ESTR 2020 统计

> Statistics for Engineers

①②③④⑤⑥⑦⑧⑨⑩⑪⑫⑬⑭⑮⑯⑰⑱⑲⑳㉑㉒㉓㉔㉕㉖㉗㉘㉙㉚㉛㉜㉝㉞㉟㊱㊲㊳㊴㊵㊶㊷㊸㊹㊺㊻㊼㊽㊾㊿

w.r.t.　　with respect to 关于（某个变量）

Course Information

> https://www.cse.cuhk.edu.hk/~sinnopan/engg2780.html

Recent Updates

- Tutorials start from Jan 15 (Week 2, Wed) and quizzes start from Jan 22 (Week 3, Wed)
- An ENGG 2780A/ESTR 2020 discussion board on piazza was set up here. Please register and sign in using your CUHK email address.

Course Description

统计学是从数据中建立概率模型并对其进行验证的科学. 我们将学习贝叶斯方法和经典方法在参数估计、置信区间、假设检验以及关于（非）独立性的推理中的应用.

> Statistics is the science of creating probabilistic models from data and validating them. We will learn about the Bayesian and classical approaches to parameter estimation, confidence intervals, hypothesis testing,and reasoning about (in)dependence.

Textbooks

The references for this course are [BT] *Introduction to Probability (2nd edition)* by Bertsekas and Tsitsiklis (Chapters 8-9) and [DS] *Probability and Statistics (4th edition)* by Morris Degroot and Mark Schervish (Chapters 7-12)

第二本书 zlibrary 下不到，请求图书馆提供了电子版链接：

https://julac-cuhk.primo.exlibrisgroup.com/permalink/852JULAC_CUHK/17kiu0g/alma991040188750503407

Grading

Your grade will be based on 4 components:

- a final exam (40%)

allowed to bring two double-sided A4 cheat sheets.

- a midterm (30%)

allowed to bring two double-sided A4 cheat sheets. The midterm exam is scheduled on Feb 24 (Week 8, Mon) during the lecture.

- exercises & quizzes (15%)

共有9组练习. 练习不会被评分. 练习将在每周一发布，并在每周三的辅导课上讨论. 详情请参考以下时间表.

> there are 9 sets of exercises. Exercises won't be graded. They will be issued on Mondays and discussed in tutorial on Wednesdays. Refer to the schedule below for details.

共有7次测验. 测验将在下周三的辅导课上进行，为一项与练习相关的单题 15 分钟开卷测验（*除计算器外，不允许使用电子设备*）. 成绩中将去掉最低的两次测验分数. 不提供补测.

> there are 7 quizzes. A 15-minute open-book (*electronic devices except calculators are NOT allowed to use*) quiz with a single question related to the exercise will be given in tutorial the next Wednesdays (refer to the schedule below for details). The lowest two quiz grades will be dropped from the count. No make-up quizzes will be offered.

Week	Date	Tutorial	TA Name	Topic	Materials
Week 2	Jan 15	T1		Solution of Exercise 1	Exercise 1; Solution 1
Week 3	Jan 22	T2		Solution of Exercise 2, quiz 1	Exercise 2; Solution 2
Week 4	Jan 29	**Lunar New Year Vacation (no tutorial)**			
Week 5	Feb 5	T3		Solution of Exercise 2	Exercise 2; Solution 2
Week 6	Feb 12	T4		Solution of Exercise 3, quiz 2	Exercise 3; Solution 3
Week 7	Feb 19	T5		Solution of Exercise 4, quiz 3	Exercise 4; Solution 4
Week 8	Feb 26	**Middterm Exam on Feb 24 (no tutorial)**			
Week 9	Mar 5	**Reading Week (no tutorial)**			
Week 10	Mar 12	T6		Solution of Exercise 5	Exercise 5; Solution 5
Week 11	Mar 19	T7		Solution of Exercise 6, quiz 4	Exercise 6; Solution 6
Week 12	Mar 26	T8		Solution of Exercise 7, quiz 5	Exercise 7; Solution 7
Week 13	Apr 2	T9		Solution of Exercise 8, quiz 6	Exercise 8; Solution 8
Week 14	Apr 9	T10		Solution of Exercise 9, quiz 7	Exercise 9; Solution 9
Week 15	Apr 16	T11		Solution of final exam exercise	Pratice Final; Solution

- attendance (15%)

Lecture 不计 Attendance

the marks you can get regarding attendance = (#quizzes you have taken / 7)* 15. In practice, to avoid introducing infinite decimal, we will the following equation to compute your attendance mark: (#quizzes you have taken / 7)* 14 + 1. E.g., if you take 5 quizzes out of 7 (miss 2 without providing a Medical Certificate (MC)), then you get 11 marks for attendance no matter your quizzes answers are correct or not.

For ESTR 2020 students

ESTR 2020 的格式与 ENGG 2780A 相同，但增加了额外的阅读材料和一个需在学期末展示的 Project. 考试和测验与 ENGG 2780A 相同.

项目可以个人完成或与搭档合作. 项目包括一份简短 report（10%）和一次 pre（5%），用于代替 ENGG 2780A 中的出勤部分. 有关项目的详细信息将稍后讨论.

不能用 AI

> ESTR 2020 follows the same format as ENGG 2780A, but with additional readings, and a project to be presented at the end of the semester. The exams and quizzes will be the same as in ENGG 2780A.

> Projects can be done individually or with a partner. They will involve a short report (10%) and a presentation (5%), which replace the attendance component as in ENGG 2780A. The details about the project will be discussed later.

Schedule

Week	Date	Lecture	Topic	Materials
Week 1	Jan 6	L1	Probability vs Statistics	pptx
Week 2	Jan 13	L2	Bayesian statistics	pptx
Week 3	Jan 20	L3	Prediction, estimation, and testing	pptx
Week 4	Jan 27	L3	Prediction, estimation, and testing	
Week 5	Feb 3	**Lunar New Year Vacation (no class)**		
Week 6	Feb 10	L4	Sampling statistics	pptx
Week 7	Feb 17	L5	Classical point estimation	pptx
Week 8	Feb 24	**Midterm Exam (during lecture)**		
Week 9	Mar 3	**Reading Week (no class)**		
Week 10	Mar 10	L6	Confidence interval I	pptx
Week 11	Mar 17	L7	Confidence interval II	pptx
Week 12	Mar 24	L8	Hypothesis test	pptx
Week 13	Mar 31	L9	Composite hypothesis test	pptx
Week 14	Apr 7	L10	Comparing populations	pptx
Week 15	Apr 14	L11	**Review**	pptx
Optional			Inference about variance	pptx

Lec 1 概率统计

> (Probability and) Statistics

1.1 概率与统计

> Probability v.s. statistics

Probability is a mathematical language for quantifying uncertainty.

关于概率论知识，见 `大二 term 1 ESTR 2018 概率论` .

In probability, we assume probability distribution is *known*

- A family of distributions
- The parameter(s) of the distribution

> 概率论是已知模型，用模型生成数据.

本课程用到的概率论知识：独立性、条件独立性、贝叶斯、基本模型

Independence

$$P(x_1,\ \ldots,\ x_n) = \prod_{i=1}^{n} P(x_i), \quad P(x_i|x_j) = P(x_i)$$

Conditional Independence

$$P(x_1,\ \ldots,\ x_n|y) = \prod_{i=1}^{n} P(x_i|y), \quad P(x_i|y, x_j) = P(x_i|y)$$

Bayes' Rule

$$P(x|y) = \frac{P(x)P(y|x)}{P(y)}$$

统计的中心法则

> The Central Dogma of Statistics

$data = independent\ samples$

We have samples of observed data, but don't know the underlining distribution

统计学是已知数据，未知分布，用统计推断知识判断最适用的模型.

> 然后用模型生成更多的未知数据（回到概率论）.

概率与统计是两块独立的知识，联系的关键是**中心极限定理**.

1.2 统计推断

> Descriptive statistics v.s. Inferential statistics

Descriptive statistics: use numbers to summarize and describe data

缺点：Do not involve generalization beyond the data at hand

> 不是本课关注内容.

本课关注统计推断.

1.2.1 经典推断

> Classical statistics

Parameters are considered as deterministic quantities that happen to unknown

Point estimation $\hat{\theta}$ with observed data x

1.2.2 贝叶斯推断

> Bayesian statistics

Parameters are considered as random variables with prior distributions

$\theta \sim f_\Theta$ or $p_\Theta(\theta)$

Bayes' rule

$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$

这里用的是概率密度的贝叶斯，见 `ESTR 2018 概率论 13.3 连续分布贝叶斯`.

准确地说，数据可能有四种情况，对应四种贝叶斯公式：

> 离散变量用概率值，连续变量用 PDF. 概率论只教了变量均离散或均连续的情况，但一个离散一个连续也可以写贝叶斯.
>
> 分母统一记为 $Z(x)$，无论是概率值 $P_X(x)$ 还是 PDF $f_X(x)$，都是常数. 这个常数来源于观测数据 x .

① Θ discrete, X discrete

$$P_{\Theta|X}(\theta|x) = \frac{P_\Theta(\theta)P_{X|\Theta}(x|\theta)}{\sum\limits_{\theta'} P_\Theta(\theta')P_{X|\Theta}(x|\theta')}$$

> 这里 θ' 是为了和 θ 区分.

② Θ discrete, X continuous

$$P_{\Theta|X}(\theta|x) = \frac{P_\Theta(\theta)f_{X|\Theta}(x|\theta)}{\sum\limits_{\theta'} P_\Theta(\theta')f_{X|\Theta}(x|\theta')}$$

③ Θ continuous, X discrete

$$f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta)P_{X|\Theta}(x|\theta)}{\int f_\Theta(\theta')P_{X|\Theta}(x|\theta')d\theta'}$$

> 证明?

④ Θ continuous, X continuous

$$f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta)f_{X|\Theta}(x|\theta)}{\int f_\Theta(\theta')f_{X|\Theta}(x|\theta')d\theta'}$$

Hypothesis testing

Confidence interval estimation

(2025.2.9) ③ 证明：

$$
\begin{aligned}
f_{\Theta|X}(\theta|x) &= \frac{d}{d\theta}P(\Theta \le \theta|X=x) \\
&= \frac{d}{d\theta}\frac{P(X=x,\Theta\le\theta)}{P(X=x)} \\
&= \frac{d}{d\theta}\frac{\int_{-\infty}^{\theta}P(X=x|\Theta=\theta)f_\Theta(\theta)d\theta}{P(X=x)} \\
&= \frac{f_\Theta(\theta)P_{X|\Theta}(x|\theta)}{\int f_\Theta(\theta')P_{X|\Theta}(x|\theta')d\theta'}
\end{aligned}
$$

④ 证明：

`ESTR 2018 概率论` 中给出了双连续变量条件 PDF 的定义：

$$
f_{X|Y}(x|y)=\frac{f_{XY}(x,y)}{f_Y(y)}
$$

该定义的合理性证明如下：

首先，对于条件而言，

$$
Y=y \Leftrightarrow y\le Y\le y+\Delta y
$$

> 并不是说这两个事件概率相等，而是它们作为条件，对于其他事件而言完全一致（在这两者之间切换，不影响其他事件基于它们的条件概率）.

小量分析：

$$
\begin{aligned}
P(x\le X\le x+\Delta x,\ y\le Y\le y+\Delta y)&\approx f_{XY}(x,y)\Delta x\Delta y \\
P(y\le Y\le y+\Delta y)&\approx f_Y(y)\Delta y
\end{aligned}
$$

条件概率：

$$
\begin{aligned}
P(x\le X\le x+\Delta x|Y=y)&=P(x\le X\le x+\Delta x|y\le Y\le y+\Delta y) \\
&\approx \frac{f_{XY}(x,y)\Delta x\Delta y}{f_Y(y)\Delta y} \\
&= \frac{f_{XY}(x,y)}{f_Y(y)}\Delta x
\end{aligned}
$$

条件概率密度：

$$
\begin{aligned}
f_{X|Y}(x|y)&=\lim_{\Delta x\to 0}\frac{P(x\le X\le x+\Delta x|Y=y)}{\Delta x} \\
&= \frac{f_{XY}(x,y)}{f_Y(y)}
\end{aligned}
$$

写出条件概率密度的两种形式：

$$
f_{X\Theta}(x,\theta)=f_{X|\Theta}(x|\theta)f_\Theta(\theta)=f_{\Theta|X}(\theta|x)f_X(x)
$$

整理得到

$$
f_{\Theta|X}(\theta|x)=\frac{f_\Theta(\theta)f_{X|\Theta}(x|\theta)}{\int f_\Theta(\theta')f_{X|\Theta}(x|\theta')d\theta'}
$$

分母由 $\int_\theta f_{\Theta|X}(\theta|x)d\theta=1$ 得到.

1.2.3 贝叶斯推断实例

Example 1：硬币是否公平?

假设有一枚硬币，想知道它是否公平（即抛掷时正面朝上的概率 θ 是否为 0.5）.

我们通过抛掷硬币 $N = 10$ 次，记录正面朝上的次数 x，并使用贝叶斯推断来估计 θ.

第一步：定义先验分布（Prior）

我们对硬币的公平性没有很强的先验知识（θ 可能是 $[0, 1]$ 之间任意值），假设硬币的正面概率 θ 在 $[0, 1]$ 之间均匀分布. 因此，选择一个均匀分布作为先验分布：

$$P(\theta) = 1, \quad \text{for } 0 \leq \theta \leq 1$$

> 这里的 P 是密度，不是概率值.
>
> 先验分布可基于领域知识，或使用非信息先验（如均匀分布）.
>
> 先验分布只是假设.

第二步：定义似然函数（Likelihood）

假设每次抛硬币是独立的，且正面朝上的概率为 θ. 这符合二项分布的假设. 于是，给定 θ，观测到 x 次正面朝上的概率为：

$$P(x|\theta) = \binom{N}{x} \theta^x (1-\theta)^{N-x}$$

> 这里的 P 是概率值，不是密度.

第三步：计算后验分布（Posterior）

根据贝叶斯定理：

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

> 这里的 $P(\theta|x), P(\theta)$ 是密度，$P(x|\theta), P(x)$ 是概率值.
>
> - $P(\theta)$：先验分布，已假设为均匀分布，即 $P(\theta) = 1$.
> - $P(x|\theta)$：似然函数，已假设为二项分布，即 $P(x|\theta) = \binom{N}{x}\theta^x(1-\theta)^{N-x}$.
> - $P(x)$：边缘似然. 这是一个归一化常数，可通过积分计算：$P(x) = \int_0^1 P(x|\theta)P(\theta)d\theta$

将 $P(\theta) = 1$ 和 $P(x|\theta)$ 代入，可得到后验分布的形式：

$$P(\theta|x) \propto \theta^x (1-\theta)^{N-x}$$

这实际上是一个 **Beta 分布** 的形式：

$$P(\theta|x) \sim Beta(x+1, N-x+1)$$

> 见 `1.2.5 Beta 分布`.
>
> 注意，和常见的 Beta-Bernoulli 不同，这里观测数据服从 Binomial.

第四步：用数据更新后验分布

假设抛硬币 $N = 10$ 次，观测到 $x = 7$ 次正面朝上. 代入后验分布公式：

$$P(\theta|x = 7) \propto \theta^7(1-\theta)^3$$

这对应于 $Beta(8, 4)$ 分布.

后验分布的期望可以计算为：

$$E[\theta|x] = \frac{\alpha}{\alpha+\beta} = \frac{8}{8+4} = \frac{2}{3}$$

因此，根据数据更新后，我们对 θ 的最佳估计是 $\frac{2}{3}$，即认为硬币更倾向于正面朝上.

第五步：贝叶斯推断的解释

- 先验分布：在观察数据前，我们相信硬币的正面概率 θ 是均匀分布的.
- 通过数据更新后验分布，并计算期望：观测到 7 次正面和 3 次反面后，我们的信念更新为 θ 更可能接近 $\frac{2}{3}$，并可以用 $Beta(8, 4)$ 来描述 θ 的不确定性.
- 可视化：后验分布 $P(\theta|x)$ 的形状可以反映数据对 θ 的影响. 例如：
  - 如果 $x = 5$，后验分布会集中在 $\theta = 0.5$ 附近.
  - 如果 $x = 10$，后验分布会更偏向 $\theta = 1$.
  - 如果 $N = 2$，数据较少，后验分布会更接近先验分布，反映出不确定性较大.

1.2.4 多观测数据

> Bayes' Rule for Multiple Random Variables

For Example 2，如果投掷多次并观测，逐次计算效率很低，可以用多变量贝叶斯公式计算：

$$\begin{aligned} f_{\Theta|X_1,\ldots,X_n}(\theta|x_1,\ldots,x_n) &= \frac{f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_n|\theta)f_\Theta(\theta)}{Z(x_1,\ldots,x_n)} \\ &\propto f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_n|\theta)f_\Theta(\theta) \\ &= f_{X_1|\Theta}(x_1|\theta)\cdots f_{X_n|\Theta}(x_n|\theta)f_\Theta(\theta) \end{aligned}$$

> 假设各个观测数据是条件独立的.
>
> 注意条件独立 $\neq$ 独立. 关于此公式的详细讨论，见 `2.2.2 硬币是否公平？`.

1.2.5 Beta 分布

Beta 分布是一种定义在 $[0,1]$ 上的连续概率分布. 在贝叶斯统计中适合作为参数（如概率）的先验分布. Beta 分布由两个正参数 α 和 β 控制. X 服从 Beta 分布，记作：

$$X \sim Beta(\alpha, \beta), \quad x \in [0,1], \quad \alpha > 0,\ \beta > 0$$

> 硬币模型中，通常用 Θ 来表示服从 Beta 分布的概率参数，而用 X 表示服从 Bernoulli 分布的成功次数（如正面朝上次数）.

概率密度函数：

$$f_X(x) = \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} & \text{for } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

实际模型中，Beta 分布多用于表示参数服从的分布，即

$$f_\Theta(\theta) = \begin{cases} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} & \text{for } 0 < \theta < 1 \\ 0 & \text{otherwise} \end{cases}$$

其中

$B(\alpha, \beta)$ 是 Beta 函数（归一化常数，Normalization Term）：

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

推导：

$$\begin{aligned} B(\alpha,\beta) &= \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx \\ &= \frac{1}{\alpha}\int_0^1 (1-x)^{\beta-1}d(x^\alpha) \\ &= \frac{1}{\alpha}[x^\alpha(1-x)^{\beta-1}\big|_0^1 + \int_0^1 x^\alpha(\beta-1)(1-x)^{\beta-2}dx] \\ &= \frac{1}{\alpha}\int_0^1 x^\alpha(\beta-1)(1-x)^{\beta-2}dx \\ &= \cdots \\ &= \frac{(\beta-1)(\beta-2)\cdots 1}{\alpha(\alpha+1)\cdots(\alpha+\beta-1)}\int_0^1 x^{\alpha+\beta-2}dx \\ &= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \end{aligned}$$

$\Gamma(\cdot)$ 是伽马函数（Gamma Function），是阶乘的推广：

$$\Gamma(n) = (n-1)! \quad \text{当 } n \text{ 为正整数}$$

> 见 1.2.7 伽马函数.

$X \sim Beta(\alpha, \beta)$ 的期望为

$$E[X] = \frac{\alpha}{\alpha+\beta}.$$

推导：

$$
\begin{aligned}
E[X] &= \int_0^1 x f_X(x) dx \\
&= \int_0^1 x \cdot \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} dx \\
&= \frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha}(1-x)^{\beta-1} dx \\
&= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} \\
&= \frac{\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)}}{\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}} \\
&= \frac{\alpha}{\alpha+\beta}
\end{aligned}
$$

> 这里顺便推导出 Beta 函数的一个性质：$B(\alpha+1,\beta) = \frac{\alpha}{\alpha+\beta} B(\alpha,\beta)$.

方差

$$
Var[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.
$$

> 待证明.

众数

$$
mode[X] = \frac{\alpha-1}{\alpha-1+\beta-1} \text{ when } \alpha,\beta > 1
$$

> 注意由于 X 是连续分布，这里的众数指概率密度函数最高点对应的 x 值.

推导:

已知概率密度函数

$$
f_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}
$$

令导数为 0，得

$$
\begin{aligned}
(\alpha-1)x^{\alpha-2}(1-x)^{\beta-1} - (\beta-1)x^{\alpha-1}(1-x)^{\beta-2} &= 0 \\
(\alpha-1)(1-x) &= (\beta-1)x \\
(\alpha-1) - (\alpha-1)x &= (\beta-1)x \\
\frac{\alpha-1}{\alpha-1+\beta-1} &= x
\end{aligned}
$$

验证最大值

$$
f'_X(x) = \frac{(\alpha-1)x^{\alpha-2}(1-x)^{\beta-1} - (\beta-1)x^{\alpha-1}(1-x)^{\beta-2}}{B(\alpha,\beta)}
$$

当 $x \in (0, \frac{\alpha-1}{\alpha-1+\beta-1})$ 时，$f'_X(x) > 0$；当 $x \in (\frac{\alpha-1}{\alpha-1+\beta-1}, 1)$ 时，$f'_X(x) < 0$. $x = \frac{\alpha-1}{\alpha-1+\beta-1}$ 时 $f_X(x)$ 取最大值.

因此，众数

$$mode[X] = \frac{\alpha - 1}{\alpha - 1 + \beta - 1} \text{ when } \alpha, \beta > 1$$

注意，判断一个连续随机变量 X 符合 Beta 分布，不需要得到定量的 PDF

$$f_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

只需要定性求出

$$f_X(x) \propto x^{\alpha-1}(1-x)^{\beta-1}$$

因为常数 $B(\alpha, \beta)$ 可以由定性关系结合概率公理（归一化）求出.

> 换句话说，只要符合正比关系就可以得到常数. 而 $B(\alpha, \beta)$ 本身又有公式，所以根本不需要计算，直接从定性关系的指数中就可以写出常数（瞪眼法）.

1.2.5.1 超参数

Beta 分布中的两个参数 α 和 β 在机器学习领域称为超参数（Hyper-parameter），指模型在训练前需要**人为指定**的参数，通常不能通过模型的训练过程自动学习得到.

> 关于超参数的作用效果，见 `2.3.2 常见共轭分布 ① Beta-Bernoulli`.

1.2.6 Gamma 分布

伽马分布是一种连续概率分布，常用于建模正值随机变量，尤其是涉及等待时间、分布形状和尺度的场景. 是许多分布（如卡方分布、指数分布）的推广.

Gamma 分布由两个正参数 α 和 β 控制. X 服从 Gamma 分布，记作：

$$X \sim Gamma(\alpha, \beta), \quad x > 0,\ \alpha > 0,\ \beta > 0$$

概率密度函数（PDF）：

$$f_X(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & x > 0 \\ 0 & x \le 0 \end{cases}$$

> $x > 0$：随机变量取值范围为正数.
>
> $\alpha > 0$：形状参数（Shape Parameter）.
>
> $\beta > 0$：尺度参数（Scale Parameter），有时也写为 $\frac{1}{\theta}$，其中 θ 是速率参数.
>
> $\Gamma(\alpha)$：伽马函数.

实际模型中，Gamma 分布多用于表示参数服从的分布，即

$$f_\Theta(\theta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} & \theta > 0 \\ 0 & \theta \le 0 \end{cases}$$

期望：

$$E[\Theta]=\int_0^{\infty}\frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha}e^{-\beta\theta}d\theta=\frac{\beta^{\alpha}}{\Gamma(\alpha)}\cdot\frac{\Gamma(\alpha+1)}{\beta^{\alpha+1}}=\frac{\alpha}{\beta}$$

方差：

$$Var[\Theta]=E[\Theta^2]-E^2[\Theta]=\frac{\alpha(\alpha+1)}{\beta^2}-\frac{\alpha^2}{\beta^2}=\frac{\alpha}{\beta^2}$$

1.2.7 伽马函数

> Gamma Function

① 定义

伽马函数是阶乘函数的推广，可以在实数和复数范围内定义.

$$\Gamma(z)=\int_0^{\infty}t^{z-1}e^{-t}dt$$

> 当 $Re(z)>0$ 时收敛.

其中，z 是复数变量，且满足实部 $Re(z)>0$.

当 n 是正整数时，

$$\Gamma(n)=(n-1)!$$

证明：当 n 是正整数时，

$$\begin{aligned}\Gamma(n)&=\int_0^{\infty}x^{n-1}e^{-x}dx\\&=-\int_0^{\infty}x^{n-1}d(e^{-x})\\&=-[x^{n-1}e^{-x}\big|_0^{\infty}-(n-1)\int_0^{\infty}x^{n-2}e^{-x}dx]\\&=(n-1)\Gamma(n-1)\\&=\cdots\\&=(n-1)!\Gamma(1)\\&=(n-1)!\end{aligned}$$

其中，

$$\Gamma(1)=\int_0^{\infty}e^{-x}dx=-e^{-x}\big|_0^{\infty}=1$$

② 性质

递推关系

$$\Gamma(z+1)=z\Gamma(z)$$

反射公式

$$\Gamma(z)\Gamma(1-z)=\frac{\pi}{sin(\pi z)}$$

乘法公式

$$\Gamma(x)\Gamma(x+\frac{1}{2})=2^{1-2x}\sqrt{\pi}\Gamma(2x)$$

斯特林公式

当 $|z|\to\infty$ 时，伽马函数的渐近公式为

$$\Gamma(z)\sim\sqrt{2\pi}z^{z-\frac{1}{2}}e^{-z}$$

对于非整数的正数 z，伽马函数提供了阶乘的推广：

$$\Gamma(\frac{1}{2})=\sqrt{\pi}$$

该值可用多种方法计算：

① 换元法

设 $x=t^2$，则 $dx=2tdt$

$$\begin{aligned}\Gamma(\frac{1}{2})&=\int_0^\infty x^{-\frac{1}{2}}e^{-x}dx\\&=2\int_0^\infty e^{-t^2}dt\\&=\sqrt{\pi}\end{aligned}$$

其中，高斯积分

$$\int_0^\infty e^{-t^2}dt=\frac{\sqrt{\pi}}{2}$$

> 证明见 附录 1.高斯积分.

② 反射公式

$$\Gamma(\frac{1}{2})\Gamma(1-\frac{1}{2})=\frac{\pi}{sin(\frac{1}{2}\pi)}=\pi\Rightarrow\Gamma(\frac{1}{2})=\sqrt{\pi}$$

③ 乘法公式

$$\Gamma(\frac{1}{2})\Gamma(\frac{1}{2}+\frac{1}{2})=\sqrt{\pi}\Gamma(2\times\frac{1}{2})\Rightarrow\Gamma(\frac{1}{2})=\sqrt{\pi}$$

Lec 2 贝叶斯推断

> Bayesian Statistical Inference
>
> 在 `1.1.3 贝叶斯推断实例` 亦有提及.

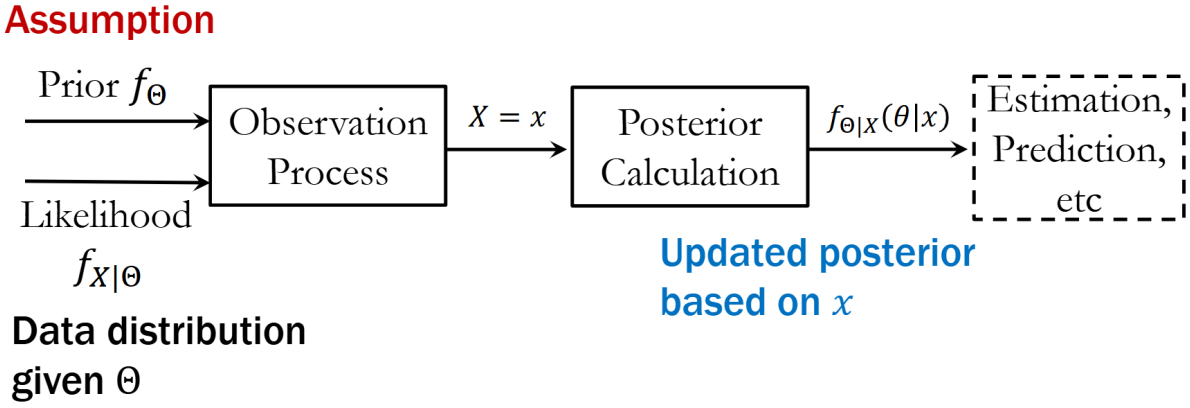
2.1 原理

> 贝叶斯法则，Bayes' Rule，是贝叶斯推断的核心.

由连续变量贝叶斯法则，可得

$$f_{\Theta|X}(\theta|x) \propto f_{\Theta}(\theta) f_{X|\Theta}(x|\theta)$$

> $f_{\Theta|X}(\theta|x)$：Posterior，后验分布
>
> $f_{\Theta}(\theta)$：Prior，先验分布
>
> $f_{X|\Theta}(x|\theta)$：Likelihood，似然函数
>
> 注意：后验的定义范围小于等于先验. 由于贝叶斯由乘法实现，先验未定义（PDF 值为 0）的部分，后验 PDF 也为 0；先验 PDF 不为 0 的部分，后验 PDF 可能由于观测数据的影响也取到 0，所以后验分布的定义范围是先验的子集.



关于后验分布的本质，见 `2.2.1 约会大作战！ （2025.2.9 思考）` .

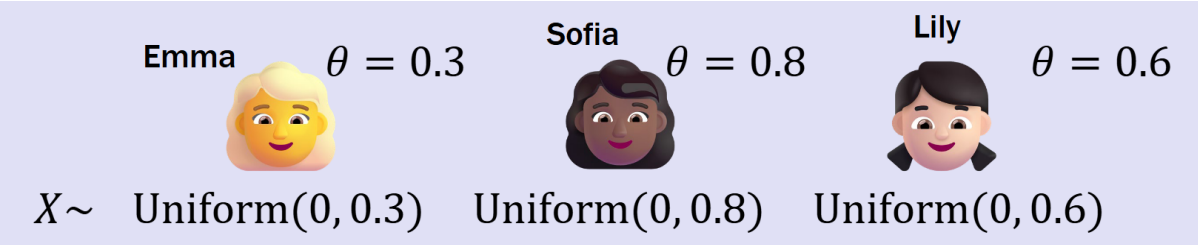
2.2 例子

2.2.1 约会大作战!

> 这个模型和概率论的两滴雨滴模型有点像.

John is waiting for Apple on their first date.

John 想对 Apple 的迟到时间建模. 根据以往经验，John 的前女友们迟到时间都服从均匀分布：



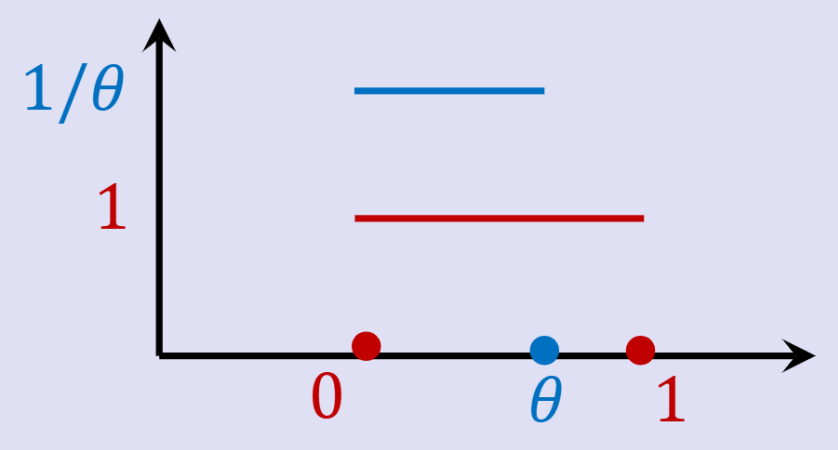
合理假设 Apple 的迟到时间也服从均匀分布：

$$X|\Theta \sim Uniform(0, \theta)$$

θ is a value of the random variable Θ. 根据以往经验，Θ 大致在 $(0, 1)$ 之间均匀取值，因此假设先验分布：

$$\Theta \sim Uniform(0, 1)$$

PDF 图像如下：



> 蓝色为迟到时间 $f_{X|\Theta}(x|\theta)$，红色为参数先验 $f_{\Theta}(\theta)$.

第一次观察：On her first date, Apple arrives $\frac{1}{2}$ hour late.

贝叶斯推断：

$$f_{\Theta|X}(\theta|\frac{1}{2}) \propto f_{\Theta}(\theta) f_{X|\Theta}(\frac{1}{2}|\theta) = \frac{1}{\theta} \quad \text{if } \frac{1}{2} \leq \theta \leq 1$$

> 注意，贝叶斯是乘法，先验取 0 后验一定取 0，所以 θ 范围只会缩小或者不变. 导致范围缩小的因素是似然，如果似然对 θ 作出额外的限制，后验的范围就会小于先验.
>
> 例如，这里 $f_{X|\Theta}(x|\theta)$ 在 $0 \leq x \leq \theta$ 时才有非零值，相当于给 θ 添加了额外限制 $\theta \geq x$，随着观察数据累加，θ 下限可能越来越高.
>
> 从题意来理解，θ 是 X 均匀分布的上限，如果观测到某个 x，显然 θ 不能低于该 x，否则该观测数据超出范围，不可能观测到.

> （2025.2.9 思考）但是，使用观测数据来预测之后的新数据是否合理，值得商榷. 例如这里第一次观测到 $\frac{1}{2}$，导致之后 Θ 永远取不到 $\frac{1}{2}$ 以下（似然带来的范围缩小），这显得有些荒谬. 原因是贝叶斯更新本质是**观测完之后，反过来马后炮地猜，让这个已观测数据出现的 Θ 分布是什么样的**，而我们使用这个更新后的分布去预测/推理新数据，实际上是假设新数据的参数分布和观测数据的后验分布一样，而这个假设不一定成立，尤其是似然函数带有硬性阈值时，仅一两个数据点出现异常也可能导致后验分布发生剧烈变化（例如这里 $x \leq \theta$ 的限制可能突然剧烈提高 θ 的下限），此时应该对异常数据进行处理，而不是把这个异常带来的后验更新直接用于进一步学习. 这种对初始数据或极端数据异常敏感的行为被一些文献称为贝叶斯推断的「脆弱性 (Brittleness)」.
>
> 以该约会模型为例，如果新女朋友第一次约会因为某些意外迟到了很久（如 0.9 hour），但是从第二次开始迟到时间就均匀分布在 $0 \sim 0.1$ hour，参数分布也永远无法拉下 0.9，随着不断更新后验 PDF 会越来越挤向 0.9，但是就是无法突破，因为被这个设计有问题的似然函数硬性锁死了取值范围.
>
> 还有个更扯淡的，这里似然函数不仅过度学习了参数的分布，好像也完全无法捕捉到数据分布信息，比如连续迟到十次 1 小时，他还会认为迟到时间是 $0 \sim 1$ 均匀分布. 也就是说只有数据的实际分布完全吻合假设的似然（均匀分布），才能进行有效学习，一旦观测数据出现了不符合似然的分布情况，学习就没有意义了. 这揭示了选择似然函数的原则之一——考虑鲁棒性，即观测数据中可能有异常值，或模型假设可能不成立时，尽量选择具有较厚尾部（heavy-tailed）的分布（如 t 分布）来作为似然函数，以增强推断的鲁棒性. 也可以理解为，异常观测值能被其他正常观测数据很好地纠正，使其不至于影响全局.

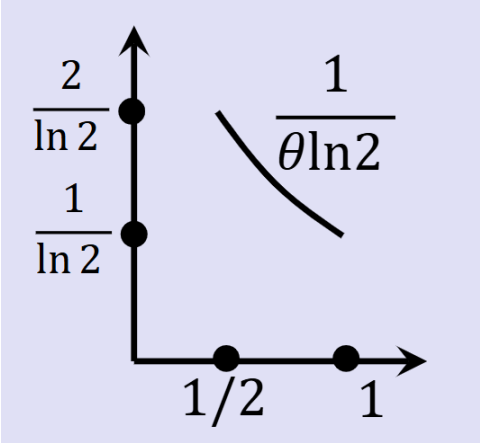
归一化常数：

$$Z(x) = \int_{\frac{1}{2}}^{1} \frac{1}{\theta} d\theta = \ln 2$$

后验分布：

$$f_{\Theta|X}(\theta|\frac{1}{2}) = \frac{1}{\theta \ln 2} \quad \text{if } \frac{1}{2} \leq \theta \leq 1$$

PDF 图像如下：



第二次观察：On her next date, Apple arrives $\frac{1}{4}$ hour late.

贝叶斯推断：

$$\begin{aligned} f_{\Theta|X_1,X_2}(\theta|\frac{1}{2},\frac{1}{4}) &\propto f_{X_2|\Theta}(\frac{1}{4}|\theta) f_{\Theta|X_1}(\theta|\frac{1}{2}) \\ &= \frac{1}{\theta^2 \ln 2} \\ &\propto \frac{1}{\theta^2} \qquad \text{if } \frac{1}{2} \leq \theta \leq 1 \end{aligned}$$

> 第一个正比号，类似 `2.2.2 硬币是否公平？` 中第一个等号的推导.

归一化常数：

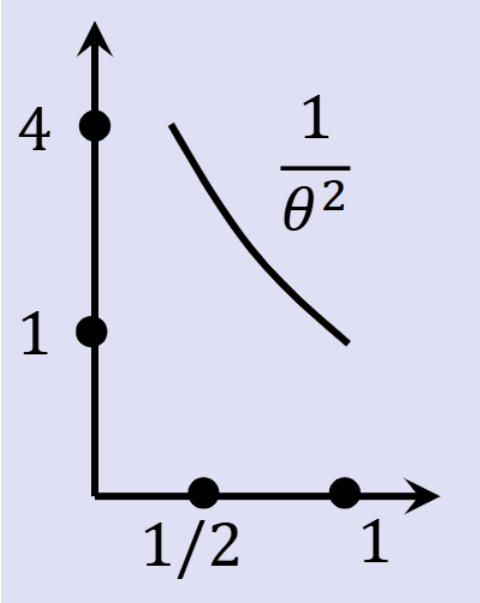
$$Z(x_1, x_2) = \int_{\frac{1}{2}}^{1} \frac{1}{\theta^2} d\theta = 1$$

> 这里的归一化常数，是将后验的所有常数部分移到分母后得到. 这样便于积分计算，按照严格定义，应该是原本常数的 $\ln 2$ 倍.

后验分布：

$$f_{\Theta|X_1,X_2}(\theta|\frac{1}{2}, \frac{1}{4}) = \frac{1}{\theta^2} \quad \text{if } \frac{1}{2} \le \theta \le 1$$

PDF 图像如下：



前三次观察：On her first 3 dates, Apple is late by $\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$ hours.

贝叶斯推断：

$$\begin{aligned} f_{\Theta|X_1,X_2,X_3}(\theta|\frac{1}{2}, \frac{1}{4}, \frac{1}{4}) &\propto f_{X_1|\Theta}(\frac{1}{2}|\theta) f_{X_2|\Theta}(\frac{1}{4}|\theta) f_{X_3|\Theta}(\frac{1}{4}|\theta) f_{\Theta}(\theta) \\ &= \frac{1}{\theta^3} \qquad \text{if } \frac{1}{2} \le \theta \le 1 \end{aligned}$$

> 第一个正比号，类似 `2.2.2 硬币是否公平？` 中的「同时更新」.

归一化常数：

$$Z(x_1, x_2, x_3) = \int_{\frac{1}{2}}^{1} \frac{1}{\theta^3} d\theta = \frac{3}{2}$$

后验分布：

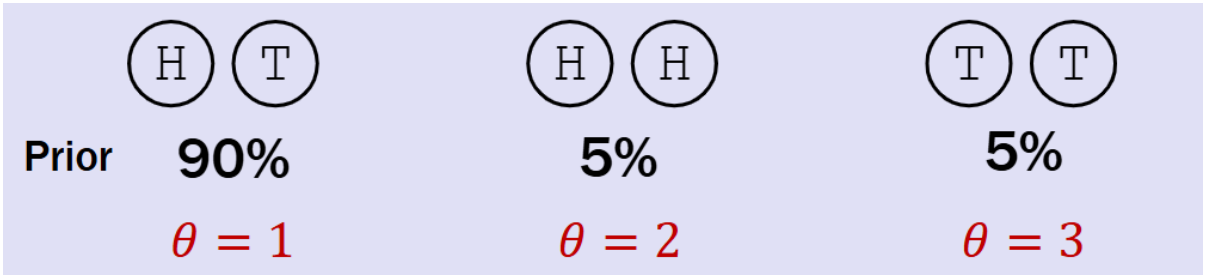
$$f_{\Theta|X_1,X_2,X_3}(\theta|\frac{1}{2}, \frac{1}{4}, \frac{1}{4}) = \frac{2}{3\theta^3} \quad \text{if } \frac{1}{2} \le \theta \le 1$$

2.2.2 硬币是否公平?

① 三类硬币模型

> 类似 `1.2.3 贝叶斯推断实例` 的 Example 1.
>
> 注意：伯努利是离散分布，注意区分离散和连续的符号.

A coin might be of the following type:



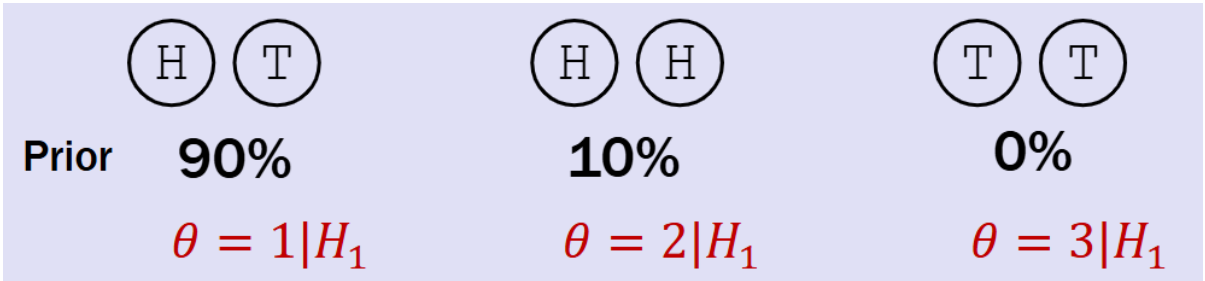
> 先验分布.

You flip a H （Head）. How do you adjust your beliefs (priors)?

$$
\begin{aligned}
P(\theta=1|H_1) &= \frac{P(H_1|\theta=1)P(\theta=1)}{Z(H_1)} = \frac{0.5\times 0.9}{Z(H_1)} = \frac{0.45}{Z(H_1)} \\
P(\theta=2|H_1) &= \frac{P(H_1|\theta=2)P(\theta=2)}{Z(H_1)} = \frac{1\times 0.05}{Z(H_1)} = \frac{0.05}{Z(H_1)} \\
P(\theta=3|H_1) &= \frac{P(H_1|\theta=3)P(\theta=3)}{Z(H_1)} = \frac{0\times 0.05}{Z(H_1)} = \frac{0}{Z(H_1)} \\
Z(H_1) &= 0.45+0.05+0=0.5
\end{aligned}
$$

后验分布

$$P(\theta=1|H_1)=0.9 \qquad P(\theta=2|H_1)=0.1 \qquad P(\theta=3|H_1)=0$$



You flip another H. How do you readjust?

> 待完成.
>
> 2025.1.29 已完成.

$$
\begin{aligned}
P(\theta=1|H_2H_1) &= \frac{P(H_2|\theta=1,\cancel{H_1})P(\theta=1|H_1)}{Z(H_2,H_1)} = \frac{0.45}{Z(H_2,H_1)} \\
P(\theta=2|H_2H_1) &= \frac{P(H_2|\theta=2,\cancel{H_1})P(\theta=2|H_1)}{Z(H_2,H_1)} = \frac{0.1}{Z(H_2,H_1)} \\
P(\theta=3|H_2H_1) &= \frac{P(H_2|\theta=3,\cancel{H_1})P(\theta=3|H_1)}{Z(H_2,H_1)} = \frac{0}{Z(H_2,H_1)} \\
Z(H_2,H_1) &= 0.45+0.1+0=0.55
\end{aligned}
$$

> 划掉 H_1，因为 H_2, H_1 在 $\theta=1$ 的条件下独立.
>
> 注意，条件独立和独立是不同概念. 这里 H_2 和 H_1 只是条件独立，但不独立.
>
> 关于条件独立，见 `ESTR 2018 概率论 4.8 条件独立`.
>
> 至于第一个等号为什么能取等，见下方推导.

$$
\begin{aligned}
P(\theta=1|H_2H_1) &= \frac{0.45}{0.55} \approx 0.82 \\
P(\theta=1|H_2H_1) &= \frac{0.1}{0.55} \approx 0.18 \\
P(\theta=3|H_2H_1) &= \frac{0}{0.55} = 0
\end{aligned}
$$

第二次更新时，

$$
P(\theta=1|H_2H_1) = \frac{P(H_2|\theta=1,\cancel{H_1})P(\theta=1|H_1)}{Z(H_2,H_1)} = \frac{0.45}{Z(H_2,H_1)}
$$

第一个等号推导如下：

$$
\begin{aligned}
P(\theta=1|H_2H_1) &= \frac{P(\theta=1,H_1,H_2)}{P(H_2H_1)} \\
&= \frac{P(H_2|\theta=1,\cancel{H_1})P(\theta=1|H_1)P(H_1)}{P(H_2H_1)} \\
&= \frac{P(H_2|\theta=1,\cancel{H_1})P(\theta=1|H_1)}{\frac{P(H_2H_1)}{P(H_1)}} \\
&= \frac{P(H_2|\theta=1,\cancel{H_1})P(\theta=1|H_1)}{\frac{P(H_2H_1)}{P(H_1)}} \\
&= \frac{P(H_2|\theta=1)P(\theta=1|H_1)}{Z(H_2,H_1)}
\end{aligned}
$$

其中，

$$
Z(H_2,H_1) = \frac{P(H_2H_1)}{P(H_1)} = P(H_2|H_1) \neq P(H_2)
$$

贝叶斯推断有两种模式. 一种是逐步更新，参数一直在变；还有一种是攒着一堆观测结果，然后同时更新，所有观测的似然都用初始参数分布. 两个模式完全等效，为什么?

对于 n 次观测结果，

逐步更新

$$
\begin{aligned}
P(\theta=1|x_1,\dots,x_n) &= \frac{P(\theta=1,x_1,\dots,x_n)}{P(x_1,\dots,x_n)} \\
&= \frac{P(x_n|\theta=1,\cancel{x_1,\dots,x_{n-1}})P(\theta=1|x_1,\dots,x_{n-1})P(x_1,\dots,x_{n-1})}{P(x_1,\dots,x_n)} \\
&= \frac{P(x_n|\theta=1,\cancel{x_1,\dots,x_{n-1}})P(\theta=1|x_1,\dots,x_{n-1})}{\frac{P(x_1,\dots,x_n)}{P(x_1,\dots,x_{n-1})}}
\end{aligned}
$$

同时更新

$$
\begin{aligned}
P(\theta=1|x_1,\dots,x_n) &= \frac{P(x_1,\dots,x_n|\theta=1)P(\theta=1)}{P(x_1,\dots,x_n)} \\
&= \frac{P(x_1|\theta=1)\cdots P(x_n|\theta=1)P(\theta=1)}{P(x_1,\cdots,x_n)}
\end{aligned}
$$

两种方式都假设观测结果在 θ 条件下独立.

> 条件独立 $\neq$ 独立. 即贝叶斯推断中，各次观测不独立. 提到的独立都是条件独立.

而逐步更新通过递归计算，最终可以得到和同时更新一样的结果，所以二者等价.

> 2025.1.29 随想：连续推断便于计算，但逐次推断便于中断和继续训练，以及实时更新.

② 连续硬币模型

> Beta-Bernoulli 模型.

A coin of unknown bias flips H,T,T. What is the bias?

抛硬币服从伯努利分布：

$$
X|\Theta \sim Bernoulli(\theta)
$$

θ is a value of the random variable Θ. 假设先验分布：

$$
\Theta \sim Uniform(0,1)
$$

> 即 $Beta(1,1)$.

贝叶斯推断：

$$
\begin{aligned}
f_{\Theta|X_1,X_2,X_3}(\theta|H,T,T) &\propto P_{X_1|\Theta}(H|\theta)P_{X_2|\Theta}(T|\theta)P_{X_3|\Theta}(T|\theta)f_\Theta(\theta) \\
&= \theta(1-\theta)^2 \qquad \text{if } 0\le\theta\le 1
\end{aligned}
$$

归一化常数：

$$
Z(x_1,x_2,x_3) = \int_0^1 \theta(1-\theta)^2 d\theta = \frac{1}{12}
$$

后验分布：

$$
f_{\Theta|X_1,X_2,X_3}(\theta|H,T,T) = 12\theta(1-\theta)^2 \quad \text{if } 0\le\theta\le 1
$$

> 即 $Beta(2,3)$.

2.2.3 iPhone 销售模型

At an Apple store, the number of iPhones sold per day is modeled as a Poisson distribution with mean θ. θ is a value of the random variable Θ. Suppose the prior distribution of Θ is $Gamma(3,2)$. Let X be the number of iPhones sold in a specific day. If $X=3$ is observed, what is the updated distribution of Θ?

似然：$X|\Theta \sim Poisson(\theta)$

$$P_{X|\Theta}(x|\theta)=\begin{cases}\frac{e^{-\theta}\theta^x}{x!} & \text{for } x=0,1,2,\dots\\ 0 & \text{otherwise}\end{cases}$$

先验：$\Theta \sim Gamma(3,2)$

$$f_{\Theta}(\theta)=\begin{cases}\frac{2^3}{\Gamma(3)}\theta^{3-1}e^{-2\theta} & \theta>0\\ 0 & \theta\le 0\end{cases}$$

> 先验分布超参数的含义：2 天平均卖 3 台.

贝叶斯推断：

$$\begin{aligned}f_{\Theta|X}(\theta|3) &\propto P_{X|\Theta}(3|\theta)f_{\Theta}(\theta)\\ &\propto \theta^5 e^{-3\theta} \qquad \text{if } \theta>0\end{aligned}$$

> 即 $Gamma(6,3)$.

归一化常数：

$$Z(x)=\int_0^{\infty}\theta^{6-1}e^{-3\theta}d\theta=\frac{\Gamma(6)}{3^6}$$

后验：

$$f_{\Theta|X}(\theta|3)=\begin{cases}\frac{3^6}{\Gamma(6)}\theta^{6-1}e^{-3\theta} & \theta>0\\ 0 & \theta\le 0\end{cases}$$

结论：Gamma 是 Poisson 的共轭先验.

If the number of iPhones sold per hour follows a Poisson distribution with unknown mean θ, then the time between two successive iPhones sold follow an exponential distribution with parameter θ.

> θ 是平均速率，台/小时. 泊松分布和指数分布共用参数.

Suppose the prior distribution of Θ is $Gamma(1,2)$. Let X be the time interval (in hour) between successive iPhones sold.

iPhone	Time
iPhone 14	9:00 am
iPhone 14	10:30 am
iPhone 14 Plus	12:30 pm
iPhone 14 Pro	3:00 pm

$$X_1 = 1.5$$
$$X_2 = 2$$
$$X_3 = 2.5$$

What is the updated distribution of Θ?

似然：$X|\Theta \sim Exp(\theta)$

$$f_{X|\Theta}(x|\theta) = \begin{cases} \theta e^{-\theta x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

先验：$\Theta \sim Gamma(1,2)$

$$f_{\Theta}(\theta) = \begin{cases} \frac{2^1}{\Gamma(1)} \theta^{1-1} e^{-2\theta} & \theta > 0 \\ 0 & \theta \leq 0 \end{cases}$$

贝叶斯推断：

$$\begin{aligned} f_{\Theta|X_1,X_2,X_3}(\theta|1.5,2,2.5) &\propto f_{X_1|\Theta}(1.5|\theta) f_{X_2|\Theta}(2|\theta) f_{X_3|\Theta}(2.5|\theta) f_{\Theta}(\theta) \\ &\propto \theta^3 e^{-8\theta} \qquad \text{if } \theta > 0 \end{aligned}$$

> 即 $Gamma(4,8)$.

归一化常数：

$$Z(x_1,x_2,x_3) = \int_0^{\infty} \theta^{4-1} e^{-8\theta} d\theta = \frac{\Gamma(4)}{8^4}$$

后验：

$$f_{\Theta|X_1,X_2,X_3}(\theta|1.5,2,2.5) = \begin{cases} \frac{8^4}{\Gamma(4)} \theta^{4-1} e^{-8\theta} & \theta > 0 \\ 0 & \theta \leq 0 \end{cases}$$

结论：Gamma 也是 Exponential 的共轭先验.

2.3 共轭分布

> Conjugate Distributions

回顾连续变量的贝叶斯法则

$$f_{\Theta|X}(\theta|x) = \frac{f_{\Theta}(\theta) f_{X|\Theta}(x|\theta)}{Z(x)}$$

其中，归一化常数 $Z(x)$ 满足

$$Z(x) = \int f_{\Theta}(\theta') f_{X|\Theta}(x|\theta') d\theta'$$

我们发现，得到定性关系

$$f_{\Theta|X}(\theta|x) \propto f_{\Theta}(\theta) f_{X|\Theta}(x|\theta)$$

并不困难；但是，要定量计算后验分布 $f_{\Theta|X}(\theta|x)$，就要在每次使用贝叶斯时计算积分 $\int f_{\Theta}(\theta') f_{X|\Theta}(x|\theta') d\theta'$，而这个积分是很难算的.

既然我们计算积分的最终目的是得到后验分布 $f_{\Theta|X}(\theta|x)$ 的定量公式，那么可否找到一个方法，能够绕过积分计算，通过结合似然（当前观测）和先验（大脑直觉、先前观测），直接得到后验？

> 瞪眼法，极大简化计算.

对于一些常用模型，引入「共轭分布」概念，能很好解决这个问题.

2.3.1 定义

$$f_{\Theta|X}(\theta|x) \propto f_{\Theta}(\theta) f_{X|\Theta}(x|\theta)$$

如果后验分布 $f_{\Theta|X}(\theta|x)$ 与先验分布 $f_{\Theta}(\theta)$ 的形式相同（属于同一个分布族），则称二者为**共轭分布**. 此时先验分布 $f_{\Theta}(\theta)$ 称为似然函数 $f_{X|\Theta}(x|\theta)$ 的**共轭先验**.

If the posterior distribution $f_{\Theta|X}(\theta|x)$ is in the same probability distribution family as the prior distribution $f_{\Theta}(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $f_{X|\Theta}(x|\theta)$.

2.3.2 常见共轭分布

似然函数（数据分布）	先验	后验
伯努利分布 / 二项分布	Beta 分布	Beta 分布
泊松分布	Gamma 分布	Gamma 分布
正态分布（已知方差）	正态分布	正态分布
正态分布（未知方差）	正态-逆伽马分布	正态-逆伽马分布
多项分布	Dirichlet 分布	Dirichlet 分布
指数分布	Gamma 分布	Gamma 分布

① Beta-Bernoulli

> 见 `ESTR 额外课程 Lec 2 常用概率分布 2.1 Beta-Bernoulli`、`2.2.2 硬币是否公平？ ② 连续硬币模型`.

Suppose $X_1, \ldots, X_n$ form a random sample from Bernoulli distribution with an unknown parameter θ $(0 < \theta < 1)$. If the prior distribution $f_{\Theta}(\theta)$ is the Beta distribution $Beta(\alpha, \beta)$ ($\alpha, \beta > 0$) , then the posterior distribution $f_{\Theta|X_1,\ldots,X_n}(\theta|x_1, \ldots, x_n)$ given observed data $\{x_i\}_{i=1}^n$, $\sum_{i=1}^n x_i = k$ is also the Beta distribution $Beta(\alpha + k, \beta + n - k)$.

> 关于 Beta-Binomial，见 `1.2.3 贝叶斯推断实例 Example 1`.

> 关于 Beta 分布，见 `1.2.5 Beta 分布`.

证明：

观测数据 $X_1, \dots, X_n$ 满足 $\sum_{i=1}^{n} x_i = k$（抛 n 次，k 次为正）

> 注意这里和二项不同，因为需要满足特定顺序（算概率不用乘二项式系数）.

先验分布 $\Theta \sim Beta(\alpha, \beta)$，概率密度函数 $f_\Theta(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$

数据分布 $X \sim Bernoulli(\theta)$，概率密度函数（似然函数）$f_{X|\Theta}(x|\theta) = \theta^X(1-\theta)^{1-X}$

后验分布 $\Theta|X_1, \dots, X_n \sim$ 未知分布，概率密度函数

$$
\begin{aligned}
f_{\Theta|X_1,\dots,X_n}(\theta|x_1,\dots,x_n) &= \frac{f_{X_1,\dots,X_n|\Theta}(x_1,\dots,x_n|\theta)f_\Theta(\theta)}{Z(x_1,\dots,x_n)} \\
&\propto f_{X_1,\dots,X_n|\Theta}(x_1,\dots,x_n|\theta)f_\Theta(\theta) \\
&= f_{X_1|\Theta}(x_1|\theta)\cdots f_{X_n|\Theta}(x_n|\theta)f_\Theta(\theta) \\
&\propto \theta^{\alpha-1+k}(1-\theta)^{\beta-1+n-k}
\end{aligned}
$$

> 见 `1.1.4 多观测数据`.

与 Beta 分布 PDF 形式一致，只是参数发生了更新：

$$
f_{\Theta|X_1,\dots,X_n}(\theta|x_1,\dots,x_n) = \frac{\theta^{(\alpha+k)-1}(1-\theta)^{(\beta+n-k)-1}}{const}
$$

其中

$$
\begin{aligned}
const &= \int_0^1 \theta^{(\alpha+k)-1}(1-\theta)^{(\beta+n-k)-1}d\theta \\
&= B(\alpha+k, \beta+n-k) \\
&= \frac{\Gamma(\alpha+k)\Gamma(\beta+n-k)}{\Gamma(\alpha+\beta+n)}
\end{aligned}
$$

> 最后一个等号的推导，见 `1.2.5 Beta 分布`.

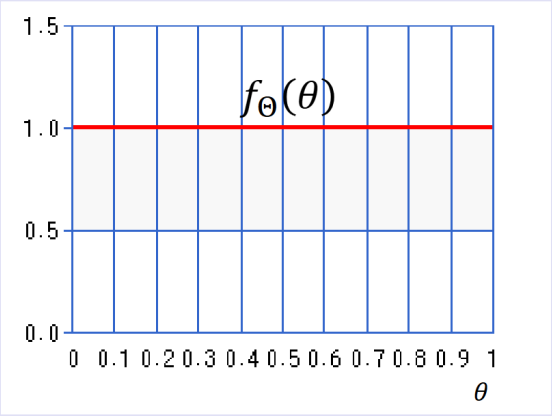
综上，后验分布 $\Theta|X_1, \dots, X_n \sim Beta(\alpha+k, \beta+n-k)$.

> 和先验分布属于同一个分布族，因此 Beta 分布是 Bernoulli 的共轭先验.

示例：

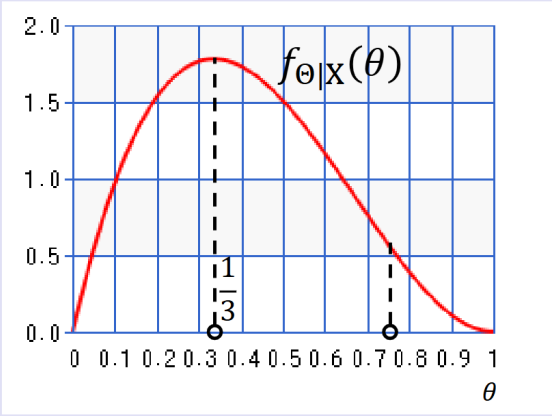
Observation: H, T, T

Prior $\Theta \sim \text{Uniform}(0,1)$



$\text{Beta}(1,1)$

Posterior $\Theta|X \sim \text{Beta}(2,3)$



$\text{Beta}(2,3)$

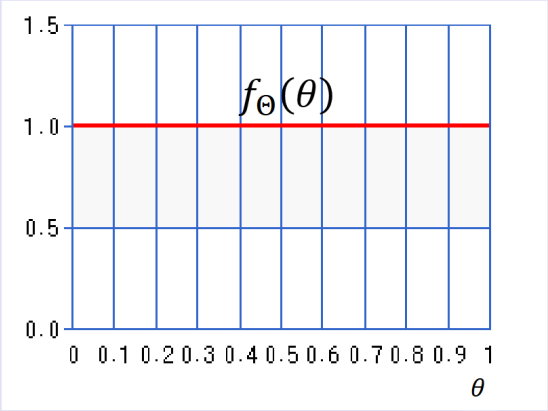
$$\text{mode}[\theta] = \frac{\alpha-1}{\alpha-1+\beta-1} \text{ when } \alpha,\beta > 1$$

众数推导见 `1.2.5 Beta 分布`.

注意 $0 \sim 1$ 之间的均匀分布是 Beta 分布的一种特殊情况，即 $Beta(1,1)$.

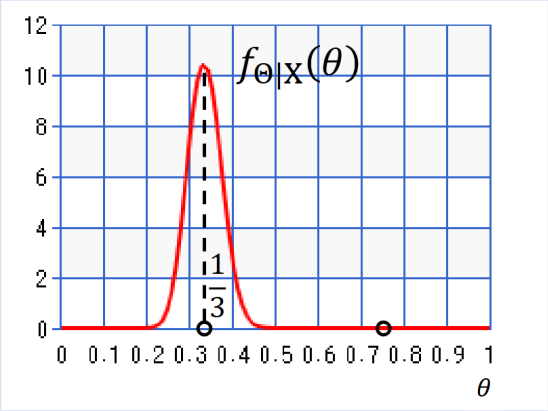
Observation: 50Hs, 100Ts

Prior $\Theta \sim \text{Uniform}(0,1)$



$\text{Beta}(1,1)$

Posterior $\Theta|X \sim \text{Beta}(51,101)$

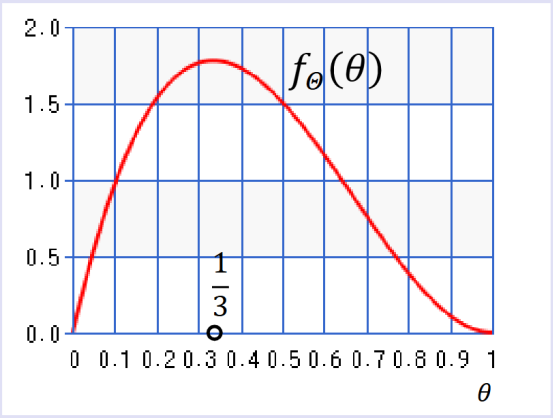


$\text{Beta}(51,101)$

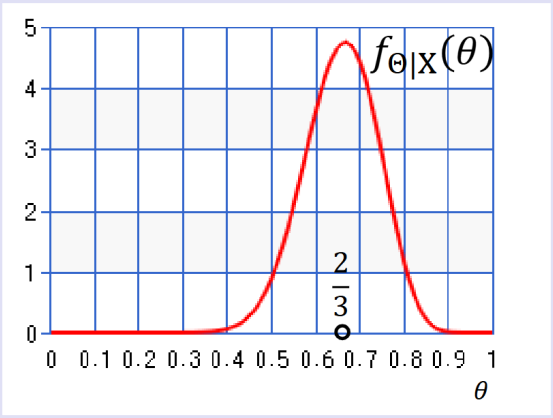
$$\text{mode}[\theta] = \frac{\alpha-1}{\alpha-1+\beta-1} \text{ when } \alpha,\beta > 1$$

Observation: 19H 8T

Prior $\Theta \sim Beta(2,3)$ **Posterior** $\Theta|X \sim Beta(21,11)$



$Beta(2,3)$



$Beta(21,11)$

$$\text{mode}[\theta] = \frac{\alpha-1}{\alpha-1+\beta-1} \text{ when } \alpha,\beta > 1$$

Do hyperparameters matter?

- If there are a lot of observed data samples, i.e., $h \gg \alpha,\ t \gg \beta$, then $Beta(\alpha+h, \beta+t) \approx Beta(h,t)$.

> The posterior mainly depends on the observed data.

- If the size of observed data samples is small, then the hyperparameters (α and β) of prior play an important role on the posterior

② Gamma-Poisson

Suppose $X_1, \ldots, X_n$ form a random sample from Poisson distribution with an unknown mean θ ($\theta > 0$) . If the prior distribution $f_\Theta(\theta)$ is the Gamma distribution $Gamma(\alpha,\beta)$ $(\alpha,\beta>0)$, then the posterior distribution $f_{\Theta|X_1,\ldots,X_n}(\theta|x_1,\ldots,x_n)$ given observed data $\{x_i\}_{i=1}^n$, $\sum_{i=1}^n x_i = k$ is also the Gamma distribution $Gamma(\alpha+k, \beta+n)$.

> 第一个超参数是发生次数的累计，第二个超参数是观测时间的累计.

Example: iPhone 销售

> 见 `2.2.3 iPhone 销售模型` .

③ Gamma-Exponential

Suppose $X_1, \ldots, X_n$ form a random sample from Exponential distribution with an unknown mean θ $(\theta > 0)$. If the prior distribution $f_\Theta(\theta)$ is the Gamma distribution $Gamma(\alpha, \beta)$ $(\alpha, \beta > 0)$, then the posterior distribution $f_{\Theta|X_1,\ldots,X_n}(\theta|x_1, \ldots, x_n)$ given observed data $\{x_i\}_{i=1}^n$, $\sum_{i=1}^n x_i = k$ is also the Gamma distribution $Gamma(\alpha + n, \beta + k)$.

> 第一个超参数是发生次数的累计，第二个超参数是观测时间的累计.

Example: iPhone 销售

> 见 `2.2.3 iPhone 销售模型`.

④ Normal-Normal

Suppose $X_1, \ldots, X_n$ form a random sample from Normal distribution with an unknown mean μ and a known variance σ^2 $(\sigma^2 > 0)$. If the prior distribution $f_{\mathrm{M}}(\mu)$ is the Normal distribution $N(\mu_0, \sigma_0^2)$, then the posterior distribution $f_{\mathrm{M}|X_1,\ldots,X_n}(\mu|x_1, \ldots, x_n)$ given observed data $\{x_i\}_{i=1}^n$, $\sum_{i=1}^n x_i = k$ is also the Normal distribution $N(\mu', \sigma'^2)$, where

$$\mu' = \frac{\sigma^2\mu_0 + \sigma_0^2 k}{\sigma^2 + n\sigma_0^2} \qquad \sigma'^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

> 待解决：如果两个参数都未知呢?

Special case: both σ_0^2 and σ^2 are 1

$$\mu' = \frac{\mu_0 + k}{n + 1} \qquad \sigma'^2 = \frac{1}{n + 1}$$

Example: A $N(\theta, 1)$ random variable takes value 3.97. θ is a value of the random variable Θ. Θ follows a standard normal. What is the posterior of Θ?

似然：$X|\Theta \sim N(\theta, 1)$

$$f_{X|\Theta}(x|\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$$

先验：$\Theta \sim N(0, 1)$

$$f_\Theta(\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\theta^2}$$

贝叶斯推断：

$$
\begin{aligned}
f_{\Theta|X}(\theta|x) &\propto f_{X|\Theta}(x|\theta)f_{\Theta}(\theta) \\
&\propto e^{-\frac{1}{2}[(x-\theta)^2+\theta^2]} \\
&\propto e^{-\frac{1}{2}(\sqrt{2}\theta-\frac{\sqrt{2}}{2}x)^2} \\
&= e^{-\frac{1}{2}(\frac{\theta-\frac{x}{2}}{\frac{1}{\sqrt{2}}})^2}
\end{aligned}
$$

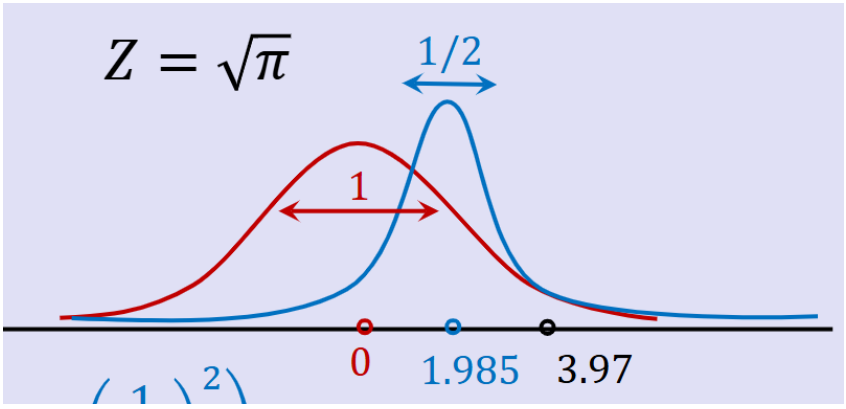
> 即 $N(\frac{x}{2},(\frac{1}{\sqrt{2}})^2)$.

归一化常数：

$$
Z(x)=\int_{-\infty}^{\infty}e^{-\frac{1}{2}(\frac{\theta-\frac{x}{2}}{\frac{1}{\sqrt{2}}})^2}d\theta=\sqrt{2\pi\sigma^2}=\sqrt{\pi}
$$

后验：$\Theta|X\sim N(\frac{x}{2},(\frac{1}{\sqrt{2}})^2)$

$$
f_{\Theta|X}(\theta|3.97)=\frac{1}{\sqrt{\pi}}e^{-\frac{1}{2}(\frac{\theta-1.985}{\frac{1}{\sqrt{2}}})^2}
$$



Example: Three independent $N(\theta,1)$ random variables take values $3.97,4.09,3.11$. What is Θ?

似然：$X_i|\Theta\sim N(\theta,1)$

$$
f_{X_i|\Theta}(x_i|\theta)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x_i-\theta)^2}
$$

先验：$\Theta\sim N(0,1)$

$$
f_{\Theta}(\theta)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\theta^2}
$$

贝叶斯推断：$x_1=3.97,\ x_2=4.09,\ x_3=3.11$

$$
\begin{aligned}
f_{\Theta|X_1,X_2,X_3}(\theta|x_1,x_2,x_3) &\propto f_{X_1|\Theta}(x_1|\theta)f_{X_2|\Theta}(x_2|\theta)f_{X_3|\Theta}(x_3|\theta)f_{\Theta}(\theta) \\
&\propto e^{-\frac{1}{2}[(x_1-\theta)^2+(x_2-\theta)^2+(x_3-\theta)^2+\theta^2]} \\
&\propto e^{-\frac{1}{2}(2\theta-\frac{x_1+x_2+x_3}{2})^2} \\
&= e^{-\frac{1}{2}(\frac{\theta-\frac{x_1+x_2+x_3}{4}}{\frac{1}{2}})^2}
\end{aligned}
$$

> 即 $N(\frac{x_1+x_2+x_3}{4},\frac{1}{4})$.

法二：由 special case, when both σ_0^2 and σ^2 are 1,

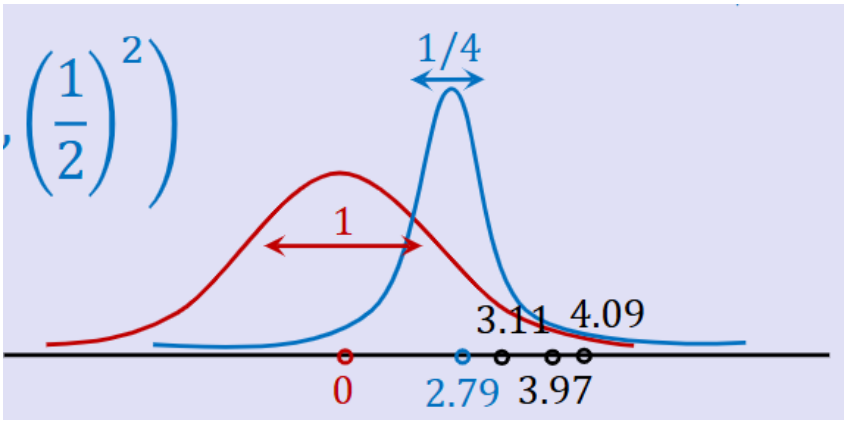
$$
\mu'=\frac{\mu_0+k}{n+1}=\frac{0+x_1+x_2+x_3}{3+1}\qquad \sigma'^2=\frac{1}{n+1}=\frac{1}{3+1}
$$

归一化常数：

$$Z(x)=\int_{-\infty}^{\infty}e^{-\frac{1}{2}(\frac{\theta-\frac{x_1+x_2+x_3}{4}}{\frac{1}{2}})^2}d\theta=\sqrt{2\pi\sigma^2}=\sqrt{\frac{\pi}{2}}$$

后验：$\Theta|X_1,X_2,X_3\sim N(\frac{x_1+x_2+x_3}{4},\frac{1}{4})$

$$f_{\Theta|X_1,X_2,X_3}(\theta|3.97,4.09,3.11)=\frac{1}{\sqrt{\frac{\pi}{2}}}e^{-\frac{1}{2}(\frac{\theta-2.7925}{\frac{1}{2}})^2}$$



A more general case

Suppose $X_1,\ldots,X_n$ form a random sample from Normal distributions with a common unknown mean μ and the known variance σ_i^2 $(\sigma_i^2>0)$. If the prior distribution $f_{\mathrm{M}}(\mu)$ is the Normal distribution $N(\mu_0,\sigma_0^2)$, then the posterior distribution $f_{\mathrm{M}|X_1,\ldots,X_n}(\mu|x_1,\ldots,x_n)$ given observed data $\{x_i\}_{i=1}^n$, $\sum_{i=1}^{n}x_i=k$ is also the Normal distribution $N(\mu',\sigma'^2)$, where

$$\frac{\mu'}{\sigma'^2}=\frac{\mu_0}{\sigma_0^2}+\frac{x_1}{\sigma_1^2}+\cdots+\frac{x_n}{\sigma_n^2}\qquad\frac{1}{\sigma'^2}=\frac{1}{\sigma_0^2}+\frac{1}{\sigma_1^2}+\cdots+\frac{1}{\sigma_n^2}$$

证明：

似然：$X_i|\mathrm{M}\sim N(\mu,\sigma_i^2)$

$$f_{X_i|\mathrm{M}}(x_i|\mu)=\frac{1}{\sqrt{2\pi\sigma_i^2}}e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma_i^2}}$$

先验：$\mathrm{M}\sim N(\mu_0,\sigma_0^2)$

$$f_M(\mu)=\frac{1}{\sqrt{2\pi\sigma_0^2}}e^{-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}}$$

贝叶斯推断：

$$\begin{aligned}f_{M|X_1,\ldots,X_n}(\mu|x_1,\ldots,x_n)&\propto f_{X_1|M}(x_1|\mu)\cdots f_{X_n|M}(x_n|\mu)f_M(\mu)\\&\propto e^{-\frac{1}{2}[\frac{(x_1-\mu)^2}{\sigma_1^2}+\cdots+\frac{(x_n-\mu)^2}{\sigma_n^2}+\frac{(\mu-\mu_0)^2}{\sigma_0^2}]}\\&\propto e^{-\frac{1}{2}\frac{(\mu-\mu')^2}{\sigma'^2}}\end{aligned}$$

其中

$$\frac{\mu'}{\sigma'^2}=\frac{\mu_0}{\sigma_0^2}+\frac{x_1}{\sigma_1^2}+\cdots+\frac{x_n}{\sigma_n^2}\qquad\frac{1}{\sigma'^2}=\frac{1}{\sigma_0^2}+\frac{1}{\sigma_1^2}+\cdots+\frac{1}{\sigma_n^2}$$

> 因为要把含 μ 的项凑平方，所以二次项和一次项系数要匹配. 常数项不匹配没关系，因为是正比号，多的或少的可以留给归一化常数一起算.

归一化常数：

$$Z(x_1,\ldots,x_n)=\int_{-\infty}^{\infty}e^{-\frac{1}{2}\frac{(\mu-\mu')^2}{\sigma'^2}}d\mu=\sqrt{2\pi\sigma'^2}$$

后验：

$$f_{M|X_1,\ldots,X_n}(\mu|x_1,\ldots,x_n)=\frac{1}{\sqrt{2\pi\sigma'^2}}e^{-\frac{1}{2}\frac{(\mu-\mu')^2}{\sigma'^2}}$$

其中

$$\frac{\mu'}{\sigma'^2}=\frac{\mu_0}{\sigma_0^2}+\frac{x_1}{\sigma_1^2}+\cdots+\frac{x_n}{\sigma_n^2}\qquad\frac{1}{\sigma'^2}=\frac{1}{\sigma_0^2}+\frac{1}{\sigma_1^2}+\cdots+\frac{1}{\sigma_n^2}$$

结论：Normal 是 Normal 的共轭先验.

Lec 3 预测/估计/假设检验

> Prediction, Estimation, Hypothesis Testing

硬币实验，给定参数先验 $f_\Theta(\theta)$, $\theta=P(H)$ 和系列观测数据 x，得到参数后验 $f_{\Theta|X}(\theta|x)$

Prediction: $P(\text{next H})=?$

> 预测主要关注如何利用现有数据和模型，来预见或推断未来可能出现的观测结果. 例如，当我们想知道「下次投掷硬币是否会出现正面」时，实际上是在根据以往数据和所建立的概率模型，对未来结果进行推断. 预测的重点是对未来随机变量的取值做出合理预判.

Estimation: $P(H)=?$

> 估计关注的是如何利用样本数据来推断潜在数据生成过程中的未知参数. 例如，在硬币实验中，我们可能需要估计「正面出现概率 $P(H)$」. 这一过程通常使用点估计（如极大似然估计）或区间估计的方法，给出一个具体数值（或区间），反映出参数的可能取值. 估计的重点在于获得对参数的数值描述，而不是直接预测未来事件.

Hypothesis testing: $P(H)=\frac{1}{3}$ or $\frac{2}{3}$?

> 假设检验则是针对某个特定的理论假设（例如 $P(H)=\frac{1}{3}$ 或 $P(H)=\frac{2}{3}$）进行验证. 通过构建原假设和备择假设，并利用样本数据计算统计量，再结合预设的显著性水平，我们可以判断数据是否支持这一假设. 如果数据与假设显著不符，则有理由拒绝原假设. 假设检验的重点在于验证我们提出的关于总体或参数的具体假设是否成立.

总结来说：

- 预测面向未来，关注如何借助已知信息预判未来数据.
- 估计则是利用现有数据去求得模型中参数的具体数值.
- 假设检验则通过对比数据与理论假设，来判断某一命题是否成立.

3.1 贝叶斯预测

> Bayesian Prediction

3.1.1 连续变量

以 `2.2.1 约会大作战!` 模型为例：

On her first date, Apple arrives $\frac{1}{2}$ hour late.

How likely to arrive more than $\frac{3}{4}$ hour late next time?

似然：

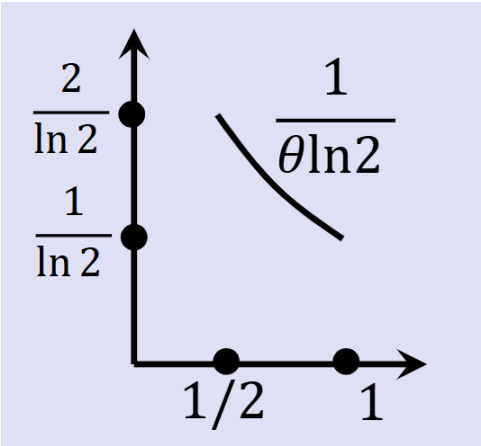
$$X_1|\Theta \sim Uniform(0,\theta)$$

先验：

$$\Theta \sim Uniform(0,1)$$

后验：

$$f_{\Theta|X_1}(\theta|\frac{1}{2})=\begin{cases}\frac{1}{\theta \ln 2} & \text{if } \frac{1}{2}\le\theta\le 1\\ 0 & \text{otherwise}\end{cases}$$



> 见 `2.2.1 约会大作战!` .

预测：

已知全概率公式及其变体

X, Y 离散:

$$\sum_y P_{X,Y}(x,y) = P_X(x)$$

$$\begin{aligned}\sum_y P_{X,Y|Z}(x,y|z) &= \sum_y \frac{P_{X,Y,Z}(x,y,z)}{P_Z(z)} \\ &= \sum_y \frac{P_{X,Z}(x,z)}{P_Z(z)} \\ &= P_{X|Z}(x|z)\end{aligned}$$

X 离散，Y 连续:

$$\int_y P(X=x, y \le Y \le y+dy) = P_X(x)$$

$$\begin{aligned}\int_y P(X=x, y \le Y \le y+dy|Z=z) &= \frac{\int_y P(X=x, Z=z, y \le Y \le y+dy)}{P_Z(z)} \\ &= \frac{\int_y P(X=x, Z=z|y \le Y \le y+dy) f_Y(y) dy}{P_Z(z)} \\ &= \frac{P_{X,Z}(x,z)}{P_Z(z)} \\ &= P_{X|Z}(x|z)\end{aligned}$$

这里 $X_2 \ge \frac{3}{4}$ 类比 $X=x$，$X_1 = \frac{1}{2}$ 类比 $Z=z$，$\theta \le \Theta \le \theta + d\theta$ 类比 $y \le Y \le y+dy$

$$
\begin{aligned}
P(X_2 \geq \frac{3}{4}|X_1 = \frac{1}{2}) &= \int_{-\infty}^{+\infty} P(X_2 \geq \frac{3}{4}, \theta \leq \Theta \leq \theta + d\theta | X_1 = \frac{1}{2}) \\
&= \int_{-\infty}^{+\infty} \frac{P(X_2 \geq \frac{3}{4}, \theta \leq \Theta \leq \theta + d\theta, X_1 = \frac{1}{2})}{P(X_1 = \frac{1}{2})} \\
&= \int_{-\infty}^{+\infty} \frac{P(X_2 \geq \frac{3}{4} | \theta \leq \Theta \leq \theta + d\theta, X_1 = \frac{1}{2}) P(\theta \leq \Theta \leq \theta + d\theta, X_1 = \frac{1}{2})}{P(X_1 = \frac{1}{2})} \\
&= \int_{-\infty}^{+\infty} P(X_2 \geq \frac{3}{4} | \theta \leq \Theta \leq \theta + d\theta, X_1 = \frac{1}{2}) P(\theta \leq \Theta \leq \theta + d\theta | X_1 = \frac{1}{2}) \\
&= \int_{-\infty}^{+\infty} P(X_2 \geq \frac{3}{4} | \theta \leq \Theta \leq \theta + d\theta, X_1 = \frac{1}{2}) f_{\Theta|X_1}(\theta|\frac{1}{2}) d\theta \\
&= \int_{-\infty}^{+\infty} P(X_2 \geq \frac{3}{4} | \Theta = \theta, X_1 = \frac{1}{2}) f_{\Theta|X_1}(\theta|\frac{1}{2}) d\theta \\
&= \int_{-\infty}^{+\infty} P(X_2 \geq \frac{3}{4} | \Theta = \theta, \cancel{X_1 = \frac{1}{2}}) f_{\Theta|X_1}(\theta|\frac{1}{2}) d\theta \\
&= \int_{\frac{1}{2}}^{1} P(X_2 \geq \frac{3}{4} | \Theta = \theta) f_{\Theta|X_1}(\theta|\frac{1}{2}) d\theta \\
&= \int_{\frac{1}{2}}^{1} P(X_2 \geq \frac{3}{4} | \Theta = \theta) \frac{1}{\theta \ln 2} d\theta \\
&= \int_{\frac{3}{4}}^{1} \int_{\frac{3}{4}}^{\theta} f_{X_2|\Theta}(x_2|\theta) dx_2 \frac{1}{\theta \ln 2} d\theta \\
&= \int_{\frac{3}{4}}^{1} \frac{1}{\theta}(\theta - \frac{3}{4}) \frac{1}{\theta \ln 2} d\theta \\
&= \frac{\ln \frac{4}{3} - \frac{1}{4}}{\ln 2} \\
&\approx 0.0544
\end{aligned}
$$

> $X_1 = \frac{1}{2}$ 可以直接消掉是因为 $X_2 \geq \frac{3}{4}$ 和 $X_1 = \frac{1}{2}$ 关于 $\Theta = \theta$ 条件独立. 注意是条件独立而不是独立.

由上述例子可总结出「贝叶斯预测」方法论：

Observation/Past data: $X = x$

If X is continuous, to predict future data $x^* \in [a, b]$

$$
\begin{aligned}
P(x^* \in [a,b] | X = x) &= \int_{-\infty}^{+\infty} P(x^* \in [a,b] | \theta) f_{\Theta|X}(\theta|x) d\theta \\
&= \int_{-\infty}^{+\infty} \int_{a}^{b} f_{X|\Theta}(x^*|\theta) dx^* f_{\Theta|X}(\theta|x) d\theta
\end{aligned}
$$

> 以更新后的参数分布 Θ 作为桥梁，巧妙地连接了观测和预测.
>
> 全流程：假设似然 & 先验 - 观测 - 贝叶斯推断 - 后验 - 预测（即在贝叶斯推断的最后加一步预测）.
>
> 这个公式有很强的可解释性，即观测数据作为条件，实际上为新数据的预测提供了后验参数分布，因此很好记忆. 完整证明见上述例子.
>
> 注意，这里外部积分的上下界为正负无穷，但实际只要对 θ 使 $P(x^* \in [a,b]|\theta)$ 和 $f_{\Theta|X}(\theta|x)$ 同时不等零的部分（交集）积分即可，多数情况下有效积分区间会缩小，详情见上述例子.
>
> 积分的意义是将 $x^* \in [a,b] \big| \theta|x$ 的概率以后验概率 $f_{\Theta|X}(\theta|x)d\theta$ 为权重加权平均，也可以理解为我们在计算 $E\left[P\left(x^* \in [a,b] \big| \theta|x\right)\right]$. 注意这里的 $\theta|x$ 是后验，但由于条件里一般不再写条件，因此简单记为 $E\left[P(x^* \in [a,b]|\theta)\right]$. 实际上这两种记法完全等价，因为 $P\left(x^* \in [a,b] \big| \theta|x\right) = P\left(x^* \in [a,b] \big| \theta, x\right) = P\left(x^* \in [a,b] \big| \theta\right)$. 这又是因为 $x^* \in [a,b]$ 和

> $X=x$ 关于 $\Theta=\theta$ 条件独立. 注意是条件独立，并不是独立. 这一步在上面那个很长的推导式里也体现了，即 $X_1=\frac{1}{2}$ 可以直接消掉.

3.1.2 离散变量

If X is discrete, to predict future data x^*

$$P(x^*|X=x)=\int_{-\infty}^{+\infty}P(x^*|\theta)f_{\Theta|X}(\theta|x)d\theta$$

> 注意 X 是离散分布时，内部不需要对 x^* 积分，只对 θ 积一次.

以 `2.2.2 硬币是否公平？ ② 连续硬币模型` 为例：观察到 n 个 H，问 $P(\text{H next flip})=?$

抛硬币服从伯努利分布：

$$X|\Theta\sim Bernoulli(\theta)$$

θ is a value of the random variable Θ. 假设先验分布：

$$\Theta\sim Uniform(0,1)=Beta(1,1)$$

后验：$\Theta|nH\sim Beta(n+1,1)$

$$f_{\Theta|X_1,\dots,X_n}(\theta|nH)=(n+1)\theta^n\quad\text{if } 0\le\theta\le 1$$

> 常见的共轭结论. 证明见 `2.3.2 常见共轭分布 ① Beta-Bernoulli`.

预测：

$$\begin{aligned}P(H^*|nH)&=\int_0^1 P(H^*|\theta)f_{\Theta|X_1,\dots,X_n}(\theta|nH)d\theta\\&=\int_0^1\theta(n+1)\theta^n d\theta\\&=\frac{n+1}{n+2}\end{aligned}$$

同理，积分的意义是将 $H^*\big|\theta|nH$ 的概率以后验概率 $f_{\Theta|X_1,\dots,X_n}(\theta|nH)d\theta$ 为权重加权平均，即计算 $E[P(H^*\big|\theta|nH)]$. 而 $H^*\big|\theta|nH$ 的概率正好是 θ（只有服从伯努利分布的数据才恰有此效果），此时贝叶斯预测公式恰好和 $E[\Theta|nH]$ 一致，即计算后验参数变量的均值. 知道这个结论后可以验证某些预测是否算对，尤其是数据服从伯努利分布的模型（例如硬币），当最新的参数分布 $\Theta\sim Beta(\alpha,\beta)$ 时（注意这里不是先验，是已经更新过的，只是记作 Θ，因为不知道更新了几次），下一个抛出正面的概率为 $E[\Theta]=\frac{\alpha}{\alpha+\beta}$.

（2025.2.12 思考）在这种硬币模型下，注意区分 Θ 预测值和 MAP 估计. 二者都给出了 Θ 的一个判断值，但贝叶斯预测整合参数不确定性和数据信息的能力比 MAP 强. 下面从数学角度分析二者区别：

首先，MAP 使用范围就比贝叶斯预测要小，MAP 只能估 θ，贝叶斯预测什么都可以测，包括 θ（测 θ 的意思是，预测一个在给定 θ 后概率恰为 θ 的事件，在观测数据条件下的概率，例如硬币模型）. 然后把贝叶斯预测中的测 θ 单独拿出来和 MAP 比，它整合参数不确定性和观测数据信息的能力比 MAP 强，公式是 $\theta f(\theta|data)d\theta$ 对 θ 积分，也就是算 θ 的均值，而 MAP 是直接让 $f(\theta|data)$ 取最大. 这里考虑一个极端的情况，先验是 $0-1$ 均匀分布，那么后验就正比于似然，MAP 令似然最大，完全没有考虑到先验，而贝叶斯预测虽然也是代入似然，但是对整个区间积分（相当于乘一个常数 1 然后积分，在不同的地方都捕捉到先验很均匀的信息）. 因此，假设先验均匀，即 $\Theta\sim Beta(1,1)$，若干观测后后验为 $Beta(\alpha,\beta)$，则预测

下一个翻出正面的概率是 $\frac{\alpha}{\alpha+\beta}$（综合了先验信息，即初始的一正一反），而最大后验估计是 $\frac{\alpha-1}{\alpha-1+\beta-1}$，只吸收了观测数据的信息（即观测数据中正面的占比，作为 θ_{MAP}）.

(2025.2.23 补充) 上面提到的极端情况，暗示了硬币模型选择均匀先验时 MAP 和 MLE 结果一致，见 `5.2 最大似然估计` .

3.2 点估计

> Point Estimation

点估计是统计推断中**依据样本估计总体**未知参数的一种方法. 主要特点：计算得到**单一的数值**，作为总体参数（如总体均值、总体方差等）的近似值.

> 与区间估计不同，点估计只提供一个「最佳猜测」.

在选择和评价点估计量时，常关注以下几个性质：

- 无偏性：如果估计量的数学期望等于真实参数值，则称该估计量为无偏的.
- 一致性：随样本量增加，估计量应收敛于真实参数，即估计精度不断提高.
- 有效性：在所有无偏估计量中，方差最小的估计量称为有效的，即最精确.
- 渐进正态性：大样本下，某些点估计量经适当标准化后，分布会趋于正态.

3.2.1 最大后验估计

> Maximum A Posteriori Estimation, MAP

How to turn conditional PDF/PMF $f_{\Theta|X}(\theta|x)$ estimate into one number?

使用 MAP 估计器

> Maximum A Posteriori (MAP) Estimator

$$\begin{aligned}\theta_{MAP} &= \arg\max_{\theta} f_{\Theta|X}(\theta|x) \\ &= \arg\max_{\theta} \frac{f_{\Theta}(\theta) f_{X|\Theta}(x|\theta)}{Z(x)} \\ &= \arg\max_{\theta} f_{\Theta}(\theta) f_{X|\Theta}(x|\theta)\end{aligned}$$

以 `2.2.1 约会大作战!` 模型为例：

On her first date, Apple arrives $\frac{1}{2}$ hour late.

似然：

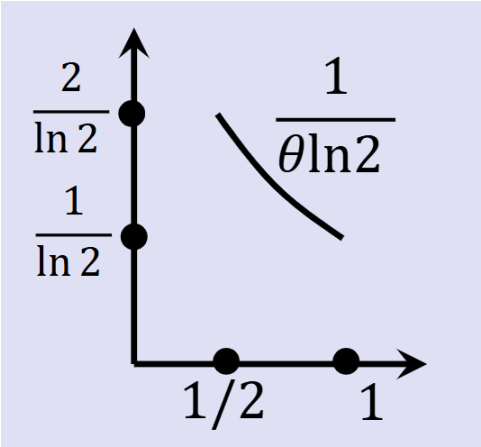
$$X|\Theta \sim Uniform(0, \theta)$$

先验：

$$\Theta \sim Uniform(0,1)$$

后验：

$$f_{\Theta|X}(\theta|\frac{1}{2}) = \begin{cases} \frac{1}{\theta \ln 2} & \text{if } \frac{1}{2} \le \theta \le 1 \\ 0 & \text{otherwise} \end{cases}$$



> 见 2.2.1 约会大作战！.

最大后验估计：

$$\theta_{MAP} = \arg\max_{\theta} \frac{1}{\theta \ln 2} = \arg\max_{\theta} \frac{1}{\theta} = \frac{1}{2}$$

> 注意求的是 θ 的估计值，不是 $f_{\Theta|X}(\theta|\frac{1}{2})$ 的.

MAP for Beta：以 2.2.2 硬币是否公平？ ② 连续硬币模型 为例

抛硬币服从伯努利分布：

$$X|\Theta \sim Bernoulli(\theta)$$

θ is a value of the random variable Θ. 假设先验分布：

$$\Theta \sim Uniform(0,1) = Beta(1,1)$$

观测：假设观测到 h 个 H，t 个 T.

后验：$\Theta|\text{h H's, t T's} \sim Beta(1+h, 1+t)$

> 常见的共轭结论. 证明见 2.3.2 常见共轭分布 ① Beta-Bernoulli.

最大后验估计：$mode[\Theta \sim Beta(\alpha,\beta)] = \frac{\alpha-1}{\alpha-1+\beta-1}$ when $\alpha, \beta > 1$

$$\theta_{MAP} = \frac{\alpha-1}{\alpha-1+\beta-1} = \frac{h}{h+t}$$

MAP for Normals

似然：$X_1|\Theta, \ldots, X_n|\Theta \sim N(\theta, 1)$

先验：$\Theta \sim N(\mu_0, 1)$

后验：$\Theta|X_1, \ldots, X_n \sim N(\frac{\mu_0+x_1+\cdots+x_n}{n+1}, \frac{1}{n+1})$

最大后验估计：$\theta_{MAP} = \frac{\mu_0+x_1+\cdots+x_n}{n+1}$

3.2.2 最大似然估计

> Maximum Likelihood Estimation, MLE
>
> 见 `5.2 最大似然估计` .

3.3 假设检验

> Hypothesis Testing

在贝叶斯统计框架下，假设检验主要是比较不同假设在观测数据下的后验概率，最终选择后验概率最大的假设. In a hypothesis testing problem, Θ takes m values, $\theta_1, \ldots, \theta_m$. Goal: to select "the optimal" hypothesis θ^*.

> 注意，先验假设 $\theta_1, \ldots, \theta_m$ 的概率可以不同，但是一开始概率高不一定后验概率也高，因为观测数据会对先验知识加以纠正. 而我们选择假设的原则是让后验概率最大.

同样采用 MAP，但注意这里 Θ 是离散变量：Choose the one for which $P_{\Theta|X}(\theta_i|x)$ is largest.

$$\begin{aligned}\theta_{MAP} &= \arg\max_{\theta} P_{\Theta|X}(\theta_i|x) \\ &= \arg\max_{\theta} \frac{P_{\Theta}(\theta_i) f_{X|\Theta}(x|\theta_i)}{Z(x)} \\ &= \arg\max_{\theta} P_{\Theta}(\theta_i) f_{X|\Theta}(x|\theta_i)\end{aligned}$$

3.3.1 二元假设检验

> Binary Hypothesis Testing

二元假设检验是一种统计学方法，用于在两种互斥的假设之间做出决策. 通常情况下，这两种假设被称为：

H_0（零假设），H_1（备择假设）

在决策过程中，可能会犯两类错误：

- 第一类错误（Type I error）：错误地拒绝了真实的零假设 H_0.
- 第二类错误（Type II error）：错误地接受了零假设 H_0，但实际上 H_1 才为真.

① MAP 决策准则

本节关注基于 MAP 的二元假设检验，即以后验为决策依据，选择后验概率最大的.

> 假设先验，基于观测数据用贝叶斯计算后验，比较各个后验即可.

Example 1: 垃圾邮件判断

Θ takes two values (e.g. $\Theta = 1$ for spam, and $\Theta = 0$ for legit)

计算是否 $f_{\Theta|X}(1|x) > f_{\Theta|X}(0|x)$

或计算是否 $\frac{P_{\Theta|X}(1|x)}{P_{\Theta|X}(0|x)} = \frac{f_{X|\Theta}(x|1)P(\Theta=1)}{f_{X|\Theta}(x|0)P(\Theta=0)} > 1$

已知先验：$P(\Theta = 1) = 20\%, P(\Theta = 0) = 80\%$

Suppose there are two patterns (independent given a specific email), X_1 and X_2, to classify whether an email is spam or legit

> 注意这里 X_1, X_2 是伯努利分布，只能取 0 或 1.

Θ	$P(X_1 = 1\|\theta)$	$P(X_2 = 1\|\theta)$
$\Theta = 0$ legit	0.03	0.0001
$\Theta = 1$ spam	0.1	0.01

In a specific email x, observe that $X_1 = 1$ and $X_2 = 0$, spam or legit?

计算后验：

$$P(\Theta = 1|X_1 = 1, X_2 = 0) \propto P(X_1 = 1, X_2 = 0|\Theta = 1)P(\Theta = 1) = 0.0198$$
$$P(\Theta = 0|X_1 = 1, X_2 = 0) \propto P(X_1 = 1, X_2 = 0|\Theta = 0)P(\Theta = 0) \approx 0.0240$$

$\Theta_{MAP} = 0$, we consider it as legit.

Example 2: 硬币类型判断

Coin A is heads with probability $\frac{2}{3}$. Coin B is tails with probability $\frac{2}{3}$.

H, H, T are 3 flips of a random coin. Which coin was it?

先验：没有先验知识的情况下，可以假设两种先验概率相等.

$$P(\Theta = A) = P(\Theta = B) = 50\%$$

后验：

$$P(\Theta = A|H, H, T) \propto P(H, H, T|\Theta = A)P(\Theta = A) = \frac{2}{27}$$
$$P(\Theta = B|H, H, T) \propto P(H, H, T|\Theta = B)P(\Theta = B) = \frac{1}{27}$$

$\Theta_{MAP} = A$, we consider it as Coin A.

错误率：What is the probability you are wrong based on MAP, given the outcome is H, H, T?

$\Theta_{MAP} = A$, we consider it as Coin A.

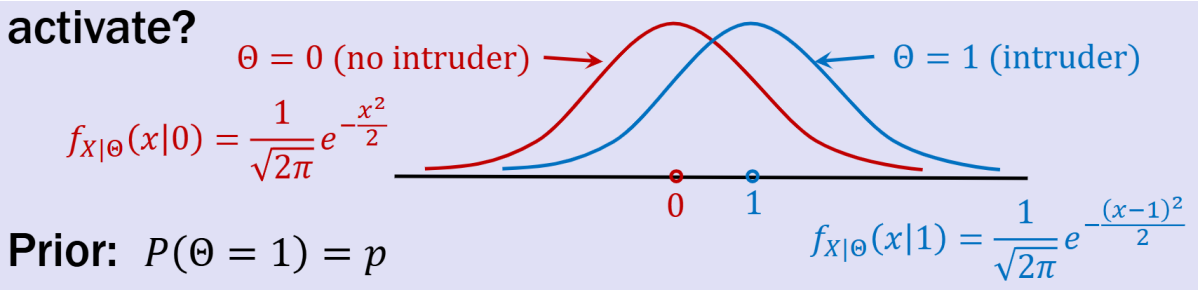
$$
\begin{aligned}
error &= P(\Theta = B|H, H, T) \\
&= \frac{P(H, H, T|\Theta = B)P(\Theta = B)}{P(H, H, T)} \\
&= \frac{\frac{1}{27}}{\frac{1}{27} + \frac{2}{27}} \\
&= \frac{1}{3}
\end{aligned}
$$

总体错误率：What is the probability you are wrong **on average** based on MAP given the outcome of 3 flips?

> 待完成.
>
> 注意这里先验一样，所以后验比例恰好和似然比例一样. 不要混淆概念.

Example 3: 劫车检测

A car-jack detector X outputs $N(0, 1)$ if there is no intruder and $N(1, 1)$ if there is. When should alarm activate?



后验：

$$
\begin{aligned}
P_{\Theta|X}(0|x^*) &\propto P_\Theta(0) f_{X|\Theta}(x^*|0) \propto (1-p) e^{-\frac{x^{*2}}{2}} \\
P_{\Theta|X}(1|x^*) &\propto P_\Theta(1) f_{X|\Theta}(x^*|1) \propto p e^{-\frac{(x^*-1)^2}{2}}
\end{aligned}
$$

MAP 决策：

$$
\frac{P_{\Theta|X}(1|x^*)}{P_{\Theta|X}(0|x^*)} = \frac{p}{1-p} e^{x^* - \frac{1}{2}} > 1 \Rightarrow x^* > \frac{1}{2} + \ln \frac{1-p}{p}
$$

当 $x^* > \frac{1}{2} + \ln \frac{1-p}{p}$ 时，警报应该激活.

What is the error?

$$
\begin{aligned}
error &= P(\hat{\theta} \neq \theta) \\
&= P(\theta = 0, x > \frac{1}{2} + \ln\frac{1-p}{p}) + P(\theta = 1, x \leq \frac{1}{2} + \ln\frac{1-p}{p}) \\
&= P(x > \frac{1}{2} + \ln\frac{1-p}{p} | \theta = 0)P(\theta = 0) + P(x \leq \frac{1}{2} + \ln\frac{1-p}{p} | \theta = 1)P(\theta = 1) \\
&= P\left(N(0,1) > \frac{1}{2} + \ln\frac{1-p}{p} | \theta = 0\right)P(\theta = 0) + P\left(N(1,1) \leq \frac{1}{2} + \ln\frac{1-p}{p} | \theta = 1\right)P(\theta = 1)
\end{aligned}
$$

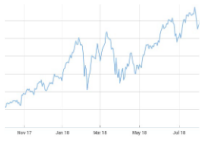
> 注意不同的 θ 条件下有不同的似然.

Lec 4 样本统计量

> Sample Statistics
>
> 本节回归经典统计学. 这里提到的统计量和 `5.1 估计量` 提到的估计量是等价概念.

$X \sim \text{Bernoulli}(p)$

$X \sim \text{Poisson}(\lambda)$



$X \sim \mathcal{N}(\mu, \sigma^2)$

Observation: a sample of n random variables $D = \{X_1, \ldots, X_n\}$

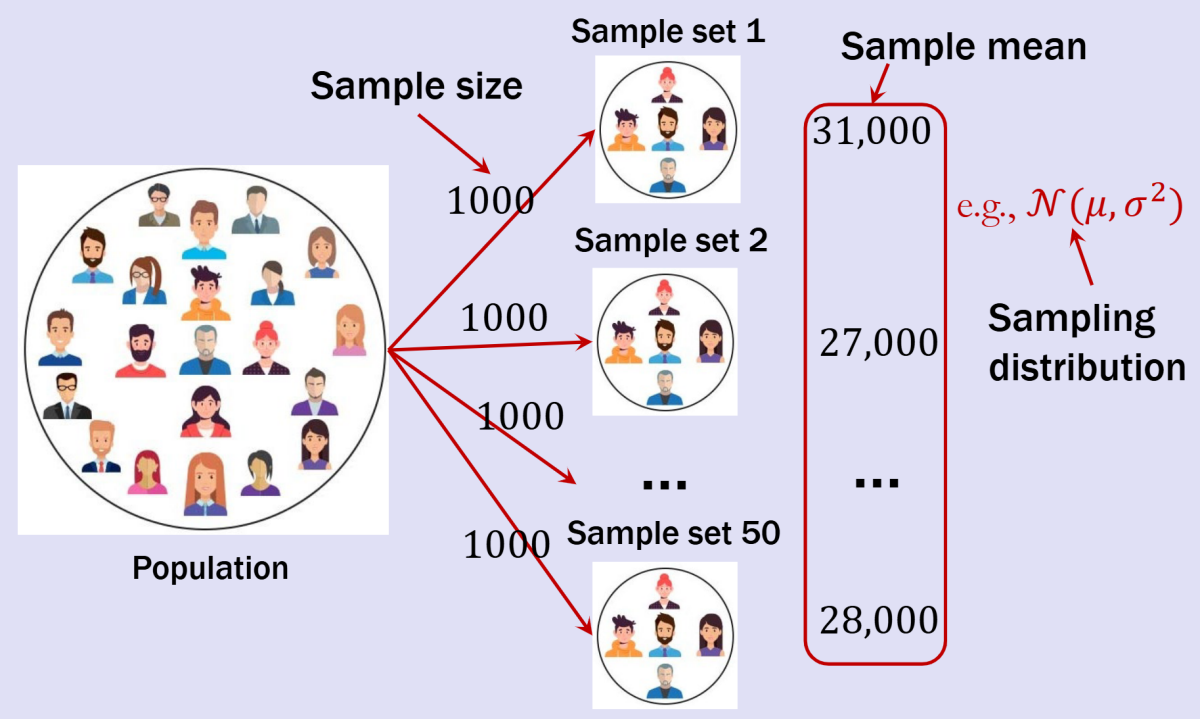
Goal: to estimate θ (p, λ, μ, σ^2, etc.)

Classical statistics:

The parameter θ is considered as a deterministic quantity that happens to unknown

Develop an **estimator** to estimate $\hat{\theta}$ based on the sample D

- With different samples $D_i = \{X_1 = x_1, \ldots, X_n = x_n\}$, the estimation $\hat{\theta}_i$ would be different
- How to evaluate of the **estimator**?



4.1 随机样本

> Random Sample

A **random sample** of size n is a joint outcome of n independent random variables $X_1, \ldots, X_n$, with same PDF/PMF

> n 个独立同分布.

$$E[X_1] = \cdots = E[X_n] = \mu$$
$$Var[X_1] = \cdots = Var[X_n] = \sigma^2$$

The process of generating a specific random sample is called **sampling**.

> 抽样.

4.2 样本统计量 & 抽样分布

> Sample Statistics & Sampling Distributions

Given a random sample of n independent random variable $X_1, \ldots, X_n$, with same PDF/PMF, the numerical descriptive measures of the sample are called **statistics**.

> 样本的数值属性.

The sample mean: $\overline{X} = \frac{X_1 + \cdots + X_n}{n}$

The sample proportion: $\hat{p} = \frac{X_1 + \cdots + X_n}{n}$ when X_i's are Bernoulli RVs.

> 样本比例，指一个样本中具备某一特定属性或特征的个体所占的比例.

The samle sum: $X = X_1 + \cdots + X_n$

The sample variance: $s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$

The probability distributions for statistics are called **sampling distributions**.

> 抽样分布.

4.2.1 样本均值

> Sample Mean
>
> 样本均值是总体均值的估计量. 关于估计量，见 `5.1 估计量` .

Derive the sampling distribution of sample mean using the laws of probability.

Consider a fair coin X. $X = 1$ means H and $X = 0$ means T. Flip the coin twice, X_1 and X_2. What is the PMF of $\overline{X}$?

X_1	X_2
0	0
0	1
1	0
1	1

Joint PMF of X_1, X_2

X_2 \ X_1	0	1
0	¼	¼
1	¼	¼

PMF of $X_1 + X_2$

x	0	1	2
$P(X_1 + X_2 = x)$	1/4	1/2	1/4

PMF of $\bar{X}$

x	0	1/2	1
$P(\bar{X} = x)$	1/4	1/2	1/4

> 注意，这里对四种样本统一分析. 假设左表四种样本概率一致，则会出现右表中的分布. 如果是确定的样本，则 $X_1 + X_2$ 和 $\overline{X}$ 也会变为确定的结果，而非分布.

Flip the coin n times:

$$X_1 + \cdots + X_n = n\overline{X} \sim Binomial(n, \frac{1}{2})$$

What if distribution of X_i is unknown?

这里的 unknown 指没有显式表达式，只有一个很大的数据表，X_i 服从从表中随机抽取的分布. 例如：

What is the average exam grade?

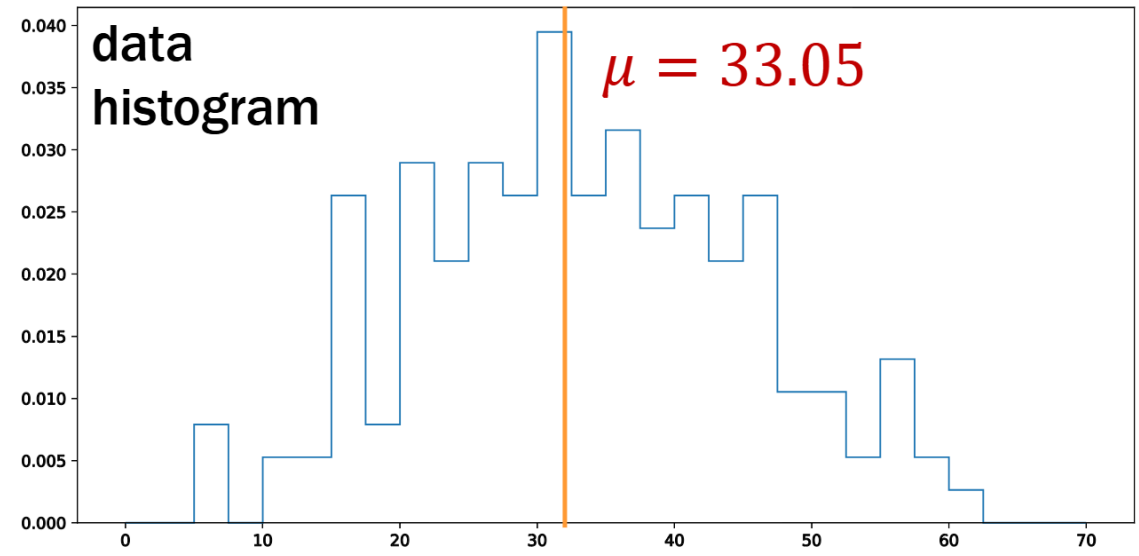
59	49	46	46	32	34	33	39	34	33	22	29	24	27
62	53	45	40	46	35	33	30	39	30	15	30	13	6
57	50	45	44	30	37	40	25	23	31	31	20	15	12
59	52	39	41	35	43	44	25	37	26	25	29	20	16
55	46	42	40	45	38	25	24	32	25	21	18	15	5
55	46	49	32	41	42	32	32	24	34	21	27	20	7
53	44	45	44	33	27	28	23	28	20	29	19	26	11
56	51	44	49	33	36	43	23	33	23	29	36	26	13
44	48	39	35	37	39	33	31	22	28	21	16	16	15
56	52	41	36	42	38	37	28	31	21	15	15	15	
46	38	37	31	39	37	40	19	24	28	20	32	29	

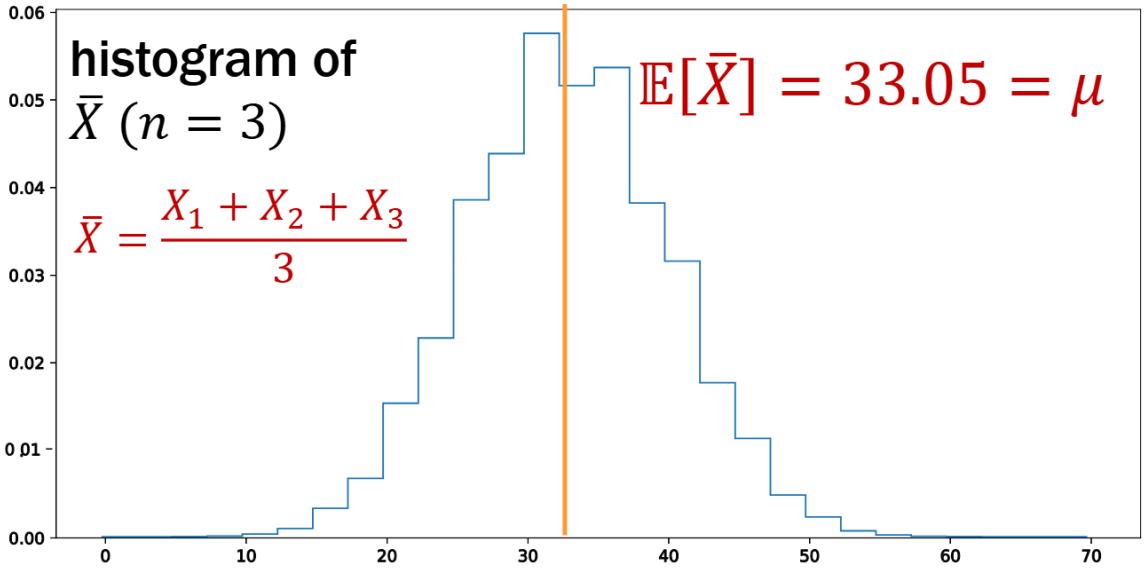
$$\bar{x} = \frac{32 + 15 + 29}{3} \approx 25.33$$

$$\bar{x} = \frac{6 + 5 + 7}{3} = 6$$ Atypical!

> 选择不同的样本，样本统计量（均值）可能不同.

虽然样本均值不固定，但仍然适合作为总体均值的估计，因为其无偏性、一致性：





① 无偏性

> 无偏性，unbiasedness，无系统性偏差，指一个估计量在重复抽样下，其期望值等于真实参数. 具体到样本均值，意味着：
>
> - 样本均值的期望值等于总体均值.
> - 没有系统性偏差：使用样本均值估计总体均值是合适的. 在大样本下，它既不会倾向于高估也不会倾向于低估总体均值，从而提供了一个可靠的统计依据.
>
> 系统性偏差：采用某个方法时，总是会将真实值偏向某一方向，且无法通过增大样本量来消除，这种偏差通常来源于测量工具的设计缺陷、抽样方法的不当或其他固定的错误因素.
>
> 和物理实验中的系统误差是同一个概念，与其相对的是偶然因素引起的随机误差. 物理实验的误差主要描述仪器精度和测量方法，统计学上的偏差主要描述统计方法.

The sample mean is an **unbiased** estimator of the population mean μ:

$$\text{For every } n, \mathbb{E}[\overline{X}] = \mu$$

Proof:

$$
\begin{aligned}
\mathbb{E}[\overline{X}] &= \mathbb{E}[\frac{X_1 + \cdots + X_n}{n}] \\
&= \frac{1}{n}\mathbb{E}[X_1 + \cdots + X_n] \\
&= \frac{1}{n}(\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]) \\
&= \frac{1}{n}(\mu + \cdots + \mu) \\
&= \mu
\end{aligned}
$$

② 一致性

> consistency

样本均值的一致性（consistency）意味着在**样本数量足够大**的情况下，通过样本均值得到的总体均值估计将非常接近于真实的总体均值，从而保证了估计方法在大样本下的可靠性.

一个估计量既可以是无偏的，又可以是一致的，两者关注的方面不同：无偏性关注「平均水平上正确」，一致性则关注「样本量增加时收敛到正确值」.

直观解释：根据大数定律，随机样本随着 size n 增加，样本均值 $\overline{X}_n$ 会越来越接近于 μ.

> 注意这里随机样本包含了由 n 个独立同分布、期望为 μ 的随机变量组成的默认前提.

Consistent: For every positive ϵ, δ and sufficiently large sample size n

$$
P(|\overline{X} - \mu| \geq \epsilon) \leq \delta
$$

> Weak law of large numbers.

Or, Suppose $X_1, X_2, \ldots$ is a sequence of IID random variables with $expected\ value\ \mathbb{E}[X_i] = \mu$ and $finite\ variance\ Var[X_i] < \infty$. Define $\overline{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)$. Then, for every $\epsilon > 0$:

$$
\lim_{n \to \infty} \mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) = 0
$$

> 见 `ESTR 2018 概率论 13.9 弱大数定律`.

4.2.2 中心极限定理

> The Central Limit Theorem

Consider IID random variables $X_1, X_2, \ldots$ with expected value $\mathbb{E}[X_i] = \mu$ and variance $Var[X_i] = \sigma^2$. Then, if we define

$$
Z_n = \frac{\frac{1}{n}(X_1 + \cdots + X_n) - \mu}{\frac{\sigma}{\sqrt{n}}}
$$

the CDF of random variable Z_n will converge to the CDF of a standard normal distribution $\phi(z)$, i.e. for every $z \in \mathbb{R}$

$$
\lim_{n \to \infty} \mathbb{P}(Z_n \leq z) = \phi(z).
$$

Therefore, according to the central limit theorem, the sum of
many independent random numbers will approximately have a *normal distribution*.

> 见 `ESTR 2018 概率论 13.11 中心极限定理`.

已知，标准化及其逆变换：

$$x \sim \mathcal{N}(\mu, \sigma^2) \quad \underset{x = \sigma z + \mu}{\overset{z = \frac{x-\mu}{\sigma}}{\rightleftarrows}} \quad z \sim \mathcal{N}(0,1)$$

PDF

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

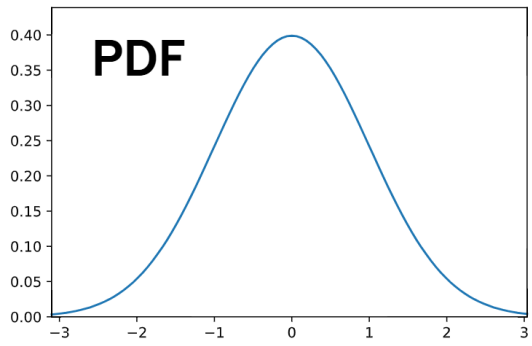
PDF

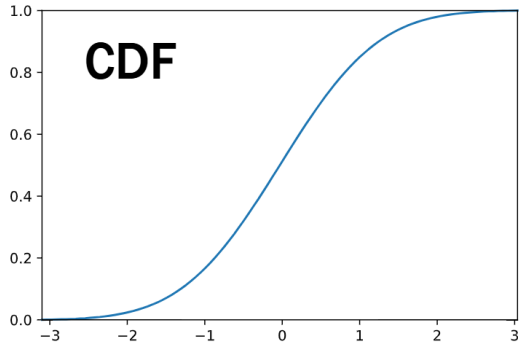
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

> 见 `ESTR 2018 概率论 13.10 标准化`.

CDF: $\phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{x^2}{2}} dx$

> Cumulative Distribution Function





定义 $X = \sum_{i=1}^{n} X_i$，则

$$\mathbb{E}[X] = n\mu, \quad Var[X] = n\sigma^2$$

$$Z_n = \frac{\frac{1}{n}(X_1 + \cdots + X_n) - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{X - n\mu}{\sqrt{n}\sigma}$$

中心极限定理

$$\lim_{n\to\infty} \mathbb{P}(Z_n \le z) = \phi(z)$$

说明当 $n \to \infty$，$Z_n \sim N(0,1) \Leftrightarrow X \sim N(n\mu, n\sigma^2) \Leftrightarrow \overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

> Note: standard deviation of a sample statistics is also called standard error.

① 应用

中心极限定理可以用来估计 样本均值/样本和 落在某个范围内的概率.

Example 1: In a population of 1000, 200 have disease: $X_i \sim Bernoulli(0.2)$, where $X_i = 1$ means having disease. For a sample of size 16, what's the probability that the sample mean is in the range 10% to 30%?

> 这里抽样在实践中应该是超几何分布，为简化问题假设可以重复抽到同一个人，即 IID 伯努利组成的随机样本.
>
> 样本均值 $\overline{X}=\frac{X_1+\cdots+X_{16}}{16}$ 为样本中的患病人数占比，则样本和 $16\overline{X}\sim Binomial(16,0.2)$. 注意二项分布是离散分布，只能取到整数.

真实概率：

$$\begin{aligned}P(0.1\le\overline{X}\le0.3)&=P(1.6\le16\overline{X}\le4.8)\\&=P(2\le16\overline{X}\le4)\\&=\sum_{i=2}^{4}\binom{16}{i}0.2^i\times(1-0.2)^{16-i}\\&\approx0.6575\end{aligned}$$

CLT 估计：$X_i\sim Bernoulli(0.2)$

法一：对均值标准化

$$\begin{aligned}\mu(\overline{X})&=\mu_{X_i}=p=0.2\\\sigma(\overline{X})&=\frac{\sigma_{X_i}}{\sqrt{n}}=\frac{\sqrt{0.2(1-0.2)}}{\sqrt{16}}=0.1\end{aligned}$$

$$\begin{aligned}P(0.1\le\overline{X}\le0.3)&=P(\frac{0.1-0.2}{0.1}\le\frac{\overline{X}-\mu(\overline{X})}{\sigma(\overline{X})}\le\frac{0.3-0.2}{0.1})\\&=P(-1\le\frac{\overline{X}-\mu(\overline{X})}{\sigma(\overline{X})}\le1)\\&\approx\phi(1)-\phi(-1)\qquad\text{Central Limit Theorem}\\&\approx0.6827\end{aligned}$$

法二：对样本和标准化

$$\begin{aligned}\mu(16\overline{X})&=16\mu_{X_i}=16p=3.2\\\sigma(16\overline{X})&=\sqrt{n}\sigma_{X_i}=\sqrt{16\times0.2(1-0.2)}=1.6\end{aligned}$$

$$\begin{aligned}P(0.1\le\overline{X}\le0.3)&=P(1.6\le16\overline{X}\le4.8)\\&=P(\frac{1.6-3.2}{1.6}\le\frac{16\overline{X}-\mu(16\overline{X})}{\sigma(16\overline{X})}\le\frac{4.8-3.2}{1.6})\\&=P(-1\le\frac{16\overline{X}-\mu(16\overline{X})}{\sigma(16\overline{X})}\le1)\\&\approx\phi(1)-\phi(-1)\qquad\text{Central Limit Theorem}\\&\approx0.6827\end{aligned}$$

Difference within 2.6%.

> 之所以有误差，是因为我们的 n 并非无穷大.

注意本课程用 CLT 估计离散变量（如二项分布）时**不考虑连续性校正**，直接标准化后代入标准正态的 CDF（查表或积分）即可.

> 关于连续性校正，见 `ESTR 2018 概率论 12.4 正态近似 ② 连续性校正` .

Example 2: In a population of 1000, 100 have disease: $Y_i \sim Bernoulli(0.1)$, where $Y_i = 1$ means having disease. For a sample of size 16, what's the probability that the sample mean is in the range 5% to 15%?

真实概率：

$$\begin{aligned} P(0.05 \le \overline{Y} \le 0.15) &= P(0.8 \le 16\overline{Y} \le 2.4) \\ &= P(1 \le 16\overline{Y} \le 2) \\ &= \sum_{i=1}^{2} \binom{16}{i} 0.1^i \times (1-0.1)^{16-i} \\ &\approx 0.6039 \end{aligned}$$

CLT 估计：$Y_i \sim Bernoulli(0.1)$

法一：对均值标准化

$$\mu(\overline{Y}) = \mu_{Y_i} = p = 0.1$$

$$\sigma(\overline{Y}) = \frac{\sigma_{Y_i}}{\sqrt{n}} = \frac{\sqrt{0.1(1-0.1)}}{\sqrt{16}} = 0.075$$

$$\begin{aligned} P(0.05 \le \overline{Y} \le 0.15) &= P(\frac{0.05-0.1}{0.075} \le \frac{\overline{Y}-\mu(\overline{Y})}{\sigma(\overline{Y})} \le \frac{0.15-0.1}{0.075}) \\ &= P(-\frac{2}{3} \le \frac{\overline{Y}-\mu(\overline{Y})}{\sigma(\overline{Y})} \le \frac{2}{3}) \\ &\approx \phi(\frac{2}{3}) - \phi(-\frac{2}{3}) \qquad \text{Central Limit Theorem} \\ &\approx 0.4950 \end{aligned}$$

> 一般用法一，法二略.
>
> 注意这里其实估计得不准，因为 n 不够大且不考虑连续性校正.

② 经验法则

> Central Rule of Thumb

If the data population is normal, then the sampling distribution of $\overline{X}$ is also normal, no matter what sample size you choose;

> 注意这个结论不显然，但本课不考察该知识点. 感兴趣可搜索关键词：`正态分布的封闭性`、`特征函数（矩母函数）`、`傅里叶变换`.

Otherwise, if $n \ge 30$, CLT usually works, BUT it

- Depends on data!
- Depends on precision!

> 经验上 $n \ge 30$ 时 CLT 就足够精确，但也要具体情况具体分析.

4.2.3 样本方差

> Sample Variance
>
> 样本方差是总体方差的估计量. 关于估计量，见 `5.1 估计量` .

Given a random sample $X_1, \dots, X_n$ with same PDF/PMF

the sample mean:

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

the sample variance:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$$

和样本均值类似，对于不同样本，样本方差也不固定. 和样本均值不同的是，上面这个样本方差公式是有偏的，不适合直接用于估计总体方差.

① 有偏性

> Biasedness

例子：$X \sim Bernoulli(\frac{1}{2})$. 总体方差 $\sigma^2 = p(1-p) = \frac{1}{4}$

> 注意：假设总体的每个数据被抽中概率相等（可以有重复数据，概率大表现为数据重复次数多），可以证明总体的均值/方差等于从总体随机抽取的随机变量的期望/方差.

Take 2 samples. What is the PMF of s^2?

Joint PMF of X_1, X_2

X_2 \ X_1	0	1
0	¼	¼
1	¼	¼

PMF of $s^2 = \frac{1}{2}\left((X_1 - \bar{X})^2 + (X_2 - \bar{X})^2\right)$

If $X_1 = X_2$, then $\bar{X} = X_1 = X_2$, and $s^2 = 0$

If $X_1 \neq X_2$, then $\bar{X} = 1/2$, and $s^2 = 1/4$

s^2	0	1/4
$P(S^2 = s^2)$	1/2	1/2

$$\mathbb{E}[s^2] = \frac{1}{8} = \frac{1}{2}\sigma^2 \neq \sigma^2$$

> Biased，样本方差采用原始定义，无法估计整体方差.

② 贝塞尔校正

> Bessel's correction

For a random sample of size n independent random variables $X_1, \dots, X_n$ with the same PDF/PMF, we have

$$\mathbb{E}[s^2] = \frac{n-1}{n}\sigma^2$$

> 这里的 σ^2 是总体方差，即 $Var[X_1] = \cdots = Var[X_n] = \sigma^2$.

记法：

只抽一个，样本方差恒为 0，期望也为 0，因此 $\mathbb{E}[s^2] = \frac{n-1}{n}\sigma^2$；

只抽一个，无法估计总体方差，分母有 $n-1$ 项，因此 $\sigma^2 = \frac{n}{n-1}\mathbb{E}[s^2]$;

样本低估了总体方差，因此要乘以一个放大系数 $\frac{n}{n-1}$ 才适合估计总体.

> 低估的原因见下面的理解.

贝塞尔校正 / Corrected sample variance / unbiased sample variance：

$$\frac{n}{n-1}s^2$$

贝塞尔校正后的样本方差公式：

$$\frac{n}{n-1}s^2 = \frac{n}{n-1}\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}$$

> 本课程中，$\frac{n}{n-1}s^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}$ 叫做 **Unbiased** Sample Variance，$s^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n}$ 叫做 Sample Variance. 如果没有强调无偏，默认用原始的分母为 n 的公式计算.

采用此公式计算样本方差，有

$$\mathbb{E}[\frac{n}{n-1}s^2] = \sigma^2$$

证明：

$$\begin{aligned}
\mathbb{E}[s^2] &= \mathbb{E}\left[\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^n \left(X_i^2 + \overline{X}^2 - 2X_i\overline{X}\right)}{n}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^n X_i^2}{n}\right] - \mathbb{E}[\overline{X}^2] \\
&= \frac{\sum_{i=1}^n \mathbb{E}[X_i^2]}{n} - \mathbb{E}[\overline{X}^2] \\
&= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

倒数第二行：

$$Var[X_i] = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 \Rightarrow \mathbb{E}[X_i^2] = \sigma^2 + \mu^2$$
$$Var[\overline{X}] = \mathbb{E}[\overline{X}^2] - \mathbb{E}[\overline{X}]^2 \Rightarrow \mathbb{E}[\overline{X}^2] = \frac{\sigma^2}{n} + \mu^2$$

> $Var[X_1] = \cdots = Var[X_n] = \sigma^2$
>
> $\mathbb{E}[X_i] = \mu$
>
> $Var[\overline{X}] = Var[\frac{X_1+\cdots+X_n}{n}] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$
>
> $\mathbb{E}[\overline{X}] = \mathbb{E}[\frac{X_1+\cdots+X_n}{n}] = \frac{n\mu}{n} = \mu$

理解：如果我们已知总体均值 μ，在样本中直接使用 μ 而不是样本均值 $\overline{X}$ 来计算方差，会发现它是无偏的，即

$$\begin{aligned}
\mathbb{E}\left[\frac{\sum_{i=1}^n (X_i-\mu)^2}{n}\right] &= \mathbb{E}\left[\frac{\sum_{i=1}^n (X_i^2+\mu^2-2X_i\mu)}{n}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^n X_i^2}{n}\right] + \mathbb{E}[\mu^2] - 2\mu\mathbb{E}[\overline{X}] \\
&= \frac{\sum_{i=1}^n \mathbb{E}[X_i^2]}{n} - \mu^2 \\
&= \sigma^2 + \mu^2 - \mu^2 \\
&= \sigma^2
\end{aligned}$$

但是我们通常不知道总体均值，因此不得不使用样本均值代替总体均值，这样通常会低估总体方差，因为

$$\begin{aligned}
\sigma^2 &= \mathbb{E}\left[\frac{\sum_{i=1}^n \left[X_i - (\overline{X}+\delta)\right]^2}{n}\right] \qquad \text{记 } \mu = \overline{X} + \delta \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^n \left(X_i - \overline{X} - \delta\right)^2}{n}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^n \left[(X_i-\overline{X})^2 + \delta^2 - 2(X_i-\overline{X})\delta\right]}{n}\right] \\
&= \mathbb{E}[s^2] + \mathbb{E}[\delta^2]
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\delta^2] &= Var[\delta] + \mathbb{E}[\delta]^2 \\
&= Var[\delta] \\
&= Var\left[\mu - \frac{X_1+\cdots+X_n}{n}\right] \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

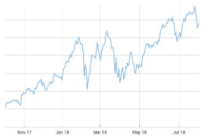
综上，$\sigma^2 = \frac{n}{n-1}\mathbb{E}[s^2]$.

Lec 5 点估计

> Point Estimation

$X \sim \text{Bernoulli}(p)$

$X \sim \text{Poisson}(\lambda)$



$X \sim \mathcal{N}(\mu, \sigma^2)$

Observation: a sample of n random variables $D = \{X_1, \dots, X_n\}$

Goal: to estimate θ (p, λ, μ, σ^2, etc.)

Classical statistics:

The parameter θ is considered as a deterministic quantity that happens to unknown

- Develop an **estimator** to estimate $\hat{\theta}$ based on the sample D

Bayesian statistics:

The parameter is considered as a random variable Θ with a known prior distribution (assumption)

- Estimate posterior probability of Θ: $f_{\Theta|D}(\theta|D)$ via Bayes' rule
- Use MAP to estimate $\hat{\theta}$

5.1 估计量

> Estimators

$X_1, \dots, X_n$ are independent samples with the same PDF/PMF parameterized by θ (unknown)

$$\hat{\Theta}_n = g(X_1, \dots, X_n) \qquad \text{Estimator}$$
$$\hat{\theta}_n = g(X_1 = x_1, \dots, X_n = x_n) \qquad \text{Estimate}$$

> Estimator 是估计量，Estimate 是估计值. 估计量是基于样本数据构造的函数，估计量本身也是一个随机变量. 估计值是在观测到具体数据之后，将这些数据代入估计量函数得到的具体数值，它是对未知参数 θ 的一个固定数值的近似.

5.1.1 无偏性

Unbiased: $\mathbb{E}[\hat{\Theta}_n] = \theta$

5.1.2 渐进无偏性

Asymptotically unbiased: $\lim_{n\to\infty} \mathbb{E}[\hat{\Theta}_n] = \theta$

5.1.3 一致性

Consistent: $\hat{\Theta}_n$ converges to θ in probability

$$\lim_{\epsilon\to 0}\lim_{n\to\infty} P(|\hat{\Theta}_n - \theta| \geq \epsilon) = 0$$

5.2 最大似然估计

> Maximum Likelihood Estimation

$X_1, \ldots, X_n$ are independent samples with the same PDF $f_{X|\Theta}(x|\theta)$ (or PMF $P_{X|\Theta}(x|\theta)$ for discrete cases)

Maximum likelihood estimate (MLE) of θ

$$\hat{\theta}_n = \arg\max_{\theta} f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\theta)$$

Motivation behind MLE: for each specific value of θ,

- We have a specified PDF $f_{X|\Theta}(x|\theta)$ for the random samples
- $f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\theta)$ is a joint PDF of $X_1, \ldots, X_n$ defined by θ

  > $f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\theta)$ measures the likelihood that $x_1, \ldots, x_n$ are observed at the same time (i.e., jointly)

- Among all the possible values, choose θ^* which achieves the maximum value of $f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\theta)$

  > the joint PDF fits the observed data best
  >
  > 意思是，取 θ^* 时出现这些观测值的概率（密度）最大.

$$\hat{\theta}_n = \arg\max_{\theta} f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\theta)$$
$$\Rightarrow f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\hat{\theta}_n) = \max_{\theta} f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\theta)$$

5.2.1 例子

Example 1: What is the MLE for θ from $Uniform(0, \theta)$ samples? Observe x_1, x_2, x_3 independently from $Uniform(0, \theta)$

$$\begin{aligned} f_{X_1,X_2,X_3|\Theta}(x_1, x_2, x_3|\theta) &= f_{X_1|\Theta}(x_1|\theta) f_{X_2|\Theta}(x_2|\theta) f_{X_3|\Theta}(x_3|\theta) \\ &= \frac{1}{\theta^3} \text{ if } \theta \geq x_1, x_2, x_3 > 0 \end{aligned}$$

$\frac{1}{\theta^3}$ is a decreasing function when $\theta > 0$

When $\theta = \max\{x_1, x_2, x_3\}$, $\frac{1}{\theta^3}$ reaches its maximum

$$\theta_{MLE} = \max\{x_1, x_2, x_3\}$$

Example 2: MLE for $Bernoulli(\theta)$. Suppose we observe k heads and $n - k$ tails. What is θ_{MLE}?

$$\begin{aligned} \theta_{MLE} &= \arg\max_\theta f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\theta) \\ &= \arg\max_\theta \theta^k (1-\theta)^{n-k} \\ &= \arg\max_\theta f_{\Theta \sim Beta(k+1, n-k+1)}(\theta) \\ &= \frac{\alpha - 1}{\alpha - 1 + \beta - 1} \\ &= \frac{k}{n} \end{aligned}$$

5.2.2 系统解法

> A systematic approach to the MLE

MLE: $\hat{\theta} = \arg\max_\theta f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\theta)$

- If θ has discrete values, for each possible value, compute $f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\theta)$ and choose the one that maximizes $f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\theta)$
- If θ has continuous values, based on the properties of $f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\theta)$ to find θ_{MLE} (as shown in the previous two examples)
- What if $f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\theta)$ is too complicated to find θ_{MLE} directly?

A systematic approach: 假设 $f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\theta)$ is differentiable w.r.t. θ. To find θ that maximizes the function $f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\theta)$:

$$\frac{\partial f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\theta)}{\partial \theta} = 0$$

- If the equation can be solved, then we get a closed-form (analytical) solution for θ_{MLE}
- Otherwise, optimization techniques need to be applied

> Out of scope

What if there are more than one parameters to estimate?

$$\{\hat{\theta}_1, \ldots, \hat{\theta}_m\} = \arg\max_{\theta_1,\ldots,\theta_m} f_{X_1,\ldots,X_n|\Theta_1,\ldots,\Theta_m}(x_1, \ldots, x_n|\theta_1, \ldots, \theta_m)$$

$$\begin{cases} \frac{\partial f_{X_1,\ldots,X_n|\Theta_1,\ldots,\Theta_m}(x_1,\ldots,x_n|\theta_1,\ldots,\theta_m)}{\partial \theta_1} = 0 \\ \ldots \\ \frac{\partial f_{X_1,\ldots,X_n|\Theta_1,\ldots,\Theta_m}(x_1,\ldots,x_n|\theta_1,\ldots,\theta_m)}{\partial \theta_m} = 0 \end{cases}$$

Closed-form (analytical) solutions for $\hat{\theta}_1, \ldots, \hat{\theta}_m$ can be obtained by solving the equations jointly

5.3 对数似然

> Log-likelihood

最大似然估计中，求解 $\frac{\partial f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_n|\theta)}{\partial \theta} = 0$ 有一些潜在的困难：

As $X_1, \ldots, X_n$ are independent

$$\frac{\partial f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_n|\theta)}{\partial \theta} = \frac{\partial \prod_{i=1}^{n} f_{X_i|\Theta}(x_i|\theta)}{\partial \theta}$$

连乘得到的函数 is complicated, especially when n is large.

定义对数似然：$\ln\left(f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_n|\theta)\right)$

> 即似然取对数.

因为对数函数在定义域上单调递增，所以最大化似然等价于最大化对数似然：

$$\hat{\theta}_n = \arg\max_{\theta} f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_n|\theta) = \arg\max_{\theta} \ln\left(f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_n|\theta)\right)$$

好处：对数函数可以将乘积转化为和的形式.

$$\ln\left(f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_n|\theta)\right) = \ln\left(\prod_{i=1}^{n} f_{X_i|\Theta}(x_i|\theta)\right) = \sum_{i=1}^{n} \ln\left(f_{X_i|\Theta}(x_i|\theta)\right)$$

此时，系统解法等价于

$$\{\hat{\theta}_1, \ldots, \hat{\theta}_m\} = \arg\max_{\theta_1,\ldots,\theta_m} f_{X_1,\ldots,X_n|\Theta_1,\ldots,\Theta_m}(x_1,\ldots,x_n|\theta_1,\ldots,\theta_m)$$

$$\begin{cases} \frac{\partial \ln\left(f_{X_1,\ldots,X_n|\Theta_1,\ldots,\Theta_m}(x_1,\ldots,x_n|\theta_1,\ldots,\theta_m)\right)}{\partial \theta_1} = \frac{\partial \sum_{i=1}^{n} \ln\left(f_{X_i|\Theta_1,\ldots,\Theta_m}(x_i|\theta_1,\ldots,\theta_m)\right)}{\partial \theta_1} = 0 \\ \ldots \\ \frac{\partial \ln\left(f_{X_1,\ldots,X_n|\Theta_1,\ldots,\Theta_m}(x_1,\ldots,x_n|\theta_1,\ldots,\theta_m)\right)}{\partial \theta_m} = \frac{\partial \sum_{i=1}^{n} \ln\left(f_{X_i|\Theta_1,\ldots,\Theta_m}(x_i|\theta_1,\ldots,\theta_m)\right)}{\partial \theta_m} = 0 \end{cases}$$

Closed-form (analytical) solutions for $\hat{\theta}_1, \ldots, \hat{\theta}_m$ can be obtained by solving the equations jointly

Example

> 待补充.

5.4 MAP vs MLE

> 待补充.

Lec 6 置信区间

6.1 置信区间 I

> Confidence Intervals

Suppose θ is an unknown parameter

Besides a single numerical estimate $\hat{\theta}_n$ of θ based on a specific set of n observed samples, we are often interested in constructing a so-called confidence interval: An interval that contains θ with a certain high probability (i.e., an interval in which we are certain confident θ falls)

Consider an estimator Θ_n of an unknown θ

For different sets of n samples, the estimates $\hat{\theta}_n$'s would be very different

Unbiasedness, asymptotical unbiasedness, and consistency are properties about an estimator NOT a specific estimate

When a specific $\hat{\theta}_n$ is used to approximate θ, can we be confident that $\hat{\theta}_n$ is close to θ? Or how close is $\hat{\theta}_n$ to θ?

High-level idea: rather than using just a point estimate $\hat{\theta}_n$, to estimate an interval of values that we are confident contains the unknown θ

Solution: based on a point estimate $\hat{\theta}_n$, create an interval $[\hat{\theta}_n^-, \hat{\theta}_n^+]$, where $\hat{\theta}_n^- < \hat{\theta}_n^+$, such that we are confident that θ falls in the interval

$$P(\hat{\theta}_n^- \leq \theta \leq \hat{\theta}_n^+) \geq 1-\alpha$$

- $1-\alpha$ 称为置信水平（Confidence level），以百分比表示. α 称为显著性水平.
- $[\hat{\theta}_n^-, \hat{\theta}_n^+]$ is called a $(1-\alpha)$-confidence interval
- $\hat{\theta}_n^-$ is called 置信下限（lower confidence limit）, and $\hat{\theta}_n^+$ is called 置信上限（upper confidence limit）.
- Width: $\hat{\theta}_n^+ - \hat{\theta}_n^-$
- Confidence parameter: α

Goal: Given a large $(1-\alpha)$, find a **narrowest** confidence interval:

$$P(\hat{\theta}_n^- \leq \theta \leq \hat{\theta}_n^+) = 1-\alpha$$

6.1.1 定义

(2025.3.10 思考) 以上是课件给出的定义，但是课件后面马上又把它证伪了. 因为这个定义是错误的. 一般而言，定义置信区间有两种方式：

① 基于随机样本的定义

考虑一个一维随机变量 $\mathcal{X}$ 服从分布 $\mathcal{F}$，又假设 θ 是 $\mathcal{F}$ 的参数之一. 独立抽样 n 次，得到一个随机样本 $\{X_1, \ldots, X_n\}$，注意这里所有 X_i 都是随机的，我们是在讨论一个尚未被观测的数据集. 如果存在统计量 $u(X_1, \ldots, X_n), v(X_1, \ldots, X_n)$ 满足 $u(X_1, \ldots, X_n) < v(X_1, \ldots, X_n)$ 使得：

$$P(u(X_1, \ldots, X_n) \leq \theta \leq v(X_1, \ldots, X_n)) = 1 - \alpha$$

则称 $[u(X_1, \ldots, X_n), v(X_1, \ldots, X_n)]$ 为一个用于估计参数 θ 的 $1 - \alpha$ 置信区间，其中：

- $1 - \alpha$ 称为置信水平.
- α 在假设检验中称为显著性水平.

> 统计量定义为样本 $\{X_1, \ldots, X_n\}$ 的一个函数，且不依赖于任何未知参数.
>
> 注意这个区间的上下限都是随机变量，并不是严格定义的区间.
>
> 这里概率的意思是，如果不断重复这样的 n 次独立抽样（即重复观测），会有 $1 - \alpha$ 的概率观测到 $\{x_1, \ldots, x_n\}$，使得 $u(x_1, \ldots, x_n) \leq \theta \leq v(x_1, \ldots, x_n)$.

② 基于观测样本的定义

接续 ① 基于随机样本的定义. 现在，对于随机变量 $\mathcal{X}$ 的一个已观测到的样本 $\{x_1, \ldots, x_n\}$，注意这里所有 x_i 都是已观测到的数值，没有随机性. 定义基于观测样本的 $1 - \alpha$ 置信区间为：

$$[u(x_1, \ldots, x_n), v(x_1, \ldots, x_n)]$$

> 这里的区间上下限都是确定的值，符合严格的区间定义，但就不能用概率来解释了.

而课件这里采用了基于观测样本的定义，但是却引入了概率，混淆概念.

注意：

- 置信区间**不是**总体参数的概率范围. 一个基于观测样本的置信区间要么包含真实值，要么不包含（概率为 0 或 1），但在多次重复抽样观测中，一定比例的区间能覆盖总体参数.
- 置信水平**不等于**覆盖真实值的概率. 例如，95% 置信区间不是说 “有 95% 的概率包含总体参数”，而是 “如果重复抽样，95% 的置信区间会包含总体参数（抽到包含总体参数的区间的概率为 95%）”.

> 下面在 `6.1.2 总体均值的置信区间` 中亦有解释.

特别地，$(-\infty, +\infty)$ is a 100%-confidence interval.

> 但是，Uninformative.

6.1.2 总体均值的置信区间

> Confidence interval for mean
>
> 总体均值是一个参数，样本均值是它的估计量.

$X_1, \dots, X_n$ are independent samples with the same PDF/PMF (i.e., have the same μ, σ^2, etc.)

If $X_1, \dots, X_n$ are normal $N(\mu, \sigma^2)$, then $\overline{X} = \frac{X_1+\cdots+X_n}{n}$ is also normal $N(\mu, (\frac{\sigma}{\sqrt{n}})^2)$

If $X_1, \dots, X_n$ are NOT normal, but n is large (e.g., $n \geq 30$), then $\overline{X} = \frac{X_1+\cdots+X_n}{n}$ can be approximated by a normal $N(\mu, (\frac{\sigma}{\sqrt{n}})^2)$ based on CLT

标准化：If $X_1, \dots, X_n$ are normal $N(\mu, \sigma^2)$ or n is large

$$\overline{X} \sim N(\mu, (\frac{\sigma}{\sqrt{n}})^2) \qquad Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

① 总体方差已知

Suppose σ^2 is known, then a $(1-\alpha)$-confidence interval for the mean μ is

$$\overline{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \Leftrightarrow [\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}]$$

> 注意这里的 $\overline{x}$ 和 $z_{\frac{\alpha}{2}}$ 都是小写字母，即确定的值，而非随机分布.

其中，

$\overline{x}$: A sample mean estimate based on observed samples

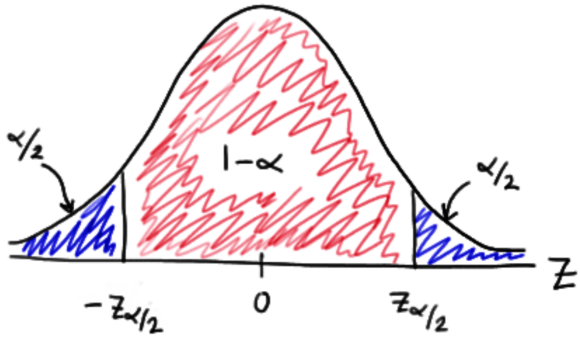
> 每次抽样都能得到不同的 $\overline{x}$，但是总体均值 μ 是固定的.

$z_{\frac{\alpha}{2}}$: z-value or z-score such that the area of the right of it under the standard normal curve is $\frac{\alpha}{2}$.

CDF table of standard normal distribution $P(Z \leq z)$

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

> 注意，$z_{\frac{\alpha}{2}}$ 的 $\frac{\alpha}{2}$ 只是一种记法，并不是查表时直接查 $\frac{\alpha}{2}$. 通常给的表是标准正态的 CDF 表，即 $Z \leq z$ 对于不同 z 的概率. 由于标准正态总面积为 1，则左侧面积为 $1-\frac{\alpha}{2}$，该面积即 $Z \leq z_{\frac{\alpha}{2}}$ 的概率，通常代入这个概率值去查表，查出对应的 $z_{\frac{\alpha}{2}}$.



> 这里也可以解释：$z = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ 如果落在 $[-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}]$ 之间，等价于 μ 在 $[\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}]$ 之间. 但是小写的 z 要么在 $[-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}]$ 之间，要么不在，概率为 0 或 1，不是一个随机事件. 只有还没观测的大写 Z 才能说在 $[-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}]$ 之间的概率为 $1-\alpha$.

注意：可以说 $[\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}]$ is a $(1-\alpha)$-confidence interval for the mean μ，但是不能说 $P(\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$，理由如下：

> 在 `6.1.1 定义` 中亦有解释.

$$
\begin{aligned}
& P(-z_{\frac{\alpha}{2}} \leq Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}) = 1 - \alpha \\
\Leftrightarrow\ & P(-z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha \\
\Leftrightarrow\ & P(-z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} - \overline{X} \leq -\mu \leq z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} - \overline{X}) = 1 - \alpha \\
\Leftrightarrow\ & P(\overline{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha
\end{aligned}
$$

注意大小写. As $\overline{X}$ is a random variable (varies across different samples), the above equation describes a random interval centered at $\overline{X}$ that has a $1 - \alpha$ probability of containing the (deterministic) population mean μ before a sample is drawn

Once a specific sample is observed, the sample mean $\overline{x}$, the interval becomes **fixed**:

$$
[\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}]
$$

This interval is no longer random: it either contains μ or it does not.

The $1 - \alpha$ confidence level: if we repeatedly sampled and computed intervals this way, $100(1 - \alpha)\%$ of those intervals would contain μ

For a specific interval, we say we are "$100(1 - \alpha)\%$ confident" it contains μ (not a probability statement)

例 1: Give a 95%-confidence interval for the mean from 30 $N(\mu, (\frac{1}{2})^2)$ samples

Solution: As $1 - \alpha = 95\%$, thus $\frac{\alpha}{2} = 0.025$.

Denote by $\overline{x}$ a sample mean estimate of the 30 samples

$\sigma = \frac{1}{2}$ is known

$z_{0.025} = P(Z \leq 0.975) \approx 1.96$

The corresponding confidence interval:

$$
\begin{aligned}
& [\overline{x} - z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}] \\
\approx\ & [\overline{x} - 1.96 \times \frac{\frac{1}{2}}{\sqrt{30}}, \overline{x} + 1.96 \times \frac{\frac{1}{2}}{\sqrt{30}}] \\
\approx\ & [\overline{x} - 0.18, \overline{x} + 0.18]
\end{aligned}
$$

例 2: How many $N(\mu, 25^2)$ samples do you need for a 95% confidence, width 10 interval?

Solution: As $1 - \alpha = 95\%$, thus $\frac{\alpha}{2} = 0.025$.

The corresponding confidence interval:

$$
[\overline{x} - z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}]
$$

$\text{Width} = 2 \cdot z_{0.025} \cdot \frac{\sigma}{\sqrt{n}} = 10 \Rightarrow n = (\frac{2 \cdot z_{0.025} \cdot \sigma}{10})^2 \approx 96$

注意:

① If the problem merely asks us to determine the interval without including terms like "at least," then simply rounding to the nearest integer is sufficient;

> 直接四舍五入，n 取 96.

② If we keep the width fixed and set a lower bound for the confidence level (for example, at least 95%), then n should be calculated using ceil rounding:

$$2z_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}}) = 10 \Rightarrow n = (\frac{2z_{\frac{\alpha}{2}} \cdot \sigma}{10})^2 \geq (\frac{2 \cdot z_{0.025} \cdot \sigma}{10})^2 = 96.04$$

> 上取整，n 取 97.

③ If we keep the confidence level fixed and set a lower bound for the width (for example, at least 10), then n should be calculated using floor rounding:

$$2z_{0.025}(\frac{\sigma}{\sqrt{n}}) \geq 10 \Rightarrow n \leq (\frac{2 \cdot z_{0.025} \cdot \sigma}{10})^2 = 96.04$$

> 下取整，n 取 96.

④ The lower bounds for the confidence level and width jointly affect the value of n, but their effects are in opposite directions.

② 总体方差未知

Suppose σ^2 is unknown, but n is large (e.g., ≥ 30) then a $(1-\alpha)$-confidence interval for the mean μ is

$$[\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{s'}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{s'}{\sqrt{n}}]$$

其中，s' is an **unbiased** sample standard deviation estimate based on observed samples

$$s'^2 = \frac{n}{n-1}s^2 = \frac{n}{n-1}\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

> 见 `4.2.3 样本方差 ② 贝塞尔校正`.
>
> 此公式描述的是估计量，上述标红的 s' 是估计值，即 $s' = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$，用小写的 x_i 和 $\overline{x}$ 代入计算.

Why? Explanation in L7.

Note: the case is unknown and n is small (small sample size) will be discussed in L7.

> L07 is `6.2 置信区间 II` in this note.
>
> 见 `6.2.4 总体均值的置信区间`.

例 1: 34 of 100 $Bernoulli(p)$ samples came out positive. Give a 95% confidence interval.

$$\overline{x} = \hat{p} = \frac{34}{100} = 0.34$$

σ is unknown, but $n = 100$ is large

We can use $s = \sqrt{\hat{p}(1-\hat{p})} \approx 0.47$ to approximate σ

> In this course, for Bernoulli distributions, we use $\hat{p}(1-\hat{p})$ to approximate σ^2 without considering the biased issue

As $1-\alpha = 95\%$, thus $\frac{\alpha}{2} = 0.025$

The corresponding confidence interval:

$$\begin{aligned} & [\overline{x} - z_{0.025} \cdot \frac{s}{\sqrt{n}}, \overline{x} + z_{0.025} \cdot \frac{s}{\sqrt{n}}] \\ \approx\ & [0.34 - 1.96 \times \frac{0.47}{\sqrt{100}}, 0.34 + 1.96 \times \frac{0.47}{\sqrt{100}}] \\ \approx\ & [0.248, 0.432] \end{aligned}$$

> 这里 0.47 如果用 $\sqrt{\hat{p}(1-\hat{p})}$ 直接代入，结果略有不同.

6.1.3 单侧置信区间

> One sided confidence intervals

Provides a bound (either upper or lower) for a population parameter with a specific confidence level $1-\alpha$. A one-sided interval addresses questions where only one directions is of interest.

e.g., Is θ at least this value? Or is θ at most this value?

① 单侧置信限

单侧置信区间（Lower one-sided confidence intervals）

$[\hat{\theta}_n^{\min}, +\infty)$: we are confident θ is at least equal to $\hat{\theta}_n^{\min}$.

其中，$\hat{\theta}_n^{\min}$ 为单侧置信下限.

单侧置信区间（Upper one-sided confidence intervals）

$(-\infty, \hat{\theta}_n^{\max}]$: we are confident θ is at most equal to $\hat{\theta}_n^{\max}$.

其中，$\hat{\theta}_n^{\max}$ 为单侧置信上限.

中文教材统称单侧置信区间，英文教材有 Lower 和 Upper 的细分.

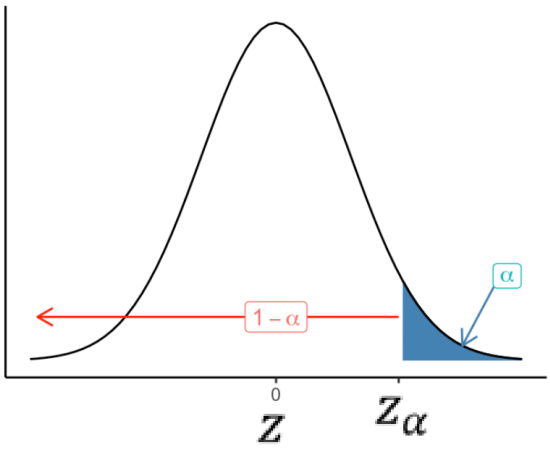
例 1: find $\hat{\theta}_n^{\min}$ such that $P(\mu \geq \hat{\theta}_n^{\min}) = 1 - \alpha$.

Solution:

$$\overline{X} \sim N(\mu, (\frac{\sigma}{\sqrt{n}})^2) \qquad Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

$$P\left(Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_\alpha\right) = 1 - \alpha \Leftrightarrow P\left(\mu \geq \overline{X} - z_\alpha \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

其中，z_α is the z-score such that the area to its right under the standard normal curve is α.



> 同理，查表不是直接查 α，而是查 $1 - \alpha$.

结论：Given a specific sample mean $\overline{x}$, the lower one-sided confidence interval is $[\overline{x} - z_\alpha \cdot \frac{\sigma}{\sqrt{n}}, +\infty)$.

> When σ^2 is unknown and $n \geq 30$, use unbiased s' to approximate σ.

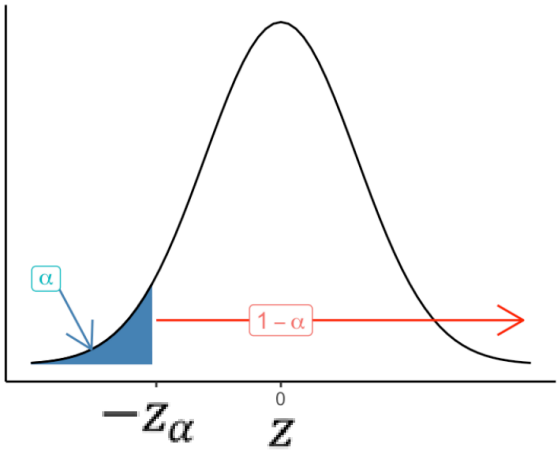
例 2: find $\hat{\theta}_n^{\max}$ such that $P(\mu \leq \hat{\theta}_n^{\max}) = 1 - \alpha$.

Solution:

$$\overline{X} \sim N(\mu, (\frac{\sigma}{\sqrt{n}})^2) \qquad Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

$$P\left(Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \geq -z_\alpha\right) = 1 - \alpha \Leftrightarrow P\left(\mu \leq \overline{X} + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

其中，z_α is the z-score such that the area to its right under the standard normal curve is α.



> 同理，查表不是直接查 α，而是查 $1-\alpha$.

结论：Given a specific sample mean $\overline{x}$, the upper one-sided confidence interval is $(-\infty, \overline{x}+z_\alpha \cdot \frac{\sigma}{\sqrt{n}}]$.

> When σ^2 is unknown and $n \geq 30$, use unbiased s' to approximate σ.

6.2 置信区间 II

> Confidence intervals II

6.2.1 卡方分布

> χ^2-distribution

$X_1, \ldots, X_n$ are independent standard normal $N(0,1)$. Then,

$$X = X_1^2 + \cdots + X_n^2 \sim \chi^2(n).$$

其中，$X = \sum_{i=1}^{n} X_i^2$ 称为 $\chi^2(n)$ random variable，$\chi^2(n)$ 称为自由度为 n 的卡方分布.

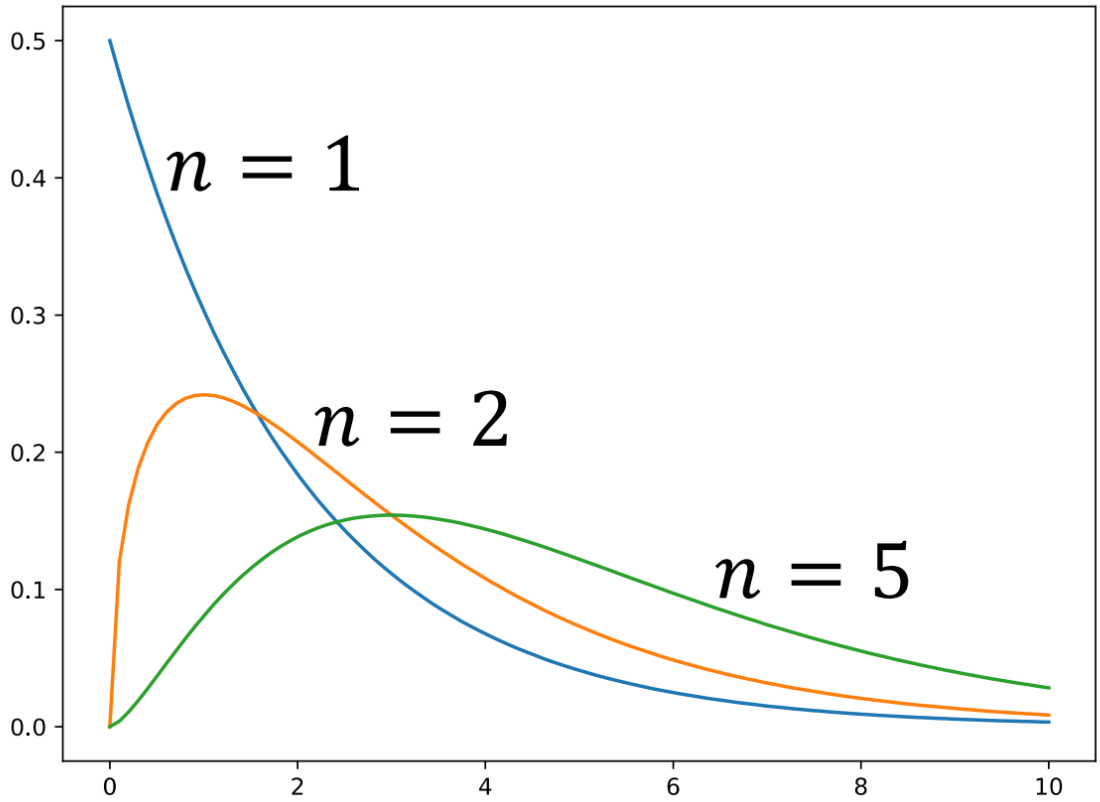
卡方分布是一种特殊的 Gamma 分布.

> $\alpha = \frac{n}{2}$, $\beta = \frac{1}{2}$.

概率密度函数：

$$f_{X\sim\chi^2(n)}(x)=\begin{cases}\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}e^{-\frac{x}{2}} & x\geq 0,\\ 0 & x<0.\end{cases}$$

> 注意，PDF 在 $x=0$ 处的具体表现需要分类讨论. 这里的公式并不严谨. 见 附录 2.2 连续分布 ④ Gamma 分布.



6.2.2 学生 t 分布

> Student's t-distribution

假设:

- $Y\sim N(0,1)$，PDF 为

$$f_Y(y)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right);$$

- $Z\sim\chi^2(n)$，PDF 为

$$f_{Z\sim\chi^2(n)}(z)=\begin{cases}\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}z^{\frac{n}{2}-1}e^{-\frac{z}{2}} & z\geq 0,\\ 0 & z<0.\end{cases}$$

- Y 与 Z 相互独立（独立抽样 / 分别抽样）.

构造变量 $X=\frac{Y}{\sqrt{\frac{Z}{n}}}=\frac{N(0,1)}{\sqrt{\frac{\chi^2(n)}{n}}}$. 随机变量 X 服从的分布，我们称（定义）它为自由度为 n 的 t 分布，记为 $X\sim t(n)$.

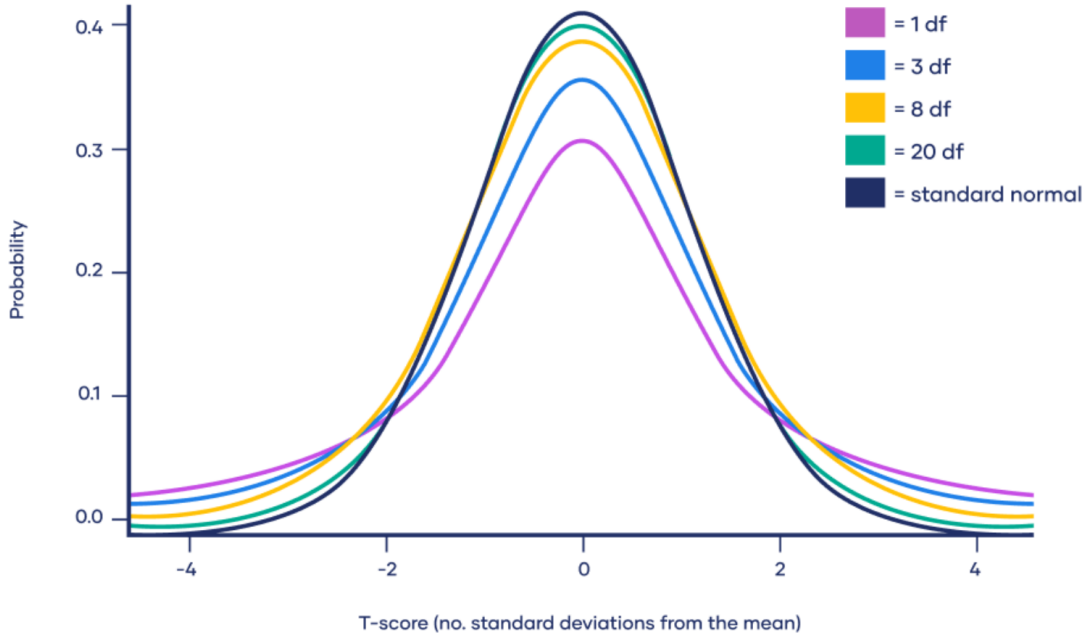
> 这是基于概率分布的定义，也可以基于 $T=\frac{\overline{X}-\mu}{\frac{S'}{\sqrt{n}}}\sim t(n-1)$ 定义. 基于哪种定义不重要，因为可以证明二者等价.
>
> 第二个等号右边的写法不严谨，分布不能直接写在等式里. 本质是基于第一个等号右边的比值构造.
>
> 注意，这里分子分母是两步独立抽样. 分子是标准正态抽样，分母是卡方抽样并进行处理（除以 n 开根号），然后通过它们的比值构造出 t 统计量. 而下一节的 $T=\frac{\overline{X}-\mu}{\frac{S'}{\sqrt{n}}}\sim t(n-1)$ 只需要一步抽样，即从一个正态总体中抽取样本，然后利用样本均值和样本标准差构造 t 统计量（注意样本均值和样本标准差不独立，不能抽两次，一次算均值一次算标准差，只能用同一批样本）.

随 n 增大，自由度为 n 的 t 分布 PDF 收敛于标准正态分布的 PDF.

> 证明见 `附录 3. 置信区间 Q & A` 第三问.

$$t(n)=\frac{\mathcal{N}(0,1)}{\sqrt{\chi^2(n)/n}}\xrightarrow[\approx]{n\to\infty}\mathcal{N}(0,1)$$

when $n\geq 30$



实践中，$n\geq 30$ 即被认为足够大，可以用标准正态近似 t 分布，因此 $n\geq 30$ 时，无论方差是否已知，我们都使用 z 分布来求置信区间.

6.2.3 t 分布定理

> Theorem regarding t-distribution

如果一组随机变量 $X_1,\dots,X_n$ 是 IID 正态随机变量，即

$$X_i\sim N(\mu,\sigma^2)$$

则无论样本量 n is large or not, we have

$$T=\frac{\overline{X}-\mu}{\frac{S'}{\sqrt{n}}}\sim t(n-1)$$

其中，

- $\overline{X}=\frac{1}{n}\sum\limits_{i=1}^{n}X_i$ 是样本均值.
- $S'^2=\frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\overline{X})^2$ 是无偏样本方差.
- $t(n-1)$ 表示自由度为 $n-1$ 的 t 分布.

> 定理证明：待补充.
>
> 注意 $T=\frac{\overline{X}-\mu}{\frac{S'}{\sqrt{n}}}$ 的 $\overline{X}$ 是样本均值，μ 是已知的总体均值；而 $S'^2=\frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\overline{X})^2$ 的 $\overline{X}$ 是样本均值. 即样本均值用到两次，总体均值用到一次.
>
> 同样要注意，$T=\frac{\overline{X}-\mu}{\frac{S'}{\sqrt{n}}}$ 用到的是 $\sqrt{n}$，而 $S'^2=\frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\overline{X})^2$ 用到的是 $n-1$.

The PDF of T is:

$$f_{T\sim t(\nu)}(t)=\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})}(1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}},$$

where ν is the number of degrees of freedom and Γ is the gamma function. This may also be written as

$$f_{T\sim t(\nu)}(t)=\frac{1}{\sqrt{\nu}B(\frac{1}{2},\frac{\nu}{2})}(1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}},$$

where B is the beta function.

> $\nu=n-1$，自由度.
>
> 这个 PDF 即自由度为 $n-1$ 的 t 分布的 PDF.
>
> PDF 推导：待补充.

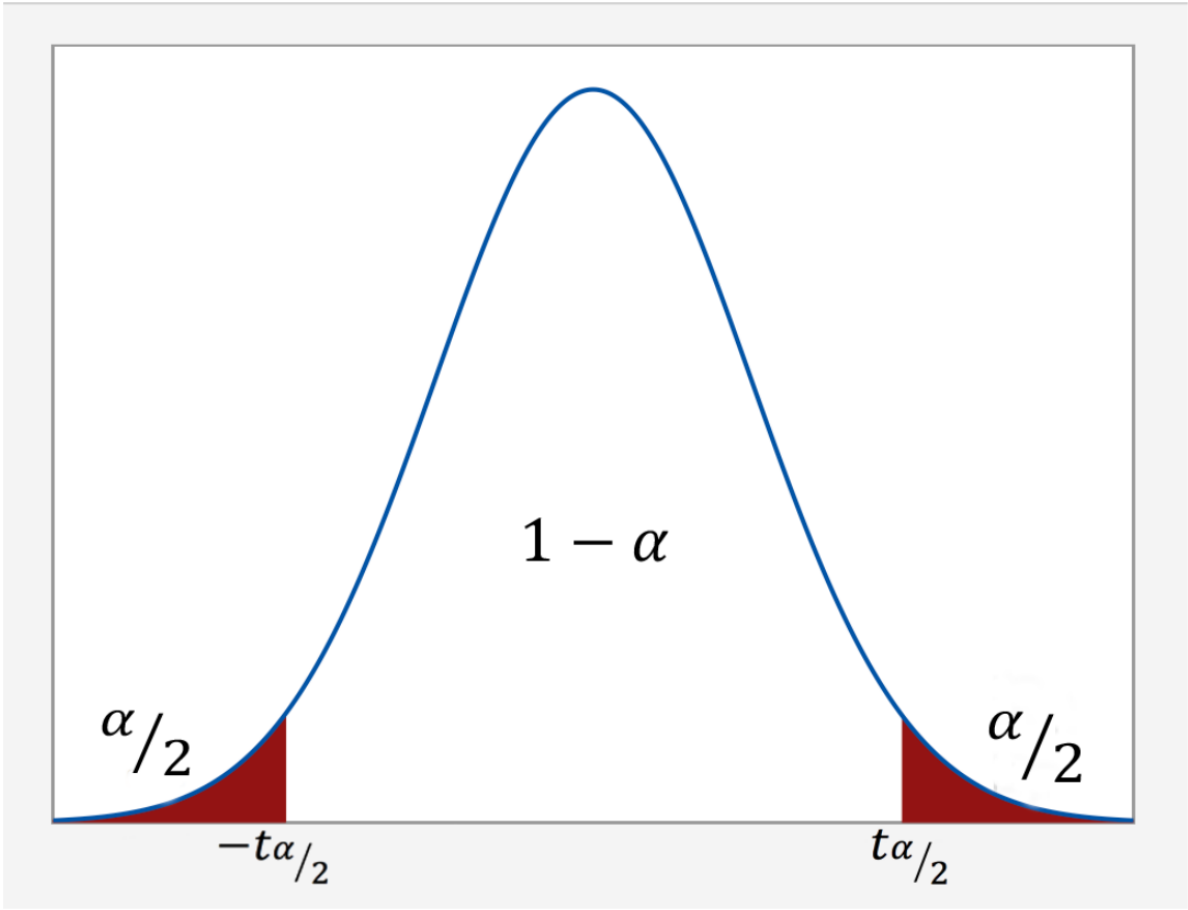
6.2.4 总体均值的置信区间

接 `6.1.2 总体均值的置信区间 ② 总体方差未知`，当 σ^2 is unknown and n is small, if $X_1,\dots,X_n$ are independent $N(\mu,\sigma^2)$, then a $(1-\alpha)$-confidence inteval for mean μ is

$$\overline{x}\pm t_{\frac{\alpha}{2}}\cdot\frac{s'}{\sqrt{n}}$$

> 这里 s' 来自无偏估计 $s'=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$.
>
> $t_{\frac{\alpha}{2}}$: t-value or t-score s.t. the area of the right of it under the t-distribution curve with degrees-of-freedom $\nu=n-1$ is $\frac{\alpha}{2}$.

> 关于 t-value 和 t-distribution，见 `6.2.2 学生 t 分布`.



证明：Given $X_1, \dots, X_n$ independent $N(\mu, \sigma^2)$ random variables, where σ^2 is unknown, and n is small ($n < 30$). 根据 t 分布定理，有

$$T = \frac{\overline{X} - \mu}{\frac{S'}{\sqrt{n}}} \sim t(n-1)$$

Then,

$$\begin{aligned} & P(-t_{\frac{\alpha}{2}} \le T \le t_{\frac{\alpha}{2}}) = 1 - \alpha \\ \Leftrightarrow\ & P(\overline{X} - t_{\frac{\alpha}{2}} \cdot \frac{S'}{\sqrt{n}} \le \mu \le \overline{X} + t_{\frac{\alpha}{2}} \cdot \frac{S'}{\sqrt{n}}) = 1 - \alpha \end{aligned}$$

> 注意，这里要满足 t 分布 PDF 关于 0 对称的前提，由 PDF 很容易得到这个结论.

Therefore, given specific $\overline{x}$ and s, we construct the $(1-\alpha)$-confidence inteval for mean μ as

$$\left[\overline{x} - t_{\frac{\alpha}{2}} \cdot \frac{s'}{\sqrt{n}}, \overline{x} + t_{\frac{\alpha}{2}} \cdot \frac{s'}{\sqrt{n}}\right]$$

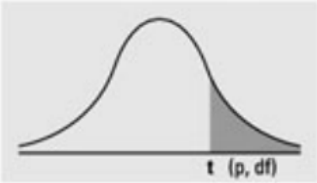
> 这里的置信水平 $1-\alpha$ 并不表示 μ 落在区间 $\left[\overline{x} - t_{\frac{\alpha}{2}} \cdot \frac{s'}{\sqrt{n}}, \overline{x} + t_{\frac{\alpha}{2}} \cdot \frac{s'}{\sqrt{n}}\right]$ 的概率为 $1-\alpha$. 原因同 `6.1.1 定义` 中的「注意」.

6.2.5 t 值表

> T-table for t-distribution

由于 $t_{\frac{\alpha}{2}}$ 难以直接计算得到，通常查表：

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%

First row: probability $P(T \geq t)$

First column: degrees of freedom

例: 5 random athletes are $152, 163, 188, 201, 192$ cm tall (following a normal distribution). Give a 95%-confidence inteval for μ.

解：$n=5$，The degrees of freedom $\nu=n-1=4$.

$\overline{x}=\frac{152+163+188+201+192}{5}=179.2$.

$s'=\sqrt{\frac{\sum_{i=1}^{5}(x_i-179.2)^2}{4}}\approx 20.73$.

$\alpha=5\%,\ \frac{\alpha}{2}=0.025\Rightarrow t_{\frac{\alpha}{2}}\approx 2.78$.

$\left[\overline{x}-t_{\frac{\alpha}{2}}\cdot\frac{s'}{\sqrt{n}},\overline{x}+t_{\frac{\alpha}{2}}\cdot\frac{s'}{\sqrt{n}}\right]\approx[153.43,204.97]$.

> 这里约等并不是使用了近似方法，而是计算上保留位数带来的误差.

6.2.6 单侧置信区间

> One-sided confidence intervals
>
> 思路同 `6.1.3 单侧置信区间` .

$X_1,\dots,X_n$ are independent $N(\mu,\sigma^2)$, $n<30$, and σ^2 is unknown.

Lower one-sided: find $\hat{\theta}_n^{\min}$ such that $P(\mu\geq\hat{\theta}_n^{\min})=1-\alpha$.

$$P(T\leq t_\alpha)=1-\alpha\Leftrightarrow P(\mu\geq\overline{X}-t_\alpha\cdot\frac{S'}{\sqrt{n}})=1-\alpha$$

结论：Given a specific $\overline{x}$ and s', the lower one-sided confidence interval is $[\overline{x}-t_\alpha\cdot\frac{s'}{\sqrt{n}},+\infty)$.

Upper one-sided: find $\hat{\theta}_n^{\max}$ such that $P(\mu\leq\hat{\theta}_n^{\max})=1-\alpha$.

$$P(T\geq -t_\alpha)=1-\alpha\Leftrightarrow P(\mu\leq\overline{X}+t_\alpha\cdot\frac{S'}{\sqrt{n}})=1-\alpha$$

结论：Given a specific $\overline{x}$ and s', the upper one-sided confidence interval is $(-\infty,\overline{x}+t_\alpha\cdot\frac{s'}{\sqrt{n}}]$.

6.3 总结

> Summary: confidence intervals for μ

$X_1,\dots,X_n$ are independent samples with same PMF/PDF, then

a $(1-\alpha)$-confidence interval for the mean μ is:

① When σ^2 is known,

sample size \ IID X_i	**normal**	**unknown**
small	$[\overline{x}-z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}},\overline{x}+z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}]$	not taught
large	$[\overline{x}-z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}},\overline{x}+z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}]$	$[\overline{x}-z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}},\overline{x}+z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}]$

② When σ^2 is unknown,

sample size \ IID X_i	normal	unknown
small	$[\overline{x}-t_{\frac{\alpha}{2}}\cdot\frac{s}{\sqrt{n}},\overline{x}+t_{\frac{\alpha}{2}}\cdot\frac{s}{\sqrt{n}}]$	not taught
large	$[\overline{x}-z_{\frac{\alpha}{2}}\cdot\frac{s}{\sqrt{n}},\overline{x}+z_{\frac{\alpha}{2}}\cdot\frac{s}{\sqrt{n}}]$	$[\overline{x}-z_{\frac{\alpha}{2}}\cdot\frac{s}{\sqrt{n}},\overline{x}+z_{\frac{\alpha}{2}}\cdot\frac{s}{\sqrt{n}}]$

Lec 7 假设检验

> Hypothesis Testing

In a hypothesis testing problem, Θ takes m values, $\theta_1,\cdots,\theta_m$. Goal: to select "the optimal" hypothesis θ^*.

7.1 二元假设检验

> Binary Hypothesis Testing Problem

In a special case that Θ only takes 2 values, e.g, 0 or 1, the problem is called binary hypothesis testing.

> 除了 $\Theta=0$ 或 1，也可以是其他只有两种可能结果的情形，例如是或不是.

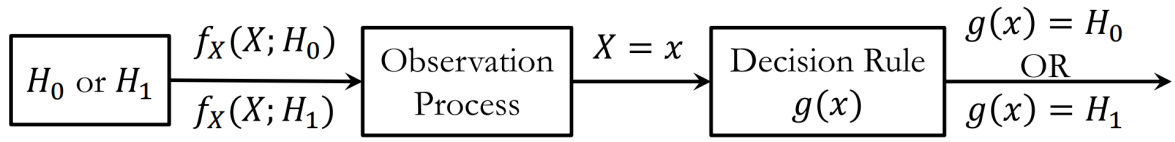
Denote the hypothesis $\Theta=0$ by H_0, called the **null hypothesis**（零假设）. Denote the hypothesis $\Theta=1$ by H_1, called the **alternative hypothesis**（备择假设）.

> 零假设与备择假设是人为设定的. H_0 is considered as a default model，即默认 H_0 成立，再根据观测数据决定是否拒绝 H_0. 在二元假设检验中，拒绝 H_0 意味着接受 H_1，不拒绝 H_0 意味着接受 H_0. 但不要认为这是理所当然，在复合假设中，不拒绝 $H_0\neq$ 接受 H_0.

Observe n independent samples $X_1,\ldots,X_n$ with the same PMF/PDF, which depends on the hypothesis.

Denote by $f_X(x;H_i)$ / $P_X(x;H_i)$ the PDF/PMF defined by the hypothesis H_i.

> 类似似然. 如果 H_i 是参数取值，如 H_0: $\Theta=0$，则 $f_X(x;H_0)=f_{X|\Theta}(x|0)$.



Classical inference framework for binary hypothesis testing

7.1.1 决策规则

> Binary Decision Rule

A binary decision rule can be represented by two disjoint regions of all the possible observations

R：拒绝域（rejection region）. Hypothesis H_0 is rejected when the observed data fall in the rejection region

R^c：接受域（acceptance region）. Complement of R. Hypothesis H_0 is accepted when the observed data fall in this region.

Designing a binary decision rule is equivalent to choosing the rejection region R.

7.1.2 两类错误

> Two Types of Errors

For a particular choice of the rejection region R

① 第一类错误

> Type I Error (false rejection / false positive，假阳性)

错误地 reject hypothesis H_0, even though H_0 is true.

The probability of Type I error:

$$\alpha(R) = P(X \in R; H_0)$$

> 注意是条件概率不是联合概率. H_0 成立条件下拒绝 H_0 的概率，是可能取 1 的.
>
> 这里的「错误拒绝概率」，不是指如果我拒绝（我认为 H_0 不成立），这个拒绝行为的错误概率（H_0 成立概率）；而是指如果本不该拒绝（H_0 成立），却被我拒绝的概率（我认为 H_0 不成立的概率）.

正确接受：$1 - \alpha(R) = P(X \notin R; H_0)$.

② 第二类错误

> Type II Error (false acceptance / false negative，假阴性)

错误地 accept hypothesis H_0, even though H_0 is false.

The probability of Type II error:

$$\beta(R) = P(X \notin R; H_1)$$

正确拒绝：$1 - \beta(R) = P(X \in R; H_1)$.

注意，错误拒绝和正确拒绝概率之和不为 1. 拒绝 H_0 确实要么正确要么错误，但那是这么写的：

$$P(H_1|X \in R) + P(H_0|X \in R) = 1.$$

但我们对错误拒绝 / 正确拒绝的定义不是这样，而是基于不同假设. 在这种定义下使用全概率公式，应该是在某个假设条件下，要么拒绝要么不拒绝，即

$$P(X \in R; H_0) + P(X \notin R; H_0) = 1$$
$$P(X \in R; H_1) + P(X \notin R; H_1) = 1$$

7.1.3 似然比

> Likelihood Ratio

Suppose $X_1, \ldots, X_n$ are independent with same PDF/PMF.

定义 Likelihood ratio: $L(x_1, \ldots, x_n) = \frac{f_X(x_1,\ldots,x_n;H_1)}{f_X(x_1,\ldots,x_n;H_0)}$.

A general decision rule: H_1 is true if $L(x_1, \ldots, x_n) > \xi$, where $\xi > 0$ is the critical value（决断值/临界比）. Otherwise, H_0 is true.

> 似然比大于决断值时，拒绝 H_0，接受 H_1.

决断值可基于 MAP 或 MLE 推导.

① MAP-based: refuse H_0 when

$$\frac{f_{\Theta|X}(1|x)}{f_{\Theta|X}(0|x)} = \frac{f_{X|\Theta}(x|1)P(\Theta = 1)}{f_{X|\Theta}(x|0)P(\Theta = 0)} > 1 \Rightarrow \xi = \frac{P(\Theta = 0)}{P(\Theta = 1)}.$$

② MLE-based: refuse H_0 when

$$\frac{f_X(x; H_1)}{f_X(x; H_0)} > 1 \Rightarrow \xi = 1.$$

注意，ξ 基于统计原则和错误容忍度人为设定，上面只是展示了两种 special cases.

> 在 LRT 中，ξ 则通过给定显著性水平 α 来设定.

例：Given a six-sided die, there are two hypotheses: fair or loaded, with the following PMFs, respectively

H_0 (fair die)

$$P_X(x; H_0) = \frac{1}{6}, \quad \text{for } x = 1, \ldots, 6.$$

H_1 (loaded die)

$$P_X(x;H_1)=\begin{cases}\frac{1}{4}, & \text{if } x=1,2\\ \frac{1}{8}, & \text{if } x=3,4,5,6\end{cases}$$

Given a single roll x of the die, the likelihood ratio is

$$L(x)=\begin{cases}\frac{\frac{1}{4}}{\frac{1}{6}}=\frac{3}{2}, & \text{if } x=1,2\\ \frac{\frac{1}{8}}{\frac{1}{6}}=\frac{3}{4}, & \text{if } x=3,4,5,6\end{cases}$$

There are 3 possibilities to consider for the critical value ξ

$\xi<\frac{3}{4}$: reject H_0 for all x

$\frac{3}{4}\le\xi<\frac{3}{2}$: reject H_0 if $x=1,2$; accept H_0 if $x=3,4,5,6$

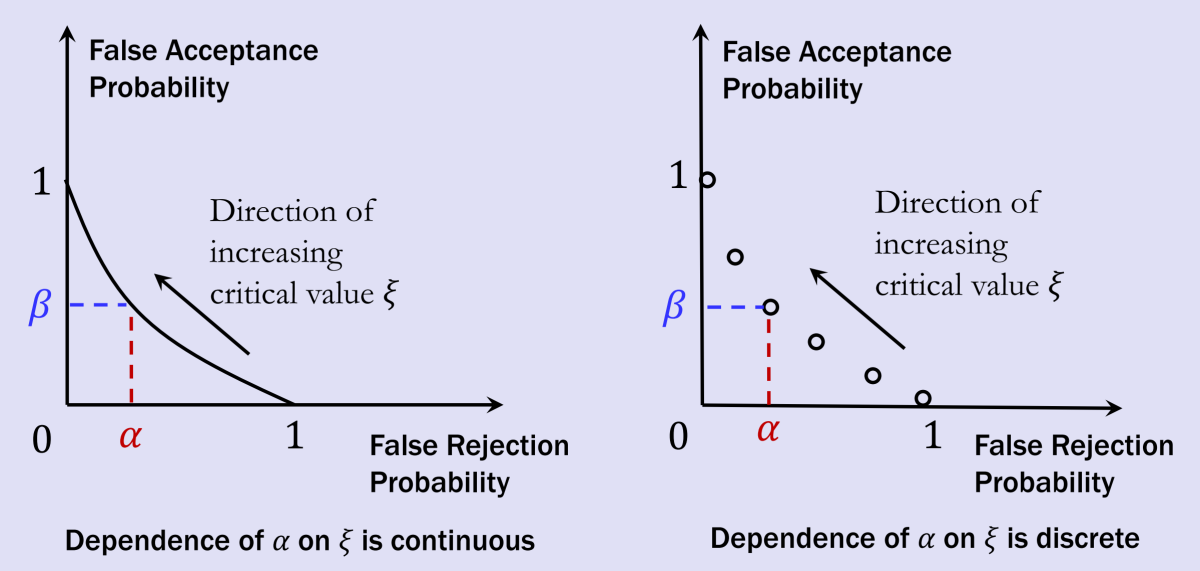
$\xi\ge\frac{3}{2}$: accept H_0 for all x

The probability of false rejection $P(\text{Reject } H_0;H_0)$:

$$\alpha(\xi)=\begin{cases}1, & \text{if } \xi<\frac{3}{4}\\ P(X=1 \text{ or } 2;H_0)=\frac{1}{3}, & \text{if } \frac{3}{4}\le\xi<\frac{3}{2}\\ 0, & \text{if } \xi\ge\frac{3}{2}\end{cases}$$

The probability of false acceptance $P(\text{Accept } H_0;H_1)$:

$$\beta(\xi)=\begin{cases}0, & \text{if } \xi<\frac{3}{4}\\ P(X=3,4,5 \text{ or } 6;H_1)=\frac{1}{2}, & \text{if } \frac{3}{4}\le\xi<\frac{3}{2}\\ 1, & \text{if } \xi\ge\frac{3}{2}\end{cases}$$



As ξ increases, the rejection region becomes smaller. Thus, the false rejection probability $\alpha(R)$ decreases while the false acceptance probability $\beta(R)$ increases

问题来了：如何选择合适的 ξ？ How to choose a trade-off ξ?

7.1.4 似然比检验

> Likelihood Ratio Test (LRT), a popular approach to choosing ξ

定义 Likelihood ratio: $L(x_1,\dots,x_n)=\frac{f_X(x_1,\dots,x_n;H_1)}{f_X(x_1,\dots,x_n;H_0)}$.

似然比检验

> 注意，默认定义好似然比，即已知零假设、备选假设以及对应的似然.

Step 1: Start with a target value α for the false rejection probability.

> α 称为显著性水平（significance level）. Typical choices for α are $0.01, 0.05, 0.10$.

Step 2: Choose a value ξ such that the false rejection probability is equal to α

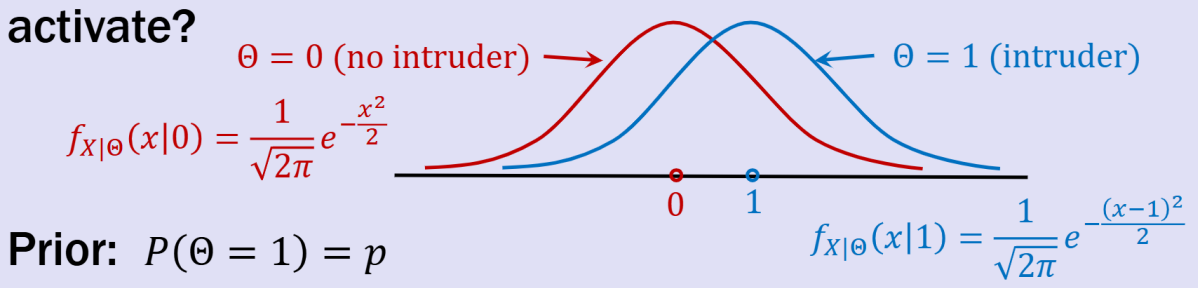
$$P(L(X)>\xi;H_0)=\alpha$$

> 这里通过指定的 false rejection probability 确定拒绝域.

Step 3: Once the $X=x$ is observed, reject H_0 if $L(x)>\xi$.

例：A car-jack detector X outputs $N(0,1)$ if there is no intruder and $N(1,1)$ if there is. When should alarm activate? How to choose ξ?

> 见 `3.3.1 二元假设检验 ① MAP 决策准则 Example 3`. 注意似然比检验相比直接用 MAP 决策准则有两个优点，一个是不需要先验知识（car-jack 先验发生率），一个是可以人为调整 α 满足自己的需要.



> 这是 `3.3.1 二元假设检验 ① MAP 决策准则 Example 3` 的图，本例中没有先验知识 $P(\Theta=1)=p$.

H_0: no intruder $f_X(x;H_0)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$

H_1: intruder $f_X(x;H_1)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-1)^2}{2}}$

似然比 $L(x)=\frac{f_X(x;H_1)}{f_X(x;H_0)}=e^{\frac{x^2-(x-1)^2}{2}}=e^{\frac{2x-1}{2}}$

For a given critical value ξ, we reject H_0 if $L(x)=e^{\frac{2x-1}{2}}>\xi$.

That is when $x>\ln\xi+\frac{1}{2}=\gamma$ we reject $H_0\Rightarrow R=\{X|X>\gamma\}$.

Suppose we set $\alpha=0.025$, where α is the false rejection probability.

$$\alpha=P(X>\gamma;H_0)=P(Z>\gamma)\Rightarrow\gamma=z_\alpha\approx 1.96\Rightarrow\xi\approx e^{1.46}.$$

拒绝域 $R=\{X|X>1.96\}$.

（2025.3.31）通解

> 没什么用，算着玩.

H_0: no intruder $f_X(x;H_0)=\frac{1}{\sqrt{2\pi\sigma_0^2}}e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}}$

H_1: intruder $f_X(x;H_1)=\frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}$

似然比 $L(x)=\frac{f_X(x;H_1)}{f_X(x;H_0)}=\frac{\sigma_0}{\sigma_1}e^{\frac{(\frac{x-\mu_0}{\sigma_0})^2-(\frac{x-\mu_1}{\sigma_1})^2}{2}}$

当 $\sigma_0=\sigma_1=\sigma$ 时，有

$$L(x)=\frac{\sigma_0}{\sigma_1}e^{\frac{(\frac{x-\mu_0}{\sigma_0})^2-(\frac{x-\mu_1}{\sigma_1})^2}{2}}=e^{\frac{2(\mu_1-\mu_0)x+\mu_0^2-\mu_1^2}{2\sigma^2}}$$

For a given critical value ξ, we reject H_0 if $L(x)=e^{\frac{2(\mu_1-\mu_0)x+\mu_0^2-\mu_1^2}{2\sigma^2}}>\xi$.

① $\mu_1>\mu_0$

When $x>\frac{\sigma^2}{\mu_1-\mu_0}\left[\ln\xi+\frac{\mu_1^2-\mu_0^2}{2\sigma^2}\right]=\gamma$ we reject $H_0\Rightarrow R=\{X|X>\gamma\}$.

② $\mu_1<\mu_0$

When $x<\frac{\sigma^2}{\mu_1-\mu_0}\left[\ln\xi+\frac{\mu_1^2-\mu_0^2}{2\sigma^2}\right]=\gamma$ we reject $H_0\Rightarrow R=\{X|X<\gamma\}$.

Set α as the false rejection probability.

$\gamma=z_\alpha$ if $\mu_1>\mu_0$ or

$\gamma=-z_\alpha$ if $\mu_1<\mu_0$.

$|\gamma|=z_\alpha\Rightarrow\xi=e^{\frac{2z_\alpha|\mu_1-\mu_0|+\mu_0^2-\mu_1^2}{2\sigma^2}}$.

7.1.5 内曼-皮尔逊引理

> Neyman-Pearson Lemma

Neyman-Pearson Lemma 指出，在给定 α 的前提下，LRT 提供最小的 β，即在所有显著性水平相同的检验中，LRT 最优.

> LRT 是在控制第一类错误的前提下最小化第二类错误的最优方法.

Consider a particular choice of ξ in LRT, which results in error probabilities

$$P(L(X)>\xi;H_0)=\alpha\qquad P(L(X)\le\xi;H_1)=\beta$$

Suppose that some other test, with rejection region R', achieves a smaller or equal false rejection probability

$$P(X\in R';H_0)\le\alpha$$

Then,

$$P(X \notin R'; H_1) \geq \beta$$

with strict inequality $P(X \notin R'; H_1) > \beta$ when $P(X \in R'; H_0) < \alpha$.

> 证明见 附录 4. 内曼-皮尔逊引理证明 .

Neyman-Pearson Lemma tells us that there is not any test such that $P(X \in R'; H_0) = \alpha$ and at the same $P(X \notin R'; H_1) < \beta$. In other words, LRT offers the smallest β when α is fixed.

Example

> to verify the Neyman-Pearson Lemma, not a proof

Given two independent samples X_1 and X_2, under H_0 they are from $N(0, 1)$; under H_1 they are from $N(2, 1)$. Suppose the false rejection probability is fixed to $\alpha = 0.05$.

Goal: to compare the false acceptance probability β obtained by LRT and that obtained by another test

Likelihood Ratio:

$$\begin{aligned} L(x) &= \frac{f_X(x_1, x_2; H_1)}{f_X(x_1, x_2; H_0)} \\ &= \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_1-2)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_2-2)^2}{2}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_2^2}{2}}} \\ &= e^{2(x_1+x_2)-4} \end{aligned}$$

Compare $L(x)$ to a critical value ξ is equivalent to comparing $x_1 + x_2$ to $\gamma = \frac{4+\ln \xi}{2}$.

Based on LRT, we aim to find γ such that the false rejection probability is equal to 0.05.

$$\alpha = P(L(x) > \xi; H_0) = P(X_1 + X_2 > \gamma; H_0) = 0.05$$

Under H_0, $X_1, X_2 \sim N(0, 1)$. We have

$$X_1 + X_2 \sim N(0, 2) \qquad Z = \frac{X_1 + X_2}{\sqrt{2}} \sim N(0, 1)$$

$$\alpha = P(X_1 + X_2 > \gamma; H_0) = P(\frac{X_1 + X_2}{\sqrt{2}} > \frac{\gamma}{\sqrt{2}}; H_0) = 0.05$$

Based on the CDF table of standard normal distribution,

$$\frac{\gamma}{\sqrt{2}} = z_{0.05} \approx 1.645 \Rightarrow \gamma \approx 2.33$$

The false acceptance probability obtained by LRT is

$$\beta = P(X_1 + X_2 \leq \gamma; H_1) \approx P(X_1 + X_2 \leq 2.33; H_1)$$

Under H_1, $X_1, X_2 \sim N(2, 1)$. We have

$$X_1 + X_2 \sim N(4, 2) \qquad Z = \frac{X_1 + X_2 - 4}{\sqrt{2}} \sim N(0, 1)$$

$$\beta \approx P(X_1 + X_2 \leq 2.33; H_1) = P(\frac{X_1 + X_2 - 4}{\sqrt{2}} \leq \frac{2.33 - 4}{\sqrt{2}}; H_1)$$

Based on the CDF table of standard normal distribution,

$$\beta \approx P(Z \leq -1.18; H_1) = P(Z \geq 1.18; H_1) \approx 0.12.$$

Let's consider another test to find a rejection region by fixing the false rejection probability $\alpha = 0.05$.

For example, we consider a rejection region of the form

$$R' = \{(X_1, X_2) | \max\{X_1, X_2\} > \lambda\}$$

$$\begin{aligned} \alpha = 0.05 &= P(\max\{X_1, X_2\} > \lambda; H_0) \\ &= 1 - P(\max\{X_1, X_2\} \leq \lambda; H_0) \\ &= 1 - P(X_1 \leq \lambda; H_0)P(X_2 \leq \lambda; H_0) \\ &= 1 - P^2(Z \leq \lambda; H_0) \end{aligned}$$

> Under H_0, $X_1, X_2 \sim N(0, 1)$.

Based on the CDF table of standard normal distribution

$$P(Z \leq \lambda; H_0) = \sqrt{1 - 0.05} \approx 0.975 \Rightarrow \lambda \approx 1.96.$$

Based on the rejection region

$$R' = \{(X_1, X_2) | \max\{X_1, X_2\} > 1.96\}$$

The false acceptance probability can be computed as

$$\begin{aligned} \beta(R') &= P(\max\{X_1, X_2\} \leq 1.96; H_1) \\ &= P(X_1 \leq 1.96; H_1)P(X_2 \leq 1.96; H_1) \end{aligned}$$

Under H_1, $X_1, X_2 \sim N(2, 1)$, thus $X_1 - 2, X_2 - 2 \sim N(0, 1)$

$$\begin{aligned} \beta(R') &= P(X_1 - 2 \leq -0.04; H_1)P(X_2 - 2 \leq -0.04; H_1) \\ &= P^2(Z \leq -0.04; H_1) \\ &= P^2(Z \geq 0.04; H_1) \\ &\approx 0.484^2 > \beta(R) \approx 0.12. \end{aligned}$$

7.2 复合假设

> Composite Hypothesis

In real-world settings, hypothesis testing problems do not always involve two well-specified parameters ($\Theta = 0$ vs. $\Theta = 1$) as alternatives.

示例：记 H_0：Null hypothesis (default)，H_1：Alternative hypothesis (complement of H_0).

- We want to claim that the average monthly income of HK residents is more than or equal to 20K HKD

> $H_0 : \mu \geq 20000$ vs. $H_1 : \mu < 20000$

- We want to assess whether the average monthly expense per family in HK is 30K HKD

> $H_0 : \mu = 30000$ vs. $H_1 : \mu \neq 30000$.

- We want to determine whether the number of cars crossing a certain intersection follows a Poisson distribution or a geometric distribution

> $H_0: X \sim Poisson(2)$ vs. $H_1: X \sim Geometric(0.5)$.

7.2.1 简单假设和复合假设

> Simple vs. Composite hypotheses

① 简单假设

> Simple hypotheses

Simple hypotheses: hypotheses where the distribution or the parameter is completely specified 参数值完全确定的假设.

如上述第二个例子中的 $H_0: \mu = 30000$，以及第三个例子的 $H_0: X \sim Poisson(2)$ 和 $H_1: X \sim Geometric(0.5)$.

② 复合假设

> Composite Hypothesis

Composite hypotheses: hypotheses where the distribution or the parameter is NOT completely specified

如上述第一个例子中的 $H_0: \mu \geq 20000$ 和 $H_1: \mu < 20000$，以及第二个例子的 $H_1: \mu \neq 30000$.

注意，从第二个例子可以看出，零假设和备择假设不要求是相同类型，即它们可以一个是简单假设，另一个是复合假设.

7.2.2 一般的假设检验

> General statistical test of hypothesis

Step 1：Specify two hypotheses 提出两个假设.

> H_0：零假设. H_1：备择假设.

Step 2：Choose a test statistic based on the random samples for the parameter in hypotheses 选择一个检验统计量.

> 这个统计量是基于样本数据计算出来的，用于判断是否拒绝 H_0.
>
> 例如，用样本均值 $\overline{X}$ 作为对总体均值 μ 的检验统计量.

Step 3：Assume H_0 is true, and look for evidence from observations to support H_1 假设 H_0 为真，然后寻找支持 H_1 的证据.

> 这一步类似于反证法（Proof by Contradiction）.
>
> 也就是说，我们先假定 H_0 是对的，然后看看数据是否强烈反对这一假设.

Step 4：Make a conclusion 作出结论.

- Reject H_0 if there is strong evidence from the test that indicates the assumption「H_0 is true」does not hold.

> 拒绝 H_0：如果数据提供了充分的证据表明 H_0 不成立.

- Not reject H_0 if there is NOT strong statistical evidence from observations to refuse the assumption.

> 不拒绝 H_0：如果数据没有足够证据反对 H_0. 注意不是接受，而是暂时无法拒绝. 在假设检验中，我们只是在检验数据是否提供了「反对 H_0」的证据，而不是在寻找支持它的证据.
>
> 这有点像法庭上的「无罪推定原则」（Presumption of Innocence）：无法定罪（Not Guilty）$\neq$ 无罪（Innocent），只是证据不足，不能定罪. 在无罪推定原则中，「被告清白 / 被告无罪（Innocent）」相当于统计学中的零假设. 被告不需要提供支持自己无罪（Innocent）的证据. 如果想要「定罪」（拒绝 H_0），检方必须提供强有力的证据反对「被告无罪」的假设. 如果证据不足，哪怕有怀疑，也必须「宣布无罪（Not Guilty Verdict / Acquittal）」，即不拒绝 H_0. 但宣布无罪（Not Guilty）$\neq$ 被告无罪（Innocent），注意中文语境里两个词都是无罪，容易造成混淆，建议 Innocent 翻译为「清白」以作区分.
>
> 无法定罪 / 宣布无罪（Not Guilty）：不拒绝 H_0
>
> 定罪 / 有罪：拒绝 H_0
>
> 无罪 / 清白（Innocent）：接受 H_0（统计上不推荐这么说）
>
> 注意，理清楚「不拒绝 H_0 $\neq$ 接受 H_0」这一概念，是掌握统计推断的逻辑本质，影响你如何正确解读数据、做决策，避免误判. 如果你说「接受 H_0」，就像在宣称你确信它是真的. 但实际上，你只是没有足够证据反对它，两者在逻辑上完全不同. 科学方法强调的是「证伪（falsifiability）」，即寻找证据推翻某个假设，而不是证明它. 如果找不到证据推翻假设，我们只是说目前没有能力否定它，并不承认它绝对正确（未来可能有新的证据推翻它）. 科学不是在寻找真理，而是在排除错误.

7.2.3 总体均值的复合假设

> Composite hypotheses on population mean μ

① 双尾检验

> Two-sided (two-tailed) test

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

② 单尾检验

> One-sided (one-tailed) test

左尾检验 Left-sided

$H_0 : \mu \geq \mu_0$

$H_1 : \mu < \mu_0$

右尾检验 Right-sided

$H_0 : \mu \leq \mu_0$

$H_1 : \mu > \mu_0$

7.2.4 临界值法

> Critical Value Approach

- 给定显著性水平 α

> Define a significant level α (or confidence of $1-\alpha$), the largest false rejection probability we can accept

- 计算拒绝域的临界值

> Find the rejection region of H_0, i.e., to find the critical values of the region s.t. $\alpha = P(H_0 \text{ is rejected}; H_0)$.

- 若检验统计量落入拒绝域，则拒绝 H_0

> If $\overline{x}$ (test statistic) is in the rejection region (i.e., strong evident to support H_1), then we consider that there is strong statistical evidence to reject H_0, otherwise, there is NOT strong statistical evidence to reject H_0.

注意，似然比检验也包含临界值的思想，但本课程的临界值法特指在复合假设检验中使用的单尾/双尾检验.
Critical value approach is applied to composite hypothesis test, while LRT is applied to simple hypothesis test.

① 双尾检验

> Two-sided test of μ

大样本，When $n \geq 30$:

Suppose $X_1, \ldots, X_n$ are independent samples with the same PDF / PMF (μ, σ^2, etc.)

Assumption: $H_0 : \mu = \mu_0$ is true

Test statistic: sample mean $\overline{X}$

Rejection region: $|\overline{X} - \mu_0| \geq \xi$, where $\xi > 0$

> Farther away from μ the sample mean $\overline{X}$ is, the more the evidence points towards $H_1 : \mu \neq \mu_0$

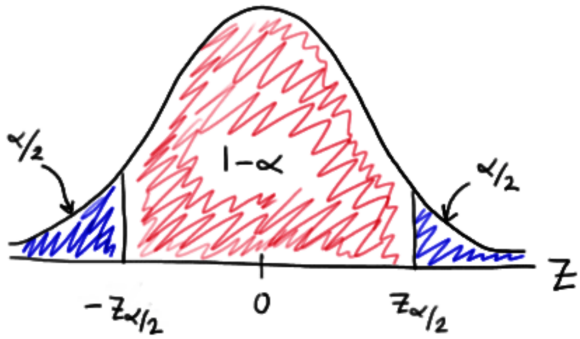
ξ is determined by solving $\alpha = P(H_0 \text{ is rejected}; H_0)$

$$\alpha = P(|\overline{X} - \mu_0| \geq \xi; \mu = \mu_0)$$

As n is large, based on CLT

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \underset{\text{assuming } H_0 \text{ is true}}{\overset{\mu=\mu_0}{\Longrightarrow}} \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$\alpha = P(|\overline{X} - \mu_0| \geq \xi) = P\left(|Z| \geq \frac{\xi}{\sigma/\sqrt{n}}\right)$$



此时有 $\frac{\xi}{\sigma/\sqrt{n}}=z_{\frac{\alpha}{2}}$. 根据 $1-\frac{\alpha}{2}$ 查表可得 $z_{\frac{\alpha}{2}}$. 但是，σ 是否已知仍需讨论.

When σ^2 is known

Given a specific estimate $\overline{x}$, if $\frac{|\overline{x}-\mu_0|}{\sigma/\sqrt{n}}\geq z_{\frac{\alpha}{2}}$, then reject H_0; otherwise, do not reject H_0.

When σ^2 is unknown, $\sigma\approx s'$

> 这里 s' 来自无偏估计 $s'=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$.

Given a specific estimate $\overline{x}$, if $\frac{|\overline{x}-\mu_0|}{s'/\sqrt{n}}\geq z_{\frac{\alpha}{2}}$, then reject H_0, otherwise, do not reject H_0.

小样本，When $n<30$:

As n is small, CLT is not applicable, but if $X_1,\dots,X_n$ are $N(\mu,\sigma^2)$, then we still have $\overline{X}\sim N(\mu,\frac{\sigma^2}{n})$

When σ^2 is known, $Z=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\sim N(0,1)$

Given a specific test statistic estimate $\overline{x}$, if $\frac{|\overline{x}-\mu_0|}{\sigma/\sqrt{n}}\geq z_{\frac{\alpha}{2}}$, then reject H_0; otherwise, do not reject H_0.

When σ^2 is unknown, $T=\frac{\overline{X}-\mu_0}{S'/\sqrt{n}}\sim t(n-1)$

Given a specific test statistic estimate $\overline{x}$, if $\frac{|\overline{x}-\mu_0|}{s'/\sqrt{n}}\geq t_{\frac{\alpha}{2}}$, then reject H_0; otherwise, do not reject H_0.

例：The average HK temperature in Feb is 18℃. Has this year been unusual? Assume temperature in Feb follows $N(\mu,\sigma^2)$ and $\sigma=3$℃. Suppose $\alpha=0.05$.

day (Feb)	1	6	11	16	21	26
temp (℃)	15	15	19	18	8	17

解：$H_0:\mu=18$ vs. $H_1:\mu\neq 18$.

方差已知，小样本，正态，得

$$Z=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\sim N(0,1)$$

Therefore, we make a conclusion based on $\frac{|\overline{x}-\mu_0|}{\sigma/\sqrt{n}}\geq z_{\frac{\alpha}{2}}$.

$$\frac{|15.33-18|}{3/\sqrt{6}}\approx 2.18>z_{0.025}=1.96$$

Reject H_0.

What if σ is unknown?

解：$H_0:\mu=18$ vs. $H_1:\mu\neq 18$.

方差未知，小样本，正态，得

$$T=\frac{\overline{X}-\mu_0}{S'/\sqrt{n}}\sim t(n-1)$$

Therefore, we make a conclusion based on $\frac{|\overline{x}-\mu_0|}{s'/\sqrt{n}}\geq t_{\frac{\alpha}{2}}$.

$$s'^2=\frac{(15-15.33)^2+\cdots+(17-15.33)^2}{6-1}\approx 15.47\Rightarrow s'\approx 3.93.$$

$$\frac{|15.33-18|}{3.93/\sqrt{6}}\approx 1.66<t_{0.025}\approx 2.57$$

Not reject H_0.

> 这说明其他条件相同情况下，方差已知 / 未知会改变判定结果.

② 单尾检验

> One-sided test of μ

左尾检验 Left-sided

$H_0:\mu\geq\mu_0$

$H_1:\mu<\mu_0$

右尾检验 Right-sided

$H_0:\mu\leq\mu_0$

$H_1:\mu>\mu_0$

To ensure the type I error does not exceed α, assume H_0 is true, and then set the false rejection probability $=\alpha$ to find the rejection region for H_0

$$\alpha=P(\text{Reject } H_0;H_0:\mu=\mu_0)\geq P(\text{Reject } H_0;H_0:\mu\leq\mu_0)$$

> Reason: $H_0:\mu=\mu_0$ is a more specific and conservative assumption than $H_0:\mu\leq\mu_0$.
> Therefore, $H_0:\mu=\mu_0$ is more likely to be falsely rejected than $H_0:\mu\leq\mu_0$.
>
> 待完成.

We transform the above hypotheses as

左尾检验 Left-sided

$H_0 : \mu = \mu_0$

$H_1 : \mu < \mu_0$

右尾检验 Right-sided

$H_0 : \mu = \mu_0$

$H_1 : \mu > \mu_0$

右尾检验

> Right-sided test of μ

$H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$

大样本，When $n \geq 30$:

$$\alpha = P(\overline{X} - \mu_0 \geq \xi) = P(Z \geq z_\alpha)$$

When σ^2 is known, given a specific test statistic estimate $\overline{x}$, if $\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \geq z_\alpha$, then reject H_0, otherwise, do not reject H_0

When σ^2 is unknown, $\sigma \approx s'$. Given a specific statistic estimate $\overline{x}$, if $\frac{\overline{x}-\mu_0}{s'/\sqrt{n}} \geq z_\alpha$, then reject H_0; otherwise, do not reject H_0

小样本，When $n < 30$:

As n is small, CLT is not applicable, but if $X_1, \ldots, X_n$ are $N(\mu, \sigma^2)$, then we still have $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

When σ^2 is known, $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

Given a specific test statistic estimate $\overline{x}$, if $\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \geq z_\alpha$, then reject H_0; otherwise, do not reject H_0.

When σ^2 is unknown, $T = \frac{\overline{X}-\mu_0}{S'/\sqrt{n}} \sim t(n-1)$

Given a specific test statistic estimate $\overline{x}$, if $\frac{\overline{x}-\mu_0}{s'/\sqrt{n}} \geq t_\alpha$, then reject H_0; otherwise, do not reject H_0.

左尾检验

$H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$

大样本，When $n \geq 30$:

When σ^2 is known, given a specific test statistic estimate $\overline{x}$, if $\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha$, then reject H_0, otherwise, do not reject H_0

When σ^2 is unknown, $\sigma \approx s'$. Given a specific statistic estimate $\overline{x}$, if $\frac{\overline{x}-\mu_0}{s'/\sqrt{n}} \leq -z_\alpha$, then reject H_0; otherwise, do not reject H_0

小样本，When $n < 30$:

As n is small, CLT is not applicable, but if $X_1, \ldots, X_n$ are $N(\mu, \sigma^2)$, then we still have $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

When σ^2 is known, $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

Given a specific test statistic estimate $\overline{x}$, if $\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha$, then reject H_0; otherwise, do not reject H_0.

When σ^2 is unknown, $T = \frac{\overline{X}-\mu_0}{S'/\sqrt{n}} \sim t(n-1)$

Given a specific test statistic estimate $\overline{x}$, if $\frac{\overline{x}-\mu_0}{s'/\sqrt{n}} \leq -t_\alpha$, then reject H_0; otherwise, do not reject H_0.

例：The average HK temperature in Feb is 18℃. Has this year been colder? Assume temperature in Feb follows $N(\mu, \sigma^2)$ and $\sigma = 3$℃. Suppose $\alpha = 0.05$.

day (Feb)	1	6	11	16	21	26
temp (℃)	15	15	19	18	8	17

$H_0 : \mu = 18$ vs. $H_1 : \mu < 18$.

$Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

Therefore, we make a conclusion based on $\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha$.

$$\frac{15.33-18}{3/\sqrt{6}} \approx -2.18 < -z_{0.05} \approx -1.645$$

Reject H_0.

What if σ is unknown?

解：$H_0 : \mu = 18$ vs. $H_1 : \mu < 18$.

方差未知，小样本，正态，得

$$T = \frac{\overline{X}-\mu_0}{S'/\sqrt{n}} \sim t(n-1)$$

Therefore, we make a conclusion based on $\frac{\overline{x}-\mu_0}{s'/\sqrt{n}} \leq -t_\alpha$.

$$s'^2 = \frac{(15-15.33)^2 + \cdots + (17-15.33)^2}{6-1} \approx 15.47 \Rightarrow s' \approx 3.93.$$

$$\frac{15.33-18}{3.93/\sqrt{6}} \approx -1.66 > -t_{0.05} \approx -2.015$$

Not reject H_0.

7.2.5 p 值法

> The p-value Approach

The p-value is the smallest probability of the type I error for which the null hypothesis would be rejected given a specific test statistic

核心思想：假设观测到 $\overline{x}$ 会拒绝 H_0，保证最小的错误拒绝率在可接受范围内.

> 就是说我准备拒绝 $\overline{x}$，如果错误拒绝率在可接受范围内，则执行拒绝；如果错误拒绝率在可接受范围外，则不拒绝.

Assume H_0 is true

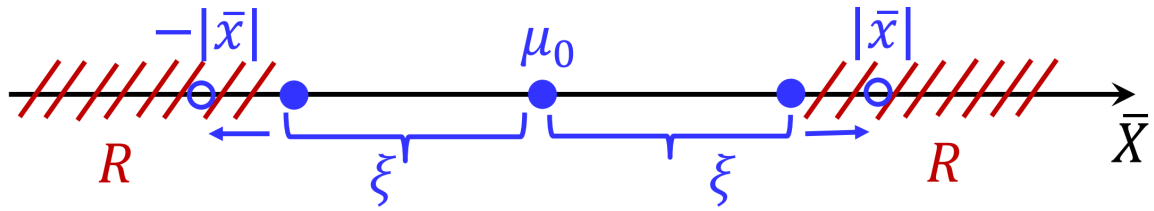
1. Estimate the sample mean $\overline{x}$ from the observed data
2. Compute the p-value
3. Compare the p-value to a predefined significant level α
    - If p-value $\leq \alpha$, then reject H_0;
    - Otherwise, NOT reject H_0.

① 双尾检验

$H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$

$$\alpha = P(\overline{X} \in R; \mu = \mu_0) \qquad R = \{\overline{X} \mid |\overline{X} - \mu_0| \geq \xi\}$$

最小错误拒绝率由最小的拒绝域确定：ξ is larger, R is smaller, α is smaller. When ξ reaches its limit, R is the smallest, thus α achieves the smallest given $\overline{x}$



The smallest rejection region given a specific μ_0 is

$$R = \{\overline{X} \mid |\overline{X} - \mu_0| \geq |\overline{x} - \mu_0|\}$$

Therefore, the p-value is computed as

$$P(|\overline{X} - \mu_0| \geq |\overline{x} - \mu_0|; \mu = \mu_0)$$

Given $n \geq 30$,

When σ^2 is known,

$$P(|\overline{X} - \mu_0| \geq |\overline{x} - \mu_0|; \mu = \mu_0) = P(Z \geq |\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}|) + P(Z \leq -|\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}|)$$

When σ^2 is unknown,

$$P(|\overline{X} - \mu_0| \geq |\overline{x} - \mu_0|; \mu = \mu_0) \approx P(Z \geq |\frac{\overline{x} - \mu_0}{s'/\sqrt{n}}|) + P(Z \leq -|\frac{\overline{x} - \mu_0}{s'/\sqrt{n}}|)$$

> 待完成.

Given $n < 30$,

As n is small, CLT is not applicable, but if $X_1, \ldots, X_n$ are $N(\mu, \sigma^2)$, then we still have $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

When σ^2 is known, the p-value is computed as

$$P(|\overline{X} - \mu_0| \geq |\overline{x} - \mu_0|; \mu = \mu_0) = P(Z \geq |\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}|) + P(Z \leq -|\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}|)$$

When σ^2 is unknown, the p-value is computed as

$$P(|\overline{X} - \mu_0| \geq |\overline{x} - \mu_0|; \mu = \mu_0) \approx P(T \geq |\frac{\overline{x} - \mu_0}{s'/\sqrt{n}}|) + P(T \leq -|\frac{\overline{x} - \mu_0}{s'/\sqrt{n}}|)$$

② 单尾检验

右尾检验

$H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$

The smallest region to reject $H_0 : \mu = \mu_0$ when H_0 is true based on the observed test statistic $\overline{x}$:

$$R = \{\overline{X} \mid \overline{X} - \mu_0 \geq \overline{x} - \mu_0\}$$

Given $n \geq 30$,

When σ^2 is known, the p-value is computed as

$$P(\overline{X} - \mu_0 \geq \overline{x} - \mu_0; \mu = \mu_0) = P(Z \geq \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}})$$

When σ^2 is unknown, the p-value is computed as

$$P(\overline{X} - \mu_0 \geq \overline{x} - \mu_0; \mu = \mu_0) \approx P(Z \geq \frac{\overline{x} - \mu_0}{s'/\sqrt{n}})$$

Given $n < 30$,

As n is small, CLT is not applicable, but if $X_1, \ldots, X_n$ are $N(\mu, \sigma^2)$, then we still have $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

When σ^2 is known, the p-value is computed as

$$P(\overline{X} - \mu_0 \geq \overline{x} - \mu_0; \mu = \mu_0) = P(Z \geq \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}})$$

When σ^2 is unknown, the p-value is computed as

$$P(\overline{X} - \mu_0 \geq \overline{x} - \mu_0; \mu = \mu_0) \approx P(T \geq \frac{\overline{x} - \mu_0}{s'/\sqrt{n}})$$

左尾检验

$H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$

The smallest region to reject $H_0 : \mu = \mu_0$ when H_0 is true based on the observed test statistic $\overline{x}$:

$$R = \{\overline{X} \mid \overline{X} - \mu_0 \leq \overline{x} - \mu_0\}$$

Given $n \geq 30$,

When σ^2 is known, the p-value is computed as

$$P(\overline{X} - \mu_0 \leq \overline{x} - \mu_0; \mu = \mu_0) = P(Z \leq \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}})$$

When σ^2 is unknown, the p-value is computed as

$$P(\overline{X} - \mu_0 \leq \overline{x} - \mu_0; \mu = \mu_0) \approx P(Z \leq \frac{\overline{x} - \mu_0}{s'/\sqrt{n}})$$

Given $n < 30$,

As n is small, CLT is not applicable, but if $X_1, \ldots, X_n$ are $N(\mu, \sigma^2)$, then we still have $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

When σ^2 is known, the p-value is computed as

$$P(\overline{X} - \mu_0 \leq \overline{x} - \mu_0; \mu = \mu_0) = P(Z \leq \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}})$$

When σ^2 is unknown, the p-value is computed as

$$P(\overline{X} - \mu_0 \leq \overline{x} - \mu_0; \mu = \mu_0) \approx P(T \leq \frac{\overline{x} - \mu_0}{s'/\sqrt{n}})$$

例：The average HK temperature in Feb is 18℃. Has this year been unusual? Assume temperature in Feb follows $N(\mu, \sigma^2)$ and $\sigma = 3$℃. Suppose $\alpha = 0.05$. Use the p-value approach

day (Feb)	1	6	11	16	21	26
temp (℃)	15	15	19	18	8	17

解：$H_0 : \mu = 18$ vs. $H_1 : \mu \neq 18$.

方差已知，小样本，正态，得

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Therefore, we make a conclusion based on p-value:

$$\begin{aligned}
\text{p-value} &= P(Z \geq |\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}|) + P(Z \leq -|\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}|) \\
&\approx 2(1 - P(Z < 2.18)) \\
&\approx 2(1 - 0.9854) \\
&= 0.0292 \\
&< \alpha = 0.05
\end{aligned}$$

Reject H_0.

Has this year been colder? What if σ is unknown?

解：$H_0 : \mu = 18$ vs. $H_1 : \mu < 18$.

方差未知，小样本，正态，得

$$T = \frac{\overline{X} - \mu_0}{S'/\sqrt{n}} \sim t(n-1)$$

Therefore, we make a conclusion based on p-value:

$$\begin{aligned}
\text{p-value} &= P(T \leq \frac{\overline{x} - \mu_0}{s'/\sqrt{n}}) \\
&\approx P(T \leq -1.66) \\
&= P(T > 1.66) \\
&\approx 0.0789 \\
&> \alpha = 0.05
\end{aligned}$$

Not reject H_0.

7.2.6 总结

> Critical value approach vs. p-value approach

Given a hypothesis testing problem about population mean μ. Assume H_0 is true. For a significant level α:

The critical value approach

1. Based on α to find a critical value
2. Estimate (normalized) sample mean $\overline{x}$ from observed data
3. Compare $\overline{x}$ to the critical value to make a decision

The p-value approach

1. Estimate (normalized) sample mean $\overline{x}$ from observed data
2. Compute the p-value based on $\overline{x}$
3. Compare the p-value to α to make a decision

7.3 比较两个总体均值

> Comparing two population means

Suppose we want to explore whether male and female college students have different driving behaviors in terms of the mean fastest driving speed

Based on a survey from 18 male students and 20 female students, we find that the mean fastest speeds driven by male and female students are 105kph and 90kph

Can we claim the mean fastest speed driven by male college students is different from the mean fastest speed driven by female college students?

> Two-sided test

OR can we claim the mean fastest speed driven by male college students is faster than that driven by female college students?

> One-sided test

7.3.1 大样本

> A large-sample test for two population means

Suppose $X_1, \dots, X_{n_x}$ are independent with same μ_x and σ_x^2, $Y_1, \dots, Y_{n_y}$ are independent with same μ_y and σ_y^2, $\{X_i\}$'s and $\{Y_i\}$'s are also independent

Suppose $n_x, n_y \geq 30$, the significant level is α

Two-sided test

$H_0 : \mu_x = \mu_y$ vs. $H_1 : \mu_x \neq \mu_y$

One-sided test

$H_0 : \mu_x = \mu_y$ vs. $H_1 : \mu_x > \mu_y$

$H_0 : \mu_x = \mu_y$ vs. $H_1 : \mu_x < \mu_y$

① 双尾检验

$H_0 : \mu_x = \mu_y$ vs. $H_1 : \mu_x \neq \mu_y$

Rewrite the above hypothesis testing problem as

$H_0 : \mu_x - \mu_y = 0$ vs. $H_1 : \mu_x - \mu_y \neq 0$

Based on CLT, we have

$$\overline{X} \sim N(\mu_x, \frac{\sigma_x^2}{n_x}),\ \overline{Y} \sim N(\mu_y, \frac{\sigma_y^2}{n_y}) \Rightarrow \overline{X} - \overline{Y} \sim N(\mu_x - \mu_y, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y})$$

Assume $H_0 : \mu_x - \mu_y = 0$ is true

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sigma_D} = \frac{\overline{X} - \overline{Y}}{\sigma_D} \sim N(0, 1) \qquad \sigma_D = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

The critical value approach:

$$\begin{aligned}\alpha &= P(|(\overline{X}-\overline{Y})-0| \geq \xi) \\ &= P(|\frac{\overline{X}-\overline{Y}}{\sigma_D}| \geq \frac{\xi}{\sigma_D}) \\ &= P(|Z| \geq z_{\frac{\alpha}{2}})\end{aligned}$$

When σ_x^2 and σ_y^2 are known

Given a specific estimate $\overline{x}$ and $\overline{y}$, if $\frac{|\overline{x}-\overline{y}|}{\sigma_D} \geq z_{\frac{\alpha}{2}}$, then reject H_0, otherwise, do not reject H_0.

When σ_x^2 and σ_y^2 are unknown, $\sigma_x^2 \approx s_x'^2$ and $\sigma_y^2 \approx s_y'^2$

Given a specific estimate $\overline{x}$ and $\overline{y}$, if $\frac{|\overline{x}-\overline{y}|}{s_D'} \geq z_{\frac{\alpha}{2}}$, then reject H_0, otherwise, do not reject H_0.

> $s_D' = \sqrt{\frac{s_x'^2}{n_x}+\frac{s_y'^2}{n_y}}$.

The p-value approach:

$$P(|Z| \geq |\frac{\overline{x}-\overline{y}}{\sigma_D}|) = P(Z \geq |\frac{\overline{x}-\overline{y}}{\sigma_D}|) + P(Z \leq -|\frac{\overline{x}-\overline{y}}{\sigma_D}|)$$

> If σ_x^2, σ_y^2 are unknown, use $s_x'^2, s_y'^2$ to compute s_D' instead.

If the p-value $\leq \alpha$, reject H_0, otherwise, NOT

② 单尾检验

Suppose $X_1, \ldots, X_{n_x}$ are independent with PDF$_x$ (μ_x and σ_x^2), $Y_1, \ldots, Y_{n_y}$ are independent with PDF$_y$ (μ_y and σ_y^2), $\{X_i\}$'s and $\{Y_i\}$'s aree also independent. Significant level is α.

以右尾为例.

$H_0 : \mu_x - \mu_y = 0$ vs. $H_1 : \mu_x - \mu_y > 0$

The critical value approach:

Given specific $\overline{x}$ and $\overline{y}$, if $\frac{\overline{x}-\overline{y}}{\sigma_D} \geq z_\alpha$, then reject H_0, otherwise, NOT, where $\sigma_D = \sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}$

> If σ_x^2 and σ_y^2 are unknown, they are replaced by $s_x'^2$ and $s_y'^2$.

The p-value approach:

Given specific $\overline{x}$ and $\overline{y}$,

$$\text{p-value} = P(Z \geq \frac{\overline{x}-\overline{y}}{\sigma_D})$$

Where $Z \sim N(0,1)$.

> If σ_x^2, σ_y^2 are unknown, use $s_x'^2, s_y'^2$ to compute s_D' instead.

If the p-value $\leq \alpha$, reject H_0, otherwise, NOT

7.3.2 小样本

> A small-sample test for two means

Suppose $n_x, n_y < 30$, the significant level is α, $X_1, \dots, X_{n_x}$ are normal; $Y_1, \dots, Y_{n_y}$ are also normal.

① 双尾检验

$H_0 : \mu_x = \mu_y$ vs. $H_1 : \mu_x \neq \mu_y$

Rewrite the above hypothesis testing problem as

$H_0 : \mu_x - \mu_y = 0$ vs. $H_1 : \mu_x - \mu_y \neq 0$

If σ_x^2 and σ_y^2 are known

$$\overline{X} \sim N(\mu_x, \frac{\sigma_x^2}{n_x}),\ \overline{Y} \sim N(\mu_y, \frac{\sigma_y^2}{n_y}) \Rightarrow \overline{X} - \overline{Y} \sim N(\mu_x - \mu_y, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y})$$

Assume $H_0 : \mu_x - \mu_y = 0$ is true

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sigma_D} = \frac{\overline{X} - \overline{Y}}{\sigma_D} \sim N(0, 1) \qquad \sigma_D = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

The critical value approach:

$$\begin{aligned} \alpha &= P(|(\overline{X} - \overline{Y}) - 0| \geq \xi) \\ &= P(|\frac{\overline{X} - \overline{Y}}{\sigma_D}| \geq \frac{\xi}{\sigma_D}) \\ &= P(|Z| \geq z_{\frac{\alpha}{2}}) \end{aligned}$$

When σ_x^2 and σ_y^2 are known

Given a specific estimate $\overline{x}$ and $\overline{y}$, if $\frac{|\overline{x} - \overline{y}|}{\sigma_D} \geq z_{\frac{\alpha}{2}}$, then reject H_0, otherwise, do not reject H_0.

When σ_x^2 and σ_y^2 are unknown

Suppose $X_1, \dots, X_{n_x}$ are independent $N(\mu_x, \sigma_x^2)$ samples, $Y_1, \dots, Y_{n_y}$ are independent $N(\mu_y, \sigma_y^2)$ samples, $\{X_i\}$'s and $\{Y_i\}$'s are also independent, and $\sigma_x^2 = \sigma_y^2 = \sigma^2$.

> When σ_x^2 and σ_y^2 are not the same, refer to Welch's t-test (not covered in this course)

Suppose $n_x, n_y < 30$, and σ^2 is unknown, we have

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{S'_D \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim t(n_x + n_y - 2)$$

其中，

$$S_D'^2 = \frac{(n_x-1)S_x'^2 + (n_y-1)S_y'^2}{n_x+n_y-2}$$

叫做合并样本方差（Pooled sample variance）.

> 注意和大样本、方差未知时的 $s'_D = \sqrt{\frac{s_x'^2}{n_x} + \frac{s_y'^2}{n_y}}$ 区分开.

Assume $H_0: \mu_x - \mu_y = 0$ is true

$$\begin{aligned}\alpha &= P(|(\overline{X}-\overline{Y}) - 0| \geq \xi) \\ &= P(|T| \geq \frac{\xi}{S'_D\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}}) \\ &= P(|T| \geq t_{\frac{\alpha}{2}})\end{aligned}$$

> Degrees of freedom: $n_x + n_y - 2$.

Given specific $\overline{x}$ and $\overline{y}$, if $\frac{|\overline{x}-\overline{y}|}{s'_D\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}} \geq t_{\frac{\alpha}{2}}$, then reject H_0, otherwise, do not reject H_0.

The p-value approach:

When σ_x^2 and σ_y^2 are known

$$P(|Z| \geq |\frac{\overline{x}-\overline{y}}{\sigma_D}|) = P(Z \geq |\frac{\overline{x}-\overline{y}}{\sigma_D}|) + P(Z \leq -|\frac{\overline{x}-\overline{y}}{\sigma_D}|)$$

If the p-value $\leq \alpha$, reject H_0, otherwise, NOT

When σ_x^2 and σ_y^2 are unknown

Suppose $X_1, \ldots, X_{n_x}$ are independent $N(\mu_x, \sigma_x^2)$ samples, $Y_1, \ldots, Y_{n_y}$ are independent $N(\mu_y, \sigma_y^2)$ samples, $\{X_i\}$'s and $\{Y_i\}$'s are also independent, and $\sigma_x^2 = \sigma_y^2 = \sigma^2$.

> When σ_x^2 and σ_y^2 are not the same, refer to Welch's t-test (not covered in this course)

Suppose $n_x, n_y < 30$, and σ^2 is unknown, we have

$$T = \frac{(\overline{X}-\overline{Y}) - (\mu_x - \mu_y)}{S'_D\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}} \sim t(n_x+n_y-2)$$

其中，

$$S_D'^2 = \frac{(n_x-1)S_x'^2 + (n_y-1)S_y'^2}{n_x+n_y-2}$$

叫做合并样本方差（Pooled sample variance）.

> 注意和大样本、方差未知时的 $s'_D = \sqrt{\frac{s_x'^2}{n_x} + \frac{s_y'^2}{n_y}}$ 区分开.

Assume $H_0: \mu_x - \mu_y = 0$ is true

$$
\begin{aligned}
& P\left(|T| \geq \left|\frac{(\overline{x}-\overline{y})-0}{s'_D\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}}\right|\right) \\
= \ & P\left(T \geq \left|\frac{\overline{x}-\overline{y}}{s'_D\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}}\right|\right) + P\left(T \leq -\left|\frac{\overline{x}-\overline{y}}{s'_D\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}}\right|\right)
\end{aligned}
$$

If the p-value $\leq \alpha$, reject H_0, otherwise, NOT

② 单尾检验

Suppose $X_1, \ldots, X_{n_x}$ are independent with PDF$_x$ (μ_x and σ_x^2), $Y_1, \ldots, Y_{n_y}$ are independent with PDF$_y$ (μ_y and σ_y^2), $\{X_i\}$'s and $\{Y_i\}$'s aree also independent. Significant level is α. PDF$_x$ and PDF$_y$ are normal.

以右尾为例.

$H_0 : \mu_x - \mu_y = 0$ vs. $H_1 : \mu_x - \mu_y > 0$

The critical value approach:

Given specific $\overline{x}$ and $\overline{y}$, if $\frac{\overline{x}-\overline{y}}{\sigma_D} \geq z_\alpha$, then reject H_0, otherwise, NOT, where $\sigma_D = \sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}$

If $\sigma_x^2 = \sigma_y^2 = \sigma^2$ are unknown

Given specific $\overline{x}$ and $\overline{y}$, if $\frac{\overline{x}-\overline{y}}{s'_D\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}} \geq t_\alpha$, then reject H_0, otherwise, NOT

> $s'^2_D = \frac{(n_x-1)s'^2_x+(n_y-1)s'^2_y}{n_x+n_y-2}$.

The p-value approach:

Given specific $\overline{x}$ and $\overline{y}$,

$$
\text{p-value} = P(Z \geq \frac{\overline{x}-\overline{y}}{\sigma_D})
$$

Where $Z \sim N(0,1)$.

If the p-value $\leq \alpha$, reject H_0, otherwise, NOT

If $\sigma_x^2 = \sigma_y^2 = \sigma^2$ are unknown

Given specific $\overline{x}$ and $\overline{y}$,

$$
\text{p-value} = P\left(T \geq \frac{\overline{x}-\overline{y}}{s'_D\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}}\right)
$$

where $T \sim t(n_x+n_y-2)$.

If the p-value $\leq \alpha$, reject H_0, otherwise, NOT

例：A thermometer reports readings

Mon	23.5	23.3	21.3	22.1	23.7
Tue	22.8	24.5	23.7		

Has the temperature increased? Suppose the readings following $N(\mu_M, \sigma^2)$ on Mon and $N(\mu_T, \sigma^2)$ on Tue are independent, with $\sigma = 1$℃. Set $\alpha = 0.05$. Use the p-value approach

解：$H_0 : \mu_T - \mu_M = 0$ vs. $H_1 : \mu_T - \mu_M > 0$

Denote by X_i the readings on Mon, and by Y_i the readings on Tue, where $n_x = 5$ and $n_y = 3$, the test statistic is $\overline{Y} - \overline{X}$

The p-value $= P(Z \geq \frac{\overline{y} - \overline{x}}{\sigma_D})$, where $\sigma_D = \sqrt{\frac{\sigma^2}{n_x} + \frac{\sigma^2}{n_y}}$

Plugging $\overline{x} = 22.78$, $\overline{y} \approx 23.67$, $\sigma = 1$, $n_x = 5$, $n_y = 3$, we have

$$\begin{aligned} \text{p-value} &\approx P(Z \geq \frac{23.67 - 22.78}{\sqrt{\frac{1}{5} + \frac{1}{3}}}) \\ &\approx P(Z \geq 1.22) \\ &\approx 0.1112 \\ &> \alpha = 0.05 \end{aligned}$$

Not reject H_0.

What about σ is unknown and use the critical value approach?

$H_0 : \mu_T - \mu_M = 0$ vs. $H_1 : \mu_T - \mu_M > 0$

Denote by X_i the readings on Mon, and by Y_i the readings on Tue, where $n_x = 5$ and $n_y = 3$, the test statistic is $\overline{Y} - \overline{X}$

The critical value approach: $T = \frac{\overline{Y} - \overline{X}}{S'_D \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim t(n_x + n_y - 2)$

Given $\overline{x}$ and $\overline{y}$, if $\frac{\overline{y} - \overline{x}}{s'_D \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \geq t_\alpha$, reject H_0, otherwise, NOT

where $s'^2_D = \frac{(n_x - 1)s'^2_x + (n_y - 1)s'^2_y}{n_x + n_y - 2}$.

Plugging $\overline{x} = 22.78$, $\overline{y} \approx 23.67$, $s'_x = 1.04$, $s'_y = 0.85$, $n_x = 5$, $n_y = 3$, we have

$$\frac{\overline{y} - \overline{x}}{s'_D \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \approx \frac{23.67 - 22.78}{0.98 \times 0.73} \approx 1.24 < t_\alpha(6) \approx 1.94$$

Not reject H_0

7.3.3 配对样本

> Paired t-test

Is there a difference in performance between midterm and final exams? Suppose the distributions of marks of exams are normal (both population means and variances are unknown)

$X_i \sim \mathcal{N}(\mu_x, \sigma_x^2)$ $Y_i \sim \mathcal{N}(\mu_y, \sigma_y^2)$

ID	Midterm	Final
1	80	90
2	60	60
3	70	55
4	73	68
5	40	70
6	90	95

$H_0 : \mu_x = \mu_y$ vs. $H_1 : \mu_x \neq \mu_y$

Can we use the test statistic: $\overline{X} - \overline{Y}$ to find a critical value or the p-value?

No, because $\{X_i\}$'s and $\{Y_i\}$'s are not independent.

Here, each pair of X_i and Y_i are dependent, as they are from student i. Therefore, $\overline{X}$ and $\overline{Y}$ are also dependent

If X and Y are dependent normal variables

$$X - Y \sim N(\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2 - 2\sigma_{xy})$$

其中，σ_{xy} 是协方差（covariance）.

引入一个新变量 Difference：

$X_i \sim \mathcal{N}(\mu_x, \sigma_x^2)$ $Y_i \sim \mathcal{N}(\mu_y, \sigma_y^2)$ $D_i = Y_i - X_i$

ID	Midterm	Final	Difference
1	80	90	10
2	60	60	0
3	70	55	-15
4	73	68	-5
5	40	70	30
6	90	95	5

Note: $D_1, \ldots, D_6$ are independent, they follow a normal distribution:

$$D_i \sim N(\mu_d, \sigma_D^2)$$

> 这里 σ_D^2 未知，变为单统计量小样本、正态分布、方差未知的复合假设双尾检验.

$$\mu_d = \mathbb{E}[D_i] = \mathbb{E}[Y_i - X_i] = \mu_x - \mu_y$$

$H_0 : \mu_d = 0$ vs. $H_1 : \mu_d \neq 0$

$\overline{D}$ follows a normal distribution

$$T = \frac{\overline{D} - 0}{S'/\sqrt{6}} \sim t(6-1)$$

Suppose $\alpha = 0.05$

$$\frac{|4.17 - 0|}{15.3/\sqrt{6}} \approx 0.67 < t_{0.025}(5) \approx 2.57$$

Not reject H_0

ESTR 额外课程

Lec 1 大纲

> 考卷、curve 与普通课一致. 唯一的不同是多了 Project.

1.1 课程内容

- **主题涵盖**:
  - 统计机器学习相关

    > Link to statistical machine learning

  - 进阶贝叶斯统计推断：多变量分布、近似推断（采样，变分推断）

    > Advanced Bayesian statistical inference：Multivariate distributions，Approximate inference (Sampling, variational inference)

  - 进阶经典统计推断（其他点估计量及其性质）

    > Other point estimators
    >
    > Properties of point estimators

1.2 Project

- **项目要求**:
  - 个人或最多两人一组完成
  - 项目主题聚焦于贝叶斯推断的近似方法

> The topic of the project is focused on approximate methods for Bayesian inference

  - 需要提交报告、代码和演示视频

> Submission: report + code + presentation video
>
> 可以使用 Library，手搓更好.

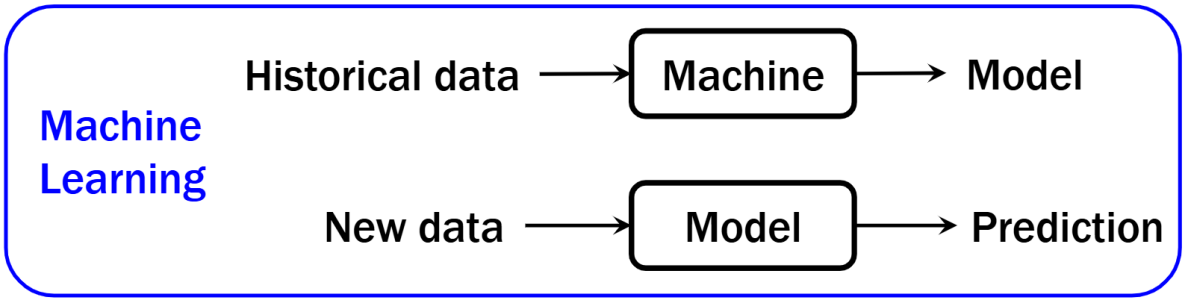
  - 详细信息将在第4周的讲座中介绍（2025年1月27日）

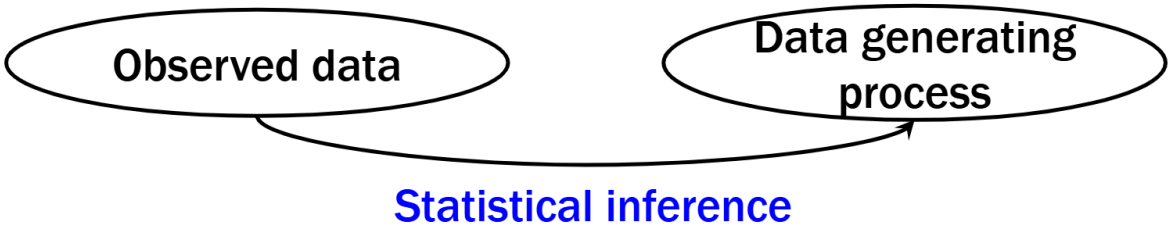
1.3 机器学习基础

- **机器学习的定义**:
  - 机器学习关注开发能够从数据中自我学习并适应新数据的计算机程序

> Focus on the development of computer programs that can teach themselves to grow from data and change when exposed to new data

  - 数据生成过程、观察数据、模型和预测的关系





Statistical inference

观察数据 - 推断模型 - 预测数据

1.4 统计机器学习

- **一般机器学习问题**:
  - 给定一组历史训练数据对 $x_1, y_1, \ldots, x_n, y_n$，其中 x_i 是m维数值向量，y_i 是标量（更一般情况下是向量）

- 目标是学习一个映射函数 $f: x \to y$，使得 $f(x_i) = y_i$，并能在新的未见数据 x^* 上做出精确预测 $f(x^*)$

1.5 假设空间

- **假设**:
  - 训练数据 $(x_1, y_1), \ldots, (x_n, y_n)$ 是独立同分布（i.i.d.）且来自未知的联合概率分布 $P_{tr}(x, y)$
  - 未见的测试数据 x^*, y^* 也是i.i.d.，且来自未知的联合概率分布 $P_{ts}(x, y)$
  - 假设 $P_{tr}(x, y) = P_{ts}(x, y)$

1.6 损失函数

- **损失函数**:
  - 用于衡量预测值 $f(x)$ 与真实值 y 之间的差异
  - 定义为 $t(f(x), y) = t(y, y)$

1.7 风险最小化

- **风险定义**:
  - 风险是损失函数在联合概率分布 $P(x, y)$ 下的期望
  - 目标是找到使风险最小的假设 $f^* = \arg\min R(f)$
  - 风险的定义为 $R(f) = \mathbb{E}_{x,y \sim P}[t(f(x), y)]$

1.8 经验风险最小化

- **经验风险**:
  - 由于联合概率分布 $P_{tr}(x, y)$ 未知，无法采样无限多的数据对
  - 实践中使用有限的训练数据对来近似期望风险，即经验风险 $R_{tr}(f) = \frac{1}{n} \sum_{i=1}^{n} t(f(x_i), y_i)$
  - 通过最小化经验风险来学习假设 $f = \arg\min R_{tr}(f)$

1.9 结构风险最小化

- **结构风险**:
  - 为了防止过拟合，引入正则化项 $d(f)$，即结构风险 $R(f) = \frac{1}{n} \sum_{i=1}^{n} t(f(x_i), y_i) + \lambda d(f)$
  - $\lambda > 0$ 是 trade-off 超参数

1.A 分布不一致

- **分布不一致**:
  - 如果 $P_{ts}(x,y) \neq P_{tr}(x,y)$，则涉及迁移学习、领域适应、领域泛化、出分布泛化等主题

这份PDF文件为统计机器学习的理论基础提供了全面的概述，并为后续的课程内容和项目工作奠定了基础.

Lec 2 常用概率分布

2.1 Beta-Bernoulli

> 在 `1.1.3 贝叶斯推断实例` 的 Example 1、`2.1 贝叶斯法则` 的 Example 2 中亦有提及.
>
> 注意 $Bernoulli$ 指的是伯努利试验中成功次数的分布，而 $Beta$ 指的是概率参数（单次试验成功可能性）的分布，二者对应硬币模型中的两个不同变量.

A coin flip $X \sim Bernoulli(\theta)$, where $P(\theta) = P(X\ is\ Head)$

θ is a value of the random variable Θ. Here we want to know the distribution of Θ.

$Prior$: $\Theta \sim Beta(\alpha, \beta)$ $(\alpha, \beta > 0)$

> 没经验就设置为均匀分布（$\alpha = 1, \beta = 1$），此时仍满足 Beta 分布.
>
> （2025.1.13 思考）超参数 α, β 的本质是，在下一批观测前，记录目前已观测数据中的正反个数，体现经验所得；如果没有任何经验，则默认一正一反，这对应哲学上的 “先验” 概念，即独立于知识和经验而存在.
>
> 但统计学上的先验含义更广，无论观察了多少次，积累了多少经验，对于下一批观察而言，前面的所有观察包括最初的哲学 “先验” 都统称为 Prior，这更符合哲学中的 “知识/认知/认识” 概念.
>
> 康德在《纯粹理性批判》中提到，认识来源于理性和经验的结合. 纯粹理性可以通过经验不断拓展对现象世界的认知，但不能认识超越经验的事物.

Observed data: $D = \{x_i\}_{i=1}^{n}$, where $\sum_{i=1}^{n} x_i = k$

> 正面取 1，反面取 0，观察 n 次抛掷结果.

$Posterior$: $\Theta|D \sim Beta(\alpha + k, \beta + n - k)$

共轭分布模型的优点：不用逐步分析，不用列贝叶斯公式，不用设置似然函数，不用计算归一化常数，一步到位由先验直接得到后验.

> 归一化常数往往难以计算.
>
> 类似背公式，这些模型都是经验总结而来，大大简化计算.

2.2 分类分布

> Categorical Distribution

硬币只有两种结果，可以用伯努利建模，如果是骰子呢（六种结果）？

- 引入分类分布作为伯努利分布的多变量版本.

> The multivariable version of Bernoulli distribution

Suppose a varibale X has m outcomes $(1, \dots, m)$ with probability $\theta_1, \dots, \theta_m$, where $\sum_{i=1}^{m} \theta_i = 1$

> $m = 2$ 时是伯努利分布.
>
> 可结合骰子模型来理解. 对于均质六面骰，$m = 6$，$\theta_1 = \theta_2 = \cdots = \theta_6 = \frac{1}{6}$

Denote a parameter vector by $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)$

X 服从分类分布，记作 $X \sim Cat(\boldsymbol{\theta})$.

> 注意这里 $\boldsymbol{\theta}$ 是一个向量.

PMF: $P(X = k) = \theta_k$

2.2.1 One-Hot

> 注意，不要和 `ENGG 2020 数字电路 1.5.2 格雷码` 混淆.

- 独热编码（One-Hot Encoding）
- 独热向量（One-Hot Vectors）

独热编码是一种常见编码方式，用于将分类数据（类别型变量）转换为机器学习模型可以处理的数值数据. 核心思想是用二进制向量（独热向量）表示每个类别，其中只有一个位置为 1，其余位置为 0.

为什么需要独热编码?

许多机器学习算法无法直接处理非数值数据（如文本类别），需要将其转换为数值形式. 直接用整数标记类别可能会引入错误的顺序关系或距离信息（例如，将类别 A、B、C 分别编码为 1、2、3 会让模型误以为 B 比 A 更接近 C）. 独热编码通过二进制向量避免了这一问题.

独热编码的原理

假设一个分类变量有 N 个不同类别（如 “苹果”、“香蕉”、“橘子”），独热编码会为每个类别创建一个长度为 N 的二进制向量，其中每个类别都有一个唯一的位置标记为 1，其他位置为 0.

示例：假设有一个水果类型变量 $[\text{苹果}, \text{香蕉}, \text{橘子}]$，独热编码结果如下

类别	独热编码
苹果	$[1, 0, 0]$
香蕉	$[0, 1, 0]$
橘子	$[0, 0, 1]$

缺点：如果类别数量非常多，独热编码会导致维度爆炸，增加存储和计算成本.

利用独热编码，可以很好地拓展观测数据量. 梯度也更好计算.

Alternatively, use a m dimensional one-hot encoding vector $\mathbf{x}$ to represent the random variable $\mathbf{X}$, if the outcome is i, only the i-th element is 1, others are 0.

独热编码时，服从分类分布的变量 $\mathbf{X}$ 是一个向量变量. 它的取值是独热向量.

e.g. throw a dice three times and get $5, 6, 2$，可以这么表示：

$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\boldsymbol{X}_3$
0	0	0
0	0	1
0	0	0
0	0	0
1	0	0
0	1	0

> 这里书写不规范，规范写法应该是小写加粗 $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}$，加粗表示向量，小写表示单个观测结果. 如果向量观测结果用大写加粗，向量分布就无法表示（没有 “大大写”）.

对于 PMF，独热编码只是变化了形式；对于多个观测结果的联合分布，独热编码则大大简化了计算.

PMF: $P(\mathbf{x}) = \prod_{i=1}^{m} \theta_i^{\mathbf{x}[i]}$

> 注意 0 次方是 1. 出现的位置计入连乘，出现次数为对应 θ_i 的指数.
>
> 这里的 $\mathbf{x}$ 是单次投掷. 如果多次投掷，且 $\mathbf{x}$ 记作各次结果的向量和，则公式左侧变量不能写 $\mathbf{x}$，而应该写为 $\mathbf{x_1}, \cdots, \mathbf{x_n}$，它的含义是各次结果分别为 $\mathbf{x_1}, \cdots, \mathbf{x_n}$ 的概率（即联合分布概率），而不能理解为 "和为 $\mathbf{x}$" 的概率（后者要用多项式定理乘以系数 $\frac{n!}{k_1!k_2!\cdots k_m!}$，其中 $k_1 + \cdots + k_m = n$）. 在 `2.5 多项分布` 中亦有提及.

联合分布（认为各次投掷结果相互独立）：

$$P(\mathbf{x_1}, \cdots, \mathbf{x_n}) = \prod_{j=1}^{n} \prod_{i=1}^{m} \theta_i^{\mathbf{x}_j[i]} = \prod_{i=1}^{m} \theta_i^{\mathbf{x}[i]}$$

其中 $\mathbf{x} = \mathbf{x_1} + \cdots + \mathbf{x_n}$.

> 作为 Dirichlet-Categorical 模型的似然函数，记作 $P_{\mathbf{X}_1,\ldots,\mathbf{X}_n|\Theta}(\mathbf{x_1}, \ldots, \mathbf{x_n}|\boldsymbol{\theta})$.

2.3 狄利克雷分布

> Dirichlet Distribution

伯努利分布仅成功不成功两种情况，成功率只需要单个参数 θ 来表示.

> 如果看作狄利克雷的特殊情况，则 $m = 2,\ \theta_1 = \theta,\ \theta_2 = 1 - \theta$.

当伯努利拓展为分类分布，对应的概率参数也拓展为 m 个. 这里有个易犯的错误，$\theta_1, \theta_2, \cdots, \theta_m$ 的分布不能用 m 个 $Beta$ 分布的联合来算，因为这 m 个参数是不独立的（有一个约束方程 $\theta_1 + \cdots + \theta_m = 1$）. 因此需要引入新的分布.

- 引入狄利克雷分布作为 Beta 分布的多变量版本.

> The multivariable version of Beta distribution

Denote a random vector by $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)$, and a parameter vector by $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_m)$

$\boldsymbol{\theta}$ 是向量随机变量 $\boldsymbol{\Theta}$ 的一个取值. $\boldsymbol{\Theta}$ 服从狄利克雷分布，记作

$$\boldsymbol{\Theta} \sim Dir(\boldsymbol{\alpha})$$

概率密度函数（PDF）：

$$Dir(\boldsymbol{\alpha}) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{m} \theta_i^{\alpha_i - 1} & \text{for } 0 < \theta_i < 1, \ \sum_{i=1}^{m} \theta_i = 1 \\ 0 & \text{otherwise} \end{cases}$$

其中 $B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{m} \Gamma(\alpha_i)}{\Gamma(\alpha_0)}$, where $\alpha_0 = \sum_{i=1}^{m} a_i$.

> 这里是超几何 Beta 函数的定义，是 Beta 函数的高维推广.

- 狄利克雷作为分类分布的共轭先验.

> Dirichlet is a conjugate prior of Categorical
>
> Dirichlet is a conjugate prior of Multinomial（见 `2.5 多项分布`）

Suppose m dimensional random vectors $\mathbf{x_1}, \cdots, \mathbf{x_n}$ form a random sample from Categorical distribution with m unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m), \ 0 < \theta_i < 1$. If the prior distribution $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ is the Dirichlet distribution $Dir(\boldsymbol{\alpha}), \ \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$, then the posterior distribution $f_{\boldsymbol{\Theta}|\mathbf{X_1}, \ldots, \mathbf{X_n}}(\boldsymbol{\theta}|\mathbf{x_1}, \ldots, \mathbf{x_n})$ given $\{\mathbf{x_i}\}_{i=1}^{n}$ is the Dirichlet distribution $Dir(\boldsymbol{\alpha'})$, where $\boldsymbol{\alpha'} = [\alpha'_1, \ldots, \alpha'_m], \ \alpha'_i = \alpha_i + \mathbf{x}[i], \ \mathbf{x} = \sum_{i=1}^{n} \mathbf{x_i}, \ \sum_{i=1}^{m} \mathbf{x}[i] = n.$

	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$	$\boldsymbol{X_4}$	$\boldsymbol{X_5}$	$\boldsymbol{X_6}$
Outcome 1	1	0	1	0	0	0
Outcome 2	0	1	0	0	0	0
Outcome 3	0	0	0	1	1	0
Outcome 4	0	0	0	0	0	1

$$\sum_{i=1}^{6} \boldsymbol{X}_i = \begin{array}{|c|}\hline \boldsymbol{X} \\ \hline 2 \\ \hline 1 \\ \hline 2 \\ \hline 1 \\ \hline \end{array}$$

> 这里规范书写应该是小写加粗.

证明：

	X_1	...	X_j	...	X_n	X	
1	1	...	1	0	0	$X_1[1]+\cdots+X_n[1]$	$X[1]$
...	...	...	...	...	...	...	
i	0	...	0	...	1	$X_1[i]+\cdots+X_n[i]$	$X[i]$
...	...	...	...	...	...	...	
m	0	...	0	...	0	$X_1[m]+\cdots+X_n[m]$	$X[m]$

> 这里规范书写是小写加粗.

Observations

$$D = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$$

> One-Hot vectors to represent m exclusive possible outcomes

Prior

$$\boldsymbol{\Theta} \sim Dir(\boldsymbol{\alpha})$$

Posterior

$$\begin{aligned} f_{\boldsymbol{\Theta}|\mathbf{X_1},\ldots,\mathbf{X_n}}(\boldsymbol{\theta}|\mathbf{x_1},\ldots,\mathbf{x_n}) &\propto f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) P_{\mathbf{X_1},\ldots,\mathbf{X_n}|\boldsymbol{\Theta}}(\mathbf{x_1},\ldots,\mathbf{x_n}|\boldsymbol{\theta}) \\ &\propto \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{m} \theta_i^{\alpha_i - 1} \prod_{j=1}^{n} \prod_{i=1}^{m} \theta_i^{\mathbf{x}_j[i]} \\ &\propto \prod_{i=1}^{m} \theta_i^{\alpha_i - 1} \prod_{i=1}^{m} \theta_i^{\mathbf{x}[i]} \\ &\propto \prod_{i=1}^{m} \theta_i^{\mathbf{x}[i] + \alpha_i - 1} \end{aligned}$$

> 这里应用了连乘符号的交换和同底数幂相乘法则（底数不变指数相加）.
>
> 此处 $\mathbf{x} = \mathbf{x_1} + \cdots + \mathbf{x_n}$.

Denote $\boldsymbol{\alpha'} = [\alpha'_1, \ldots, \alpha'_m]$，其中 $\alpha'_i = \alpha_i + \mathbf{x}[i]$. 由 Dirichlet 分布可知，

$$\int_{\Delta_m} \prod_{i=1}^{m} \theta_i^{\mathbf{x}[i] + \alpha_i - 1} d\theta_1 \cdots d\theta_m = B(\boldsymbol{\alpha'})$$

其中 Δ_m 表示 m 维单纯形（simplex）：

$$\Delta_m = \{(\theta_1, \ldots, \theta_m) | \theta_i \geq 0, \sum_{i=1}^{m} \theta_i = 1\}.$$

综上，

$$f_{\boldsymbol{\Theta}|\mathbf{X_1},\ldots,\mathbf{X_n}}(\boldsymbol{\theta}|\mathbf{x_1},\ldots,\mathbf{x_n}) = \frac{1}{B(\boldsymbol{\alpha'})} \prod_{i=1}^{m} \theta_i^{\alpha'_i - 1}$$

后验分布满足狄利克雷，即

$$\boldsymbol{\Theta}|\mathbf{X_1},\ldots,\mathbf{X_n} \sim Dir(\boldsymbol{\alpha'})$$

2.4 二项分布

> Binomial Distribution

- 引入二项分布作为伯努利分布的推广.

Suppose there are n Bernoulli independent variable X_i with parameter p (the probability of success)

$$X_1 + \cdots + X_n = X \sim Binomial(n, p)$$

概率质量函数

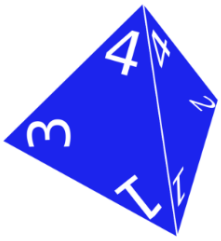
$$\begin{aligned} PMF(k, n, p) &= P(X = k; n, p) \\ &= \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

2.5 多项分布

> Multinomial Distribution

- 讨论多项分布作为二项分布的推广.
- Multinomial 分布也可看作 n 次重复独立 Categorical 试验的结果分布.

> 重复独立分类分布试验的联合分布.



4 sided dice

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	$\sum_{i=1}^{7} X_i =$	X
Outcome 1	1	0	1	0	0	1	0		3
Outcome 2	0	1	0	0	0	0	0		1
Outcome 3	0	0	0	1	1	0	0		2
Outcome 4	0	0	0	0	0	0	1		1

Probabilities for outcomes: $\theta_1, \theta_2, \theta_3$,and θ_4

What is the probability of this outcome?

> 三个 1，一个 2，两个 3，一个 4.

$$\begin{aligned} P(\mathbf{x}) &= \binom{7}{3} \cdot \binom{7-3}{1} \cdot \binom{4-1}{2} \cdot \binom{2-1}{1} \theta_1^3 \cdot \theta_2^1 \cdot \theta_3^2 \cdot \theta_4^1 \\ &= \frac{7!}{3!1!2!1!} \theta_1^3 \cdot \theta_2^1 \cdot \theta_3^2 \cdot \theta_4^1 \\ &= \binom{7}{3,1,2,1} \theta_1^3 \cdot \theta_2^1 \cdot \theta_3^2 \cdot \theta_4^1 \end{aligned}$$

多项分布

Suppose there are m outcomes, with probabilities $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ respectively, where $\sum_{i=1}^{m} \theta_i = 1$. Suppose we have n independent trials, and let $\mathbf{x} = [k_1, \ldots, k_m]$ be the random vector of counts of each outcome.

$\mathbf{x}$ 是向量随机变量 $\mathbf{X}$ 的一个取值. $\mathbf{X}$ 满足多项分布

$$\mathbf{X} \sim Mult(n, \boldsymbol{\theta})$$

PMF of $\mathbf{X}$ is

$$P(\mathbf{x}) = \binom{n}{k_1, \ldots, k_m} \prod_{i=1}^{m} \theta_i^{k_i}$$

where $k_1, \ldots, k_m \geq 0$ and $\sum_{i=1}^{m} k_i = n$.

- 狄利克雷作为多项分布的共轭先验.

> Dirichlet is a conjugate prior of Multinomial
>
> 待证明.

Lec 3 贝叶斯预测

Lec 4 贝叶斯推断

回顾贝叶斯推断：

$$f_{\Theta|X}(\theta|x) = \frac{f_{\Theta}(\theta) f_{X|\Theta}(x|\theta)}{Z(x)}$$

- $f_{\Theta|X}(\theta|x)$：后验（Posterior）.
- $f_{\Theta}(\theta)$：先验（Prior）.
- $f_{X|\Theta}(x|\theta)$：似然（Likelihood）.
- $Z(x)$：归一化常数.

$$Z(x) = \int_{\theta} f_{\Theta}(\theta) f_{X|\Theta}(x|\theta) d\theta$$

对于多观测数据，记 $D = \{X_1, \ldots, X_n\}$ 为联合随机变量，$d = \{x_1, \ldots, x_n\}$ 为一个取值. 贝叶斯推断可写作

$$f_{\Theta|D}(\theta|d) = \frac{f_\Theta(\theta) f_{D|\Theta}(d|\theta)}{Z(d)}$$
$$Z(d) = \int_\theta f_\Theta(\theta) f_{D|\Theta}(d|\theta) d\theta$$

贝叶斯预测：

$$\begin{aligned} f_{X|D}(x^*|d) &= \int_{-\infty}^{+\infty} f_{X|\Theta}(x^*|\theta) f_{\Theta|D}(\theta|d) d\theta \\ &= \mathbb{E}_{\Theta|D=d}[f_{X|\Theta}(x^*|\theta)] \end{aligned}$$

4.1 计算挑战

> Computational Challenges

4.1.1 计算后验分布

计算贝叶斯预测必须先算出后验 $f_{\Theta|D}(\theta|d)$. 注意，是要算出准确的后验公式，而不只是得到正比于先验乘似然的关系. 而为了计算准确的后验，则不得不计算归一化常数 $Z(d)$. 然而，Computing $Z(d)$ requires computing very high dimensional integrals as θ is multivariate in practice.

> 参数往往是向量甚至矩阵形式，使积分变得十分复杂.

4.1.2 计算预测分布

即使解决了后验分布问题，贝叶斯预测本身也包含对 θ 的积分：

$$f_{X|D}(x^*|d) = \int_{-\infty}^{+\infty} f_{X|\Theta}(x^*|\theta) f_{\Theta|D}(\theta|d) d\theta$$

同理，也难以计算.

总之，后验和预测的计算都涉及了对 θ 的积分. 因此，计算准确的后验，乃至准确的贝叶斯预测都是理论可行，但实际不可行的（计算能力有限）. 为此，科学家提出两种解决方案：

- Sample from $f_{\Theta|D}(\theta|d)$ and use sample mean to approximate expectation.
- Approximate $f_{\Theta|D}(\theta|d)$ by $q(\theta)$, which is a simple distribution.

> 见 `Lec 6 近似推断`.

4.2 采样算法

贝叶斯预测：

$$\begin{aligned} f_{X|D}(x^*|d) &= \int_{-\infty}^{+\infty} f_{X|\Theta}(x^*|\theta) f_{\Theta|D}(\theta|d) d\theta \\ &= \mathbb{E}_{\Theta|D=d}[f_{X|\Theta}(x^*|\theta)] \end{aligned}$$

如 `4.1 计算挑战` 所示，贝叶斯预测面临两个计算问题. 我们分开解决，先解决预测分布本身的这个积分. 虽然我们很难对 θ 的整个定义域进行积分，但这个积分的本质是求期望，利用样本均值估计期望的无偏性，我们可以 draw independent sample $\{\theta_1, \dots, \theta_n\}$ from posterior distribution $f_{\Theta|D}(\theta|d)$，计算一系列 $f_{X|\Theta}(x^*|\theta)$，and use sample mean to approximate expectation:

> 假设我们已经实现了某些方法对后验分布进行采样，如何实现见下面.

$$\overline{f_{X|\Theta}(x^*|\theta)} = \frac{1}{n}\sum_{i=1}^{n} f_{X|\Theta}(x^*|\theta_i)$$

> 这步也叫蒙特卡洛近似（Monte Carlo Approximation）.

无偏性：

$$\mathbb{E}_{\Theta|D=d}[f_{X|\Theta}(x^*|\theta)] = \mathbb{E}_{\Theta|D=d}\left[\overline{f_{X|\Theta}(x^*|\theta)}\right]$$

抽象成一般问题，我们有 Problem Setup：

Suppose $\mathbf{z}$ is multi-variate random variable, and we are interested in evaluating the expectation:

$$\mathbb{E}_{\mathbf{Z}}\left[h(\mathbf{z})\right] = \int_{\mathbf{z}} h(\mathbf{z}) f(\mathbf{z}) d\mathbf{z}$$

> $f(\mathbf{z})$ 对应贝叶斯推断的参数后验.
>
> $h(\mathbf{z})$ 对应贝叶斯预测的基于后验的似然.
>
> 整个积分相当于似然函数对后验参数分布的期望，结果是预测数据分布，即基于已观测数据对总体数据分布的预测分布（最大该分布即可得到预测值）.

其中，

$$f(\mathbf{z}) = \frac{g(\mathbf{z})}{Z}$$

> $g(\mathbf{z})$ 对应贝叶斯推断的先验乘似然，好计算；
>
> Z 对应贝叶斯推断的归一化常数，难计算.

Our objective is to draw independent samples $\{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ from $f(\mathbf{z})$ to approximate $\mathbb{E}_{\mathbf{Z}}\left[h(\mathbf{z})\right]$:

$$\mathbb{E}_{\mathbf{Z}}\left[h(\mathbf{z})\right] \approx \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{z}_i)$$

接下来将进一步介绍如何从 $f(\mathbf{z})$ 中采样.

4.2.1 基本采样

> Basic Sampling Algorithm: Inverse Transform Sampling

假设我们已经能够从均匀分布中生成随机样本，本节介绍如何从非均匀分布中生成随机样本.

目标：已实现从均匀分布 $Z \sim \text{Uniform}(0,1)$ 中采样 z，希望实现从给定非均匀分布 $Y \sim \text{Given Non-Uniform Distribution}$ 中采样 y.

方法：$y = F_Y^{-1}(z)$.

> 先采样 z，然后代入 Y 的累积分布函数的反函数生成 y.

证明：假设 $y = a(z)$，$a(\cdot)$ 是未知变换.

$$\begin{aligned} & P(y \le Y \le y+dy) = P(z \le Z \le z+dz) \\ \Leftrightarrow\ & f_Y(y)\,|dy| = f_Z(z)\,|dz| \\ \Leftrightarrow\ & f_Y(y) = f_Z(z)\left|\frac{dz}{dy}\right| = \left|\frac{dz}{dy}\right| \end{aligned}$$

方便讨论起见，设 $dy,\ dz > 0$，有

$$z = \int_0^z dz = \int_{-\infty}^{y} f_Y(y)dy = F_Y(y) \Leftrightarrow y = F_Y^{-1}(z).$$

可知 $a(\cdot)$ 即 Y 的累积分布函数的反函数.

> 注意，这里的给定分布 $Y \sim \text{Given Non-Uniform Distribution}$ 必须满足概率分布的性质，即 PDF 已归一化，对所有可能取值积分为 1.

Multivariable case: $\mathbf{y} = a(\mathbf{z})$

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Z}}(\mathbf{z})\,|\det \mathbf{J}|$$

其中，$|\det \mathbf{J}|$ 是雅可比行列式的绝对值.

> Absolute Value of the Jacobian Determinant.

雅可比行列式：雅可比矩阵为方阵时的行列式.

雅可比矩阵：假设某函数 $f : \mathbb{R}^n \to \mathbb{R}^m$，有 $\mathbf{y} = f(\mathbf{x})$. 其中 $\mathbf{y} \in \mathbb{R}^m$，$\mathbf{x} \in \mathbb{R}^n$. 定义一个 $m \times n$ 的矩阵

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{y}}{\partial x_1} & \cdots & \frac{\partial \mathbf{y}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

为该函数的雅可比矩阵.

4.2.2 拒绝采样

> Rejection Sampling

回归正题. 我们想从目标分布 $f(\mathbf{z}) = \frac{g(\mathbf{z})}{Z}$ 中采样，但归一化常数 Z 算不出来，难以直接采样.

> 注意，虽然 `4.2.1 基本采样` 介绍了如何从非均匀分布中采样，但那也要满足「分布已知」的前提. 未知分布是无法直接采样的.

虽然我们不知道 Z，因而无法作出 $f(\mathbf{z})$ 的图像. 但 $g(\mathbf{z})$ 等于先验乘似然是相对好计算的，可以直接求出 $g(\mathbf{z})$ 图像. 于是，引入拒绝采样法.

拒绝采样法

目标:

从目标分布 $f(\mathbf{z})=\frac{g(\mathbf{z})}{Z}$ 中采样，其中 $g(\mathbf{z})$ 是某个未归一化的密度函数，易求；Z 是归一化常数，难求.

方法:

① 引入一个易于采样的提议分布 $q(\mathbf{z})$.

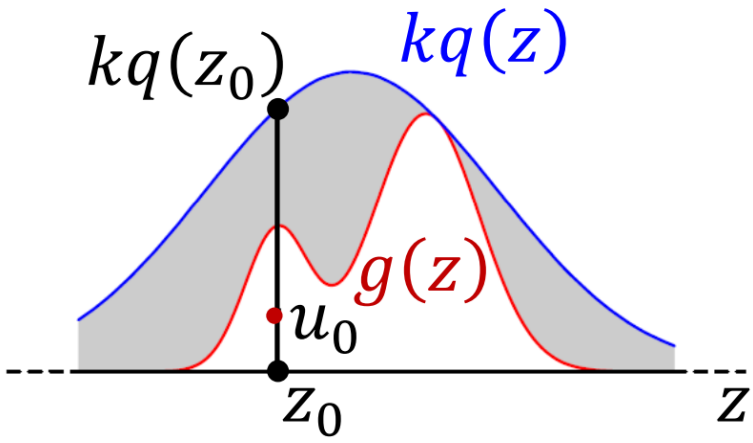
② 找到常数 k，使得 $kq(\mathbf{z}) \geq g(\mathbf{z})$ 对任意 $\mathbf{z}$ 成立.

③ 从 $q(\mathbf{z})$ 采样一个样本 $\mathbf{z}_0$.

④ 从均匀分布 $\mathrm{Uniform}(0, kq(\mathbf{z}_0))$ 中采样一个值 u_0.

⑤ 若 $u_0 > g(\mathbf{z}_0)$，拒绝该样本；否则接受.

⑥ 重复 ③ ∼ ⑤，直到接受的样本数达到指定数量.



下面逐步说明该方法的合理性.

① 提议分布

> proposal distribution

提议分布 $q(\mathbf{z})$ 没有固定选择，要根据以下原则综合考量:

- 易于采样.
- 与 $g(\mathbf{z})$ 形状相似.

> 可以减小 k，提高接受率.

常用选择:

形状	推荐的
近似正态	正态分布 $N(\mu, \sigma^2)$

$g(\mathbf{z})$ 形状	推荐的 $q(\mathbf{z})$
近似均匀	均匀分布 $\mathrm{Uniform}(a, b)$
近似单侧	指数分布 $\mathrm{Exp}(\lambda)$
近似长尾	柯西分布、学生 t 分布
复杂分布	混合多个分布进行优化

② 常数 k

$$kq(\mathbf{z}) \geq g(\mathbf{z}), \qquad \forall \mathbf{z}$$

$kq(\mathbf{z})$ 形成了 $g(\mathbf{z})$ 的上界. 而上界有无穷多个，理论上 k 有无穷多解. 但我们想要最紧的上界，使 k 尽可能小，从而提高采样效率. 那么，理想的 k 满足

$$k = \sup_{\mathbf{z}} \frac{g(\mathbf{z})}{q(\mathbf{z})}$$

> 注意这里的 k 比 $f(\mathbf{z}) = \frac{g(\mathbf{z})}{Z}$ 中的 Z 要好算很多，因为导数比积分好算.

③ 一次采样

从提议分布 $q(\mathbf{z})$ 中采样一个样本 $\mathbf{z}_0$. 关于如何从给定非均匀分布中采样，见 `4.2.1 基本采样` .

④ 二次采样

得到 $\mathbf{z}_0$ 后，从均匀分布 $\mathrm{Uniform}(0, kq(\mathbf{z}_0))$ 中再采样一个值 u_0.

> 默认我们能够从均匀分布中生成随机样本.

⑤ 拒绝判定

若 $u_0 > g(\mathbf{z}_0)$，拒绝该样本；否则接受.

⑥ 重复 ③ ∼ ⑤

重复 ③ ∼ ⑤，直到接受的样本数达到指定数量. 最终抽取的样本集等效于从目标分布 $f(\mathbf{z}) = \frac{g(\mathbf{z})}{Z}$ 中采样得到的样本集.

证明：

由于 u_0 在 $[0, kq(\mathbf{z}_0)]$ 内均匀分布，其密度为

$$f_{U \sim \mathrm{Uniform}(0, kq(\mathbf{z}_0)) | \mathbf{z}}(u_0 | \mathbf{z}_0) = \frac{1}{kq(\mathbf{z}_0)}, \quad 0 \leq u_0 \leq kq(\mathbf{z}_0)$$

因此，接受某个 $\mathbf{z}_0$ 的概率是

$$P(\text{accept}|\mathbf{z}_0) = \frac{g(\mathbf{z}_0)}{kq(\mathbf{z}_0)}$$

最终接受的样本 PDF 满足

$$f(\mathbf{z}_0|\text{accept}) \propto q(\mathbf{z}_0)P(\text{accept}|\mathbf{z}_0) = \frac{g(\mathbf{z}_0)}{k} \propto g(\mathbf{z}_0)$$

归一化后即目标分布 $f(\mathbf{z}_0|\text{accept}) = f(\mathbf{z}) = \frac{g(\mathbf{z})}{Z}$.

虽然拒绝采样最终生成的样本符合目标分布，但

- 采样效率取决于 k：如果 $kq(\mathbf{z})$ 紧密包裹 $g(\mathbf{z})$，那么拒绝率较低，采样效率高；反之大部分采样点都被拒绝，效率低下.
- 难以用于高维场景：在高维空间中，构造合适的 $kq(\mathbf{z})$ 非常困难，采样效率极低.

因此，拒绝采样适用于低维、容易找到合适提议分布的情况，但在高维场景下，通常使用其他方法（如重要性采样、马尔可夫链蒙特卡洛等）.

> 见 `4.2.3 重要性采样`，`4.2.4 MCMC 方法` 和 `Lec 5 马尔可夫链蒙特卡洛`.

4.2.3 重要性采样

> Importance Sampling

回顾 Problem Setup：

Suppose $\mathbf{z}$ is multi-variate random variable, and we are interested in evaluating the expectation:

$$\mathbb{E}_{\mathbf{Z}}\left[h(\mathbf{z})\right] = \int_{\mathbf{z}} h(\mathbf{z})f(\mathbf{z})d\mathbf{z}$$

> $f(\mathbf{z})$ 对应贝叶斯推断的参数后验.
>
> $h(\mathbf{z})$ 对应贝叶斯预测的基于后验的似然.
>
> 整个积分相当于似然函数对后验参数分布的期望，结果是预测数据分布，即基于已观测数据对总体数据分布的预测分布（最大该分布即可得到预测值）.

其中，

$$f(\mathbf{z}) = \frac{g(\mathbf{z})}{Z}$$

> $g(\mathbf{z})$ 对应贝叶斯推断的先验乘似然，好计算；
>
> Z 对应贝叶斯推断的归一化常数，难计算.

Our objective is to draw independent samples $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ from $f(\mathbf{z})$ to approximate $\mathbb{E}_{\mathbf{Z}}\left[h(\mathbf{z})\right]$:

$$\mathbb{E}_{\mathbf{Z}}\left[h(\mathbf{z})\right] \approx \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{z}_i)$$

注意，我们最终目标其实不是实现对 $f(\mathbf{z})=\frac{g(\mathbf{z})}{Z}$ 的抽样，而是计算期望 $\mathbb{E}_{\mathbf{Z}}[h(\mathbf{z})]=\int_{\mathbf{z}}h(\mathbf{z})f(\mathbf{z})d\mathbf{z}$. Draw independent samples $\{\mathbf{z}_1,\dots,\mathbf{z}_n\}$ from $f(\mathbf{z})$ 并计算 $\mathbb{E}_{\mathbf{Z}}[h(\mathbf{z})]\approx\frac{1}{n}\sum_{i=1}^{n}h(\mathbf{z}_i)$ 只是一种估计方法. 如果我们只需计算期望，而不关心生成的样本分布，可以引入重要性采样.

重要性采样

目标：

计算 $\mathbb{E}_{\mathbf{Z}}[h(\mathbf{z})]=\int_{\mathbf{z}}h(\mathbf{z})f(\mathbf{z})d\mathbf{z}$，其中 $f(\mathbf{z})=\frac{g(\mathbf{z})}{Z}$. $g(\mathbf{z})$ 是某个未归一化的密度函数，易求；Z 是归一化常数，难求.

方法：

① 引入一个易于采样的提议分布 $q(\mathbf{z})$.

② 从 $q(\mathbf{z})$ 采样 n 个样本 $\mathbf{z}_1,\dots,\mathbf{z}_n$.

③ 计算期望 $\mathbb{E}_{\mathbf{Z}}[h(\mathbf{z})]\approx\sum_{i=1}^{n}w_ih(\mathbf{z}_i)$.

> 其中 $w_i=\frac{\frac{g(\mathbf{z}_i)}{q(\mathbf{z}_i)}}{\sum_{i=1}^{n}\frac{g(\mathbf{z}_i)}{q(\mathbf{z}_i)}}$ 为重要性权重，下面会详细解释.

① 提议分布

> proposal distribution

和拒绝采样类似，引入一个易于采样的提议分布 $q(\mathbf{z})$.

② n 次采样

从 $q(\mathbf{z})$ 采样 n 个样本 $\mathbf{z}_1,\dots,\mathbf{z}_n$. 关于如何从给定非均匀分布中采样，见 `4.2.1 基本采样` .

③ 计算期望

计算期望 $\mathbb{E}_{\mathbf{Z}}[h(\mathbf{z})]\approx\sum_{i=1}^{n}w_ih(\mathbf{z}_i)$，其中 $w_i=\frac{\frac{g(\mathbf{z}_i)}{q(\mathbf{z}_i)}}{\sum_{i=1}^{n}\frac{g(\mathbf{z}_i)}{q(\mathbf{z}_i)}}$ 为重要性权重.

这么算的合理性解释如下：

重写期望，有

$$\mathbb{E}_{\mathbf{Z}}[h(\mathbf{z})]=\int_{\mathbf{z}}h(\mathbf{z})f(\mathbf{z})d\mathbf{z}=\int_{\mathbf{z}}h(\mathbf{z})\frac{f(\mathbf{z})}{q(\mathbf{z})}q(\mathbf{z})d\mathbf{z}.$$

原来是基于目标分布 $f(\mathbf{z})$ 抽样计算 $h(\mathbf{z})$ 的期望；重写后可看作基于提议分布 $q(\mathbf{z})$ 抽样计算 $h(\mathbf{z})\frac{f(\mathbf{z})}{q(\mathbf{z})}$ 的期望. 用不规范的写法即

$$
\begin{aligned}
\mathbb{E}_{\mathbf{Z}\sim f(\mathbf{z})}\left[h(\mathbf{z})\right] &= \mathbb{E}_{\mathbf{Z}\sim q(\mathbf{z})}\left[h(\mathbf{z})\frac{f(\mathbf{z})}{q(\mathbf{z})}\right] \\
&\approx \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{z}_i)\frac{f(\mathbf{z}_i)}{q(\mathbf{z}_i)} \qquad \text{蒙特卡洛近似} \\
&= \sum_{i=1}^{n} h(\mathbf{z}_i)\frac{f(\mathbf{z}_i)}{nq(\mathbf{z}_i)}
\end{aligned}
$$

记 $w_i = \frac{f(\mathbf{z}_i)}{nq(\mathbf{z}_i)}$ 为基于提议分布 $q(\mathbf{z})$ 抽样的 $h(\mathbf{z}_i)$ 对应权重. 计算目标期望等价为计算 $h(\mathbf{z}_i)$ 以 w_i 为权重的加权平均值. 注意，若提议分布 $q(\mathbf{z})$ 恰好等于目标分布 $f(\mathbf{z})$，则所有权重均为 $\frac{1}{n}$，每个样本同等重要；若 $q(\mathbf{z}) \neq f(\mathbf{z})$，则 $\frac{f(\mathbf{z})}{q(\mathbf{z})}$ 越大的地方抽出来的样本越重要. 这样，通过对 $q(\mathbf{z})$ 大于 $f(\mathbf{z})$ 的地方生成的样本降低权重，对 $q(\mathbf{z})$ 小于 $f(\mathbf{z})$ 的地方生成的样本提高权重，就能用 $q(\mathbf{z})$ 加权样本来模拟 $f(\mathbf{z})$ 不加权样本的表现.

> 例如，$q(\mathbf{z})$ 比 $f(\mathbf{z})$ 大的地方，本来不需要生成这么多这附近的样本，但是 $q(\mathbf{z})$ 生成了过多样本，就可以通过降低权重的方式抵消过量样本的影响，这样就能模拟 $f(\mathbf{z})$ 抽出的样本对 $h(\mathbf{z})$ 的贡献了.

这里，可以证明 w_i 是已归一化的权重：

$$
\begin{aligned}
\sum_{i=1}^{n} w_i &= \sum_{i=1}^{n} \frac{f(\mathbf{z}_i)}{nq(\mathbf{z}_i)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{f(\mathbf{z}_i)}{q(\mathbf{z}_i)} \\
&\approx \int_{\mathbf{z}} \frac{f(\mathbf{z})}{q(\mathbf{z})} q(\mathbf{z}) d\mathbf{z} \qquad \text{蒙特卡洛近似逆运算} \\
&= 1
\end{aligned}
$$

但是，有个棘手的问题：对于每个权重 $w_i = \frac{f(\mathbf{z}_i)}{nq(\mathbf{z}_i)} = \frac{g(\mathbf{z}_i)}{nZq(\mathbf{z}_i)}$，由于 Z 未知，权重也无法算出. 但我们已知各权重比例：

$$
w_1 : w_2 : \cdots : w_n = \frac{g(\mathbf{z}_1)}{q(\mathbf{z}_1)} : \frac{g(\mathbf{z}_2)}{q(\mathbf{z}_2)} : \cdots : \frac{g(\mathbf{z}_n)}{q(\mathbf{z}_n)}
$$

以及它们的和：

$$
\sum_{i=1}^{n} w_i \approx 1
$$

于是，可以对比例 $\frac{g(\mathbf{z}_1)}{q(\mathbf{z}_1)} : \frac{g(\mathbf{z}_2)}{q(\mathbf{z}_2)} : \cdots : \frac{g(\mathbf{z}_n)}{q(\mathbf{z}_n)}$ 归一化计算出各权重：

$$
w_i = \frac{\frac{g(\mathbf{z}_i)}{q(\mathbf{z}_i)}}{\sum\limits_{i=1}^{n} \frac{g(\mathbf{z}_i)}{q(\mathbf{z}_i)}}, \qquad i = 1, 2, \ldots, n.
$$

> 定义归定义，计算归计算. 不一定要按照定义来计算，可以另辟蹊径，只要证明你的方法和定义等效即可.

同样是归一化常数，有限次采样求和 $\sum\limits_{i=1}^{n} \frac{g(\mathbf{z}_i)}{q(\mathbf{z}_i)}$ 是可计算的，积分 $Z = \int_{\mathbf{z}} g(\mathbf{z}) d\mathbf{z}$ 是难以计算的. 实际上，我们正是利用蒙特卡洛近似，用有限次采样得到的归一化常数 $\sum\limits_{i=1}^{n} \frac{g(\mathbf{z}_i)}{q(\mathbf{z}_i)}$ 来近似代替积分的归一化常数 nZ：

> 注意这里 nZ 是因为定义中 $w_i = \frac{f(\mathbf{z}_i)}{nq(\mathbf{z}_i)} = \frac{g(\mathbf{z}_i)}{nZq(\mathbf{z}_i)} = \frac{\frac{g(\mathbf{z}_i)}{q(\mathbf{z}_i)}}{nZ}$ 分母为 nZ.

$$
\begin{aligned}
\sum_{i=1}^{n} \frac{g(\mathbf{z}_i)}{q(\mathbf{z}_i)} &= \frac{1}{n} \sum_{i=1}^{n} \frac{ng(\mathbf{z}_i)}{q(\mathbf{z}_i)} \\
&\approx \int_{\mathbf{z}} \frac{ng(\mathbf{z})}{q(\mathbf{z})} q(\mathbf{z}) d\mathbf{z} \qquad \text{蒙特卡洛近似逆运算} \\
&= nZ.
\end{aligned}
$$

或者，既然我们已经得到 $\sum_{i=1}^{n} w_i \approx 1$，代入 $w_i = \frac{f(\mathbf{z}_i)}{nq(\mathbf{z}_i)} = \frac{g(\mathbf{z}_i)}{nZq(\mathbf{z}_i)}$ 即

$$
\sum_{i=1}^{n} \frac{g(\mathbf{z}_i)}{nZq(\mathbf{z}_i)} \approx 1 \Leftrightarrow nZ \approx \sum_{i=1}^{n} \frac{g(\mathbf{z}_i)}{q(\mathbf{z}_i)}.
$$

对比：拒绝采样 vs. 重要性采样

相同点：

- 都使用一个提议分布 $q(\mathbf{z})$ 来近似目标分布 $f(\mathbf{z})$，因为直接从 $f(\mathbf{z})$ 采样困难.
- 都需要计算比率 $\frac{g(\mathbf{z})}{q(\mathbf{z})}$.
- 都使用了蒙特卡洛近似.

不同点：

方法	目标	采样方式	核心计算	适用场景
拒绝采样	生成符合 $f(\mathbf{z})$ 的样本	先从 $q(\mathbf{z})$ 采样，再用接受-拒绝机制筛选	$k = \sup_{\mathbf{z}} \frac{g(\mathbf{z})}{q(\mathbf{z})}$	低维，目标分布形状常见
重要性采样	估计期望值 $\mathbb{E}_{\mathbf{Z}}[h(\mathbf{z})] = \int_{\mathbf{z}} h(\mathbf{z}) f(\mathbf{z}) d\mathbf{z}$	直接从 $q(\mathbf{z})$ 采样，所有样本都被利用	$w_i = \frac{\frac{g(\mathbf{z}_i)}{q(\mathbf{z}_i)}}{\sum_{i=1}^{n} \frac{g(\mathbf{z}_i)}{q(\mathbf{z}_i)}}$	高维

注意，拒绝采样在高维情况下效率极低，因为合适的 k 和 $q(\mathbf{z})$ 都不好找，导致大多数样本被拒绝. 重要性采样弥补了这一缺点，它不需要 $q(\mathbf{z})$ 和 $f(\mathbf{z})$ 形状相似或将其包裹，并且所有样本都被利用，采样效率高.

当然，重要性采样也有缺点：首先它只适用于不关注 $f(\mathbf{z})$ 分布的情况，因为它只计算期望，不生成符合 $f(\mathbf{z})$ 的样本；其次，如果 $q(\mathbf{z})$ 和 $f(\mathbf{z})$ 形状差距过大，会导致权重方差很大，影响计算精度.

4.2.4 MCMC 方法

> 马尔可夫链蒙特卡洛（Markov Chain Monte Carlo, MCMC）

如上所示，高维场景下，使用重要性采样虽然一定程度解决了拒绝采样的低效率问题，但是又引入了新的精度问题. 因此，我们仍然需要引入更高级的采样方法，如马尔可夫链蒙特卡洛方法.

> 见 `Lec 5 马尔可夫链蒙特卡洛`.

Lec 5 马尔可夫链蒙特卡洛

> Markov Chain Monte Carlo, MCMC

马尔可夫链蒙特卡洛是一类用于从复杂概率分布中采样的算法，广泛应用于贝叶斯统计、物理模拟、机器学习等领域. MCMC 结合了马尔可夫链（Markov Chain）和蒙特卡洛方法（Monte Carlo Method），利用马尔可夫链的性质构造一个易于采样的序列，使其逐渐收敛到目标分布.

5.1 马尔可夫链

> Markov Chains

一组随机变量 $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$ 是一个一阶马尔可夫链（first-order Markov chain）if the following conditional independence holds

$$P(\mathbf{z}_{k+1}|\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k) = P(\mathbf{z}_{k+1}|\mathbf{z}_k) \qquad \text{for } k \in \{1, \ldots, n-1\}$$

> 当前状态只依赖于前一个状态，而与更早的状态无关.

若为连续变量，则为

$$f(\mathbf{z}_{k+1}|\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k) = f(\mathbf{z}_{k+1}|\mathbf{z}_k) \qquad \text{for } k \in \{1, \ldots, n-1\}$$

5.1.1 生成

要生成一个马尔可夫链，需要：

① 定义初始状态的概率分布 $P(\mathbf{z}_1)$.

② 构造转移核（Transition Kernels）：

$$T_k(\mathbf{z}_{k+1} \leftarrow \mathbf{z}_k) \equiv P(\mathbf{z}_{k+1}|\mathbf{z}_k), \quad k \in \{1, \ldots, n-1\}.$$

若为连续变量，则转移核记为

$$T_k(\mathbf{z}_{k+1} \leftarrow \mathbf{z}_k) \equiv f(\mathbf{z}_{k+1}|\mathbf{z}_k), \quad k \in \{1, \ldots, n-1\}.$$

5.1.2 齐次马尔可夫链

> homogeneous

A Markov chain is called homogeneous if the transition kernels are the same for all k.

> 注意，转移核在所有 k 上相同，意味着它不随时间变化，而不是对于任意 $\mathbf{z}_{k+1}$ 和 $\mathbf{z}_k$ 都会给出相同的值. 转移核本身还是一个分布，不一定是常函数.

5.1.3 边缘概率

> the marginal probability

The marginal probability of a specific state can be computed via product and sum rules，即全概率公式

$$P(\mathbf{z}_{k+1}) = \sum_{\mathbf{z}_k} T(\mathbf{z}_{k+1} \leftarrow \mathbf{z}_k) P(\mathbf{z}_k)$$

或连续变量，以 PDF 形式表示：

$$f(\mathbf{z}_{k+1}) = \int_{\mathbf{z}_k} T(\mathbf{z}_{k+1} \leftarrow \mathbf{z}_k) f(\mathbf{z}_k) d\mathbf{z}_k$$

注意，这里是对 $\mathbf{z}_k$ 所有可能的取值求和或积分，而不是对 k 从 1 到 $n-1$ 求和. 因为 $\mathbf{z}_{k+1}$ 的概率（分布）只取决于 $\mathbf{z}_k$ 和转移核，与更早的状态无关.

注意，这里 $f(\mathbf{z}_{k+1})$ 和 $f(\mathbf{z}_k)$ 可能不是同一个分布，因为没有任何条件说明它们是该马尔可夫链的平稳分布.

5.1.4 平稳分布

> Stationary Distribution

A distribution $P^*(\cdot)$ is said to be stationary or invariant with respect to a Markov chain if each step in the chain leaves $P^*(\cdot)$ invariant

$$P^*(\mathbf{z}) = \sum_{\mathbf{z}'} T(\mathbf{z} \leftarrow \mathbf{z}') P^*(\mathbf{z}'), \qquad \forall \mathbf{z}$$

5.1.5 细致平衡

> Detailed Balance

A sufficient (but not necessary) condition for ensuring $P^*(\cdot)$ is stationary or invariant is to choose a transition kernel to satisfy the property of detailed balance, defined by

$$T(\mathbf{z}' \leftarrow \mathbf{z}) P^*(\mathbf{z}) = T(\mathbf{z} \leftarrow \mathbf{z}') P^*(\mathbf{z}') \qquad \forall \mathbf{z}, \mathbf{z}'$$

若满足细致平衡，则称该链是可逆的（Reversible）.

证明：细致平衡是平稳分布的充分条件

① 离散变量

给定细致平衡条件

$$T(\mathbf{z}' \leftarrow \mathbf{z}) P^*(\mathbf{z}) = T(\mathbf{z} \leftarrow \mathbf{z}') P^*(\mathbf{z}') \qquad \forall \mathbf{z}, \mathbf{z}'$$

两边对 $\mathbf{z}$ 求和

$$\begin{aligned} \sum_{\mathbf{z}} T(\mathbf{z}' \leftarrow \mathbf{z}) P^*(\mathbf{z}) &= \sum_{\mathbf{z}} T(\mathbf{z} \leftarrow \mathbf{z}') P^*(\mathbf{z}') \\ &= P^*(\mathbf{z}') \sum_{\mathbf{z}} T(\mathbf{z} \leftarrow \mathbf{z}') \\ &= P^*(\mathbf{z}'), \qquad \forall \mathbf{z}' \end{aligned}$$

这正是平稳分布的定义.

② 连续变量

给定细致平衡条件

$$f(\mathbf{z})T(\mathbf{z}'|\mathbf{z}) = f(\mathbf{z}')T(\mathbf{z}|\mathbf{z}'), \qquad \forall \mathbf{z}, \mathbf{z}'$$

两边对 $\mathbf{z}$ 积分

$$\begin{aligned}\int_{\mathbf{z}} f(\mathbf{z})T(\mathbf{z}'|\mathbf{z})d\mathbf{z} &= \int_{\mathbf{z}} f(\mathbf{z}')T(\mathbf{z}|\mathbf{z}')d\mathbf{z} \\ &= f(\mathbf{z}')\int_{\mathbf{z}} T(\mathbf{z}|\mathbf{z}')d\mathbf{z} \\ &= f(\mathbf{z}'), \qquad \forall \mathbf{z}'\end{aligned}$$

这正是平稳分布的定义.

5.2 MCMC 方法

> Markov Chain Monte Carlo: Basic Idea

回顾 Problem Setup：

Suppose $\mathbf{z}$ is multi-variate random variable, and we are interested in evaluating the expectation:

$$\mathbb{E}_{\mathbf{Z}}\left[h(\mathbf{z})\right] = \int_{\mathbf{z}} h(\mathbf{z})f(\mathbf{z})d\mathbf{z}$$

> $f(\mathbf{z})$ 对应贝叶斯推断的参数后验.
>
> $h(\mathbf{z})$ 对应贝叶斯预测的基于后验的似然.
>
> 整个积分相当于似然函数对后验参数分布的期望，结果是预测数据分布，即基于已观测数据对总体数据分布的预测分布（最大该分布即可得到预测值）.

其中，

$$f(\mathbf{z}) = \frac{g(\mathbf{z})}{Z}$$

> $g(\mathbf{z})$ 对应贝叶斯推断的先验乘似然，好计算；
>
> Z 对应贝叶斯推断的归一化常数，难计算.

Our objective is to draw independent samples $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ from $f(\mathbf{z})$ to approximate $\mathbb{E}_{\mathbf{Z}}\left[h(\mathbf{z})\right]$:

$$\mathbb{E}_{\mathbf{Z}}\left[h(\mathbf{z})\right] \approx \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{z}_i)$$

使用**马尔可夫链蒙特卡洛方法**实现采样的核心思想：

① 构造一个以 $f(\mathbf{z})$ 为平稳分布的马尔可夫链；

② 通过模拟该链来生成样本；

③ 经过若干步后，样本近似服从 $f(\mathbf{z})$ 分布.

> 大多数 MCMC 算法中，构造的转移核通常是固定的，即构造的马尔可夫链是齐次的.

当马尔可夫链运行足够久，即使初始状态不是从 $f(\mathbf{z})$ 抽取的，最终生成的样本也会收敛到这个分布，从而可以用于近似抽样.

To guarantee the convergence to $f(\mathbf{z})$, the Markov chain needs to be ergodic.

> 遍历性（Ergodicity）是收敛的保证.
>
> 遍历性：任取一个时间段，所有状态都有出现的可能. 数学表示为 $T(\mathbf{z}' \leftarrow \mathbf{z}) > 0$ for any $\mathbf{z}'$ and $\mathbf{z}$.

注意，马尔可夫链的样本是有序列依赖性的，前一个状态会影响下一个状态，因此样本之间存在相关性. The sequence $\{\mathbf{z}_1, \mathbf{z}_2, \ldots\}$ is not a set of independent samples.

解决方法：

① 丢弃前期样本（burn-in）

② 抽稀（thinning）

Discard most of samples and just keep every T-th sample

$$\mathbf{z}_1,\ \mathbf{z}_2,\ \ldots,\ \mathbf{z}_T,\ \mathbf{z}_{T+1},\ \ldots,\ \mathbf{z}_{2T},\ \mathbf{z}_{2T+1},\ \ldots,\ \mathbf{z}_{nT}$$

5.2.1 MH 算法

> Metropolis-Hasting Algorithm

Metropolis-Hastings 算法是一种常用的 MCMC 方法.

核心思想：构造提议分布 $q(\mathbf{z}'|\mathbf{z})$ 生成候选样本，使用接受率决定是否接受该样本.

详细步骤：记目标分布 $f(\mathbf{z}) = \frac{g(\mathbf{z})}{Z}$.

① 构造提议分布 $q(\mathbf{z}'|\mathbf{z})$.

② 选择初始状态 $\mathbf{z}_0$. 设置初始时间 $t = 0$.

③ 从 $q(\mathbf{z}'|\mathbf{z}_t)$ 采样一个样本 $\mathbf{z}'$.

④ 计算接受率 $\alpha(\mathbf{z}_t, \mathbf{z}') = \min\left(1, \frac{g(\mathbf{z}')q(\mathbf{z}_t|\mathbf{z}')}{g(\mathbf{z}_t)q(\mathbf{z}'|\mathbf{z}_t)}\right)$.

⑤ 以概率 α 接受新样本. 若接受，令 $\mathbf{z}_{t+1} = \mathbf{z}'$；否则，令 $\mathbf{z}_{t+1} = \mathbf{z}_t$.

⑥ 更新时间 $t \leftarrow t + 1$.

⑦ 重复 ③ ~ ⑥，直到样本符合要求.

下面逐步说明该方法的合理性.

① 提议分布

> Proposal Distribution

MH 算法通过提议 + 接受机制构造转移核，从而构造以目标分布为平稳分布的马尔可夫链. 因此，提议分布是转移核的一个组件. 提议分布 $q(\mathbf{z}'|\mathbf{z})$ 的选择直接影响 MH 算法的采样效率、收敛速度和样本质量.

> 注意，提议分布是一个条件分布.

构造原则

- 易于采样：必须能高效地从 $q(\mathbf{z}'|\mathbf{z})$ 中采样.
- 易于计算：$\frac{q(\mathbf{z}|\mathbf{z}')}{q(\mathbf{z}'|\mathbf{z})}$ 要易于计算，用于接受率.
- 与目标分布匹配：如果 q 太差，会导致接受率低或收敛慢.

常见提议分布

提议分布类型	特点	适用场景
高斯分布 / 对称分布	简单，通用	低维，目标分布较平滑
独立提议分布	不依赖当前状态	提议与目标分布足够接近
自适应分布	动态调整，提高采样效率	中维，不确定最优提议结构
梯度引导分布	考虑目标分布结构，效率高	高维，目标分布复杂或多峰

简单情况选高斯提议即可；高维或结构复杂时，考虑自适应或基于梯度的方法；提议分布的尺度（如步长）需要调得恰当，接受率 α 通常控制在 20%–50% 之间为佳.

例：高斯提议分布

一维情形：$z \in \mathbb{R}$.

假设基于当前状态 $Z = z$，下一个提议状态随机变量 $Z'|Z = z \sim N(z, \sigma^2)$. 它表示，以当前状态 z 为均值、σ^2 为方差的高斯分布. 即提议分布 PDF 为

$$q(z'|z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z'-z)^2}{2\sigma^2}\right).$$

多维情形：$\mathbf{z} \in \mathbb{R}^d$. $\mathbf{Z}'|\mathbf{Z} = \mathbf{z} \sim N(\mathbf{z}, \Sigma)$. PDF 为

$$q(\mathbf{z}'|\mathbf{z}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{z}'-\mathbf{z})^T\Sigma^{-1}(\mathbf{z}'-\mathbf{z})\right)$$

其中，Σ 是协方差矩阵，控制每个方向的步长与相关性.

② 初始化

> Initializing the Markov Chain

选择初始状态 $\mathbf{z}_0$. 设置初始时间 $t = 0$.

初始状态可以随机选择，例如从均匀分布或高斯分布中采样；也可以基于目标分布的高密度区域进行初始化. 多峰分布中，初始点不同可能导致链探索不同的模式.

注意，包括初始状态 $\mathbf{z}_0$ 在内的前若干步样本通常用于烧入期（Burn-in），最终结果要剔除，用于去除随机选择的初始状态对最终样本的影响. 烧入期长度根据目标分布复杂程度而定，通常为总迭代数的 $10\% \sim 30\%$.

③ 采样

> Sampling from the Proposal Distribution

在每一步，MH 算法从当前状态 $\mathbf{z}_t$ 出发，依据提议分布 $q(\mathbf{z}'|\mathbf{z}_t)$ 采样一个候选状态 $\mathbf{z}'$.

④ 接受率

> Acceptance Probability

计算接受率 $\alpha(\mathbf{z}_t, \mathbf{z}') = \min\left(1, \frac{g(\mathbf{z}')q(\mathbf{z}_t|\mathbf{z}')}{g(\mathbf{z}_t)q(\mathbf{z}'|\mathbf{z}_t)}\right)$.

对称提议分布（如高斯分布）时简化为 $\alpha(\mathbf{z}_t, \mathbf{z}') = \min\left(1, \frac{g(\mathbf{z}')}{g(\mathbf{z}_t)}\right)$.

> 注意，α 括号内的两个参数不能对调. 计算的是从 $\mathbf{z}_t$ 转移到 $\mathbf{z}'$ 的接受率.

提议分布与接受率共同构成转移核. 即

$$\begin{cases} T(\mathbf{z}'|\mathbf{z}_t) = q(\mathbf{z}'|\mathbf{z}_t) \cdot \alpha(\mathbf{z}_t, \mathbf{z}') & \mathbf{z}' \neq \mathbf{z}_t \\ T(\mathbf{z}_t|\mathbf{z}_t) = 1 - \int_{\mathbf{z}' \neq \mathbf{z}_t} q(\mathbf{z}'|\mathbf{z}_t) \cdot \alpha(\mathbf{z}_t, \mathbf{z}') d\mathbf{z}' \end{cases}$$

注意：拒绝候选样本（即 $\mathbf{z}' = \mathbf{z}_t$）的转移概率并不是由提议分布 $q(\mathbf{z}'|\mathbf{z}_t)$ 和接受率 $\alpha(\mathbf{z}_t, \mathbf{z}')$ 的某个具体值直接给出的，而是由所有被拒绝的候选样本的**累计概率**决定的. 这里利用全概率公式用 1 减去候选样本被接受的累计概率得到.

> $T(\mathbf{z}_t|\mathbf{z}_t)$ 本身是一个概率值，而不是概率分布 $T(\mathbf{z}'|\mathbf{z}_t)$ 取 $\mathbf{z}' = \mathbf{z}$ 时的密度值.

⑤ 状态更新

> State Update

以概率 α 接受新样本. 若接受，令 $\mathbf{z}_{t+1} = \mathbf{z}'$；否则，令 $\mathbf{z}_{t+1} = \mathbf{z}_t$.

⑥ 时间更新

> Time Step Increment

更新时间 $t \leftarrow t + 1$.

⑦ **重复 ③ ∼ ⑥**

> Iterate Until Convergence

重复 ③ ∼ ⑥，直到样本符合要求，即满足以下条件之一：

- 达到预设的采样次数（如 10000 步）；
- 完成烧入期 + 有效样本采集；
- 使用收敛诊断指标（如 Gelman-Rubin 指标）判断已收敛.

通过提议分布 $q(\mathbf{z}'|\mathbf{z}_t)$ 和接受率 $\alpha(\mathbf{z}_t, \mathbf{z}')$ 构造的马尔可夫链，以目标分布 $f(\mathbf{z})$ 为平稳分布. 证明如下：

要证明目标分布 $f(\mathbf{z}) = \frac{g(\mathbf{z})}{Z}$ 是马尔可夫链的平稳分布，通常的方法是验证细致平衡条件，即

$$f(\mathbf{z})T(\mathbf{z}'|\mathbf{z}) = f(\mathbf{z}')T(\mathbf{z}|\mathbf{z}'), \qquad \forall \mathbf{z}, \mathbf{z}'.$$

其中，转移核定义为

$$\begin{cases} T(\mathbf{z}'|\mathbf{z}_t) = q(\mathbf{z}'|\mathbf{z}_t) \cdot \alpha(\mathbf{z}_t, \mathbf{z}') & \mathbf{z}' \neq \mathbf{z}_t \\ T(\mathbf{z}_t|\mathbf{z}_t) = 1 - \int_{\mathbf{z}' \neq \mathbf{z}_t} q(\mathbf{z}'|\mathbf{z}_t) \cdot \alpha(\mathbf{z}_t, \mathbf{z}') d\mathbf{z}' \end{cases}$$

接受率定义为

$$\alpha(\mathbf{z}_t, \mathbf{z}') = \min\left(1, \frac{g(\mathbf{z}')q(\mathbf{z}_t|\mathbf{z}')}{g(\mathbf{z}_t)q(\mathbf{z}'|\mathbf{z}_t)}\right)$$

代入细致平衡

$$\frac{g(\mathbf{z})}{Z} \cdot q(\mathbf{z}'|\mathbf{z}) \cdot \alpha(\mathbf{z}, \mathbf{z}') = \frac{g(\mathbf{z}')}{Z} \cdot q(\mathbf{z}|\mathbf{z}') \cdot \alpha(\mathbf{z}', \mathbf{z}), \qquad \forall \mathbf{z}, \mathbf{z}'.$$

分两类讨论，可以得到 $\frac{g(\mathbf{z}')q(\mathbf{z}|\mathbf{z}')}{g(\mathbf{z})q(\mathbf{z}'|\mathbf{z})} \leq 1$ 和 $\frac{g(\mathbf{z}')q(\mathbf{z}|\mathbf{z}')}{g(\mathbf{z})q(\mathbf{z}'|\mathbf{z})} > 1$ 时上式都成立. 证毕.

5.2.2 Gibbs 采样

> Gibbs Sampling

待完成.

Lec 6 近似推断

6.1 拉普拉斯近似

6.2 变分推断

附录

1. 高斯积分

$$\int_0^\infty e^{-x^2}dx=\frac{\sqrt{\pi}}{2}$$

证明：

先考虑完整的高斯积分.

$$I=\int_{-\infty}^{\infty}e^{-x^2}dx$$

平方，得到

$$I^2=(\int_{-\infty}^{\infty}e^{-x^2}dx)(\int_{-\infty}^{\infty}e^{-y^2}dy)$$

在两个积分中，x 和 y 的积分是独立的，可以将二重积分写为

$$I^2=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{-(x^2+y^2)}dxdy$$

转为极坐标，

$$\begin{cases}x=r\cos\theta\\y=r\sin\theta\end{cases}$$

有

$$x^2+y^2=r^2$$
$$dxdy=rdrd\theta$$

积分区域为

$$r\in[0,\infty),\quad \theta\in[0,2\pi).$$

于是

$$I^2=\int_0^{2\pi}\int_0^{\infty}e^{-r^2}rdrd\theta=\pi$$

> 这步是简单的积分，不再赘述.

$$I=\sqrt{\pi}$$

由于 e^{-x^2} 关于 y 轴对称，因此

$$\int_0^{\infty}e^{-x^2}dx=\frac{1}{2}\int_{-\infty}^{\infty}e^{-x^2}dx=\frac{\sqrt{\pi}}{2}$$

2. 分布表

> 见 `ESTR 2018 概率论`.

2.1 离散分布

① 伯努利分布

> Bernoulli Random Variable，又名两点分布，0-1分布
>
> 见 `ESTR 2018 概率论 9.4 伯努利分布`.

A $Bernoulli(p)$ $random\ variable$ X shows the result of a $trial$ where $X=1$ for the $success\ outcome$ with $probability$ p and $X=0$ for the $failure\ outcome$ with $probability$ $1-p$.

> 注意只有一个参数 p，因为只进行一次实验.

The PMF of a $Bernoulli(p)$ random variable is

x	0	1
$p(x)$	$1-p$	p

The expected value of a $Bernoulli(p)$ random variable is

$\mathbb{E}[X]=0\times(1-p)+1\times p=p$

方差：$p(1-p)$

> 见 `ESTR 2018 概率论 10.12 二项分布的方差`.

② 二项分布

> Binomial Random Variable
>
> 见 `ESTR 2018 概率论 6.1 二项分布`.

We call X a $Binomial(n,p)$ $Random\ Variable$ when X represents $the\ number\ of\ successes$ over n $independent\ trials$, each with a $success\ probability\ of\ p$.

The **probability mass function (PMF)** of a $Binomial(n,p)$ $Random\ Variable$ is

$$p(k)=\mathbb{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

期望：np

> 见 `ESTR 2018 概率论 9.5 二项分布的期望`.

众数：$\lfloor(n+1)p\rfloor$

> 若 $(n+1)p$ 是整数，则有 $(n+1)p$，$(n+1)p-1$ 两个众数

方差：$np(1-p)$

> 见 `ESTR 2018 概率论 10.12 二项分布的方差`.

③ 几何分布

> Geometric Random Variable（注意，不是超几何分布）
>
> 见 `ESTR 2018 概率论 6.2 几何分布` .

We call X a $Geometric(p)\ Random\ Variable$ when X represents $the\ first\ time\ of\ success$ over $a\ series\ of\ independent\ trials\ X_1, X_2, \cdots$, each with a $success\ probability\ of\ p$.

$X =$ first (smallest) n such that $X_n = 1$ (success).

> 描述 n 次实验中，第一次成功在第 k 次的概率.
>
> X 满足几何分布，简记：$X \sim Geometric(p)$.

The **probability mass function (PMF)** of a $Geometric(p)\ Random\ Variable$ is

$$p(k) = \mathbb{P}(X = k) = p(1-p)^{k-1}$$

期望：$\frac{1}{p}$

> 见 `ESTR 2018 概率论 10.6 几何分布的期望`.

众数：1

方差：$\frac{1-p}{p^2}$

> 见 `ESTR 2018 概率论 10.7 几何分布的方差`.

④ 泊松分布

> Poisson Random Variable
>
> 见 `ESTR 2018 概率论 7.2 泊松分布` .

二项分布试验次数 n 足够大（相比期望）时，Binomial Random Variable X 可以近似为 $Poisson\ Random\ Variable\ X$. 这时 n 与 p 不再是变量，只有一个变量 $\lambda = np$.

A $Poisson(\lambda)\ random\ variable\ X$ has the PMF:

$$p(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

> k 是从 0 到正无穷的整数.

期望：λ

> 见 `ESTR 2018 概率论 9.6 泊松分布的期望`.

众数：$\lfloor \lambda \rfloor$

> 若 λ 是整数，则有 λ，$\lambda - 1$ 两个众数.

方差：λ

> 见 `ESTR 2018 概率论 10.13 泊松分布的方差`.

⑤ 多项分布

⑥ 负二项分布

⑦ 超几何分布

⑧ 零膨胀泊松分布

2.2 连续分布

① 均匀分布

> Uniform Random Variable
>
> 见 `ESTR 2018 概率论 11.1 均匀分布` .

A uniform random variable T over internal $[0, 1]$ satisfies

$$\mathbb{P}(T \leq t) = \begin{cases} 0 & \text{if } t \leq 0, \\ t & \text{if } 0 < t \leq 1, \\ 1 & \text{if } 1 < t \end{cases}$$

Consider a Uniform random variable X over internal $[a, b]$. Then,

The PDF of X is

$$f(x) = \begin{cases} 0 & \text{if } x < a, \\ \frac{1}{b-a} & \text{if } a \leq x \leq b, \\ 0 & \text{if } b < x \end{cases}$$

期望：$\mathbb{E}[X] = \int_a^b x \frac{1}{b-a} dx = \frac{a+b}{2}$

方差：$Var[X] = \int_a^b (x - \mathbb{E}[X])^2 \cdot \frac{1}{b-a} \cdot dx = \frac{1}{3(b-a)} (x - \frac{a+b}{2})^3 \Big|_a^b = \frac{(b-a)^2}{12}$

(2024.12.12) 法二：

$\mathbb{E}[X] = \int_a^b x \frac{1}{b-a} dx = \frac{a+b}{2}$

$\mathbb{E}[X^2] = \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{3}(a^2 + b^2 + ab)$

$Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{(b-a)^2}{12}$

② Beta 分布

> 见 `1.2.5 Beta 分布`.

Beta 分布是一种定义在 $[0, 1]$ 上的连续概率分布. 在贝叶斯统计中适合作为参数（如概率）的先验分布. Beta 分布由两个正参数 α 和 β 控制. X 服从 Beta 分布，记作：

$$X \sim Beta(\alpha, \beta), \quad x \in [0, 1], \quad \alpha > 0,\ \beta > 0$$

> 硬币模型中，通常用 Θ 来表示服从 Beta 分布的概率参数，而用 X 表示服从 Bernoulli 分布的成功次数（如正面朝上次数）.

概率密度函数：

$$f_X(x) = \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} & \text{for } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

实际模型中，Beta 分布多用于表示参数服从的分布，即

$$f_\Theta(\theta) = \begin{cases} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} & \text{for } 0 < \theta < 1 \\ 0 & \text{otherwise} \end{cases}$$

其中

$B(\alpha, \beta)$ 是 Beta 函数（归一化常数，Normalization Term）：

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$\Gamma(\cdot)$ 是伽马函数（Gamma Function），是阶乘的推广：

$$\Gamma(n) = (n-1)! \quad \text{当 } n \text{ 为正整数}$$

> 见 `1.2.7 伽马函数`.

$X \sim Beta(\alpha, \beta)$ 的期望为

$$E[X] = \frac{\alpha}{\alpha + \beta}.$$

方差

$$Var[X] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

> 待证明.

众数

$$mode[X] = \frac{\alpha - 1}{\alpha - 1 + \beta - 1} \text{ when } \alpha, \beta > 1$$

> 注意由于 X 是连续分布，这里的众数指概率密度函数最高点对应的 x 值.
>
> 见 `3.2.1 最大后验估计` .

③ 指数分布

> Exponential Random Variable
>
> 见 `ESTR 2018 概率论 12.1 指数分布` .

Rain is falling on your head at a rate of λ drops/sec. How long do we wait until the next drop?

An $Exponential(\lambda)$ random variable has the following PDF:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

The $Exponential(\lambda)$ random variable X satisfies:

The CDF of X is $F(x) = 1 - e^{-\lambda x}$

> $\mathbb{P}(X \geq x) = e^{-\lambda x}.$

The expected value of X is $\mathbb{E}[X] = \frac{1}{\lambda}$

> $\mathbb{E}[X] = \int_0^\infty x\lambda e^{-\lambda x}dx = \int_0^\infty -xd(e^{-\lambda x}) = -xe^{-\lambda x}\big|_0^\infty - \int_0^\infty e^{-\lambda x}d(-x) = \int_0^\infty e^{-\lambda x}dx = \frac{1}{\lambda}$

The variance of X is $Var[X] = \frac{1}{\lambda^2}$

> $\mathbb{E}[X^2] = \int_0^\infty x^2\lambda e^{-\lambda x}dx = \int_0^\infty -x^2d(e^{-\lambda x}) = -x^2e^{-\lambda x}\big|_0^\infty - \int_0^\infty e^{-\lambda x}d(-x^2) = \frac{2\mathbb{E}[X]}{\lambda} = \frac{2}{\lambda^2}$
>
> $Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{\lambda^2}$

The standard deviation of X is $\sigma = \frac{1}{\lambda}$

④ Gamma 分布

> 见 `1.2.6 Gamma 分布`.

伽马分布是一种连续概率分布，常用于建模正值随机变量，尤其是涉及等待时间、分布形状和尺度的场景. 是许多分布（如卡方分布、指数分布）的推广.

Gamma 分布由两个正参数 α 和 β 控制. X 服从 Gamma 分布，记作：

$$X \sim Gamma(\alpha, \beta), \quad x > 0,\ \alpha > 0,\ \beta > 0$$

概率密度函数（PDF）：

$$f_X(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

> $x > 0$：随机变量取值范围为正数.
>
> $\alpha > 0$：形状参数（Shape Parameter）.
>
> $\beta > 0$：尺度参数（Scale Parameter），有时也写为 $\frac{1}{\theta}$，其中 θ 是速率参数.
>
> $\Gamma(\alpha)$：伽马函数.

（2025.3.24）注意，Gamma 分布的 PDF 在 $x = 0$ 处的具体表现需要分类讨论. 这里的公式并不严谨.

> $\alpha > 1$ 时，$f_X(0) = 0$;
>
> $\alpha = 1$ 时，$f_X(0) = \beta$;
>
> $0 < \alpha < 1$ 时，$\lim\limits_{x \to 0^+} f_X(0) = \infty$，$\lim\limits_{x \to 0^-} f_X(0) = 0$. 此时 PDF 在 $x = 0$ 处具有奇异性，$x = 0$ 处为无穷间断点.

虽然 $0 < \alpha < 1$ 时，PDF 在 $x = 0$ 处具有奇异性，但这种奇异性是可积的，因为 $\int_0^\delta x^{\alpha-1} dx = \frac{x^\alpha}{\alpha}\Big|_0^\delta = \frac{\delta^\alpha}{\alpha}$ 对 $0 < \alpha < 1$ 是有限的. 故该 PDF 仍然是合法的.

实际模型中，Gamma 分布多用于表示参数服从的分布，即

$$f_\Theta(\theta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} & \theta > 0 \\ 0 & \theta \leq 0 \end{cases}$$

⑤ 正态分布

> Normal Distribution as the Limit of Binomial Distribution，也叫 $Gaussian\ distribution$.
>
> 见 `ESTR 2018 概率论 12.2 正态分布`.

We define the normal (Gaussian) probability density function (PDF) with parameters μ and σ^2 as

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

Note that the parameters μ and σ^2 in the above definition are equal to the $mean$ and $variance$ of the **normal random variable** X:

$$\mathbb{E}[X] = \mu, \qquad Var[X] = \sigma^2$$

⑥ 卡方分布

⑦ 学生 t 分布

⑧ F 分布

⑨ 对数正态分布

⑩ Weibull 分布

⑪ Cauchy 分布

⑫ Laplace 分布

⑬ 对数 Gamma 分布

⑭ Pareto 分布

⑮ 三角分布

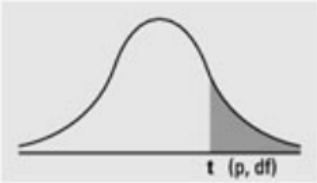
2.3 z 值表

CDF table of standard normal distribution $P(Z \leq z)$

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

2.4 t 值表

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%

3. 置信区间 Q & A

Regarding Confidence Interval II, I have 7 questions:

1. We know that when the population variance σ^2 is unknown, the sample size n is small, and the sample follows an IID normal distribution, the t-distribution and the unbiased sample variance S^2 (we use its square root S) can be used to calculate the confidence interval. **This is an exact calculation without any estimation/approximation, right?** (Here, we disregard any computational error due to rounding in the t-distribution integration.)

Answer: It is based on the Theorem on Page 7 of L7. No approximation is needed.

2. Now, following question 1, the t-distribution applies whether the sample size n is large or small. In principle, even when the sample size n is large—as long as the sample follows an IID normal distribution—we could still use the t-distribution to obtain an exact confidence interval. However, the L07 PPT tells us that when the sample size $n \geq 30$ and the population variance σ^2 is unknown, regardless of whether the sample follows a normal distribution, we uniformly use the z-distribution to approximate/estimate via the CLT.

My understanding is: due to the special nature of the T-table, it is impossible to list every combination of degrees of freedom and probability (for example, if the T-table only goes up to 30 and my sample degrees of freedom is 50, I cannot find that row). Meanwhile, the z-table—being independent of the sample size n (as long as it's above 30)—contains enough density of information to cover nearly all cases with a precision of 0.01 (from 0.00 to 3.09, enough for most cases). Therefore, even though we lose some precision when the sample size n is large, we gain a unified solution and can quickly look up the answer, so we choose the z-distribution approximation over the t-distribution calculation. Am I right?

Answer: Yes, when $n \geq 30$, the t values are very close to z values. There is no need to use the T-table anymore.

3. Concerning question 2, there is another issue: **in Appendix A L07**, when the sample size $n \geq 30$ and the population variance σ^2 is unknown, whether or not the sample follows a normal distribution (both case I and case II), we use the CLT to justify the use of the z-distribution for estimating the confidence interval. But earlier, we mentioned that when the population variance σ^2 is known and the sample is normal (Case I), there is no need to use the CLT because the calculation is exact. So, my question is: in Appendix A, for Case I, does the initial step using the t-distribution followed by the use of the CLT imply that, because the population variance is σ^2 unknown, **the CLT approximation is**

unavoidable? Is it because the χ^2-square distribution, unlike the sum $X_1 + X_2 + \cdots + X_n$, does not strictly follow a normal distribution (especially when n is small) so must rely on the CLT?

Answer: For case I, we don't use CLT. It is just based on the property of t distribution, when $n \geq 30$, the curve of t-distribution can be approximated by that of standard normal.

Additional question 1: For "the curve of t-distribution can be approximated by that of standard normal", which can be written as $t(n) = \frac{N(0,1)}{\sqrt{\frac{\chi^2(n)}{n}}} \overset{n\to\infty}{\longrightarrow} N(0,1) \Leftrightarrow \frac{\chi^2(n)}{n} = \frac{X_1^2+X_2^2+\cdots+X_n^2}{n} \overset{n\to\infty}{\longrightarrow} 1$, I thought that this step of asymptotic approximation still employs the CLT: $\frac{\frac{X_1^2+X_2^2+\cdots+X_n^2}{n}-1}{\sqrt{\frac{2}{n}}} \overset{n\to\infty}{\longrightarrow} N(0,1) \Leftrightarrow \frac{X_1^2+X_2^2+\cdots+X_n^2}{n} \overset{n\to\infty}{\longrightarrow} 1 + \sqrt{\frac{2}{n}}N(0,1) \approx 1.$

Here, the mean of 1 and the standard deviation $\sqrt{\frac{2}{n}}$ are calculated using Gamma integrals from PDF, but it's not our focus, so in case I we skip the proof. But it still use CLT (or say, Law of Large Numbers $\frac{X_1+X_2+\cdots+X_n}{n} \overset{n\to\infty}{\longrightarrow} \mu$), right?

Additional answer 1: Indeed, to show that t-distribution converges to the standard normal distribution, we can show that the PDF of t-distribution converges with the standard normal PDF. We do not need to use CLT.

The PDF of t-distribution can be found at https://en.wikipedia.org/wiki/Student%27s_t-distribution.

Additional question 2: Got it! May we prove like this:

The PDF of t-distribution is:

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})}(1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}},$$

where ν is the number of degrees of freedom and Γ is the gamma function. This may also be written as

$$f(t) = \frac{1}{\sqrt{\nu}B(\frac{1}{2},\frac{\nu}{2})}(1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}},$$

where B is the beta function.

Then, we have $\lim_{\nu\to\infty}(1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}} = e^{-\frac{t^2}{2}} \Leftrightarrow f(t) \overset{\nu\to\infty}{\longrightarrow} N(0,1).$

Here the constant $\frac{1}{\sqrt{\nu}B(\frac{1}{2},\frac{\nu}{2})}$ doesn't need to be computed, because the integral of the PDF over the entire real domain is 1, so the one whose PDF proportional to $e^{-\frac{t^2}{2}}$ is the standard normal distribution.

Additional answer 2: The proof concerns the limits of functions. For more details, please refer to the online or office reference.

4. As mentioned in question 3, when the sample size $n \geq 30$ and the population variance σ^2 is unknown, **if the CLT is unavoidable, could we combine the proofs for Case I and Case II?** That is, could the Case I scenario be merged into Case II for a combined proof, since both methods rely on an asymptotic approximation based on the CLT?

Answer: Refer to the answer of Q3.

5. Following question 4, on one hand, I think these 2 cases may be combined together; on the other hand, I feel that Case I and Case II are somewhat different: **Case II uses Slutsky's Theorem**, which appears to simultaneously approximate both S and $\sqrt{n}(\overline{X} - \mu)$, then takes the ratio of their approximations as an approximation for the true ratio. **Does this approach incur a greater approximation error than the single-step approximation in Case I?** In other words, **does it converge more slowly?** I'm not very clear on the methods and terminology for evaluating approximations.

Answer: Refer to answer of Q4.

6. As of the end of L07, can the methods we have learned for constructing confidence intervals for the population mean μ be summarized as follows:

Categories: Whether the population variance σ^2 is known; whether n is greater than or equal to 30; whether the sample follows a normal distribution. This gives a total of $2 \times 2 \times 2 = 8$ combinations.

Out of these 8, we have taught 6 cases, and 2 cases have not been covered:

① Population variance σ^2 known, $n \geq 30$, sample follows a normal distribution: z-distribution + population variance σ^2, exact solution.

② Population variance σ^2 known, $n \geq 30$, sample does not follow a normal distribution: z-distribution + population variance σ^2, approximate solution.

③ Population variance σ^2 known, $n < 30$, sample follows a normal distribution: z-distribution + population variance σ^2, exact solution.

④ Population variance σ^2 known, $n < 30$, sample does not follow a normal distribution: not taught.

⑤ Population variance σ^2 unknown, $n \geq 30$, sample follows a normal distribution: z-distribution + sample variance s^2, approximate solution.

⑥ Population variance σ^2 unknown, $n \geq 30$, sample does not follow a normal distribution: z-distribution + sample variance s^2, approximate solution.

⑦ Population variance σ^2 unknown, $n < 30$, sample follows a normal distribution: t-distribution + sample variance s^2, exact solution.

⑧ Population variance σ^2 unknown, $n < 30$, sample does not follow a normal distribution: not taught.

Or, write it as two tables. a $(1 - \alpha)$-confidence interval for the mean μ is:

① When σ^2 is known,

sample size \ IID X_i	normal	unknown
small	$[\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}]$	not taught
large	$[\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}]$	$[\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}]$

② When σ^2 is unknown,

sample size \ IID X_i	normal	unknown
small	$[\overline{x} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \overline{x} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}]$	not taught
large	$[\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}]$	$[\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}]$

Answer: A summary has been provided on Page 13 of L7.

7. For the two "not taught", would you mind providing some information for me for further study?

Answer: The untaught case is an open problem as we have no idea about the distribution of the sample mean. As you can see, for all the cases, we use the fact that the sample mean follows a normal distribution or can be approximated by a normal distribution.

4. 内曼-皮尔逊引理证明

> 见 `7.1.5 内曼–皮尔逊引理`.

Neyman-Pearson 引理：在给定第一类错误概率（显著性水平）α 下，对于简单假设 H_0 和 H_1，拒绝域

$$R^* = \{x : L(x) = \frac{f_1(x)}{f_0(x)} > \xi\} \text{（以及在 } L(x) = \xi \text{ 处适当随机化）}$$

构造的似然比检验（LRT）具有最大功效，也就是说，对于任一其他检验 $\phi(x)$ 满足

$$\int \phi(x) f_0(x) dx \leq \alpha,$$

> $\int \phi(x) f_0(x) dx$ 即错误拒绝 H_0 的概率.

都有

$$\int \phi(x) f_1(x) dx \leq \int \phi^*(x) f_1(x) dx.$$

其中 $\phi^*(x)$ 是 LRT 的拒绝函数.

> $\int \phi(x) f_1(x) dx$ 是正确拒绝 H_0 的概率，即 $1-\beta$.

证明:

设 X 在 H_0 下的 PDF / PMF 为 $f_0(x)$，在 H_1 下为 $f_1(x)$. 定义似然比

$$L(x) = \frac{f_1(x)}{f_0(x)}.$$

构造 LRT 的拒绝函数 $\phi^*(x)$ 为

$$\phi^*(x) = \begin{cases} 1, & \text{当 } L(x) > \xi, \\ \gamma, & \text{当 } L(x) = \xi, \\ 0, & \text{当 } L(x) < \xi. \end{cases}$$

> 检验函数 / 拒绝函数 $\phi^*(x)$ 表示在样本 x 下拒绝 H_0 的概率.

其中，选取合适的常数 ξ 和随机化比例 $\gamma \in [0,1]$ 使得

$$\int \phi^*(x) f_0(x) dx = \alpha.$$

> 注意，当数据是离散分布，可能无法找到一个确定的 ξ 使得 $P(L(x) > \xi | H_0) = \alpha$ 成立. 为了解决此问题，我们对边界 $L(x) = \xi$ 上的样本引入随机化.

设另有一个检验函数 $\phi(x)$，满足

$$\int \phi(x) f_0(x) dx \le \alpha.$$

考察积分

$$I = \int (\phi^*(x) - \phi(x)) \left[f_1(x) - \xi f_0(x) \right] dx$$

分三个区域讨论:

① $L(x) = \xi$

此区域 $f_1(x) - \xi f_0(x) = 0$，对积分贡献为 0.

② $L(x) > \xi$

$\phi^*(x) - \phi(x) = 1 - \phi(x) \ge 0$; $f_1(x) - \xi f_0(x) > 0$.

此区域对积分贡献非负.

③ $L(x) < \xi$

$\phi^*(x) - \phi(x) = -\phi(x) \le 0$; $f_1(x) - \xi f_0(x) < 0$.

此区域对积分贡献非负.

综上

$$\begin{aligned} &\int (\phi^*(x) - \phi(x)) \left[f_1(x) - \xi f_0(x) \right] dx \ge 0 \\ \Leftrightarrow &\int (\phi^*(x) - \phi(x)) f_1(x) dx \ge \xi \int (\phi^*(x) - \phi(x)) f_0(x) dx. \end{aligned}$$

注意到

$$\begin{cases}\int \phi^*(x)f_0(x)dx = \alpha \\ \int \phi(x)f_0(x)dx \leq \alpha\end{cases} \Rightarrow \int (\phi^*(x) - \phi(x))f_0(x)dx \geq 0.$$

因此

$$\int (\phi^*(x) - \phi(x))f_1(x)dx \geq \xi \int (\phi^*(x) - \phi(x))f_0(x)dx \geq 0$$

即

$$\begin{aligned}&\int \phi^*(x)f_1(x)dx \geq \int \phi(x)f_1(x)dx \\ \Leftrightarrow\ &1 - \beta(R^*) \geq 1 - \beta(R) \\ \Leftrightarrow\ &\beta(R^*) \leq \beta(R).\end{aligned}$$