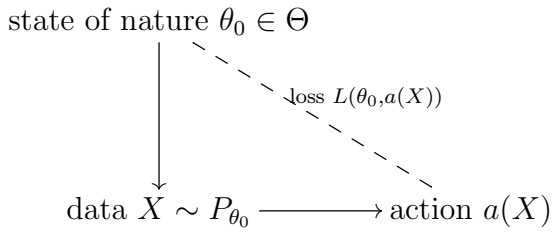
Statistics catch-up lecture

Tengyao Wang
t.wang@statslab.cam.ac.uk

October 2014

1 Introduction

Statistics is the practice of making inference or decisions based on data. It is concerned with the interaction between *states of nature*, *data* and human *actions*. The relationship can be best summarised in the following diagram.



There is some state of nature θ_0 unknown to the statistician, that belongs to a space Θ. Data X is generated randomly from a distribution P_{θ_0} that depends on the state of nature. An action $a(X)$ (e.g. an inference or a decision) is required based on the observation X. A *loss function* $L(\cdot, \cdot)$ is used to measure the cost of the action with respect to the (hidden) state of nature.

Example 1. In a genetic study, $n = 2217$ patients are recruited, among which 1127 have bladder cancer. After whole genome sequencing, $p = 475290$ single nucleatide polymorphisms (or SNPs, a possible variation of the DNA sequence at one location) are measured on each patient. We construct a matrix $X = (X_{ij})_{1 \le i \le n, 1 \le j \le p}$ where $X_{ij} = 1$ if the ith patient has a mutation at jth location and 0 otherwise. Denote $X_i^\top$ for the ith row of X. Let $Y = (Y_i)_{1 \le i \le n}$ be a response vector, where $Y_i = 1$ if the ith patient has bladder cancer and 0 otherwise. The goal is to understand how SNPs can be used to predict bladder cancer. One possible way to analyse the data is to assume a logistic model, i.e.

$$\log \frac{\mathbb{P}(Y_i = 1)}{\mathbb{P}(Y_i = 0)} = X_i^\top \beta_0, \quad i = 1, \ldots, n.$$

Here $\beta_0 = \big((\beta_0)_j\big)_{1\le j\le p}$ is our (unknown) state of nature. Distribution of the data Y is completely determined by β_0. Our "action" is to come up with an estimator for β_0. Suppose we believe that at most $k = 50$ SNPs are relevant to bladder cancer. One possible way to do so is then to let $\hat{\beta} = \hat{\beta}(Y)$ be a solution to the optimisation problem

$$\max_{\beta} \quad \prod_{i=1}^{n} \frac{Y_i e^{X_i^\top \beta} + (1 - Y_i)}{e^{X_i^\top \beta} + 1} \quad \text{subject to} \quad \sum_{i=1}^{p} \mathbf{1}_{\{\beta_j \neq 0\}} \le k.$$

A possible loss function in this case will be $L(\beta_0, \hat{\beta}) = \|\beta_0 - \hat{\beta}\|^2$.

The above example illustrates various aspects of statistics. Some of which are discussed below.

1. Given a real world problem, we make some simplifying assumptions to enable us to analyse it statistically. We may want to validate those assumptions or choose between competing assumptions using data. (statistical modelling, model selection)

2. One important assumption to make in the model is the space Θ of the possible states of nature. In the above example, $\Theta = \{\beta : \sum_{i=1}^{p} \mathbf{1}_{\{\beta_j \neq 0\}} \le k\}$. Naturally, the larger Θ is, the harder it will be to estimate the state of nature, or make other decisions. Sometimes, Θ can be an infinite-dimensional space. (parametric statistics vs. nonparametric statistics)

3. With a statistical model and data at hand, statisticians need to come up with good rules for actions. (statistical methodology)

4. For a particular action, e.g. $\hat{\beta}$ in the example, we may want to evaluate quantitatively how well it performs in terms of the loss function. (statistical theory)

2 Probability

A formal treatment of basics of probability theory can be found in the Probability and Measure catch-up lecture. Here, we review several important concepts useful in statistics.

2.1 Discrete random variables

Formally, a *random variable* is a measurable map $X : \Omega \to \mathcal{X}$ from a sample space Ω to the state space $\mathcal{X}$. For example, the smallest number of dice rolls to see all six values at least once is a random variable, where the probability space is $\Omega = \big\{(a_1, a_2, a_3, \ldots) : a_i \in \{1, 2, 3, 4, 5, 6\} \text{ for all } i\big\}$ and the state space is $\mathcal{X} = \mathbb{N}$.

If $\mathcal{X}$ is finite or countably infinite, we say the random variable X is *discrete*. A discrete random variable is completely specified by its probability mass function

$$p_X(x) := \mathbb{P}(X = x), \quad x \in \mathcal{X}.$$

A *binomial* random variable X with parameters $n \in \mathbb{N}$ and $p \in [0, 1]$ (written $X \sim \text{Bin}(n, p)$) is the number of successes in n trials where each trial has probability p of being successful. It has probability mass function

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, \ldots, n\}.$$

In the special case when $n = 1$ we have a *Bernoulli* random variable with parameter p (written $X \sim \text{Bern}(p)$).

A *negative binomial* random variable X with parameters $k \in \mathbb{N}$ and $p \in (0, 1]$ (written $X \sim \text{NegBin}(k, p)$) is the number of trials needed to observe k successes, where each trial has success probability p. It has probability mass function

$$\mathbb{P}(X = n) = \binom{n-1}{k-1} p^k (1-p)^{n-k}, \quad n \in \{k, k+1, \ldots\}.$$

The special case where $k = 1$ is called a *geometric* random variable with parameter p (written $X \sim \text{Geom}(p)$).

A *Poisson* random variable X with parameter λ (written $X \sim \text{Poi}(\lambda)$) has probability mass function

$$\mathbb{P}(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad k \in \mathbb{N}.$$

2.2 Continuous random variables

Often, the random variable X takes value in $\mathcal{X} \subseteq \mathbb{R}$, in which case we can define its *cumulative distribution function* (cdf) by

$$F_X(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}.$$

The cdf is always right-continuous and nondecreasing. If it is continuous, we call X a *continuous random variable*. Furthermore, if $F_X(x)$ is differentiable, the derivative $f_X(x) = F_X'(x)$ is called the *probability density function* (pdf). Continuous random variables are completely determined by their cumulative distribution function, or (if exist) probability density function.

The most important continuous distribution is the normal distribution (or Gaussian distribution). A *normal random variable* X with mean parameter μ and variance parameter σ^2 (written $X \sim N(\mu, \sigma^2)$) has pdf

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

A *uniform random variable* X on a set $A \subseteq \mathbb{R}$ (written $X \sim \text{Unif}(A)$) has pdf

$$f_X(x) = \begin{cases} \lambda(A)^{-1} & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases},$$

where $\lambda(A)$ is the Lebesgue measure of A.

A *Gamma random variable* X with shape parameter α and rate parameter λ (written $X \sim \text{Gamma}(\alpha, \lambda)$) has pdf

$$f_X(x) = \frac{e^{-\lambda x}\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)}, \quad x \in [0, \infty),$$

where $\Gamma(\cdot)$ is the Gamma function. In the special case where $\alpha = 1$, we have an *exponential random variable* with rate parameter λ (written $X \sim \text{Exp}(\lambda)$). In the special case where $\alpha = k/2$ and $\lambda = 1/2$, we have a *chi-square random variable* with degree of freedom parameter k (written $X \sim \chi^2_k$). A chi-square random variable with k degree of freedom can also be characterised by being the sum of squares of k independent $N(0, 1)$ random variables.

Closely related to the chi-square distributions are *t-distribution* and *F-distribution*. If $Z \sim N(0, 1)$, $U \sim \chi^2_k$ and $V \sim \chi^2_\ell$ are independent, then $T = Z/\sqrt{V/\ell}$ follows a t-distribution with ℓ degrees of freedom (written $T \sim t_\ell$), and $F = (U/k)/(V/\ell)$ follows an F-distribution with parameters (k, ℓ) (written $F \sim F_{k,\ell}$). Note that square of a t_ℓ distribution is an $F_{1,\ell}$ distribution.

2.3 Random vectors

If $X_1, \ldots, X_p$ are random variables, then $X = (X_1, \ldots, X_p)^\top$ is a random vector. The pdf f_X of X is the nonnegative function satisfying

$$\mathbb{P}(X \in A) = \int_{(x_1,\ldots,x_p)\in A} f_X(x_1, \ldots, x_n)\, dx_1 \cdots dx_p.$$

The most important random vector is a multivariate normal random vector. The desnity function of a p-variate normal random vector with mean $\mu \in \mathbb{R}^p$ and variance $\Sigma \in \mathbb{R}^{p\times p}$ (written $X \sim N_p(\mu, \Sigma)$) is

$$f_X(x) = \frac{1}{(2\pi)^{p/2}(\det \Sigma)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}, \quad x \in \mathbb{R}^p.$$

The normality of a random vector is preserved under linear transformation and translation. If $X \sim N_p(\mu, \Sigma)$, and $A \in \mathbb{R}^{q\times p}$ and $\alpha \in \mathbb{R}^q$, then $AX + \alpha \sim N_p(A\mu + \alpha, A\Sigma A^\top)$.

2.4 Moments and related quantities

The kth moment of a random variable X (if exists) is defined as $\mathbb{E}(X^k)$. The first moment $\mu = \mathbb{E}X$ is its mean. The "centred" second moment is its variance $\sigma^2 = \mathbb{E}\{(X-\mu)^2\}$. The "centred" third and fourth moments are related to the skewness ($\mathbb{E}\{(X-\mu)^3\}/\sigma^3$) and kurtosis ($\mathbb{E}\{(X-\mu)^4\}/\sigma^4$) respectively. The covariance of two random variables X, Y is defined by $\mathrm{Cov}(X,Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)$.

Expectation, variance and covariance naturally generalise to multivariate cases. If $X = (X_1, \ldots, X_p)^\top$ and $Y = (Y_1, \ldots, Y_q)^\top$ are a random vector, then $\mathrm{Var}(X) = \mathbb{E}\{(X - \mathbb{E}X)(X - \mathbb{E}X)^\top\}$, and $\mathrm{Cov}(X,Y) = \mathbb{E}\{(X - \mathbb{E}X)(Y - \mathbb{E}Y)^\top\}$. Note that $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$. Furthermore, it can be shown directly from definition that if $A \in \mathbb{R}^{r\times p}$ and $B \in \mathbb{R}^{s\times q}$ are matrices, then $\mathrm{Cov}(AX, BY) = A\mathrm{Cov}(X,Y)B^\top$. Also, if $p = q$, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + \mathrm{Cov}(X,Y) + \mathrm{Cov}(Y,X)$.

2.5 Notions of convergence

Since random variables are functions from sample space Ω to some state space $\mathcal{X}$, it is not surprising that convergence is a more delicate notion for a sequence of random variables than a sequence of non-random points in $\mathcal{X}$. There are three notions of convergence commonly used in probability and statistics. Let $X, X_1, X_2, X_3, \ldots$ be random variables defined on the same sample space Ω. Then we say (X_n) *converges almost surely* to X, and write $X_n \xrightarrow{\text{a.s.}} X$, if

$$\mathbb{P}\{\omega \in \Omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty\} = 1.$$

In other words, if we view the random variables as functions from Ω to $\mathcal{X}$, then almost sure convergence is a pointwise convergence at almost all points in Ω with respect to the probability measure $\mathbb{P}$. The second notion of convergence is *convergence in probability*. We say (X_n) converges in probability to X, and write $X_n \xrightarrow{\text{p}} X$, if

$$\forall \epsilon, \quad \mathbb{P}\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \le \epsilon\} \to 1, \quad \text{as } n \to \infty.$$

Intuitively, convergence in probability says that large deviations from the limit X becomes increasingly rare as n grows. The last notion of convergence is *convergence in distribution*, sometimes also called convergence in law or weak convergence. Here, $X_1, X_2, \ldots$ need not be defined on the same sample space (but their images still share the same state space). Let $X_n : \Omega_n \to \mathcal{X}$ for $n = 1, 2, \ldots$ and $X : \Omega \to \mathcal{X}$. Then we say (X_n) converges to X in distribution, and write $X_n \xrightarrow{\text{d}} X$, if for all bounded continuous functions $f : \mathcal{X} \to \mathbb{R}$,

$$\mathbb{E}f(X_n) \to \mathbb{E}f(X), \quad \text{as } n \to \infty.$$

The three notions of convergence satisfy the following relation:

$$X_n \xrightarrow{\text{a.s.}} X \quad \Rightarrow \quad X_n \xrightarrow{\text{p}} X \quad \Rightarrow \quad X_n \xrightarrow{\text{d}} X.$$

Furthermore, neither of the above implications can be reversed.

2.6 Limit theorems

Many theoretical statistics results concern with the performance of estimators in the asymptotic regime (i.e. what happens if we are allowed to perform the estimation based on increasingly larger samples). Such analysis often makes use of the limit theorems from probability theory.

Theorem 1 (Strong Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be independent and identically distributed (i.i.d.) random variables such that $\mu := \mathbb{E}(X_1) < \infty$. Let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ be the sample mean of first n random variables. Then*

$$\bar{X}_n \xrightarrow{\text{a.s.}} \mu.$$

Theorem 2 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables such that $\mathbb{E}X_1 = \mu$ and $\mathrm{Var}(X_1) = \sigma^2$. Then*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\text{d}} N(0, \sigma^2).$$

The Strong Law of Large Numbers states that the sample mean of i.i.d. samples is a "consistent" estimator for the first moment. The Central Limit Theorem then characterises how large the fluctuation from the true mean μ can be.

2.7 Order in probability notation

Related to the concept of convergence is the "order in probability" notation. For a sequence of random variables $(X_n : \Omega \to \mathcal{X})$ and a sequence of non-random points $a_n \in \mathcal{X}$. We write $X_n = o_{\text{p}}(a_n)$ if $X_n/a_n \xrightarrow{\text{p}} 0$. We write $X_n = O_{\text{p}}(a_n)$ if for every ϵ, there exists $M > 0$ such that for all n, $\mathbb{P}(|X_n/a_n| \geq M) \leq \epsilon$.

Example 2. Let $X_1, X_2, \ldots$ be i.i.d. random variables such that the first two moments of X_1 is finite. Then the Strong Law of Large Numbers implies that $\bar{X}_n - \mu = o_{\text{p}}(1)$, whereas the Central Limit Theorem gives the more precise statement that $\bar{X}_n - \mu = O_{\text{p}}(1/\sqrt{n})$.

3 Linear Model

We now focus on arguably one of the most important statistical models and use it to illustrate several ideas in statistics.

Suppose we have data $(X_1, y_1), \ldots, (X_n, y_n)$, where $X_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ for $1 \leq i \leq n$. The *regression* model in statistics view X_is as the predictors and y_is the responses. It assumes that y_i is a particular realisation of a random variable Y_i (in other words, $y_i = Y_i(\omega_0)$ for some $\omega_0 \in \Omega$), whose distribution is determined by X_i and an unknown parameter θ_0. The linear model is a specific regression model. It assumes that $Y_i =$

$X_i^\top \beta_0 + \epsilon_i$ for all i, where $\beta_0 \in \mathbb{R}^p$, $\epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} N(0, \sigma_0^2)$. The aim is to use X_i to explain Y_i; in other words, we want to estimate β_0.

To simplify notation, we collect $X_1, \ldots, X_n$ into a matrix X whose ith row is $X_i^\top$, and combine $Y_1, \ldots, Y_n$ into a column vector Y. If we assume the "design matrix" X is fixed, then the only randomness in the data comes from Y. The distribution of Y is completely determined by the state of nature (β_0, σ_0^2) in the sense that $Y \sim N_n(X\beta_0, \sigma_0^2 I_n)$. For a given β, σ^2, and a particular realisation y for the random variable Y, the likelihood function $L(\beta, \sigma^2; y)$ of β, σ^2 with respect to y is equal to the density function of y with respect to β, σ^2:

$$L(\beta, \sigma^2; y) := f(y; \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\|y - X\beta\|^2}.$$

The *maximum likelihood estimator* for β_0 and (the nuisance parameter) σ_0^2 is obtained by maximising $L(\beta, \sigma^2; y)$ with respect to β and σ^2 when $y = (y_1, \ldots, y_n)^\top$. The optimisation problem can be solved uniquely if and only if X has rank p (i.e. all columns of X are linearly independent). The maximisers are $\hat{\beta} = (X^\top X)^{-1} X^\top y$ and $\hat{\sigma}^2 = \frac{1}{n}\|y - X\hat{\beta}\|^2$.

A remarkable fact about the linear model is that we have perfect understanding of the estimators $\hat{\beta}$ and $\hat{\sigma}^2$. To be completely clear about our terminology, we call $\hat{\beta}(y) = (X^\top X)^{-1} X^\top y$ an "estimate", which is a fixed quantity, and $\hat{\beta}(Y) = (X^\top X)^{-1} X^\top Y$ an "estimator", which is a random variable. From properties of multivariate normal distribution, $\hat{\beta} \sim N_p(\beta_0, \sigma_0^2 (X^\top X)^{-1})$. The fitted values (as random variables) $\hat{Y} = X\hat{\beta} = X(X^\top X)^{-1} X^\top Y$ can be interpreted as the projection of Y to the column space of matrix X. We call $P = X(X^\top X)^{-1} X^\top$ the "projection matrix" or the "hat matrix". Finally, using Cochran's theorem, we deduce that $\hat{\sigma}^2 \sim \frac{\sigma_0^2}{n} \chi^2_{n-p}$ and $\hat{\sigma}^2$ is independent of $\hat{\beta}$.

Theorem 3 (Cochran)**.** *Let $Y \sim N_n(0, I_n)$, and $A_1, \ldots, A_k$ symmetric $n \times n$ matrices where $\operatorname{rank}(A_i) = r_i$, $\sum_{i=1}^k A_i = I_n$ and $\sum_{i=1}^k r_i = n$. Then $Y^\top A_i Y \sim \chi^2_{r_i}$ for $1 \le i \le k$ independently.*

From the above calculation, we know that the jth coefficient of the maximum likelihood estimator $\hat{\beta}_j \sim N\big((\beta_0)_j, \sigma_0^2 [(X^\top X)^{-1}]_{jj}\big)$. Thus,

$$\frac{(\hat{\beta}_j - (\beta_0)_j)/\sqrt{\sigma_0^2 [(X^\top X)^{-1}]_{jj}}}{\sqrt{n\hat{\sigma}^2/(\sigma_0^2(n-p))}} = \frac{\hat{\beta}_j - (\beta_0)_j}{\hat{\sigma}} \sqrt{\frac{n-p}{n[(X^\top X)^{-1}]_{jj}}} \sim t_{n-p}.$$

Let α be the upper-0.025 quantile of a t_{n-p} distribution (for $n - p$ large, $\alpha \approx 2$). Then

$$\text{CI} = \Big[\hat{\beta}_j - \alpha\hat{\sigma}\sqrt{n[(X^\top X)^{-1}]_{jj}/(n-p)},\ \hat{\beta}_j + \alpha\hat{\sigma}\sqrt{n[(X^\top X)^{-1}]_{jj}/(n-p)}\ \Big]$$

is a 95% confidence interval for $(\beta_0)_j$. If we want to test the null hypothesis $H_0 : (\beta_0)_j = 0$ against the alternative hypothesis $H_1 : (\beta_0)_j \neq 0$, then we can reject the null hypothesis at 0.05 significant level if $0 \notin \text{CI}$.

4 Maximum Likelihood Estimation

As mentioned before, the likelihood function is just the probability density function viewed from a different perspective: $L(\theta; y) = f(y; \theta)$. It is often easier to work with the log-likelihood function $\ell(\theta; y) = \log L(\theta; y)$. Suppose $\{P_\theta : \theta \in \Theta\}$ is a family of distributions indexed by θ and data Y has distribution P_{θ_0} for $\theta_0 \in \Theta$. Maximum likelihood estimator aims to estimate θ_0 by maximising $L(\theta; Y)$, or equivalently maximising $\ell(\theta; Y)$.

If the log-likelihood function is differentiable in θ, we can define the *score function* as its first derivative in θ:

$$u(\theta; Y) = \frac{\partial \ell(\theta; Y)}{\partial \theta}.$$

When score function exists, the maximum likelihood estimation is equivalent to solving the score equation $u(\theta; Y) = 0$. If the log-likelihood function is twice differentiable in θ, we call its negative Hessian matrix the *observed information matrix*

$$j(\theta; Y) = -\frac{\partial}{\partial \theta}\left(\frac{\partial \ell(\theta; Y)}{\partial \theta}\right)^\top.$$

Note that $u(\theta; Y)$ is a random vector and $j(\theta; Y)$ is a random matrix. The expectation (over $Y \sim P_{\theta_0}$) of j at θ_0 is called the *Fisher information matrix*:

$$i(\theta_0) = \mathbb{E}_{\theta_0}(j(\theta_0; Y))$$

Fisher information matrix measures on average how sharp the log-likelihood function peaks at its maximum, which in turn gives a measure of the ease of estimating θ using the MLE.

Maximum likelihood estimators are popular among statisticians primarily because of its many nice properties when the sample size is large. Suppose we are in a setting where the data Y is a vector of n i.i.d. observations $Y_1, \ldots, Y_n$ with marginal densities $f(\cdot; \theta_0)$. Then the joint log-likelihood is

$$\ell(\theta; Y_1, \ldots, Y_n) = \log f(Y_1, \ldots, Y_n; \theta) = \log \prod_{i=1}^n f(Y_i; \theta) = \sum_{i=1}^n \ell(\theta; Y_i).$$

Example 3. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poi}(\lambda_0)$ for some $\lambda_0 > 0$. The joint log-likelihood function is

$$\ell(\lambda; X_1, \ldots, X_n) = \sum_{i=1}^n \big(-\lambda + X_i \log \lambda - \log(X_i!)\big).$$

Hence

$$\frac{\partial \ell}{\partial \lambda} = -n + \frac{\sum_{i=1}^n X_i}{\lambda}.$$

Thus, the MLE is $\hat{\lambda} = n^{-1} \sum_{i=1}^n X_i$.

In this setting, under mild regularity conditions (such as assuming smoothness conditions on ℓ and permitting interchange of the order of integration and differentiation), the maximum likelihood estimator $\theta_n = \hat{\theta}_n(Y_1, \ldots, Y_n)$ can be shown to be consistent (so in particular asymptotically unbiased)

$$\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0.$$

and asymptotically normally distributed

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\text{d}} N(0, i^{(1)}(\theta_0),$$

where $i^{(1)}(\theta_0) = \mathbb{E}_{\theta_0}(j(\theta_0; Y_1))$ is the Fisher information matrix of the first observation (which is the same as the information at all other observations). Furthermore, the Cramer–Rao lower bound implies that the MLE has asymptotically the smallest variance among all asymptotically unbiased estimators.

Example 4. In the linear model, $\theta_0 = (\beta_0, \sigma_0^2)$, $\ell(\beta, \sigma^2; Y_i) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(Y_i - X_i^\top \beta)^2$. We calculate that

$$\frac{\partial \ell(\beta, \sigma^2; Y_i)}{\partial \beta} = \frac{1}{\sigma^2}(Y_i - X_i\beta)X_i^\top, \quad \frac{\partial \ell(\beta, \sigma^2; Y_i)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(Y_i - X_i^\top \beta)^2.$$
$$\frac{\partial}{\partial \beta}\left(\frac{\partial \ell(\beta, \sigma^2; Y_i)}{\partial \beta}\right)^\top = \frac{-X_i X_i^\top}{\sigma^2}, \quad \frac{\partial}{\partial \sigma^2}\left(\frac{\partial \ell(\beta, \sigma^2; Y_i)}{\partial \beta}\right)^\top = -\frac{1}{\sigma^4}(Y_i - X_i\beta)X_i^\top,$$
$$\frac{\partial}{\partial \sigma^2}\left(\frac{\partial \ell(\beta, \sigma^2; Y_i)}{\partial \sigma^2}\right)^\top = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(Y_i - X_i^\top \beta)^2.$$

Thus the Fisher information matrix of the ith observation is

$$i(\theta_0; Y_i) = \begin{pmatrix} (X_i X_i^\top)/\sigma^2 & 0 \\ 0 & (2\sigma^4)^{-1} \end{pmatrix}$$

The observations $Y_1, \ldots, Y_n$ are not exactly i.i.d.. However, if $n^{-1}X^\top X = \sum_{i=1}^n X_i X_i^\top \to L$ as $n \to \infty$, then a slight modification of the argument for the MLE shows that

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{\text{d}} N_p(0, \sigma^2 L^{-1}), \quad \sqrt{n}(\hat{\sigma}^2 - \sigma_0^2) \xrightarrow{\text{d}} N(0, 2\sigma^4).$$

5 Confidence interval

The MLE is an example of a point estimator. Often, it is desirable to indicate the reliability of the point estimator by supplying a range of values in Θ. This range is called a *confidence set*. In the case where the parameter space Θ is a subset of the real line, and the range of values is taken to be an interval on $\mathbb{R}$, the confidence set is called a *confidence interval*.

Example 5. In a recent poll, $n = 15724$ Americans are asked whether they approval the Barack Obama's job as the President. Among them, $y = 6916$ answered yes. If the population approval rate is p_0, then the number of yes answers Y among n random Americans has distribution $Y \sim \text{Bin}(n, p_0)$. The maximum likelihood estimator for p_0 is $\hat{p} = Y/n$, which has mean p_0 and variance $p_0(1-p_0)/n$. From central limit theorem, $\sqrt{n}(p_0 - \hat{p}_0)$ is approximately $N(0, p_0(1-p_0))$. The approximate 95% confidence interval is $[\hat{p} - 2n^{-1/2}\sqrt{\hat{p}(1-\hat{p})}, \hat{p} + 2n^{-1/2}\sqrt{\hat{p}(1-\hat{p})}]$. Plugging in the numbers, we find the maximum likelihood estimate is $\hat{p} = y/n = 0.440$ and the realisation of confidence interval is $[0.432, 0.448]$.

A common and quite tempting mistake is say in the preceding example that the true value p_0 lies in $[0.432, 0.448]$ approximately 95% of the time. This is not true since the population parameter p_0 either belongs to the interval or it does not. This common misconception is caused by the unfortunate fact that statisticians use the term "confidence interval" to mean both the random interval constructed from the random data and the particular realisation of it given the observed data. It would be much clearer if we can call them "confidence interval estimator" (random) and "confidence interval estimate" (non-random) respectively.

In the above example, the "confidence interval estimator" is

$$\big[\hat{p}(Y) - 2n^{-1/2}\sqrt{\hat{p}(Y)(1-\hat{p}(Y))},\ \hat{p}(Y) + 2n^{-1/2}\sqrt{\hat{p}(Y)(1-\hat{p}(Y))}\big],$$

and it contains p_0 with probability 95%. On the other hand, the "confidence interval estimate" is $[0.432, 0.448]$.

6 Hypothesis testing

Another important area of statistical inference is hypothesis testing. A statistical *hypothesis* is an assertion about the distribution of some random variables, and a *test* is a procedure that decides whether to reject the hypothesis.

In the general framework, the data X is randomly generated with a density $f(\cdot; \theta_0)$ where θ_0 is some point in the parameter space Θ. We have two competing hypothesis. We want to test the null hypothesis $H_0 : \theta_0 \in \Theta_0$ against the alternative hypothesis $H_1 : \theta_0 \in \Theta_0$, where $\Theta_0 \cap \Theta_1 = \emptyset$ (but it is not necessary for $\Theta_0 \cup \Theta_1 = \Theta$). The test ψ is a $\{0, 1\}$-valued function of the random variable X. We reject the null hypothesis if and only if $\psi(X) = 1$. There are two possible types of errors one can make using the test ψ. *Type I error* occurs if we reject the null hypothesis when it is true. The maximum probability of Type I error is called the *size* or *significance level* of the test, denoted by

$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\psi(X) = 1).$$

Type II error occurs if we fail to reject the null hypothesis when it is false. The maximum probability of Type II error is 1 minus the *power* of the test, denoted by

$$\beta = \sup_{\theta \in \Theta_1} \mathbb{P}_\theta(\psi(X) = 0).$$

A common mistake is to say things like "we do not reject the null hypothesis at size 0.05, which means the null hypothesis is true with 95% probability. Similar to confidence intervals, any probabilistic interpretation for a hypothesis test is always a prior probability, which is probabilistic statement solely about the procedure before any actual data is seen.

Usually, H_0 is the "default" hypothesis, e.g. a new drug has no treatment value for a disease, or the Higgs boson has not been observed. It is usually more costly to make a Type I error than a Type II error. Hence, the general practice in hypothesis testing is to construct a test with a particular size α and then analyse its power $1 - \beta$. To get a test with a fixed size, say $\alpha = 0.05$, is equivalent to find an event with at least 95% probability for any $\theta \in \Theta_0$. Confidence intervals are often useful in constructing such events.

Example 6. Following Example 5, if we want to test the null hypothesis H_0 : Obama's job approval rate is 50%. Then we can define the test to be

$$\psi(Y) = \mathbf{1}\Big\{0.5 \notin \big[\hat{p}(Y) - 2n^{-1/2}\sqrt{\hat{p}(Y)(1-\hat{p}(Y))},\ \hat{p}(Y) + 2n^{-1/2}\sqrt{\hat{p}(Y)(1-\hat{p}(Y))}\big]\Big\}$$

The test result using the observed data $y = 6916$ is

$$\psi(y) = \mathbf{1}\Big\{0.5 \notin [0.432, 0.448]\Big\} = 1,$$

which says that there is less than 5% chance of observing anything equal or more extreme than the poll result if the true approval rate is 50%.

7 Bayesian inference

In both confidence interval estimation and hypothesis testing, we say it is incorrect to say things like "the probability of $[0.432, 0.448]$ containing θ_0 is 95%" or "the probability of H_0 being true is smaller than 5%". Any statement as such would require us to treat the true parameter θ_0 as a random variable. This is handled in Bayesian inference, which is built upon the Bayes' rule from conditional probability:

$$f_{\text{posterior}}(\theta|y) = \frac{f_Y(y|\theta) f_{\text{prior}}(\theta)}{\int_{\tilde{\theta}} f_Y(y|\tilde{\theta}) f_{\text{prior}}(\tilde{\theta})\, d\tilde{\theta}},$$

where f_{prior} is the prior density for θ, $f_Y(y|\theta) = L(\theta; y)$ is the conditional density of y given θ (which is exactly the likelihood function) and $f_{\text{posterior}}$ is the posterior density for θ conditional on having seen data y. In short, the Bayes' rule tells us that the posterior is proportional to prior times likelihood.

Example 7. The level of a tumour marker is elevated in 96% of pancreatic cancer patients (sensitivity) and not elevated in 99.5% of patients without pancreatic cancer (specificity). In a random screening, John finds his level of this tumour marker elevated. Let $\theta_0 = \mathbf{1}\{\text{John has pancreatic cancer}\}$, and $Y = \mathbf{1}\{\text{John's tumour marker level is elevated}\}$. The observation is $y = 1$. The likelihood for $\theta_0 = 0$ is $f(y = 1|\theta_0 = 0) = 0.005$ and the likelihood for $\theta_0 = 1$ is $f(y = 1|\theta_0 = 1) = 0.96$. We can reject the null hypothesis that John does not have pancreatic cancer at size 0.05. However, to make any probabilistic statement about whether John has pancreatic cancer, we need to know the prior probability of a random person having pancreatic cancer in addition to the sensitivity and specificity of the test itself. Suppose we know the prevalence of pancreatic cancer for people in John's age group is 0.00015, then

$$f_{\text{posterior}}(\theta = 1|y = 1) = \frac{0.96 \times 0.00015}{0.96 \times 0.00015 + 0.005 \times 0.99985} \approx 0.028.$$

Note the posterior probability takes into account two unlikely events: i) a random person has pancreatic cancer and ii) a non-cancerous patient has elevated tumour marker level.

Example 8. In the linear model, if the prior distribution for β_0 is $N(0, I_p)$ and the prior distribution for σ_0^2 is χ_1^2. Then the posterior distribution for (β_0, σ_0^2) is

$$f_{\text{posterior}}(\beta, \sigma_0^2|y) = \frac{(2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\|y - X\beta\|^2} (2\pi)^{-p/2} e^{-\frac{1}{2}\|\beta\|^2} (2\pi\sigma^2)^{-1/2} e^{-\sigma^2/2}}{\int_{\tilde{\beta},\tilde{\sigma}^2} (2\pi\tilde{\sigma}^2)^{-n/2} e^{-\frac{1}{2\tilde{\sigma}^2}\|y - X\tilde{\beta}\|^2} (2\pi)^{-p/2} e^{-\frac{1}{2}\|\tilde{\beta}\|^2} (2\pi\tilde{\sigma}^2)^{-1/2} e^{-\tilde{\sigma}^2/2}\, d\tilde{\beta} d\tilde{\sigma}^2}$$

Appendix: review of vector differentiation

We review some vector differentiation techniques that are useful in statistical calculations. Let $x \in \mathbb{R}^p, y \in \mathbb{R}^q, z \in \mathbb{R}^r$ be vectors, and $A \in \mathbb{R}^{s \times q}$ a constant matrix, where $p, q, r, s \in \mathbb{N}$. The following are some rules to perform differentiation with respect to a vector.

- (Jacobian) $\frac{\partial y}{\partial x} = \left(\frac{\partial y_i}{\partial x_j}\right)_{1 \le i \le q, 1 \le j \le p} \in \mathbb{R}^{q \times p}$.
- (Linearity) $\frac{\partial Ay}{\partial x} = A\frac{\partial y}{\partial x}$ and when $q = r$, $\frac{\partial (y+z)}{\partial x} = \frac{\partial y}{\partial x} + \frac{\partial z}{\partial x}$.
- (Product rule) When $q = r$, $\frac{\partial (y^\top z)}{\partial x} = y^\top \frac{\partial z}{\partial x} + z^\top \frac{\partial y}{\partial x}$.
- (Chain rule) When z is a function of y and y a function of z, $\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}$.

Example 9. We calculate $\frac{\partial}{\partial\beta}\|y - X\beta\|^2$.

$$\begin{aligned}\frac{\partial}{\partial\beta}\|y - X\beta\|^2 &= \frac{\partial}{\partial\beta}(y - X\beta)^\top(y - X\beta) = \frac{\partial(y - X\beta)^\top(y - X\beta)}{\partial(y - X\beta)}\frac{\partial(y - X\beta)}{\partial\beta} \\ &= \big((y - X\beta)^\top I + (y - X\beta)^\top I\big)(-X) = -2(y - X\beta)^\top X.\end{aligned}$$