Chapter 6
Principle of Data Deduction
Overview

Outline

Introductory Statistics

statistics	probability
universe of information	sample space - Ω and probability - P
$\Downarrow$	$\Downarrow$
ask a question and collect data	define a random variable X
$\Downarrow$	$\Downarrow$
organize into the empirical cumulative distribution function	organize into the cumulative distribution function
$\Downarrow$	$\Downarrow$
compute sample means and variances	compute distributional means and variances

Inferential Statistics

In many cases, statistical inference uses the following structure:

We observe a realization of random variables on a probability space Ω,

$$X(\omega) = (X_1(\omega), , X_n(\omega))$$

where each of the X_i has the same distribution, F.

- The random variable may be independent in the case of sampling with replacement or
- more generally exchangeable in the case of sampling without replacement from a finite population.

The aim of the inference is to say something about which distribution it is.

Inferential Statistics

We often restrict ourselves to statistical models based on limiting considerations to distributions from some class $\mathcal{F}$.

- If $\mathcal{F}$ can be indexed by a set $\Theta \subset \mathbb{R}^d$, then $\mathcal{F}$ is called a parametric model. We generally set up this indexing so that the parameterization is identifiable, i.e., the mapping from $\Theta \to \mathcal{F}$ is one to one.
  - For a parameter choice $\theta = (\theta_1, \ldots, \theta_d)$, we denote the probability of the observations by P_θ and the expectation by E_θ.
  - If only a subset of the parameters are the subject of inference, then the rest of the parameters are called nuisance parameters.
- Those situations where the class $\mathcal{F}$ cannot be indexed is called a nonparametric model. For example, $\mathcal{F}$, the collection of distributions having a continuous density on $[0, 1]$ is a nonparametric model.

Inferential Statistics

Often, the distributions of X have a density $\mathbf{f}_{X|\Theta}(\mathbf{x}|\theta)$ with respect to some reference measure ν on the state space $\mathcal{X}$ for each value of the parameter θ,

$$P_\theta\{X \in B\} = \int_B \mathbf{f}_{X|\Theta}(\mathbf{x}|\theta)\ \nu(dx).$$

The choice for the measure ν depends on the state space $\mathcal{X}$ for X.

The most typical choices are

- X is a subset of $\mathcal{X} \subset \mathbb{R}^k$ and ν is Lebesgue measure, then
$$\int_B \mathbf{f}_{X|\Theta}(\mathbf{x}|\theta)\nu(dx) = \int_B \mathbf{f}_{X|\Theta}(\mathbf{x}|\theta)dx.$$
- X is a subset of $\mathcal{X} \subset \mathbb{Z}^k$ and ν is counting measure, then
$$\int_B \mathbf{f}_{X|\Theta}(\mathbf{x}|\theta)\nu(dx) = \sum_{x \in B} \mathbf{f}_{X|\Theta}(\mathbf{x}|\theta).$$

Densities and Likelihoods

Our analysis is based on the distribution of the random variables that underlie the data under any value θ. For each $\theta \in \Theta$, we have a density function

$$\mathbf{f}_X(\mathbf{x}|\theta).$$

For experimental designs based on a simple random sample, the observations $X_1, \ldots, X_n$, are drawn from a family of distributions each having density $f_X(x|\theta)$. For independent random variables, the joint density is the product of the marginal densities

$$\mathbf{f}_X(\mathbf{x}|\theta) = \prod_{k=1}^{n} f_X(x_k|\theta) = f_X(x_1|\theta) f_X(x_2|\theta) \cdots f_X(x_n|\theta).$$

In this circumstance, the data $\mathbf{x}$ are known and the parameter θ is unknown. Thus, we write the density function as

$$L(\theta|\mathbf{x}) = \mathbf{f}_X(\mathbf{x}|\theta)$$

and call L the likelihood function.

Densities and Likelihoods

- For Bernoulli trials with a known number of trials n but unknown success probability parameter p has joint density

$$\begin{aligned}\mathbf{f}_X(\mathbf{x}|p) &= p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}\cdots p^{x_n}(1-p)^{1-x_n} = p^{\sum_{k=1}^n x_k}(1-p)^{\sum_{k=1}^n (1-x_k)} \\ &= p^{\sum_{k=1}^n x_k}(1-p)^{n-\sum_{k=1}^n x_k} = p^{n\bar{x}}(1-p)^{n(1-\bar{x})}\end{aligned}$$

- Normal random variables unknown mean μ and standard deviation σ has joint density

$$\begin{aligned}\mathbf{f}_X(\mathbf{x}|\mu,\sigma) &= \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x_1-\mu)^2}{2\sigma^2}\right)\cdots\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^n (x_k-\mu)^2\right)\end{aligned}$$

Exercise. Find the joint density of n independent $\Gamma(\alpha,\beta)$ random variables.

Exponential Families

Recall that family of continuous random variables is called an exponential family if the probability density functions can be expressed in the form

$$f_X(x|\eta) = h(x)\exp\left(\sum_{i=1}^{k}\eta_i t_i(x) - A(\eta)\right), \quad x \in \mathcal{X}.$$

Thus, $\mathcal{X}$ is the common domain of the $f_X(x|\eta)$.

- $h(x)$ is a non-negative function.
  - The definition of an exponential family does not depend on the reference measure ν. If another measure $k\nu$ is chosen then $h(x)$ is replace by $h(x)/k(x)$.
- $t_i(x)$ are real-valued functions on the state space.
- $\eta = (\eta_1, \ldots, \eta_k)$ is called the natural parameter.

Exponential Families

Also,

$$\frac{\partial A}{\partial \eta_j}(\eta) = Et_j(X) \qquad \frac{\partial^2 A}{\partial \eta_j \partial \eta_k}(\eta) = \mathrm{Cov}(t_j(X), t_k(X)).$$

The likelihood for a simple random sample $X_1, \ldots, X_n$,

$$\begin{aligned}
\mathbf{f}_X(\mathbf{x}|\eta) &= \prod_{j=1}^{n} h(x_j) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x_j) - A(\eta)\right) \\
&= \prod_{j=1}^{n} h(x_i) \cdot \exp\left(\sum_{j=1}^{n}\sum_{i=1}^{k} \eta_i t_i(x_j)\right) e^{-nA(\eta)} \\
&= \prod_{j=1}^{n} h(x_i) \cdot \exp\left(\sum_{j=1}^{n} \langle \eta, t(x_j)\rangle\right) e^{-nA(\eta)}
\end{aligned}$$

Parameter Estimation

In the simplest possible terms, the goal of estimation theory is to answer the question:

What is that number?

Statistics has provided two distinct approaches this question - typically called

- classical or frequentist, and
- Bayesian.

Definition. A statistic is a function of the data that does not depend on any unknown parameter.

Exercise. Give a listing of statistics seen to this point.

Parameter Estimation

For parameter estimation, we consider $X = (X_1, \ldots, X_n)$, independent random variables chosen according to one of a family of probabilities P_θ where θ is element from the parameter space Θ. Based on our analysis, we choose an estimator $\hat{\theta}(X)$. If the data $\mathbf{x}$ takes on the values $x_1, x_2, \ldots, x_n$, then

$$\hat{\theta}(x_1, x_2, \ldots, x_n)$$

is called the estimate of θ. Thus we have three closely related objects.

1. θ - the parameter, an element of the parameter space, is a number or a vector.
2. $\hat{\theta}(x_1, x_2, \ldots, x_n)$ - the estimate, is a number or a vector obtained by evaluating the estimator on the data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.
3. $\hat{\theta}(X_1, \ldots, X_n)$ - the estimator, is a random variable. We will analyze the distribution of this random variable to decide how well it performs in estimating θ.

Coin Tosses

For Bernoulli trials $X = (X_1, \ldots, X_n)$, we have

1. p, a single parameter, the probability of success, with parameter space $[0, 1]$.
2. $\hat{p}(x_1, \ldots, x_n)$ is the sample proportion of successes in the data and can be used to make a point estimate for p.
3. $\hat{p}(X_1, \ldots, X_n)$, the sample mean of the random variables

$$\hat{p}(X_1, \ldots, X_n) = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n} S_n$$

is an estimator of p. We can give the distribution of this estimator because S_n is a binomial random variable.

Classical Statistics

In classical statistics, the state of nature is assumed to be fixed, but unknown to us. Thus, one goal of estimation is to determine which of the P_θ is the source of the data. The estimate is a statistic

$$\hat{\theta} : \text{data} \to \Theta.$$

For estimation procedures, the classical approach to statistics is based on two fundamental questions:

- How do we determine estimators?
- How do we evaluate estimators?
  - Does this estimator in any way systematically under or over estimate the parameter?
  - Does it has large or small variance?
  - How does it compare to a notion of best possible estimator?
  - How easy is it to determine and to compute?
  - How does the procedure improve with increased sample size?

Bayesian Statistics

In Bayesian statistics, (X, Ψ) is a random variable on the cross product of the state space and the parameter space. The density π of Ψ on the parameter space Θ is called the prior density. Thus, the prior density and the family $\{P_\theta; \theta \in \Theta\}$ determine the joint distribution of (X, Ψ).

In this approach, both the parameter and the data are modeled as random. Inference is based on posterior density derived from Bayes formula

$$f_{\Theta|X}(\theta|\mathbf{x}) = \frac{f_{X,\Theta}(\mathbf{x}, \theta)}{f_X(\mathbf{x})} = \frac{f_X(\mathbf{x}|\theta)\pi(\theta)}{f_X(\mathbf{x})}.$$

where the denominator is the continuous mixture

$$f_X(\mathbf{x}) = \int_\Theta f_{X,\Theta}(\mathbf{x}, \theta)\ d\theta = \int_\Theta f_{X|\Theta}(\mathbf{x}|\theta)\pi(\theta)\ d\theta$$

Coin Tosses

We consider independent flips of a biased coin and use a Bayesian approach to make some inference for the probability of heads. We first set a prior distribution for $\tilde{P}$. The beta family $Beta(\alpha, \beta)$ of distributions takes values in the interval $[0, 1]$ and provides a convenient prior density π. Thus,

$$\pi(p) = c_{\alpha,\beta} p^{(\alpha-1)}(1-p)^{(\beta-1)}, \quad 0 < p < 1.$$

The $Beta(\alpha, \beta)$ distribution has

$$\text{mean } \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \text{variance } \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Coin Tosses

If we perform n Bernoulli trials $\mathbf{x} = (x_1, \ldots, x_n)$, then the joint density

$$\mathbf{f}_{X|\tilde{P}}(\mathbf{x}|p) = p^{\sum_{k=1}^n x_k}(1-p)^{n-\sum_{k=1}^n x_k}.$$

Thus the posterior distribution of the parameter $\tilde{P}$ given the data $\mathbf{x}$,

$$\begin{aligned} f_{\tilde{P}|X}(p|\mathbf{x}) \propto \mathbf{f}_{X|\tilde{P}}(\mathbf{x}|p)\pi(p) &= p^{\sum_{k=1}^n x_k}(1-p)^{n-\sum_{k=1}^n x_k} \cdot c_{\alpha,\beta} p^{(\alpha-1)}(1-p)^{(\beta-1)}. \\ &= c_{\alpha,\beta} p^{\alpha+\sum_{k=1}^n x_k - 1}(1-p)^{\beta+n-\sum_{k=1}^n x_k - 1}. \end{aligned}$$

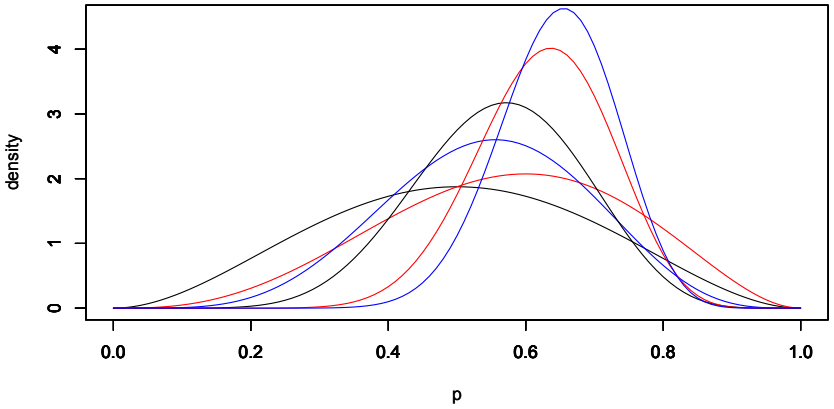
Consequently, the posterior distribution is also from the beta family with parameters

$$\alpha + \sum_{k=1}^n x_k \quad \text{and} \quad \beta + n - \sum_{k=1}^n x_k = \beta + \sum_{k=1}^n (1 - x_k).$$

$$\alpha + \#\text{ successes} \quad \text{and} \quad \beta + \#\text{ failures}.$$

Coin Tosses

Posterior densities based on $Beta(3, 3)$ prior.



Data : *H T H T H T H H T H H T H T H H H H T T H T H H H*

Coin Tosses

Notice that the posterior mean can be written as

$$\begin{aligned}\frac{\alpha+\sum_{k=1}^{n} x_k}{\alpha+\beta+n} &= \frac{\alpha}{\alpha+\beta+n}+\frac{\sum_{k=1}^{n} x_k}{\alpha+\beta+n}\\ &= \frac{\alpha}{\alpha+\beta}\cdot\frac{\alpha+\beta}{\alpha+\beta+n}+\frac{1}{n}\sum_{k=1}^{n} x_k\cdot\frac{n}{\alpha+\beta+n}\\ &= \frac{\alpha}{\alpha+\beta}\cdot\frac{\alpha+\beta}{\alpha+\beta+n}+\bar{x}\cdot\frac{n}{\alpha+\beta+n}.\end{aligned}$$

This expression allow us to see that the posterior mean can be expresses as a weighted average. The relative weights are

$\alpha+\beta$ from the prior and n, the number of observations.

Bayesian Estimation

This brings forward two central issues in the use of the Bayesian approach to estimation.

- The posterior mean, $E[P|X = x]$ can be used to give a point estimate for p.
- If the number of observations is small, then the estimate relies heavily on the quality of the choice of the prior distribution π. Thus, an unreliable choice for π leads to an unreliable estimate.
- As the number of observations increases, the estimate relies less and less on the prior distribution. In this circumstance, the prior may simply be playing the roll of a catalyst that allows the machinery of the Bayesian methodology to proceed.

Hypothesis Testing, Classical Statistics

Suppose we are interested in deciding if the parameter θ lies in one portion Θ_0 of the parameter space. We can then set a hypothesis

$$H_0 : \theta \in \Theta_0$$

versus the alternative hypothesis

$$H_1 : \theta \in \Theta_0^c$$

In classical statistics, a test of this hypothesis would be to choose a rejection region $\mathcal{R}$, and a decision function d of the data $\mathbf{x}$ and reject H_0 if $d(\mathbf{x}) \in \mathcal{R}$. The power function

$$\beta(\theta) = P_\theta\{d(X) \in \mathcal{R}\}$$

gives, for each value of the parameter θ, the probability that the hypothesis is rejected.

Classical Statistics

Example. Suppose that, under the probability P_θ , X consists of n independent $N(\theta, \sigma)$ random variables. Consider the two-sided test

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

To determine $\mathcal{R}$, note that the sample mean $\bar{X} \sim N(\theta, \sigma^2/n)$and let Z be a standard normal. Set $z_{\alpha/2}$ satisfy $\alpha = P\{Z > z_{\alpha/2}\}$, then

$$\mathcal{R} = \left\{ \mathbf{x}; \frac{|\bar{x} - \theta_0|}{\sigma/\sqrt{n}} > z_{\alpha/2} \right\}.$$

Exercise. $\beta(\theta_0) = \alpha$ and $\beta(\theta) > \alpha$, $\theta \neq \theta_0$

Hypothesis Testing,. Bayesian Statistics

Recall the example of a normal prior on Ψ of normal observations X. We take

- The prior density to be $N(\theta_1, 1/\lambda_0)$
- The observations $X_1, . , . , X_n$ are independent $N(\theta, \sigma^2)$
- Their mean $\bar{X} \sim N(\theta, \sigma^2/n)$
- The posterior distribution is $N(\theta_1(\bar{x}), \sigma^2/(n + \lambda_0\sigma^2))$ where

$$\theta_1(\bar{x}) = \frac{\lambda_0}{\lambda_0 + n/\sigma^2}\theta_1 + \frac{n/\sigma^2}{\lambda_0 + n/\sigma^2}\bar{x}.$$

We reject the null hypothesis if θ_0 falls too far away from the posterior mean.

Estimation

For classical statistics, we use $\bar{X}$ to estimate the true parameter value θ.

$$E_\theta \bar{X} = \theta, \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

The first is a statement that $\bar{X}$ is an unbiased estimator of θ.

For Bayesian statistics we use $\theta_1(\bar{X})$ to estimate θ.

$$\theta_1(\bar{X}) = \frac{\lambda_0}{\lambda_0 + n/\sigma^2}\theta_1 + \frac{n/\sigma^2}{\lambda_0 + n/\sigma^2}\bar{X}$$

is a convex combination of the prior mean θ_1 and the mean of the observations $\bar{X}$.
Note that the weights move towards $\bar{X}$ as n increases.