脳波解析のための数学シリーズ
統計編

後藤 優仁

2021 年 9 月 4 日

目 次

第 1 章　はじめに **3**

第 2 章　確率分布 **4**
2.1　確率分布と確率密度関数 4
2.2　一様分布/ベルヌーイ分布/2 項分布/ポアソン分布 5
2.2.1　一様分布 5
2.2.2　ベルヌーイ分布 7
2.2.3　2 項分布 8
2.2.4　ポアソン分布 9
2.3　正規分布 10
2.3.1　正規分布とは 10
2.3.2　確率密度関数の導出 11
2.3.3　100p%点 14
2.4　t 分布 16
2.5　F 分布 17
2.6　χ^2 分布 18

第 3 章　大数の法則と中心極限定理 **20**
3.1　大数の法則 20
3.1.1　マルコフの不等式 21
3.1.2　チェビシェフの不等式 23
3.1.3　大数の法則 24
3.2　中心極限定理 24

第 4 章　統計的検定 **26**
4.1　統計的仮説検定とは 26
4.2　詳しく説明 30
4.3　t 検定 31

第 5 章　ベイズ統計 **32**
5.1　頻度主義とベイズ主義 32
5.2　加法定理と乗法定理 32
5.3　ベイズの定理 34
5.4　無情報事前分布 38

5.5 共役事前分布 38
5.6 ベイズの応用 38

第 6 章 統計的推定 **40**
6.1 点推定と区間推定 40
6.2 最尤推定 41
6.3 最大事後確率推定 (MAP 推定) 43
6.4 ベイズ推定 44
6.5 逐次推定 45
6.6 推定のまとめ 46

第 7 章 予測 **47**
7.1 ベイズ予測 47

第1章　はじめに

統計処理とは, 実際に実験やアンケートを通して集めたデータを解釈する際に必要になる処理です. 論文を読むためにも必須, というか一番大事な項目なので頑張りましょう. analysis の様々な処理を行った結果，実験で計測した脳波を何らかの形で解釈しやすいようなデータに変形する事が出来ました．つまり解析をしました．次に必要なのは，この解析で得られたデータを解釈する行程です．ここで必要になるのが統計的な処理ですね．個人的には楽しくないですが，仕方ないのでやっていきましょう．

また，結局脳の計算をモデル化しようとか発展的な議論をしていこうと思った時にはここら辺が非常に重要になってきます．先に勉強しておくとだいぶ後が楽だと思いますが，何に役立つのか分からないとモチベは上がらないし勉強できないのは自分も痛いほど分かるので，逆に先に advanced.pdf とかでチラ見してくるのも良いかもしれませんね.

あと検定．正味これがいっちゃん大事なのですがいっちゃん勉強する気になれないんですよねぇ... がんばります.

第2章　確率分布

基礎編に乗っているような基礎統計量だけの議論では必要ないため，(多分) 高校数学ではやりませんが，それ以上の統計学をやる上で必須になってくるのが確率分布の概念です．

世の中には無数の統計的データがある，というか収集できますが，そいつらの統計的な分布，つまりヒストグラムの形は概ねいくつかの典型的なパターンのどれかに当てはまる事が知られています．中でも正規分布という分布はつよつよで，中心極限定理という定理が示すように世の中の多くの事象に適用できる分布の形です．

こいつらを知っている事で，その性質から考えて，実際に得られたデータの平均や分散といった値はどうなのか（推定），そもそも普通にありえる話なのか，何か特殊な事情で一般的ではないデータが取られたのか（統計的検定）や，では次に来る事象はどうなる可能性が高いか（予測）など様々な議論を展開していく事が可能になります．

そのため，確率分布はそれだけで学んだところで特に面白味がありませんが，今後使っていくという意味では重要です．ここでは，代表的な確率分布を載せていきます．覚える必要はないので，どんな奴らがいるのかなんとなく見て，3 章以降で出てきた際に参照できるようにしておきましょう．

2.1　確率分布と確率密度関数

まずは定義の確認から改めて行っておきましょう！

確率とは, 起こりうる全体の事象のうち, ある事象がどれくらいの割合で起こるのかを表します．

確率分布とは, 起こりうる事象それぞれの確率がどのような分布をしているかです．

確率密度関数とは, 実際に確率分布をプロット (横軸に事象, 縦軸に確率) した際に現れる分布の形を表現する関数の事です．場合によっては事象は離散分布になったりするため，その場合はこの分布を連続と見立てて確率密度関数が定義されます．

2.2　一様分布/ベルヌーイ分布/2 項分布/ポアソン分布

まずは簡単な分布から見て, 徐々に複雑なものを見ていきます. 今回は脳の数学であまり使わない気がするやつらはまとめて雑に確認します.

2.2.1　一様分布

例えばサイコロを振った時に出る目の確立分布を考えると, 当然 6 つの事象それぞれがすべて $\frac{1}{6}$ で表せるため, プロットすると以下のようになります.

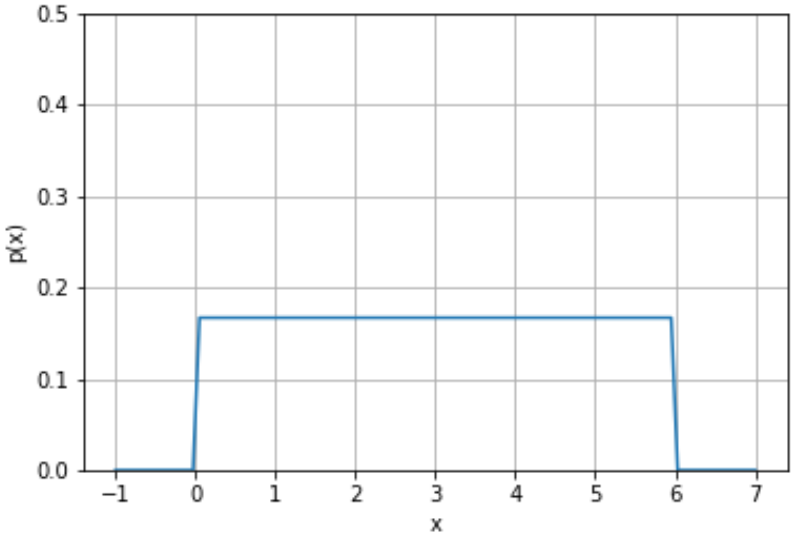

図 2.1: サイコロの確立分布 (一様分布)

一様分布は確率密度関数を定義するまでもないのですが, 勉強のためにあえてやるとすると以下のようになります. まずさいころの例だとそれぞれの目が出てくる確率はこうですね.

$$f(x) = \frac{1}{6}(1 \leq X \leq 6) \tag{2.1}$$

定義域が 0 から 6 で, いずれの点においても $\frac{1}{6}$ の確立で目がでるよってことですね. 簡単です. また確率を考える際, 式 (2.1) のように左辺の (x) と右辺定義域内にいる (X) で使い分けがなされます. X は確率変数を表し, 実際の値は取りません. 観測されて確定した時には x として表現されます. 今回の X と x の関係は

$$X = [x_1, x_2, x_3, x_4, x_5, x_6] \tag{2.2}$$

のような感じです.

より一般には離散一様分布に従う変数の確率関数は, N を変数が取りうる値 (さいころの目=6) として,

$$P(X = x) = \frac{1}{N} \qquad (x = 1, 2, ..., N) \tag{2.3}$$

となります. これはさすがに良いですよね. 次.

離散一様分布の期待値と分散は以下です.

$$\mathbb{E}(X) = \sum_{k=1}^{N} kP(X = k) \tag{2.4}$$
$$= \sum_{k=1}^{N} k\frac{1}{N} \tag{2.5}$$
$$= \frac{(1+N)N}{2}\frac{1}{N} \tag{2.6}$$
$$= \frac{1+N}{2} \tag{2.7}$$

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \tag{2.8}$$
$$= \sum_{k=1}^{N} k^2\frac{1}{N} - (\frac{1+N}{2})^2 \tag{2.9}$$
$$= \frac{1}{N}N\frac{(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4} \tag{2.10}$$
$$= \frac{N^2-1}{12} \tag{2.11}$$

一見複雑?な気もするかもしれませんが, 普通に計算してみればわかるとおもいます. ほら,「1 から 100 までの数の和は?」みたいな問題って工夫して暗算出来ましたよね? どうやるんでしたっけ.

連続一様分布については確率密度関数は以下のようになります.

$$f(x) = \begin{cases} \frac{1}{b-a} & (a \leq X \leq b) \\ 0 & (otherwise) \end{cases} \tag{2.12}$$

これは面積で考えると分かりやすいかも．確率密度関数は積分すると 1 にならないといけません．b-a は定義域なので，つまり面積でいうところの横の長さです．縦の長さはそれぞれの値が観測される確率，つまり $f(x)$ ですよね？

そう，なので定義の $f(x)$ を定義域で積分すると 1 になります．当たり前の事なのですが，確率密度関数の考え方にいまいち慣れていない人ように丁寧に言いました．以下はこのような話は省略します．

とにかく，積分すると 1 になるのが確率密度関数です．
平均 $\mathbb{E}(X)$ 及び分散 $\mathbb{V}(X)$ は

$$\mathbb{E}(X) = \frac{a+b}{2} \tag{2.13}$$

$$\mathbb{V}(X) = \frac{(b-a)^2}{12} \tag{2.14}$$

一応，期待値くらいは証明もしておきます．でもまあ簡単です．

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)dx \tag{2.15}$$

$$= \int_{-\infty}^{a} xf(x)dx + \int_{a}^{b} xf(x)dx + \int_{b}^{\infty} xf(x)dx \tag{2.16}$$

$$= 0 + \int_{a}^{b} x\frac{1}{b-a}dx + 0 \tag{2.17}$$

$$= \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b \tag{2.18}$$

$$= \frac{b-a}{2} \tag{2.19}$$

2.2.2 ベルヌーイ分布

ベルヌーイ分布は, 1 か 0 かです. ある事が起きるか起きないか, 成功するかしないかなどの 2 値分類ですね. 超簡単です！！

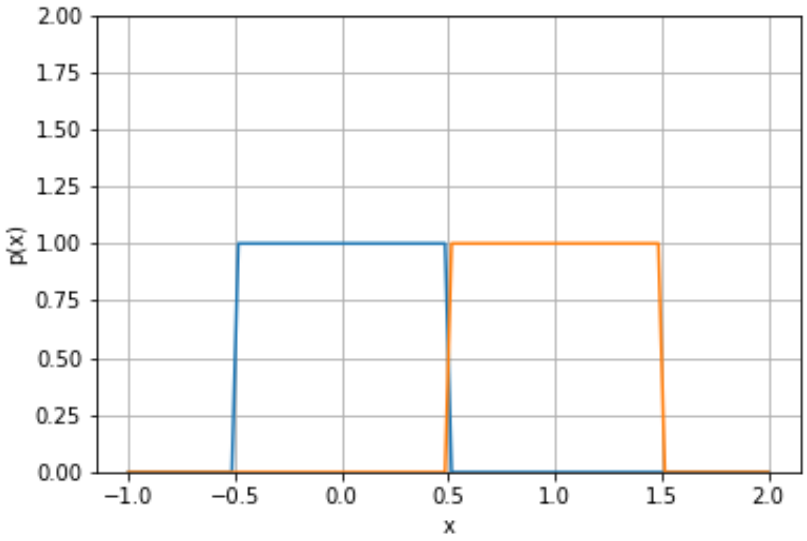

図 2.2: ベルヌーイ分布

今回は 1:1 の図になっていますが, もちろん 1:5 くらいの割合になる事もあります. たとえばサイコロで 1 が出るか出ないかとか. 以上.

関数はこんな感じかな

$$f(x) = \begin{cases} p & (x = 1) \\ q(= 1 - p) & (x = 0) \end{cases} \tag{2.20}$$

平均 $\mathbb{E}(X)$ 及び分散 $\mathbb{V}(X)$ は

$$\mathbb{E}(X) = p \tag{2.21}$$

$$\mathbb{V}(X) = pq \tag{2.22}$$

です. 興味ないので証明とかやりません. 僕も知らない.

2.2.3 2 項分布

2 項分布は「同じことを何回も繰り返した時, ある事柄が何回おこるか」の確立分布です. こいつは式を見れば早いですね.

$$f(x) = {}_n\mathrm{C}_x p^x (1-p)^{n-x} \tag{2.23}$$

式 (2.23) の右辺左側にある ${}_n\mathrm{C}_x p^x$ は, 確率 p で起こる事象 A が, 全体 n のうち x 回でたという意味で, 右側の $(1-p)^{n-x}$ は確率 (1-p) で, A が起きなかった回数が (n-x) 回出たという意味です.

これらを掛け合わせているので, 何回当たって何回外れたかを表す式ですね.

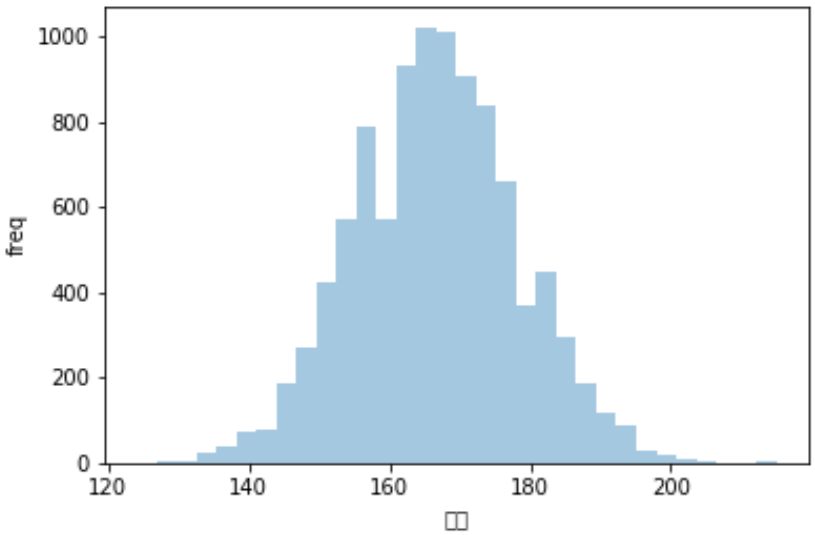

図 2.3: ベルヌーイ分布

平均 $\mathbb{E}(X)$ 及び分散 $\mathbb{V}(X)$ は

$$\mathbb{E}(X) = np \tag{2.24}$$
$$\mathbb{V}(X) = np(1-p) \tag{2.25}$$

です. 興味ないので証明とかやりません. 僕も知らない.

2.2.4 ポアソン分布

ポアソン分布は「稀な事象が一定時間内にどれくらい起きるか」です.

たとえば, ある月にある地域で起きた交通死亡事故の件数とかです. 修羅の国やグンマ―, あるいは名古屋でもない限り, 普通は 0 か多くて 2 件とかですよね.

確率密度関数は以下です.

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda} \tag{2.26}$$

平均 $\mathbb{E}(X)$ 及び分散 $\mathbb{V}(X)$ は

$$\mathbb{E}(X) = \lambda \tag{2.27}$$

$$\mathbb{V}(X) = \lambda \tag{2.28}$$

です. こいつの場合, 平均も分散も同じ定数 λ なので, こいつの値だけで形が決まります. λ がどんな値なのか, どうやって決まるのかは知りません. 使う事があれば足します.

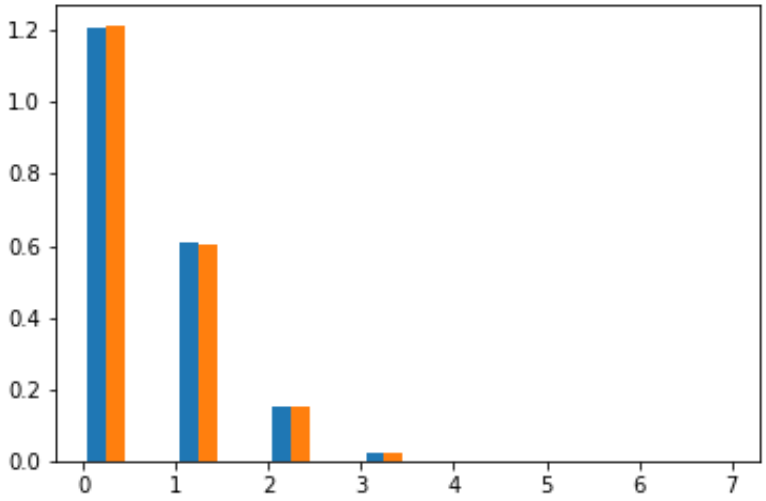

図 2.4: ポアソン分布

2.3 正規分布

2.3.1 正規分布とは

さて本題です. 数ある確率分布の中でも抜きんでて重要な分布, 正規分布, またの名をガウス分布です. 世の中の多くの事象がこの分布に従っていて, また従っていると仮定して統計的処理がされています. 今後の統計学の基礎になるのでしっかり理解しましょう.

正規分布は, 平均 $\mathbb{E}(X)$ が母集団の平均 μ と等しく, 分散 $\mathbb{V}(X)$ も母集団の分散 σ^2 と等しくなる分布で, 確率密度関数は以下になります.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.29}$$

いや, うん. 分かります. やばそうですよね.

ただまぁ, なんでこんな殺意高めな式になるのかは正規分布のグラフを見れば理解できます. 母集団の平均が 50, 標準偏差 20 の正規分布だとこうなります.

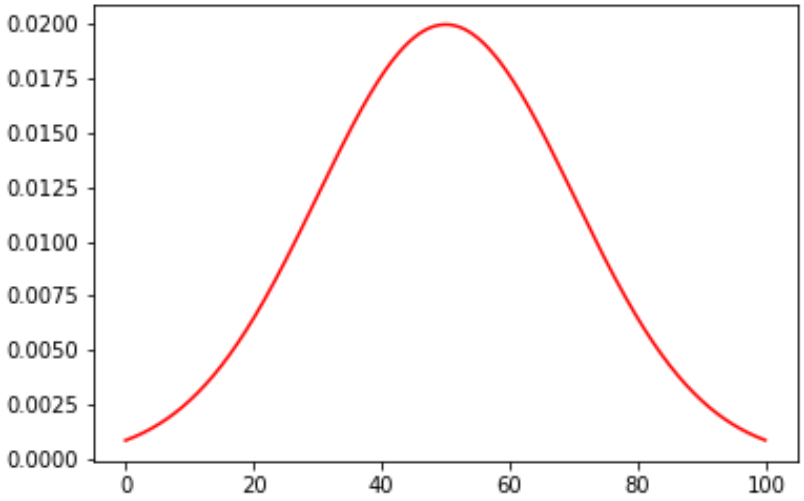

図 2.5: 正規分布

定義通り, 確率変数の平均も 50, 分散も 400(20 の二乗) になっていますね！！これが正規分布です. 綺麗ですね.

2.3.2 確率密度関数の導出

さて, 式 (2.29) の解説です. まずややこしいとこを全て消し飛ばし, 以下のように変形しましょう.

$$f(x) = e^{-x^2} \tag{2.30}$$

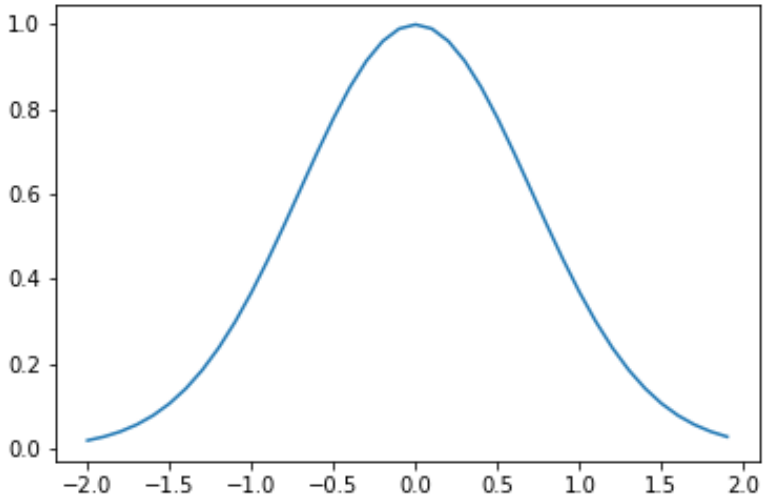

図 2.6: 式 (2.30) の図

この関数の説明はいりませんね？ 指数関数の二乗の負版です.

まず, 世の中の多くの事象は平均値を取る確率が一番大きく, 平均値から離れるにつれその値を取る確率は小さくなることが知られています. んで, これをどう数式で表現すれば良いかと悩んだ末考えだされたのが正規分布関数だと思ってください.

そうすると, とりあえず釣り鐘型の関数が欲しいという事で式 (2.30) を考えました. 実際これでほぼ完成です.

ただ, これは任意定数がないため平均が 0, 分散 (幅) も一定ですね. これでは実用できません. そこで任意定数を導入し, 平均値と分散を可変にします.

まずは平均値, つまりこのグラフの頂点の位置を動かします.

$$f(x) = e^{-(x-\mu)^2} \tag{2.31}$$

二次関数で高校の時にやりましたね.

次に, 分散... つまりこの釣り鐘の幅というか広がり方を変えます.

$$f(x) = e^{-\frac{(x-\mu)^2}{\sigma^2}} \tag{2.32}$$

平均に比べて一寸難解かもしれません. σ, つまり標準偏差で割る事で, σ の値が小さければ鋭く, 大きければ扁平なグラフになります.

ただ, σ は正負が定まらないため, 二乗して分散の形にする事で符号を一定にするわけです.

ともあれ ... これで, 母平均と母分散によって形を変える釣り鐘型分布が完成です！！めでたい！！

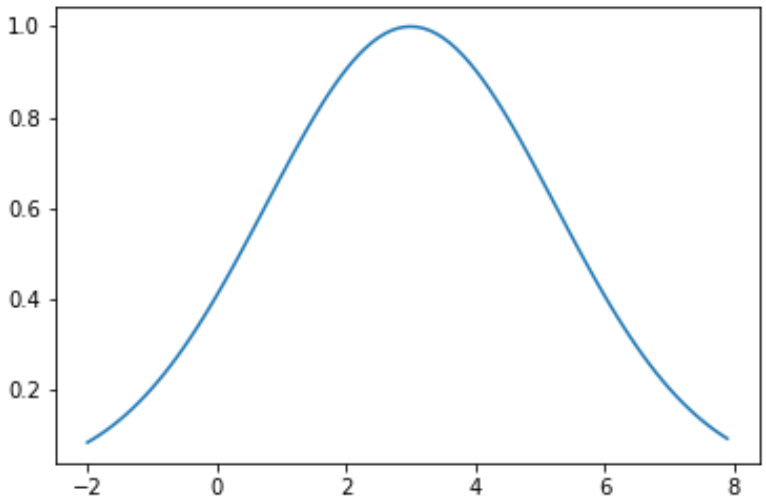

図 2.7: 式 (2.32) の図.(平均 3, 分散 10 の場合)

え？式 (2.29) と違うじゃないかって？
せっかちだなぁ..... 余裕がない人間はモテないですよ.

まあ, グラフの形は実際これで完成なのです. ただ, 今我々が求めていたのは正規分布の「確率密度関数」でしたね？

確率の密度なのですから, 当然合計して 1 にならないといけないわけで, 先程の式 (2.32) ではその点がダメなのです. ちゃんと全確率点での値を足して 1 になるように正規化する必要があります.

つーわけで積分方程式を解きますがその前に, 出てくる計算結果を綺麗にするために式 (2.32) に細工をし, σ^2 を $2\sigma^2$ にしておきます. 2 が付きましたが, グラフの性質自体は変わりませんね？

では積分方程式.

$$\int_{-\infty}^{\infty} ce^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1 \tag{2.33}$$

を解きます... 出来たものがこちらです.

$$c = \frac{1}{\sqrt{2\pi\sigma^2}} \tag{2.34}$$

式 (2.34) で得られた c を改めて式 (2.32 の変形版) に代入すると

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.35}$$

が出てきましたね！良かった良かった. 以上が正規分布の確率密度関数の導出工程でした. フーリエなんかに比べれば雑魚ですね！ワンパンでした！

え？積分方程式の解き方ですか？なんか MATLAB にお願いしたら解いてくれました...

2.3.3 100p%点

次. 正規分布を学ぶ意味は, この概念があるからといっても過言じゃない重要な性質です. 正規分布は平均が丁度真ん中で, 広がり方は分散によって定義される左右対称な特殊な分布でしたね？

つまり, 平均の周辺であるほど高確率で観測され, 裾野ほど「レア」な事象というわけです.

この性質を利用して, 正規分布には「100p%点」と呼ばれる指標を導入する事が出来ます. これは「その点以上 (以下) の部分の確率の合計が p になる境界」を指します. x 軸の正方向なら上側, 負方向なら下側, 両方を指すなら両側 100p%点です.

言葉だとややこしいですが, グラフで見れば一瞬で分かります.

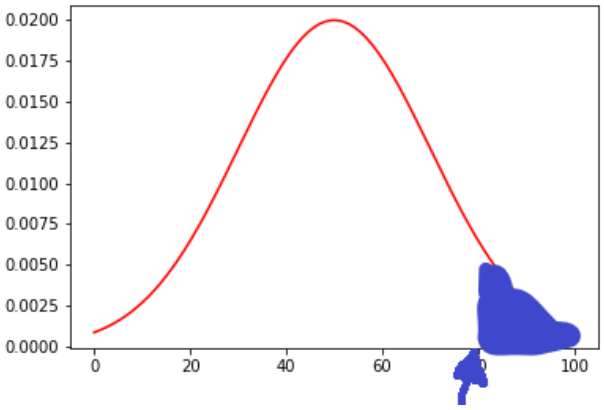

図 2.8: 正規分布の上側 5%点

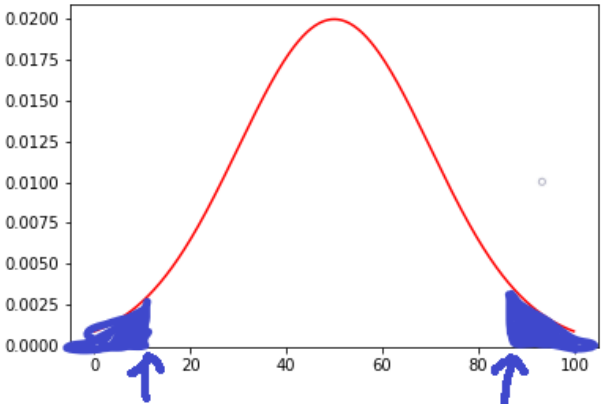

図 2.9: 正規分布の両側 5%点

上側 5%に比べ, 両側 5%は左右に領域が分散した分上側の領域が狭くなってるのに注意です.

だいたい, 正規分布でよく用いられるのは上下両側の 5%と 1%点です. 何に使うのかはあとで説明するので, ひとまず概念だけ覚えておいてください.

2.4　t 分布

少ない標本数をもとに母分散がわかっていない母集団の母平均推定に使われるのが t 分布です. 詳しくは推定の項で触れるので, ここでは確率密度関数と性質の確認をします.

$$f(x) = k(1 + \frac{x^2}{\alpha})^{-\frac{\alpha+1}{2}} \tag{2.36}$$

式 (2.36) が t 分布の確率密度関数です. ここで k は定数, α は自由度という指標で, この形を「自由度 α の t 分布」と表現します. 自由度とは何かはここでは説明しませんが, 母集団によって算出できる値です. それ以外の部分では正規分布に似ていますね. 実際, 自由度 α が十分に大きい場合には正規分布になります. ただこいつの平均と分散は

$$\mathbb{E}(X) = 0 \tag{2.37}$$

$$\mathbb{V}(X) = \frac{\alpha}{\alpha - 2} \tag{2.38}$$

です.

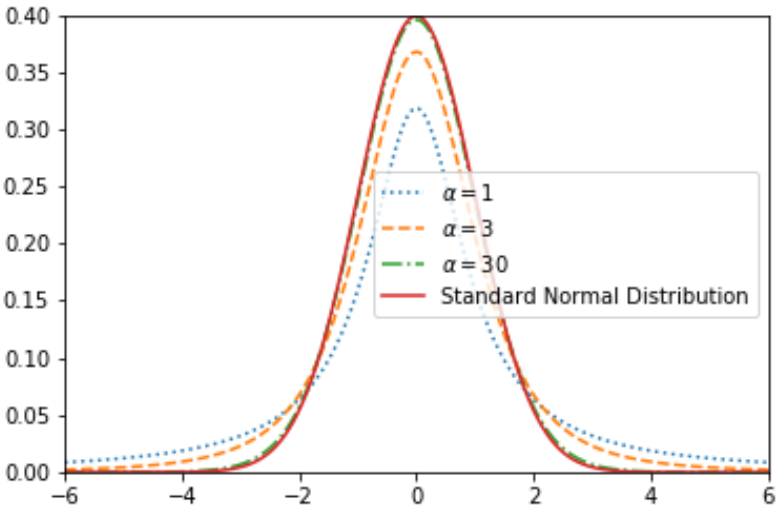

図 2.10: 自由度の異なる t 分布と標準正規分布

自由度の異なる t 分布と標準正規分布の比較です. 標準正規分布とは, 正規分布のうち平均が 0, 分散が 1 のやつです. α の値が大きいとほぼ同じ形になる事が分かると思います.

では何に使うのかですが, そこはひとまず置いておきます. とにかくこういう分布があるのです.

2.5　F 分布

F 分布は, 少し特殊な形をしていますが非常に重要です. どう重要かというと F 分布はこれまでの分布と異なり, 2 つの母集団から得られた標本分散の比の確率分布になります. これの何がすごいのかというと, 例えば図 (2.5) は自由度が 3 の母集団 A と自由度が 10 の母集団 B の F 分布ですが, 「普通は」この組み合わせだと 0 から 2 あたりの値になる事が多いのです.

しかしもし今, 自由度 3 と 10 のとある母集団 A,B 間で F 値を取ったら 6 とかが出たとします.

通常はありえない値, つまり統計的に考えると「自由度だけでなく分散そのものに差がある = A,B の母集団は異なる」となるわけですね. こちらも詳しくは後程. 式はえぐえぐです. 震えろ.

$$f(x) = \frac{kx^{\frac{m}{2}-1}}{\{1+(\frac{m}{n})x\}^{\frac{m+n}{2}}} \tag{2.39}$$

こいつの導出はいいです. 僕もわからん. 気が向いたら勉強してまとめるかも. 例によって k は定数, m,n は標本集団 2 つのそれぞれの自由度です. こいつは「自由度 m,n の F 分布」と表現されます. 今はそれだけ. 続いてグラフです. 幸いこっちは簡単.

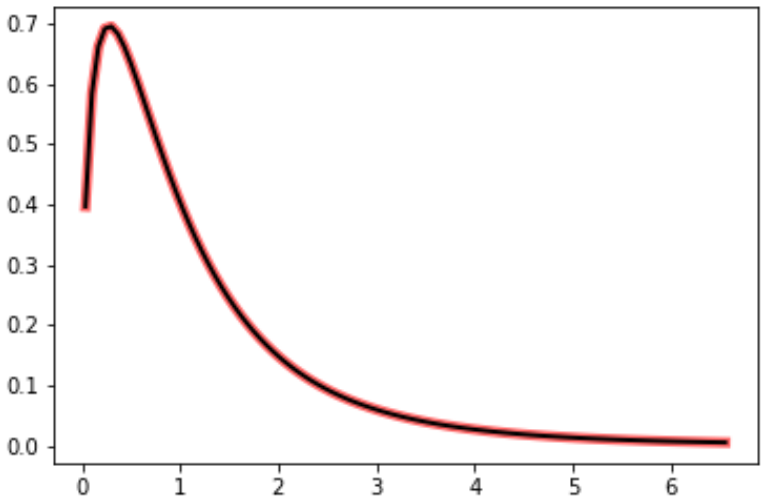

図 2.11: 自由度 3.10 の F 分布

恒例の平均値と分散です. こいつらも殺意高めだけど覚える必要は今のところ皆無だと思ってます.

$$\mathbb{E}(X) = \frac{n}{n-2} \tag{2.40}$$

$$\mathbb{V}(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \tag{2.41}$$

2.6 χ^2 分布

ラスト. χ^2 分布です. こいつも重要です. 我々は結果を理論に落とし込む時, なんらかの数式に近似したり回帰したりするわけですが, 実際問題理論と観測値が完全に一致するなんていう事はまれで, 様々な要因で若干の誤差が生じます. その誤差が「誤差」なのか「理論の誤り」なのかを評価する時に使うのがこの分布とそれに基づく検定です.

式はこちらです.

$$f(x) = k\chi^{\frac{\alpha}{2}-1}e^{-\frac{\chi}{2}} \tag{2.42}$$

またえぐえぐですね. えぐえぐと言えば, 筆者は声優の江口拓也さんが「やはり俺の青春ラブコメは間違っている」のラジオでやっていた「ぼっちラジオ」が大好きです. どうせこんな数学の同人誌読んでる人は陰キャだし, 楽しめると思います. 是非 YouTube で検索してみてください.

さて, 例によってここの α は自由度, k は定数です.

幸い, どこの説明を見てもこの式は使わないと言われているので覚えなくていいでしょう.

平均値と分散です.

$$\mathbb{E}(X) = \alpha \tag{2.43}$$

$$\mathbb{V}(X) = 2\alpha \tag{2.44}$$

簡単すぎワロタ ww ついでグラフです.

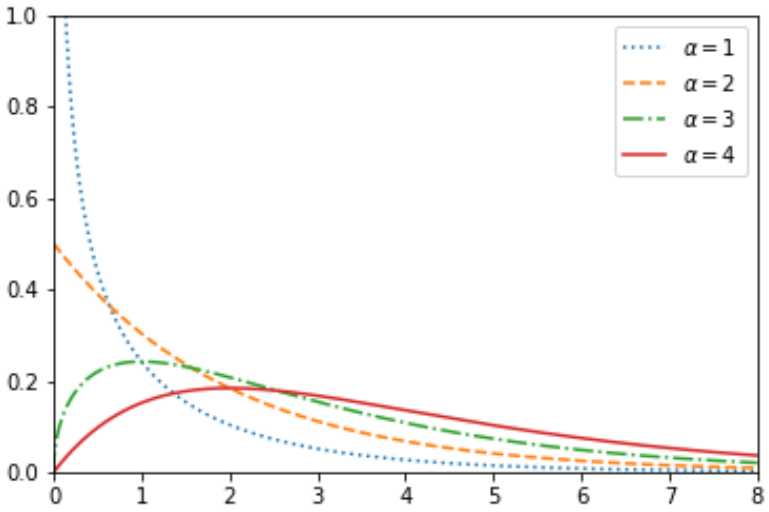

図 2.12: 自由度の異なる χ^2 分布

こいつは結構 F 分布に似てますね. ただ注意が必要なのは, F 分布は標本分散の比に関する分布ですが χ^2 分布は標本分散そのものの分布です.

第3章　大数の法則と中心極限定理

さて，こうして様々な確率分布の形を見てきたわけですが，そもそも全体のデータを観測して，そいつらの分布を見るならまだしも未知の集団の中から一部を抜き出して分布を見たところで，そいつの性質 (平均とか) をどれだけ信じて良いものでしょうか.

3.1　大数の法則

よくある問題が，政党支持率を街頭インタビューやてきとうな電話で 100 人や 1000 人に聞いたデータを元に算出しているのはどれだけ信用できるのか問題とかですよね.「あんなたった 100 人に聞いただけじゃ信用できん！印象操作だ！」みたいな話になりかねません．そこで出てくるのが，大数の法則です.

今，コインを n 回投げて，その表が出る確率がちゃんと 1/2 なのかを確かめる作業を考えます．i 回目の試行で表が出た場合 1，裏の場合 0 を取る確率変数 x_i を考えると，n 回投げた時の表が出た回数は

$$r = \sum_{i}^{n} x_i \tag{3.1}$$

この量を頻度と呼び，更にこれを試行数 n で割った r/n を相対頻度と言います．r は確率変数なので，二項分布

$$f(x) = {}_nC_x(p)^n \tag{3.2}$$

に従います．こいつらの期待値と分散は二項分布のとこで確認したように

$$E(r) = np \tag{3.3}$$
$$V(r) = np(1-p) \tag{3.4}$$

である事が分かります．相対頻度の期待値と分散は

$$E(r/n) = np/n = p \tag{3.5}$$
$$V(r/n) = p(1-p)/n \tag{3.6}$$

となる事も分かると思います．で，こいつの試行数 n の値を変えた時，E(r/n) の値がどう変化していくのかを見てみましょう．直観的，というか常識的に，n=3 とかみたいに極端に試行数が少ないと信用できない事が分かるかと思います．逆に，n=1000000 とかといった膨大なデータがあれば，そいつらがちゃんと分布の真の期待値である p に収束する事が分かるでしょう．で，もしここで n=100 とかで見ても全然収束してないのであれば，街頭インタビューで 100 人のデータを用いて結論付けられた発表は信じるに値しなさそうですね．

では，コイン投げをしていきます．表が出る確率はちゃんと 1/2 としておきます．ここで，試行回数を変化させていった時それぞれに対応する，r/n の 0.5 付近 (± 0.1) の値を見ていきます．つまり，ちゃんと真の期待値に迫れているかという事を評価します．

$$P(0.4 \leq r/10 \leq 0.6) = 0.656 \tag{3.7}$$
$$P(0.4 \leq r/20 \leq 0.6) = 0.737 \tag{3.8}$$
$$P(0.4 \leq r/50 \leq 0.6) = 0.881 \tag{3.9}$$
$$P(0.4 \leq r/100 \leq 0.6) = 0.968 \tag{3.10}$$

はい，という事で，n が増えるにつれて確率が上がっていき，真の確率 (表が出る確率) に迫っている事が分かります．実際，もう 100 回もやれば十分ですね．というのが感覚的な説明でしたが，これを数式で表現すると以下のようになります．

$$P(|r/n - 0.5| > 0.1) \longrightarrow 0 \qquad (n \to \infty) \tag{3.11}$$

相対頻度が真の値である 0.5 から 0.1 以上離れている確率が 0 に収束していく過程，ですね．これこそが大数の法則 (の一例) です．

では，より一般に大数の法則を導いて行きましょう．そのためには，チェビシェフの不等式というものが必要で，さらにそいつはマルコフの不等式をベースに考える必要があります．ということでまずはマルコフの不等式から行きましょう．

3.1.1 マルコフの不等式

マルコフの不等式
任意の確率変数 X と $a > 0$ に対して

$$P(|X| \geq a) \leq \frac{E[|X|]}{a} \tag{3.12}$$

これがマルコフの不等式です．よく分からんと思うのでまずは証明しましょう．確率変数 X に対して，ある範囲として a がでてきてるので X の期待値 $E[|X|]$ は以下のように書けます．

$$\begin{aligned} E[X] &= \int_0^\infty xf(x)dx \\ &= \int_0^a xf(x)dx + \int_a^\infty xf(x)dx \end{aligned} \tag{3.13}$$

式 (3.13) のうち，第一項を意図的に消した場合，下のように変形されますね．X は絶対値を取られるやつなので，第一項は必ず正の値を取っていることに注意です．

$$E[X] \geq \int_a^\infty xf(x)dx \tag{3.14}$$

ここで，x が取りうる範囲は a から ∞ までとなっているので，この中で最小の a ですべての x を置き換えてみます．すると

$$E[X] \geq a\int_a^\infty f(x)dx \tag{3.15}$$

$$\therefore E[X] \geq aP[X \geq a] \tag{3.16}$$

となります．何をやってるのかちょっと分からなくなってくるかもしれませんが，まず a から ∞ で f(x) を積分してるんだから a 以上の値をとる確率密度関数ですよね．で，こいつの a 倍をしているものなので，$aP[X \geq a]$ で書き換えられるわけです．したらこの式を移項して生理してあげればマルコフの定理，即ち

$$P[|X| \geq a] \leq \frac{E[|X|]}{a} \tag{3.17}$$

がでてきます．証明終わり．

では肝心の，この不等式が何を指しているのかですが，まあ実際数字を入れてみれ考えれば分かると思います．さいころで考えてみます．出る値は 1 6 なので，期待値は 3.5 です．この時，a をたとえば 1 にした時，1 より大きい値を取る可能性は当然高く，それは右辺が 1 で割られている事からも明らかです．が，3 とかになると割った結果がほぼ 1 になりますね．3.5/3 だから．雲行きが怪しくなってきました．さて，ここに 5 なんかを入れると，もはや分母の方が大きいので，値は 1 よりも小さくなります．一方の左辺も，a の値が大きくなるにつれて取れる値の範囲が狭くなっていくので，それらの値を取る確率もやはり下がっていきます．

この関係を表すのがマルコフの不等式でした．

3.1.2 チェビシェフの不等式

ふう．マルコフの不等式が分かれば，チェビシェフの不等式なんてワンパンです，ワンパン．秒で沈めてやりましょう！

チェビシェフの不等式
期待値 $E[Y] = \mu$, 分散 $V[Y] = \sigma^2$ とするとき，任意の実数 $k > 0$ に対して

$$P[|Y - \mu| \geq k\sigma] \leq \frac{1}{k^2} \tag{3.18}$$

が成り立つ

めちゃくちゃ見覚えありますね．そう，これはマルコフの不等式の特別な場合を考えればすぐ証明できます．

まずマルコフの不等式において，$X = (Y - \mu)^2, a = k^2\sigma^2$ とすると

$$P[(Y - \mu)^2 \geq k^2\sigma^2] \leq \frac{E[(Y - \mu)^2]}{k^2\sigma^2} \tag{3.19}$$

$$\therefore P[|Y - \mu| \geq k\sigma] \leq \frac{1}{k^2} \tag{3.20}$$

となります．ただしここで，$E[(Y - \mu)^2] = \sigma^2$ を使っています．

さて，こいつの何がすごいのかを考えてみます．結論からすると，チェビシェフの不等式がそのまま大数の法則を示すものになっています．というのも，まず式 (3.18) を見てみると「得られたデータが，$k\sigma$ 以上離れた位置，正規分布とかで考えるならめっちゃ裾の方ですね，を取る確率」である左辺は，k^2 の逆数以下になるよって話ですね．

たとえば，あとで出てきますが正規分布の時によく使う 3σ 以上の値を取る確率は 0.3%程度なのですが，実際この値 $(1/k^2)$ は $1/9$ なので 0.3 以下になりますね．

すごいのは，得られたデータがどんな分布に従っていたのかという母集団の情報が全くなくても，こいつの収まる範囲が分かる事です．どんな分布でも絶対，$1/k^2$ 以下になるってわけですね．正規分布の時実際そうなるのは上で確認した通り．めっちゃすごいっすねこれ．分布知らなくても言えるわけですからね！！

3.1.3 大数の法則

で，肝心の大数の法則です．証明していきます．チェビシェフの不等式の証明において，k=a^2 としてみると，

$$P[|X-\mu| \geq a] \leq \frac{\sigma^2}{a^2} \tag{3.21}$$

となります．これも計算自体は全く同じなので，チェビシェフの不等式と呼ばれます．ここで標本平均 $\bar{X}$ を考えると，$E[\bar{X}] = \mu, V[\bar{X}] = \frac{\sigma^2}{n}$ となります．これを式 (3.21) に代入すると

$$P[|\bar{X}-\mu| \geq a] \leq \frac{\sigma^2}{a^2 n} \quad \xrightarrow{n\to\infty} 0 \tag{3.22}$$

となると思います．良いでしょうか．n の値が大きくなるにつれ，右辺の分母が大きくなるので値は 0 に収束していくわけですね．そう，これが大数の法則です！！

n の値，つまり得られた標本の数 ($\bar{X}$は$\sum X_i/n$) が多くなるほど，真の値との誤差は 0 に向かって小さくなっていくことを指しています．

その主張することは何かというと，こうです．

「大標本では，観測された標本平均を母集団の真の平均とみなしてよい」

これがあるからこそ，我々はある程度のサイズのデータを使って一般化した議論が出来るわけですね．しっかり確認しておきましょう．

3.2 中心極限定理

大数の法則同様，統計学においてとても重要な定理としてもう一つ，中心極限定理があります．というかこちらの方がどちらかと言えば大事かも．定理は以下です．

中心極限定理
期待値 μ，標準偏差 σ の分布に従う確率変数列 $X_1, X_2 ... X_n$ に対し，$S_n := \sum_{k=1}^{n} X_k$ とすると

$$P(a \leq \frac{S_n - n\mu}{\sqrt{n}\sigma} \leq b) \xrightarrow{n\to\infty} \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \tag{3.23}$$

が成り立つ．

はい，なんも分からんと思います．

ここで言いたいのは，与えられたデータから求めたそれぞれの標本平均 X の総和である S_n を標準化 (左辺の真ん中のごちゃごちゃしたやつ．平均を引いて標準偏差で割ってる) したやつは n が十分に大きいと，つまり得られるデータが多いと，期待値 0，分散 1 の正規分布に収束するということです．

正規分布の確率密度関数は以下でしたね．

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.24}$$

ここで μ を 0，σ は 1 を代入した形が中心極限定理の右辺に来ています．こいつを a から b の範囲で積分したものが，標準化変数である左辺の項が a から b に収まる確率と同等になるという事を表していますね．

即ち，n が十分に大きいなら部分和である S_n は平均が $n\mu$，分散が $n\sigma^2$ の正規分布に従う事になり，よって標本平均である $\bar{X_n} = (X_1 + ... + X_n)/n$ も平均 μ，標準偏差 $\sigma/\sqrt{n}$ に従う事になります．

大事なのは，今回定理のところでそもそも，母集団の分布を何も仮定していないところです．大数の法則同様，母集団の分布に関わらず従う法則なわけですね．

母集団がどんな形であれ，n が大きい時にはそこから得られた標本平均は正規分布に従うという性質を意味します．

ゆえに，正規分布は数ある確率分布の中でも最も重要な分布と言われるわけですね．今後学んでいく様々な検定や推定なども，正規分布を仮定する事が多いのはこれが理由になっています．

(証明... モーメント関数とか必要で自分がちゃんと整理しきれてないのと，どこにその説明いれるべきなんや... てか必要かな... と悩んでしまっているのでとりあえず保留！ごめんなさい！！)

第4章　統計的検定

今の今まで，面倒だなあ面白くないんだよなあと逃げに逃げまくっていましたが...「はやく検定書けよ使えねえオタクだな」って色々な人に尻を蹴りあげられたので，その勢いで腰をあげて頑張ろうと思います．

しかしまぁ，最近の神経科学というか実験心理学では統計の問題が色々と指摘されていて，日々 Twitter で神経科学のキラー細胞みたいな御仁が色々な人の研究を晒上げて燃やし尽くしている状況なので，統計の記述をするのマジで怖いですね... 自浄作用を働かせるのは大事だけど，それはあくまで界隈の発展のためであって他者を攻撃して気持ちよくなるためのものではないと思うのでもう少し考えてみてほしいものですね．
さて，がんばるぞい

4.1　統計的仮説検定とは

統計的仮説検定とは何か，いつ使うものなのかですが，結論から言えばどの論文にもほぼ必ず出てくる作業ですよね．研究者の掲げる「ある仮説」の有意性を検定するものです．「ある仮説」の元で期待する結果と，実際に得られて観測された結果とを見比べた時，そこにある差が偶然起こった程度のものだと言えるか，否かを評価します．偶然じゃないのなら，きっとその差には何か意味があるはず，つまり仮説が間違っているという事になります．

これを使って，たとえば 1 軍と 2 軍の能力は同等であるという仮説の元に能力テストを行って，結果を比べると明らかに 1 軍の方が高かったとなれば，残念ながら仮説は間違っていて 1 軍より 2 軍の方が優秀であったんだ！のような主張を出来るわけですね．

この発想自体は，我々人間が日頃からやっているようなごく普通の思考プロセスですよね．

与えられた仮説のもとで考えられる範囲を超えた誤差が生じた場合，それは何らかの意味があると考え，このずれは「有意である」，と言います．ここから，統計的仮説検定とは即ち仮説の有意性検定であるとも言えます．

ではここで問題なのが，どうやってその有意性を判断するのかです．有意性とは，標本が有意なずれを示す確率であるので，ここで標本分布を使う事になります．コインを 10 回投げて，表が何回出るかを

使ってコインがいかさまコインじゃないかを判定する事を考えましょう．一般のコインであれば，確率的に平均を取れば 5 回表が出るはずですね．

5 回表が出たら，普通の人ならイカサマコインではない判定をすると思います．では 4 回ならどうでしょう？たった 10 回の試行中なので，僕なら良しとします．では 3 回は？ここらへんから少し怪しいですね... 僕は何とも言えないなぁくらいの評価です．2 回まで行くと，ちょっとイカサマを疑っちゃいますね．1 回とか 0 回は舐めてる．許せんわ．

と，普通に我々が行う思考はこんな感じですね．表が出る確率 p は $p(\text{表}) = 1/2$ であると考えて，二項分布 $(Bi(10, 1/2))$ の確率分布に従う事を仮説として，その中で与えられたデータ (表が出た回数) が分布のどのあたりにあるのかを考えるわけです．

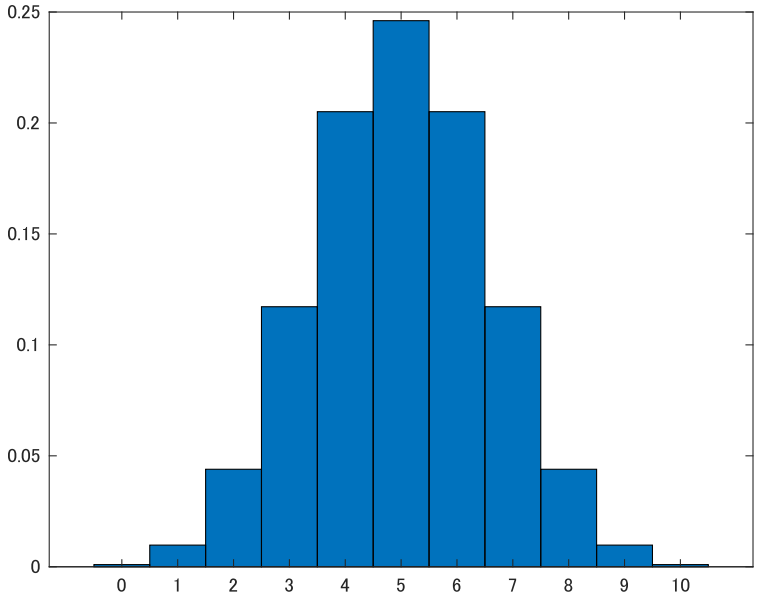

図 4.1: 表が出る回数の二項分布

確率の議論なので，ここでさっきの例をちゃんと計算してみましょう．n=10 の 2 項分布で表が出る回数別にそれぞれの確率を計算すると以下のようになりました．

$$
\begin{aligned}
x = 5 \qquad & p = 0.2461 \\
x = 4 \qquad & p = 0.2051 \\
x = 3 \qquad & p = 0.1172 \\
x = 2 \qquad & p = 0.0439
\end{aligned}
$$

こう見ると，やはり 3 あたりから怪しくて，2 はやばいってのがなんとなく分かりますね．2 回だけ表が出る確率は二項分布の元でわずか 4.3%しかないわけですからね．

この時，仮説の元では出るはずがない結果がでたということで，仮説が誤っていたと判断します．この事を，仮説を棄却する，と言います．

また，実用では基本的にこの仮説は否定したい，というか否定される事を想定して設定されるものです (研究であれば，Ａグループとｂグループに差があると言いたいから，差がないという仮説を立てて否定しますね)．故に，こうして否定される仮説の事を特に，帰無仮説とも言います．

反対に，対立的な仮説（コインの例なら，表が出る確率は 1/2 未満である，とか）を立てて置くこともあります．この仮説を特に，対立仮説と言います．

帰無仮説と対立仮説は互いに反する事象なので，帰無仮説を否定する事は対立仮説を採択する事とほぼ同義，というふうにして使われる事が多いですね．がしかし，注意が必要なのは帰無仮説が $p = 0.5$ だとして，対立仮説に $p < 0.5$ を置いたとしたら第三の可能性として $p > 0.5$ もあります．こうした場合は帰無仮説の否定＝対立仮説の採択とは違うので気を付ける必要があります．

以上の内容を言い換え，改めて仮説検定とは何か考えると，
「用意された仮説が有意であるか否かを判断し，それに応じて仮説を棄却するか採用するか判定する作業」
であると言えます．

ここまでの内容で，ある程度くらいの脳を持った人なら気付くと思いますが，確率やらを持ってきてそれっぽい空気を出してはいるけど結局最後の有意性の判定が恣意的でしたよね．さっきは「わずか 4.3% しか起こりえないはずの事象を観測するのはおかしい」と決めつけて，コインはイカサマであると判断しました．

その時々の気分で勝手に変えるのも困るので，こうした有意性の判定のためにある基準値を設ける必要があります．それが有意水準 (α) と呼ばれるやつです．

ほら，よく論文とかでも見る $p < 0.05$ とかのやつです．これは p 値が 0.05 という有意水準を下回っていますよ，つまり有意ですよ，だから帰無仮説を棄却しますよ，という意思表示なわけですね．一般に，有意水準は 0.1,0.05,0.01 などが使われます．

しかし有意水準を下回ったからといって，絶対にありえないわけじゃありません．5%の確率でしかありえない事は，20 回に 1 回程度は起こりうる事でもあります．そのため，帰無仮説や対立仮説をそれぞれ採択したけれども，実際は誤りで他方が正解であったという事が起こってしまいます．これらの誤りをまとめると以下のようになります．

		判断	
		差あり	差なし
真実	差あり	正しい判断	**第２種の過誤**
	差なし	**第１種の過誤**	正しい判断

図 4.2: 統計的仮説検定の結果と過誤の対応

帰無仮説を棄却したけど実際は偶然で 5%とかの確率を引き当ててしまっていた場合を第一種の過誤，帰無仮説を採択したけど実際は仮説が間違っている場合を第二種の過誤と言います．

また，第一種の加護は偽陽性，第二種は偽陰性とも言います．個人的にはこっちのが分かりやすくて好きです．
もう一つ注意が必要なのは，統計的仮説検定は帰無仮説が否定される事によって対立仮説を採択するといういわゆる背理法を用いているので，厳密には帰無仮説が棄却されなかったからと言ってそれが正しいという事にもならないし，対立仮説が間違っているとも言えません．あくまで，帰無仮説と得られた結果が矛盾しない，というだけにとどまります．

以上の性質を踏まえ，検定の用いる目的として主に 3 つの状況が考えられます．

- 帰無仮説を反証し，棄却する目的
- 異常の検知
- 数学的に扱いやすいためおく便宜的な仮定

まず一つ目，これが最も我々が使うものかもですね．先に説明したように背理法的な使い方をして，データに見られる誤差が確率的に偶然で許される範囲かを考え，逸脱していた場合は帰無仮説を棄却して対立仮説を採択するやつ．新薬の効果があるって主張したいから，まずは薬を投与してもしなくても変わりませんよって帰無仮説を設定してそいつを論破してやろうって使い方です．

次に二つ目，異常の検知です．こいつも検定の例だとよく出てきます．本書でも書くと思うけど，工場の生産品の不良品率の話が例によくでてきますね．通常の状態であれば観測されるデータが属しているはずのデータ群として帰無仮説を設定しておいて，こいつが棄却されないならヨシ，されてしまうのなら何か異常があるとするやつです．

最後がちょっと特殊ですが，統計モデルを考える際に使うものぽい．確率変数が正規分布に従うって仮定をいろんなところで置くと思うんですけど，まあこれは中心極限定理があるからある程度信用でき

るとは家やはり分からない事なので，あくまで仮定は仮定，仮説です．なので，仮説検定を使ってちゃんとデータがこの分布に従っているのか，つまりここで置いたモデルは妥当なものなのかを検証する際にも使うことが出来ます．便利ですね．

4.2 詳しく説明

と，ここまでの内容は皆さん大体，既に知ってると思います．個人的な話になりますが，統計を勉強する上で特に検定でやる気が起きないの，とにかく数式での説明がないからよく分からないというかイメージしきれないってところなんですよね．という事で，以上の内容を数式使って表していきます．

まず，検定の問題では帰無仮説と対立仮説を置いていました．こいつらは相反するべきものなので，以下のように定義します．

$$\Theta = \Theta_0 \cup \Theta_1, \Theta_0 \cap \Theta_1 = \emptyset(\text{空集合}) \tag{4.1}$$

$$H_0 : \theta \in \Theta_0 vs. H_1 : \theta \in \Theta_1 \tag{4.2}$$

ここで，式 (4.1) では Θ は母数空間であり排反な部分集合である Θ_0, Θ_1 に分けられています．こいつらがそれぞれ帰無仮説と対立仮説に対応する母数の集合です．ここで，未知のパラメータである θ がそのどちらに属しているかを表すのが式 (4.2) あり，これが帰無仮説と対立仮説 (H) に相当します．

例で考えます．ある工場の製品の不良生産品率を p としたとき，それが許容範囲かどうかを検定する問題で, p がある p_0 以下であればヨシ，それを超えている場合には生産工程に問題があるとして責任者を減給する事にしましょう．さらに，一般に帰無仮説は Θ_0，対立仮説は Θ_1 とする事が多いようなので従うと，ここで考える検定問題は

$$H_0 : p \leq p_0 \quad vs. \quad H_1 : p > p_0 \tag{4.3}$$

となりますね．閾値として設定した p_0 を以下だったら有意さなし，セーフ，許すとして，超えていたら対立仮説を採択，この工場は問題ありとするわけです．

それから前節で説明を忘れていた大事な概念に，両側検定と片側検定があります．これは検定問題の式を考えると分かりやすいですが，

$$H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta \neq \theta_0 \tag{4.4}$$

のように，特定の値を決めてそれぴったしなのが帰無仮説，そうでない場合を対立仮説とする場合を両側検定と言います．つまり，分布の右でも左でも関係なく，とにかく帰無仮説で定めた値から逸脱した範囲の場合対立仮説を採択するものです．

一方で，

$$H_0 : \theta \leq \theta_0 \quad vs. \quad H_1 : \theta > \theta_0 \tag{4.5}$$

のように（あるいは不等号が逆でも），決めた値に対してどちらかに超えた場合を検出するものを片側検定と言います．たとえば，ある薬の治療効果を確かめたいという目的で検定をするなら，なんらかの健康の指標が薬の投与によって上昇すれば良いので，上側の片側検定で良さそうですね．このように目的に応じて検定は使い分けられます．

で，我々？検定を使う人？がやる決定 d は，H_0 が正しいのか H_1 が正しいのかを判断する事です．帰無仮説 H_0 を採択する事を d=0，棄却する事を d=1 としておきましょう．

すると決定の結果生じる損失について次のようにまとめられます．これを使って偽陽性や偽陰性について考えていきます．

$$\begin{aligned} L(\theta, d=0) &= \begin{cases} 0, & if \quad \theta \in \Theta_0 \\ 1, & if \quad \theta \in \Theta_1 \end{cases} \\ L(\theta, d=1) &= 1 - L(\theta, d=0) \end{aligned} \tag{4.6}$$

損失関数については統計編で触れてないと思いますが，まあ「状況」と「判断」を変数としたときの，間違いによって生じる「損失」についての関数です．判断があってれば損失は生まれないし，間違っていれば生まれます．

ここでは，偽陽性や偽陰性の時に損失が生じれば良いわけですね．たしかに式 (4.6) では正解が Θ_0 なのに d=1，つまり $\theta \in \Theta_0$ を否定しちゃったり，逆に Θ_1 なのに d=0 として $\theta \in \Theta_1$ を否定しちゃったりしたときに 1 を取る，つまり損失が生まれる事になっていますね．

そう，偽陰性や偽陰性とはつまり，こうして判断の結果損失が生まれてしまう状態を指します．より一般には，偽陽性と偽陰性はどちらも同じ損失である必要はなく，偽陽性よりも偽陰性の方を厳しくみたいとか色々あると思いますが，ここではとりあえず等価としています．

4.3 t 検定

第5章　ベイズ統計

5.1　頻度主義とベイズ主義

これに触れるのは気が引けるというか，ややこしいのですがどの統計本もこの話題から触っているので一応．確率を考える方法には，どうやら “主義” があるようです．確率や統計はその性質上，絶対的な正解というものは存在しません．「こんな傾向があるよね」なんて議論を数学的にやろうという話なので，絶対はないからです．それ故，考え方の基礎というか根っこの部分で置く仮定のようなもので派閥が別れるのかななんて思っています．

とかく，確率には頻度主義と呼ばれる派閥とベイズ主義と呼ばれる派閥があります．この派閥という表現だと対立しているようにも思えますし，実際対立している面倒な人達もいるようですが実際の現場ではどちらも便利なところがあるし，要は使い分けだと思います．よってどちらが優れているとか，どちらが正解とかいう議論は無意味だし，答えもないと思います．注意してください．

ということで，筆者はこうした宗教戦争には興味がないのと下手に触れると炎上しそうなので紹介だけです．興味があるようなら自分で勉強してみてください．正直自分はよく分かってないです．

5.2　加法定理と乗法定理

ベイズ統計に入る前に，確率に関する基本的な法則を二つ導入しておきます．加法定理と乗法定理です．三角関数とかでおなじみのあれです．確率の加法と乗法について考えてみましょう．

まずは加法定理から．確率変数 X, Y を導入します．それぞれ $X = x_i, Y = y_j$ を取る確率を $p(X = x_i, Y = y_j)$ と書き，これを $X = x_i, Y = y_j$ の同時確率と言います．

ここまでは良いですよね．x_i と y_j が同時に観測される確率だから同時確率です．

さて，この X, Y の両方についてそれぞれサンプルを取る作業を N 回行う事にします．この時，$X = x_i, Y = y_j$ となる試行の回数を n_{ij} と表します．こうすると，以下の式が成り立ちますね．

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} \tag{5.1}$$

これは普通に，「大小のさいころ 2 つを振って大 3, 小 6 を引く確率（同時確率）を求めよ」とかと同じです．全試行分の当該試行ですね．

では次に，$X = x_i$ となる確率を考えます．Y が取る値と関係なく，$X = x_i$ となる試行数を c_i としておきましょう．そうすると，$p(X = x_i)$ は

$$p(X = x_i) = \frac{c_i}{N} \tag{5.2}$$

になります．「大小 2 つのさいころを振って，大きい方が 6 になる確率は？」です．うーん？この例あまり良くないですね．どちらにせよ試行数 1 なので... まあいいや．

この式 (5.2) ですが，式 (5.1) を使ってあえて複雑に書けばこうなります．

$$p(X = x_i) = \sum_{j=1} p(X = x_i, Y = y_j) = \sum_{j=1} \frac{n_{ij}}{N} \tag{5.3}$$

Y の値はなんでも良いから，X が x_i になる確率を出したのが式 (5.2) なので，これは全ての $Y = y_j$ について式 (5.1) を足し合わせたものと同じだよねという式になります．

これが確率の加法定理です．また，捉え方を変えるとこの処理は x 以外の変数 (ここでは y）についての周辺化をするとも言い，式 (5.3) を x の周辺確率とも言います．

次に，$X = x_i$ が既に観測されている状況に限って，その中で $Y = y_j$ が観測される可能性を考えます．これを条件付き確率と言い，今の例だと以下のようになります．

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i} \tag{5.4}$$

左辺の記法，最初は慣れないかもしれませんが慣れましょう．縦線は条件付きである事を表し，「右側の条件の元の左についての〜〜」を意味します．逆にならないように注意．筆者は最初よく逆に捉えていてわけわからんになっていましたが，これ以降頻繁に出てくる表現です．ここでは，上で述べたように “$X = x_i$の時の $Y = Y_i$の確率 (p)” という意味です．

式に戻りますが，右辺は分母に $X = x_i$ になる試行数である c_i が来ていて，分子は X と Y が x_i, y_j になる試行数である n_{ij} ですね．言葉の通り，$X = x_i$ が確定した状況での $Y = y_j$ が成り立つ可能性になっている事が分かるかと思います．

以上の事から確認されるように，加法定理同様，式 (5.1，5.2) を考慮すると

$$p(X = x_i, Y = y_j) = p(Y = y_j|X = x_i)p(X = x_i) \tag{5.5}$$

が言えます．つまり，$X = x_i, Y = y_j$ の同時確率は $X = x_i$ が観測された上での $Y = y_j$ の確率であり，これらの乗算によって求められるという事です．一応計算しても，

$$p(Y = y_j|X = x_i)p(X = x_i) = \frac{n_{ij}}{c_i}\frac{c_i}{N} = \frac{n_{ij}}{N} \tag{5.6}$$

で式 (5.1) が確かに求められていますね．これが確率の乗法定理です．意味するところは，$X = x_i, Y = y_j$ の同時確率は $X = x_i$ の確率と，$X = x_i$ の元での $Y = y_j$ の条件付確率との積である，ということになります．

まとめると，確率論の基本法則２つは以下になります．

確率の基本法則

$$p(X) = \sum_Y p(X, Y) \qquad \text{加法定理}$$

$$p(X, Y) = p(Y|X)p(X) \qquad \text{乗法定理}$$

ここで p(X) は確率変数 X の確率分布を指します．なので先程までの記法に従うと $p(X = x_1, x_2, x_3, ...,$ となります．面倒なので以降はこうやって省略します．

5.3 ベイズの定理

いよいよベイズに行きます．ここではベイズ統計の歴史だとか主義がどうとかは一切触れないので，ぬるっと導入していきます．筆者がベイズいまいち分からん気がする理由は，ここでぬるっと導入できちゃうところにあったりもします．つまり主義がだの仮定がだの言っている割に，いわゆる頻度統計 (?) で使ってる普通の式の変形でだせちゃうんですよね．いきます．

まず，同時確率は対称です．つまり $p(X, Y) = p(Y, X)$ です．式 (5.1) より自明ですね．この性質を使って遊んでいるとベイズの定理が出てきちゃいます．

$$p(X, Y) = p(Y|X)p(X) \tag{5.7}$$

$$\therefore p(Y|X) = \frac{p(X, Y)}{p(X)} \tag{5.8}$$

$$\therefore p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \qquad \because p(X, Y) = p(Y, X) \tag{5.9}$$

できちゃいました．$\therefore, \because$ の使い方あってるかな？あってるよね．$\therefore$ は tex だと therefore って書き，$\because$ は because って書きます．これで分かりますね？

さて，あっさりとベイズの定理が出せちゃったわけですが，その意味を解釈する前にもう一つ導きたい式があります．分母にいる $p(X)$ に関して，加法定理と同時確率の対称性を使って

$$p(X) = \sum_Y p(X,Y) = \sum_Y p(Y,X) = \sum_Y p(X|Y)p(Y) \tag{5.10}$$

が言えますね．これを使って書き直した特殊な表現と原型を以下にまとめます．

ベイズの定理

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \tag{5.11}$$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)} \tag{5.12}$$

上の式がよく見るベイズの形．下は式 (5.10) を使って定義しなおした式です．下を使って，ベイズの定理の意味するところを見ていきましょう．右辺の分母は，分子の計算を全ての Y について足し合わせたものになっています．つまりこれは，左辺の条件付確率を全ての Y について足し合わせると 1 になるように正規化しているのだという事がわかります．確率だもんね，A である確率は 60%，B である確率が 90%，あわせて 150%です！なんてあほな話は困ります．

大小 2 つのさいころの例で考えるなら，こんな感じ

$$p(\text{小}=1|\text{大}=3) = \frac{p(\text{大}=3|\text{小}=1)p(\text{小}=1)}{p(\text{大}=3|\text{小}=1)p(\text{小}=1) + ... + p(\text{大}=3|\text{小}=6)p(\text{小}=6))} \tag{5.13}$$

$$= \frac{\frac{1}{6}\times\frac{1}{6}}{\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}} \tag{5.14}$$

$$= \frac{1}{6} \tag{5.15}$$

このことから，ベイズの定理の大事な部分は分母以外の $p(Y|X), p(X|Y)$, $p(Y)$ になります．こいつらについてそれぞれ何を意味しているのか考えていきましょう．

まず $p(Y)$ と $p(Y|X)$ です．$p(Y)$ は単純ですね，確率変数 Y の分布です．これは，計算を行う前から我々が「知っている」分布なので，Y の事前分布といいます．

知っているとは何かというのがベイズ確率の面白い？批判されている？ところで，ここで置く分布はなんでもいいです．「さいころなんだから全部 1/6 の一様分布でしょ」としてもいいし，「さっきから 3 ばかり出てるから 3 が 60%くらいな気がする」でも，「3 が全然出てこないからそろそろ出てきそう」でもいいし，とにかく，Xを観測する前に想定されているYの確率です．

一方，$p(Y|X)$ は何かというと，Xを観測した元での，という条件付確率分布なので事後分布と呼びます．ベイズで求めるのは事後分布ですね．事前分布 $p(Y)$ に何か $(p(X|Y), p(X))$ をかけたり割ったりし

て，条件付確率 $p(Y|X)$ にしているので，事前分布を得られた X で条件づけた何らかの処理によって修正する過程とも捉えられることが分かります．

残りは $p(X|Y)$ ですね．こいつが多分一番納得しにくいのかなと思います．対称性使ってひょっこり持ってきた項だしね．これは X の Y の元での条件付確率です．先程確認したように，ベイズの定理は事前確率 $p(Y)$ を，X を観測した上での $p(Y|X)$ に修正する過程でした．

ここで Y の元での X，というのが何を指すかですが，事前確率 Y の元でデータ X が観測される可能性，ですね？言い換えると，「Y の値は Y_j である！」と仮定したときに X_i が得られる確率，となります．

なので，事前確率が正解に近いほど大きな値になり，逆に遠いほど小さい値になります．事前分布が，得られた X に対してどれだけもっともらしいのかを表すので尤度と呼ばれる量です．

以上を踏まえて，改めてベイズの定理を眺めると，正規化項を無視すれば

$$\text{事後分布} \propto \text{尤度} \times \text{事前分布} \tag{5.16}$$

という式であると解釈する事ができます．

事前分布で持ってきていた “予想” の元で得られたデータがどれだけありえるかで予想を修正していくわけですね．あ，$\propto$ は比例を意味する記号です．

理解を深めてもらうため，一つちゃんとベイズっぽい例題を考えてみます．

研究室 A と研究室 B があります．研究室 A は男 4 人，女 6 人，研究室 B は男 8 人女 2 人だとします．多分　研究室 A は心理系のわりかしほんわかしたところで，研究室 B はくさいオタクが多そうですね．女性 2 人の心労はすごそうです．

それはさておき，今ここからどちらかの研究室を選択し，無作為に 3 回授業の手伝いのために雑用を選んで来たら，男男女だったとします．さて，最初に選んだ研究室はどちらでしょう？という問題を考え，ベイズ的に解いてみましょう．

研究室が A，B である事象をそれぞれ $L = L_a, L_b$ とし，選んだのが男女である事象を $S = S_m, S_f$ とします．

今，1 回目に男性が選択された時の研究室 A，B それぞれの事後確率を計算してみましょう．ベイズの定理を使って表してみます．

$$p(L_a|S_m) = \frac{p(S_m|L_a)p(L_a)}{\sum_L p(S|L)p(L)} \tag{5.17}$$

$$p(L_b|S_m) = \frac{p(S_m|L_b)p(L_b)}{\sum_L p(S|L)p(L)} \tag{5.18}$$

こうなりますね．次に各変数に実際の値を代入していきます．まず研究室の選択方法は無作為であったと仮定し，事前分布は一様，つまり $p(L_a) = p(L_b) = \frac{1}{2}$ とします．$p(S_m|L_a), p(S_m|L_b), p(S_m)$ は上記の設定より普通に計算できるので，1 回目の試行での事後確率は以下のようになります．

$$p(L_a|S_m) = \frac{\frac{4}{10} \times \frac{1}{2}}{\frac{4}{10} \times \frac{1}{2} + \frac{8}{10} \times \frac{1}{2}} = \frac{1}{3} \tag{5.19}$$

$$p(L_b|S_m) = \frac{\frac{8}{10} \times \frac{1}{2}}{\frac{4}{10} \times \frac{1}{2} + \frac{8}{10} \times \frac{1}{2}} = \frac{2}{3} \tag{5.20}$$

これはまあ普通ですよね．単純に全体のうち男性の比率が研究室 A と B で 1:2 になっているのを反映しています．では次に 2 回目の試行について考慮します．1 回目に男性が出たという事から，研究室 AorB という事前分布が更新されます．

先程は一様にどちらも 50%という仮定でしたが，今回は研究室 B である確率が高いと考えて事前分布を設定します．注意が必要なのは，分子の事前分布のみでなく分母に含まれている $p(Y)$ の値もそれぞれ更新されている事です．この点，あえて $p(S_m)$ と書かない今回の記法の方が理解しやすいと思います．

$$p(L_a|S_m) = \frac{\frac{4}{10} \times \frac{1}{3}}{\frac{4}{10} \times \frac{1}{3} + \frac{8}{10} \times \frac{2}{3}} = \frac{1}{5} \tag{5.21}$$

$$p(L_b|S_m) = \frac{\frac{8}{10} \times \frac{2}{3}}{\frac{4}{10} \times \frac{1}{3} + \frac{8}{10} \times \frac{2}{3}} = \frac{4}{5} \tag{5.22}$$

確率が更新されました．これがベイズの「母数自体がパラメータ」という考え方を反映していて，唯一無二の確率を出すというよりも「今まで得られているデータからはこんな事が言えそう」という主張です．よってデータを得るたびに逐次的に更新する事が出来ます．

これを使ったのが，後で触れる逐次学習だったりになりますがそれはとりあえず置いておきます．

さて，最後．今まで男男と連続で引いたので，やはり男性率の圧倒的に高い研究室 B である確率が高いという事になっています．しかし最後は女性を引きます．どうなるでしょう？

$$p(L_a|S_f) = \frac{\frac{6}{10} \times \frac{1}{5}}{\frac{6}{10} \times \frac{1}{5} + \frac{2}{10} \times \frac{4}{5}} = \frac{6}{14} \tag{5.23}$$

$$p(L_b|S_f) = \frac{\frac{2}{10} \times \frac{4}{5}}{\frac{6}{10} \times \frac{1}{5} + \frac{2}{10} \times \frac{4}{5}} = \frac{8}{14} \tag{5.24}$$

ありゃ，意外とがっつり削られたな．こんなもんなのか．3 回目に女性を引いたことによって，今まで高まっていた研究室 B である確率ががくりと落ち，逆に A である確率が上がりましたね．とはいえまだ B が優勢... という状況です．

例のごとく事前分布 $p(Y)$ は更新されているので注意．ベイズ統計はこのようにして予想を適宜修正していくような使い方が出来るわけですね．この，少しずつ分布を更新していく過程をベイズ更新とも言います．これゆえ機械学習だとか脳のモデルと相性が良いわけです．つまり，得られたデータを毎瞬反映させて少しずつ内部モデルを更新していけるわけです．つよい．

面倒なので計算はしませんが，一つ皆さんも気になるだろう話題について問題提起だけしておきます．

今回の問題では研究室 A か B のどちらが選ばれたのか最初に一様分布を仮定しましたが，そうじゃない場合はどうなるでしょう？

雑用を選ぶ問題だったので，なんとなく先生の力が弱い方の研究室が押し付けられそうですよね．研究室 A は新任の特任准教授，研究室 B は学部長の主催するラボだったとしましょう．そうするとなんとなく，押し付けらそうなので事前分布に研究室 A75%, B が 25%とかって仮定をおいてみたくなります．

アカデミア，政治色強いですよね．研究は大した事なくても政治だけでのしあがっていく先生も多いし，闇を感じるところは多々あります．いや，筆者もおそらく政治強い側の人間になりそうなので強く言えませんが... コネって大事ですよね．

閑話休題．

さて，暇な人はこのように事前分布を変えて計算してみてください．当然結果は変わります．そしてこの点がまさに，ベイズに批判的な人達の武器で，事前分布に何を用いるかが結局強く影響しちゃうんですよね．なのでベイズを用いる際は，事前分布の妥当性はちゃんと検証しなければならないです．

5.4 無情報事前分布

わすれてた

5.5 共役事前分布

ごめんなさい！！！！！！！！！！！

5.6 ベイズの応用

応用，というか実際にベイズを何にどうやって使っていくのかという話です．まず，ここまでベイズは離散で考えてきましたが当然連続確率分布でも計算できます．むしろそっちの利用の方が多いでしょう．まぁ例のごとく基本は $\sum$ を $\int$ に変えるだけです．

これを使ってどんなアプリケーションが使えるのかというと，先程の逐次学習がまずあります．次々観測されるデータを元に，内部モデルを更新していく過程ですね．パターン認識，機械学習，それから脳のモデルなんかに頻繁に適用されて議論されます．

自分の専門でもあるから少し述べるなら，脳は常に環境に適用するため，次に何が起きるのかや，得られた刺激がどういったものなのかを推定していく必要があります．

同時に，もしその推定が間違っているなら修正していく必要があるし，うまくいったなら強化していくと嬉しい事がありそうです．まさにこの過程がベイズ使った更新なんじゃない？みたいな話になっていくわけですね．

まぁ実際はそんな単純じゃないし，上に書いたような内容は古いというか単純な話です．もう少し複

雑だったり，厳密にはちょっと違ったりしてきます．がこれ以降は万人に共通して必要な知識というほどでもないと思うので advanced の方に引き渡します．興味ある人は読んでみてください．

第6章　統計的推定

なぜ確率分布のあとにベイズに突然触れたかというと，この章で扱う推定に密接に関わってくるからです．実世界で統計を使っていく上で重要なのは，あるデータ群が形成する分布の種類を特定する事ではありません．ただ分布型が分かっただけでは大抵，なんの役にも立ちません．

重要なのは，その分布の平均や分散です．クラスのテスト成績が正規分布に従う事を知っていても，肝心の平均点などが分からないと学生は喜んでいいのか落ち込むべきなのか分からないし，教員も授業内容の補修をするべきかもっと踏み込んだ議論をするべきか分かりません．

という事で，母集団の確率分布を規定する “母数” を推定したくなります．母数は，たとえば母平均や母分散といった量があります．母平均の推定はわりとよくやっていますね．

あるクラスでてきとうに選んだ5人の点数 $(X_1, X_2 ... X_5)$ があれば，彼らの平均からクラス全体の平均を推定しようとするでしょう．これが母数の推定です．

$$\hat{\mu} = (X_1 + X_2 + ... X_5)/5 \tag{6.1}$$

この式では，母平均 (μ) の推定を行っています．μ の上についている $\hat{\mu}$ は推定値である事を意味します．当然，母集団全てを加算平均してだした $\hat{\mu}$ は μ と同じ値になります．しかし現実には母集団全てを観測し計算するなど不可能なことが多いため，その一部のみを使ってうまく母数に近づけた推定値を求めていく必要があります．

あるいは，統計的ないわゆる○○分布ではないけれどもなんらかの関数で説明できるような変化，分布も多くあります．これらを説明する際に，多項式曲線を使ってフィッティングしていったりするのですが，この際にもやはりこのモデルの形を定めるパラメータの適切な値を推定したくなります．

こうした量を推定していく方法を勉強していきましょう．

6.1　点推定と区間推定

母数の推定値を求める際に，方法が大きく分けて2種類あります．それが点推定と区間推定です．簡単ですが一応．点推定は先程の式 (6.1) のように母数をある一つの値 $\hat{\theta}$ で指定する方法を指します．なおここで θ は母数を表す一般化記号で，実際には μ, σ などがあてはまります．

単一の点で推定をする以上，どうしても実際の母数との誤差が生じます．そのため，この誤差を最小

化していきたいというのが点推定の主なモチベーションです.

区間推定は，点推定と異なり推定に幅を持たせる方法です．たとえば，A 組のテストの平均点は 50-60 点の間で，56 くらいの可能性がめっちゃ高い．などといった推定です.

6.2 最尤推定

まずは点推定のうち，頻度主義でよく使われる最尤推定を考えます．そのために，まずはベイズの時に出てきていた尤度についてもう少し考えます．尤度は何もベイズ主義の用語ではありません．頻度主義でもよく使うもので，結局ベイズと頻度とのちがいは何かというと事前分布を用いるかの方にあります.

で，尤度ですが，事前確率を評価しているという風に言っていましたが推定の枠組みで考え直します．推定は分布の母数だったり，あるいはより複雑な多項式曲線フィッティングのパラメータだったりするので，こいつらを確率変数ベクトル $\mathbf{w}$ とします．たとえば $\mathbf{w} = [\mu, \quad \sigma]$ とかです.

観測データを D でおくと，ベイズの定理は以下の形を取ります.

$$p(\mathbf{w}|D) = \frac{p(D|\mathbf{w})p(\mathbf{w})}{p(D)} \tag{6.2}$$

ここで，尤度 $p(D|\mathbf{w})$ について考えるとデータ集合 D に対するベクトル $\mathbf{w}$ による評価関数と捉えられます．事前分布を更新するという話の時と同様です．例のごとく p(D) は正規化しているだけなのでどうでもいいです.

さて，最尤推定ですが，読んで字のごとく尤度を最適化する推定方法です．尤度はパラメータベクトルの元での，得られたデータの尤もらしさを評価する関数なので，これを最大化する事で最も尤もらしいパラメータの推定値，最尤推定値を求める手法です．実際に例を考えながら説明してみます.

ある研究室 (A) に所属する学振特別研究員の数を推定したいとします．学生 (S) の数は 5 人にしておきましょう．推定したいパラメータ (DC 研究員数) を i とします．そうすると

$$p(S = DC|i) = \frac{i}{5} \tag{6.3}$$

$$p(S = \text{凡人}\,|i) = \frac{5-i}{5} \tag{6.4}$$

と尤度を仮定できます．最尤推定の基本は，尤度が最大の事象が観測されたという仮定を置くことにあります．よって，観測された事象分の尤度をかけ合わせて，最大になるパラメータ i の値を採用します.

たとえば今回，研究室 A から適当に人を選んできて DC に採用されているか確認したところ，採用採用不採用となっていたとします．その場合尤度関数 L(S; i) は

$$L(S;i) = \frac{i}{5}^2 \frac{5-i}{5} \tag{6.5}$$

となります．こいつを最大化すれば良いわけですね．ちなみに記法ですが，f(x;y) という記法は数学ではパラメータ y によって規定される変数 x の関数，という意味です．ここでは y は変数ではなく，f を定義する際の固定値になります．だからこそ，今回は固定値として使う i の値を最適化したいというモチベになります．最尤法について多くの資料ではこっちの記法を使っていますが，違いが特になさそうなのと面倒なのでここでは $L(S|i)$ としておくことにします．

せっかくなんで手計算．

$$L(S|i) \propto -i^3 + 5i^2 \tag{6.6}$$

$$\therefore L(S|i)' \propto i(-3i + 10) \tag{6.7}$$

$$i = \frac{10}{3} \tag{6.8}$$

最大化は微分して 0 になる点を探せば良いんでしたね．今回のだとだいたい 3 が尤度最大の点でした．研究室 A の DC 学生は 5 人中 3 人であるという推定になります．優秀や...

なお，今回は大した計算じゃなかったのでそのままやりましたが実際は 100 試行とかになると尤度を掛け算でだすと値がすごい事になってしまうので対数に変換した対数尤度関数の最大化を行う事が多いです．対数に変換した場合，掛け算は足し算になるので計算が楽ですね．

さて，計算方法と気持ちが分かったところで以上の手順を一般化します．まずは尤度関数の定義から．

尤度関数

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta) \qquad \text{尤度関数} \tag{6.9}$$

$$logL(\theta|\mathbf{x}) = \sum_{i=1}^{n} f(x_i|\theta) \qquad \text{対数尤度関数} \tag{6.10}$$

大丈夫でしょうか．尤度関数のかけざんを対数尤度関数では足し算に変換しました．ここで θ は母数だったり，とにかくパラメータです．パラメータ θ によって規定され，変数 x を変数とする尤度を全ての x についてかけあわせたのが尤度関数でした．

対数尤度は計算を簡単にするため，それを対数変換したものですね．

で，最尤推定はこいつらを最大化する処理なので，最尤推定値 $\hat{\theta}$ は

$$\hat{\theta}_{ML} = \arg\max_{\theta} L(\theta) \tag{6.11}$$

$$= \arg\max_{\theta} logL(\theta) \tag{6.12}$$

となります．$\arg\max_x f(x)$ は関数 $f(x)$ を最大にする x の集合を意味します．つまり一番高い上に凸の山の頂点です．勿論集合の要素は 1 つのみでも問題ないので，解は一つだったり複数だったりします．

余談ですが，対数尤度関数は最大化する事で尤もらしい値をだすものでしたが，機械学習とかだとよくこれを符号反転させた $-logL(\theta)$ を誤差関数と呼び，これを最小化する方向にフィッティングしているものも多いようです．単調減少していくので，0 に近付けるという意味で扱いやすいんでしょうね．

6.3 最大事後確率推定 (MAP 推定)

しかし本当に尤度最大化で良いのでしょうか？

さっきの例だと，自分のラボを優秀に見せようとした PI がわざと採択されている学生だけ選ばれるように仕向けていたのかもしれません．本当は 1 人しかいないかもしれないし，極端な話 2 回目の試行で止めていたら採用採用だったので採用率 100%，最尤推定の結果は当然，採択者 5/5 名となってしまいます．

これでは困りますよね．このように，最尤推定の弱点として十分な試行回数がないと極端すぎる結果を招きかねない点があります．どうにか，尤度だけじゃなくて「あの教授は信用ならないしなぁ」とかといった事も考慮できないのでしょうか...

そこでベイズに立ち返りましょう．

ベイズの定理の右辺には尤度だけじゃなく，事前分布もありました！これも使えば良いわけですね！

改めてベイズの式．

$$p(\mathbf{w}|\mathbf{D}) = \frac{p(\mathbf{D}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{D})} \tag{6.13}$$

$$\therefore p(\mathbf{w}|\mathbf{D}) \propto p(\mathbf{D}|\mathbf{w})p(\mathbf{w}) \tag{6.14}$$

最尤推定はこのうち尤度 ($p(\mathbf{D}|\mathbf{w})$) にのみ注目して，こいつを最大化させていました．しかしそれでは十分なサンプル数がないと偶然の影響を消しきれないのでしたね．ではそれに加えて事前分布 $p(\mathbf{w})$ も用いてみましょう．

という事で, 最尤推定と同じ問題に取り組むもう一つの方法として, MAP 推定 (Maximum a posteriori estimation) があります．

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|\mathbf{D}) \tag{6.15}$$

$$= \arg\max_{\theta} \frac{p(\mathbf{D}|\theta)p(\theta)}{p(\mathbf{D})} \tag{6.16}$$

$$= \arg\max_{\theta} p(\mathbf{D}|\theta)p(\theta) \tag{6.17}$$

MAP 推定は日本語では最大事後確率推定，そのままですね．つまりベイズの定理の左辺を最大化しようという話で，事後確率を最大にする θ を求めます．

つまり事後分布の一番山になっている部分に対応する θ を選択します．簡単ですね．

ここまでは良いと思いますが，問題が一つ．
右辺で尤度に事前分布をかける事は，すなわち得られるパラメータ θ 自体も分布をもっていて，唯一つの真の値ではないという考えの元行うものです．ではそのパラメータが従う分布をどうするか？という事が問題になるわけで方法は大きく 2 種類です．
一つ，先行研究や何かしらの資料から明らかになっている分布を引っ張ってくる．これは分かりやすいですね．学振の例だったら，そのラボのこれまでの採択率だったり，国全体の採択率だったりのデータをもってきても良いかも．
次，そういった傾向すら分からない場合．この場合は無情報事前分布を持ってきます．無情報事前分布は事前に情報がない場合や事前分布を設定するにあたって根拠がない場合などに用いる分布でしたね．

無情報事前分布のうち，「多分一様やろ」，として特に重みづけをしない分布が一様分布でした．勿論こいつも MAP 推定に使えます．その場合，MAP 推定値と最尤推定値は等しくなります．

しかし無情報事前分布として一様分布を利用する問題点として，連続分布にはなれない事などがあげられます．そのため，やはり多いのは共役事前分布を用いる事だと思います．尤度関数の形が分かっている場合，それに対応した共役事前分布を用いる事が多いです．
二項分布に対してはベータ分布，とかね．

6.4 ベイズ推定

MAP 推定は事後確率の最大にする θ を求める方法でした．言い換えると，事後確率分布が最大になるような θ を求める事です．よって，MAP 推定値が同じ θ になった二つの分布があったとしてもその違いは評価されません．たとえば (計算すると若干違うかも？)，100 戦やって 75 勝しているベテランと 4 戦やって 3 勝している新人とを同じ勝率，強さで評価してしまったりすることになります．

勝率が同じになるのは良いとしても，その信頼度が全然違いますよね．ベテランの方がデータが多いので信用できます．新人の方は単純にビギナーズラックかも．

そこでベイズ推定は，事後分布そのものを使っちゃいます．
つまり，単純に arg max しないでそのまま分布を使えばいいわけです．ただ注意しないといけないのは，arg max ではないので MAP 推定のように $p(\mathbf{D})$ は無視できません．

手順は普通にベイズを使って，

$$p(\theta|\mathbf{D}) = \frac{p(\mathbf{D}|\theta)p(\theta)}{p(\mathbf{D})} \tag{6.18}$$

を計算し，事後分布 $p(\theta|\mathbf{D})$ を求めるだけです．この時，別に θ の点推定をしているわけじゃないので $\hat{\theta}_{bayes}$ のような量が求められるわけではないです．それをするのは MAP 推定でしたね．

こうすることで，θ の分布を求められます．裾野が広く背の低い分布なら信頼度は低いし，裾野が狭く背の高い分布なら信頼度が高い，という風に見ることが出来ます．一般に，データ数が多いほど信頼度は上がっていく傾向にあります．

6.5 逐次推定

分布を使えるということに加え，もう一つベイズ推定の便利な点，それは求めた事後分布を事前分布として用いて，また次のデータを使って推定を計算しなおす，という事を容易に可能にする点です．

ここでは，そんな逐次推定として逐次ベイズ推定を紹介します．

とはいえ，それっぽい事は実は既にやっています．ベイズの定理の説明で無作為に選んだ男女の数から研究室を推定する問題を考えましたよね．一回目男性，二回目男性，三回目女性，とデータを重ねるにつれて事前確率を更新しつつ計算しました．基本的にはあの操作の事です．

$$\begin{aligned} p(\theta|x_1) &= \frac{p(x_1|\theta)p(\theta)}{p(x_1)} \\ \therefore p(\theta|x_1) &\propto p(x_1|\theta)p(\theta) \end{aligned} \tag{6.19}$$

これは一つのデータを観測して未知の母数の事後分布を推定する式でした．ここで，$p(x_1|\theta)$ を特に尤度関数 $L(\theta|x_1)$ と言いました．次に，もういくつかデータを観測した場合と，そこに更にもう一つデータをプラスした尤度関数は

$$L(\theta|x_1, ..., x_n) = \Pi_{i=1}^{n} p(x_i|\theta) \tag{6.20}$$

$$L(\theta|x_1, ..., x_n, x_{n+1}) = L(\theta|x_1, ..., x_n)p(x_{n+1}|\theta) \tag{6.21}$$

となりました．よってデータ $X = x_1, ..., x_n$ で計算された事後分布

$$p(\theta|x_1, ..., x_n) \propto L(\theta|x_1, ..., x_n)p(\theta) \tag{6.22}$$

$$\tag{6.23}$$

と，そこにデータ x_{n+1} を新たに加えて計算された事後分布とは

$$\begin{aligned} p(\theta|x_1, ..., x_n, x_{n+1}) &\propto L(\theta|x_1, ..., x_n)p(x_{n+1}|\theta)p(\theta) && (6.24) \\ &\propto L(\theta|x_{n+1})L(\theta|x_1, ..., x_n)p(\theta) && (6.25) \\ &\propto L(\theta|x_{n+1})p(\theta|x_1, ..., x_n) && (6.26) \end{aligned}$$

と証明できるように，これまでの観測値を得て計算した事後分布は新たな観測値を得たときの事前分布になる関係にある事がわかります．これを用いて，新しくデータを観測するたびに $p(x|\theta)$ をかけていく事で分布を更新していく事ができます．

一方，最尤推定などの手法は事前分布を考えません．つまり「さっき計算したデータ」を利用することが出来ないわけですね．もう一度最初から計算しなおす必要があります．この点，新しく得たデータ分だけ分布をちょっと修正すれば言い訳ですから，ベイズ推定は計算コスト的にも考え方的にも楽だし便利ですね．

6.6 推定のまとめ

推定はあくまで，パラメータのもつ不確実性を実測されたデータで条件づけて更新する作業です．得られたデータから，そのデータの母集団が従っている分布は，母数 (母平均とか) は，パラメータは，なんだろうと考えるものです．

最尤推定は実測値が尤も観測されやすそうなパラメータを尤度のみを用いて推定する手法で，ベイズ推定は事前分布も用いてそのパラメータの事後分布を推定する手法でした．MAP 推定はベイズ推定の事後分布のうち，事後確率が最大になるパラメータを点推定する手法でした．

これらの手法を用いる事で，得られたデータ分布の特徴を掴むことができます．よって，脳の学習モデルの基本になっていたりします．これまでに得られた感覚データを元に，外界についてのモデルを内部に構築していく過程のことを我々は学習と呼んでいます．よって，脳は (逐次) ベイズ推定によって環境に適用しているよね！なんてベイズ脳理論なんていうものが盛んだった過去があります．

というよりも，ベイズ推定自体が人間っぽく統計しようって考えられたみたいな背景があるようです．

今ではもっと拡張というか盛り盛りになった自由エネルギー原理なんていうお化けに吸収されていたりしますが，いずれにせよ重要な概念です．頭のかたすみに置いておくくらいしておくと嬉しいかもですね．

第7章　予測

実測データで条件づけたパラメータの不確実性を減少させる推定に対し，実測データで条件づけた未観測のデータについての不確実性を更新するのが予測です．実際の現場だとこっちを使いたい事も多いですよね．

機械学習なんかはこれをいかにセクシーでエレガントにやるかを競っているわけです．神経科学的には，BMI の実装だったり，脳の知覚モデルなんかを理解していく上で必要になります．しかし絶対というわけでもないのでさらっと，とりあえずベイズの予測だけいれておきます．

7.1　ベイズ予測

まずは分かりやすいベイズ予測．これまでに観測されたデータを用いたベイズ推定で求めた事後分布を使って，

ベイズ予測の手順

$$p(\theta|\mathbf{D}) = \frac{p(\mathbf{D}|\theta)p(\theta)}{p(\mathbf{D})} \qquad \text{事後分布の推定}$$

$$p(D^{new}|\mathbf{D}) = \int p(D^{new}|\theta)p(\theta|\mathbf{D})d\theta \qquad \text{予測} \tag{7.1}$$

のような事をして，未観測のデータを予測します．$p(D^{new}|\mathbf{D})$ はデータ集合 $\mathbf{D}$ を観測した状態で，次に観測されると思われるデータの予測値の分布で，これを事後予測分布と言います．

式を解釈すると，パラメータ θ が取りうる値の全ての範囲で，パラメータ θ がある値を取る確率 $p(\theta)$ で重みづけした $p(D_1^{new}|\theta)$ の足し合わせを D_1^{new} の確率，$p(D_2^{new}|\theta)$ の足し合わせを D_2^{new} の確率，というように計算することで D^{new} が取りうる値の予測分布を求めるものです．

ここで，尤度 $p(D^{new}|\theta)$ は未観測データの尤度になりますが，事前に観測されているデータと同じ生成過程で生成されているとすれば事後分布の計算の際に用いた尤度 $p(\mathbf{D}|\theta)$ と同じものを使うことが出来ます．

こうして得られた事後予測分布から，期待値を取ったりして点推定しても良いし，そのまま分布として扱ってもいいです．ちょっとお漏らしすると，この予測分布による点推定？のような事を脳は常におこなっていて，ゆえに「我々が感じているのは厳密には実際の外界じゃなくて脳の予測結果 (*正確じゃない表現です) である，だからたまに予測がミスる事があって，それが錯覚である」，なーんて話があったりもします．楽しいですよね．

詳しくはいずれ advanced で書いていきましょう．ぼくの学士卒論は視聴覚統合錯覚だったので，当

時はこういった話を聞いてとてもわくわく，というか興味をそそられました．計算論的に神経科学に向き合うのも良いな！と思ったきっかけです．当時は何書いてるのか何も分からなかったのに，今は自分で簡単に解説できるくらいにはなったのかなと思うと感慨深いです．