Lecture 10: Statistical Inference

Ziyu Shao

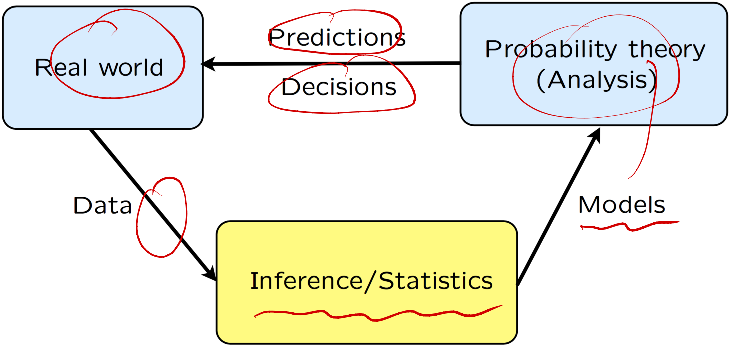
School of Information Science and Technology
ShanghaiTech University

December 24, 2024

Outline

Outline

Probability & Statistics



Statistical Inference

- The process of extracting information from available data
- Called "Learning" in CS
- Called "Signal Processing" in EE

Important Concepts

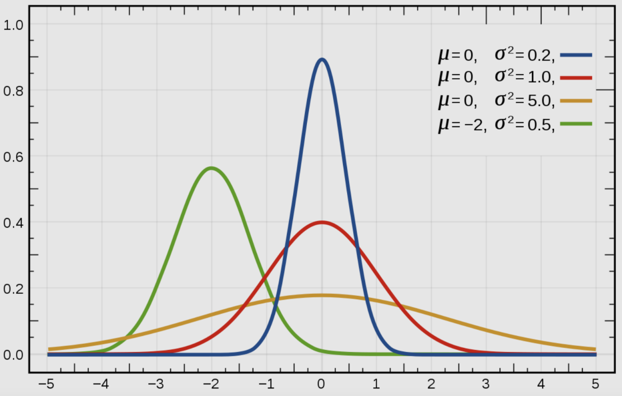
- Population: distribution $\mathcal{F}$
- Sample: random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ where n is the sample size
- Random Sample: $\{X_i\}$ are i.i.d random variables and $X_i \sim \mathcal{F}$
- Data: real vector $\boldsymbol{x} = (x_1, \ldots, x_n)$, the value of sample $\boldsymbol{X}$
- From sample to infer property of Population
- Statistic: a function of sample $\boldsymbol{X}$

Our Focus: Parameterized Statistical Inference

- Given a parametric distribution model (a family of PMFs or PDFs) $\mathcal{F} = \{p(x;\theta) : \theta \in \mathcal{R}\}$
- θ is an unknown parameter in a parameter space $\mathcal{R}$
- Now given (random)sample from such model: $\boldsymbol{X} = (X_1, \ldots, X_n)$
- How to make parameterized statistical inference?

Example: Parameterized Distribution Model

$$\left\{ p(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2 \right), \mu \in \mathbb{R}, \sigma > 0 \right\}.$$



Parameterized Statistical Inference: Bayesian versus Frequentist

- Difference relates to the nature of the unknown parameter θ
- Treated as an random variable Θ with prior (known) distribution: **Bayesian approach**
- Treated as an unknown constant θ: **frequentist approach**

Core Tasks of Statistical Inference

- Point Estimation: Our Focus
- Interval Estimation (Confidence Interval)
- Hypothesis Testing

Outline

Statistical Inference: Frequentist Perspective

- Classical statistics: unknown constant θ

$$\xrightarrow{\theta} \boxed{p_X(x;\theta)} \xrightarrow{X} \boxed{\text{Estimator}} \xrightarrow{\widehat{\Theta}}$$

- Hypothesis testing: $H_0 : \theta = 1/2$ versus $H_1 : \theta = 3/4$

- Composite hypotheses: $H_0 : \theta = 1/2$ versus $H_1 : \theta \neq 1/2$

- Estimation: design an **estimator** $\widehat{\Theta}$, to "keep estimation **error** $\widehat{\Theta} - \theta$ small"

Point Estimation: Frequentist Perspective

- Given parameterized distribution model $p(x;\theta)$ (PMF or PDF)
- θ: unknown parameter.
- Random sample: $\boldsymbol{X} = (X_1, \ldots, X_n)$
- Point Estimation refers to providing a single "best guess" of parameter θ based on random sample $\boldsymbol{X}$
- Estimator $\hat{\Theta} = g(\boldsymbol{X})$: a function of sample $\boldsymbol{X}$
- Estimate $\hat{\theta} = g(\boldsymbol{x})$: when observed date is $\boldsymbol{x}$, i.e., $\boldsymbol{X} = \boldsymbol{x}$

Estimation Method: Maximum Likelihood Estimation (MLE)

$p(x;\theta) = L(\theta)$

- We observe a particular data $\boldsymbol{x} = (x_1, \ldots, x_n)$,
- Likelihood: the probability (or probability density) of seeing data $\boldsymbol{x}$ under different values of parameter θ, i.e., $p(\boldsymbol{x};\theta)$
- A **maximum likelihood estimate** (MLE) is a value of the parameter θ that maximizes the likelihood $p(\boldsymbol{x};\theta)$ over all possible values:

$$\hat{\theta} = \arg\max_{\theta} p(\boldsymbol{x};\theta)$$

$$= \arg\max_{\theta} \log p(x;\theta)$$

MLE under Independent Case

- Random Sample: $\{X_i\}$ are i.i.d., we have

$$\log[p(\boldsymbol{x};\theta)] = \log \prod_{i=1}^{n} p(x_i;\theta) = \sum_{i=1}^{n} \log[p(x_i;\theta)]$$

- Thus a **maximum likelihood estimate** (MLE) under independent case is shown as follows:

$$\hat{\theta} = \arg\max_{\theta} p(\boldsymbol{x};\theta) = \arg\max_{\theta} \log[p(\boldsymbol{x};\theta)]$$

$$= \arg\max_{\theta} \sum_{i=1}^{n} \log[p(x_i;\theta)]$$

Example: Biased Coin Problem

Coin $\underline{\text{Bern}(\theta)}$

$\underline{p: \text{unknow constant.}}$

1°. n coin tosses;

Random Sample $\mathcal{X} = (X_1, \dots X_n)$, $X_i \sim \text{Bern}(\theta)$

$P(X_i = 1) = p$
$P(X_i = 0) = 1-p$

2°. $\underline{X_i} = \underline{x_i}$, $x_i = 0$ or 1.

$\underline{P(X_i = x) = \theta^x (1-\theta)^{1-x}}$

$\underline{x = 0 \text{ or } 1}$

$$L(\theta) = P(\mathbf{x}; \theta) = \prod_{i=1}^{n} p(x_i; \theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}$$

$\underline{S_n = X_1 + \dots + X_n}$

$$= \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n - \sum_{i=1}^{n} x_i} = \theta^{S_n} (1-\theta)^{n-S_n}$$

$\underline{\#\text{ of heads in n coin tosses;}}$

3°. $\log L(\theta) = \underline{S_n \log\theta + (n-S_n)\log(1-\theta)} = \underline{g(\theta)}$

$\left[\begin{array}{l} g'(\theta) = 0 \\ g''(\theta) \le 0 \end{array} \right]$

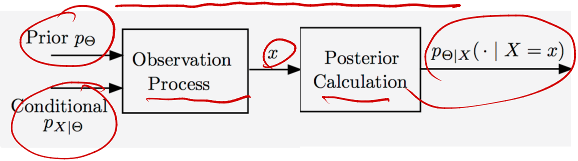
$$\hat{\theta}_{MLE} = \arg\max_{\theta} g(\theta) = \frac{1}{n} S_n = \frac{1}{n}(X_1 + \dots + X_n)$$

Example: Biased Coin Problem

Outline

Statistical Inference: The Bayesian Perspective

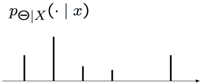
- Unknown Θ
  - treated as a random variable
  - prior distribution p_Θ or f_Θ
- Observation X Likelihood function
  - observation model $p_{X|\Theta}$ or $f_{X|\Theta}$
- Use appropriate version of the Bayes rule to find $p_{\Theta|X}(\cdot \mid X = x)$ or $f_{\Theta|X}(\cdot \mid X = x)$

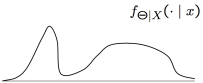


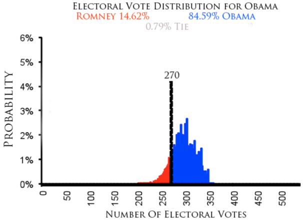
- Where does the prior come from?
  - symmetry
  - known range
  - earlier studies
  - subjective or arbitrary

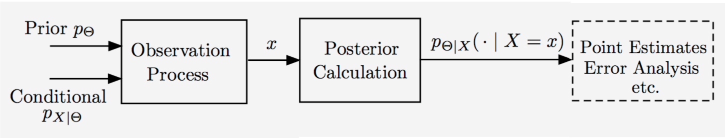
The Output of Bayesian Statistical Inference

The complete answer is a posterior distribution:
PMF $p_{\Theta|X}(\cdot \mid x)$ or PDF $f_{\Theta|X}(\cdot \mid x)$









Recall: General LOTP

	Y **discrete**	Y **continuous**
X **discrete**	$P(X=x)=\sum_y P(X=x\|Y=y)P(Y=y)$	$P(X=x)=\int_{-\infty}^{\infty} P(X=x\|Y=y)f_Y(y)dy$
X **continuous**	$f_X(x)=\sum_y f_X(x\|Y=y)P(Y=y)$	$f_X(x)=\int_{-\infty}^{\infty} f_{X\|Y}(x\|y)f_Y(y)dy$

Recall: General Bayes' Rule

	Y **discrete**	Y **continuous**
X **discrete**	$P(Y=y\|X=x)=\frac{P(X=x\|Y=y)P(Y=y)}{P(X=x)}$	$f_Y(y\|X=x)=\frac{P(X=x\|Y=y)f_Y(y)}{P(X=x)}$
X **continuous**	$P(Y=y\|X=x)=\frac{f_X(x\|Y=y)P(Y=y)}{f_X(x)}$	$f_{Y\|X}(y\|x)=\frac{f_{X\|Y}(x\|y)f_Y(y)}{f_X(x)}$

Bayesian Posterior Calculation

The Four Versions of Bayes' Rule

- Θ discrete, X discrete:

$$p_{\Theta|X}(\theta \,|\, x) = \frac{p_\Theta(\theta) p_{X|\Theta}(x \,|\, \theta)}{\sum_{\theta'} p_\Theta(\theta') p_{X|\Theta}(x \,|\, \theta')}.$$

- Θ discrete, X continuous:

$$p_{\Theta|X}(\theta \,|\, x) = \frac{p_\Theta(\theta) f_{X|\Theta}(x \,|\, \theta)}{\sum_{\theta'} p_\Theta(\theta') f_{X|\Theta}(x \,|\, \theta')}.$$

- Θ continuous, X discrete:

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

- Θ continuous, X continuous:

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

LOTP

Estimation Methods

estimator

$\hat{\Theta} = E[\Theta | X]$

- Method 1 is Posterior Mean: Given the observation data $\boldsymbol{x}$, the estimation of θ is $\hat{\theta} = E[\Theta|\boldsymbol{X} = \boldsymbol{x}]$
- Method 2 is The Maximum A Posteriori Probability (MAP)
- Given the observation value x, the MAP rule selects a value $\hat{\theta}$ that maximizes the posterior probability(probability density) $p_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$:

$$\hat{\theta} = \arg\max_{\theta} p_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$$

- Equivalently,

$$\hat{\theta} = \arg\max_{\theta} p_{\Theta}(\theta) p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)$$

Outline

Beta Distribution

$a=b=1$; $f(x)=c$

$\text{Beta}(1,1) = \text{Unif}(0,1)$ $\int_0^1 f(x)dx=1$

Definition

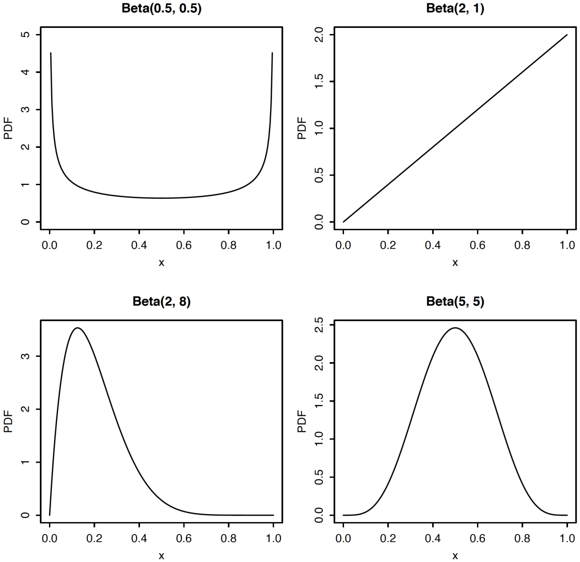
$\Rightarrow c=1$

An r.v. X is said to have the *Beta distribution* with parameters a and b, $a > 0$ and $b > 0$, if its PDF is

$$f(x) = \frac{1}{\beta(a,b)} x^{a-1}(1-x)^{b-1}, \; 0 < x < 1,$$

where the constant $\beta(a,b)$ is chosen to make the PDF integrate to 1. We write this as $X \sim \mathrm{Beta}(a,b)$. Beta distribution is a generalization of uniform distribution.

PDF of Beta Distribution



Expectation of Beta Distribution

$$\beta(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx.$$

Gamma Function

Definition

The gamma function Γ is defined by

$$\Gamma(a) = \int_0^\infty x^a e^{-x} \frac{dx}{x},$$

for real numbers $a > 0$.

Property of Gamma Function

- $\Gamma(a+1) = a\Gamma(a)$ for all $a > 0$.
- $\Gamma(n) = (n-1)!$ if n is a positive integer.

Gamma Distribution

$a = 1$; $f(y) = \lambda e^{-\lambda y}$, $y > 0$

$\mathrm{Expo}(\lambda)$

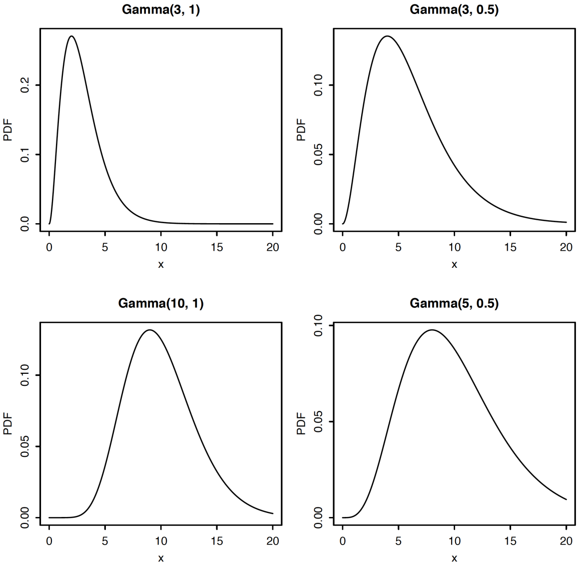
Definition

An r.v. Y is said to have the *Gamma distribution* with parameters a and λ, $a > 0$ and $\lambda > 0$, if its PDF is

$$f(y) = \frac{1}{\Gamma(a)} (\lambda y)^a e^{-\lambda y} \frac{1}{y}, \; y > 0.$$

We write $Y \sim \mathrm{Gamma}(a, \lambda)$. Gamma distribution is a generalization of the exponential distribution.

PDF of Gamma Distribution



Moments of Gamma Distribution

$Y \sim \text{Gamma}(a, \lambda)$

$E(Y) = \frac{a}{\lambda}$

$Var(Y) = \frac{a}{\lambda^2}$

if $\lambda = 1$

$E(Y) = Var(Y)$

$= a$

Pois

Poisson - Gamma Duality

$X \sim Pois(\lambda)$, PMF $P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$

$X \sim \text{Gamma}(k+1, 1)$ PDF $f(x) = \frac{x^k e^{-x}}{k!}$

Gamma: Convolution of Exponential

Theorem

Let $X_1, ..., X_n$ *be i.i.d.* $\mathrm{Expo}(\lambda)$. *Then*

$$\underline{X_1 + \cdots + X_n} \sim \underline{\mathrm{Gamma}(n, \lambda)}.$$

MGF

Beta-Gamma Connection

$X+Y$ independent of $\frac{X}{X+Y}$.

When we add independent Gamma r.v.s X and Y with the same rate λ, the total $X+Y$ has a Gamma distribution, the fraction $\frac{X}{X+Y}$ has a Beta distribution, and the total is independent of the fraction.

Story: Bank–post Office

While running errands, you need to go to the bank, then to the post office. Let $X \sim \mathrm{Gamma}(a, \lambda)$ be your waiting time in line at the bank, and let $Y \sim \mathrm{Gamma}(b, \lambda)$ be your waiting time in line at the post office (with the same λ for both). Assume X and Y are independent. What is the joint distribution of $T = X + Y$ (your total wait at the bank and post office) and $W = \frac{X}{X+Y}$ (the fraction of your waiting time spent at the bank)?

Story: Bank–post Office

$T = X+Y$, $W = \frac{X}{X+Y}$.

① $t>0$, $w>0$; $t = x+y$, $w = \frac{x}{x+y}$ $\Rightarrow$ $x = tw$, $y = t(1-w)$ $\Rightarrow$ $\frac{\partial(x,y)}{\partial(t,w)} = \begin{bmatrix} w & t \\ 1-w & -t \end{bmatrix}$

$\Rightarrow \det(\quad) = -t < 0.$

② $f_{T,W}(t,w) = f_{X,Y}(x,y) \cdot |-t| = f_X(x) \cdot f_Y(y) \cdot t$

$= \frac{1}{\Gamma(a)} (\lambda x)^a e^{-\lambda x} \cdot \frac{1}{x} \cdot \frac{1}{\Gamma(b)} (\lambda y)^b e^{-\lambda y} \cdot \frac{1}{y} \cdot t \quad \left(\begin{matrix} x = tw \\ y = t(1-w) \end{matrix} \right)$

$= \underbrace{\frac{\Gamma(a+b)}{\Gamma(a) \cdot \Gamma(b)} w^{a-1} (1-w)^{b-1}}_{W \sim Beta(a,b)} \cdot \underbrace{\frac{1}{\Gamma(a+b)} (\lambda t)^{a+b} \cdot e^{-\lambda t} \cdot \frac{1}{t}}_{T \sim Gamma(a+b,\lambda)}$

$\frac{1}{\beta(a,b)}$

Story: Bank–post Office

$$\Rightarrow \quad \beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

$W \sim \text{Beta}(a,b)$ $\quad a>0,\ b>0,$ $\quad E(W)$?

$T = X+Y,$ $\quad W = \frac{X}{X+Y},$ $\quad X \sim \text{Gamma}(a,\lambda)$ $\quad E(X) = \frac{a}{\lambda}$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad Y \sim \cdots (b,\lambda)$ $\quad E(Y) = \frac{b}{\lambda}$

T, W are independent $\quad \Rightarrow \quad E[TW] = E[T]\cdot E[W]$

$$\Rightarrow \quad E[W] = \frac{E[W\cdot T]}{E[T]} = \frac{E[X]}{E[X]+E[Y]} = \frac{\frac{a}{\lambda}}{\frac{a}{\lambda}+\frac{b}{\lambda}} = \frac{a}{a+b}$$

Story: Bank–post Office

Outline

Conjugate Prior

- Before Monte Carlo, posterior calculation is hard
- Conjugate Prior: reduce the computing complexity of posterior distribution
- Loosely speaking, a prior distribution is conjugate to the likelihood model if both the prior and posterior distribution stay in the same distribution family.

Story: Beta-Binomial Conjugacy

p

We have a coin that lands Heads with probability p, but we don't know what p is. Our goal is to infer the value of p after observing the outcomes of n tosses of the coin. The larger that n is, the more accurately we should be able to estimate p.

k heads out of n tosses

Story: Beta-Binomial Conjugacy

p r.v. Prior $\text{Beta}(a,b)$

X : # of heads out of n tosses.

likelihood $X \mid p = p \sim \text{Bin}(n, k)$

$$f(p \mid X=k) = \frac{Pr(X=k \mid p)}{Pr(X=k)} \cdot f(p)$$

$$= \binom{n}{k} p^k (1-p)^{n-k} \cdot \frac{1}{\beta(a,b)} p^{a-1} (1-p)^{b-1}$$

$$Pr(X=k) \overset{LOTP}{=} \int_0^1 Pr(X=k \mid p) f(p)\,dp$$

④ $f(p \mid X=k)$ is a function of p. $0 < p < 1$

$$= C \cdot p^k (1-p)^{n-k} \cdot p^{a-1} \cdot (1-p)^{b-1} = C \cdot p^{k+a-1} (1-p)^{n-k+b-1}$$

$$\propto p^{k+a-1} (1-p)^{n-k+b-1} \sim \text{Beta}(k+a, n-k+b).$$

Story: Beta-Binomial Conjugacy

$p \sim \text{Beta}(a,b)$ → $\dfrac{\overset{\text{Likelihood}}{X|p=p \sim \text{Bin}(n,p)}}{X=k \text{ out of } n \text{ tosses}}$ → $p|X=k$ $\sim \text{Beta}(a+k, b+n-k)$

↑ Prior ↑ Postrior

Pseudo counts

(a, b)

↙ # of prior successes ↘ # of prior failures

in earlier experiments

$\xrightarrow{X=k}$

k successes

$n-k$ failures

$(a+k, b+n-k)$

↓ $X'=m$ out of n tosses

$(a+k+m, b+n-k+n-m)$

Story: Beta-Binomial Conjugacy

Story: Beta-Binomial Conjugacy

- Furthermore, notice the very simple formula for updating the distribution of p.
- We just add the number of observed successes, k, to the first parameter of the Beta distribution.
- We also add the number of observed failures, $n - k$, to the second parameter of the Beta distribution.
- So a and b have a concrete interpretation in this context:
  - a as the number of prior successes in earlier experiments
  - b as the number of prior failures in earlier experiments
  - a, b: pseudo counts

Mean vs. Bayesian Average

$Y \sim Beta(a,b)$

$E(Y) = \frac{a}{a+b}$

- Infer the value of p (probability of coin lands heads)
- Observed k heads out of n tosses of the coin

$Beta(a+k, b+n-k)$

MLE

- Mean: $\frac{k}{n}$
- Bayesian Average: $E(p|X=k) = \frac{a+k}{a+b+n}$
- Suppose the prior distribution is Unif(0,1): $a = 1, b = 1$
- Bayesian Average: $\frac{k+1}{n+2}$
- When $k = n$, we have: 1 (mean) vs. $\frac{n+1}{n+2}$ (Bayesian average)

$n=k=3$

MLE

$n \to \infty$ $\frac{n+1}{n+2} \to 1$

Story: Beta-Binomial Conjugacy

If we have a Beta prior distribution on p and data that are conditionally Binomial given p, then when going from prior to posterior, we don't leave the family of Beta distributions. We say that **the Beta is the conjugate prior of the Binomial**.

Example: Inference of A Biased Coin

1°. $\Theta \sim Unif(0,1) = Beta(1,1)$ # of heads $X|\Theta=\theta \sim Bin(n,\theta)$

By Beta-Binomial Conjugacy, $\Theta|X=k \sim Beta(1+k, 1+n-k)$

$\hat{\Theta}_{BA} = E[\Theta|X=k] = \frac{1+k}{1+k+1+n-k} = \frac{k+1}{n+2} \Rightarrow \hat{\Theta}_{BA} = \frac{X+1}{n+2}$.

We wish to estimate the probability of landing heads, denoted by θ, of a biased coin. We model θ as the value of a random variable Θ with a known prior PDF $f_\Theta \sim Unif(0,1)$. We consider n independent tosses and let X be the number of heads observed. Find the MAP estimator of Θ.

posterior PDF

$f_{\Theta|X=k}(\theta) \propto \theta^k(1-\theta)^{n-k}$

2°. MAP Estimator.

$\hat{\Theta}_{MAP} = \arg\max_\theta f_{\Theta|X=k}(\theta) = \arg\max_\theta \theta^k(1-\theta)^{n-k}$.

$\Rightarrow \hat{\Theta}_{MAP} = \frac{k}{n}$. $\Rightarrow \hat{\Theta}_{MAP}(X) = \frac{X}{n}$ $\hat{\Theta}_{MLE} = \frac{X}{n}$

Solution

Solution

Example: Revisit Biased Coin Problem

1° MMSE : $E[\Theta|X] = \frac{X+1}{n+2}$.

estimator under Bayesian Average

2°. LLSE : $L[\Theta|X] = E[\Theta] + \frac{\mathrm{cov}(\Theta, X)}{\mathrm{Var}(X)}(X - E(X))$

We wish to estimate the probability of landing heads, denoted by θ, of a biased coin. We model θ as the value of a random variable Θ with a known prior PDF $f_\Theta \sim Unif(0,1)$. We consider n independent tosses and let X be the number of heads observed. Find the MMSE $E(\Theta|X)$ and LLSE $L(\Theta|X)$.

$\Theta \sim \mathrm{Unif}(0,1) \Rightarrow E(\Theta) = \frac{1}{2}$; $\mathrm{Var}(\Theta) = \frac{1}{12}$; $E(\Theta^2) = \frac{1}{3}$

$X|\Theta=\theta \sim \mathrm{Bin}(n,\theta)$

$\Rightarrow E[X|\Theta=\theta] = n\theta \Rightarrow E[X|\Theta] = n\Theta$

$\mathrm{Var}[X|\Theta=\theta] = n\theta(1-\theta) \Rightarrow \mathrm{Var}[X|\Theta] = n\Theta(1-\Theta)$

Solution

$E[X|\theta] = n\theta$; $Var[X|\theta] = n\theta(1-\theta)$;

$$\Rightarrow \quad E[X] = E[E[X|\theta]] = E[n\theta] = nE[\theta] = \frac{n}{2};$$

$$Var[X] = E[Var(X|\theta)] + Var[\underline{E[X|\theta]}]$$

$$= E[n\theta(1-\theta)] + Var[n\theta]$$

$$= n(E[\theta] - E[\theta^2]) + n^2 \underline{Var(\theta)}$$

$$= n\left[\frac{1}{2} - \frac{1}{3}\right] + n^2 \cdot \frac{1}{12}$$

$$= \frac{n}{12}(n+2)$$

Solution

$$\mathrm{Cov}(X,\theta) = \underline{E[\theta X]} - \underline{E[\theta]\cdot E[X]}$$

$$= \underline{E[E[\theta X|\theta]]} - \underline{\tfrac{1}{2}\cdot\tfrac{n}{2}}$$

$$= E[\theta\, \underline{E[X|\theta]}] \quad \cdots$$

$$= E[\theta\cdot n\theta] \quad \cdots$$

$$= n\,E[\theta^2] - \frac{n}{4}$$

$$= n\cdot\frac{1}{3} - \frac{n}{4} = \frac{1}{12}n$$

$$\Rightarrow \text{LLSE } L[\theta|X] = E[\theta] + \frac{\mathrm{Cov}(\theta,X)}{\mathrm{Var}(X)}[X - E(X)]$$

$$= \frac{1}{2} + \frac{\frac{1}{12}n}{\frac{n}{12}(n+2)}\left(X - \frac{n}{2}\right) = \frac{X+1}{n+2} = \underline{E[\theta|X]}$$

Solution

Recall: Story of Multinomial Distribution

Each of n objects is independently placed into one of k categories. An object is placed into category j with probability p_j, where the p_j are nonnegative and $\sum_{j=1}^{k} p_j = 1$. Let X_1 be the number of objects in category 1, X_2 the number of objects in category 2, etc., so that $X_1 + \ldots + X_k = n$. Then $X = (X_1, \ldots, X_k)$ is said to have the Multinomial distribution with parameters n and $\mathbf{p} = (p_1, \ldots, p_k)$. We write this as $\mathbf{X} \sim Mult_k(n, \mathbf{p})$.

Recall: Multinomial Joint PMF

$\propto p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$

Theorem

If $\mathbf{X} \sim \mathrm{Mult}_k(n, \mathbf{p})$, *then the joint PMF of* $\mathbf{X}$ *is*

$$P(X_1 = n_1, ..., X_k = n_k) = \frac{n!}{n_1! n_2! ... n_k!} \cdot p_1^{n_1} p_2^{n_2} ... p_k^{n_k}$$

for $n_1, ..., n_k$ *satisfying* $n_1 + \cdots + n_k = n$.

Dirichlet Distribution

Beta(a,b) PDF $\propto p^{a-1}(1-p)^{b-1}$

$\frac{a}{a+b}$

The Dirichlet distribution is parameterized by a vector α of positive real numbers.

- The PDF is:

$\propto p_1^{\alpha_1-1} p_2^{\alpha_2-1} \cdots p_k^{\alpha_k-1}$

$$f(p_1, p_2, \ldots, p_k; \alpha_1, \alpha_2, \ldots, \alpha_K) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}$$

where $p_1 + \ldots + p_k = 1$ and $0 < p_i < 1$.

- The marginal mean of P_j is:

$$E(P_j) = \frac{\alpha_j}{\sum_{i=1}^{K} \alpha_i}.$$

Story: Dirichlet-Multinomial Conjugacy

Beta – Binomial Conjugacy

If we have a Dirichlet prior distribution on $\mathbf{p}$ and data that are conditionally Multinomial given $\mathbf{p}$, then when going from prior to posterior, we don't leave the family of Dirichlet distributions. We say that the Dirichlet is the conjugate prior of the Multinomial.

Likelihood Model: Discrete

$(a+k, \; b+n-k)$

$(\alpha_1+n_1, \alpha_2+n_2, \ldots \alpha_k+n_k)$

Sample Space	Sampling Dist.	Conjugate Prior	Posterior
$\mathcal{X} = \{0,1\}$	Bernoulli(θ)	Beta(α,β)	Beta($\alpha+n\overline{X}, \beta+n(1-\overline{X})$)
$\mathcal{X} = \mathbb{Z}_+$	Poisson(λ)	Gamma(α,β)	Gamma($\alpha+n\overline{X}, \beta+n$)
$\mathcal{X} = \mathbb{Z}_{++}$	Geometric(θ)	Gamma(α,β)	Gamma($\alpha+n, \beta+n\overline{X}$)
$\mathcal{X} = \mathbb{H}_K$	Multinomial(θ)	Dirichlet(α)	Dirichlet($\alpha+n\overline{X}$)

Likelihood Model: Continuous

Sampling Dist.	Conjugate Prior	Posterior
Uniform(θ)	Pareto(ν_0, k)	Pareto $\left(\max\{\nu_0, X_{(n)}\}, n+k\right)$
Exponential(θ)	Gamma(α, β)	Gamma$(\alpha+n, \beta+n\overline{X})$
$N(\mu, \sigma^2)$, known σ^2	$N(\mu_0, \sigma_0^2)$	$N\left(\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}\left(\frac{\mu_0}{\sigma_0^2}+\frac{n\overline{X}}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}\right)$
$N(\mu, \sigma^2)$, known μ	InvGamma(α, β)	InvGamma $\left(\alpha+\frac{n}{2}, \beta+\frac{n}{2}\overline{(X-\mu)^2}\right)$
$N(\mu, \sigma^2)$, known μ	ScaledInv-$\chi^2(\nu_0, \sigma_0^2)$	ScaledInv-$\chi^2\left(\nu_0+n, \frac{\nu_0\sigma_0^2}{\nu_0+n}+\frac{n\overline{(X-\mu)^2}}{\nu_0+n}\right)$
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, known $\boldsymbol{\Sigma}$	$N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$	$N\left(\mathbf{K}\left(\boldsymbol{\Sigma}_0^{-1}\mu_0+n\boldsymbol{\Sigma}^{-1}\overline{X}\right), \mathbf{K}\right)$, $\mathbf{K}=\left(\boldsymbol{\Sigma}_0^{-1}+n\boldsymbol{\Sigma}^{-1}\right)^{-1}$
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, known $\boldsymbol{\mu}$	InvWishart$(\nu_0, \mathbf{S}_0)$	InvWishart$(\nu_0+n, \mathbf{S}_0+n\overline{\mathbf{S}})$, $\overline{\mathbf{S}}$ sample covariance

Outline

Rating System

- Consumers rely on the collective intelligence of other consumers: rating
- A common metric: 5 star rating
- Requirement: many ratings are needed to make this system work
- Quality of rating system depends on
    - average number of stars
    - average number of reviews

Which One to Choose?

- 1. Presto Coffee Pot - average rating of 5 (1 review).
- 2. Cuisinart Brew Central - average rating of 4.1 (78 reviews).

Example: Movie Ranking

- Data Set : http://grouplens.org/datasets/movielens/
- Top Ten Movies

Top 10 Movies chosen by Mean

title	count	mean
Aiqing wansui (1994)	1	5
They Made Me a Criminal (1939)	1	5
Great Day in Harlem, A (1994)	1	5
Saint of Fort Washington, The (1993)	2	5
Entertaining Angels: The Dorothy Day Story (1996)	1	5
Someone Else's America (1995)	1	5
Star Kid (1997)	3	5
Santa with Muscles (1996)	2	5
Prefontaine (1997)	3	5
Marlene Dietrich: Shadow and Light (1996)	1	5

Tool: Bayesian Estimation

- Mean of star reviews with a limited number of observations
- Useful for recommender services and other predictive algorithms that use preference space measures like star reviews.

Joint Distribution

- To use Bayesian estimation to compute the posterior probability for star ratings, we must use a joint distribution.
- We are not estimating the distribution of some scalar value X but, rather, the joint distributions of the probability estimate of whether or not the reviewer will give the movie a 1, 2, 3, 4, or 5 star rating (not just a simple thumbs up or down).
- In this case, the random variable is a categorical distribution because it can take some value within 1,2,3,4,5 with probabilities as follows:

$$p_1 + p_2 + p_3 + p_4 + p_5 = 1$$

Multinomial Distribution

- We can compute our posterior probability with N observations for five categories with corresponding numbers K_1, K_2, K_3, K_4, K_5 as follows:

  Likelihood model

  $$Pr(O|p_1, p_2, p_3, p_4, p_5) \propto p_1^{K_1} p_2^{K_2} p_3^{K_3} p_4^{K_4} p_5^{K_5}$$

  where $K_1 + \ldots + K_5 = N$.
- This is a multinomial distribution.

Dirichlet Distribution: Prior

$\alpha^0 = (\alpha_1^0, \ldots, \alpha_5^0)$

prior joint PDF $\propto p_1^{\alpha_1^0 - 1} p_2^{\alpha_2^0 - 1} \cdots p_5^{\alpha_5^0 - 1}$

- If we include our prior as a distribution of the exact same form in the proportionality equation (e.g. a Dirichlet distribution with parameter α^0), then

$\propto Pr(O \mid p_1 \ldots p_5) \times Pr(p_1, \ldots p_5)$

$p_1^{K_1} p_2^{K_2} \cdots p_5^{K_5}$

$$Pr(p_1, p_2, p_3, p_4, p_5 | O) \propto \prod_{j=1}^{5} p_j^{K_j + \alpha_j^0 - 1}$$

$\alpha^1 = (\alpha_1^1, \ldots \alpha_5^1)$

- This is another Dirichlet distribution with another parameter α^1:

$$\alpha_j^1 = K_j + \alpha_j^0, \forall j$$

Expected Average

- What is the expected value of the average rating given a posterior in the shape of our Dirichlet distribution?
- The expected value of the average rating based on the posterior is then computed for our star ratings as follows:

$$E(p_1 + 2p_2 + 3p_3 + 4p_4 + 5p_5|O) = \sum_{i=1}^{5} iE(p_i|O)$$

- Using our Dirichlet distribution we can compute the probability of a star value given our observations as the ratio of the Dirichlet parameter for that star to the sum of the Dirichlet parameters:

$$E(p_i|O) = \frac{\alpha_i^1}{\sum_{j=1}^{5} \alpha_j^1}$$

Intra-Item: Bayesian Average Rating

$$\sum_{i=1}^{5} i \cdot \frac{\alpha_i'}{\sum_{j=1}^{5}\alpha_j'} = \frac{\sum_{i=1}^{5} i \cdot \alpha_i'}{\sum_{j=1}^{5}\alpha_j'} = \frac{\sum_{i=1}^{5} i(\alpha_i^0 + K_i)}{\sum_{j=1}^{5}(\alpha_j^0 + K_j)} = \frac{\sum_{i=1}^{5} i\cdot\alpha_i^0 + \sum_{i=1}^{5} i\cdot K_i}{\sum_{j=1}^{5}\alpha_j^0 + \sum_{j=1}^{5} K_j}$$

$$\text{Bayes Average Rating} = \frac{\sum_{i=1}^{5} i\alpha_i^0 + \sum_{i=1}^{5} iK_i}{N + \sum_{i=1}^{5}\alpha_i^0}$$

$$= \frac{\sum_{i=1}^{5} i\cdot\alpha_i^0 + \sum_{i=1}^{5} i\cdot K_i}{\sum_{i=1}^{5}\alpha_j^0 + N}$$

- N: the number of reviews
- $\sum_{i=1}^{5} iK_i$: sum of all review scores
- $\sum_{i=1}^{5} \alpha_i^0$: prior(given) number of reviews
- $\sum_{i=1}^{5} i\alpha_i^0$: prior sum of all review scores

Intra-Item: Bayesian Average Rating

1°. $C = 0 \longrightarrow \frac{\sum(\text{ratings})}{N}$ (mean)

2°. $N = 0 \longrightarrow m$ (prior average)

$$\text{Bayes Average Rating} = \frac{C \cdot m + \sum(\text{ratings})}{C + N}$$

- N: the number of reviews
- m: a prior for the average of review scores
- C: a prior for the number of reviews

Example: Movie Ranking

- Data Set : http://grouplens.org/datasets/movielens/
- Top Ten Movies

Case 1: $m = 3.25$ & $C = 50$

title	bayes	count	mean
One Flew Over the Cuckoo's Nest (1975)	4.125796	264	4.291667
Raiders of the Lost Ark (1981)	4.145745	420	4.252381
Rear Window (1954)	4.167954	209	4.387560
The Silence of the Lambs (1991)	4.171591	390	4.289744
The Godfather (1972)	4.171706	413	4.283293
The Usual Suspects (1995)	4.206625	267	4.385768
Casablanca (1942)	4.250853	243	4.456790
The Shawshank Redemption (1994)	4.265766	283	4.445230
Star Wars (1977)	4.270932	583	4.358491
Schindler's List (1993)	4.291667	298	4.466443

Case 2: $m = 2$ & $C = 6$

title	count	bayes	mean
One Flew Over the Cuckoo's Nest (1975)	264	4.244526	4.291667
The Godfather (1972)	413	4.252955	4.283293
The Silence of the Lambs (1991)	390	4.257500	4.289744
Star Wars (1977)	583	4.335582	4.358491
The Usual Suspects (1995)	267	4.335740	4.385768
The Wrong Trousers (1993)	118	4.351562	4.466102
A Close Shave (1995)	112	4.368852	4.491071
The Shawshank Redemption (1994)	283	4.395904	4.445230
Casablanca (1942)	243	4.399209	4.456790
Schindler's List (1993)	298	4.418831	4.466443

Inter-Items: **Pseudo** Bayesian Average Rating

$$\bar{m}_i = \frac{C_i \cdot m_i + \sum(\text{ratings})}{C_i + N}$$

- $\bar{m}_i$: bayesian average rating for item i
- N: the number of reviews for all items
- m_i: average of review scores for item i
- C_i: the number of reviews for item i

Example: Bayesian Changes Order

$N = 10+15+228+150+124 = 527$

$\Sigma(\text{ratings}) = 10\times4.920 + 15\times4.667 + 228\times4.535 + 150\times4.310 + 124\times4.298 = 2332.67$

MacBook	No. Ratings	Ave. Rating	Rank	Bayesian Rating	Bayesian Rank
MB991LL	10	4.920	1	4.436	2
MB403LL	15	4.667	2	4.433	3
MB402LL	228	4.535	3	4.459	1
MC204LL	150	4.310	4	4.401	5
MB061LL	124	4.298	5	4.402	4

$$\overline{m}_1 = \frac{C_1 \cdot m_1 + \Sigma\text{ratings}}{C_1 + N} = \frac{10\times4.920 + 2332.67}{10+527}$$

$= 4.436$

Reverse Engineering Amazon

- Bayesian adjustment
- Recency of view
- Reputation score

Key Factors

- Bayesian ranking
- Too few or too outdated reviews penalized
- Very high quality reviews help a lot

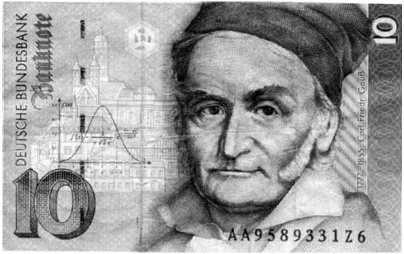
Summary

- Average ratings scalarize a vector and ranks
- Number of ratings should matter, Bayesian ranking does that
- Other statistical methods help too

Outline

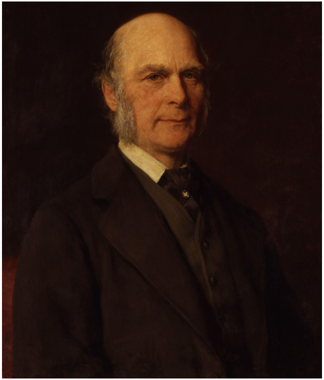
Classical Statistics

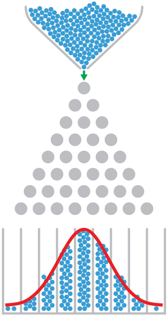
- 1800s: Linear Statistical Model and the method of least squares for estimation is often credited to Gauss (1777-1855) (1809), Adrien-Marie Legendre (1752-1833) (1805), Robert Adrain (1775-1843).
- Gauss also showed the optimality of the least-square approach (Gauss-Markov Theorem, 1823).



Classical Statistics

- 1888: Sir Francis Galton proposed the concept of correlation
- 1889: Sir Francis Galton proposed the concept of regression
- 1889: Sir Francis Galton proposed the Galton Board





Classical Statistics

- Carl
Karl Pearson (1857-1936) is credited for the establishment of the discipline of statistics. He contributed to theory of linear regression, correlation, Pearson curve, chi-square test, and the method of moments for estimation.

Modern Statistics: Frequency Perspective

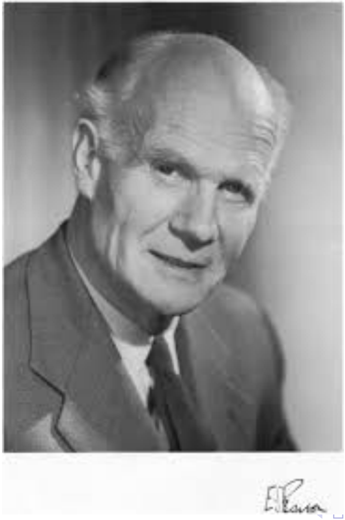
- 1908: William Gosset (Student) (1876-1937) proposed Student t-distribution and t-test statistics
- Precursor of small-sample statistics and hypothesis testing.

Modern Statistics: Frequency Perspective

- 1912-1922: Sir Ronald Aylmer Fisher (1890-1962) developed the notion of maximum likelihood estimator.
- He also worked on the analysis of variance (ANOVA), F-distribution, Fisher information and design of experiment.
- Co-founder of Modern Statistics (Mathematical Statistics or Statistical Inference)

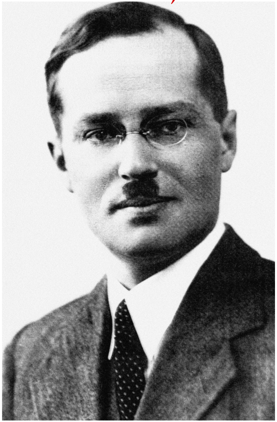
Modern Statistics: Frequency Perspective

- Egon Sharpe Pearson (1895-1980): co-founder of Neyman-Pearson Theory for hypothesis testing.



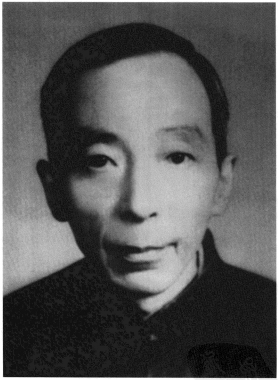
Modern Statistics: Frequency Perspective

- Jerzy Neyman (1894-1981): Co-founder of Modern Statistics (Mathematical Statistics or Statistical Inference)
- 1928-1938: Theoretical foundations of testing hypothesis, point estimation, confidence interval and survey sampling.



Modern Statistics: Frequency Perspective

- 1940s: Pao-Lu Hsu (1910-1970) obtained several exact or asymptotic distributions of important statistics in the theory of multivariate analysis.



Modern Statistics: Bayesian Perspective

- 1937: Bruno de Finetti proposed a predictive inference approach to statistics, emphasizing the prediction of future observations based on past observations.
- 1939: Harold Jeffreys applied Bayesian analysis for geophysics data.
- 1941-1944: Alan Turing applied Bayesian analysis for breaking the German code (Enigma)
- 1954s:Jimmie Savage proposed Bayesian statistics systematically
- 1950s: Bayesian econometrics originated from Harvard business school prevailed in economics society
- 1950s-1988: Efficient Monte carlo methods such as Metropolis and Gibbs sampling appeared.
- 1990-present: Bayesian statistics become the focus of mathematical statistics

References

- Chapters 9 of **BH**
- Chapters 4 & 6 & 8of **BT**