STAT 801: Mathematical Statistics

Course outline:

- Distribution Theory.
  - Basic concepts of probability.
  - Distributions
  - Expectation and moments
  - Transforms (such as characteristic functions, moment generating functions)
  - Distribution of transformations

- Point estimation
  - Maximum likelihood estimation.
  - Method of moments.
  - Optimality Theory.
  - Bias, mean squared error.
  - Sufficiency.
  - Uniform Minimum Variance (UMV) Unbiased Estimators.

- Hypothesis Testing
  - Neyman Pearson optimality theory.
  - Most Powerful, Uniformly Most Powerful, Unbiased tests.
- Confidence sets
  - Pivots
  - Associated Hypothesis Tests
  - Inversion of hypothesis tests to get confidence sets.
- Decision Theory.

Statistics versus Probability

Standard view of scientific inference has a set of theories which make predictions about the outcomes of an experiment:

Theory	Prediction
A	1
B	2
C	3

Conduct experiment, see outcome 2: **infer** B is correct (or at least A and C are wrong).

Add **Randomness**

Theory	Prediction
A	Usually 1 sometimes 2 never 3
B	Usually 2 sometimes 1 never 3
C	Usually 3 sometimes 1 never 2

See outcome 2: infer Theory B probably correct, Theory A probably not correct, Theory C is wrong.

Probability Theory: construct table: compute likely outcomes of experiments.

Statistics: inverse process. Use table to draw inferences from outcome of experiment. How should we do it and how wrong are our inferences likely to be? Notice: hopeless task unless different theories make different predictions.

Start with Probability; switch after about 5 weeks to statistics.

Probability Definitions

Probability Space (or **Sample Space**): ordered triple $(\Omega, \mathcal{F}, P)$.

- Ω is a set (possible outcomes); elements are ω called elementary outcomes.

- $\mathcal{F}$ is a family of subsets (**events**) of Ω with the property that $\mathcal{F}$ is a σ-field (or Borel field or σ-algebra):

  1. Empty set $\emptyset$ and Ω are members of $\mathcal{F}$.

  2. $A \in \mathcal{F}$ implies $A^c = \{\omega \in \Omega : \omega \not\in A\} \in \mathcal{F}$.

  3. $A_1, A_2, \cdots$ in $\mathcal{F}$ implies $A = \cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

- P a function, domain $\mathcal{F}$, range a subset of $[0,1]$ satisfying:

  1. $P(\emptyset) = 0$ and $P(\Omega) = 1$.

  2. **Countable additivity**: $A_1, A_2, \cdots$ **pairwise disjoint** $(j \neq k \; A_j \cap A_k = \emptyset)$

  $$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Axioms guarantee can compute probabilities by usual rules, including approximation.

Consequences of axioms:

$$A_i \in \mathcal{F}; i = 1, 2, \cdots \text{ implies } \cap_i A_i \in \mathcal{F}$$

$$A_1 \subset A_2 \subset \cdots \text{ implies } P(\cup A_i) = \lim_{n\to\infty} P(A_n)$$

$$A_1 \supset A_2 \supset \cdots \text{ implies } P(\cap A_i) = \lim_{n\to\infty} P(A_n)$$

Vector valued random variable: function $X : \Omega \mapsto R^p$ such that, writing $X = (X_1, \ldots, X_p)$,

$$P(X_1 \le x_1, \ldots, X_p \le x_p)$$

is defined for any constants $(x_1, \ldots, x_p)$. Formally the notation

$$X_1 \le x_1, \ldots, X_p \le x_p$$

is a subset of Ω or **event**:

$$\{\omega \in \Omega : X_1(\omega) \le x_1, \ldots, X_p(\omega) \le x_p\} .$$

Remember X is a function on Ω so X_1 is also a function on Ω.

In almost all of probability and statistics the dependence of a random variable on a point in the probability space is hidden! You almost always see X not $X(\omega)$.

Borel σ-field in R^p: smallest σ-field in R^p containing every open ball.

Every common set is a Borel set, that is, in the Borel σ-field.

An R^p valued **random variable** is a map $X : \Omega \mapsto R^p$ such that when A is Borel then $\{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F}$.

Fact: this is equivalent to

$$\{\omega \in \Omega : X_1(\omega) \le x_1, \ldots, X_p(\omega) \le x_p\} \in \mathcal{F}$$

for all $(x_1, \ldots, x_p) \in R^p$.

Jargon and notation: we write $P(X \in A)$ for $P(\{\omega \in \Omega : X(\omega) \in A\})$ and define the **distribution** of X to be the map

$$A \mapsto P(X \in A)$$

which is a probability on the set R^p with the Borel σ-field rather than the original Ω and $\mathcal{F}$.

Cumulative Distribution Function (CDF) of X: function F_X on R^p defined by

$$F_X(x_1,\ldots,x_p) = P(X_1 \le x_1,\ldots,X_p \le x_p)\,.$$

Properties of F_X (usually just F) for $p = 1$:

1. $0 \le F(x) \le 1$.

2. $x > y \Rightarrow F(x) \ge F(y)$ (monotone non-decreasing).

3. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

4. $\lim_{x\searrow y} F(x) = F(y)$ (right continuous).

5. $\lim_{x\nearrow y} F(x) \equiv F(y-)$ exists.

6. $F(x) - F(x-) = P(X = x)$.

7. $F_X(t) = F_Y(t)$ for all t implies that X and Y have the same distribution, that is, $P(X \in A) = P(Y \in A)$ for any (Borel) set A.

Defn: Distribution of rv X is **discrete** (also call X discrete) if $\exists$ countable set $x_1, x_2, \cdots$ such that

$$P(X \in \{x_1, x_2 \cdots\}) = 1 = \sum_i P(X = x_i) \,.$$

In this case the **discrete density** or **probability mass function** of X is

$$f_X(x) = P(X = x) \,.$$

Defn: Distribution of rv X is **absolutely continuous** if there is a function f such that

$$P(X \in A) = \int_A f(x) dx \tag{1}$$

for any (Borel) set A. This is a p dimensional integral in general. Equivalently

$$F(x) = \int_{-\infty}^{x} f(y) \, dy \,.$$

Defn: Any f satisfying (1) is a **density** of X.

For most x F is differentiable at x and

$$F'(x) = f(x) \,.$$

Example: X is Uniform[0,1].

$$F(x) = \begin{cases} 0 & x \le 0 \\ x & 0 < x < 1 \\ 1 & x \ge 1\,. \end{cases}$$

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ \text{undefined} & x \in \{0, 1\} \\ 0 & \text{otherwise}\,. \end{cases}$$

Example: X is exponential.

$$F(x) = \begin{cases} 1 - e^{-x} & x > 0 \\ 0 & x \le 0\,. \end{cases}$$

$$f(x) = \begin{cases} e^{-x} & x > 0 \\ \text{undefined} & x = 0 \\ 0 & x < 0\,. \end{cases}$$

Distribution Theory

Basic Problem:

Start with assumptions about f or CDF of random vector $X = (X_1, \ldots, X_p)$.

Define $Y = g(X_1, \ldots, X_p)$ to be some function of X (usually some statistic of interest).

Compute distribution or CDF or density of Y?

Univariate Techniques

Method 1: compute the CDF by integration and differentiate to find f_Y.

Example: $U \sim \text{Uniform}[0,1]$ and $Y = -\log U$.

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(-\log U \le y) \\ &= P(\log U \ge -y) = P(U \ge e^{-y}) \\ &= \begin{cases} 1 - e^{-y} & y > 0 \\ 0 & y \le 0 \end{cases} \end{aligned}$$

so Y has standard exponential distribution.

Example: $Z \sim N(0,1)$, i.e.

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

and $Y = Z^2$. Then

$$\begin{aligned} F_Y(y) &= P(Z^2 \le y) \\ &= \begin{cases} 0 & y < 0 \\ P(-\sqrt{y} \le Z \le \sqrt{y}) & y \ge 0\,. \end{cases} \end{aligned}$$

Now differentiate

$$P(-\sqrt{y} \le Z \le \sqrt{y}) = F_Z(\sqrt{y}) - F_Z(-\sqrt{y})$$

to get

$$f_Y(y) = \begin{cases} 0 & y < 0 \\ \frac{d}{dy}\left[F_Z(\sqrt{y}) - F_Z(-\sqrt{y})\right] & y > 0 \\ \text{undefined} & y = 0\,. \end{cases}$$

Then

$$\begin{aligned}
\frac{d}{dy}F_Z(\sqrt{y}) &= f_Z(\sqrt{y})\frac{d}{dy}\sqrt{y} \\
&= \frac{1}{\sqrt{2\pi}}\exp\left(-\left(\sqrt{y}\right)^2/2\right)\frac{1}{2}y^{-1/2} \\
&= \frac{1}{2\sqrt{2\pi y}}e^{-y/2}\,.
\end{aligned}$$

(Similar formula for other derivative.) Thus

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi y}}e^{-y/2} & y > 0 \\ 0 & y < 0 \\ \text{undefined} & y = 0\,. \end{cases}$$

We will find **indicator** notation useful:

$$1(y > 0) = \begin{cases} 1 & y > 0 \\ 0 & y \le 0 \end{cases}$$

which we use to write

$$f_Y(y) = \frac{1}{\sqrt{2\pi y}}e^{-y/2}1(y > 0)$$

(changing definition unimportantly at $y = 0$).

Notice: I never evaluated F_Y before differentiating it. In fact F_Y and F_Z are integrals I can't do but I can differentiate them anyway. Remember fundamental theorem of calculus:

$$\frac{d}{dx} \int_a^x f(y)\, dy = f(x)$$

at any x where f is continuous.

Summary: for $Y = g(X)$ with X and Y each real valued

$$\begin{aligned} P(Y \le y) &= P(g(X) \le y) \\ &= P(X \in g^{-1}(-\infty, y])\,. \end{aligned}$$

Take d/dy to compute the density

$$f_Y(y) = \frac{d}{dy} \int_{\{x: g(x) \le y\}} f_X(x)\, dx\,.$$

Often can differentiate without doing integral.

Method 2: Change of variables.

Assume g is one to one. I do: g is increasing and differentiable. Interpretation of density (based on density $= F'$):

$$\begin{aligned} f_Y(y) &= \lim_{\delta y \to 0} \frac{P(y \le Y \le y + \delta y)}{\delta y} \\ &= \lim_{\delta y \to 0} \frac{F_Y(y+\delta y) - F_Y(y)}{\delta y} \end{aligned}$$

and

$$f_X(x) = \lim_{\delta x \to 0} \frac{P(x \le X \le x + \delta x)}{\delta x}.$$

Assume $y = g(x)$. Define δy by $y + \delta y = g(x + \delta x)$. Then

$$P(y \le Y \le y + \delta y) = P(x \le X \le x + \delta x).$$

Get

$$\frac{P(y \le Y \le y + \delta y))}{\delta y} = \frac{P(x \le X \le x + \delta x)/\delta x}{\{g(x+\delta x) - y\}/\delta x}.$$

Take limit to get

$$f_Y(y) = f_X(x)/g'(x)$$

or

$$f_Y(g(x))g'(x) = f_X(x).$$

Alternative view:

Each probability is integral of a density:

First is integral of f_Y over the small interval from $y = g(x)$ to $y = g(x + \delta x)$. The interval is narrow so f_Y is nearly constant and

$$P(y \le Y \le g(x+\delta x)) \approx f_Y(y)(g(x+\delta x)-g(x)) .$$

Since g has a derivative the difference

$$g(x + \delta x) - g(x) \approx \delta x g'(x)$$

and we get

$$P(y \le Y \le g(x + \delta x)) \approx f_Y(y)g'(x)\delta x .$$

Same idea applied to $P(x \le X \le x + \delta x)$ gives

$$P(x \le X \le x + \delta x) \approx f_X(x)\delta x$$

so that

$$f_Y(y)g'(x)\delta x \approx f_X(x)\delta x$$

or, cancelling the δx in the limit

$$f_Y(y)g'(x) = f_X(x) .$$

If you remember $y = g(x)$ then you get

$$f_X(x) = f_Y(g(x))g'(x)\,.$$

Or solve $y = g(x)$ to get x in terms of y, that is, $x = g^{-1}(y)$ and then

$$f_Y(y) = f_X(g^{-1}(y))/g'(g^{-1}(y))$$

This is just the change of variables formula for doing integrals.

Remark: For g decreasing $g' < 0$ but then the interval $(g(x), g(x+\delta x))$ is really $(g(x+\delta x), g(x))$ so that $g(x) - g(x+\delta x) \approx -g'(x)\delta x$. In both cases this amounts to the formula

$$f_X(x) = f_Y(g(x))|g'(x)|\,.$$

Mnemonic:

$$f_Y(y)dy = f_X(x)dx\,.$$

Example: $X \sim \text{Weibull(shape } \alpha\text{, scale } \beta)$ or

$$f_X(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left\{-(x/\beta)^\alpha\right\} 1(x > 0)\,.$$

Let $Y = \log X$ or $g(x) = \log(x)$.

Solve $y = \log x$: $x = \exp(y)$ or $g^{-1}(y) = e^y$.

Then $g'(x) = 1/x$ and

$$1/g'(g^{-1}(y)) = 1/(1/e^y) = e^y\,.$$

Hence

$$f_Y(y) = \frac{\alpha}{\beta}\left(\frac{e^y}{\beta}\right)^{\alpha-1} \exp\left\{-(e^y/\beta)^\alpha\right\} 1(e^y > 0)e^y\,.$$

For any y, $e^y > 0$ so indicator = 1. So

$$f_Y(y) = \frac{\alpha}{\beta^\alpha} \exp\left\{\alpha y - e^{\alpha y}/\beta^\alpha\right\}\,.$$

Define $\phi = \log\beta$ and $\theta = 1/\alpha$; then,

$$f_Y(y) = \frac{1}{\theta} \exp\left\{\frac{y-\phi}{\theta} - \exp\left\{\frac{y-\phi}{\theta}\right\}\right\}\,.$$

Extreme Value density with **location** parameter ϕ and **scale** parameter θ. (Note: several distributions are called Extreme Value.)

Marginalization

Simplest multivariate problem:

$$X = (X_1, \ldots, X_p), \qquad Y = X_1$$

(or in general Y is any X_j).

Theorem 1 *If X has density $f(x_1, \ldots, x_p)$ and $q < p$ then $Y = (X_1, \ldots, X_q)$ has density*

$$f_Y(x_1, \ldots, x_q) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_p)\, dx_{q+1} \ldots dx_p$$

$f_{X_1,\ldots,X_q}$ is the **marginal** density of $X_1, \ldots, X_q$ and f_X the **joint** density of X but they are both just densities. "Marginal" just to distinguish from the joint density of X.

Example: The function

$f(x_1, x_2) = Kx_1x_2 1(x_1 > 0, x_2 > 0, x_1 + x_2 < 1)$

is a density provided

$$P(X \in R^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2)\, dx_1\, dx_2 = 1\,.$$

The integral is

$$\begin{aligned} K\int_0^1\int_0^{1-x_1} x_1x_2\, dx_2\, dx_1 & \\ &= K\int_0^1 x_1(1-x_1)^2\, dx_1/2 \\ &= K(1/2 - 2/3 + 1/4)/2 \\ &= K/24 \end{aligned}$$

so $K = 24$. The marginal density of x_1 is

$$\begin{aligned} f_{X_1}(x_1) = & \int_{-\infty}^{\infty} 24x_1x_2 \\ & \quad \times 1(x_1 > 0, x_2 > 0, x_1 + x_2 < 1)\, dx_2 \\ =& 24\int_0^{1-x_1} x_1x_2 1(0 < x_1 < 1) dx_2 \\ =& 12x_1(1-x_1)^2 1(0 < x_1 < 1)\,. \end{aligned}$$

This is a Beta$(2, 3)$ density.

General case: $Y = (Y_1, \ldots, Y_q)$ with components $Y_i = g_i(X_1, \ldots, X_p)$.

Case 1: $q > p$.

Y **won't** have density for "smooth" g.

Y will have a **singular** or discrete distribution. Problem rarely of real interest.

But, e.g., residuals in regression problems have singular distribution.

Case 2: $q = p$.

Use change of variables formula which generalizes the one derived above for the case $p = q = 1$. (See below.)

Case 3: $q < p$.

Pad out Y: add on $p - q$ more variables (carefully chosen) say $Y_{q+1}, \ldots, Y_p$.

Find functions $g_{q+1}, \ldots, g_p$. Define for $q < i \le p$, $Y_i = g_i(X_1, \ldots, X_p)$ and $Z = (Y_1, \ldots, Y_p)$.

Choose g_i so that we can use change of variables on $g = (g_1, \ldots, g_p)$ to compute f_Z.

Find f_Y by integration:

$$f_Y(y_1,\ldots,y_q) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_Z(y_1,\ldots,y_q,z_{q+1},\ldots,z_p) dz_{q+1} \ldots dz_p$$

Change of Variables

Suppose $Y = g(X) \in R^p$ with $X \in R^p$ having density f_X. **Assume g is a one to one ("injective") map,** i.e., $g(x_1) = g(x_2)$ if and only if $x_1 = x_2$. Find f_Y:

Step 1: Solve for x in terms of y: $x = g^{-1}(y)$.

Step 2: Use basic equation:

$$f_Y(y)dy = f_X(x)dx$$

and rewrite it in the form

$$f_Y(y) = f_X(g^{-1}(y))\frac{dx}{dy}$$

Interpretation of derivative $\frac{dx}{dy}$ when $p > 1$:

$$\frac{dx}{dy} = \left|\det\left(\frac{\partial x_i}{\partial y_j}\right)\right|$$

which is the so called **Jacobian**.

Equivalent formula inverts the matrix:

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{\left|\frac{dy}{dx}\right|}.$$

This notation means

$$\left|\frac{dy}{dx}\right| = \left|\det\begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_p} \\ & \vdots & & \\ \frac{\partial y_p}{\partial x_1} & \frac{\partial y_p}{\partial x_2} & \cdots & \frac{\partial y_p}{\partial x_p} \end{bmatrix}\right|$$

but with x replaced by the corresponding value of y, that is, replace x by $g^{-1}(y)$.

Example: The density

$$f_X(x_1, x_2) = \frac{1}{2\pi} \exp\left\{-\frac{x_1^2 + x_2^2}{2}\right\}$$

is the **standard bivariate normal density**. Let $Y = (Y_1, Y_2)$ where $Y_1 = \sqrt{X_1^2 + X_2^2}$ and $0 \le Y_2 < 2\pi$ is angle from the positive x axis to the ray from the origin to the point (X_1, X_2). I.e., Y is X in polar co-ordinates.

Solve for x in terms of y:

$$\begin{aligned} X_1 &= Y_1 \cos(Y_2) \\ X_2 &= Y_1 \sin(Y_2) \end{aligned}$$

so that

$$\begin{aligned} g(x_1, x_2) &= (g_1(x_1, x_2), g_2(x_1, x_2)) \\ &= (\sqrt{x_1^2 + x_2^2}, \text{argument}(x_1, x_2)) \\ g^{-1}(y_1, y_2) &= (g_1^{-1}(y_1, y_2), g_2^{-1}(y_1, y_2)) \\ &= (y_1 \cos(y_2), y_1 \sin(y_2)) \\ \left|\frac{dx}{dy}\right| &= \left|\det \begin{pmatrix} \cos(y_2) & -y_1 \sin(y_2) \\ \sin(y_2) & y_1 \cos(y_2) \end{pmatrix}\right| \\ &= y_1 \,. \end{aligned}$$

It follows that

$$\begin{aligned} f_Y(y_1, y_2) &= \frac{1}{2\pi} \exp\left\{-\frac{y_1^2}{2}\right\} y_1 \times \\ & \quad 1(0 \le y_1 < \infty) 1(0 \le y_2 < 2\pi) \,. \end{aligned}$$

Next: marginal densities of Y_1, Y_2?

Factor f_Y as $f_Y(y_1, y_2) = h_1(y_1)h_2(y_2)$ where

$$h_1(y_1) = y_1 e^{-y_1^2/2} 1(0 \le y_1 < \infty)$$

and

$$h_2(y_2) = 1(0 \le y_2 < 2\pi)/(2\pi)\,.$$

Then

$$\begin{aligned} f_{Y_1}(y_1) &= \int_{-\infty}^{\infty} h_1(y_1)h_2(y_2)\,dy_2 \\ &= h_1(y_1)\int_{-\infty}^{\infty} h_2(y_2)\,dy_2 \end{aligned}$$

so marginal density of Y_1 is a multiple of h_1. Multiplier makes $\int f_{Y_1} = 1$ but in this case

$$\int_{-\infty}^{\infty} h_2(y_2)\,dy_2 = \int_0^{2\pi} (2\pi)^{-1} dy_2 = 1$$

so that

$$f_{Y_1}(y_1) = y_1 e^{-y_1^2/2} 1(0 \le y_1 < \infty)\,.$$

(Special Weibull or Rayleigh distribution.)

Similarly

$$f_{Y_2}(y_2) = 1(0 \le y_2 < 2\pi)/(2\pi)$$

which is the **Uniform**($(0, 2\pi)$ density. Exercise: $W = Y_1^2/2$ has standard exponential distribution. Recall: by definition $U = Y_1^2$ has a χ^2 distribution on 2 degrees of freedom. Exercise: find χ_2^2 density.

Note: We show below factorization of density is equivalent to independence.

Independence, conditional distributions

So far density of X specified explicitly. Often modelling leads to a specification in terms of marginal and conditional distributions.

Def'n: Events A and B are independent if

$$P(AB) = P(A)P(B)\,.$$

(Notation: AB is the event that both A and B happen, also written $A \cap B$.)

Def'n: A_i, $i = 1, \ldots, p$ are **independent** if

$$P(A_{i_1} \cdots A_{i_r}) = \prod_{j=1}^{r} P(A_{i_j})$$

for any $1 \le i_1 < \cdots < i_r \le p$.

Example: $p = 3$

$$\begin{aligned} P(A_1 A_2 A_3) &= P(A_1)P(A_2)P(A_3) \\ P(A_1 A_2) &= P(A_1)P(A_2) \\ P(A_1 A_3) &= P(A_1)P(A_3) \\ P(A_2 A_3) &= P(A_2)P(A_3)\,. \end{aligned}$$

All these equations needed for independence!

Example: Toss a coin twice.

$A_1 = \{$first toss is a Head$\}$
$A_2 = \{$second toss is a Head$\}$
$A_3 = \{$first toss and second toss different$\}$

Then $P(A_i) = 1/2$ for each i and for $i \neq j$

$$P(A_i \cap A_j) = \frac{1}{4}$$

but

$$P(A_1 \cap A_2 \cap A_3) = 0 \neq P(A_1)P(A_2)P(A_3)\,.$$

Def'n: X and Y are **independent** if

$$P(X \in A; Y \in B) = P(X \in A)P(Y \in B)$$

for all A and B.

Def'n: Rvs $X_1, \ldots, X_p$ **independent**:

$$P(X_1 \in A_1, \cdots, X_p \in A_p) = \prod P(X_i \in A_i)$$

for any $A_1, \ldots, A_p$.

Theorem:

1. If X and Y are independent then for all x, y
$$F_{X,Y}(x,y) = F_X(x)F_Y(y)\,.$$
2. If X and Y are independent with joint density $f_{X,Y}(x,y)$ then X and Y have densities f_X and f_Y, and
$$f_{X,Y}(x,y) = f_X(x)f_Y(y)\,.$$
3. If X and Y independent with marginal densities f_X and f_Y then (X,Y) has joint density
$$f_{X,Y}(x,y) = f_X(x)f_Y(y)\,.$$
4. If $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for **all** x, y then X and Y are independent.
5. If (X,Y) has density $f(x,y)$ and there exist $g(x)$ and $h(y)$ st $f(x,y) = g(x)h(y)$ for (almost) **all** (x,y) then X and Y are independent with densities given by
$$f_X(x) = g(x)/\int_{-\infty}^{\infty} g(u)du$$
$$f_Y(y) = h(y)/\int_{-\infty}^{\infty} h(u)du\,.$$

Proof:

1: Since X and Y are independent so are the events $X \le x$ and $Y \le y$; hence

$$P(X \le x, Y \le y) = P(X \le x)P(Y \le y)\,.$$

2: Suppose X and Y real valued.

Asst 2: existence of $f_{X,Y}$ implies that of f_X and f_Y (marginal density formula). Then for any sets A and B

$$\begin{aligned} P(X \in A, Y \in B) &= \int_A \int_B f_{X,Y}(x,y) dy dx \\ P(X \in A)P(Y \in B) &= \int_A f_X(x) dx \int_B f_Y(y) dy \\ &= \int_A \int_B f_X(x) f_Y(y) dy dx \end{aligned}$$

Since $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$

$$\int_A \int_B [f_{X,Y}(x,y) - f_X(x)f_Y(y)] dy dx = 0$$

It follows (measure theory) that the quantity in [] is 0 (almost every pair (x,y)).

3: For any A and B we have

$$\begin{aligned} P(X \in A, Y \in B) & \\ &= P(X \in A)P(Y \in B) \\ &= \int_A f_X(x)dx \int_B f_Y(y)dy \\ &= \int_A \int_B f_X(x)f_Y(y)dydx \end{aligned}$$

Define $g(x,y) = f_X(x)f_Y(y)$; we have proved that for $C = A \times B$

$$P((X,Y) \in C) = \int_C g(x,y)dydx$$

To prove that g is $f_{X,Y}$ we need only prove that this integral formula is valid for an arbitrary Borel set C, not just a rectangle $A \times B$.

Use *monotone class* argument: collection $\mathcal{C}$ of sets C for which identity holds has closure properties which guarantee that $\mathcal{C}$ includes the Borel sets.

4: Another monotone class argument.

5: We are given

$$\begin{aligned} P(X \in A, Y \in B) &= \int_A \int_B g(x)h(y)dydx \\ &= \int_A g(x)dx \int_B h(y)dy \end{aligned}$$

Take $B = R^1$ to see that

$$P(X \in A) = c_1 \int_A g(x)dx$$

where $c_1 = \int h(y)dy$. So $c_1 g$ is the density of X. Since $\int\int f_{X,Y}(xy)dxdy = 1$ we see that $\int g(x)dx \int h(y)dy = 1$ so that $c_1 = 1/\int g(x)dx$. Similar argument for Y.

Theorem: If $X_1, \ldots, X_p$ are independent and $Y_i = g_i(X_i)$ then $Y_1, \ldots, Y_p$ are independent. Moreover, $(X_1, \ldots, X_q)$ and $(X_{q+1}, \ldots, X_p)$ are independent.

Conditional probability

Def'n: $P(A|B) = P(AB)/P(B)$ if $P(B) \neq 0$.

Def'n: For discrete X and Y the conditional probability mass function of Y given X is

$$\begin{aligned} f_{Y|X}(y|x) &= P(Y=y|X=x) \\ &= f_{X,Y}(x,y)/f_X(x) \\ &= f_{X,Y}(x,y)/\sum_t f_{X,Y}(x,t) \end{aligned}$$

For absolutely continuous X $P(X = x) = 0$ for all x. What is $P(A|X = x)$ or $f_{Y|X}(y|x)$? Solution: use limit

$$P(A|X=x) = \lim_{\delta x \to 0} P(A|x \le X \le x + \delta x)$$

If, e.g., X, Y have joint density $f_{X,Y}$ then with $A = \{Y \le y\}$ we have

$$\begin{aligned} & P(A|x \le X \le x + \delta x) \\ & \qquad = \frac{P(A \cap \{x \le X \le x + \delta x\})}{P(x \le X \le x + \delta x)} \\ & \qquad = \frac{\int_{-\infty}^{y} \int_{x}^{x+\delta x} f_{X,Y}(u,v)\,du\,dv}{\int_{x}^{x+\delta x} f_X(u)\,du} \end{aligned}$$

Divide top, bottom by δx; let $\delta x \to 0$. Denom converges to $f_X(x)$; numerator converges to

$$\int_{-\infty}^{y} f_{X,Y}(x,v)dv$$

Define conditional cdf of Y given $X=x$:

$$P(Y \le y|X=x) = \frac{\int_{-\infty}^{y} f_{X,Y}(x,v)dv}{f_X(x)}$$

Differentiate wrt y to get def'n of conditional density of Y given $X=x$:

$$f_{Y|X}(y|x) = f_{X,Y}(x,y)/f_X(x)\,;$$

in words "conditional = joint/marginal".

The Multivariate Normal Distribution

Defn: $Z \in R^1 \sim N(0,1)$ iff

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,.$$

Defn: $Z \in R^p \sim MVN(0,I)$ if and only if $Z = (Z_1,\ldots,Z_p)^t$ with the Z_i independent and each $Z_i \sim N(0,1)$.

In this case according to our theorem

$$\begin{aligned} f_Z(z_1,\ldots,z_p) &= \prod \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} \\ &= (2\pi)^{-p/2} \exp\{-z^t z/2\}\,; \end{aligned}$$

superscript t denotes matrix transpose.

Defn: $X \in R^p$ has a multivariate normal distribution if it has the same distribution as $AZ+\mu$ for some $\mu \in R^p$, some $p \times p$ matrix of constants A and $Z \sim MVN(0,I)$.

Matrix A singular: X does not have a density.

A invertible: derive multivariate normal density by change of variables:

$$X = AZ + \mu \Leftrightarrow Z = A^{-1}(X - \mu)$$

$$\frac{\partial X}{\partial Z} = A \qquad \frac{\partial Z}{\partial X} = A^{-1}\,.$$

So

$$\begin{aligned} f_X(x) &= f_Z(A^{-1}(x-\mu))|\det(A^{-1})| \\ &= \frac{\exp\{-(x-\mu)^t(A^{-1})^tA^{-1}(x-\mu)/2\}}{(2\pi)^{p/2}|\det A|}\,. \end{aligned}$$

Now define $\Sigma = AA^t$ and notice that

$$\Sigma^{-1} = (A^t)^{-1}A^{-1} = (A^{-1})^tA^{-1}$$

and

$$\det\Sigma = \det A \det A^t = (\det A)^2\,.$$

Thus f_X is

$$\frac{\exp\{-(x-\mu)^t\Sigma^{-1}(x-\mu)/2\}}{(2\pi)^{p/2}(\det\Sigma)^{1/2}}\,;$$

the $MVN(\mu,\Sigma)$ density. Note density is the same for all A such that $AA^t = \Sigma$. This justifies the notation $MVN(\mu,\Sigma)$.

For which μ, Σ is this a density?

Any μ but if $x \in R^p$ then

$$\begin{aligned} x^t \Sigma x &= x^t A A^t x \\ &= (A^t x)^t (A^t x) \\ &= \sum_1^p y_i^2 \ge 0 \end{aligned}$$

where $y = A^t x$. Inequality strict except for $y = 0$ which is equivalent to $x = 0$. Thus Σ is a positive definite symmetric matrix.

Conversely, if Σ is a positive definite symmetric matrix then there is a square invertible matrix A such that $AA^t = \Sigma$ so that there is a $MVN(\mu, \Sigma)$ distribution. (A can be found via the Cholesky decomposition, e.g.)

When A is singular X will not have a density: $\exists a$ such that $P(a^t X = a^t \mu) = 1$; X is confined to a hyperplane.

Still true: distribution of X depends only on $\Sigma = AA^t$: if $AA^t = BB^t$ then $AZ + \mu$ and $BZ + \mu$ have the same distribution.

Properties of the MVN distribution

1: All margins are multivariate normal: if

$$X = \left[\begin{array}{c} X_1 \\ X_2 \end{array} \right]$$

$$\mu = \left[\begin{array}{c} \mu_1 \\ \mu_2 \end{array} \right]$$

and

$$\Sigma = \left[\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array} \right]$$

then $X \sim MVN(\mu, \Sigma) \Rightarrow X_1 \sim MVN(\mu_1, \Sigma_{11})$.

2: All conditionals are normal: the conditional distribution of X_1 given $X_2 = x_2$ is $MVN(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$

3: $MX + \nu \sim MVN(M\mu + \nu, M\Sigma M^t)$: affine transformation of MVN is normal.

Normal samples: Distribution Theory

Theorem: Suppose $X_1, \ldots, X_n$ are independent $N(\mu, \sigma^2)$ random variables. Then

1. $\bar{X}$ (sample mean)and s^2 (sample variance) independent.

2. $n^{1/2}(\bar{X} - \mu)/\sigma \sim N(0, 1)$.

3. $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$.

4. $n^{1/2}(\bar{X} - \mu)/s \sim t_{n-1}$.

Proof: Let $Z_i = (X_i - \mu)/\sigma$.

Then $Z_1, \ldots, Z_p$ are independent $N(0, 1)$.

So $Z = (Z_1, \ldots, Z_p)^t$ is multivariate standard normal.

Note that $\bar{X} = \sigma\bar{Z} + \mu$ and $s^2 = \sum(X_i - \bar{X})^2/(n - 1) = \sigma^2 \sum(Z_i - \bar{Z})^2/(n - 1)$ Thus

$$\frac{n^{1/2}(\bar{X} - \mu)}{\sigma} = n^{1/2}\bar{Z}$$

$$\frac{(n-1)s^2}{\sigma^2} = \sum(Z_i - \bar{Z})^2$$

and

$$T = \frac{n^{1/2}(\bar{X} - \mu)}{s} = \frac{n^{1/2}\bar{Z}}{s_Z}$$

where $(n - 1)s_Z^2 = \sum(Z_i - \bar{Z})^2$.

So: reduced to $\mu = 0$ and $\sigma = 1$.

Step 1: Define

$$Y = (\sqrt{n}\bar{Z}, Z_1 - \bar{Z}, \ldots, Z_{n-1} - \bar{Z})^t \,.$$

(So Y has same dimension as Z.) Now

$$Y = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix}$$

or letting M denote the matrix

$$Y = MZ \,.$$

It follows that $Y \sim MVN(0, MM^t)$ so we need to compute MM^t:

$$MM^t = \left[\begin{array}{c|cccc} 1 & 0 & 0 & \cdots & 0 \\ \hline 0 & 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & -\frac{1}{n} & \ddots & \cdots & -\frac{1}{n} \\ 0 & \vdots & \cdots & & 1-\frac{1}{n} \end{array}\right]$$

$$= \left[\begin{array}{c|c} 1 & 0 \\ \hline & \\ 0 & Q \end{array}\right] .$$

Solve for Z from Y: $Z_i = n^{-1/2}Y_1 + Y_{i+1}$ for $1 \le i \le n-1$. Use the identity

$$\sum_{i=1}^{n}(Z_i - \bar{Z}) = 0$$

to get $Z_n = -\sum_{i=2}^{n} Y_i + n^{-1/2}Y_1$. So M invertible:

$$\Sigma^{-1} \equiv (MM^t)^{-1} = \left[\begin{array}{c|c} 1 & 0 \\ \hline & \\ 0 & Q^{-1} \end{array}\right].$$

Use change of variables to find f_Y. Let $\mathbf{y}_2$ denote vector whose entries are $y_2, \ldots, y_n$. Note that

$$y^t\Sigma^{-1}y = y_1^2 + \mathbf{y}_2^t Q^{-1}\mathbf{y}_2\,.$$

Then

$$\begin{aligned} f_Y(y) =& (2\pi)^{-n/2}\exp[-y^t\Sigma^{-1}y/2]/|\det M| \\ =& \frac{1}{\sqrt{2\pi}}e^{-y_1^2/2}\times \\ & \frac{(2\pi)^{-(n-1)/2}\exp[-\mathbf{y}_2^t Q^{-1}\mathbf{y}_2/2]}{|\det M|}\,. \end{aligned}$$

Note: f_Y is ftn of y_1 times a ftn of $y_2, \ldots, y_n$.

Thus $\sqrt{n}\bar{Z}$ is independent of $Z_1 - \bar{Z}, \ldots, Z_{n-1} - \bar{Z}$.

Since s_Z^2 is a function of $Z_1 - \bar{Z}, \ldots, Z_{n-1} - \bar{Z}$ we see that $\sqrt{n}\bar{Z}$ and s_Z^2 are independent.

Also, density of Y_1 is a multiple of the function of y_1 in the factorization above. But factor is standard normal density so $\sqrt{n}\bar{Z} \sim N(0,1)$.

First 2 parts done. Third part is a homework exercise.

Derivation of the χ^2 density:

Suppose $Z_1, \ldots, Z_n$ are independent $N(0,1)$. Define χ_n^2 distribution to be that of $U = Z_1^2 + \cdots + Z_n^2$. Define angles $\theta_1, \ldots, \theta_{n-1}$ by

$$\begin{aligned} Z_1 &= U^{1/2} \cos\theta_1 \\ Z_2 &= U^{1/2} \sin\theta_1 \cos\theta_2 \\ \vdots &= \vdots \\ Z_{n-1} &= U^{1/2} \sin\theta_1 \cdots \sin\theta_{n-2} \cos\theta_{n-1} \\ Z_n &= U^{1/2} \sin\theta_1 \cdots \sin\theta_{n-1} \, . \end{aligned}$$

(Spherical co-ordinates in n dimensions. The θ values run from 0 to π except last θ from 0 to 2π.) Derivative formulas:

$$\frac{\partial Z_i}{\partial U} = \frac{1}{2U} Z_i$$

and

$$\frac{\partial Z_i}{\partial \theta_j} = \begin{cases} 0 & j > i \\ -Z_i \tan\theta_i & j = i \\ Z_i \cot\theta_j & j < i \, . \end{cases}$$

Fix $n = 3$ to clarify the formulas. Use shorthand $R = \sqrt{U}$

Matrix of partial derivatives is

$$\begin{bmatrix} \frac{\cos\theta_1}{2R} & -R\sin\theta_1 & 0 \\ \frac{\sin\theta_1\cos\theta_2}{2R} & R\cos\theta_1\cos\theta_2 & -R\sin\theta_1\sin\theta_2 \\ \frac{\sin\theta_1\sin\theta_2}{2R} & R\cos\theta_1\sin\theta_2 & R\sin\theta_1\cos\theta_2 \end{bmatrix}.$$

Find determinant by adding $2U^{1/2}\cos\theta_j/\sin\theta_j$ times col 1 to col $j+1$ (no change in determinant).

Resulting matrix lower triangular; diagonal entries:

$$\frac{\cos\theta_1}{R}, \frac{R\cos\theta_2}{\cos\theta_1}, \frac{R\sin\theta_1}{\cos\theta_2}$$

Multiply these together to get

$$U^{1/2}\sin(\theta_1)/2$$

(non-negative for all U and θ_1).

General n: every term in the first column contains a factor $U^{-1/2}/2$ while every other entry has a factor $U^{1/2}$.

FACT: multiplying a column in a matrix by c multiplies the determinant by c.

SO: Jacobian of transformation is

$$u^{(n-1)/2}u^{-1/2}/2 \times h(\theta_1, \theta_{n-1})$$

for some function, h, which depends only on the angles.

Thus joint density of $U, \theta_1, \ldots \theta_{n-1}$ is

$$(2\pi)^{-n/2} \exp(-u/2) u^{(n-2)/2} h(\theta_1, \cdots, \theta_{n-1})/2 \,.$$

To compute the density of U we must do an $n-1$ dimensional multiple integral $d\theta_{n-1} \cdots d\theta_1$.

Answer has the form

$$cu^{(n-2)/2} \exp(-u/2)$$

for some c.

Evaluate c by making

$$\begin{aligned}\int f_U(u)du &= c\int_0^\infty u^{(n-2)/2}\,\mathsf{exp}(-u/2)du \\ &= 1.\end{aligned}$$

Substitute $y = u/2$, $du = 2dy$ to see that

$$\begin{aligned}c2^{n/2}\int_0^\infty y^{(n-2)/2}e^{-y}dy &= c2^{n/2}\Gamma(n/2) \\ &= 1.\end{aligned}$$

CONCLUSION: the χ^2_n density is

$$\frac{1}{2\Gamma(n/2)}\left(\frac{u}{2}\right)^{(n-2)/2} e^{-u/2}1(u>0)\,.$$

Fourth part: consequence of first 3 parts and def'n of t_ν distribution.

Defn: $T \sim t_\nu$ if T has same distribution as

$$Z/\sqrt{U/\nu}$$

for $Z \sim N(0,1)$, $U \sim \chi^2_\nu$ and Z, U independent.

Derive density of T in this definition:

$$\begin{aligned} P(T \le t) &= P(Z \le t\sqrt{U/\nu}) \\ &= \int_0^\infty \int_{-\infty}^{t\sqrt{u/\nu}} f_Z(z) f_U(u) dz du \end{aligned}$$

Differentiate wrt t by differentiating inner integral:

$$\frac{\partial}{\partial t} \int_{at}^{bt} f(x)dx = bf(bt) - af(at)$$

by fundamental thm of calculus. Hence

$$\frac{d}{dt} P(T \le t) = \int_0^\infty \frac{f_U(u)}{\sqrt{2\pi}} \left(\frac{u}{\nu}\right)^{1/2} \exp\left(-\frac{t^2 u}{2\nu}\right) du\,.$$

Plug in

$$f_U(u) = \frac{1}{2\Gamma(\nu/2)}(u/2)^{(\nu-2)/2}e^{-u/2}$$

to get

$$f_T(t) = \frac{\int_0^\infty (u/2)^{(\nu-1)/2}e^{-u(1+t^2/\nu)/2}du}{2\sqrt{\pi\nu}\Gamma(\nu/2)}.$$

Substitute $y = u(1+t^2/\nu)/2$, to get

$$dy = (1+t^2/\nu)du/2$$

$$(u/2)^{(\nu-1)/2} = [y/(1+t^2/\nu)]^{(\nu-1)/2}$$

leading to

$$f_T(t) = \frac{(1+t^2/\nu)^{-(\nu+1)/2}}{\sqrt{\pi\nu}\Gamma(\nu/2)}\int_0^\infty y^{(\nu-1)/2}e^{-y}dy$$

or

$$f_T(t) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)}\frac{1}{(1+t^2/\nu)^{(\nu+1)/2}}.$$

Expectation, moments

Two elementary definitions of expected values:

Defn: If X has density f then

$$E\{g(X)\} = \int g(x)f(x)\,dx\,.$$

Defn: If X has discrete density f then

$$E\{g(X)\} = \sum_x g(x)f(x)\,.$$

FACT: if $Y = g(X)$ for a smooth g

$$\begin{aligned} E(Y) &= \int y f_Y(y)\,dy \\ &= \int g(x)f_Y(g(x))g'(x)\,dx \\ &= E\{g(X)\} \end{aligned}$$

by change of variables formula for integration. This is good because otherwise we might have two different values for $E(e^X)$.

In general, there are random variables which are neither absolutely continuous nor discrete. Here's how probabilists define E in general.

Defn: RV X is simple if we can write

$$X(\omega) = \sum_1^n a_i 1(\omega \in A_i)$$

for some constants $a_1, \ldots, a_n$ and events A_i.

Defn: For a simple rv X define

$$E(X) = \sum a_i P(A_i)\,.$$

For positive random variables which are not simple extend definition by approximation:

Defn: If $X \ge 0$ then

$$E(X) = \text{sup}\{E(Y) : 0 \le Y \le X, Y \text{ simple}\}\,.$$

Defn: X is **integrable** if

$$E(|X|) < \infty\,.$$

In this case we define

$$E(X) = E\{\text{max}(X,0)\} - E\{\text{max}(-X,0)\}\,.$$

Facts: E is a linear, monotone, positive operator:

1. **Linear**: $E(aX+bY) = aE(X)+bE(Y)$ provided X and Y are integrable.

2. **Positive**: $P(X \ge 0) = 1$ implies $E(X) \ge 0$.

3. **Monotone**: $P(X \ge Y) = 1$ and X, Y integrable implies $E(X) \ge E(Y)$.

Major technical theorems:

Monotone Convergence: If $0 \le X_1 \le X_2 \le \cdots$ and $X = \lim X_n$ (which has to exist) then

$$E(X) = \lim_{n\to\infty} E(X_n)\,.$$

Dominated Convergence: If $|X_n| \le Y_n$ and $\exists$ rv X such that $X_n \to X$ (technical details of this convergence later in the course) and a random variable Y such that $Y_n \to Y$ with $E(Y_n) \to E(Y) < \infty$ then

$$E(X_n) \to E(X)\,.$$

Often used with all Y_n the same rv Y.

Fatou's Lemma: If $X_n \ge 0$ then

$$E(\limsup X_n) \le \limsup E(X_n)\,.$$

Theorem: With this definition of E if X has density $f(x)$ (even in R^p say) and $Y = g(X)$ then

$$E(Y) = \int g(x)f(x)dx\,.$$

(Could be a multiple integral.) If X has pmf f then

$$E(Y) = \sum_x g(x)f(x)\,.$$

First conclusion works, e.g., even if X has a density but Y doesn't.

Defn: The r^{th} moment (about the origin) of a real rv X is $\mu'_r = E(X^r)$ (provided it exists). We generally use μ for $E(X)$.

Defn: The r^{th} central moment is

$$\mu_r = E[(X-\mu)^r]$$

We call $\sigma^2 = \mu_2$ the variance.

Defn: For an R^p valued random vector X

$$\mu_X = E(X)$$

is the vector whose i^{th} entry is $E(X_i)$ (provided all entries exist).

Defn: The $(p \times p)$ variance covariance matrix of X is

$$\text{Var}(X) = E\left[(X-\mu)(X-\mu)^t\right]$$

which exists provided each component X_i has a finite second moment.

Moments and probabilities of rare events are closely connected as will be seen in a number of important probability theorems.

Example: Markov's inequality

$$\begin{aligned} P(|X-\mu| \ge t) &= E[1(|X-\mu| \ge t)] \\ &\le E\left[\frac{|X-\mu|^r}{t^r} 1(|X-\mu| \ge t)\right] \\ &\le \frac{E[|X-\mu|^r]}{t^r} \end{aligned}$$

Intuition: if moments are small then large deviations from average are unlikely.

Special case is Chebyshev's inequality:

$$P(|X-\mu| \ge t) \le \frac{\mathrm{Var}(X)}{t^2}\,.$$

Example moments: If $Z \sim N(0,1)$ then

$$\begin{aligned} E(Z) &= \int_{-\infty}^{\infty} z e^{-z^2/2} dz/\sqrt{2\pi} \\ &= \left. \frac{-e^{-z^2/2}}{\sqrt{2\pi}} \right|_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

and (integrating by parts)

$$\begin{aligned} E(Z^r) =& \int_{-\infty}^{\infty} z^r e^{-z^2/2} dz/\sqrt{2\pi} \\ =& \left. \frac{-z^{r-1} e^{-z^2/2}}{\sqrt{2\pi}} \right|_{-\infty}^{\infty} \\ & + (r-1) \int_{-\infty}^{\infty} z^{r-2} e^{-z^2/2} dz/\sqrt{2\pi} \end{aligned}$$

so that

$$\mu_r = (r-1)\mu_{r-2}$$

for $r \geq 2$. Remembering that $\mu_1 = 0$ and

$$\mu_0 = \int_{-\infty}^{\infty} z^0 e^{-z^2/2} dz/\sqrt{2\pi} = 1$$

we find that

$$\mu_r = \begin{cases} 0 & r \text{ odd} \\ (r-1)(r-3)\cdots 1 & r \text{ even}. \end{cases}$$

If now $X \sim N(\mu,\sigma^2)$, that is, $X \sim \sigma Z + \mu$, then $E(X) = \sigma E(Z) + \mu = \mu$ and

$$\mu_r(X) = E[(X-\mu)^r] = \sigma^r E(Z^r)$$

In particular, we see that our choice of notation $N(\mu,\sigma^2)$ for the distribution of $\sigma Z + \mu$ is justified; σ is indeed the variance.

Similarly for $X \sim MVN(\mu,\Sigma)$ we have $X = AZ + \mu$ with $Z \sim MVN(0,I)$ and

$$E(X) = \mu$$

and

$$\begin{aligned}\text{Var}(X) &= E\left\{(X-\mu)(X-\mu)^t\right\} \\ &= E\left\{AZ(AZ)^t\right\} \\ &= AE(ZZ^t)A^t \\ &= AIA^t = \Sigma\,.\end{aligned}$$

Note use of easy calculation: $E(Z) = 0$ and

$$\text{Var}(Z) = E(ZZ^t) = I\,.$$

Moments and independence

Theorem: If $X_1,\ldots,X_p$ are independent and each X_i is integrable then $X = X_1 \cdots X_p$ is integrable and

$$E(X_1 \cdots X_p) = E(X_1) \cdots E(X_p)\,.$$

Proof: Suppose each X_i is simple:

$$X_i = \sum_j x_{ij} 1(X_i = x_{ij})$$

where the x_{ij} are the possible values of X_i. Then

$$
\begin{aligned}
&E(X_1 \cdots X_p) \\
&\quad = \sum_{j_1 \cdots j_p} x_{1j_1} \cdots x_{pj_p} \\
&\qquad \times E(1(X_1 = x_{1j_1}) \cdots 1(X_p = x_{pj_p})) \\
&\quad = \sum_{j_1 \cdots j_p} x_{1j_1} \cdots x_{pj_p} \\
&\qquad \times P(X_1 = x_{1j_1} \cdots X_p = x_{pj_p}) \\
&\quad = \sum_{j_1 \cdots j_p} x_{1j_1} \cdots x_{pj_p} \\
&\qquad \times P(X_1 = x_{1j_1}) \cdots P(X_p = x_{pj_p}) \\
&\quad = \sum_{j_1} x_{1j_1} P(X_1 = x_{1j_1}) \times \cdots \\
&\qquad \times \sum_{j_p} x_{pj_p} P(X_p = x_{pj_p}) \\
&\quad = \prod E(X_i) \,.
\end{aligned}
$$

General $X_i \ge 0$:

Let X_{in} be X_i rounded down to nearest multiple of 2^{-n} (to maximum of n).

That is: if

$$\frac{k}{2^n} \le X_i < \frac{k+1}{2^n}$$

then $X_{in} = k/2^n$ for $k = 0, \ldots, n2^n$. For $X_i > n$ put$X_{in} = n$.

Apply case just done:

$$E(\prod X_{in}) = \prod E(X_{in}) \,.$$

Monotone convergence applies to both sides.

For general case write each X_i as difference of positive and negative parts:

$$X_i = \max(X_i, 0) - \max(-X_i, 0) \,.$$

Apply positive case.

Moment Generating Functions

Defn: The moment generating function of a real valued X is

$$M_X(t) = E(e^{tX})$$

defined for those real t for which the expected value is finite.

Defn: The moment generating function of $X \in R^p$ is

$$M_X(u) = E[e^{u^t X}]$$

defined for those vectors u for which the expected value is finite.

Formal connection to moments:

$$\begin{aligned} M_X(t) &= \sum_{k=0}^{\infty} E[(tX)^k]/k! \\ &= \sum_{k=0}^{\infty} \mu_k' t^k/k!\,. \end{aligned}$$

Sometimes can find power series expansion of M_X and read off the moments of X from the coefficients of $t^k/k!$.

Theorem: If M is finite for all $t \in [-\epsilon, \epsilon]$ for some $\epsilon > 0$ then

1. Every moment of X is finite.

2. M is C^∞ (in fact M is analytic).

3. $\mu'_k = \frac{d^k}{dt^k} M_X(0)$.

Note: C^∞ means has continuous derivatives of all orders. Analytic means has convergent power series expansion in neighbourhood of each $t \in (-\epsilon, \epsilon)$.

The proof, and many other facts about mgfs, rely on techniques of complex variables.

MGFs and Sums

If $X_1, \ldots, X_p$ are independent and $Y = \sum X_i$ then the moment generating function of Y is the product of those of the individual X_i:

$$E(e^{tY}) = \prod_i E(e^{tX_i})$$

or $M_Y = \prod M_{X_i}$.

Note: also true for multivariate X_i.

Problem: power series expansion of M_Y not nice function of expansions of individual M_{X_i}.

Related fact: first 3 moments (meaning μ, σ^2 and μ_3) of Y are sums of those of the X_i:

$$\begin{aligned} E(Y) &= \sum E(X_i) \\ \text{Var}(Y) &= \sum \text{Var}(X_i) \\ E[(Y - E(Y))^3] &= \sum E[(X_i - E(X_i))^3] \end{aligned}$$

but

$$\begin{aligned} &E[(Y - E(Y))^4] = \\ &\quad \sum \{E[(X_i - E(X_i))^4] - 3E^2[(X_i - E(X_i))^2]\} \\ &\qquad\qquad + 3\left\{\sum E[(X_i - E(X_i))^2]\right\}^2 \end{aligned}$$

Related quantities: **cumulants** add up properly.

Note: log of the mgf of Y is sum of logs of mgfs of the X_i.

Defn: the cumulant generating function of a variable X by

$$K_X(t) = \log(M_X(t))\,.$$

Then

$$K_Y(t) = \sum K_{X_i}(t)\,.$$

Note: mgfs are all positive so that the cumulative generating functions are defined wherever the mgfs are.

SO: K_Y has power series expansion:

$$K_Y(t) = \sum_{r=1}^{\infty} \kappa_r t^r/r!\,.$$

Defn: the κ_r are the cumulants of Y.

Observe

$$\kappa_r(Y) = \sum \kappa_r(X_i)\,.$$

Relation between cumulants and moments:

Cumulant generating function is

$$\begin{aligned} K(t) &= \log(M(t)) \\ &= \log(1 + [\mu_1 t + \mu_2' t^2/2 + \mu_3' t^3/3! + \cdots]) \end{aligned}$$

Call quantity in $[\ldots]$ x; expand

$$\log(1+x) = x - x^2/2 + x^3/3 - x^4/4 \cdots .$$

Stick in the power series

$$x = \mu t + \mu_2' t^2/2 + \mu_3' t^3/3! + \cdots ;$$

Expand out powers of x; collect together like terms. For instance,

$$\begin{aligned} x^2 &= \mu^2 t^2 + \mu \mu_2' t^3 \\ &\qquad + [2\mu_3' \mu/3! + (\mu_2')^2/4] t^4 + \cdots \\ x^3 &= \mu^3 t^3 + 3\mu_2' \mu^2 t^4/2 + \cdots \\ x^4 &= \mu^4 t^4 + \cdots . \end{aligned}$$

Now gather up the terms. The power t^1 occurs only in x with coefficient μ. The power t^2 occurs in x and in x^2 and so on.

Putting these together gives

$$K(t) = \\ \mu t + [\mu_2' - \mu^2]t^2/2 \\ + [\mu_3' - 3\mu\mu_2' + 2\mu^3]t^3/3! \\ + [\mu_4' - 4\mu_3'\mu - 3(\mu_2')^2 + 12\mu_2'\mu^2 - 6\mu^4]t^4/4! \cdots$$

Comparing coefficients to $t^r/r!$ we see that

$$\begin{aligned} \kappa_1 &= \mu \\ \kappa_2 &= \mu_2' - \mu^2 = \sigma^2 \\ \kappa_3 &= \mu_3' - 3\mu\mu_2' + 2\mu^3 = E[(X-\mu)^3] \\ \kappa_4 &= \mu_4' - 4\mu_3'\mu - 3(\mu_2')^2 + 12\mu_2'\mu^2 - 6\mu^4 \\ &= E[(X-\mu)^4] - 3\sigma^4 \,. \end{aligned}$$

Check the book by Kendall and Stuart (or the new version called *Kendall's Theory of Advanced Statistics* by Stuart and Ord) for formulas for larger orders r.

Example: If $X_1,\ldots,X_p$ are independent and X_i has a $N(\mu_i,\sigma_i^2)$ distribution then

$$\begin{aligned}M_{X_i}(t) &= \int_{-\infty}^{\infty} e^{tx} e^{-\frac{1}{2}(x-\mu_i)^2/\sigma_i^2} dx/(\sqrt{2\pi}\sigma_i) \\ &= \int_{-\infty}^{\infty} e^{t(\sigma_i z+\mu_i)} e^{-z^2/2} dz/\sqrt{2\pi} \\ &= e^{t\mu_i}\int_{-\infty}^{\infty} e^{-(z-t\sigma_i)^2/2+t^2\sigma_i^2/2} dz/\sqrt{2\pi} \\ &= e^{\sigma_i^2 t^2/2+t\mu_i} .\end{aligned}$$

So cumulant generating function is:

$$K_{X_i}(t) = \log(M_{X_i}(t)) = \sigma_i^2 t^2/2 + \mu_i t \,.$$

Cumulants are $\kappa_1 = \mu_i$, $\kappa_2 = \sigma_i^2$ and every other cumulant is 0.

Cumulant generating function for $Y = \sum X_i$ is

$$K_Y(t) = \sum \sigma_i^2 t^2/2 + t \sum \mu_i$$

which is the cumulant generating function of $N(\sum \mu_i, \sum \sigma_i^2)$.

Example: Homework: derive moment and cumulant generating function and moments of a Gamma rv.

Now suppose $Z_1, \ldots, Z_\nu$ independent $N(0,1)$ rvs.

By definition: $S_\nu = \sum_1^\nu Z_i^2$ has χ^2_ν distribution. It is easy to check $S_1 = Z_1^2$ has density

$$(u/2)^{-1/2} e^{-u/2} / (2\sqrt{\pi})$$

and then the mgf of S_1 is

$$(1-2t)^{-1/2} \,.$$

It follows that

$$M_{S_\nu}(t) = (1-2t)^{-\nu/2}$$

which is (homework) moment generating function of a Gamma$(\nu/2, 2)$ rv.

SO: χ^2_ν dstbn has Gamma$(\nu/2, 2)$ density:

$$(u/2)^{(\nu-2)/2} e^{-u/2} / (2\Gamma(\nu/2)) \,.$$

Example: The Cauchy density is

$$\frac{1}{\pi(1+x^2)};$$

corresponding moment generating function is

$$M(t) = \int_{-\infty}^{\infty} \frac{e^{tx}}{\pi(1+x^2)} dx$$

which is $+\infty$ except for $t = 0$ where we get 1.

Every t distribution has exactly same mgf. So: can't use mgf to distinguish such distributions.

Problem: these distributions do not have infinitely many finite moments.

So: develop substitute for mgf which is defined for every distribution, namely, the characteristic function.

Characteristic Functions

Definition: The characteristic function of a real rv X is

$$\phi_X(t) = E(e^{itX})$$

where $i = \sqrt{-1}$ is the imaginary unit.

Aside on complex arithmetic.

Complex numbers: add $i = \sqrt{-1}$ to the real numbers.

Require all the usual rules of algebra to work.

So: if i and any real numbers a and b are to be complex numbers then so must be $a + bi$.

Multiplication: If we multiply a complex number $a + bi$ with a and b real by another such number, say $c + di$ then the usual rules of arithmetic (associative, commutative and distributive laws) require

$$\begin{aligned}(a+bi)(c+di) &= ac + adi + bci + bdi^2 \\ &= ac + bd(-1) + (ad+bc)i \\ &= (ac - bd) + (ad+bc)i\end{aligned}$$

so this is precisely how we define multiplication.

Addition: follow usual rules to get

$$(a+bi)+(c+di)=(a+c)+(b+d)i\,.$$

Additive inverses: $-(a+bi)=-a+(-b)i$.

Multiplicative inverses:

$$\begin{aligned}\frac{1}{a+bi} &= \frac{1}{a+bi}\frac{a-bi}{a-bi}\\ &= \frac{a-bi}{a^2-abi+abi-b^2i^2}\\ &= \frac{a-bi}{a^2+b^2}\,.\end{aligned}$$

Division:

$$\begin{aligned}\frac{a+bi}{c+di} &= \frac{(a+bi)}{(c+di)}\frac{(c-di)}{(c-di)}\\ &= \frac{ac-bd+(bc+ad)i}{c^2+d^2}\,.\end{aligned}$$

Notice: usual rules of arithmetic don't require any more numbers than

$$x+yi$$

where x and y are real.

Transcendental functions: For real x have $e^x = \sum x^k/k!$ so

$$e^{x+iy} = e^x e^{iy}\,.$$

How to compute e^{iy}?

Remember $i^2 = -1$ so $i^3 = -i$, $i^4 = 1$ $i^5 = i^1 = i$ and so on. Then

$$\begin{aligned}
e^{iy} =& \sum_0^\infty \frac{(iy)^k}{k!} \\
=& 1 + iy + (iy)^2/2 + (iy)^3/6 + \cdots \\
=& 1 - y^2/2 + y^4/4! - y^6/6! + \cdots \\
& \quad + iy - iy^3/3! + iy^5/5! + \cdots \\
=& \cos(y) + i\sin(y)
\end{aligned}$$

We can thus write

$$e^{x+iy} = e^x(\cos(y) + i\sin(y))$$

Identify $x + yi$ with the corresponding point (x, y) in the plane. Picture the complex numbers as forming a plane.

Now every point in the plane can be written in polar co-ordinates as $(r \cos \theta, r \sin \theta)$ and comparing this with our formula for the exponential we see we can write

$$x + iy = \sqrt{x^2 + y^2}\, e^{i\theta} = re^{i\theta}$$

for an angle $\theta \in [0, 2\pi)$.

Multiplication revisited: $x+iy = re^{i\theta}$, $x'+iy' = r'e^{i\theta'}$.

$$(x + iy)(x' + iy') = re^{i\theta} r' e^{i\theta'} = rr' e^{i(\theta+\theta')} .$$

We will need from time to time a couple of other definitions:

Definition: The **modulus** of $x + iy$ is

$$|x + iy| = \sqrt{x^2 + y^2}\,.$$

Definition: The **complex conjugate** of $x + iy$ is $\overline{x + iy} = x - iy$.

Some identities: $z = x + iy = re^{i\theta}$ and $z' = x' + iy' = r'e^{i\theta'}$. Then

$$z\overline{z} = x^2 + y^2 = r^2 = |z|^2$$

$$\frac{z'}{z} = \frac{z'\overline{z}}{|z|^2} = rr'e^{i(\theta' - \theta)}$$

$$\overline{re^{i\theta}} = re^{-i\theta}.$$

Notes on calculus with complex variables.

Essentially usual rules apply so, for example,

$$\frac{d}{dt}e^{it} = ie^{it}\,.$$

We will (mostly) be doing only integrals over the real line; the theory of integrals along paths in the complex plane is a very important part of mathematics, however.

FACT: (not used explicitly in course). If $f : \mathbb{C} \mapsto \mathbb{C}$ is differentiable then f is analytic (has power series expansion).

End of Aside

Characteristic Functions

Definition: The characteristic function of a real rv X is

$$\phi_X(t) = E(e^{itX})$$

where $i = \sqrt{-1}$ is the imaginary unit.

Since

$$e^{itX} = \cos(tX) + i \sin(tX)$$

we find that

$$\phi_X(t) = E(\cos(tX)) + iE(\sin(tX))\,.$$

Since the trigonometric functions are bounded by 1 the expected values must be finite for all t.

This is precisely the reason for using characteristic rather than moment generating functions in probability theory courses.

Theorem 2 *For any two real rvs X and Y the following are equivalent:*

1. *X and Y have the same distribution, that is, for any (Borel) set A we have*
$$P(X \in A) = P(Y \in A)\,.$$

2. *$F_X(t) = F_Y(t)$ for all t.*

3. *$\phi_X(t) = E(e^{itX}) = E(e^{itY}) = \phi_Y(t)$ for all real t.*

Moreover, all of these are implied if there is a positive ϵ such that for all $|t| \le \epsilon$

$$M_X(t) = M_Y(t) < \infty\,.$$

Inversion

Previous theorem is non-constructive characterization.

Can get from ϕ_X to F_X or f_X by **inversion**.

See homework for basic **inversion** formula:

If X is a random variable taking only integer values then for each integer k

$$\begin{aligned} P(X=k) &= \frac{1}{2\pi}\int_0^{2\pi} \phi_X(t)e^{-itk}dt \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \phi_X(t)e^{-itk}dt\,. \end{aligned}$$

The proof proceeds from the formula

$$\phi_X(t) = \sum_k e^{ikt} P(X=k)\,.$$

Now suppose X has continuous bounded density f. Define

$$X_n = [nX]/n$$

where $[a]$ denotes the integer part (rounding down to the next smallest integer). We have

$$\begin{aligned}
P(k/n \le X < (k+1)/n) \\
&= P([nX] = k) \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} \phi_{[nX]}(t) \times e^{-itk} dt\,.
\end{aligned}$$

Make the substitution $t = u/n$, and get

$$\begin{aligned}
nP(k/n \le X < (k+1)/n) = \frac{1}{2\pi} \\
\times \int_{-n\pi}^{n\pi} \phi_{[nX]}(u/n) e^{-iuk/n} du\,.
\end{aligned}$$

Now, as $n \to \infty$ we have

$$\phi_{[nX]}(u/n) = E(e^{iu[nX]/n}) \to E(e^{iuX}) \,.$$

(Dominated convergence: $|e^{iu}| \le 1$.)

Range of integration converges to the whole real line.

If $k/n \to x$ left hand side converges to density $f(x)$ while right hand side converges to

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_X(u)e^{-iux}du$$

which gives the inversion formula

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_X(u)e^{-iux}du \,.$$

Many other such formulas are available to compute things like $F(b) - F(a)$ and so on.

All such formulas called **Fourier inversion formulas**.

Characteristic ftn also called **Fourier transform** of f or F.

Inversion of the Moment Generating Function

MGF and characteristic function related formally:

$$M_X(it) = \phi_X(t)\,.$$

When M_X exists this relationship is not merely formal; the methods of complex variables mean there is a “nice” (analytic) function which is $E(e^{zX})$ for any complex $z = x + iy$ for which $M_X(x)$ is finite.

SO: there is an inversion formula for M_X using a complex *contour integral*:

If z_1 and z_2 are two points in the complex plane and C a path between these two points we can define the path integral

$$\int_C f(z)dz$$

by the methods of line integration.

Do algebra with such integrals via usual theorems of calculus.

The Fourier inversion formula was

$$2\pi f(x) = \int_{-\infty}^{\infty} \phi(t) e^{-itx} dt$$

so replacing ϕ by M we get

$$2\pi f(x) = \int_{-\infty}^{\infty} M(it) e^{-itx} dt\,.$$

If we just substitute $z = it$ then we find

$$2\pi i f(x) = \int_C M(z) e^{-zx} dz$$

where the path C is the imaginary axis.

Complex contour integration: replace C by any other path which starts and ends at the same place.

Sometimes can choose path to make it easy to do the integral approximately; this is what **saddlepoint approximations** are.

Inversion formula is called the **inverse Laplace transform**; the mgf is also called the Laplace transform of f or F.

Applications of Inversion

1): Numerical calculations

Example: Many statistics have a distribution which is approximately that of

$$T = \sum \lambda_j Z_j^2$$

where the Z_j are iid $N(0,1)$. In this case

$$\begin{aligned} E(e^{itT}) &= \prod E(e^{it\lambda_j Z_j^2}) \\ &= \prod (1-2it\lambda_j)^{-1/2}\,. \end{aligned}$$

Imhof (*Biometrika*, 1961) gives a simplification of the Fourier inversion formula for

$$F_T(x) - F_T(0)$$

which can be evaluated numerically:

$$\begin{aligned} &F_T(x) - F_T(0) \\ &\qquad = \int_0^x f_T(y)dy \\ &\qquad = \int_0^x \frac{1}{2\pi}\int_{-\infty}^{\infty} \prod (1-2it\lambda_j)^{-1/2} e^{-ity} dt dy\,. \end{aligned}$$

Multiply

$$\phi(t) = \left[\frac{1}{\prod(1-2it\lambda_j)}\right]^{1/2}$$

top and bottom by the complex conjugate of the denominator:

$$\phi(t) = \left[\frac{\prod(1+2it\lambda_j)}{\prod(1+4t^2\lambda_j^2)}\right]^{1/2} .$$

The complex number $1+2it\lambda_j$ is $r_j e^{i\theta_j}$ where

$$r_j = \sqrt{1+4t^4\lambda_j^2}$$

and

$$\tan(\theta_j) = 2t\lambda_j\,.$$

This allows us to rewrite

$$\phi(t) = \left[\frac{\prod r_j e^{i\sum\theta_j}}{\prod r_j^2}\right]^{1/2}$$

or

$$\phi(t) = \frac{e^{i\sum\tan^{-1}(2t\lambda_j)/2}}{\prod(1+4t^2\lambda_j^2)^{1/4}}.$$

Assemble this to give

$$F_T(x) - F_T(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{i\theta(t)}}{\rho(t)}\int_0^x e^{-iyt}dydt$$

where

$$\theta(t) = \sum\tan^{-1}(2t\lambda_j)/2$$

and $\rho(t) = \prod(1+4t^2\lambda_j^2)^{1/4}$. But

$$\int_0^x e^{-iyt}dy = \frac{e^{-ixt}-1}{-it}.$$

We can now collect up the real part of the resulting integral to derive the formula given by Imhof. I don't produce the details here.

2): The central limit theorem (in some versions) can be deduced from the Fourier inversion formula: if $X_1,\ldots,X_n$ are iid with mean 0 and variance 1 and $T = n^{1/2}\bar{X}$ then with ϕ denoting the characteristic function of a single X we have

$$\begin{aligned} E(e^{itT}) &= E(e^{in^{-1/2}t\sum X_j}) \\ &= \left[\phi(n^{-1/2}t)\right]^n \\ &\approx \left[\phi(0) + \frac{t\phi'(0)}{\sqrt{n}} + \frac{t^2\phi''(0)}{2n} + o(n^{-1})\right]^n \end{aligned}$$

But now $\phi(0) = 1$ and

$$\phi'(t) = \frac{d}{dt}E(e^{itX_1}) = iE(X_1 e^{itX_1})\,.$$

So $\phi'(0) = E(X_1) = 0$. Similarly

$$\phi''(t) = i^2 E(X_1^2 e^{itX_1})$$

so that

$$\phi''(0) = -E(X_1^2) = -1\,.$$

It now follows that

$$\begin{aligned} E(e^{itT}) &\approx [1 - t^2/(2n) + o(1/n)]^n \\ &\to e^{-t^2/2}\,. \end{aligned}$$

With care we can then apply the Fourier inversion formula and get

$$\begin{aligned} f_T(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}[\phi(tn^{-1/2})]^n dt \\ &\to \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}e^{-t^2/2} dt \\ &= \frac{1}{\sqrt{2\pi}}\phi_Z(-x) \end{aligned}$$

where ϕ_Z is the characteristic function of a standard normal variable Z. Doing the integral we find

$$\phi_Z(x) = \phi_Z(-x) = e^{-x^2/2}$$

so that

$$f_T(x) \to \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

which is a standard normal density.

Proof of the central limit theorem not general: requires T to have bounded continuous density.

Central limit theorem: statement about cdfs not densities:

$$P(T \le t) \to P(Z \le t)\,.$$

3) Saddlepoint approximation from MGF inversion formula

$$2\pi i f(x) = \int_{-i\infty}^{i\infty} M(z)e^{-zx}dz$$

(limits of integration indicate contour integral running up imaginary axis.)

Replace contour (using complex variables) with line $Re(z) = c$. ($Re(Z)$ denotes the real part of z, that is, x when $z = x + iy$ with x and y real.) Must choose c so that $M(c) < \infty$. Rewrite inversion formula using cumulant generating function $K(t) = \log(M(t))$:

$$2\pi i f(x) = \int_{c-i\infty}^{c+i\infty} \exp(K(z) - zx)dz\,.$$

Along the contour in question we have $z = c + iy$ so we can think of the integral as being

$$i \int_{-\infty}^{\infty} \exp(K(c+iy) - (c+iy)x) dy \,.$$

Now do a Taylor expansion of the exponent:

$$\begin{aligned} &K(c+iy) - (c+iy)x = \\ &K(c) - cx + iy(K'(c) - x) - y^2 K''(c)/2 + \cdots \,. \end{aligned}$$

Ignore the higher order terms and select a c so that the first derivative

$$K'(c) - x$$

vanishes. Such a c is a saddlepoint. We get the formula

$$\begin{aligned} 2\pi f(x) \approx \exp(&K(c) - cx) \\ &\times \int_{-\infty}^{\infty} \exp(-y^2 K''(c)/2) dy \,. \end{aligned}$$

Integral is normal density calculation; gives

$$\sqrt{2\pi / K''(c)} \,.$$

Saddlepoint approximation is

$$f(x) = \frac{\exp(K(c) - cx)}{\sqrt{2\pi K''(c)}} \,.$$

Essentially same idea: Laplace's approximation.

Example: Sterling's approximation to factorial:

$$n! = \int_0^\infty \exp(n\log(x) - x)dx\,.$$

Exponent maximized when $x = n$.

For n large approximate $f(x) = n\log(x) - x$ by

$$f(x) \approx f(x_0) + (x - x_0)f'(x_0) + (x - x_0)^2 f''(x_0)/2$$

and choose $x_0 = n$ to make $f'(x_0) = 0$. Then

$$n! \approx \int_0^\infty \exp[n\log(n) - n - (x - n)^2/(2n)]dx\,.$$

Substitute $y = (x - n)/\sqrt{n}$; get approximation

$$n! \approx n^{1/2} n^n e^{-n} \int_{-\infty}^\infty e^{-y^2/2} dy$$

or

$$n! \approx \sqrt{2\pi} n^{n+1/2} e^{-n}\,.$$

Note: sloppy about limits of integration.

Real proof must show integral over x not near n is negligible.

Convergence in Distribution

Undergraduate version of central limit theorem: if $X_1, \ldots, X_n$ are iid from a population with mean μ and standard deviation σ then $n^{1/2}(\bar{X} - \mu)/\sigma$ has approximately a normal distribution.

Also Binomial(n, p) random variable has approximately a $N(np, np(1-p))$ distribution.

Precise meaning of statements like "X and Y have approximately the same distribution"?

Desired meaning: X and Y have nearly the same cdf.

But care needed.

Q1) If n is a large number is the $N(0,1/n)$ distribution close to the distribution of $X \equiv 0$?

Q2) Is $N(0,1/n)$ close to the $N(1/n,1/n)$ distribution?

Q3) Is $N(0,1/n)$ close to $N(1/\sqrt{n},1/n)$ distribution?

Q4) If $X_n \equiv 2^{-n}$ is the distribution of X_n close to that of $X \equiv 0$?

Answers depend on how close close needs to be so it's a matter of definition.

In practice the usual sort of approximation we want to make is to say that some random variable X, say, has nearly some continuous distribution, like $N(0,1)$.

So: want to know probabilities like $P(X > x)$ are nearly $P(N(0,1) > x)$.

Real difficulty: case of discrete random variables or infinite dimensions: not done in this course.

Mathematicians' meaning of close:

Either they can provide an upper bound on the distance between the two things or they are talking about taking a limit.

In this course we take limits.

Definition: A sequence of random variables X_n converges in distribution to a random variable X if

$$E(g(X_n)) \to E(g(X))$$

for every bounded continuous function g.

Theorem 3 *The following are equivalent:*

1. *X_n converges in distribution to X.*

2. *$P(X_n \le x) \to P(X \le x)$ for each x such that $P(X = x) = 0$.*

3. *The limit of the characteristic functions of X_n is the characteristic function of X:*

   $$E(e^{itX_n}) \to E(e^{itX})$$

   for every real t.

These are all implied by

$$M_{X_n}(t) \to M_X(t) < \infty$$

for all $|t| \le \epsilon$ for some positive ϵ.

Now let's go back to the questions I asked:

- $X_n \sim N(0, 1/n)$ and $X = 0$. Then

  $$P(X_n \le x) \to \begin{cases} 1 & x > 0 \\ 0 & x < 0 \\ 1/2 & x = 0 \end{cases}$$

  Now the limit is the cdf of $X = 0$ except for $x = 0$ and the cdf of X is not continuous at $x = 0$ so yes, X_n converges to X in distribution.

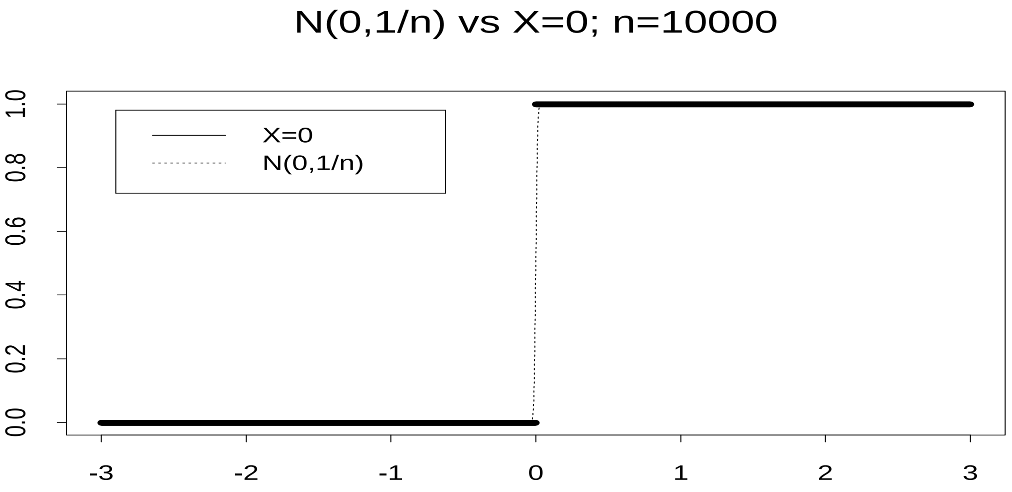
- I asked if $X_n \sim N(1/n, 1/n)$ had a distribution close to that of $Y_n \sim N(0, 1/n)$. The definition I gave really requires me to answer by finding a limit X and proving that both X_n and Y_n converge to X in distribution. Take $X = 0$. Then

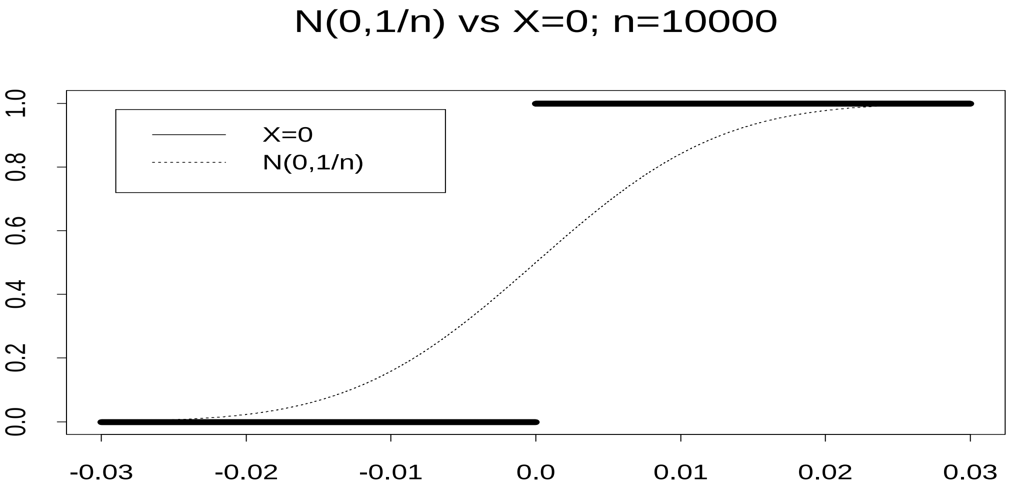
  $$E(e^{tX_n}) = e^{t/n+t^2/(2n)} \to 1 = E(e^{tX})$$

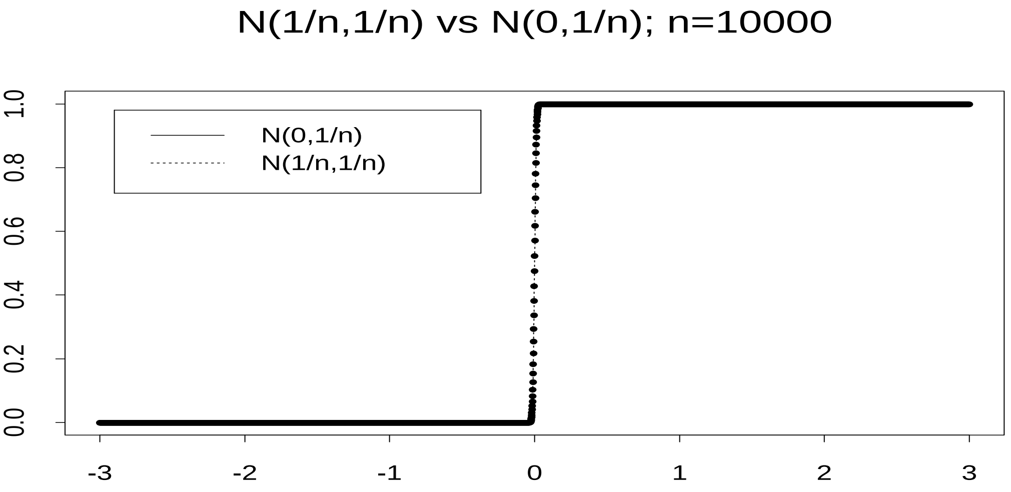
  and

  $$E(e^{tY_n}) = e^{t^2/(2n)} \to 1$$

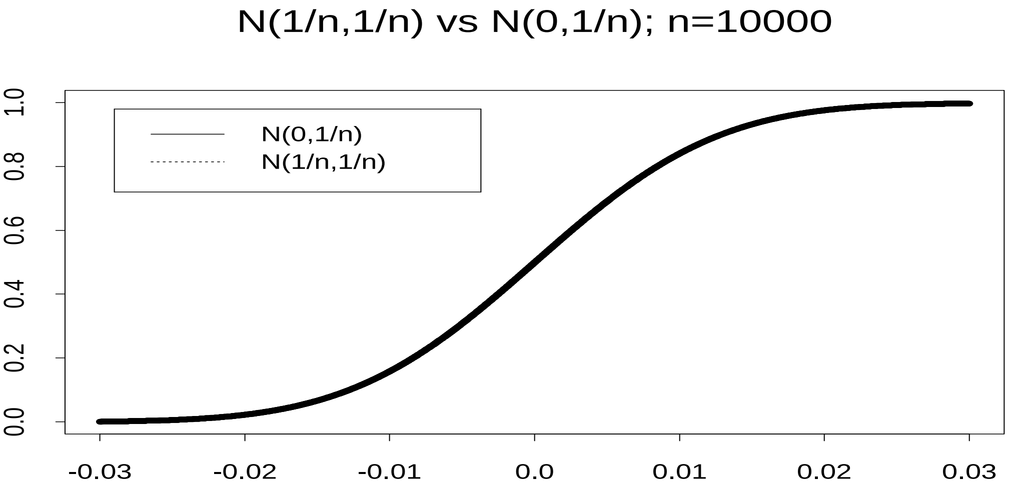
  so that both X_n and Y_n have the same limit in distribution.


N(0,1/n) vs X=0; n=10000
X=0
N(0,1/n)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3


N(0,1/n) vs X=0; n=10000
X=0
N(0,1/n)
0.0
0.2
0.4
0.6
0.8
1.0
-0.03
-0.02
-0.01
0.0
0.01
0.02
0.03



N(1/n,1/n) vs N(0,1/n); n=10000



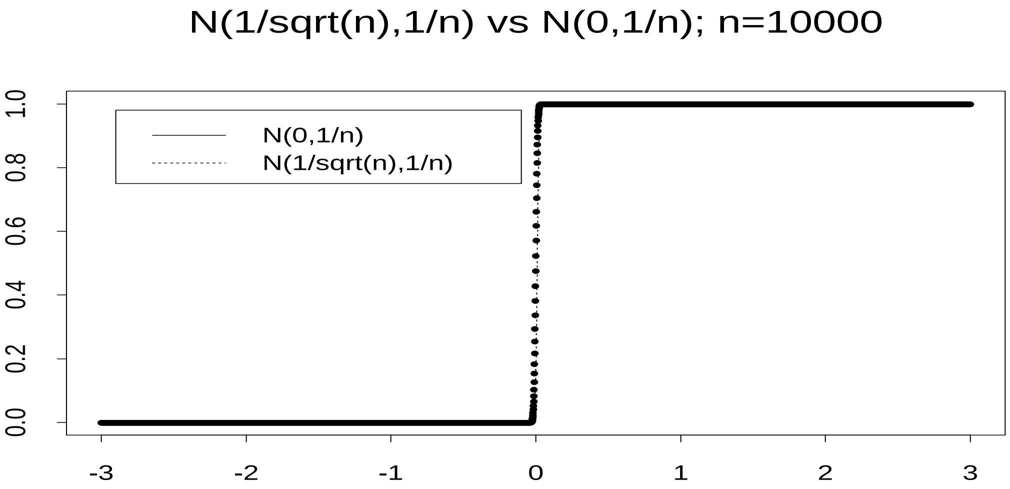
N(1/n,1/n) vs N(0,1/n); n=10000

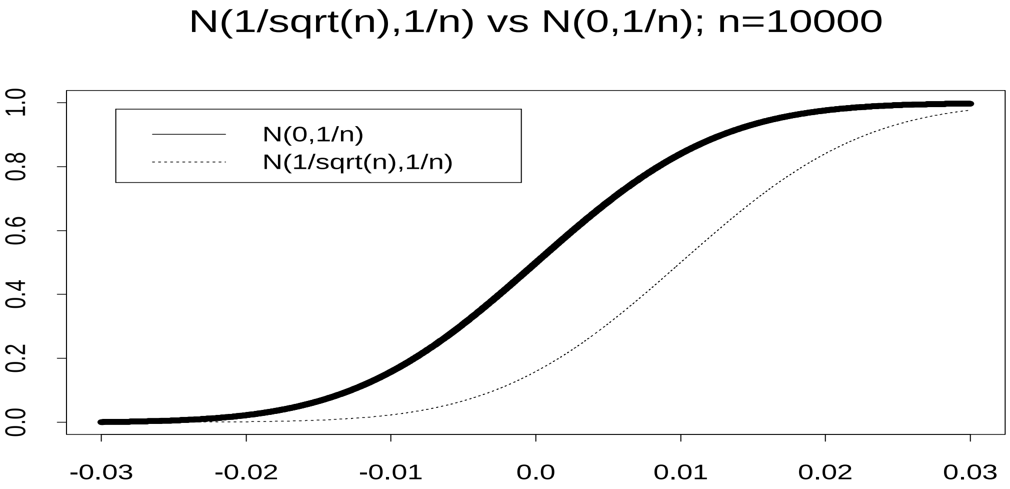
- Multiply both X_n and Y_n by $n^{1/2}$ and let $X \sim N(0,1)$. Then $\sqrt{n}X_n \sim N(n^{-1/2}, 1)$ and $\sqrt{n}Y_n \sim N(0,1)$. Use characteristic functions to prove that both $\sqrt{n}X_n$ and $\sqrt{n}Y_n$ converge to $N(0,1)$ in distribution.

- If you now let $X_n \sim N(n^{-1/2}, 1/n)$ and $Y_n \sim N(0, 1/n)$ then again both X_n and Y_n converge to 0 in distribution.

- If you multiply X_n and Y_n in the previous point by $n^{1/2}$ then $n^{1/2}X_n \sim N(1,1)$ and $n^{1/2}Y_n \sim N(0,1)$ so that $n^{1/2}X_n$ and $n^{1/2}Y_n$ are **not** close together in distribution.

- You can check that $2^{-n} \to 0$ in distribution.


N(1/sqrt(n),1/n) vs N(0,1/n); n=10000


N(1/sqrt(n),1/n) vs N(0,1/n); n=10000

Summary: to derive approximate distributions:

Show sequence of rvs X_n converges to some X.

The limit distribution (i.e. dstbon of X) should be non-trivial, like say $N(0,1)$.

Don't say: X_n is approximately $N(1/n, 1/n)$.

Do say: $n^{1/2}(X_n - 1/n)$ converges to $N(0,1)$ in distribution.

The Central Limit Theorem

If $X_1, X_2, \cdots$ are iid with mean 0 and variance 1 then $n^{1/2}\bar{X}$ converges in distribution to $N(0,1)$. That is,

$$P(n^{1/2}\bar{X} \le x) \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2} dy \, .$$

Proof: As before

$$E(e^{itn^{1/2}\bar{X}}) \to e^{-t^2/2} \, .$$

This is the characteristic function of $N(0,1)$ so we are done by our theorem.

Edgeworth expansions

In fact if $\gamma = E(X^3)$ then

$$\phi(t) \approx 1 - t^2/2 - i\gamma t^3/6 + \cdots$$

keeping one more term. Then

$$\log(\phi(t)) = \log(1+u)$$

where

$$u = -t^2/2 - i\gamma t^3/6 + \cdots .$$

Use $\log(1+u) = u - u^2/2 + \cdots$ to get

$$\begin{aligned}\log(\phi(t)) \approx &\\ &[-t^2/2 - i\gamma t^3/6 + \cdots] \\ &\qquad - [\cdots]^2/2 + \cdots\end{aligned}$$

which rearranged is

$$\log(\phi(t)) \approx -t^2/2 - i\gamma t^3/6 + \cdots .$$

Now apply this calculation to

$$\log(\phi_T(t)) \approx -t^2/2 - iE(T^3)t^3/6 + \cdots .$$

Remember $E(T^3) = \gamma/\sqrt{n}$ and exponentiate to get

$$\phi_T(t) \approx e^{-t^2/2} \exp\{-i\gamma t^3/(6\sqrt{n}) + \cdots\} .$$

You can do a Taylor expansion of the second exponential around 0 because of the square root of n and get

$$\phi_T(t) \approx e^{-t^2/2}(1 - i\gamma t^3/(6\sqrt{n}))$$

neglecting higher order terms. This approximation to the characteristic function of T can be inverted to get an **Edgeworth** approximation to the density (or distribution) of T which looks like

$$f_T(x) \approx \frac{1}{\sqrt{2\pi}} e^{-x^2/2}[1 - \gamma(x^3 - 3x)/(6\sqrt{n}) + \cdots] .$$

Remarks:

1. The error using the central limit theorem to approximate a density or a probability is proportional to $n^{-1/2}$.

2. This is improved to n^{-1} for symmetric densities for which $\gamma = 0$.

3. These expansions are **asymptotic**. This means that the series indicated by $\cdots$ usually does **not** converge. When $n = 25$ it may help to take the second term but get worse if you include the third or fourth or more.

4. You can integrate the expansion above for the density to get an approximation for the cdf.

Multivariate convergence in distribution

Definition: $X_n \in R^p$ converges in distribution to $X \in R^p$ if

$$E(g(X_n)) \to E(g(X))$$

for each bounded continuous real valued function g on R^p.

This is equivalent to either of

Cramér Wold Device: $a^t X_n$ converges in distribution to $a^t X$ for each $a \in R^p$.

or

Convergence of characteristic functions:

$$E(e^{ia^t X_n}) \to E(e^{ia^t X})$$

for each $a \in R^p$.

Extensions of the CLT

1. $Y_1, Y_2, \cdots$ iid in R^p, mean μ, variance covariance Σ then $n^{1/2}(\bar{Y} - \mu)$ converges in distribution to $MVN(0, \Sigma)$.

2. Lyapunov CLT: for each n $X_{n1}, \ldots, X_{nn}$ independent rvs with
$$\begin{aligned} E(X_{ni}) &= 0 \\ \text{Var}(\sum_i X_{ni}) &= 1 \\ \sum E(|X_{ni}|^3) &\to 0 \end{aligned}$$
then $\sum_i X_{ni}$ converges to $N(0,1)$.

3. Lindeberg CLT: 1st two conds of Lyapunov and
$$\sum E(X_{ni}^2 1(|X_{ni}| > \epsilon)) \to 0$$
each $\epsilon > 0$. Then $\sum_i X_{ni}$ converges in distribution to $N(0,1)$. (Lyapunov's condition implies Lindeberg's.)

4. Non-independent rvs: m-dependent CLT, martingale CLT, CLT for mixing processes.

5. Not sums: Slutsky's theorem, δ method.

Slutsky's Theorem: If X_n converges in distribution to X and Y_n converges in distribution (or in probability) to c, a constant, then $X_n + Y_n$ converges in distribution to $X + c$. More generally, if $f(x, y)$ is continuous then $f(X_n, Y_n) \Rightarrow f(X, c)$.

Warning: the hypothesis that the limit of Y_n be constant is essential.

Definition: We say Y_n converges to Y in probability if $\forall \epsilon > 0$:

$$P(|Y_n - Y| > \epsilon) \to 0 \, .$$

Fact: for Y constant convergence in distribution and in probability are the same.

Always convergence in probability implies convergence in distribution.

Both are weaker than almost sure convergence:

Definition: We say Y_n converges to Y almost surely if

$$P(\{\omega \in \Omega : \lim_{n \to \infty} Y_n(\omega) = Y(\omega)\}) = 1 \, .$$

The delta method: Suppose:

- Sequence Y_n of rvs converges to some y, a constant.

- $X_n = a_n(Y_n - y)$ then X_n converges in distribution to some random variable X.

- f is differentiable ftn on range of Y_n.

Then $a_n(f(Y_n) - f(y))$ converges in distribution to $f'(y)X$.

If $X_n \in R^p$ and $f : R^p \mapsto R^q$ then f' is $q \times p$ matrix of first derivatives of components of f.

Example: Suppose $X_1, \ldots, X_n$ are a sample from a population with mean μ, variance σ^2, and third and fourth central moments μ_3 and μ_4. Then

$$n^{1/2}(s^2 - \sigma^2) \Rightarrow N(0, \mu_4 - \sigma^4)$$

where $\Rightarrow$ is notation for convergence in distribution. For simplicity I define $s^2 = \overline{X^2} - \bar{X}^2$.

How to apply δ method:

1) Write statistic as a function of averages:

Define

$$W_i = \left[\begin{array}{c} X_i^2 \\ X_i \end{array} \right] .$$

See that

$$\bar{W}_n = \left[\begin{array}{c} \overline{X^2} \\ \overline{X} \end{array} \right]$$

Define

$$f(x_1, x_2) = x_1 - x_2^2$$

See that $s^2 = f(\bar{W}_n)$.

2) Compute mean of your averages:

$$\mu_W \equiv \mathsf{E}(\bar{W}_n) = \left[\begin{array}{c} \mathsf{E}(X_i^2) \\ \mathsf{E}(X_i) \end{array} \right] = \left[\begin{array}{c} \mu^2 + \sigma^2 \\ \mu \end{array} \right] .$$

3) In δ method theorem take $Y_n = \bar{W}_n$ and $y = \mathsf{E}(Y_n)$.

4) Take $a_n = n^{1/2}$.

5) Use central limit theorem:

$$n^{1/2}(Y_n - y) \Rightarrow MVN(0, \Sigma)$$

where $\Sigma = \text{Var}(W_i)$.

6) To compute Σ take expected value of

$$(W - \mu_W)(W - \mu_W)^t$$

There are 4 entries in this matrix. Top left entry is

$$(X^2 - \mu^2 - \sigma^2)^2$$

This has expectation:

$$\text{E}\left\{(X^2 - \mu^2 - \sigma^2)^2\right\} = \text{E}(X^4) - (\mu^2 + \sigma^2)^2 .$$

Using binomial expansion:

$$\begin{aligned}\mathrm{E}(X^4) &= \mathrm{E}\{(X-\mu+\mu)^4\} \\ &= \mu_4 + 4\mu\mu_3 + 6\mu^2\sigma^2 \\ &\quad + 4\mu^3\mathrm{E}(X-\mu) + \mu^4\,.\end{aligned}$$

So

$$\Sigma_{11} = \mu_4 - \sigma^4 + 4\mu\mu_3 + 4\mu^2\sigma^2$$

Top right entry is expectation of

$$(X^2-\mu^2-\sigma^2)(X-\mu)$$

which is

$$\mathrm{E}(X^3) - \mu\mathrm{E}(X^2)$$

Similar to 4th moment get

$$\mu_3 + 2\mu\sigma^2$$

Lower right entry is σ^2.

So

$$\Sigma = \left[\begin{array}{cc} \mu_4 - \sigma^4 + 4\mu\mu_3 + 4\mu^2\sigma^2 & \mu_3 + 2\mu\sigma^2 \\ \mu_3 + 2\mu\sigma^2 & \sigma^2 \end{array}\right]$$

7) Compute derivative (gradient) of f: has components $(1, -2x_2)$. Evaluate at $y = (\mu^2 + \sigma^2, \mu)$ to get

$$a^t = (1, -2\mu) \, .$$

This leads to

$$n^{1/2}(s^2 - \sigma^2) \approx$$
$$n^{1/2}[1, -2\mu] \left[\begin{array}{c} \overline{X^2} - (\mu^2 + \sigma^2) \\ \bar{X} - \mu \end{array} \right]$$

which converges in distribution to

$$(1, -2\mu) MVN(0, \Sigma) \, .$$

This rv is $N(0, a^t \Sigma a) = N(0, \mu_4 - \sigma^4)$.

Alternative approach worth pursuing. Suppose c is constant.

Define $X_i^* = X_i - c$.

Then: sample variance of X_i^* is same as sample variance of X_i.

Notice all central moments of X_i^* same as for X_i. Conclusion: no loss in $\mu = 0$. In this case:

$$a^t = (1,0)$$

and

$$\Sigma = \left[\begin{array}{cc} \mu_4 - \sigma^4 & \mu_3 \\ \mu_3 & \sigma^2 \end{array} \right] .$$

Notice that

$$a^t\Sigma = [\mu_4 - \sigma^4, \mu_3]$$

and

$$a^t\Sigma a = \mu_4 - \sigma^4 .$$

Special case: if population is $N(\mu,\sigma^2)$ then $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$. Our calculation has

$$n^{1/2}(s^2-\sigma^2) \Rightarrow N(0,2\sigma^4)$$

You can divide through by σ^2 and get

$$n^{1/2}(\frac{s^2}{\sigma^2}-1) \Rightarrow N(0,2)$$

In fact ns^2/σ^2 has a χ^2_{n-1} distribution and so the usual central limit theorem shows that

$$(n-1)^{-1/2}[ns^2/\sigma^2-(n-1)] \Rightarrow N(0,2)$$

(using mean of χ^2_1 is 1 and variance is 2).

Factor out n to get

$$\sqrt{\frac{n}{n-1}}n^{1/2}(s^2/\sigma^2-1)+(n-1)^{-1/2} \Rightarrow N(0,2)$$

which is δ method calculation except for some constants.

Difference is unimportant: Slutsky's theorem.

Monte Carlo

Given rvs $X_1, \ldots, X_n$; distbn specified.

Statistic $T(X_1, \ldots, X_n)$ whose dstbn wanted.

To compute $P(T > t)$:

1. Generate $X_1, \ldots, X_n$ from the density f.

2. Compute $T_1 = T(X_1, \ldots, X_n)$.

3. Repeat N times getting $T_1, \ldots, T_N$.

4. Estimate $p = P(T > t)$ as $\hat{p} = M/N$ where M is number of repetitions where $T_i > t$.

5. Estimate accuracy of $\hat{p}$ using $\sqrt{\hat{p}(1-\hat{p})/N}$.

Note: accuracy inversely proportional to $\sqrt{N}$.

Next: tricks to make method more accurate.
Warning: tricks only change constant — SE still inversely proportional to $\sqrt{N}$.

Generating the Sample

Transformation

Basic computing tool: pseudo uniform random numbers — variables U which have (approximately) a Uniform$[0,1]$ distribution.

Other dstbns generated by transformation:

Exponential: $X = -\log U$ has an exponential distribution:

$$\begin{aligned} P(X > x) &= P(-\log(U) > x) \\ &= P(U \le e^{-x}) = e^{-x} \end{aligned}$$

Pitfall: Random uniforms generated on computer sometimes have only 6 or 7 digits.

Consequence: tail of generated dstbn grainy.

Explanation: suppose U multiple of 10^{-6}.

Largest possible value of X is $6\log(10)$.

Improved algorithm:

- Generate U a Uniform[0,1] variable.
- Pick a small ϵ like 10^{-3} say. If $U > \epsilon$ take $Y = -\log(U)$.
- If $U \le \epsilon$: conditional dstbn of $Y - y$ given $Y > y$ is exponential. Generate new U'. Compute $Y' = -\log(U')$. Take $Y = Y' - \log(\epsilon)$.

Resulting Y has exponential distribution.

Exercise: check by computing $P(Y > y)$.

General technique: inverse probability integral transform.

If Y is to have cdf F:

Generate $U \sim Uniform[0,1]$.

Take $Y = F^{-1}(U)$:

$$\begin{aligned} P(Y \le y) &= P(F^{-1}(U) \le y) \\ &= P(U \le F(y)) = F(y) \end{aligned}$$

Example: X exponential. $F(x) = 1 - e^{-x}$ and $F^{-1}(u) = -\log(1-u)$.

Compare to previous method. (Use U instead of $1-U$.)

Normal: $F = \Phi$ (common notation for standard normal cdf).

No closed form for F^{-1}.

One way: use numerical algorithm to compute F^{-1}.

Alternative: Box Müller

Generate U_1, U_2 two independent Uniform[0,1] variables.

Define

$$Y_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2)$$

and

$$Y_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)\,.$$

Exercise: (use change of variables) Y_1 and Y_2 are independent $N(0,1)$ variables.

Acceptance Rejection

Suppose: can't calculate F^{-1} but know f.

Find density g and constant c such that

1) $f(x) \le cg(x)$ for each x and

2) G^{-1} is computable or can generate observations $W_1, W_2, \ldots$ independently from g.

Algorithm:

1) Generate W_1.

2) Compute $p = f(W_1)/(cg(W_1)) \le 1$.

3) Generate uniform[0,1] random variable U_1 independent of all Ws.

4) Let $Y = W_1$ if $U_1 \le p$.

5) Otherwise get new W, U; repeat until you find $U_i \le f(W_i)/(cg(W_i))$.

6) Make Y be last W generated.

This Y has density f.

Markov Chain Monte Carlo

Recently popular tactic, particularly for generating multivariate observations.

Theorem Suppose $W_1, W_2, \ldots$ is an (ergodic) Markov chain with stationary transitions and the stationary initial distribution of W has density f. Then starting the chain with *any* initial distribution

$$\frac{1}{n}\sum_{i=1}^{n} g(W_i) \to \int g(x)f(x)dx\,.$$

Estimate things like $\int_A f(x)dx$ by computing the fraction of the W_i which land in A.

Many versions of this technique including Gibbs Sampling and Metropolis-Hastings algorithm.

Technique invented in 1950s: Metropolis et al.

One of the authors was Edward Teller "father of the hydrogen bomb".

Importance Sampling

If you want to compute

$$\theta \equiv E(T(X)) = \int T(x)f(x)dx$$

you can generate observations from a different density g and then compute

$$\widehat{\theta} = n^{-1}\sum T(X_i)f(X_i)/g(X_i)$$

Then

$$\begin{aligned} E(\widehat{\theta}) &= n^{-1}\sum E\left\{T(X_i)f(X_i)/g(X_i)\right\} \\ &= \int \{T(x)f(x)/g(x)\}g(x)dx \\ &= \int T(x)f(x)dx \\ &= \theta \end{aligned}$$

Variance reduction

Example: estimate dstbn of sample mean for a Cauchy random variable.

Cauchy density is

$$f(x) = \frac{1}{\pi(1+x^2)}$$

Generate $U_1, \ldots, U_n$ uniforms.

Define $X_i = \tan^{-1}(\pi(U_i - 1/2))$.

Compute $T = \bar{X}$.

To estimate $p = P(T > t)$ use

$$\hat{p} = \sum_{i=1}^{N} 1(T_i > t)/N$$

after generating N samples of size n.

Estimate is unbiased.

Standard error is $\sqrt{p(1-p)/N}$.

Improvement: $-X_i$ also has Cauchy dstbn.

Take $S_i = -T_i$.

Remember that S_i has same dstbn as T_i.

Try (for $t > 0$)

$$\tilde{p} = [\sum_{i=1}^{N} 1(T_i > t) + \sum_{i=1}^{N} 1(S_i > t)]/(2N)$$

which is the average of two estimates like $\hat{p}$.

The variance of $\tilde{p}$ is

$$\begin{aligned}(4N)^{-1}\text{Var}(1(T_i > t) + 1(S_i > t))& \\ &= (4N)^{-1}\text{Var}(1(|T| > t))\end{aligned}$$

which is

$$\frac{2p(1-2p)}{4N} = \frac{p(1-2p)}{2N}$$

Variance has extra 2 in denominator and numerator is also smaller – particularly for p near 1/2.

So need only half the sample size to get the same accuracy.

Regression estimates

Suppose $Z \sim N(0,1)$. Compute

$$\theta = E(|Z|)\,.$$

Generate N iid $N(0,1)$ variables $Z_1,\ldots,Z_N$.

Compute $\hat\theta = \sum |Z_i|/N$.

But know $E(Z_i^2) = 1$.

Also: $\hat\theta$ is positively correlated with $\sum Z_i^2/N$.

So we try

$$\tilde\theta = \hat\theta - c(\sum Z_i^2/N - 1)$$

Notice that $E(\tilde\theta) = \theta$ and

$$\begin{aligned}\text{Var}(\tilde\theta) = \\ &\text{Var}(\hat\theta) - 2c\text{Cov}(\hat\theta, \sum Z_i^2/N) \\ &\qquad + c^2\text{Var}(\sum Z_i^2/N)\end{aligned}$$

The value of c which minimizes this is

$$c = \frac{\text{Cov}(\hat\theta, \sum Z_i^2/N)}{\text{Var}(\sum Z_i^2/N)}$$

Estimate c by regressing $|Z_i|$ on Z_i^2!

Statistical Inference

Definition: A **model** is a family $\{P_\theta; \theta \in \Theta\}$ of possible distributions for some random variable X.

WARNING: Data set is X, so X will generally be a big vector or matrix or even more complicated object.)

Assumption in this course: true distribution P of X is P_{θ_0} for some $\theta_0 \in \Theta$.

JARGON: θ_0 is *true value* of the parameter.

Notice: this assumption is wrong; we hope it is not wrong in an important way.

If it's wrong: enlarge model, put in more distributions, make Θ bigger.

Goal: observe value of X, guess θ_0 or some property of θ_0.

Classic mathematical versions of guessing:

1. Point estimation: compute estimate $\hat{\theta} = \hat{\theta}(X)$ which lies in Θ (or something close to Θ).

2. Point estimation of ftn of θ: compute estimate $\hat{\phi} = \hat{\phi}(X)$ of $\phi = g(\theta)$.

3. Interval (or set) estimation: compute set $C = C(X)$ in Θ which we think will contain θ_0.

4. Hypothesis testing: choose between $\theta_0 \in \Theta_0$ and $\theta_0 \notin \Theta_0$ where $\Theta_0 \subset \Theta$.

5. Prediction: guess value of an observable random variable Y whose distribution depends on θ_0. Typically Y is the value of the variable X in a repetition of the experiment.

Several schools of statistical thinking. Main schools of thought summarized roughly as follows:

- **Neyman Pearson**: A statistical procedure is evaluated by its long run frequency performance. Imagine repeating the data collection exercise many times, independently. Quality of procedure measured by its average performance when true distribution of X values is P_{θ_0}.

- **Bayes**: Treat θ as random just like X. Compute conditional law of unknown quantities given knowns. In particular ask how procedure will work on the data we actually got – no averaging over data we might have got.

- **Likelihood**: Try to combine previous 2 by looking only at actual data while trying to avoid treating θ as random.

We use Neyman Pearson approach to evaluate quality of likelihood and other methods.

Likelihood Methods of Inference

Toss coin 6 times and get Heads twice.

p is probability of getting H.

Probability of getting exactly 2 heads is

$$15p^2(1-p)^4$$

This function of p, is **likelihood** function.

Definition: The likelihood function is map L: domain Θ, values given by

$$L(\theta) = f_\theta(X)$$

Key Point: think about how the density depends on θ not about how it depends on X.

Notice: X, observed value of the data, has been plugged into the formula for density.

Notice: coin tossing example uses the discrete density for f.

We use likelihood for most inference problems:

1. Point estimation: we must compute an estimate $\widehat{\theta} = \widehat{\theta}(X)$ which lies in Θ. The **maximum likelihood estimate (MLE)** of θ is the value $\widehat{\theta}$ which maximizes $L(\theta)$ over $\theta \in \Theta$ if such a $\widehat{\theta}$ exists.

2. Point estimation of a function of θ: we must compute an estimate $\widehat{\phi} = \widehat{\phi}(X)$ of $\phi = g(\theta)$. We use $\widehat{\phi} = g(\widehat{\theta})$ where $\widehat{\theta}$ is the MLE of θ.

3. Interval (or set) estimation. We must compute a set $C = C(X)$ in Θ which we think will contain θ_0. We will use
$$\{\theta \in \Theta : L(\theta) > c\}$$
for a suitable c.

4. Hypothesis testing: decide whether or not $\theta_0 \in \Theta_0$ where $\Theta_0 \subset \Theta$. We base our decision on the likelihood ratio
$$\frac{\text{sup}\{L(\theta); \theta \in \Theta \setminus \Theta_0\}}{\text{sup}\{L(\theta); \theta \in \Theta_0\}}.$$

Maximum Likelihood Estimation

To find MLE maximize L.

Typical function maximization problem:

Set gradient of L equal to 0

Check root is maximum, not minimum or saddle point.

Examine some likelihood plots in examples:

Cauchy Data

IID sample $X_1, \ldots, X_n$ from Cauchy(θ) density

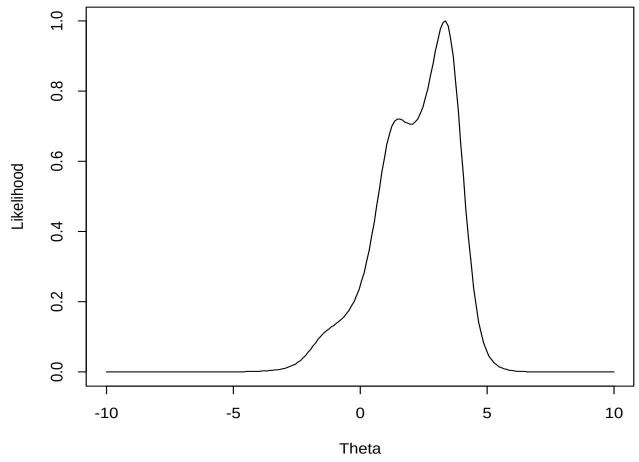
$$f(x;\theta) = \frac{1}{\pi(1+(x-\theta)^2)}$$

The likelihood function is

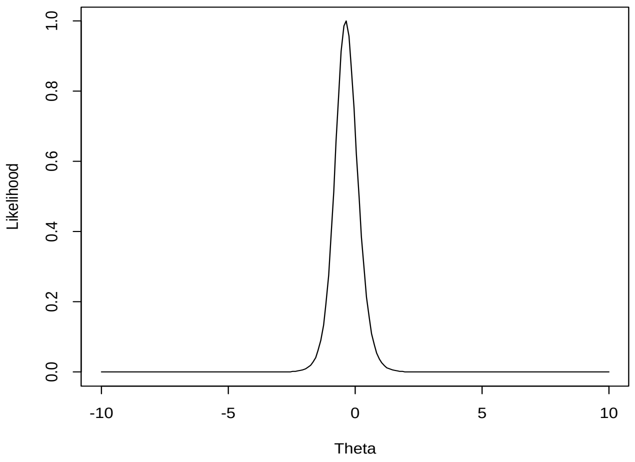
$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\pi(1+(X_i-\theta)^2)}$$

[Examine likelihood plots.]

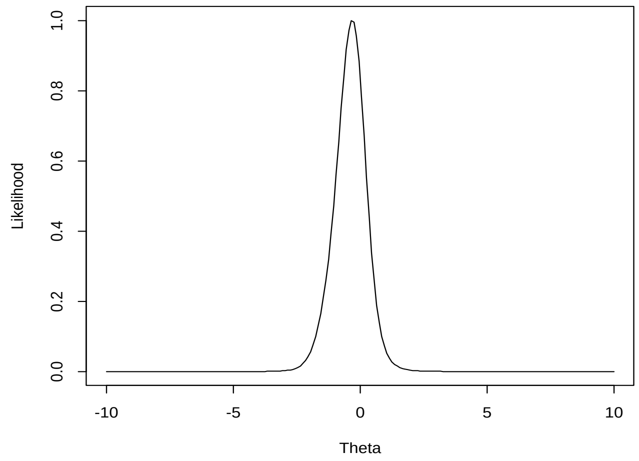
Likelihood Function: Cauchy, n=5



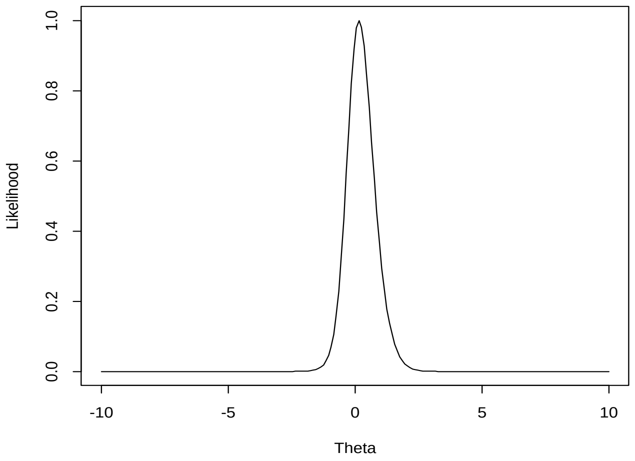
Likelihood Function: Cauchy, n=5



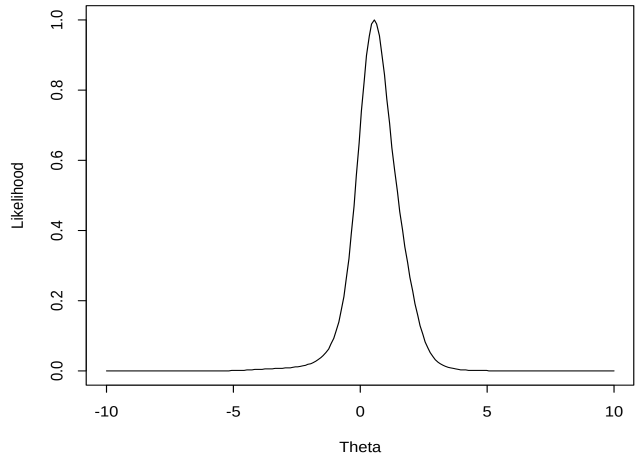
Likelihood Function: Cauchy, n=5



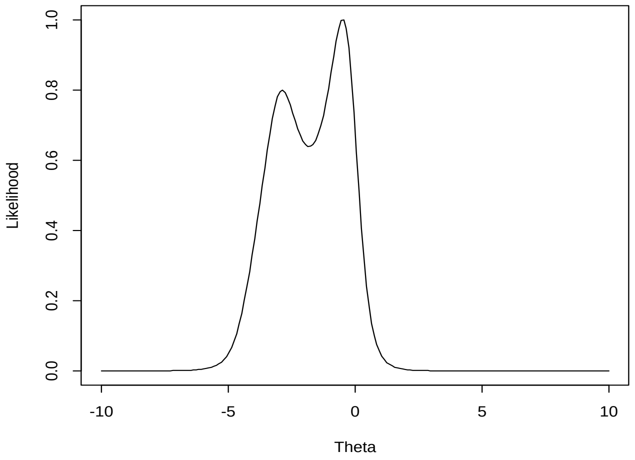
Likelihood Function: Cauchy, n=5



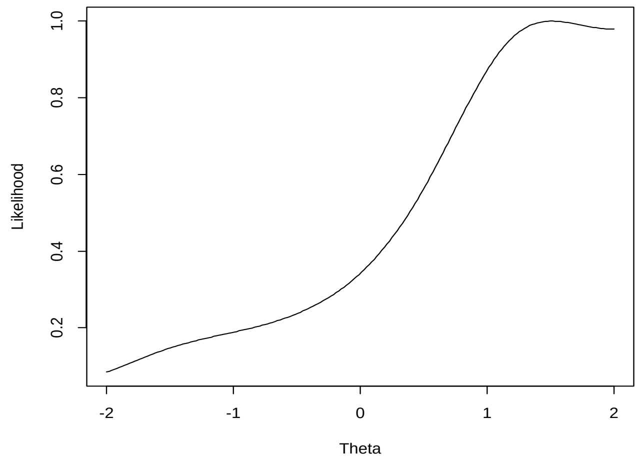
Likelihood Function: Cauchy, n=5



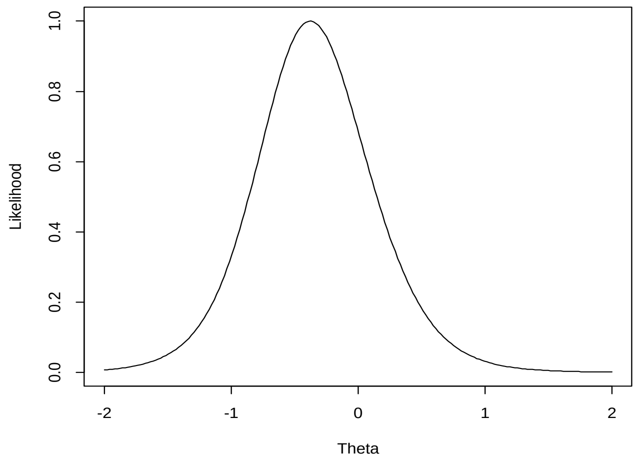
Likelihood Function: Cauchy, n=5



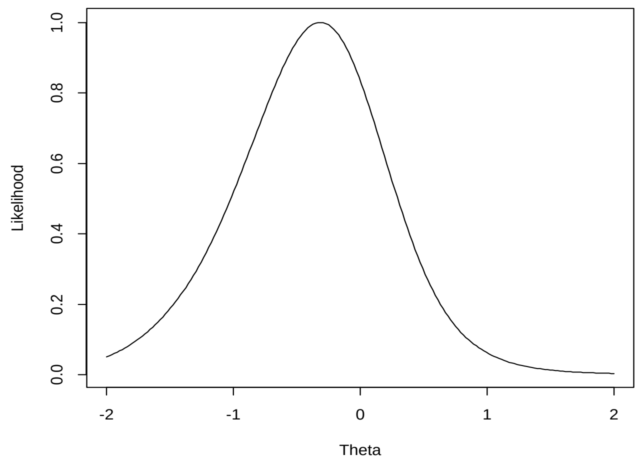
Likelihood Function: Cauchy, n=5



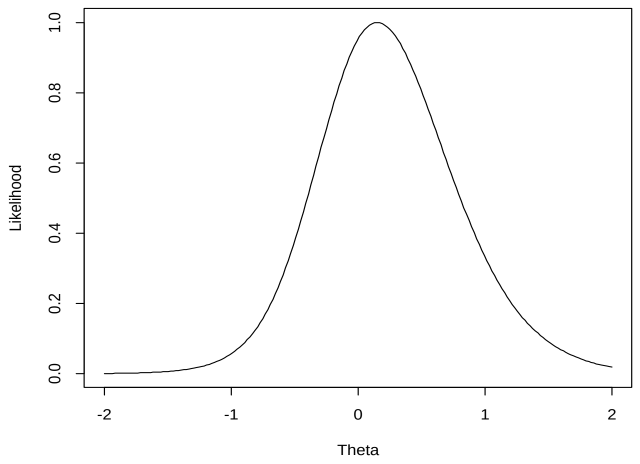
Likelihood Function: Cauchy, n=5



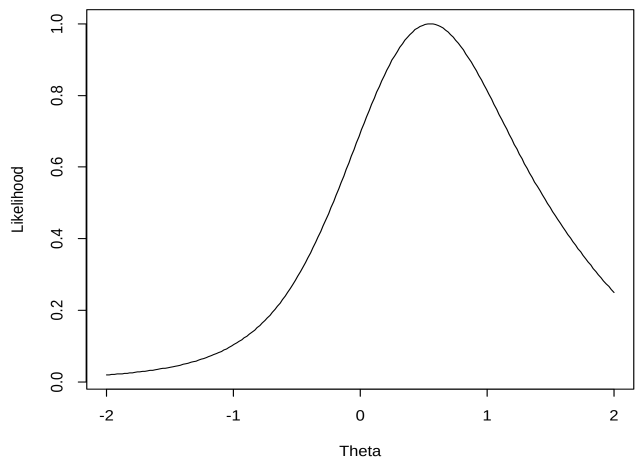
Likelihood Function: Cauchy, n=5



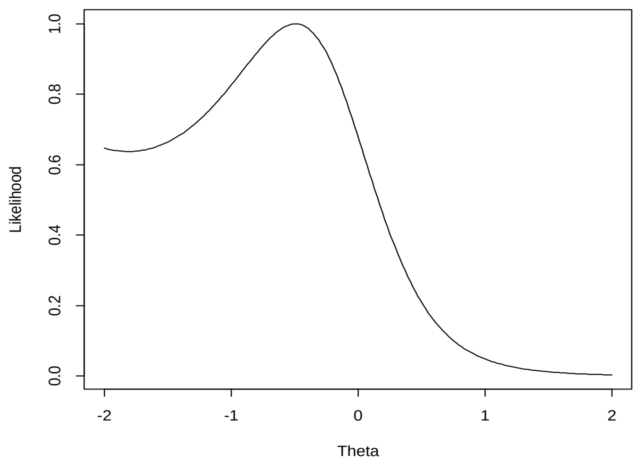
Likelihood Function: Cauchy, n=5



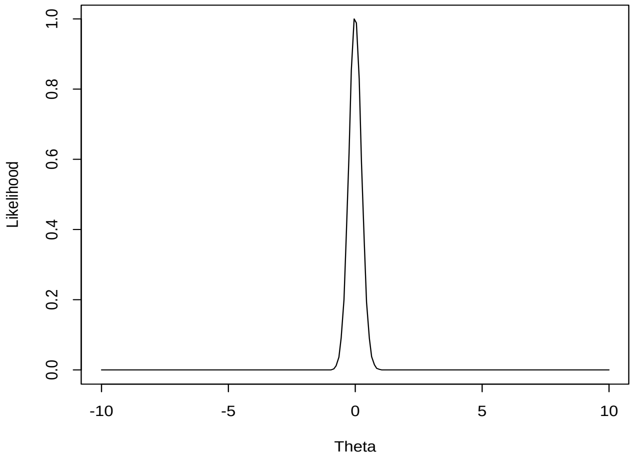
Likelihood Function: Cauchy, n=5



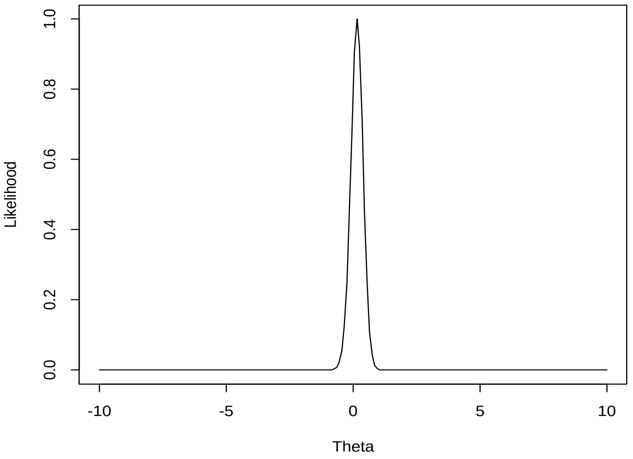
Likelihood Function: Cauchy, n=5



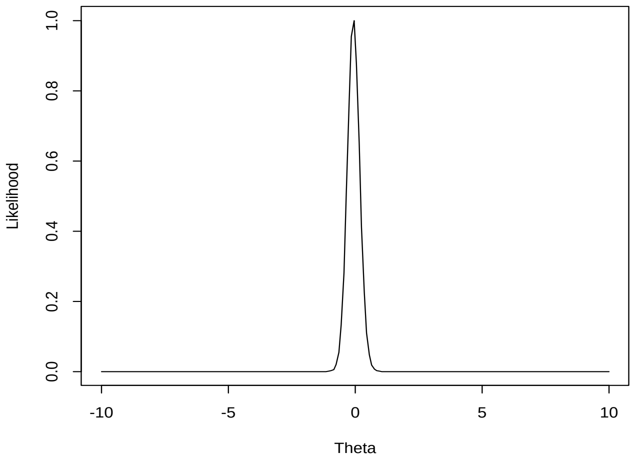
Likelihood Function: Cauchy, n=25



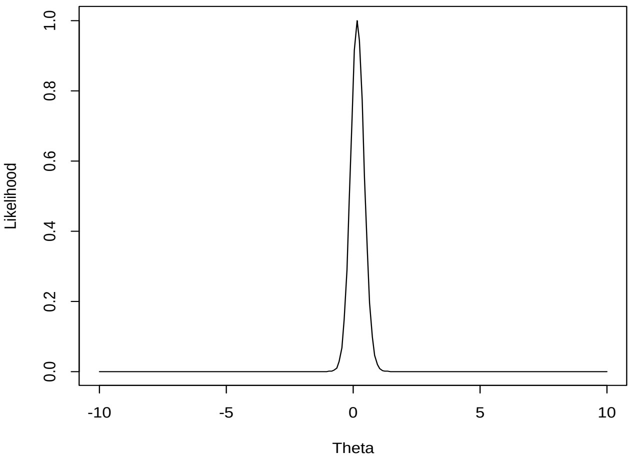
Likelihood Function: Cauchy, n=25



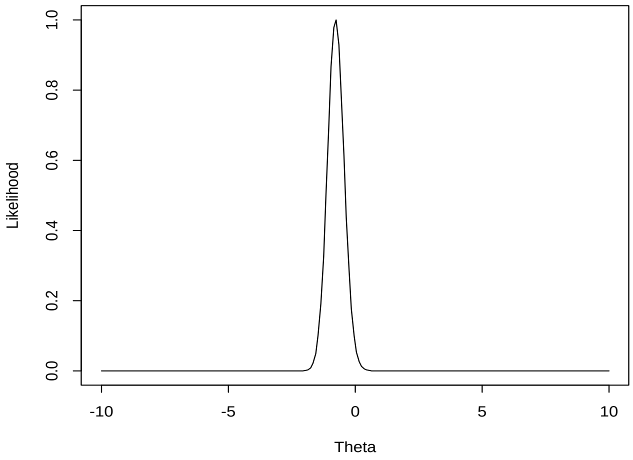
Likelihood Function: Cauchy, n=25



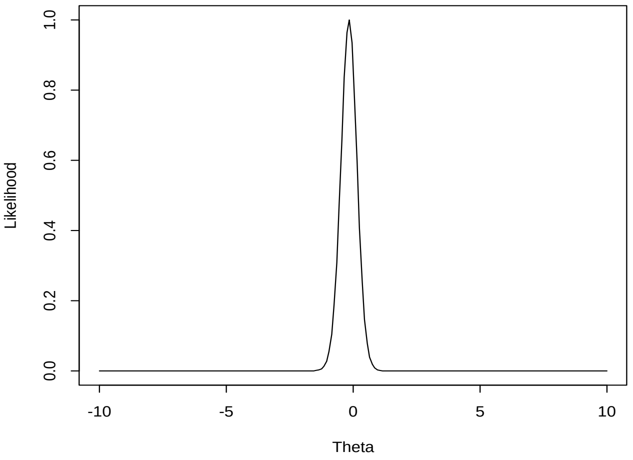
Likelihood Function: Cauchy, n=25



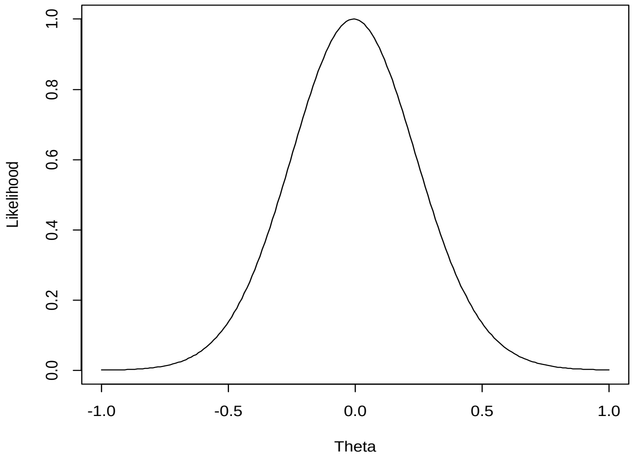
Likelihood Function: Cauchy, n=25



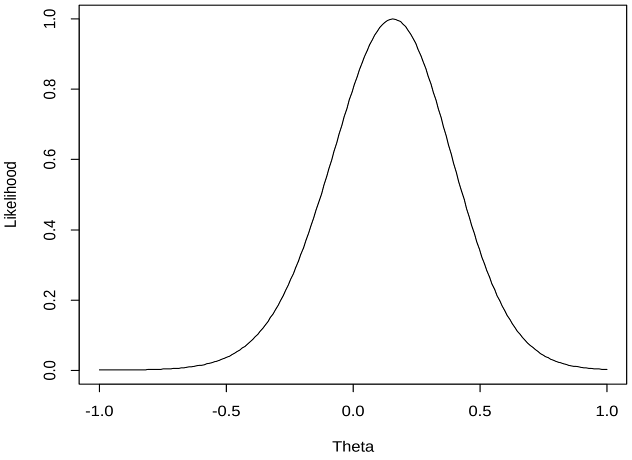
Likelihood Function: Cauchy, n=25



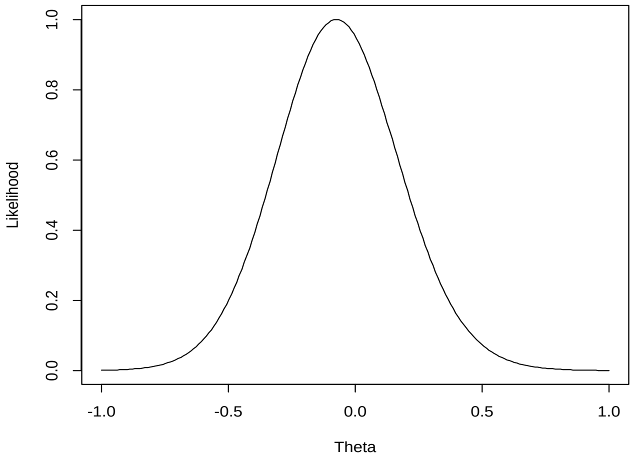
Likelihood Function: Cauchy, n=25



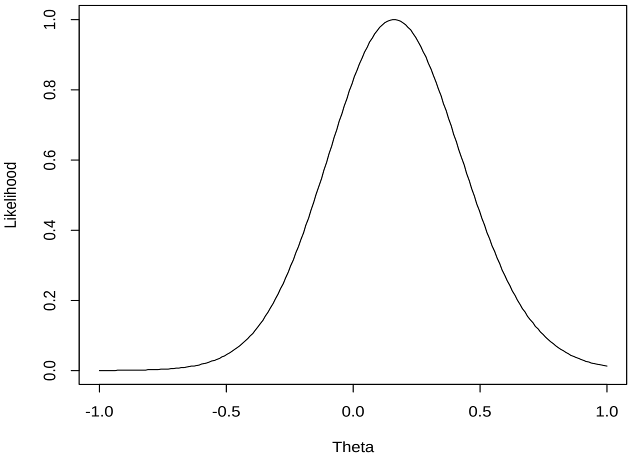
Likelihood Function: Cauchy, n=25



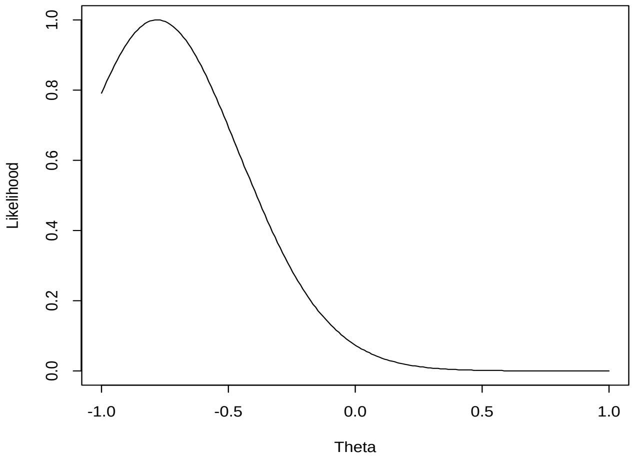
Likelihood Function: Cauchy, n=25



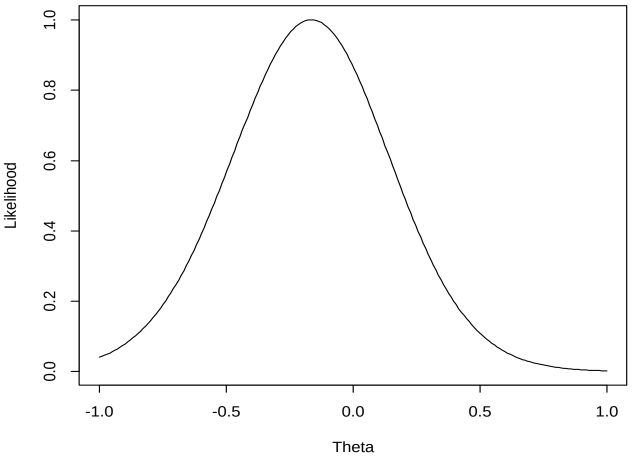
Likelihood Function: Cauchy, n=25



Likelihood Function: Cauchy, n=25



Likelihood Function: Cauchy, n=25



I want you to notice the following points:

- The likelihood functions have peaks near the true value of θ (which is 0 for the data sets I generated).

- The peaks are narrower for the larger sample size.

- The peaks have a more regular shape for the larger value of n.

- I actually plotted $L(\theta)/L(\hat{\theta})$ which has exactly the same shape as L but runs from 0 to 1 on the vertical scale.

To maximize this likelihood: differentiate L, set result equal to 0.

Notice L is product of n terms; derivative is

$$\sum_{i=1}^n \prod_{j\neq i} \frac{1}{\pi(1+(X_j-\theta)^2)} \frac{2(X_i-\theta)}{\pi(1+(X_i-\theta)^2)^2}$$

which is quite unpleasant.

Much easier to work with logarithm of L: log of product is sum and logarithm is monotone increasing.

Definition: The **Log Likelihood** function is

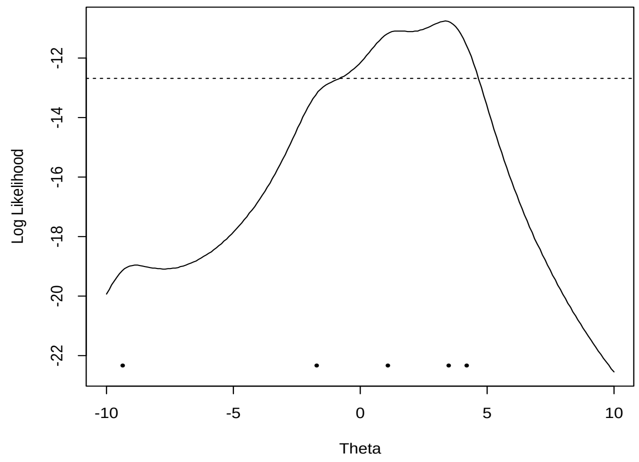
$$\ell(\theta) = \log\{L(\theta)\} .$$

For the Cauchy problem we have

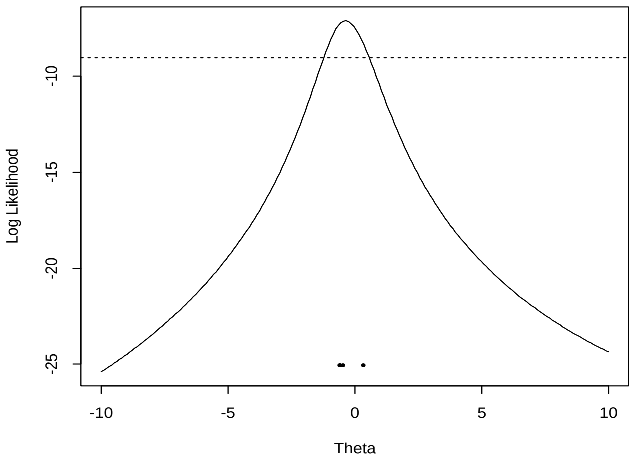
$$\ell(\theta) = -\sum \log(1+(X_i-\theta)^2) - n\log(\pi)$$

[Examine log likelihood plots.]

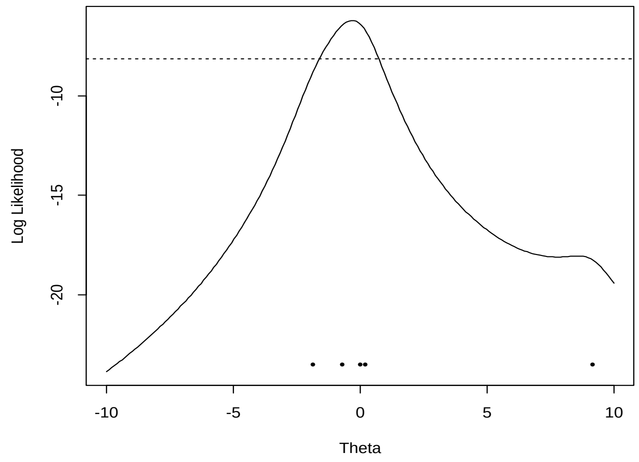
Likelihood Ratio Intervals: Cauchy, n=5



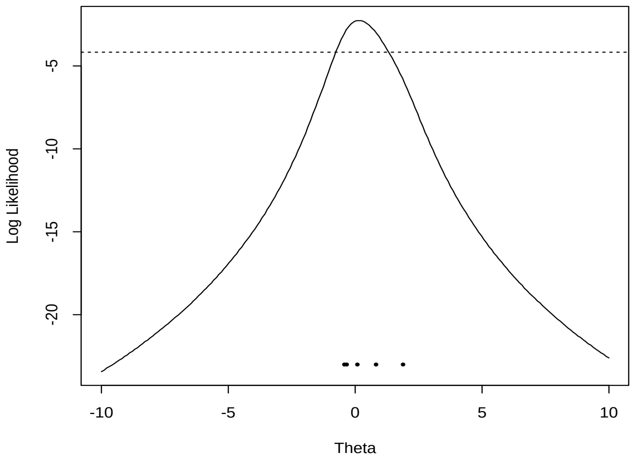
Likelihood Ratio Intervals: Cauchy, n=5



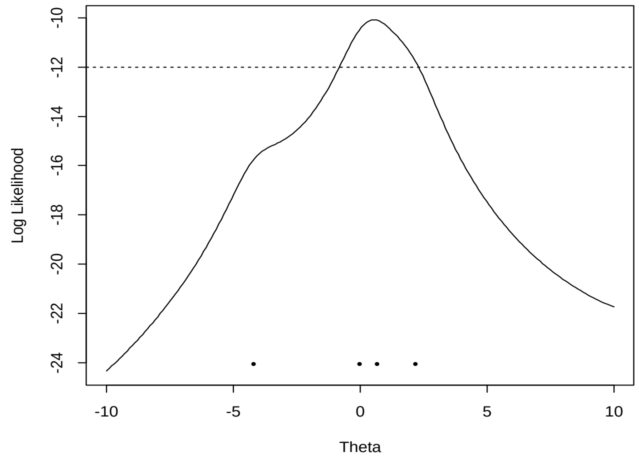
Likelihood Ratio Intervals: Cauchy, n=5



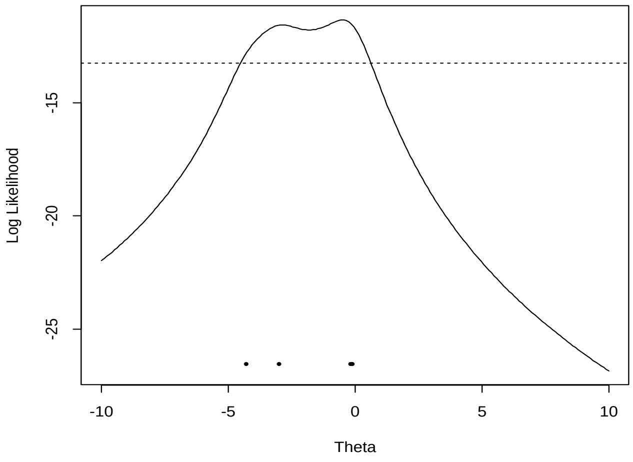
Likelihood Ratio Intervals: Cauchy, n=5

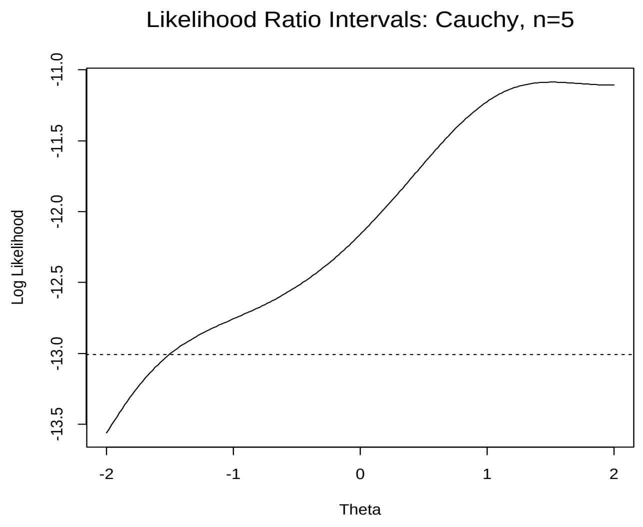


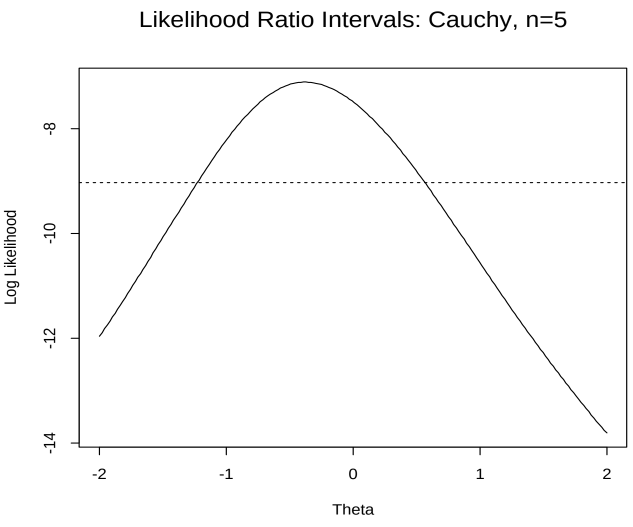
Likelihood Ratio Intervals: Cauchy, n=5

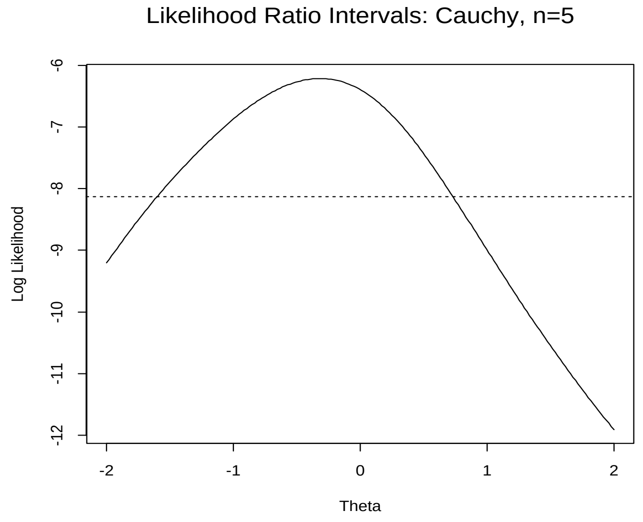


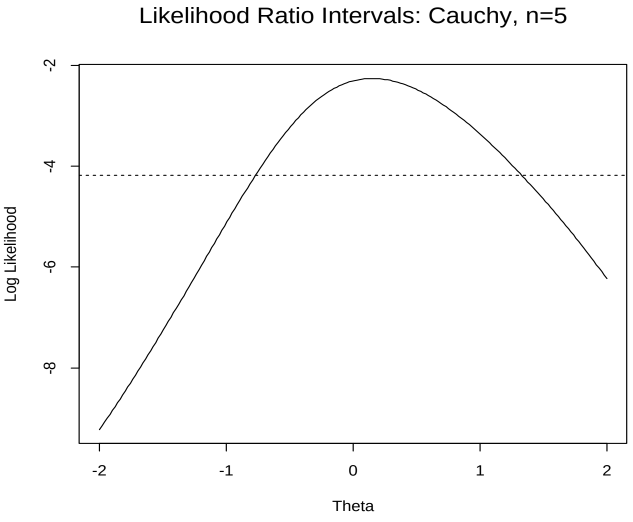
Likelihood Ratio Intervals: Cauchy, n=5

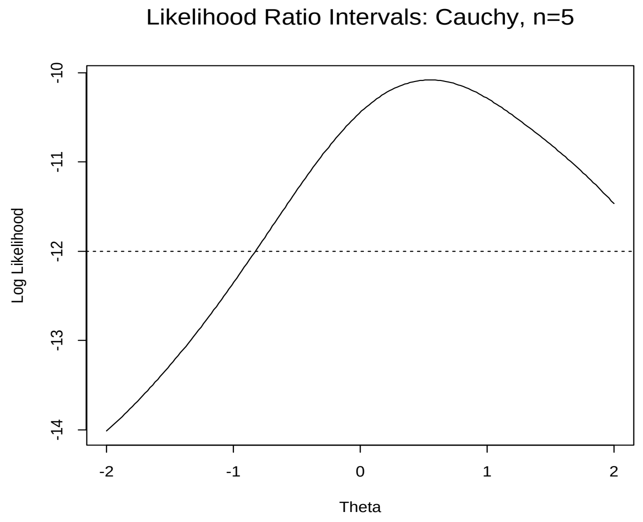


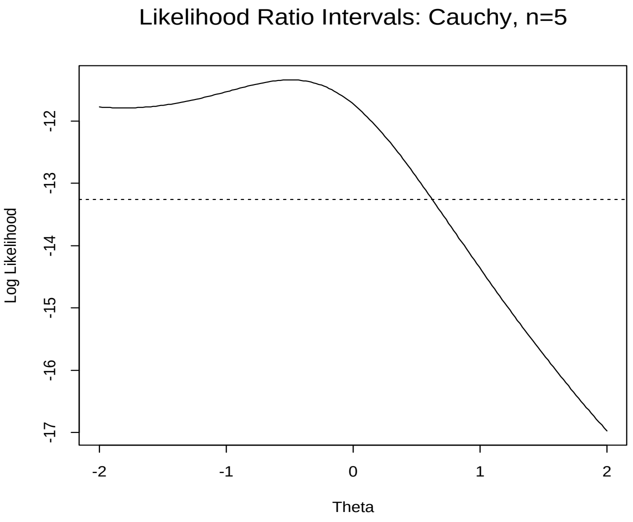

Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
-11.0
-11.5
-12.0
-12.5
-13.0
-13.5
Theta
-2
-1
0
1
2


Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
-8
-10
-12
-14
Theta
-2
-1
0
1
2

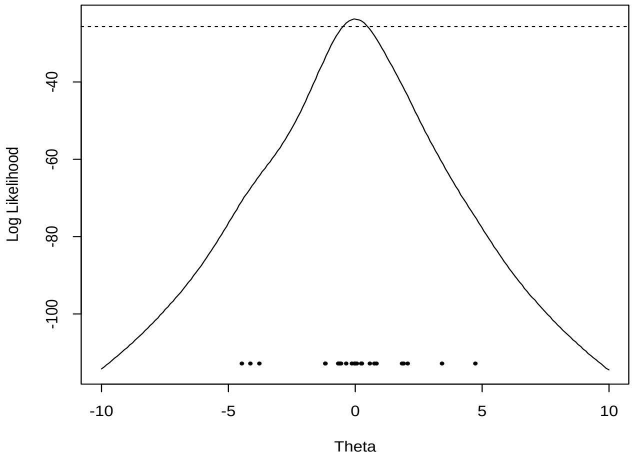

Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
-6
-7
-8
-9
-10
-11
-12
Theta
-2
-1
0
1
2


Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
-2
-4
-6
-8
Theta
-2
-1
0
1
2

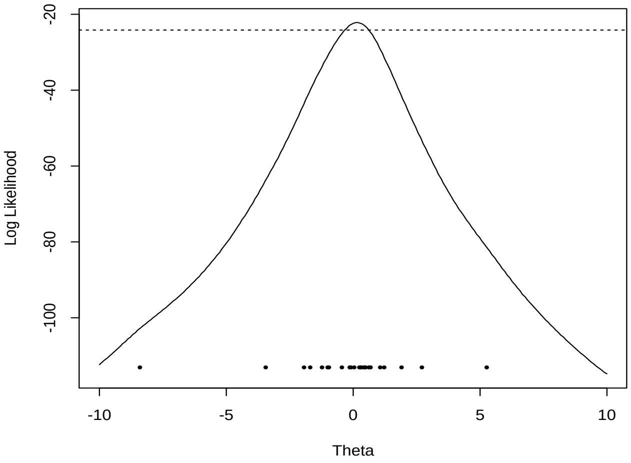

Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
-10
-11
-12
-13
-14
Theta
-2
-1
0
1
2


Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
-12
-13
-14
-15
-16
-17
Theta
-2
-1
0
1
2

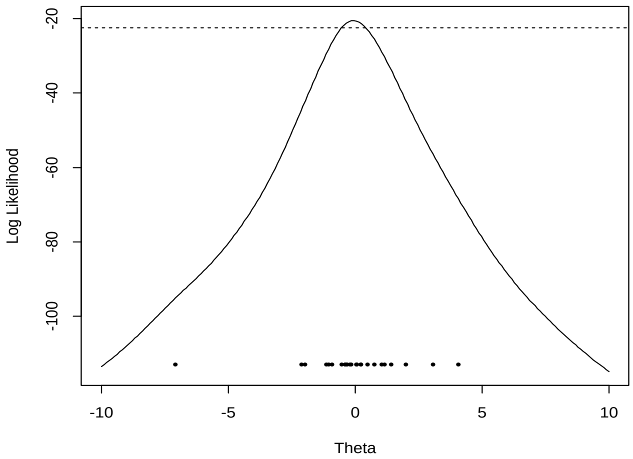
Likelihood Ratio Intervals: Cauchy, n=25



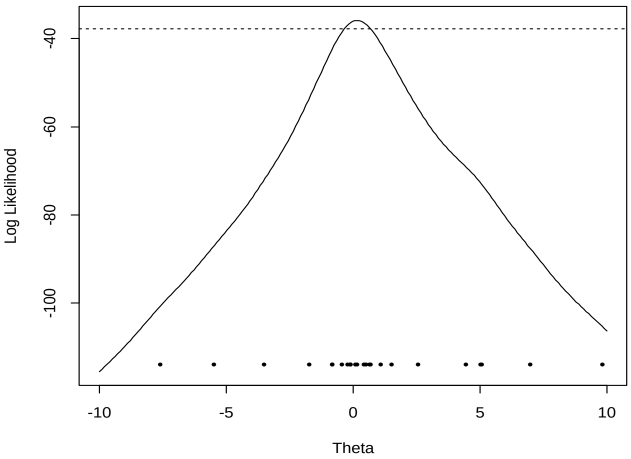
Likelihood Ratio Intervals: Cauchy, n=25



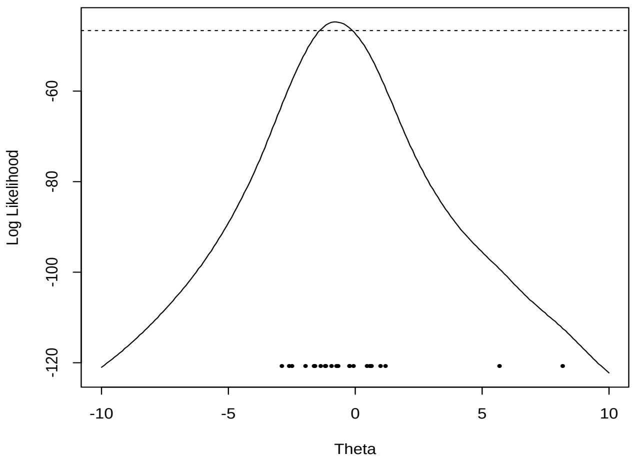
Likelihood Ratio Intervals: Cauchy, n=25



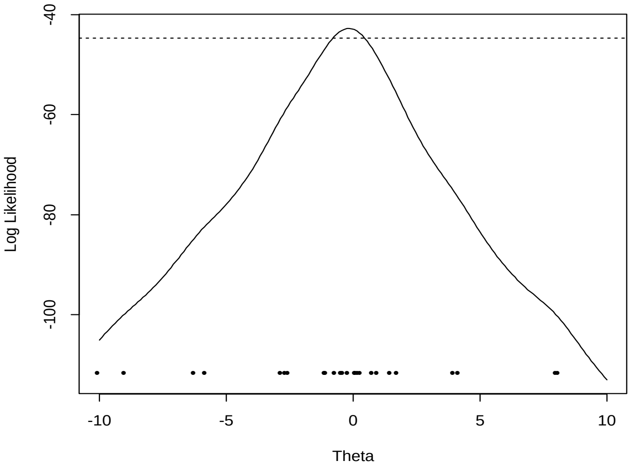
Likelihood Ratio Intervals: Cauchy, n=25

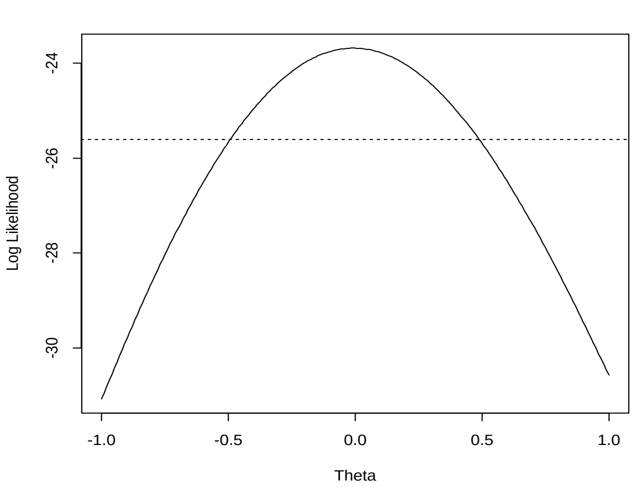


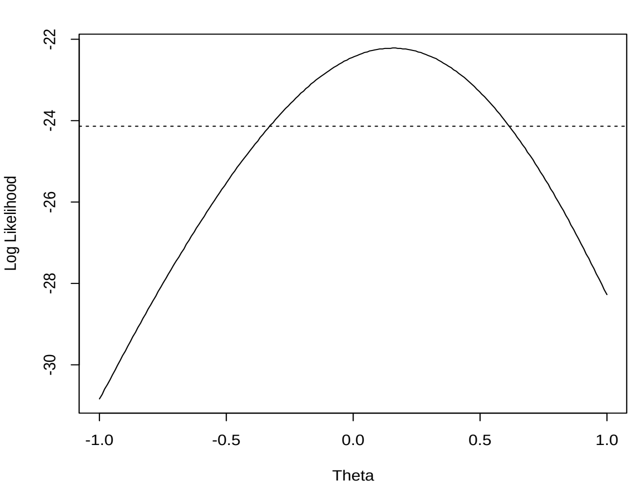
Likelihood Ratio Intervals: Cauchy, n=25

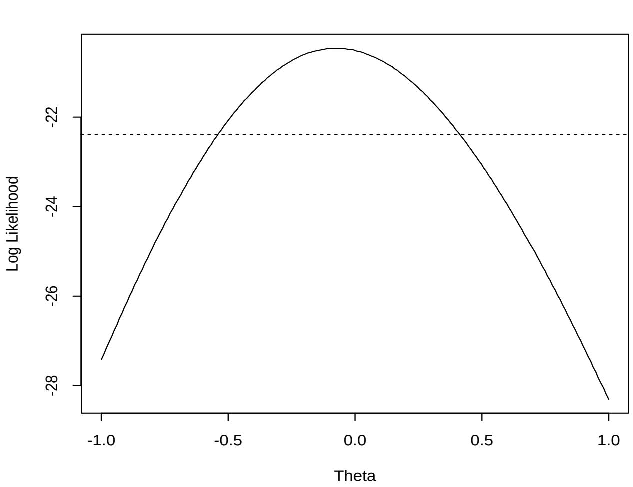


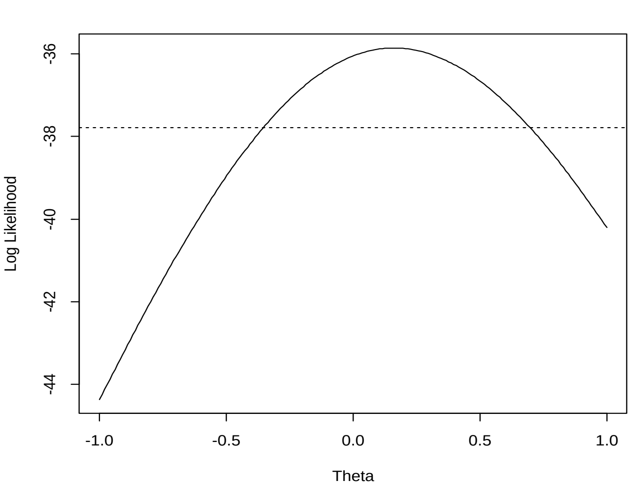
Likelihood Ratio Intervals: Cauchy, n=25

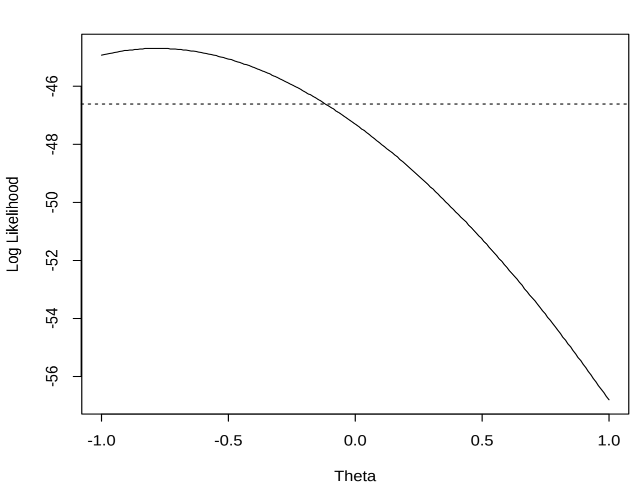


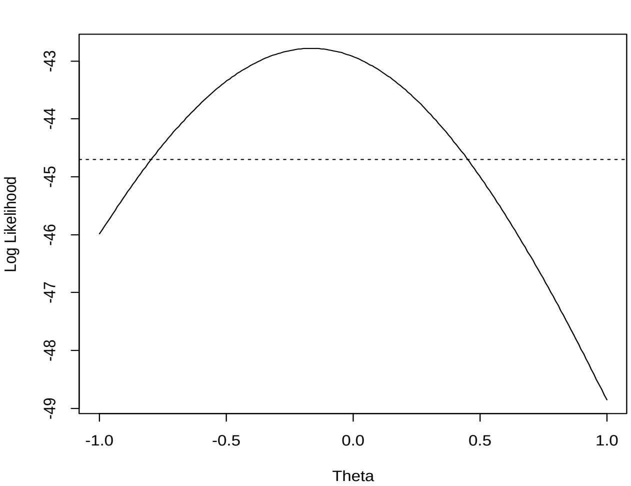

Likelihood Ratio Intervals: Cauchy, n=25


Likelihood Ratio Intervals: Cauchy, n=25


Likelihood Ratio Intervals: Cauchy, n=25


Likelihood Ratio Intervals: Cauchy, n=25


Likelihood Ratio Intervals: Cauchy, n=25


Likelihood Ratio Intervals: Cauchy, n=25

Notice the following points:

- Plots of ℓ for $n = 25$ quite smooth, rather parabolic.

- For $n = 5$ many local maxima and minima of ℓ.

Likelihood tends to 0 as $|\theta| \to \infty$ so max of ℓ occurs at a root of ℓ', derivative of ℓ wrt θ.

Def'n: **Score Function** is gradient of ℓ

$$U(\theta) = \frac{\partial \ell}{\partial \theta}$$

MLE $\widehat{\theta}$ usually root of **Likelihood Equations**

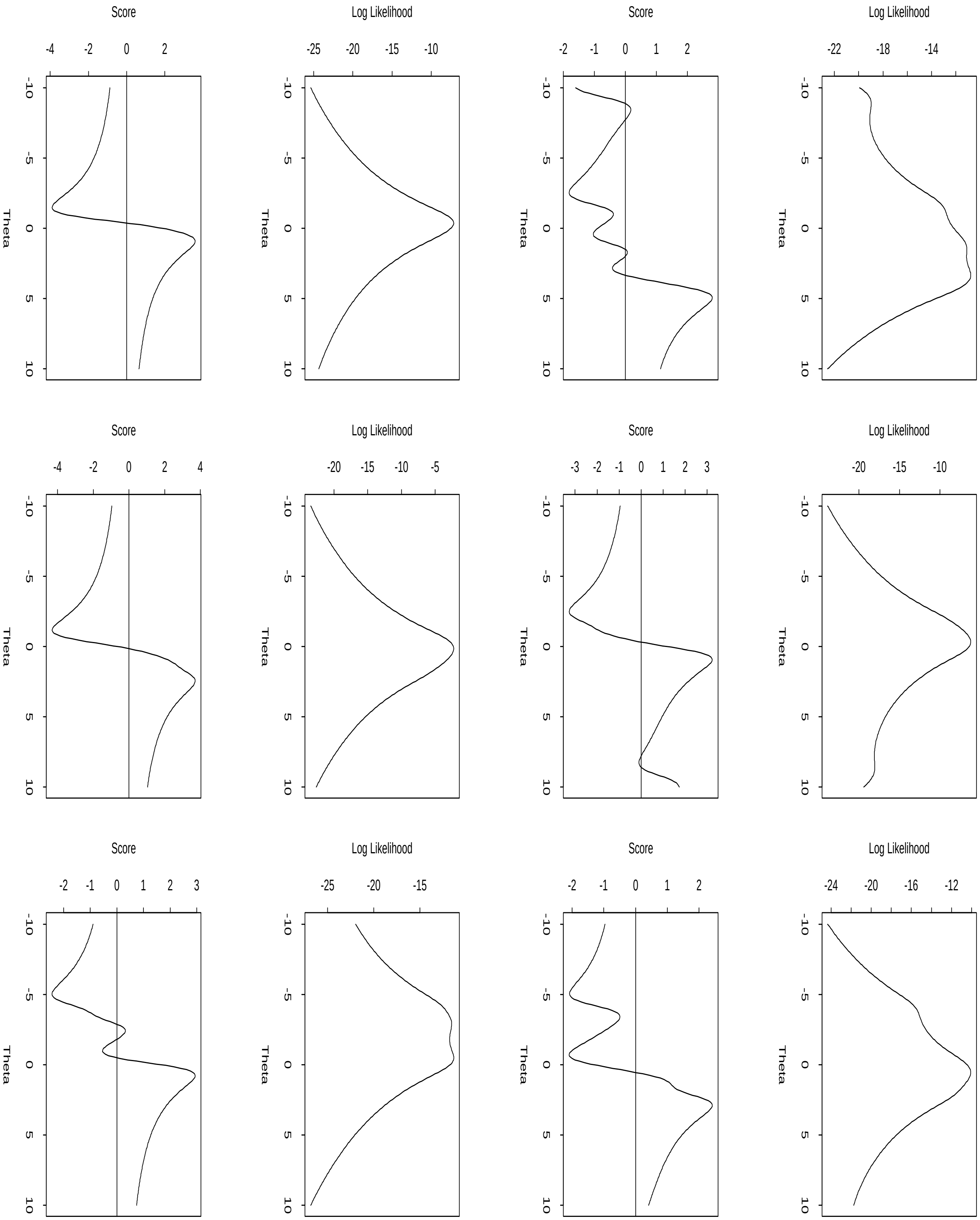
$$U(\theta) = 0$$

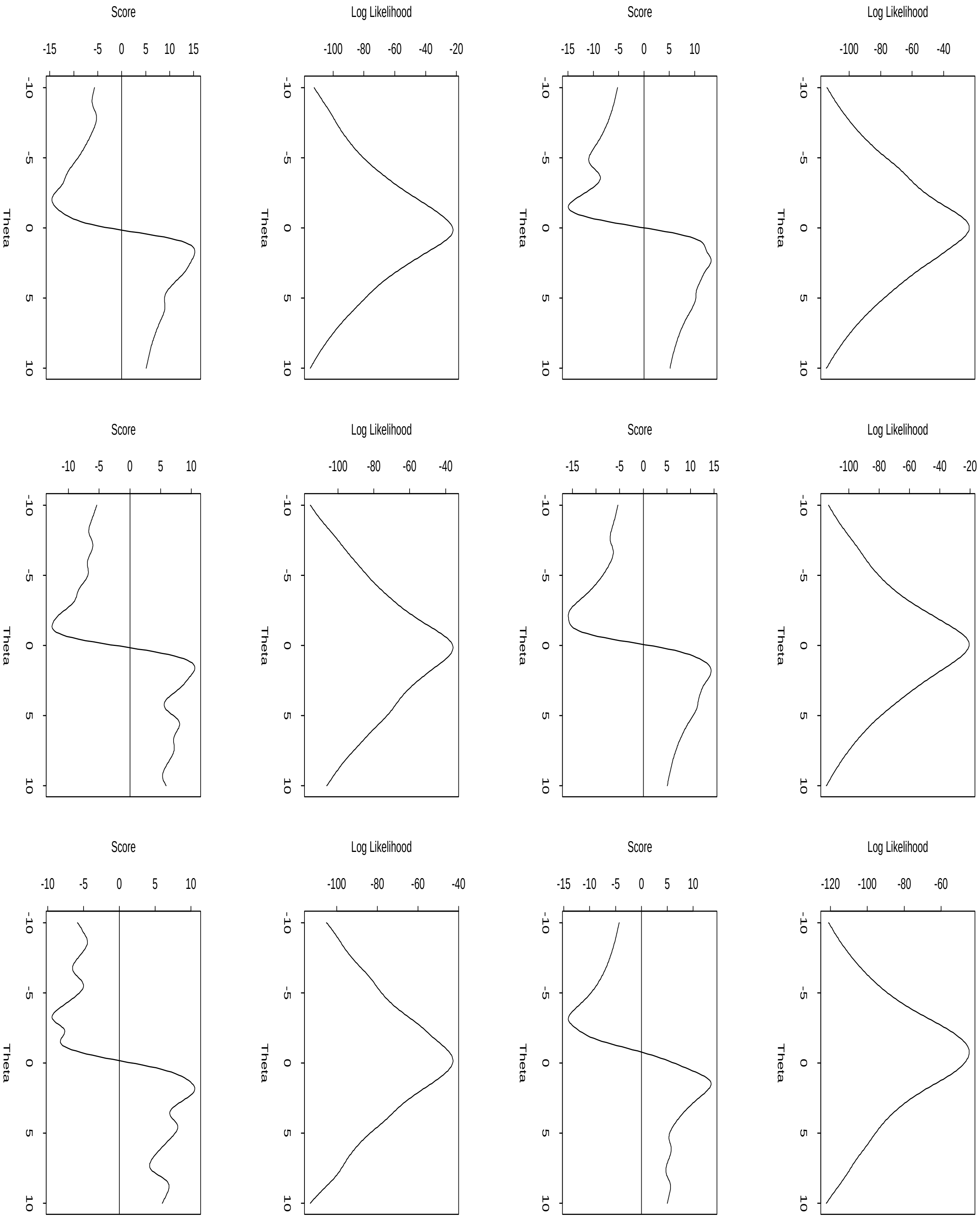
In our Cauchy example we find

$$U(\theta) = \sum \frac{2(X_i - \theta)}{1 + (X_i - \theta)^2}$$

[Examine plots of score functions.]

Notice: often multiple roots of likelihood equations.


Score
Log Likelihood
Score
Log Likelihood
Theta
-10
-5
0
5
10
-4
-2
0
2
-25
-20
-15
-10
-2
-1
0
1
2
-22
-18
-14
Score
Log Likelihood
Score
Log Likelihood
-4
-2
0
2
4
-20
-15
-10
-5
-3
-2
-1
0
1
2
3
-20
-15
-10
Score
Log Likelihood
Score
Log Likelihood
-2
-1
0
1
2
3
-25
-20
-15
-2
-1
0
1
2
-24
-20
-16
-12


Score
-15
-5
0
5
10
15
Theta
-10
-5
0
5
10
Log Likelihood
-100
-80
-60
-40
-20
Score
-15
-10
-5
0
5
10
Log Likelihood
-100
-80
-60
-40
Score
-10
-5
0
5
10
Log Likelihood
-100
-80
-60
-40
Score
-15
-5
0
5
10
15
Log Likelihood
-100
-80
-60
-40
-20
Score
-10
-5
0
5
10
Log Likelihood
-100
-80
-60
-40
Score
-15
-10
-5
0
5
10
Log Likelihood
-120
-100
-80
-60

Example : $X \sim$ Binomial(n,θ)

$$L(\theta) = \binom{n}{X}\theta^X(1-\theta)^{n-X}$$

$$\ell(\theta) = \log\binom{n}{X} + X\log(\theta) + (n-X)\log(1-\theta)$$

$$U(\theta) = \frac{X}{\theta} - \frac{n-X}{1-\theta}$$

The function L is 0 at $\theta = 0$ and at $\theta = 1$ unless $X = 0$ or $X = n$ so for $1 \le X \le n$ the MLE must be found by setting $U = 0$ and getting

$$\hat\theta = \frac{X}{n}$$

For $X = n$ the log-likelihood has derivative

$$U(\theta) = \frac{n}{\theta} > 0$$

for all θ so that the likelihood is an increasing function of θ which is maximized at $\hat\theta = 1 = X/n$. Similarly when $X = 0$ the maximum is at $\hat\theta = 0 = X/n$.

The Normal Distribution

Now we have $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$. There are two parameters $\theta = (\mu, \sigma)$. We find

$$L(\mu, \sigma) = \frac{e^{-\sum(X_i - \mu)^2/(2\sigma^2)}}{(2\pi)^{n/2}\sigma^n}$$

$$\ell(\mu, \sigma) = -\frac{n}{2}\log(2\pi) - \frac{\sum(X_i - \mu)^2}{2\sigma^2} - n\log(\sigma)$$

and that U is

$$\left[\begin{array}{c} \frac{\sum(X_i - \mu)}{\sigma^2} \\ \frac{\sum(X_i - \mu)^2}{\sigma^3} - \frac{n}{\sigma} \end{array}\right]$$

Notice that U is a function with two components because θ has two components.

Setting the likelihood equal to 0 and solving gives

$$\hat{\mu} = \bar{X}$$

and

$$\hat{\sigma} = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n}}$$

Check this is maximum by computing one more derivative. Matrix H of second derivatives of ℓ is

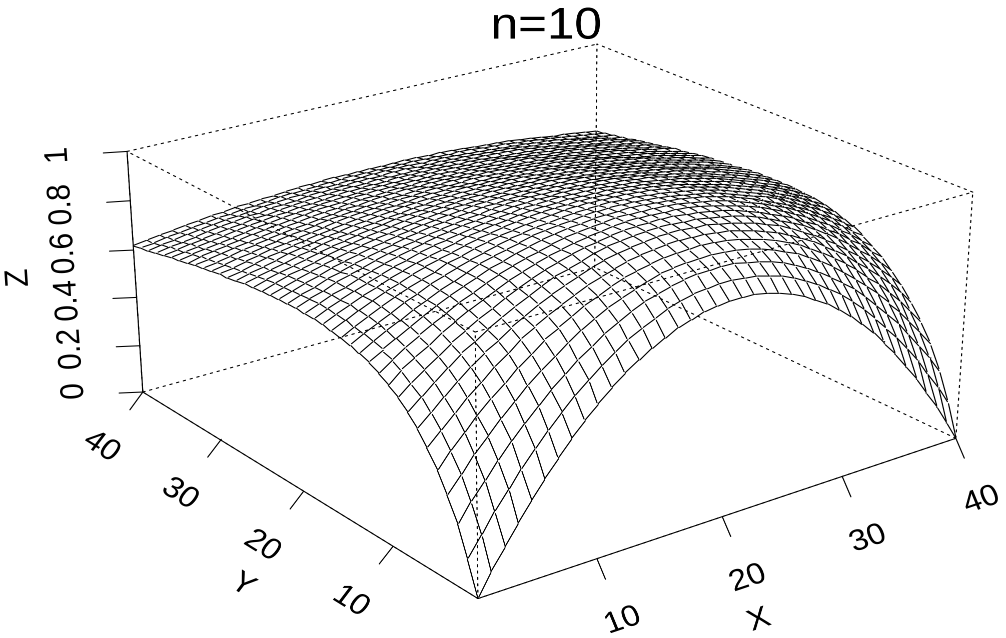
$$\left[\begin{array}{cc} \frac{-n}{\sigma^2} & \frac{-2\sum(X_i-\mu)}{\sigma^3} \\ \frac{-2\sum(X_i-\mu)}{\sigma^3} & \frac{-3\sum(X_i-\mu)^2}{\sigma^4}+\frac{n}{\sigma^2} \end{array}\right]$$

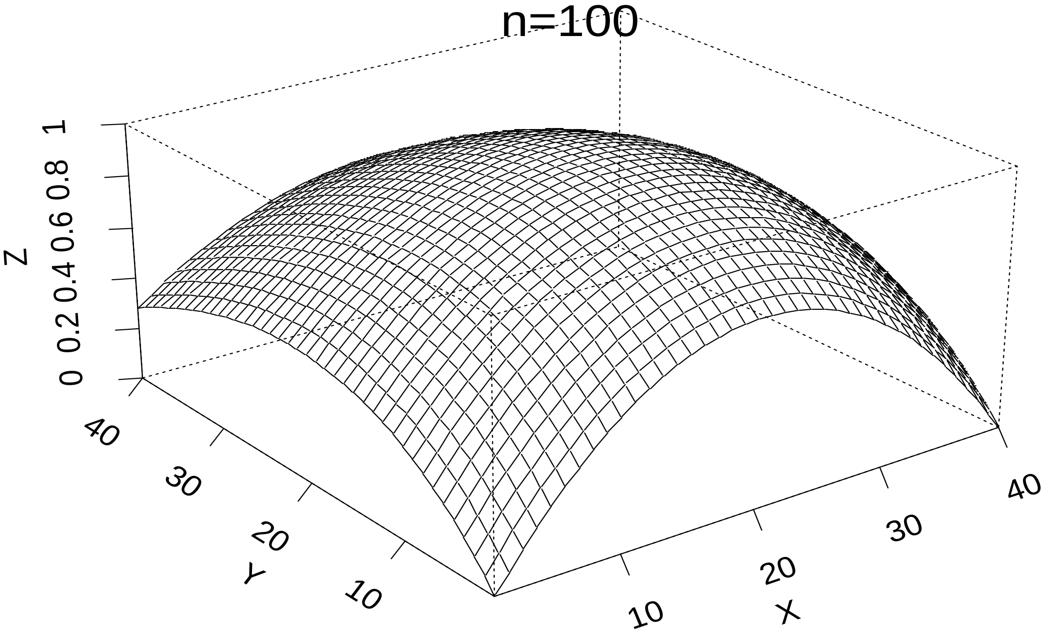
Plugging in the mle gives

$$H(\hat\theta) = \left[\begin{array}{cc} \frac{-n}{\hat\sigma^2} & 0 \\ 0 & \frac{-2n}{\hat\sigma^2} \end{array}\right]$$

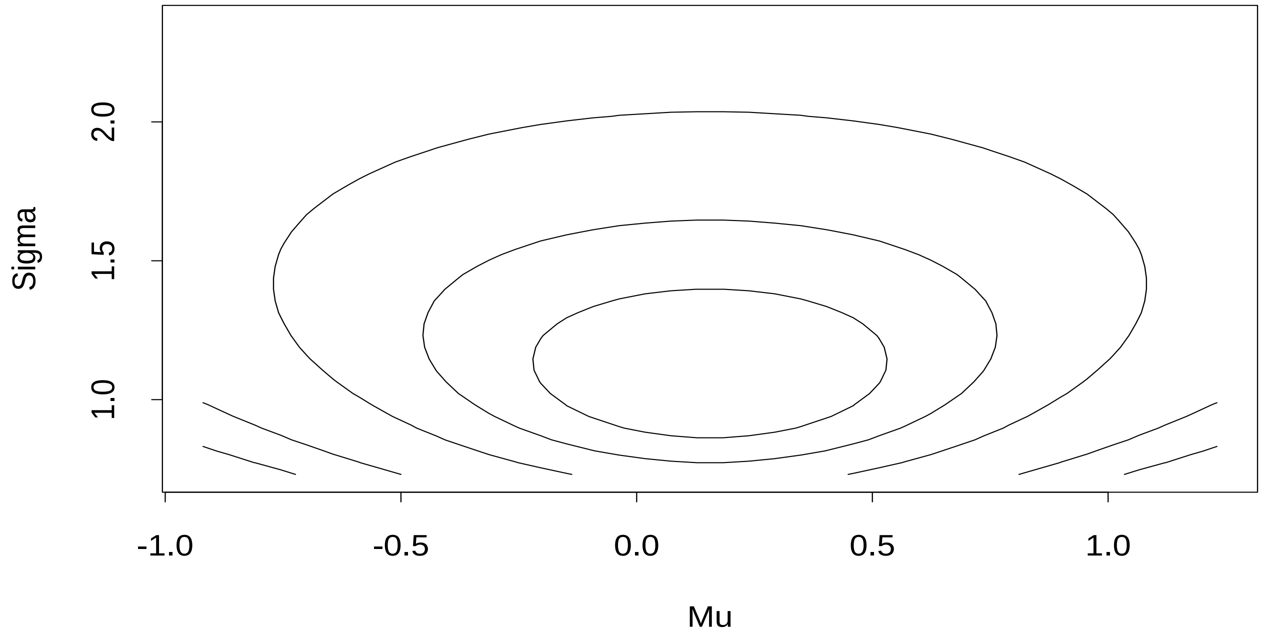
which is negative definite. Both its eigenvalues are negative. So $\hat\theta$ must be a local maximum.

[Examine contour and perspective plots of ℓ.]


n=10
Z
0
0.2
0.4
0.6
0.8
1
Y
10
20
30
40
X
10
20
30
40

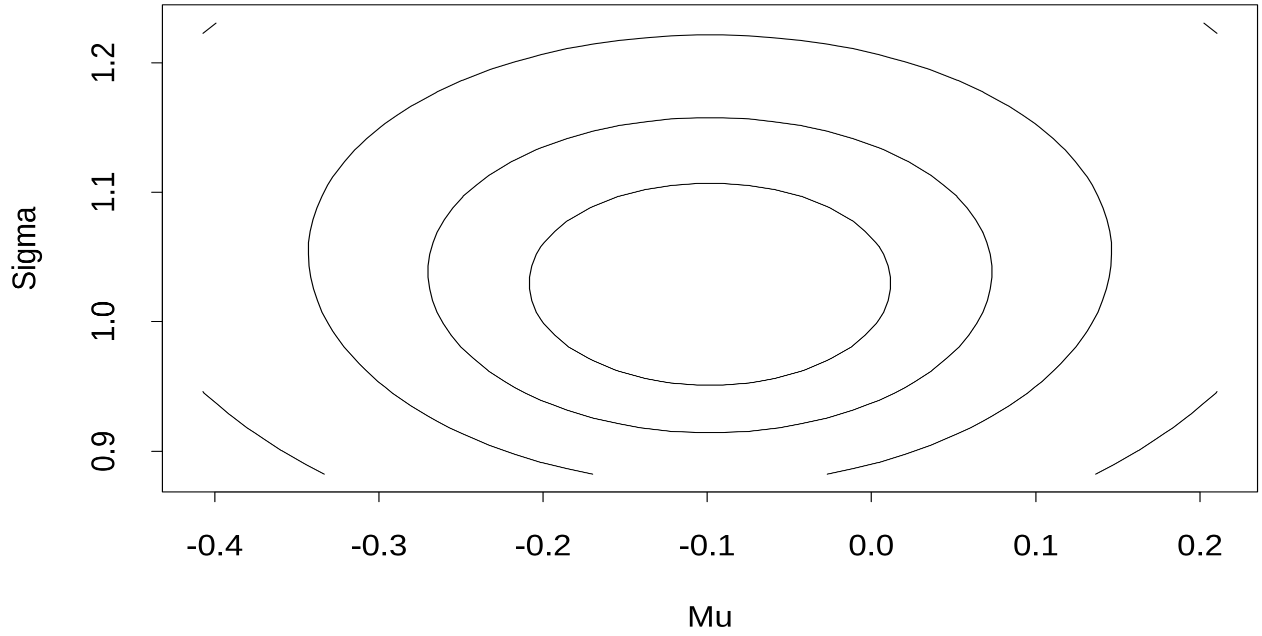

n=100
Z
0
0.2
0.4
0.6
0.8
1
Y
10
20
30
40
X
10
20
30
40

n=10



Sigma
1.0
1.5
2.0
-1.0
-0.5
0.0
0.5
1.0
Mu

n=100



Sigma
0.9
1.0
1.1
1.2
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
Mu

Notice that the contours are quite ellipsoidal for the larger sample size.

For $X_1, \ldots, X_n$ iid log likelihood is

$$\ell(\theta) = \sum \log(f(X_i, \theta)) .$$

The score function is

$$U(\theta) = \sum \frac{\partial \log f}{\partial \theta}(X_i, \theta) .$$

MLE $\hat\theta$ maximizes ℓ. If maximum occurs in interior of parameter space and the log likelihood continuously differentiable then $\hat\theta$ solves the likelihood equations

$$U(\theta) = 0 .$$

Some examples concerning existence of roots:

Solving $U(\theta) = 0$: Examples

N(μ, σ^2)**)**

Unique root of likelihood equations is a global maximum.

[Remark: Suppose we called $\tau = \sigma^2$ the parameter. Score function still has two components: first component same as before but second component is

$$\frac{\partial}{\partial \tau}\ell = \frac{\sum (X_i - \mu)^2}{2\tau^2} - \frac{n}{2\tau}$$

Setting the new likelihood equations equal to 0 still gives

$$\hat{\tau} = \hat{\sigma}^2$$

General **invariance** (or **equivariance**) principal: If $\phi = g(\theta)$ is some reparametrization of a model (a one to one relabelling of the parameter values) then $\hat{\phi} = g(\hat{\theta})$. Does not apply to other estimators.]

Cauchy: location θ

At least 1 root of likelihood equations but often several more. One root is a global maximum; others, if they exist may be local minima or maxima.

Binomial(n, θ)

If $X = 0$ or $X = n$: no root of likelihood equations; likelihood is monotone. Other values of X: unique root, a global maximum. Global maximum at $\widehat{\theta} = X/n$ even if $X = 0$ or n.

The 2 parameter exponential

The density is

$$f(x;\alpha,\beta) = \frac{1}{\beta} e^{-(x-\alpha)/\beta} 1(x > \alpha)$$

Log-likelihood is $-\infty$ for $\alpha > \min\{X_1, \ldots, X_n\}$ and otherwise is

$$\ell(\alpha,\beta) = -n\log(\beta) - \sum (X_i - \alpha)/\beta$$

Increasing function of α till α reaches

$$\hat{\alpha} = X_{(1)} = \min\{X_1, \ldots, X_n\}$$

which gives mle of α. Now plug in $\hat{\alpha}$ for α; get so-called profile likelihood for β:

$$\ell_{\text{profile}}(\beta) = -n\log(\beta) - \sum (X_i - X_{(1)})/\beta$$

Set β derivative equal to 0 to get

$$\hat{\beta} = \sum (X_i - X_{(1)})/n$$

Notice mle $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ does *not* solve likelihood equations; we had to look at the edge of the possible parameter space. α is called a *support* or *truncation* parameter. ML methods behave oddly in problems with such parameters.

Three parameter Weibull

The density in question is

$$f(x;\alpha,\beta,\gamma) = \frac{1}{\beta}\left(\frac{x-\alpha}{\beta}\right)^{\gamma-1} \\ \times \exp[-\{(x-\alpha)/\beta\}^{\gamma}]1(x>\alpha)$$

Three likelihood equations:

Set β derivative equal to 0; get

$$\hat{\beta}(\alpha,\gamma) = \left[\sum (X_i-\alpha)^{\gamma}/n\right]^{1/\gamma}$$

where $\hat{\beta}(\alpha,\gamma)$ indicates mle of β could be found by finding the mles of the other two parameters and then plugging in to the formula above.

It is not possible to find explicitly the remaining two parameters; numerical methods are needed.

However putting $\gamma < 1$ and letting $\alpha \to X_{(1)}$ will make the log likelihood go to ∞.

MLE is not uniquely defined: any $\gamma < 1$ and any β will do.

If the true value of γ is more than 1 then the probability that there is a root of the likelihood equations is high; in this case there must be two more roots: a local maximum and a saddle point! For a true value of $\gamma > 1$ the theory we detail below applies to the local maximum and not to the global maximum of the likelihood equations.

Large Sample Theory

Study approximate behaviour of $\widehat{\theta}$ by studying the function U.

Notice U is sum of independent random variables.

Theorem: If $Y_1, Y_2, \ldots$ are iid with mean μ then

$$\frac{\sum Y_i}{n} \to \mu$$

Called law of large numbers. Strong law

$$P(\lim \frac{\sum Y_i}{n} = \mu) = 1$$

and the weak law that

$$\lim P(|\frac{\sum Y_i}{n} - \mu| > \epsilon) = 0$$

For iid Y_i the stronger conclusion holds; for our heuristics ignore differences between these notions.

Now suppose θ_0 is true value of θ. Then

$$U(\theta)/n \to \mu(\theta)$$

where

$$\begin{aligned}\mu(\theta) =& E_{\theta_0}\left[\frac{\partial \log f}{\partial\theta}(X_i,\theta)\right] \\ =& \int \frac{\partial \log f}{\partial\theta}(x,\theta)f(x,\theta_0)dx\end{aligned}$$

Example: $N(\mu,1)$ data:

$$U(\mu)/n = \sum (X_i-\mu)/n = \bar{X} - \mu$$

If the true mean is μ_0 then $\bar{X} \to \mu_0$ and

$$U(\mu)/n \to \mu_0 - \mu$$

Consider $\mu < \mu_0$: derivative of $\ell(\mu)$ is likely to be positive so that ℓ increases as μ increases.

For $\mu > \mu_0$: derivative is probably negative and so ℓ tends to be decreasing for $\mu > 0$.

Hence: ℓ is likely to be maximized close to μ_0.

Repeat ideas for more general case. Study rv

$$\log[f(X_i,\theta)/f(X_i,\theta_0)].$$

You know the inequality

$$E(X)^2 \le E(X^2)$$

(difference is Var(X) ≥ 0.)

Generalization: Jensen's inequality: for g a convex function ($g'' \ge 0$ roughly) then

$$g(E(X)) \le E(g(X))$$

Inequality above has $g(x) = x^2$. Use $g(x) = -\log(x)$: convex because $g''(x) = x^{-2} > 0$. We get

$$\begin{aligned}
-\log(E_{\theta_0}[f(X_i,\theta)/f(X_i,\theta_0)]& \\
&\le E_{\theta_0}[-\log\{f(X_i,\theta)/f(X_i,\theta_0)\}]
\end{aligned}$$

But

$$\begin{aligned}
E_{\theta_0}\left[\frac{f(X_i,\theta)}{f(X_i,\theta_0)}\right] &= \int \frac{f(x,\theta)}{f(x,\theta_0)} f(x,\theta_0)dx \\
&= \int f(x,\theta)dx \\
&= 1
\end{aligned}$$

We can reassemble the inequality and this calculation to get

$$E_{\theta_0}[\log\{f(X_i,\theta)/f(X_i,\theta_0)\}] \le 0$$

Fact: inequality is strict unless the θ and θ_0 densities are actually the same.

Let $\mu(\theta) < 0$ be this expected value.

Then for each θ we find

$$\frac{\ell(\theta) - \ell(\theta_0)}{n} = \frac{\sum \log[f(X_i, \theta)/f(X_i, \theta_0)]}{n} \to \mu(\theta)$$

This proves likelihood probably higher at θ_0 than at any other single θ.

Idea can often be stretched to prove that the mle is **consistent**; need **uniform** convergence in θ.

Definition A sequence $\widehat{\theta}_n$ of estimators of θ is consistent if $\widehat{\theta}_n$ converges weakly (or strongly) to θ.

Proto theorem: In regular problems the mle $\widehat{\theta}$ is consistent.

More precise statements of possible conclusions. Use notation

$$N(\epsilon) = \{\theta : |\theta - \theta_0| \le \epsilon\} .$$

Suppose:

$\widehat{\theta}_n$ is global maximizer of ℓ.

$\widehat{\theta}_{n,\delta}$ maximizes ℓ over $N(\delta) = \{|\theta - \theta_0| \le \delta\}$.

$$A_\epsilon = \{|\widehat{\theta}_n - \theta_0| \le \epsilon\}$$

$$B_{\delta,\epsilon} = \{|\widehat{\theta}_{n,\delta} - \theta_0| \le \epsilon\}$$

$$C_L = \{\exists! \theta \in N(L/n^{1/2}) : U(\theta) = 0, U'(\theta) < 0\}$$

Theorem:

1. Under conditions **I** $P(A_\epsilon) \to 1$ for each $\epsilon > 0$.

2. Under conditions **II** there is a $\delta > 0$ such that for all $\epsilon > 0$ we have $P(B_{\delta,\epsilon}) \to 1$.

3. Under conditions **III** for all $\delta > 0$ there is an L so large and an n_0 so large that for all $n \ge n_0$, $P(C_L) > 1 - \delta$.

4. Under conditions **III** there is a sequence L_n tending to ∞ so slowly that $P(C_{L_n}) \to 1$.

Point: conditions get weaker as conclusions get weaker. Many possible conditions in literature. See book by Zacks for some precise conditions.

Asymptotic Normality

Study shape of log likelihood near the true value of θ.

Assume $\hat\theta$ is a root of the likelihood equations close to θ_0.

Taylor expansion (1 dimensional parameter θ):

$$\begin{aligned}U(\hat\theta) =& 0 \\ =& U(\theta_0)+U'(\theta_0)(\hat\theta-\theta_0) \\ &+U''(\tilde\theta)(\hat\theta-\theta_0)^2/2\end{aligned}$$

for some $\tilde\theta$ between θ_0 and $\hat\theta$.

WARNING: This form of the remainder in Taylor's theorem is not valid for multivariate θ.

Derivatives of U are sums of n terms.

So each derivative should be proportional to n in size.

Second derivative is multiplied by the square of the small number $\hat\theta - \theta_0$ so should be negligible compared to the first derivative term.

Ignoring second derivative term get

$$-U'(\theta_0)(\hat\theta - \theta_0) \approx U(\theta_0)$$

Now look at terms U and U'.

Normal case:

$$U(\theta_0) = \sum (X_i - \mu_0)$$

has a normal distribution with mean 0 and variance n (SD $\sqrt{n}$).

Derivative is

$$U'(\mu) = -n\,.$$

Next derivative U'' is 0.

Notice: both U and U' are sums of iid random variables.

Let

$$U_i = \frac{\partial \log f}{\partial \theta}(X_i, \theta_0)$$

and

$$V_i = -\frac{\partial^2 \log f}{\partial \theta^2}(X_i, \theta)$$

In general, $U(\theta_0) = \sum U_i$ has mean 0 and approximately a normal distribution.

Here is how we check that:

$$\begin{aligned}
E_{\theta_0}(U(\theta_0)) &= nE_{\theta_0}(U_1) \\
&= n\int \frac{\partial \log(f(x,\theta_0))}{\partial\theta} f(x,\theta_0)dx \\
&= n\int \frac{\partial f(x,\theta_0)/\partial\theta}{f(x,\theta_0)} f(x,\theta_0)dx \\
&= n\int \frac{\partial f}{\partial\theta}(x,\theta_0)dx \\
&= n\frac{\partial}{\partial\theta}\int f(x,\theta)dx\Big|_{\theta=\theta_0} \\
&= n\frac{\partial}{\partial\theta}1 \\
&= 0
\end{aligned}$$

Notice: interchanged order of differentiation and integration at one point.

This step is usually justified by applying the dominated convergence theorem to the definition of the derivative.

Differentiate identity just proved:

$$\int \frac{\partial \log f}{\partial \theta}(x,\theta) f(x,\theta) dx = 0$$

Take derivative of both sides wrt θ; pull derivative under integral sign:

$$\int \frac{\partial}{\partial \theta}\left[\frac{\partial \log f}{\partial \theta}(x,\theta) f(x,\theta)\right] dx = 0$$

Do the derivative and get

$$\begin{aligned} -\int \frac{\partial^2 \log(f)}{\partial \theta^2} f(x,\theta) dx \\ = \int \frac{\partial \log f}{\partial \theta}(x,\theta) \frac{\partial f}{\partial \theta}(x,\theta) dx \\ = \int \left[\frac{\partial \log f}{\partial \theta}(x,\theta)\right]^2 f(x,\theta) dx \end{aligned}$$

Definition: The **Fisher Information** is

$$I(\theta) = -E_\theta(U'(\theta)) = nE_{\theta_0}(V_1)$$

We refer to $\mathcal{I}(\theta_0) = E_{\theta_0}(V_1)$ as the information in 1 observation.

The idea is that I is a measure of how curved the log likelihood tends to be at the true value of θ. Big curvature means precise estimates. Our identity above is

$$I(\theta) = Var_\theta(U(\theta)) = n\mathcal{I}(\theta)$$

Now we return to our Taylor expansion approximation

$$-U'(\theta_0)(\widehat{\theta} - \theta_0) \approx U(\theta_0)$$

and study the two appearances of U.

We have shown that $U = \sum U_i$ is a sum of iid mean 0 random variables. The central limit theorem thus proves that

$$n^{-1/2}U(\theta_0) \Rightarrow N(0, \sigma^2)$$

where $\sigma^2 = \text{Var}(U_i) = E(V_i) = \mathcal{I}(\theta)$.

Next observe that

$$-U'(\theta) = \sum V_i$$

where again

$$V_i = -\frac{\partial U_i}{\partial \theta}$$

The law of large numbers can be applied to show

$$-U'(\theta_0)/n \to E_{\theta_0}[V_1] = \mathcal{I}(\theta_0)$$

Now manipulate our Taylor expansion as follows

$$n^{1/2}(\hat\theta - \theta_0) \approx \left[\frac{\sum V_i}{n}\right]^{-1} \frac{\sum U_i}{\sqrt{n}}$$

Apply Slutsky's Theorem to conclude that the right hand side of this converges in distribution to $N(0, \sigma^2/\mathcal{I}(\theta)^2)$ which simplifies, because of the identities, to $N\{0, 1/\mathcal{I}(\theta)\}$.

Summary

In regular families: assuming $\hat{\theta} = \hat{\theta}_n$ is a consistent root of $U(\theta) = 0$.

- $n^{-1/2}U(\theta_0) \Rightarrow MVN(0, \mathcal{I})$ where

  $$\mathcal{I}_{ij} = \mathrm{E}_{\theta_0}\left\{V_{1,ij}(\theta_0)\right\}$$

  and

  $$V_{k,ij}(\theta) = -\frac{\partial^2 \log f(X_k,\theta)}{\partial\theta_i\partial\theta_j}$$

- If $\mathbf{V}_k(\theta)$ is the matrix $[V_{k,ij}]$ then

  $$\frac{\sum_{k=1}^n \mathbf{V}_k(\theta_0)}{n} \to \mathcal{I}$$

- If $\mathbf{V}(\theta) = \sum_k \mathbf{V}_k(\theta)$ then

  $$\{\mathbf{V}(\theta_0)/n\}n^{1/2}(\hat{\theta} - \theta_0) - n^{-1/2}U(\theta_0) \to 0$$

  in probability as $n \to \infty$.

- Also
$$\{\mathbf{V}(\hat\theta)/n\} n^{1/2}(\hat\theta - \theta_0) - n^{-1/2}U(\theta_0) \to 0$$
in probability as $n \to \infty$.

- $n^{1/2}(\hat\theta - \theta_0) - \{\mathcal{I}(\theta_0)\}^{-1}U(\theta_0) \to 0$ in probability as $n \to \infty$.

- $n^{1/2}(\hat\theta - \theta_0) \Rightarrow MVN(0, \mathcal{I}^{-1})$.

- In general (not just iid cases)
$$\sqrt{I(\theta_0)}(\hat\theta - \theta_0) \Rightarrow N(0,1)$$
$$\sqrt{I(\hat\theta)}(\hat\theta - \theta_0) \Rightarrow N(0,1)$$
$$\sqrt{V(\theta_0)}(\hat\theta - \theta_0) \Rightarrow N(0,1)$$
$$\sqrt{V(\hat\theta)}(\hat\theta - \theta_0) \Rightarrow N(0,1)$$
where $V = -\ell''$ is the so-called *observed information*, the negative second derivative of the log-likelihood.

Note: If the square roots are replaced by matrix square roots we can let θ be vector valued and get $MVN(0, I)$ as the limit law.

Why all these different forms? Use limit laws to test hypotheses and compute confidence intervals. Test $H_o : \theta = \theta_0$ using one of the 4 quantities as test statistic. Find confidence intervals using quantities as *pivots*. E.g.: second and fourth limits lead to confidence intervals

$$\hat\theta \pm z_{\alpha/2}/\sqrt{I(\hat\theta)}$$

and

$$\hat\theta \pm z_{\alpha/2}/\sqrt{V(\hat\theta)}$$

respectively. The other two are more complicated. For iid $N(0, \sigma^2)$ data we have

$$V(\sigma) = \frac{3\sum X_i^2}{\sigma^4} - \frac{n}{\sigma^2}$$

and

$$I(\sigma) = \frac{2n}{\sigma^2}$$

The first line above then justifies confidence intervals for σ computed by finding all those σ for which

$$\left|\frac{\sqrt{2n}(\hat\sigma - \sigma)}{\sigma}\right| \le z_{\alpha/2}$$

Similar interval can be derived from 3rd expression, though this is much more complicated.

Usual summary: mle is consistent and asymptotically normal with an asymptotic variance which is the inverse of the Fisher information.

Problems with maximum likelihood

1. Many parameters lead to poor approximations. MLEs can be far from right answer. See homework for Neyman Scott example where MLE is not consistent.

2. Multiple roots of the likelihood equations: you must choose the right root. Start with different, consistent, estimator; apply iterative scheme like Newton Raphson to likelihood equations to find MLE. Not many steps of NR generally required if starting point is a reasonable estimate.

Finding (good) preliminary Point Estimates

Method of Moments

Basic strategy: set sample moments equal to population moments and solve for the parameters.

Definition: The r^{th} sample moment (about the origin) is

$$\frac{1}{n}\sum_{i=1}^{n} X_i^r$$

The r^{th} population moment is

$$\mathsf{E}(X^r)$$

(**Central** moments are

$$\frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^r$$

and

$$\mathsf{E}\left[(X-\mu)^r\right] .$$

If we have p parameters we can estimate the parameters $\theta_1, \ldots, \theta_p$ by solving the system of p equations:

$$\mu_1 = \bar{X}$$

$$\mu_2' = \overline{X^2}$$

and so on to

$$\mu_p' = \overline{X^p}$$

You need to remember that the population moments μ_k' will be formulas involving the parameters.

Gamma Example

The Gamma(α, β) density is

$$f(x;\alpha,\beta) = \frac{1}{\beta\Gamma(\alpha)}\left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left[-\frac{x}{\beta}\right] 1(x > 0)$$

and has

$$\mu_1 = \alpha\beta$$

and

$$\mu_2' = \alpha(\alpha+1)\beta^2.$$

This gives the equations

$$\begin{aligned} \alpha\beta &= \overline{X} \\ \alpha(\alpha+1)\beta^2 &= \overline{X^2} \end{aligned}$$

or

$$\begin{aligned} \alpha\beta &= \overline{X} \\ \alpha\beta^2 &= \overline{X^2} - \overline{X}^2. \end{aligned}$$

Divide the second equation by the first to find the method of moments estimate of β is

$$\tilde{\beta} = (\overline{X^2} - \overline{X}^2)/\overline{X}\,.$$

Then from the first equation get

$$\tilde{\alpha} = \overline{X}/\tilde{\beta} = (\overline{X})^2/(\overline{X^2} - \overline{X}^2)\,.$$

The method of moments equations are much easier to solve than the likelihood equations which involve the function

$$\psi(\alpha) = \frac{d}{d\alpha}\log(\Gamma(\alpha))$$

called the digamma function.

Score function has components

$$U_\beta = \frac{\sum X_i}{\beta^2} - n\alpha/\beta$$

and

$$U_\alpha = -n\psi(\alpha) + \sum \log(X_i) - n\log(\beta)\,.$$

You can solve for β in terms of α to leave you trying to find a root of the equation

$$-n\psi(\alpha) + \sum \log(X_i) - n\log(\sum X_i/(n\alpha)) = 0$$

To use Newton Raphson on this you begin with the preliminary estimate $\hat\alpha_1 = \tilde\alpha$ and then compute iteratively

$$\hat\alpha_{k+1} = \frac{\overline{\log(X)} - \psi(\hat\alpha_k) - \log(\overline{X})/\hat\alpha_k}{1/\alpha - \psi'(\hat\alpha_k)}$$

until the sequence converges. Computation of ψ', the trigamma function, requires special software. Web sites like *netlib* and *statlib* are good sources for this sort of thing.

Estimating Equations

Same large sample ideas arise whenever estimates derived by solving some equation.

Example: large sample theory for **Generalized Linear Models**.

Suppose Y_i is number of cancer cases in some group of people characterized by values x_i of some covariates.

Think of x_i as containing variables like age, or a dummy for sex or average income or ....

Possible parametric regression model: Y_i has a Poisson distribution with mean μ_i where the mean μ_i depends somehow on x_i.

Typically assume $g(\mu_i) = \beta_0 + x_i\beta$; g is **link** function.

Often $g(\mu) = \log(\mu)$ and $x_i\beta$ is a matrix product: x_i row vector, β column vector.

"Linear regression model with Poisson errors".

Special case $\log(\mu_i) = \beta x_i$ where x_i is a scalar.

The log likelihood is simply

$$\ell(\beta) = \sum (Y_i \log(\mu_i) - \mu_i)$$

ignoring irrelevant factorials. The score function is, since $\log(\mu_i) = \beta x_i$,

$$U(\beta) = \sum (Y_i x_i - x_i \mu_i) = \sum x_i (Y_i - \mu_i)$$

(Notice again that the score has mean 0 when you plug in the true parameter value.)

The key observation, however, is that it is not necessary to believe that Y_i has a Poisson distribution to make solving the equation $U = 0$ sensible. Suppose only that $\log(E(Y_i)) = x_i\beta$. Then we have assumed that

$$E_\beta(U(\beta)) = 0$$

This was the key condition in proving that there was a root of the likelihood equations which was consistent and here it is what is needed, roughly, to prove that the equation $U(\beta) = 0$ has a consistent root $\hat{\beta}$.

Ignoring higher order terms in a Taylor expansion will give

$$V(\beta)(\hat{\beta}-\beta) \approx U(\beta)$$

where $V = -U'$. In the mle case we had identities relating the expectation of V to the variance of U. In general here we have

$$\text{Var}(U) = \sum x_i^2 \text{Var}(Y_i)\,.$$

If Y_i is Poisson with mean μ_i (and so $\text{Var}(Y_i) = \mu_i$) this is

$$\text{Var}(U) = \sum x_i^2 \mu_i\,.$$

Moreover we have

$$V_i = x_i^2 \mu_i$$

and so

$$V(\beta) = \sum x_i^2 \mu_i\,.$$

The central limit theorem (the Lyapunov kind) will show that $U(\beta)$ has an approximate normal distribution with variance $\sigma_U^2 = \sum x_i^2 \text{Var}(Y_i)$ and so

$$\hat{\beta} - \beta \approx N(0, \sigma_U^2 / (\sum x_i^2 \mu_i)^2)$$

If $\text{Var}(Y_i) = \mu_i$, as it is for the Poisson case, the asymptotic variance simplifies to $1/\sum x_i^2 \mu_i$.

Other estimating equations are possible, popular. If w_i is any set of deterministic weights (possibly depending on μ_i) then could define

$$U(\beta) = \sum w_i (Y_i - \mu_i)$$

and still conclude that $U = 0$ probably has a consistent root which has an asymptotic normal distribution.

Idea widely used:

Example: Generalized Estimating Equations, Zeger and Liang.

Abbreviation: GEE.

Called by econometricians Generalized Method of Moments.

An estimating equation is unbiased if

$$E_\theta(U(\theta)) = 0$$

Theorem: Suppose $\hat\theta$ is a consistent root of the unbiased estimating equation

$$U(\theta) = 0.$$

Let $V = -U'$. Suppose there is a sequence of constants $B(\theta)$ such that

$$V(\theta)/B(\theta) \to 1$$

and let

$$A(\theta) = Var_\theta(U(\theta))$$

and

$$C(\theta) = B(\theta)A^{-1}(\theta)B(\theta).$$

Then

$$\sqrt{C(\theta_0)}(\hat\theta - \theta_0) \Rightarrow N(0,1)$$
$$\sqrt{C(\hat\theta)}(\hat\theta - \theta_0) \Rightarrow N(0,1)$$

Other ways to estimate A, B and C lead to same conclusions. There are multivariate extensions using matrix square roots.

Optimality theory for point estimates

Why bother doing the Newton Raphson steps?

Why not just use the method of moments estimates?

Answer: method of moments estimates not usually as close to right answer as MLEs.

Rough principle: A good estimate $\hat{\theta}$ of θ is usually close to θ_0 if θ_0 is the true value of θ. Closer estimates, more often, are better estimates.

This principle must be quantified if we are to "prove" that the mle is a good estimate. In the Neyman Pearson spirit we measure average closeness.

Definition: The **Mean Squared Error** (MSE) of an estimator $\hat{\theta}$ is the **function**

$$MSE(\theta) = E_\theta[(\hat{\theta} - \theta)^2]$$

Standard identity:

$$MSE = \mathsf{Var}_\theta(\widehat{\theta}) + Bias^2_{\widehat{\theta}}(\theta)$$

where the bias is defined as

$$Bias_{\widehat{\theta}}(\theta) = E_\theta(\widehat{\theta}) - \theta\,.$$

Primitive example: I take a coin from my pocket and toss it 6 times. I get $HTHTTT$. The MLE of the probability of heads is

$$\widehat{p} = X/n$$

where X is the number of heads. In this case I get $\widehat{p} = \frac{1}{3}$.

Alternative estimate: $\tilde{p} = \frac{1}{2}$.

That is, $\tilde{p}$ ignores data; guess coin is fair.

The MSEs of these two estimators are

$$MSE_{\text{MLE}} = \frac{p(1-p)}{6}$$

and

$$MSE_{0.5} = (p - 0.5)^2$$

If $0.311 < p < 0.689$ then 2nd MSE is smaller than first.

For this reason I would recommend use of $\tilde{p}$ for sample sizes this small.

Same experiment with a thumbtack: tack can land point up (U) or tipped over (O).

If I get $UOUOOO$ how should I estimate p the probability of U?

Mathematics is identical to above but is $\tilde{p}$ is better than $\hat{p}$?

Less reason to believe $0.311 \le p \le 0.689$ than with a coin.

Unbiased Estimation

Binomial problem shows general phenomenon.

An estimator can be good for some values of θ and bad for others.

To compare $\hat{\theta}$ and $\tilde{\theta}$, two estimators of θ: Say $\hat{\theta}$ is better than $\tilde{\theta}$ if it has *uniformly* smaller MSE:

$$MSE_{\hat{\theta}}(\theta) \le MSE_{\tilde{\theta}}(\theta)$$

for **all** θ.

Normally we also require that the inequality be strict for at least one θ.

Question: is there a *best* estimate – one which is better than every other estimator?

Answer: NO. Suppose $\hat{\theta}$ were such a best estimate. Fix a θ^* in Θ and let $\tilde{\theta} \equiv \theta^*$.

Then MSE of $\tilde{\theta}$ is 0 when $\theta = \theta^*$. Since $\hat{\theta}$ is better than $\tilde{\theta}$ we must have

$$MSE_{\hat{\theta}}(\theta^*) = 0$$

so that $\hat{\theta} = \theta^*$ with probability equal to 1.

So $\hat{\theta} = \tilde{\theta}$.

If there are actually two different possible values of θ this gives a contradiction; so no such $\hat{\theta}$ exists.

Principle of Unbiasedness: A good estimate is unbiased, that is,

$$E_\theta(\hat\theta) \equiv \theta \, .$$

WARNING: In my view the Principle of Unbiasedness is a load of hog wash.

For an unbiased estimate the MSE is just the variance.

Definition: An estimator $\hat\phi$ of a parameter $\phi = \phi(\theta)$ is **Uniformly Minimum Variance Unbiased** (UMVU) if, whenever $\tilde\phi$ is an unbiased estimate of ϕ we have

$$\text{Var}_\theta(\hat\phi) \le \text{Var}_\theta(\tilde\phi)$$

We call $\hat\phi$ the UMVUE. ('E' is for Estimator.)

The point of having $\phi(\theta)$ is to study problems like estimating μ when you have two parameters like μ and σ for example.

Cramér Rao Inequality

If $\phi(\theta) = \theta$ we can derive some information from the identity

$$E_\theta(T) \equiv \theta$$

When we worked with the score function we derived some information from the identity

$$\int f(x,\theta)dx \equiv 1$$

by differentiation and we do the same here. If $T = T(X)$ is some function of the data X which is unbiased for θ then

$$E_\theta(T) = \int T(x)f(x,\theta)dx \equiv \theta$$

Differentiate both sides to get

$$\begin{aligned}
1 &= \frac{d}{d\theta}\int T(x)f(x,\theta)dx \\
&= \int T(x)\frac{\partial}{\partial\theta}f(x,\theta)dx \\
&= \int T(x)\frac{\partial}{\partial\theta}\log(f(x,\theta))f(x,\theta)dx \\
&= E_\theta(T(X)U(\theta))
\end{aligned}$$

where U is the score function.

Since score has mean 0

$$\text{Cov}_\theta(T(X), U(\theta)) = 1$$

Remember correlations between -1 and 1 or

$$\begin{aligned} 1 &= |\text{Cov}_\theta(T(X), U(\theta))| \\ &\le \sqrt{\text{Var}_\theta(T)\text{Var}_\theta(U(\theta))}\,. \end{aligned}$$

Squaring gives Cramér Rao Lower Bound:

$$\text{Var}_\theta(T) \ge \frac{1}{I(\theta)}\,.$$

Inequality is strict unless corr = 1 so that

$$U(\theta) = A(\theta)T(X) + B(\theta)$$

for non-random constants A and B (may depend on θ.) This would prove that

$$\ell(\theta) = A^*(\theta)T(X) + B^*(\theta) + C(X)$$

for other constants A^* and B^* and finally

$$f(x, \theta) = h(x)e^{A*(\theta)T(x)+B^*(\theta)}$$

for $h = e^C$.

Summary of Implications

- You can recognize a UMVUE sometimes. If $\text{Var}_\theta(T(X)) \equiv 1/I(\theta)$ then $T(X)$ is the UMVUE. In the $N(\mu, 1)$ example the Fisher information is n and $\text{Var}(\overline{X}) = 1/n$ so that $\overline{X}$ is the UMVUE of μ.

- In an asymptotic sense the MLE is nearly optimal: it is nearly unbiased and (approximate) variance nearly $1/I(\theta)$.

- Good estimates are highly correlated with the score.

- Densities of exponential form (called *exponential family*) given above are somehow special.

- Usually inequality is strict — strict unless score is affine function of a statistic T and T (or T/c for constant c) is unbiased for θ.

What can we do to find UMVUEs when the CRLB is a strict inequality?

Example: Suppose X has a Binomial(n,p) distribution. The score function is

$$U(p) = \frac{1}{p(1-p)}X - \frac{n}{1-p}$$

CRLB will be strict unless $T = cX$ for some c. If we are trying to estimate p then choosing $c = n^{-1}$ does give an unbiased estimate $\hat{p} = X/n$ and $T = X/n$ achieves the CRLB so it is UMVU.

Different tactic: Suppose $T(X)$ is some unbiased function of X. Then we have

$$E_p(T(X) - X/n) \equiv 0$$

because $\hat{p} = X/n$ is also unbiased. If $h(k) = T(k) - k/n$ then

$$E_p(h(X)) = \sum_{k=0}^{n} h(k)\binom{n}{k}p^k(1-p)^{n-k} \equiv 0$$

LHS of $\equiv$ sign is polynomial function of p as is RHS.

Thus if the left hand side is expanded out the coefficient of each power p^k is 0.

Constant term occurs only in term $k = 0$; its coefficient is

$$h(0)\binom{n}{0} = h(0)\,.$$

Thus $h(0) = 0$.

Now $p^1 = p$ occurs only in term $k = 1$ with coefficient $nh(1)$ so $h(1) = 0$.

Since terms with $k = 0$ or 1 are 0 the quantity p^2 occurs only in $k = 2$ term; coefficient is

$$n(n-1)h(2)/2$$

so $h(2) = 0$.

Continue to see that $h(k) = 0$ for each k.

So *only* unbiased function of X is X/n.

A Binomial random variable is a sum of n iid Bernoulli(p) rvs. If $Y_1, \ldots, Y_n$ iid Bernoulli(p) then $X = \sum Y_i$ is Binomial(n, p).

Could we do better by than $\hat{p} = X/n$ by trying $T(Y_1, \ldots, Y_n)$ for some other function T?

Try $n = 2$. There are 4 possible values for Y_1, Y_2. If $h(Y_1, Y_2) = T(Y_1, Y_2) - [Y_1 + Y_2]/2$ then

$$E_p(h(Y_1, Y_2)) \equiv 0$$

and we have

$$\begin{aligned} E_p(h(Y_1, Y_2)) &= h(0,0)(1-p)^2 \\ &\quad + [h(1,0) + h(0,1)]p(1-p) \\ &\quad + h(1,1)p^2 \,. \end{aligned}$$

This can be rewritten in the form

$$\sum_{k=0}^{n} w(k)\binom{n}{k}p^k(1-p)^{n-k}$$

where

$$\begin{aligned} w(0) &= h(0,0) \\ 2w(1) &= h(1,0) + h(0,1) \\ w(2) &= h(1,1)\,. \end{aligned}$$

So, as before $w(0) = w(1) = w(2) = 0$.

Argument can be used to prove:

For any unbiased estimate $T(Y_1, \ldots, Y_n)$:

Average value of $T(y_1, \ldots, y_n)$ over $y_1, \ldots, y_n$ which have exactly k 1s and $n-k$ 0s is k/n.

Now let's look at the variance of T:

$$\begin{aligned}
&\text{Var}(T)\\
&\quad= E_p([T(Y_1,\ldots,Y_n) - p]^2)\\
&\quad= E_p([T(Y_1,\ldots,Y_n) - X/n + X/n - p]^2)\\
&\quad= E_p([T(Y_1,\ldots,Y_n) - X/n]^2) +\\
&\qquad 2E_p([T(Y_1,\ldots,Y_n) - X/n][X/n - p])\\
&\quad+ E_p([X/n - p]^2)
\end{aligned}$$

Claim cross product term is 0 which will prove variance of T is variance of X/n plus a non-negative quantity (which will be positive unless $T(Y_1,\ldots,Y_n) \equiv X/n$). Compute the cross product term by writing

$$\begin{aligned}
&E_p([T(Y_1,\ldots,Y_n) - X/n][X/n - p])\\
&\quad= \sum_{y_1,\ldots,y_n} [T(y_1,\ldots,y_n) - \sum y_i/n][\sum y_i/n - p]\\
&\qquad\qquad \times p^{\sum y_i}(1-p)^{n-\sum y_i}
\end{aligned}$$

Sum over those $y_1,\ldots,y_n$ whose sum is an integer x; then sum over x:

$$\begin{aligned}
&E_p([T(Y_1,\ldots,Y_n) - X/n][X/n-p]) \\
&\qquad = \sum_{x=0}^{n} \sum_{\sum y_i = x} [T(y_1,\ldots,y_n) - \sum y_i/n] \\
&\qquad\qquad \times [\sum y_i/n - p] p^{\sum y_i}(1-p)^{n-\sum y_i} \\
&\quad = \sum_{x=0}^{n} \left[\sum_{\sum y_i = x} [T(y_1,\ldots,y_n) - x/n] \right] [x/n - p] \\
&\qquad\qquad\qquad\qquad\qquad\qquad \times p^x (1-p)^{n-x}
\end{aligned}$$

We have already shown that the sum in [] is 0!

This long, algebraically involved, method of proving that $\hat{p} = X/n$ is the UMVUE of p is one special case of a general tactic.

To get more insight rewrite

$$
\begin{aligned}
E_p\{T(Y_1,\ldots,Y_n)\}
&= \sum_{x=0}^{n} \sum_{\sum y_i = x} T(y_1,\ldots,y_n) \\
&\quad \times P(Y_1=y_1,\ldots,Y_n=y_n) \\
&= \sum_{x=0}^{n} \sum_{\sum y_i = x} T(y_1,\ldots,y_n) \\
&\quad \times P(Y_1=y_1,\ldots,Y_n=y_n|X=x)P(X=x) \\
&= \sum_{x=0}^{n} \frac{\sum_{\sum y_i = x} T(y_1,\ldots,y_n)}{\binom{n}{x}} \binom{n}{x} p^x (1-p)^{n-x}
\end{aligned}
$$

Notice: large fraction is average value of T over y such that $\sum y_i = x$.

Notice: weights in average do not depend on p.

Notice: this average is actually

$$
\begin{aligned}
E\{T(Y_1,\ldots,Y_n)|X=x\}
&= \sum_{y_1,\ldots,y_n} T(y_1,\ldots,y_n) \\
&\quad \times P(Y_1=y_1,\ldots,Y_n=y_n|X=x)
\end{aligned}
$$

Notice: conditional probabilities do not depend on p.

In a sequence of Binomial trials if I tell you that 5 of 17 were heads and the rest tails the actual trial numbers of the 5 Heads are chosen at random from the 17 possibilities; all of the 17 choose 5 possibilities have the same chance and this chance does not depend on p.

Notice: with data $Y_1, \ldots, Y_n$ log likelihood is

$$\ell(p) = \sum Y_i \log(p) - (n - \sum Y_i) \log(1-p)$$

and

$$U(p) = \frac{1}{p(1-p)} X - \frac{n}{1-p}$$

as before. Again CRLB is strict except for multiples of X. Since only unbiased multiple of X is $\hat{p} = X/n$ UMVUE of p is $\hat{p}$.

Sufficiency

In the binomial situation the conditional distribution of the data $Y_1, \ldots, Y_n$ given X is the same for all values of θ; we say this conditional distribution is **free** of θ.

Defn: Statistic $T(X)$ is sufficient for the model $\{P_\theta; \theta \in \Theta\}$ if conditional distribution of data X given $T = t$ is free of θ.

Intuition: Data tell us about θ **if** different values of θ give different distributions to X. If two different values of θ correspond to same density or cdf for X we cannot distinguish these two values of θ by examining X. Extension of this notion: if two values of θ give same conditional distribution of X given T then observing T in addition to X doesn't improve our ability to distinguish the two values.

Mathematically Precise version of this intuition: Suppose $T(X)$ is sufficient statistic and $S(X)$ is any estimate or confidence interval or ... If you only know value of T then:

- Generate an observation X^* (via some sort of Monte Carlo program) from the conditional distribution of X given T.
- Use $S(X^*)$ instead of $S(X)$. Then $S(X^*)$ has the same performance characteristics as $S(X)$ because the distribution of X^* is the same as that of X.

You can carry out the first step **only** if the statistic T is sufficient; otherwise you need to know the true value of θ to generate X^*.

Example 1: $Y_1,\ldots,Y_n$ iid Bernoulli(p). Given $\sum Y_i = y$ the indexes of the y successes have the same chance of being any one of the $\binom{n}{y}$ possible subsets of $\{1,\ldots,n\}$. Chance does not depend on p so $T(Y_1,\ldots,Y_n) = \sum Y_i$ is sufficient statistic.

Example 2: $X_1,\ldots,X_n$ iid $N(\mu,1)$. Joint distribution of $X_1,\ldots,X_n,\overline{X}$ is MVN. All entries of mean vector are μ. Variance covariance matrix partitioned as

$$\left[\begin{array}{cc} I_{n\times n} & \mathbf{1}_n/n \\ \mathbf{1}_n^t/n & 1/n \end{array}\right]$$

where $\mathbf{1}_n$ is column vector of n 1s and $I_{n\times n}$ is $n\times n$ identity matrix.

Compute conditional means and variances of X_i given $\overline{X}$; use fact that conditional law is MVN. Conclude conditional law of data given $\overline{X} = x$ is MVN. Mean vector has all entries x. Variance-covariance matrix is $I_{n\times n} - \mathbf{1}_n\mathbf{1}_n^t/n$. No dependence on μ so $\overline{X}$ is sufficient.

WARNING: Whether or not statistic is sufficient depends on density function and on Θ.

Theorem: [Rao-Blackwell] Suppose $S(X)$ is a sufficient statistic for model $\{P_\theta, \theta \in \Theta\}$. If T is an estimate of $\phi(\theta)$ then:

1. $E(T|S)$ is a statistic.

2. $E(T|S)$ has the same bias as T; if T is unbiased so is $E(T|S)$.

3. $\text{Var}_\theta(E(T|S)) \le \text{Var}_\theta(T)$ and the inequality is strict unless T is a function of S.

4. MSE of $E(T|S)$ is no more than MSE of T.

Proof: Review conditional distributions: abstract definition of conditional expectation is:

Defn: $E(Y|X)$ is any function of X such that

$$E\left[R(X)E(Y|X)\right] = E\left[R(X)Y\right]$$

for any function $R(X)$. $E(Y|X = x)$ is a function $g(x)$ such that

$$g(X) = E(Y|X)$$

Fact: If X, Y has joint density $f_{X,Y}(x, y)$ and conditional density $f(y|x)$ then

$$g(x) = \int y f(y|x) dy$$

satisfies these definitions.

Proof:

$$\begin{aligned} E(R(X)g(X)) &= \int R(x)g(x)f_X(x)dx \\ &= \int\int R(x)yf_X(x)f(y|x)dydx \\ &= \int\int R(x)yf_{X,Y}(x,y)dydx \\ &= E(R(X)Y) \end{aligned}$$

Think of $E(Y|X)$ as average Y holding X fixed. Behaves like ordinary expected value but functions of X only are like constants:

$$E(\sum A_i(X)Y_i|X) = \sum A_i(X)E(Y_i|X)$$

Example: $Y_1,\ldots,Y_n$ iid Bernoulli(p). Then $X = \sum Y_i$ is Binomial(n,p). Summary of conclusions:

- Log likelihood function of X only not of $Y_1,\ldots,Y_n$.

- Only function of X which is unbiased estimate of p is $\hat{p} = X/n$.

- If $T(Y_1,\ldots,Y_n)$ is unbiased for p then average value of $T(y_1,\ldots,y_n)$ over $y_1,\ldots,y_n$ for which $\sum y_i = x$ is x/n.

- Distribution of T given $\sum Y_i = x$ does not depend on p.

- If $T(Y_1,\ldots,Y_n)$ is unbiased for p then
$$\text{Var}(T) = \text{Var}(\hat{p}) + E[(T-\hat{p})^2]$$

- $\hat{p}$ is the UMVUE of p.

This proof that $\hat{p} = X/n$ is UMVUE of p is special case of general tactic.

Proof of the Rao Blackwell Theorem

Step 1: The definition of sufficiency is that the conditional distribution of X given S does not depend on θ. This means that $E(T(X)|S)$ does not depend on θ.

Step 2: This step hinges on the following identity (called Adam's law by Jerzy Neyman – he used to say it comes before all the others)

$$E[E(Y|X)] = E(Y)$$

which is just the definition of $E(Y|X)$ with $R(X) \equiv 1$.

From this we deduce that

$$E_\theta[E(T|S)] = E_\theta(T)$$

so that $E(T|S)$ and T have the same bias. If T is unbiased then

$$E_\theta[E(T|S)] = E_\theta(T) = \phi(\theta)$$

so that $E(T|S)$ is unbiased for ϕ.

Step 3: relies on very useful decomposition. (Total sum of squares = regression sum of squares + residual sum of squares.)

$$\text{Var(Y)} = \text{Var}\{E(Y|X)\} + E[\text{Var}(Y|X)]$$

The conditional variance means

$$\text{Var}(Y|X) = E[\{Y - E(Y|X)\}^2|X]$$

Square out right hand side:

$$\begin{aligned}\text{Var}(E(Y|X)) &= E[\{E(Y|X) - E[E(Y|X)]\}^2] \\ &= E[\{E(Y|X) - E(Y)\}^2]\end{aligned}$$

and

$$E[\text{Var}(Y|X)] = E[\{Y - E(Y|X)\}^2]$$

Adding these together gives

$$E\Big[Y^2 - 2YE[Y|X] + 2(E[Y|X])^2 \\ -2E(Y)E[Y|X] + E^2(Y)\Big]$$

Simplify remembering $E(Y|X)$ is function of X — constant when holding X fixed. So

$$E[Y|X]E[Y|X] = E[YE(Y|X)|X]$$

taking expectations gives

$$\begin{aligned} E[(E[Y|X])^2] &= E[E[YE(Y|X)|X]] \\ &= E[YE(Y|X)] \end{aligned}$$

So 3rd term above cancels with 2nd term.

Fourth term simplifies

$$E[E(Y)E[Y|X]] = E(Y)E[E[Y|X]] = E^2(Y)$$

so that

$$\text{Var}(E(Y|X)) + E[\text{Var}(Y|X)] = E[Y^2] - E^2(Y)$$

Apply to Rao Blackwell theorem to get

$$\text{Var}_\theta(T) = \text{Var}_\theta(E(T|S)) + E[(T - E(T|S))^2]$$

Second term ≥ 0 so variance of $E(T|S)$ is no more than that of T; will be strictly less unless $T = E(T|S)$. This would mean that T is already a function of S. Adding the squares of the biases of T (or of $E(T|S)$) gives the inequality for MSE.

Examples:

In the binomial problem $Y_1(1-Y_2)$ is an unbiased estimate of $p(1-p)$. We improve this by computing

$$E(Y_1(1-Y_2)|X)$$

We do this in two steps. First compute

$$E(Y_1(1-Y_2)|X=x)$$

Notice that the random variable $Y_1(1-Y_2)$ is either 1 or 0 so its expected value is just the probability it is equal to 1:

$$\begin{aligned}
&E(Y_1(1-Y_2)|X=x)\\
&\quad= P(Y_1(1-Y_2)=1|X=x)\\
&\quad= P(Y_1=1,Y_2=0|Y_1+Y_2+\cdots+Y_n=x)\\
&\quad= \frac{P(Y_1=1,Y_2=0,Y_1+\cdots+Y_n=x)}{P(Y_1+Y_2+\cdots+Y_n=x)}\\
&\quad= \frac{P(Y_1=1,Y_2=0,Y_3+\cdots+Y_n=x-1)}{\binom{n}{x}p^x(1-p)^{n-x}}\\
&\quad= \frac{p(1-p)\binom{n-2}{x-1}p^{x-1}(1-p)^{(n-2)-(x-1)}}{\binom{n}{x}p^x(1-p)^{n-x}}\\
&\quad= \frac{\binom{n-2}{x-1}}{\binom{n}{x}}\\
&\quad= \frac{x(n-x)}{n(n-1)}
\end{aligned}$$

This is simply $n\hat{p}(1-\hat{p})/(n-1)$ (can be bigger than 1/4, the maximum value of $p(1-p)$).

Example: If $X_1, \ldots, X_n$ are iid $N(\mu, 1)$ then $\bar{X}$ is sufficient and X_1 is an unbiased estimate of μ. Now

$$\begin{aligned} E(X_1|\bar{X}) &= E[X_1 - \bar{X} + \bar{X}|\bar{X}] \\ &= E[X_1 - \bar{X}|\bar{X}] + \bar{X} \\ &= \bar{X} \end{aligned}$$

which is the UMVUE.

Finding Sufficient statistics

Binomial(n, θ): log likelihood $\ell(\theta)$ (part depending on θ) is function of X alone, not of $Y_1, \ldots, Y_n$ as well.

Normal example: $\ell(\mu)$ is, ignoring terms not containing μ,

$$\ell(\mu) = \mu \sum X_i - n\mu^2/2 = n\mu\bar{X} - n\mu^2/2\,.$$

Examples of the **Factorization Criterion**:

Theorem: If the model for data X has density $f(x, \theta)$ then the statistic $S(X)$ is sufficient if and only if the density can be factored as

$$f(x, \theta) = g(S(x), \theta)h(x)$$

Proof: Find statistic $T(X)$ such that X is a one to one function of the pair S, T. Apply change of variables to the joint density of S and T. If the density factors then

$$f_{S,T}(s,t) = g(s,\theta)h(x(s,t))J(s,t)$$

where J is the Jacobian, so conditional density of T given $S = s$ does not depend on θ. Thus the conditional distribution of (S,T) given S does not depend on θ and finally the conditional distribution of X given S does not depend on θ.

Conversely if S is sufficient then the $f_{T|S}$ has no θ in it so joint density of S, T is

$$f_S(s,\theta)f_{T|S}(t|s)$$

Apply change of variables formula to get

$$f_X(x) = f_S(S(x),\theta)f_{T|S}(t(x)|S(x))J(x)$$

where J is the Jacobian. This factors.

Example: If $X_1,\ldots,X_n$ are iid $N(\mu,\sigma^2)$ then the joint density is

$$(2\pi)^{-n/2}\sigma^{-n}\times$$
$$\exp\{-\sum X_i^2/(2\sigma^2)+\mu\sum X_i/\sigma^2-n\mu^2/(2\sigma^2)\}$$

which is evidently a function of

$$\sum X_i^2, \sum X_i$$

This pair is a sufficient statistic. You can write this pair as a bijective function of $\bar{X}, \sum(X_i-\bar{X})^2$ so that this pair is also sufficient.

Example: If $Y_1,\ldots,Y_n$ are iid Bernoulli(p) then

$$\begin{aligned} f(y_1,\ldots,y_p;p) &= \prod p^{y_i}(1-p)^{1-y_i} \\ &= p^{\sum y_i}(1-p)^{n-\sum y_i} \end{aligned}$$

Define $g(x,p)=p^x(1-p)^{n-x}$ and $h\equiv 1$ to see that $X=\sum Y_i$ is sufficient by the factorization criterion.

Minimal Sufficiency

In any model $S(X) \equiv X$ is sufficient. (Apply the factorization criterion.) In any iid model the vector $X_{(1)}, \ldots, X_{(n)}$ of order statistics is sufficient. (Apply the factorization criterion.) In $N(\mu, 1)$ model we have 3 sufficient statistics:

1. $S_1 = (X_1, \ldots, X_n)$.

2. $S_2 = (X_{(1)}, \ldots, X_{(n)})$.

3. $S_3 = \bar{X}$.

Notice that I can calculate S_3 from the values of S_1 or S_2 but not vice versa and that I can calculate S_2 from S_1 but not vice versa. It turns out that $\bar{X}$ is a **minimal** sufficient statistic meaning that it is a function of any other sufficient statistic. (You can't collapse the data set any more without losing information about μ.)

Recognize minimal sufficient statistics from ℓ:

Fact: If you fix some particular θ^* then the log likelihood ratio function

$$\ell(\theta) - \ell(\theta^*)$$

is minimal sufficient. WARNING: the function is the statistic.

Subtraction of $\ell(\theta^*)$ gets rid of irrelevant constants in ℓ. In $N(\mu, 1)$ example:

$$\ell(\mu) = -n\log(2\pi)/2 - \sum X_i^2/2 + \mu\sum X_i - n\mu^2/2$$

depends on $\sum X_i^2$, not needed for sufficient statistic. Take $\mu^* = 0$ and get

$$\ell(\mu) - \ell(\mu^*) = \mu\sum X_i - n\mu^2/2$$

This function of μ is minimal sufficient. Notice: from $\sum X_i$ can compute this minimal sufficient statistic and vice versa. Thus $\sum X_i$ is also minimal sufficient.

Completeness

In Binomial(n, p) example only one function of X is unbiased. Rao Blackwell shows UMVUE, if it exists, will be a function of any sufficient statistic.

Q: Can there be more than one such function?

A: Yes in general but no for some models like the binomial.

Definition: A statistic T is complete for a model $P_\theta; \theta \in \Theta$ if

$$E_\theta(h(T)) = 0$$

for all θ implies $h(T) = 0$.

We have already seen that X is complete in the Binomial(n, p) model. In the $N(\mu, 1)$ model suppose

$$E_\mu(h(\bar{X})) \equiv 0 \,.$$

Since $\bar{X}$ has a $N(\mu, 1/n)$ distribution we find that

$$E(h(\bar{X})) = \frac{\sqrt{n}e^{-n\mu^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} h(x) e^{-nx^2/2} e^{n\mu x} dx$$

so that

$$\int_{-\infty}^{\infty} h(x) e^{-nx^2/2} e^{n\mu x} dx \equiv 0 \,.$$

Called Laplace transform of $h(x)e^{-nx^2/2}$.

Theorem: Laplace transform is 0 if and only if the function is 0 (because you can invert the transform).

Hence $h \equiv 0$.

How to Prove Completeness

Only one general tactic: suppose X has density

$$f(x,\theta) = h(x)\,\mathrm{exp}\{\sum_1^p a_i(\theta)S_i(x) + c(\theta)\}$$

If the range of the function $(a_1(\theta),\ldots,a_p(\theta))$ as θ varies over Θ contains a (hyper-) rectangle in R^p then the statistic

$$(S_1(X),\ldots,S_p(X))$$

is complete and sufficient.

You prove the sufficiency by the factorization criterion and the completeness using the properties of Laplace transforms and the fact that the joint density of $S_1,\ldots,S_p$

$$g(s_1,\ldots,s_p;\theta) = h^*(s)\,\mathrm{exp}\{\sum a_k(\theta)s_k + c^*(\theta)\}$$

Example: $N(\mu,\sigma^2)$ model density has form

$$\frac{\exp\left\{\left(-\frac{1}{2\sigma^2}\right)x^2+\left(\frac{\mu}{\sigma^2}\right)x-\frac{\mu^2}{2\sigma^2}-\log\sigma\right\}}{\sqrt{2\pi}}$$

which is an exponential family with

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

$$a_1(\theta)=-\frac{1}{2\sigma^2}$$

$$S_1(x)=x^2$$

$$a_2(\theta)=\frac{\mu}{\sigma^2}$$

$$S_2(x)=x$$

and

$$c(\theta)=-\frac{\mu^2}{2\sigma^2}-\log\sigma\,.$$

It follows that

$$(\sum X_i^2,\sum X_i)$$

is a complete sufficient statistic.

Remark: The statistic $(s^2, \bar{X})$ is a one to one function of $(\sum X_i^2, \sum X_i)$ so it must be complete and sufficient, too. Any function of the latter statistic can be rewritten as a function of the former and vice versa.

FACT: A complete sufficient statistic is also minimal sufficient.

The Lehmann-Scheffé Theorem

Theorem: If S is a complete sufficient statistic for some model and $h(S)$ is an unbiased estimate of some parameter $\phi(\theta)$ then $h(S)$ is the UMVUE of $\phi(\theta)$.

Proof: Suppose T is another unbiased estimate of ϕ. According to Rao-Blackwell, T is improved by $E(T|S)$ so if $h(S)$ is not UMVUE then there must exist another function $h^*(S)$ which is unbiased and whose variance is smaller than that of $h(S)$ for some value of θ. But

$$E_\theta(h^*(S) - h(S)) \equiv 0$$

so, in fact $h^*(S) = h(S)$.

Example: In the $N(\mu,\sigma^2)$ example the random variable $(n-1)s^2/\sigma^2$ has a χ^2_{n-1} distribution. It follows that

$$E\left[\frac{\sqrt{n-1}s}{\sigma}\right] = \frac{\int_0^\infty x^{1/2}\left(\frac{x}{2}\right)^{(n-1)/2-1} e^{-x/2}dx}{2\Gamma((n-1)/2)}\,.$$

Make the substitution $y=x/2$ and get

$$E(s) = \frac{\sigma}{\sqrt{n-1}}\frac{\sqrt{2}}{\Gamma((n-1)/2)}\int_0^\infty y^{n/2-1}e^{-y}dy\,.$$

Hence

$$E(s) = \sigma\frac{\sqrt{2}\Gamma(n/2)}{\sqrt{n-1}\Gamma((n-1)/2)}\,.$$

The UMVUE of σ is then

$$s\frac{\sqrt{n-1}\Gamma((n-1)/2)}{\sqrt{2}\Gamma(n/2)}$$

by the Lehmann-Scheffé theorem.

Criticism of Unbiasedness

- UMVUE can be **inadmissible for squared error loss** meaning there is a (biased, of course) estimate whose MSE is smaller for every parameter value. An example is the UMVUE of $\phi = p(1-p)$ which is $\hat\phi = n\hat{p}(1-\hat{p})/(n-1)$. The MSE of
$$\tilde\phi = \min(\hat\phi, 1/4)$$
is smaller than that of $\hat\phi$.

- Unbiased estimation may be impossible.

  Binomial(n,p) log odds is
$$\phi = \log(p/(1-p))\,.$$
  Since the expectation of any function of the data is a polynomial function of p and since ϕ is **not** a polynomial function of p there is no unbiased estimate of ϕ

- The UMVUE of σ is not the square root of the UMVUE of σ^2. This method of estimation does not have the parameterization equivariance that maximum likelihood does.

- Unbiasedness is irrelevant (unless you average together many estimators).

  Property is an average over possible values of the estimate in which positive errors are allowed to cancel negative errors.

  Exception to criticism: if you average a number of estimators to get a single estimator then it is a problem if all the estimators have the same bias.

  See assignment 5, one way layout example: mle of the residual variance averages together many biased estimates and so is very badly biased. That assignment shows that the solution is not really to insist on unbiasedness but to consider an alternative to averaging for putting the individual estimates together.

Hypothesis Testing

Hypothesis testing: a statistical problem where you must choose, on the basis of data X, between two alternatives. We formalize this as the problem of choosing between two *hypotheses*: $H_o : \theta \in \Theta_0$ or $H_1 : \theta \in \Theta_1$ where Θ_0 and Θ_1 are a partition of the model $P_\theta; \theta \in \Theta$. That is $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$.

A rule for making the required choice can be described in two ways:

1. In terms of the set

   $$R = \{X : \text{we choose } \Theta_1 \text{ if we observe } X\}$$

   called the *rejection* or *critical* region of the test.

2. In terms of a function $\phi(x)$ which is equal to 1 for those x for which we choose Θ_1 and 0 for those x for which we choose Θ_0.

For technical reasons which will come up soon I prefer to use the second description. However, each ϕ corresponds to a unique rejection region $R_\phi = \{x : \phi(x) = 1\}$.

Neyman Pearson approach treats two hypotheses asymmetrically. Hypothesis H_o referred to as the *null* hypothesis (traditionally the hypothesis that some treatment has no effect).

Definition: The power function of a test ϕ (or the corresponding critical region R_ϕ) is

$$\pi(\theta) = P_\theta(X \in R_\phi) = E_\theta(\phi(X))$$

Interested in **optimality** theory, that is, the problem of finding the best ϕ. A good ϕ will evidently have $\pi(\theta)$ small for $\theta \in \Theta_0$ and large for $\theta \in \Theta_1$. There is generally a trade off which can be made in many ways, however.

Simple versus Simple testing

Finding a best test is easiest when the hypotheses are very precise.

Definition: A hypothesis H_i is **simple** if Θ_i contains only a single value θ_i.

The simple versus simple testing problem arises when we test $\theta = \theta_0$ against $\theta = \theta_1$ so that Θ has only two points in it. This problem is of importance as a technical tool, not because it is a realistic situation.

Suppose that the model specifies that if $\theta = \theta_0$ then the density of X is $f_0(x)$ and if $\theta = \theta_1$ then the density of X is $f_1(x)$. How should we choose ϕ? To answer the question we begin by studying the problem of minimizing the total error probability.

Type I error: the error made when $\theta = \theta_0$ but we choose H_1, that is, $X \in R_\phi$.

Type II error: when $\theta = \theta_1$ but we choose H_0.

The **level** of a simple versus simple test is

$$\alpha = P_{\theta_0}(\text{We make a Type I error})$$

or

$$\alpha = P_{\theta_0}(X \in R_\phi) = E_{\theta_0}(\phi(X))$$

Other error probability denoted β is

$$\beta = P_{\theta_1}(X \not\in R_\phi) = E_{\theta_1}(1 - \phi(X)).$$

Minimize $\alpha+\beta$, the total error probability given by

$$\begin{aligned} \alpha + \beta &= E_{\theta_0}(\phi(X)) + E_{\theta_1}(1 - \phi(X)) \\ &= \int [\phi(x) f_0(x) + (1 - \phi(x)) f_1(x)] dx \end{aligned}$$

Problem: choose, for each x, either the value 0 or the value 1, in such a way as to minimize the integral. But for each x the quantity

$$\phi(x)f_0(x) + (1-\phi(x))f_1(x)$$

is between $f_0(x)$ and $f_1(x)$. To make it small we take $\phi(x) = 1$ if $f_1(x) > f_0(x)$ and $\phi(x) = 0$ if $f_1(x) < f_0(x)$. It makes no difference what we do for those x for which $f_1(x) = f_0(x)$. Notice: divide both sides of inequalities to get condition in terms of **likelihood ratio** $f_1(x)/f_0(x)$.

Theorem: For each fixed λ the quantity $\beta+\lambda\alpha$ is minimized by any ϕ which has

$$\phi(x) = \begin{cases} 1 & \frac{f_1(x)}{f_0(x)} > \lambda \\ 0 & \frac{f_1(x)}{f_0(x)} < \lambda \end{cases}$$

Neyman and Pearson suggested that in practice the two kinds of errors might well have unequal consequences. They suggested that rather than minimize any quantity of the form above you pick the more serious kind of error, label it **Type I** and require your rule to hold the probability α of a Type I error to be no more than some prespecified level α_0. (This value α_0 is typically 0.05 these days, chiefly for historical reasons.)

The Neyman and Pearson approach is then to minimize β subject to the constraint $\alpha \le \alpha_0$. Usually this is really equivalent to the constraint $\alpha = \alpha_0$ (because if you use $\alpha < \alpha_0$ you could make R larger and keep $\alpha \le \alpha_0$ but make β smaller. For discrete models, however, this may not be possible.

Example: Suppose X is Binomial(n,p) and either $p=p_0=1/2$ or $p=p_1=3/4$.

If R is any critical region (so R is a subset of $\{0,1,\ldots,n\}$) then

$$P_{1/2}(X\in R)=\frac{k}{2^n}$$

for some integer k. Example: to get $\alpha_0=0.05$ with $n=5$: possible values of α are $0, 1/32 = 0.03125, 2/32=0.0625$, etc.

Possible rejection regions for $\alpha_0=0.05$:

Region	α	β
$R_1=\emptyset$	0	1
$R_2=\{x=0\}$	0.03125	$1-(1/4)^5$
$R_3=\{x=5\}$	0.03125	$1-(3/4)^5$

So R_3 minimizes β subject to $\alpha<0.05$.

Raise α_0 slightly to 0.0625: possible rejection regions are R_1, R_2, R_3 and $R_4=R_2\cup R_3$.

The first three have the same α and β as before while R_4 has $\alpha = \alpha_0 = 0.0625$ an $\beta = 1 - (3/4)^5 - (1/4)^5$. Thus R_4 is optimal!

Problem: if all trials are failures "optimal" R chooses $p = 3/4$ rather than $p = 1/2$.

But: $p = 1/2$ makes 5 failures much more likely than $p = 3/4$.

Problem is discreteness. Solution:

Expand set of possible values of ϕ to $[0, 1]$. Values of $\phi(x)$ between 0 and 1 represent the chance that we choose H_1 given that we observe x; the idea is that we actually toss a (biased) coin to decide! This tactic will show us the kinds of rejection regions which are sensible.

In practice: restrict our attention to levels α_0 for which best ϕ is always either 0 or 1. In the binomial example we will insist that the value of α_0 be either 0 or $P_{\theta_0}(X \geq 5)$ or $P_{\theta_0}(X \geq 4)$ or ...

Smaller example: 4 possible values of X and 2^4 possible rejection regions. Here is a table of the levels for each possible rejection region R:

R	α
$\emptyset$	0
{3}, {0}	1/8
{0,3}	2/8
{1}, {2}	3/8
{0,1}, {0,2}, {1,3}, {2,3}	4/8
{0,1,3}, {0,2,3}	5/8
{1,2}	6/8
{0,1,2}, {1,2,3}	7/8
{0,1,2,3}	1

Best level 2/8 test has rejection region $\{0,3\}$, $\beta = 1 - [(3/4)^3 + (1/4)^3] = 36/64$. Best level 2/8 test using randomization rejects when $X = 3$ and, when $X = 2$ tosses a coin with $P(H) = 1/3$, then rejects if you get H. Level is $1/8 + (1/3)(3/8) = 2/8$; probability of Type II error is $\beta = 1 - [(3/4)^3 + (1/3)(3)(3/4)^2(1/4)] = 28/64$.

Definition: A hypothesis test is a function $\phi(x)$ whose values are always in $[0,1]$. If we observe $X = x$ then we choose H_1 with conditional probability $\phi(X)$. In this case we have

$$\pi(\theta) = E_\theta(\phi(X))$$

$$\alpha = E_0(\phi(X))$$

and

$$\beta = 1 - E_1(\phi(X))$$

Note that a test using a rejection region C is equivalent to

$$\phi(x) = 1(x \in C)$$

The Neyman Pearson Lemma: In testing f_0 against f_1 the probability β of a type II error is minimized, subject to $\alpha \le \alpha_0$ by the test function:

$$\phi(x) = \begin{cases} 1 & \frac{f_1(x)}{f_0(x)} > \lambda \\ \gamma & \frac{f_1(x)}{f_0(x)} = \lambda \\ 0 & \frac{f_1(x)}{f_0(x)} < \lambda \end{cases}$$

where λ is the largest constant such that

$$P_0(\frac{f_1(X)}{f_0(X)} \ge \lambda) \ge \alpha_0$$

and

$$P_0(\frac{f_1(X)}{f_0(X)} \le \lambda) \ge 1 - \alpha_0$$

and where γ is any number chosen so that

$$\begin{aligned} E_0(\phi(X)) &= P_0(\frac{f_1(X)}{f_0(X)} > \lambda) \\ &+ \quad \gamma P_0(\frac{f_1(X)}{f_0(X)} = \lambda) \\ &= \alpha_0 \end{aligned}$$

The value of γ is unique if $P_0(\frac{f_1(X)}{f_0(X)} = \lambda) > 0$.

Example: Binomial(n, p) with $p_0 = 1/2$ and $p_1 = 3/4$: ratio f_1/f_0 is

$$3^x 2^{-n}$$

If $n = 5$ this ratio is one of 1, 3, 9, 27, 81, 243 divided by 32.

Suppose we have $\alpha = 0.05$. λ must be one of the possible values of f_1/f_0. If we try $\lambda = 243/32$ then

$$\begin{aligned} P_0(3^X 2^{-5} \geq 243/32) &= P_0(X = 5) \\ &= 1/32 < 0.05 \end{aligned}$$

and

$$\begin{aligned} P_0(3^X 2^{-5} \geq 81/32) &= P_0(X \geq 4) \\ &= 6/32 > 0.05 \end{aligned}$$

So $\lambda = 81/32$.

Since

$$P_0(3^X 2^{-5} > 81/32) = P_0(X = 5) = 1/32$$

we must solve

$$P_0(X = 5) + \gamma P_0(X = 4) = 0.05$$

for γ and find

$$\gamma = \frac{0.05 - 1/32}{5/32} = 0.12$$

NOTE: No-one ever uses this procedure. Instead the value of α_0 used in discrete problems is chosen to be a possible value of the rejection probability when $\gamma = 0$ (or $\gamma = 1$). When the sample size is large you can come very close to any desired α_0 with a non-randomized test.

If $\alpha_0 = 6/32$ then we can either take λ to be 243/32 and $\gamma = 1$ or $\lambda = 81/32$ and $\gamma = 0$. However, our definition of λ in the theorem makes $\lambda = 81/32$ and $\gamma = 0$.

When the theorem is used for continuous distributions it can be the case that the cdf of $f_1(X)/f_0(X)$ has a flat spot where it is equal to $1-\alpha_0$. This is the point of the word "largest" in the theorem.

Example: If $X_1, \ldots, X_n$ are iid $N(\mu, 1)$ and we have $\mu_0 = 0$ and $\mu_1 > 0$ then

$$\frac{f_1(X_1,\ldots,X_n)}{f_0(X_1,\ldots,X_n)}$$

$$=$$

$$\exp\{\mu_1 \sum X_i - n\mu_1^2/2 - \mu_0 \sum X_i + n\mu_0^2/2\}$$

which simplifies to

$$\exp\{\mu_1 \sum X_i - n\mu_1^2/2\}$$

Now choose λ so that

$$P_0(\exp\{\mu_1 \sum X_i - n\mu_1^2/2\} > \lambda) = \alpha_0$$

Can make it equal because $f_1(X)/f_0(X)$ has a continuous distribution. Rewrite probability as

$$P_0(\sum X_i > [\log(\lambda) + n\mu_1^2/2]/\mu_1) \\ = \\ 1 - \Phi\left(\frac{\log(\lambda) + n\mu_1^2/2}{n^{1/2}\mu_1}\right)$$

Let z_α be upper α critical point of $N(0,1)$; then

$$z_{\alpha_0} = [\log(\lambda) + n\mu_1^2/2]/[n^{1/2}\mu_1] \,.$$

Solve to get a formula for λ in terms of z_{α_0}, n and μ_1.

The rejection region looks complicated: reject if a complicated statistic is larger than λ which has a complicated formula. But in calculating λ we re-expressed the rejection region in terms of

$$\frac{\sum X_i}{\sqrt{n}} > z_{\alpha_0}$$

The key feature is that this rejection region is the same for any $\mu_1 > 0$. [WARNING: in the algebra above I used $\mu_1 > 0$.] This is why the Neyman Pearson lemma is a lemma!

Definition: In the general problem of testing Θ_0 against Θ_1 the level of a test function ϕ is

$$\alpha = \sup_{\theta \in \Theta_0} E_\theta(\phi(X))$$

The power function is

$$\pi(\theta) = E_\theta(\phi(X))$$

A test ϕ^* is a Uniformly Most Powerful level α_0 test if

1. ϕ^* has level $\alpha \le \alpha_o$

2. If ϕ has level $\alpha \le \alpha_0$ then for every $\theta \in \Theta_1$ we have

$$E_\theta(\phi(X)) \le E_\theta(\phi^*(X))$$

Proof of Neyman Pearson lemma: Given a test ϕ with level strictly less than α_0 we can define the test

$$\phi^*(x) = \frac{1-\alpha_0}{1-\alpha}\phi(x) + \frac{\alpha_0-\alpha}{1-\alpha}$$

has level α_0 and β smaller than that of ϕ. Hence we may assume without loss that $\alpha = \alpha_0$ and minimize β subject to $\alpha = \alpha_0$. However, the argument which follows doesn't actually need this.

Lagrange Multipliers

Suppose you want to minimize $f(x)$ subject to $g(x) = 0$. Consider first the function

$$h_\lambda(x) = f(x) + \lambda g(x)$$

If x_λ minimizes h_λ then for any other x

$$f(x_\lambda) \le f(x) + \lambda[g(x) - g(x_\lambda)]$$

Now suppose you can find a value of λ such that the solution x_λ has $g(x_\lambda) = 0$. Then for any x we have

$$f(x_\lambda) \le f(x) + \lambda g(x)$$

and for any x satisfying the constraint $g(x) = 0$ we have

$$f(x_\lambda) \le f(x)$$

This proves that for this special value of λ the quantity x_λ minimizes $f(x)$ subject to $g(x) = 0$.

Notice that to find x_λ you set the usual partial derivatives equal to 0; then to find the special x_λ you add in the condition $g(x_\lambda) = 0$.

Return to proof of NP lemma

For each $\lambda > 0$ we have seen that ϕ_λ minimizes $\lambda\alpha + \beta$ where $\phi_\lambda = 1(f_1(x)/f_0(x) \ge \lambda)$.

As λ increases the level of ϕ_λ decreases from 1 when $\lambda = 0$ to 0 when $\lambda = \infty$. There is thus a value λ_0 where for $\lambda > \lambda_0$ the level is less than α_0 while for $\lambda < \lambda_0$ the level is at least α_0. Temporarily let $\delta = P_0(f_1(X)/f_0(X) = \lambda_0)$. If $\delta = 0$ define $\phi = \phi_\lambda$. If $\delta > 0$ define

$$\phi(x) = \begin{cases} 1 & \frac{f_1(x)}{f_0(x)} > \lambda_0 \\ \gamma & \frac{f_1(x)}{f_0(x)} = \lambda_0 \\ 0 & \frac{f_1(x)}{f_0(x)} < \lambda_0 \end{cases}$$

where $P_0(f_1(X)/f_0(X) > \lambda_0) + \gamma\delta = \alpha_0$. You can check that $\gamma \in [0, 1]$.

Now ϕ has level α_0 and according to the theorem above minimizes $lambda_0\alpha + \beta$. Suppose ϕ^* is some other test with level $\alpha^* \le \alpha_0$. Then

$$\lambda_0\alpha_\phi + \beta_\phi \le \lambda_0\alpha_{\phi^*} + \beta_{\phi^*}$$

We can rearrange this as

$$\beta_{\phi^*} \ge \beta_\phi + (\alpha_\phi - \alpha_{\phi^*})\lambda_0$$

Since

$$\alpha_{\phi^*} \le \alpha_0 = \alpha_\phi$$

the second term is non-negative and

$$\beta_{\phi^*} \ge \beta_\phi$$

which proves the Neyman Pearson Lemma.

Example application of NP: Binomial(n, p) to test $p = p_0$ versus p_1 for a $p_1 > p_0$ the NP test is of the form

$$\phi(x) = 1(X > k) + \gamma 1(X = k)$$

where we choose k so that

$$P_{p_0}(X > k) \le \alpha_0 < P_{p_0}(X \ge k)$$

and $\gamma \in [0, 1)$ so that

$$\alpha_0 = P_{p_0}(X > k) + \gamma P_{p_0}(X = k)$$

This rejection region depends only on p_0 and not on p_1 so that this test is UMP for $p = p_0$ against $p > p_0$. Since this test has level α_0 even for the larger null hypothesis it is also UMP for $p \le p_0$ against $p > p_0$.

Application of the NP lemma: In the $N(\mu,1)$ model consider $\Theta_1 = \{\mu > 0\}$ and $\Theta_0 = \{0\}$ or $\Theta_0 = \{\mu \le 0\}$. The UMP level α_0 test of $H_0 : \mu \in \Theta_0$ against $H_1 : \mu \in \Theta_1$ is

$$\phi(X_1, \ldots, X_n) = 1(n^{1/2}\bar{X} > z_{\alpha_0})$$

Proof: For either choice of Θ_0 this test has level α_0 because for $\mu \le 0$ we have

$$\begin{aligned} P_\mu(n^{1/2}\bar{X} &> z_{\alpha_0}) \\ &= P_\mu(n^{1/2}(\bar{X} - \mu) > z_{\alpha_0} - n^{1/2}\mu) \\ &= P(N(0,1) > z_{\alpha_0} - n^{1/2}\mu) \\ &\le P(N(0,1) > z_{\alpha_0}) \\ &= \alpha_0 \end{aligned}$$

(Notice the use of $\mu \le 0$. The central point is that the critical point is determined by the behaviour on the edge of the null hypothesis.)

Now if ϕ is any other level α_0 test then we have

$$E_0(\phi(X_1,\ldots,X_n)) \le \alpha_0$$

Fix a $\mu > 0$. According to the NP lemma

$$E_\mu(\phi(X_1,\ldots,X_n)) \le E_\mu(\phi_\mu(X_1,\ldots,X_n))$$

where ϕ_μ rejects if

$$f_\mu(X_1,\ldots,X_n)/f_0(X_1,\ldots,X_n) > \lambda$$

for a suitable λ. But we just checked that this test had a rejection region of the form

$$n^{1/2}\bar{X} > z_{\alpha_0}$$

which is the rejection region of ϕ^*. The NP lemma produces the same test for every $\mu > 0$ chosen as an alternative. So we have shown that $\phi_\mu = \phi^*$ for any $\mu > 0$.

Fairly general phenomenon: for any $\mu > \mu_0$ the likelihood ratio f_μ/f_0 is an increasing function of $\sum X_i$. The rejection region of the NP test is thus always a region of the form $\sum X_i > k$. The value of the constant k is determined by the requirement that the test have level α_0 and this depends on μ_0 not on μ_1.

Definition: The family $f_\theta; \theta \in \Theta \subset R$ has monotone likelihood ratio with respect to a statistic $T(X)$ if for each $\theta_1 > \theta_0$ the likelihood ratio $f_{\theta_1}(X)/f_{\theta_0}(X)$ is a monotone increasing function of $T(X)$.

Theorem: For a monotone likelihood ratio family the Uniformly Most Powerful level α test of $\theta \le \theta_0$ (or of $\theta = \theta_0$) against the alternative $\theta > \theta_0$ is

$$\phi(x) = \begin{cases} 1 & T(x) > t_\alpha \\ \gamma & T(X) = t_\alpha \\ 0 & T(x) < t_\alpha \end{cases}$$

where

$$P_{\theta_0}(T(X) > t_\alpha) + \gamma P_{\theta_0}(T(X) = t_\alpha) = \alpha_0 \, .$$

Typical family where this works: one parameter exponential family. Usually there is no UMP test.

Example: test $\mu = \mu_0$ against two sided alternative $\mu \neq \mu_0$. There is no UMP level α test.

If there were its power at $\mu > \mu_0$ would have to be as high as that of the one sided level α test and so its rejection region would have to be the same as that test, rejecting for large positive values of $\bar{X} - \mu_0$. But it also has to have power as good as the one sided test for the alternative $\mu < \mu_0$ and so would have to reject for large negative values of $\bar{X} - \mu_0$. This would make its level too large.

Favourite test: usual 2 sided test rejects for large values of $|\bar{X} - \mu_0|$. Test maximizes power subject to two constraints: first, level α; second power is minimized at $\mu = \mu_0$. Second condition means power on alternative is larger than on the null.

Definition: A test ϕ of Θ_0 against Θ_1 is unbiased level α if it has level α and, for every $\theta \in \Theta_1$ we have

$$\pi(\theta) \geq \alpha\,.$$

When testing a point null hypothesis like $\mu = \mu_0$ this requires that the power function be minimized at μ_0 which will mean that if π is differentiable then

$$\pi'(\mu_0) = 0$$

Example: $N(\mu,1)$: data $X=(X_1,\ldots,X_n)$. If ϕ is any test function then

$$\pi'(\mu)=\frac{\partial}{\partial\mu}\int\phi(x)f(x,\mu)dx$$

Differentiate under the integral and use

$$\frac{\partial f(x,\mu)}{\partial\mu}=\sum(x_i-\mu)f(x,\mu)$$

to get the condition

$$\int\phi(x)\bar{x}f(x,\mu_0)dx=\mu_0\alpha_0$$

Minimize $\beta(\mu)$ subject to two constraints

$$E_{\mu_0}(\phi(X))=\alpha_0$$

and

$$E_{\mu_0}(\bar{X}\phi(X))=\mu_0\alpha_0.$$

Fix two values $\lambda_1 > 0$ and λ_2 and minimize

$$\lambda_1\alpha + \lambda_2 E_{\mu_0}[(\bar{X} - \mu_0)\phi(X)] + \beta$$

The quantity in question is just

$$\int [\phi(x) f_0(x)(\lambda_1 + \lambda_2(\bar{x} - \mu_0)) \\ + (1 - \phi(x)) f_1(x)] dx \,.$$

As before this is minimized by

$$\phi(x) = \begin{cases} 1 & \frac{f_1(x)}{f_0(x)} > \lambda_1 + \lambda_2(\bar{x} - \mu_0) \\ 0 & \frac{f_1(x)}{f_0(x)} < \lambda_1 + \lambda_2(\bar{x} - \mu_0) \end{cases}$$

The likelihood ratio f_1/f_0 is simply

$$\exp\{n(\mu_1 - \mu_0)\bar{X} + n(\mu_0^2 - \mu_1^2)/2\}$$

and this exceeds the linear function

$$\lambda_1 + \lambda_2(\bar{X} - \mu_0)$$

for all $\bar{X}$ sufficiently large or small. That is,

$$\lambda_1\alpha + \lambda_2 E_{\mu_0}[(\bar{X} - \mu_0)\phi(X)] + \beta$$

is minimized by a rejection region of the form

$$\{\bar{X} > K_U\} \cup \{\bar{X} < K_L\}$$

Satisfy constraints: adjust K_U and K_L to get level α and $\pi'(\mu_0) = 0$. 2nd condition shows rejection region symmetric about μ_0 so test rejects for

$$\sqrt{n}|\bar{X} - \mu_0| > z_{\alpha/2}$$

Mimic Neyman Pearson lemma proof to check that if λ_1 and λ_2 are adjusted so that the unconstrained problem has the rejection region given then the resulting test minimizes β subject to the two constraints.

A test ϕ^* is a Uniformly Most Powerful Unbiased level α_0 test if

1. ϕ^* has level $\alpha \le \alpha_0$.

2. ϕ^* is unbiased.

3. If ϕ has level $\alpha \le \alpha_0$ and ϕ is unbiased then for every $\theta \in \Theta_1$ we have
$$E_\theta(\phi(X)) \le E_\theta(\phi^*(X))$$

Conclusion: The two sided z test which rejects if
$$|Z| > z_{\alpha/2}$$
where
$$Z = n^{1/2}(\bar{X} - \mu_0)$$
is the uniformly most powerful unbiased test of $\mu = \mu_0$ against the two sided alternative $\mu \neq \mu_0$.

Nuisance Parameters

The t-test is UMPU.

Suppose $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$. Test $\mu = \mu_0$ or $\mu \le \mu_0$ against $\mu > \mu_0$. Parameter space is two dimensional; boundary between the null and alternative is

$$\{(\mu, \sigma); \mu = \mu_0, \sigma > 0\}$$

If a test has $\pi(\mu, \sigma) \le \alpha$ for all $\mu \le \mu_0$ and $\pi(\mu, \sigma) \ge \alpha$ for all $\mu > \mu_0$ then $\pi(\mu_0, \sigma) = \alpha$ for all σ because the power function of any test must be continuous. (Uses dominated convergence theorem; power function is an integral.)

Think of $\{(\mu, \sigma); \mu = \mu_0\}$ as parameter space for a model. For this parameter space

$$S = \sum (X_i - \mu_0)^2$$

is complete and sufficient. Remember definitions of both completeness and sufficiency depend on the parameter space.

Suppose $\phi(\sum X_i, S)$ is an unbiased level α test. Then we have

$$E_{\mu_0,\sigma}(\phi(\sum X_i, S)) = \alpha$$

for all σ. Condition on S and get

$$E_{\mu_0,\sigma}[E(\phi(\sum X_i, S)|S)] = \alpha$$

for all σ. Sufficiency guarantees that

$$g(S) = E(\phi(\sum X_i, S)|S)$$

is a statistic and completeness that

$$g(S) \equiv \alpha$$

Now let us fix a single value of σ and a $\mu_1 > \mu_0$. To make our notation simpler I take $\mu_0 = 0$. Our observations above permit us to condition on $S = s$. Given $S = s$ we have a level α test which is a function of $\bar{X}$.

If we maximize the conditional power of this test for each s then we will maximize its power. What is the conditional model given $S = s$? That is, what is the conditional distribution of $\bar{X}$ given $S = s$? The answer is that the joint density of $\bar{X}, S$ is of the form

$$f_{\bar{X},S}(t,s) = h(s,t)\,\mathrm{exp}\{\theta_1 t + \theta_2 s + c(\theta_1,\theta_2)\}$$

where $\theta_1 = n\mu/\sigma^2$ and $\theta_2 = -1/\sigma^2$.

This makes the conditional density of $\bar{X}$ given $S = s$ of the form

$$f_{\bar{X}|s}(t|s) = h(s,t)\,\mathrm{exp}\{\theta_1 t + c^*(\theta_1,s)\}$$

Note disappearance of θ_2 and null is $\theta_1 = 0$. This permits application of NP lemma to the conditional family to prove that UMP unbiased test has form

$$\phi(\bar{X}, S) = 1(\bar{X} > K(S))$$

where $K(S)$ chosen to make conditional level α. The function $x \mapsto x/\sqrt{a - x^2}$ is increasing in x for each a so that we can rewrite ϕ in the form

$$\phi(\bar{X}, S) = \\ 1(n^{1/2}\bar{X}/\sqrt{n[S/n - \bar{X}^2]/(n-1)} > K^*(S))$$

for some K^*. The quantity

$$T = \frac{n^{1/2}\bar{X}}{\sqrt{n[S/n - \bar{X}^2]/(n-1)}}$$

is the usual t statistic and is exactly independent of S (see Theorem 6.1.5 on page 262 in Casella and Berger). This guarantees that

$$K^*(S) = t_{n-1,\alpha}$$

and makes our UMPU test the usual t test.

Optimal tests

- A good test has $\pi(\theta)$ large on the alternative and small on the null.

- For one sided one parameter families with MLR a UMP test exists.

- For two sided or multiparameter families the best to be hoped for is UMP Unbiased or Invariant or Similar.

- Good tests are found as follows:
  1. Use the NP lemma to determine a good rejection region for a simple alternative.
  2. Try to express that region in terms of a statistic whose definition does not depend on the specific alternative.
  3. If this fails impose an additional criterion such as unbiasedness. Then mimic the NP lemma and again try to simplify the rejection region.

Likelihood Ratio tests

For general composite hypotheses optimality theory is not usually successful in producing an optimal test. instead we look for heuristics to guide our choices. The simplest approach is to consider the likelihood ratio

$$\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)}$$

and choose values of $\theta_1 \in \Theta_1$ and $\theta_0 \in \Theta_0$ which are reasonable estimates of θ assuming respectively the alternative or null hypothesis is true. The simplest method is to make each θ_i a maximum likelihood estimate, but maximized only over Θ_i.

Example 1: $N(\mu, 1)$: test $\mu \le 0$ against $\mu > 0$. (Remember UMP test.) Log likelihood is

$$-n(\bar{X}-\mu)^2/2$$

If $\bar{X} > 0$ then global maximum in Θ_1 at $\bar{X}$. If $\bar{X} \le 0$ global maximum in Θ_1 at 0. Thus $\hat{\mu}_1$ which maximizes $\ell(\mu)$ subject to $\mu > 0$ is $\bar{X}$ if $\bar{X} > 0$ and 0 if $\bar{X} \le 0$. Similarly, $\hat{\mu}_0$ is $\bar{X}$ if $\bar{X} \le 0$ and 0 if $\bar{X} > 0$. Hence

$$\frac{f_{\hat{\theta}_1}(X)}{f_{\hat{\theta}_0}(X)} = \exp\{\ell(\hat{\mu}_1) - \ell(\hat{\mu}_0)\}$$

which simplifies to

$$\exp\{n\bar{X}|\bar{X}|/2\}$$

Monotone increasing function of $\bar{X}$ so rejection region will be of the form $\bar{X} > K$. To get level α reject if $n^{1/2}\bar{X} > z_\alpha$. Notice simpler statistic is *log likelihood ratio*

$$\lambda \equiv 2\log\left(\frac{f_{\hat{\mu}_1}(X)}{f_{\hat{\mu}_0}(X)}\right) = n\bar{X}|\bar{X}|$$

Example 2: In the $N(\mu,1)$ problem suppose we make the null $\mu = 0$. Then the value of $\hat\mu_0$ is simply 0 while the maximum of the log-likelihood over the alternative $\mu \neq 0$ occurs at $\bar X$. This gives

$$\lambda = n\bar X^2$$

which has a χ^2_1 distribution. This test leads to the rejection region $\lambda > (z_{\alpha/2})^2$ which is the usual UMPU test.

Example 3: For the $N(\mu,\sigma^2)$ problem testing $\mu = 0$ against $\mu \neq 0$ we must find two estimates of μ, σ^2. The maximum of the likelihood over the alternative occurs at the global mle $\bar X, \hat\sigma^2$. We find

$$\ell(\hat\mu,\hat\sigma^2) = -n/2 - n\log(\hat\sigma)$$

Maximize ℓ over null hypothesis. Recall

$$\ell(\mu,\sigma) = -\frac{1}{2\sigma^2}\sum (X_i-\mu)^2 - n\log(\sigma)$$

On null $\mu = 0$ so find $\hat\sigma_0$ by maximizing

$$\ell(0,\sigma) = -\frac{1}{2\sigma^2}\sum X_i^2 - n\log(\sigma)$$

This leads to

$$\hat\sigma_0^2 = \sum X_i^2/n$$

and

$$\ell(0,\hat\sigma_0) = -n/2 - n\log(\hat\sigma_0)$$

This gives

$$\lambda = -n\log(\hat\sigma^2/\hat\sigma_0^2)$$

Since

$$\frac{\hat\sigma^2}{\hat\sigma_0^2} = \frac{\sum(X_i-\bar X)^2}{\sum(X_i-\bar X)^2 + n\bar X^2}$$

we can write

$$\lambda = n\log(1+t^2/(n-1))$$

where

$$t = \frac{n^{1/2}\bar X}{s}$$

is the usual t statistic. Likelihood ratio test rejects for large values of $|t|$ — the usual test.

Notice that if n is large we have

$$\lambda \approx n[1 + t^2/(n-1) + O(n^{-2})] \approx t^2 .$$

Since the t statistic is approximately standard normal if n is large we see that

$$\lambda = 2[\ell(\hat{\theta}_1) - \ell(\hat{\theta}_0)]$$

has nearly a χ^2_1 distribution.

This is a general phenomenon when the null hypothesis being tested is of the form $\phi = 0$. Here is the general theory. Suppose that the vector of $p+q$ parameters θ can be partitioned into $\theta = (\phi, \gamma)$ with ϕ a vector of p parameters and γ a vector of q parameters. To test $\phi = \phi_0$ we find two mles of θ. First the global mle $\hat{\theta} = (\hat{\phi}, \hat{\gamma})$ maximizes the likelihood over $\Theta_1 = \{\theta : \phi \neq \phi_0\}$ (because typically the probability that $\hat{\phi}$ is exactly ϕ_0 is 0).

Now we maximize the likelihood over the null hypothesis, that is we find $\hat\theta_0 = (\phi_0, \hat\gamma_0)$ to maximize

$$\ell(\phi_0, \gamma)$$

The log-likelihood ratio statistic is

$$2[\ell(\hat\theta) - \ell(\hat\theta_0)]$$

Now suppose that the true value of θ is ϕ_0, γ_0 (so that the null hypothesis is true). The score function is a vector of length $p + q$ and can be partitioned as $U = (U_\phi, U_\gamma)$. The Fisher information matrix can be partitioned as

$$\left[\begin{array}{cc} \mathcal{I}_{\phi\phi} & \mathcal{I}_{\phi\gamma} \\ \mathcal{I}_{\gamma\phi} & \mathcal{I}_{\gamma\gamma} \end{array} \right].$$

According to our large sample theory for the mle we have

$$\hat{\theta} \approx \theta + \mathcal{I}^{-1}U$$

and

$$\hat{\gamma}_0 \approx \gamma_0 + \mathcal{I}_{\gamma\gamma}^{-1}U_\gamma$$

If you carry out a two term Taylor expansion of both $\ell(\hat{\theta})$ and $\ell(\hat{\theta}_0)$ around θ_0 you get

$$\ell(\hat{\theta}) \approx \ell(\theta_0) + U^t\mathcal{I}^{-1}U + \frac{1}{2}U^t\mathcal{I}^{-1}V(\theta)\mathcal{I}^{-1}U$$

where V is the second derivative matrix of ℓ. Remember that $V \approx -\mathcal{I}$ and you get

$$2[\ell(\hat{\theta}) - \ell(\theta_0)] \approx U^t\mathcal{I}^{-1}U\,.$$

A similar expansion for $\hat{\theta}_0$ gives

$$2[\ell(\hat{\theta}_0) - \ell(\theta_0)] \approx U_\gamma^t\mathcal{I}_{\gamma\gamma}^{-1}U_\gamma\,.$$

If you subtract these you find that

$$2[\ell(\widehat{\theta}) - \ell(\widehat{\theta}_0)]$$

can be written in the approximate form

$$U^t M U$$

for a suitable matrix M. It is now possible to use the general theory of the distribution of $X^t M X$ where X is $MVN(0, \Sigma)$ to demonstrate that

Theorem: The log-likelihood ratio statistic

$$\lambda = 2[\ell(\widehat{\theta}) - \ell(\widehat{\theta}_0)]$$

has, under the null hypothesis, approximately a χ_p^2 distribution.

Aside:

Theorem: Suppose $X \sim MVN(0, \Sigma)$ with Σ non-singular and M is a symmetric matrix. If $\Sigma M \Sigma M \Sigma = \Sigma M \Sigma$ then $X^t M X$ has a χ^2_ν distribution with df $\nu = trace(M\Sigma)$.

Proof: We have $X = AZ$ where $AA^t = \Sigma$ and Z is standard multivariate normal. So $X^t M X = Z^t A^t M A Z$. Let $Q = A^t M A$. Since $AA^t = \Sigma$ condition in the theorem is

$$AQQA^t = AQA^t$$

Since Σ is non-singular so is A. Multiply by A^{-1} on left and $(A^t)^{-1}$ on right; get $QQ = Q$.

Q is symmetric so $Q = P\Lambda P^t$ where Λ is diagonal matrix containing the eigenvalues of Q and P is orthogonal matrix whose columns are the corresponding orthonormal eigenvectors. So rewrite

$$Z^t Q Z = (P^t Z)^t \Lambda (PZ)\,.$$

$W = P^t Z$ is $MVN(0, P^t P = I)$; i.e. W is standard multivariate normal. Now

$$W^t \Lambda W = \sum \lambda_i W_i^2$$

We have established that the general distribution of any quadratic form $X^t M X$ is a linear combination of χ^2 variables. Now go back to the condition $QQ = Q$. If λ is an eigenvalue of Q and $v \neq 0$ is a corresponding eigenvector then $QQv = Q(\lambda v) = \lambda Q v = \lambda^2 v$ but also $QQv = Qv = \lambda v$. Thus $\lambda(1-\lambda)v = 0$. It follows that either $\lambda = 0$ or $\lambda = 1$. This means that the weights in the linear combination are all 1 or 0 and that $X^t M X$ has a χ^2 distribution with degrees of freedom, ν, equal to the number of λ_i which are equal to 1. This is the same as the sum of the λ_i so

$$\nu = trace(\Lambda)$$

But

$$\begin{aligned} trace(M\Sigma) &= trace(MAA^t) \\ &= trace(A^tMA) \\ &= trace(Q) \\ &= trace(P\Lambda P^t) \\ &= trace(\Lambda P^t P) \\ &= trace(\Lambda) \end{aligned}$$

In the application Σ is $\mathcal{I}$ the Fisher information and $M = \mathcal{I}^{-1} - J$ where

$$J = \left[\begin{array}{cc} 0 & 0 \\ 0 & \mathcal{I}_{\gamma\gamma}^{-1} \end{array}\right]$$

It is easy to check that $M\Sigma$ becomes

$$\left[\begin{array}{cc} I & 0 \\ -\mathcal{I}_{\gamma\phi}\mathcal{I}_{\phi\phi} & 0 \end{array}\right]$$

where I is a $p \times p$ identity matrix. It follows that $\Sigma M \Sigma M \Sigma = \Sigma M \Sigma$ and $trace(M\Sigma) = p$.

Confidence Sets

Defn: A level β confidence set for a parameter $\phi(\theta)$ is a random subset C, of the set of possible values of ϕ such that for each θ

$$P_\theta(\phi(\theta) \in C) \ge \beta$$

Confidence sets are very closely connected with hypothesis tests:

From confidence sets to tests

Suppose C is a level $\beta = 1 - \alpha$ confidence set for ϕ.

To test $\phi = \phi_0$: reject if $\phi \notin C$. This test has level α.

From tests to confidence sets

Conversely, suppose that for each ϕ_0 we have available a level α test of $\phi = \phi_0$ who rejection region is say R_{ϕ_0}.

Define $C = \{\phi_0 : \phi = \phi_0 \text{ is not rejected}\}$; get level $1 - \alpha$ confidence set for ϕ.

Example: Usual t test gives rise in this way to the usual t confidence intervals

$$\bar{X} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}.$$

Confidence sets from Pivots

Definition: A **pivot** (pivotal quantity) is a function $g(\theta, X)$ whose distribution is the same for all θ. (θ in pivot is same θ as being used to calculate distribution of $g(\theta, X)$.

Using pivots to generate confidence sets:

Pick a set A in space of possible values for g.

Let $\beta = P_\theta(g(\theta, X) \in A)$; since g is pivotal β is the same for all θ.

Given data X solve the relation

$$g(\theta, X) \in A$$

to get

$$\theta \in C(X, A)\,.$$

Example: $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ is a pivot in the $N(\mu,\sigma^2)$ model.

Given $\beta = 1-\alpha$ consider the two points

$$\chi^2_{n-1,1-\alpha/2} \text{ and } \chi^2_{n-1,\alpha/2}.$$

Then

$$P(\chi^2_{n-1,1-\alpha/2} \le (n-1)s^2/\sigma^2 \le \chi^2_{n-1,\alpha/2}) = \beta$$

for all μ, σ.

Solve this relation:

$$P(\frac{(n-1)^{1/2}s}{\chi_{n-1,\alpha/2}} \le \sigma \le \frac{(n-1)^{1/2}s}{\chi_{n-1,1-\alpha/2}}) = \beta$$

so interval

$$\left[\frac{(n-1)^{1/2}s}{\chi_{n-1,\alpha/2}}, \frac{(n-1)^{1/2}s}{\chi_{n-1,1-\alpha/2}}\right]$$

is a level $1-\alpha$ confidence interval.

In the same model we also have

$$P(\chi^2_{n-1,1-\alpha} \le (n-1)s^2/\sigma^2) = \beta$$

which can be solved to get

$$P(\sigma \le \frac{(n-1)^{1/2}s}{\chi_{n-1,1-\alpha}}) = \beta$$

This gives a level $1-\alpha$ interval

$$(0, (n-1)^{1/2}s/\chi_{n-1,1-\alpha}) \, .$$

The right hand end of this interval is usually called a confidence upper bound.

In general the interval from

$$(n-1)^{1/2}s/\chi_{n-1,\alpha_1} \text{ to } (n-1)^{1/2}s/\chi_{n-1,1-\alpha_2}$$

has level $\beta = 1-\alpha_1-\alpha_2$. For fixed β can minimize length of resulting interval numerically — rarely used. See homework for an example.

Decision Theory and Bayesian Methods

Example: Decide between 4 modes of transportation to work:

- B = Ride my bike.
- C = Take the car.
- T = Use public transit.
- H = Stay home.

Costs depend on weather: R = Rain or S = Sun.

Ingredients of Decision Problem: No data case.

- Decision space $D = \{B, C, T, H\}$ of possible actions.

- Parameter space $\Theta = \{R, S\}$ of possible "states of nature".

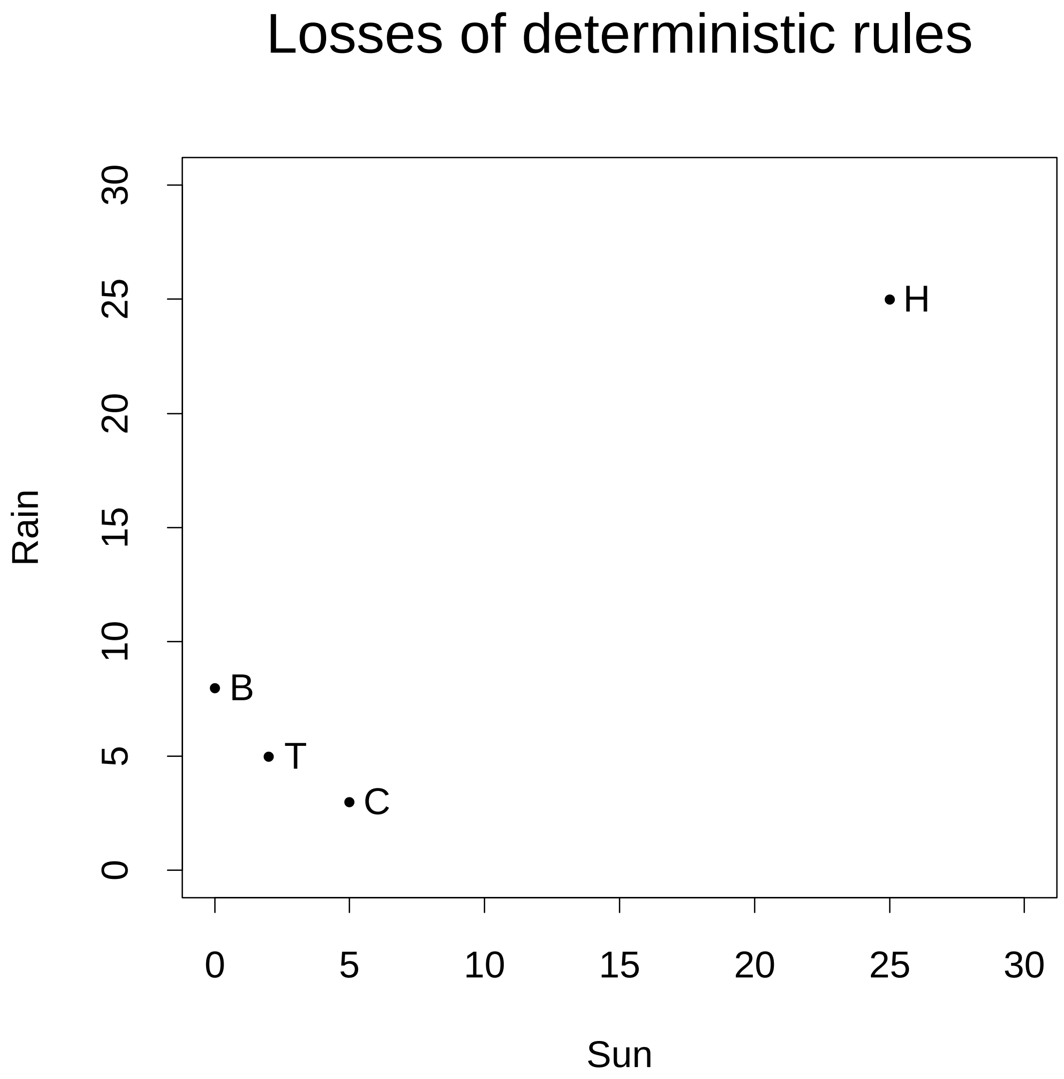
- Loss function $L = L(d, \theta)$ loss incurred if do d and θ is true state of nature.

In the example we might use the following table for L:

	C	B	T	H
R	3	8	5	25
S	5	0	2	25

Notice that if it rains I will be glad if I drove. If it is sunny I will be glad if I rode my bike. In any case staying at home is expensive.

In general we study this problem by comparing various functions of θ. In this problem a function of θ has only two values, one for rain and one for sun and we can plot any such function as a point in the plane. We do so to indicate the geometry of the problem before stating the general theory.


Losses of deterministic rules
Rain
Sun
0
5
10
15
20
25
30
0
5
10
15
20
25
30
H
B
T
C

Statistical Decision Theory

Statistical problems have another ingredient, the data. We observe X a random variable taking values in say $\mathcal{X}$. We may make our decision d depend on X. A **decision rule** is a function $\delta(X)$ from $\mathcal{X}$ to D. We will want $L(\delta(X), \theta)$ to be small for all θ. Since X is random we quantify this by averaging over X and compare procedures δ in terms of the **risk function**

$$R_\delta(\theta) = E_\theta(L(\delta(X), \theta))$$

To compare two procedures we must compare two functions of θ and pick "the smaller one". But typically the two functions will cross each other and there won't be a unique 'smaller one'.

Example: In estimation theory to estimate a real parameter θ we used $D = \Theta$,

$$L(d, \theta) = (d - \theta)^2$$

and find that the risk of an estimator $\hat{\theta}(X)$ is

$$R_{\hat{\theta}}(\theta) = E[(\hat{\theta} - \theta)^2]$$

which is just the Mean Squared Error of $\hat{\theta}$. We have already seen that there is no unique best estimator in the sense of MSE. How do we compare risk functions in general?

- **Minimax methods** choose δ to minimize the worst case risk:

  $$\text{sup}\{R_\delta(\theta);\theta \in \Theta)\}.$$

  We call δ^* minimax if

  $$\sup_\theta R_{\delta^*}(\theta) = \inf_\delta \sup_\theta R_\delta(\theta)$$

  Usually the sup and inf are achieved and we write max for sup and min for inf. This is the source of "minimax".

- **Bayes methods** choose δ to minimize an average

  $$r_\pi(\delta) = \int R_\delta(\theta)\pi(\theta)d\theta$$

  for a suitable density π. We call π a **prior** density and r the **Bayes** risk of δ for the prior π.

Example: Transport problem has no data so the only possible (non-randomized) decisions are the four possible actions B, C, T, H. For B and T the worst case is rain. For the other two actions Rain and Sun are equivalent. We have the following table:

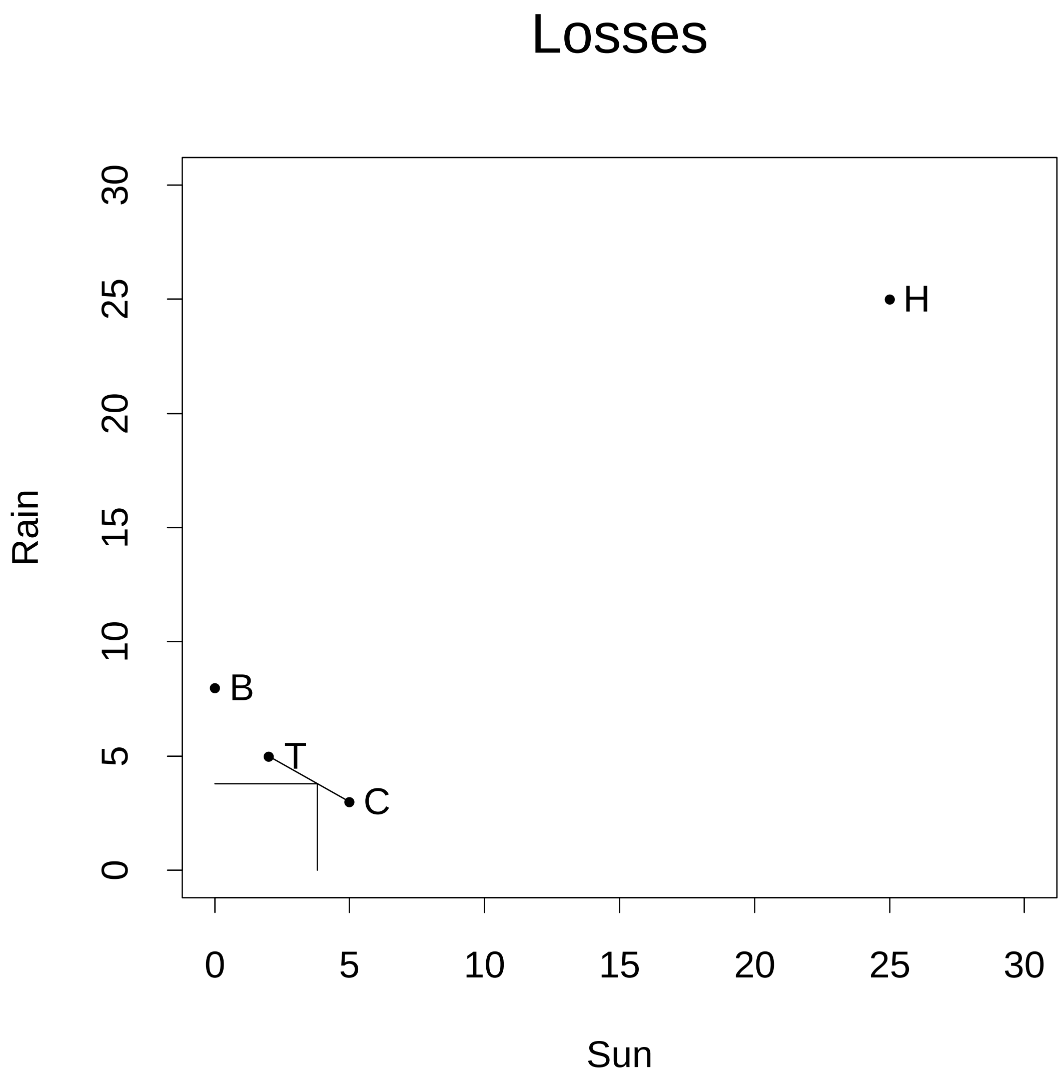
	C	B	T	H
R	3	8	5	25
S	5	0	2	25
Maximum	5	8	5	25

Smallest maximum: take car, or transit.

Minimax action: take car or public transit.

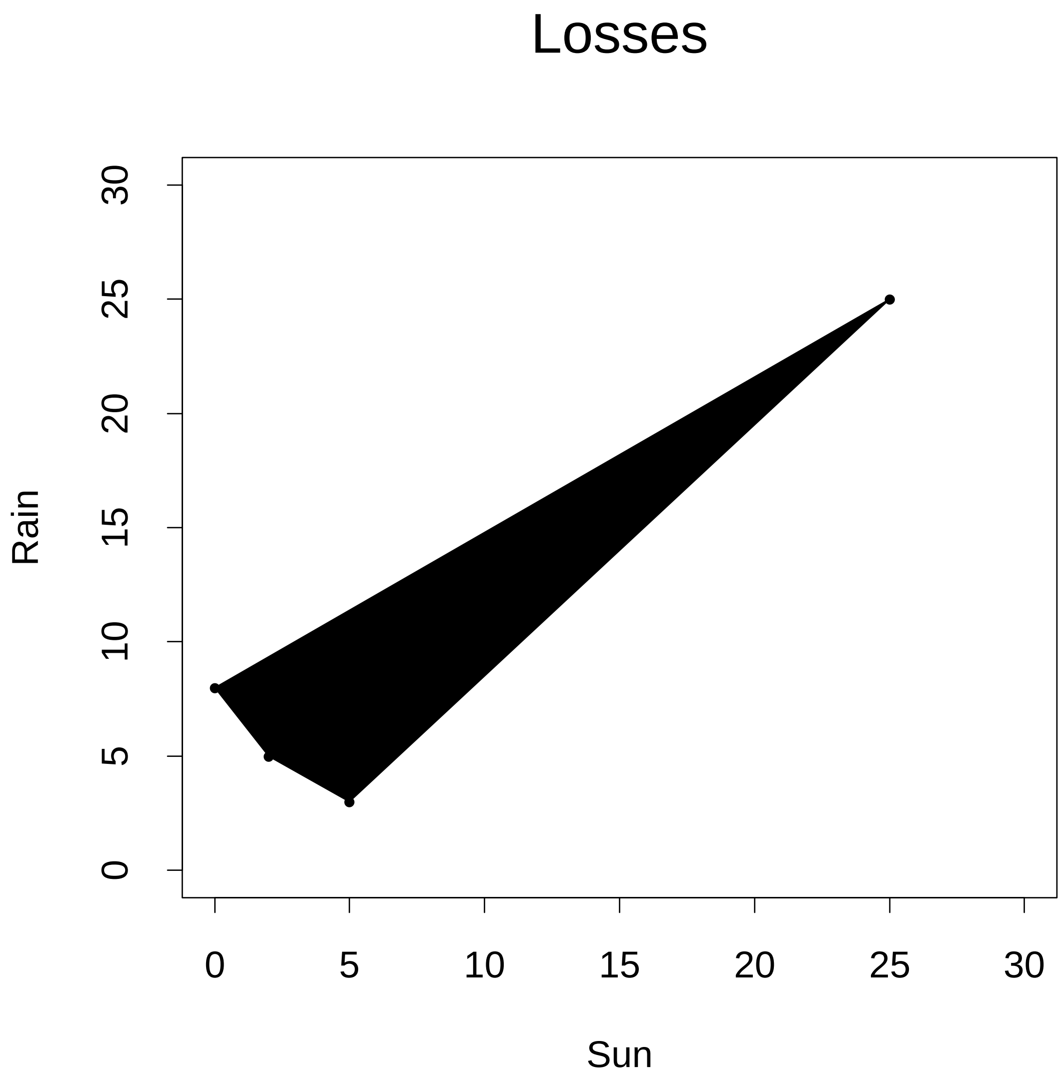
Now imagine: toss coin with probability λ of getting Heads, take my car if Heads, otherwise take transit. Long run average daily loss would be $3\lambda+5(1-\lambda)$ when it rains and $5\lambda+2(1-\lambda)$ when it is Sunny. Call this procedure d_λ; add it to graph for each value of λ. Varying λ from 0 to 1 gives a straight line running from $(3,5)$ to $(5,2)$. The two losses are equal when $\lambda = 3/5$. For smaller λ worst case risk is for sun; for larger λ worst case risk is for rain.

Added to graph: loss functions for each d_λ, (straight line) and set of (x,y) pairs for which $\min(x,y) = 3.8$ — worst case risk for d_λ when $\lambda = 3/5$.


Losses
Rain
Sun
0
5
10
15
20
25
30
0
5
10
15
20
25
30
H
B
T
C

The figure then shows that $d_{3/5}$ is actually the minimax procedure when randomized procedures are permitted.

In general we might consider using a 4 sided coin where we took action B with probability λ_B, C with probability λ_C and so on. The loss function of such a procedure is a convex combination of the losses of the four basic procedures making the set of risks achievable with the aid of randomization look like the following:


Losses
Rain
Sun
0
5
10
15
20
25
30
0
5
10
15
20
25
30

Randomization in decision problems permits assumption that set of possible risk functions is convex — an important technical conclusion used to prove many basic decision theory results.

Graph shows many points in the picture correspond to bad decision procedures. Rain or not taking my car to work has a lower loss than staying home; the decision to stay home is inadmissible.

Definition: A decision rule δ is **inadmissible** if there is a rule δ^* such that

$$R_{\delta^*}(\theta) \le R_\delta(\theta)$$

for all θ and there is at least one value of θ where the inequality is strict. A rule which is not inadmissible is called **admissible**.

Admissible procedures have risks on lower left of graphs, i.e., lines connecting B to T and T to C are the admissible procedures.

Connection between Bayes procedures and admissible procedures:

Prior distribution in example specified by two probabilities, π_S and π_R which add up to 1.

If $L = (L_S, L_R)$ is the risk function for some procedure then the Bayes risk is

$$r_\pi = \pi_R L_R + \pi_S L_S.$$

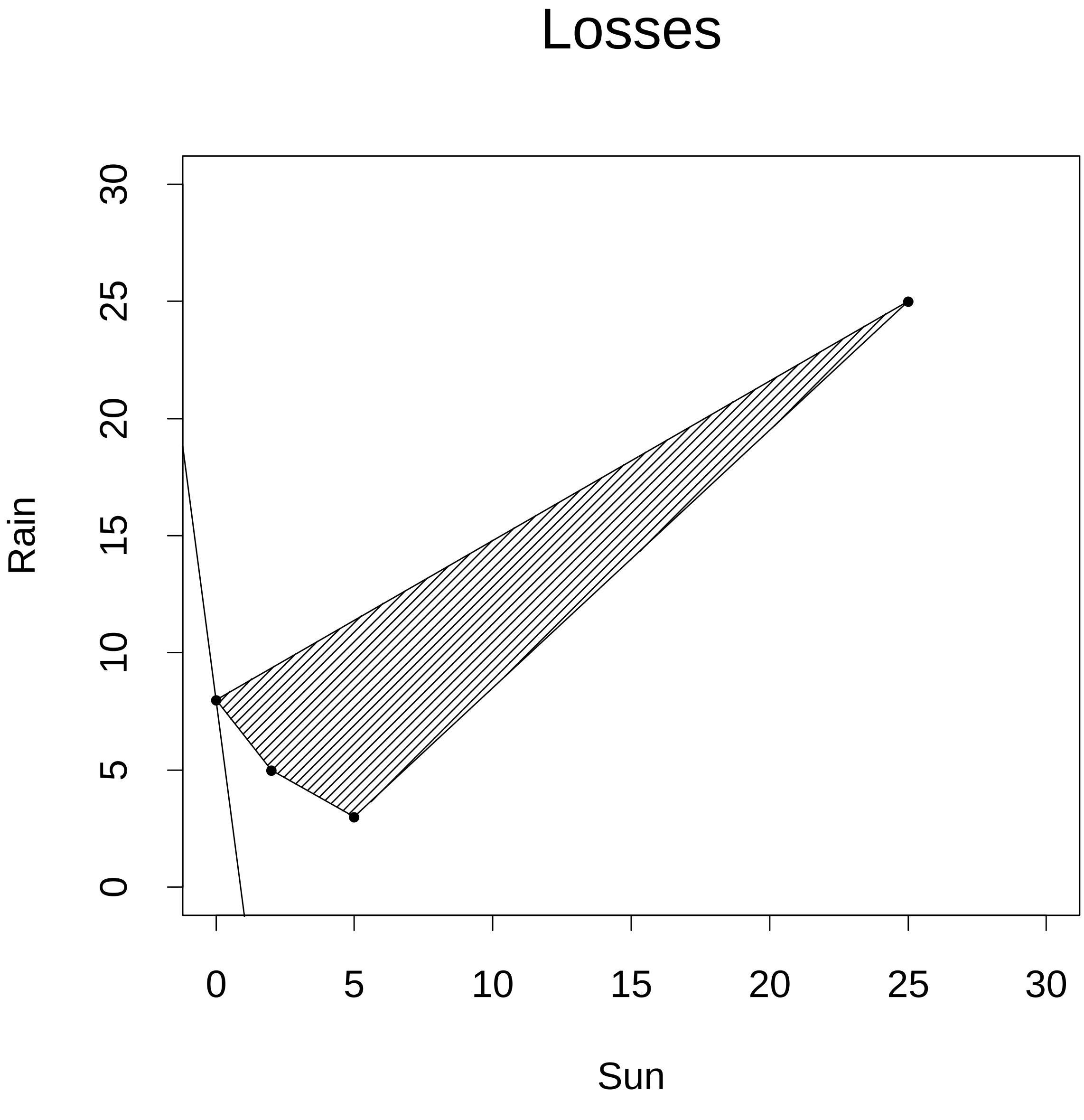
Consider set of L such that this Bayes risk is equal to some constant.

On picture this is line with slope $-\pi_S/\pi_R$.

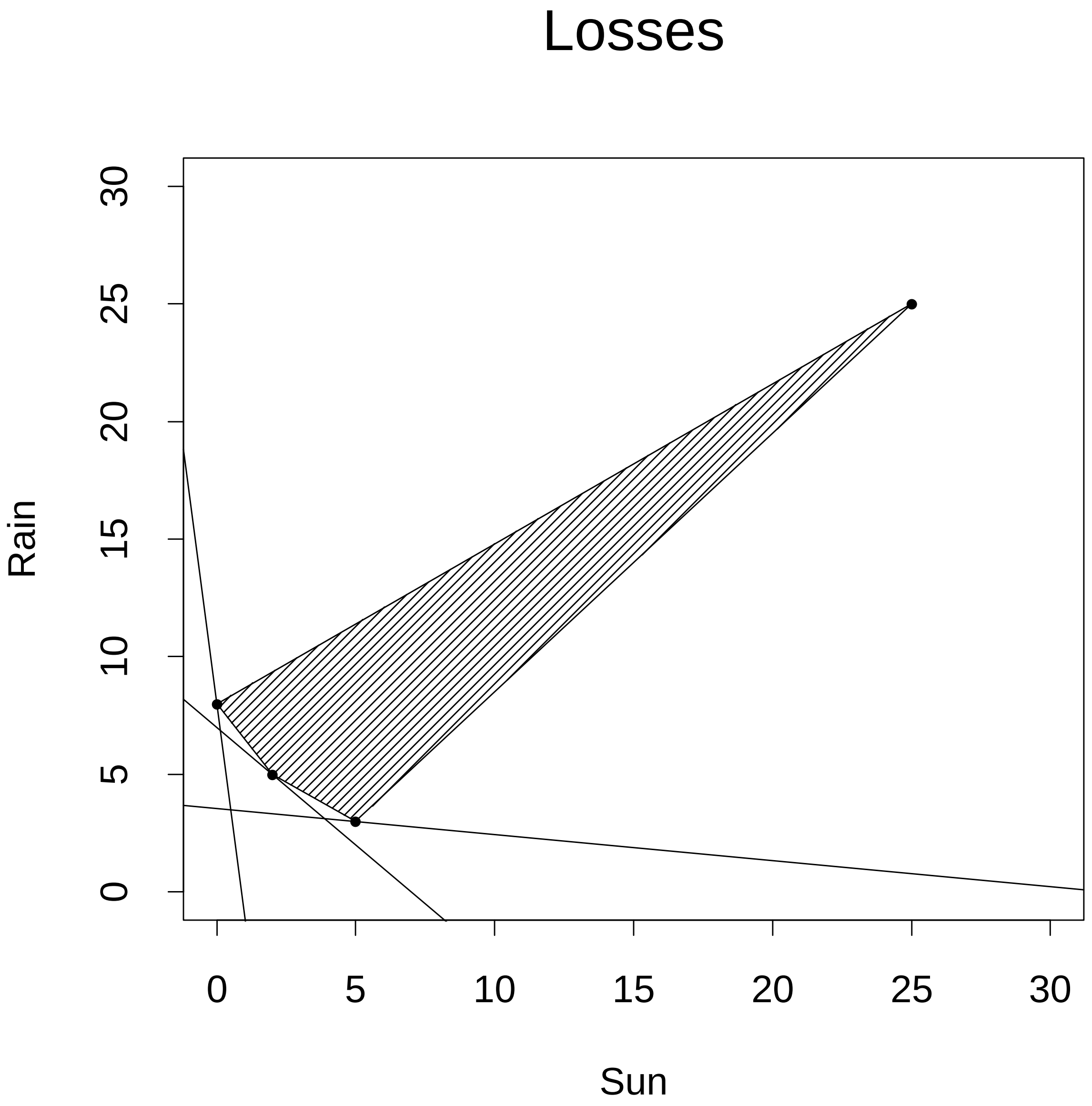
Consider three priors: $\pi_1 = (0.9, 0.1)$, $\pi_2 = (0.5, 0.5)$ and $\pi_3 = (0.1, 0.9)$.

For π_1: imagine a line with slope -9 =0.9/0.1 starting on the far left of the picture and sliding right until it bumps into the convex set of possible losses in the previous picture. It does so at point B as shown in the next graph.

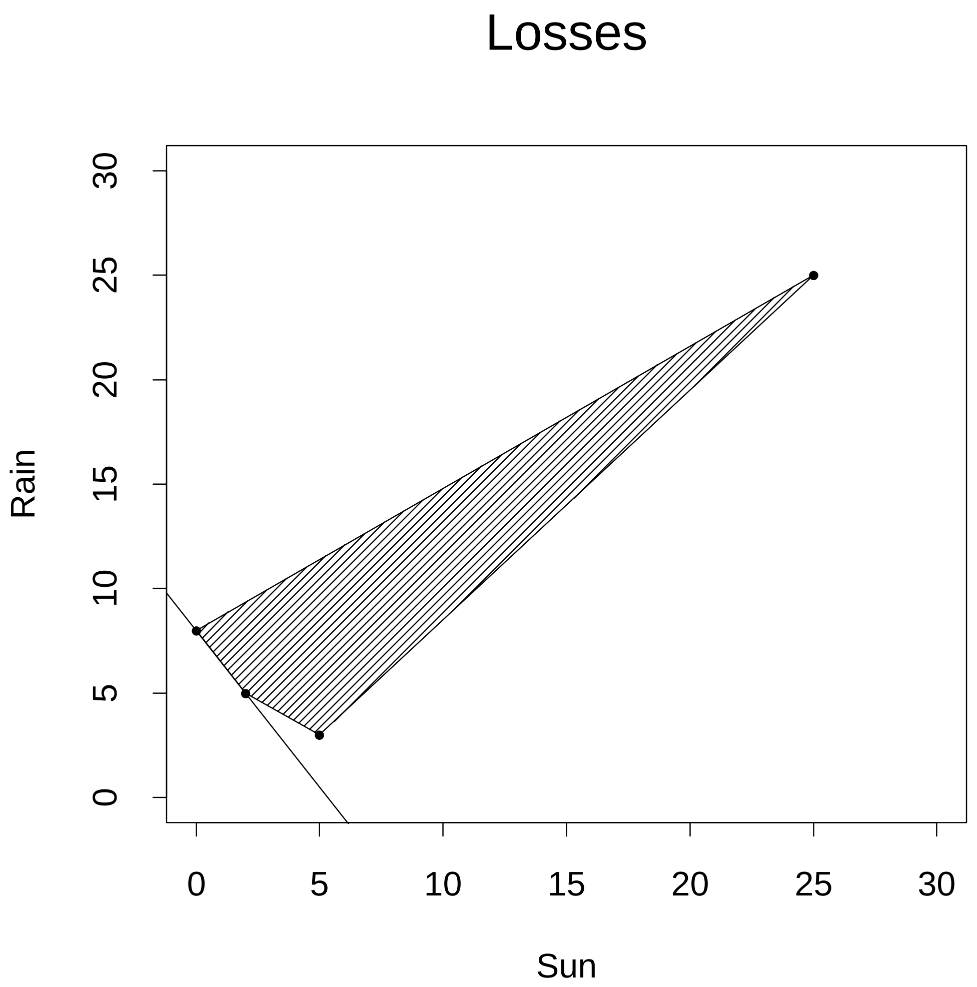
Sliding this line to the right corresponds to making r_π larger and larger so that when it just touches the convex set we have found the Bayes procedure.


Losses
Rain
Sun
0
5
10
15
20
25
30
0
5
10
15
20
25
30

Here is a picture showing the same lines for the three priors above.



Bayes procedure for π_1 (you're pretty sure it will be sunny) is to ride your bike. If it's a toss up between R and S you take the bus. If R is very likely you take your car. Prior $(0.6, 0.4)$ produces the line shown here:



Any point on line BT is Bayes for this prior.

Decision Theory and Bayesian Methods Summary for no data case

- Decision space is the set of possible actions I might take. We assume that it is convex, typically by expanding a basic decision space D to the space $\mathcal{D}$ of all probability distributions on D.

- Parameter space Θ of possible "states of nature".

- Loss function $L = L(d, \theta)$ which is the loss I incur if I do d and θ is the true state of nature.

- We call δ^* minimax if
$$\max_\theta L(\delta^*, \theta) = \min_\delta \max_\theta L(\delta, \theta)\,.$$

- A **prior** is a probability distribution π on Θ,.

- The Bayes risk of a decision δ for a prior π is

  $$r_\pi(\delta) = E_\pi(L(\delta,\theta)) = \int L(\delta,\theta)\pi(\theta)d\theta$$

  if the prior has a density. For finite parameter spaces Θ the integral is a sum.

- A decision δ^* is Bayes for a prior π if

  $$r_\pi(\delta^*) \le r_\pi(\delta)$$

  for any decision δ.

- For infinite parameter spaces: $\pi(\theta) > 0$ on Θ is a proper prior if $\int \pi(\theta) d\theta < \infty$; divide π by integral to get a density. If $\int \pi(\theta) d\theta = \infty$ π is an **improper** prior density.

- Decision δ is **inadmissible** if there is δ^* such that

  $$L(\delta^*, \theta) \le L(\delta, \theta)$$

  for all θ and there is at least one value of θ where the inequality is strict. A decision which is not inadmissible is called **admissible**.

- Every admissible procedure is Bayes, perhaps only for an improper prior. (Proof uses the Separating Hyperplane Theorem in Functional Analysis.)

- Every Bayes procedure with finite Bayes risk (for prior with density > 0 for all θ) is admissible.

  Proof: If δ is Bayes for π but not admissible there is a δ^* such that

  $$L(\delta^*, \theta) \le L(\delta, \theta)$$

  Multiply by the prior density; integrate:

  $$r_\pi(\delta^*) \le r_\pi(\delta)$$

  If there is a θ for which the inequality involving L is strict and if the density of π is positive at that θ then the inequality for r_π is strict which would contradict the hypothesis that δ is Bayes for π.

  Notice: theorem actually requires the extra hypotheses: positive density, and risk functions of δ and δ^* continuous.

- A minimax procedure is admissible. (Actually there can be several minimax procedures and the claim is that at least one of them is admissible. When the parameter space is infinite it might happen that set of possible risk functions is not closed; if not then we have to replace the notion of admissible by some notion of nearly admissible.)

- The minimax procedure has constant risk. Actually the admissible minimax procedure is Bayes for some π and its risk is constant on the set of θ for which the prior density is positive.

Decision Theory and Bayesian Methods Summary when there is data

- Decision space is the set of possible actions I might take. We assume that it is convex, typically by expanding a basic decision space D to the space $\mathcal{D}$ of all probability distributions on D.

- Parameter space Θ of possible "states of nature".

- Loss function $L = L(d, \theta)$: loss I incur if I do d and θ is true state of nature.

- Add data $X \in \mathcal{X}$ with model $\{P_\theta; \theta \in \Theta\}$: model density is $f(x|\theta)$.

- A *procedure* is a map $\delta : \mathcal{X} \mapsto \mathcal{D}$.

- The risk function for δ is the expected loss:

$$R_\delta(\theta) = R(\delta, \theta) = \mathsf{E}\left[L\{\delta(X), \theta\}\right].$$

- We call δ^* minimax if

$$\max_\theta R(\delta^*, \theta) = \min_\delta \max_\theta R(\delta, \theta)\,.$$

- A **prior** is a probability distribution π on Θ,.

- **Bayes risk** of decision δ for prior π is

$$\begin{aligned} r_\pi(\delta) &= E_\pi(R(\delta, \theta)) \\ &= \int L(\delta(x), \theta) f(x|\theta)\pi(\theta) dx d\theta \end{aligned}$$

  if the prior has a density. For finite parameter spaces Θ the integral is a sum.

- A decision δ^* is Bayes for a prior π if

$$r_\pi(\delta^*) \le r_\pi(\delta)$$

  for any decision δ.

- For infinite parameter spaces: $\pi(\theta) > 0$ on Θ is a **proper** prior if $\int \pi(\theta) d\theta < \infty$; divide π by integral to get a density. If $\int \pi(\theta) d\theta = \infty$ π is an **improper** prior density.

- Decision δ is **inadmissible** if there is δ^* such that

$$R(\delta^*, \theta) \le R(\delta, \theta)$$

  for all θ and there is at least one value of θ where the inequality is strict. A decision which is not inadmissible is called **admissible**.

- Every admissible procedure is Bayes, perhaps only for an improper prior.

- If every risk function is continuous then every Bayes procedure with finite Bayes risk (for prior with density > 0 for all θ) is admissible.

- A minimax procedure is admissible.

- The minimax procedure has constant risk. The admissible minimax procedure is Bayes for some π; its risk is constant on the set of θ for which the prior density is positive.

Bayesian estimation

Focus on problem of estimation of 1 dimensional parameter.

Mean Squared Error corresponds to using

$$L(d,\theta) = (d-\theta)^2 \,.$$

Risk function of procedure (estimator) $\widehat{\theta}$ is

$$R_{\widehat{\theta}}(\theta) = E_\theta[(\widehat{\theta}-\theta)^2]$$

Now consider prior with density $\pi(\theta)$.

Bayes risk of $\widehat{\theta}$ is

$$\begin{aligned} r_\pi &= \int R_{\widehat{\theta}}(\theta)\pi(\theta)d\theta \\ &= \int\int (\widehat{\theta}(x)-\theta)^2 f(x;\theta)\pi(\theta)dxd\theta \end{aligned}$$

Choose $\hat{\theta}$ to minimize r_π?

Recognize that $f(x;\theta)\pi(\theta)$ is really a joint density

$$\int\int f(x;\theta)\pi(\theta)dxd\theta = 1$$

For this joint density: conditional density of X given θ is just the model $f(x;\theta)$.

Justifies notation $f(x|\theta)$.

Compute r_π a different way by factoring the joint density a different way:

$$f(x|\theta)\pi(\theta) = \pi(\theta|x)f(x)$$

where now $f(x)$ is the marginal density of x and $\pi(\theta|x)$ denotes the conditional density of θ given X.

Call $\pi(\theta|x)$ the **posterior density.**

Found via Bayes theorem (which is why this is Bayesian statistics):

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\phi)\pi(\phi)d\phi}$$

With this notation we can write

$$r_\pi(\hat\theta) = \int \left[\int (\hat\theta(x)-\theta)^2 \pi(\theta|x)d\theta\right] f(x)dx$$

Can choose $\hat\theta(x)$ separately for each x to minimize the quantity in square brackets (as in the NP lemma).

Quantity in square brackets is a quadratic function of $\hat\theta(x)$; minimized by

$$\hat\theta(x) = \int \theta\pi(\theta|x)d\theta$$

which is

$$E(\theta|X)$$

and is called the **posterior expected mean** of θ.

Example: estimating normal mean μ.

Imagine, for example that μ is the true speed of sound.

I think this is around 330 metres per second and am pretty sure that I am within 30 metres per second of the truth with that guess.

I might summarize my opinion by saying that I think μ has a normal distribution with mean ν =330 and standard deviation $\tau = 10$.

That is, I take a prior density π for μ to be $N(\nu, \tau^2)$.

Before I make any measurements best guess of μ minimizes

$$\int (\hat{\mu} - \mu)^2 \frac{1}{\tau\sqrt{2\pi}} \exp\{-(\mu - \nu)^2/(2\tau^2)\} d\mu$$

This quantity is minimized by the prior mean of μ, namely,

$$\hat{\mu} = E_\pi(\mu) = \int \mu\pi(\mu) d\mu = \nu\,.$$

Now collect 25 measurements of the speed of sound.

Assume: relationship between the measurements and μ is that the measurements are unbiased and that the standard deviation of the measurement errors is $\sigma = 15$ which I assume that we know.

So model is: given μ, $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$.

The joint density of the data and μ is then

$$(2\pi)^{-n/1}\sigma^{-n}\exp\{-\sum(X_i-\mu)^2/(2\sigma^2)\} \\ \times \\ (2\pi)^{-1/2}\tau^{-1}\exp\{-(\mu-\nu)^2/\tau^2\}.$$

Thus $(X_1, \ldots, X_n, \mu) \sim MVN$. Conditional distribution of θ given $X_1, \ldots, X_n$ is normal.

Use standard MVN formulas to calculate conditional means and variances.

Alternatively: exponent in joint density has form

$$-\frac{1}{2}\left[\mu^2/\gamma^2 - 2\mu\psi/\gamma^2\right]$$

plus terms not involving μ where

$$\frac{1}{\gamma^2} = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)$$

and

$$\frac{\psi}{\gamma^2} = \frac{\sum X_i}{\sigma^2} + \frac{\nu}{\tau^2}$$

So: conditional of μ given data is $N(\psi, \gamma^2)$.

In other words the posterior mean of μ is

$$\frac{\frac{n}{\sigma^2}\bar{X} + \frac{1}{\tau^2}\nu}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

which is a weighted average of the prior mean ν and the sample mean $\bar{X}$.

Notice: weight on data is large when n is large or σ is small (precise measurements) and small when τ is small (precise prior opinion).

Improper priors: When the density does not integrate to 1 we can still follow the machinery of Bayes' formula to derive a posterior.

Example: $N(\mu,\sigma^2)$; consider prior density

$$\pi(\mu) \equiv 1.$$

This "density" integrates to ∞; using Bayes' theorem to compute the posterior would give

$$\pi(\mu|X) = \frac{(2\pi)^{-n/2}\sigma^{-n}\exp\{-\sum(X_i-\mu)^2/(2\sigma^2)\}}{\int(2\pi)^{-n/2}\sigma^{-n}\exp\{-\sum(X_i-\nu)^2/(2\sigma^2)\}d\nu}$$

It is easy to see that this cancels to the limit of the case previously done when $\tau \to \infty$ giving a $N(\bar{X},\sigma^2/n)$ density.

I.e., Bayes estimate of μ for this improper prior is $\bar{X}$.

Admissibility: Bayes procedures corresponding to proper priors are admissible. It follows that for each $w \in (0,1)$ and each real ν the estimate

$$w\bar{X} + (1-w)\nu$$

is admissible. That this is also true for $w = 1$, that is, that $\bar{X}$ is admissible is much harder to prove.

Minimax estimation: The risk function of $\bar{X}$ is simply σ^2/n. That is, the risk function is constant since it does not depend on μ. Were $\bar{X}$ Bayes for a proper prior this would prove that $\bar{X}$ is minimax. In fact this is also true but hard to prove.

Example: Given p, X has a Binomial(n, p) distribution.

Give p a Beta(α, β) prior density

$$\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

The joint "density" of X and p is

$$\binom{n}{X} p^X (1-p)^{n-X} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1};$$

posterior density of p given X is of the form

$$c p^{X+\alpha-1}(1-p)^{n-X+\beta-1}$$

for a suitable normalizing constant c.

This is Beta$(X+\alpha, n-X+\beta)$ density.

Mean of Beta(α, β) distribution is $\alpha/(\alpha+\beta)$.

So Bayes estimate of p is

$$\frac{X+\alpha}{n+\alpha+\beta} = w\hat{p} + (1-w)\frac{\alpha}{\alpha+\beta}$$

where $\hat{p} = X/n$ is the usual mle.

Notice: again weighted average of prior mean and mle.

Notice: prior is proper for $\alpha > 0$ and $\beta > 0$.

To get $w = 1$ take $\alpha = \beta = 0$; use improper prior

$$\frac{1}{p(1-p)}$$

Again: each $w\hat{p} + (1-w)p_o$ is admissible for $w \in (0,1)$.

Again: it is true that $\hat{p}$ is admissible but our theorem is not adequate to prove this fact.

The risk function of $w\hat{p} + (1-w)p_0$ is

$$R(p) = E[(w\hat{p} + (1-w)p_0 - p)^2]$$

which is

$$w^2\text{Var}(\hat{p}) + (wp + (1-w)p - p)^2$$
$$=$$
$$w^2 p(1-p)/n + (1-w)^2(p-p_0)^2$$

Risk function constant if coefficients of p^2 and p in risk are 0.

Coefficient of p^2 is

$$-w^2/n + (1-w)^2$$

so $w = n^{1/2}/(1+n^{1/2})$.

Coefficient of p is then

$$w^2/n - 2p_0(1-w)^2$$

which vanishes if $2p_0 = 1$ or $p_0 = 1/2$.

Working backwards: to get these values for w and p_0 require $\alpha = \beta$. Moreover

$$w^2/(1-w)^2 = n$$

gives

$$n/(\alpha+\beta) = \sqrt{n}$$

or $\alpha = \beta = \sqrt{n}/2$. Minimax estimate of p is

$$\frac{\sqrt{n}}{1+\sqrt{n}}\hat{p} + \frac{1}{1+\sqrt{n}}\frac{1}{2}$$

Example: $X_1, \ldots, X_n$ iid $MVN(\mu, \Sigma)$ with Σ known.

Take improper prior for μ which is constant.

Posterior of μ given X is then $MVN(\bar{X}, \Sigma/n)$.

Multivariate estimation: common to extend the notion of squared error loss by defining

$$L(\hat\theta,\theta)=\sum(\hat\theta_i-\theta_i)^2=(\hat\theta-\theta)^t(\hat\theta-\theta)\,.$$

For this loss risk is sum of MSEs of individual components.

Bayes estimate is again posterior mean. Thus $\bar X$ is Bayes for an improper prior in this problem.

It turns out that $\bar X$ is minimax; its risk function is the constant $trace(\Sigma)/n$.

If the dimension p of θ is 1 or 2 then $\bar X$ is also admissible but if $p \ge 3$ then it is inadmissible.

Fact first demonstrated by James and Stein who produced an estimate which is better, in terms of this risk function, for every μ.

So-called **James Stein** estimator is essentially never used.

Hypothesis Testing and Decision Theory

Decision analysis of hypothesis testing takes $D = \{0, 1\}$ and

$$L(d, \theta) = 1(\text{make an error})$$

or more generally $L(0, \theta) = \ell_1 1(\theta \in \Theta_1)$ and $L(1, \theta) = \ell_2 1(\theta \in \Theta_0)$ for two positive constants ℓ_1 and ℓ_2. We make the decision space convex by allowing a decision to be a probability measure on D. Any such measure can be specified by $\delta = P(\text{reject})$ so $\mathcal{D} = [0, 1]$. The loss function of $\delta \in [0, 1]$ is

$$L(\delta, \theta) = (1 - \delta)\ell_1 1(\theta \in \Theta_1) + \delta \ell_0 1(\theta \in \Theta_0) \, .$$

Simple hypotheses: Prior is $\pi_0 > 0$ and $\pi_1 > 0$ with $\pi_0 + \pi_1 = 1$.

Procedure: map from sample space to $\mathcal{D}$ – a test function.

Risk function of procedure $\phi(X)$ is a pair of numbers:

$$R_\phi(\theta_0) = E_0(L(\delta, \theta_0))$$

and

$$R_\phi(\theta_1) = E_1(L(\delta, \theta_1))$$

We find

$$R_\phi(\theta_0) = \ell_0 E_0(\phi(X)) = \ell_0 \alpha$$

and

$$R_\phi(\theta_1) = \ell_1 E_1(1 - \phi(X)) = \ell_1 \beta$$

The Bayes risk of ϕ is

$$\pi_0\ell_0\alpha + \pi_1\ell_1\beta$$

We saw in the hypothesis testing section that this is minimized by

$$\phi(X) = 1(f_1(X)/f_0(X) > \pi_0\ell_0/(\pi_1\ell_1))$$

which is a likelihood ratio test. These tests are Bayes and admissible. The risk is constant if $\beta\ell_1 = \alpha\ell_0$; you can use this to find the minimax test in this context.

Hypothesis Testing and Decision Theory

Decision analysis of hypothesis testing takes $D = \{0, 1\}$ and

$$L(d, \theta) = 1(\text{make an error})$$

or more generally $L(0, \theta) = \ell_1 1(\theta \in \Theta_1)$ and $L(1, \theta) = \ell_2 1(\theta \in \Theta_0)$ for two positive constants ℓ_1 and ℓ_2. We make the decision space convex by allowing a decision to be a probability measure on D. Any such measure can be specified by $\delta = P(\text{reject})$ so $\mathcal{D} = [0, 1]$. The loss function of $\delta \in [0, 1]$ is

$$L(\delta, \theta) = (1 - \delta)\ell_1 1(\theta \in \Theta_1) + \delta \ell_0 1(\theta \in \Theta_0)\,.$$

Simple hypotheses: Prior is $\pi_0 > 0$ and $\pi_1 > 0$ with $\pi_0 + \pi_1 = 1$.

Procedure: map from sample space to $\mathcal{D}$ – a test function.

Risk function of procedure $\phi(X)$ is a pair of numbers:

$$R_\phi(\theta_0) = E_0(L(\delta, \theta_0))$$

and

$$R_\phi(\theta_1) = E_1(L(\delta, \theta_1))$$

We find

$$R_\phi(\theta_0) = \ell_0 E_0(\phi(X)) = \ell_0 \alpha$$

and

$$R_\phi(\theta_1) = \ell_1 E_1(1 - \phi(X)) = \ell_1 \beta$$

The Bayes risk of ϕ is

$$\pi_0 \ell_0 \alpha + \pi_1 \ell_1 \beta$$

We saw in the hypothesis testing section that this is minimized by

$$\phi(X) = 1(f_1(X)/f_0(X) > \pi_0\ell_0/(\pi_1\ell_1))$$

which is a likelihood ratio test. These tests are Bayes and admissible. The risk is constant if $\beta\ell_1 = \alpha\ell_0$; you can use this to find the minimax test in this context.