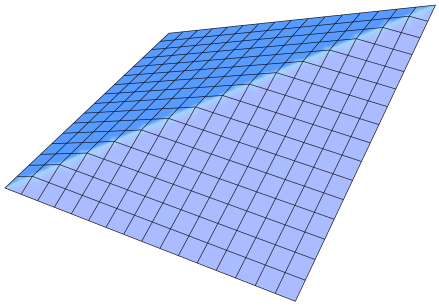


STAT 241: Statistics (Advanced Level)

Prof. Mu Zhu • Winter 2014 • University of Waterloo
TR 11:30AM–12:50PM in MC 4007

Transcribed, LaTeXed, and edited by Rob Zimmerman

Note: These are personal lecture notes, not official course materials. They may contain errors (which by default you should attribute to me, Rob Zimmerman, rather than the course instructor). The permanent home of these — and all of my other lecture notes — is https://rob-zimmerman.github.io/coursenotes.

Contents

Lecture 1 — Tuesday, January 07, 2014

1. Probability Review and Point Estimation

In a probability problem, the model $X \sim f(x;\theta)$ and its parameter θ are usually given. We might then calculate $\mathbb{E}[X]$, $\mathrm{Var}(X)$, or $\mathbb{P}(a \le X \le b)$. Statistics runs this logic in reverse: we observe $X_1, X_2, \dots, X_n$ from the model and use the data to learn the unknown parameter θ.

There are several forms of **statistical inference**, which we study in the first half of the course. First, **point estimation** seeks a function $\hat{\theta}(X_1, X_2, \dots, X_n)$, called an estimator, that approximates the unknown value θ. We ask how to construct such a function and how to compare competing estimators. Second, **interval estimation** seeks lower and upper bounds $\ell(X_1, X_2, \dots, X_n)$ and $u(X_1, X_2, \dots, X_n)$, typically designed so that

$$\mathbb{P}_\theta\big(\ell(X_1, \dots, X_n) \le \theta \le u(X_1, \dots, X_n)\big) \ge 1 - \alpha$$

for every admissible θ, where $\alpha > 0$ is small. Exact procedures attain equality; conservative or discrete procedures may have slightly greater coverage. Third, **hypothesis testing** compares claims such as $H_0 : \theta = \theta_0$ and $H_A : \theta = \theta_A$: which is more plausible given the observations $X_1, \dots, X_n$? These three topics make up the first half of the course.

The second half of the course develops several important **applied models**. We begin with the classical normal model $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, first with σ^2 known and then with σ^2 unknown. We next study the **two-sample case**, with independent samples from $\mathcal{N}(\mu_A, \sigma^2)$ and $\mathcal{N}(\mu_B, \sigma^2)$; **simple linear regression**, where $Y_i \overset{\text{indep}}{\sim} \mathcal{N}(\alpha + \beta X_i, \sigma^2)$; and the **multinomial distribution** $(X_1, X_2, \dots, X_k) \sim \text{Multinomial}(n; p_1, p_2, \dots, p_k)$, where $p_1 + \cdots + p_k = 1$.

We end with **Bayesian statistics**, whose starting point is to treat θ as a random variable. We write $\pi(\theta)$ for the prior density, which represents our beliefs before collecting data. For a continuous scalar parameter with parameter space Θ, Bayes' theorem gives the posterior density

$$\pi(\theta \mid X_1, \dots, X_n) = \frac{f(X_1, \dots, X_n \mid \theta)\pi(\theta)}{\displaystyle\int_\Theta f(X_1, \dots, X_n \mid u)\pi(u)\,\mathrm{d}u}.$$

For a discrete parameter, the denominator is instead a sum. The corresponding integral can be difficult to compute, especially when θ is high-dimensional, so more advanced methods often study

the posterior indirectly. For this reason, Bayesian methods are not a major focus of this course.

Important Probability Distributions

The most basic distribution is the **Bernoulli distribution**. If $X \sim \text{Bernoulli}(p)$, where $0 \leq p \leq 1$, then X takes the value 1 with probability p and the value 0 with probability $1-p$. The probability mass function is $f(x) = p^x(1-p)^{1-x}$ for $x = 0, 1$. We have $\mathbb{E}[X] = p$ and $\text{Var}(X) = p(1-p)$.

The **binomial distribution** arises as the number of successes in n independent Bernoulli trials. If $X \sim \text{Bin}(n, p)$, where n is a nonnegative integer and $0 \leq p \leq 1$, then X can take on the values $0, 1, \ldots, n$. The probability mass function is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

The expectation is $\mathbb{E}[X] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = np$, and the variance is $\text{Var}(X) = np(1-p)$.

Lecture 2 — Thursday, January 09, 2014

Let us study the **multinomial distribution**. Consider a vector of random variables

$$(X_1, \ldots, X_k) \sim \text{Multinomial}(n; p_1, \ldots, p_k).$$

The probability mass function is

$$f_{X_1,\ldots,X_k}(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

subject to

$$p_i \geq 0 \quad (i = 1, \ldots, k), \qquad \sum_{i=1}^{k} p_i = 1, \qquad x_1, \ldots, x_k \in \{0, 1, \ldots, n\}, \qquad \text{and} \qquad \sum_{i=1}^{k} x_i = n.$$

This is the joint distribution of the random vector $(X_1, \ldots, X_k)$. As a special case, consider the trinomial model $k = 3$. Because $X_3 = n - X_1 - X_2$, only two coordinates are free, and their joint mass function is

$$f_{X_1,X_2}(x_1, x_2) = \frac{n!}{x_1! x_2! (n - x_1 - x_2)!} p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n - x_1 - x_2}.$$

Two natural questions arise. First, what are the marginal distributions of X_1 and X_2? Marginalizing out X_2 gives

$$\begin{aligned} f_{X_1}(x_1) &= \sum_{x_2=0}^{n-x_1} f(x_1, x_2) = \sum_{x_2=0}^{n-x_1} \frac{n!}{x_1! x_2! (n-x_1-x_2)!} p_1^{x_1} p_2^{x_2} (1-p_1-p_2)^{n-x_1-x_2} \\ &= \frac{n!}{(n-x_1)! x_1!} p_1^{x_1} \sum_{x_2=0}^{n-x_1} \frac{(n-x_1)!}{x_2! (n-x_1-x_2)!} p_2^{x_2} (1-p_1-p_2)^{n-x_1-x_2} \\ &= \binom{n}{x_1} p_1^{x_1} (1-p_1)^{n-x_1} \text{ by the binomial theorem.} \end{aligned}$$

Thus

$$X_1 \sim \text{Bin}(n, p_1).$$

By symmetry, $X_2 \sim \text{Bin}(n, p_2)$.

Second, what is the conditional distribution of $X_2 \mid X_1 = x_1$? Assume $p_1 < 1$ and that the conditioning event has positive probability. We have

$$\begin{aligned} f_{X_2|X_1}(x_2 \mid x_1) &= \frac{f(x_1, x_2)}{f_{X_1}(x_1)} = \frac{(n-x_1)!}{x_2! (n-x_1-x_2)!} \frac{p_2^{x_2} (1-p_1-p_2)^{n-x_1-x_2}}{(1-p_1)^{n-x_1}} \\ &= \binom{n-x_1}{x_2} \left(\frac{p_2}{1-p_1}\right)^{x_2} \left(1 - \frac{p_2}{1-p_1}\right)^{n-x_1-x_2}. \end{aligned}$$

Thus,

$$X_2 \mid X_1 = x_1 \sim \text{Bin}\left(n - x_1, \frac{p_2}{1-p_1}\right).$$

More generally, if $p_k < 1$ and $\mathbb{P}(X_k = x_k) > 0$,

$$(X_1, X_2, \ldots, X_{k-1}) \mid X_k = x_k \sim \text{Multinomial}\left(n - x_k, \frac{p_1}{1-p_k}, \frac{p_2}{1-p_k}, \ldots, \frac{p_{k-1}}{1-p_k}\right).$$

Next, recall the **Poisson distribution**. For $\lambda > 0$, a random variable $X \sim \text{Poisson}(\lambda)$ takes values $0, 1, 2, \ldots$ with probability mass function

$$f_X(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \qquad x \in \{0, 1, 2, \ldots\}.$$

This distribution commonly models rare events. To motivate it, start from the binomial mass function $f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$ and let $n \to \infty$ and $p \to 0$ while keeping $np = \lambda$. In other words, the number of trials grows while each trial becomes less likely to succeed. For a fixed

nonnegative integer x, substitute $p = \lambda/n$ and take n sufficiently large that $n \geq x$ and $n > \lambda$. Then

$$\begin{aligned} f_X(x) = \binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} &= \frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \frac{n(n-1)\dots(n-x+1)}{n\cdot n\dots n}\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^{-x}\left(1-\frac{\lambda}{n}\right)^{n} \\ &\xrightarrow{n\to\infty} \frac{\lambda^x}{x!}e^{-\lambda}, \end{aligned}$$

using the exponential limit. Thus, the Poisson approximation to the binomial distribution is most appropriate when p is small and n is large.

Now we can compute the expectation:

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} x\cdot f(x) = \sum_{x=0}^{\infty} x\cdot\frac{e^{-\lambda}\lambda^x}{x!} = 0 + \sum_{x=1}^{\infty}\frac{e^{-\lambda}\lambda^{(x-1)}}{(x-1)!}\cdot\lambda = e^{-\lambda}\cdot\lambda\sum_{y=0}^{\infty}\frac{\lambda^y}{y!} = \lambda$$

(with $y = x-1$).

The **exponential distribution** is very important. For $\lambda > 0$, a random variable $X \sim \text{Exp}(\lambda)$ has density

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

This distribution is often used as a model of failure time.

Let us construct a **Poisson process** with rate $\lambda > 0$. Consider a stochastic process $\{X(t) : t \geq 0\}$ which satisfies three properties:

1. $X(0) = 0$.
2. For $0 \leq s < t$, $X(t) - X(s) \sim \text{Poisson}(\lambda(t-s))$, which models the number of events between times s and t.
3. Increments over disjoint time intervals, such as $X(t_2) - X(t_1)$ and $X(t_4) - X(t_3)$ for $0 \leq t_1 < t_2 \leq t_3 < t_4$, are independent.

Let T be the waiting time from one event to the next. Then $\mathbb{P}(T > t)$ is the probability of no

event on a time interval of length t, so

$$\mathbb{P}(T > t) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t}.$$

Hence, for $t \geq 0$,

$$F_T(t) = \mathbb{P}(T \leq t) = 1 - e^{-\lambda t} \qquad \text{and} \qquad f_T(t) = \lambda e^{-\lambda t};$$

the cdf is zero for $t < 0$, as is the density. Therefore $T \sim \text{Exp}(\lambda)$, so the exponential distribution models waiting times between events.

Finally, recall the **normal distribution**. For $\sigma > 0$, a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ has density function

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

We have $\mathbb{E}[X] = \mu$ and $\text{Var}(X) = \sigma^2$. Recall that the **standard normal** model is $Z \sim \mathcal{N}(0, 1)$. Some rules of thumb are useful:

$$\mathbb{P}(\mu - \sigma < X < \mu + \sigma) \approx 68\% \qquad \text{and} \qquad \mathbb{P}(\mu - 2\sigma < X < \mu + 2\sigma) \approx 95\%.$$

Example 2.1 Consider the following application. Suppose that $X \sim \mathcal{N}(\mu_A, \sigma^2)$ if the observation is of type A, and $X \sim \mathcal{N}(\mu_B, \sigma^2)$ if it is of type B, where $\mu_A > \mu_B$. Given an observation X from either type, we would like to infer which type was more likely to have generated X. A natural **decision rule** is to claim type A if $X > c$ for some $c \in \mathbb{R}$. We need to optimize c to minimize the error. The probability of an erroneous classification is given by

$$\mathbb{P}(\text{error}) = \mathbb{P}(X > c \mid B) \cdot \mathbb{P}(B) + \mathbb{P}(X \leq c \mid A) \cdot \mathbb{P}(A),$$

by the law of total probability. Assume $\mathbb{P}(A) = \mathbb{P}(B) = \frac{1}{2}$. Then

$$\mathbb{P}(\text{error}) = \frac{1}{2}\left(1 - F_B(c) + F_A(c)\right).$$

Differentiating with respect to c, we get

$$\frac{\mathrm{d}}{\mathrm{d}c}\mathbb{P}(\text{error}) = \frac{1}{2}\big(f_A(c) - f_B(c)\big).$$

The two equal-variance densities cross only at their midpoint, and the derivative changes

from negative to positive there. Thus the unique minimizing threshold is

$$c = \frac{\mu_A + \mu_B}{2}.$$

Lecture 3 — Tuesday, January 14, 2014

Maximum Likelihood Estimation

The first step in statistical inference is to obtain data and use it to estimate the parameters of the distribution.

Assume we have already chosen a model, meaning the distributional family from which the data arise. If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$, independence gives the joint density or mass function $\prod_{i=1}^{n} f(X_i;\theta)$. Viewed as a function of θ, this is the **likelihood function**

$$L(\theta) = \prod_{i=1}^{n} f(X_i;\theta).$$

The **maximum likelihood estimate** (**MLE**) is

$$\widehat{\theta}_{\text{MLE}} = \operatorname*{argmax}_{\theta} L(\theta).$$

In short, the MLE is the parameter value that makes the observed data most likely under the model.

In practice, we often maximize $l(\theta) = \log(L(\theta))$, called the **log-likelihood function**, using the natural logarithm. This transformation changes the scale but not the maximizer. Differentiating a large product is cumbersome, while the log-likelihood converts products into sums which are much easier to handle.

Example 3.1 Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$, where $0 \leq p \leq 1$. Then the likeli-

hood function is

$$L(p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^{n} X_i}(1-p)^{n-\sum_{i=1}^{n} X_i}.$$

Let us maximize this. We have

$$l(p) = \sum_{i=1}^{n} X_i \log(p) + (n - \sum_{i=1}^{n} X_i)\log(1-p).$$

If $0 < \sum_i X_i < n$, differentiating yields

$$\frac{\mathrm{d}}{\mathrm{d}p} l(p) = \frac{\sum_{i=1}^{n} X_i}{p} - \frac{n - \sum_{i=1}^{n} X_i}{1-p} = 0 \implies \sum_{i=1}^{n} X_i = np \implies \widehat{p}_{\mathrm{MLE}} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}.$$

The second derivative is strictly negative on $(0,1)$, so this critical point is the unique maximum. If $\sum_i X_i = 0$, the likelihood is maximized at $p = 0$; if $\sum_i X_i = n$, it is maximized at $p = 1$. Thus $\widehat{p}_{\mathrm{MLE}} = \bar{X}$ in every case.

Example 3.2 Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Exp}(\lambda)$, where $\lambda > 0$. Then the likelihood function is

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i},$$

and

$$l(\lambda) = n\log(\lambda) - \lambda \sum_{i=1}^{n} X_i.$$

We have

$$l'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0 \implies \widehat{\lambda}_{\mathrm{MLE}} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}},$$

and it is easy to perform a second derivative test to verify that $\widehat{\lambda}_{\mathrm{MLE}}$ is a global maximum.

This makes intuitive sense because

$$\mathbb{E}[X_i] = \int_0^\infty x\lambda e^{-\lambda x}\,\mathrm{d}x = \frac{1}{\lambda}$$

for all i, and it is reasonable to estimate λ by setting $\bar{X} = \frac{1}{\lambda}$.

Method-of-Moments Estimator (MOM)

Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$. We can calculate

$$\mathbb{E}[X_i] = \int x \cdot f(x;\theta)\,\mathrm{d}x = g_1(\theta) \qquad \text{and} \qquad \mathbb{E}\left[X_i^2\right] = \int x^2 \cdot f(x;\theta)\,\mathrm{d}x = g_2(\theta),$$

and so on. These are the **moments** of X_i. The **method of moments** (**MOM**) consists of substituting $\mathbb{E}[X_i^m]$ with

$$\frac{1}{n}\sum_{i=1}^n X_i^m$$

thereby replacing the population moments by their sample analogues. Thus we solve equations such as

$$\frac{1}{n}\sum_{i=1}^n X_i = g_1(\theta) \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^n X_i^2 = g_2(\theta).$$

We may choose whichever moments are convenient. That freedom makes the method somewhat arbitrary and less systematic than maximum likelihood. Sometimes the two methods agree, and sometimes they do not.

Example 3.3 Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\sigma > 0$. Let us apply both methods to estimate (μ, σ^2). For the maximum likelihood approach, we have

$$L(\mu,\sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma}\right)^n e^{-\frac{\sum_{i=1}^n (X_i-\mu)^2}{2\sigma^2}}.$$

Therefore

$$l(\mu,\sigma^2) = \text{constant} - n\log(\sigma) - \frac{\sum_{i=1}^n (X_i-\mu)^2}{2\sigma^2}.$$

Differentiating with respect to μ, we get

$$\begin{aligned}\frac{\partial}{\partial\mu}l(\mu,\sigma^2) &= \frac{\sum_{i=1}^{n}2(X_i-\mu)}{2\sigma^2} = 0\\ &\implies \sum_{i=1}^{n}(X_i-\mu)=0\\ &\implies \sum_{i=1}^{n}X_i - n\mu = 0\end{aligned}$$

which implies that

$$\widehat{\mu}_{\text{MLE}} = \frac{\sum_{i=1}^{n}X_i}{n} = \bar{X}.$$

Substituting $\widehat{\mu}_{\text{MLE}} = \bar{X}$ and differentiating the resulting profile log-likelihood with respect to σ gives

$$-\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^3} = 0.$$

For a nonconstant sample, the profile log-likelihood increases before this critical point and decreases after it, so

$$\widehat{\sigma}^2_{\text{MLE}} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}.$$

If every observation is identical, the likelihood is unbounded as $\sigma \downarrow 0$, so no MLE exists in the parameter space $\sigma > 0$; this event has probability zero under the model when $n \geq 2$. Now for the method of moments. We have

$$\mathbb{E}\left[X_i\right] = \mu \qquad \text{and} \qquad \mathbb{E}\left[X_i^2\right] = \text{Var}(X_i) + \mathbb{E}\left[X_i\right]^2 = \sigma^2 + \mu^2.$$

This gives $\widehat{\mu}_{\text{MOM}} = \bar{X} = \widehat{\mu}_{\text{MLE}}$ and

$$\frac{1}{n}\sum_{i=1}^{n}X_i^2 = \sigma^2 + \mu^2,$$

which implies

$$\widehat{\sigma}^2_{\text{MOM}} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2.$$

Officially we are done. But we can verify that the two estimates of σ^2 agree:

$$\begin{aligned}
\widehat{\sigma}^2_{\text{MOM}} &= \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n} \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \bar{X}^2 - \frac{2\bar{X}}{n}\sum_{i=1}^{n} X_i \\
&= \frac{1}{n}\sum_{i=1}^{n} (X_i^2 + \bar{X}^2 - 2X_i\bar{X}) \\
&= \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2 \\
&= \widehat{\sigma}^2_{\text{MLE}}.
\end{aligned}$$

The MLE and the MOM estimator are not the same in general.

Example 3.4 Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$, where $\theta > 0$. For the method of moments, we have

$$\mathbb{E}[X_i] = \int_0^\theta x \cdot \frac{1}{\theta}\,\mathrm{d}x = \frac{\theta}{2}$$

for all i, which implies $\bar{X} = \frac{\theta}{2}$, and hence $\widehat{\theta}_{\text{MOM}} = 2\bar{X}$. Now for maximum likelihood, we have

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta) = \theta^{-n} \mathbb{1}_{\{\theta \geq X_{(n)}\}},$$

where $X_{(n)} = \max\{X_1, \ldots, X_n\}$ and the observed data are nonnegative. We do not need the logarithm here, since the likelihood has a simple form. On its support the likelihood is decreasing in θ, so $\widehat{\theta}_{\text{MLE}} = X_{(n)}$. This is a constrained maximization problem. The estimator is sensible; it is strictly below the true parameter with probability one, but we will later see why it is still preferable.

The key point is that both procedures produce estimators that are functions of the data $X_1, \ldots, X_n$.

Lecture 4 — Thursday, January 16, 2014

Constrained Optimization and Applied Examples

When we perform maximum likelihood estimation, we are sometimes constrained by the setup or parameter space of the statistical model.

Example 4.1 Suppose we observe n independent trials, where each trial falls into exactly one of K categories with probabilities $p_1, p_2, \ldots, p_K$. Let X_k denote the number of observations in category k. Then

$$(X_1, X_2, \ldots, X_K) \sim \text{Multinomial}(n; p_1, p_2, \ldots, p_K),$$

with constraints $p_k \geq 0$ for $k = 1, \ldots, K$ and

$$\sum_{k=1}^{K} p_k = 1.$$

Let us determine the MLE of $(p_1, \ldots, p_K)$. For the count vector $(X_1, \ldots, X_K)$, satisfying

$$\sum_{k=1}^{K} X_k = n,$$

the likelihood is

$$L(p_1, \ldots, p_K) = \frac{n!}{X_1! \cdots X_K!} \prod_{k=1}^{K} p_k^{X_k},$$

so, when every $X_k > 0$, the log-likelihood is

$$l(p_1, \ldots, p_K) = \text{constant} + \sum_{k=1}^{K} X_k \log(p_k).$$

We cannot maximize this by setting the partial derivatives of l equal to zero independently, because the parameters must satisfy

$$\sum_{k=1}^{K} p_k = 1.$$

Instead, we introduce a Lagrange multiplier and consider

$$\mathcal{L}(p_1, \ldots, p_K, \lambda) = \sum_{k=1}^{K} X_k \log(p_k) + \lambda \left(\sum_{k=1}^{K} p_k - 1 \right).$$

Differentiating gives

$$\frac{\partial \mathcal{L}}{\partial p_k} = \frac{X_k}{p_k} + \lambda = 0, \quad k = 1, \ldots, K,$$

and

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{k=1}^{K} p_k - 1 = 0.$$

Hence

$$p_k = -\frac{X_k}{\lambda}, \quad k = 1, \ldots, K.$$

Summing over k, we obtain

$$1 = \sum_{k=1}^{K} p_k = -\frac{1}{\lambda} \sum_{k=1}^{K} X_k = -\frac{n}{\lambda},$$

so $\lambda = -n$. Substituting this back yields

$$\widehat{p}_k^{\mathrm{MLE}} = \frac{X_k}{n}, \quad k = 1, \ldots, K.$$

This is intuitive: the MLE of each category probability is the observed sample proportion, and it also agrees with $\widehat{p}_k^{\mathrm{MOM}}$. If some $X_k = 0$, the preceding logarithmic calculation does not apply at $p_k = 0$. Nevertheless, an unobserved category must receive probability 0 at the maximum. Indeed, if it received positive probability, reallocating that mass among the observed categories in proportion to their current probabilities would strictly increase every factor that appears in the likelihood. Thus the same formula holds on the boundary.

Example 4.2 (Hardy–Weinberg model) Suppose we have genotypes AA, Aa, and aa with probabilities $(1-\theta)^2$, $2\theta(1-\theta)$, and θ^2, respectively, where $0 \leq \theta \leq 1$; this is a specifically parametrized multinomial model. We want to estimate θ. Suppose we have data X_1, X_2, X_3. Then

$$L(\theta) = \frac{n!}{X_1!X_2!X_3!}\left[(1-\theta)^2\right]^{X_1}\left[2\theta(1-\theta)\right]^{X_2}\left[\theta^2\right]^{X_3} = (\text{const.}) \cdot (1-\theta)^{2X_1+X_2} \cdot \theta^{X_2+2X_3}.$$

Therefore

$$l(\theta) = (\text{const.}) + (2X_1 + X_2)\log(1-\theta) + (X_2 + 2X_3)\log(\theta),$$

and, for an interior solution,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} l(\theta) = 0$$

implies

$$\widehat{\theta}_{\text{MLE}} = \frac{X_2 + 2X_3}{2n}.$$

The same formula gives the appropriate endpoint when all observed alleles are of one type.

For example, suppose we observe $X_1 = 37$, $X_2 = 46$, and $X_3 = 17$. Then $\widehat{\theta}_{\text{MLE}} = \frac{46+2\cdot 17}{2(37+46+17)} = 0.4$.

We can check the Hardy–Weinberg model by comparing actual observations with model predictions. The model predicts that among 100 people (properly chosen), we would have

$$n(1-\widehat{\theta})^2 = 100(1-0.4)^2 = 36$$

individuals with genotype AA, 48 with Aa, and 16 with aa. These are quite close to the observed values. The difference is tolerable, and we could say that the Hardy–Weinberg model is reasonable for this data. Note that $p_1 + p_2 + p_3 = (1-\theta+\theta)^2 = 1$.

Example 4.3 Suppose $Y_i \overset{\text{indep}}{\sim} \text{Poisson}(\theta x_i)$ for $i = 1, \ldots, n$, where $\theta \geq 0$ and the fixed covariates satisfy $x_i > 0$. These variables are independent but not identically distributed; however, they still belong to the same distributional family. The values x_i are **covariates** (or **predictors**). We have

$$L(\theta) = \prod_{i=1}^{n} \frac{e^{-(\theta x_i)}(\theta x_i)^{Y_i}}{Y_i!},$$

and

$$l(\theta) = \sum_{i=1}^{n} (-\theta x_i + Y_i \log(\theta x_i)) + \text{constant} = -\theta \sum_{i=1}^{n} x_i + \log(\theta) \sum_{i=1}^{n} Y_i + \text{constant}.$$

If $\sum_i Y_i > 0$, differentiating yields

$$-\sum_{i=1}^{n} x_i + \frac{1}{\theta} \sum_{i=1}^{n} Y_i = 0,$$

which implies

$$\widehat{\theta}_{\text{MLE}} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i}.$$

If every $Y_i = 0$, the likelihood is instead maximized at the boundary $\widehat{\theta} = 0$, so the same formula still applies.

Example 4.4 Suppose $X_i \overset{\text{indep}}{\sim} \text{Poisson}(V_i\theta)$ for $i = 1, \ldots, n$, where $V_i > 0$ and $\theta \geq 0$ (for example, X_i could be the count of bacteria from the ith laboratory sample of volume V_i). Let

$$Y_i = \begin{cases} 1, & X_i > 0 \\ 0, & X_i = 0 \end{cases}.$$

We want to estimate θ using only the data $(V_1, Y_1), \ldots, (V_n, Y_n)$, thereby bypassing direct measurements of $X_1, X_2, \ldots$. This is a realistic measurement scenario.

Clearly $Y_i \sim \text{Bernoulli}(p_i)$, where

$$p_i = \mathbb{P}(X_i > 0) = 1 - \mathbb{P}(X_i = 0) = 1 - \frac{e^{-(V_i\theta)}(V_i\theta)^0}{0!} = 1 - e^{-V_i\theta}.$$

Therefore

$$L(\theta) = \prod_{i=1}^{n} (1 - e^{-V_i\theta})^{Y_i} (e^{-V_i\theta})^{1-Y_i},$$

and

$$l(\theta) = \sum_{i=1}^{n} \left[Y_i \log\left(1 - e^{-V_i\theta}\right) + (1 - Y_i)(-V_i\theta) \right].$$

Differentiating yields

$$l'(\theta) = \sum_{i=1}^{n} \left[Y_i \frac{V_i e^{-V_i\theta}}{1 - e^{-V_i\theta}} - (1 - Y_i)V_i \right],$$

If the data contain at least one zero and one one, the log-likelihood is strictly concave, its derivative changes sign, and its unique maximizer is the numerical solution of $l'(\theta) = 0$. If every $Y_i = 0$, the MLE is $\widehat{\theta} = 0$; if every $Y_i = 1$, the likelihood increases toward its supremum as $\theta \to \infty$, so there is no finite MLE.

When $l'(\theta) = 0$ has no closed-form solution, we turn to numerical optimization, for example by using a root-finding procedure such as **Newton's method**. To solve $f(x) = 0$, we start from an initial guess x_0 and iterate

$$x_{m+1} = x_m - \frac{f(x_m)}{f'(x_m)}, \qquad m = 0, 1, 2, \ldots.$$

In our example, this becomes

$$\theta_{m+1} = \theta_m - \frac{l'(\theta_m)}{l''(\theta_m)}.$$

A variation of Newton's method, the **Fisher scoring algorithm**, replaces $l''(\theta)$ with $\mathbb{E}\left[l''(\theta)\right]$. This approach is often more stable for maximum likelihood estimation problems.

Lecture 5 — Tuesday, January 21, 2014

Properties of Estimators: Bias, Efficiency, and MSE

We now ask a central question: how do we assess the quality of an estimator?

To illustrate, let $Y \sim \text{Bin}(n, p)$. *Before* we flip a coin n times, Y is a random variable; *after* observing the flips, it becomes data. It is important to keep this connection between data and distribution in mind.

Recall Example 3.4, where the method of moments and maximum likelihood give different estimators. At this point, it is not immediately clear which estimator should be preferred. In that example, the MLE has the smaller mean squared error, and we can justify the comparison analytically. This is not a universal finite-sample guarantee for every model.

The estimator $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ is a function of the random variables $X_1, \ldots, X_n$ and hence is a random variable itself. The distribution of $\widehat{\theta}$ is called the **sampling distribution**.

Definition 5.1 An estimator $\widehat{\theta}$ is **unbiased** if it is correctly centered; that is, if

$$\mathbb{E}\left[\widehat{\theta}\right] = \theta.$$

The **bias** of an estimator is

$$\text{Bias}_\theta(\widehat{\theta}) = \mathbb{E}_\theta[\widehat{\theta}] - \theta.$$

Definition 5.2 Suppose $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are both unbiased estimators of θ. We say $\widehat{\theta}_1$ is more **efficient** than $\widehat{\theta}_2$ if $\text{Var}(\widehat{\theta}_1) < \text{Var}(\widehat{\theta}_2)$.

Definition 5.3 Assuming the second moment is finite, the **mean squared error** (**MSE**) of an estimator $\widehat{\theta}$ is defined as

$$\text{MSE}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right].$$

Theorem 5.4 $\text{MSE}(\widehat{\theta}) = \text{Bias}(\widehat{\theta})^2 + \text{Var}(\widehat{\theta})$.

Proof. A direct calculation gives

$$\begin{aligned}
\text{MSE}(\widehat{\theta}) &= \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right] \\
&= \mathbb{E}\left[(\widehat{\theta} - \mathbb{E}\left[\widehat{\theta}\right] + \mathbb{E}\left[\widehat{\theta}\right] - \theta)^2\right] \\
&= \mathbb{E}\left[(\widehat{\theta} - \mathbb{E}\left[\widehat{\theta}\right])^2\right] + 2(\mathbb{E}\left[\widehat{\theta}\right] - \theta)\mathbb{E}\left[\widehat{\theta} - \mathbb{E}\left[\widehat{\theta}\right]\right] + (\mathbb{E}\left[\widehat{\theta}\right] - \theta)^2 \\
&= \text{Var}(\widehat{\theta}) + \text{Bias}(\widehat{\theta})^2.
\end{aligned}$$

The middle term vanishes because $\mathbb{E}\left[\widehat{\theta} - \mathbb{E}\left[\widehat{\theta}\right]\right] = 0$. □

Lecture 6 — Thursday, January 23, 2014

Sampling Distributions and Order Statistics

The estimator $\widehat{\theta} = g(X_1, X_2, \ldots, X_n)$ is a random variable, and we want to find its distribution. Recall that we often had $\widehat{\theta} = \bar{X}$ or $\widehat{\theta} = X_{\max}$. In the former case, to understand the sampling distribution of $\widehat{\theta}$ it usually suffices to know the distribution of $X_1 + \cdots + X_n$. Let us start with a simple example.

Example 6.1 Suppose $X, Y \overset{\text{iid}}{\sim} \text{Unif}(0,1)$. We want the distribution of $Z = X + Y$. We have

$$F_Z(z) = \mathbb{P}(Z \le z) = \mathbb{P}(X + Y \le z) = \iint\limits_{x+y\le z} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \iint\limits_{x+y\le z} f_X(x) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y$$

$$= \iint\limits_{[0,1]^2 \cap \{x+y\le z\}} 1\,\mathrm{d}x\,\mathrm{d}y = \begin{cases} 0, & z < 0, \\ z^2/2, & 0 \le z \le 1, \\ 1 - \dfrac{(2-z)^2}{2}, & 1 < z \le 2, \\ 1, & z > 2. \end{cases}$$

Differentiating gives the density

$$f_Z(z) = \begin{cases} z, & 0 < z < 1, \\ 2 - z, & 1 \le z < 2, \\ 0, & \text{otherwise.} \end{cases}$$

Figure 1 illustrates the geometric calculation of the cdf $F_Z(z) = \mathbb{P}(X + Y \le z)$ for $z \le 1$ and $z > 1$.

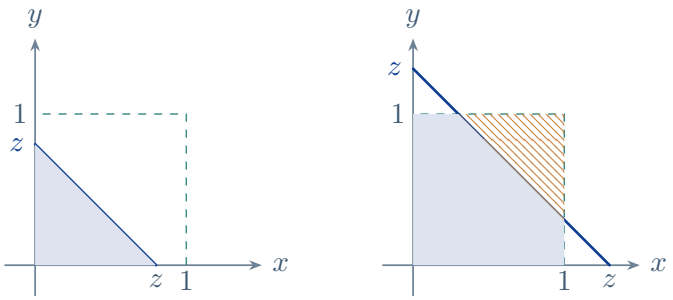


FIGURE 1

The resulting density function $f_Z(z)$, shown in Figure 2, is that of a **triangular distribution**:

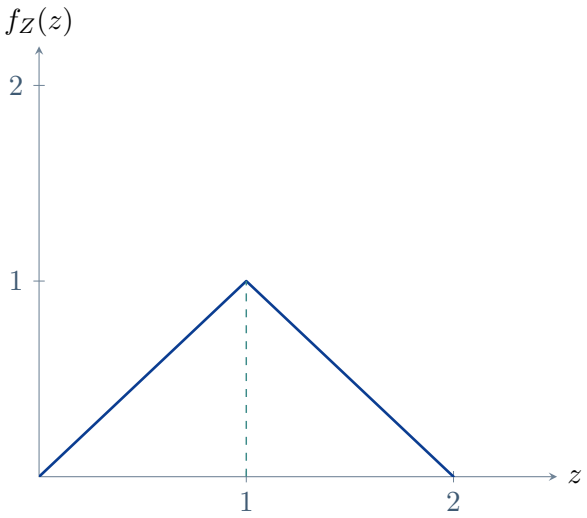


FIGURE 2

Adding two uniform random variables gives a joined piecewise linear density. Adding three gives a piecewise quadratic density. Adding four, five, and so on gives piecewise cubic, quartic, and higher-order polynomial densities, which join smoothly. This behavior is consistent with the central limit theorem. More generally, finding the distribution of $g(X_1, \ldots, X_n)$ typically requires n-dimensional integration, which is difficult in practice. This is why asymptotic approximations, such as the central limit theorem, are often used.

Example 6.2 Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_X(x)$ where f_X is a density, and let $Y = \max\{X_1, \ldots, X_n\}$. What is the density $f_Y(y)$? We have

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X_1, \ldots, X_n \leq y) = \mathbb{P}(X_1 \leq y) \cdot \mathbb{P}(X_2 \leq y) \cdots \mathbb{P}(X_n \leq y) = [F_X(y)]^n.$$

Differentiating, we obtain

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = n \cdot [F_X(y)]^{n-1} \cdot f_X(y).$$

This formula is quite general.

Example 6.3 Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$. We know from the method of moments that $\widehat{\theta}_{\text{MOM}} = 2\bar{X}$, and from maximum likelihood estimation that $\widehat{\theta}_{\text{MLE}} = \max\{X_1, \ldots, X_n\} = Y$. We have $f_X(y) = \frac{1}{\theta}$ for $y \in (0, \theta)$ and

$$F_X(y) = \mathbb{P}(X \leq y) = \frac{y}{\theta}$$

for $y \in (0, \theta)$. From the general formula, the sampling distribution of $\widehat{\theta}_{\text{MLE}}$ is $f_Y(y) = n \left(\frac{y}{\theta}\right)^{n-1} \cdot \frac{1}{\theta}$. We can also calculate the bias using

$$\begin{aligned}
\mathbb{E}[Y] &= \int_0^\theta y \cdot f_Y(y) \, \mathrm{d}y \\
&= \int_0^\theta y \cdot n \left(\frac{y}{\theta}\right)^{n-1} \cdot \frac{1}{\theta} \, \mathrm{d}y \\
&= \frac{n}{\theta^n} \cdot \left[\frac{y^{n+1}}{n+1}\right]_0^\theta \\
&= \frac{n}{n+1}\theta \neq \theta,
\end{aligned}$$

so

$$\text{Bias}(\widehat{\theta}_{\text{MLE}}) = \mathbb{E}\left[\widehat{\theta}_{\text{MLE}}\right] - \theta = -\frac{\theta}{n+1}.$$

The estimator $\widehat{\theta}_{\text{MLE}}$ is strictly below θ with probability one, but it converges to θ in probability as $n \to \infty$. For the variance of $\widehat{\theta}_{\text{MLE}}$, recall that $\text{Var}(Y) = \mathbb{E}\left[Y^2\right] - \mathbb{E}[Y]^2$. We compute

$$\mathbb{E}\left[Y^2\right] = \int_0^\theta y^2 \cdot n \left(\frac{y}{\theta}\right)^{n-1} \cdot \frac{1}{\theta} \, \mathrm{d}y = \frac{n}{\theta^n} \left[\frac{y^{n+2}}{n+2}\right]_0^\theta = \frac{n}{n+2}\theta^2.$$

Thus,

$$\begin{aligned}\mathrm{MSE}(\widehat{\theta}_{\mathrm{MLE}}) &= \mathrm{Bias}^2(\widehat{\theta}_{\mathrm{MLE}}) + \mathrm{Var}(\widehat{\theta}_{\mathrm{MLE}}) \\ &= \left[\frac{n}{n+1}\theta - \theta\right]^2 + \left[\frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2\right] \\ &= \frac{2\theta^2}{(n+1)(n+2)} \\ &= O\left(\frac{1}{n^2}\right).\end{aligned}$$

Now let us examine $\widehat{\theta}_{\mathrm{MOM}} = 2\bar{X}$. For the bias:

$$\mathbb{E}\left[2\bar{X}\right] = 2\mathbb{E}\left[\bar{X}\right] = 2\mathbb{E}\left[X_i\right] = 2 \cdot \frac{\theta}{2} = \theta,$$

so it is unbiased. For the variance:

$$\mathrm{Var}(2\bar{X}) = 4 \cdot \mathrm{Var}(\bar{X}) = 4 \cdot \frac{\mathrm{Var}(X_i)}{n} = \frac{4}{n}\left(\frac{(b-a)^2}{12}\right) = \frac{4}{n}\left(\frac{(\theta-0)^2}{12}\right) = \frac{\theta^2}{3n} = O\left(\frac{1}{n}\right).$$

Thus, the MSE shrinks faster for the MLE than for the MOM estimator, as expected.

The next example is just for fun.

Example 6.4 Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_X(x)$ with continuous cumulative distribution function $F_X(x)$. Now consider an independent observation X_{n+1} from the same distribution. What is the probability that it is the largest of all $n+1$ observations?

Let $Y = \max\{X_1, \ldots, X_n\}$ and write $W = X_{n+1}$. We have $f_Y(y) = nF_X(y)^{n-1}f_X(y)$. Then, using Fubini's theorem,

$$\begin{aligned}\mathbb{P}(W \geq Y) &= \iint\limits_{w \geq y} f_{W,Y}(w,y)\,\mathrm{d}w\,\mathrm{d}y \\ &= \iint\limits_{w \geq y} f_X(w) n F_X(y)^{n-1} f_X(y)\,\mathrm{d}w\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} f_X(w)\left[\int_{-\infty}^{w} n F_X(y)^{n-1} f_X(y)\,\mathrm{d}y\right]\mathrm{d}w \\ &= \int_{-\infty}^{\infty} f_X(w)\left[F_X(y)^n\right]_{-\infty}^{w}\,\mathrm{d}w\end{aligned}$$

$$= \int_{-\infty}^{\infty} f_X(w) F_X(w)^n \, \mathrm{d}w$$
$$= \left[\frac{F_X(w)^{n+1}}{n+1} \right]_{-\infty}^{\infty}$$
$$= \frac{1}{n+1}.$$

There is no need for any integration here. Since $X_1, \ldots, X_{n+1}$ are independent and identically distributed and their common distribution is continuous, the probability of obtaining any ordering (or equivalently, any permutation) of them is identical to that of obtaining any other ordering. There are $n!$ permutations of $X_1, \ldots, X_n$, and $(n+1)!$ if X_{n+1} is added in. With X_{n+1} fixed and $X_1, \ldots, X_n$ free, the desired probability is

$$\frac{n!}{(n+1)!} = \frac{1}{n+1}.$$

Lecture 7 — Tuesday, January 28, 2014

2. Asymptotic Theory

Asymptotics and Moment Generating Functions

So far, we have found distributions directly: determine a density or distribution function, then integrate or differentiate as needed. This approach quickly becomes intractable for complicated functions of many random variables. Asymptotic theory gives us another route.

Definition 7.1 We say that X_n **converges in distribution** to X and write $X_n \xrightarrow{D} X$ if $F_{X_n}(x) \to F_X(x)$ for all continuity points x of F_X as $n \to \infty$.

Definition 7.2 We say that X_n **converges in probability** to $c \in \mathbb{R}$ and write $X_n \xrightarrow{\mathbb{P}} c$ if for all $\varepsilon > 0$, we have $\mathbb{P}(|X_n - c| \geq \varepsilon) \to 0$ as $n \to \infty$.

Definition 7.3 The **moment generating function (mgf)** of a random variable X is

$$\operatorname{mgf}_X(t) \coloneqq \mathbb{E}\left[e^{tX}\right].$$

We can verify that

$$\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{mgf}_X(t)\bigg|_{t=0} = \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}t}e^{tX}\right]\bigg|_{t=0} = \mathbb{E}\left[e^{tX}\cdot X\right]\big|_{t=0} = \mathbb{E}\left[X\right],$$

under appropriate regularity conditions. This gives the first moment. Similarly, the second moment is given by

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\operatorname{mgf}_X(t)\bigg|_{t=0} = \mathbb{E}\left[X^2\right].$$

A common sufficient condition for $X_n \xrightarrow{D} X$ is that $\operatorname{mgf}_{X_n}(t) \to \operatorname{mgf}_X(t)$ on a neighborhood of 0, which we will soon use in our heuristic proof of the central limit theorem.

Proposition 7.4 Suppose $Y = X_1 + \cdots + X_n$, where the random variables X_i are independent. Then

$$\begin{aligned}
\operatorname{mgf}_Y(t) &= \mathbb{E}\left[e^{tY}\right] \\
&= \mathbb{E}\left[e^{t(X_1+\cdots+X_n)}\right] \\
&= \mathbb{E}\left[e^{tX_1}e^{tX_2}\ldots e^{tX_n}\right] \\
&= \prod_{i=1}^{n}\mathbb{E}\left[e^{tX_i}\right] \\
&= \prod_{i=1}^{n}\operatorname{mgf}_{X_i}(t).
\end{aligned}$$

Proposition 7.4 is powerful because it makes no assumption on the distribution of each X_i. The X_i may have completely different distributions, yet we can still compute the moment generating function of their sum through simple multiplication.

Proposition 7.5 If $Y = aX + b$, then

$$\operatorname{mgf}_Y(t) = \mathbb{E}\left[e^{t(aX+b)}\right] = \mathbb{E}\left[e^{atX}\cdot e^{bt}\right] = e^{bt}\cdot\operatorname{mgf}_X(at).$$

Example 7.6 Let $Z \sim \mathcal{N}(0,1)$ with $f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$, the standard bell curve. The mgf of Z is

$$\begin{aligned}
\operatorname{mgf}_Z(t) &= \mathbb{E}\left[e^{tZ}\right] \\
&= \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \, \mathrm{d}z \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(z^2-2tz+t^2)/2} e^{t^2/2} \, \mathrm{d}z \\
&= e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(z-t)^2/2} \, \mathrm{d}z \\
&= e^{t^2/2}
\end{aligned}$$

because the last integral is the density function of the $\mathcal{N}(t,1)$ distribution, and hence integrates to 1.

Example 7.7 Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$. Let $\frac{X-\mu}{\sigma} =: Z \sim \mathcal{N}(0,1)$, so $X = \sigma Z + \mu$. By Proposition 7.5,

$$\operatorname{mgf}_X(t) = e^{\mu t} e^{\sigma^2 t^2/2} = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

Example 7.8 Suppose $X_1, \ldots, X_n$ are independent and identically distributed as $\mathcal{N}(\mu, \sigma^2)$. Let $S_n \coloneqq X_1 + \cdots + X_n$ and $\bar{X}_n = \frac{S_n}{n}$. Then by Proposition 7.4, we have

$$\operatorname{mgf}_{S_n}(t) = \prod_{i=1}^{n} \operatorname{mgf}_{X_i}(t) = e^{n\mu t + \frac{n\sigma^2 t^2}{2}},$$

which shows that $S_n \sim \mathcal{N}(n\mu, n\sigma^2)$. Similarly,

$$\operatorname{mgf}_{\bar{X}}(t) = \operatorname{mgf}_{S_n}\left(\frac{t}{n}\right) = e^{\mu t + \frac{\sigma^2 t^2}{2n}},$$

which implies that

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Theorem 7.9 (Central limit theorem) Let $X_1, \ldots, X_n$ be independent and identically distributed random variables from some distribution with $\mathbb{E}[X_i] = \mu$ and $0 < \operatorname{Var}(X_i) =$

$\sigma^2 < \infty$. Define $S_n = X_1 + \cdots + X_n$. Then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1).$$

Although the central limit theorem is stated for S_n, the corresponding result for the sample mean follows immediately:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1).$$

The theorem therefore gives the asymptotic distribution of $\bar{X}_n$, and it often serves as the starting point for approximating more complicated estimators.

Heuristic mgf proof of the central limit theorem. We can write the left-hand side as

$$\frac{1}{\sqrt{n}}\left[\frac{X_1 - \mu}{\sigma} + \cdots + \frac{X_n - \mu}{\sigma}\right] =: \frac{1}{\sqrt{n}}[X_1^* + \cdots + X_n^*].$$

Note that $\mathbb{E}[X_i^*] = 0$ and $\mathrm{Var}(X_i^*) = 1$. Therefore it suffices to prove $\frac{S_n}{\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1)$ under the assumption that $\mu = 0$ and $\sigma^2 = 1$. This mgf argument requires the additional assumption that the common mgf exists on a neighborhood of zero; the theorem itself does not. Suppose $m(t)$ is the common moment-generating function of the standardized variables X_i^*, so that $\mathrm{mgf}_{S_n}(t) = [m(t)]^n$. Then $\mathrm{mgf}_{\frac{S_n}{\sqrt{n}}}(t) = [m\left(\frac{t}{\sqrt{n}}\right)]^n$. As $n \to \infty$, we have $\frac{t}{\sqrt{n}} \to 0$. We expand $m(\frac{t}{\sqrt{n}})$ about 0 using a second-order Taylor expansion:

$$m\left(\frac{t}{\sqrt{n}}\right) \approx m(0) + m'(0)\frac{t}{\sqrt{n}} + m''(0)\frac{t^2}{2n}.$$

Observe that $m'(0) = \mathbb{E}[X_i^*] = 0$,

$$m''(0) = \mathbb{E}\left[(X_i^*)^2\right] = \mathrm{Var}(X_i^*) + \mathbb{E}[X_i^*]^2 = 1,$$

and

$$m(0) = \mathbb{E}\left[e^{tX^*}\right]\Big|_{t=0} = 1.$$

Therefore (heuristically, though not entirely rigorously), we have

$$\left[m\left(\frac{t}{\sqrt{n}}\right)\right]^n \approx \left[1 + \frac{t^2}{2n}\right]^n \to e^{t^2/2},$$

which is the moment generating function of $\mathcal{N}(0,1)$. Since convergence of moment generating functions implies convergence in distribution, we are done. □

The Weak Law of Large Numbers

We now return to the notion of convergence in probability.

Proposition 7.10 (Chebyshev's inequality) Let X be a random variable with $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2 < \infty$. Then for all $\varepsilon > 0$, we have

$$\mathbb{P}(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Proof. On the event $\{|X - \mu| \geq \varepsilon\}$, we have $(X-\mu)^2 \geq \varepsilon^2$. Therefore

$$\sigma^2 = \mathbb{E}[(X-\mu)^2] \geq \mathbb{E}\left[(X-\mu)^2 \mathbb{1}_{\{|X-\mu|\geq\varepsilon\}}\right] \geq \varepsilon^2 \mathbb{P}(|X-\mu| \geq \varepsilon).$$

Rearranging yields the claimed inequality. This argument applies to discrete, continuous, and mixed distributions alike. □

Theorem 7.11 (Weak law of large numbers) Let $X_1, \ldots, X_n$ be independent and identically distributed random variables from some distribution with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then $\bar{X}_n \xrightarrow{\mathbb{P}} \mu$.

Proof. Applying Chebyshev's inequality to $\bar{X}$, we have for all $\varepsilon > 0$ that

$$\mathbb{P}(\left|\bar{X} - \mu\right| \geq \varepsilon) \leq \frac{\mathrm{Var}(\bar{X})}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0$$

as $n \to \infty$. Since $\mu = \mathbb{E}[X_i] = \mathbb{E}\left[\bar{X}\right]$, we conclude that $\bar{X} \xrightarrow{\mathbb{P}} \mu$. □

How does the central limit theorem compare to the weak law of large numbers? Recall that a random variable has variance zero if and only if it is almost surely constant.

Now, without loss of generality, assume $\mathbb{E}[X_i] = 0$ and $\text{Var}(X_i) = 1$. Theorem 7.11 states that

$$\bar{X}_n \xrightarrow{\mathbb{P}} 0,$$

while Theorem 7.9 states that

$$\frac{\bar{X} - 0}{1/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1),$$

which is equivalent to $\sqrt{n}\bar{X}_n \xrightarrow{D} \mathcal{N}(0,1)$.

Note that $\text{Var}(\bar{X}_n) = \frac{1}{n} \to 0$, so $\bar{X}_n$ essentially becomes nonrandom as $n \to \infty$. In contrast,

$$\text{Var}(\sqrt{n}\bar{X}_n) = (\sqrt{n})^2 \cdot \text{Var}(\bar{X}_n) = n \cdot \frac{1}{n} = 1,$$

which is constant for all n. Therefore $\sqrt{n}\bar{X}_n$ remains nondegenerate for every sample size.

Lecture 8 — Thursday, January 30, 2014

3. Interval Estimation and Hypothesis Testing

Confidence Intervals and Pivotal Quantities

In the canonical model, we assume $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with known σ^2 (an unrealistic but instructive starting point). From Example 3.3, $\bar{X}$ is both the MLE and MOM for μ, with $\mathbb{E}\left[\bar{X}\right] = \mu$ and $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$. Standardizing gives

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Hence

$$\mathbb{P}\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) \approx 95\%,$$

which is equivalent to

$$\mathbb{P}\left(\bar{X} - \frac{1.96\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right) \approx 95\%.$$

Thus a 95% **confidence interval** for μ is

$$\left(\bar{X} - \frac{1.96\sigma}{\sqrt{n}}, \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right).$$

A **pivotal quantity** q is a function that (1) depends only on the data $(X_1, \ldots, X_n)$, the parameter of interest, and known constants, and (2) has a fully specified distribution. For example,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is pivotal because $\bar{X}$ comes from the data, σ is known, and the distribution is $\mathcal{N}(0,1)$.

Knowing the distribution of a pivotal quantity allows us to write

$$\mathbb{P}\left(L_\alpha \leq q(X_1, \ldots, X_n; \theta) \leq U_\alpha\right) = 1 - \alpha$$

for a given small α, since knowing the distribution of $q(X_1, \ldots, X_n; \theta)$ means that we can determine the endpoints L_α and U_α. In general, the endpoints are not unique, since the only requirement is that they cover $1 - \alpha$ of the total mass of the distribution.

For the canonical case, the $(1-\alpha)\cdot 100\%$ confidence interval for μ is

$$\left(\bar{X} - \zeta_\alpha \frac{\sigma}{\sqrt{n}}, \bar{X} + \zeta_\alpha \frac{\sigma}{\sqrt{n}}\right)$$

where ζ_α satisfies $\mathbb{P}(Z \leq \zeta_\alpha) = 1 - \alpha/2$ for $Z \sim \mathcal{N}(0,1)$.

Example 8.1 For $\theta > 0$, consider the density $f_X(x) = \frac{2(\theta - x)}{\theta^2}$ on $x \in (0, \theta)$ and zero elsewhere. First, we verify that $Y = q(X; \theta) = X/\theta$ is a pivotal quantity. Condition (1) is satisfied. For condition (2), we apply some distribution theory. For $0 < y < 1$, the cdf of Y is

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}\left(\frac{X}{\theta} \leq y\right) \\ &= \mathbb{P}(X \leq \theta y) \\ &= \int_0^{\theta y} \frac{2(\theta - x)}{\theta^2}\, \mathrm{d}x \end{aligned}$$

$$= \left[\frac{2x}{\theta} - \frac{x^2}{\theta^2}\right]_0^{\theta y}$$
$$= 2y - y^2.$$

Together with $F_Y(y) = 0$ for $y \leq 0$ and $F_Y(y) = 1$ for $y \geq 1$, this is completely known and independent of the parameter θ, which confirms (2).

Next, we construct a confidence interval for θ based on X. On $0 < y < 1$, we have

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = 2 - 2y,$$

and the density is zero elsewhere. The equal-tail cutoffs are shown in Figure 3.

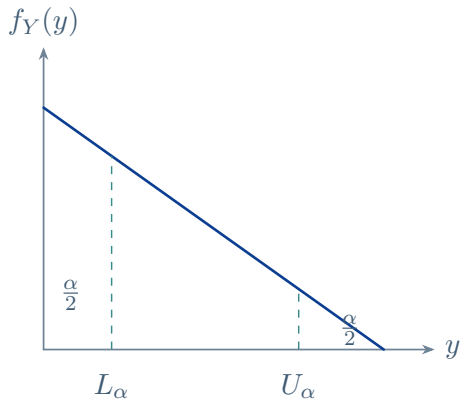


FIGURE 3

For an **equal-tail confidence interval**, we want L_α, U_α such that $F_Y(L_\alpha) = \alpha/2$ and $1 - F_Y(U_\alpha) = \alpha/2$. Solving $2y - y^2 = A$ gives

$$y = \frac{2 \pm \sqrt{4 - 4A}}{2} = 1 \pm \sqrt{1 - A}.$$

Since $y \leq 1$, we take $y = 1 - \sqrt{1 - A}$. Therefore

$$L_\alpha = 1 - \sqrt{1 - \alpha/2} \qquad \text{and} \qquad U_\alpha = 1 - \sqrt{\alpha/2}.$$

Now we use the pivotal quantity:

$$\mathbb{P}\left(1 - \sqrt{1-\alpha/2} \leq \frac{X}{\theta} \leq 1 - \sqrt{\alpha/2}\right) = 1 - \alpha,$$

which implies

$$\mathbb{P}\left(\frac{X}{1-\sqrt{\alpha/2}} \leq \theta \leq \frac{X}{1-\sqrt{1-\alpha/2}}\right) = 1 - \alpha.$$

Therefore, a $(1-\alpha)\cdot 100\%$ confidence interval for θ is

$$\left(\frac{X}{1-\sqrt{\alpha/2}}, \frac{X}{1-\sqrt{1-\alpha/2}}\right).$$

For a concrete example, suppose we want a 28% confidence interval with $X = 1$ and $\alpha = 0.72$. Then a 28% confidence interval based on $X = 1$ is

$$\left(\frac{1}{0.4}, \frac{1}{0.2}\right) = (2.5, 5).$$

Observe that if $\alpha = 0$, we obtain a 100% confidence interval; but substituting this value yields a confidence interval of (X, ∞), which is useless in practice.

Fisher Information and the Cramér–Rao Bound

We now present an important fact about maximum likelihood estimators.

Theorem 8.2 (Asymptotic normality of the MLE) Let $\hat{\theta}_n$ denote the maximum likelihood estimator of θ based on an independent and identically distributed sample of size n, where θ represents the unknown parameter. Under suitable regularity conditions, we have

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

where

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

is the **Fisher information**.

This is a highly abstract result, so we illustrate it with a simple example.

Example 8.3 Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$ with density

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

Then

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

and

$$\frac{\partial}{\partial \mu} \log(f(x)) = \frac{x-\mu}{\sigma^2} \qquad \text{and} \qquad \frac{\partial^2}{\partial \mu^2} \log(f(x)) = -\frac{1}{\sigma^2}.$$

Theorem 8.2 says that $\hat{\theta}_n = \bar{X}$, $\theta = \mu$, and

$$I(\theta) = \frac{1}{\sigma^2}.$$

Therefore,

$$\sqrt{n \cdot \frac{1}{\sigma^2}}(\bar{X} - \mu) \xrightarrow{D} \mathcal{N}(0,1).$$

Observe that

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

so the convergence is in fact exact. The distribution does not merely converge; it is already $\mathcal{N}(0,1)$ for all n.

We note several important consequences of Theorem 8.2. Convergence in distribution alone does not guarantee convergence of moments. Under the additional regularity and uniform-integrability conditions that justify such convergence, for large n,

$$\mathbb{E}\left[\sqrt{n \cdot I(\theta)}(\hat{\theta}_n - \theta)\right] \approx 0,$$

which implies

$$\mathbb{E}\left[\hat{\theta}_n\right] \approx \theta.$$

Thus regular maximum likelihood estimators are asymptotically unbiased. Under suitable regularity conditions, Theorem 8.2 remains valid if we replace $I(\theta)$ with the plug-in quantity $I(\hat{\theta}_n)$.

Second, for large n,

$$\mathrm{Var}(\sqrt{n \cdot I(\theta)}(\hat{\theta}_n - \theta)) \approx 1,$$

which implies

$$n \cdot I(\theta) \cdot \mathrm{Var}(\hat{\theta}_n) \approx 1,$$

and therefore

$$\mathrm{Var}(\hat{\theta}_n) \approx \frac{1}{n \cdot I(\theta)}.$$

This quantity is called the **asymptotic variance**.

Third,

$$\sqrt{n \cdot I(\theta)}(\hat{\theta}_n - \theta)$$

is an approximate pivotal quantity. This allows us to construct an approximate confidence interval for θ of the form

$$\hat{\theta}_n \pm \zeta_\alpha \frac{1}{\sqrt{n \cdot I(\hat{\theta}_n)}},$$

where ζ_α is determined from the $\mathcal{N}(0,1)$ distribution.

Finally, under the usual differentiability and support regularity conditions, the **Cramér–Rao bound** states that if $\hat{\theta}_n$ is unbiased for θ and has finite variance, then

$$\mathrm{Var}(\hat{\theta}_n) \geq \frac{1}{n \cdot I(\theta)}.$$

Regular maximum likelihood estimators have this variance to first order and are therefore asymptotically efficient, which helps explain why maximum likelihood estimation is so widely used in practice.

Lecture 9 — Tuesday, February 04, 2014

Sample Size Calculations

Example 9.1 Suppose $X \sim \mathrm{Bin}(n, p)$. For instance, X might represent the number of "yes" answers to a certain yes-or-no survey question. We want to estimate p with a high degree of accuracy. How large should n be? More precisely, what value of n gives the approximate planning target $\mathbb{P}(|\hat{p} - p| < 0.01) \approx 99\%$?

We know that $\hat{p} = \frac{X}{n}$ is both the maximum likelihood estimator and the method of moments estimator for p. Using a normal approximation, we target

$$\mathbb{P}\left(\left|\frac{X}{n} - p\right| < 0.01\right) \approx 99\%,$$

which implies

$$\mathbb{P}(-0.01n < X - np < 0.01n) \approx 99\%.$$

We can apply Theorem 7.9 to X, yielding

$$\mathbb{P}\left(\frac{-0.01n}{\sqrt{np(1-p)}} < \frac{X - np}{\sqrt{np(1-p)}} < \frac{0.01n}{\sqrt{np(1-p)}}\right) \approx 99\%.$$

Theorem 7.9 tells us that the middle term converges in distribution to $\mathcal{N}(0, 1)$. The central 99% target area is shown in Figure 4.

FIGURE 4

From standard normal distribution tables or statistical software, we find that we require

$$\frac{0.01n}{\sqrt{np(1-p)}} \geq 2.576.$$

This expression still depends on p, but we can use a conservative bound in practice. Since $p(1-p)$ is a quadratic function with maximum value $1/4$ at $p = 1/2$, we substitute this value to obtain

$$0.01\sqrt{n} = (2.576)(0.5),$$

which yields

$$n \approx 16{,}590.$$

Rounding up gives the conservative approximate planning value $n = 16{,}590$. This is surprisingly large. Political polls usually tolerate a margin of error closer to three percentage points than one, which requires a much smaller sample size. An exact finite-sample guarantee would require a binomial calculation and could differ slightly because of discreteness.

Hypothesis Testing

Consider a situation where we have two **hypotheses** for a parameter: $H_0 : \theta = \theta_0$ (the **null hypothesis**) and $H_A : \theta = \theta_A$ (the **alternative hypothesis**).

Example 9.2 Suppose $X \sim \text{Bin}(n, p)$ with hypotheses $H_0 : p = 0.5$ and $H_A : p = 0.9$. Intuitively, we would reject H_0 if $X \geq c$ for some threshold c (to be determined), and otherwise fail to reject it. For concreteness, suppose $n = 10$ and consider $c = 7$.

Before proceeding, we need to establish some definitions.

A **type I error** occurs when we reject H_0 even though H_0 is true. A **type II error** occurs when we fail to reject H_0 even though H_A is true. The **significance level** is $\mathbb{P}(\text{type I error})$, and the **power** is $1 - \mathbb{P}(\text{type II error})$.

Example 9.3 Returning to Example 9.2, the significance level equals

$$\mathbb{P}(\text{type I error}) = \mathbb{P}(X \geq 7 \mid H_0) = \sum_{x=7}^{10} \binom{10}{x} 0.5^x (1 - 0.5)^{10-x} = 0.172.$$

Similarly, the power equals

$$\begin{aligned} 1 - \mathbb{P}(\text{type II error}) &= 1 - \mathbb{P}(X < 7 \mid H_A) \\ &= 1 - \sum_{x=0}^{6} \binom{10}{x} (0.9)^x (1 - 0.9)^{10-x} \\ &= 1 - 0.013 \\ &= 0.987. \end{aligned}$$

Observe that a hypothesis test is fully characterized by its **rejection region** R, the subset of the sample space for which an observation of data in R would lead us to reject H_0 in favour of H_A.

Likelihood Ratio Tests and the Neyman–Pearson Lemma

The **likelihood ratio test** provides a systematic approach to hypothesis testing. We reject the null hypothesis H_0 if

$$\Lambda = \frac{L(\theta_A)}{L(\theta_0)} > c_\alpha$$

for a suitable critical value c_α. We choose c_α so that the significance level is a predetermined value α (often $\alpha = 0.05$). Note that the **likelihood ratio** Λ is also written as $\frac{f_A(x)}{f_0(x)}$.

Theorem 9.4 (Neyman–Pearson lemma) For testing one simple hypothesis against another, a likelihood ratio test of level α is most powerful among all tests of level at most α. In a discrete model, randomization on the likelihood-ratio boundary may be needed to attain size exactly α.

Proof. For simplicity, suppose a nonrandomized threshold attains size exactly α. Let R_1 be its rejection region:

$$R_1 = \left\{ x \mid \frac{f_A(x)}{f_0(x)} \geq c_\alpha \right\}.$$

Let R_2 be another rejection region with level at most α, so

$$\mathbb{P}(R_2 \mid H_0) \leq \mathbb{P}(R_1 \mid H_0) = \alpha,$$

equivalently,

$$\int_{R_2} f_0(x)\,\mathrm{d}x \leq \int_{R_1} f_0(x)\,\mathrm{d}x.$$

Define $C_1 = R_1 \setminus R_2$ and $C_2 = R_2 \setminus R_1$. Then

$$\int_{C_2} f_0(x)\,\mathrm{d}x \leq \int_{C_1} f_0(x)\,\mathrm{d}x.$$

To prove optimality, it suffices to show

$$\int_{C_1} f_A(x)\,\mathrm{d}x \geq \int_{C_2} f_A(x)\,\mathrm{d}x,$$

since this is equivalent to $\mathbb{P}(R_1 \mid H_A) \geq \mathbb{P}(R_2 \mid H_A)$. On $C_1 \subseteq R_1$, we have $f_A \geq c_\alpha f_0$. On $C_2 \subseteq R_1^c$, we have $f_A \leq c_\alpha f_0$. Therefore,

$$\int_{C_1} f_A(x)\,\mathrm{d}x \geq \int_{C_1} c_\alpha f_0(x)\,\mathrm{d}x \geq \int_{C_2} c_\alpha f_0(x)\,\mathrm{d}x \geq \int_{C_2} f_A(x)\,\mathrm{d}x.$$

Hence the likelihood ratio test is most powerful among level-α tests. □

Example 9.5 Suppose $X_1, \ldots, X_n$ are independent and identically distributed as $\mathcal{N}(\mu, \sigma^2)$ with σ^2 known, and consider $H_0 : \mu = \mu_0$ versus $H_A : \mu = \mu_A$. The likelihood ratio is

$$\Lambda = \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu_A)^2}{2\sigma^2}}}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu_0)^2}{2\sigma^2}}}.$$

Rejecting for large Λ is equivalent to requiring

$$\begin{aligned} -\sum_{i=1}^{n}(X_i - \mu_A)^2 + \sum_{i=1}^{n}(X_i - \mu_0)^2 &\text{ to be large} \\ \iff 2(\mu_A - \mu_0)\sum_{i=1}^{n} X_i &\text{ to be large} \\ \iff 2(\mu_A - \mu_0)n\bar{X} &\text{ to be large.} \end{aligned}$$

Therefore, if $\mu_A > \mu_0$, we reject for large $\bar{X}$; if $\mu_A < \mu_0$, we reject for small $\bar{X}$.

Lecture 10 — Thursday, February 06, 2014

Power Analyses

Recall from Example 9.5 that, in the canonical example $\mathcal{N}(\mu, \sigma^2)$ with σ^2 known, the likelihood ratio test prescribes the following:

1. If $\mu_A > \mu_0$, reject H_0 when $\bar{X}$ is large.
2. If $\mu_A < \mu_0$, reject H_0 when $\bar{X}$ is small.

We choose thresholds for "large" and "small" to make the significance level equal to α. Without loss of generality, assume $\mu_A > \mu_0$. We wish to choose a threshold c such that

$$\mathbb{P}(\bar{X} > c \mid H_0) = \alpha,$$

where H_0 is the hypothesis that $\mu = \mu_0$. Under H_0, we have

$$\bar{X} \sim \mathcal{N}(\mu_0, \sigma^2/n) \qquad \text{and} \qquad \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Therefore,

$$\mathbb{P}\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{c - \mu_0}{\sigma/\sqrt{n}}\right) = \alpha.$$

Let $z_{1-\alpha}$ satisfy $\mathbb{P}(Z > z_{1-\alpha}) = \alpha$ for $Z \sim \mathcal{N}(0,1)$. Then $c = \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}$. For instance, if we want $\alpha = 0.05$, then we reject H_0 when

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > 1.645,$$

or equivalently, when

$$\bar{X} > \mu_0 + 1.645\frac{\sigma}{\sqrt{n}}.$$

The power of the likelihood ratio test equals

$$\mathbb{P}\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha} \mid H_A\right) = \mathbb{P}\left(\frac{\bar{X} - \mu_A}{\sigma/\sqrt{n}} > \frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} + z_{1-\alpha}\right).$$

By convention, we denote $\Phi(\cdot)$ as the cumulative distribution function of $\mathcal{N}(0,1)$. Define $\Delta := \mu_A - \mu_0 > 0$. Then the power equals

$$1 - \Phi\left(\frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} + z_{1-\alpha}\right) = 1 - \Phi\left(z_{1-\alpha} - \frac{\Delta\sqrt{n}}{\sigma}\right).$$

Observe that if Δ increases, then power increases (because Φ is increasing). If σ increases, power decreases. The ratio

$$\frac{\Delta}{\sigma}$$

is commonly called the **signal-to-noise ratio**. If n increases, power increases; if the critical quantile increases, power decreases. With the model and sample size fixed, moving the rejection threshold trades type I error against type II error; both cannot be reduced simultaneously merely by moving that threshold. In Figure 5, region A is the type I error under H_0, regions $B + C$ form the type II error under H_A, and region D is the power under H_A.

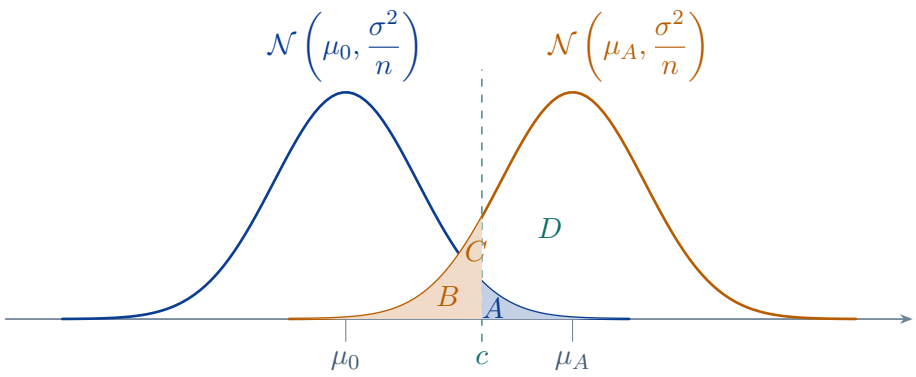


FIGURE 5

Composite Hypotheses

So far we have considered hypotheses of the form $H_0 : \mu = \mu_0$ and $H_A : \mu = \mu_A$. These are called **simple hypotheses**. We now consider the **composite** alternative

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_A : \mu \neq \mu_0 \end{cases}$$

Here, H_A simply states that the null hypothesis is wrong but does not specify a particular alternative value.

For composite hypotheses, we can construct a **generalized likelihood ratio test**. Let Θ be the unrestricted parameter space under the full model, including both the null and alternative possibilities, and let

$$\Theta_0 := \{\theta \mid \theta \text{ is allowed under } H_0\}.$$

Then the generalized likelihood ratio statistic is

$$\Lambda = \frac{\sup_{\theta \in \Theta} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)}.$$

As before, we reject H_0 if $\Lambda > c_\alpha$.

Example 10.1 Returning to Example 9.5, we have $\Lambda = \frac{L(\hat{\mu}_{\text{MLE}})}{L(\mu_0)}$. We know that $\hat{\mu}_{\text{MLE}} = \bar{X}$.

Therefore

$$\begin{aligned}
\Lambda &= \frac{(\sqrt{2\pi}\sigma)^{-n} e^{-\sum_{i=1}^{n}(X_i-\bar{X})^2/(2\sigma^2)}}{(\sqrt{2\pi}\sigma)^{-n} e^{-\sum_{i=1}^{n}(X_i-\mu_0)^2/(2\sigma^2)}} \\
&= \exp\left(\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2 - \sum_{i=1}^{n}(X_i-\bar{X})^2}{2\sigma^2}\right) \\
&= \exp\left(\frac{\sum_{i=1}^{n}(X_i^2 - 2\mu_0 X_i + \mu_0^2) - \left(\sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2\right)}{2\sigma^2}\right) \\
&= \exp\left(\frac{n(\bar{X}^2 - 2\mu_0\bar{X} + \mu_0^2)}{2\sigma^2}\right) \\
&= \exp\left(\frac{n(\bar{X}-\mu_0)^2}{2\sigma^2}\right).
\end{aligned}$$

We reject H_0 if Λ is large, which is equivalent to rejecting when $2\log\Lambda$ is large. This occurs when

$$\frac{n(\bar{X}-\mu_0)^2}{\sigma^2}$$

is large, or equivalently when

$$\frac{|\bar{X}-\mu_0|}{\sigma/\sqrt{n}}$$

is large. Thus we reject H_0 if

$$\bar{X} > \mu_0 + c_\alpha\frac{\sigma}{\sqrt{n}} \qquad \text{or} \qquad \bar{X} < \mu_0 - c_\alpha\frac{\sigma}{\sqrt{n}}.$$

If $\mu_A > \mu_0$ (or $\mu_A < \mu_0$), we have a **one-sided test**. If $\mu_A \neq \mu_0$, we have a **two-sided test**. For $\alpha = 0.05$, we have $c_\alpha = 1.645$ for a one-sided test and $c_\alpha = 1.96$ for a two-sided test. For a one-sided test, we reject if

$$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > 1.645,$$

whereas for a two-sided test, we reject if

$$\frac{|\bar{X} - \mu_0|}{\sigma/\sqrt{n}} > 1.96.$$

Interval Estimation versus Hypothesis Testing

For the canonical case where $X_1, \ldots, X_n$ are independent and identically distributed as $\mathcal{N}(\mu, \sigma^2)$ with σ^2 known, our confidence interval for μ is

$$\left(\bar{X} - c_\alpha \frac{\sigma}{\sqrt{n}}, \bar{X} + c_\alpha \frac{\sigma}{\sqrt{n}}\right).$$

Our two-sided rejection region for testing $\mu = \mu_0$ consists of

$$\bar{X} > \mu_0 + c_\alpha \frac{\sigma}{\sqrt{n}} \qquad \text{or} \qquad \bar{X} < \mu_0 - c_\alpha \frac{\sigma}{\sqrt{n}},$$

which is equivalent to

$$\mu_0 < \bar{X} - c_\alpha \frac{\sigma}{\sqrt{n}} \qquad \text{or} \qquad \mu_0 > \bar{X} + c_\alpha \frac{\sigma}{\sqrt{n}}.$$

Hence we reject $H_0 : \mu = \mu_0$ if and only if μ_0 is not in the confidence interval. Equivalently, the complement of our rejection region is precisely our confidence interval. Therefore interval estimation and hypothesis testing are fundamentally equivalent in this situation.

Generalized Likelihood Ratio Tests and Wilks' Theorem

It turns out that the quantity $2 \log \Lambda$ is very important. To motivate the role, recall that if $Z \sim \mathcal{N}(0,1)$, then $Z^2 \sim \chi^2_{(1)}$; and if $Z_1, \ldots, Z_n \overset{iid}{\sim} \mathcal{N}(0,1)$, then

$$\sum_{i=1}^{n} Z_i^2 \sim \chi^2_{(n)}.$$

Theorem 10.2 (Wilks' theorem) Under the usual smoothness, identifiability, and interior-point regularity conditions,

$$2 \log \Lambda \xrightarrow{D} \chi^2_{(df)},$$

where $df := \dim(\Theta) - \dim(\Theta_0)$. Here, the dimension is the number of free parameters in the corresponding model.

Example 10.3 For $\mathcal{N}(\mu, \sigma^2)$ with σ^2 known, we have

$$2\log\Lambda = \frac{(\bar{X} - \mu_0)^2}{\sigma^2/n} \sim (\mathcal{N}(0,1))^2 = \chi^2_{(1)}.$$

We have $\dim(\Theta) = 1$ and $\dim(\Theta_0) = 0$.

Under H_0, in this case $2\log\Lambda \sim \chi^2_{(1)}$ exactly (not just asymptotically). We would reject H_0 if $2\log\Lambda > c_\alpha$, where c_α is obtained from a $\chi^2_{(1)}$ table. For instance, when $\alpha = 0.05$, we get $c_\alpha = 3.84$. Observe that $3.84 = (1.96)^2$, which is not a coincidence.

Lecture 11 — Tuesday, February 11, 2014

Recall from Theorem 10.2 that $2\log\Lambda \xrightarrow{D} \chi^2_{(df)}$. This makes

$$2\ell(\hat{\theta}_{\text{MLE}}) - 2\ell(\theta)$$

an approximate pivotal quantity. For a composite two-sided test (where $H_0 : \theta = \theta_0$ and $H_A : \theta \neq \theta_0$), the set

$$\{\theta \mid 2\ell(\hat{\theta}_{\text{MLE}}) - 2\ell(\theta) < c_\alpha\}$$

therefore forms an approximate $(1-\alpha)\cdot 100\%$ confidence set.

Example 11.1 Suppose we have two independent samples with $\lambda_1, \lambda_2 \geq 0$:

$$X_1, \ldots, X_m \overset{\text{iid}}{\sim} \text{Poisson}(\lambda_1) \qquad \text{and} \qquad Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda_2).$$

We test $H_0 : \lambda_1 = \lambda_2$ against $H_A : \lambda_1 \neq \lambda_2$.

The likelihood ratio is

$$\Lambda = \frac{\sup_{\lambda_1,\lambda_2} L(\lambda_1, \lambda_2)}{\sup_{\lambda_1=\lambda_2} L(\lambda_1, \lambda_2)}.$$

For the Poisson distribution,

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{Z_i}}{Z_i!} \propto e^{-n\lambda}\lambda^{\sum_{i=1}^{n} Z_i},$$

where $Z_1, \ldots, Z_n \sim \text{Poisson}(\lambda)$.

For a positive total count, the maximizer is interior. Then

$$\ell(\lambda) = -n\lambda + \sum_{i=1}^{n} Z_i \log(\lambda),$$

so

$$\ell'(\lambda) = -n + \frac{\sum_{i=1}^{n} Z_i}{\lambda} = 0,$$

and therefore

$$\hat{\lambda}_{\text{MLE}} = \bar{Z}.$$

If every count is zero, the likelihood is maximized at the boundary $\widehat{\lambda} = 0$, so the same formula remains valid.

Under the unrestricted model, we have $\hat{\lambda}_1 = \bar{X}$ and $\hat{\lambda}_2 = \bar{Y}$. Under H_0, we obtain the **pooled estimate**

$$\hat{\lambda}_1 = \hat{\lambda}_2 = \frac{\sum_{i=1}^{m} X_i + \sum_{j=1}^{n} Y_j}{m+n}.$$

Letting $\hat{\lambda}$ be the common value of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ above, the likelihood ratio test statistic is

$$\begin{aligned}\Lambda &= \frac{\prod_{i=1}^{m} \frac{e^{-\bar{X}}(\bar{X})^{X_i}}{X_i!} \prod_{j=1}^{n} \frac{e^{-\bar{Y}}(\bar{Y})^{Y_j}}{Y_j!}}{\prod_{i=1}^{m} \frac{e^{-\hat{\lambda}}(\hat{\lambda})^{X_i}}{X_i!} \prod_{j=1}^{n} \frac{e^{-\hat{\lambda}}(\hat{\lambda})^{Y_j}}{Y_j!}} \\ &= \frac{e^{-m\bar{X}}(\bar{X})^{m\bar{X}} e^{-n\bar{Y}}(\bar{Y})^{n\bar{Y}}}{e^{-(m+n)\hat{\lambda}}(\hat{\lambda})^{(m+n)\hat{\lambda}}}\end{aligned}$$

$$= \frac{(\bar{X})^{m\bar{X}}(\bar{Y})^{n\bar{Y}}}{\left(\frac{m\bar{X}+n\bar{Y}}{m+n}\right)^{m\bar{X}+n\bar{Y}}}$$

where 0^0 is interpreted by continuity as 1. Thus

$$\log\Lambda = m\bar{X}\log\left(\frac{(m+n)\bar{X}}{m\bar{X}+n\bar{Y}}\right) + n\bar{Y}\log\left(\frac{(m+n)\bar{Y}}{m\bar{X}+n\bar{Y}}\right),$$

with the convention $0\log 0 = 0$. We reject H_0 if $2\log\Lambda > 3.84$ at an approximate significance level of 0.05. The chi-square calibration presumes that the true common rate is positive and that the expected counts are not too small.

If we assume $n = m$, we can simplify $\log\Lambda$ to

$$\log\Lambda = n\left[\bar{X}\log\frac{2\bar{X}}{\bar{X}+\bar{Y}} + \bar{Y}\log\frac{2\bar{Y}}{\bar{X}+\bar{Y}}\right].$$

Let us verify our intuition that $\log\Lambda \geq 0$, with equality if and only if $\bar{X} = \bar{Y}$. If $\bar{X}+\bar{Y} > 0$, dividing by the positive quantity $n(\bar{X}+\bar{Y})$ does not change the minimizer. If both means are zero, then $\Lambda = 1$ directly.

We can write

$$p = \frac{\bar{X}}{\bar{X}+\bar{Y}} \qquad \text{and} \qquad 1-p = \frac{\bar{Y}}{\bar{X}+\bar{Y}}.$$

Then

$$g(p) := \frac{\log\Lambda}{n(\bar{X}+\bar{Y})} = p\log(2p) + (1-p)\log(2(1-p)).$$

Therefore

$$g'(p) = \log(2p) + 1 - \log(2(1-p)) - 1 = \log p - \log(1-p).$$

Thus

$$g'(p) = 0 \implies p = 1-p \implies p = \frac{1}{2},$$

and $p = 1/2$ means the same thing as $\bar{X} = \bar{Y}$. Moreover,

$$g''(p) = \frac{1}{p} + \frac{1}{1-p} > 0,$$

and hence $g(p)$ is minimized at $p = \frac{1}{2}$, with

$$g\left(\frac{1}{2}\right) = 0.$$

We still need the asymptotic distribution of $2\log\Lambda$ to decide *how different* $\bar{X}$ and $\bar{Y}$ must be before we reject H_0.

P-values

In a hypothesis test, choosing a significance level gives a rejection rule, but it does not indicate how strongly the observed data contradict H_0. To measure this strength of evidence, we compare the observed test statistic to its null distribution.

Let T be a test statistic, and suppose that larger values of T provide stronger evidence against H_0. After observing T_{obs}, define the ***p*-value**

$$p\text{-value} := \mathbb{P}(T \geq T_{\text{obs}} \mid H_0).$$

In words, this is the probability under H_0 of obtaining a value of the test statistic at least as extreme as the one observed. Equivalently, it is the smallest significance level at which H_0 would be rejected.

For the likelihood ratio statistic Λ used above,

$$p\text{-value} = \mathbb{P}(\Lambda \geq \Lambda_{\text{obs}} \mid H_0).$$

Example 11.2 Suppose $X_1, X_2, \ldots, X_5$ are independent and identically distributed as $\mathcal{N}(\mu, 5)$, so $\bar{X} \sim \mathcal{N}(\mu, 1)$. We test $H_0 : \mu = 0$ against $H_A : \mu \neq 0$, and reject when $|\bar{X}| > 1.96$ at significance level 0.05. There is clearly more evidence against H_0 if $|\bar{X}_{\text{obs}}| = 3$ than if $|\bar{X}_{\text{obs}}| = 2$. Therefore, the p-value here is

$$p\text{-value} = \mathbb{P}(|\bar{X}| \geq |\bar{X}_{\text{obs}}| \mid H_0).$$

Figures 6 and 7 compare the two-sided tail regions when $|\bar{X}_{\text{obs}}|$ equals 2 and 3, respectively.

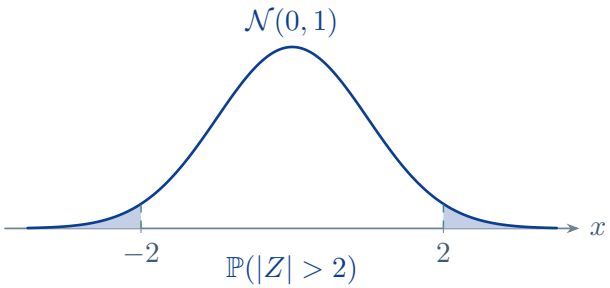


Figure 6

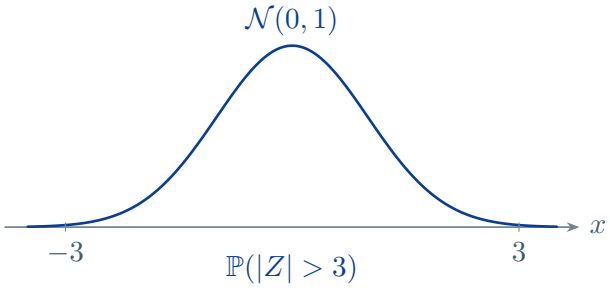


Figure 7

Lecture 12 — Thursday, February 13, 2014

4. Classical Inference for Normal Models

Student's t-Distribution and the One-Sample t-Test

So far, we have been studying the canonical case where $X_1, \ldots, X_n$ are independent and identically distributed as $\mathcal{N}(\mu, \sigma^2)$ with σ^2 known. In this setting,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is a pivotal quantity. We now consider a more realistic scenario where σ^2 is unknown. In this case, the pivotal quantity for the known-variance normal model is no longer valid.

Let

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

be the **sample variance**. The pivotal quantity

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}.$$

For testing $H_0 : \mu = \mu_0$, the **one-sample *t*-test** substitutes μ_0 for μ in this statistic.

What is this distribution?

Definition 12.1 Suppose $Z \sim \mathcal{N}(0,1)$ and $U \sim \chi^2_{(n)}$ are independent random variables. Then

$$T = \frac{Z}{\sqrt{U/n}}$$

is distributed according to $t_{(n)}$, the **(Student's) *t*-distribution** with n degrees of freedom.

Theorem 12.2 Let $X_1, \ldots, X_n$ be independent and identically distributed as $\mathcal{N}(\mu, \sigma^2)$, where $n \geq 2$ and $\sigma > 0$. Then the following hold:

1. $$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{(n-1)}.$$

2. $$\bar{X} \qquad \text{and} \qquad \sum_{i=1}^{n}(X_i - \bar{X})^2$$

   are independent, which we denote as

   $$\bar{X} \perp \sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Given Theorem 12.2, we can write

$$T = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{S/\sigma}.$$

The numerator has distribution $\mathcal{N}(0,1)$. The denominator equals

$$\frac{S}{\sigma} = \sqrt{\frac{S^2}{\sigma^2}} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)\sigma^2}} \sim \sqrt{\frac{\chi^2_{(n-1)}}{n-1}},$$

by (1) of Theorem 12.2. Furthermore, the numerator and denominator are independent by (2). Hence

$$T \sim t_{(n-1)}.$$

Proof. Let $Z_i = (X_i - \mu)/\sigma$, so $Z = (Z_1, \ldots, Z_n)^\mathsf{T} \sim \mathcal{N}(0, I_n)$. Choose an orthogonal matrix Q whose first row is $n^{-1/2}(1, \ldots, 1)$, and set $W = QZ$. Orthogonal invariance of the standard multivariate normal distribution gives $W \sim \mathcal{N}(0, I_n)$, so its coordinates are independent standard normal random variables. Moreover,

$$W_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}.$$

Orthogonality also preserves squared length, and subtracting the squared component in the all-ones direction gives

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} = \sum_{i=1}^{n} Z_i^2 - W_1^2 = \sum_{j=2}^{n} W_j^2 \sim \chi^2_{(n-1)}.$$

This proves (1). The sample mean depends only on W_1, while the residual sum of squares depends only on $W_2, \ldots, W_n$; their independence proves (2). □

Let us summarize what we know about the unknown-variance normal model.

1. Recall the identity

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2.$$

We have already shown that

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

Therefore, an unbiased estimator of σ^2 is

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

whereas

$$\hat{\sigma}^2_{\mathrm{MLE}} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}$$

is slightly biased.

2. The quantity

$$\frac{\bar{X}-\mu}{S/\sqrt{n}}$$

is a new pivotal quantity. Consequently,

$$\left(\bar{X}-c_\alpha\frac{S}{\sqrt{n}}, \bar{X}+c_\alpha\frac{S}{\sqrt{n}}\right)$$

is a confidence interval for μ, where c_α is obtained from a $t_{(n-1)}$ table. For $\alpha = 0.05$, we have $c_\alpha \approx 2$ when n is relatively large, and $c_\alpha \to 1.96$ as $n \to \infty$.

3. The one-sample t-test rejects $H_0 : \mu = \mu_0$ if

$$\frac{|\bar{X}-\mu_0|}{S/\sqrt{n}} > c_\alpha.$$

4. The one-sample t-test is equivalent to the likelihood ratio test. Under the unrestricted model,

$$\hat{\mu}_{\mathrm{MLE}} = \bar{X} \qquad \text{and} \qquad \hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

Under H_0,

$$\tilde{\mu}_{\text{MLE}} = \mu_0 \qquad \text{and} \qquad \tilde{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2 = \hat{\sigma}_0^2.$$

Thus

$$\begin{aligned}
\Lambda &= \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\hat{\sigma}_A} \exp\left(-\frac{(X_i - \bar{X})^2}{2\hat{\sigma}_A^2}\right)}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\hat{\sigma}_0} \exp\left(-\frac{(X_i - \mu_0)^2}{2\hat{\sigma}_0^2}\right)} \\
&= \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_A^2}\right)^{n/2} \frac{\exp\left(-\dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{2\hat{\sigma}_A^2}\right)}{\exp\left(-\dfrac{\sum_{i=1}^{n}(X_i - \mu_0)^2}{2\hat{\sigma}_0^2}\right)} \\
&= \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_A^2}\right)^{n/2}.
\end{aligned}$$

Therefore,

$$2\log\Lambda = 2\log\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_A^2}\right)^{n/2} = n\log\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_A^2}\right) = n\log\left[\frac{\sum_{i=1}^{n}(X_i - \mu_0)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right].$$

Also,

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \mu_0)^2 &= \sum_{i=1}^{n}\left[(X_i - \bar{X}) + (\bar{X} - \mu_0)\right]^2 \\
&= \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(\bar{X} - \mu_0)^2 + 2\sum_{i=1}^{n}(X_i - \bar{X})(\bar{X} - \mu_0) \\
&= \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2,
\end{aligned}$$

since $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$ and $\sum_{i=1}^{n}(\bar{X} - \mu_0)^2 = n(\bar{X} - \mu_0)^2$. Hence

$$2\log\Lambda = n\log\left[1 + \frac{n(\bar{X} - \mu_0)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right].$$

Now recall that

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}},$$

so

$$T^2 = \frac{n(\bar{X} - \mu_0)^2}{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}.$$

Hence

$$2\log\Lambda = n\log\left[1 + \frac{T^2}{n-1}\right].$$

It follows that $2\log\Lambda$ is large if and only if T^2 is large, which holds if and only if $|T|$ is large. Therefore, the likelihood ratio test for this model is equivalent to the t-test.

We also have

$$2\log\Lambda \xrightarrow{D} \chi^2_{(1)}.$$

Here $\dim(\Theta) = 2$ (that is, μ and σ^2) and $\dim(\Theta_0) = 1$ (that is, σ^2). A rigorous proof of the limiting chi-square statement combines the limiting normal distribution of the t-statistic with a Taylor expansion and Slutsky's theorem. That convergence theorem is beyond the scope of this course; the exact identity above is all that is needed to see that the likelihood ratio test and the t-test have the same rejection regions.

Lecture 13 — Tuesday, February 25, 2014

Two-Sample t-Test

Suppose

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu_A, \sigma^2) \qquad \text{and} \qquad Y_1, Y_2, \ldots, Y_m \overset{\text{iid}}{\sim} \mathcal{N}(\mu_B, \sigma^2),$$

where $\sigma^2 > 0$ is unknown, $n, m \geq 2$, and the two samples are independent of each other. We are interested in testing whether their population means agree:

$$H_0 : \mu_A = \mu_B \qquad \text{versus} \qquad H_A : \mu_A \neq \mu_B.$$

In practice, the null hypothesis often receives preferential treatment (analogous to the presumption of innocence). Note that we can equivalently write $H_0 : \mu_A - \mu_B = 0$.

We use an approach similar to the one-sample t-test. Recall that for the one-sample t-test, we first compute the maximum likelihood estimators of μ and σ^2. We then replace the maximum likelihood estimator of σ^2 with the unbiased estimator

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}.$$

Finally, we observe that

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Since

$$\operatorname{Var}(\bar{X}) = \frac{\sigma^2}{n},$$

the standard deviation of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$. Replacing σ with S yields the **standard error**

$$\frac{S}{\sqrt{n}}.$$

This final step amounts to using a test statistic of the form

$$\frac{\text{MLE}}{\text{standard error of maximum likelihood estimator}} \sim t_{(\text{degrees of freedom})}.$$

For the **two-sample *t*-test**, we proceed analogously. First, we construct the likelihood function:

$$L(\mu_A, \mu_B, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X_i - \mu_A)^2}{2\sigma^2}\right) \cdot \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y_j - \mu_B)^2}{2\sigma^2}\right).$$

The log-likelihood is

$$\ell(\mu_A, \mu_B, \sigma^2) = -n \log \sigma - \frac{\sum_{i=1}^{n}(X_i - \mu_A)^2}{2\sigma^2} - m \log \sigma - \frac{\sum_{j=1}^{m}(Y_j - \mu_B)^2}{2\sigma^2} + \text{constant}.$$

Setting $\frac{\partial \ell}{\partial \mu_A} = 0$ yields

$$\sum_{i=1}^{n}(X_i - \mu_A) = 0 \implies \hat{\mu}_A = \bar{X}.$$

Similarly,

$$\frac{\partial \ell}{\partial \mu_B} = 0 \implies \sum_{j=1}^{m}(Y_j - \mu_B) = 0 \implies \hat{\mu}_B = \bar{Y}.$$

Finally,

$$\frac{\partial \ell}{\partial \sigma} = \frac{-n - m}{\sigma} + \frac{\sum_{i=1}^{n}(X_i - \mu_A)^2}{\sigma^3} + \frac{\sum_{j=1}^{m}(Y_j - \mu_B)^2}{\sigma^3} = 0,$$

which gives

$$\hat{\sigma}^2_{\text{MLE}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2}{n + m}.$$

Next, we construct an unbiased estimator of σ^2. Recall that

$$\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = (n-1)\sigma^2 \qquad \text{and} \qquad \mathbb{E}\left[\sum_{j=1}^{m}(Y_j - \bar{Y})^2\right] = (m-1)\sigma^2.$$

Therefore, an unbiased estimator of σ^2 is

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2}{n + m - 2}.$$

Finally, we compute the test statistic. We have $\hat{\mu}_A - \hat{\mu}_B = \bar{X} - \bar{Y}$. Since

$$\text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m},$$

the standard deviation is

$$\sigma\sqrt{\frac{1}{n} + \frac{1}{m}},$$

and the standard error is

$$S\sqrt{\frac{1}{n} + \frac{1}{m}}.$$

Therefore

$$\frac{(\bar{X} - \bar{Y}) - (\mu_A - \mu_B)}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{(n+m-2)}.$$

Under H_0, the centering term $\mu_A - \mu_B$ is zero. Here the residual degrees of freedom equal the total sample size $n + m$ minus the two fitted means, giving $n + m - 2$.

For the two-sample t-test, the same equivalence with the likelihood ratio test can be derived analogously to the one-sample case.

Lecture 14 — Tuesday, March 04, 2014

5. Regression and Categorical Data Analysis

Simple Linear Regression

Consider the model $Y_i \overset{\text{indep}}{\sim} \mathcal{N}(\alpha + \beta X_i, \sigma^2)$ for $i = 1, \ldots, n$, where $\sigma > 0$, $n \geq 3$, and the observed covariates are not all equal. We condition throughout on the observed X_i values, so all expectations and variances below are conditional on them, with that conditioning suppressed from the notation. The key hypothesis test of interest is $H_0 : \beta = 0$ versus $H_A : \beta \neq 0$.

We begin by constructing the maximum likelihood estimators for α and β. The likelihood and log-likelihood are

$$L(\alpha, \beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y_i - \alpha - \beta X_i)^2}{2\sigma^2}\right)$$

and

$$\ell(\alpha, \beta, \sigma^2) = -n \log(\sigma) - \frac{\sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2}{2\sigma^2} + \text{constant},$$

respectively. Setting $\frac{\partial \ell}{\partial \alpha} = 0$ yields

$$\sum_{i=1}^{n} (Y_i - \alpha - \beta X_i) = 0 \implies \hat{\alpha}_{\text{MLE}} = \bar{Y} - \hat{\beta}\bar{X}.$$

Setting $\frac{\partial \ell}{\partial \beta} = 0$ yields

$$-\sum_{i=1}^{n} (Y_i - \alpha - \beta X_i) X_i = 0.$$

After substituting $\alpha = \bar{Y} - \hat{\beta}\bar{X}$, this becomes

$$\sum_{i=1}^{n} (Y_i - \bar{Y} + \hat{\beta}\bar{X} - \hat{\beta} X_i) X_i = 0,$$

and therefore

$$\hat{\beta}_{\text{MLE}} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y}) X_i}{\sum_{i=1}^{n} (X_i - \bar{X}) X_i}.$$

Remark 14.1 As a useful identity, note that

$$\sum_{i=1}^{n} (Y_i - \bar{Y}) \bar{X} = 0.$$

Indeed,

$$\bar{X} \sum_{i=1}^{n} (Y_i - \bar{Y}) = \bar{X} \left(\sum_{i=1}^{n} Y_i - n\bar{Y} \right) = 0.$$

Similarly,

$$\sum_{i=1}^{n} (X_i - \bar{X}) \bar{X} = 0.$$

Therefore we can rewrite

$$\hat{\beta}_{\text{MLE}} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})X_i - \sum_{i=1}^{n}(Y_i - \bar{Y})\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})X_i - \sum_{i=1}^{n}(X_i - \bar{X})\bar{X}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

This alternative form has a nice interpretation: the covariance is

$$\text{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

and the variance is

$$\text{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

Essentially, the numerator approximates $\text{Cov}(X, Y)$ while the denominator approximates $\text{Var}(X)$.

Next, setting $\frac{\partial \ell}{\partial \sigma} = 0$ yields

$$\frac{-n}{\sigma} + \frac{\sum_{i=1}^{n}(Y_i - \hat{\alpha}_{\text{MLE}} - \hat{\beta}_{\text{MLE}}X_i)^2}{\sigma^3} = 0,$$

which gives

$$\hat{\sigma}^2_{\text{MLE}} = \frac{\sum_{i=1}^{n}(Y_i - \hat{\alpha}_{\text{MLE}} - \hat{\beta}_{\text{MLE}}X_i)^2}{n}.$$

An unbiased version of this estimator is

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2}{n - 2},$$

where the denominator $(n - 2)$ equals the sample size minus the number of parameters estimated from the mean part. The formal proof requires linear algebra, so we omit it here.

Next, we compute the standard error of $\hat{\beta}_{\text{MLE}}$, which requires first finding its variance. We have

$$\begin{aligned}\mathbb{E}\left[\hat{\beta}_{\text{MLE}}\right] &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})\mathbb{E}[Y_i]}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(\alpha + \beta X_i)}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ &= \beta,\end{aligned}$$

because $\sum_i (X_i - \bar{X}) = 0$ and $\sum_i (X_i - \bar{X})X_i = \sum_i (X_i - \bar{X})^2$.

Hence $\hat{\beta}_{\text{MLE}}$ is unbiased for β. For the variance, write

$$\text{Var}(\hat{\beta}_{\text{MLE}}) = \text{Var}\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) = \text{Var}\left(\sum_{i=1}^{n} a_i Y_i\right),$$

where

$$a_i := \frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Finally,

$$\begin{aligned}\text{Var}(\hat{\beta}_{\text{MLE}}) &= \sum_{i=1}^{n} a_i^2 \,\text{Var}(Y_i) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} a_i a_j \cdot 0 = \sum_{i=1}^{n} a_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^{n} a_i^2 \\ &= \sigma^2 \sum_{i=1}^{n}\left[\frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]^2 = \sigma^2 \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.\end{aligned}$$

Thus the standard deviation of $\hat{\beta}_{\text{MLE}}$ is

$$\frac{\sigma}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}},$$

and the corresponding standard error is

$$\frac{S}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}},$$

where

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2}{n-2}.$$

More generally,

$$\frac{\hat{\beta}_{\text{MLE}} - \beta}{S/\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} \sim t_{(n-2)}.$$

Thus, to test $\beta = 0$ versus $\beta \neq 0$, we substitute the null value 0 and reject for large absolute values of this statistic.

If the researcher has control over the data collection process, the X_i values should be spread out as much as the feasible design region allows to maximize the denominator

$$\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

which in turn minimizes $\text{Var}(\hat{\beta}_{\text{MLE}})$.

Lecture 15 — Thursday, March 06, 2014

Fitted Values and Prediction Intervals

Recall from the simple linear regression estimator formulas that

$$\hat{\beta}_{\text{MLE}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

and

$$\hat{\alpha}_{\text{MLE}} = \bar{Y} - \hat{\beta}_{\text{MLE}}\bar{X}.$$

$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ denotes the **fitted value**, an estimate of the conditional mean at X_i, and $R_i = Y_i - \hat{Y}_i$ denotes the **residual**. The key observation is that $\hat{Y}_i$ is a random variable, so we can quantify its uncertainty by computing its variance. At $X = X_0$, we want to compute

$$\operatorname{Var}(\hat{Y}_0) = \operatorname{Var}(\hat{\alpha} + \hat{\beta}X_0) = \operatorname{Var}(\hat{\alpha}) + X_0^2 \operatorname{Var}(\hat{\beta}) + 2X_0 \operatorname{Cov}(\hat{\alpha}, \hat{\beta}).$$

The fitted line and its vertical residuals are illustrated in Figure 8. Each observed response is decomposed into its fitted value on the line and a signed residual.

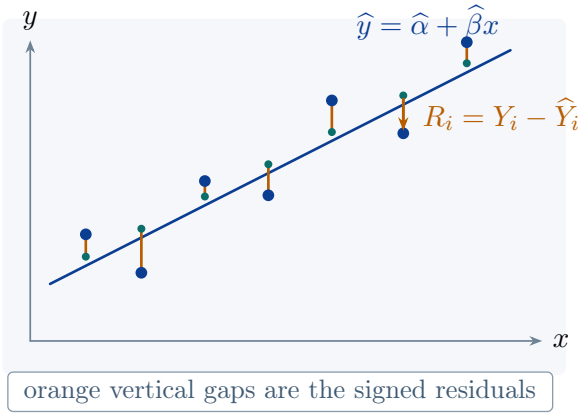


Figure 8

Earlier, in the variance calculation for the slope estimator, we computed

$$\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Now we compute the remaining terms. First,

$$\begin{aligned}\mathrm{Var}(\hat{\alpha}) &= \mathrm{Var}(\bar{Y} - \hat{\beta}\bar{X}) \\ &= \mathrm{Var}(\bar{Y}) + \bar{X}^2 \,\mathrm{Var}(\hat{\beta}) - 2\bar{X}\,\mathrm{Cov}(\bar{Y}, \hat{\beta}) \\ &= \frac{\sigma^2}{n} + \bar{X}^2 \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} - 2\bar{X}\,\mathrm{Cov}(\bar{Y}, \hat{\beta}).\end{aligned}$$

Also,

$$\mathrm{Cov}(\bar{Y}, \hat{\beta}) = \mathrm{Cov}\left(\sum_{i=1}^{n}\frac{1}{n}Y_i, \sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n}\frac{1}{n}a_i \,\mathrm{Var}(Y_i) = \sum_{i=1}^{n}\frac{1}{n}a_i\sigma^2 = \frac{\sigma^2}{n}\sum_{i=1}^{n} a_i = 0.$$

Therefore,

$$\mathrm{Var}(\hat{\alpha}) = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right].$$

Similarly,

$$\mathrm{Cov}(\hat{\alpha}, \hat{\beta}) = \mathrm{Cov}(\bar{Y} - \hat{\beta}\bar{X}, \hat{\beta}) = \mathrm{Cov}(\bar{Y}, \hat{\beta}) - \bar{X}\,\mathrm{Cov}(\hat{\beta}, \hat{\beta}) = -\bar{X}\,\mathrm{Var}(\hat{\beta}) = \frac{-\bar{X}\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

So

$$\mathrm{Var}(\hat{Y}_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] + X_0^2\frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + 2X_0\frac{-\bar{X}\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2 + X_0^2 - 2X_0\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right]$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right].$$

This variance is minimized at $X_0 = \bar{X}$. As X_0 moves away from $\bar{X}$ (the center of the observed X values), $\text{Var}(\hat{Y}_0)$ increases, so estimation of the mean response becomes less certain. Predicting a new observation requires an additional independent error term and therefore adds σ^2 to this variance.

Observe that $\text{Cov}(\hat{\alpha}, \hat{\beta}) \geq 0$ when $\bar{X} \leq 0$ and $\text{Cov}(\hat{\alpha}, \hat{\beta}) \leq 0$ when $\bar{X} \geq 0$. Since $\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X}$, the point $(\bar{X}, \bar{Y})$ lies on the fitted line. When $\bar{X} \geq 0$, increasing the slope tends to decrease the intercept; when $\bar{X} \leq 0$, they tend to increase together. In this sense, $\bar{X}$ acts as a pivot.

Example 15.1 If $Y = \alpha X^{\beta}$, then X and Y have a **power law** relationship. Taking logarithms gives

$$\log(Y) = \log(\alpha) + \beta \log(X).$$

So we can fit a linear regression of $\log(Y)$ on $\log(X)$, obtain estimates α' and β', and then transform back:

$$\hat{\beta} = \beta' \qquad \text{and} \qquad \hat{\alpha} = e^{\alpha'}.$$

Lecture 16 — Tuesday, March 11, 2014

Standardized Regression

In the simple linear regression setup, consider the standardized variables

$$U_i = \frac{X_i - \bar{X}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}} \qquad \text{and} \qquad V_i = \frac{Y_i - \bar{Y}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

Then

$$\sum_{i=1}^{n} U_i = \frac{\sum_{i=1}^{n}(X_i - \bar{X})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}} = 0,$$

which implies $\bar{U} = 0$. Similarly, $\bar{V} = 0$. Furthermore, the sample variance of the U_i satisfies

$$\frac{1}{n-1}\sum_{i=1}^{n}(U_i - \bar{U})^2 = \frac{1}{n-1}\sum_{i=1}^{n} U_i^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot (n-1)} = 1.$$

Likewise,

$$\frac{1}{n-1}\sum_{i=1}^{n}(V_i - \bar{V})^2 = 1.$$

Thus U_i measures how many sample standard deviation units X_i lies above or below $\bar{X}$.

Regressing the standardized values V_i on U_i by least squares gives

$$\tilde{\alpha} = \bar{V} - \tilde{\beta}\bar{U} = 0 \qquad \text{and} \qquad \tilde{\beta} = \frac{\sum_{i=1}^{n}(U_i - \bar{U})(V_i - \bar{V})}{\sum_{i=1}^{n}(U_i - \bar{U})^2} = \frac{\sum_{i=1}^{n} U_i V_i}{\sum_{i=1}^{n} U_i^2}.$$

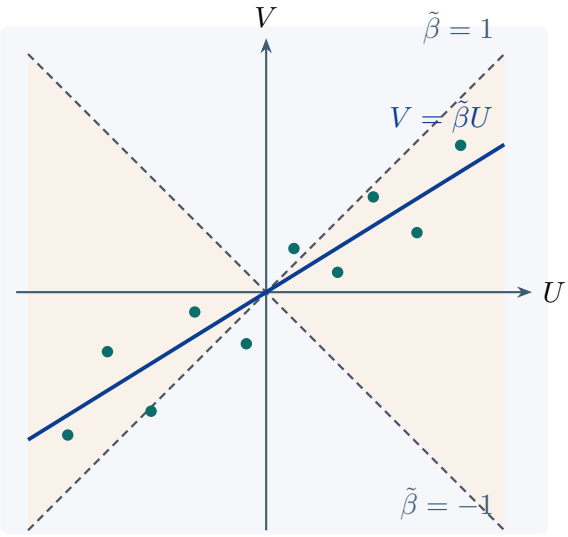
Observe that

$$\tilde{\beta}^2 = \frac{\left(\sum_{i=1}^{n} U_i V_i\right)^2}{\left(\sum_{i=1}^{n} U_i^2\right)^2} \leq \frac{\left(\sum_{i=1}^{n} U_i^2\right)\left(\sum_{i=1}^{n} V_i^2\right)}{\left(\sum_{i=1}^{n} U_i^2\right)^2} = \frac{(n-1)(n-1)}{(n-1)^2} = 1.$$

Therefore $\tilde{\beta} \in [-1, 1]$; in fact, $\tilde{\beta}$ is the sample correlation. When both axes use the same graphical scale, the fitted line lies between slopes -1 and 1. If $|\tilde{\beta}| < 1$, fitted standardized responses are less extreme than their corresponding standardized predictors, which is the algebra behind **regression to the mean**.

Because both axes are measured in standard-deviation units, the fitted slope is confined between

the two diagonal bounds in Figure 9. The sign of $\tilde{\beta}$ determines the direction of the association, while its magnitude determines how closely the fitted line approaches a diagonal.



standardization places the fitted slope between -1 and 1

FIGURE 9

Pearson's Goodness-of-Fit Test

Consider

$$H_0 : p_1 = p_1^0, p_2 = p_2^0, \ldots, p_K = p_K^0,$$

with the alternative H_A that at least one equality above fails. The likelihood ratio is

$$\Lambda = \frac{\mathcal{L}(\tilde{p}_1, \ldots, \tilde{p}_K)}{\mathcal{L}(p_1^0, \ldots, p_K^0)} = \frac{\dfrac{n!}{X_1! \ldots X_K!} \tilde{p}_1^{X_1} \ldots \tilde{p}_K^{X_K}}{\dfrac{n!}{X_1! \ldots X_K!} (p_1^0)^{X_1} \ldots (p_K^0)^{X_K}}.$$

Hence

$$2 \log \Lambda = 2 \sum_{j=1}^{K} X_j \log \left(\frac{\tilde{p}_j}{p_j^0} \right).$$

Since $\tilde{p}_j = \frac{X_j}{n}$, we can write

$$O_j := X_j = n\tilde{p}_j \qquad \text{and} \qquad E_j := np_j^0,$$

to denote the observed and expected counts, respectively, so

$$2\log\Lambda = 2\sum_{j=1}^{K} O_j \log\left(\frac{O_j}{E_j}\right),$$

where $p_j^0 > 0$ and terms with $O_j = 0$ are interpreted as zero. We reject H_0 if $2\log\Lambda > C_\alpha$, where C_α is obtained from an asymptotic $\chi^2_{(K-1)}$ reference distribution. This calibration requires the expected counts to grow sufficiently large.

Example 16.1 Consider data on 63 sudden heart attacks:

M	Tu	W	Th	F	Sat	Sun
22	7	6	13	5	4	6

Here H_0 states that heart attacks are equally likely on all seven days, and H_A states otherwise. Under H_0,

$$E_j = 63 \cdot \frac{1}{7} = 9, \qquad j = 1, \ldots, 7.$$

Then

$$2\log\Lambda = 2\sum_{j=1}^{7} O_j \log\left(\frac{O_j}{E_j}\right) = 2\left[22\log\left(\frac{22}{9}\right) + 7\log\left(\frac{7}{9}\right) + \cdots + 6\log\left(\frac{6}{9}\right)\right] = 23.27.$$

The 5% cutoff from χ^2_6 is 12.59. If H_0 were true, values above 12.59 would occur only about 5% of the time. Since $23.27 > 12.59$, we reject H_0.

The statistic used in **Pearson's test** is

$$\sum_{j=1}^{K} \frac{(O_j - E_j)^2}{E_j} \xrightarrow{D} \chi^2_{(K-1)}$$

under the null hypothesis, for a fixed number of cells whose expected counts grow with the sample size. To derive the relationship to the likelihood ratio test, let

$$f(x) = x\log\left(\frac{x}{x_0}\right),$$

where x_0 is known. We use a second-order Taylor expansion. Since

$$f'(x) = \log\left(\frac{x}{x_0}\right) + 1 \qquad \text{and} \qquad f''(x) = \frac{1}{x},$$

we have

$$\begin{aligned} f(x) &\approx f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2 \\ &= 0 + 1(x - x_0) + \frac{1}{2x_0}(x - x_0)^2 \\ &= x - x_0 + \frac{(x - x_0)^2}{2x_0}. \end{aligned}$$

Therefore,

$$2\log\Lambda \approx 2\sum_{j=1}^{K}\left[(O_j - E_j) + \frac{1}{2E_j}(O_j - E_j)^2\right].$$

Since

$$\sum_{j=1}^{K} O_j = n$$

and

$$\sum_{j=1}^{K} E_j = n,$$

we obtain

$$2\log\Lambda \approx \sum_{j=1}^{K}\frac{(O_j - E_j)^2}{E_j}.$$

Example 16.2 Recall Example 4.2, with genotype probabilities

$$\mathbb{P}(\text{AA}) = (1-\theta)^2, \qquad \mathbb{P}(\text{Aa}) = 2\theta(1-\theta), \qquad \text{and} \qquad \mathbb{P}(\text{aa}) = \theta^2.$$

Given data with 37 individuals of type AA, 46 of type Aa, and 17 of type aa in a sample of 100, we test whether the Hardy–Weinberg model holds. The null hypothesis is that the population follows the Hardy–Weinberg parametrization. In Example 4.2, we found that

$$\hat{\theta}_{\text{MLE}} = \frac{X_2 + 2X_3}{2n} = \frac{46 + 2\cdot 17}{2\cdot 100} = 0.4.$$

This yields expected counts

$$E_1 = 36, \qquad E_2 = 48, \qquad \text{and} \qquad E_3 = 16,$$

which are close to the observed values. To quantify the fit, we compute

$$2\log\Lambda = 2\left[37\log\left(\frac{37}{36}\right) + 46\log\left(\frac{46}{48}\right) + 17\log\left(\frac{17}{16}\right)\right] = 0.1733.$$

The Pearson test statistic equals

$$\frac{(37-36)^2}{36} + \frac{(46-48)^2}{48} + \frac{(17-16)^2}{16} = 0.1736,$$

which is nearly identical. However, we must be careful with the degrees of freedom:

$$\dim(\Theta) = 2, \qquad \dim(\Theta_0) = 1, \qquad \text{and} \qquad df = 2 - 1 = 1.$$

For 1 degree of freedom, the 5% cutoff is 3.84. Since

$$0.1733 < 3.84,$$

we fail to reject the null hypothesis; the data provide no evidence against the Hardy–Weinberg model.

We note three important observations. First, Pearson's test is a second-order Taylor approximation to the likelihood ratio test. Second, each summand is the square of a standardized cell-count residual, so the statistic aggregates discrepancies between observed and expected counts. Third, Pearson's statistic is substantially easier to calculate by hand than the likelihood ratio test statistic, which explains its widespread use until the 1960s when computers became widely available.

Lecture 17 — Thursday, March 13, 2014

Applications of the Multinomial Model

We now look at several applications of the multinomial model: hospital quality assessment, authorship signatures in Jane Austen's writing (with analogous applications in music), and comparisons between two distributions.

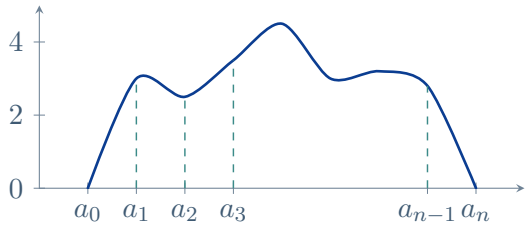


FIGURE 10

If we divide a continuous distribution into n intervals, the resulting counts of independent and identically distributed observations from each interval form a multinomial model. This can help identify which distribution is plausible, provided the required parameters are known or can be estimated.

To make this precise, suppose $X_1, \ldots, X_m$ are independent and identically distributed observations and we choose cut points

$$-\infty = a_0 < a_1 < \cdots < a_n = \infty.$$

The partition is shown in Figure 10. Define interval counts

$$O_i = \sum_{r=1}^{m} \mathbb{1}_{\{a_{i-1} < X_r \le a_i\}}, \quad i = 1, \ldots, n.$$

Under a candidate model F_θ, the cell probabilities are

$$p_i(\theta) = \mathbb{P}(a_{i-1} < X \le a_i) = \int_{a_{i-1}}^{a_i} f_\theta(x)\,\mathrm{d}x = F_\theta(a_i) - F_\theta(a_{i-1}),$$

so that

$$(O_1, \ldots, O_n) \sim \text{Multinomial}\big(m; p_1(\theta), \ldots, p_n(\theta)\big).$$

If θ is unknown, we can estimate it by $\hat{\theta}$, and then compare observed counts O_i with expected counts $E_i = m p_i(\hat{\theta})$. A standard summary is Pearson's statistic

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

or the likelihood ratio version

$$G^2 = 2 \sum_{i=1}^{n} O_i \log\left(\frac{O_i}{E_i}\right).$$

Small values indicate that the observed bin counts are consistent with the candidate distribution;

unusually large values indicate that the model allocates mass very differently from the data and is therefore implausible. With a fixed partition, regular parameter estimation, and expected cell counts that grow sufficiently large, these statistics are compared asymptotically with a $\chi^2_{(n-1-k)}$ distribution, where k is the number of estimated parameters.

Another key application is testing independence. For discrete random variables, this is done with **contingency tables** (for example, smoking status versus cancer status). Under ordinary multinomial sampling in an $I \times J$ table, the independence model has dimension $(I-1)+(J-1)$, while the unrestricted model has dimension $IJ-1$. Hence

$$(IJ-1)-(I-1)-(J-1) = \boxed{(I-1)(J-1)}.$$

This degrees-of-freedom formula is common, but it is not universally valid in every modelling setup.

Lecture 18 — Tuesday, March 18, 2014

6. Study Design and Sampling

Stratified Sampling and Paired Analysis

A researcher wishes to choose 10 blocks from 100 possible options. Selecting the most "representative" ones produces a **judgment sample**. Using a random number generator instead produces a **simple random sample**. A third approach divides the blocks into strata and takes a simple random sample from each (say, five small blocks and five large ones) to obtain a **stratified random sample**. Judgment sampling is vulnerable to systematic human bias (psychologists continue to investigate why). Random sampling removes that source of selection bias, and well-chosen strata can reduce variance further.

To compare the variances cleanly, consider a superpopulation made of two equally weighted strata. Let a draw from the overall mixture have mean μ and variance σ^2; a draw from Stratum 1 has mean μ_1 and variance σ_1^2, while a draw from Stratum 2 has mean μ_2 and variance σ_2^2. A simple random sample $Z_1, \ldots, Z_{2n}$ of $2n$ independent draws from the mixture has

$$\mathbb{E}\left[\bar{Z}\right] = \mu \qquad \text{and} \qquad \operatorname{Var}(\bar{Z}) = \frac{\sigma^2}{2n}.$$

For a stratified design, independently draw $X_1, \ldots, X_n$ from Stratum 1 and $Y_1, \ldots, Y_n$ from Stratum 2. The estimator $\frac{\bar{X}+\bar{Y}}{2}$ satisfies

$$\mathbb{E}\left[\frac{\bar{X}+\bar{Y}}{2}\right] = \frac{\mu_1+\mu_2}{2} = \mu,$$

while

$$\mathrm{Var}\left(\frac{\bar{X}+\bar{Y}}{2}\right) = \frac{1}{4}\left[\mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y})\right] = \frac{\sigma_1^2+\sigma_2^2}{4n},$$

with no covariance term because the two strata are sampled independently. We can have

$$\frac{\sigma_1^2+\sigma_2^2}{4n} < \frac{\sigma^2}{2n}$$

when strata are internally homogeneous.

Example 18.1 Consider the equally weighted distribution on $\{1,2,3,4\}$, divided into the strata small $= \{1,2\}$ and large $= \{3,4\}$. We have

$$\mu = \frac{1+2+3+4}{4} = 2.5, \qquad \mu_1 = \frac{1+2}{2} = 1.5, \qquad \text{and} \qquad \mu_2 = \frac{3+4}{2} = 3.5,$$

which satisfies

$$\frac{\mu_1+\mu_2}{2} = \frac{1.5+3.5}{2} = 2.5 = \mu.$$

The variances are

$$\begin{aligned}
\sigma^2 &= \frac{(1-2.5)^2+(2-2.5)^2+(3-2.5)^2+(4-2.5)^2}{4} = \frac{5}{4} = 1.25,\\
\sigma_1^2 &= \frac{(1-1.5)^2+(2-1.5)^2}{2} = 0.25,\\
\sigma_2^2 &= \frac{(3-3.5)^2+(4-3.5)^2}{2} = 0.25.
\end{aligned}$$

With $n = 2$ draws from each stratum, the stratified estimator has variance

$$\frac{\sigma_1^2+\sigma_2^2}{4n} = \frac{0.25+0.25}{8} = 0.0625,$$

whereas the mean of four simple random draws from the mixture has variance

$$\frac{\sigma^2}{2n} = \frac{1.25}{4} = 0.3125.$$

This demonstrates the variance reduction under equal allocation. For sampling without replacement from a finite population, both formulas also require their corresponding finite-population corrections.

Paired Analysis

We now demonstrate mathematically how pairing can reduce variance. For an unpaired analysis, suppose

$$X_1, \dots, X_n \sim \mathcal{N}(\mu_1, \sigma^2) \qquad \text{and} \qquad Y_1, \dots, Y_n \sim \mathcal{N}(\mu_2, \sigma^2)$$

are independent samples. Then the estimator $\bar{X} - \bar{Y}$ has variance

$$\operatorname{Var}(\bar{X} - \bar{Y}) = \operatorname{Var}(\bar{X}) + \operatorname{Var}(\bar{Y}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \frac{2\sigma^2}{n}.$$

(There is no covariance term.) For a paired analysis, the pairs (X_i, Y_i) are independent across i, but the two observations within a pair may be correlated. Let $D_i = X_i - Y_i$ and estimate $\bar{D}$. We have

$$\mathbb{E}\left[\bar{D}\right] = \mathbb{E}\left[\bar{X} - \bar{Y}\right] = \mu_1 - \mu_2,$$

so $\bar{D}$ is also unbiased for $\mu_1 - \mu_2$. The variance is

$$\operatorname{Var}(\bar{D}) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \operatorname{Var}(D_i),$$

where we note that $\operatorname{Cov}(D_i, D_j) = 0$ for $i \neq j$. Since

$$\operatorname{Var}(D_i) = \operatorname{Var}(X_i - Y_i) = \operatorname{Var}(X_i) + \operatorname{Var}(Y_i) - 2\operatorname{Cov}(X_i, Y_i),$$

and using the correlation

$$\operatorname{Corr}(X_i, Y_i) = \frac{\operatorname{Cov}(X_i, Y_i)}{\sqrt{\operatorname{Var}(X_i)\operatorname{Var}(Y_i)}} = \rho,$$

we obtain

$$\operatorname{Var}(D_i) = \sigma^2 + \sigma^2 - 2\rho\sigma^2 = (2 - 2\rho)\sigma^2.$$

Therefore,

$$\operatorname{Var}(\bar{D}) = \frac{2(1-\rho)\sigma^2}{n}.$$

When $\rho > 0$ (which is the rationale for pairing), we have

$$\mathrm{Var}(\bar{D}) < \mathrm{Var}(\bar{X} - \bar{Y}).$$

Example 18.2 Suppose the correlation is 50%, that is, $\rho = 0.5$, which represents a mild correlation for "good pairs." This yields a 50% reduction in variance, which translates to doubling the effective precision.

Lecture 19 — Thursday, March 20, 2014

Graphical Methods

Histograms and **scatter plots** are valuable visualization tools. For histograms, a common rule of thumb is that larger sample sizes permit more bins, but typically no more than $\sqrt{n}$ bins should be used. Box-and-whisker plots are particularly effective for comparing the distributions of several variables.

Good plotting practices are essential. For example, we should always use consistent scales across comparable plots. The principle of *garbage in, garbage out* applies: mathematics cannot remedy data of poor quality. There is no fully automatic method to detect problematic data; careful human inspection of plots and datasets remains necessary.

Empirical Studies

Chapter 3 of the official course notes discusses empirical study methodology. The **PPDAC cycle** describes the five stages of statistical inquiry: *Problem* (define the research question), *Plan* (design the study), *Data* (collect observations), *Analysis* (apply statistical methods), and *Conclusion* (interpret results and communicate findings).

In a study, the **population** is the group about which we wish to make inferences. An **attribute** is a characteristic of the population, such as average age, distribution of household income, or even a functional relationship between two other attributes.

The distinction between the target population and the study population is important. The **target population** is the full group to which we want our conclusions to apply, while the **study**

population is the group that is actually accessible for sampling. In an ideal study these coincide, but in practice the study population is often narrower because of geography, timing, eligibility restrictions, or nonresponse. When the two differ, we must be careful about how far the results can be generalized beyond the group that was actually studied.

An **observational study** involves collecting data without controlling or manipulating variables, while an **experimental study** involves actively controlling variables.

Lecture 21 — Tuesday, March 25, 2014

7. Causation and Bayesian Statistics

Causation versus Association

We now turn to the interpretation of statistical relationships.

Causation means that, *all else being equal*, changing X produces changes in Y. These changes are typically probabilistic, representing shifts in the distribution of Y rather than deterministic outcomes. For example, a cancer drug may significantly improve survival rates without guaranteeing recovery for every patient. The key principle in establishing causation is **randomization**, which minimizes violations of the "all else equal" requirement.

Association, in contrast, can arise from three scenarios: X causes Y, or Y causes X, or an unseen **confounding factor** Z causes both X and Y.

An **experimental** (or **randomized**) **study** controls the randomization process. In a **double-blind study**, neither the physician nor the patient knows whether the treatment or placebo is being administered. This matters because subtle cues such as facial expressions can influence patient outcomes.

Blocking is a technique where participants are divided into blocks that control for known potential confounding factors (for example, age groups in a drug study).

In **observational studies**, we cannot randomize. For instance, to study whether smoking causes cancer, we cannot ethically assign people to smoke for 20 years. We can only analyze observational

data. Chi-squared tests can assess association, but they cannot establish causation. Probabilistic outcomes mean that some heavy smokers live long lives, while some healthy nonsmokers develop cancer early.

Matching attempts to make comparison groups as similar as possible (for example, when studying gender differences, match participants by health status, weight, and other relevant factors).

Example 21.1 (Berkeley admissions data; Bickel, Hammel, and O'Connell (1975)) For fall 1973, aggregate UC Berkeley graduate-admissions data showed that about 44% of male applicants were admitted, compared with about 35% of female applicants. Department-level analysis, however, found few statistically significant sex differences and about as many departments favouring women as favouring men. The aggregate discrepancy arose largely because women applied disproportionately to more selective departments, while men applied more frequently to less selective ones. Department selectivity was therefore a confounding factor. This phenomenon is known as **Simpson's paradox**.

Example 21.1 illustrates a fundamental limitation of observational studies: if data broken down by department had been unavailable, the misleading aggregate pattern would have been impossible to detect. This lesson is perhaps more important than technical details like the likelihood ratio test: always keep these principles in mind when interpreting statistical associations.

Bayesian Statistics

The Bayesian framework treats the parameter θ as a random variable. We write $\pi(\theta)$ for the **prior density** and $\pi(\theta \mid x_1, \ldots, x_n)$ for the **posterior density**; these represent our beliefs before and after observing the data. For a continuous scalar parameter with parameter space Θ, Bayes' theorem gives

$$\pi(\theta \mid x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n \mid \theta)\pi(\theta)}{\displaystyle\int_\Theta f(x_1, \ldots, x_n \mid u)\pi(u)\,\mathrm{d}u}.$$

The numerator involves $f(x_1, \ldots, x_n \mid \theta)$, which is the likelihood of θ. The denominator is often computationally challenging, especially when θ is a vector with thousands of components, making even numerical evaluation difficult.

Once the posterior is obtained, however, we can compute point estimates (such as $\mathbb{E}[\theta \mid x_1, \ldots, x_n]$ or the posterior mode) or **credible regions**. Unlike frequentist confidence intervals, Bayesian credible regions directly represent our uncertainty about θ given the data, since we have the entire posterior distribution.

Example 21.2 Suppose $X \sim \text{Bin}(n, p)$ and we are interested in the unknown parameter p. We first need to specify a prior distribution. Without prior knowledge about p, a reasonable choice is the uniform prior $p \sim \text{Unif}(0, 1)$. Then

$$\pi(p \mid x) = \frac{f(x \mid p) \cdot \pi(p)}{\displaystyle\int f(x \mid p) \cdot \pi(p) \, \mathrm{d}p} = \frac{\binom{n}{x} p^x (1-p)^{n-x} \cdot 1}{\displaystyle\int_0^1 \binom{n}{x} p^x (1-p)^{n-x} \cdot 1 \, \mathrm{d}p}.$$

To evaluate this expression, we need a few properties of the beta distribution.

For $\alpha, \beta > 0$, we say that Z has a **beta distribution** and write $Z \sim \text{Beta}(\alpha, \beta)$ if

$$f(z) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} z^{\alpha - 1} (1 - z)^{\beta - 1}$$

for $z \in (0, 1)$, with density zero elsewhere. The expected value is

$$\begin{aligned}
\mathbb{E}[Z] &= \int_0^1 z \cdot f(z) \, \mathrm{d}z = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 z \cdot z^{\alpha - 1} (1 - z)^{\beta - 1} \, \mathrm{d}z \\
&= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \underbrace{\int_0^1 z^{(\alpha+1)-1} (1 - z)^{\beta - 1} \, \mathrm{d}z}_{\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)} \text{ since the Beta}(\alpha+1,\beta) \text{ density integrates to 1}} \\
&= \frac{\alpha}{\alpha + \beta}.
\end{aligned}$$

Example 21.3 Returning to the denominator in Example 21.2, which resembles the beta distribution, it is equal to

$$\begin{aligned}
&\int_0^1 \binom{n}{x} p^x (1-p)^{n-x} \cdot 1 \, \mathrm{d}p \\
&= \frac{\Gamma(x+1) \cdot \Gamma(n-x+1)}{\Gamma(n+2)} \binom{n}{x} \cdot \underbrace{\int_0^1 \frac{\Gamma(n+2)}{\Gamma(x+1) \cdot \Gamma(n-x+1)} p^{(x+1)-1} (1-p)^{(n-x+1)-1} \, \mathrm{d}p}_{=1} \\
&= \binom{n}{x} \frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)}
\end{aligned}$$

$$= \frac{n!}{x!(n-x)!} \cdot \frac{x! \cdot (n-x)!}{(n+1)!}$$
$$= \frac{n!}{(n+1)!}$$
$$= \frac{1}{n+1}.$$

Thus

$$\pi(p \mid x) = (n+1)\binom{n}{x} p^x (1-p)^{n-x} \cdot 1 = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} p^{(x+1)-1}(1-p)^{(n-x+1)-1}.$$

We conclude that the posterior distribution is $p \mid x \sim \text{Beta}(x+1, n-x+1)$.

Lecture 22 — Thursday, March 27, 2014

Conjugate Priors and Jeffreys' Rule

Example 22.1 Continuing from Example 21.2, we saw that when $X \sim \text{Bin}(n, p)$ and the prior on p is $\text{Unif}(0, 1)$, the posterior is $\text{Beta}(x+1, n-x+1)$.

The posterior expectation is therefore

$$\mathbb{E}[p \mid x] = \frac{x+1}{n+2}.$$

By comparison, $\hat{p}_{\text{MLE}} = \frac{x}{n}$. These are close for large n, but can differ noticeably for small n. For one observation (i.e., $n = 1$) we have the following tabulation:

x	0	1
MLE	0	1
Bayes	1/3	2/3

Observe that the $\text{Unif}(0, 1)$ distribution is exactly the $\text{Beta}(1, 1)$ distribution. The prior mean is $\frac{1}{2}$, and the posterior mean $\mathbb{E}[p \mid x]$ combines information from both the prior mean and the likelihood.

The beta distribution thus is a **conjugate prior** for the binomial family, meaning the posterior remains a beta distribution. In particular, if the prior is $\text{Beta}(\alpha, \beta)$ so that

$$\pi(p) \propto p^{\alpha-1}(1-p)^{\beta-1},$$

and the model for the data is $\text{Bin}(n, p)$ so that

$$f(x \mid p) = \binom{n}{x} p^x (1-p)^{n-x},$$

then

$$\pi(p \mid x) \propto f(x \mid p)\pi(p) \propto p^{x+\alpha-1}(1-p)^{n-x+\beta-1},$$

so the posterior is $\text{Beta}(x+\alpha, n-x+\beta)$. The posterior mean is

$$\mathbb{E}[p \mid x] = \frac{x+\alpha}{n+\alpha+\beta}.$$

Common conjugate pairs include beta–binomial, gamma–Poisson, and normal–normal; in each case, conjugacy is straightforward to check.

How do we choose a prior? A common idea is to choose a "most ignorant" prior that encodes minimal information. For binomial p, the $\text{Beta}(1,1)$ distribution (or equivalently, the $\text{Unif}(0,1)$ distribution) is a flat prior. A key issue is that flat priors are generally not invariant under reparametrization.

Example 22.2 Let $\pi_p(p) = 1$ for $p \in (0,1)$ and set $Q = -\log(p) \in [0, \infty)$. Then

$$F_Q(q) = \mathbb{P}(Q \le q) = \mathbb{P}(-\log(p) \le q) = \mathbb{P}(p \ge e^{-q}) = 1 - e^{-q},$$

so

$$\pi_Q(q) = \frac{\mathrm{d}}{\mathrm{d}q} F_Q(q) = e^{-q}.$$

So the prior on Q is $\text{Exp}(1)$, which is decidedly not flat. Thus a flat prior is not invariant to a change of scale. More generally, for a differentiable one-to-one monotone transformation $Q = g(p)$, the density obeys the usual **change-of-variables formula**

$$\pi_Q(q) = \pi_p(g^{-1}(q)) \left| \frac{\mathrm{d}}{\mathrm{d}q} g^{-1}(q) \right|.$$

For the cdf, $F_Q(q) = F_p(g^{-1}(q))$ when g is increasing, while $F_Q(q) = 1 - F_p(g^{-1}(q))$ for a

continuous p when g is decreasing.

Jeffreys' rule says that to obtain invariance under reparametrization, choose

$$\pi(\theta) \propto \sqrt{-\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} l(\theta)\right]}.$$

This quantity is based on the Fisher information. If we reparametrize with $\phi = g(\theta)$, the chain rule gives

$$\frac{\partial^2 l}{\partial\phi^2} = \frac{\partial^2 l}{\partial\theta^2}\left(\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right)^2 + \frac{\partial l}{\partial\theta}\frac{\mathrm{d}^2\theta}{\mathrm{d}\phi^2}.$$

Under the usual regularity conditions, the expected score $\mathbb{E}[\partial l/\partial\theta]$ is zero. Therefore Jeffreys' rule transforms as

$$\begin{aligned}
\pi(\phi) &= \sqrt{-\mathbb{E}\left[\frac{\partial^2}{\partial\phi^2} l(\phi)\right]} \\
&= \sqrt{-\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} l(\theta) \cdot \left(\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right)^2\right]} \\
&= \sqrt{-\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} l(\theta)\right]} \cdot \left|\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right| \\
&= \pi(\theta) \cdot \left|\frac{\mathrm{d}}{\mathrm{d}\phi} g^{-1}(\phi)\right|,
\end{aligned}$$

which matches the change-of-variables formula.

Example 22.3 For $X \sim \mathrm{Bin}(n, p)$,

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{and} \qquad l(p) = \text{const} + x\log(p) + (n-x)\log(1-p),$$

so

$$l'(p) = \frac{x}{p} - \frac{n-x}{1-p} \qquad \text{and} \qquad l''(p) = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2},$$

and therefore

$$-\mathbb{E}\left[l''(p)\right] = \frac{\mathbb{E}\left[X\right]}{p^2} + \frac{n - \mathbb{E}\left[X\right]}{(1-p)^2} = \frac{n}{p(1-p)}.$$

Jeffreys' rule yields

$$\pi(p) \propto \frac{1}{\sqrt{p(1-p)}},$$

so Jeffreys' prior is Beta $\left(\frac{1}{2}, \frac{1}{2}\right)$. Under this prior, the posterior mean is

$$\mathbb{E}\left[p \mid x\right] = \frac{x + 1/2}{n+1},$$

compared with

$$\mathbb{E}\left[p \mid x\right] = \frac{x+1}{n+2}$$

under the uniform prior. One interpretation of a Beta(α, β) prior is that it acts like $(\alpha + \beta)$ pseudo-observations, of which α are successes.

Lecture 23 — Tuesday, April 01, 2014

Empirical Bayes

Conditionally on $\theta_1, \ldots, \theta_n$, let the Y_i be independent with densities $f(y_i \mid \theta_i)$, and suppose the θ_i are independent draws from $\pi(\theta \mid \lambda)$ with common **hyperparameter** λ. In simple Bayes, λ is fixed in advance. In **hierarchical Bayes**, we add a prior on λ. Here we focus instead on **empirical Bayes**.

The marginal (unconditional) distribution of Y_i is

$$m(y_i) = \int f(y_i \mid \theta_i)\pi(\theta_i \mid \lambda)\,\mathrm{d}\theta_i.$$

This is an instance of the law of total probability. After integrating out θ_i, $m(y_i)$ depends only on λ. The empirical Bayes procedure first estimates λ from $\prod_{i=1}^n m(y_i)$ by, for example, maximum likelihood or the method of moments, and then uses the posterior $\pi(\theta_i \mid y_i, \hat{\lambda})$, which depends on the data $Y_1, \ldots, Y_n$ through the estimate $\hat{\lambda}$.

Example 23.1 If $Y_i \sim \mathrm{Bin}(n, p_i)$ and $p_i \sim \mathrm{Beta}(\alpha, \beta)$, then the posterior expectation is

$$\mathbb{E}\left[p_i \mid Y_i\right] = \frac{Y_i + \alpha}{n + \alpha + \beta}.$$

The vanilla Bayes formulation fixes α and β (for example at 1 and 1 or 1/2 and 1/2), while empirical Bayes estimates them as $\hat{\alpha}$ and $\hat{\beta}$, respectively, and then uses

$$\frac{Y_i + \hat{\alpha}}{n + \hat{\alpha} + \hat{\beta}}.$$

Example 23.2 Let $Y_i \overset{\text{indep}}{\sim} \mathcal{N}(\theta_i, \sigma^2)$ with known σ^2, and prior $\theta_i \sim \mathcal{N}(\mu, \tau^2)$. The posterior is

$$\begin{aligned}\pi(\theta_i \mid Y_i) &= \frac{f(y_i \mid \theta_i)\pi(\theta_i \mid \mu, \tau^2)}{\displaystyle\int f(y_i \mid u)\pi(u \mid \mu, \tau^2)\,\mathrm{d}u} \\ &= \frac{\dfrac{1}{\sqrt{2\pi}\sigma}\exp\left(-\dfrac{(y_i-\theta_i)^2}{2\sigma^2}\right)\cdot\dfrac{1}{\sqrt{2\pi}\tau}\exp\left(-\dfrac{(\theta_i-\mu)^2}{2\tau^2}\right)}{\displaystyle\int \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-u)^2}{2\sigma^2}\right)\cdot\frac{1}{\sqrt{2\pi}\tau}\exp\left(-\frac{(u-\mu)^2}{2\tau^2}\right)\mathrm{d}u}.\end{aligned}$$

The denominator works out to the density of the $\mathcal{N}(\mu, \sigma^2 + \tau^2)$ distribution, which is the marginal distribution of $y_i \mid \mu$. Using this, we obtain

$$\pi(\theta_i \mid Y_i) = \frac{\dfrac{1}{\sqrt{2\pi}\sigma\tau}}{\dfrac{1}{\sqrt{2\pi}\sqrt{\sigma^2+\tau^2}}}\exp\left(-\left[\frac{(y_i-\theta_i)^2}{2\sigma^2} + \frac{(\theta_i-\mu)^2}{2\tau^2} - \frac{(y_i-\mu)^2}{2(\sigma^2+\tau^2)}\right]\right).$$

Let

$$B = \frac{\sigma^2}{\sigma^2+\tau^2} \qquad \text{and} \qquad v = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}.$$

Completing the square gives

$$\frac{(y_i-\theta_i)^2}{2\sigma^2} + \frac{(\theta_i-\mu)^2}{2\tau^2} - \frac{(y_i-\mu)^2}{2(\sigma^2+\tau^2)} = \frac{(\theta_i - [(1-B)y_i + B\mu])^2}{2v}.$$

Therefore

$$\pi(\theta_i \mid Y_i) = \frac{1}{\sqrt{2\pi v}}\exp\left(-\frac{(\theta_i - [(1-B)y_i + B\mu])^2}{2v}\right),$$

so the posterior is

$$\theta_i \mid Y_i \sim \mathcal{N}\left((1-B)Y_i + B\mu, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right).$$

In simple Bayes, μ is specified in advance, giving the posterior mean

$$(1-B)Y_i + B\mu,$$

where $B = \frac{\sigma^2}{(\sigma^2+\tau^2)}$. On the other hand, in empirical Bayes, we estimate μ from $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2+\tau^2)$, so $\hat{\mu} = \bar{Y}$ if we use maximum likelihood or the method of moments. Then the posterior mean of θ_i is

$$(1-B)Y_i + B\bar{Y} = \bar{Y} + (1-B)(Y_i - \bar{Y}).$$

The key feature is that the posterior of θ_i depends not only on Y_i, but also on all $Y_1, \ldots, Y_n$ through $\bar{Y}$.

Example 23.3 In Example 23.2, if τ^2 is unknown and $n > 3$, we could instead estimate $\sigma^2+\tau^2$ from the marginal model $Y_i \sim \mathcal{N}(\mu, \sigma^2+\tau^2)$. From the one-sample t-test framework,

$$\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\sigma^2+\tau^2} \sim \chi^2_{(n-1)} \qquad \text{and} \qquad \frac{\sigma^2+\tau^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \sim \text{Inv-}\chi^2_{(n-1)}.$$

The latter distribution is an **inverse-chi-squared** distribution with $n-1$ degrees of freedom. If $V \sim \text{Inv-}\chi^2_{(n)}$, then $\mathbb{E}[V] = \frac{1}{(n-2)}$, so

$$\mathbb{E}\left[\frac{\sigma^2+\tau^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}\right] = \frac{1}{n-3}.$$

Hence

$$\frac{n-3}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

is unbiased for $\frac{1}{(\sigma^2+\tau^2)}$, leading to the empirical Bayes estimator

$$\hat{\theta}_i = \bar{Y} + \left(1 - \frac{\sigma^2(n-3)}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\right)(Y_i - \bar{Y}).$$

This is the usual empirical-Bayes form of the **James–Stein estimator** for θ_i. The shrinkage factor can be negative for unusually concentrated data; replacing it by its positive part gives the commonly used positive-part James–Stein estimator.

Lecture 24 — Thursday, April 03, 2014

Posterior Odds

Frequentist hypothesis tests are calibrated by the significance level and power, but they do not directly assign probabilities to competing hypotheses after observing data. In a Bayesian framework, we begin with prior probabilities for H_0 and H_A, and then update them using the observed result of a test. The key quantity is the **posterior odds**

$$\frac{\mathbb{P}(H_A \text{ true} \mid E)}{\mathbb{P}(H_0 \text{ true} \mid E)},$$

where E denotes the event that the data favor H_A (equivalently, that the testing procedure rejects H_0).

Using Bayes' theorem,

$$\mathbb{P}(H_A \text{ true} \mid E) = \frac{\mathbb{P}(E \mid H_A \text{ true})\mathbb{P}(H_A \text{ true})}{\mathbb{P}(E)}.$$

An analogous expression holds for H_0. Taking the ratio gives

$$\frac{\mathbb{P}(H_A \text{ true} \mid E)}{\mathbb{P}(H_0 \text{ true} \mid E)} = \frac{\mathbb{P}(E \mid H_A)}{\mathbb{P}(E \mid H_0)} \cdot \frac{\mathbb{P}(H_A)}{\mathbb{P}(H_0)} = \frac{\text{power}}{\text{significance level}} \cdot \frac{\mathbb{P}(H_A)}{\mathbb{P}(H_0)}.$$

The term

$$\frac{\mathbb{P}(H_A)}{\mathbb{P}(H_0)}$$

is called the **prior odds**, and the full ratio above is the **posterior odds**. Thus the binary event E contributes an evidence multiplier $\frac{\text{power}}{\text{significance level}}$ that updates prior odds into posterior odds. Here "power" refers to a simple alternative; for a composite alternative it must be averaged over the prior distribution within H_A. Using the complete data rather than only the reject-or-not event generally retains more evidence.

For instance, if the prior odds are $1:1$, the power is 0.80, and the significance level is 0.05, then the posterior odds become $16:1$ in favor of H_A.

Appendix: Lecture and Tutorial Index