IB Statistics

Ishan Nath, Lent 2023

Based on Lectures by Dr. Sergio Bacallado

March 9, 2023

Contents

1 Introduction

Statistics is the science of making informed decisions. It can include:

- Design of experiments,
- Graphical exploration of data,
- Formal statistical inference (part of Decision theory),
- Communication of results.

Let $X_1, X_2, \ldots, X_n$ be independent observations from a distribution $f(x \mid \theta)$, with parameter θ. We wish to make inferences about the value of θ from $X_1, X_2, \ldots, X_n$. Such inference can include:

- Estimating θ,
- Quantifying uncertainty in estimates,
- Testing a hypothesis about θ.

1.1 Probability Review

Let Ω be the *sample space* of outcomes in an experiment. A measurable subset of Ω is called an *event*. We denote the set of events as $\mathcal{F}$.

A function $\mathbb{P} : \mathcal{F} \to [0, 1]$ is called a *probability measure* if:

- $\mathbb{P}(\emptyset) = 0$,
- $\mathbb{P}(\Omega) = 1$,
- $\mathbb{P}(\bigcup_i A_i) = \sum_i \mathbb{P}(A_i)$, if (A_i) are disjoint and countable.

A *random variable* is a (measurable) function $X : \Omega \to \mathbb{R}$.

The *distribution function* of X is

$$F_X(x) = \mathbb{P}(X \leq x).$$

A *discrete random variable* takes values in a countable subset $E \subset \mathbb{R}$, and its *probability mass function* or pmf is $p_X(x) = \mathbb{P}(X = x)$.

We say X has *continuous* distribution if it has a *probability density function* or pdf, satisfying

$$\mathbb{P}(X \in A) = \int_A f_X(x) \, \mathrm{d}x,$$

for any measurable A. The *expectation* of X is defined

$$\mathbb{E}[X] = \begin{cases} \sum_{x \in X} x \cdot p_X(x) & X \text{ discrete,} \\ \int x \cdot f_X(x)\,\mathrm{d}x & X \text{ continuous.} \end{cases}$$

If $g : \mathbb{R} \to \mathbb{R}$, then

$$\mathbb{E}[g(x)] = \int g(x) f_X(x)\,\mathrm{d}x.$$

The *variance* of X is

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

We say that $X_1, X_2, \ldots, X_n$ are *independent* if for all $x_1, x_2, \ldots, x_n$,

$$\mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \mathbb{P}(X_1 \leq x_1) \cdots \mathbb{P}(X_n \leq x_n).$$

If the variables have probability density functions, then

$$f_X(x) = \prod_{i=1}^{n} f_{X_i}(x_i),$$

where X is the vector of variables $(X_1, \ldots, X_n)$ and x is the vector $(x_1, \ldots, x_n)$.

Importantly, if $a_1, \ldots, a_n \in \mathbb{R}$,

$$\mathbb{E}[a_1 X_1 + \cdots + a_n X_n] = a_1 \mathbb{E}[X_1] + \cdots + a_n \mathbb{E}[X_n].$$

Moreover,

$$\mathrm{Var}(a_1 X_1 + \cdots + a_n X_n) = \sum_{i,j} a_i a_j \,\mathrm{Cov}(X_i, X_j).$$

Here the *covariance* of X_i and X_j is

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])].$$

If $X = (X_1, \ldots, X_n)^T$ and $\mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n])$, then the linearity of expectation can be rewritten as

$$\mathbb{E}[a^T X] = a^T \mathbb{E}[X],$$

and moreover

$$\mathrm{Var}(a^T X) = a^T \,\mathrm{Var}(X) a,$$

where $\mathrm{Var}(X)$ is the *covariance matrix*: $(\mathrm{Var}(X))_{ij} = \mathrm{Cov}(X_i, X_j)$.

1.2 Moment Generating Functions

The *moment generating function* of a variable X is

$$M_X(t) = \mathbb{E}[e^{tx}].$$

This may only exist for t in some neighbourhood of 0. The important properties of MGFs is that

$$\mathbb{E}[X^n] = \frac{\mathrm{d}^n}{\mathrm{d}t^n} M_X(0),$$

and from this we obtain $M_X = M_Y \iff F_x = F_y$.

MGFs also make it easy to find the distribution function of sums of iid variables.

Example 1.1.

Let $X_1, \dots, X_n$ be iid Poisson(μ). Then

$$\begin{aligned} M_{X_1}(t) = \mathbb{E}[e^{tX_1}] &= \sum_{x=0}^{\infty} e^{tx} \cdot \frac{e^{-\mu}\mu^x}{x!} \\ &= e^{-\mu} \sum_{x=0}^{\infty} \frac{(e^t\mu)^x}{x!} = e^{-\mu} e^{\mu \exp(t)} = e^{-\mu(1-e^t)}. \end{aligned}$$

If $S_n = X_1 + \cdots + X_n$, then

$$\begin{aligned} M_{S_n}(t) &= \mathbb{E}[e^{t(X_1+\cdots+X_n)}] = \prod_{i=1}^{n} \mathbb{E}[e^{tX_i}] \\ &= e^{-\mu(1-e^t)n} \end{aligned}$$

This is the same as a Poisson(μn) MGF, so $S_n \sim$ Poisson($\mu \cdot n$).

1.3 Limit Theorems

We list some important limit theorems, starting with the *weak law of large numbers* (WLLN). This says if $X_1, \dots, X_n$ are iid with $\mathbb{E}[X_1] = \mu$, then let $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean. WLLN says that for all $\varepsilon > 0$,

$$\mathbb{P}(|\overline{X_n} - \mu| > \varepsilon) \to 0,$$

as $n \to \infty$.

The *strong law of large numbers* (SLLN) says a stronger result, namely

$$\mathbb{P}(\overline{X_n} \to \mu) = 1,$$

i.e. $\overline{X_n}$ converges to μ almost surely.

The *central limit theorem* is another important limit theorem. If we take

$$Z_n = \frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma},$$

where $\sigma^2 = \mathrm{Var}(X_i)$, then Z_n is "approximately" $N(0,1)$ as $n \to \infty$.

What this means is that $\mathbb{P}(Z_n \le z) \to \Phi(z)$ as $n \to \infty$ for all $z \in \mathbb{R}$, where Φ is the distribution function of a $N(0,1)$ variable.

1.4 Conditioning

Let X and Y be discrete random variables. Their *joint pmf* is

$$p_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y).$$

The *marginal pmf* is

$$p_X(x) = \mathbb{P}(X = x) = \sum_{y \in Y} p_{X,Y}(x,y).$$

The *conditional pmf* of X given $Y = y$ is

$$p_{X|Y}(x \mid y) = \mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{p_{X,Y}(x,y)}{p_Y(y)}.$$

This is defined to be 0 if $p_Y(y) = 0$.

For continuous random variables X, Y, the *joint pdf* $f_{X,Y}$ has

$$\mathbb{P}(X \le x', y \le y') = \int_{-\infty}^{x'} \int_{-\infty}^{y'} f_{X,Y}(x,y) \, \mathrm{d}y \, \mathrm{d}x.$$

The *marginal pdf* of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \, \mathrm{d}x.$$

The *conditional pdf* of X given Y is

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

The *conditional expectation* is given by

$$\mathbb{E}[X \mid Y] = \begin{cases} \sum_x x \cdot p_{X|Y}(x \mid y) & X, Y \text{ discrete,} \\ \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x \mid y)\, \mathrm{d}x & X, Y \text{ continuous.} \end{cases}$$

This is a random variable, which is a function of Y. The *tower property* says that

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X].$$

Hence we can write the variance of X as follows:

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= \mathbb{E}[\mathbb{E}[X^2 \mid Y]] - (\mathbb{E}[\mathbb{E}[X \mid Y]])^2 \\ &= \mathbb{E}[\mathbb{E}[X^2 \mid Y] - (\mathbb{E}[X \mid Y])^2] + \mathbb{E}[\mathbb{E}[X \mid Y]^2] - \mathbb{E}[\mathbb{E}[X \mid Y]]^2 \\ &= \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y]). \end{aligned}$$

1.5 Change of Variables

The *change of variables* formula is as follows:

Let $(x, y) \mapsto (u, v)$ be a differentiable bijection. Then,

$$\begin{aligned} f_{U,V}(u, v) &= f_{X,Y}(x(u, v), y(u, v)) \cdot |\det J|, \\ J &= \frac{\partial(x, y)}{\partial(u, v)} = \begin{pmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{pmatrix}. \end{aligned}$$

1.6 Important Distributions

$X \sim \mathrm{Negbin}(k, p)$ if X models the time in successive iid $\mathrm{Ber}(p)$ trials to achieve k successes. If $k = 1$, this is the same as a geometric distribution.

$X \sim \mathrm{Poisson}(\lambda)$ is the limit of $\mathrm{Bin}(n, \lambda/n)$ random variables, as $n \to \infty$.

If $X_i \sim \Gamma(\alpha_i, \lambda)$ for $i = 1, \ldots, n$ with $X_1, \ldots, X_n$ independent, then if $S_n = X_1 + \cdots + X_n$,

$$M_{S_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left(\frac{\lambda}{\lambda - 1}\right)^{\alpha_1 + \cdots + \alpha_n}$$

which is the mgf of a $\Gamma(\sum \alpha_i, \lambda)$ random variable. Hence $S_n \sim \Gamma(\sum \alpha_i, \lambda)$.

Also, if $X \sim \Gamma(a, \lambda)$, then for any $b \in (0, \infty)$, $bX \sim \Gamma(a, \lambda/b)$.

Special cases of the Gamma distribution include $\Gamma(1, \lambda) = \mathrm{Exp}(\lambda)$, and $\Gamma(\frac{k}{2}, \frac{1}{2}) = \chi^2_k$, the Chi-squared distribution with k degrees of freedom. This can be thought of as the sum of k independent squared $N(0, 1)$ random variables.

2 Estimation

Suppose we observe data $X_1, X_2, \ldots, X_n$, which are iid from some pdf (or pmf) $f_X(x \mid \theta)$, with θ unknown. We let $X = (X_1, \ldots, X_n)$.

Definition 2.1. An *estimator* is a statistic or a function of the data $T(X) = \hat{\theta}$, which we use to approximate the true parameter θ. The distribution of $T(X)$ is called the *sampling distribution*.

Example 2.1.

If $X_1, \ldots, X_n$ are iid $N(\mu, 1)$, we can define an estimator for the mean as

$$\hat{\mu} = T(X) = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The sampling distribution of $\hat{\mu}$ is $N(\mu, \frac{1}{n})$.

Definition 2.2. The *bias* of $\hat{\theta} = T(X)$ is

$$\operatorname{bias}(\hat{\theta}) = \mathbb{E}_\theta[\hat{\theta}] - \theta.$$

Remark. In general, the bias is a function of θ, even if the notation $\operatorname{bias}(\hat{\theta})$ does not make that explicit.

Definition 2.3. We say that $\hat{\theta}$ is *unbiased* if $\operatorname{bias}(\hat{\theta}) = 0$ for all $\theta \in \Theta$.

Example 2.2.

Out previous estimator

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

is unbiased because $\mathbb{E}_\mu[\hat{\mu}] = \mu$ for all $\mu \in \mathbb{R}$.

Definition 2.4. The *mean squared error* (mse) of $\hat{\theta}$ is

$$\operatorname{mse}(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2].$$

Like the bias, the mean squared error of $\hat{\theta}$ is a function of θ.

2.1 Bias-Variance Decomposition

We can write the mean squared error as

$$\begin{aligned}\operatorname{mse}(\hat{\theta}) &= \mathbb{E}_\theta[(\hat{\theta}-\theta)^2] = \mathbb{E}_\theta[(\hat{\theta}-\mathbb{E}_\theta[\hat{\theta}]+\mathbb{E}_\theta[\hat{\theta}]-\theta)^2] \\ &= \operatorname{Var}_\theta(\hat{\theta}) + \operatorname{bias}^2(\hat{\theta}) + 2\underbrace{[\mathbb{E}_\theta[(\hat{\theta}-\mathbb{E}_\theta[\hat{\theta}])]]}_{0}(\mathbb{E}_\theta[\hat{\theta}]-\theta).\end{aligned}$$

The two terms on the right hand side are non-negative, so there is a trade off between bias and variance.

Example 2.3.

Let $X \sim \operatorname{Bin}(n,\theta)$, where n is known, and we wish to estimate θ. The standard estimator is

$$T_u = \frac{X}{n}, \quad \mathbb{E}_\theta[T_u] = \frac{\mathbb{E}_\theta[X]}{n} = \theta.$$

Hence T_u is unbiased. We can also calculate the mean squared error as

$$\operatorname{mse}(T_u) = \operatorname{Var}_\theta(T_u) = \frac{\operatorname{Var}_\theta(X)}{n^2} = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n}.$$

Consider a second estimator

$$T_B = \frac{X+1}{n+2} = w\frac{X}{n} + (1-w)\frac{1}{2},$$

for $w = \frac{n}{n+2}$. In this case T_B is interpolating between our unbiased estimator, and the constant estimator. The bias of T_B is

$$\operatorname{bias}(T_B) = \mathbb{E}_\theta[T_B] - \theta = \mathbb{E}[\frac{X+1}{n+2}] - \theta = \frac{1}{n+2} - \frac{2}{n+2}\theta.$$

This is not equal to zero for all but one value of θ. Hence, T_B is biased. We can also calculate the variance

$$\begin{aligned}\operatorname{Var}_\theta(T_B) &= \frac{1}{(n+2)^2}n\theta(1-\theta) - w^2\frac{\theta(1-\theta)}{n}, \\ \operatorname{mse}(T_B) &= \operatorname{Var}_\theta(T_B) + \operatorname{bias}^2(T_B) \\ &= w^2\frac{\theta(1-\theta)}{n} + (1-w)^2\left(\frac{1}{2}-\theta\right)^2.\end{aligned}$$

Hence the mse of the biased estimator is a weighted average of the mse of the unbiased estimator, and a parabola. For θ around $1/2$, the biased estimator has a lower mse than the unbiased estimator.

The message here is that our prior judgements about θ affect our choice of estimator, and unbiasedness is not always desirable.

Example 2.4.

Suppose $X \sim \text{Poisson}(\lambda)$. We wish the estimate $\theta = \mathbb{P}(X = 0)^2 = e^{-2\lambda}$. For an estimator $T(X)$ to be unbiased, we must have for all λ,

$$\begin{aligned} \mathbb{E}_\lambda[\hat{\theta}] = \sum_{x=0}^{\infty} T(x) \frac{e^{-\lambda}\lambda^x}{x!} &= e^{-2\lambda} = \theta \\ \iff \sum_{x=0}^{\infty} T(x) \frac{\lambda^x}{x!} &= e^{-\lambda} = \sum_{x=0}^{\infty} (-1)^x \frac{\lambda^x}{x!}. \end{aligned}$$

For this to hold for all $\lambda \geq 0$, we should take $T(X) = (-1)^X$. But this estimator makes no sense.

2.2 Sufficiency

Suppose $X_1, \ldots, X_n$ are iid random variables from a distribution with pdf (or pmf) $f_X(\cdot \mid \theta)$. Let $X = (X_1, \ldots, X_n)$.

The question is: is there a statistic $T(X)$ which contains all the information in X needed to estimate θ?

Definition 2.5. A statistic T is *sufficient* for θ if the conditional distribution of X given $T(X)$ does not depend on θ.

Note θ and $T(X)$ may be vector-valued.

Example 2.5.

Let $X_1, \ldots, X_n$ be iid $\text{Ber}(\theta)$ for $\theta \in [0, 1]$. Then,

$$f_X(\cdot \mid \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{x_1+\cdots+x_n}(1-\theta)^{n-x_1-\cdots-x_n}.$$

This only depends on X through

$$T(X) = \sum_{i=1}^{n} x_i.$$

Indeed, for x with $x_1 + \cdots + x_n = t$,

$$f_{X|T=t}(x \mid T(x) = t) = \frac{\mathbb{P}_\theta(X = x, T(X) = t)}{\mathbb{P}_\theta(T(X) = t)} = \frac{\mathbb{P}_\theta(X = x)}{\mathbb{P}\theta(T(x) = t)}$$
$$= \frac{\theta^{x_1+\cdots+x_n}(1-\theta)^{n-x_1-\cdots-x_n}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \binom{n}{t}^{-1},$$

and otherwise this probability is 0. As this doesn't depend on θ, $T(X)$ is sufficient for θ.

Theorem 2.1 (Factorization criterion)**.** *T is sufficient for θ if and only if*

$$f_X(x \mid \theta) = g(T(x), \theta) \cdot h(x),$$

for suitable functions g, h.

Proof: We only do the discrete case.

Suppose that $f_X(x \mid \theta) = g(T(x), \theta)h(x)$. If $T(x) = t$, then

$$f_{X|T=t}(x \mid T = t) = \frac{\mathbb{P}_\theta(X = x, T(X) = t)}{\mathbb{P}_\theta(T(X) = t)}$$
$$= \frac{g(T(x), \theta)h(x)}{\sum_{T(x')=t} g(T(x'), \theta)h(x')}$$
$$= \frac{g(t, \theta)}{g(t, \theta)} \cdot \frac{h(x)}{\sum_{T(x')=t} h(x')}.$$

This doesn't depend on θ, so $T(X)$ is sufficient. Conversely, if $T(X)$ is sufficient, then

$$\mathbb{P}_\theta(X = x) = \mathbb{P}_\theta(X = x, T(X) = t)$$
$$= \underbrace{\mathbb{P}_\theta(T(X) = t)}_{g(t,\theta)} \cdot \underbrace{\mathbb{P}_\theta(X = x \mid T(X) = t)}_{h(x)}.$$

Therefore the pmf of X factorizes.

Example 2.6.

Return to our example from before, where $X_1, \ldots, X_n$ are iid Ber(θ). Then

$$f_X(x \mid \theta) = \theta^{x_1+\cdots+x_n}(1-\theta)^{n-x_1-\cdots-x_n}.$$

Hence if we take $g(t,\theta) = \theta^t(1-\theta)^{n-t}$, and $h(x) = 1$, we immediately get that $T(X) = \sum x_i$ is sufficient.

Example 2.7.

Let $X_1, \ldots, X_n$ be iid $U([0,\theta])$, for $\theta > 0$. Then,

$$\begin{aligned} f_X(x \mid \theta) &= \prod_{i=1}^n \frac{1}{\theta} \mathbb{1}(X_i \in [0,\theta]) \\ &= \underbrace{\frac{1}{\theta^n} \mathbb{1}(\max_i x_i \leq \theta)}_{g(T(x),\theta)} \underbrace{\mathbb{1}(\min_i x_i \geq 0)}_{h(x)} . \end{aligned}$$

Hence $T(x) = \max_i x_i$ is a sufficient statistic for θ.

2.3 Minimal Sufficiency

Sufficient statistics are not unique. Indeed, any one-to-one function of a sufficient statistic is also sufficient. Also $T(X) = X$ is always sufficient, but not very useful.

Definition 2.6. A sufficient statistic T is *minimal sufficient* if it is a function of any other sufficient statistic, so if T' is also sufficient, then

$$T'(x) = T'(y) \implies T(x) = T(y),$$

for all x, y in our space.

By this definition, any two minimal sufficient statistics T, T' are in bijection with each other, so

$$T(x) = T(y) \iff T'(x) = T'(y).$$

Theorem 2.2. *Suppose that $T(X)$ is a statistic such that*

$$\frac{f_X(x \mid \theta)}{f_X(y \mid \theta)}$$

is constant as a function of θ, if and only if $T(x) = T(y)$. Then T is minimal sufficient.

Let $x \overset{1}{\sim} y$ if

$$\frac{f_X(x \mid \theta)}{f_X(y \mid \theta)}$$

is constant in θ. It is easy to check that $\overset{1}{\sim}$ is an equivalence relation.

Similarly, for a given statistic T, $x \overset{2}{\sim} y$ if $T(x) = T(y)$ defines another equivalence relation.

The condition of the theorem says that $\overset{1}{\sim}$ and $\overset{2}{\sim}$ are the same for minimal sufficient statistics.

Remark. We can always construct a statistic T which is constant on the equivalence classes of $\overset{1}{\sim}$, which by the theorem is minimal sufficient.

Proof: For any value of T, let z_t be a representative from the equivalence class

$$\{x \mid T(x) = t\}.$$

Then,

$$f_X(x \mid \theta) = f_X(z_{T(x)} \mid \theta)\frac{f_X(x, \theta)}{f_X(z_{T(x)} \mid \theta)}.$$

This is exactly in the form $g(T(x), t)h(x)$, so by the factorization criterion T is sufficient.

To prove that T is minimal, take any other sufficient statistic S. We want to show that if $S(x) = S(y)$, then $T(x) = T(y)$.

By the factorization criterion, there are functions g_s, h_s such that

$$f_X(x, \theta) = g_s(S(x), \theta)h_s(x).$$

Suppose $S(x) = S(y)$. Then the ratio

$$\frac{f_X(x \mid \theta)}{f_X(y \mid \theta)} = \frac{g_s(S(x), \theta)h_s(x)}{g_s(S(y), \theta)h_s(y)} = \frac{h_s(x)}{h_s(y)},$$

is independent of θ. Hence $x \overset{1}{\sim} y$. By the hypothesis, we get that $T(x) = T(y)$.

Remark. Sometimes the range of X depends on θ. In this case we can interpret

$$\frac{f_X(x \mid \theta)}{f_Y(y \mid \theta)} \text{ constant in } \theta,$$

to mean that

$$f_X(x \mid \theta) = c(x, y)f_X(y \mid \theta),$$

for some function c which does not depend on θ.

Example 2.8.

Suppose that $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, with parameters (μ, σ^2) unknown. Then,

$$\begin{aligned}\frac{f_X(x \mid t)}{f_X(y \mid t)} &= \frac{(2\pi\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2)}{(2\pi\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2)} \\ &= \exp\left[-\frac{1}{2\sigma^2}\left(\sum x_i^2 - \sum y_i^2\right) + \frac{\mu}{\sigma^2}\left(\sum x_i - \sum y_i\right)\right].\end{aligned}$$

Hence if $\sum x_i^2 = \sum y_i^2$ and $\sum x_i = \sum y_i$, this ratio does not depend on (μ, σ^2). The converse is also true: if the ratio does not depend on (μ, σ^2), then we must have $\sum x_i^2 = \sum y_i^2$ and $\sum x_i = \sum y_i$. By the theorem, $T(x) = (\sum x_i^2, \sum x_i)$ is minimal sufficient.

Recall that bijections of T are also minimal sufficient. A more common way of expressing a minimal sufficient statistic in this model is $S(X) = (\bar{X}, S_{xx})$, where

$$\bar{X} = \frac{1}{n}\sum_i X_i, \qquad S_{xx} = \sum_i (X_i - \bar{X})^2.$$

In this example, (μ, σ^2) and $T(X)$ are both 2-dimensional. In general, the parameter and sufficient statistic can have different dimensions.

For example, if $X_1, \ldots, X_n$ are iid $N(\mu, \mu^2)$, where $\mu \geq 0$, then the minimal sufficient statistic is $S(X) = (\bar{X}, S_{xx})$.

2.4 Rao-Blackwell Theorem

So far we have written $\mathbb{E}_\theta$ and $\mathbb{P}_\theta$ to denote the expectations and probabilities in the model where $X_1, \ldots, X_n$ are iid drawn from $f_X(\cdot \mid \theta)$. From now on, we drop the subscript θ.

Theorem 2.3 (Rao-Blackwell Theorem)**.** *Let T be a sufficient statistic for θ. Let $\tilde{\theta}$ be some estimator for θ, with $\mathbb{E}[\tilde{\theta}^2] < \infty$ for all θ. Define a new estimator $\hat{\theta} = \mathbb{E}[\tilde{\theta} \mid T(X)]$. Then, for all θ,*

$$\mathbb{E}[(\hat{\theta} - \theta)^2] \leq \mathbb{E}[(\tilde{\theta} - \theta)^2],$$

with equality if and only if $\tilde{\theta}$ is a function of $T(X)$.

Remark. $\hat{\theta}$ is a valid estimator, as it does not depend on θ, only on X, as T is sufficient:

$$\hat{\theta}(T(x)) = \int \tilde{\theta}(x) f_{X|T}(x|T)\,\mathrm{d}x,$$

where neither $\tilde{\theta}$ nor the conditional distribution depend on θ.

The message is that we can improve the mean squared error of any estimator $\tilde{\theta}$ by taking a conditional expectation given $T(X)$.

Proof: By the tower property,

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[\mathbb{E}[\tilde{\theta} \mid T]] = \mathbb{E}[\tilde{\theta}].$$

So $\text{bias}(\hat{\theta}) = \text{bias}(\tilde{\theta})$ for all θ. By the conditional variance formula,

$$\begin{aligned}\text{Var}(\tilde{\theta}) &= \mathbb{E}[\text{Var}(\tilde{\theta} \mid T)] + \text{Var}(\mathbb{E}[\tilde{\theta} \mid T]) \\ &= \mathbb{E}[\text{Var}(\tilde{\theta} \mid T)] + \text{Var}(\hat{\theta}).\end{aligned}$$

Hence $\text{Var}(\tilde{\theta}) \geq \text{Var}(\hat{\theta})$ for all θ. Hence $\text{mse}(\tilde{\theta}) \geq \text{mse}(\hat{\theta})$.

Note that $\text{Var}(\tilde{\theta} \mid T) > 0$ with some positive probability unless $\tilde{\theta}$ is a function of $T(X)$. So $\text{mse}(\tilde{\theta}) > \text{mse}(\hat{\theta})$ unless $\tilde{\theta}$ is a function of $T(X)$.

Example 2.9.

Say $X_1, \ldots, X_n$ are iid Poisson(λ). We wish to estimate $\theta = \mathbb{P}(X_1 = 0) = e^{-\lambda}$. Then

$$\begin{aligned}f_X(x \mid \lambda) &= \frac{e^{-n\lambda}\lambda^{x_1+\cdots+x_n}}{x_1!\cdots x_n!} \\ &= \frac{\theta^n(-\log\theta)^{x_1+\cdots+x_n}}{x_1!\cdots x_n!}\end{aligned}$$

Letting $h(x) = 1/(x_1!\cdots x_n!)$, $g(T(x), \theta) = \theta^n(-\log\theta)^{T(x)}$, by the factorization criterion, $T(x) = \sum x_i$ is a sufficient statistic. Let $\tilde{\theta} = \mathbb{1}(X_i = 0)$. This is unbiased, but only uses one observation X_1. Using Rao-Blackwell, we can find

$$\begin{aligned}\hat{\theta} &= \mathbb{E}[\tilde{\theta} \mid T = t] = \mathbb{P}\left(X_1 = 0 \,\middle|\, \sum_{i=1}^n X_i = t\right) \\ &= \frac{\mathbb{P}(X_1 = 0, X_1 + \cdots + X_n = t)}{\mathbb{P}(X_1 + \cdots + X_n = t)} = \frac{\mathbb{P}(X_1 = 0, X_2 + \cdots + X_n = t)}{\mathbb{P}(X_1 + \cdots + X_n = t)} \\ &= \frac{\mathbb{P}(X_1 = 0)\mathbb{P}(X_2 + \cdots + X_n = t)}{\mathbb{P}(X_1 + \cdots + X_n = t)} = \frac{e^{-\lambda}\mathbb{P}(\text{Poisson}((n-1)\lambda) = t)}{\mathbb{P}(\text{Poisson}(n\lambda) = t)} \\ &= \frac{e^{-n\lambda}((n-1)\lambda)^t/t!}{e^{-n\lambda}(n\lambda)^t/t!} = \left(1 - \frac{1}{n}\right)^t.\end{aligned}$$

So $\hat{\theta} = (1 - \frac{1}{n})^{x_1 + \cdots + x_n}$ is an estimator which by the Rao-Blackwell theorem has $\text{mse}(\hat{\theta}) < \text{mse}(\tilde{\theta})$.

As $n \to \infty$,

$$\hat{\theta} = \left(1 - \frac{1}{n}\right)^{n\bar{x}} \overset{n\to\infty}{\to} e^{-\bar{x}},$$

and by the strong law of large numbers

$$\bar{x} \to \mathbb{E}[X_1] = \lambda.$$

so $\hat{\theta} \to e^{-\lambda}$.

Example 2.10.

Let $X_1, \ldots, X_n$ be iid $U([0, \theta])$ where θ is unknown and $\theta \geq 0$. Then recall $T(X) = \max_i X_i$ is sufficient for θ.

Let $\tilde{\theta} = 2X_1$, which is unbiased. Then,

$$\begin{aligned}
\hat{\theta} &= \mathbb{E}[\tilde{\theta} \mid T = t] = 2\mathbb{E}[X_1 \mid \max_i X_i = t] \\
&= 2\mathbb{E}[X_1 \mid \max_i X_i = t, \max_i X_i = X_1]\mathbb{P}(\max_i X_i = X_1 \mid \max_i X_i = t) \\
&+ 2\mathbb{E}[X_1 \mid \max_i X_i = t, \max_i X_i \neq X_1]\mathbb{P}(\max_i X_i \neq X_1 \mid \max_i X_i = t) \\
&= \frac{2t}{n} + \frac{2(n-1)}{n}\mathbb{E}[X_1 \mid X_1 \leq t, \max_{i>1} X_i = t] = \frac{2t}{n} + \frac{2(n-1)}{n}\frac{t}{2} = \frac{n+1}{n}t.
\end{aligned}$$

So $\hat{\theta} = \frac{n+1}{n} \max_i X_i$ is a valid estimator with $\text{mse}(\hat{\theta}) < \text{mse}(\tilde{\theta})$.

2.5 Maximum Likelihood Estimation

Let $X = (X_1, \ldots, X_n)$ have joint pdf (or pmf) $f_X(X \mid \theta)$.

Definition 2.7. The likelihood function is

$$L : \theta \mapsto f_X(X \mid \theta).$$

The *maximum likelihood estimator* is any value of θ maximizing $L(\theta)$.

If $X_1, \ldots, X_n$ are iid each with pdf (or pmf) $f_X(\cdot \mid \theta)$, then

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i \mid \theta).$$

We will denote the logarithm

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f_X(x_i \mid \theta).$$

Example 2.11.

If $X_1, \dots, X_n$ are iid Ber(θ), then

$$\ell(\theta) = \left(\sum x_i\right) \log \theta = \left(n - \sum_{x_i}\right) \log(1 - \theta),$$

and the derivative

$$\frac{\partial \ell}{\partial \theta} = \frac{\sum x_i}{\theta} - \frac{n - \sum x_i}{1 - \theta}.$$

This is zero if and only if $\theta = \frac{1}{n} \sum x_i = \bar{X}$.

Hence $\bar{X}$ is the maximum likelihood estimator for θ, and is unbiased as $\mathbb{E}[\bar{X}] = 0$.

Example 2.12.

If $X_1, \dots, X_n$ are iid $N(\mu, \sigma^2)$, then

$$\log(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

This is maximized when $\partial \ell / \partial \mu = \partial \ell / \partial \sigma^2 = 0$. First, we get

$$\frac{\partial \ell}{\partial \mu} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu),$$

which is equal to zero when $\mu = \bar{X}$. Then

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2.$$

This is zero when

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n} S_{xx}.$$

Hence $(\hat{\mu}, \hat{\sigma}^2) = (\bar{X}, S_{xx}/n)$ give the maximum likelihood estimator in this model.

Note that $\hat{\mu} = \bar{X}$ is unbiased. Now we want to see if $\hat{\sigma}^2$ is biased. We could compute it directly, but later in the course we will show that

$$\frac{S_{xx}}{\sigma^2} = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1},$$

hence

$$\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}[\chi^2_{n-1}]\frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2 \neq \sigma^2,$$

which is biased, but asymptotically unbiased.

Example 2.13.

Let $X_1, \ldots, X_n$ be iid $U([0,\theta])$. Then

$$L(\theta) = \frac{1}{\theta^n}\mathbb{1}(\max_i X_i \leq \theta).$$

We can see from the plot that $\hat{\theta}_{\text{mle}} = \max_i X_i$ is the maximum likelihood estimator for θ. We also started from an unbiased estimator, and using Rao-Blackwellization we found an estimator

$$\hat{\theta} = \frac{n+1}{n}\max_i X_i.$$

This is also unbiased. So in this model the mle is biased as

$$\mathbb{E}[[\hat{\theta}_{\text{mle}}] = \mathbb{E}\left[\frac{n}{n+1}\hat{\theta}\right] = \frac{n}{n+1}\theta,$$

however it is asymptotically unbiased.

The maximum likelihood estimator has the following properties:

1. If T is a sufficient statistic, then the maximum likelihood estimator is a function of $T(X)$. By the factorization criterion,

$$L(\theta) = g(T(X),\theta)h(X).$$

   If $T(x) = T(y)$, then the likelihood function with data x and y is the same up to a multiplicative constant. Hence the maximum likelihood estimator in each case is the same.

2. If $\phi = h(\theta)$ where h is a bijection, then the maximum likelihood estimator of

ϕ is

$$\hat{\phi} = h(\hat{\theta}),$$

where $\hat{\theta}$ is the maximum likelihood estimator of θ.

3. Asymptotically, we have normality. This says $\sqrt{n}(\hat{\theta} - \theta)$ is approximately normal with mean 0 when n is large. Under some regularity conditions, for a measurable set A,

$$\mathbb{P}(\sqrt{n}(\hat{\theta} - \theta) \in A) \overset{n\to\infty}{\to} \mathbb{P}(Z \in A),$$

where $Z \sim N(0, \Sigma)$. This holds for all regular values of θ.

Here Σ is some function of ℓ, and there is a theorem (Cramer-Rao) which says this is the smallest variance attainable.

4. Sometimes, if the maximum likelihood estimator is not available analytically, we can find it numerically.

2.6 Confidence Intervals

Definition 2.8. A $(100 \cdot \gamma)\%$ *confidence interval* for a parameter θ is a random interval $(A(X), B(X))$ such that

$$\mathbb{P}(A(X) \leq \theta \leq B(X)) = \gamma,$$

for all values of θ.

The frequentist interpretation of the confidence interval is:

There exists some fixed true parameter θ. We repeat the experiment many times. On average, $100 \cdot \gamma\%$ of the time the interval $(A(X), B(X))$ contains θ.

The incorrect interpretation is:

Having observed $X = x$, there is a probability γ that θ is in $(A(x), B(x))$.

Example 2.14.

Let $X_1, \ldots, X_n$ be iid $N(\theta, 1)$. To find a 95% confidence interval for θ, we know that

$$\bar{X} = \frac{1}{n}\sum_{x_i} \sim N\left(\theta, \frac{1}{n}\right).$$

Hence

$$Z = \sqrt{n}(\bar{X} - \theta) \sim N(0, 1).$$

Z has this distribution for all θ. Then let z_1, z_2 be any two numbers such that $\Phi(z_2) - \Phi(z_1) = 0.95$. Then,

$$\mathbb{P}(z_1 \leq \sqrt{n}(\bar{X} - \theta) \leq z_2) = 0.95.$$

Rearranging,

$$\mathbb{P}\left(\bar{X} - \frac{z_2}{\sqrt{n}} \leq \theta \leq \bar{X} - \frac{z_1}{\sqrt{n}}\right) = 0.95.$$

Therefore $(\bar{X} - \frac{z_2}{\sqrt{n}}, \bar{X} - \frac{z_1}{\sqrt{n}})$ is a 95% confidence interval.

There are multiple ways to choose z_1, z_2. Usually we minimise the width of the interval, which is achieved by $z_1 = \Phi^{-1}(0.025), z_2 = \Phi^{-1}(0.975)$.

To find a confidence interval, we can do the following:

1. Find some quantity $R(X, \theta)$ such that the $\mathbb{P}_\theta$ distribution of $R(X, \theta)$ does not depend on θ. This is called a *pivot*.

   For example, we chose $Z = \sqrt{n}(\bar{X} - \mu) \sin N(0, 1)$.

2. Write down a probabilistic statement about the pivot of the form

   $$\mathbb{P}(c_1 \leq R(x, \theta) \leq c_2) = \gamma,$$

   by using quantiles c_1, c_2 of the distribution of $R(X, \theta)$ (typically $N(0, 1)$ or χ^2_p).

3. Rearrange the inequalities to leave θ in the middle.

Proposition 2.1. *If T is a monotone increasing function $T : \mathbb{R} \to \mathbb{R}$, and $(A(X), B(X))$ is a $100 \cdot \gamma$% confidence interval for θ, then $(T(A(X)), T(B(X)))$ is a confidence interval for $T(\theta)$.*

Remark. When θ is a vector, we talk about confidence sets.

Example 2.15.

Let $X_1, \ldots, X_n$ be iid $N(0, \sigma^2)$. We want to find a 95% confidence interval for σ^2.

Note that $\frac{X_i}{\sigma} \sim N(0, 1)$, so using all the data points,

$$R(X, \sigma^2) = \sum_{i=1}^{n} \frac{X_i^2}{\sigma^2} \sim \chi^2_n$$

is a pivot. Let $c_1 = F^{-1}_{\chi^2_n}(0.025)$, $c_2 = F^{-1}_{\chi^2_n}(0.975)$. Then,

$$\mathbb{P}(c_1 \leq R(X, \sigma^2) \leq c_2) = 0.95.$$

Rearranging,

$$\mathbb{P}\left(\frac{\sum x_i^2}{c_2} \leq \sigma^2 \leq \frac{\sum x_i^2}{c_1}\right) = 0.95.$$

Hence $[\sum x_i^2/c_2, \sum x_i^2/c_1]$ is a 95% confidence interval σ^2.

Applying the proposition, $[\sqrt{\sum x_i^2/c_2}, \sqrt{\sum x_i^2/c_1}]$ is a 95% confidence interval for σ.

Example 2.16.

Let $X_1, \ldots, X_n$ be $\mathrm{Ber}(p)$, for n. To find an approximate 95% for confidence interval p.

Recall that the maximum likelihood estimator for p is

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

By the central limit theorem, when n is large, $\hat{p}$ is approximately $N(p, \frac{p(1-p)}{n})$. Hence,

$$\sqrt{n}\frac{(\hat{p} - p)}{\sqrt{p(1-p)}} \sim N(0, 1),$$

approximately. If $z = \Phi^{-1}(0.975)$, then

$$\mathbb{P}\left(-z \leq \frac{\sqrt{n}(\hat{p} - p)}{\sqrt{p(1-p)}} \leq z\right) \approx 0.95.$$

Rearranging this is tricky. Instead, we argue that $n \to \infty$, $\hat{p}(1-\hat{p}) \to p(1-p)$. So replacing this in the denominator,

$$\mathbb{P}\left(-z \leq \frac{\sqrt{n}(\hat{p} - p)}{\sqrt{\hat{p}(1-\hat{p})}} \leq z\right) \approx 0.95.$$

Rearranging this, we get

$$\mathbb{P}\left(\hat{p} - z\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \leq p \leq \hat{p} + z\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right) \approx 0.95.$$

Hence this is an approximate 95% confidence interval for p.

Note that $z \approx 1.96$ and $\sqrt{\hat{p}(1-\hat{p})} \le \frac{1}{2}$ for all $\hat{p} \in (0,1)$. So a conservative confidence interval is $[\hat{p} \pm 1.96 \cdot \frac{1}{2} \cdot \frac{1}{\sqrt{n}}]$.

2.7 Interpreting Confidence Intervals

Suppose X_1, X_2 are iid $U[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$. We find a sensible 50% confidence interval for θ. Consider

$$\begin{aligned}
\mathbb{P}(\theta \text{ between } X_1, X_2) &= \mathbb{P}(\min(X_1, X_2) \le \theta \le \max(X_1, X_2)) \\
&= \mathbb{P}(X_1 \le \theta \le X_2) + \mathbb{P}(X_2 \le \theta \le X_1) \\
&= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.
\end{aligned}$$

Hence we can immediately conclude that $(\min(X_1, X_2), \max(X_1, X_2))$ is a 50% confidence interval for θ.

This does not mean for specific $X_1 = x_1$, $X_2 = x_2$ the value θ lies in the interval $(\min(x_1, x_2), \max(x_1, x_2))$ with probability $\frac{1}{2}$: consider when $|x_1 - x_2| > \frac{1}{2}$.

In this case, we can be sure that θ is in $(\min(x_1, x_2), \max(x_1, x_2))$, as the value θ can be at most distance $\frac{1}{2}$ from x_1 and x_2.

However the frequentist interpretation makes sense: if we repeat the experiment many times, we see $\theta \in (\min(X_1, X_2), \max(X_1, X_2))$ exactly 50% of the time. We cannot say, given a specific observation that we are 50% certain that θ is in the confidence interval.

3 Bayesian Inference

So far, we have assumed there is some true parameter θ. That data X has pdf (or pmf) $f_X(\cdot \mid \theta)$.

Bayesian analysis is a different framework, where we treat θ as a random variable, taking values in Θ.

We begin by assigning to θ a *prior distribution* $\pi(\theta)$, which represents the opinions or information about θ before seeing on any data.

Conditional on θ, the data X has pdf (or pmf) $f_X(x \mid \theta)$. Having observed a specific value of $X = x$, this information is combined with the prior to form the *posterior distribution* $\pi(\theta \mid x)$, which is the conditional distribution of θ given $X = x$. By Bayes' rule,

$$\pi(\theta \mid x) = \frac{\pi(\theta) \cdot f_X(x \mid \theta)}{f_X(x)},$$

where $f_X(x)$ is the marginal probability of X, and

$$f_X(x) = \begin{cases} \int_\Theta f_X(x \mid \theta)\pi(\theta)\,\mathrm{d}\theta & \theta \text{ continuous,} \\ \sum_\Theta f_X(x \mid \theta)\pi(\theta) & \theta \text{ discrete.} \end{cases}$$

Example 3.1.

Consider a patient getting a COVID test. Then the possible values are $\theta \in \{0, 1\}$, corresponding to the patient not having COVID, and the patient having COVID, respectively.

We also have data $X \in \{0, 1\}$, corresponding to the patient getting a negative test, or positive test, respectively.

We also know the sensitivity of the test as $f_X(X = 1 \mid \theta = 1)$, and the specificity $f_X(X = 0 \mid \theta = 0)$.

To run Bayesian analysis, we can take a prior $\pi(\theta = 1) = p$, if we know the proportion p of people infected. Then the chance of infection given a true test is

$$\pi(\theta = 1 \mid X = 1) = \frac{\pi(\theta = 1) f_X(X = 1 \mid \theta = 1)}{\pi(\theta = 0) f_X(X = 1 \mid \theta = 0) + \pi(\theta = 1) f_X(X = 1 \mid \theta = 1)}.$$

If $\pi(\theta = 0) \gg \pi(\theta = 1)$, then this posterior can still be very small.

Example 3.2.

Let $\theta \in [0, 1]$ be the mortality rate for a new surgery. In the first 10 operations, there were not deaths.

If we have a model $X_i \sim \text{Ber}(\theta)$, where $X_i = 1$ if the i'th operation is fatal, and 0 otherwise, then

$$f_X(x \mid \theta) = \theta^{x_1+\cdots+x_{10}}(1-\theta)^{10-x_1+\cdots+x_{10}}.$$

For our prior, we are told that the surgery is performed in other hospital with a mortality rate ranging from 0.03 to 0.2, with an average of 0.1. We can take $\pi(\theta) \sim \text{Beta}(a, b)$, with $a = 3$ and $b = 27$, so that the mean of $\pi(\theta)$ is 0.1 and $\pi(0.03 < \theta < 0.2) = 0.9$.

The posterior distribution is then

$$\begin{aligned}\pi(\theta \mid x) &\propto \pi(\theta) f_X(x \mid \theta) \\ &\propto \theta^{a-1}(1-\theta)^{b-1}\theta^{x_1+\cdots+x_{10}}(1-\theta)^{10-x_1-\cdots-x_{10}} \\ &= \theta^{x_1+\cdots+x_{10}+a-1}(1-\theta)^{b+10-x_1-\cdots-x_{10}-1}.\end{aligned}$$

We can deduce that the posterior distribution is a $\text{Beta}(\sum x_i + a, 10 - \sum x_i + b)$ distribution. In our case, since there are no deaths, the posterior distribution is $\text{Beta}(3, 37)$.

Note that the prior and posterior distributions are in the same family of distributions. This is known as *conjugacy*.

With the information gained from the posterior, we can make decisions under uncertainty. The formal process is:

1. We must pick a decision $\delta \in D$.
2. The loss function $L(\theta, \delta)$ is the loss incurred when we make decision δ and the true parameter has value θ.
3. We pick the decision which minimizes the posterior expected loss:

$$\delta^* = \operatorname*{argmin}_{\delta \in D} \int_\Theta L(\theta, \delta)\pi(\theta \mid x)\, \mathrm{d}\theta.$$

For point estimation, the decision is a “best guess” for the true parameter, so $\delta \in \Theta$.

The *Bayes estimator* $\hat{\theta}^{(k)}$ minimizes

$$h(\delta) = \int_\Theta L(\theta, \delta)\pi(\theta \mid x)\,\mathrm{d}\theta.$$

Example 3.3.

Consider quadratic loss $L(\theta, \delta) = (\theta - \delta)^2$. Then

$$h(\delta) = \int_\Theta (\theta - d)^2 \pi(\theta \mid x)\,\mathrm{d}\theta.$$

Now $h'(\delta) = 0$ if

$$\int_\Theta (\theta - d)\pi(\theta \mid x)\,\mathrm{d}\theta = 0.$$

Hence

$$\int_\Theta \pi(\theta \mid x)\,\mathrm{d}\theta = \delta \int_\Theta \pi(\theta \mid x)\,\mathrm{d}\theta = \delta.$$

So the Bayes estimator is the posterior mean of θ.

Example 3.4.

Consider an absolute error loss $L(\theta, \delta) = |\theta - \delta|$. Then,

$$\begin{aligned} h(\delta) &= \int_\Theta |\theta - \delta| \pi(\theta \mid x)\,\mathrm{d}\theta \\ &= \int_{-\infty}^{\delta} -(\theta - \delta)\pi(\theta \mid x)\,\mathrm{d}\theta + \int_\delta^\infty (\theta - \delta)\pi(\theta \mid x)\,\mathrm{d}\theta \\ &= -\int_{-\infty}^{\delta} \theta\pi(\theta \mid x)\,\mathrm{d}\theta + \int_\delta^\infty \theta\pi(\theta \mid x)\,\mathrm{d}\theta \\ &\quad + \delta \int_{-\infty}^{\delta} \pi(\theta \mid x)\,\mathrm{d}\theta - \delta \int_\delta^\infty \pi(\theta \mid x)\,\mathrm{d}\theta. \end{aligned}$$

Taking the derivative with respect to δ, by the fundamental theorem of calculus,

$$h'(\delta) = \int_{-\infty}^{\delta} \pi(\theta \mid x)\,\mathrm{d}\theta - \int_\delta^\infty \pi(\theta \mid x)\,\mathrm{d}\theta.$$

Hence $h'(\delta) = 0$ if and only if

$$\int_{-\infty}^{\delta} \pi(\theta \mid x)\,\mathrm{d}\theta = \int_d^\infty \pi(\theta \mid x)\,\mathrm{d}\theta.$$

In this case, the Bayes estimator is the median of the posterior.

3.1 Credible Interval

A $100 \cdot \gamma\%$ *credible interval* $(A(x), B(x))$ is one which satisfies

$$\pi(A(x) \leq \theta \leq B(x) \mid x) = \gamma.$$

Hence,

$$\int_{A(x)}^{B(x)} \pi(\theta \mid x) \, \mathrm{d}\theta = \gamma.$$

Note that we can interpret credible intervals conditionally.

If T is a sufficient statistic, then $\pi(\theta \mid x)$ only depends on x through $T(x)$, as

$$\begin{aligned} \pi(\theta \mid x) &\propto \pi(\theta) f_X(x \mid \theta) \\ &= \pi(\theta) g(T(x), \theta) h(x) \\ &\propto \pi(\theta) g(T(x), \theta). \end{aligned}$$

Example 3.5.

Let $X_1, \ldots, X_n$ be iid $N(\mu, 1)$. We assign a prior for μ as $\pi(\mu) \sim N(0, 1/\tau^2)$. Then

$$\begin{aligned} \pi(\mu \mid x) &\propto f_X(x \mid \mu) \cdot \pi(\mu) \\ &\propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \exp\left[\frac{-\mu^2\tau^2}{2}\right] \\ &\propto \exp\left[-\frac{1}{2}(n + \tau^2)\left(\mu - \frac{x_1 + \cdots + x_n}{n + \tau^2}\right)^2\right]. \end{aligned}$$

We recognise that this is a normal distribution, namely

$$N\left(\frac{x_1 + \cdots + x_n}{n + \tau^2}, \frac{1}{n + \tau^2}\right).$$

The Bayes estimator is $\hat{\mu}^{(b)} = \frac{x_1 + \cdots + x_n}{n + \tau^2}$ for both the quadratic loss and absolute error loss. Contrast this to the maximum likelihood estimator $\frac{x_1 + \cdots + x_n}{n}$. A 95% credible interval is

$$\left(\hat{\mu}^{(b)} - \frac{1.96}{\sqrt{n + \tau^2}}, \hat{\mu}^{(b)} + \frac{1.96}{\sqrt{n + \tau^2}}\right).$$

This is close to a 95% confidence interval when $n \gg \tau^2$.

Example 3.6.

Take $X_1, \ldots, X_n$ iid Poisson(λ). Then we take a prior for λ as $\pi(\lambda) \sim \text{Exp}(1)$. Hence

$$\begin{aligned}
\pi(\lambda \mid x) &\propto f_X(x \mid \lambda) \cdot \pi(\lambda) \\
&\propto e^{-n\lambda} \lambda^{x_1 + \cdots + x_n} \cdot e^{-\lambda} \\
&= e^{-(n+1)\lambda} \lambda^{x_1 + \cdots + x_n}.
\end{aligned}$$

This is a $\Gamma(x_1 + \cdots + x_n + 1, n+1)$ distribution. The Bayes estimator under quadratic loss is the posterior mean,

$$\hat{\lambda}^{(b)} = \frac{x_1 + \cdots + x_n + 1}{n+1} \overset{n \to \infty}{\to} \frac{x_1 + \cdots + x_n}{n} = \hat{\lambda}^{(mle)}.$$

Under the absolute error loss, the Bayes estimator $\tilde{\lambda}^{(b)}$ has property

$$\int_0^{\tilde{\lambda}^{(b)}} \frac{(n+1)^{x_1 + \cdots + x_n - 1}}{(x_1 + \cdots + x_n)!} \lambda^{x_1 + \cdots + x_n} e^{-(n+1)\lambda} \, \mathrm{d}\lambda = \frac{1}{2}.$$

This has no closed form solution.

4 Simple Hypotheses

A *hypothesis* is some assumption about the distribution of the data X. Scientific questions are phrased as a choice between a *null hypothesis* H_0 (also known as a base case, simple model, or no effect) and an *alternative hypothesis* H_1 (also known as a complex model, interesting case, positive or negative effect).

Example 4.1.

1. Let $X_1, \ldots, X_n$ be iid $\mathrm{Ber}(\theta)$. Consider two hypothesis, $H_0 : \theta = \frac{1}{2}$ (i.e. we have a fair coin), and $H_1 : \theta = \frac{3}{4}$.
2. We also may consider $H_0 : \theta = \frac{1}{2}$, $H_1 : \theta \neq \frac{1}{2}$.
3. Let $X_1, \ldots, X_n$ take values in $\mathbb{N}_0$. Then we can consider $H_0 : X_i \sim \mathrm{Poisson}(\lambda)$ for some $\lambda > 0$, and $H_1 : X_1 \sim f_1$ for some other mass function f_1.
4. Finally, if X has probability distribution function $f(\cdot \mid \theta)$, where $\theta \in \Theta$, then we can consider $H_0 : \theta \in \Theta_0 \subset \Theta$, and $H_1 : \theta \notin \Theta_0$. This is a goodness-of-fit test.

A hypothesis is said to be *simple* if it fully specifies the distribution of X.

Example 4.2.

We look at the above hypotheses.

1. Both H_0 and H_1 are simple.
2. H_0 is simple, but H_1 is not, as it does not determine the value of θ, hence does not determine the distribution of X.
3. Neither H_0 nor H_1 are simple.
4. H_0 is simple if and only if Θ_0 contains exactly one value. Similarly H_1 is simple if and only if Θ_0^c contains exactly once value.

A test of H_0 is defined by a *critical region* $C \subset \mathcal{X}$. When $X \in C$, we “reject” H_0, and when $X \notin C$ we say we “fail to reject” or “find no evidence against” H_0.

To each test, we can associate two types of errors:

Type 1 error: We reject H_0 when H_0 is true.

Type 2 error: We fail to reject H_0 when H_0 is false.

When H_0 and H_1 are simple, we define

$$\begin{aligned}\alpha &= \mathbb{P}_{H_0}(H_0 \text{ is rejected}) = \mathbb{P}_{H_0}(X \in C),\\ \beta &= \mathbb{P}_{H_1}(H_0 \text{ is not rejected}) = \mathbb{P}_{H_1}(X \notin C).\end{aligned}$$

Here α is the probability of a type 1 error, and β is the probability of a type 2 error. The *size* of a test is α, and the *power* of the test is $1-\beta$.

There is a trade-off between minimizing size and maximizing power. Usually, we fix an acceptable size, then pick a test of size α which maximizes the power.

4.1 Neyman-Pearson Lemma

Let H_0, H_1 be simple, and let X have probability distribution function f_i under H_0, for $i = 0, 1$.

The *likelihood ratio statistic* is

$$\Lambda_x(H_0, H_1) = \frac{f_1(X)}{f_0(X)}.$$

A *likelihood ratio test* (or LRT) rejects H_0 when

$$X \in C = \{x \mid \Lambda_x(H_0, H_1) > k\},$$

for some threshold or critical value k.

Theorem 4.1 (Neyman-Pearson Lemma). *Suppose that f_0, f_1 are non-zero on the same sets. Suppose there exists k such that the likelihood ratio test with critical region*

$$C = \{x \mid \Lambda_x(H_0, H_1) > k\}$$

has size α.

Then, this is the test with the smallest β (highest power) out of all tests of size less than or equal to α.

Remark. A LRT of size α may not exist. Even then, there is a "randomized LRT" with size α.

Proof: Let $\bar{C}$ be the complement of C. The LRT has

$$\alpha = \mathbb{P}_{H_0}(X \in C) = \int_C f_0(x)\,\mathrm{d}x,$$
$$\beta = \mathbb{P}_{H_1}(X \notin C) = \int_{\bar{C}} f_1(x)\,\mathrm{d}x.$$

Let C^* be the critical region of another test with size α^* and power $1-\beta^*$, with $\alpha^* \le \alpha$. Then we will prove $\beta \le \beta^*$, or $\beta - \beta^* \le 0$. Now,

$$\begin{aligned}
\beta - \beta^* &= \int_{\bar{C}} f_1(x)\,\mathrm{d}x - \int_{\bar{C^*}} f_1(x)\,\mathrm{d}x \\
&= \int_{\bar{C}\cap C^*} f_1(x)\,\mathrm{d}x - \int_{\bar{C^*}\cap C} f_1(x)\,\mathrm{d}x \\
&= \int_{\bar{C}\cap C^*} \frac{f_1(x)}{f_0(x)} f_0(x)\,\mathrm{d}x - \int_{\bar{C^*}\cap C} \frac{f_1(x)}{f_0(x)} f_0(x)\,\mathrm{d}x \\
&\le k\left[\int_{\bar{C}\cap C^*} f_0(x)\,\mathrm{d}x - \int_{\bar{C^*}\cap C} f_0(x)\,\mathrm{d}x\right] \\
&= k\left[\int_{C^*} f_0(x)\,\mathrm{d}x - \int_C f_0(x)\,\mathrm{d}x\right] \\
&= k(\alpha^* - \alpha) \le 0.
\end{aligned}$$

Example 4.3.

Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma_0^2)$, where the variance σ_0^2 is known. We want the best size α test for the hypotheses $H_0 : \mu = \mu_0$, and $H_1 : \mu = \mu_1$, for some fixed $\mu_1 > \mu_0$.

The likelihood ratio statistic is

$$\begin{aligned}
\Lambda_x(H_0, H_1) &= \frac{\exp(-\frac{1}{2\sigma_0^2}\sum(x_i - \mu_1)^2)}{\exp(-\frac{1}{2\sigma_0^2}\sum(x_i - \mu_0^2)} \\
&= \exp\left(\frac{\mu_1 - \mu_0}{\sigma_0^2} n\bar{x} + \frac{n(\mu_0^2 - \mu_1^2)}{2\sigma_0^2}\right).
\end{aligned}$$

This is monotone in $\bar{x}$, the sample mean. Hence, for any k, there is a c such that

$$\Lambda_x(H_0, H_1) > k \iff \bar{x} > c.$$

Thus the likelihood critical region is $\{x \mid \bar{x} > c\}$ for some constant c.

By the same logic, the likelihood ratio test is of the form

$$C = \left\{ \sqrt{n}\frac{(\bar{x} - \mu_0)}{\sigma_0} > c' \right\}.$$

We want to pick c' such that

$$\mathbb{P}_{H_0}\left(\sqrt{n}\frac{(\bar{x} - \mu_0)}{\sigma_0} > c' \right) = \alpha.$$

But we know that

$$\sqrt{n}\frac{(\bar{x} - \mu_0)}{\sigma_0} \sim N(0, 1),$$

so if we take $c' = \Phi^{-1}(1 - \alpha) = z_\alpha$, the LRT has critial region

$$\left\{ x \mid \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma_0} > z_\alpha \right\}.$$

By the Neyman-Pearson lemma, this is the most powerful test of size α.

This is called a z-test, because we use a z statistic to define the critical region.

4.2 P-value

For any test with critical region of the form $\{x \mid T(x) > k\}$ for some statistic T, a *p-value* or observed significance level is

$$p = \mathbb{P}_{H_0}(T(X) > T(x^*)),$$

where x^* is the observed data. In the above example, if we let $\mu_0 = 5$, $\mu_1 = 6$, $\sigma_0 = 1$ and $\alpha = 0.05$, and we observe

$$x^* = (5.1, 5.5, 4.9, 5.3),$$

then $\bar{x^*} = 5.2$, and $z^* = 0.4$. The value $z_\alpha = \Phi^{-1}(1 - \alpha) = 1.645$, and so in this case we fail to reject $H_0 : \mu = 5$, with p-value 0.35.

Proposition 4.1. *Under H_0, p has $U[0, 1]$ distribution, where p is a function of x^*, and the null distribution assumes $x^* \sim \mathbb{P}_{H_0}$.*

Proof: Let F be the cdf of T. Then,

$$\begin{aligned}\mathbb{P}_{H_0}(p < u) &= \mathbb{P}_{H_0}(1 - F(T) < u) = \mathbb{P}_{H_0}(F(T) > 1 - u)\\ &= \mathbb{P}_{H_0}(T > F^{-1}(1-u)) = 1 - F(F^{-1}(1-u)) = u,\end{aligned}$$

for all $u \in [0,1]$. Hence $p \sim U[0,1]$.

4.3 Composite Hypotheses

Let $x \sim F_X(\cdot \mid \theta)$, for $\theta \in \Theta$. Then we can consider composite hypotheses $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta_1$.

The type 1 and type 2 error probabilities depend on the value of θ within Θ_0 or Θ_1, respectively.

Let C be some critical region.

Definition 4.1. The *power function* of the test C is

$$W(\theta) = \mathbb{P}_\theta(X \in C).$$

The *size* of C is the worst case type 1 error probability:

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

We say that C is *uniformly most powerful* (or UMP) of size α for H_0 against H_1 if

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

and for any other test C^* of size $\leq \alpha$ with power function W^*, we have

$$W(\theta) \geq W^*(\theta),$$

for all $\theta \in \Theta_1$.

Note that the UMP does not need to exist. But in some simple cases, the LRT is the UMP.

Example 4.4.

Again let $X_1, \ldots, X_n$ be $N(\mu, \sigma_0^2)$ with σ_0^2 known, and we wish to test $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$ for some fixed μ_0.

We have studied the simple hypothesis where $H_0' : \mu = \mu_0$, $H_1' : \mu = \mu_1$, with $\mu_1 > \mu_0$, and found the LRT was

$$C = \left\{ x \middle| z = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma_0} > z_\alpha \right\}.$$

Now we claim the same test C is the UMP for H_0 against H_1. Indeed, the power function for C is

$$\begin{aligned} W(\mu) = \mathbb{P}_\mu(X \in C) &= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma_0} > z_\alpha \right) \\ &= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\bar{x} - \mu)}{\sigma_0} > z_\alpha + \sqrt{n}\frac{(\mu_0 - \mu)}{\sigma_0} \right) \\ &= 1 - \Phi\left(z_\alpha + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0} \right). \end{aligned}$$

This is monotone increasing in $\mu = (-\infty, \infty)$. Therefore the test C has size α as

$$\sup_{\mu \in \Theta_0} W(\mu) = \alpha.$$

It remains to show that if C^* is another test of size $\leq \alpha$ with power function W^*, then $W(\mu_1) \geq W^*(\mu_1)$ for all $\mu_1 > \mu_0$.

The main observation is that the critical region depends only on μ_0, and C is the LRT for the simple hypotheses H_0', H_1'. Hence any other test C^* of H_0 versus H_1 of size $\leq \alpha$ also has size $\leq \alpha$ for H_0' versus H_1'. Thus, by the Neyman-Pearson lemma, we know that $W(\mu_1) \geq W^*(\mu_1)$.

As we can apply this argument for any $\mu_1 > \mu_0$, we have

$$W^*(\mu_1) \leq W(\mu_1),$$

for all $\mu_1 > \mu_0$.

4.4 Generalized Likelihood Ratio Tests

Again, let $X \sim f_X(\cdot \mid \theta)$, and $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta_1$.

The *generalized likelihood ratio statistic* is

$$\Lambda_x(H_0; H_1) = \frac{\sup_{\theta \in \Theta_1} f_X(x \mid \theta)}{\sup_{\theta \in \Theta_0} f_X(x \mid \theta)}.$$

Large values of Λ_x indicate a larger departure from the null H_0.

Example 4.5.

Let $X_1, \dots, X_n$ be iid $N(\mu, \sigma_0^2)$ with σ_0 fixed. We wish to test

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0,$$

for fixed μ_0. Here $\Theta_0 = \{\mu_0\}$, $\Theta_1 = \mathbb{R} \setminus \{\mu_0\}$. The generalized likelihood ratio test (GLR) is

$$\Lambda_x(H_0; H_1) = \frac{(2\pi\sigma_0)^{-n/2} \exp(-\frac{1}{2\sigma_0^2} \sum (x_i - \bar{x})^2)}{(2\pi\sigma_0)^{-n/2} \exp(-\frac{1}{2\sigma_0^2} \sum (x_i - \mu_0)^2)}.$$

Taking twice the logarithm of Λ_x,

$$2 \log \Lambda_x = \frac{n}{\sigma_0^2} (\bar{x} - \mu_0)^2.$$

The GLR rejects when Λ_x is large (or when $2 \log \Lambda_x$ is large), i.e. when

$$\left| \sqrt{n} \frac{(\bar{x} - \mu_0)}{\sigma_0} \right|$$

is large. Under H_0, this has a $N(0, 1)$ distribution. For a test of size α, we reject if

$$\left| \sqrt{n} \frac{(\bar{x} - \mu_0)}{\sigma_0} \right| > z_{\sigma/2} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

As $2 \log \Lambda_x = n \frac{(\bar{x} - \mu_0)}{\sigma_0^2} \sim \chi_1^2$ under H_0, we can also define the critical region of the GLR test as

$$\left\{ x \mid n \frac{(\bar{x} - \mu_0)}{\sigma_0^2} > \chi_1^2(\alpha) \right\}.$$

In general, we can approximate the distribution of $2 \log \Lambda_x$ with a χ^2 distribution when n is large:

4.5 Wilk's theorem

Suppose θ is k-dimensional, $\theta = (\theta_1, \dots, \theta_n)$.

The *dimension* of a hypothesis $H_0 : \theta \in \Theta_0$ is the number of free parameters in Θ_0.

Example 4.6.

1. If
$$\Theta_0 = \{\theta \in \mathbb{R}^k \mid \theta_1 = \theta_2 = \cdots = \theta_p = 0\}$$
for some $p < k$, then $\dim(\Theta_0) = k - p$.
2. Let $A \in \mathbb{R}^{p \times k}$, and $b \in \mathbb{R}^p$, with $p < k$. Let
$$\Theta_0 = \{\theta \in \mathbb{R}^k \mid A\theta = b\}.$$
Then $\dim(\Theta_0) = k - p$, if the rows of A are linearly independent, and Θ_0 is a hyperplane.
3. Let
$$\Theta_0 = \{\theta \in \mathbb{R}^k \mid \theta_0 = f_i(\phi), \phi \in \mathbb{R}^p\},$$
for $p < k$. Here ϕ are the free parameters, and f_i need not be linear. Under these conditions, $\dim(\Theta_0) = p$.

Theorem 4.2 (Wilk's theorem)**.** *Suppose* $\Theta_0 \subset \Theta_1$. *Let*
$$\dim(\Theta_1) - \dim(\Theta_0) = p.$$
If $X_1, \ldots, X_n$ *are iid from* $f_X(\cdot \mid \theta)$*, then as* $n \to \infty$*, the limiting distribution of* $2\log\Lambda_x$ *under* H_0 *is* χ_p^2.

I.e. for any $\theta \in \Theta_0$*, and any* $l > 0$*,*
$$(2\log\Lambda_x \leq l) \overset{n\to\infty}{\to} \mathbb{P}(\Xi \leq l),$$
where $\Xi \sim \chi_p^2$.

We can use this as follows: if we reject H_0 when $2\log\Lambda_x \geq \chi_p^2(\alpha)$, then when n is large, the size of the test is approximately α.

Example 4.7.

In the two-sided normal mean test,
$$\Theta_0 = \{\mu_0\}, \qquad \Theta_1 = \mathbb{R} \setminus \{\mu_0\},$$
we found $2\log\Lambda_x \sim \chi_1^2$.

If we take $\Theta_1 = \mathbb{R}$, the GRL statistic doesn't change, so $2\log\Lambda_x \sim \chi_1^2$, and
$$\dim(\Theta_1) - \dim(\Theta_0) = 1 - 0 - 1.$$

Here, the prediction of Wilk's theorem is exact.

4.6 Tests of Goodness-of-fit

Let $X_1, \dots, X_n$ be iid samples from a distribution on $\{1, 2, \dots, k\}$.

Let $p_i = \mathbb{P}(X_1 = i)$, and let N_i be the number of observations equal to i. Hence,

$$\sum_{i=1}^{n} p_i = 1, \qquad \sum_{i=1}^{n} N_i = n.$$

For a goodness-of-fit test, $H_0 : p = \tilde{p}$, for some fixed distribution $\tilde{p}$ on $\{1, \dots, k\}$.

Let $H_1 : p$ is any distribution with

$$\sum_{i=1}^{n} p_i = 1, \qquad p_i \geq 0.$$

Example 4.8.

Mendel's experiment involved crossing $n = 556$ smooth yellow peas with wrinkly green peas.

Each member of the progeny can have any combination of the two features. Let (p_1, p_2, p_3, p_4) be the probabilities of each type, and (N_1, N_2, N_3, N_4) the number of each progeny of each type. Then Mendel's hypothesis is

$$H_0 : p = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right) = \tilde{p}.$$

Then we want to see if there is any evidence in N to reject H_0. The model can be written as

$$(N_1, N_2, N_3, N_4) \sim \text{Multinomial}(n; p_1, p_2, p_3, p_4).$$

The likelihood is

$$L(p) \propto p_1^{N_1} p_2^{N_2} p_3^{N_3} p_4^{N_4},$$

hence

$$l(p) = C + \sum_{i=1}^{4} N_i \log p_i.$$

We can test H_0 against H_1 using a GLR test:

$$2 \log \Lambda_x = 2(\sup_{p \in \Theta_1} l(p) - \sup_{p \in \Theta_0} l(p)).$$

The latter term is $l(\tilde{p})$. In the alternative, p must satisfy $\sum p_i = 1$. So

$$\sup_{p\in\Theta_1} l(p) = \sup_{\sum p_i = 1} \sum_{i=1}^{4} N_i \log p_i.$$

Use the Lagrangian

$$\mathcal{L}(p,\lambda) = \sum_{i=1}^{4} N_i \log p_i - \lambda\left(\sum_{i=1}^{4} p_i - 1\right).$$

We find that $\hat{p}_i = \frac{N_i}{n}$, the observed proportion of samples of type i. Hence

$$2\log\Lambda_x = 2(l(\hat{p}) - l(\tilde{p})) = 2\sum_{i=1}^{4} N_i \log\left(\frac{N_i}{n\tilde{p}_i}\right).$$

Wilk's theorem tells us that $2\log\Lambda_x$ is approximately χ^2_p with

$$p = \dim(\Theta_1) - \dim(\Theta_0) = (k-1) - 0 = k-1.$$

So we can reject H_0 with size approximately α when

$$2\log\Lambda_x > \chi^2_{k-1}(\alpha).$$

It is common to write

$$2\log\Lambda = 2\sum_i o_i \log\left(\frac{o_i}{e_i}\right),$$

where $o_i = N_i$ is the observed number of type i, and $e_i = n\tilde{p}_i$ is the expected number of type i under the null hypothesis.

4.7 Pearson's Statistic

Let $\delta_i = o_i - e_i$. Then,

$$\begin{aligned} 2\log\Lambda &= 2\sum_i (e_i + o_i)\log\left(1 + \frac{\delta_i}{e_i}\right) \approx 2\sum_i \left(\delta_i + \frac{\delta_i^2}{e_i} - \frac{\delta_i^2}{2e_i}\right) \\ &= \sum_i \frac{\delta_i^2}{e_i} = \sum_i \frac{(o_i - e_i)^2}{e_i}. \end{aligned}$$

This is called *Pearson's statistic*. This also tends to a χ^2_{k-1} distribution when n is large.

Example 4.9.

We return to Mendel's experiment, this time with the data

$$(n_1, n_2, n_3, n_4) = (315, 108, 102, 31).$$

Then the GLR and Pearson's statistics are

$$2 \log \Lambda \approx 0.618, \qquad \sum_i \frac{(o_i - e_i)^2}{e_i} \approx 0.604.$$

We refer each statistic to a $\chi^2_{k-1} = \chi^2_3$ distribution. We get $\chi^2_3(0.05) = 7.815$. Thus we don't reject H_0 at size 5%.

The p-value is $\mathbb{P}(\chi^2_3 > 0.6) \approx 0.96$. In fact, the data fits the null model almost too well.

We can also have a goodness-of-fit test for a composite null, i.e.

$$\begin{aligned} &H_0 : p_i = p_i(\theta), \\ &H_1 : p \text{ any distribution on } \{1, \dots, k\}. \end{aligned}$$

Example 4.10.

Individuals can have three genotypes. We have a null-hypothesis

$$H_0 : p_1 = \theta^2, \quad p_2 = 2\theta(1-\theta), \quad p_3 = (1-\theta)^2,$$

for some $\theta \in [0, 1]$. Then

$$2 \log \Lambda = 2\big(\sup_{\text{any } p} l(p) - \sup_{\theta} l(p(\theta))\big) = 2\big(l(\hat{p}) - l(p(\hat{\theta}))\big),$$

where $\hat{p}$ is the maximum likelihood estimator in the alternative H_1, and $\hat{\theta}$ is the maximum likelihood estimator in null H_θ.

Last time we found that $\hat{p}_i = \frac{N_i}{n}$. Then $\hat{\theta}$ would need to be computed for the null model in question. We get that

$$2 \log \Lambda = 2 \sum_i N_i \log\left(\frac{N_i}{np_i(\hat{\theta})}\right) = 2 \sum_i o_i \log\left(\frac{o_i}{e_i}\right),$$

where again $o_i = N_i$ is the observed number of type i, and $e_i = np_i(\hat{\theta})$ is the expected number of type i under the null hypothesis.

We can similarly define a Pearson statistic

$$\sum_i \frac{(o_i - e_i)^2}{e_i}$$

using the same argument as before.

Each statistic can be referred to a χ^2_d when n is large by Wilk's theorem, where

$$d = \dim(\Theta_1) - \dim(\Theta_0) = k - 1 - \dim(\Theta_0).$$

Example 4.11.

Going back to our example, we have

$$l(\theta) = \sum_i N_i \log p_i(\theta) = 2N_1 \log \theta + N_2 \log(2\theta(1-\theta)) + 2N_3 \log(1-\theta).$$

Maximizing over $\theta \in [0, 1]$ gives

$$\hat{\theta} = \frac{2N_1 + 2N_2}{2n}.$$

In this model $2 \log \Lambda$ and the Pearson statistic have a χ^2_d distribution with

$$d = k - 1 - \dim(\Theta_0) = 3 - 1 - 1 = 1.$$

4.8 Testing Independence in Contingency Tables

Suppose $(X_1, Y_1), \dots, (X_n, Y_n)$ are independent with X_i taking values in $\{1, \dots, r\}$, and Y_i taking values in $\{1, \dots, c\}$.

The entries in a contingency table are

$$N_{ij} = |\{l \mid 1 \le l \le n, (X_l, Y_l) = (i, j)\}|,$$

which is the number of samples of type (i, j).

Example 4.12.

For COVID-19 deaths, we can take X_i to be the age group of the i'th death, and Y_i the week on which it fell.

From these statistics, we wish to see if deaths are decreasing faster for an older age group that had been vaccinated.

Now we can construct the probability model. Assume n is fixed. Then a sample (X_l, Y_l) has probability p_{ij} of having value (i, j). Thus

$$(N_{11}, \dots, N_{1c}, N_{21}, \dots, N_{rc}) \sim \text{Multinomial}(n; p_{11}, \dots, p_{1c}, p_{21}, \dots, p_{rc}).$$

Remark. Fixing n may not be natural; we will consider other models as well.

The null hypothesis is that X_i is independent of Y_i for each sample. Formalizing, let

$$p_{i+} = \sum_{j=1}^{c} p_{ij}, \qquad p_{+j} = \sum_{i=1}^{r} p_{ij}.$$

Then the hypotheses are

$$\begin{aligned} &H_0 : p_{ij} = p_{i+}p_{+j}, \\ &H_1 : (p_{ij}) \text{ is unconstrained, except for } p_{ij} \geq 0, \sum p_{ij} = 1. \end{aligned}$$

The generalized LRT is

$$2 \log \Lambda = 2 \sum_{i,j} o_{ij} \log \left(\frac{o_{ij}}{e_{ij}} \right),$$

where $o_{ij} = N_{ij}$, and $e_{ij} = n\hat{p}_{ij}$, and $\hat{p}$ is the maximum likelihood estimator under the independence model H_0. Using Lagrange multipliers, we can find $\hat{p}_{ij} = \hat{p}_{i+}\hat{p}_{+j}$, where

$$\begin{aligned} \hat{p}_{i+} &= \frac{N_{i+}}{n}, & \hat{p}_{+j} &= \frac{N_{+j}}{n}, \\ N_{i+} &= \sum_j N_{ij}, & N_{j+} &= \sum_i N_{ij}. \end{aligned}$$

Hence the GLR is

$$2 \log \Lambda = 2 \sum_{i,j} N_{ij} \log \left(\frac{N_{ij}}{n\hat{p}_{i+}\hat{p}_{+j}} \right) \approx \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}.$$

From Wilk's, the asymptotic distribution of these statistic is χ^2_d with

$$d = \dim(\Theta_1) - \dim(\Theta_0) = (rc - 1) - [(r - 1) + (c - 1)] = (r - 1)(c - 1).$$

Recall: if X is a random vector,

$$\mathbb{E}[AX + b] = A\mathbb{E}[X] + b, \qquad \text{Var}(AX + b) = A \, \text{Var}(X) A^T.$$

Definition 4.2. We say X has multivariate normal distribution if for any $t \in \mathbb{R}^n$, $t^T X$ is normal.

Proposition 4.2. *If X if MVN, then $AX + b$ is MVN.*

Proof: Say $AX + b$ is in $\mathbb{R}^m$ Take $t \in \mathbb{R}^m$, then

$$t^T(AX + b) = (A^T t)^t X + t^t b,$$

where the first term is $N(\mu, \sigma^2)$ for some μ, σ^2, and the latter term is constant. Thus,

$$t^T(AX + b) \sim N(\mu + t^T b, \sigma^2).$$

Proposition 4.3. *A MVN distribution is fully specified by its mean and variance.*

Proof: Take X_1, X_2 both MVN with same mean μ, variance Σ. We will show that their mgf's are the same, hence X_1, X_2 have the same distribution:

$$\begin{aligned}\mathbb{E}[e^{t^T X_1}] = M_{t^T X_1}(1) &= \exp\left(1 = \mathbb{E}[t^T X_1] + \frac{1}{2}\operatorname{Var}(t^T X_1)\right) \\ &= \exp\left(t^t \mu + \frac{1}{2} t^T \Sigma t\right).\end{aligned}$$

This only depends on μ, Σ, so it is the same for X_1, X_2.

4.9 Orthogonal Projection

Definition 4.3. We say $P \in \mathbb{R}^{n \times n}$ is an *orthogonal projection* if it is:

- independent: $PP = P$,
- symmetric: $P^T = P$.

Equivalently, $P \in \mathbb{R}^{n \times n}$ is an *orthogonal projection* if for any v in the column space, $Pv = v$, and for any w perpendicular to the column vectors, $Pw = 0$.

Proposition 4.4. *These two definitions are equivalent.*

Proof: To show the first definition equals the second, take v a column vector of P, so $v = Pa$ for some $a \in \mathbb{R}^n$. Then,

$$Pv = PPa = Pa = v.$$

Take w perpendicular to the column space. Then $P^T w = 0$. Then,

$$Pw = P^T w = 0.$$

To show the second definition implies the first, we can write any $a \in \mathbb{R}^n$ uniquely as $a = v + w$, where v is in the column space, and w is perpendicular to the column space. Then

$$PPa = PP(v+w) = Pv = P(v+w) = Pa.$$

As a was arbitrary, $P = P^2$. For symmetry, take $u_1, u_2 \in \mathbb{R}^n$. Then,

$$(Pu_1)^T((I-P)u_2) = 0.$$

Hence,

$$u_1^T(P^T - P^TP)u_2 = 0.$$

Since this holds for all u_1, u_2, $P^T = P^TP$. Hence $P^T = P$.

Corollary 4.1. *If P is an orthogonal projection, then so is $I-P$.*

Proof: We have $(I-P)^T = I^T - P^T = I - P$, and

$$(I-P)^2 = I - 2P + P^2 = I - P.$$

Proposition 4.5. *If $P \in \mathbb{R}^{n\times n}$ is an orthogonal projection, then $P = UU^T$, where the columns of U form an orthogonal basis for the columns space of P.*

Proposition 4.6. *UU^T is clearly symmetric and idempotent,*

$$UU^TUU^T = UU^T.$$

So UU^T is an orthogonal projection. To show that it is equal to P, note the column space is equal to the column space of P by construction.

Corollary 4.2. $n = \mathrm{rank}(P) = \mathrm{tr}(U^TU) = \mathrm{tr}(UU^T) = \mathrm{tr}(P)$.

Theorem 4.3. *If X is MVN, $X \sim N(0, \sigma^2 I)$ and P is an orthogonal projection, then*

1. *$PX \sim N(0, \sigma^2 P)$, $(I-P)X \sim N(0, \sigma^2(I-P))$, PX and $(I-P)X$ independent.*
2. *$\frac{\|PX\|^2}{\sigma^2} \sim \chi^2_{\mathrm{rank}(P)}$.*

Proof: The vector

$$\begin{pmatrix} P \\ (I-P) \end{pmatrix} X = AX$$

is MVN, because it is a linear function of X. The distribution is specified by the mean and variance:

$$\mathbb{E}[AX] = \begin{pmatrix} P \\ I - P \end{pmatrix} \mathbb{E}[X] = 0,$$

and

$$\begin{aligned} \operatorname{Var}(AX) &= \begin{pmatrix} P \\ I - P \end{pmatrix} X \begin{pmatrix} P \\ I - P \end{pmatrix}^T = \begin{pmatrix} P \\ I - P \end{pmatrix} \sigma^2 I \begin{pmatrix} P \\ I - P \end{pmatrix}^T \\ &= \sigma^2 \begin{pmatrix} P & 0 \\ 0 & I - P \end{pmatrix}. \end{aligned}$$

Let $Z \sim N(0, \sigma^2 P)$, and $Z' \sim N(0, \sigma^2(I - P))$, with Z, Z' independent. Then,

$$\begin{pmatrix} Z \\ Z' \end{pmatrix} \sim N\left(0, \sigma^2 \begin{pmatrix} P & 0 \\ 0 & I - P \end{pmatrix}\right).$$

Hence we have

$$\begin{pmatrix} PX \\ (I - P)X \end{pmatrix} = \begin{pmatrix} Z \\ Z' \end{pmatrix}.$$

Therefore $PX, (1 - P)X$ independent.

To show 2, note

$$\frac{\|PX\|^2}{\sigma^2} = \frac{(PX)^T PX}{\sigma^2} = \frac{X^T (UU^T)^T UU^T X}{\sigma^2} = \frac{X^T UU^T X}{\sigma^2}.$$

The columns of U form an orthogonal basis for the column space of P, so

$$\frac{\|PX\|^2}{\sigma^2} = \frac{\|U^T X\|^2}{\sigma^2} = \sum_{i=1}^{\operatorname{rank}(P)} \frac{(U^T X)_i}{\sigma^2}.$$

But $U^T X \sim N(0, \sigma^2 I)$, so

$$\operatorname{Var}(U^T X) = U^T \operatorname{Var}(X) U = \sigma^2 U^T U = \sigma^2 I.$$

Therefore $(U^T X)_i$, for $i = 1, \ldots, \operatorname{rank}(P)$, and iid $N(0, \sigma^2)$. Thus, $\frac{\|PX\|^2}{\sigma^2}$ is the sum of $\operatorname{rank}(P)$ squared independent $N(0, 1)$ variables, meaning it is $\chi^2_{\operatorname{rank}(P)}$.

Let's look at an example. Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, with μ, σ^2 unknown.

Recall that the MLE for μ is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The MLE for σ^2 is

$$\hat{\sigma}^2 = \frac{S_{xx}}{n},$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Theorem 4.4.

(i) $\bar{X} \sim N(\mu, \sigma^2/n)$,

(ii) $\frac{S_{xx}}{\sigma^2} \sim \chi^2_{n-1}$,

(iii) $\bar{x}$, S_{xx} *independent.*

Proof: Let $\mathbf{1} = (1, \dots, 1)^T$. Let $P = \frac{1}{n}\mathbf{1}\mathbf{1}^T$ be an orthogonal projection. It is easy to check that $P = P^T = P^2$. We can write

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \mu\mathbf{1} + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2 I)$. Now note,

- $\bar{X}$ is a function of PX,

  $$PX = \mu\mathbf{1} + P\varepsilon.$$

  Because $\bar{X} = (PX)_1$. In particular, $\bar{X}$ is a function of $P\varepsilon$.

- We also have

  $$S_{xx} = \sum_{i=1}^{n} (X_i - \bar{X})^2 = \|X - \mathbf{1}\bar{X}\|^2 = \|(I-P)X\|^2 = \|(I-P)\varepsilon\|^2,$$

  so S_{xx} is a function of $(I-P)\varepsilon$.

By the previous theorem, $P\varepsilon$ and $(I-P)\varepsilon$ are independent, so $\bar{X}$ and S_{xx} are independent. Also,

$$\frac{S_{xx}}{\sigma^2} = \frac{\|(I-P)\varepsilon\|^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Consider a linear model $Y = X\beta + \varepsilon$, where

$$\hat{\beta} = \text{argmin}_\beta \|Y = X\beta\|^2 = (X^T X)^{-1} X^T Y.$$

Then $P = X(X^T X)^{-1} X^T$ is an orthogonal projection onto the column space of X, and also

$$\hat{Y} = X\hat{\beta} = PY, \qquad Y - \hat{Y} = (I - P)Y.$$

Theorem 4.5. *Let $\varepsilon \sim N(0, \sigma^2 I)$. Then,*

1. $P\varepsilon \sim N(0, \sigma^2 P)$*,* $(I - P)\varepsilon \sim N(0, \sigma^2(I - P))$*,*
2. $P\varepsilon$ *independent of* $(I - P)\epsilon$*.*
3. *We have*
$$\frac{\|P\epsilon\|^2}{\sigma^2} \sim \chi^2_{\text{rank}(P)}.$$

4.10 Normal Linear Model

Take $Y = X\beta + \varepsilon$, where $\epsilon \sim N(0, \sigma^2 I)$. The MLE has two parameters: $\beta \in \mathbb{R}^p$ and $\sigma^2 \in \mathbb{R}_+$. The log-likelihood is

$$l(\beta, \sigma^2) = c + \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \|Y - X\beta\|^2.$$

For any $\sigma^2 > 0$, we can see that $l(\sigma^2, \beta)$ is maximized as a function of β at the minimizer of $\|Y - X\beta\|^2$, i.e. the least-squares estimator $\hat{\beta}$. Now we find

$$\frac{\partial l}{\partial \sigma^2}(\hat{\beta}, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \|Y - X\hat{\beta}\|^2.$$

As $\sigma^2 \mapsto l(\hat{B}, \sigma^2)$ is unique, there is a unique maximizer when the derivative is 0, i.e.

$$\hat{\sigma}^2 = \frac{\|Y - X\hat{\beta}\|^2}{n} = \frac{\|(I - P)Y\|^2}{n}.$$

Theorem 4.6.

1. $\hat{\beta} \sim N(\beta, \sigma^2(X^T X)^{-1})$*,*
2. $\frac{\hat{\sigma}^2}{\sigma} n \sim \chi^2_{n-p}$*,*
3. $\hat{\beta}, \hat{\sigma}^2$ *are independent.*

Proof: As $\hat{\beta}$ is linear in Y, it is a multivariate normal. We already know $\mathbb{E}[\hat{\beta}] = \beta$, and $\mathrm{Var}(\hat{\beta}) = \sigma^2(X^TX)^{-1}$, so this proves (1).

For (2), note

$$\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{\|(I-P)Y\|^2}{\sigma^2} = \frac{\|(I-P)(X\beta+\epsilon)\|^2}{\sigma^2} = \frac{\|(I-P)\epsilon\|^2}{\sigma^2} \sim \chi^2_{\mathrm{rank}(I-P)}.$$

As $\mathrm{rank}(I-P) = n-p$, this proves (2).

Finally for (3), note $\hat{\sigma}^2$ is a function of $(I-P)\epsilon$. We will show that $\hat{\beta}$ is a function of $P\epsilon$, which implies they are independent. Indeed,

$$\begin{aligned}\hat{\beta} &= (X^TX)^{-1}X^TY = (X^TX)^{-1}X^T(X\beta+\varepsilon) = \beta + (X^TX)^{-1}X^T\varepsilon \\ &= \beta + (X^TX)^{-1}X^TP\varepsilon,\end{aligned}$$

as $X^TP = X^T$

Corollary 4.3. *$\hat{\sigma}^2$ is biased:*

$$\mathbb{E}\left[\frac{\hat{\sigma}^2 n}{\sigma^2}\right] = n-p \implies \mathbb{E}[\hat{\sigma}^2] = \left(\frac{n-p}{n}\right)\sigma^2.$$

4.10.1 Student's t-distribution

If $U \sim N(0,1)$, $V \sim \chi^2_n$ and U, V independent, then we say that

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

has a t_n distribution.

4.10.2 The F distribution

If $V \sim \chi^2_n$, $W \sim \chi^2_m$ and V, W independent, then we say

$$F = \frac{V/n}{W/m}$$

has a $F_{n,m}$ distribution.

4.10.3 Confidence Sets for β

Suppose we want a $100(1-\alpha)\%$ confidence interval for one of the coefficients, say β_1. Note that

$$\frac{\beta_1 - \hat{\beta}_1}{\sqrt{\sigma^2 (X^T X)^{-1}_{11}}} \sim N(0,1),$$

because $\hat{\beta}_1 = N(\beta_1, \sigma^2 (X^T X)^{-1}_{11})$. Also,

$$\frac{\hat{\sigma}^2}{\sigma^2} n \sim \chi^2_{n-p},$$

and these two statistic are independent. Hence,

$$\frac{\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 (X^T X)^{-1}_{11}}}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2} \frac{n}{n-p}}} \sim \frac{N(0,1)}{\sqrt{\chi^2_{n-p}/(n-p)}} \sim t_{n-p},$$

which depends only on β_1 and not on σ^2. Hence we can use this as a pivot:

$$\mathbb{P}_{\beta,\sigma^2}\left(-t_{n-p}(\alpha/2) \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{(X^T X)^{-1}_{11}}} \sqrt{\frac{n-p}{n\hat{\sigma}^2}} \leq t_{n-p}(\alpha/2) \right) = 1 - \alpha.$$

We can use the fact that t_n distribution is symmetric around 0. Rearranging the inequalities, we get

$$\mathbb{P}_{\beta,\sigma^2}(\hat{\beta}_1 - M \leq \beta_1 \leq \hat{\beta}_1 + M) = 1 - \alpha,$$

where

$$M = t_{n-p}(\alpha/2) \sqrt{\frac{(X^T X)^{-1}_{11} \hat{\sigma}^2}{(n-p)/n}}.$$

We conclude that $[\hat{\beta}_1 \pm M]$ is a $(1-\alpha)100\%$ confidence interval for β_1.

Note that this is not asymptotic.

By the duality between tests of significance and confidence intervals, we can find a size α test for

$$H_0 : \beta_1 = \beta^*,$$
$$H_1 : \beta_1 \neq \beta^*.$$

We simply reject H_0 if β^* is not contained in the $100(1-\alpha)\%$ confidence interval for β_1.

4.11 Confidence Ellipsoids for β

Note that $\hat{\beta} - \beta \sim N(0, \sigma^2(X^TX)^{-1})$. As X has full rank X^TX is positive definite. So it has eigendecomposition

$$X^TX = UDU^T,$$

where D is diagonal and U is unitary. Define

$$(X^TX)^\alpha = UD^\alpha U^T,$$

where

$$D^\alpha = \begin{pmatrix} D_{11}^\alpha & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & D_{pp}^\alpha \end{pmatrix}.$$

Then

$$(X^TX)^{1/2}(\hat{\beta} - \beta) \sim N(0, \sigma^2 I).$$

Hence,

$$\frac{\|X(\hat{\beta} - \beta)\|^2}{\sigma^2} = \frac{\|(X^TX)^{1/2}(\hat{\beta} - \beta)\|^2}{\sigma^2} \sim \chi_p^2.$$

This is a function of $\hat{\beta}$, so is independent of

$$\frac{\hat{\sigma}^2 n}{\sigma^2} \sim \chi_{n-p}^2.$$

Hence,

$$\frac{\|X(\hat{\beta} - \beta)\|^2/p}{\hat{\sigma}^2 n/(n-p)} \sim F_{p,n-p}.$$

This only depends on β, and not on σ^2, so it can be used as a pivot. For all β, σ^2

$$\mathbb{P}_{\sigma^2,\beta}\left(\frac{\|X(\hat{\beta} - \beta)\|^2/p}{\hat{\sigma}^2 n/(n-p)} \leq F_{p,n-p}(\alpha)\right) = 1 - \alpha.$$

So we can say that the set

$$\left\{\beta \in \mathbb{R}^p \;\middle|\; \frac{\|X(\hat{\beta} - \beta)\|^2/p}{\hat{\sigma}^2 n/(n-p)} \leq F_{p,n-p}(\alpha)\right\}$$

is a $100(1-\alpha)\%$ confidence set for β. The principal axes are given by the eigenvectors of X^TX.

Index