Statistics 201B:
Introduction of Statistics at an Advanced Level

Haiyan Huang <hhuang@stat.berkeley.edu>
Department of Statistics

UC Berkeley, Fall 2014

Syllabus

M/W/F 10-12p Lecture, 160 DWINELLE
Tu/Th 9:30am-11am (201 Discussion Section), 11-12:30pm (202 Discussion Section), 344 Evans

Instructor: Dr. Haiyan Huang
Office: 317 Evans
Office Hours: 1-3pm M
e-mail: hhuang@stat.berkeley.edu

GSI: Yongdong Liu
email: liuyd@berkeley.edu
Office Hours: Tu/Th 1:30-3:30pm

Syllabus

About This Course: This course is the second part of a sequence of courses on probability (201A) and statistics (201B) at the masters level. We will cover the fundamentals of statistical inference, testing, and modeling. Prerequisites: multivariable calculus; basic linear algebra (vectors and matrices); probability topics covered in 201A.

Textbooks

There is no required textbook for this course. But there are several recommended ones.

1. "All of Statistics" by Larry Wasserman. We'll cover the topics listed in Section II and parts of Section III. Even though we will follow the topics in this book, this book cannot serve as our textbook. The material content of the book is not detailed enough.
2. "Statistical Inference" by Casella and Berger. It is much more detailed and present "traditional" topics in more depth. But it lacks materials on modern topics.
3. "Theoretical Statistics: Topics for a Core Course" by Robert Keener. This book is detailed and has relatively more modern topics than the Casella and Berger book.

Homework and exam questions will be based on the lectures!

Grading

Your final grade will be a weighted average of your average homework score (20%), midterm (35%), and the final exam (45%).

- **Homeworks:** Problem sets will be assigned each Wednesday, for a total of 6 assignments. You should download the assignments from the bSpace page for this class (https://bspace.berkeley.edu/). Each problem set is to be turned in on Thursday a week later at the beginning of Discussion sessions. No late assignments will be accepted. The homework with lowest score will not be included in the final homework grade. Some problems may not be graded, and you should review the solutions carefully for those problems. Students can discuss homework assignments in groups of at most three. Each student must write up his/her own solutions individually, and must explicitly name any

collaborators at the top of the homework. Any evidence of cheating will be subject to disciplinary action.

- **Midterm:** The in-class midterm is tentatively scheduled on Wednesday November 12 (could be modified later based on the class pace). You are allowed to bring one double sided A4 page of handwritten notes to the exam.

- **Final:** The final exam will be cumulative. The exam is on Monday, December 15, 2014 (8-11am). You are allowed to bring two double sided A4 pages of handwritten notes to the exam.

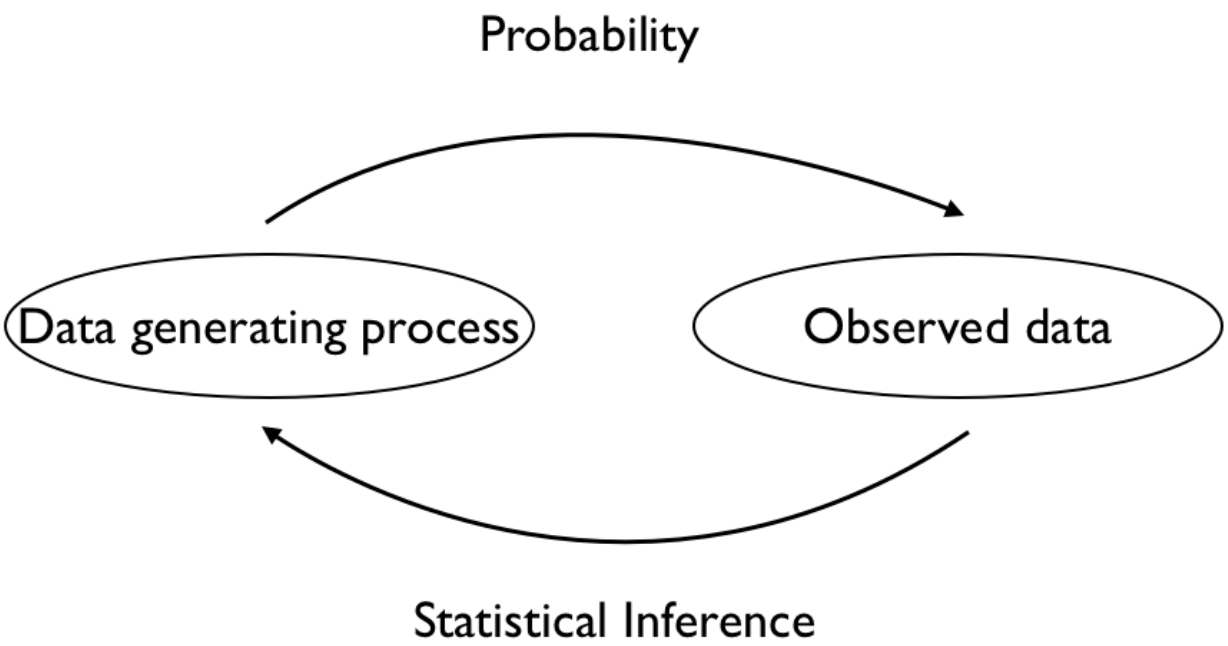
Other logistics

Computing: The assignments will involve some computing. We will use `R`, which is available for free at http://cran.r-project.org/. Your GSI will teach you everything you need to know about `R` for this course.

Email: I will not answer detailed homework questions through emails. Please bring questions to office hours.

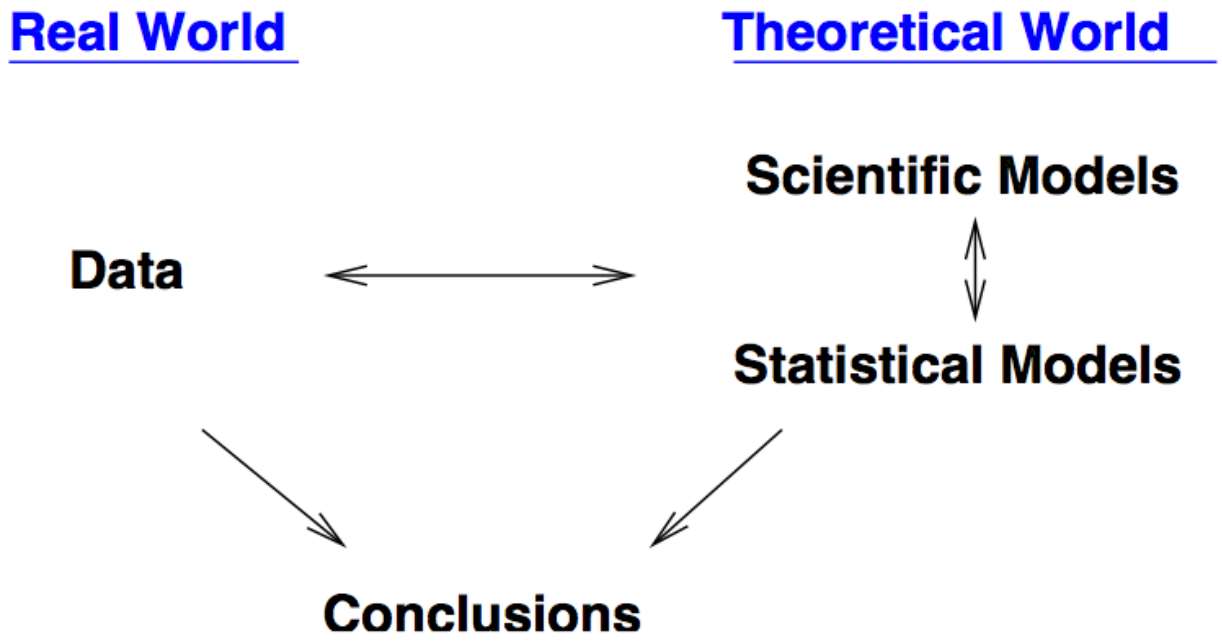
Students with disabilities: If you need accommodations for any disabilities, please speak to me after class or during office hours so that we can make the necessary arrangements.

The Big Picture



adapted from Wasserman, 2004

A Slightly Bigger Picture



from Kass, 2009

Course topics

- A brief review of probability
- Methods for statistical inference (estimation and testing)
  - Nonparametric inference
  - Parametric inference
    - Frequentist inference
    - Bayesian inference
  - Decision (e.g. estimation) evaluation; Statistical decision theory
- Statistical models/methods for prediction and classification

A detailed list of topics to be covered

- Introduction to inference
  - Bias, standard error, MSE of estimators
  - Consistent estimators
  - (asymptotic) Confidence intervals and coverage probability
- Empirical CDF and statistical functionals
  - Statistical properties of ECDF
  - Plug-in estimator
- The bootstrap
  - Monte Carlo integration
  - Importance Sampling

- Parametric inference
  - Sufficient, minimal sufficient statistics; Rao-Blackwell theorem
  - Method of moments estimator
  - Maximum likelihood estimator (MLE)
  - Finding the MLE when the parameter space is restricted
  - The equivariance property of the MLE
  - Exponential family
  - Fisher information
  - Delta method

- Hypothesis testing
  - The power function and size of a test
  - Type I and Type II error probabilities
  - Constructing a level α test, given a $1-\alpha$ confidence interval
  - Wald test

  - The likelihood ratio test statistic (T or λ); the limiting distribution of the LRT statistic
  - Calculating the exact or approximate p-value of a test

- Bayesian statistics

  - Frequentist and subjective interpretations of probability
  - Posterior distribution; conjugate prior; Jeffreys prior
  - Credible interval
  - Rejection sampling
  - Bayes factor

- Decision theory

  - Risk functions
  - Admissibility
  - Bayes rules, minimax actions

- Linear regression
  - Ordinary Least Square estimators in simple linear regression
  - Various sums of squares that arise in linear regression
  - Model selection techniques

- Density estimation
  - Histogram estimator and its statistical properties; the tradeoff between bias and variance occurs in a histogram as a function of bin width
  - Kernel density estimator (KDE); the tradeoff between bias and variance occurs in a KDE as a function of bandwidth

- Nonparametric regression (if time permits)
  - The k-nearest neighbors method of nonparametric regression; the Nadaraya-Watson kernel estimator

- Classification (if time permits)
  - Bayes classification rule in terms of the true, unknown functions
  - Quadratic or linear classification boundary

Non-Parametric Inference

October 21, 2014

Plug–in (or Substitution) Principle: a non-parametric estimation method

Let $X_1, ..., X_n \overset{iid}{\sim} F$, where F can be parametric or nonparametric. Assume that we are interested in estimating the quantities that are related to F, such as the mean, median, variance, quantities, etc, by a nonparametric way. No matter F is parametric or non-parametric, we can write the quantities of interest as a function of F, $\theta(F)$.

The substitution (plug–in) method is to estimate $\theta(F)$ with $\theta(\hat{F}_n)$, where $\hat{F}_n$ is the empirical distribution of F.

Note: We also use F to denote the CDF of the distribution.

The Empirical CDF and Statistical Functionals

Let $X_1, \ldots, X_n \overset{iid}{\sim} F$. The empirical CDF $\hat{F}_n$ puts mass $1/n$ at each datapoint.

$$\begin{aligned} \hat{F}_n(x) &= \frac{\sum_{i=1}^n I(X_i \leq x)}{n} \\ &= \#\{X_i \leq x\}/n \end{aligned}$$

Note that $P_{\hat{F}_n}(X \leq x) = \hat{F}_n(x)$ is often different from $P_F(X \leq x) = F(x)$.

It's also helpful to note that $Y_i = I(X_i \leq x), i = 1, \ldots, n$ are iid Bernoulli r.v.'s, with

$$p = P(Y_i = 1) = P(X_i \leq x) = F(x)$$

Examples of Plug–in Estimators

- $\theta(F) = \mathrm{E}_F(X)$. Then plug–in estimator will be

$$\begin{aligned} \theta(\hat{F}_n) &= \mathrm{E}_{\hat{F}_n}(X) = \textstyle\sum_t t P_{\hat{F}_n}(X = t) \\ &= \textstyle\sum_t t \cdot \frac{\sum_{i=1}^n I(X_i = t)}{n} \\ &= \frac{\sum_{i=1}^n X_i}{n} \\ &= \bar{X}_n \end{aligned} \tag{1}$$

(This result is independent of the distribution of X.)

- $\theta(F) = var_F(x)$. Then the plug–in estimate will be

$$\begin{aligned} \theta(\hat{F}_n) &= var_{\hat{F}_n}(X) = \mathrm{E}_{\hat{F}_n}(X^2) - (\mathrm{E}_{\hat{F}_n}(X))^2 \\ &= \frac{\sum_{i=1}^n X_i^2}{n} - (\frac{\sum_{i=1}^n X_i}{n})^2 \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \end{aligned} \tag{2}$$

- $\theta(F) = median(X) = \inf_t\{t|F(t) \geq \frac{1}{2}\}$. Then the plug–in estimate for the median is $\theta(\hat{F}_n) = \inf_t\{t|\hat{F}_n(t) \geq \frac{1}{2}\}$.

Note: In general, how plug-in estimator works depending on the properties of $\hat{F}_n$ and also the property of θ function.

Properties of the Empirical CDF

For any fixed x,

$$\begin{aligned} E[\hat{F}_n(x)] &= F(x) \\ V[\hat{F}_n(x)] &= \frac{F(x)[1-F(x)]}{n} \\ MSE[\hat{F}_n(x)] &= V[\hat{F}_n(x)] \to 0 \\ \hat{F}_n(x) &\xrightarrow{P} F(x) \end{aligned}$$

The Glivenko-Cantelli Theorem is even stronger, giving uniform convergence almost surely:
Let $X_1, \ldots X_n \overset{iid}{\sim} F$. Then

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{as} 0$$

A theorem on Plug–in estimator

Theorem. Suppose the function $\theta(F)$ is continuous in the sup-norm:

$\forall \epsilon > 0, \exists \delta > 0$ such that "$||G - F||_\infty < \delta$ implies $\theta(G) - \theta(F) < \epsilon$".

[That is for any ϵ, if there is some G close enough to F, then $\theta(G)$ is close to $\theta(F)$.]

Then,

$$\theta(\hat{F}_n) \xrightarrow{P} \theta(F).$$

Linear function of F

A statistical functional $T(F)$ (or $\theta(F)$) is any function of F. Some examples are the mean $\int x dF(x)$, variance $\int x^2 dF(x) - \left(\int x dF(x)\right)^2$, and p^{th} quantile

$$F^{-1}(p) = \inf\{x : F(x) \geq p\}$$

A linear functional can be written as $T(F) = \int r(x) dF(x)$. The mean is a linear functional, but the variance and quantile function are not.

The plug-in estimator of $T(F)$ is just $T(\hat{F}_n)$. When T is a linear functional,

$$T(\hat{F}_n) = \int r(x) d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} r(X_i)$$

.

Let $X_1, \ldots, X_n \overset{iid}{\sim} F$. Find the plug-in estimators for

- the expected value of X_1
- the expected value of $\exp(X_1)$
- the variance of X_1
- the median of F

Confidence Interval of the Empirical CDF

Dvoretzky-Kiefer-Wolfowitz Inequality: Let $X_1, \ldots, X_n \overset{iid}{\sim} F$. For any $\epsilon > 0$,

$$P\left(\sup_x |F(x) - \hat{F}_n(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

It follows that the functions

$$\begin{aligned} L(x) &= \max\{\hat{F}_n(x) - \epsilon_n, 0\} \\ U(x) &= \min\{\hat{F}_n(x) + \epsilon_n, 1\} \\ &\quad \text{for } \epsilon_n = \sqrt{\log(2/\alpha)/(2n)} \end{aligned}$$

form a global $1 - \alpha$ confidence band for F. That is,

$$P\left(L(x) \leq F(x) \leq U(x) \text{ for all } x\right) \geq 1 - \alpha$$

Often we have $T(\hat{F}_n) \approx N(T(F), \widehat{se}^2)$, which allows us to form an approximate $1-\alpha$ confidence interval for $T(F)$ of

$$T(\hat{F}_n) \pm z_{\alpha/2}\widehat{se}$$

Example: Verify that the `R` expression

```
mean(x) + c(-2, 2) * sd(x)/sqrt(length(x))
```

produces an approximate 95% confidence interval for the mean waiting time for Old Faithful Geyser Data (built-in data in R).

The Bootstrap

The bootstrap is a computer-intensive method for estimating measures of uncertainty in problems for which no analytical solution is available.

There are technically two classes of bootstrap methods: parametric and nonparametric.

The nonparametric bootstrap uses two main ideas:

- Monte Carlo integration
- The empirical CDF

Monte Carlo integration is based on the following approximation:

$$\begin{aligned} E[h(Y)] &= \int h(y) dF_Y(y) \\ &\approx \frac{1}{B} \sum_{j=1}^{B} h(Y_j) \end{aligned}$$

where $Y_1, \ldots, Y_B \overset{iid}{\sim} F_Y$. Note that if $E[|h(Y)|] < \infty$,

$$\frac{1}{B} \sum_{j=1}^{B} h(Y_j) \overset{as}{\to} E[h(Y)]$$

as $B \to \infty$. Typically we have control over B, so we can make the approximation arbitrarily good.

A simple example: Use Monte Carlo integration to approximate

$$\int_{-\infty}^{\infty} \sin^2(x) e^{-x^2} dx$$

Solution: We can write this as $\sqrt{\pi} \int_{-\infty}^{\infty} \sin^2(x) f(x) dx$, where $f(x)$ is the PDF of a $N(0, 1/2)$ r.v. Therefore, we can

1. Draw $Y_1, \ldots, Y_B \overset{iid}{\sim} N(0, 1/2)$.

   ```
   > B <- 10000; y <- 1/sqrt(2) * rnorm(B)
   ```

2. Approximate $\sqrt{\pi} \int_{-\infty}^{\infty} \sin^2(x) f(x) dx \approx \frac{\sqrt{\pi}}{B} \sum_{j=1}^{B} \sin^2(Y_j)$.

   ```
   > sqrt(pi) * mean(sin(y)^2)
   [1] 0.5509956
   ```

Importance sampling is an adaptation to the usual Monte Carlo integration that allows us to sample from an "importance function" g rather than the target density h. Note that

$$\begin{aligned} E_h[q(\theta)] &= \int q(\theta)h(\theta)d\theta \\ &= \int q(\theta)\frac{h(\theta)}{g(\theta)}g(\theta)d\theta \\ &\approx \frac{1}{B}\sum_{i=1}^{B} q(\theta_i)\frac{h(\theta_i)}{g(\theta_i)} \end{aligned}$$

where $\theta_1, \ldots, \theta_B \overset{iid}{\sim} g(\theta)$.

A more complicated example: Use Monte Carlo integration to approximate $V_\lambda[median(X_1, \ldots, X_n)]$ when $X_1, \ldots, X_n \overset{iid}{\sim} Exp(\lambda)$.

This is more complicated in two ways:

1. Unlike an analytical calculation, on the computer we need particular values of n and λ. To see how $V_\lambda[median(X_1, \ldots, X_n)]$ changes with n and λ, we need to use Monte Carlo integration many times for different combinations.

2. For each combination, we need to sample B times from the *sampling distribution* of $median(X_1, \ldots, X_n)$. That is, for each $j = 1, \ldots, B$, we need to sample $X_1, \ldots, X_n \overset{iid}{\sim} Exp(\lambda)$ and calculate the median. Don't confuse n and B: n is the sample size, while B is the number of MC samples.

One combination: Let $n = 10$ and $\lambda = 5$. Then

- Draw $Y_1, \ldots, Y_B \overset{iid}{\sim} F_Y$, where F_Y is the CDF of $median(X_1, \ldots, X_n)$.

```
> n <- 10; lambda <- 5; B <- 10000
> samples <- matrix(rexp(n*B, rate = 1/lambda),
+    nrow = B, ncol = n)
> y <- apply(samples, MARGIN = 1, FUN = median)
```

- Approximate $V_\lambda[median(X_1, \ldots, X_n)] \approx \frac{1}{B}\sum_{j=1}^{B}(Y_j - \bar{Y})^2$.

```
> var(y)
[1] 2.402400
```

Back to the bootstrap...

Suppose we have data $X_1, \ldots, X_n$ and we compute statistic $T_n = g(X_1, \ldots, X_n)$.

It's not always possible to calculate $V_F[T_n]$ analytically, which is where the bootstrap comes in.

If we knew F, we could use MC integration to approximate $V_F[T_n]$. However, we don't in practice, so we make an initial approximation of F with the empirical CDF $\hat{F}_n$.

$$V_F[T_n] \quad \underset{\text{ECDF; depends on } n}{\approx} \quad V_{\hat{F}_n}(T_n) \quad \underset{\text{MC integration; depends on } B}{\approx} \quad \widehat{V}_{\hat{F}_n}(T_n)$$

Sampling from $\hat{F}_n$ is easy: just draw one observation at random from $X_1, \ldots, X_n$. Repeated sampling is "with replacement."

The algorithm:

1. Repeat the following B times to obtain $T^*_{n,1}, \ldots, T^*_{n,B}$, an iid sample from the sampling distribution for T_n implied by $\hat{F}_n$.
   (a) Draw $X^*_1, \ldots, X^*_n \sim \hat{F}_n$.
   (b) Compute $T^*_n = g(X^*_1, \ldots, X^*_n)$.

2. Use this sample to approximate $V_{\hat{F}_n}(T_n)$ by MC integration. That is, let

$$v_{boot} = \widehat{V}_{\hat{F}_n}(T_n) = \frac{1}{B}\sum_{j=1}^{B}\left(T^*_{n,j} - \frac{1}{B}\sum_{k=1}^{B} T^*_{n,k}\right)^2$$

Confidence intervals can also be constructed from the bootstrap samples.

Method 1: Normal-based interval

$$C_n = T_n \pm z_{\alpha/2} \widehat{se}_{boot}$$

where $\widehat{se}_{boot} = \sqrt{v_{boot}}$; this only works well if the distribution of T_n is close to Normal. Note that asymptotic normality of T_n is a property involving n, not B.

Method 2: Quantile intervals

$$C_n = \left(T^*_{\alpha/2}, T^*_{1-\alpha/2} \right)$$

where T^*_β is the β quantile of the bootstrap sample $T^*_{n,1}, \ldots, T^*_{n,B}$.

Bootstrapping method for estimating bias

$X_1, ..., X_n \sim F_0$. Let F_1 be the corresponding empirical CDF (i.e., $\hat{F}_n$). Then $\theta(F_1)$ is an empirical Plug–In estimator of $\theta(F_0)$. How to estimate the following bias?

$$t_0 = E_{F_0}(\theta(F_0) - \theta(F_1))$$

Answer: We draw a sample $Y_1, ..., Y_n$ from F_1 and derive the empirical CDF F_2. We can estimate t_0 by

$$\hat{t}_0 = E_{F_1}(\theta(F_1) - \theta(F_2))$$

Example (Bias correction). We want to estimate $\theta(F_0) = (E_{F_0}X)^2 = \mu^2$, where X follows F_0 with mean μ and variance σ^2. The EPI estimator is $\theta(F_1) = (E_{F_1}Y)^2 = \bar{X}^2$, where Y follows F_1. The bias is

$$t_0 = E_{F_0}(\theta(F_0) - \theta(F_1)) = \theta(F_0) - E_{F_0}[\theta(F_1)] = -\sigma^2/n.$$

Now we consider the estimator

$$\tilde{\theta} = \theta(F_1) + \hat{t}_0 = \theta(F_1) + [\theta(F_1) - E_{F_1}[\theta(F_2)]]$$

Note that $Y_1, ..., Y_n \sim F_1$ with mean $\bar{X}$ and variance $\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n}$, and $\theta(F_2) = (E_{F_2}[Z])^2 = (\bar{Y})^2$, where Z follows F_2.

$$E_{F_1}[\theta(F_2)] = E_{F_1}[(\bar{Y})^2] = (\bar{X})^2 + \frac{1}{n}\left(\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n}\right)$$

Then for the corrected estimator

$$\tilde{\theta} = \theta(F_1) + \hat{t}_0 = \theta(F_1) + [\theta(F_1) - E_{F_1}[\theta(F_2)]] = (\bar{X})^2 - \frac{1}{n}\left(\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n}\right),$$

We have

$$\begin{aligned} E_{F_0}(\tilde{\theta}) &= \left(\mu^2 + \frac{\sigma^2}{n}\right) - E_{F_0}\left[\sum_{i=1} \frac{(X_i - \bar{X}^2)}{n^2}\right] \\ &= \mu^2 + \frac{\sigma^2}{n} - \frac{(n-1)\sigma^2}{n^2} \\ &= \mu^2 + \frac{\sigma^2}{n^2} \end{aligned}$$

Note that $E_{F_0}[\theta(F_1)] = E_{F_0}[\bar{X}^2] = \mu^2 + \sigma^2/n$.

Pivotal intervals

In parametric statistics, a pivot is a function $R(X_1, \ldots, X_n, \theta)$ whose distribution doesn't depend on θ. This is useful because we can construct a confidence interval for $R_n = R(X_1, \ldots, X_n, \theta)$ without knowing θ and then manipulate it to construct a confidence interval for θ.

In nonparametric statistics, we typically can't find a quantity that is exactly pivotal, i.e., whose distribution doesn't depend on the unknown F.

If $\theta = T(F)$ is a location parameter, then $R_n = \hat{\theta}_n - \theta$ is approximately pivotal. If we knew the CDF H of R_n, we could construct an exact $1 - \alpha$ confidence interval for θ of (a, b), where

$$\begin{aligned} a &= \hat{\theta}_n - H^{-1}(1 - \alpha/2) \\ b &= \hat{\theta}_n - H^{-1}(\alpha/2) \end{aligned}$$

Since we don't know H, we estimate it using the bootstrap samples.

$$\hat{H}(r) = \frac{1}{B}\sum_{j=1}^{B} I(R^*_{n,j} \leq r)$$

where $R^*_{n,j} = \hat{\theta}^*_{n,j} - \hat{\theta}_n$. In other words, we form the empirical CDF of H using the bootstrap samples of the pivot. Therefore, the plug-in estimates of $H^{-1}(1-\alpha/2)$ and $H^{-1}(\alpha/2)$ are just the $1-\alpha/2$ and $\alpha/2$ sample quantiles of these samples.

This gives a $1-\alpha$ bootstrap pivotal interval of

$$C_n = \left(2\hat{\theta}_n - \hat{\theta}^*_{1-\alpha/2}, 2\hat{\theta}_n - \hat{\theta}^*_{\alpha/2}\right)$$

Location and Scale Family

Let Y be a random variable with distribution F. Let F_μ be the distribution function of $Y+\mu$; F_σ be the distribution of σY; and $F_{\mu,\sigma}$ be the distribution of $\sigma Y + \mu$. Then,

1. The family $\{F_\mu : -\infty < \mu < \infty\}$ is called a location family (e.g. $\mathcal{N}(\mu, 1)$). μ is the location parameter.

2. The family $\{F_\sigma : \sigma > 0\}$ is called a scale family (e.g. $\mathcal{N}(0, \sigma^2)$). σ^2 is the scale parameter.

3. The family $\{F_{\sigma,\mu} : -\infty < \mu < \infty, \sigma > 0\}$ is called a location scale family (e.g. $\mathcal{N}(\mu, \sigma^2)$).

Without loss of generality we assume $E[Y] = 0$ and $Var(Y) = 1$.

Parametric Inference

A parametric model has the form

$$\mathcal{F} = \{F(x;\theta) : \theta \in \Theta\}$$

where $\Theta \subseteq \mathbb{R}^k$ is the parameter space.

We typically choose a class $\mathcal{F}$ based on knowledge about the particular problem. We might say we're making certain assumptions about the data generating mechanism. It's good practice when using a parametric model to look for violations of these assumptions.

We'll begin with two methods for constructing estimators of θ: the method of moments and maximum likelihood estimation.

Suppose $\theta = (\theta_1, \ldots, \theta_k)$. For $j = 1, \ldots, k$, define the j^{th} moment

$$\alpha_j \equiv \alpha_j(\theta) = E_\theta[X^j] = \int x^j dF_\theta(x)$$

and the j^{th} sample moment $\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^n X_i^j$.

The method of moments estimator $\hat{\theta}_n$ is defined to be the value of θ s.t.

$$\begin{aligned} \alpha_1(\hat{\theta}_n) &= \hat{\alpha}_1 \\ \alpha_2(\hat{\theta}_n) &= \hat{\alpha}_2 \\ \vdots \quad &\vdots \quad \vdots \\ \alpha_k(\hat{\theta}_n) &= \hat{\alpha}_k \end{aligned}$$

MOM generalization: Instead of using $\alpha_j(\theta) = E_\theta[X^j]$, we can consider $\alpha_j(\theta) = E_\theta[g(X)^j]$ and find $\hat{\theta}_n$ s.t. $\alpha_j(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^n g(X_i)^j$, $j = 1, ..., n$.

The maximum likelihood estimator (MLE) is obtained by maximizing the likelihood function

$$\begin{aligned}\mathcal{L}_n(\theta) &= f(X_1, \ldots, X_n; \theta) \\ &= \prod_{i=1}^{n} f(X_i; \theta) \quad \text{if the data are independent}\end{aligned}$$

That is, the likelihood is just the joint density of the data, but viewed as a function of θ.

It's often easier to work with the log-likelihood function

$$\ell_n(\theta) = \log \mathcal{L}_n(\theta) = \sum_{i=1}^{n} \log f(X_i; \theta)$$

If the log-likelihood is differentiable with respect to θ, possible candidates for the MLE are those in the interior of Θ that solve

$$\frac{\partial}{\partial \theta_j} \ell_n(\theta) = 0, \quad j = 1, \ldots, k$$

We still need to check that we've found the global maximum. Also note that if the maximum occurs on the boundary of Θ, the first derivative may not be zero.

It's not always possible to maximize the likelihood analytically, and in these cases we turn to numerical maximization methods.

Examples:

- Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$. Find the MLE for θ.
- Now solve the same problem, but with the restriction $\Theta = [0, \infty)$.
- Let $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0, \theta)$. Find the MLE for θ.
- Let $X_1, \ldots, X_n \overset{iid}{\sim} Unif(\theta, \theta + 1)$. Find the MLE for θ.
- Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$. Show that MOM and MLE are equivalent for this distribution family. Can this result be generalized?

Some properties of the MLE that we will explore are:

1. Equivariance: If $\hat{\theta}_n$ is the MLE of θ, then $g(\hat{\theta}_n)$ is the MLE of $g(\theta)$.

2. Consistency: $\hat{\theta}_n \xrightarrow{P} \theta^*$, where θ^* is the true value of the parameter.

3. Asymptotic normality: $(\hat{\theta}_n - \theta^*)/se(\hat{\theta}_n) \xrightarrow{D} N(0,1)$.

4. Asymptotic efficiency: The MLE has the smallest asymptotic variance among asymptotically normal estimators.

Conditions for the last three can be somewhat technical, so we'll start with the case that $\theta \in \Theta \subseteq \mathbb{R}$ and will focus more on intuition than on details.

Equivariance: Let $\tau = g(\theta)$ be a function of θ. Let $\hat{\theta}_n$ be the MLE of θ. Then $\hat{\tau}_n = g(\hat{\theta}_n)$ is the MLE of τ.

Proof: Suppose that g is one-to-one. Then it possesses an inverse g^{-1}, and we can define the induced likelihood $\mathcal{L}^*(\tau) = \mathcal{L}(g^{-1}(\tau))$. But for any τ,

$$\mathcal{L}^*(\tau) = \mathcal{L}(g^{-1}(\tau)) \leq \mathcal{L}(\hat{\theta}_n) = \mathcal{L}^*(g(\hat{\theta}_n))$$

so $\hat{\tau} = g(\hat{\theta})$ maximizes $\mathcal{L}^*$.

The general case is only slightly more complicated; we define

$$\mathcal{L}^*(\tau) = \sup_{\theta : g(\theta) = \tau} \mathcal{L}(\theta)$$

The following conditions are sufficient for consistency of the MLE:

1. $X_1, \ldots, X_n$ are iid with density $f(x;\theta)$.

2. Identifiability, i.e. if $\theta \neq \theta'$, then $f(x;\theta) \neq f(x;\theta')$.

3. The densities $f(x;\theta)$ have common support, i.e. $\{x : f(x;\theta) > 0\}$ is the same for all θ.

4. The parameter space Θ contains an open set ω of which the true parameter value θ^* is an interior point.

5. The function $f(x;\theta)$ is differentiable with respect to θ in ω.

These conditions ensure uniform convergence in probability of a normalized form of the log-likelihood to its expected value.

Note that

$$\begin{aligned}
\ell_n(\theta) &= \sum_{i=1}^{n} \log f(X_i;\theta) \\
&\propto \frac{1}{n}\sum_{i=1}^{n} \log f(X_i;\theta) \\
&\xrightarrow{P} E_{\theta^*}[\log f(X_1;\theta)] \text{ for any fixed } \theta \text{ by WLLN}
\end{aligned}$$

where θ^* denotes the true value of θ. Showing consistency requires that the convergence is uniform in θ. We also need to show that $E_{\theta^*}[\log f(X_1;\theta)]$ is maximized at $\theta = \theta^*$.

One class of distributions that satisfies the conditions is known as the **exponential family**. For $\Theta \subseteq \mathbb{R}$, these have densities that can be written as

$$f(x;\theta) = h(x)c(\theta)\exp\{\eta(\theta)T(x)\}$$

Example: Show that each belongs to the exponential family

- $Binomial(n,p)$ with n known
- $Exponential(\lambda)$

Show that $Unif(0,\theta)$ does not.

Define the score function $s(X;\theta) = \frac{\partial}{\partial\theta}\log f(X;\theta)$.

Then the **Fisher information** (based on n observations) is

$$
\begin{aligned}
I_n(\theta) &= V_\theta\left(\frac{\partial}{\partial\theta}\ell_n(\theta)\right) \\
&= V_\theta\left(\sum_{i=1}^{n} s(X_i;\theta)\right) \\
&= \sum_{i=1}^{n} V_\theta(s(X_i;\theta)) \quad \text{(if } X_1,\ldots,X_n \text{ are independent)} \\
&= nV_\theta(s(X_1;\theta)) \quad \text{(if } X_1,\ldots,X_n \text{ are identically distributed)} \\
&= nI_1(\theta) \\
&\equiv nI(\theta)
\end{aligned}
$$

In addition, under a condition satisfied for exponential family models, we can calculate

$$I(\theta) = -E_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f(X; \theta) \right]$$

Example: Suppose $X_1, \ldots, X_n \overset{iid}{\sim} Pois(\lambda)$. Calculate $I_n(\lambda)$.

The "observed" Fisher information

$$I_n^{obs}(\theta) = \frac{-\partial^2}{\partial \theta^2} \sum_{i=1}^{n} \log f(X_i; \theta)$$

measures the curvature of the log-likelihood function. In particular $I_n^{obs}(\hat{\theta})$ measures the curvature at the MLE. The more peaked $\ell_n(\theta)$ is around $\hat{\theta}$, the more "information" the likelihood gives us. $I(\theta)$ measures the average value of this quantity.

Under two additional conditions (also satisfied by iid observations under exponential family models), we have

- Asymptotic normality: $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0, 1/I(\theta))$
- Asymptotic efficiency: If $\tilde{\theta}_n$ is some other estimator s.t. $\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{D} N(0, v(\theta))$, then $v(\theta) \geq 1/I(\theta)$ for all θ.

Asymptotic normality still holds replacing $I(\theta)$ by $I(\hat{\theta})$, that is,

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{1/I(\hat{\theta}_n)}} \xrightarrow{D} N(0, 1)$$

We can use this to construct approximate $1 - \alpha$ confidence intervals for θ.

Under each of the following models, find the MLE for θ and calculate an approximate 95% confidence interval using the limiting normal distribution.

1. $X_1, \ldots, X_n \overset{iid}{\sim} Exp(\theta)$

2. $X_1, \ldots, X_n \overset{iid}{\sim} Binomial(m, \theta)$ for known m

3. $X_1, \ldots, X_n \overset{iid}{\sim} Normal(\theta, \sigma^2)$ for known σ^2

Summary

For an exponential family,

1. MOM and MLE are equivalent (check by yourself).

2. MLE is consistent.

3. MLE is asymptotically normal after an appropriate linear transformation.

4. MLE is asymptotically efficient (smallest asymptotic variance defined through fisher information).

Fisher Information matrix

When $\theta = (\theta_1, \ldots, \theta_k)$, we define the Fisher information matrix as follows.

The Hessian matrix is the matrix of second partial derivatives of the log-likelihood, with

$$H_{jj} = \frac{\partial^2}{\partial \theta_j^2} \ell_n(\theta); \quad H_{jk} = \frac{\partial^2}{\partial \theta_j \partial \theta_k} \ell_n(\theta)$$

The Fisher information matrix is

$$I_n(\theta) = - \begin{bmatrix} E_\theta(H_{11}) & \cdots & E_\theta(H_{1k}) \\ E_\theta(H_{21}) & \cdots & E_\theta(H_{2k}) \\ \vdots & \vdots & \vdots \\ E_\theta(H_{k1}) & \cdots & E_\theta(H_{kk}) \end{bmatrix}$$

Let $\hat{\theta}_n$ be the (vector valued) MLE, and let $J_n(\theta) = I_n(\theta)^{-1}$. Then under appropriate regularity conditions and for large n,

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{D}{\approx} N(0, nJ_n(\theta))$$

We can use the marginal densities ($\hat{\theta}_{n,i} \overset{D}{\approx} N(\theta_i, J_{n,ii}(\theta))$ to construct 95% confidence intervals for the individual parameters.

Example: Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. The MLEs for μ and σ are $\hat{\mu}_n = \bar{X}_n$ and $\hat{\sigma}_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$. In addition...

$$I_n(\mu, \sigma) = \begin{bmatrix} n/\sigma^2 & 0 \\ 0 & 2n/\sigma^2 \end{bmatrix}$$

$$J_n(\mu, \sigma) = I_n(\mu, \sigma)^{-1} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2/(2n) \end{bmatrix}$$

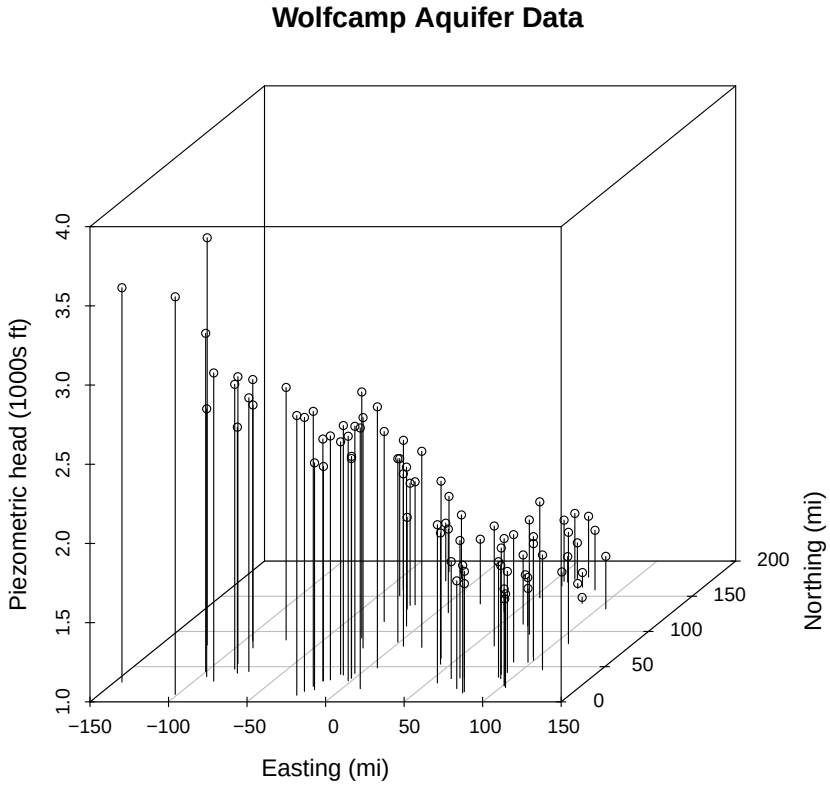
Using the fact that both $\hat{\mu}_n$ and $\hat{\sigma}_n$ are consistent, we can plug in to get

$$\hat{X}_n \pm 2\sqrt{\frac{\hat{\sigma}_n^2}{n}} \text{ and } \hat{\sigma}_n \pm 2\sqrt{\frac{\hat{\sigma}_n^2}{2n}}$$

as approximate 95% confidence intervals for μ and σ.

A more complicated likelihood problem

Given measurements of hydraulic head from an aquifer, how to create a predicted surface.



The Wolfcamp Aquifer lies below Deaf Smith County, Texas, once under consideration by DOE as a nuclear waste repository site.

Creating a smooth surface from the measurements would allow us to predict the path of potential contaminants.

Aside: Multivariate normal distribution

The multivariate normal distribution for a vector $Z = (Z_1, Z_2, \ldots, Z_n)'$ with mean vector μ and covariance matrix Σ has pdf

$$f(z; \mu, \Sigma) = (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)\right\}$$

Let μ_i denote the i^{th} element of μ, and Σ_{ij} the element of Σ in the i^{th} row and j^{th} column. Then

- $Z_i \sim N(\mu_i, \Sigma_{ii})$
- $Cov(Z_i, Z_j) = \Sigma_{ij}$

In the aquifer example, we could fit a multivariate normal model where μ and Σ have special structure.

In this example, we could take

$$\mu_i = E[Z_i] = \beta_0 + \beta_1 x_i + \beta_2 y_i$$

where (x_i, y_i) is the location of observation i.

The observations are clearly not independent, so Σ is not diagonal. One model would be to have correlation decay with distance, such as

$$\Sigma_{ij} = Cov(Z_i, Z_j) = \sigma^2 \exp\{-d_{ij}/\rho\}$$

where $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$.

There is no closed form expression for the MLE of $\theta = (\beta_0, \beta_1, \beta_2, \sigma^2, \rho)$.

In many cases, it's not possible to find a closed-form expression for the MLE in multiparameter models. This is true even for some common distributions like the Gamma and Beta distributions.

However, numerical optimization is a highly developed field that comes to our rescue in applied problems (that is, when we have actual values for $X_1, \ldots, X_n$).

Most of these algorithms are written for minimization, so we need to

- Write a function for the negative log-likelihood
- Minimize it numerically
- Examine the behavior of the negative log-likelihood at the minimum
- Optionally, get a numerical approximation of the Hessian and compute the observed Fisher information matrix

Multiparameter delta method

Suppose $\tau = g(\theta_1, \ldots, \theta_k)$ is a differentiable function. Let $\nabla g = (\frac{\partial}{\partial \theta_1} g(\theta) \cdots \frac{\partial}{\partial \theta_k} g(\theta))'$ be the gradient of g and suppose that ∇g evaluated at $\hat{\theta}_n$ is not zero. Then

$$\frac{\hat{\tau}_n - \tau}{\widehat{se}(\hat{\tau}_n)} \xrightarrow{D} N(0,1)$$

where

$$\widehat{se}(\hat{\tau}_n) = \sqrt{(\hat{\nabla} g)' J_n(\hat{\theta}_n)(\hat{\nabla} g)}$$

and $\hat{\nabla} g$ is ∇g evaluated at $\hat{\theta}_n$.

Example: Continuing the example on page 7, let $\tau = g(\mu, \sigma) = \mu/\sigma$. Find the MLE for τ and its limiting normal distribution.

Sufficiency

Motivation. We hope to separate the information contained in the data into the information relevant for making inference about θ and the information irrelevant for these inferences. In other words, we would like to compress the data to, e.g. $T(X)$, without loss of information. (Actually, it often turns out that some part of the data carries no information about the unknown distribution that produces the data)

Benefits:

1. increasing computational efficiency and decreasing storage requirements
2. involving irrelevant information may increase an estimator's risk (see Rao-Blackwell Theorem)
3. Improving the scientific interpretability of our data

Definition of Sufficient Statistic

Suppose X has a distribution from $\mathcal{P} = \{ P_\theta : \theta \in \Omega \}$. A statistic T is **sufficient** for θ if, for every t in the range $\mathcal{T}$ of T, the conditional distribution $P_\theta(X \mid T(X) = t)$ is independent of θ.

Example: Let $X_i \sim Ber(\theta)$ i.i.d., $i = 1, ..., n$. Show that $T = \sum_{i=1}^{n} X_i$ is sufficient for θ.

Neyman Fractorization Theorem. Suppose a family $\{ P_\theta : \theta \in \Omega \}$ of distributions have joint mass functions or densities $\{ p(x; \theta) : \theta \in \Omega \}$. Then a statistic T is sufficient for θ if and only if there are functions h and g such that the density/mass function can be written

$$p(x; \theta) = h(x) \cdot g(T(x), \theta).$$

Proof: To be presented in class (for the discrete case).

Examples:

1. Let $Y_i \sim Uniform(0, \theta)$ i.i.d., $i = 1, ..., n$. Show that $T = Y_{(n)}$ is sufficient for θ.

2. Let $X_i \sim N(\theta, 1)$ i.i.d., $i = 1, ..., n$. Show that $T = \sum_{i=1}^{n} X_i$ is sufficient for θ.

The Rao-Blackwell Theorem

Suppose X is distributed according to $P_\theta(x) \in \{P_\theta : \theta \in \Omega\}$ and a statistic $T(X)$ is sufficient for θ. Given any estimator $\delta(X)$ of θ, define $\eta(T) = E_\theta[\delta(X)|T(X)]$. If the loss function $\mathcal{L}(\theta, \delta(X))$ is convex and the risk function $R(\theta, \delta(X)) = E[\mathcal{L}(\theta, \delta(X))] < \infty$, then $R(\theta, \eta) \leq R(\theta, \delta)$. If $\mathcal{L}$ is strictly convex, then the inequality is strict unless $\delta = \eta$.

Note that the loss function reflects the degree of wrongness of an estimate. The commonly used quadratic loss function is defined as $\mathcal{L}(\theta, \delta) = (\theta - \delta(X))^2$.

Proof of Rao-Blackwell: by Jensen's inequality and iterated expectation.

Minimal Sufficiency

Definition. Suppose $T(X)$ is sufficient for $P = \{P_\theta : \theta \in \Omega\}$. For any other sufficient statistic $S(X)$, if we can always find a function f such that $T = f(S)$, then T is minimally sufficient.

($T = f(S)$ means (i) the knowledge of S implies the knowledge of T, and (ii) T provides a greater reduction of data unless f is one-to-one.)

A d-parameter exponential family has pdf in the following form

$$p(x, \theta) = h(x) \exp[\sum_{i=1}^{d} \eta_i(\theta) T_i(x) - A(\theta)],$$

which is of full rank if $\eta(\Theta) = \{\eta_1(\theta), ..., \eta_d(\theta)\}$ has non-empty interior in $\Re^d$ and $T_1(x), ..., T_d(x)$ are linearly independent. In a full rank exponential family, the natural sufficient statistic $T = (T_1, ..., T_d)$ is minimally sufficient.

Examples

- Let $X_1, ..., X_n$ be iid and follow a normal distribution $N(\mu, \sigma^2)$. Find the minimal sufficient statistic for μ and σ^2.

Relevant Readings on Sufficient Statistics

Chapters 2-5 of the Robert Keener book.

Hypothesis Testing

A statistical hypothesis is a statement about a parameter (or a statistical functional in nonparametric models).

A hypothesis test partitions the parameter space Θ into two disjoint sets Θ_0 and Θ_1, and produces a decision rule for choosing between

$$H_0 : \theta \in \Theta_0 \quad \text{and} \quad H_1 : \theta \in \Theta_1$$

H_0 is called the null hypothesis and H_1 is called the alternative hypothesis. The possible choices are

- Retain H_0
- Reject H_0 and accept H_1

The decision of whether to reject H_0 is determined by whether the sample $X = (X_1, \ldots, X_n)$ falls into a predefined rejection region R.

Usually, the rejection region R has the form

$$R = \{x_1, \ldots, x_n : T(x_1, \ldots, x_n) > c\}$$

where T is called a test statistic and c is called a critical value.

The idea is to construct R so that the probability of the data falling into it when H_0 is true is small.

Example: Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, and let $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ be the MLEs. If $H_0 : \mu = 0$, one test statistic we might consider is $T = |\hat{\mu}_n / \hat{\sigma}_n|$, reasoning that if H_0 is true, T will tend to be small.

Note: c is just a placeholder. It usually will depend on n and/or our choice of Θ_0 and Θ_1.

We evaluate a test using its power function. This is defined by

$$\beta(\theta) = P_\theta(X \in R)$$

Ideally, we would like $\beta(\theta)$ to be 0 when $\theta \in \Theta_0$ and 1 when $\theta \in \Theta_1$, but that is typically impossible to achieve.

Qualitatively, a good test has small $\beta(\theta)$ when $\theta \in \Theta_0$ and large $\beta(\theta)$ when $\theta \in \Theta_1$.

However, there are typically tradeoffs between the two, so that the researcher must choose between tests based on what kind of error probabilities he/she is willing to accept. More on this shortly.

Example 1: Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, where σ^2 is known. Consider testing $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$, using rejection region

$$R = \{x_1, \ldots, x_n : |\bar{X}_n| > c\}$$

Find and plot $\beta(\mu)$.

Example 2: Let $X \sim Bin(5, p)$. Consider testing $H_0 : p \leq 1/2$ versus $H_1 : p > 1/2$. Consider two different rejection regions:

$$\begin{aligned} R_1 &= \{x : x = 5\} \\ R_2 &= \{x : x \geq 3\} \end{aligned}$$

Plot and compare the corresponding power functions $\beta_1(p)$ and $\beta_2(p)$.

To make the problem of comparing tests better defined, we restrict ourselves to tests of a certain level, and then we try to find a test within that class that has large $\beta(\theta)$ for $\theta \in \Theta_1$.

A test is said to have level α if its size is less than or equal to α. The size of a test is defined to be

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$$

In words, the size of a test is the largest probability of rejecting H_0 when H_0 is true. This is called a Type I error.

	Retain H_0	Reject H_0
H_0 true	Correct	Type I error
H_1 true	Type II error	Correct

Since H_0 usually represents a "default" hypothesis, first guaranteeing that the probability of this is small is a scientifically conservative strategy.

Continuation of Example 1: Find the size of the test as a function of c (and possibly other things). What should c be to produce a size α test?

Continuation of Example 2: Consider a rejection region of the form $R = \{x : x \geq c\}$.

- What values of c do we need to consider?
- For each of these, find the size of the corresponding test.
- What c should we choose if we want a probability of Type I error of no more than 10%?

Theorem: Correspondence between point-null tests and confidence sets

1. For each $\theta_0 \in \Theta$, let $A(\theta_0) = R^C(\theta_0)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$. For each possible sample $x_1, \ldots, x_n$, define a set $C_n(x_1, \ldots, x_n)$ in Θ by

$$C_n(x_1, \ldots, x_n) = \{\theta_0 : x_1, \ldots, x_n \in A(\theta_0)\}$$

   Then $C_n(X_1, \ldots, X_n)$ is a $1 - \alpha$ confidence set for θ_0. That is, $P_{\theta_0}(\theta_0 \in C_n(X_1, \ldots, X_n)) \geq 1 - \alpha$.

2. Let $C_n(X_1, \ldots, X_n)$ be a $1 - \alpha$ confidence set. For any $\theta_0 \in \Theta$, define

$$A(\theta_0) = \{x_1, \ldots, x_n : \theta_0 \in C_n(x_1, \ldots, x_n)\}$$

   Then $R(\theta_0) = A^C(\theta_0)$ is the rejection region of a level α test of $H_0 : \theta = \theta_0$.

Practically speaking, this means that if we already have a $1-\alpha$ confidence interval for θ and we want to test $H_0 : \theta = \theta_0$, a level α test is just to reject H_0 if θ_0 falls outside the interval.

Often we can get *approximate* $1-\alpha$ confidence intervals using an estimator of θ that is asymptotically normal.

Wald Test:

Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Let $\hat{\theta}_n$ be an estimator such that $(\hat{\theta}_n - \theta_0)/\widehat{se}(\hat{\theta}_n) \xrightarrow{D} N(0,1)$. The size α Wald test rejects H_0 when $T > z_{\alpha/2}$, where

$$T = \left| \frac{\hat{\theta}_n - \theta_0}{\widehat{se}(\hat{\theta}_n)} \right|$$

We can show that asymptotically, the Wald test has size α, and that it is obtained by inverting the approximate $1-\alpha$ normal-based CI for θ.

Note that this method is quite general; we just need asymptotic normality.

For example

- Consider a multi-parameter problem in which $\hat{\theta}_n$ is the MLE and g is an invertible function. Then we can form a Wald test based on

$$\frac{g(\hat{\theta}) - g(\theta_0)}{\widehat{se}(g(\hat{\theta}_n))} \xrightarrow{D} N(0,1)$$

  where $\widehat{se}(g(\hat{\theta}_n))$ is found using the Delta method.

- Consider the case that $\theta = T(F)$ for some unknown distribution F. If T is a linear functional, the plug-in estimator is a mean of iid random variables, so we can use the CLT. In the case of a nonlinear functional, we could approximate $\widehat{se}(\hat{\theta}_n)$ using the bootstrap.

Examples

- Consider again $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where σ^2 is known. Show that the size α Wald test for $H_0 : \mu = 0$ produces a rejection region as in Example 1 above. (Actually the size is exactly α in this case).

- Now suppose that σ^2 is unknown. Construct a size α Wald test for $H_0 : \mu = 0$.

- Suppose that $X \sim Bin(m, p_1)$ and $Y \sim Bin(n, p_2)$. Construct a size α Wald test for $H_0 : p_1 = p_2$.

- Let $F(u, v)$ be the joint distribution of two r.v. U and V. Let $\theta = T(F) = \rho(U, V)$, where ρ denotes the correlation. Describe how to construct a size α Wald test for $H_0 : \rho = 0$ using the plug-in estimator and the bootstrap.

Likelihood ratio test (LRT)

Another broadly applicable class of tests is the likelihood ratio test (LRT). Let

$$T(X) = \frac{\sup_{\theta \in \Theta} \mathcal{L}_n(\theta)}{\sup_{\theta \in \Theta_0} \mathcal{L}_n(\theta)}$$

If $T(X)$ is large, it means there are values of θ in Θ_1 which are larger than for any in Θ_0. A likelihood ratio test is a test for which

$$R = \{x : T(x) > c\}$$

If $\hat{\theta}_n$ is the MLE and $\hat{\theta}_{n,0}$ is the MLE restricting $\theta \in \Theta_0$, then

$$T(X) = \frac{\mathcal{L}_n(\hat{\theta}_n)}{\mathcal{L}_n(\hat{\theta}_{n,0})}$$

Sometimes we can calculate the power function for the LRT exactly.

Example: Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Find $T(X)$ and find a simplified expression for the form of the rejection region. Use it to find the size α LRT.

When the power function can not be calculated exactly, and Θ_0 consists of fixing certain elements of θ (e.g., as in a point-null hypothesis), we can use the limiting distribution

$$\lambda(X) = 2 \log T(X) \xrightarrow{D} \chi^2_{r-q}$$

where r is the dimension of Θ and q is the dimension of Θ_0 .

Aside: The χ^2_k distribution (read "chi squared with k degrees of freedom") is the distribution of the sum of squares of k independent standard normal random variables. That is, if $Z_1, \ldots, Z_k \overset{iid}{\sim} N(0,1)$, then

$$Y = \sum_{i=1}^{k} Z_i^2 \sim \chi^2_k$$

We can use this approximation to find an appropriate critical value.

Example: Suppose $X_1, \ldots, X_n \overset{iid}{\sim} Poisson(\theta)$. Let $\hat{\theta}_n = \sum_{i=1}^n X_i/n$ be the MLE for θ. For testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, we have

$$\begin{aligned}
\lambda &= 2 \log \frac{\mathcal{L}(\hat{\theta}_n)}{\mathcal{L}(\theta_0)} \\
&= 2 \log \frac{e^{-n\hat{\theta}_n} \hat{\theta}_n^{\sum x_i}}{e^{-n\theta_0} \theta_0^{\sum x_i}} \\
&= 2n[(\theta_0 - \hat{\theta}_n) - \hat{\theta}_n \log(\theta_0/\hat{\theta}_n)]
\end{aligned}$$

Since for large n, $\lambda \overset{D}{\approx} \chi^2_1$, to construct an approximate size α LRT, we find $\chi^2_{1,\alpha}$ s.t. $P(\chi^2_1 < \chi^2_{1,\alpha}) = 1 - \alpha$ and reject H_0 if $\lambda > \chi^2_\alpha$.

P-value

Suppose that for every $\alpha \in (0,1)$ we have a size α test with rejection region R_α. When $R_\alpha = \{x : T(x) \geq c_\alpha\}$,

$$\text{p-value} = \sup_{\theta \in \Theta_0} P_\theta(T(X) \geq T(x))$$

where x is the observed data.

Therefore, the p-value is the probability under H_0 of observing a value $T(X)$ the same as or more extreme than what was actually observed.

Equivalently,

$$\text{p-value} = \inf\{\alpha : T(x) \in R_\alpha\}$$

That is, the p-value is the smallest level at which we can reject H_0 with x observed.

In the case of the Wald test, the (approximate) p-value is

$$\text{p-value} = P_{\theta_0}(|W| > |w|) \approx P(|Z| > |w|) = 2\Phi(-|w|)$$

where w is the observed value of the statistic and $Z \sim N(0, 1)$.

In the case of the LRT with point null hypothesis and limiting χ^2_{r-q} distribution, the (approximate) p-value is

$$\text{p-value} = P_{\theta_0}(\lambda(X) > \lambda(x)) \approx P(\chi^2_{r-q} > \lambda(x))$$

Theorem: If the test statistic has a continuous distribution, then under $H_0 : \theta = \theta_0$, the p-value has a $Unif(0, 1)$ distribution. Therefore, if we reject H_0 when the p-value is less than α, the probability of a Type I error is α.

Note! It is very tempting to think that $P(H_0|Data)$, but this is not the case. We have calculated the p-value *assuming H_0 is true.* Moreover, this kind of quantity doesn't make sense in frequentist statistics, in which we think of the parameters (determining H_0) as being fixed. However, we will see soon that this quantity does make sense (and can be calculated) in a Bayesian framework.

Neyman-Pearson Theorem

Suppose we test $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$. Let

$$T(X) = \frac{\mathcal{L}_n(\theta_1)}{\mathcal{L}_n(\theta_0)} = \frac{f(x_1, ..., x_n; \theta_1)}{f(x_1, ..., x_n; \theta_0)}.$$

Suppose we reject H_0 when $T > c$. If we choose c so that $P_{\theta_0}(T > c) = \alpha$, then this test the most powerful, size α test. That is, among all tests with size α, this test maximize the power $\beta(\theta_1)$.

We'll next discuss some tests when the data are multinomial.

Aside: The Multinomial Distribution

Suppose $Z \in \{1, ..., k\}$ and let $p_j = P(Z = j)$. The parameter $p = (p_1, \ldots, p_k)$ is really only $k - 1$ dimensional, since $\sum_{j=1}^{k} p_j = 1$. Suppose we observe an iid sample $Z_1, \ldots, Z_n$. Let $X_j = \#\{Z_i : Z_i = j\}$. Then we say $X = (X_1, \ldots, X_k)$ has $Multinomial(n, p)$ distribution.

The PDF is

$$f(x_1, \ldots, x_k; p) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

Note that the labels $1, \ldots, k$ for the Z's are arbitrary, and that $Binomial(n, p)$ distribution is just a special case.

The MLE is $(\hat{p}_1, \ldots, \hat{p}_k) = (X_1/n, \ldots, X_k/n)$.

Consider testing $H_0 : (p_1, \ldots, p_k) = (p_{01}, \ldots, p_{ok})$ versus the alternative that they are not equal. The LRT rejects when

$$T(X) = \frac{\mathcal{L}_n(\hat{p})}{\mathcal{L}(p_0)} = \prod_{j=1}^{k} \left(\frac{\hat{p}_j}{p_{0j}} \right)^{X_j}$$

is large. Since we don't know how to calculate the exact probability of this, we'll use the limiting χ^2. That is,

$$\lambda(X) = 2 \log T(X) = 2 \sum_{j=1}^{k} X_j \log \left(\frac{\hat{p}_j}{p_{0j}} \right) \xrightarrow{D} \chi^2_{k-1}$$

The degrees of freedom is $k-1$ because the dimension of Θ is $k-1$ and the dimension of Θ_0 is zero (a point). The approximate size α LRT rejects H_0 when $\lambda(X) \geq \chi^2_{k-1,\alpha}$.

Example: Consider the following data on 2009 freshman admissions at Berkeley.

	California Residents	Non-Residents	International Students
Applicants	38,082	6,309	4,259
Admitted	11,252	1,110	666
(% Admitted)	(29.5%)	(17.6%)	(15.6%)
Enrolled	4,262	216	301

Treat the enrolled students as a sample from the Multinomial distribution, and test the hypothesis that the proportion of the three groups among the enrolled students is the same as it was for admitted students, i.e., that

$$p = \left(\frac{11252}{13028}, \frac{1110}{13028}, \frac{666}{13028}\right)$$

Another popular test for this situation is called Pearson's χ^2 test. The statistic is defined as

$$T = \sum_{j=1}^{k} \frac{(X_j - np_{0j})^2}{np_{0j}} = \sum_{j=1}^{k} \frac{(X_j - E_j)^2}{E_j}$$

Here $E_j = E[X_j] = np_{0j}$ is the expected value of X_j under H_0.

This statistic also has a limiting χ^2_{k-1} distribution under H_0.

Example: Test the Berkeley data again using Pearson's χ^2 test.

The LRT and Pearson's χ^2 are asymptotically equivalent, so they give similar answers for large n. However, Pearson's χ^2 statistic tends to converge to χ^2_{k-1} in distribution faster, so it is preferable for small n.

The LRT and Pearson's χ^2 also arise in tests of independence. Consider a simple case first, that two r.v.'s Y and Z are binary. The data are shown in the left table and the corresponding probabilities in the right table.

	$Y=0$	$Y=1$	
$Z=0$	X_{00}	X_{01}	$X_{0.}$
$Z=1$	X_{10}	X_{11}	$X_{1.}$
	$X_{.0}$	$X_{.1}$	n

	$Y=0$	$Y=1$	
$Z=0$	p_{00}	p_{01}	$p_{0.}$
$Z=1$	p_{10}	p_{11}	$p_{1.}$
	$p_{.0}$	$p_{.1}$	1

We treat $X = (X_{00}, X_{01}, X_{1,0}, X_{11})$ as a sample from a multinomial distribution. Under the null hypothesis that Y and Z are independent, the cell probabilities are the product of the row and column probabilities:

$$p_{ij} = p_{i.}p_{.j}$$

We can use this to construct either a LRT (doing a constrained maximization) or Pearson's χ^2.

Consider now a table with I rows and J columns. The unconstrained MLEs are $\hat{p}_{ij} = X_{ij}/n$, and under H_0, the constrained MLEs are

$$\hat{p}_{0ij} = \hat{p}_{0i.}\hat{p}_{0.j} = \frac{X_{i.}}{n}\frac{X_{.j}}{n}$$

Therefore for the LRT we have $\lambda = 2\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}\log\left(\frac{nX_{ij}}{X_{i.}X_{j.}}\right)$ and for Pearson's χ^2 we have

$$T = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(X_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i-1}^{I}\sum_{j=1}^{J}\frac{(X_{ij}-n\hat{p}_{0ij})^2}{n\hat{p}_{0ij}}$$

Both test statistics have a limiting χ^2_ν distribution, where $\nu = (I-1)(J-1)$.

Finally, we can adapt these ideas to form a test of goodness of fit. Here the null hypothesis is that the data come from an assumed parametric model. The idea is to "discretize" both the data and the model.

First, define k disjoint intervals $I_1, \ldots, I_k$. Define

$$p_j(\theta) = P_\theta(X \in I_j) = \int_{I_j} f(x;\theta)dx$$

Let $N_j = \#\{X_i \in I_j\}$, the number of observations that fall into I_j. Treat $N = (N_1, \ldots, N_k)$ as a sample from a multinomial distribution with $p(\theta) = (p_1(\theta), \ldots, p_k(\theta))$, and maximize the likelihood to get $\tilde{\theta}$.

Then under H_0 that the data are iid draws from $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$, the test statistic $Q = \sum_{j=1}^{k} \frac{(N_j - np_j(\tilde{\theta}))^2}{np_j(\tilde{\theta})} \xrightarrow{D} \chi^2_{k-1-s}$, where s is the dimension of θ.

"Multiple testing" refers to the problem of testing $m > 1$ hypotheses, and wanting to control something more than the error rate for each test.

The Bonferroni Method controls the probability of having at least one false rejection. If α is the upper bound placed on this probability, the method achieves this by using level α/m for each of the tests.

$$
\begin{aligned}
P(\text{at least one Type I error}) &= P(\bigcup_{i=1}^{m} \text{Type I error in the } i^{th} \text{ test}) \\
&\leq \sum_{i=1}^{m} P(\text{Type I error in the } i^{th} \text{ test}) \\
&= \sum_{i=1}^{m} \alpha/m = \alpha
\end{aligned}
$$

In many cases, Bonferroni is too conservative. Another option is to control the False Discovery Rate (FDR), which is

$$FDR = E\left(\frac{\text{Number of false rejections}}{\text{Total number of rejections}}\right)$$

Benjamini and Hochberg suggested the following procedure, which guarantees $FDR \leq \alpha$:

1. For each test, compute the $p-value$. Let $P_{(1)} < \cdots < P_{(m)}$ denote the ordered p-values.

2. Select $R = \max\{i : P_{(i)} < \frac{i\alpha}{m}\}$, when the p-values are independent.

3. Reject all null hypotheses for which the p-value $\leq P_{(R)}$.

Bayesian Statistics

Bayesian statistics is built upon a subjective interpretation of probability. What exactly is meant by "subjective" is a source of controversy. It can mean simply that probability statements are judgements made by the statistical practitioner. However others accuse Bayesian statistics of allowing the practitioner to impose his/her own biases.

Another way of saying this is that a Bayesian statistician uses the language of probability to reflect two different kinds of uncertainty about a problem:

- aleatory uncertainty: due to inherent randomness in a system or observations of the system; used in frequentist statistics too

- epistemic uncertainty: due to our own incomplete understanding of the system; the point of scientific inquiry is to reduce this

Bayesian statistics is built upon Bayes Theorem. If $x^n = (x_1, \ldots, x_n)$ represents the observed data, we have

$$f(\theta|x^n) = \frac{f(x^n|\theta)f(\theta)}{f(x^n)}$$

To use Bayes Theorem for inference, we attach interpretations.

- $f(\theta)$: the prior density – reflects knowledge of θ before seeing the data
- $f(x^n|\theta)$: the likelihood – joint density of the data for particular θ
- $f(\theta|x^n)$: the posterior density – reflects knowledge of θ after seeing the data
- $f(x^n)$: the normalizing constant – the marginal distribution for the data; can be hard to calculate

First consider a special class of problems in which the calculations on the previous page can be done in closed form.

A conjugate prior distribution for θ is one for which $f(\theta)$ and $f(\theta|x^n)$ belong to the same parametric family. In these cases, the key to calculation is to identify the kernel of the density and see how it is changed by the likelihood.

The kernel of a density for θ is the part that depends on θ, ignoring any constant multiplicative terms.

Example: Suppose $\theta \sim N(m, v)$ where v is known. What is the kernel?

Examples using conjugate priors

1. Suppose $X_1, \ldots, X_n|\lambda \overset{iid}{\sim} Poisson(\lambda)$, and the prior is $\lambda \sim Gamma(a, b)$. Find the posterior distribution for λ.

2. Suppose $X_1, \ldots, X_n|\theta \overset{iid}{\sim} N(\theta, \sigma^2)$, where σ^2 is known. Let the prior be $\theta \sim N(a, b^2)$. Find the posterior distribution for θ.

3. Suppose $X_1, \ldots, X_n|\sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$, where θ is known. Let the prior distribution for σ^2 be inverse gamma with parameters a and b. The prior PDF is

$$f(\sigma^2; a, b) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-\alpha-1}\exp\{-\beta/\sigma^2\}$$

   Find the posterior distribution for σ^2.

In Bayesian statistics, all inference is based on the posterior distribution. We can use the posterior to calculate quantities similar to those under frequentist statistics (point estimates and intervals), or we can examine the posterior probability of *any* event of interest.

The posterior mean is a commonly used point estimator:

$$E[\theta|X_1,\ldots,X_n] = \int \theta f(\theta|X_1,\ldots,X_n)d\theta$$

It can often be written as a weighted average of the prior mean and the MLE. For example, in the second example on the previous page,

$$\begin{aligned} E[\theta|X_1,\ldots,X_n] &= \frac{b^2\sum_{i=1}^n X_i + a\sigma^2}{nb^2+\sigma^2} \\ &= \frac{nb^2}{nb^2+\sigma^2}\bar{X}_n + \frac{\sigma^2}{nb^2+\sigma^2}a \end{aligned}$$

A $1-\alpha$ credible interval for θ (also called a posterior interval) is an interval C_n satisfying

$$P(\theta \in C_n | X_1, \ldots, X_n) = 1 - \alpha$$

Note a few differences compared to a confidence interval:

- The probability statement is about θ, not C_n. C_n is a function of $X_1, \ldots, X_n$, which we are conditioning on in the probability statement.

- The statement is an equality. This is different from a frequentist interval, which puts a lower bound on the probability of coverage. Here we're not making a guarantee; we're just providing one summary of the posterior distribution.

- The intervals constructed this way may or may not have good frequentist coverage rates.

Note that C_n is not uniquely defined. There are several popular methods for finding such intervals.

A $1-\alpha$ equal-tail credible interval is an interval (a, b) such that

$$\int_{-\infty}^{a} f(\theta|x^n)d\theta = \int_{b}^{\infty} f(\theta|x^n)d\theta = \alpha/2$$

A $1-\alpha$ highest posterior density (HPD) region R_n is defined such that

1. $P(\theta \in R_n|x^n) = 1-\alpha$

2. $R_n = \{\theta : f(\theta|x^n) > k\}$ for some k.

When $f(\theta|x^n)$ is unimodal, R_n is an interval.

Often it is more informative to plot $f(\theta|x^n)$ than it is to report an interval.

It is typically the case that the posterior distribution can't be calculated in closed form. This difficulty was a major roadblock for Bayesian statistics until the last 30 years or so, during which Monte Carlo sampling from the posterior became widespread.

We'll consider two basic methods for sampling from the posterior:

- Rejection sampling: produces an exact, iid sample, but may be difficult to tailor to a given problem

- Importance sampling: more generally applicable, but only approximates the distribution

Markov Chain Monte Carlo (MCMC) methods construct a Markov chain that has the posterior distribution as its stationary distribution. These are very flexible but are beyond the scope of this course.

Suppose we have $\theta_1, \ldots, \theta_B \overset{iid}{\sim} f(\theta|x^n)$. The basic Monte Carlo approximation to the posterior mean of any function $q(\theta)$ is

$$\begin{aligned} E[q(\theta)|x^n] &= \int q(\theta) f(\theta|x^n) d\theta \\ &\approx \frac{1}{B} \sum_{i=1}^{B} q(\theta_i) \end{aligned}$$

This is broader than it might seem at first glance. For example, q could be an indicator function, giving us a way of approximating the posterior probability of any event.

We can also construct the ECDF to estimate the posterior CDF, or use a histogram or kernel density estimate (later in the course) to estimate the posterior PDF.

Now consider rejection sampling where we first sample from the prior, with $g(\theta) = f(\theta)$, and the target is the posterior distribution, with $h(\theta) \propto k(\theta) = f(x^n|\theta)f(\theta)$. Note that by definition,

$$\frac{k(\theta)}{g(\theta)} = \frac{f(x^n|\theta)f(\theta)}{f(\theta)} = f(x^n|\theta) \leq f(x^n|\hat{\theta}_n) \equiv M$$

where $\hat{\theta}_n$ is the MLE. So the rejection sampling algorithm becomes

1. Draw $\theta^{cand} \sim f(\theta)$.
2. Generate $u \sim Unif(0, 1)$.
3. If $u \leq f(x^n|\theta^{cand})/f(x^n|\hat{\theta}_n)$, accept θ^{cand}; otherwise reject it.

Repeat 1-3 until B values of θ^{cand} have been accepted.

Importance sampling is an adaptation to the usual Monte Carlo integration that allows us to sample from an "importance function" g rather than the target density h. Note that

$$\begin{aligned} E_h[q(\theta)] &= \int q(\theta)h(\theta)d\theta \\ &= \int q(\theta)\frac{h(\theta)}{g(\theta)}g(\theta)d\theta \\ &\approx \frac{1}{B}\sum_{i=1}^{B} q(\theta_i)\frac{h(\theta_i)}{g(\theta_i)} \end{aligned}$$

where $\theta_1, \ldots, \theta_B \overset{iid}{\sim} g(\theta)$.

We can use this principle to obtain an approximation to $E[q(\theta)|x^n]$.

Sample from the prior: $\theta_1, \ldots, \theta_B \overset{iid}{\sim} f(\theta)$, then for each $i = 1, \ldots, B$, calculate

$$w_i = \frac{\mathcal{L}_n(\theta_i)}{\sum_{i=1}^{B} \mathcal{L}_n(\theta_i)}$$

Then $E[q(\theta)|x^n] \approx \sum_{i=1}^{B} q(\theta_i) w_i$.

How should we choose a prior distribution? Several schools of thought answer this question differently:

- Subjective Bayesianism: The prior should reflect in as much detail as possible the researcher's prior knowledge of and uncertainties about the problem. These should be determined through *prior elicitiation*.

- Objective Bayesianism: The prior should incorporate as little subjective information as possible. Priors with this property are known as *non-informative*.

- Robust Bayesianism: Reasonable people may hold different priors, and it is difficult to precisely express even one person's prior; we should therefore consider the *sensitivity* of our inferences to changes in the prior.

A Bayesian analysis will often incorporate more than one of these ideas.

The simplest kind of non-informative prior places a uniform distribution on θ. When the range of θ is bounded, this prior gives a valid PDF, since it integrates to 1.

It is also possible to assign a uniform prior when the range of θ is not bounded. For example, suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$. We could take $f(\theta) \propto 1$. This prior is called "improper," since $\int_{-\infty}^{\infty} f(\theta) d\theta = \infty$.

However, we can still apply the Bayesian machinery to get

$$\begin{aligned} f(\theta|x^n) &\propto f(x^n|\theta) f(\theta) \\ &\propto \exp\left\{ -\frac{1}{2}[n\theta^2 - 2n\theta\bar{X}_n] \right\} \end{aligned}$$

which is the kernel of a $N(\bar{X}_n, 1/n)$ distribution for θ. Therefore we still have a "proper posterior."

In a Bayesian analysis, hypotheses, like parameters, can be described using probability distributions.

The simplest case is when the hypotheses describe regions into which θ can fall, and these all have positive prior probability. If $H_0 : \theta \in \Theta_0$, then

$$\begin{aligned} \text{Prior probability: } P(H_0) &= \int_{\Theta_0} f(\theta) d\theta \\ \text{Posterior probability: } P(H_0|x^n) &= \int_{\Theta_0} f(\theta|x^n) d\theta \end{aligned}$$

Suppose $H_0, \ldots, H_{K-1}$ are K hypotheses under consideration. (Typically $K = 2$, but in theory we can have more.) Suppose that under H_k, $\theta \sim f(\theta|H_k)$. θ may mean different things under the various hypotheses.

Note that

$$P(H_k|x^n) = \frac{f(x^n|H_k)P(H_k)}{\sum_{k=1}^{K} f(x^n|H_k)P(H_k)}$$

Therefore, the posterior odds of H_i relative to H_j equals

$$\frac{P(H_i|x^n)}{P(H_j|x^n)} = \frac{f(x^n|H_i)}{f(x^n|H_j)} \times \frac{P(H_i)}{P(H_j)}$$

The term $f(x^n|H_i)/f(x^n|H_j)$ is called the Bayes Factor for comparing H_i to H_j. I'll denote it BF_{ij}.

Computing the Bayes Factor

When H_i and H_j represent regions of the parameter space, it's easier to calculate the prior and posterior odds, and from this compute the Bayes Factor.

If $H_i : \theta = \theta_i$ and $H_j : \theta = \theta_j$, then the Bayes Factor is just the ratio of likelihoods under the two values.

More generally,

$$f(x^n|H_i) = \int_\Theta f(x^n|\theta, H_i) f(\theta|H_i) d\theta,$$

which is called the marginal likelihood. If $f(\theta|H_i)$ is conjugate, it can be calculated in closed form. Otherwise, we use sampling to approximate it. For example, we could use MC integration, sampling from $f(\theta|H_i)$.

Example: Albert Pujols (St. Louis Cardinals) was voted the "most feared hitter in baseball" in 2010. However, Ichiro Suzuki (Seattle Mariners) has a very similar batting average. Here are their career statistics from 2001 to 2010, when they both played major league baseball.

Pujols: 5146 at bats; 1717 hits
Suzuki: 6099 at bats; 2030 hits

If we consider that each player has a "true" batting average p, around which their actual batting average fluctuates, we might be interested in looking at evidence for/against the hypothesis $p_{Pujols} = p_{Suzuki}$.

Suppose $X|p_1 \sim Bin(n, p_1)$ and $Y|p_2 \sim (m, p_2)$. Under $H_0 : p_1 = p_2$, we assign prior distribution $p_1 \sim Unif(0, 1)$ (and $p_2 = p_1$) and under $H_1 : p_1 \neq p_2$, we assign independent priors $p_1 \sim Unif(0, 1)$ and $p_2 \sim Unif(0, 1)$.

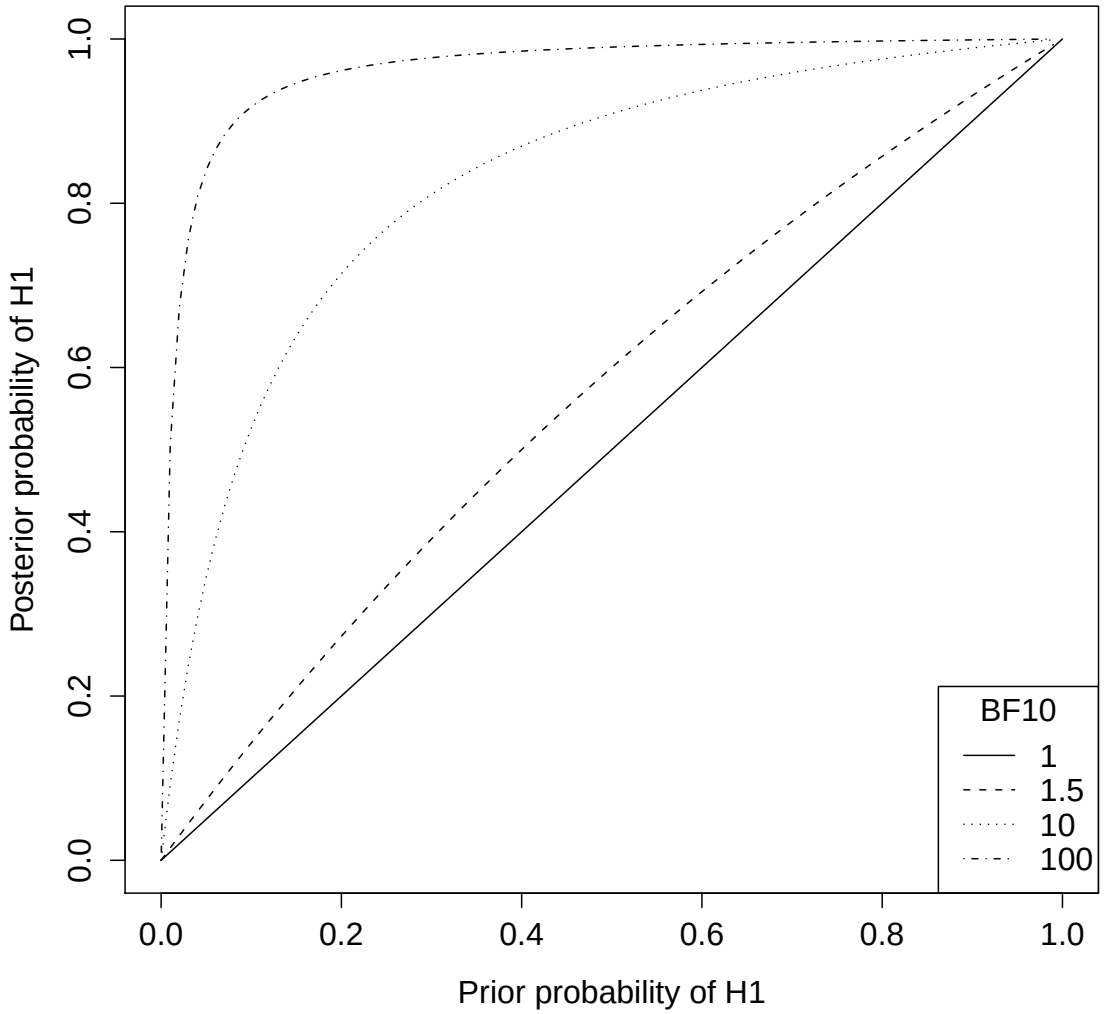
Calculate $f(x, y|H_1)$. (The rest of the problem is in the homework.)

Let BF_{10} be the Bayes Factor for comparing H_1 to H_0. We might classify BF_{10} as a measure of evidence against H_0 and in favor of H_1 as follows

$log_{10}(BF_{10})$	BF_{10}	Evidence
$0 - 1/2$	$1 - 3/2$	Weak
$1/2 - 1$	$3.2 - 10$	Moderate
$1 - 2$	$10 - 100$	Strong
> 2	> 100	Decisive

The key to this interpretation is to note that if $p = P(H_1)$ and $p^* = P(H_1|Data)$, then

$$p^* = \frac{\frac{p}{1-p}BF_{10}}{1 + \frac{p}{1-p}BF_{10}}$$


Posterior probability of H1
0.0
0.2
0.4
0.6
0.8
1.0
Prior probability of H1
0.0
0.2
0.4
0.6
0.8
1.0
BF10
1
1.5
10
100

More on noninformative priors

One property we might want a noninformative prior to possess is that it be **transformation invariant**. For example, if θ represents a distance, our inference shouldn't depend on whether θ is expressed in miles or kilometers.

That is, if instead of expressing the likelihood given θ, we express it given $\phi = g(\theta)$, we want a rule for choosing the priors f_θ and f_ϕ such that

$$f_\phi(\phi) = f_\theta(g^{-1}(\phi)) \left| \frac{dg^{-1}(\phi)}{d\phi} \right|$$

The Jeffreys prior does just this. For 1-dimensional θ, we have $f(\theta) \propto I(\theta)^{1/2}$.

Decision Theory

Statistical decision theory is concerned with making decisions under uncertainty. We express our uncertainties around the problem in terms of an unknown quantity or "state of nature" θ.

The particular decision made is also referred to as an "action," and we'll denote it by a, with the collection of all possible actions denoted by $\mathcal{A}$.

A loss function

$$L(\theta, a) : (\Theta \times \mathcal{A}) \to [0, \infty)$$

describes the consequences of taking action a when the true state of nature is θ. In reality, we never know the true value of the loss (at least not at the time of the decision).

Example: A drug company has developed a new pain reliever. They are trying to determine how much of the drug to produce, but they are uncertain about the proportion of the market the drug will capture (θ).

Suppose a is an estimate of θ. The company plans to produce an amount proportional to a. One possible loss function is

$$L(\theta, a) = \begin{cases} K(\theta - a) & a - \theta < 0 \\ 2K(a - \theta) & a - \theta \geq 0 \end{cases}$$

for some constant K. This loss function implies that an overestimate of demand (leading to overproduction of the drug) is considered twice as costly as an underestimate. The loss is also taken to be linear, which may be reasonable if the total cost is proportional to the number of units produced.

Many results are based on the following "standard" loss functions. These are expressed in generic "units of utility."

- Squared error loss: $L(\theta, a) = (\theta - a)^2$
- Linear loss: $L(\theta, a) = \begin{cases} K_1(\theta - a) & a - \theta < 0 \\ K_2(a - \theta) & a - \theta \geq 0 \end{cases}$
- Absolute error loss: $L(\theta, a) = |\theta - a|$ (linear loss with $K_1 = K_2$)
- L^p loss: $L(\theta, a) = |\theta - a|^p$
- Zero-one loss: $L(\theta, a) = \begin{cases} 0 & a = \theta \\ 1 & a \neq \theta \end{cases}$

Since we don't know the actual loss, we may consider an "expected loss" and then choose an "optimal" decision with respect to this. This "expected loss" is known as risk. However, there are several ways of thinking about the expectation; hence, several different risks.

Note: in what follows, we'll consider estimation problems only, that is, actions $a = \hat{\theta}(x)$.

1. The posterior risk

$$r(\hat{\theta}|x) = \int L(\theta, \hat{\theta}(x)) f(\theta|x) d\theta$$

averages over uncertainty in θ after conditioning on observations x. We may think of it as a function of x, as well as the particular form of $\hat{\theta}$. Another way to think of this is that, conditional on the x we observed, we just get a single number for each estimator $\hat{\theta}$ we might consider.

2. The frequentist risk (or sometimes just "risk")

$$R(\theta, \hat{\theta}) = \int L(\theta, \hat{\theta}(x)) f(x|\theta) dx$$

averages over different possible realizations x of the random variable, given that the true "state of nature" is θ. It is a function of θ, as well as the particular form of $\hat{\theta}$.

Consider two estimators, $\hat{\theta}$ and $\hat{\theta}'$. We say $\hat{\theta}'$ dominates $\hat{\theta}$ if

$$\begin{aligned} R(\theta, \hat{\theta}') &\leq R(\theta, \hat{\theta}) \text{ for all } \theta \\ R(\theta, \hat{\theta}') &< R(\theta, \hat{\theta}) \text{ for at least one } \theta \end{aligned}$$

The estimator $\hat{\theta}$ is called inadmissible if there is at least one other estimator $\hat{\theta}'$ that dominates it. Otherwise it is called admissible.

Example: Suppose $X \sim N(\theta, 1)$ and we are estimating θ under squared error loss. Consider $\hat{\theta}_c(x) = cx$.

- Calculate the risk in terms of c and θ.
- Calculate the risk when $c = 1$.
- Show that $\hat{\theta}_c$ is inadmissible when $c > 1$.
- Make a plot comparing the risk when $c = 1/2$ and $c = 1$.

3. The Bayes risk:

$$
\begin{aligned}
r(f,\hat{\theta}) &= \int R(\theta,\hat{\theta}) f(\theta) d\theta \\
&= \int \left[\int L(\theta,\hat{\theta}(x)) f(x|\theta) dx \right] f(\theta) d\theta \\
&= \int \int L(\theta,\hat{\theta}(x)) f(x,\theta) dx \, d\theta \\
&= \int \left[\int L(\theta,\hat{\theta}(x)) f(\theta|x) d\theta \right] f(x) dx \\
&= \int r(\hat{\theta}|x) f(x) dx
\end{aligned}
$$

averages over both θ and X. It depends on the particular form of $\hat{\theta}$.

A decision rule that minimizes the Bayes risk is called a Bayes rule. The estimator $\hat{\theta}$ is a Bayes rule, or Bayes estimator (under a particular model and loss function) if

$$r(f, \hat{\theta}) = \inf_{\tilde{\theta}} r(f, \tilde{\theta})$$

We can also find the Bayes estimator using the posterior risk. For each x, let $\hat{\theta}(x)$ be the value of $\hat{\theta}$ that minimizes $r(\hat{\theta}|x)$. (Recall that for each x, $r(\hat{\theta}|x)$ returns a single number for each $\hat{\theta}$.) The estimator defined in this way is the Bayes estimator. This is because

$$r(f, \hat{\theta}) = \int r(\hat{\theta}|x) f(x) dx$$

and we have defined this $\hat{\theta}$ to minimize the quantity being integrated for each x; hence we've also minimized the whole integral.

Example (continued): Suppose $X \sim N(\theta, 1)$ and we are estimating θ under squared error loss. Consider $\hat{\theta}_c(x) = cx$.

- Calculate the Bayes risk using the prior $\theta \sim N(0, \tau^2)$.
- Find the Bayes rule among estimators $\hat{\theta}_c$.
- Find the Bayes risk of this estimator.

We can calculate the Bayes rule explicitly for several standard loss functions.

- Squared error loss: posterior mean
- Absolute error loss: posterior median
- Zero-one loss: posterior mode

Recall the different risk functions:

1. Posterior risk (depends on x and the form of $\hat{\theta}$)

$$r(\hat{\theta}|x) = \int L(\theta, \hat{\theta}(x)) f(\theta|x) d\theta$$

2. Frequentist risk (depends on θ and the form of $\hat{\theta}$)

$$R(\theta, \hat{\theta}) = \int L(\theta, \hat{\theta}(x)) f(x|\theta) dx$$

3. Bayes risk (depends on the form of $\hat{\theta}$)

$$r(f, \hat{\theta}) = \int \int L(\theta, \hat{\theta}(x)) f(x, \theta) dx \, d\theta$$

Completely equivalently, we could write

1. Posterior risk: $r(\hat{\theta}|x) = E_{\theta|X}[L(\theta, \hat{\theta}(x))]$

2. Frequentist risk: $R(\theta, \hat{\theta}) = E_{X|\theta}[L(\theta, \hat{\theta}(X))]$

3. Bayes risk: $r(f, \hat{\theta}) = E_{\theta,X}[L(\theta, \hat{\theta}(X))]$

By iterated expectation, we also have that

$$r(f, \hat{\theta}) = E_{\theta}[E_{X|\theta}[L(\theta, \hat{\theta}(X))]] = E_{\theta}[R(\theta, \hat{\theta})]$$

and

$$r(f, \hat{\theta}) = E_{X}[E_{\theta|X}[L(\theta, \hat{\theta}(X))]] = E_{X}[r(\hat{\theta}|X)]$$

Example: Suppose $X_1, \ldots, X_n | \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$, where θ is known. Let the prior distribution for σ^2 be inverse gamma with parameters a and b. The prior PDF is

$$f(\sigma^2; a, b) = \frac{b^a}{\Gamma(a)} (\sigma^2)^{-a-1} \exp\{-b/\sigma^2\}$$

- Find the posterior distribution for σ^2.
- What is the Bayes estimator under squared error loss?
- What is the Bayes estimator under absolute error loss?
- What is the Bayes estimator under zero-one loss?

You may use the fact the mean of an $InverseGamma(a, b)$ distribution is $b/(a-1)$ when $a > 1$, the mode is $b/(a+1)$, and the median is not available in closed form.

One final note about Bayes rules: under weak conditions, they are admissible. The intuition for this is that if there existed a rule that had lower risk, it would also have lower Bayes risk.

Here is one set of conditions:

Suppose that $\Theta \subseteq \mathbb{R}$ and that $R(\theta, \hat{\theta})$ is a continuous function of θ for every $\hat{\theta}$. Let f be a prior density that assigns positive probability to any open subset of Θ. Let $\hat{\theta}^f$ be a Bayes rule, with finite Bayes risk. Then $\hat{\theta}^f$ is admissible.

We'll now consider a different strategy for choosing an action, called a **minimax rule**. To motivate this, consider the following example.

An investor is deciding whether or not to purchase \$1000 of risky ZZZ bonds. If the investor buys the bonds, they can be redeemed at maturity for a net gain of \$500. There could, however, be a default on the bonds, in which case the original \$1000 investment would be lost. If the investor doesn't buy the bonds, she will put her money in a "safe" investment, for which she will be guaranteed a net gain of \$300 over the same time period. She estimates the probability of a default to be 0.1.

- Describe the parameter space Θ and the space of possible actions $\mathcal{A}$.
- What is the prior distribution?
- For each possible $\theta \in \Theta$ and $a \in \mathcal{A}$, compute the loss.
- Is any action inadmissible?

In the previous example, the Bayes rule is for the investor to buy the bonds, since this minimizes her expected loss (maximizes her expected gain) relative to the prior distribution for a default occurring.

However, suppose the investor is very conservative, and wants to choose a strategy to minimize the "worst case scenario." This is known as the minimax strategy – it minimizes the maximum loss that could occur.

Writing the frequentist risk of action a as $R(\theta, a)$, the maximum risk

$$\bar{R}(a) = \sup_{\theta} R(\theta, a)$$

Which action in the example minimizes $\bar{R}(a)$?

In the estimation context, our possible actions are estimators $\hat{\theta}$. Then the maximum risk is

$$\bar{R}(\hat{\theta}) = \sup_{\theta} R(\theta, \hat{\theta})$$

A decision rule that minimizes the maximum frequentist risk is called a minimax rule. The estimator $\hat{\theta}$ is a minimax rule (under a particular loss function) if

$$\sup_{\theta} R(\theta, \hat{\theta}) = \inf_{\tilde{\theta}} \sup_{\theta} R(\theta, \tilde{\theta})$$

Example (continued): Suppose $X \sim N(\theta, 1)$ and we are estimating θ under squared error loss. Consider $\hat{\theta}_c(x) = cx$.

- Calculate $\sup_\theta R(\theta, \hat{\theta}_c)$.
- Use this to determine the minimax estimator of θ.
- Let the prior distribution for θ be $N(a, b)$. Determine the Bayes estimator of θ.

Geometry of Bayes and Minimax Points for Finite Ω

Given a finite parameter space $\Omega = \{\theta_1, \cdots, \theta_k\}$, we define the risk set as $S \subseteq \mathbb{R}^k$ such that

$$S = \{(y_1, \cdots, y_k) : y_i = R(\theta_i, \delta) \text{ for } \delta \in \mathcal{A}\}$$

Lemma. The risk set S is always convex when $\mathcal{A}$ has randomized estimators.

In this setting, a prior of θ can be considered as a finite vector

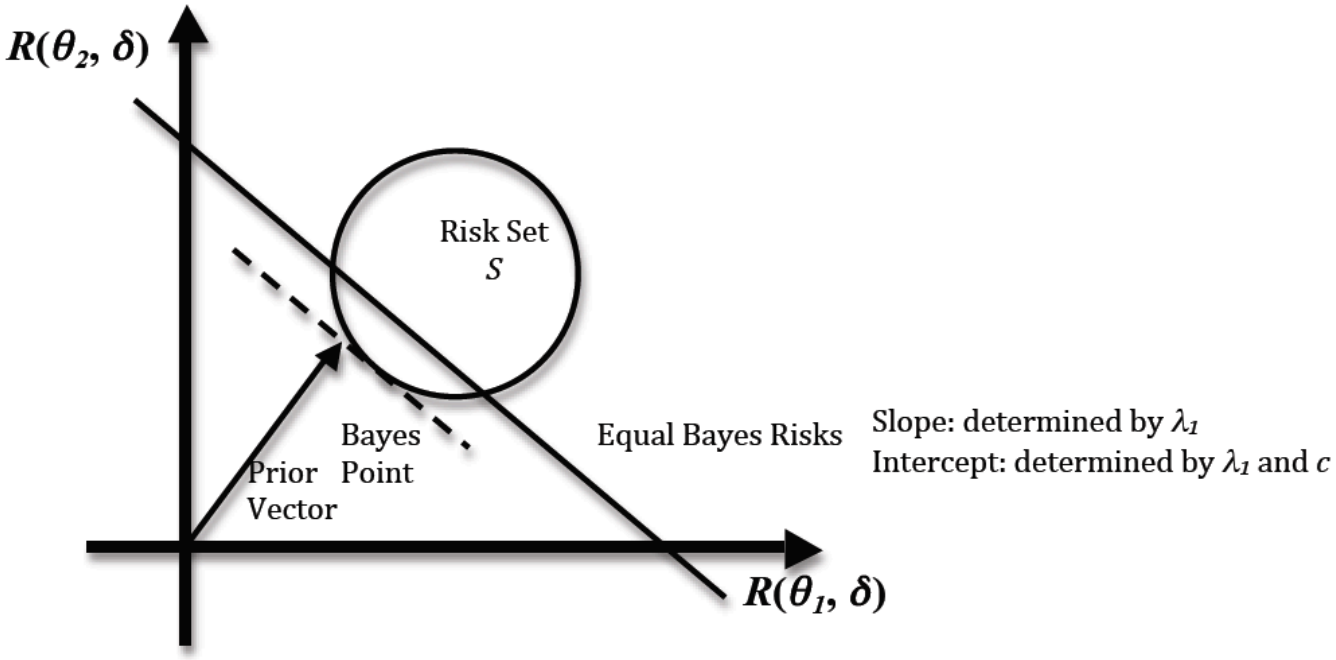
$$\lambda(\theta) = (\lambda_1, \cdots, \lambda_k) = (\lambda(\theta_1), \cdots, \lambda(\theta_k)),$$

with $\sum_{i=1}^{k} \lambda_i = 1$ and $\lambda \underset{k}{\geqslant} 0$. The Bayes risk is

$$r(\Lambda, \delta) = \sum_{i=1} \lambda_i R(\theta_i, \delta) = (\lambda_1, \cdots, \lambda_k) \begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix}$$

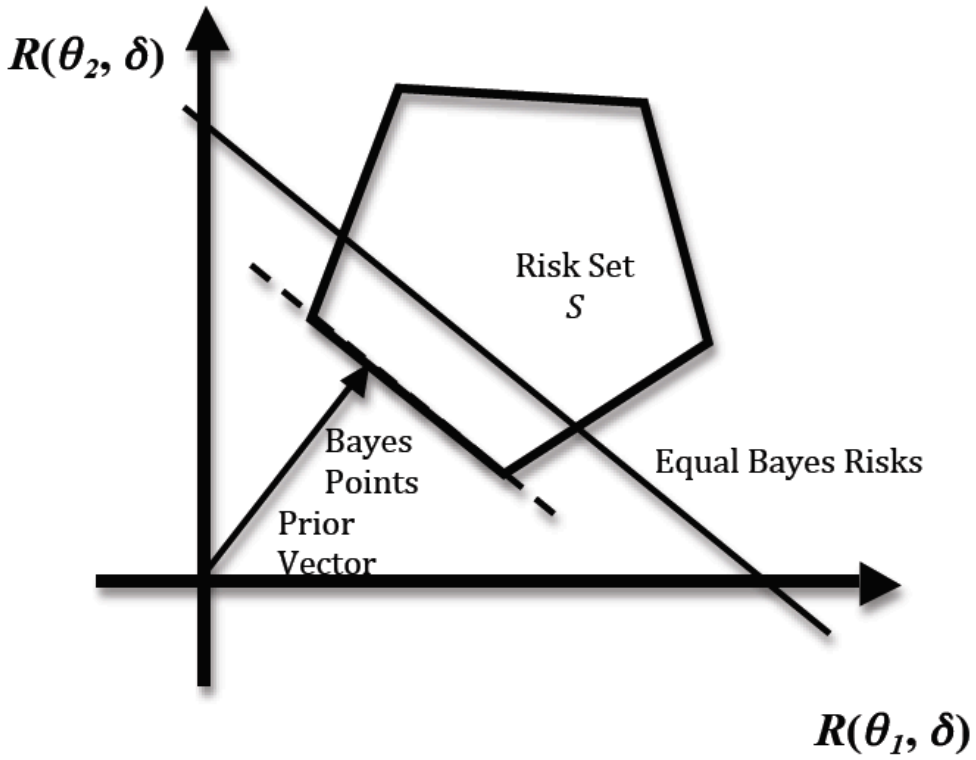
Any given prior vector $(\lambda_1, \cdots, \lambda_k)$ is normal to hyperplanes of constant Bayes risks in $\mathbb{R}^k$:

$$\text{Hyperplane:} \{(y_1, \cdots, y_k) : (\lambda_1, \cdots, \lambda_k) \begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix} = c \text{ for } \delta \in \mathcal{A}\}$$



Geometry of a Bayes Point for $k = 2$.

For any given Λ, the Bayes points should be on the hyperplane that is tangent to the risk set (ie. gives the smallest c). But Bayes points may not be unique.

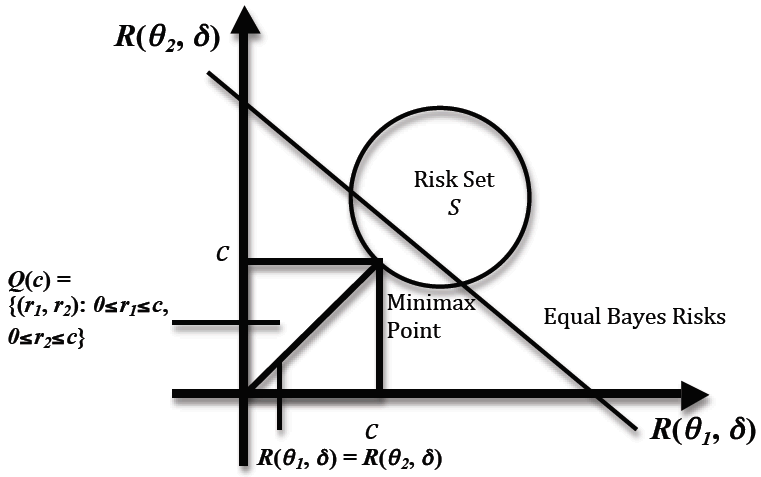


Where is the minimax point?

Where is the minimax point?

$$\sup_{\theta_1,\theta_2}(\theta,\delta^*) = \inf_{\delta\in\mathcal{A}}\sup_{\theta_1,\theta_2} R(\theta,\delta)$$

Consider the points on the vertical and horizontal segments: the points corresponding to $\sup_{\theta\in\Omega} R(\theta,\delta) = c$. The points which give the smallest c in that risk set are the minimax points. That should be the first points of contact between the squares $Q(c)$ and the risk set S.



In general it can be difficult to find minimax rules when the parameter space is infinite. One connection to Bayes rules is the following:

Suppose that $\hat{\theta}$ is the Bayes rule with respect to some prior f. Suppose further that $\hat{\theta}$ has constant risk: $R(\theta, \hat{\theta}) = c$ for some c. Then $\hat{\theta}$ is minimax.

Example: Suppose $X|p \sim Bin(n, p)$ and the loss is squared error.

- Show $\hat{p} = X/n$ is not minimax. *Hint: Consider the randomized estimator*

$$\tilde{p} = \left\{ \begin{array}{ll} X/n & \text{with probability } 1 - \frac{1}{n+1} \\ 1/2 & \text{with probability } \frac{1}{n+1} \end{array} \right\}$$

- Consider the Bayes estimator when $p \sim Beta(a, b)$. Find a and b so that the Bayes estimator has constant frequentist risk. This estimator is then minimax.

Summary: We are considering taking possible actions $a \in \mathcal{A}$, and unknown quantities affecting our decision are represented by $\theta \in \Theta$.

In the estimation context, the action is just an estimate of θ, $\hat{\theta}(x)$.

The loss function describes the consequences of taking action a when the true state of nature is θ. We write it $L(\theta, a)$ or $L(\theta, \hat{\theta}(x))$.

Ultimately, we want to *choose* an action a or an estimate $\hat{\theta}(x)$. Our choice is driven by looking at a particular *risk function.*

So far we have seen two strategy:
(1) the Bayes rule, which chooses $\hat{\theta}(x)$ to minimize the Bayes risk; (2) the minimax rule, which minimizes the maximum frequentist risk.

Under squared error loss, the Bayes rule is posterior mean; under absolute error loss, the Bayes rule is posterior median; under zero-one loss, the Bayes rule is posterior mode.

In general it can be difficult to find minimax rules when the parameter space is infinite. One connection to Bayes rules is the following:

Suppose that $\hat{\theta}$ is the Bayes rule with respect to some prior f. Suppose further that $\hat{\theta}$ has constant risk: $R(\theta, \hat{\theta}) = c$ for some c. Then $\hat{\theta}$ is minimax.

Example: Suppose $X|p \sim Bin(n, p)$ and the loss is squared error.

- Consider the Bayes estimator when $p \sim Beta(a, b)$. Find a and b so that the Bayes estimator has constant frequentist risk. This estimator is then minimax.

Example: Consider a decision problem with possible states of nature θ_1 and θ_2 . Let X be a random variable with probability function $p(x|\theta)$:
$P(X = 0|\theta_1) = 0.2, P(X = 1|\theta_1) = 0.8;$
$P(X = 0|\theta_2) = 0.4, P(X = 1|\theta_2) = 0.6.$

Two non-randomized actions a_1 and a_2 are considered with the following loss function:
$L(\theta_1, a_1(0)) = 1, L(\theta_1, a_1(1)) = 2, L(\theta_1, a_2(0)) = 4, L(\theta_1, a_2(1)) = 0;$
$L(\theta_2, a_1(0)) = 3, L(\theta_2, a_1(1)) = 1, L(\theta_2, a_2(0)) = 1, L(\theta_2, a_2(1)) = 4.$

1. Give and plot the risk set $S = \{(r_1, r_2) : r_1 = \lambda R(\theta_1, a_1) + (1 - \lambda)R(\theta_1, a_2), r_2 = \lambda R(\theta_2, a_1) + (1 - \lambda)R(\theta_2, a_2), \lambda \in [0, 1]\}$.

2. Suppose θ has the prior distribution $\Lambda(\theta)$ defined by $P(\theta = \theta_1) = 0.9, P(\theta = \theta_2) = 0.1$. What is the Bayes rule with respect to $\Lambda(\theta)$?

3. Find the minimax rule(s).

Linear Regression

The term "regression" describes a class of models for studying the relationship between a response variable Y and covariates (also called explanatory variables or regressors) $X^{(1)}, \ldots, X^{(p)}$.

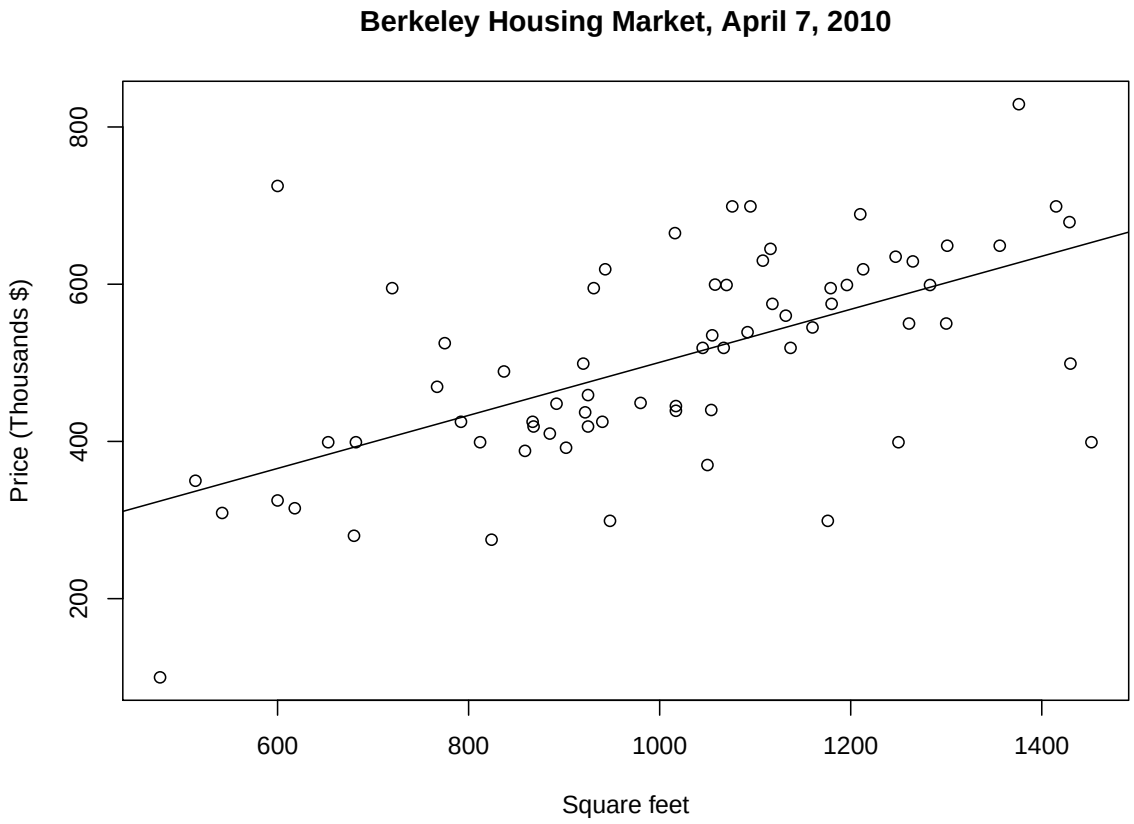
The assumption of linearity is less restrictive than it might seem, since the $X's$ can consist of nonlinear transformations of other variables of interest.

We'll start with the simple linear regression model, which means $p = 1$ and

$$\begin{aligned} E[Y|X = x] &= \beta_0 + \beta_1 x \\ V[Y|X = x] &= \sigma^2 \end{aligned}$$

We're not (yet) assuming anything else about $p(Y|X)$.

We observe pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, and based on this we estimate β_0, β_1, and σ^2. For example, here is some data from www.zillow.com:



The model for an individual observation is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $E[\epsilon_i] = 0$ and $V[\epsilon_i] = \sigma^2$.

The fitted regression line is $\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$, and the fitted values are $\hat{Y}_i = \hat{r}(X_i)$. The residuals are

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

A classical way of estimating β_0 and β_1 is by minimizing the residual sum of squares

$$RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

The least squares estimates are

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}\end{aligned}$$

Once we have $\hat{\beta}_0$ and $\hat{\beta}_1$, we may form an unbiased estimator of σ^2 via

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n \hat{\epsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

In the housing example, $\hat{\beta}_0 = 163.3$ and $\hat{\beta}_1 = 0.337$. We may interpret $\hat{\beta}_1$ to mean that for every additional square foot, the average price increases by \$337.

Now add the assumption that $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2)$. Equivalently, $Y_i | X_i \sim N(\mu_i, \sigma^2)$, where $\mu_i = \beta_0 + \beta_1 X_i$.

Conditioning on X, we have a likelihood

$$\mathcal{L}(\beta_0, \beta_1, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \right\}$$

The MLEs for β_0 and β_1 are the same as the least squares estimates. The MLE for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Some basic properties of β_0 and β_1:

1. They are unbiased: $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$.

2. They are consistent: $\hat{\beta}_0 \xrightarrow{P} \beta_0$ and $\hat{\beta}_1 \xrightarrow{P} \beta_1$.

3. They are asymptotically normal:

$$\frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)} \xrightarrow{D} N(0,1) \quad \text{and} \quad \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \xrightarrow{D} N(0,1)$$

The variances are

$$\begin{aligned} V[\hat{\beta}_0] &= \frac{\sigma^2}{n} \frac{\sum_{i=1}^n X_i^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ V[\hat{\beta}_1] &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ Cov[\hat{\beta}_0, \hat{\beta}_1] &= -\frac{\sigma^2 \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \end{aligned}$$

To estimate standard errors, we plug in $\hat{\sigma}^2$ (either unbiased or MLE) for σ^2. This allows us to construct confidence intervals and carry out tests.

Usually the test we're interested in is for $H_0 : \beta_1 = 0$. For this we can construct a Wald test using $W = \hat{\beta}_1 / \widehat{se}(\hat{\beta}_1)$.

In R, much of this calculation can be carried out using the `lm` function.

```
> linmod <- lm(price~sqft, data = berkhousing)
> summary(linmod)

Call:
lm(formula = price ~ sqft, data = berkhousing)

Residuals:
     Min       1Q   Median       3Q      Max
-260.983  -51.817    3.214   46.845  359.347

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 163.22699   57.16475   2.855  0.00572 **
sqft          0.33738    0.05517   6.115  5.6e-08 ***

... (more stuff)
```

Plotting the points and adding the fitted line:

```
plot(berkhousing$sqft, berkhousing$price,
     xlab = "Square feet", ylab = "Price (Thousands $)")
abline(linmod) # Add the fitted line to the plot
```

Here is some code to compute the p-value for the Wald test for $\beta_1 = 0$. In this case it is very small, as was the p-value for the t-test that `lm` computed.

```
> beta1 <- linmod$coefficients[2]
> se.beta1 <- summary(linmod)$coefficients[2,2]
> W <- beta1/se.beta1
> 2*pnorm(-abs(W))
        sqft
9.670514e-10
```

`linmod` and `summary(linmod)` are lists, but they print in special ways. To see what's inside the list, use `names(linmod)` and `names(summary(linmod))`.

What if we want to predict Y from X? We need to be careful what we mean by this: are we talking about

- the fit $\hat{r}(x_*) = \widehat{E}[Y|X = x_*]$? This is the mean of a distribution.
- a new observation Y_* for $X = x_*$? This is a sample from a distribution.

Our confidence intervals are different, depending on which one we want.

```
> predict(linmod, newdata = data.frame(sqft = 1000), interval = "confidence")
       fit      lwr      upr
1 500.6042 474.5419 526.6664
> predict(linmod, newdata = data.frame(sqft = 1000), interval = "prediction")
       fit     lwr      upr
1 500.6042 282.698 718.5103
```

The coefficient of variation, usually just called R^2, is the ratio of "explained" sum of squares to total sums of squares:

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \\ &= 1 - \frac{\sum_{i=1}^{n}\hat{\epsilon}_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \end{aligned}$$

R^2 ranges from zero (no variance explained) to one (all variance explained – a perfect fit).

Now consider including more than one covariate, so the model is

$$\begin{aligned} Y_i &= \sum_{j=1}^{k} \beta_j X_{ij} + \epsilon_i \\ &= \mathbf{X}_i' \beta + \epsilon_i \end{aligned}$$

where $\mathbf{X}_i = (X_{i1}\ X_{i2}\ \ldots\ X_{ik})'$ and $\beta = (\beta_1\ \beta_2\ \ldots\ \beta_k)'$.

Then the model for the vector of the observations $Y = (Y_1\ Y_2\ \ldots\ Y_n)'$ is

$$Y = X\beta + \epsilon$$

where X is the $n \times k$ matrix with i^{th} row x_i' and $\epsilon = (\epsilon_1\ \epsilon_2\ \ldots\ \epsilon_n)'$. Usually the model will contain an intercept, with $X_{i1} = 1$ for all i. That is, the first column of X contains all 1's.

Example (Hamilton, 1983)

In the late 1970's, the city of Concord New Hampshire experienced a growing demand for water, despite having roughly stable population. In 1979 and 1980 there was a shortage of water, leading to a media campaign to persuade citizens to use less. Over the next year, water use declined by about 15%. The 1981 Concord Water Study examined what variables were associated with water use.

$Y =$ Summer 1981 water use (cubic feet)
$X_1 =$ Household income (thousands of dollars)
$X_2 =$ Summer 1980 water use (cubic feet)
$X_3 =$ Highest household education (years)
$X_4 =$ Retired: 1 for yes, 0 for no
$X_5 =$ People living in the house in 1981
$X_6 =$ Increase in number of people from 1980

How do we fit the model, interpret it, and decide which variables to keep? We will consider the case that the ϵ_i's are iid normal. However, it will be convenient to write distributions in terms of the multivariate normal distribution.

Recall that we write $Z \sim N(\mu, \Sigma)$ to denote that Z is multivariate normal with $E[Z_i] = \mu_i$ and $Cov(Z_i, Z_j) = \Sigma_{ij}$.

We will make use of the fact that if $Z \sim N(\mu, \Sigma)$, then

$$AZ \sim N(A\mu, A\Sigma A').$$

The multivariate regression model with normal errors is

$$Y \sim N(X\beta, \sigma^2 I_n)$$

where I_n is the $n \times n$ identity matrix.

The likelihood is

$$f(y|\beta,\sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right\}$$

Note that the likelihood involves $(Y-X\beta)'(Y-X\beta)$, which is the residual sum of squares (RSS) for this model.

$$RSS = (Y-X\beta)'(Y-X\beta) = ||Y-X\beta||^2 = \sum_{i=1}^{N}(Y_i - x_i'\beta)^2$$

As before, maximizing the likelihood with respect to β is equivalent to minimizing RSS.

Setting $\frac{\partial}{\partial \beta} RSS = -2X'Y + 2X'X\beta \equiv 0_k$ implies

$$X'X\beta = X'Y$$

These are called the "normal equations." They have a unique solution if and only if $X'X$ is nonsingular, which is true if and only if the rank of X is k. That is, the regressors may not be linear combinations of one another.

The unique solution to the normal equations in this case (and correspondingly, the MLE for β when the data are normal) is

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Note that $\hat{\beta} = AY$, where $A = (X'X)^{-1}X'$. Therefore,

$$\hat{\beta} \sim N(AX\beta, A[\sigma^2 I_n]A') = N(\beta, \sigma^2(X'X)^{-1})$$

The interpretation for $\hat{\beta}_j$ is a change in the mean of Y, holding all the other covariates constant. This is important, because it means that the interpretation of $\hat{\beta}_j$ will change, depending on which variables are in the model.

To form confidence intervals for the elements of β, we can use the result on the previous slide, but replacing σ^2 by an estimate, which we will derive next. Take

$$\hat{\beta}_j \pm z_{\alpha/2} \widehat{se}(\hat{\beta}_j),$$

where

$$\widehat{se}(\hat{\beta}_j) = \hat{\sigma}^2 [(X'X)^{-1}]_{jj}.$$

Slutzky's Theorem gives us $\frac{\hat{\beta}_j - \beta_j}{\widehat{se}(\hat{\beta}_j)} \xrightarrow{D} N(0,1)$, which we can use to construct a Wald test.

The MLE for σ^2 is $\hat{\sigma}^2 = \frac{1}{n}(Y - X\hat{\beta})'(Y - X\hat{\beta})$.

Note that

$$\begin{aligned} Y - X\hat{\beta} &= Y - X(X'X)^{-1}X'Y \\ &= (I_n - X(X'X)^{-1}X')Y \\ &\equiv PY \end{aligned}$$

P is a symmetric and idempotent matrix, meaning $P' = P$ and $PP = P$. Therefore,

$$\hat{\sigma}^2 = \frac{1}{n}(PY)'(PY) = \frac{1}{n}Y'P'PY = \frac{1}{n}Y'PY$$

The term $Y'PY$ is known as a quadratic form in Y. A general result for quadratic forms gives us that

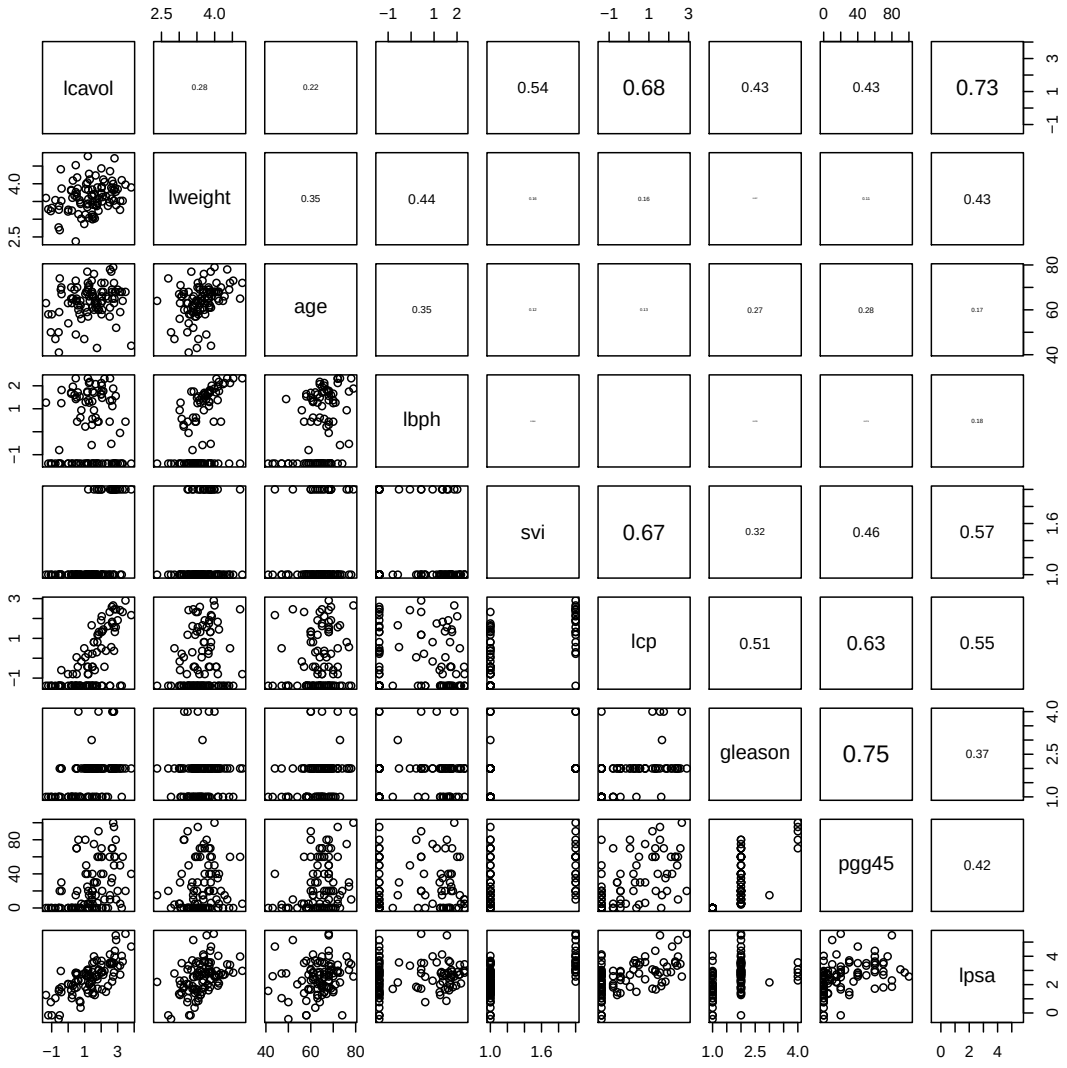
$$\begin{aligned}
E[Y'PY] &= tr\{PCov(Y)\} + E[Y]'PE[Y] = \sigma^2(n-k) \\
&= tr\{P\sigma^2 I_n\} + (X\beta)'P(X\beta) \\
&= \sigma^2 tr\{I_n - X(X'X)^{-1}X'\} + (X\beta)'[I - X(X'X)^{-1}X'](X\beta) \\
&= \sigma^2[tr\{I_n\} - tr\{X(X'X)^{-1}X'\}] \\
&= \sigma^2[n - tr\{X'X(X'X)^{-1}\}] \\
&= \sigma^2(n-k)
\end{aligned}$$

Therefore, $E[\hat{\sigma}^2] = \frac{n-k}{n}\sigma^2$, and we can also use this to form an unbiased estimator.

The term "model selection" refers to choosing a single model from within a class of models under consideration, in this case, linear regression models.

Here is an example, taken from "The Elements of Statistical Learning" by Hastie, Tibshirani, and Friedman. Stamey et al. (1989) measured the level of prostate-specific antigen and other clinical measures in men who were about to receive a radical prostatectomy. The variables are

`lpsa` = log prostate-specific antigen
`lcavol` = log cancer volume
`lweight` = log prostate weight
`age` = patient's age
`lbph` = log benign prostatic hyperplasia
`svi` = seminal vesicle invasion (categories 0 or 1)
`lcp` = log capsular penetration
`gleason` = Gleason score (categories 6, 7, 8, 9)
`ppg45` = percent of Gleason scores 4 or 5


lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45
lpsa
0.28
0.22
0.54
0.68
0.43
0.43
0.73
0.35
0.44
0.43
0.35
0.27
0.28
0.17
0.18
0.67
0.32
0.46
0.57
0.51
0.63
0.55
0.75
0.37
0.42

Here is some output from fitting a multiple regression model in R, including all the covariates.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.486274   0.885929   0.549  0.58451
lcavol       0.549532   0.090087   6.100 2.94e-08 ***
lweight      0.623816   0.200463   3.112  0.00252 **
age         -0.023103   0.011282  -2.048  0.04364 *
lbph         0.091523   0.058454   1.566  0.12108
svi1         0.744537   0.244602   3.044  0.00310 **
lcp         -0.124612   0.094565  -1.318  0.19109
gleason7     0.253874   0.216141   1.175  0.24341
gleason8     0.480520   0.760895   0.632  0.52938
gleason9    -0.036003   0.494866  -0.073  0.94217
pgg45        0.004902   0.004622   1.061  0.29184
```

We already have one possible tool for model selection, which is hypothesis testing. Particularly if the number of possible regressors is small, we can consider testing $H_0 : \beta_j = 0 \; \forall j \in J_0$ for a set of terms J_0.

Here is an R function for the likelihood ratio test. Can you see what it is doing?

```
lrt <- function(mod, mod0){
  ll <- sum(dnorm(mod$residuals,
            sd = summary(mod)$sigma, log = TRUE))
  ll.0 <- sum(dnorm(mod0$residuals,
              sd = summary(mod0)$sigma, log = TRUE))
  lambda <- 2*(ll-ll.0)
  return(1 - pchisq(lambda, df = mod$rank - mod0$rank))
}
```

```
> full <- lm(lpsa~., data = prostate)
> reduced <- lm(lpsa~lcavol+lweight+svi, data = prostate)
> lrt(full, reduced)
[1] 0.1964859
```

Actually, we do not need the limiting χ^2 distribution in this case; we can modify the test statistic slightly to one that has an exact F distribution. (Similar reasoning applies to using the t distribution for the test statistic for individual regressors.) When the sample size is large, both tests will give similar answers.

```
> anova(full, reduced) # do the F test
...
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     86 41.811
2     93 46.568 -7   -4.7575 1.3979 0.2166
```

Hypothesis testing is not necessarily a good model selection technique, however. Some possible issues are

- Multiple testing: There are many possible combinations of regressors. How should we adjust for exploring different possibilities? Usually we choose which models to test based on looking at preliminary results, which changes the type I error rate.

- Failing to reject the null hypothesis is not the same as finding evidence for the null hypothesis. It may be that the test has low power, e.g. due to small sample size.

- There is no reason to think that decisions based on testing parameters will lead to good predictions, if that is our goal. On the other hand, if model interpretation is our goal, we often have substantive reasons for keeping non-significant regressors in the model.

Consider predicting a new observation Y^*, for covariates X^*, and let S denote a subset of the covariates in the model. If our loss function is squared error, we could choose S to minimize the frequentist risk, which is $MSPE = E[(\hat{Y}^*(S) - Y^*)^2]$, where the expectation is over both the observed Y and the new Y^*. $MSPE$ stands for "mean squared prediction error."

However, $MSPE$ may vary depending on X_*. What Wasserman calls the "prediction risk",

$$R(S) = \sum_{i=1}^{n} E[(\hat{Y}_i(S) - Y_i^*)^2]$$

sums the MSPE at all of observed values of X. This is somewhat similar to integrating MSPE over the distribution of X^*.

A natural estimate for $R(S)$ is the training error

$$\hat{R}_{tr}(S) = \sum_{i=1}^{n} (\hat{Y}_i(S) - Y_i)^2,$$

which is related to R^2 (for a particular S) by

$$R^2(S) = 1 - \frac{\sum_{i=1}^{n} (\hat{Y}_i(S) - Y_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = 1 - \frac{\hat{R}_{tr}(S)}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

However $\hat{R}_{tr}(S)$ (or, equivalently, $R^2(S)$) is bad criterion for model selection. In fact, by construction, $\hat{R}_{tr}(S)$ can only decrease (R^2 can only increase) as we add terms to S, whereas the true $R(S)$ tends to eventually increase. More complicated selection criteria penalize extra model complexity and thereby decrease this bias.

"Adjusted R^2" is reported by many statistical packages. This is just

$$1 - \frac{n-1}{n-k}\frac{RSS}{TSS}$$

where n is the number of observations and k is the number of regressors in the model (the dimension of S). However, there is no particular theory guiding this choice.

Mallow's C_p statistic is

$$\hat{R}(S) = \hat{R}_{tr}(S) + 2k\hat{\sigma}^2$$

where k is the dimension of S and $\hat{\sigma}^2$ is the unbiased estimate of σ^2 *under the full model*. The second term is a bias correction.

The Akaike Information Criterion (AIC) is

$$AIC(S) = \ell_n(\hat{\beta}_S, \hat{\sigma}^2_S) - k$$

where $\hat{\beta}_S$ and $\hat{\sigma}^2_S$ are the MLEs under model S, ℓ_n is the log-likelihood, and k is the number of regressors in S. It is an estimate of the risk function under a different loss function, which is the Kullback Leibler distance between the estimated and true probability distributions.

The Bayesian Information Criterion (BIC) imposes a stronger penalty, with

$$BIC(S) = \ell_n(\hat{\beta}_S, \hat{\sigma}^2_S) - \frac{k}{2}\log n$$

Maximizing BIC is approximately equivalent to choosing the model with highest posterior probability, if the prior distribution assigns equal probability to all models.

Another way of estimating the risk is to set aside a (randomly selected) subset of our data and see how well we can predict it using a model constructed from the rest. In the machine learning literature, these are called the "validation" and "training" data sets.

The risk estimator is

$$\hat{R}_V(S) = \sum_{i=1}^{m} (\hat{Y}_i^*(S) - Y_i^*)^2$$

where m is the size of the validation data set. How big m is will depend on how much data you have "to spare" from your total data set. Common practice is to take m to be $1/4$ to $1/2$ the total data size.

This approach is really only feasible if you have a lot of data to begin with, otherwise the parameters in each model may not be well estimated.

An adaptation of this idea uses almost all the data for fitting, but it approximates the risk by repeatedly fitting the model to all but one data point. This is called leave-one-out cross-validation. The risk estimator is

$$\hat{R}_{CV}(S) = \sum_{i=1}^{n} (Y_i - \hat{Y}_{(i)})^2$$

where $\hat{Y}_{(i)}$ is the prediction for Y_i obtained by fitting the model using all the data *except for* the i^{th} observation.

Luckily, we don't actually have to recalculate the model n times, since

$$\hat{R}_{CV}(S) = \sum_{i=1}^{n} \left(\frac{Y_i - \hat{Y}_i(S)}{1 - U_{ii}(S)} \right)^2$$

where $U(S) = X_S(X_S'X_S)^{-1}X_S$.

When the number of possible regressors is large, it can be computationally infeasible to calculate a criterion for all the possible models. Two popular methods for exploring the space of models are the forward and backward stepwise selection procedures.

Forward selection starts with the intercept-only model. The criterion is then computed for each model with a single covariate, and the one with the best possible criterion is chosen. (The way we've defined each criterion, smaller is better.) Then the criterion is computed for each model with that covariate plus another possible covariate, and so on. Backward selection starts with the full model, with all possible covariates, and at each step deletes the one that leads to the most improvement.

They will not necessarily end up in the same place, and neither is guaranteed to find the best out of all possible models.

Nonparametric Regression

Consider observing

$$Y_i = r(x_i) + \epsilon_i,$$

where the ϵ_i's are iid with $E[\epsilon_i] = 0$ and $V[\epsilon_i] = \sigma^2$. The function r is unknown, and we want to estimate it under minimal assumptions.

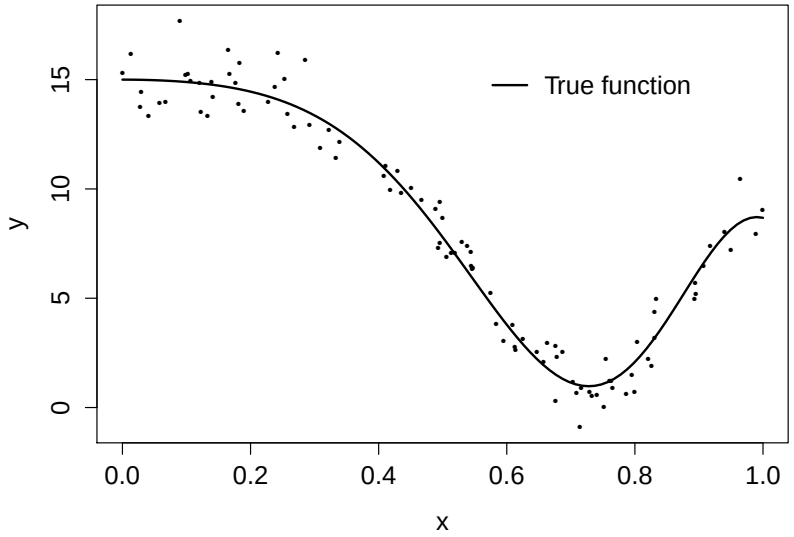
When $r(x) = E[Y|X = x]$, this extends the linear regression model we looked at recently.

(For simplicity we'll only consider the univariate case, with $x_i \in \mathbb{R}$).

Finding a good estimate of $r(x)$ will involve understanding the role of *smoothing* in the *bias-variance tradeoff*. The methods we'll consider fall into two categories:

1. "Localize" the problem: use observations close to x to estimate $r(x)$.
2. Turn the problem back into something we know how to solve: multiple regression, using orthonormal functions in x.

Example:



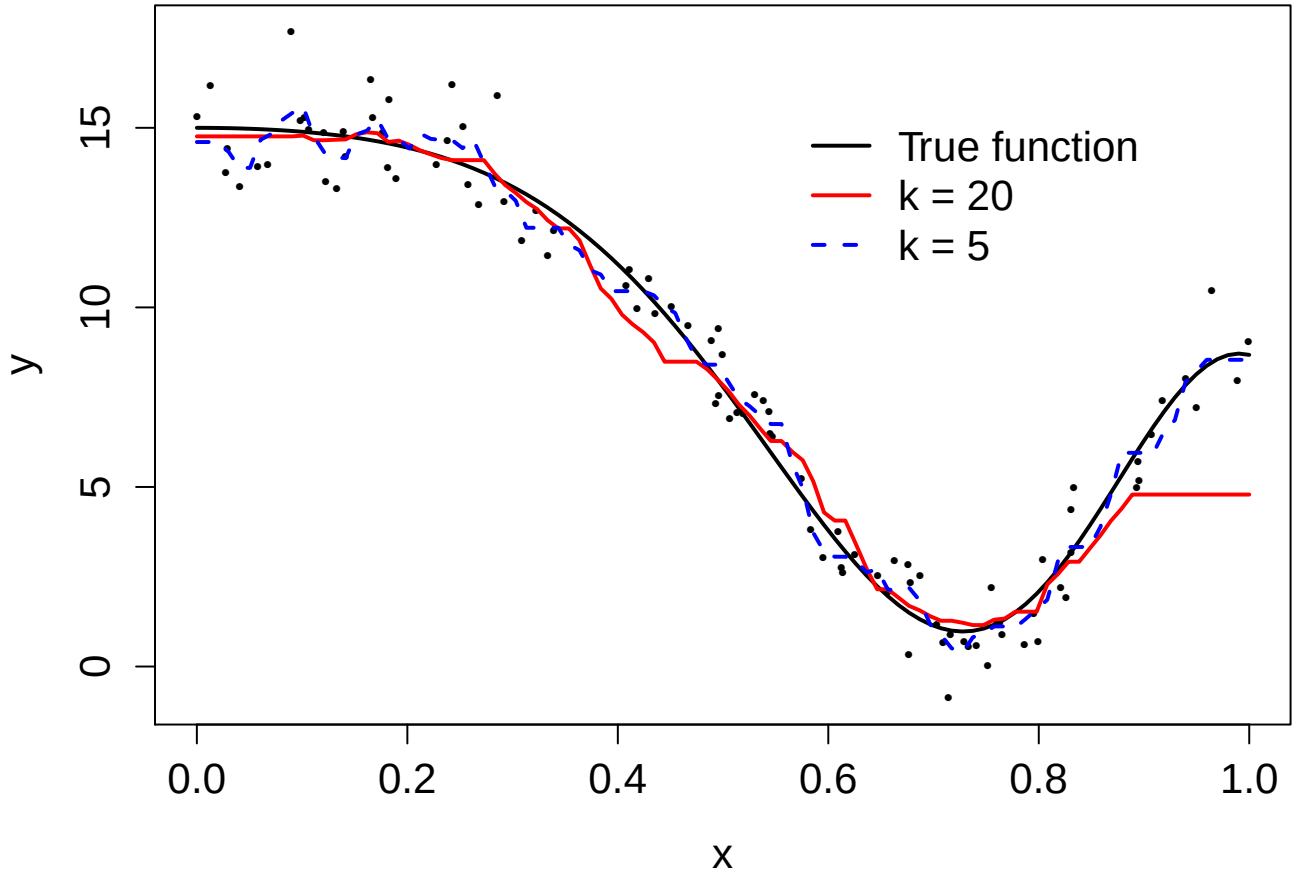
One simple but popular method for estimating r is called $k-$nearest neighbors. Let $N_k(x)$ denote the k values of $x_1, \ldots, x_n$ that are closest to x. Then the estimator is just the mean over the corresponding Y_i values:

$$\hat{r}(x) = \frac{1}{k} \sum_{i: x_i \in N_k(x)} Y_i$$

The motivation for this estimator is that nearby observations "stand in" for repeated samples at a particular x. The function $r(x)$ is treated as if it were constant over the neighborhood $N_k(x)$.

Although this is not usually true, it may be a good approximation if r is smooth and the neighborhood is small.

The degree of smoothing depends on k.



Let f be the function we're trying to estimate. As with the ECDF $\hat{F}_n(x)$, in studying a particular estimator $\hat{f}_n(x)$, we need to keep track of two things that can vary: the observed data used to construct $\hat{f}_n$, and the value of x at which we evaluate the function.

We will (primarily for tractability) use the integrated squared error (ISE) as our loss function:

$$L(f, \hat{f}_n) = \int [f(x) - \hat{f}_n(x)]^2 g(x) dx,$$

where $g(x)$ is pdf for X.

This gives us a frequentist risk function

$$R(f, \hat{f}_n) = E[L(f, \hat{f}_n)]$$

We can rewrite the risk as

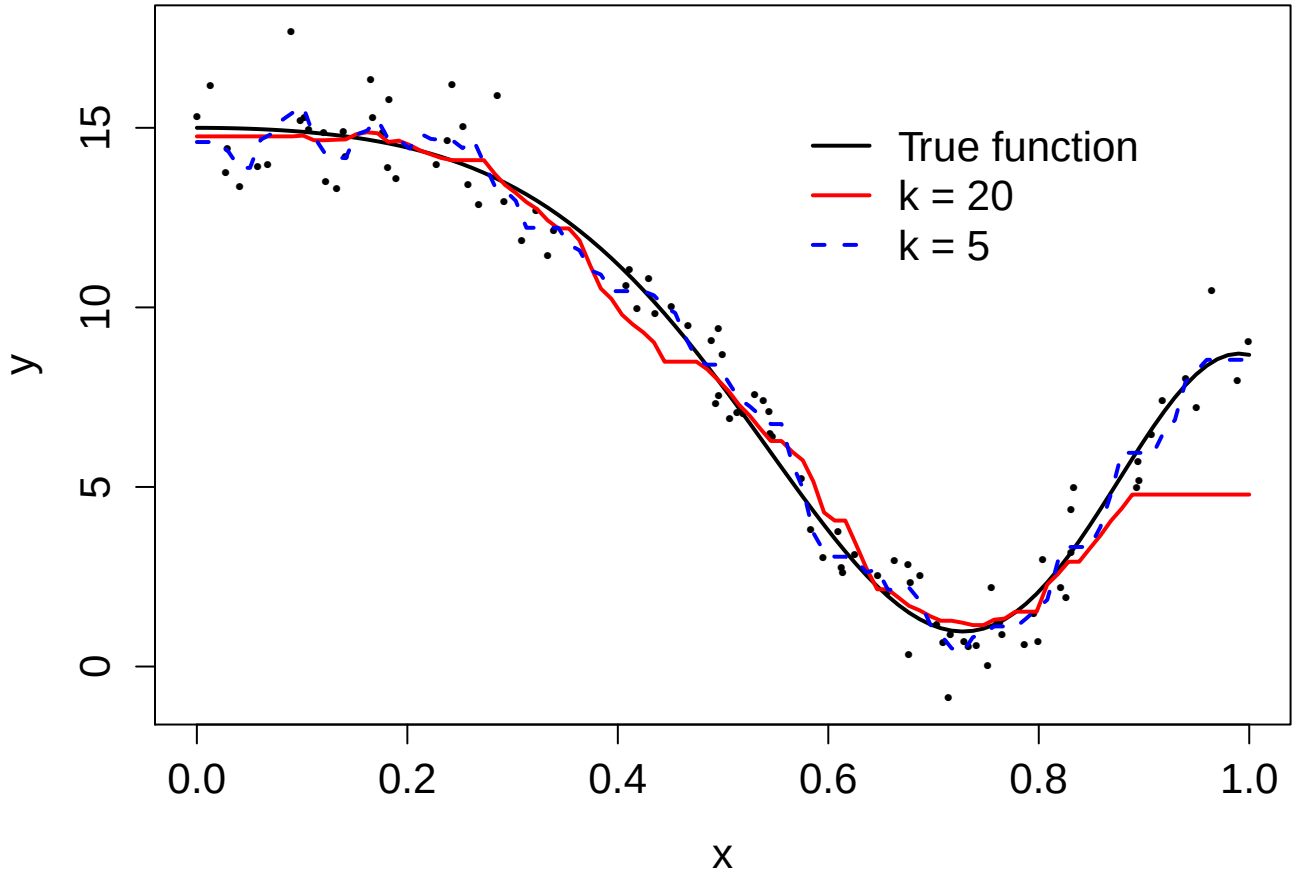
$$R(f, \hat{f}_n) = \int b^2(x)g(x)dx + \int v(x)g(x)dx$$

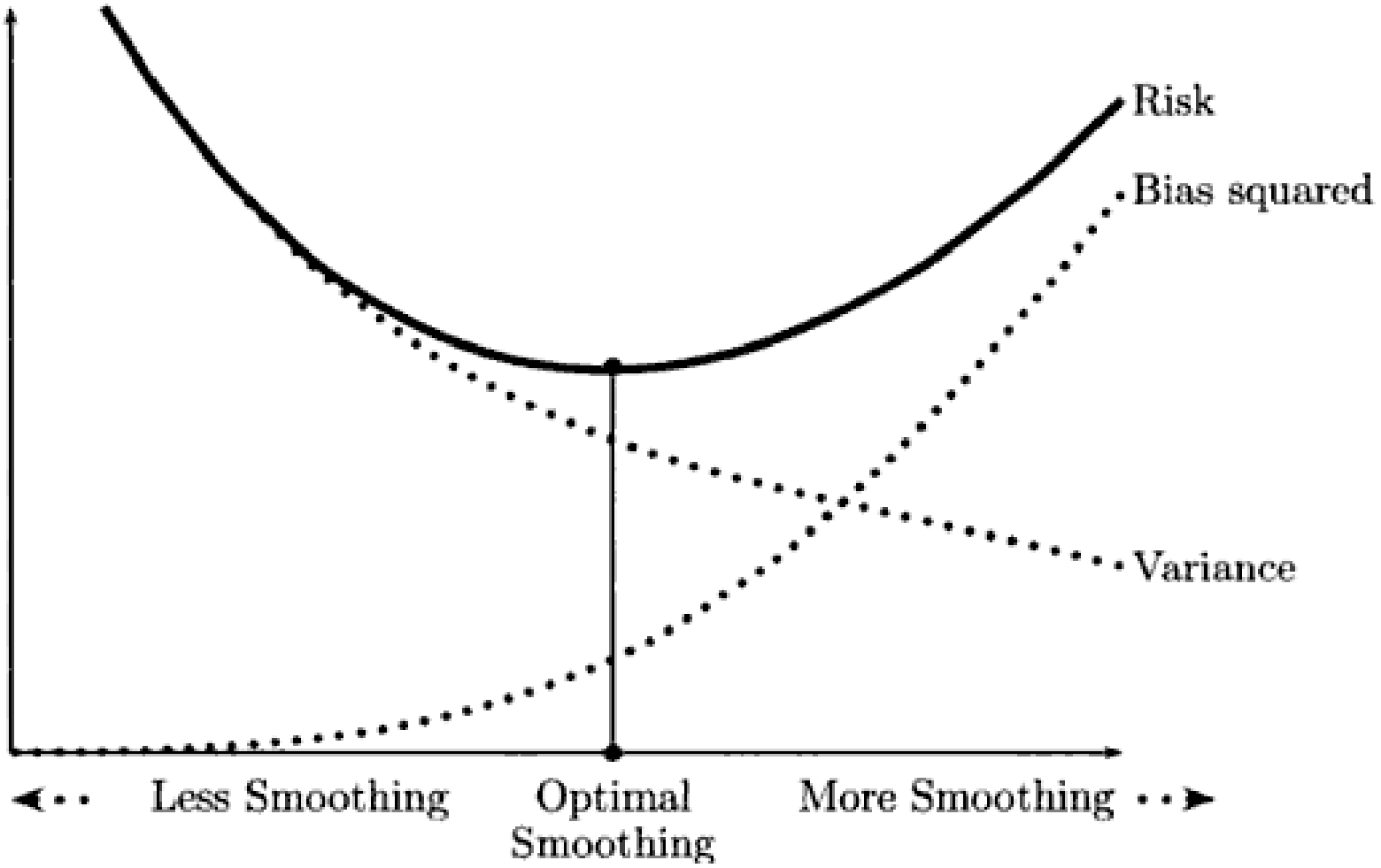
where

$$\begin{aligned} b(x) &= E[\hat{f}_n(x)] - f(x) \\ v(x) &= V[\hat{f}_n(x)] \end{aligned}$$

In the cases we will study, $\hat{f}_n$ will depend not only on the data, but also on a choice of *smoothing parameter*. We want to choose the smoothing parameter to minimize $R(f, \hat{f}_n)$. Typically we have that $b(x)$ increases with more smoothing and $v(x)$ decreases, so that the sum $R(f, \hat{f}_n)$ involves a *tradeoff* between bias and variance.

The degree of smoothing depends on k.




Risk
Bias squared
Variance
Less Smoothing
Optimal Smoothing
More Smoothing

One drawback to the $k-$nearest neighbor estimator is that it is discontinuous. We can think of this estimator as a weighted average, where each Y_i is given a weight of either zero or $1/k$, depending on whether it's in the neighborhood.

The Nadaraya-Watson kernel estimator allows the weights to decay smoothly with distance.

$$\hat{r}(x) = \sum_{i=1}^{n} w_i(x) Y_i$$

where K is a kernel and

$$w_i(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)}$$

A kernel is a function K such that

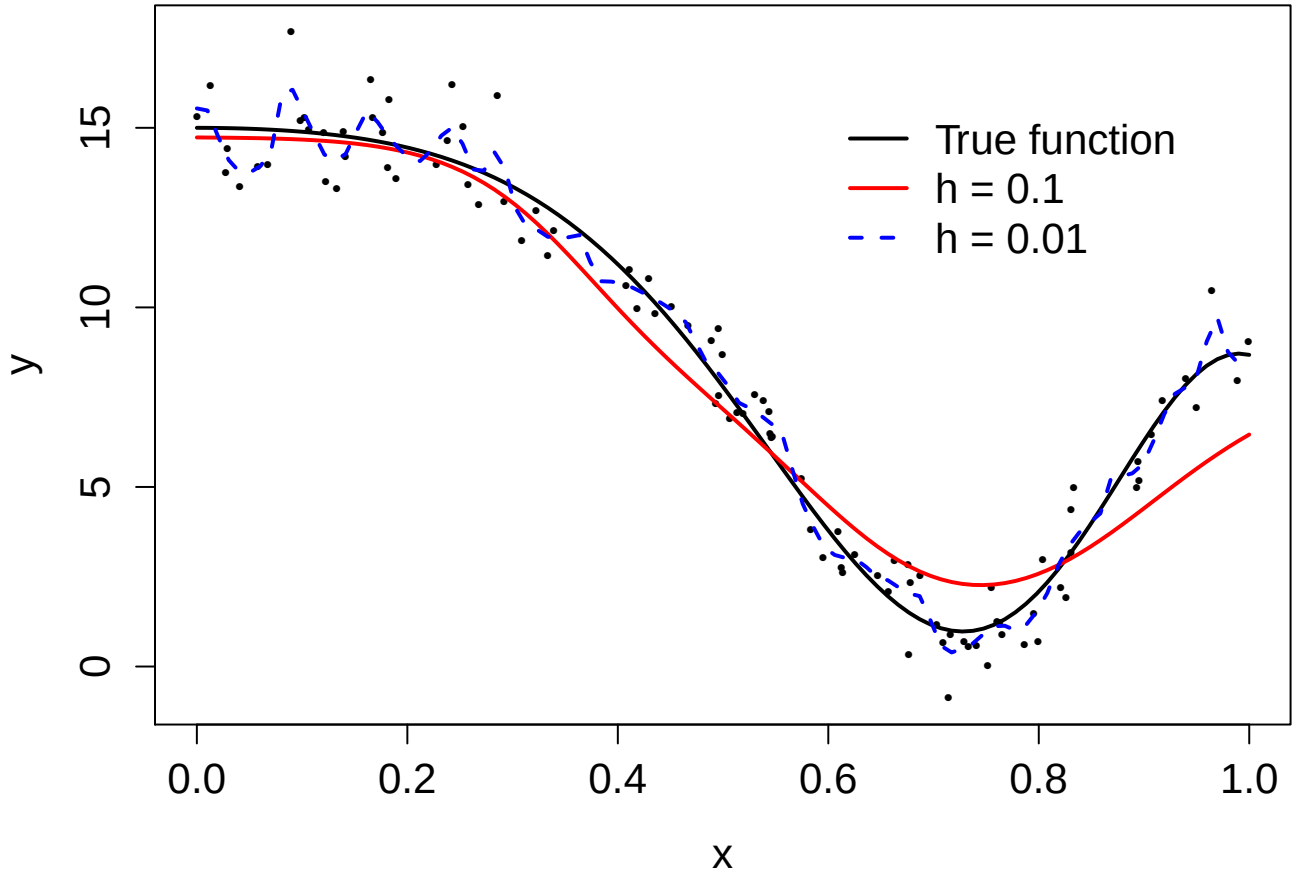
- $K(x) \geq 0$
- $\int K(x)dx = 1$
- $\int xK(x)dx = 0$
- $\int x^2K(x)dx \equiv \sigma_K^2 > 0$

The Nadaraya-Watson kernel estimator

$$\hat{r}(x) = \sum_{i=1}^{n} \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)} Y_i,$$

where h is called the bandwidth. The choice of h matters more than the functional form of K. One popular kernel function is Gaussian kernel.

The degree of smoothing depends on h.



We can choose the bandwidth h to minimize the cross-validation estimate of the risk (using squared error loss)

$$\begin{aligned}\hat{J}(h) &= \sum_{i=1}^{n}(Y_i - \hat{r}_{-i}(x_i))^2 \\ &= \sum_{i=1}^{n}\frac{(Y_i - \hat{r}(x_i))^2}{\left(1 - \frac{K(0)}{\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right)}\right)^2}\end{aligned}$$

(See `npreg.R` for implementation.)

Now consider a completely different idea, appropriate for functions r in

$$L_2(a,b) = \left\{ f : [a,b] \to \mathbb{R}, \quad \int_a^b f(x)^2 dx < \infty \right\}$$

These functions may be written as

$$r(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x), \quad \text{where } \beta_j = \int_a^b r(x)\phi_j(x)dx$$

and the functions $\phi_1, \phi_2, \ldots$ form an *orthonormal basis* for $L_2(a,b)$.

Before we define an orthonormal basis, note what we've done: we've given ourselves a way to approximate

$$\begin{aligned} r(x) &= \sum_{j=1}^{\infty} \beta_j \phi_j(x) \\ &\approx \sum_{j=1}^{J} \beta_j \phi_j(x) \end{aligned}$$

This approximation is accurate when r is smooth, since in this case β_j is small for large j.

Since we don't know the true function, we can't calculate β_j exactly, but we can estimate it, treating $\phi_1(x), \ldots, \phi_J(x)$ as covariates in a multiple regression.

A sequence of functions $\phi_1, \phi_2, \phi_3, \ldots$ forms an *orthonormal basis* for $L_2(a, b)$ if

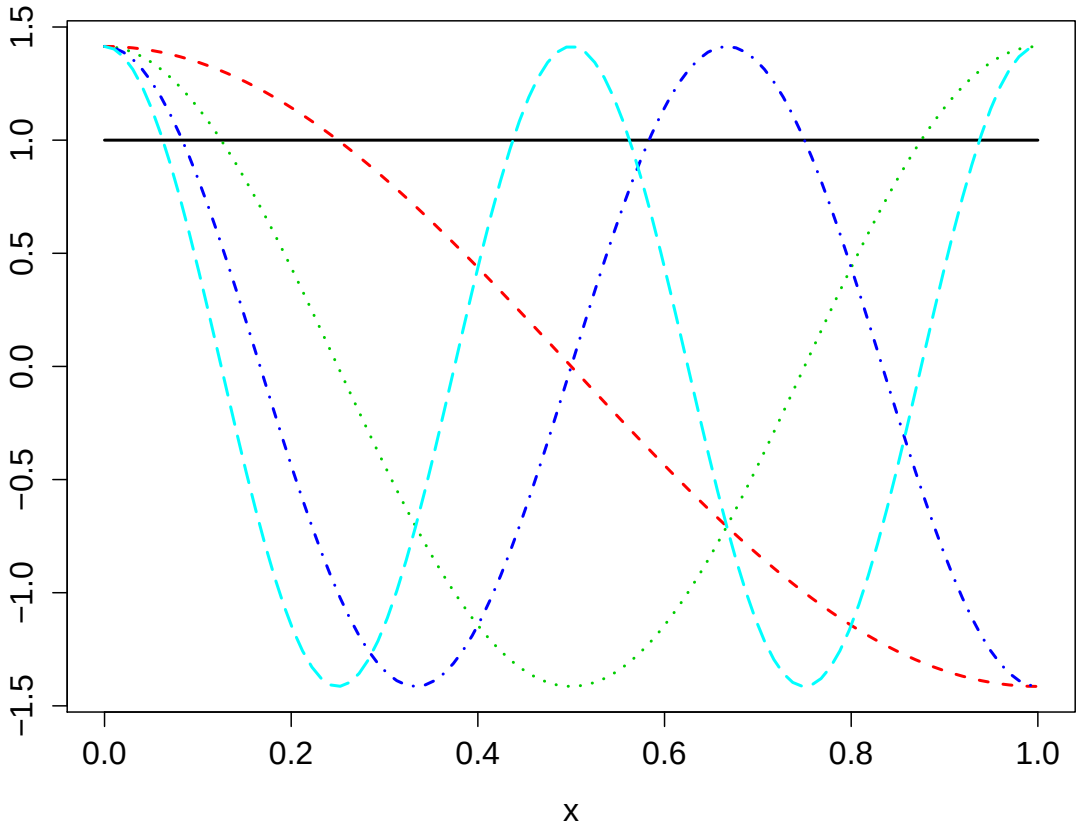
- $\int_a^b \phi_i(x)\phi_j(x)dx = 0$ for $i \neq j$ (orthogonal)
- $\int_a^b \phi_i(x)dx = 1$ for all i (normal)
- The only function orthogonal to each $\phi_j(x)$ is the zero function (complete)

For example, the *cosine basis* for $L_2(0, 1)$ takes $\phi_0(x) = 1$ and

$$\phi_j(x) = \sqrt{2}\cos(j\pi x)$$

for $j \geq 1$.

First five cosine basis functions on $[0, 1]$



The Legendre polynomials on $[-1, 1]$ are given by the recursive relationship $P_0(x) = 1$, $P_1(x) = x$, and

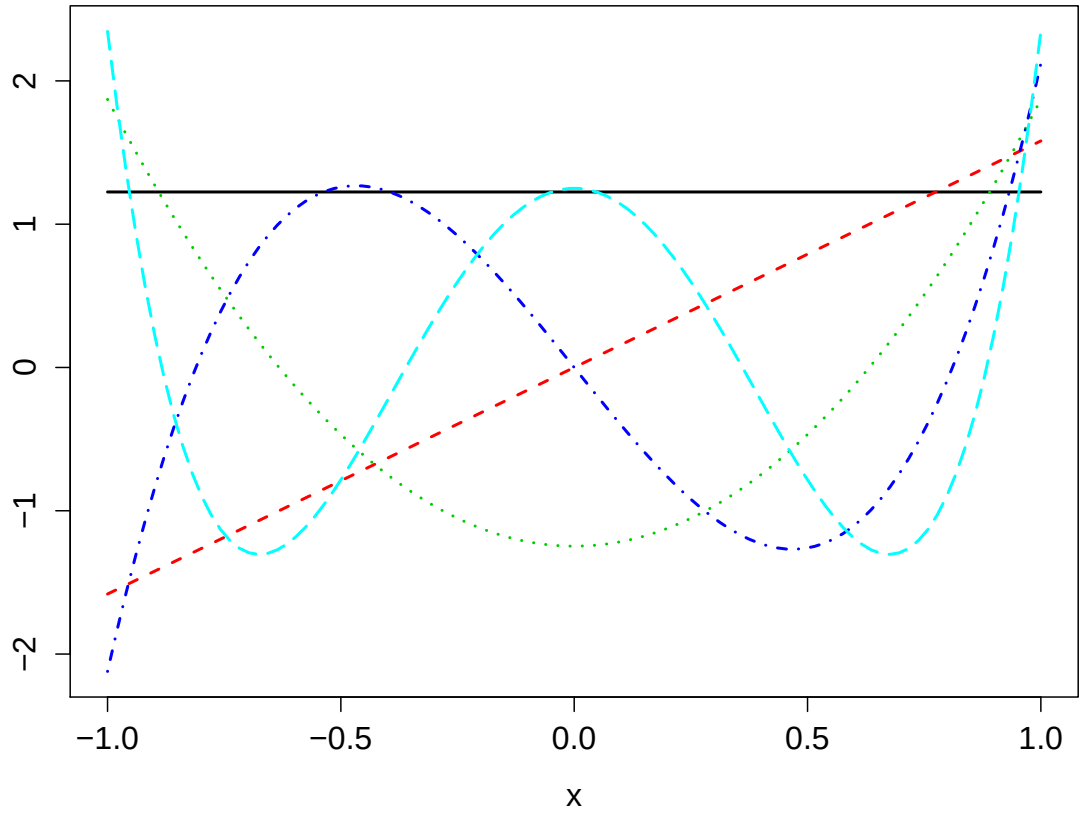
$$P_{j+1}(x) = \frac{(2j+1)xP_j(x) - jP_{j-1}(x)}{j+1}$$

They can be used to form an orthonormal basis for $L_2(-1, 1)$, setting

$$\phi_j(x) = \sqrt{(2j+1)/2}P_j(x)$$

(Note: The particular interval doesn't really matter, since it's easy to rescale the x's.)

First five Legendre polynomial basis functions on $[-1, 1]$



Recall our approximation $r(x) \approx \sum_{j=1}^{J} \beta_j \phi_j(x)$. For a particular choice of $J \leq n$, define

$$\Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_J(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \cdots & \phi_J(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_n) & \phi_1(x_n) & \cdots & \phi_J(x_n) \end{pmatrix}$$

We have recast the nonparametric regression problem $Y_i = r(x_i) + \epsilon_i$ as the multivariate regression

$$Y = \Phi\beta + \eta$$

where $\eta_i = \epsilon_i$ plus some error from the approximation, hopefully small.

The least squares estimator of β is just

$$\hat{\beta} = (\Phi'\Phi)^{-1}\Phi'Y$$