Aの場合も $x=0$ の薬Bの場合も男女の区別をやめると薬Aと薬Bで効果に変わりがないことを意味している.

(5) 薬Aと薬Bのどちらを与えたかと男女のどちらであるかは独立である. そのことは効果の有無を意味する $y=1,0$ について和を取って得られる確率質量函数 $P(x,z)$ の表

	$z=1=$ 女性	$z=0=$ 男性
$x=1=$ 薬A	$P(x=1,z=1)=1/4$	$P(x=1,z=0)=1/4$
$x=0=$ 薬B	$P(x=0,z=1)=1/4$	$P(x=0,z=0)=1/4$

からわかる.

(6) 薬Aと薬Bのどちらを与えたかを無視すると, 男女のどちらであるかと薬の効果の有無は独立である. そのことは薬Aと薬Bのどちらを与えたかを意味する $x=1,0$ について和を取って得られる確率質量函数 $P(y,z)$ の表

	$y=1=$ 効果有	$y=0=$ 効果無
$z=1=$ 女性	$P(y=1,z=1)=1/4$	$P(y=0,z=1)=1/4$
$z=0=$ 男性	$P(y=1,z=0)=1/4$	$P(y=0,z=0)=1/4$

からわかる.

(7) しかし, 男女を区別すると全然違う結果が見えて来る. 薬Aは女性だけに効果があり, 薬Bは男性だけに効果がある. $z=1$ の女性の場合に制限した確率質量函数 $P(x,y,z)$ $P(x,y,z)$ の表

	$z=1=$ 女性	
	$y=1=$ 効果有	$y=0=$ 効果無
$x=1=$ 薬A	$P(1,1,1)=1/4$	$P(1,0,1)=0$
$x=0=$ 薬B	$P(0,1,1)=0$	$P(0,0,1)=1/4$

より, $x=1$ の薬Aの場合には $y=1$ の効果有の確率が正であるが, $x=0$ の薬Bの場合には $y=1$ の効果有の確率が0になっている. $z=0$ の男性の場合に制限した確率質量函数 $P(x,y,z)$ の表

	$z=0=$ 男性	
	$y=1=$ 効果有	$y=0=$ 効果無
$x=1=$ 薬A	$P(1,1,0)=0$	$P(1,0,0)=1/4$
$x=0=$ 薬B	$P(0,1,0)=1/4$	$P(0,0,0)=0$

より, $x=1$ の薬Aの場合には $y=1$ の効果有の確率が0で効果がないが, $x=0$ の薬Bの場合には $y=1$ の効果有の確率が正になっている.

このように現実世界において確率変数達が独立か否かは重大な意味を持ち得る. ある重要な条件(上の場合には女性か男性か)を無視して, 「XとYは独立である」(どちらの薬を与えても効果は同じである)と結論すると大変なことになってしまうかもしれない. XとYも, XとZも, YとZも独立であっても, XとYとZの全体は独立でないかもしれない.

2 確率変数達の共分散と相関係数と無相関性

分散と共分散と相関係数と無相関性に関する理論は線形代数における有限実ベクトル空間における内積の理論と本質的に同じものであると考えられる. 分散, 共分散, 相関係数, 無相関性はそれぞれベクトルのノルムの2乗, 内積, ベクトルのあいだの角度 θ に関する $\cos\theta$, ベクトルの直交に対応している.

確率変数	線形代数
分散	ノルムの2乗
共分散	内積
相関係数	$\cos\theta$
無相関	直交

2.1 確率変数達の共分散と相関係数の定義

確率函数 X, Y の期待値をそれぞれ $\mu_X = E[X]$, $\mu_Y = E[Y]$ と書くことにする.

確率変数 X, Y の **共分散** (covariance) $\sigma_{XY} = \mathrm{cov}(X, Y)$ を次のように定義する:

$$\sigma_{XY} = \mathrm{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

$X = Y$ のときこれは X の分散になる. 分散の場合と同様に

$$\begin{aligned}\mathrm{cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[XY] - \mu_X E[Y] - E[X]\mu_Y + \mu_X\mu_Y \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y].\end{aligned}$$

$X = Y$ の場合から $\mathrm{var}(X) = E[X^2] - E[X]^2$ が得られる.

線形代数の言葉を使えば, 共分散 $\sigma_{XY} = \mathrm{cov}(X, Y)$ は内積に対応しており, 分散 $\sigma_X^2 = \mathrm{var}(X)$, $\sigma_Y^2 = \mathrm{var}(Y)$ はノルム(ベクトルの長さ)の二乗に対応しており, 標準偏差 $\sigma_X = \mathrm{std}(X)$, $\sigma_Y = \mathrm{std}(Y)$ はノルム(ベクトルの長さ)に対応している.

さらに確率変数 X, Y の **相関係数** (correlation coefficient) $\rho_{XY} = \mathrm{cov}(X, Y)$ はベクトルのあいだのなす角度 θ に対する $\cos\theta$ の対応物として次のように定義される:

$$\rho_{XY} = \mathrm{cor}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\mathrm{std}(X)\,\mathrm{std}(Y)} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$

相関係数という言葉を見たら $\cos\theta$ を想像し, 分散, 標準偏差, 共分散という言葉を見たら, ベクトルの長さの2乗, ベクトルの長さ, ベクトルの内積を想像すればよい. $X - \mu_X$ や $Y - \mu_Y$ がベクトルに対応している.

非常に重要な注意・警告: 以上で定義したのは確率変数達の共分散と相関係数であり, 後で定義する標本(サンプル, データ)の共分散と相関係数とは異なる数学的対象になるので注意が必要である. この点は平均と分散の場合と同様である. 我々は複雑な現実世界に立ち向かうための道具として数学的に統計モデルを設定し, 統計モデル内の住人として確率変数を考える. そのとき, 確率変数達の平均(期待値), 分散, 共分散, 相関係数を計算することは, 統計モデル内の確率変数の特徴を調べていることになる. 統計モデルを数学的フィクションである現実世界のデータと比較することは以下のように行われる:

1. 数学的フィクションである統計モデル内に住んでる現実世界のデータのモデル化になっている確率変数の特徴を調べる.
2. 現実世界で取得したデータの特徴を調べる.
3. 統計モデル内の確率変数の特徴と現実世界で取得したデータの特徴を比較してみる.

上で定義したの確率変数達の共分散と相関係数は1の段階で使用される. 2の段階で使われる道具については後で説明する.

数学的な注意: 統計学的な位置付けとして, 確率変数達の平均, 分散, 共分散, 相関係数などと, 標本(サンプル, データ)平均, 分散, 共分散, 相関係数は異なる対象だとみなす必要があることはすぐ上で説明した. しかし, 純粋に数学的性質を調べる場合には, 後者は前者の特別な場合になっていることから, 前者の数学的性質を調べれば後者の数学的性質もわかるという仕組みになっている. 初学者が混乱しがちになる原因の1つは, 数学的に鋭い初学者が確率変数達の扱いとデータ(サンプル, 標本)の取り扱いが数学的に似ていることに気付いてしまうせいで, それらの位置付けを混同してしまいがちなことにもあるように思われる.

2.2 確率変数達の無相関の定義

X, Y の相関係数 $\rho_{XY} = \mathrm{cor}(X, Y)$ が 0 になるとき, 確率変数達 X, Y は **無相関** であるという. X と Y が無相関であることは $X - \mu_X$ と $Y - \mu_Y$ が「直交している」と考えてもよい.

注意: X, Y の相関係数 $\rho_{XY} = \mathrm{cor}(X, Y)$ が **近似的に** 0 **になっている場合** には($\cos\theta$ に対応する量が近似的に 0 になっている場合には), 確率変数達 X, Y は **近似的に無相関** であると考える. そのとき注意するべきことは共分散 $\sigma_{XY} = \mathrm{cov}(X, Y)$ が近似的に 0 になっていても(内積に当たる量が近似的に 0 になっていても), 相関係数が 0 からほど遠い値になっていて, 近似的に無相関とは言えない場合があることである. 下にある関連問題の節の内容を参照せよ.

確率変数達 $X_1, \dots, X_n$ が **無相関** であるとは, そのうちの互いに異なる任意の2つが無相関であることだと定める. $\mu_i = E[X_i]$, $\sigma_i^2 = \mathrm{var}(X_i)$ のとき, $X_1, \dots, X_n$ が無相関であることは,

$$\mathrm{cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] = \sigma_i^2\delta_{ij}$$

$$\mathrm{cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] = \sigma_{ij}$$

と書けることと同値である. ここで δ_{ij} は $i = j$ の場合にのみ 1 になり, それ以外の場合に 0 になるKroneckerのデルタである.

2.3 問題: 確率変数の相関係数の計算例

確率変数の組 (X, Y) は確率 $0.1, 0.2, 0.3, 0.4$ でそれぞれ $(10, -50), (-4, -10), (0, 20), (2, 0)$ という値になると仮定する:

probability	0.1	0.2	0.3	0.4
X	10	-4	0	2
Y	-50	-10	20	0

確率変数 X, Y の相関係数を小数点以下第2桁まで求めよ.

解答例: まず X, Y の期待値を求める:

$$\begin{aligned}\mu_X &= E[X] = 10 \cdot 0.1 + (-4) \cdot 0.2 + 0 \cdot 0.3 + 2 \cdot 0.4 = 1,\\ \mu_Y &= E[Y] = (-50) \cdot 0.1 + (-10) \cdot 0.2 + 20 \cdot 0.3 + 0 \cdot 0.4 = -1.\end{aligned}$$

ゆえに期待値を X, Y から引いた値の分布は次の表のようになる:

probability	0.1	0.2	0.3	0.4
$X - \mu_X$	9	-5	-1	1
$Y - \mu_Y$	-49	-9	21	1

この表を用いて確率変数達としての X, Y の分散と共分散を求める:

$$\begin{aligned}\mathrm{var}(X) &= 9^2 \cdot 0.1 + (-5)^2 \cdot 0.2 + (-1)^2 \cdot 0.3 + 1^2 \cdot 0.4 = 13.8,\\ \mathrm{var}(Y) &= (-49)^2 \cdot 0.1 + (-9)^2 \cdot 0.2 + 21^2 \cdot 0.3 + 1^2 \cdot 0.4 = 389,\\ \mathrm{cov}(X, Y) &= 9(-49) \cdot 0.1 + (-5)(-9) \cdot 0.2 + (-1)21 \cdot 0.3 + 1 \cdot 1 \cdot 0.4 = -41.\end{aligned}$$

相関係数を求めよう:

$$\mathrm{cor}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}} \approx -0.56.$$

解答終

In [26]:
```
@show E_X = 10*0.1 + (-4)*0.2 + 0*0.3 + 2*0.4
@show E_Y = (-50)*0.1 + (-10)*0.2 + 20*0.3 + 0*0.4
```

```
E_X = 10 * 0.1 + -4 * 0.2 + 0 * 0.3 + 2 * 0.4 = 1.0
E_Y = -50 * 0.1 + -10 * 0.2 + 20 * 0.3 + 0 * 0.4 = -1.0
```

Out[26]:
```
-1.0
```

In [27]:
```
@show var_X = 9^2*0.1 + (-5)^2*0.2 + (-1)^2*0.3 + 1^2*0.4
@show var_Y = (-49)^2*0.1 + (-9)^2*0.2 + 21^2*0.3 + 1^2*0.4
@show cov_XY = 9*(-49)*0.1 + (-5)*(-9)*0.2 + (-1)*21*0.3 + 1*1*0.4;
```

```
var_X = 9 ^ 2 * 0.1 + (-5) ^ 2 * 0.2 + (-1) ^ 2 * 0.3 + 1 ^ 2 * 0.4 = 13.8
var_Y = (-49) ^ 2 * 0.1 + (-9) ^ 2 * 0.2 + 21 ^ 2 * 0.3 + 1 ^ 2 * 0.4 = 389.0
cov_XY = 9 * -49 * 0.1 + -5 * -9 * 0.2 + -1 * 21 * 0.3 + 1 * 1 * 0.4 = -41.0
```

In [28]:
```
@show cor_XY = cov_XY/(√var_X * √var_Y);
```

```
cor_XY = cov_XY / (√var_X * √var_Y) = -0.5595894432479921
```

In [29]:

```
X = [10, -4, 0, 2]
Y = [-50, -10, 20, 0]
P = [0.1, 0.2, 0.3, 0.4]
@show E_X = sum(X[i]*P[i] for i in eachindex(P))
@show E_Y = sum(Y[i]*P[i] for i in eachindex(P))
@show var_X = sum((X[i] - E_X)^2*P[i] for i in eachindex(P))
@show var_Y = sum((Y[i] - E_Y)^2*P[i] for i in eachindex(P))
@show cov_XY = sum((X[i] - E_X) * (Y[i] - E_Y) * P[i] for i in eachindex(P))
@show cor_XY = cov_XY/(√var_X * √var_Y);
```

```
E_X = sum((X[i] * P[i] for i = eachindex(P))) = 1.0
E_Y = sum((Y[i] * P[i] for i = eachindex(P))) = -1.0
var_X = sum(((X[i] - E_X) ^ 2 * P[i] for i = eachindex(P))) = 13.8
var_Y = sum(((Y[i] - E_Y) ^ 2 * P[i] for i = eachindex(P))) = 389.0
cov_XY = sum(((X[i] - E_X) * (Y[i] - E_Y) * P[i] for i = eachindex(P))) = -41.0
cor_XY = cov_XY / (√var_X * √var_Y) = -0.5595894432479921
```

2.4 問題: Cauchy-Schwarzの不等式

確率変数達 X, Y について, 以下の不等式を示せ:

$$\mathrm{cov}(X, Y) \leq \mathrm{std}(X)\,\mathrm{std}(Y).$$

簡単のため X, Y が同時確率質量函数を持つと仮定してよい.

注意: Cauchy-Schwarzの不等式より, $\mathrm{std}(X)\,\mathrm{std}(Y) \neq 0$ のとき, 相関係数

$$\rho_{XY} = \mathrm{cor}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\mathrm{std}(X)\,\mathrm{std}(Y)}$$

が -1 以上 1 以下になることがわかる. これより, 0 以上 π 以下の θ を

$$\rho_{XY} = \mathrm{cor}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\mathrm{std}(X)\,\mathrm{std}(Y)} = \cos\theta$$

によって定めることができる. この θ はベクトルのあいだの角度の確率変数の場合の類似物になっている.

解答例1: 右辺の2乗から左辺の2乗を引いた結果が 0 以上になること, すなわち, 次を示せばよい:

$$\mathrm{var}(X)\,\mathrm{var}(Y) - \mathrm{cov}(X, Y)^2 \geq 0. \tag{$*$}$$

確率変数の分散や共分散は確率変数から定数を引いても変わらないので, X, Y からそれらの期待値を引いたものを考えることによって, X, Y の期待値は 0 であると仮定してよい. このとき,

$$\mathrm{var}(X) = E[X^2], \quad \mathrm{var}(Y) = E[Y^2], \quad \mathrm{cov}(X, Y) = E[XY].$$

X, Y の同時確率質量函数を $P(x, y)$ と書く:

$$E[f(X, Y)] = \sum_{x,y} f(x, y)P(x, y).$$

このとき,

$$E[f(X, Y)]E[g(X, Y)] = \sum_{x,y,x',y'} f(x, y)g(x', y')P(x, y)P(x', y').$$

以上の準備のもとで,

$$\begin{aligned}
2\,\mathrm{var}(X)\,\mathrm{var}(Y) &= E[X^2]E[Y^2] + E[Y^2]E[X^2] \\
&= \sum_{x,y,x',y'} (x^2y'^2 + y^2x'^2)P(x, y)P(x', y'), \\
2\,\mathrm{cov}(X, Y)^2 0 &= 2E[XY]E[XY] = \sum_{x,y,x',y'} 2xyx'y'\,P(x, y)P(x', y') \\
&= \sum_{x,y,x',y'} 2xy'yx'\,P(x, y)P(x', y').
\end{aligned}$$

ゆえに,

$$
\begin{aligned}
2\operatorname{var}(X)\operatorname{var}(Y) - 2\operatorname{cov}(X,Y)^2 &= \sum_{x,y,x',y'} (x^2y'^2 + y^2x'^2 - 2xy'yx')P(x,y)P(x',y') \\
&= \sum_{x,y,x',y'} (xy' - yx')^2 P(x,y)P(x',y') \geq 0.
\end{aligned}
$$

これより$(*)$が成立することがわかる.

解答終

上の解答例でやっていることは本質的に (X,Y) と独立で同分布な確率変数 (X',Y') に対応する離散変数 x', y' を導入したことである. そのことに気付けば, 同時確率質量函数の存在に頼らない一般的な証明法も得られる.

解答例2(より一般的な場合): 右辺の2乗から左辺の2乗を引いた結果が 0 以上になること, すなわち, 次を示せばよい:

$$
\operatorname{var}(X)\operatorname{var}(Y) - \operatorname{cov}(X,Y)^2 \geq 0. \tag{$*$}
$$

確率変数の分散や共分散は確率変数から定数を引いても変わらないので, X, Y からそれらの期待値を引いたものを考えることによって, X, Y の期待値は 0 であると仮定してよい. このとき,

$$
\operatorname{var}(X) = E[X^2], \quad \operatorname{var}(Y) = E[Y^2], \quad \operatorname{cov}(X,Y) = E[XY].
$$

さらに, (X',Y') は (X,Y) と同じ分布を持つ (X,Y) と独立な確率変数であるとする. このとき, 次の公式を使用できる:

$$
E[f(X,Y)] = E[f(X',Y')], \quad E[f(X,Y)g(X',Y')] = E[f(X,Y)]E[g(X',Y')].
$$

以上の準備のもとで,

$$
\begin{aligned}
&2\operatorname{var}(X)\operatorname{var}(Y) = 2E[X^2]E[Y^2] = E[X^2]E[Y'^2] + E[Y^2]E[X'^2] = E[X^2Y'^2 + Y^2X'^2], \\
&2\operatorname{cov}(X,Y)^2 = 2E[XY]^2 = 2E[XY]E[X'Y'] = E[2XY'YX'].
\end{aligned}
$$

ゆえに,

$$
\begin{aligned}
2\operatorname{var}(X)\operatorname{var}(Y) - 2\operatorname{cov}(X,Y)^2 &= E[X^2Y'^2 + Y^2X'^2] - E[2XY'YX'] \\
&= E[X^2Y'^2 + Y^2X'^2 - 2XY'YX'] \\
&= E[(XY' - X'Y)^2] \geq 0.
\end{aligned}
$$

これより$(*)$が成立することがわかる.

解答終

注意: 線形代数の教科書におけるCauchy-Schwarzの不等式の証明と以上で紹介した証明を比較してみよ. もしも比較した線形代数の教科書の証明が以上で紹介した証明と全く違っているならば, その教科書の文脈において上で紹介した証明の方法が使えないかどうかを考えてみよ. 参照にした教科書に上で紹介した証明と全く異なる証明が載っているならば, 上で証明した確率変数に関するCauchy-Schwarzの不等式をその教科書の方法で証明してみよ.

2.5 問題: 等確率有限離散分布の相関係数と $\cos\theta$ の関係

$a_i, b_i \in \mathbb{R}$ であるとし, 確率変数の組 (X,Y) は同じ確率 $1/n$ で $(a_1,b_1),\dots,(a_n,b_n)$ のどれかの値になるものと仮定する. この仮定は

$$
E[f(X,Y)] = \frac{1}{n}\sum_{i=1}^{n} f(a_i,b_i)
$$

の成立と同値である(確率変数の組が定義されていることはその函数の期待値が定義されていることと同じことだと考えてよい). さらに,

$$
\mu_X = E[X] = \frac{1}{n}\sum_{i=1}^{n} a_i, \quad \mu_Y = E[Y] = \frac{1}{n}\sum_{i=1}^{n} b_i
$$

とおき, ベクトル v_X, v_Y を次のように定める:

$$
v_X = \frac{1}{\sqrt{n}}\begin{bmatrix} a_1 - \mu_X \\ \vdots \\ a_n - \mu_X \end{bmatrix}, \quad v_Y = \frac{1}{\sqrt{n}}\begin{bmatrix} b_1 - \mu_Y \\ \vdots \\ b_n - \mu_y \end{bmatrix}
$$

これらの内積(対応する成分の積の和)を (v_X, v_Y) と書き, ノルム(成分の2乗の和の平方根)を $\|v_X\|, \|v_Y\|$ と書き, v_X, v_Y のあいだの角度を θ と書くことにし, $\|v_X\| > 0, \|v_Y\| > 0$ であると仮定する. このとき, X, Y の相関係数 $\operatorname{cor}(X,Y)$ について次が成立することを示せ:

$$\mathrm{cor}(X,Y) = \frac{(v_X, v_Y)}{\|v_X\|\|v_Y\|} = \cos\theta.$$

注意: 相関係数とベクトルのあいだの角度の関係についてはこの場合をイメージしておけばよい.

解答例: 内積とノルムの定義より,

$$\begin{aligned}
\|v_X\|^2 &= \frac{1}{n}\sum_{i=1}^{n}(a_i - \mu_X)^2 = E[(X-\mu_X)^2] = \mathrm{var}(X),\\
\|v_Y\|^2 &= \frac{1}{n}\sum_{i=1}^{n}(b_i - \mu_Y)^2 = E[(Y-\mu_Y)^2] = \mathrm{var}(Y),\\
(v_X, v_Y) &= \frac{1}{n}\sum_{i=1}^{n}(a_i - \mu_X)(b_i - \mu_Y) = E[(X-\mu_X)(Y-\mu_Y)] = \mathrm{cov}(X,Y).
\end{aligned}$$

ゆえに,

$$\mathrm{cor}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}} = \frac{(v_X, v_Y)}{\|v_X\|\|v_Y\|} = \cos\theta.$$

解答終

2.6 問題: 相関係数の計算

確率変数達 X, Y は無相関であるとし, それぞれの期待値と分散はどちらも 0 と 1 であるとし, $a, b, c, d \in \mathbb{R}$, $a^2 + b^2 > 0$, $c^2 + d^2 > 0$ であると仮定する. 確率変数 A, B を次のように定める:

$$A = aX + bY, \quad B = cX + dY.$$

確率変数達 A, B の相関係数を求めよ.

解答例: X, Y の期待値はどちらも 0 なのでそれらの一次結合である A, B の期待値も 0 になる. さらに, X, Y の分散はどちらも 1 でそれらは無相関だと仮定したので,

$$\begin{aligned}
&\mathrm{var}(A) = a^2\,\mathrm{var}(X) + b^2\,\mathrm{var}(Y) = a^2 + b^2,\\
&\mathrm{var}(B) = c^2\,\mathrm{var}(X) + d^2\,\mathrm{var}(Y) = c^2 + d^2,\\
&\mathrm{cov}(A,B) = E[AB] = E[acX^2 + (ad+bc)XY + bdY^2]\\
&\quad = ac\,\mathrm{var}(X) + (ad+bc)\,\mathrm{cov}(X,Y) + bd\,\mathrm{var}(Y) = ac + bd.
\end{aligned}$$

ゆえに

$$\mathrm{cor}(A,B) = \frac{\mathrm{cov}(A,B)}{\sqrt{\mathrm{var}(A)}\sqrt{\mathrm{var}(B)}} = \frac{ac+bd}{\sqrt{a^2+b^2}\sqrt{c^2+d^2}}.$$

解答終

注意: ベクトル v_A, v_B を

$$v_A = \begin{bmatrix} a \\ b \end{bmatrix}, \quad v_B = \begin{bmatrix} c \\ d \end{bmatrix}$$

と定め, これらの標準内積(対応する成分の積の和)を (v_A, v_B) と書き, それらのノルム(すべての成分の2乗の和の平方根)を $\|v_A\|, \|v_B\|$ と書き, それらの間の角度を θ と書くとき,

$$\mathrm{cor}(A,B) = \frac{(v_A, v_B)}{\|v_A\|\|v_B\}} = \cos\theta$$

が成立している. 上の問題の状況において, 確率変数 A, B の相関係数が 1 に近いことはベクトル v_A, v_B の向きがほぼ同じであることと同値であり, A, B の相関係数が -1 に近いことは v_A, v_B の向きがほぼ正反対であることに同値であり, A, B が無相関に近いことは v_A, v_B がほぼ直交していることと同値である.

2.7 問題: 共分散が 0 に近くても相関係数が 0 から遠い場合がある

平均が 0 の具体的な確率変数達 X, Y の組で, 共分散 $\sigma_{XY} = \mathrm{cov}(X,Y) = E[XY]$ は 0 に近い値だが, 相関係数 $\rho_{XY} = \mathrm{cov}(X,Y) = \mathrm{cov}(X,Y)/(\mathrm{std}(X)\,\mathrm{std}(Y))$ が 0 から遠い値になっているものを構成せよ.

解答例: $\varepsilon, \delta > 0$ であるとし, 確率変数の組 (X,Y) は同じ確率 1/3 で $(\varepsilon+\delta, \varepsilon)$, $(-\varepsilon, \delta)$, $(-\delta, -\varepsilon-\delta)$ のどれかの値になるものとする. このとき, X, Y の平均はそれぞれ

$$E[X] = \frac{(\varepsilon+\delta)+(-\varepsilon)+(-\delta)}{3} = 0, \quad E[Y] = \frac{\varepsilon+\delta+(-\varepsilon-\delta)}{3} = 0$$

となる. X と Y の分散と共分散と相関係数は以下のようになる:

$$\begin{aligned}
\mathrm{var}(X) &= E[X^2] = \frac{(\varepsilon+\delta)^2+(-\varepsilon)^2+(-\delta)^2}{3} = \frac{2(\varepsilon^2+\varepsilon\delta+\delta^2)}{3}, \\
\mathrm{var}(Y) &= E[Y^2] = \frac{\varepsilon^2+\delta^2+(-\varepsilon-\delta)^2}{3} = \frac{2(\varepsilon^2+\varepsilon\delta+\delta^2)}{3}, \\
\mathrm{cov}(X,Y) &= E[XY] = \frac{(\varepsilon+\delta)\varepsilon+(-\varepsilon)\delta+(-\delta)(-\varepsilon-\delta)}{3} = \frac{(\varepsilon^2+\varepsilon\delta+\delta^2)}{3}, \\
\mathrm{cor}(X,Y) &= \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}} = \frac{1}{2}.
\end{aligned}$$

$\varepsilon, \delta > 0$ の値が小さいとき, X, Y の共分散は小さくなるが, 相関係数は 1/2 で全然 0 に近くない. **解答終**

注意: 3次元ベクトル v_X, v_Y を以下のように定める:

$$v_X = \frac{1}{\sqrt{3}}\begin{bmatrix} \varepsilon+\delta \\ -\varepsilon \\ -\delta \end{bmatrix}, \quad v_Y = \frac{1}{\sqrt{3}}\begin{bmatrix} \varepsilon \\ \delta \\ -\varepsilon-\delta \end{bmatrix}.$$

このとき, これらの標準内積(対応する成分の積の和)を (v_X, v_Y) と書き, それらのノルム(すべての成分の2乗の和の平方根)を $\|v_X\|, \|v_Y\|$ と書き, それらの間の角度を θ と書くとき,

$$\begin{aligned}
&\mathrm{var}(X) = \|v_X\|^2, \quad \mathrm{var}(Y) = \|v_Y\|^2, \quad \mathrm{cov}(X,Y) = (v_X, v_Y), \\
&\mathrm{cor}(X,Y) = \frac{(v_X, v_Y)}{\|v_X\|\|v_Y\|} = \cos\theta
\end{aligned}$$

が成立していることに注意せよ.

2.8 問題: 独立ならば無相関である (実質1行で解ける)

確率変数達 X, Y が独立ならば確率変数達 X, Y は無相関であることを示せ.

解答例: X, Y は独立なので $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ となる. ゆえに, $\mu_X = E[X]$, $\mu_Y = E[Y]$ とおくと,

$$E[(X-\mu_X)(Y-\mu_Y)] = E[X-\mu_X]E[Y-\mu_Y] = (E[X]-\mu_X)(E[Y]-\mu_Y) = 0 \cdot 0 = 0.$$

解答終

2.9 問題: 無相関でも独立とは限らない

確率変数 X, Y で無相関だが独立でないものを具体的に構成せよ.

解答例1: 確率質量函数 $P(x, y)$ を次の表の通りに定める:

	$y=1$	$y=2$	$y=3$	
$x=1$	$P(1,1)=0$	$P(1,2)=1/4$	$P(1,3)=0$	$P(x=1)=1/4$
$x=2$	$P(2,1)=1/4$	$P(2,2)=0$	$P(2,3)=1/4$	$P(x=2)=2/4$
$x=3$	$P(3,1)=0$	$P(3,2)=1/4$	$P(3,3)=0$	$P(x=3)=1/4$
	$P(y=1)=1/4$	$P(y=2)=2/4$	$P(y=3)=1/4$	

確率変数 X, Y は同時確率質量函数 $P(x, y)$ を持つとする. このとき,

$$E[X] = E[Y] = 1 \cdot (1/4) + 2 \cdot (2/4) + 3 \cdot (1/4) = 2.$$

であり, 確率が 0 でない (x, y) については x または y が X と Y の期待値 2 になるので, 共分散は 0 になる:

$$\begin{aligned}
\mathrm{cov}(X,Y) = E[(X-2)(Y-2)] &= \sum_{x,y}(x-2)(y-2)P(x,y) \\
&= (2-2)(1-2)(1/4) + (1-2)(2-2)(1/4) \\
&\quad + (3-2)(2-2)(1/4) + (2-2)(3-2)(1/4) = 0.
\end{aligned}$$

これで X, Y は無相関であることがわかった.

しかし, たとえば $P(1,1) = 0$ と $P(x=1)P(y=1) = (1/4)(1/4)$ が等しくないので X, Y は独立ではない.

解答終

解答例2: 確率質量函数 $P(x, y)$ を次の表の通りに定めても上と同様に, X, Y は無相関だが独立ではないことを示せる:

	$y = 1$	$y = 2$	$y = 3$	
$x = 1$	$P(1, 1) = 1/8$	$P(1, 2) = 1/8$	$P(1, 3) = 1/8$	$P(x = 1) = 3/8$
$x = 2$	$P(2, 1) = 1/8$	$P(2, 2) = 0$	$P(2, 3) = 1/8$	$P(x = 2) = 2/8$
$x = 3$	$P(3, 1) = 1/8$	$P(3, 2) = 1/8$	$P(3, 3) = 1/8$	$P(x = 3) = 3/8$
	$P(y = 1) = 3/8$	$P(y = 2) = 2/8$	$P(y = 3) = 3/8$	

解答終

注意: 3×3 の表をすべて 1/9 で埋めてしまうと, X, Y は独立になってしまう.

解答例3: $\mathbb{R}^2$ における単位円盤上の一様分布の確率密度函数を次のように定める:

$$p(x, y) = \begin{cases} 1/\pi & (x^2 + y^2 \leq 1) \\ 0 & (x^2 + y^2 > 1) \end{cases}$$

これを同時確率密度函数として持つ確率変数 X, Y を考える:

$$E[f(X, Y)] = \iint_{\mathbb{R}^2} f(x, y)p(x, y)\, dx\, dy.$$

単位円盤の対称性から $E[X] = E[Y] = 0$ となることがわかる. (具体的に積分を計算しても易しい.) それらの共分散は

$$\mathrm{cov}(X, Y) = E[XY] = \frac{1}{\pi} \iint_{x^2+y^2 \leq 1} xy\ dx\, dy$$

と書けるが, $xy \geq 0$ の部分の積分と $xy \leq 0$ の部分の積分が単位円盤の対称性より互いに打ち消しあって $\mathrm{cov}(X, Y) = 0$ となることがわかる. X 単独の密度函数は

$$p(x) = \frac{1}{\pi} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} dy = \frac{2}{\pi}\sqrt{1 - x^2}$$

となり, 同様にして $p(y) = (2/\pi)\sqrt{1 - y^2}$ となるが, $-1 < x, y < 1$ かつ $x^2 + y^2 > 1$ のとき, $p(x, y) = 0$ となるが, $p(x)p(y) \neq 0$ となるので, それらは等しくない. ゆえに X, Y は独立ではない.

解答終

この解答例3については, 下の方でプロットした (X, Y) の分布のサンプルの散布図も参照して欲しい. 散布図を見れば直観的にどうなっているかを把握しやすい.

In [30]:

```
n = 10^4
XY = [(r = √rand(); t = 2π*rand(); (r*cos(t), r*sin(t))) for _ in 1:n]
X, Y = first.(XY), last.(XY)
@show cov(X, Y) cor(X, Y)
P1 = scatter(X, Y; msc=:auto, ms=2, alpha=0.7, label="", xlabel="x", ylabel="y")
P2 = histogram(X; norm=true, alpha=0.3, bin=41, label="X")
plot!(x -> 2/π*√(1 - x^2), -1, 1; label="", lw=2)
P3 = histogram(X; norm=true, alpha=0.3, bin=41, label="Y")
plot!(y -> 2/π*√(1 - y^2), -1, 1; label="", lw=2)
plot(P1, P2, P3; size=(800, 400), layout=@layout [a [b; c]])
```

```
cov(X, Y) = -0.0005083072927225841
cor(X, Y) = -0.0020186738360558273
```

Out[30]:

単位円盤上の一様分布は「無相関だが独立ではない場合」の例になっている.

この「無相関だが独立ではない場合」の例のサンプルに含まれる点の個数を減らすと以下のように見える.

In [31]:

```
n = 100
XY = [(r = √rand(); t = 2π*rand(); (r*cos(t), r*sin(t))) for _ in 1:n]
X, Y = first.(XY), last.(XY)
@show cov(X, Y) cor(X, Y)
P1 = scatter(X, Y; msc=:auto, label="", xlabel="x", ylabel="y")
plot!(; xlim=(-1.05, 1.05), ylim=(-1.05, 1.05))
plot!(; size=(400, 400))
```

```
cov(X, Y) = 0.011653127995250636
cor(X, Y) = 0.046497931398409825
```

Out[31]:

上の場合において X, Y はどちらも **半円分布** (semicircle distribution)に従っているが, 独立ではない. X, Y の分布が以上と同じ半円分布のときと X, Y が独立になるならば, (X, Y) の分布のサンプルは以下のようになる.

In [32]:

```
n = 10^4
sc = Semicircle(1)
X, Y = rand(sc, n), rand(sc, n)
@show cov(X, Y) cor(X, Y)
P1 = scatter(X, Y; msc=:auto, ms=2, alpha=0.7, label="", xlabel="x", ylabel="y")
P2 = histogram(X; norm=true, alpha=0.3, bin=41, label="X")
plot!(x -> 2/π*√(1 - x^2), -1, 1; label="", lw=2)
P3 = histogram(X; norm=true, alpha=0.3, bin=41, label="Y")
plot!(y -> 2/π*√(1 - y^2), -1, 1; label="", lw=2)
plot(P1, P2, P3; size=(800, 400), layout=@layout [a [b; c]])
```

```
cov(X, Y) = 0.004473060988369731
cor(X, Y) = 0.0179942903137646
```

Out[32]:

これは X, Y が独立な場合であり, ゆえに無相関にもなっている.

サンプルが含む点の個数を減らすと以下のように見える.

In [33]:

```
n = 100
sc = Semicircle(1)
X, Y = rand(sc, n), rand(sc, n)
@show cov(X, Y)
P1 = scatter(X, Y; msc=:auto, label="", xlabel="x", ylabel="y")
plot!(; xlim=(-1.05, 1.05), ylim=(-1.05, 1.05))
plot!(; size=(400, 400))
```

```
cov(X, Y) = 0.0030840296722668315
```

Out[33]:

2.10 問題: 無相関な確率変数達の和の分散はそれぞれの分散の和になる

$X_1, \ldots, X_n$ は無相関とは限らない(ゆえに独立とは限らない)確率変数達であるとする. このとき, 期待値を取る操作の線形性より,

$$E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n]$$

となる. 確率変数達の和の期待値は, 独立性や無相関性が成立していなくても, それぞれの期待値の和になる.

無相関性を仮定すると, 分散についても簡明な結果を得ることができる.

$X_1, \ldots, X_n$ が **無相関な確率変数達** ならば(特に独立な確率変数ならば), 次が成立することを示せ:

$$\mathrm{var}(X_1 + \cdots + X_n) = \mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n).$$

この結果は今後空気のごとく使われる.

ヒント: 互いに直交するベクトル達 $v_1, \ldots, v_n$ について, 内積を $(\ ,\ )$ と書くとき, $(v_i, v_j) = \delta_{ij}\|v_i\|^2$ が成立することを使えば, $\|v_1 + \cdots + v_n\|^2 = \|v_1\|^2 + \cdots + \|v_n\|^2$ を示せることと本質的に同じ. この結果はPythagorasの定理(平面の場合は三平方の定理)そのものである.

解答例: $\mu_i = E[X_i]$ とおくと,

$$E\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n \mu_i$$

が成立しており, さらに, $\sigma_i^2 = \mathrm{var}(X_i) = E[(X_i - \mu_i)^2]$ とおくと, $X_1, \ldots, X_n$ が無相関であることより,

$$E[(X_i - \mu_i)(X_j - \mu_j)] = \mathrm{cov}(X_i, X_j) = \sigma_i^2 \delta_{ij}$$

が成立しているので(δ_{ij} は $i = j$ の場合にのみ 1 でそれ以外のとき 0), 一般に

$$\left(\sum_{i=1}^n a_i\right)^2 = \sum_{i=1}^n a_i \sum_{i=j}^n a_j = \sum_{i,j=1}^n a_i a_j$$

と計算できることを使うと,

$$
\begin{aligned}
\mathrm{var}\left(\sum_{i=1}^n X_i\right) &= E\left[\left(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i\right)^2\right] = E\left[\left(\sum_{i=1}^n (X_i - \mu_i)\right)^2\right] \\
&= E\left[\sum_{i,j=1}^n (X_i - \mu_i)(X_j - \mu_j)\right] = \sum_{i,j=1}^n E\left[(X_i - \mu_i)(X_j - \mu_j)\right] \\
&= \sum_{i,j=1}^n \sigma_i^2 \delta_{ij} = \sum_{i=1}^n \sigma_i^2 = \sum_{i=1}^n \mathrm{var}(X_i).
\end{aligned}
$$

解答終

2.11 問題(メタアナリシスの出発点): 共通の期待値と異なる分散を持つ確率変数の荷重平均

$X_1, \dots, X_n$ は無相関な確率変数達であるとし, X_i 達は共通の期待値 μ と共通でない分散 $\sigma_i^2 > 0$ を持つと仮定する:

$$
E[X_1] = \cdots = E[X_n] = \mu, \quad E[X_i] = \sigma_i^2 \quad (i = 1, \dots, n)
$$

このとき, 共通の期待値 μ の推定量 $\hat{\mu}$ を荷重 w_i による荷重平均

$$
\hat{\mu} = \sum_{i=1}^n w_i X_i, \quad \sum_{i=1}^n w_i = 1
$$

の形で定義したい. 条件 $\sum_{i=1}^n w_i = 1$ より, 不偏性 $E[\hat{\mu}] = \mu$ がただちに導かれる. このとき, $\hat{\mu}$ の分散 $\mathrm{var}(\hat{\mu})$ を最小化する荷重 w_i は

$$
w_i = \frac{1/\sigma_i^2}{\sum_{i=1}^n 1/\sigma_i^2}, \quad \text{i.e.} \quad \hat{\mu} = \frac{\sum_{i=1}^n X_i/\sigma_i^2}{\sum_{i=1}^n 1/\sigma_i^2}
$$

になることを示せ. さらに, このとき

$$
\mathrm{var}(\hat{\mu}) = \frac{1}{\sum_{i=1}^n 1/\sigma_i^2} < \min(\sigma_1^2, \dots, \sigma_n^2)
$$

となることも示せ.

注意: この結果はことなる分散を持つ複数の推定量をうまくまとめることによって, 分散がより小さな推定量を作れることを意味している. これは同一の量に関する異なる研究の結果をまとめるメタアナリシスの出発点になる.

解答例: まず $\hat{\mu}$ の分散の式を整理しよう:

$$
\mathrm{var}(\hat{\mu}) = E\left[\left(\sum_{i=1}^n w_i X_i - \mu\right)^2\right] = E\left[\left(\sum_{i=1}^n w_i (X_i - \mu)\right)^2\right] = \sum_{i=1}^n w_i^2 \sigma_i^2.
$$

2つめの等号で $\sum_{i=1}^n w_i = 1$ であることを使い, 3つめの等号で $w_i X_i$ 達が無相関であることを使った(一般に無相関な確率変数の和の分散がそれぞれの分散の和になるのであった). $w_n = 1 - \sum_{i=1}^{n-1} w_i$ とおいて, $w_1, \dots, w_{n-1}$ を独立変数とみなすとき, それらによる $\mathrm{var}(\hat{\mu})$ の偏微分が 0 になるという方程式の解が求める w_i 達になる. $w_n = 1 - \sum_{i=1}^{n-1} w_i$ に注意すれば, $i = 1, \dots, n-1$ について

$$
0 = \frac{1}{2}\frac{\partial}{\partial w_i}\mathrm{var}(\hat{\mu}) = w_i \sigma_i^2 - \left(1 - \sum_{i=1}^{n-1} w_i\right)\sigma_i^2 = w_i \sigma_i^2 - w_n \sigma_i^2
$$

なので, この方程式の解について, $w_i \sigma_i$ 達は互いにすべて等しくなり, w_i 達の総和は 1 になる. そのとき, 以下のようになることがわかる:

$$
w_i = \frac{1/\sigma_i^2}{\sum_{i=1}^n 1/\sigma_i^2}, \quad \mathrm{var}(\hat{\mu}) = \sum_{i=1}^n w_i^2 \sigma_i^2 = \frac{1}{\sum_{i=1}^n 1/\sigma_i^2} < \frac{1}{1/\sigma_i^2} = \sigma_i^2.
$$

解答終

2.12 問題: 二項分布と負の二項分布の平均と分散のBernoulli分布と幾何分布の場合への帰着

Bernoulli分布 $\mathrm{Bernoulli}(p)$ の平均と分散がそれぞれ p, $p(1-p)$ であることと, 幾何分布 $\mathrm{Geometric}(p)$ の平均と分散がそれぞれ $(1-p)/p$, $(1-p)/p^2$ であることを認めて, 二項分布 $\mathrm{Binomial}(n, p)$ と負の二項分布 $\mathrm{NegativeBinomial}(k, p)$ の平均と分散を平易な計算で求めてみよ. 以下を示せ:

(1) $K \sim \mathrm{Binomial}(n, p)$ ならば $E[K] = np$, $\mathrm{var}(K) = np(1-p)$.

(2) $M \sim \mathrm{NegativeBinomial}(k, p)$ ならば $E[M] = k(1-p)/p$, $\mathrm{var}(M) = k(1-p)/p^2$.

解答例:

二項分布は試行回数 n の成功確率 p のBernoulli試行で生成された 1 と 0 からなる長さ n の数列中に含まれる 1 の個数の分布であった. Bernoulli試行の確率質量函数は

$$P(x_1, \dots, x_n) = P(x_1) \cdots P(x_n), \quad P(x_i) = p^{x_i}(1-p)^{1-x_i} \quad (x_i = 1, 0)$$

とBernoulli分布の確率質量函数 $P(x_i)$ の積で書けるのであった. この事実はBernoulli分布に従う確率変数 $X_1, \dots, X_n$ が独立であることを意味する. そして, Bernoulli試行で生成された 1 と 0 からなる長さ n の数列中に含まれる 1 の個数を意味する確率変数は $K = \sum_{i=1}^n X_i$ と書ける. このことから,

$$E[K] = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n p = np,$$
$$\mathrm{var}(K) = \sum_{i=1}^n \mathrm{var}(X_i) = \sum_{i=1}^n p(1-p) = np(1-p).$$

となることがわかる.

幾何分布は成功確率 p のBernoulli試行を 1 が1つ出るまで続けたときに出た 0 の個数の分布であった. $M_1, \dots, M_k$ はそれぞれが成功確率 p の幾何分布に従う独立な確率変数であるとする. このとき, $M = \sum_{i=1}^k M_i$ はBernoulli試行を 1 が k 回出るまで続けたときに 0 が出た個数に等しい. M_i は $i-1$ 番目の 1 から i 番目の 1 が出るまでに出た 0 の個数を意味する確率変数だと解釈される. このことは, M が負の二項分布 $\mathrm{NegativeBinomial}(k, p)$ に従う確率変数になることを意味する. このことから,

$$E[M] = \sum_{i=1}^k E[M_i] = \sum_{i=1}^k \frac{1-p}{p} = \frac{k(1-p)}{p},$$
$$\mathrm{var}(M) = \sum_{i=1}^k \mathrm{var}(M_i) = \sum_{i=1}^k \frac{1-p}{p^2} = \frac{k(1-p)}{p^2}.$$

となることがわかる.

解答終

注意: 計算が大幅に簡単になった!

2.13 問題: 番号が異なる確率変数達が無相関なときの確率変数の和の共分散

確率変数達 $X_1, Y_1, \dots, X_n, Y_n$ について次が成立していると仮定する:

$$\mathrm{cov}(X_i, Y_j) = \delta_{ij}\,\mathrm{cov}(X_i, Y_i).$$

このとき, 次が成立することを示せ:

$$\mathrm{cov}(X_1 + \cdots + X_n, Y_1 + \cdots + Y_n) = \mathrm{cov}(X_1, Y_1) + \cdots + \mathrm{cov}(X_n, Y_n).$$

ヒント: ベクトル達 $u_1, v_2, \dots, u_n, v_n$ の中の2つの異なる添え字を持つ u_i と v_j が互いに直交するならば, 内積 (,) について

$$(u_1 + \cdots + u_n, v_1 + \cdots + v_n) = (u_1, v_1) + \cdots + (u_n, v_n)$$

が成立することと本質的に同じことである.

解答例:

記号の簡単のため $A_i = X_i - E[X_i]$, $B_i = Y_i - E[Y_i]$ とおく. このとき, $E[A_i] = E[B_i] = 0$ より,

$$\mathrm{var}(A_i) = E[A_i^2], \quad \mathrm{var}(B_i) = E[B_i^2], \quad \mathrm{cov}(A_i, B_j) = E[A_i B_j].$$

A_i, B_i 達について上の問題を解けばよい. 問題の仮定より, A_i, B_i 達について次が成立している:

$$E[A_i B_j] = \delta_{ij} E[A_i B_i].$$

ゆえに,

$$E\left[\left(\sum_{i=1}^n A_i\right)\left(\sum_{j=1}^n B_j\right)\right] = \sum_{i,j=1}^n E[A_i B_j] = \sum_{i,j=1}^n \delta_{ij} E[A_i B_i] = \sum_{i=1}^n E[A_i B_i].$$

これは

$$\mathrm{cov}(A_1 + \cdots + A_n, B_1 + \cdots + B_n) = \mathrm{cov}(A_1, B_1) + \cdots + \mathrm{cov}(A_n, B_n).$$

の成立を意味する.

解答終

2.14 分散共分散行列とその半正定値性

n 個の確率変数を縦に並べてできるベクトル X と実数を成分に持つベクトル a を考える:

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}, \quad a = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \in \mathbb{R}^n.$$

このとき, X, a の転置をそれぞれ X, a^T と書くと,

$$a^T X = X^T a = \sum_{i=1}^{n} a_i X_i$$

なので

$$a^T X X^T a = \left(\sum_{i=1}^{n} a_i X_i \right)^2 \geq 0.$$

ゆえに, 期待値を取る操作の線形性と単調性より,

$$0 \leq E[a^T X X^T a] = a^T E[X X^T] a$$

でかつ, XX^T は $X_i X_j$ を (i, j) 成分とする $n \times n$ 行列になるので, $E[XX^T]$ は $E[X_i X_j]$ を (i, j) 成分とする $n \times n$ の実対称行列になる. 上の計算より, その実対称行列の固有値はすべて 0 以上になることがわかる.

ここで $E[X_i] = 0$ と仮定する. このとき, $E[XX^T]$ は i 番目の対角成分が X_i の分散 $\mathrm{var}(X_i)$ で, $i \neq j$ のときの (i, j) 成分が共分散 $\mathrm{cov}(X_i, X_j)$ であるような $n \times n$ の実対称行列になる.

このとき, 行列 $E[XX^T]$ を **分散共分散行列** と呼ぶ.

分散共分散行列の固有値はすべて 0 以上になる. この結果を分散共分散行列の **半正定値性** と呼ぶことにする. 特に分散共分散行列式は 0 以上になる.

3 標本(サンプル, データ)の平均と分散と共分散と相関係数

3.1 標本平均の定義

n 個の数値 $x_1, \ldots, x_n$ で構成されたデータ(標本, サンプル)に対して,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n}$$

をその **標本平均** (sample mean)と呼ぶ.

統計モデルを設定するとは現実世界で得られるデータ(標本, サンプル)の生じ方を確率分布の言葉でモデル化することである. そのとき, 統計モデル内部では, データ(標本, サンプル)を, n 個の数値ではなく, n 個の確率変数達 $X_1, \ldots, X_n$ でモデル化することになる. その場合には, 確率変数達の加法平均

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{X_1 + \cdots + X_n}{n}$$

をその **標本平均** (sample mean)と呼ぶ. X_i 達が確率変数であることより, それらの標本平均 $\bar{X}$ も確率変数になることに注意せよ. この文脈での標本平均 $\bar{X}$ は確率モデル内部における確率変数になる.

統計学入門の文脈では, 標本のモデル化とみなされるn 個の確率変数達 $X_1, \ldots, X_n$ は独立同分布だと仮定することが多い.

以下では, X_i 達は同じ期待値と分散を持ち, 無相関であると仮定する(この条件は独立同分布よりも弱い).

注意・警告: 標本のモデル化としての確率変数達の **標本平均と確率変数の平均(期待値)を明瞭に区別しなければいけない**. 各確率変数 X_i の期待値 $E[X_i]$ や確率変数としての標本平均 $\bar{X}$ の期待値

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu$$

と標本平均 $\bar{X}$ を明瞭に区別することが必要になる.

統計モデル内での標本のモデル化としての確率変数達 $X_1, \ldots, X_n$ の標本平均 $\bar{X}$ は確率変数なので, 数学的には函数とみなされ, 直観的にはランダムに値が決まる変数だとみなされるものになる. それに対して, 標本平均 $\bar{X}$ の期待値 $E[\bar{X}]$ は単なる数値になる. このように, 確率変数としての標本平均 $\bar{X}$ とその期待値 $E[\bar{X}]$ は全く異なる数学的対象になる.

注意: 上で期待値を取る操作 $E[\ ]$ の線形性を用いて, 標本平均の期待値が X_i 達の共通の期待値 μ に一致すること

$$E[\bar{X}] = \mu$$

も示されていることに注意せよ. これは標本平均を取る操作が平均の **不偏推定量** を作る操作になっていることも意味している. この点については後で詳しく説明する.

3.2 問題: 無相関な確率変数達の標本平均の分散

同じ期待値 μ と分散 σ^2 を持つ無相関な n 個の確率変数達 $X_1, \ldots, X_n$ の標本分散 $\bar{X}$ について次が成立していることを示せ:

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}$$

注意: この問題で扱っている確率変数としての $\bar{X}$ の分散と次の節で説明する標本のモデル化 $X_1, \ldots, X_n$ の分散は異なる数学的対象であることに注意せよ.

注意: $\mathrm{std}(\bar{X}) = \sigma/\sqrt{n}$ は **標準誤差** (standard error)と呼ばれ, SE と書かれることがある. 未知の真の標準誤差の推定量をも標準誤差と呼ぶことがあるが, このノートではそのスタイルを採用せずに, 未知の真の標準誤差の推定量を「標準誤差の推定量」と丁寧に呼ぶことにする.

解答例: 無相関な確率変数達の和の分散はそれぞれの確率変数の分散の和になるのであった. X_i 達が無相関であるという仮定より, X_i/n 達も無相関になる. そして,

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

であることより,

$$\mathrm{var}(\bar{X}) = \sum_{i=1}^{n} \mathrm{var}\left(\frac{X_i}{n}\right) = \sum_{i=1}^{n} \frac{\mathrm{var}(X_i)}{n^2} = \sum_{i=1}^{n} \frac{\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

解答終

3.3 標本分散と不偏分散の定義

n 個の数値 $x_1, \ldots, x_n$ で構成されたデータ(標本, サンプル)の標本平均を

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n}$$

と書くとき,

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

を **標本分散** (sample variance, 標本の無補正分散)と呼ぶ. **慣習的には** 補正無し標本分散が使用されることは少なく, 次のように定義される **不偏分散** (unbiased variance, 標本の不偏補正分散)を使うことが多い:

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

どうして n ではなく $n-1$ で割るかについては後で説明する. このとき,

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

とおくと,

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 = n\overline{x^2} - 2n\bar{x}^2 + n\bar{x}^2 = n\left(\overline{x^2} - \bar{x}^2\right)$$

なので,

$$s^2 = \frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right).$$

$X_1, \dots, X_n$ は標本のモデル化とみなされる n 個の確率変数達であるとする. X_i 達は同じ期待値と分散を持ち, 無相関であると仮定し, それらの標本平均を

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + \cdots + X_n}{n}$$

と書く. このとき, $X_1, \dots, X_n$ の **不偏分散** S^2 を次のように定義する:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

このとき,

$$\overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

とおくと,

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2 = n\overline{X^2} - 2n\bar{X}^2 + n\bar{X}^2 = n\left(\overline{X^2} - \bar{X}^2\right)$$

なので,

$$S^2 = \frac{n}{n-1}\left(\overline{X^2} - \bar{X}^2\right).$$

この場合には, 不偏分散 S^2 は確率変数になることに注意せよ. この S^2 と単なる数値になる $\bar{X}$ の確率変数としての分散 $\mathrm{var}(\bar{X})$ は異なる数学的対象になるので混乱しないようにして欲しい.

3.4 不偏推定量について: 不偏分散の定義ではどうして n ではなく $n-1$ で割るか

$X_1, \dots, X_n$ に共通の期待値と分散をそれぞれ μ, σ^2 と書くことにする. $X_1, \dots, X_n$ を現実で得られる標本のモデル化とみなす. そのとき, 統計モデル内での μ, σ^2 の値は未知の真の平均と分散の値であって欲しい. この設定では未知の μ と σ^2 を標本から推定する方法が欲しくなる.

未知の μ の推定量として優れているのは標本平均である. 標本のモデル化とみなされる $X_1, \dots, X_n$ の標本平均 $\bar{X}$ についてはその期待値 $E[\bar{X}]$ が未知の真の平均値 μ に等しくなる. この性質を標本平均の **不偏性** という. 未知の値 θ の推定量を $E[T] = \theta$ を満たす確率変数 T とするとき, T は θ の **不偏推定量** (unbiased estimator) と呼ぶ.

次に, 不偏分散の定義でどうして n ではなく $n-1$ で割るかを理解するために, $n-1$ で割る前の量の期待値を計算してみよう.

$X_1, \dots, X_n$ 達の標本平均 $\bar{X}$ の期待値と分散がそれぞれ次のようになることはわかっている:

$$E[\bar{X}] = \mu, \quad \mathrm{var}(\bar{X}) = E[(\bar{X} - \mu)^2] = \frac{\sigma^2}{n}.$$

そして, 以下が成立している:

$$
\begin{aligned}
\sum_{i=1}^{n}(X_i-\bar{X})^2 &= \sum_{i=1}^{n}((X_i-\mu)-(\bar{X}-\mu))^2 \\
&= \sum_{i=1}^{n}\left((X_i-\mu)^2-2(X_i-\mu)(\bar{X}-\mu)+(\bar{X}-\mu)^2)^2\right) \\
&= \sum_{i=1}^{n}(X_i-\mu)^2-2\left(\sum_{i=1}^{n}(X_i-\mu)\right)(\bar{X}-\mu)+n(\bar{X}-\mu)^2)^2 \\
&= \sum_{i=1}^{n}(X_i-\mu)^2-2n(\bar{X}-\mu)^2+n(\bar{X}-\mu)^2)^2 \\
&= \sum_{i=1}^{n}(X_i-\mu)^2-n(\bar{X}-\mu)^2.
\end{aligned}
$$

4つめの等号で $\sum_{i=1}^{n} X_i = n\bar{X}$ を用いた. ゆえに,

$$
E\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = \sum_{i=1}^{n}E[(X_i-\mu)^2]-nE[(\bar{X}-\mu)^2] = n\sigma^2-n\frac{\sigma^2}{n} = (n-1)\sigma^2.
$$

これより, σ^2 の不偏推定量を作るためには $\sum_{i=1}^{n}(X_i-\bar{X})^2$ を $n-1$ で割らなければいけないことがわかる. 実際にそのように定義した σ^2 の不偏推定量が不偏分散 S^2 である:

$$
S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2, \quad E[S^2]=\sigma^2.
$$

このように, $\sum_{i=1}^{n}(X_i-\bar{X})^2$ の確率変数としての分散を計算して $(n-1)\sigma^2$ になることを確認すれば, n ではなく $n-1$ で割る理由は明瞭である.

注意: 不偏分散の平方根 $\sqrt{S^2}$ は標準偏差 σ の不偏推定量になって **いない** ことに注意せよ. 推定量の不偏性は座標に依存する.

標準正規分布の場合に不偏分散と補正されていない標本分散の分布を同時プロットしてみよう.

In [34]:

```
n = 10
L = 10^5
X = rand(Normal(), n, L)
S² = var.(eachcol(X))
BV = (n-1)/n * S²
stephist(S²; norm=true, label="unbiased", c=1)
vline!([mean(S²)]; label="E[unbiased]", c=:blue)
stephist!(BV; norm=true, label="biased", c=2)
vline!([mean(BV)]; label="E[biased]", c=:red)
vline!([1]; label="true variance", c=:black, ls=:dash)
```

Out[34]:

確かに不偏分散(unbiased variance)の分布の平均値は真の値に一致しているが, 補正されていない標本分散(biased)の平均値は真の値よりも少し小さくなっている.

さらにこのグラフから, 不偏分散の場合も非補正標本分散の場合も確率密度函数は真の値よりも小さな値で最大になることにも注意せよ.

注意: 標準正規分布のサイズ n の標本分布において, 不偏分散の $n-1$ 倍は自由度 $n-1$ のχ²分布 $\mathrm{Chisq}(n-1)$ に従うことを示せる. 分布 $\mathrm{Chisq}(n-1)$ の期待値は $n-1$ なので, その $n-1$ 分の 1 の期待値は 1 になる.

注意: 標準誤差の推定量 として

$$\widehat{\mathrm{SE}} = \sqrt{\frac{S^2}{n}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

を採用しよう. このとき,

$$E\left[\widehat{\mathrm{SE}}^2\right] = \frac{E[S^2]}{n} = \frac{\sigma^2}{n} = \mathrm{var}(\bar{X}) = \mathrm{SE}^2$$

なので, $\widehat{\mathrm{SE}}^2$ は標準誤差の2乗 $\mathrm{SE}^2 = \mathrm{var}(\bar{X}) = \sigma^2/n$ の不偏推定量になっている. ただし, $\widehat{\mathrm{SE}}$ 自身は標準誤差 SE の不偏推定量になって **いない** ことに注意せよ. 推定量の不偏性は座標に依存する.

注意: 不偏推定量の方が不偏でない推定量より常に優れていると考えるのは誤りである. 推定量に不偏性の条件を課すと平均二乗誤差が大きくなってしまうことが多い(不偏分散についても実際にそうなることを後で示す). 不偏推定量を使うか否かはそうすることのメリットとデメリットの両方を考慮に入れて決定するべきである. ただし, 上で紹介した不偏分散については害は相当に小さく, 慣習的に非常によく使われている.

注意: $X_1, \ldots, X_n$ は標本のモデル化とみなされる独立同分布な確率変数達であるとし, X_i 達に共通の期待値と分散をそれぞれ μ, σ^2 と書くことにする. このとき, その標本平均 $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$ の期待値と分散がそれぞれ

$$E[\bar{X}] = \mu, \quad \mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}$$

となり, 不偏分散 $S^2 = (1/(n-1))\sum_{i=1}^{n}(X_i - \bar{X})^2$ の期待値が

$$E[S^2] = \sigma^2$$

となることまではすでにわかっている. この一連のノートではほぼ常に与えられた確率変数の期待値と分散を同時に計算して来た. ここでも, 確率変数としての不偏分散の分散も計算するべきである. そして, さらに標本平均と不偏分散の共分散も計算するべきである. しかし, そのためには X_i 達共通の分布の **歪度** (わいど, skewness)と **尖度** (せんど, kurtosis)の定義を先にしておいた方がよい. 次の節で歪度と尖度を定義する.

3.5 データの共分散の定義

n 個の数値の対 $(x_1, y_1), \ldots, (x_n, y_n)$ で構成されたデータ(標本, サンプル)について, x_i, y_i 達各々の標本平均を

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

と書くとき,

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

を **標本共分散** (sample covariance, 標本の無補正共分散)と呼ぶ. **慣習的には** 補正無しの標本共分散が使用されることは少なく, 次のように定義される **不偏共分散** (unbiased variance, 標本の不偏補正共分散)を使うことが多い:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

n ではなく $n-1$ で割る理由は不偏分散の場合と同様である. 後で詳しく説明する. このとき,

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

とおくと,

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i\,\bar{y} + n\bar{x}\bar{y} \\ &= n\overline{xy} - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} = n\left(\overline{xy} - \bar{x}\bar{y}\right)\end{aligned}$$

なので,

$$s_{xy} = \frac{n}{n-1}\left(\overline{xy} - n\bar{x}\bar{y}\right).$$

$(X_1, Y_1), \ldots, (X_n, Y_n)$ は上の型のデータのモデル化とみなされる n 個の確率変数対達であるとする. X_i 達は無相関で共通の同じ

期待値 μ_X と分散 σ_X^2 を持ち, Y_i 達も無相関で共通の同じ期待値 μ_Y と分散 σ_Y^2 を持ち, X_i と Y_i の共分散は i によらない共通の値 σ_{XY} になっているとし, 異なる番号 i, j について X_i と Y_j は無相関であると仮定する. これらの仮定は次のように式で書ける:

$$
\begin{aligned}
&E[X_i] = \mu_X, \quad E[Y_i] = \mu_Y \\
&E[(X_i - \mu_X)(X_j - \mu_X)] = \delta_{ij}\sigma_X^2, \\
&E[(Y_i - \mu_Y)(Y_j - \mu_Y)] = \delta_{ij}\sigma_Y^2, \\
&E[(X_i - \mu_X)(Y_j - \mu_Y)] = \delta_{ij}\sigma_{XY}.
\end{aligned}
$$

それらの X_i 達と Y_i 達それぞれの標本平均を次のように書くことにする:

$$
\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.
$$

このとき, $(X_1, Y_1), \ldots, (X_n, Y_n)$ の **不偏共分散** S_{XY} を次のように定義する:

$$
S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})
$$

このとき,

$$
\overline{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i
$$

とおくと,

$$
\begin{aligned}
\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_{i=1}^{n} X_i Y_i - \bar{X}\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_i\,\bar{Y} + n\bar{X}\bar{Y} \\
&= n\overline{XY} - n\bar{X}\bar{Y} - n\bar{X}\bar{Y} + n\bar{X}\bar{Y} = n\left(\overline{XY} - \bar{X}\bar{Y}\right)
\end{aligned}
$$

なので,

$$
s_{XY} = \frac{n}{n-1}\left(\overline{XY} - n\bar{X}\bar{Y}\right).
$$

この場合には, 不偏分散 S_{XY} は確率変数になることに注意せよ. この S_{XY} と単なる数値になる $\bar{X}, \bar{Y}$ の確率変数としての共分散 $\mathrm{cov}(\bar{X}, \bar{Y})$ は異なる数学的対象になるので混乱しないようにして欲しい.

3.6 問題: 標本平均達の共分散

前節の設定のもとで次が成立することを示せ:

$$
\mathrm{cov}(\bar{X}, \bar{Y}) = \frac{\sigma_{XY}}{n}.
$$

解答例: 標本平均の分散の計算とほぼ同じになる.

番号が異なる確率変数達が無相関なときの確率変数達の和の共分散は各々の対の共分散の和になるのであった. これを確率変数の和達

$$
\bar{X} = \sum_{i=1}^{n}\frac{X_i}{n}, \quad \bar{Y} = \sum_{i=1}^{n}\frac{Y_i}{n}
$$

に適用すると,

$$
\mathrm{cov}(\bar{X}, \bar{Y}) = \sum_{i=1}^{n}\mathrm{cov}\left(\frac{X_i}{n}, \frac{Y_i}{n}\right) = \sum_{i=1}^{n}\frac{\mathrm{cov}(X_i, Y_i)}{n^2} = \sum_{i=1}^{n}\frac{\sigma_{XY}}{n^2} = \frac{\sigma_{XY}}{n}.
$$

解答終

3.7 問題: 不偏共分散の定義ではどうして n ではなく $n-1$ で割るか

前々の設定のもとで, 不偏共分散 S_{XY} が X_i, Y_i の共分散 σ_{XY} の不偏推定量になっていることを示せ. すなわち次が成立することを示せ:

$$
E[S_{XY}] = \sigma_{XY}.
$$

注意: これが不偏共分散の定義で n ではなく $n-1$ で割る理由である.

解答例: 不偏分散の場合の議論とほぼ同じ.

$\bar{X}, \bar{Y}$ の期待値と共分散がそれぞれ次のようになることはわかっている:

$$E[\bar{X}] = \mu_X, \quad E[\bar{Y}] = \mu_Y, \quad \mathrm{cov}(\bar{X}, \bar{Y}) = E[(\bar{X} - \mu_Y)(\bar{Y} - \mu_Y)] = \frac{\sigma_{XY}}{n}.$$

そして, 以下が成立している:

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_{i=1}^{n}((X_i - \mu_X) - (\bar{X} - \mu_X))((Y_i - \mu_Y) - (\bar{Y} - \mu_Y)) \\
&= \sum_{i=1}^{n}\Big((X_i - \mu_X)(Y_i - \mu_Y) - (X_i - \mu_X)(\bar{Y} - \mu_Y) \\
&\qquad - (\bar{X} - \mu_X)(Y_i - \mu_Y) + (\bar{X} - \mu_X)(\bar{Y} - \mu_Y)\Big) \\
&= \sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y) - \left(\sum_{i=1}^{n}(X_i - \mu_X)\right)(\bar{Y} - \mu_Y) \\
&- (\bar{X} - \mu_X)\left(\sum_{i=1}^{n}(Y_i - \mu_Y)\right) + n(\bar{X} - \mu_X)(\bar{Y} - \mu_Y) \\
&= \sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y) - n(\bar{X} - \mu_X)(\bar{Y} - \mu_Y) \\
&- n(\bar{X} - \mu_X)(\bar{Y} - \mu_Y) + n(\bar{X} - \mu_X)(\bar{Y} - \mu_Y) \\
&= \sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y) - n(\bar{X} - \mu_X)(\bar{Y} - \mu_Y).
\end{aligned}$$

4つめの等号で $\sum_{i=1}^{n} X_i = n\bar{X}$, $\sum_{i=1}^{n} Y_i = n\bar{Y}$ を用いた. ゆえに,

$$\begin{aligned}
E\left[\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\right] &= \sum_{i=1}^{n} E[(X_i - \mu_X)(Y_i - \mu_Y)] - nE[(\bar{X} - \mu_X)(\bar{Y} - \mu_Y)] \\
&= n\sigma_{XY} - n\frac{\sigma_{XY}}{n} = (n-1)\sigma_{XY}.
\end{aligned}$$

ゆえに

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

より

$$E[S_{XY}] = \sigma_{XY}.$$

解答終

3.8 データの相関係数の定義 (以上の定義のまとめにもなっている)

データの **相関係数** (correlation coefficient of sample, Pearsonの相関係数)を定義しよう.

数値データの場合の共分散の定義の設定のもとで, x_i, y_i 達の標本平均, 不偏分散達と不偏共分散を以下のように書くことにする:

$$\begin{aligned}
&\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \\
&s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2, \\
&s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).
\end{aligned}$$

s_x^2, s_y^2 の平方根をそれぞれ s_x, s_y と書く. x_i 達と y_i 達の **相関係数** r_{xy} を次のように定める:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

この r_{xy} は数値になる.

データのモデル化である確率変数達の共分散の定義の設定のもとで, X_i, Y_i 達の標本平均, 不偏分散達と不偏共分散を以下のように書くことにする:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$
$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$
$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

S_X^2, S_Y^2 の平方根をそれぞれ S_X, S_Y と書く. X_i 達と Y_i 達の **相関係数** R_{XY} を次のように定める:

$$R_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

この R_{XY} は確率変数になる. (確率変数とは数学的には「その函数の期待値が定義されている変数」のことであり, 直観的には値がランダムに決まる変数だと思える.)

4 最小二乗法による線形回帰

最小二乗法による単純な線形回帰について説明しよう. 以下, $x_1, \ldots, x_n$ のうち2つは互いに異なると仮定する.

与えられたデータ $(x_1, y_1), \ldots, (x_n, y_n)$ における y_i 達の値達を x_i 達の値の共通の一次函数 $\alpha + \beta x_i$ (α, β は i によらない)で近似することを考える. そのとき,

$$\varepsilon_i = y_i - (\alpha + \beta x_i)$$

を残差と呼ぶ. このとき, **平均二乗残差**

$$f(\alpha, \beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2$$

を最小化することを **最小二乗法** と呼び, 最小二乗法で求めた係数 α, β の値を **回帰係数** と呼び, $\hat{\alpha}, \hat{\beta}$ と書き, それに対応する一次函数

$$y = \hat{\alpha} + \hat{\beta}$$

を **回帰直線** (regression line)と呼ぶ. このとき, x を **説明変数** や **独立変数** などと呼び, y を **目的変数**, **従属変数**, **反応変数** などと呼ぶ.

これは一次函数によるデータの近似になっているので, **線形回帰** (ordinary least squares, OLS, linear regression)と呼ばれる. この場合には説明変数が x の1つだけなので **単回帰** (simple regression, simple linear regression)と呼ぶこともある. 説明変数が多変数になった場合には **重回帰** (multiple regression)と呼ぶ.

以下ではさらに, **最小平均二乗残差** (もしくは単に **平均二乗残差**)の値を次のように書くことにする:

$$\hat{\sigma}^2 = f(\hat{\alpha}, \hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\hat{\alpha} + \hat{\beta} x_i))^2$$

注意: 最小二乗法は説明変数 x と目的変数 y について対称ではないので, それらの変数の立場を逆転して作った回帰直線は元の回帰直線に一致しない. 下の方で紹介するGalton (1886)の回帰直線の節も参照せよ.

4.1 問題: 最小二乗法による線形回帰の公式の導出

データ x_i の標本平均と不偏分散をそれぞれ $\bar{x}$, s_x^2 と書き, y_i の標本平均と不偏分散をそれぞれ $\bar{y}$, s_y^2 と書き, x_i と y_i の不偏共分散を s_{xy} と書くとき, 以上の $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2$ は次のように表されることを示せ:

$$\hat{\beta} = \frac{s_{xy}}{s_x^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad \hat{\sigma}^2 = \frac{n-1}{n}\frac{s_x^2 s_y^2 - s_{xy}^2}{s_x^2}.$$

4.1.1 解答例1: 単回帰の公式の素朴な導出

解答例1: 複雑な計算をシンプルに記述するために以下のようにおく:

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \quad \overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2, \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i.$$

このとき,

$$s_x^2 = \frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right), \quad s_y^2 = \frac{n}{n-1}\left(\overline{y^2} - \bar{y}^2\right), \quad s_{xy} = \frac{n}{n-1}\left(\overline{xy} - \bar{x}\bar{y}\right).$$

平均二乗残差 $f(\alpha, \beta)$ を最小化する $\alpha = \hat{\alpha}$, $\beta = \hat{\beta}$ は f の導函数を 0 する. ゆえに, $\hat{\alpha}, \hat{\beta}$ は次の方程式の解である:

$$\begin{aligned}
0 &= \frac{1}{2}\frac{\partial f}{\partial \alpha} = \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i - y_i) = \alpha + \bar{x}\beta - \bar{y}, \\
0 &= \frac{1}{2}\frac{\partial f}{\partial \beta} = \frac{1}{n}\sum_{i=1}^{n} x_i(\alpha + \beta x_i - y_i) = \bar{x}\alpha + \overline{x^2}\beta - \overline{xy}.
\end{aligned}$$

後者から前者に $\bar{x}$ をかけたものを引くと

$$\left(\overline{x^2} - \bar{x}^2\right)\beta = \overline{xy} - \bar{x}\bar{y}.$$

これの解は

$$\hat{\beta} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{s_{xy}}{s_x^2}.$$

上の方程式の前者より, 次によって $\hat{\alpha}$ も求めることができる:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

このとき

$$\begin{aligned}
\hat{\sigma}^2 &= f(\hat{\alpha}, \hat{\beta}) = f(\bar{y} - \hat{\beta}\bar{x}, \hat{\beta}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - (\bar{y} - \hat{\beta}\bar{x} + \beta x_i))^2 = \frac{1}{n}\sum_{i=1}^{n}((y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x}))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left((y_i - \bar{y})^2 - 2\hat{\beta}(x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}^2(x_i - \bar{x})^2\right) \\
&= \frac{n-1}{n}\left(s_y^2 - 2\hat{\beta}s_{xy} + \hat{\beta}^2 s_x^2\right) = \frac{n-1}{n}\left(s_y^2 - 2\frac{s_{xy}}{s_x^2}s_{xy} + \frac{s_{xy}^2}{(s_x^2)^2}s_x^2\right) \\
&= \frac{n-1}{n}\frac{s_x^2 s_y^2 - s_{xy}^2}{s_x^2}.
\end{aligned}$$

解答終

注意: 上の解答例ではわざと素朴な方法で計算してみた.

4.1.2 解答例2: データの5つの要約値だけで平均二乗残差を表示

解答例2: 平均二乗残差 $f(\alpha, \beta)$ の和の二乗の内側に $\bar{y} - (\alpha + \beta\bar{x})$ を引いて足して, 適当に整理して, 展開して, 整理すると以下のようになる:

$$\begin{aligned}
f(\alpha, \beta) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - (\alpha + \beta x_i) - (\bar{y} - (\alpha + \beta\bar{x})) + (\bar{y} - (\alpha + \beta\bar{x})))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}((y_i - \bar{y}) - \beta(x_i - \bar{x}) + (\bar{y} - (\alpha + \beta\bar{x})))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big((y_i - \bar{y})^2 + \beta^2(x_i - \bar{x})^2 + 2\beta(x_i - \bar{x})(y_i - \bar{y}) + (\bar{y} - (\alpha + \beta\bar{x}))^2 \\
&\qquad + 2\underbrace{(y_i - \bar{y})}_{\text{sum}=0}(\bar{y} - \alpha + \beta\bar{x})) - 2\beta\underbrace{(x_i - \bar{x})}_{\text{sum}=0}(\bar{y} - (\alpha + \beta\bar{x}))\Big) \\
&= \frac{n-1}{n}\left(s_y^2 + \beta^2 s_x^2 - 2\beta s_{xy}\right) + (\bar{y} - (\alpha + \beta\bar{x}))^2 \\
&= \frac{n-1}{n}\left(s_x^2\left(\beta - \frac{s_{xy}}{s_x^2}\right)^2 + \frac{s_x^2 s_y^2 - s_{xy}^2}{s_x^2}\right) + (\bar{y} - (\alpha + \beta\bar{x}))^2.
\end{aligned}$$

これより, 平均二乗残差 $f(\alpha, \beta)$ は, β, α がそれぞれ

$$\hat{\beta} = \frac{s_{xy}}{s_x^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

に等しいときに, 最小値

$$\hat{\sigma}^2 = \frac{n-1}{n}\frac{s_x^2 s_y^2 - s_{xy}^2}{s_x^2}$$

になることがわかる.

解答終

注意: 上の計算では最初に, 回帰係数 α, β の函数 $f(\alpha, \beta)$ (平均二乗残差)が $2n$ 個の数値で構成されているデータ $(x_1, y_1), \ldots, (x_n, y_n)$ を要約している標本平均と分散共分散の5つの数値 $\bar{x}, \bar{y}, s_x^2, s_y^2, s_{xy}^2$ だけで記述されることを示した. 特に **回帰直線と最小平均二乗残差の値はそれらの5つの要約値だけで決まってしまう.** データのそれら5つの値への要約が不適切な場合には最小二乗法を使うべきではない.

注意: 以上とはまったく別の **線形代数** における **直交射影** で最小二乗法を理解することもできる. 最小二乗法は数学的には直交射影の理論そのものだと言ってよい. この点に関する詳しい解説は別の機会に行いたい. 以上では最も素朴な最小二乗法の扱い方をした.

注意: 最小二乗法は **正規分布モデルの最尤法の一種** だとみなされる. これも非常に重要な視点である. この点についての解説も別の機会に行いたい. 最小二乗法を与える正規分布モデルの確率密度函数は

$$\begin{aligned} p(y_1, \ldots, y_n | x_1, \ldots, x_n, \alpha, \beta, \sigma) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2 \right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{n}{2\sigma^2} \left[\frac{n-1}{n} \left(s_x^2 \left(\beta - \frac{s_{xy}}{s_x^2} \right)^2 + \frac{s_x^2 s_y^2 - s_{xy}^2}{s_x^2} \right) + (\bar{y} - (\alpha + \beta\bar{x}))^2 \right] \right) \end{aligned}$$

であり, $\bar{x}, \bar{y}, s_x^2, s_y^2, s_{xy}^2$ はこの統計モデルの **十分統計量** になっている. この統計モデルは以下のようにも記述される:

$$y_i - (\alpha + \beta x_i) \sim \mathrm{Norma}(0, \sigma) \quad (i = 1, \ldots, n)$$

4.1.3 コンピュータによる最小二乗法の計算例

WolframAlpha (https://www.wolframalpha.com/)では次のように入力すれば単回帰を実行できる.

[実行 (https://www.wolframalpha.com/input?i=linear+fit+%2811%2C22%29%2C+%2812%2C20%29%2C+%2813%2C25%29%2C+%2814%2C+19%29%2C+%2815%2C+24%2
→ 7.35238 + 1.17143 x

```
linear fit (11,22), (12,20), (13,25), (14, 19), (15, 24), (16, 29)
```

不偏分散と不偏共分散も以下のように入力すれば計算できる.

[実行 (https://www.wolframalpha.com/input?i=variance+%2811%2C+12%2C+13%2C+14%2C+15%2C+16%29)] → 3.5

```
variance (11, 12, 13, 14, 15, 16)
```

[実行 (https://www.wolframalpha.com/input?i=covariance+%2811%2C+12%2C+13%2C+14%2C+15%2C+16%29%2C+%2822%2C+20%2C+25%2C+19%2C+24%2C+29%29)]
→ 4.1

```
covariance (11, 12, 13, 14, 15, 16), (22, 20, 25, 19, 24, 29)
```

他にも便利な道具があるので, 自分で色々試してみて研究しておくとよい.

以下のコードはJulia言語 (https://julialang.org/).

In [35]:
```
[ones(6) [11, 12, 13, 14, 15, 16]] \ [22, 20, 25, 19, 24, 29]
```

Out[35]:
```
2-element Vector{Float64}:
 7.352380952380989
 1.171428571428569
```

In [36]:
```
var([11, 12, 13, 14, 15, 16])
```
Out[36]: 3.5

In [37]:
```
cov([11, 12, 13, 14, 15, 16], [22, 20, 25, 19, 24, 29])
```
Out[37]: 4.1

不偏分散と不偏共分散については以下のように素朴に計算してもよい. (素朴な計算も一度はやってみるべきである.)

In [38]:
```
n = 6
```
Out[38]: 6

In [39]:
```
# xの標本平均
a = (11 + 12 + 13 + 14 + 15 + 16)/n
```
Out[39]: 13.5

In [40]:
```
# xの不偏分散
((11 - a)^2 + (12 - a)^2 + (13 - a)^2 + (14 - a)^2 + (15 - a)^2 + (16 - a)^2)/(n - 1)
```
Out[40]: 3.5

次の公式を使った方が入力が楽になる:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right).$$

In [41]:
```
(11^2 + 12^2 + 13^2 + 14^2 + 15^2 + 16^2 - n*a^2)/(n - 1) # こちらの方が入力が楽
```
Out[41]: 3.5

In [42]:
```
# yの標本平均
b = (22 + 20 + 25 + 19 + 24 + 29)/n
```
Out[42]: 23.166666666666668

標本分散についても次の公式を使って計算してみよう:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_u - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right).$$

In [43]:
```
# xとyの不偏共分散
(11*22 + 12*20 + 13*25 + 14*19 + 15*24 + 16*29 - n*a*b)/(n - 1)
```
Out[43]: 4.1

グラフもプロットしてみよう. さらに色々な計算方法も試してみよう.

In [44]:
```
# プロット用函数
function plot_ols(x, y, α̂, β̂; xlim=nothing,
        title="data and regression line", kwargs...)
    a, b = extrema(x)
    isnothing(xlim) && (xlim = (a - 0.1(b-a), b + 0.1(b-a)))
    scatter(x, y; label="", msw=0)
    plot!(x -> α̂ + β̂*x, xlim...; label="", lw=2)
    plot!(; xlabel="x", ylabel="y", msw=0)
    plot!(; size=(400, 300))
    plot!(; title, kwargs...)
end
```
Out[44]: plot_ols (generic function with 1 method)

In [45]:
```
# 「よろしくみなさん」に擬似乱数のシードを設定(テスト用)
Random.seed!(4649373)
```

Out[45]:
```
TaskLocalRNG()
```

In [46]:
```
# create test data
n = 100
α₀, β₀, σ₀ = 10, 2, 4
x = rand(Normal(10, 2), n)
noise = rand(Normal(0, σ₀), n)
y = @. α₀ + β₀*x + noise
@show α₀ β₀ σ₀
println()

# ordinary least squares
@show x̄ = sum(x[i] for i in 1:n)/n
@show ȳ = sum(y[i] for i in 1:n)/n
@show sx² = sum((x[i] - x̄)^2 for i in 1:n)/(n - 1)
@show sy² = sum((y[i] - ȳ)^2 for i in 1:n)/(n - 1)
@show sxy = sum((x[i] - x̄)*(y[i] - ȳ) for i in 1:n)/(n - 1)
println()
β̂ = sxy/sx²
α̂ = ȳ - β̂*x̄
σ̂ = √((n - 1)/n*(sx²*sy² - sxy^2)/sx²)
@show α̂ β̂ σ̂

plot_ols(x, y, α̂, β̂)
```

```
α₀ = 10
β₀ = 2
σ₀ = 4

x̄ = sum((x[i] for i = 1:n)) / n = 9.930093374143759
ȳ = sum((y[i] for i = 1:n)) / n = 29.76287823674706
sx² = sum(((x[i] - x̄) ^ 2 for i = 1:n)) / (n - 1) = 3.9902128888981885
sy² = sum(((y[i] - ȳ) ^ 2 for i = 1:n)) / (n - 1) = 37.32614282138912
sxy = sum(((x[i] - x̄) * (y[i] - ȳ) for i = 1:n)) / (n - 1) = 8.920847575776838

α̂ = 7.562346117161635
β̂ = 2.235682111246986
σ̂ = 4.148269982085684
```

Out[46]:

不偏分散と不偏共分散は次のようなコードでも計算できる.

In [47]:
```
@show sx² = var(x)
@show sy² = var(y)
@show sxy = cov(x, y);
```

```
sx² = var(x) = 3.990212888898188
sy² = var(y) = 37.3261428213891
sxy = cov(x, y) = 8.920847575776842
```

以下のように線形代数的な実装にすればシンプルなコードになる.

In [48]:

```
# ordinary least squares
X = [ones(n) x] # design matrix
α̂, β̂ = X \ y
σ̂ = norm(y - (α̂ .+ β̂*x))/√n
@show α̂ β̂ σ̂

plot_ols(x, y, α̂, β̂)
```

```
α̂ = 7.562346117161626
β̂ = 2.235682111246988
σ̂ = 4.148269982085681
```

Out[48]:

正規分布モデルの最尤法でも同じことをできる.

In [49]:

```
# negative log likelihood
function f(x, y, α, β, t)
    -sum(logpdf(Normal(0, exp(t)), y[i] - (α + β*x[i])) for i in eachindex(x))
end

# minimize negative log likelihood = maximize likelihood
o = optimize(((α, β, t),) -> f(x, y, α, β, t), zeros(3), LBFGS())
α̂, β̂, t̂ = o.minimizer
σ̂ = exp(t̂)
@show α̂ β̂ σ̂

plot_ols(x, y, α̂, β̂)
```

```
α̂ = 7.562346115104934
β̂ = 2.2356821114394743
σ̂ = 4.148269982255
```

Out[49]:

4.2 必修問題: 最小二乗法の計算例 (Anscombe's quartet)

以下に示した $(x_1, y_1), \ldots, (x_n, y_n)$ 型の4つのデータのそれぞれについて,

- x_i 達の標本平均 $\bar{x}$
- y_i 達の標本平均 $\bar{y}$
- x_i 達の不偏分散 s_x^2,
- y_i 達の不偏分散 s_y^2,
- x_i 達と y_i 達の不偏共分散 s_{xy},
- 回帰直線 $y = \hat{\alpha} + \hat{\beta}x$ の係数 $\hat{\alpha}$, $\hat{\beta}$ と最小平均二乗残差の平方根 $\hat{\sigma}$

を求めよ. それぞれについて小数点以下第1桁まで求めよ.

1. (10, 8.04), (8, 6.95), (13, 7.58), (9, 8.81), (11, 8.33), (14, 9.96), (6, 7.24), (4, 4.26), (12, 10.84), (7, 4.82), (5, 5.68)
2. (10, 9.14), (8, 8.14), (13, 8.74), (9, 8.77), (11, 9.26), (14, 8.1), (6, 6.13), (4, 3.1), (12, 9.13), (7, 7.26), (5, 4.74)
3. (10, 7.46), (8, 6.77), (13, 12.74), (9, 7.11), (11, 7.81), (14, 8.84), (6, 6.08), (4, 5.39), (12, 8.15), (7, 6.42), (5, 5.73)
4. (8, 6.58), (8, 5.76), (8, 7.71), (8, 8.84), (8, 8.47), (8, 7.04), (8, 5.25), (19, 12.5), (8, 5.56), (8, 7.91), (8, 6.89)

注意: 実はこれは有名なアンスコムの例 (https://ja.wikipedia.org/wiki/%E3%82%A2%E3%83%B3%E3%82%B9%E3%82%B3%E3%83%A0%E3%81%AE%E4%BE%8B)である. この例の教訓はデータの散布図をプロットすることの重要性である. 回帰直線を求める統計分析は2,3,4のデータについては適切ではない.

コピー＆ペースト用: Anscombe's quartet

データ1

- x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
- y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68

データ2

- x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
- y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.13, 7.26, 4.74

データ3

- x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
- y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73

データ4

- x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
- y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91, 6.89

解答例:

データ1:

$$\begin{aligned}
&\bar{x} \approx 9.0, \quad \bar{y} \approx 7.5, \\
&s_x^2 \approx 11.0, \quad s_y^2 \approx 4.1, \quad s_{xy} \approx 5.5 \\
&\hat{\alpha} \approx 3.0, \quad \hat{\beta} \approx 0.5, \quad \hat{\sigma} \approx 1.1
\end{aligned}$$

データ2:

$$\begin{aligned}
&\bar{x} \approx 9.0, \quad \bar{y} \approx 7.5, \\
&s_x^2 \approx 11.0, \quad s_y^2 \approx 4.1, \quad s_{xy} \approx 5.5 \\
&\hat{\alpha} \approx 3.0, \quad \hat{\beta} \approx 0.5, \quad \hat{\sigma} \approx 1.1
\end{aligned}$$

データ3:

$$\begin{aligned}
&\bar{x} \approx 9.0, \quad \bar{y} \approx 7.5, \\
&s_x^2 \approx 11.0, \quad s_y^2 \approx 4.1, \quad s_{xy} \approx 5.5 \\
&\hat{\alpha} \approx 3.0, \quad \hat{\beta} \approx 0.5, \quad \hat{\sigma} \approx 1.1
\end{aligned}$$

データ4:

$$\begin{aligned}
&\bar{x} \approx 9.0, \quad \bar{y} \approx 7.5, \\
&s_x^2 \approx 11.0, \quad s_y^2 \approx 4.1, \quad s_{xy} \approx 5.5 \\
&\hat{\alpha} \approx 3.0, \quad \hat{\beta} \approx 0.5, \quad \hat{\sigma} \approx 1.1
\end{aligned}$$

解答終

注意: 小数点以下第1桁までの数値は4つのデータについてすべて一致している! データの散布図を描くと, 回帰直線を描くことが適切なのは最初のデータだけで, 残りの3つのデータについては不適切なことがわかる. グラフは下の方にある.

4.2.1 WaolframAlphaでAnscombeの例1を扱う方法

データ1に関するWolframAlphaの使用例: xの標本平均 (https://www.wolframalpha.com/input?i=mean+10%2C+8%2C+13%2C+9%2C+11%2C+14%2C+6%2C+4%2C+12%2C+7%2C+5&lang=ja), yの標本平均 (https://www.wolframalpha.com/input?i=mean+8.04%2C+6.95%2C+7.58%2C+8.81%2C+8.33%2C+9.96%2C+7.24%2C+4.26%2C+10.84%2C+4.82%2C+5.68), xの不偏分散 (https://www.wolframalpha.com/input?i=var+10%2C+8%2C+13%2C+9%2C+11%2C+14%2C+6%2C+4%2C+12%2C+7%2C+5), yの不偏分散 (https://www.wolframalpha.com/input?i=var+8.04%2C+6.95%2C+7.58%2C+8.81%2C+8.33%2C+9.96%2C+7.24%2C+4.26%2C+10.84%2C+4.82%2C+5.68), x,yの不偏共分散 (https://www.wolframalpha.com/input?i=covar+%2810%2C+8%2C+13%2C+9%2C+11%2C+14%2C+6%2C+4%2C+12%2C+7%2C+5%29%2C+%288.04%2C+6.95%2C+7
回帰直線 (https://www.wolframalpha.com/input?i=linear+fit+%2810%2C+8.04%29%2C+%288%2C+6.95%29%2C+%2813%2C+7.58%29%2C+%289%2C+8.81%29%2C+%2811%2
最小平均二乗残差の平方根 (https://www.wolframalpha.com/input?i=sqrt%28%28n-1%29%2Fn*%28a*c+-+b%5E2%29%2Fa%29+where+n%3D11%2C+a%3D11%2C+b%3D5.5%2C+c%3D4.127)

4.2.2 Julia言語でAnscombeの例を扱う方法

RDatatsets.jl パッケージを入れればその中にAnscombeの例が入っている.

In [50]:
```julia
using RDatasets
anscombe = dataset("datasets", "anscombe")
```

Out[50]: 11 rows × 8 columns

	X1	X2	X3	X4	Y1	Y2	Y3	Y4
	Int64	Int64	Int64	Int64	Float64	Float64	Float64	Float64
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.1	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.1	5.39	12.5
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89

In [51]:
```julia
# コピー＆ペーストに利用するための表示
data1 = collect(zip(anscombe.X1, anscombe.Y1))
data2 = collect(zip(anscombe.X2, anscombe.Y2))
data3 = collect(zip(anscombe.X3, anscombe.Y3))
data4 = collect(zip(anscombe.X4, anscombe.Y4))
@show data1 data2 data3 data4;
```

```
data1 = [(10, 8.04), (8, 6.95), (13, 7.58), (9, 8.81), (11, 8.33), (14, 9.96), (6, 7.24), (4, 4.
26), (12, 10.84), (7, 4.82), (5, 5.68)]
data2 = [(10, 9.14), (8, 8.14), (13, 8.74), (9, 8.77), (11, 9.26), (14, 8.1), (6, 6.13), (4, 3.
1), (12, 9.13), (7, 7.26), (5, 4.74)]
data3 = [(10, 7.46), (8, 6.77), (13, 12.74), (9, 7.11), (11, 7.81), (14, 8.84), (6, 6.08), (4,
5.39), (12, 8.15), (7, 6.42), (5, 5.73)]
data4 = [(8, 6.58), (8, 5.76), (8, 7.71), (8, 8.84), (8, 8.47), (8, 7.04), (8, 5.25), (19, 12.
5), (8, 5.56), (8, 7.91), (8, 6.89)]
```

In [52]:

```julia
# コピー＆ペーストに利用するための表示
@show anscombe.X1 anscombe.Y1
@show anscombe.X2 anscombe.Y2
@show anscombe.X3 anscombe.Y3
@show anscombe.X4 anscombe.Y4;
```

```
anscombe.X1 = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
anscombe.Y1 = [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]
anscombe.X2 = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
anscombe.Y2 = [9.14, 8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.13, 7.26, 4.74]
anscombe.X3 = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
anscombe.Y3 = [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]
anscombe.X4 = [8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8]
anscombe.Y4 = [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91, 6.89]
```

In [53]:

```julia
function solve_anscombe(x, y;
        title="", xlim=(3, 20), ylim=(2, 14), size=(250, 225))
    @show x̄ = mean(x)
    @show ȳ = mean(y)
    @show sx² = var(x)
    @show sy² = var(y)
    @show sxy = cov(x, y)
    β̂ = sxy/sx²
    α̂ = ȳ - β̂*x̄
    σ̂ = √((n - 1)/n*(sx²*sy² - sxy^2)/sx²)
    @show α̂ β̂ σ̂

    plot_ols(x, y, α̂, β̂; title, xlim, ylim, size)
end
```

Out[53]:

```
solve_anscombe (generic function with 1 method)
```

In [54]:

```julia
PA1 = solve_anscombe(anscombe.X1, anscombe.Y1, title="Anscombe 1")
```

```
x̄ = mean(x) = 9.0
ȳ = mean(y) = 7.500909090909093
sx² = var(x) = 11.0
sy² = var(y) = 4.127269090909091
sxy = cov(x, y) = 5.501
α̂ = 3.0000909090909103
β̂ = 0.5000909090909091
σ̂ = 1.1672644558967773
```

Out[54]:

In [55]:
```
PA2 = solve_anscombe(anscombe.X2, anscombe.Y2, title="Anscombe 2")
```

```
x̄ = mean(x) = 9.0
ȳ = mean(y) = 7.500909090909091
sx² = var(x) = 11.0
sy² = var(y) = 4.127629090909091
sxy = cov(x, y) = 5.5
α̂ = 3.000909090909091
β̂ = 0.5
σ̂ = 1.167841085079644
```

Out[55]:

In [56]:
```
PA3 = solve_anscombe(anscombe.X3, anscombe.Y3, title="Anscombe 3")
```

```
x̄ = mean(x) = 9.0
ȳ = mean(y) = 7.500000000000001
sx² = var(x) = 11.0
sy² = var(y) = 4.12262
sxy = cov(x, y) = 5.497
α̂ = 3.0024545454545466
β̂ = 0.49972727272727274
σ̂ = 1.1669888559879227
```

Out[56]:

In [57]:
```
PA4 = solve_anscombe(anscombe.X4, anscombe.Y4, title="Anscombe 4")
```

```
x̄ = mean(x) = 9.0
ȳ = mean(y) = 7.50090909090909
sx² = var(x) = 11.0
sy² = var(y) = 4.12324909090909
sxy = cov(x, y) = 5.499
α̂ = 3.0017272727272726
β̂ = 0.49990909090909086
σ̂ = 1.1664075231238866
```

Out[57]:

In [58]:
```
plot(PA1, PA2, PA3, PA4; size=(500, 450), layout=(2,2))
```

Out[58]:

アンスコムの例では1以外のデータについて単純に最小二乗法で回帰直線を引くことは不適切である.

さらに, データを要約せずにその全体の様子を見るためにグラフを描いてみるとことが非常に大事なこともこの例からわかる.

4.3 Galton (1886)が描いた2つの回帰直線

下の方にある図は次の論文の5ページ目からの転載である:

- Galton, Francis. Regression towards mediocrity in hereditary stature. Journal of the Anthropological Institute, Vol. 5 (1886), 246-263. [link (https://galton.org/bib/JournalItem.aspx_action=view_id=157)]

この論文の著者の Francis Galton (https://en.wikipedia.org/wiki/Francis_Galton) (フランシス・ゴルトン (https://ja.wikipedia.org/wiki/%E3%83%95%E3%83%A9%E3%83%B3%E3%82%B7%E3%82%B9%E3%83%BB%E3%82%B4%E3%
は「平均への回帰の記述」や「相関係数の導入」で有名である.

このグラフにおける縦軸は親の身長の「平均値」であり, 横軸は成人した子供の身長である(単位はインチ). 親の身長の「平均値」の正確な意味は論文の2ページ目で以下のように説明されている:

> My data consisted of th e heights of 930 adult children and of their respective parentages, 205 in number. In every case I transmuted the female statures to their corresponding male equivalents and used them in their transmuted form, so that no objection grounded on the sexual difference of stature need be raised when I speak of averages. The factor I used was 1.08, which is equivalent to adding a little less than one-twelfth to each female height.

翻訳: 私のデータは930人の成人した子供とその親205人の身長である. どの場合でも, 女性の身長を対応する男性の身長に変換し, 変換された状態で使用した. それによって, 平均値について語るときに, 身長の性差に基づく異論を唱える必要を無くした. 使用した係数は1.08であり, これは各々の女性の身長にその12分の1弱を足すことに相当する.

上のグラフに描かれた楕円は, データの分布が多変量正規分布に近いことを表していると考えてよい. さらに, 親と子の身長のあいだに相関関係があることがわかる.

直線 ON が説明変数を親の身長, 目的変数を子の身長としたときの回帰直線であり, 直線 OM は2つの変数の立場を逆転させて作った回帰直線である. **2つの回帰直線は互いに一致しないし, 楕円の長軸とも一致しない.**

点 O はデータの平均値(数式で書くと $(\bar{x}, \bar{y})$ もしくは $(\bar{y}, \bar{x})$)であり, 回帰直線 ON が通る点 N は楕円と横軸に平行な直線が接する点になっており, 回帰直線 OM が通る点 M は楕円と縦軸に平行な直線が接する点になっている. (そうなる理由を自分で考えることは数学に関するちょっとした練習になるので, 興味がある人は自分で考えてみるとよい. 少し下の方に解答例がある.)

4.3.1 Galtonのグラフの再現

以下のセルのデータは上の論文の3ページ目にある Table I である. (数値の入力間違いが残っている可能性があるので, 原論文を参照して各自確認すること. 私自身は入力間違いはないと思っているが, それは私がそう思っているだけに過ぎない. このようなことはよくある.)

In [59]:

```
galton = [
    0  0  0  0  0  0  0  0  0  0  0  1  3  0
    0  0  0  0  0  0  0  1  2  1  2  7  2  4
    0  0  0  0  1  3  4  3  5 10  4  9  2  2
    1  0  1  0  1  1  3 12 18 14  7  4  3  3
    0  0  1 16  4 17 27 20 33 25 20 11  4  5
    1  0  7 11 16 25 31 34 48 21 18  4  3  0
    0  3  5 14 15 36 38 28 38 19 11  4  0  0
    0  3  3  5  2 17 17 14 13  4  0  0  0  0
    1  0  9  5  7 11 11  7  7  5  2  1  0  0
    1  1  4  4  1  5  5  0  2  0  0  0  0  0
    1  0  2  4  1  2  2  1  1  0  0  0  0  0
]
# 上の行列の縦軸は親の身長であり，行列の下ほど身長は小さくなる．
# 各種函数の仕様に合わせるために，行列の縦方向を逆転させて扱う．
galton = reverse(galton; dims=1)
# 行列の横軸が子の身長(単位はインチ)
galton_x = 61.2:74.2
# 行列の縦軸が親の身長(単位はインチ)
galton_y = 63.5:73.5

# データのヒートマップによる可視化
heatmap(galton_x, galton_y, galton; size=(320, 220))
```

Out[59]:

以下のセルでは楕円の長軸と短軸と2本の回帰直線をプロットしている．

In [60]:

```
# 適当にランダマイズしてデータを作成
Random.seed!(4649373)
data = hcat(([x-0.5+rand(), y-0.5+rand()]
        for (i, y) in enumerate(galton_y)
            for (j, x) in enumerate(galton_x)
                for _ in 1:galton[i,j])...)

# 多変量正規分布の最尤法
mvnormal = fit_mle(MvNormal, data)

# 楕円の長軸と短軸を求める
μ = mean(mvnormal) # 楕円の中心 [x̄, ȳ]
Σ = cov(mvnormal)  # 分散共分散行列 ((n-2)/n)[sx² sxy; sxy sy²]
d, U = eigen(Σ)    # U[:,j] = Σの固有値 d[j] の固有ベクトル
a = [-U[1,1]/U[2,1], -U[1,2]/U[2,2]] # 長軸・短軸の傾き
@show Σ μ a
g1(x) = a[1]*(x - μ[1]) + μ[2] # 長軸
g2(x) = a[2]*(x - μ[1]) + μ[2] # 短軸

# 最小二乗法による線形回帰
x, y = data[1,:], data[2,:]
X = [x.^0 x]
Y = [y.^0 y]
@show β = X\y # OLS(x, y)の回帰係数
@show γ = Y\x # OLS(y, x)の回帰係数
f1(x) = β[1] + β[2]*x # OLS(x, y)の回帰直線
f2(y) = γ[1] + γ[2]*y # OLS(y, x)の回帰直線

# グラフをプロット
widenlim((a, b)) = (c = 0.05(b-a); (a-c, b+c))
xlim = widenlim(extrema(galton_x))
ylim = widenlim(extrema(galton_y))
xs = range(xlim..., 1000)
ys = range(ylim..., 1000)

plot(legend=:topleft)
scatter!(x, y; label="data", msw=0, alpha=0.7)
contour!(xs, ys, (x,y) → pdf(mvnormal, [x,y]); colorbar=false, ls=:dot)
plot!(xs, g1.(xs); label="major axis", color=:blue)
plot!(xs, g2.(xs); label="minor axis", color=:blue, ls=:dash)
plot!(f2.(ys), ys; label="OLS(y, x)", color=:red)
plot!(xs, f1.(xs); label="OLS(x, y)", color=:red, ls=:dash)
plot!(; xlim, ylim, aspectratio=1, xtick=0:100, ytick=0:100)
plot!(; size=(560, 440))
title!("Ellipse axes and regression lines of Galton (1886)")
```

```
Σ = [6.496979569005935 2.0836915591542615; 2.0836915591542615 3.350202984520065]
μ = [68.09572980291252, 68.30018786422937]
a = [0.4979680162348105, -2.008161101512316]
β = X \ y = [46.460732175978194, 0.320716963478627]
γ = Y \ x = [25.61575900743164, 0.6219597943116172]
```

Out[60]:

注意・警告: 2つの回帰直線は楕円の長軸に一致しない. この点を誤解している人もいるようなので注意すること. 目分量で大雑把に回帰直線を手で引くときには, 楕円の長軸に線を引きたくなるかもしれないが, 誤りになる. 回帰直線は楕円と横軸または縦軸に平行な直線が接する点を通る直線になる.

4.3.2 グラフ中の楕円と2変量正規分布モデルや回帰直線の関係

上の計算での多変量正規分布の最尤法では, 次の確率密度函数で定まる2変量正規分布の標本分布を統計モデルとして採用している:

$$p(x, y|\mu, \Sigma) = \frac{1}{\sqrt{\det(2\pi\Sigma)}}\exp\left(-\frac{1}{2}\left[\,x-\mu_x,\ y-\mu_y\,\right]\Sigma^{-1}\begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix}\right).$$

ここで $\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \in \mathbb{R}^2$ で Σ は固有値が正の 2×2 の実対称行列である. 独立なパラメータは5つある. この統計モデルのデータ $(x_1, y_1), \ldots, (x_n, y_n)$ に関する最尤法はその5つのパラメータを次のように決定する:

$$\hat{\mu} = \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}, \quad \hat{\Sigma} = \frac{n-1}{n}\begin{bmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{bmatrix}.$$

単回帰の結果を決定したデータの5つの要約値 $\bar{x}, \bar{y}, s_x^2, s_{xy}, s_y^2$ が2変量正規分布モデルの最尤法の解も決定する.

$\hat{\Sigma}$ の式の右辺の $(n-1)/n$ は分散と共分散の推定値として不偏分散と不偏共分散を採用したので必要になった因子である. 分散と共分散の推定値として $n-1$ ではなく n で割った方を採用すればその因子はなくなる. このように正規分布モデルの最尤法の解の公式と不偏分散・不偏共分散は相性は悪い. 慣習的に使用されている不偏分散と不偏共分散にはこのようなトレードオフもある. (どのような場合であっても都合のよい流儀は存在しない.)

以上の統計モデルと楕円の関係について説明しよう.

分散共分散行列 Σ を $\Sigma = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ と書くと,

$$\Sigma^{-1} = \frac{1}{ac-b^2}\begin{bmatrix} c & -b \\ -b & a \end{bmatrix}$$

なので

$$\left[\,x-\mu_x,\ y-\mu_y\,\right]\Sigma^{-1}\begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix} = \frac{c(x-\mu_x)^2 - 2b(x-\mu_x)(y-\mu_y) +}{ac-b^2}.$$

前節で描いたグラフ中の楕円はこれが一定の値になるという条件で定義された図形である.

以下, 計算を簡単に記述するために $\mu_x = \mu_y = 0$ と仮定する. このとき, その楕円の方程式は

$$ay^2 - 2bxy + cx^2 = \text{constant}.$$

になる. この条件を保ったまま (x, y) を微小に動かすと,

$$2ay\,dy - 2bx\,dy - 2by\,dx + 2cx\,dx = 0$$

すなわち

$$(ay - bx)\,dy - (by - cx)\,dx = 0$$

が得られる(これは点 (x, y) を通る楕円の接線の方程式だとみなされる).

点 (x, y) が楕円と y 軸と平行な直線の接点のとき, そのような接点 (x, y) を楕円上に乗せたままで微小に動かすと x 座標は高次の微小量しか変化しないので $dx = 0$ が得られ, $y = (b/a)x$ が得られる. これは, $\Sigma = \hat{\Sigma}$ の場合には $a = ((n-1)/n)s_x^2$, $b = ((n-1)/n)s_{xy}$ なので $y = (s_{xy}/s_x^2)x$ を意味する. その式で定義される直線の傾き s_{xy}/s_x^2 は説明変数が x で目的変数が y のときの単回帰直線の傾き $\bar{\beta}$ に一致している.

これで, 説明変数が x で目的変数が y のときの単回帰直線が楕円と y 軸に平行な直線の接点を通ることも証明された.

注意: このことから, 楕円の内側を通る y 軸に平行な直線 ℓ について, 説明変数が x で目的変数が y のときの単回帰直線と直線 ℓ の交点と楕円と直線 ℓ の2つの交点の距離が等しくなることもわかる. なぜならば, 変換 $(x, y) \mapsto (x, y - (b/a)x)$ によって, 回帰直線 $y = (b/a)x$ が楕円と y 軸に平行な直線の接点を通るという条件も保たれるので($\mu_x = \mu_y = 0$ と仮定している), その変換でうつした先の楕円の軸の片方は x 軸に一致し, うつした先の楕円が x 軸について線対称になることがわかるからである. 下の方で描いたグラフを見よ.

注意: 以上の数学的議論を追い切れない人は前節で描いたグラフやGolton氏自身が描いたグラフをよく見て, 単回帰直線が楕円と横軸(もしくは縦軸)と平行な直線の接点を通っていることを目で確認せよ. 数学的な議論が追い切れない場合であっても, 結論をグラフを見て確認できるならば, 記憶に残って役に立つ知識になることが多い.

In [61]:

```
f(x, y, h) = 6y^2 - 4x*y + 3x^2 - h
g(x, h, s) = (2x + s*√(4x^2 - 6*(3x^2 - h)))/6
h = f(1.5, 0.5, 0)
x = range(-2, 2, 200)
y = range(-1.5, 1.5, 200)
plot(; aspectratio=1, size=(400, 300))
plot!(x, x -> x/3; label="regression line", legend=:topleft)
vline!([-1.5, 1.5]; label="")
k = -0.4
plot!([k, k], [g(k, h, -1), g(k, h, 1)]; label="", c=3)
plot!([k-0.05, k+0.05], fill((k/3 + g(k, h, -1))/2, 2); label="", c=3)
plot!([k-0.05, k+0.05], fill((k/3 + g(k, h,  1))/2, 2); label="", c=3)
k = 0.8
plot!([k, k], [g(k, h, -1), g(k, h, 1)]; label="", c=4)
plot!([k-0.05, k+0.05], fill((k/3 + g(k, h, -1))/2 + 0.02, 2); label="", c=4)
plot!([k-0.05, k+0.05], fill((k/3 + g(k, h, -1))/2 - 0.02, 2); label="", c=4)
plot!([k-0.05, k+0.05], fill((k/3 + g(k, h,  1))/2 + 0.02, 2); label="", c=4)
plot!([k-0.05, k+0.05], fill((k/3 + g(k, h,  1))/2 - 0.02, 2); label="", c=4)
contour!(x, y, (x, y) -> f(x, y, h); colorbar=false, levels=[0])
plot!(; xlabel="x", ylabel="y")
```

Out[61]:

確かに, 楕円と縦軸に平行な直線の接点を通る直線(=回帰直線)が楕円を上下の長さについて二等分していることがこの図を見ればわかる.

5 モーメントとキュムラントと歪度と尖度

5.1 モーメントとその母函数と特性函数とキュムラント母函数の定義

確率変数 X と $m = 0, 1, 2, \ldots$ について

$$\mu_m(X) = E[X^m]$$

を X の m 次の **モーメント**(moment, 積率) と呼び,

$$M_X(t) = E[e^{tX}] = E\left[\sum_{m=0}^{\infty} X^m \frac{t^m}{m!}\right] = \sum_{m=0}^{\infty} E[X^m]\frac{t^m}{m!} = \sum_{m=0}^{\infty} \mu_m(X)\frac{t^m}{m!}$$

を **モーメント母函数** (moment generating function, mgf)と呼ぶ.

X が従う確率分布の名前が Dist のとき, これらを **分布** Dist **のモーメントとモーメント母函数** と呼ぶ. 以下も同様である.

モーメント母函数の定義で t を $it = \sqrt{-1}\,t$ で置き換えたもの

$$\begin{aligned}\varphi_X(t) = E[e^{itX}] &= E\left[\sum_{m=0}^{\infty} i^m X^m \frac{t^m}{m!}\right] \\ &= \sum_{m=0}^{\infty} i^m E[X^m]\frac{t^m}{m!} = \sum_{m=0}^{\infty} i^m \mu_m(X)\frac{t^m}{m!}\end{aligned}$$

を **特性函数** (characteristic function)と呼ぶ. 特性函数を扱う場合には $i = \sqrt{-1}$ としたいので, i を番号の意味で使わないように気を付ける必要がある.

モーメント母函数だけではなく, 特性函数もモーメント達の母函数になっている.

モーメント母函数の対数

$$K_X(t) = \log M(t) = \log E[e^{tX}] = \sum_{m=1}^{\infty} \kappa_m(X)\frac{t^m}{m!} = \sum_{m=1}^{\infty} \kappa_m \frac{t^m}{m!}$$

を **キュムラント母函数** (cumulant generating function, cgf) と呼び, その展開係数 $\kappa_m = \kappa_m(X)$ を X の m 次のキュムラントと呼ぶ.

注意: 取り得る値が実数になる確率変数 X について $|e^{itX}| = 1$ となるので, $E[e^{itX}]$ は常に絶対収束しており, 特性函数は常にうまく定義されている. それに対して e^{tX} の値は巨大になる可能性があり, $E[e^{tX}]$ が収束しない場合が出て来る. モーメント母函数やキュムラント母函数の取り扱いではこの点に注意する必要がある.

注意: モーメント母函数とキュムラント母函数はそれぞれ物理での統計力学における **分配函数** と **自由エネルギー** (もしくは **Massieu函数**)の確率論的な類似物になっている. ただし, 逆温度 β について $t = -\beta$ とおく必要がある. 逆に言えば, モーメント母函数とキュムラントの表示における $\beta = -t$ の逆数は絶対温度の確率論的類似物になっていることになる.

5.2 特性函数による期待値の表示

X が確率密度函数 $p(x)$ を持つとき, 函数 $f(x)$ のFourier変換を

$$\hat{f}(t) = \int_{-\infty}^{\infty} f(x)e^{-itx}\,dx$$

と書くと, $f(t)$ がそう悪くない函数ならば逆Fourier変換によってもとの函数に戻せる:

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(t)e^{itx}\,dt.$$

Fourier解析の基礎については次のリンク先を参照せよ(逆Fourier変換に関する結果はこのノート内では認めて使ってよい):

- 12 Fourier解析 (https://nbviewer.org/github/genkuroki/Calculus/blob/master/12%20Fourier%20analysis.ipynb)

ゆえに, x に確率変数 X を代入して両辺の期待値を取り, 期待値を取る操作と積分を交換すると,

$$E[f(X)] = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(t)E[e^{itX}]\,dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(t)\varphi_X(t)\,dt.$$

ここで $\varphi_X(t) = E[e^{itX}]$ は X の特性函数である.

確率変数 X が従う分布は様々な函数 $f(x)$ に関する期待値 $E[f(X)]$ から決まるので, $E[f(X)]$ が X の特性函数 $\varphi_X(t)$ を用いて表せたということは, X **の特性函数** $\varphi_X(t)$ **から** X **が従う分布が唯一つに決まる** ことを意味している.

さらに, 分布Aの特性函数が分布Bの特性函数で近似されていれば, 分布Aにおける期待値が分布Bにおける期待値で近似されることもわかる. これは **分布の近似を特性函数の近似で確認できる** ことを意味する.

モーメント母函数 $M_X(t) = E[e^{tX}]$ が t の函数として, 定義域を自然に複素数まで拡張できているとき(正確には解析接続できていれば), $\varphi_X(t) = M_X(it)$ が成立する. キュムラント母函数はモーメント母函数の対数である. これらより, **モーメント母函数やキュムラント母函数からも分布が唯一つに決まる** ことがわかる.

5.3 問題: 分布のアフィン変換のキュムラント

確率変数 X と $a, b \in \mathbb{R}$, $a \neq 0$ について, $aX + b$ のモーメント母函数とキュムラント母函数とキュムラントが X のそれらで次のように表されることを示せ:

$$\begin{aligned} &M_{aX+b}(t) = e^{bt} M_X(at), \quad K_{aX+b}(t) = K_X(at) + bt, \\ &\kappa_1(aX+b) = a\kappa_1(X) + b, \quad \kappa_m(aX+b) = a^m \kappa_m(X) \quad (m = 2, 3, 4, \ldots) \end{aligned}$$

注意: キュムラントの変換公式は非常に単純な形になる. $\kappa_1(aX+b) = a\kappa_1(X) + b$ は $\kappa_1(X) = E[X]$ だったので当然である. 2次以上のキュムラントは a^m 倍されるだけになる. モーメント母函数の対数を取ってキュムラント母函数を定義し, その展開によってキュムラントを定義することにはこのような利点がある. キュムラント母函数を定義することには, 物理の統計力学で分配函数の対数を取って自由エネルギーを定義することと同様の利点がある.

注意: この結果は空気のごとく使われる.

解答例: $aX + b$ のモーメント母函数を X のモーメント母函数であらわそう:

$$M_{aX+b}(t) = E[e^{t(aX+b)}] = e^{bt} E[e^{atX}] = e^{bt} M_X(at).$$

ゆえに, $aX + b$ のキュムラントは次の形になる:

$$K_{aX+b}(t) = \log M_{aX+b}(t) = \log(e^{bt} M_X(at)) = K_X(at) + bt.$$

$X \mapsto aX + b$ によってキュムラント母函数は $K_X(t) \mapsto K_X(at) + bt$ に似た形式で変換される.

X のキュムラント $\kappa_m(X)$ は次のようにキュムラント母函数を展開することによって定義されるのであった:

$$K_X(t) = \sum_{m=1}^{\infty} \kappa_m(X) \frac{t^m}{m!}.$$

$K_{aX+b}(t)$ の展開結果は

$$\begin{aligned} K_{aX+b}(t) = K_X(at) + bt &= \sum_{m=0}^{\infty} \kappa_m(X) \frac{(at)^m}{m!} + bt \\ &= (a\kappa_1(X) + b)t + \sum_{m=2}^{\infty} a^m \kappa_m(X) \frac{t^m}{m!} \end{aligned}$$

になるので,

$$\kappa_1(aX+b) = a\kappa_1(X) + b, \quad \kappa_m(aX+b) = a^m \kappa_m(X) \quad (m = 2, 3, 4, \ldots)$$

となることがわかる.

解答終

5.4 問題: 標準正規分布のモーメント母函数と特性函数とキュムラント母函数

標準正規分布に従う確率変数 Z のモーメント母函数と特性函数とキュムラント母函数が次のようになることを示せ:

$$M_Z(t) = e^{t^2/2}, \quad \varphi_Z(t) = M_Z(it) = e^{-t^2/2}, \quad K_Z(t) = \log M_Z(t) = \frac{t^2}{2}.$$

この結果は中心極限定理の証明で使われる.

解答例: $Z \sim \mathrm{Normal}(0, 1)$ と仮定する. このとき,

$$tz - \frac{z^2}{2} = -\frac{z^2 - 2tz}{2} = -\frac{(z-t)^2 - t^2}{2} = -\frac{(z-t)^2}{2} + \frac{t^2}{2}$$

なので,

$$
\begin{aligned}
M_Z(t) = E[e^{tZ}] &= \int_{-\infty}^{\infty} e^{tz} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(z-t)^2/2+t^2/2}\, dz \\
&= \frac{e^{t^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-w^2/2}\, dw = e^{t^2/2}.
\end{aligned}
$$

4つめの等号で $z = w + t$ とおいた. ゆえに,

$$
\varphi_Z(t) = E[e^{itZ}] = M_Z(it) = e^{-t^2/2}, \quad K_Z(t) = \log M_Z(t) = \frac{t^2}{2}.
$$

解答終

注意: 標準正規分布のキュムラント母函数は $t^2/2$ というたったの一項だけになってしまう. 標準でない正規分布のキュムラント母函数は t について1次と2次の項だけになる. 1次の項の係数は分布の期待値で, $t^2/2$ の係数は分散になる. 実際, 平均 μ, 分散 σ^2 を持つ確率変数 X について,

$$
\begin{aligned}
K_X(t) &= \log E[e^{tX}] = \log E[e^{t(X-\mu)+t\mu}] \\
&= t\mu + \log\left(1 + E[X-\mu]t + E[(X-\mu)^2]\frac{t^2}{2} + O(t^3)\right) \\
&= t\mu + \log\left(1 + \sigma^2\frac{t^2}{2} + O(t^3)\right) = t\mu + \sigma^2\frac{t^2}{2} + O(t^3).
\end{aligned}
$$

そして, $\sigma Z + \mu \sim \mathrm{Normal}(\mu, \sigma)$ であり,

$$
M_{\sigma Z+\mu}(t) = E[e^{t(\sigma Z+\mu)}] = e^{\mu t} E[e^{t\sigma Z}] = e^{\mu t} M_Z(\sigma t) = e^{\mu t + \sigma^2 t^2/2}.
$$

ゆえに

$$
K_{\sigma Z+\mu}(t) = \log M_{\sigma Z+\mu}(t) = \mu t + \sigma^2\frac{t^2}{2}.
$$

他の分布のキュムラント母函数を計算したときに出て来る t について3次以上の項はその分布が正規分布からどれだけ離れているかを表している.

5.5 確率変数の標準化と標準化キュムラントと歪度と尖度

確率変数 X は確率変数であるとし, $\mu = E[X]$, $\sigma = \sqrt{E[(X-\mu)^2]}$ とおく. このとき,

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

を確率変数の **標準化** (standardization)と呼ぶ. Z の期待値と分散はそれぞれ 0 と 1 になる.

X の標準化のモーメントやキュムラントおよびそれらの母函数をそれぞれ **標準化モーメント**, **標準化キュムラント** と呼び, それぞれを $\bar{\mu}_m(X)$, $\bar{\kappa}_m(X)$ と表す. 詳しくは以下の通り:

$$
\begin{aligned}
&\bar{\mu}_m(X) = \mu_m(Z) = E\left[\left(\frac{X-\mu}{\sigma}\right)^m\right], \\
&M_Z(t) = E\left[\exp\left(t\frac{X-\mu}{\sigma}\right)\right] = \sum_{m=0}^{\infty} \bar{\mu}_m(X)\frac{t^m}{m!} = 1 + \frac{t^2}{2} + \bar{\mu}_3(X)\frac{t^3}{3!} + \bar{\mu}_4(X)\frac{t^4}{4!} + \cdots, \\
&K_Z(t) = \log M_Z(t) = \sum_{m=1}^{\infty} \bar{\kappa}_m(X)\frac{t^m}{m!} = \frac{t^2}{2} + \bar{\kappa}_3(X)\frac{t^3}{3!} + \bar{\kappa}_4(X)\frac{t^4}{4!} + \cdots.
\end{aligned}
$$

$\bar{\kappa}_3(X)$ と $\bar{\kappa}_4(X)$ は次のように表される:

$$
\bar{\kappa}_3(X) = \bar{\mu}_3(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right], \quad \bar{\kappa}_4(X) = \bar{\mu}_4(X) - 3 = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3.
$$

このことは, $\log(1+a) = a - a^2/2 + O(a^3)$ を使って以下のようにして確認される:

$$
\begin{aligned}
\log\left(1 + \frac{t^2}{2} + \bar{\mu}_3(X)\frac{t^3}{3!} + \bar{\mu}_4(X)\frac{t^4}{4!} + O(t^5)\right) &= \frac{t^2}{2} + \bar{\mu}_3(X)\frac{t^3}{3!} + \bar{\mu}_4(X)\frac{t^4}{4!} - \frac{1}{2}\left(\frac{t^2}{2}\right)^2 + O(t^5) \\
&= \frac{t^2}{2} + \bar{\mu}_3(X)\frac{t^3}{3!} + (\bar{\mu}_4(X) - 3)\frac{t^4}{4!} + O(t^5).
\end{aligned}
$$

$\bar{\kappa}_3(X)$ を X もしくは X が従う分布の **歪度** (わいど, skewness) と呼び, $\bar{\kappa}_4(X)$ を **尖度** (せんど, kurtosis)と呼び, 次のようにも書くことにする:

$$\mathrm{skewness}(X) = \bar{\kappa}_3(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right], \quad \mathrm{kurtosis}(X) = \bar{\kappa}_4(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3.$$

歪度は左右の非対称性の尺度であり, 尖度は分布の尖り具合が正規分布とどれだけ違うかの尺度になっている.

X が正規分布に従う確率変数の場合にはその標準化 $Z = (X-\mu)/\sigma$ は標準正規分布に従う確率変数になるので, その標準化キュムラント達は $\bar{\kappa}_2(X) = 1$, $\bar{\kappa}_m(X) = 0$ $(m \neq 0)$ となる. 2次の標準化キュムラントは常に 1 になるが, 3 次以上の標準化キュムラントは X が正規分布でなければ 0 でなくなる.

このことから, 3 次以上の標準化キュムラントは分布が正規分布からどれだけ離れているかを表していると考えられる. それらのうち最初の2つが上で定義した歪度 $\bar{\kappa}_3(X)$ と尖度 $\bar{\kappa}_4(X)$ になっている.

分布の歪度 $\bar{\kappa}_3(X)$ と尖度 $\bar{\kappa}_4(X)$ は分布がどれだけ正規分布から離れているかを表す最も基本的な量である.

注意: $\bar{\kappa}_4(X) = \bar{\mu}_4(X) - 3$ ではなく, 3 を引く前の $\bar{\mu}_4(X)$ を尖度と定義する流儀もあるが, このノートでは正規分布の扱いでキュムラントが非常に便利なことを重視したいので, 3 を引いた側の $\bar{\kappa}_4$ を尖度の定義として採用する. 3 を引いた方の $\bar{\kappa}_4(X) = \bar{\mu}_4(X) - 3$ は **過剰尖度** (かじょうせんど, **excess kurtosis**)と呼ばれることも多い. 正規分布の尖度を($\bar{\kappa}_4$ の方の 0 ではなく) $\bar{\mu}_4$ の方の 3 とするときに, そこからどれだけ分布の尖り具合が増したかを $\bar{\kappa}(X)$ が表していることを「過剰」と表現している.

In [62]:
```
@vars t μ3 μ4 μ5 κ3 κ4 κ5
Mt = 1 + t^2/2 + μ3*t^3/6 + μ4*t^4/24 + μ5*t^5
expr = series(log(Mt), t)
```

Out[62]: $\frac{t^2}{2} + \frac{t^3\mu_3}{6} + t^4\left(\frac{\mu_4}{24} - \frac{1}{8}\right) + t^5\left(-\frac{\mu_3}{12} + \mu_5\right) + O\left(t^6\right)$

5.6 問題: 正規分布の歪度と尖度

X が正規分布に従う確率変数であるとき, その歪度 $\bar{\kappa}_3(X)$ と尖度 $\bar{\kappa}_4(X)$ が以下のようになることを示せ:

$$\bar{\kappa}_3(X) = \bar{\kappa}_4(X) = 0.$$

解答例: X が正規分布 $\mathrm{Normal}(\mu, \sigma)$ に従う確率変数のとき, その標準化 $Z = (X-\mu)/\sigma$ は標準正規分布に従うので, $Z = (X-\mu)/\sigma$ のキュムラント母函数は $t^2/2$ になり, その 3 次以上の項がすべて 0 になる. ゆえに $\bar{\kappa}_m(X) = 0$ ($m = 3, 4, 5, \ldots$). 特に $\bar{\kappa}_3(X) = \bar{\kappa}_4(X) = 0$.

解答終

5.7 問題: 一様分布のキュムラント母函数と歪度と尖度

一様分布 $\mathrm{Uniform}(0, 1)$ に従う確率変数 X について, そのキュムラント母函数 $K_X(t) = \log E[e^{tX}]$ と歪度 $\bar{\kappa}_3(X)$ と尖度 $\bar{\kappa}_4(X)$ が以下のようになることを示せ:

$$K_X(t) = \log\frac{e^t - 1}{t}, \quad \bar{\kappa}_3(X) = 0, \quad \bar{\kappa}_4(X) = -\frac{6}{5}.$$

一様分布は歪度が 0 でかつ尖度がかなり小さな連続分布の例になっている.

解答例: モーメント母函数とキュムラント母函数は以下のように計算される:

$$E[e^{tX}] = \int_0^1 e^{tx}\,dx = \left[\frac{e^{tx}}{t}\right]_0^1 = \frac{e^t - 1}{t},$$
$$K_X(t) = \log\frac{e^t - 1}{t} = \frac{1}{2}t + \frac{1}{12}\frac{t^2}{2} - \frac{1}{120}\frac{t^4}{4!} + O(t^5).$$

ゆえに, $\mu = 1/2$, $\sigma = \sqrt{1/12}$, $Z = (X-\mu)/\sigma$ とおくと,

$$K_Z(t) = K_X\left(\frac{t}{\sigma}\right) - \frac{\mu}{\sigma}t = \frac{t^2}{2} - \frac{6}{5}\frac{t^4}{4!} + O(t^5).$$

ゆえに $\bar{\kappa}_3(X) = 0$, $\bar{\kappa}_4 = -6/5$.

解答終

In [63]:
```
@vars t
K_X = log((exp(t) - 1)/t)
κ = [limit(diff(K_X, t, m) , t, 0) for m in 1:4]
```

Out[63]:
$$\begin{bmatrix} \frac{1}{2} \\ \frac{1}{12} \\ 0 \\ -\frac{1}{120} \end{bmatrix}$$

In [64]:
```
μ, σ = κ[1], √κ[2]
κ = [limit(diff(K_X(t⇒t/σ) - μ/σ*t, t, m) , t, 0) for m in 1:4]
```

Out[64]:
$$\begin{bmatrix} 0 \\ 1 \\ 0 \\ -\frac{6}{5} \end{bmatrix}$$

5.8 問題: 独立な確率変数達の和のモーメント母函数と特性函数とキュムラント母函数

独立な確率変数達 $X_1, \dots, X_n$ の和のモーメント母函数と特性函数とキュムラント母函数が次のように表されることを示せ:

$$\begin{aligned} M_{X_1+\cdots+X_n}(t) &= M_{X_1}(t) \cdots M_{X_n}(t), \\ \varphi_{X_1+\cdots+X_n}(t) &= \varphi_{X_1}(t) \cdots \varphi_{X_n}(t), \\ K_{X_1+\cdots+X_n}(t) &= K_{X_1}(t) + \cdots + K_{X_n}(t). \end{aligned}$$

注意: この結果は空気のごとく使われる.

解答例: 独立な確率変数達 $X_1, \dots, X_n$ について

$$E[f_1(X_1) \cdots f_n(X_n)] = E[f_1(X_1)] \cdots E[f_n(X_n)]$$

が成立することより,

$$\begin{aligned} M_{X_1+\cdots+X_n}(t) &= E[e^{t(X_1+\cdots+X_n)}] = E[e^{tX_1} \cdots e^{tX_n}] = M_{X_1}(t) \cdots M_{X_n}(t), \\ \varphi_{X_1+\cdots+X_n}(t) &= E[e^{it(X_1+\cdots+X_n)}] = E[e^{itX_1} \cdots e^{itX_n}] = \varphi_{X_1}(t) \cdots \varphi_{X_n}(t). \end{aligned}$$

ゆえに

$$\begin{aligned} K_{X_1+\cdots+X_n}(t) &= \log M_{X_1+\cdots+X_n}(t) = \log(M_{X_1}(t) \cdots M_{X_n}(t)) \\ &= \log M_{X_1}(t) + \cdots + \log M_{X_n}(t) = K_{X_1}(t) + \cdots + K_{X_n}(t). \end{aligned}$$

解答終

5.9 問題: ガンマ分布のキュムラント母函数と歪度と尖度

ガンマ分布 $\mathrm{Gamma}(\alpha, \theta)$ に従う確率変数 X について, そのキュムラント母函数 $K_X(t) = \log E[e^{tX}]$ と歪度 $\bar{\kappa}_3(X)$ と尖度 $\bar{\kappa}_4(X)$ が以下のようになることを示せ:

$$K_X(t) = -\alpha \log(1 - \theta t), \quad \bar{\kappa}_3(X) = \frac{2}{\sqrt{\alpha}}, \quad \bar{\kappa}_3(X) = \frac{6}{\alpha}.$$

解答例: $t < 1/\theta$ と仮定する. このとき,

$$\mu = \alpha\theta, \quad \sigma = \sqrt{\alpha}\,\theta, \quad Z = \frac{X - \mu}{\sigma}$$

とおくと,

$$
\begin{aligned}
E[e^{tX}] &= \int_0^\infty e^{tx} \frac{e^{-x/\theta} x^{\alpha-1}}{\theta^\alpha \Gamma(\alpha)}\, dx = \frac{1}{\theta^\alpha \Gamma(\alpha)} \int_0^\infty e^{-((1-\theta t)/\theta)x} x^{\alpha-1}\, dx \\
&= \frac{1}{\theta^\alpha \Gamma(\alpha)} \frac{\theta^\alpha}{(1-\theta t)^\alpha} \Gamma(\alpha) = \frac{1}{(1-\theta t)^\alpha}, \\
K_X(t) &= -\alpha \log(1-\theta t) \\
&= \alpha\theta t + \alpha\theta^2 \frac{t^2}{2} + \alpha\theta^3 \frac{t^3}{3} + \alpha\theta^4 \frac{t^4}{4} + O(t^5) \\
&= \mu t + \sigma^2 \frac{t^2}{2} + \frac{2\sigma^3}{\sqrt{\alpha}} \frac{t^3}{3!} + \frac{6\sigma^4}{\alpha} \frac{t^4}{4!} + O(t^5), \\
K_Z(t) &= K_X\left(\frac{t}{\sigma}\right) - \frac{\mu}{\sigma} t = \frac{t^2}{2} + \frac{2}{\sqrt{\alpha}} \frac{t^3}{3!} + \frac{6}{\alpha} \frac{t^4}{4!} + O(t^5).
\end{aligned}
$$

ゆえに,

$$
\bar{\kappa}_3(X) = \frac{2}{\sqrt{\alpha}}, \quad \bar{\kappa}_3(X) = \frac{6}{\alpha}.
$$

解答終

In [65]:
```
@vars t
@vars α θ positive=true
K_X = -α*log(1 - θ*t)
κ = [diff(K_X, t, m)(t⇒0).simplify().factor() for m in 1:4]
```

Out[65]:

$$
\begin{bmatrix} \alpha\theta \\ \alpha\theta^2 \\ 2\alpha\theta^3 \\ 6\alpha\theta^4 \end{bmatrix}
$$

In [66]:
```
μ, σ = κ[1], √κ[2]
[diff(K_X(t⇒t/σ) - μ/σ*t, t, m)(t⇒0).simplify().factor() for m in 1:4]
```

Out[66]:

$$
\begin{bmatrix} 0 \\ 1 \\ \frac{2}{\sqrt{\alpha}} \\ \frac{6}{\alpha} \end{bmatrix}
$$

5.10 問題: 二項分布のキュムラント母函数と歪度と尖度

二項分布 Binomial(n, p) に従う確率変数 X について, そのキュムラント母函数 $K_X(t) = \log E[e^{tX}]$ と歪度 $\bar{\kappa}_3(X)$ と尖度 $\bar{\kappa}_4(X)$ が以下のようになることを示せ:

$$
K_X(t) = n \log(1 + p(e^t - 1)), \quad \bar{\kappa}_3(X) = \frac{1-2p}{\sqrt{np(1-p)}}, \quad \bar{\kappa}_4(X) = \frac{1-6p(1-p)}{np(1-p)}.
$$

注意: $n = 1$, $p = 1/2$ のとき $\bar{\kappa}_4(X) = -2$ となる. 実はその場合の分布 Bernoulli(1/2) が確率分布全体の中で最小の尖度を持つ分布になっている.

解答例: $X_1, \ldots, X_n$ をそれぞれがBernoulli分布 Binomial(p) に従う独立同分布な確率変数達であるとする. このとき, 二項分布に従う確率変数は $X = \sum_{i=1}^n X_i$ と作れる. ゆえに

$$
K_X(t) = \sum_{i=1}^n K_{X_i}(t) = nK_{X_1}(t).
$$

$X_1 \sim$ Bernoulli(p) について,

$$
\begin{aligned}
&E[e^{tX_1}] = e^{t1}p + e^{t0}(1-p) = 1 + p(e^t - 1), \\
&K_{X_1}(t) = \log(1 + p(e^t - 1)), \\
&K_X(t) = n \log(1 + p(e^t - 1))
\end{aligned}
$$

なので,

$$
\mu = np, \quad \sigma = \sqrt{np(1-p)}
$$

とおくと,

$$
\begin{aligned}
K_X(t) &= n\log(1+p(e^t-1)) \\
&= npt + np(1-p)\frac{t^2}{2} + np(1-p)(1-2p)\frac{t^3}{3!} + np(1-p)(6p^2-6p+1)\frac{t^4}{4!} + O(t^5) \\
&= \mu t + \sigma^2\frac{t^2}{2} + \sigma^2(1-2p)\frac{t^3}{3!} + \sigma^2(1-6p(1-p))\frac{t^4}{4!} + O(t^5), \\
K_{(X-\mu)/\sigma}(t) &= K_X\left(\frac{t}{\sigma}\right) - \frac{\mu}{\sigma}t = \frac{t^2}{2} + \frac{1-2p}{\sigma}\frac{t^3}{3!} + \frac{1-6p(1-p)}{\sigma^2}\frac{t^4}{4!} + O(t^5).
\end{aligned}
$$

ゆえに

$$
\bar{\kappa}_3(X) = \frac{1-2p}{\sigma} = \frac{1-2p}{\sqrt{np(1-p)}}, \quad \bar{\kappa}_4(X) = \frac{1-6p(1-p)}{\sigma^2} = \frac{1-6p(1-p)}{np(1-p)}.
$$

解答終

In [67]:
```
kurtosis(Bernoulli(0.5))
```

Out[67]: -2.0

In [68]:
```
@vars n p t
K_X = n * log(1 + p*(exp(t) - 1))
κ = [diff(K_X, t, m)(t⇒0).simplify().factor() for m in 1:4]
```

Out[68]:

$$
\begin{bmatrix} np \\ -np\,(p-1) \\ np\,(p-1)\,(2p-1) \\ -np\,(p-1)\left(6p^2-6p+1\right) \end{bmatrix}
$$

In [69]:
```
μ, σ = κ[1], √κ[2]
[diff(K_X(t⇒t/σ) - μ/σ*t, t, m)(t⇒0).simplify().factor() for m in 1:4]
```

Out[69]:

$$
\begin{bmatrix} 0 \\ 1 \\ -\frac{2p-1}{\sqrt{-np(p-1)}} \\ -\frac{6p^2-6p+1}{np(p-1)} \end{bmatrix}
$$

6 独立同分布な確率変数達の不偏分散の分散

以下の節においては, $X_1, \ldots, X_n$ は独立同分布な確率変数達であるとし, それらの標本分散と不偏分散を以下のように書くことにする:

$$
\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2.
$$

さらに, X_i の期待値と分散をそれぞれ μ, σ^2 と書くことにする.

このとき, 以下が成立することはすでに示した:

$$
E[\bar{X}] = \mu, \quad E[S_X^2] = \sigma^2, \quad \mathrm{var}(X) = E[(\bar{X}-\mu)^2] = \frac{\sigma^2}{n}.
$$

さらに以下では, X_i の **歪度** (わいど, skewness)と **尖度** (せんど, kurtosis)をそれぞれ $\bar{\kappa}_3 = \bar{\kappa}_3(X_i)$, $\bar{\kappa}_4 = \bar{\kappa}_4(X_i)$ と書くことにする:

$$
\bar{\kappa}_3 = E\left[\left(\frac{X_i-\mu}{\sigma}\right)^3\right], \quad \bar{\kappa}_4 = E\left[\left(\frac{X_i-\mu}{\sigma}\right)^4\right] - 3.
$$

このとき, 標本平均と不偏分散の共分散と不偏分散の分散は X_i の歪度 $\bar{\kappa}_3$ と尖度 $\bar{\kappa}_4$ を使って次のように書ける:

$$
\mathrm{cov}(\bar{X}, S_X^2) = \sigma^3\frac{\bar{\kappa}_3}{n}, \quad \mathrm{var}(S_X^2) = \sigma^4\left(\frac{\bar{\kappa}_4}{n} + \frac{2}{n-1}\right).
$$

以下ではこの結果を証明する.

6.1 期待値 0, 分散 1 の場合への帰着

$Z = (X - \mu)/\sigma$ とおくと, Z の期待値と分散はそれぞれ 0 と 1 になる.

$Z_i = (X_i - \mu)/\sigma$ とおくと, $Z_1, \dots, Z_n$ はそれぞれが Z と同じ分布に従う独立同分布な確率変数達になる. そして, $X_i = \sigma Z_i + \mu$ より,

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i, \quad S_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2$$

とおくと,

$$\bar{X} = \sigma\bar{Z} + \mu, \quad S_X^2 = \sigma^2 S_Z^2$$

なので,

$$\mathrm{cov}(\bar{X}, S_X^2) = \sigma^3\,\mathrm{cov}(\bar{Z}, S_Z^2), \quad \mathrm{var}(S_X^2) = \sigma^4\,\mathrm{var}(S_Z^2).$$

これより, 標本分散と不偏分散の分散と共分散の計算は $\mu = 0, \sigma = 1$ の場合に帰着する.

6.2 標本平均と不偏分散の共分散の計算

標本平均と不偏分散の共分散は X_i の歪度 $\bar{\kappa}_3$ を使って次のように書ける:

$$\mathrm{cov}(\bar{X}, S_X^2) = \frac{\bar{\kappa}_3}{n}\sigma^3.$$

証明: $\mu = 0, \sigma = 1$ と仮定する. このとき, $\mu = 0$ という仮定より,

$$\mathrm{cov}(\bar{X}, S_X^2) = E[\bar{X}(S_X^2 - \sigma^2)] = E[\bar{X}S_X^2] - \sigma^2\underbrace{E[\bar{X}]}_{=0} = E[\bar{X}S_X^2].$$

そして,

$$\begin{aligned}
n\bar{X}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \frac{2}{n}\sum_{1\le i<j\le n} X_iX_j, \\
(n-1)S_X^2 &= \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i^2 - 2\bar{X}X_i + \bar{X}^2) = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \\
&= \sum_{i=1}^{n} X_i^2 - \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{2}{n}\sum_{1\le i<j\le n} X_iX_j = \frac{n-1}{n}\sum_{i=1}^{n} X_i^2 - \frac{2}{n}\sum_{1\le i<j\le n} X_iX_j, \\
\therefore \quad S_X^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{2}{n(n-1)}\sum_{1\le i<j\le n} X_iX_j.
\end{aligned}$$

ゆえに

$$\bar{X}S_X^2 = \frac{1}{n}\sum_{i=1}^{n} X_i\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{2}{n(n-1)}\sum_{1\le i<j\le n} X_iX_j\right).$$

$E[X_i] = \mu = 0$ という仮定と X_i 達の独立性より, i, j, k がすべて等しいとき以外には $E[X_iX_jX_k] = 0$ となり, $\mu = 0, \sigma = 1$ という仮定より $E[X_i^3] = \bar{\kappa}_3$ となるので,

$$\mathrm{cov}(\bar{X}, S_X^2) = E[\bar{X}S_X^2] = \frac{1}{n^2}\sum_{i=1}^{n} E[X_i^3] = \frac{\bar{\kappa}_3}{n}.$$

$\mu = 0, \sigma = 1$ と仮定しない一般の場合には, これが σ^3 倍されて,

$$\mathrm{cov}(\bar{X}, S_X^2) = \sigma^3\frac{\bar{\kappa}_3}{n}.$$

証明終

6.3 不偏分散の分散の計算

不偏分散の分散は X_i の尖度 $\bar{\kappa}_4$ を使って次のように書ける:

$$\mathrm{var}(S_X^2) = \sigma^4 \left(\frac{\bar{\kappa}_4}{n} + \frac{2}{n-1} \right).$$

証明: $\mu = 0, \sigma = 1$ と仮定する. このとき, $E[S_X^2] = \sigma^2 = 1$ より,

$$\mathrm{var}(S_X^2) = E[(S_X^2)^2] - E[S_X^2]^2 = E[(S_X^2)^2] - 1.$$

そして, 前節の計算より

$$S_X^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} X_i X_j$$

なので,

$$\begin{aligned}
(S_X^2)^2 = & \frac{1}{n^2} \left(\sum_{i=1}^{n} X_i^4 + 2 \sum_{1 \le i < j \le n} X_i^2 X_j^2 \right) \\
& - \frac{2}{n^2(n-1)} \sum_{i=1}^{n} \sum_{1 \le j < k \le n} X_i^2 X_j X_k \\
& + \frac{4}{n^2(n-1)^2} \sum_{1 \le i < j \le n} \sum_{1 \le k < l \le n} X_i X_j X_k X_l.
\end{aligned}$$

$\mu = 0, \sigma = 1$ でかつ $X_1, \dots, X_n$ は独立であるという仮定と, $E[X_i] = 0$, $E[X_i^2] = 1$, $E[X_i^4] = \bar{\kappa}_4 + 3$ であることを使うと,

$$\begin{aligned}
& E\left[\sum_{i=1}^{n} X_i^4 \right] = n(\bar{\kappa}_4 + 3), \\
& E\left[2 \sum_{1 \le i < j \le n} X_i^2 X_j^2 \right] = 2 \frac{n(n-1)}{2} = n(n-1), \\
& E\left[\sum_{1 \le j < k \le n} X_i^2 X_j X_k \right] = 0, \\
& E\left[\sum_{1 \le i < j \le n} \sum_{1 \le k < l \le n} X_i X_j X_k X_l \right] = \frac{n(n-1)}{2}.
\end{aligned}$$

最後の等式は $i = k, j = l$ の場合のみが和に寄与することを使った. ゆえに

$$\begin{aligned}
\mathrm{var}(S_X^2) &= \frac{n(\bar{\kappa}_4 + 3) + n(n-1)}{n^2} + \frac{4}{n^2(n-1)^2} \frac{n(n-1)}{2} - 1 \\
&= \frac{\bar{\kappa}_4 + n + 2}{n} + \frac{2}{n(n-1)} - 1 \\
&= \frac{\bar{\kappa}_4 + n + 2}{n} + \frac{2}{n-1} - \frac{2}{n} - \frac{n}{n} \\
&= \frac{\bar{\kappa}_4}{n} + \frac{2}{n-1}.
\end{aligned}$$

$\mu = 0, \sigma = 1$ と仮定しない一般の場合には, これが σ^4 倍されて,

$$\mathrm{var}(S_X^2) = \sigma^4 \left(\frac{\bar{\kappa}_4}{n} + \frac{2}{n-1} \right).$$

証明終

6.4 歪度と尖度に関する不等式

以上の結果を用いて, 歪度 $\bar{\kappa}_3$ と尖度 $\bar{\kappa}_4$ のあいだには

$$\bar{\kappa}_4 \ge \bar{\kappa}_3^2 - 2 \ge -2$$

が成立することを示そう. 標本平均 $\bar{X}$ と不偏分散 S_X^2 の分散共分散行列

$$\begin{bmatrix} \mathrm{var}(\bar{X}) & \mathrm{cov}(\bar{X}, S_X^2) \\ \mathrm{cov}(S_X^2, \bar{X}) & \mathrm{var}(S_X^2) \end{bmatrix} = \begin{bmatrix} \sigma^2/n & \sigma^3 \bar{\kappa}_3/n \\ \sigma^3 \bar{\kappa}_3/n & \sigma^4(\bar{\kappa}_4/n + 2/(n-1)) \end{bmatrix}$$

は半正定値になる. 特にその行列式は 0 以上になる:

$$0 \leq \frac{\sigma^2}{n}\sigma^4\left(\frac{\bar{\kappa}_4}{n}+\frac{2}{n-1}\right)-\left(\frac{\sigma^3\bar{\kappa}_3}{n}\right)^2=\frac{\sigma^6}{n^2}\left(\bar{\kappa}_4+\frac{2}{1-1/n}-\bar{\kappa}_3^2\right).$$

ゆえに,

$$\bar{\kappa}_4 \geq \bar{\kappa}_3^2-\frac{2}{1-1/n}.$$

$n \to \infty$ とすることによって,

$$\bar{\kappa}_4 \geq \bar{\kappa}_3^2-2.$$

を得る.

注意: 上の不等式は **Karl Pearson の不等式** と呼ばれることがある.

- Pearson, Karl. IX. Mathematical contributions to the theory of evolution.—XIX. Second supplement to a memoir on skew variation. Philosophical Transactions of the Royal Society of London A, 216 (546): 429–457. [doi (https://doi.org/10.1098/rsta.1916.0009)]

6.5 問題: 不偏分散と不偏補正されていない標本分散の平均二乗誤差の比較

不偏補正されていない($n-1$ ではなく $n-1$ で割ってできる)標本分散

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2=\frac{n-1}{n}S_X^2$$

と推定先の値 σ^2 の差の二乗の期待値(平均二乗誤差)が次のようになることを示せ:

$$E\left[\left(\frac{n-1}{n}S_X^2-\sigma^2\right)^2\right]=\left(\frac{n-1}{n}\right)^2\mathrm{var}(S_X^2)+\frac{\sigma^4}{n^2}.$$

これより, 次の公式が導かれることを示せ:

$$\mathrm{var}(S_X^2)-E\left[\left(\frac{n-1}{n}S_X^2-\sigma^2\right)^2\right]=\frac{\sigma^4}{n^2}\left(\left(2-\frac{1}{n}\right)\bar{\kappa}_4+3+\frac{2}{n-1}\right).$$

これより, $\bar{\kappa}_4 \geq -3/2$ ならば不偏分散の平均二乗誤差の方が不偏補正されていない標本平均の平均二乗誤差よりも大きくなることがわかる.

注意: $\bar{\kappa}_4 \geq \bar{\kappa}_3^2-2 \geq -2$ なので多くの場合に $\bar{\kappa}_4 > -3/2$ となる. 例えば正規分布では $\bar{\kappa}_4=0$ なので,

$$\mathrm{var}(S_X^2)-E\left[\left(\frac{n-1}{n}S_X^2-\sigma^2\right)^2\right]=\frac{\sigma^4}{n^2}\frac{3n-1}{n-1}$$

となり, 不偏分散の平均二乗誤差は補正されていない標本分散より大きくなる. 不偏分散の採用にはこのようなトレードオフも存在するが, この害は小さいと考えられる.

解答例: $(n-1)/n-1=-1/n$ と $E[S_X^2]=\sigma^2$, $E[(S_X^2-\sigma^2)^2]=\mathrm{var}(S_X^2)$ より,

$$\begin{aligned}
E\left[\left(\frac{n-1}{n}S_X^2-\sigma^2\right)^2\right] &= E\left[\left(\frac{n-1}{n}(S_X^2-\sigma^2)-\frac{\sigma^2}{n}\right)^2\right] \\
&= E\left[\left(\frac{n-1}{n}\right)^2(S_X^2-\sigma^2)^2-2\frac{\sigma^2}{n}\frac{n-1}{n}(S_X^2-\sigma^2)+\frac{\sigma^4}{n^2}\right] \\
&= \left(\frac{n-1}{n}\right)^2\mathrm{var}(S_X^2)+\frac{\sigma^4}{n^2}.
\end{aligned}$$

これより,

$$\mathrm{var}(S_X^2)-E\left[\left(\frac{n-1}{n}S_X^2-\sigma^2\right)^2\right]=\frac{2n-1}{n^2}\mathrm{var}(S_X^2)-\frac{\sigma^4}{n^2}.$$

これに $\mathrm{var}(S_X^2)=\sigma^4\left(\bar{\kappa}_4/n+2/(n-1)\right)$ を代入すると,

$$
\begin{aligned}
\mathrm{var}(S_X^2) - E\left[\left(\frac{n-1}{n}S_X^2 - \sigma^2\right)^2\right] &= \frac{2n-1}{n^2}\sigma^4\left(\frac{\bar{\kappa}_4}{n} + \frac{2}{n-1}\right) - \frac{\sigma^4}{n^2} \\
&= \frac{\sigma^4}{n^2}\left(\frac{2n-1}{n}\bar{\kappa}_4 + \frac{4n-2}{n-1} - 1\right) = \frac{\sigma^4}{n^2}\left(\frac{2n-1}{n}\bar{\kappa}_4 + \frac{3n-1}{n-1}\right) \\
&= \frac{\sigma^4}{n^2}\left(\left(2 - \frac{1}{n}\right)\bar{\kappa}_4 + 3 + \frac{2}{n-1}\right).
\end{aligned}
$$

解答終

In [70]:

```
function print_varS²_minus_MSEV²(dist, n; L=10^6)
    @show dist
    @show μ = mean(dist)
    @show σ = std(dist)
    @show κ = kurtosis(dist)
    @show n
    @show L = 10^6
    X = rand(dist, n, L)
    S² = var.(eachcol(X))
    V² = var.(eachcol(X); corrected=false)
    var_S² = mean((s² - σ^2)^2 for s² in S²)
    MSE_V² = mean((v² - σ^2)^2 for v² in V²)
    theoretical_val = σ^4/n^2*((2n-1)/n * κ + (3n-1)/(n-1))
    theoretical_value = theoretical_val
    @show theoretical_val
    @show var_S² - MSE_V²
    @show theoretical_value / σ^4
    @show (var_S² - MSE_V²) / σ^4
    nothing
end
```

Out[70]:

```
print_varS²_minus_MSEV² (generic function with 1 method)
```

In [71]:

```
print_varS²_minus_MSEV²(Normal(1, 0.5), 5)
```

```
dist = Normal{Float64}(μ=1.0, σ=0.5)
μ = mean(dist) = 1.0
σ = std(dist) = 0.5
κ = kurtosis(dist) = 0.0
n = 5
L = 10 ^ 6 = 1000000
theoretical_val = 0.00875
var_S² - MSE_V² = 0.008711816245110061
theoretical_value / σ ^ 4 = 0.14
(var_S² - MSE_V²) / σ ^ 4 = 0.13938905992176098
```

正規分布の場合に確かに理論通りに不偏分散の平均二乗誤差の方が非補正標本分散の平均二乗誤差よりも大きくなっている.

In [72]:

```
print_varS²_minus_MSEV²(Bernoulli(0.5), 5)
```

```
dist = Bernoulli{Float64}(p=0.5)
μ = mean(dist) = 0.5
σ = std(dist) = 0.5
κ = kurtosis(dist) = -2.0
n = 5
L = 10 ^ 6 = 1000000
theoretical_val = -0.0002500000000000002
var_S² - MSE_V² = -0.00025235920018789205
theoretical_value / σ ^ 4 = -0.0040000000000000036
(var_S² - MSE_V²) / σ ^ 4 = -0.004037747203006273
```

これが尖度が最小の場合. 不偏分散の平均二乗誤差の方を確かに小さくできているが, ほんの少しだけである.

7 標本分布における標本平均と不偏分散の同時分布の視覚化

以下の節においては, $X_1, \ldots, X_n$ は独立同分布な確率変数達であるとし, それらの標本分散と不偏分散を以下のように書くことにする:

$$
\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.
$$

さらに, 各 X_i の期待値と分散と歪度と尖度をそれぞれ $\mu, \sigma^2, \bar{\kappa}_3, \bar{\kappa}_4$ と書くことにする. このとき, 標本平均 $\bar{X}$ と不偏分散 S^2 の同時分布の期待値と分散・共分散は以下のようになるのであった:

$$
\begin{aligned}
&E[\bar{X}] = \mu, \quad E[S^2] = \sigma^2, \\
&\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}, \quad \mathrm{cov}(\bar{X}, S^2) = \sigma^3 \frac{\bar{\kappa}_3}{n}, \quad \mathrm{var}(S^2) = \sigma^4 \left(\frac{\bar{\kappa}_4}{n} + \frac{2}{n-1} \right).
\end{aligned}
$$

これを, 期待値ベクトルと分散共分散行列の形式で書くと,

$$
\begin{aligned}
&\begin{bmatrix} E[\bar{X}] \\ E[S^2] \end{bmatrix} = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}, \\
&\begin{bmatrix} \mathrm{var}(\bar{X}) & \mathrm{cov}(\bar{X}, S^2) \\ \mathrm{cov}(S^2, \bar{X}) & \mathrm{var}(S^2) \end{bmatrix} = \frac{1}{n} \begin{bmatrix} \sigma^2 & \sigma^3 \bar{\kappa}_3 \\ \sigma^3 \bar{\kappa}_3 & \sigma^4 \left(\bar{\kappa}_4 + \frac{2}{1-1/n} \right) \end{bmatrix} \\
&\qquad = \frac{1}{n} \begin{bmatrix} \sigma & 0 \\ 0 & \sigma^2 \end{bmatrix} \begin{bmatrix} 1 & \bar{\kappa}_3 \\ \bar{\kappa}_3 & \bar{\kappa}_4 + \frac{2}{1-1/n} \end{bmatrix} \begin{bmatrix} \sigma & 0 \\ 0 & \sigma^2 \end{bmatrix}.
\end{aligned}
$$

以下では X_i の従う分布が色々な場合に $\bar{X}, S^2$ の同時分布を視覚化してみる. その結果を見ると, 一般に以下が成立していそうなことがわかる:

- 標本平均 $\bar{X}$ と不偏分散 S^2 の組 $(\bar{X}, S^2)$ の平面上での分布は, n が十分に大きければ (μ, σ^2) を中心とする楕円状の小さな領域内に集中する分布(実は多変量正規分布で近似される)になり(中心極限定理!), $n \to \infty$ で (μ, σ^2) に集中するようになる(大数の法則!).
- もとの分布が対称ならば $(\bar{X}, S^2)$ の分布も左右対称になり, もとの分布が非対称ならば $(\bar{X}, S^2)$ の分布も左右非対称になり, n が大きなとき斜めに傾いた楕円状の領域内に集中する分布になる.

In [73]:

```
myskewness(dist) = skewness(dist)
myskewness(dist::MixtureModel) = _myskewness(dist)
function _myskewness(dist)
    μ, σ = mean(dist), std(dist)
    quadgk(x -> ((x-μ)/σ)^3*pdf(dist, x), extrema(dist)...)[1]
end

mykurtosis(dist) = kurtosis(dist)
mykurtosis(dist::MixtureModel) = _mykurtosis(dist)
function _mykurtosis(dist)
    μ, σ = mean(dist), std(dist)
    quadgk(x -> ((x-μ)/σ)^4*pdf(dist, x), extrema(dist)...)[1] - 3
end
```

Out[73]:

```
_mykurtosis (generic function with 1 method)
```

In [74]:

```
skewness(Gamma(2, 3)), _myskewness(Gamma(2,3))
```

Out[74]:

```
(1.414213562373095, 1.4142135623731023)
```

In [75]:

```
kurtosis(Gamma(2, 3)), _mykurtosis(Gamma(2,3))
```

Out[75]:

```
(3.0, 2.999999999999999)
```

In [76]:

```
function plot_X̄_and_SX²(dist; n = 10, L = 10^4,
        sk = myskewness(dist), ku = mykurtosis(dist),
        size = (400, 350), legend = :topleft, alpha = 0.4, kwargs...)
    μ, σ = mean(dist), std(dist)
    Σ = (1/n)*[σ^2 σ^3*sk; σ^3*sk σ^4*(ku+2/(1-1/n))]
    A = inv(Σ)
    f(x, y) = A[1,1]*(x-μ)^2 + 2A[1,2]*(x-μ)*(y-σ^2) + A[2,2]*(y-σ^2)^2
    h = quantile(Chisq(2), 0.95)
    X = rand(dist, n, L)
    X̄ = mean.(eachcol(X))
    S² = var.(eachcol(X))
    scatter(X̄, S²; ms=1.5, msw=0, alpha, label="")
    scatter!([μ], [σ^2]; msc=:auto, label="(μ, σ²)")
    x = range(extrema(X̄)..., 200)
    y = range(extrema(S²)..., 200)
    contour!(x, y, f; label="", levels=[h], c=2, colorbar=false)
    plot!(; xlabel="x̄", ylabel="s²")
    plot!(; size, legend, kwargs...)
end

function plot_X̄_and_SX²_2x2(dist; ns = (10, 40, 160, 640),
        size=(800, 700), kwargs...)
    μ, σ², sk, ku = mean(dist), var(dist), myskewness(dist), mykurtosis(dist)
    println(dist)
    @show μ σ²
    println("skewness = ", sk)
    println("kurtosis = ", ku)

    PP = []
    for n in ns
        P = plot_X̄_and_SX²(dist; n, sk, ku, kwargs..., title="n = $n")
        push!(PP, P)
    end
    plot(PP...; size, layout=(2,2),
        titlefontsize=12, guidefontsize=10, tickfontsize=8)
end
```

Out[76]: plot_X̄_and_SX²_2x2 (generic function with 1 method)

7.1 正規分布の標本分布における $\bar{X}, S^2$ の同時分布

In [77]:
```
plot_X̄_and_SX²_2x2(Normal(); xlim=(-1.1, 1.1), ylim=(-0.1, 3.0))
```

```
Normal{Float64}(μ=0.0, σ=1.0)
μ = 0.0
σ² = 1.0
skewness = 0.0
kurtosis = 0.0
```

Out[77]:

7.2 一様分布の標本分布における$\bar{X}, S^2$ の同時分布

In [78]:
```
plot_X̄_and_SX²_2x2(Uniform(); xlim=(0.2, 0.8), ylim=(0.01, 0.18))
```

```
Uniform{Float64}(a=0.0, b=1.0)
μ = 0.5
σ² = 0.08333333333333333
skewness = 0.0
kurtosis = -1.2
```

Out[78]:

7.3 単峰型のガンマ分布の標本分布における $\bar{X}, S^2$ の同時分布

In [79]:
```
plot_X̄_and_SX²_2x2(Gamma(3, 4); xlim=(5, 20), ylim=(0, 200))
```

```
Gamma{Float64}(α=3.0, θ=4.0)
μ = 12.0
σ² = 48.0
skewness = 1.1547005383792517
kurtosis = 2.0
```

Out[79]:

7.4 指数分布の標本分布における $\bar{X}, S^2$ の同時分布

In [80]:
```
plot_X̄_and_SX²_2x2(Exponential(); xlim=(0.2, 2.3), ylim=(-0.2, 7))
```

```
Exponential{Float64}(θ=1.0)
μ = 1.0
σ² = 1.0
skewness = 2.0
kurtosis = 6.0
```

Out[80]:

7.5 非対称なベータ分布の標本分布における $\bar{X}$, S^2 の同時分布

In [81]:
```
plot_X̄_and_SX²_2x2(Beta(0.1, 0.9); xlim=(-0.02, 0.4), ylim=(-0.05, 0.2))
```

```
Beta{Float64}(α=0.1, β=0.9)
μ = 0.1
σ² = 0.045000000000000005
skewness = 2.514157444218836
kurtosis = 5.611111111111112
```

Out[81]:

7.6 非対称な2つ山の混合正規分布の標本分布における $\bar{X}, S^2$ の同時分布

In [82]:

```
plot_X̄_and_SX²_2x2(MixtureModel([Normal(), Normal(20)], [0.95, 0.05]);
    xlim=(-1, 8), ylim=(-2, 120))
```

```
MixtureModel{Normal{Float64}}(K = 2)
components[1] (prior = 0.9500): Normal{Float64}(μ=0.0, σ=1.0)
components[2] (prior = 0.0500): Normal{Float64}(μ=20.0, σ=1.0)

μ = 1.0
σ² = 20.0
skewness = 3.823676241524649
kurtosis = 13.58499999999097
```

Out[82]:

7.7 非対称なBernoulli分布の標本分布における $\bar{X}, S^2$ の同時分布

In [83]:
```
plot_X̄_and_SX²_2x2(Bernoulli(0.1); xlim=(-0.02, 0.6), ylim=(-0.01, 0.3))
```

```
Bernoulli{Float64}(p=0.1)
μ = 0.1
σ² = 0.09000000000000001
skewness = 2.6666666666666665
kurtosis = 5.111111111111109
```

Out[83]:

7.8 対数正規分布の標本分布における $\bar{X}, S^2$ の同時分布

In [84]:
```
plot_X̄_and_SX²_2x2(LogNormal(); xlim=(-0.2, 5), ylim=(-2, 50))
```

```
LogNormal{Float64}(μ=0.0, σ=1.0)
μ = 1.6487212707001282
σ² = 4.670774270471604
skewness = 6.184877138632554
kurtosis = 110.9363921763115
```

Out[84]:

対数正規分布では $(\bar{X}, S^2)$ の分布の正規分布近似の精度をよくするためには相当に n を大きくする必要がある.

8 正規分布の標本分布の場合

この節では, $X_1, \ldots, X_n$ はそれぞれが正規分布 $\mathrm{Normal}(\mu, \sigma)$ に従う独立な確率変数達であると仮定する:

$$(X_1, \ldots, X_n) \sim \mathrm{Normal}(\mu, \sigma)^n.$$

さらに, それらの標本分散と不偏分散を

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

と書く. このとき, 正規分布の標本分布の特殊性によって, $\bar{X}$ と S^2 は独立になり, $\bar{X}$ は平均 μ, 分散 σ^2/n の正規分布に従い,

$$Y = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

は自由度 $\nu = n-1$ のχ²分布に従う:

$$\left(\bar{X}, \frac{(n-1)S^2}{\sigma^2}\right) \sim \mathrm{Normal}\left(\mu, \sigma/\sqrt{n}\right) \times \mathrm{Chisq}(n-1).$$

以下ではこの結果を証明する.

8.1 準備: $\sqrt{n}\,\bar{x}$ を含む正規直交座標系

$\mathbb{R}^n$ を縦ベクトルの空間とみなし, その標準座標系を $x_1, \dots, x_n$ と書く. このとき, $\mathbb{R}^n$ の任意の正規直交基底 $v_1, \dots, v_n$ について

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} y_i v_i$$

で定義される $\mathbb{R}^n$ の座標系 $y_1, \dots, y_n$ をその直交基底が定める $\mathbb{R}^n$ の **正規直交座標系** と呼ぶ. このとき, ベクトル $a, b \in \mathbb{R}^n$ の標準内積を (a, b) と書く, それに対応するノルムを $\|a\| = \sqrt{(a,a)}$ と書くと,

$$y_i = (v_i, x), \quad \|x\|^2 = \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i^2$$

などが成立している.

以下では $v_n \in \mathbb{R}^n$ は

$$v_n = \frac{1}{\sqrt{n}}\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

と定められていると仮定し, $\bar{x}$ を次のように定める:

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

このとき, $\|v_n\|^2 = (1^2 + \cdots + 1^2)/n = 1$ なので v_n は長さ 1 の単位ベクトルになり,

$$(v_n, x) = \frac{x_1 + \cdots + x_n}{\sqrt{n}} = \sqrt{n}\,\bar{x}.$$

さらに, v_n を含む $\mathbb{R}^n$ の任意の正規直交基底 $v_1, \dots, v_{n-1}, v_n$ を取り, それが定める正規直交座標系を $y_1, \dots, y_{n-1}, y_n$ と書くことにする. このとき, $y_n = (v_n, x) = \sqrt{n}\,\bar{x}$ なので, $y_1, \dots, y_{n-1}, y_n$ は $y_n = \sqrt{n}\,\bar{x}$ の正規直交座標系への拡張になっている.

例: v_n を含む正規直交基底が任意に取られているせいで議論が抽象的過ぎるように感じる人達は, 以下の $v_1, \dots, v_{n-1}$ 上の v_n を合わせたものがそのような正規直交基底の具体例になっていることを確認して, それらを使用すればよい:

$$v_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad v_2 = \frac{1}{\sqrt{6}}\begin{bmatrix} 1 \\ 1 \\ -2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \dots, \quad v_{n-1} = \frac{1}{\sqrt{n(n-1)}}\begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ -(n-1) \end{bmatrix}.$$

これらが内積について $(v_i, v_j) = \delta_{ij}$ を満たしていること(正規直交基底であること)は, この定義式を眺めて暗算すればすぐに確かめられる. **例終**

ベクトル v が

$$v = \begin{bmatrix} x_1 - \mu \\ \vdots \\ x_n - \mu \end{bmatrix} = x - \sqrt{n}\,\mu v_n$$

と定義されているとき,

$$(v_i, v) = (v_i, x) = y_i \quad (i = 1, \dots, n-1), \quad (v_n, v) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (x_i - \mu) = \sqrt{n}(\bar{x} - \mu)$$

なので,

$$\|v\|^2 = \sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n-1} y_i^2 + n(\bar{x} - \mu)^2.$$

さらに不偏分散関係でよく出て来る公式

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} ((x_i - \mu) - (\bar{x} - \mu))^2 = \sum_{i=1}^{n} (x_i - \mu)^2 - n(\bar{x} - \mu)^2$$

より,

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2, \quad \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n-1} y_i^2$$

であることもわかる. これらの公式は後で正規分布の確率密度函数達の積に適用される.

8.2 準備 $n-2$ 次元単位球面上の微小面積要素

前節の設定をそのまま引き継ぐ.

$\mathbb{R}^n$ における $\mathbb{R}v_n$ の直交補空間上の正規直交座標系として $y_1, \dots, y_{n-1}$ を採用できる. その直交補空間内での $n-2$ 単位球面 $S^{n-2} : y_1^2 + \cdots + y_{n-1}^2 = 1$ 上の積分で使用可能な座標系 $\theta = (\theta_1, \dots, \theta_{n-2})$ を任意に取って, その座標系での球面上の微小面積要素を $d\theta$ と略して書くことにする. このとき, $n-2$ 次元単位球面 S^{n-2} の面積 A_{n-2} は

$$A_{n-2} = \int_{S^{n-2}} d\theta = \frac{2\pi^{(n-1)/2}}{\Gamma((n-1)/2)}$$

と表される.

例: 具体的に座標系を取っていないせいで議論が抽象的過ぎると感じる人は, 以下の座標系を利用すればよい:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_{n-1} \end{bmatrix} = \begin{bmatrix} \cos\theta_1 \cos\theta_2 \cos\theta_3 \cdots \cos\theta_{n-2} \\ \sin\theta_1 \cos\theta_2 \cos\theta_3 \cdots \cos\theta_{n-2} \\ \sin\theta_2 \cos\theta_3 \cdots \cos\theta_{n-2} \\ \sin\theta_3 \cdots \cos\theta_{n-2} \\ \ddots \quad \vdots \\ \sin\theta_{n-2} \end{bmatrix} \qquad \begin{pmatrix} -\pi \le \theta_1 \le \pi, \\ -\frac{\pi}{2} \le \theta_2, \dots, \theta_{n-2} \le \frac{\pi}{2} \end{pmatrix}$$

この座標系は「Gauss積分, ガンマ函数, ベータ函数」のノートの「$n-1$ 次元単位球面上の座標を使った面積の計算」の節で紹介したものと本質的に同じものである. この座標系で微小面積要素は次のように表される:

$$d\theta = \cos^0\theta_1 \cos^1\theta_2 \cos^2\theta_3 \cdots \cos^{n-3}\theta_{n-2}\, d\theta_1\, d\theta_2\, d\theta_3 \cdots d\theta_{n-2}.$$

しかし, この例で紹介した複雑な式は後で使用されない. **例終**

このとき, さらに,

$$r = \sqrt{y_1^2 + \cdots + y_{n-1}^2}$$

とおくと, $\mathbb{R}^n$ における積分のための座標として $(\bar{x}, r, \theta)$ を取れて, その微小体積要素 dV は

$$dV = dy_n\, r^{n-2}\, dr\, d\theta = \sqrt{n}\, r^{n-2}\, d\bar{x}\, dr\, d\theta$$

と書ける. さらに, 不偏分散 S^2 に対応する変数 $u = s^2$ を

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

と定めると, $r = \sqrt{(n-1)u}$ なので, これを上の式に代入して,

$$dV = \sqrt{n}\,(n-1)^{(n-1)/2}\,u^{(n-1)/2-1}\,d\bar{x}\,du\,d\theta.$$

これの右辺の定数因子の形は重要ではない.

8.3 正規分布の標本分布における $\bar{X}, S^2$ の同時分布の計算

前節までの設定をそのまま引き継ぐ.

このとき, $(X_1, \dots, X_n) \sim \mathrm{Normal}(\mu, \sigma)^n$ の同時密度函数は以上で容易した座標系で

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 = (n-1)u + n(\bar{x} - \mu)^2$$

を代入すれば

$$\mathrm{const.}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) = \mathrm{const.}\exp\left(-\frac{(n-1)u}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right)$$

と表される. 確率変数 S^2 に対応する積分変数を u と書くことにしたのであった. ゆえに,

$$\begin{aligned}
&E[f(\bar{X}, S^2)] \\
&= \mathrm{const.}\int_{S^{n-2}}\left(\int_0^\infty\left(\int_{\mathbb{R}} f(\bar{x}, u)\exp\left(-\frac{(n-1)u}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right)u^{(n-1)/2-1}\,d\bar{x}\right)du\right)d\theta \\
&= \mathrm{const.}\int_0^\infty\left(\int_{\mathbb{R}} f(\bar{x}, u)\exp\left(-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right)\exp\left(-\frac{(n-1)u}{2\sigma^2}\right)u^{(n-1)/2-1}\,d\bar{x}\right)du.
\end{aligned}$$

const. の部分は $f = 1$ の場合にこれが 1 になることから決定される. この式から S^2 に対応する通常の変数を $u > 0$ と書くことにすれば, $\bar{X}, S^2$ の同時確率密度函数は次の形になることがわかる:

$$\mathrm{const.}\exp\left(-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right)\exp\left(-\frac{(n-1)u}{2\sigma^2}\right)u^{(n-1)/2-1}.$$

これは $\bar{X}$ と S^2 が独立であり,

$$\begin{aligned}
\bar{X} &\sim \mathrm{Normal}\left(\mu, \sqrt{\frac{\sigma^2}{n}}\right), \\
S^2 &\sim \mathrm{Gamma}\left(\frac{n-1}{2}, \frac{2\sigma^2}{n-1}\right) = \frac{\sigma^2\,\mathrm{Chisq}(n-1)}{n-1}
\end{aligned}$$

となっていることを意味している. これらは次と同値である:

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim \mathrm{Normal}(0, 1), \qquad \frac{(n-1)S^2}{\sigma^2} \sim \mathrm{Chisq}(n-1).$$

8.4 正規分布の標本分布から自然に t 分布に従う確率変数が得られること

一般に, Z, Y が独立な確率変数達で, Z は標準正規分布に従い, Y が自由度 ν のχ²分布に従うとき, $T = Z/\sqrt{Y/\nu}$ は自由度 ν の t 分布に従うのであった.

この結果を標本平均 $\bar{X}$ の標準化

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

とこれと独立に自由度 $\nu = n - 1$ のχ²分布に従う

$$Y = \frac{(n-1)S^2}{\sigma^2} \sim \mathrm{Chisq}(n-1)$$

に適用すると, $Y/(n-1) = S^2/\sigma^2$ なので,

$$T = \frac{Z}{\sqrt{Y/(n-1)}} = \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}\sqrt{\frac{\sigma^2}{S^2}} = \frac{\bar{X}-\mu}{\sqrt{S^2/n}}$$

が自由度 $n-1$ の t 分布に従うことがわかる:

$$T = \frac{\bar{X}-\mu}{\sqrt{S^2/n}} \sim \mathrm{TDist}(\nu).$$

注意: この T は Z が含む分散パラメータ σ^2 を標本の不偏分散 S^2 に置き換えた形をしており, Z と違って T は σ^2 を使わずに書けている. T は μ のみを含み σ^2 を含まないので, 分散 σ^2 の設定によらない μ に関する結果を得たいときに, この T は使える. 実際, この T は, 正規分布モデルによって, 平均 μ に関する区間推定や検定を σ^2 の存在に煩わされずに行うために使用される.

8.5 t 分布の使いどころ

以上では, 正規分布のサイズ n の標本分布においては, 標本平均の標準化の式に含まれる分散のパラメータ σ^2 を不偏分散 S^2 で置き換えると, 自由度 $n-1$ の t 分布に従う確率変数 T が得られることを長くて精密な計算によって示すことができた.

しかし, 次の疑問は重要である:

> このようにがんばって t 分布に従う確率変数を正規分布から作っても, 現実世界のデータがぴったり正規分布の標本分布に従っていることがありそうだとは思えない. 現実世界におけるデータは近似的にも全然正規分布に従っていない場合が多いだろう. 以上の計算は現実での統計分析に役に立つのだろうか?

この疑問への答えを得るためには中心極限定理について学ぶ必要がある. (グラフのプロットの形ではすでに中心極限定理の例をすでに我々は目で確認している.)

以上で構成した t 分布の自由度は $n-1$ だったので, 標本サイズ n が大きな場合にはその t 分布は標準正規分布で近似される. 実は正規分布の標本分布に限らず, 適当な緩い条件を満たす任意の確率分布の標本分布について, 中心極限定理より, T と同じ式で定義された統計量が標準正規分布で近似されることを示せる. そのような場合には t 分布を使うことに実践的に大して意味があるとは言えないが, t 分布を使用しても害はない(正規分布以外の標本分布で使っても害がない).

一方, 左右が対称でおとなしめの確率分布の標本分布については, かなり小さな n であっても中心極限定理がよく効いてくれて, 標本平均 $\bar{X}$ の標準化 Z の分布の標準正規分布による近似の誤差が小さくなる. しかし, 実践的な状況では真の分散の値は不明なので, それを不偏分散で置き換える近似が必要になる. その近似は式の上では以上で構成した T を使うのと同じことになる.

その場合の T はもとの分布が正規分布ではないので t 分布には従わない. しかし, 標準正規分布でよく近似された分布の分散を適当に確率的に揺らがせることによって得られた分布に従うことになる. だから, その分布をどんぶり勘定で t 分布で近似してやれば統計分析の安全性を高めることができるだろうと考えられる. t 分布を使用することにはこういう意味がある.

注意・警告: 以上で説明したように, データが正規分布の標本分布に従っていなくても, 幾つか条件が揃えば, t 分布を使った統計分析は実践的に意味を持ち得る. しかし, 極めて多くの解説で「正規分布の仮定が成立していない場合には t 分布を使った統計分析法を使ってはいけない」と解説されている. それは誤解である. 他の解説を読むときには注意して欲しい.

8.6 問題: T の計算例

In [ ]:
```
1
```