Statistics 2

Hadrien Montanelli

1 Estimation

Delta method Suppose $X_1, \ldots, X_n$ are iid with mean μ and variance σ^2. The CLT yields

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \overset{d}{\approx} \mathrm{N}(0,1).$$

Suppose we need the asymptotic distribution of $g(\overline{X})$, for some function g. Expanding $g(\overline{X})$ yields

$$g(\overline{X}) \approx g(\mu) + (\overline{X} - \mu)g'(\mu),$$

asymptotic mean and *asymptotic variance*

$$\mathrm{E}(g(\overline{X})) \approx g(\mu) \quad \text{and} \quad \mathrm{var}(g(\overline{X})) \approx \frac{g'(\mu)^2\sigma^2}{n},$$

and *asymptotic distribution*

$$g(\overline{X}) \overset{d}{\approx} \mathrm{N}\left(g(\mu), \frac{g'(\mu)^2\sigma^2}{n}\right).$$

Order statistics (data) The *order statistics* of data $x_1, \ldots, x_n$ are their values in increasing order, which we denote $x_{(1)} \leq \ldots \leq x_{(n)}$.

Sample median The *sample median* is $m = x_{([n+1]/2)}$ if n is odd, or $m = 1/2(x_{(n/2)} + x_{(n/2+1)})$ if n is even.

Lower and upper quartile The *lower quartile* has $1/4$ of the sample that is less than it, and the *upper quartile* has $3/4$ of the sample that is less than it.

Inter-quartile range (IQR) The *inter-quartile range* is defined by IQR $=$ upper quartile $-$ lower quartile.

Order statistic (random sample) The *rth order statistic* of the random sample $X_1, \ldots, X_n$ is the random variable $X_{(r)}$, where $X_{(1)} \leq \ldots \leq X_{(n)}$.

Pdf of order statistic Suppose $X_1, \ldots, X_n$ are iid and continuous, each having cdf F and pdf f. Then, the pdf of $X_{(r)}$ is

$$f_{(r)}(x) = \frac{n!}{(r-1)!(n-r)!}F(x)^{r-1}[1-F(x)]^{n-r}f(x).$$

Quantile For a distribution with cdf F and pdf f, the pth *quantile* is the value x_p such that

$$F(x_p) = \int_{-\infty}^{x_p} f(u)du = p, \quad 0 \leq p \leq 1.$$

Probability integral transform Suppose X is a continuous random variable taking values in (a, b), with strictly increasing cdf F. Then, $F(X) \sim \mathrm{U}(0,1)$ is called the *probability integral transform* of X.

Lemma If $U_{(1)}, \ldots, U_{(n)}$ are the order statistics of a random sample of size n from a $\mathrm{U}(0,1)$ distribution, then

$$\mathrm{E}(U_{(r)}) = \frac{r}{n+1},$$
$$\mathrm{var}(U_{(r)}) = \frac{r}{(n+1)(n+2)}\left(1 - \frac{r}{n+1}\right).$$

Q-Q plot If data $x_1, \ldots, x_n$ are from a distribution with cdf F, then Q-Q plots use the approximation

$$F(x_{(k)}) \approx \frac{k}{n+1}.$$

Normal Q-Q plot If data $x_1, \ldots, x_n$ are from a $\mathrm{N}(\mu, \sigma^2)$ distribution, for some unknown μ and σ^2, then

$$x_{(k)} \approx \sigma\phi^{-1}\left(\frac{k}{n+1}\right) + \mu.$$

Exponential Q-Q plot If data $x_1, \ldots, x_n$ are from a $\mathrm{Exp}(\mu)$ distribution, for some unknown μ, then

$$x_{(k)} \approx -\mu \log\left(1 - \frac{k}{n+1}\right).$$

Pareto Q-Q plot If data $x_1, \ldots, x_n$ are from a $\mathrm{Par}(\alpha, \theta)$ distribution, for some unknown α and θ, then

$$\log x_{(k)} \approx \log\alpha - \frac{1}{\theta}\log\left(1 - \frac{k}{n+1}\right).$$

Observed/Fisher information In a model with scalar θ and log-likelihood $\ell(\theta; \boldsymbol{X})$, the *observed information* $J(\theta)$ and the *Fisher information* $I(\theta)$ are

$$J(\theta) = -\frac{d^2\ell}{d\theta^2}, \quad I(\theta) = \mathrm{E}\left(-\frac{d^2\ell}{d\theta^2}\right).$$

When $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, the *observed information matrix* and the *Fisher information matrix* are $p \times p$ symmetric matrices $J(\theta)$ and $I(\theta)$ whose (j,k) elements are

$$J(\boldsymbol{\theta})_{j,k} = -\frac{\partial^2\ell}{\partial\theta_j\partial\theta_k}, \quad I(\boldsymbol{\theta})_{j,k} = \mathrm{E}\left(-\frac{\partial^2\ell}{\partial\theta_j\partial\theta_k}\right).$$

Note that expectations are taken over $\boldsymbol{X}$.

Fisher information (iid) If $X_1, \ldots, X_n$ are iid from $f(\boldsymbol{x}; \boldsymbol{\theta})$, then $I(\boldsymbol{\theta}) = ni(\boldsymbol{\theta})$, where $i(\boldsymbol{\theta})$ is the Fisher information in a sample of size 1; that is, in the scalar case,

$$I(\theta) = ni(\theta), \quad i(\theta) = \mathrm{E}\left(-\frac{d^2\log f(X_1;\theta)}{d\theta^2}\right).$$

Properties of MLEs (scalar case)

- $\sqrt{I(\theta)}(\hat{\theta} - \theta) \xrightarrow{d} \mathrm{N}(0,1)$, that is, $\hat{\theta} \overset{d}{\approx} \mathrm{N}(\theta, I(\theta)^{-1})$;
- $\hat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$;
- $\hat{\theta}$ is asymptotically unbiased;
- the variance of $\hat{\theta}$ is $\sim 1/n$, and as small as possible;
- if $\psi = g(\theta)$ then $\hat{\psi} = g(\hat{\theta})$.

2 Confidence intervals

CI using MLE Using $\sqrt{I(\theta)}(\hat{\theta}-\theta) \overset{d}{\approx} \mathrm{N}(0,1)$,

$$\mathrm{P}\left(-z_{\alpha/2} < \sqrt{I(\theta)}(\hat{\theta}-\theta) < z_{\alpha/2}\right) \approx 1-\alpha.$$

One can either solve for θ, or use $I(\theta) \approx I(\hat{\theta})$.

Student t-distribution Let $Z \sim \mathrm{N}(0,1)$ and $Y \sim \chi^2_n$ be independent. We say that $T = Z/(\sqrt{Y/n})$ has a *student t-distribution*, and write $T \sim t_n$.

Independence of $\overline{X}$ and S Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$. Then, $\overline{X}$ and S are independent, and

$$\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right), \quad \frac{(n-1)S}{\sigma^2} \sim \chi^2_{n-1}.$$

Pivot A *pivot* is a random variable, function of both $\boldsymbol{X}$ and θ, whose distribution does not depend on θ.

Pivot (normal) Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$. Examples of pivot include

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}, \quad \frac{(n-1)S}{\sigma^2} \sim \chi^2_{n-1}.$$

CI using pivot Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$. We have

$$\mathrm{P}\left(-t_{n-1}(\alpha/2) < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{n-1}(\alpha/2)\right) = 1-\alpha,$$

where $t_{n-1}(\alpha/2)$ satisfies $\mathrm{P}(t_{n-1} > t_{n-1}(\alpha/2)) = \alpha/2$.

3 Hypothesis testing

General setup Let $X_1, \ldots, X_n$ be a random sample from $f(x;\theta)$ where $\theta \in \Theta$ is a scalar or vector parameter. Suppose we are interested in testing, for $\Theta_0 \cap \Theta_1 = \emptyset$,

- the null hypothesis H_0: $\theta \in \Theta_0$;
- against the alternative hypothesis H_1: $\theta \in \Theta_1$.

Consider a statistic $t(\boldsymbol{X})$ such that large values of $t(\boldsymbol{X})$ cast doubt on H_0, and let $t_{obs} = t(\boldsymbol{x})$ be the value observed. Then, the *p-value* is $p = \mathrm{P}(t(\boldsymbol{X}) \geq t_{obs}|H_0)$.

Critical region The *critical region* C is such that if we have $\boldsymbol{x} \in C \subset \mathbb{R}^n$, we reject H_0, and keep it otherwise.

Errors in hypothesis testing

- type I error: rejecting H_0 when H_0 is true;
- type II error: not rejecting H_0 when H_0 is false.

Size and power The *type I error probability*, also called the *size*, is

$$\alpha = \mathrm{P}(\text{reject } H_0|H_0 \text{ true}) = \sup_{\theta\in\Theta_0} \mathrm{P}(\boldsymbol{X} \in C|\theta),$$

while the *type II error probability* is

$$\beta(\theta) = \mathrm{P}(\text{don't reject } H_0|\text{true value is } \theta) = \mathrm{P}(\boldsymbol{X} \notin C|\theta),$$

and $w(\theta) = 1-\beta(\theta) = \mathrm{P}(\boldsymbol{X} \in C|\theta)$ is called the *power*. We want $w(\theta) \approx 1$ for $\theta \in \Theta_1$, and $w(\theta) \approx 0$ for $\theta_0 \in \Theta_0$.

Neyman–Pearson lemma Consider testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. Define the critical region C by

$$C = \left\{\boldsymbol{x} : \frac{L(\theta_0;\boldsymbol{x})}{L(\theta_1;\boldsymbol{x})} \leq k\right\},$$

and suppose k and α are such that $\mathrm{P}(\boldsymbol{X} \in C|H_0) = \alpha$. Then, among all tests of size $\leq \alpha$, the test with critical region C has maximum power.

Uniformly most powerful Consider testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta_1$. If the critical region is the same for all $\theta_1 \in \Theta_1$, then C is said to be *uniformly most powerful*.

Likelihood ratio Consider testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta \supset \Theta_0$. The *likelihood ratio* is defined by

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\theta\in\Theta_0} L(\theta;\boldsymbol{x})}{\sup_{\theta\in\Theta} L(\theta;\boldsymbol{x})}.$$

A *likelihood ratio test* (LRT) of H_0 against H_1 has critical region $C = \{\boldsymbol{x} : \lambda(\boldsymbol{x}) \leq k\}$. For a test of size α, we must choose k such that $\sup_{\theta\in\Theta_0} \mathrm{P}(\lambda(\boldsymbol{X}) \leq k|\theta) = \alpha$.

Likelihood ratio statistic The *likelihood ratio statistic* is defined by $\Lambda(\boldsymbol{X}) = -2\log\lambda(\boldsymbol{X})$. For Λ, the LRT has critical region $C = \{\boldsymbol{x} : \Lambda(\boldsymbol{x}) \geq k\}$. If H_0 is true, then

$$\Lambda(\boldsymbol{x}) \xrightarrow{d} \chi^2_p \text{ as } n \to \infty, \quad p = \dim\Theta - \dim\Theta_0.$$

For a size-α test, we choose k such that $\mathrm{P}(\chi^2_p \geq k) = \alpha$.

Goodness of fit tests Consider n independent observations of categories $i = 1, \ldots, k$. Let n_i be the number of observations in category i, with $\sum_i n_i = n$, and π_i the probability of being category i, with $\sum_i \pi_i = 1$. We test

- the null hypothesis $H_0 : \pi_i = \pi_i(\theta)$, where $\theta \in \Theta$, and $\dim H_0 = q < k-1$;
- against the general alternative H_1 : the π_i are unrestricted except for $\sum_i \pi_i = 1$, and $\dim H_1 = k-1$.

In this case, the likelihood ratio statistic is given by

$$\Lambda = -2\log\frac{\sup_{H_0} L(\theta;\boldsymbol{x})}{\sup_{H_1} L(\theta;\boldsymbol{x})} = 2\sum_{i=1}^{k} n_i \log\left(\frac{n_i}{n_i\pi_i(\hat{\theta})}\right),$$

and $\Lambda \approx \chi^2_{k-1-q}$ when H_0 is true.

4 Bayesian inference

Notation We write $f(\boldsymbol{x}|\theta)$ instead of $f(\boldsymbol{x};\theta)$ to emphasize that we have a model for data $\boldsymbol{x}$ given the value θ.

Prior We summarize our beliefs about θ in a *prior* pmf/pdf $\pi(\theta)$; we treat θ as a random variable.

Posterior Using the continuous Bayes' theorem,

$$f_{Z|Y}(z|y) = \frac{f_{Z|Y}(y|z)f_Z(z)}{f_Y(y)},$$

we define the *posterior* pmf/pdf as

$$\pi(\theta|\boldsymbol{x}) \propto f(\boldsymbol{x}|\theta) \times \pi(\theta),$$
$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

Posterior summaries

- the *posterior mode* (θ that maximizes $\pi(\theta|\boldsymbol{x})$);
- the *posterior mean* (expectation over θ);
- the *posterior variance* (variance over θ);
- the *posterior median* (satisfies $\int_{-\infty}^{m} \pi(\theta|x)d\theta = 1/2$);
- other quantiles of $\pi(\theta|\boldsymbol{x})$.

Credible interval A $100(1-\alpha)\%$ *credible interval* for θ is an interval (θ_a, θ_b) such that

$$\mathrm{P}(\theta_a \le \theta \le \theta_b|\boldsymbol{x}) = \int_{\theta_a}^{\theta_b} \pi(\theta|\boldsymbol{x})d\theta = 1-\alpha.$$

If $\mathrm{P}(\theta \le \theta_a|\boldsymbol{x}) = \mathrm{P}(\theta \ge \theta_b|\boldsymbol{x}) = \alpha/2$, then the interval is called *equal-tailed*.

Highest posterior density A credible interval I is called a *highest posterior density* interval if $\pi(\theta|\boldsymbol{x}) \ge \pi(\theta'|\boldsymbol{x})$, for all $\theta \in I$ and $\theta' \notin I$. (For the normal, every equal-tailed credible interval is a highest posterior density interval.)

Posterior predictive density Let X_{n+1} represent a future observation, independent of $X_1, \ldots, X_n$, and let $\boldsymbol{x} = (x_1, \ldots, x_n)$ denote the observed data. The pdf of X_{n+1}, the *posterior predictive density*, is defined by

$$f(x_{n+1}|\boldsymbol{x}) = \int_\Theta f(x_{n+1}, \theta|\boldsymbol{x})d\theta = \int_\Theta f(x_{n+1}|\theta)\pi(\theta|\boldsymbol{x})d\theta,$$

where we used $\mathrm{P}(A \cap B|C) = \mathrm{P}(A|B \cap C)\mathrm{P}(B|C)$.

Proper prior A prior is *proper* if $\int \pi = 1$, and is *improper* if the integral cannot be normalized.

Jeffreys prior The *Jeffreys prior* is defined by

$$\pi(\theta) \propto I(\theta)^{1/2},$$

where $I(\theta)$ is the expected information.

Prior/posterior odds Suppose we want to compare two hypotheses H_0 and H_1, exactly one of which is true. The *prior* and *posterior odds* of H_0 relative to H_1 are

$$\text{prior odds} = \frac{\mathrm{P}(H_0)}{\mathrm{P}(H_1)}, \quad \text{posterior odds} = \frac{\mathrm{P}(H_0|\boldsymbol{x})}{\mathrm{P}(H_1|\boldsymbol{x})}.$$

Bayes factor The *Bayes factor* B_{01} is defined via

$$\frac{\mathrm{P}(H_0|\boldsymbol{x})}{\mathrm{P}(H_1|\boldsymbol{x})} = \frac{P(\boldsymbol{x}|H_0)}{\mathrm{P}(\boldsymbol{x}|H_1)} \times \frac{\mathrm{P}(H_0)}{\mathrm{P}(H_1)},$$
$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}.$$

General setup We are assuming we have:

- prior probabilities $\mathrm{P}(H_0)$ and $\mathrm{P}(H_1)$, which satisfy $\mathrm{P}(H_0) + \mathrm{P}(H_1) = 1$;
- prior distributions for $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$ under H_0 and H_1, which we write as $\pi(\theta_0|H_0)$ and $\pi(\theta_1|H_1)$;
- models for data $\boldsymbol{x}$ under H_0 and H_1, which we write as $f(\boldsymbol{x}|\theta_0, H_0)$ and $f(\boldsymbol{x}|\theta_1, H_1)$.

Computing the Bayes factor Use

$$\mathrm{P}(H_i) = \int_{\Theta_i} \pi(\theta_i)d\theta_i \quad \text{and} \quad \mathrm{P}(H_i|\boldsymbol{x}) = \int_{\Theta_i} \pi(\theta_i|\boldsymbol{x})d\theta_i.$$

Alternatively, use

$$\pi(\theta_i|H_i) = \pi(\theta_i) \Big/ \int_{\Theta_i} \pi(\theta_i)d\theta_i,$$

and

$$\mathrm{P}(\boldsymbol{x}|H_i) = \int_{\Theta_i} f(\boldsymbol{x}|\theta_i, H_i)\pi(\theta_i|H_i)d\theta_i.$$

Note that if $H_i : \theta = \theta_i$, then $\mathrm{P}(\boldsymbol{x}|H_i) = f(\boldsymbol{x}|\theta_i)$.

Assessing evidence The quantity $2\log B_{01}$ is used to summarize the evidence for H_0 compared to H_1:

B_{01}	$2\log B_{01}$	Evidence for H_0
< 1	< 0	Negative
$1-3$	$0-2$	Hardly worth a mention
$3-20$	$2-6$	Positive
$20-150$	$6-10$	Strong
> 150	> 10	Very strong

Asymptotic normality Let $\tilde{\ell}(\theta) = \log \pi(\theta|\boldsymbol{x})$ and $\tilde{\theta}$ be the posterior mode. Expanding $\tilde{\ell}(\theta)$ yields

$$\tilde{\ell}(\theta) \approx \tilde{\ell}(\tilde{\theta}) + (\theta - \tilde{\theta})\tilde{\ell}'(\tilde{\theta}) + \frac{1}{2}(\theta - \tilde{\theta})^2\tilde{\ell}''(\tilde{\theta}),$$

with $\tilde{\ell}'(\tilde{\theta}) = 0$. Therefore,

$$\pi(\theta|\boldsymbol{x}) = \exp(\tilde{\ell}(\theta)) \propto \exp\left(-\frac{1}{2}\tilde{J}(\tilde{\theta})(\theta - \tilde{\theta})^2\right),$$

i.e., $\theta|\boldsymbol{x} \overset{d}{\approx} \mathrm{N}(\tilde{\theta}, \tilde{J}(\tilde{\theta})^{-1})$. (Similarly, $\theta|\boldsymbol{x} \overset{d}{\approx} \mathrm{N}(\hat{\theta}, J(\hat{\theta})^{-1})$.)