Basic Biostatistics for Beginners

Risa Kawaguchi

CiRA Bioinformatics Study Meeting
Thursday 22nd June, 2023

Kyoto University

Contents

About presentation materials

- Made by Beamer on overleaf
- Available at `https://github.com/carushi/cb_lab/code_collection/230622_cb_bio_stat/`
- 誤りなど見つけましたらご連絡頂けると幸いです

Probability theory for statistics

Resources

- 解析学 - 確率論・測度論 - 統計学
- Mathematical analysis - Probability and measure theory - Statistics
- 統計学入門 (基礎統計学 1) - 自然科学の統計学
- *https://bellcurve.jp/statistics/course/*
- *http://ibisforest.org/index.php?FrontPage*
- *https://www.statskingdom.com/index.html*
- *https://github.com/tsg-ut/awesome-prml-ja* PRML

重要な概念

- 標本 ω・標本空間 Ω - サイコロの目・とりうる目全体
- 事象 A・事象空間 F - 偶数・奇数、3 以上など確率測度で可測な部分集合の和
- 確率測度 $P(A)$・確率空間 (Ω, F, P) - それぞれの事象に対しその確率（実数）を返す関数
- 確率変数 $X(\omega)$ - 事象を表す変数
- 確率分布 $Pr(X)$ - 確率変数がある値となるときの確率を返す関数

確率の性質を満たすには、様々な条件が必要とされている。例えば事象が無限個あった場合（連続値など）の確率や、離散的な分布において、期待値などの計算はどのようになされるのか？それらを厳密に定義するためには、数学の理解が必要。以下は上辺の理解。

- 和が 1 $- P(\Omega) = 1$
- F に対するそれぞれの確率測度が 0-1 の範囲内 $- P : F \to [0, 1]$

期待値と平均値

期待値とは一般に、確率変数のとりうる値に確率の重みをかけた値。n 回試行した場合のサンプルの観測確率を $\frac{1}{n}$ とすれば、サンプル平均は標本集団における X の期待値となり、期待値 = 平均の一般化と考えることも出来る。

- 離散の場合 $E(X) = \sum_i x_i \times Pr(x_i)$
- 標本平均 $mean(\bar{x}) = \sum_{i=0}^{n} x_i \times \frac{1}{n}$
- 連続の場合 $E(X) = \int_{-\infty}^{\infty} x \times Pr(x)dx$
- 分散 $V(X) = E[(X - E[X])^2] = \sum_i x_i^2 \times Pr(x_i) - E(X)^2$
- 標本分散 $Var = \sum_i^n (x_i - \bar{x}) \times \frac{1}{n}$
- 三次、四次の平均値周りの期待値を σ^3 と σ^4 でわったものは歪度、尖度

ガウス分布の場合の平均・分散

ガウス分布・正規分布 $N = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$

- Z-score や t 検定の計算の際には、母集団が正規分布に従うと仮定
- ガウス積分の公式 $\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}$
- 平均 $E(X) = \int_{-\infty}^{\infty} x \times Pr(x) dx = \mu$
- 分散 $V(X) = E(x^2) - E(x)^2 = \sigma^2$
- 正規分布は歪度が 0、尖度が 3 になる関数
- 1 次・2 次モーメントの値で関数全体が規定される
- 中心極限定理により、サンプルの数を増やしていくとサンプルの平均は正規分布に近づくことが知られる
- 性質的にも扱いやすいためによく用いられる

おまけ：便利なモーメント母関数

モーメント母関数は、t で n 回微分して t=0 を代入すると $E[x^n]$ となる関数。存在しないこともある。正規分布においては、

- モーメント母関数 $M_X(t) = E[e^{tX}] = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$
- $\mu = 0$ のとき（簡単のため）
- 一階微分＝平均：$M'_X(t) = t\sigma^2 e^{\frac{\sigma^2 t^2}{2}}, M'_X(0) = 0$
- 二階微分＝分散：$M''_X(t) = \sigma^2(1+\sigma^2 t^2)e^{\frac{\sigma^2 t^2}{2}}, M''_X(0) = \sigma^2$
- 三階微分＝歪度：$M'''_X(t) = \sigma^4(3t+\sigma^2 t^3)e^{\frac{\sigma^2 t^2}{2}}, M'''_X(0) = 0$
- 以下 n 回続く…

おまけ：便利な特性関数

特性関数は確率分布を完全に定義する関数で、確率分布のフーリエ変換後の関数とも言える。

- 特性関数 $\psi_X(t) = E[e^{itX}] = \int_{-\infty}^{\infty} e^{itX} dF(X)$
- $= \int_{-\infty}^{\infty} e^{itx} f_X(x) dx$
- 正規分布の場合 $\psi_X(t) = \exp(i\mu t) \exp\left(-\sigma^2 t^2/2\right)$

Hypothesis Testing - 仮説検定

統計

何かの効果や特性を知りたいが、すべて（母集団）を観測できないとき、そして観測にノイズやばらつきが存在するときに、より信頼性の高い結論を得るために様々なデータを集めたり、モデルに基づく確率・期待値を利用して仮説を検証する。

- ナイチンゲールによる病院の衛生状態と戦死者・傷病者の関係性の証明
- 選挙における投票と出口調査
- モンティ・ホール問題
- ギャンブルに勝てるかどうか？
- このレアガチャは当たるのか？

母集団の仮定

- 経験分布 - 得られたサンプルの中での外れ値を探す *
- 確率分布 - 正規分布・ポワソン分布・ベータ分布・ロジスティック分布など
- ノンパラメトリックな方法
- より厳しい仮定を置くほど有意差を鋭敏に検出できる

平均値・分散・分布の違いなどを検出

正規分布を仮定した検定

- t 検定 - 2 集団の平均の差。分散が等しい正規分布を仮定（異分散の場合は Welch’s t test）
- F 検定 - 正規分布に従う分布の標準偏差の違いを検出（諸説あり）

ノンパラメトリックな検定

- Wilcoxon の符号順位和検定・Mann-Whitney U 検定 - ２集団の順位差
- Kolmogorov-Smirnov 検定 - 分布全体の差
- χ^2 検定 - 期待値への当てはまり度合いなどでよく使われる $\sum \frac{(O-E)^2}{E}$

仮説検定とは

- 1. モデルをもとに帰無仮説 H_0 を設定する（ランダム、同じ母集団、差がない）
- 2. 対立仮説 H_1 を証明したい事柄とする（ランダムではない、別の母集団、差がある）
- 3. 帰無仮説に従うときの確率を計算する
- 4. 帰無仮説に従う確率が十分に低いとき、帰無仮説を棄却する
- 5. 帰無仮説が棄却されなかった場合、対立仮説が正しいとする統計的な有意性はない
- モデルに従うとき利用してそれぞれの値のときの確率を計算する
- これにより正規分布表・t 分布表などが作れる
- ただしこれらは特定の自由度・標本サイズのときの近似値であり、

Multiple Testing - 多重検定

検定を何度も行う場合

- 一回の検定で誤って帰無仮説が棄却する可能性は $\alpha(=0.05)$ の値で制御される
- しかし、この検定を繰り返すと、誤って棄却される可能性は上昇する
  $1-(1-0.05)\times(1-0.05)...$
- cf) 低確率のガチャ
- これをどのように補正するか？

よく利用される補正方法

- Bonferroni 法
- - FWER（family-wise error rate）を制御する＝少なくとも一回真の仮説を誤って棄却する確率により制御。
- - やることは α を検定回数 n で割るもので、厳しい基準に基づく。
- - GWAS など大量の検定を行う場合、何も検出できないことも多い
- Benjamini-Hochberg(BH) 法
- - FDR (false-discovery rate) を制御する＝棄却された仮説の中での誤って棄却された仮説の割合の期待値（α, q）により制御。
- - p 値を昇順に並べて $p_i \leq \frac{i}{n}q$ を $i = n$ から 1 まで条件を満たすまで探索（満たした場合 $1 \sim i$ まで棄却する）
- - 補正した値は q 値と呼ばれる
- - 計算には便利な関数を使おう

実践編

*https://colab.research.google.com/github/carushi/cb_lab/blob/main/code_collection/230622_cb_bio_stat/sample_biostat_template.ipynb*にアクセス

gmail アカウントでログイン

自分でプログラムを Run！

他にも github にあがっているコードは Google collab で利用可能

https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb

https://colab.research.google.com/github/lexfridman/mit-deep-learning/blob/master/tutorial_deep_learning_basics/deep_learning_basics.ipynb