Statistics

Notes by Finley Cooper

26th February 2026

Contents

1 Parametric Estimation

1.1 Review of IA Probability

1.1.1 Starting axioms

We observe some data $X_1, \ldots, X_n$ iid random variables taking values in a sample space $\mathcal{X}$. Let $X = (X_1, \ldots, X_n)$. We assume that X_1 belongs to a *statistical model* $\{p(x;\theta) : \theta \in \Theta\}$ with θ unknown. For example $p(x;\theta)$ could be a pdf.

Let's see some examples

(i) Suppose that $X_1 \sim \text{Poisson}(\lambda)$ where $\theta = \lambda \in \Theta = (0, \infty)$.

(ii) Suppose that $X_1 \sim \mathcal{N}(\mu, \sigma^2)$, where $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$.

We have some common questions about these statistical models.

(i) We want to give an estimate $\hat{\theta} : \mathcal{X}^n \to \Theta$ of the true value of θ.

(ii) We also want to give an interval estimator $(\hat{\theta}_1(X), \hat{\theta}_2(X))$ of θ.

(iii) Further we want to test of hypothesis about θ. For example we might make the hypothesis that $H_0 : \theta = 0$.

Let's do a quick reivew of IA Probability. Let $X : \Omega \to \mathbb{R}$ be a random variable defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. So Ω is the sample space, $\mathcal{F}$ is the set of events, and $\mathbb{P} : \mathcal{F} \to [0, 1]$ is the probability measure.

The cumulative distribution function (cdf) of X is $F_X(s) = \mathbb{P}(X \leq x)$. A discrete random variable takes values in a countable set $\mathcal{X}$ and has probability mass function (pmf) given by $p_X(x) = \mathbb{P}(X = x)$. A continuous random variable has probability density function (pdf) f_X satisfying $P(X \in A) = \int_A f_X(x)\mathrm{d}x$ (for measurable sets A). We say that $X_1, \ldots, X_n$ are independent if $\mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod_{i=1}^n \mathbb{P}(X_i \leq x_i)$ for all choices $x_1, \ldots, x_n$. If $X_1, \ldots, X_n$ have pdfs (or pmfs) $f_{X_1}, \ldots, f_{X_n}$, then this is equivalent to $f_X(x) = \prod_{i=1}^n f_{X_i}(x_i)$ for all x_i. The expectation of X is,

$$\mathbb{E}(x) = \begin{cases} \sum_{x \in \mathcal{X}} x p_X(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f_X(x) & \text{if } X \text{ is continuous} \end{cases}.$$

The variance of X is $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$. The moment generating function of X is $M(t) = \mathbb{E}[e^{tX}]$ and can be used to generate the momentum of a random varaible by taking derivatives. If two random variables have the same moment generating functions, then they have the same distribution.

The expectation operator is linear and

$$\text{Var}(a_1 X_1 + \cdots + a_n X_n) = \sum_{i,j=1}^{n} a_i a_j \,\text{Cov}(X_i, X_j),$$

where $\text{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}(X_i)(X_j - \mathbb{E}(X_j))]$. In vector notation writing X as the column vector of X_i and a as the column vector for a_i we get that

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

Similar for the variance we get that

$$\text{Var}\left(a^T X\right) = a^T \,\text{Var}(X)\, a$$

where $\mathrm{Var}(X)$ is the covariance matrix for X with entries $\mathrm{Cov}(X_i, X_j)$.

1.1.2 Joint random variables

If X is a discrete random variable with pmf $P_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$ and marginal pmf $P_Y(y) = \sum_{x \in X} P_{X,Y}(x,y)$, then the conditional pmf is

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

If X, Y are continuous then the join pdf $f_{X,Y}$ satifies

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

and 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 of X given Y is

$$E(X \mid Y) = \begin{cases} \sum_{x \in X} x \mathbb{P}_{X|Y}(x \mid Y) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f_{X|Y}(x \mid Y) \mathrm{d}y & \text{if } Y \text{ is continuous} \end{cases}.$$

Remark. $\mathbb{E}(X \mid Y)$ is a function of Y so $\mathbb{E}(X \mid Y)$ is a random variable.

We also have the law of total expectation,

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

This is a consequence of the law of total probability which is

$$p_X(x) = \sum_y p_{X|Y}(x \mid y) p_Y(y).$$

Now we have a new (but less useful) theorem similar to the tower property of expectation.

Theorem. (Law of total variance)

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y]).$$

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

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

We also have the change of variables formula. If we have a mapping $(x,y) \to (u,v)$, a bijection from $\mathbb{R}^2 \to \mathbb{R}^2$, then

$$f_{U,V}(u,v) = f_{X,Y}(x(u,v), y(u,v)) |\det J|,$$

where J is the Jacobian matrix.

1.1.3 Limit theorems

Suppose $X_1, \ldots, X_n$ are iid random variables with mean μ and variance σ^2. Define the sum $S = \sum_{i=1}^n X_i$ and the sample mean $\bar{X}_n = \frac{S_n}{n}$. We have the following theorems.

Theorem. (Weak Law of Large Numbers)

$$\bar{X}_n \to \mu$$

where $\to$ means that $\mathbb{P}\left(|\bar{X}_n - \mu| > \varepsilon\right) \to 0$ as $n \to \infty$ for all $\varepsilon > 0$.

Theorem. (Strong Law of Large Numbers)

$$\bar{X}_n \to \mu$$

almost surely. So $\mathbb{P}\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1$.

Theorem. (Central Limit Theorem) The random variables

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

is approximately $\mathcal{N}(0,1)$ for large n. Or we can write this as

$$S_n \approx \mathcal{N}(n\mu, n\sigma^2).$$

Formally this means that $\mathbb{P}\left(Z_n \leq z\right) \to \Phi(z)$ for all $z \in \mathbb{R}$ where $\Phi(z)$ is the cdf of $\mathcal{N}(0,1)$.

1.2 Estimators

Suppose that $X_1, \ldots, X_n$ are iid with pdf $f_X(x \mid \theta)$ and parameter θ unknown.

Definition. (Estimator) A function of the data $T(X) \to \hat{\theta}$ which is used to approximate the true parameter θ is called an *estimator* (or sometimes a *statistic*). The distribution of $T(X)$ is the *sampling distribution*

For an example suppose that $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$ and let $\hat{\mu} = T(x) = \frac{1}{n}\sum_{i=1}^n X_i$. The sampling distribution of $\hat{\mu}$ is $T(X) \sim \mathcal{N}(\mu, \frac{1}{n})$.

Definition. (Bias) The *bias* of a random variable $\hat{\theta} = T(X)$ is

$$\text{bias}(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta}) - \theta,$$

where the expectation is taken over the model $X_1 \sim f_X(\cdot \mid \theta)$.

Remark. In general the bias might be a function of θ which is not explicit in the notation.

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

So for our estimator from before, $\hat{\mu}$, is unbiased since

$$\mathbb{E}_\mu(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\mu(X_i) = \mu.$$

1.2.1 Bias-variance decomposition

Definition. (Mean squared error) The *mean squared error* of an estimator $\hat{\theta}$ is

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

Remark. Note that the MSE is generally a function of θ like the bias. Again this is not clear from the notation.

Proposition. (Bias-variance decomposition) For an estimator $\hat{\theta}$ of a parameter θ, we have that

$$\text{mse}(\hat{\theta}) = \Big(\text{bias}(\hat{\theta})\Big)^2 + \text{Var}_\theta(\hat{\theta}).$$

Proof.

$$\begin{aligned}
\text{mse}(\hat{\theta}) &= \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] \\
&= \mathbb{E}_\theta\left[\Big(\hat{\theta} - \mathbb{E}_\theta(\hat{\theta}) + \mathbb{E}_\theta(\hat{\theta}) - \theta\Big)^2\right] \\
&= \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta(\hat{\theta}))^2] + (\mathbb{E}_\theta(\hat{\theta}) - \theta)^2 + 2(\mathbb{E}_\theta(\hat{\theta}) - \theta)\cdot\mathbb{E}_\theta[\hat{\theta} - \mathbb{E}_\theta(\hat{\theta})] \\
&= \Big(\text{bias}(\hat{\theta})\Big)^2 + \text{Var}_\theta(\hat{\theta}). \quad \square
\end{aligned}$$

Let's see an example. Suppose that $X \sim \text{Binomial}(n, \theta)$ where is n is known and we want to estimate $\theta \in [0, 1]$. Let $T_u = \frac{X}{n}$ be an estimator, so $\mathbb{E}_\theta(T_u) = \frac{\mathbb{E}(X)}{n} = \frac{n\theta}{n} = \theta$, hence this estimator is unbiased. And $\text{mse}(T_u) = \text{Var}\,(T_u) + \text{bias}(T_u) = \frac{\theta(1-\theta)}{n}$.

Instead if we used the estimator $T_b = \frac{X+1}{n+2} = \omega\frac{X}{n} + (1-\omega)\frac{1}{2}$ where $\omega = \frac{n}{n+2}$. We get that

$$\begin{aligned}
\text{bias}(T_b) &= (1-\omega)(\frac{1}{2} - \theta) \\
\text{Var}\,(T_b) &= \omega^2\frac{\theta(1-\theta)}{n}.
\end{aligned}$$

Giving that

$$\text{mse}(T_b) = \omega^2\frac{\theta(1-\theta)}{n} + (1-\omega)^2(\frac{1}{2} - \theta)^2$$

1.3 Sufficient statistics

Suppose $X_1, \dots, X_n$ are iid random variables taking values in χ with pdf $f_{X_1}(\cdot \mid \theta)$. Consider θ as fixed. Denote $X = (X_1, \dots, X_n)$.

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

Remark. The parameter θ may be a vector, and $T(X)$ may be a vector.

Suppose $X_1, \dots, X_n \sim \text{Binomial}(1, \theta)$ iid for some $\theta \in [0, 1]$. Then

$$\begin{aligned} f_X(x \mid \theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \end{aligned}$$

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

$$\begin{aligned} f_{X|T=t}(x \mid T(x) = t) &= \frac{\mathbb{P}_\theta(X = x, T(X) = t)}{\mathbb{P}_\theta(T(X) = t)} \\ &= \frac{\mathbb{P}_\theta(X = x)}{\mathbb{P}_\theta(T(X) = t)} = \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}. \end{aligned}$$

Theorem. (Factorisation criterion) The statistics T is sufficient for θ if and only if $f_X(x \mid \theta) = g(T(x), \theta)h(x)$ for some suitable g and h.

Proof. Suppose that $f_X(x \mid \theta) = g(T(x), \theta)h(x)$. We can compute

$$\begin{aligned} f_{X|T=t}(x \mid T = t) &= \frac{\mathbb{P}_\theta(X = x, T(x) = t)}{\mathbb{P}_\theta(T(x) = t)} \\ &= \frac{g(T(x), \theta)h(x)}{\sum_{x';T(x')=t} g(t, \theta)h(x')} \\ &= \frac{h(x)}{\sum_{x';T(x')=t} h(x')} \end{aligned}$$

which doesn't depend on θ, so $T(X)$ is sufficient.

Conversely, suppose $T(X)$ is sufficient. We can write

$$\begin{aligned} \mathbb{P}_\theta(X = x) &= \mathbb{P}_\theta(X = x, T(X) = T(x)) \\ &= \mathbb{P}_\theta(X = x \mid T(X) = T(x))\mathbb{P}_\theta(T(X) = T(x)) \\ &= h(x)g(T(X), \theta). \end{aligned}$$

So we're done. □

Remark. For our example before we can define $T(x) = \sum x_i$ and $g(t, \theta) = \theta^t(1-\theta)^{n-t}$ and $h(x) = 1$.

Let's see another example. Let $X_1, \dots, X_n$ be iid uniform on $[0, \theta]$ for some $\theta \in (0, \infty)$. So

$$\begin{aligned} f_X(x = \theta) &= \prod_{i=1}^{n} \frac{1}{\theta} \mathbf{1}\{x_i \in [0, \infty]\} \\ &= \frac{1}{\theta^n} \mathbf{1}\{\max x_i \leq \theta\} \mathbf{1}\{\min x_i \geq 0\} \\ &= g(T(x), \theta) h(x). \end{aligned}$$

1.4 Minimal sufficiency

Definition. (Minimal sufficient) A sufficient statistics $T(X)$ is *minimal sufficient* if it is a function of every other sufficient statistic. So if $T'(X)$ is also sufficient, then $T'(x) = T'(y) \implies T(x) = T(y)$ for all $x, y \in \chi$.

Remark. Minimal sufficient statistics are unique up to bijection.

Theorem. Suppose $T(X)$ is a statistics such that $\frac{f_X(x|\theta)}{f_X(y|\theta)}$ is constant a function of θ if and only if $T(x) = T(y)$. Then T is minimal sufficient.

Let's see an example before we prove this. Suppose that $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Then

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

This is constant in (μ, σ^2) if and only if $\sum_i x_i = \sum_i y_i$ and $\sum_i x_i^2 = \sum_i y_i^2$ therefore $T(X) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is minimal sufficient.

Proof. Need to show that such a statistics is sufficient and minimal. First we'll show sufficiency. For each t pick a x_t such that $T(x_t) = t$. Now let $x \in \chi_N^2$ and let $T(x) = t$. So $T(x) = T(x_t)$, so by the hypothesis $\frac{f_X(x,\theta)}{f_X(x_t,\theta)}$ does not depend on θ. Let this be $h(x)$ and let $g(t, \theta) = f_X(x, \theta)$ then we have that $f_X(x, \theta) = g(t, \theta)h(x)$ so sufficient.

Now let S be any other sufficient statistic. By the factorisation criterion, there exists g_S, h_S such that $f_X(x \mid \theta) = G_S(S(x), \theta) h_S(x)$. Suppose $S(x) = S(y)$. Then

$$\frac{f_X(x \mid \theta)}{f_X(y \mid \theta)} = \frac{g_S(S(x), \theta) h_S(x)}{g_S(S(y), \theta) h_S(y)} = \frac{h_S(x)}{h_S(y)}$$

which does not depend on θ so $T(x) = T(y)$ so T is minimal sufficient. □

We know that bijections of minimal sufficient statistics are still minial sufficient statistics, so we can write our minimal sufficient statistic for $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ as

$$S(X) = (\overline{X}, S_{XX})$$

where $\overline{X} = \frac{1}{n} \sum_i X_i$ and $S_{XX} = \sum_i (X_i - \overline{X})^2$, since there is a bijection between them.

Until now we used $\mathbb{E}_\theta$ and $\mathbb{P}_\theta$ to denote expectation and probability when $X_1, \dots, X_n$ are iid from a distribution with pdf $f_X(x \mid \theta)$. From now on we drop the subscript θ to simplify notation.

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

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

This inequality is strict unless $\tilde{\theta}$ is a function of T.

Remark. We have that $\hat{\theta}(T) = \int \tilde{\theta}(x) f_{X|T}(x \mid T)\mathrm{d}x$. By sufficiency of T, the conditional pdf does *not* depend on θ so $\hat{\theta}$ does not depend on θ, and is valid estimator.

Proof. By the tower property of expectation,

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

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

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

So

$$\mathrm{mse}(\tilde{\theta}) \geq \mathrm{mse}(\hat{\theta}).$$

Equality is achieved only when $\mathrm{Var}\left(\tilde{\theta} \mid T\right) = 0$ with probability 1 which requiers $\tilde{\theta}$ to be a function of T. □

Let's see an example of this. Suppose that $X_1, \ldots, X_n \sim \mathrm{Poisson}(\lambda)$ iid. Let $\theta = \mathbb{P}(X_1 = 0) = e^{-\lambda}$. Then

$$f_X(x \mid \theta) = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod_i x_i!} = \frac{\theta^n(-\log\theta)^{\sum x_i}}{\prod_i x_i!}.$$

By the factorisation criterion, $T(X) = \sum_i x_i$ is sufficient. Recall that $\sum x_i \sim \mathrm{Poisson}(n\lambda)$. Let $\tilde{\theta} = \mathbf{1}\{X_1 = 0\}$. Then

$$\begin{aligned}
\hat{\theta} = \mathbb{E}[\tilde{\theta} \mid T = t] &= \mathbb{P}\left(X_1 = 0 \mid \sum_{i=1}^n X_i = t\right) \\
&= \frac{\mathbb{P}\left(X_1 = 0, \sum_{i=2}^n X_i = t\right)}{\mathbb{P}\left(\sum_{i=1}^n X_i = t\right)} \\
&= \frac{\mathbb{P}\left(X_1 = 0\right)\mathbb{P}\left(\sum_{i=2}^n X_i = t\right)}{\mathbb{P}\left(\sum_{i=1}^n X_i = t\right)} \\
&= \frac{e^{-\lambda}e^{-(n-1)\lambda}\frac{((n-1)\lambda)^t}{t!}}{e^{-n\lambda}\frac{(n\lambda)^t}{t!}} = \left(\frac{n-1}{n}\right)^t
\end{aligned}$$

Hence $\hat{\theta} = \left(1 - \frac{1}{n}\right)^{\sum x_i}$ has $\mathrm{mse}(\hat{\theta}) < \mathrm{mse}(\tilde{\theta})$ for all θ. We can see that as $n \to \infty$, $\hat{\theta} \to e^{-\overline{X}} = e^{-\lambda} = \theta$.

Let $X_1, \dots, X_n \sim \text{Uniform}([0, \theta])$ and suppose we want to estimate $\theta \geq 0$. Last time we saw that $T = \max X_i$ is sufficient for θ. Let $\tilde{\theta} = 2X_1$ be an estimator (unbias). Then

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

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

1.5 Likelihood

Definition. (Likelihood) Let $X = (X_1, \dots X_n)$ have a joint pdf $f_X(x \mid \theta)$. The *likelihood* of θ is the function

$$L : \theta \to f_X(x \mid \theta).$$

The max likelihood estimator (MLE) is the value of θ maximizing L.

If $X_1, \dots, X_n \sim f_X(\cdot \mid \theta)$ iid, then $L(\theta) = \prod_{i=1}^n f_X(x_i \mid \theta)$.

It's usually easier to work with the log-likelihood, since this reduces to a sum. So in the iid case,

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

For example let $X_1, \dots, X_n \sim \text{Bernoulli}(p)$ iid. Then we get that

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

Taking the derivative with respect to p,

$$\frac{\partial \ell}{\partial p} = \frac{\sum_i X_i}{p} - \frac{n - \sum_i X_i}{1-p}.$$

So setting the derivative to zero we get that

$$p = \frac{\sum X_i}{n}.$$

Hence the MLE is

$$\hat{p} = \frac{\sum_i X_i}{n},$$

and since $\mathbb{E}[\hat{p}] = p$, this is unbiased.

Now suppose $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.

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

So

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

which is zero when $\mu = \frac{\sum_i X_i}{n}$ regardless of σ. Also

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

If we set $\mu = \frac{\sum_i X_i}{n}$ then we get $\frac{\partial \ell}{\partial \sigma^2} = 0$ if $\sigma^2 = \frac{1}{n}\sum(X_i - \bar{X})^2 = \frac{S_{xx}}{n}$. Hence the MLE is

$$(\hat{\mu}, \hat{\sigma}^2 = (\bar{X}, \frac{S_{xx}}{n}).$$

Notte that μ is unbiased, but we will see later that

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

So $\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}\left(\frac{S_{xx}}{n}\right) = \frac{n-1}{n}\sigma^2$. So $\hat{\sigma}^2$ is not unbiased, but is asymptotically unbiased as $n \to \infty$.

Suppose now that $X_1, \ldots, X_n \sim \text{Uniform}([0, \theta])$ iid. Then

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

Hence the MLE is $\hat{\theta}_{\text{MLE}} = \max_i X_i$. Recall that last time, we had an aunbiased estimator $\tilde{\theta}$ and by the Rao-Blackwell Theorem we found the estimator $\hat{\theta} = \mathbb{E}[\tilde{\theta} \mid T] = \frac{n+1}{n}\max_i X_i$. Note that $\hat{\theta}_{\text{MLE}} = \frac{n}{n+1}\hat{\theta}$, so $\mathbb{E}[\hat{\theta}_{\text{MLE}}] = \frac{n}{n+1}\mathbb{E}[\hat{\theta}] = \frac{n}{n+1}\theta$. Again this not unbiased, but is asymptotically unbiased.

Let's see some properties of the MLE.

(i) If T is a sufficient statistic, the MLE is a function of T. We can factorise $L(\theta) = g(T(x), \theta)h(x)$.

(ii) If $\phi = h(\theta)$ where h is a bijection, the MLE of ϕ is $\hat{\phi} = h(\hat{\theta})$ where $\hat{\theta}$ is the MLE of θ.

(iii) Asymptotic normality: $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta)$ is approximately normal with mean 0 for large n. The covariance matrix is the "smallest attainable" (see II Principles of Statistics).

1.6 Confidence intervals

Definition. (Confidence intervals) A $(100\gamma)\%$ confidience interval for a paramter θ is a random interval $(A(X), B(X))$ such that $\mathbb{P}(A(X) \leq \theta \leq B(X)) = \gamma$ for some $\gamma \in (0, 1)$ and all values of the true parameter θ.

Remark. The incorrect interperatation: Having observed $X = x$, there is a $1 - \gamma$ probability that θ is in $(A(X), B(X))$. This is wrong.

Suppose that $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$ iid. We want to find a 95% confidence interval for θ. We know that $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\theta, \frac{1}{n})$. If we define $Z = \sqrt{n}(\overline{X} - \theta)$ Then $Z \sim \mathcal{N}(0, 1)$ no matter

the value of θ. Let z_1, z_2 be numbers with $\Phi(z_1) - \Phi(z_2) = 0.95$ where Φ is the cdf of the standard normal. $\mathbb{P}\left(z_1 \leq \sqrt{n}(\overline{X} - \theta) \leq z_2\right) = 0.95$ rearranging we get that

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

hence

$$\left(\overline{X} - \frac{z_2}{\sqrt{n}}, \overline{X} - \frac{z_1}{\sqrt{n}}\right)$$

is a 95% confidence interval.

This is the recipe for confidence intervals.

(i) Find a quantity $R(X, \theta)$ such that this $\mathbb{P}_\theta$ distribution of $R(X, \theta)$ does not depend on θ. This is called a *pivot* for example $R(X, \theta) = \sqrt{n}(\overline{X} - \theta)$.

(ii) Write down the statement

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

where (c_1, c_2) are quantiles of the distribution of $R(X, \theta)$.

(iii) Rearranging the above to leave θ in the middle of the inequality, so we get something in the form

$$\mathbb{P}\left(A(X) \leq \theta \leq B(X)\right).$$

Remark. When θ is a vector, we talk about *confidence sets* rather than intervals.

Suppose that $X_1, \ldots, X_n \sim \mathcal{N}(0, \sigma^2)$ iid. We want a 95% confidence interval for σ^2. Note that

$$\frac{X_i}{\sigma} \sim \mathcal{N}(0, 1),$$

so $\sum_{i=1}^n \frac{X_i^2}{\sigma^2} \sim \chi_n^2$. Hence

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

is a *pivot*. Let $F^{-1}_{\chi_n^2}(0.025)$ and $F^{-1}_{\chi_n^2}(0.975)$. Then

$$\mathbb{P}\left(c_1 \leq \sum_{i=1}^n \frac{X_i^2}{\sigma^2} \leq c_2\right) = 0.95,$$

so rearranging we get that

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

gives our confidence interval.

Now suppose that $X_1, \ldots, X_n \sim$ Bernoulli(p) for large n. We will find an approximate 95% confidence interval for p. The maximium likelihood estimator of p is $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i$. By the central limit theorem $\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$ approximately. Thus

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

for large n. So we have our pivot, which gives

$$\mathbb{P}\left(-z_{0.025} \leq \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}} \leq z_{0.025}\right) \approx 0.95$$

Instead of inverting directly, if n is large $\hat{p} \approx p$, so switching p with $\hat{p}$ on the denominator we get that

$$\mathbb{P}\left(\hat{p} - z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 0.95$$

which gives our confidence interval. Since for all $\hat{p} \in [0,1]$ we have $\hat{p}(1-\hat{p}) \leq \frac{1}{4}$ we would also report a conservative confidence interval of $\hat{p} \pm z_{0.025}\sqrt{\frac{1}{4n}}$.

1.7 Bayesian estimation

So far we've been using frequentist methods treating $\theta \in \Theta$ as fixeed. For Bayesian methods, we treat θ as random with a prior distribution $\pi(\theta)$. Conditional on θ the data X has pdf $f_X(\cdot \mid \theta)$. Having observed that $X = x$, we combine with the prior to form a posterior distribution $\pi(\theta \mid X)$. By Bayes' rule,

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

where $f_X(x)$ is the marginal distribution of X, so

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

More simply,

$$\pi(\theta \mid x) \propto \pi(\theta) f_X(x \mid \theta).$$

Often is it easier to reconginise the RHS as propotional to a known distribution.

Remark. By the factorisation criterion, the posterior only depends on X through a sufficient statistic.

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

Suppose $\theta \in [0,1]$ is the mortality rate for some procedure at Addenbrooks. In the first 10 operations there are no deaths. In other hospitals across the country the mortality rate is between $3 - 20\%$, with average of 10%. Consider the prior distribution $\pi(\theta) \sim \text{Beta}(a,b)$ we can choose $(a,b) = (3,27)$ so $\pi(\theta)$ as mean 0.1 and $\pi(0.03 < \theta < 0.2) = 0.9$.

Let $X_i \sim \text{Bernoulli}(\theta)$ be indicator for whether ith patient at Addenbrookes dies.

$$f_X(x \mid \theta) = \theta^{\sim x_i}(1-\theta)^{n - \sum x_i}.$$

The posterior is

$$\begin{aligned} \pi(\theta \mid X) &\propto \pi(\theta) f_X(x \mid \theta) \\ &\propto \theta^{a-1}(1-\theta)^{b-1}\theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \\ &= \theta^{\sum x_i + a - 1}(1-\theta)^{b+n-\sum x_i - 1}. \end{aligned}$$

Hence

$$\pi(\theta \mid X) \sim \text{Beta}(a + \sum x_i, b + n - \sum x_i)$$

so pluggin in $a = 3, b = 27, n = 10, \sum X_i = 0$, so the posterior is $\text{Beta}(3, 37)$.

Remark. In the example the prior and posterior were from the same family of distributions known as conjugacy.

Supposewe put a $\text{Beta}(a, b)$ prior on the parameter θ of kidney cancer death rates in each county. We can estimate $(a, b) = (27, 58000)$ with $\frac{a}{a+b} \approx 4.65 \times 10^{-9}$ being the kidney cancer death rate in the United States. The previous example shows that if we observe $\sum_{i=1}^{n} X_i$ deaths in a county, the posterior mean estimate is $\frac{a+\sum X_i}{a+b-n}$. This is equal to

$$\frac{n}{a+b+n} \cdot \frac{\sum X_i}{n} + \frac{a+b}{a+b+n} \cdot \frac{a}{a+b}.$$

For large n, we use $\approx \frac{\sum X_i}{n}$ as our estimate, for small n we use $\frac{a}{a+b}$ and in between we shrink our estimate between them.

What is the use of the posterior distribution? This opens us to decision theory.

(i) We must pick a decision $\delta \in D$;

(ii) We have a loss function $L(\theta, \delta)$ which gives loss incurred in making decision δ when the true paramter value is θ.

(iii) Von-Neumann-Morgenstern Theorem: Under axioms of rational behaviour, pick δ that minimises expected loss under posterior.

Definition. (Bayes estimator) The *Bayes estimator* $\hat{\theta}^{(b)}$ is defined by

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

and

$$\hat{\theta}^{(b)} = \arg\min h(\delta)$$

Consider the case where we have quadartic loss, so $L(\theta, \delta) = (\theta - \delta)^2$. Then we have that

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

Differentiating with resepct to δ we get that $h'(\delta) = 0$ if

$$\int_{\Theta} (\theta - \delta)\pi(\theta \mid X)\mathrm{d}\theta = 0$$

so

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

is the posterior mean. Now suppose we have absolute loss, so $L(\theta, \delta) = |\theta - \delta|$. So

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

Taking derivatives and applying FTC we get that

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

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

$$\int_{-\infty}^{\delta} \pi(\theta \mid X)\mathrm{d}\theta = \int_{\delta}^{\infty} \pi(\theta \mid X)\mathrm{d}\theta$$

so $\hat{\theta}^{(b)}$ is the posterior median.

Supose we have $X_1, \ldots, X_n \overset{\text{idd}}{\sim} \mathcal{N}(\mu, 1)$ and prior $\pi(\mu)$ that is $\mathcal{N}(0, \frac{1}{\tau^2})$ for some known $\tau > 0$. Then

$$\begin{aligned} \pi(\mu \mid X) &\propto f_X(x \mid \mu)\pi(\mu) \\ &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2\right)\exp\left(-\frac{-\mu^2\tau^2}{2}\right) \\ &\propto \exp\left(-\frac{1}{2}(n+\tau^2)\left(\mu - \frac{\sum X_i}{n+\tau^2}\right)^2\right) \end{aligned}$$

This is the pdf of a $\mathcal{N}\left(\frac{\sum X_I}{n+\tau^2}, \frac{1}{n+\tau^2}\right)$ distribution. The posterior mean and median are both $\frac{\sum X_i}{n+\tau^2}$.

Definition. (Credible interval) A $100\gamma\%$ *credible interval* satifies that

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

2 Hypothesis Testing

Definition. (Hypothesis) A *hypothesis* is an assumption about a distribution of data X taking values in χ.

Definition. (Null/Alternative hypothesis) The *null hypothesis* H_0 is the base case. The *alternative hypothesis* is the positive or negative effect the interesting case, denoted by H_1.

For example let $X_1, \ldots, X_n \overset{\text{idd}}{\sim} \text{Bernoulli}(\theta)$. We may have the null hypothesis $H_0 : \theta = \frac{1}{2}$ and then make alternative hypothesis $H_1 : \theta = \frac{3}{4}$ or $H_1 : \theta \neq \frac{1}{2}$ for example.

Suppose that $X_1, \ldots, X_n$ are iid. Then we have the hypotheses:

$$H_0 : X_i \text{ has pdf } f_0$$
$$H_1 : X_i \text{ has pdf } f_1$$

This is called a goodness of fit test.

Now suppose that X has pdf $f(\cdot \mid \theta)$ for some $\theta \in \Theta$.

$$H_0 : \theta \in \Theta_0 \subsetneq \Theta$$
$$H_1 : \theta \notin \Theta_0$$

2.1 Simple hypotheses

Definition. (Simple/composite hypothesis) A *simple hypothesis* fully specifies the distribution of X. Otherwise we say the hypothesis is *composite*.

Definition. (Test and critical regions) A *test* of H_0 is defined by a *critical region*, C. When $X \in C$, we reject H_0, otherwise we do not reject H_1.

Definition. (Type I Error) A *Type I Error* occurs when we reject H_0 when H_0 is true.

Definition. (Type II Error) A *Type II Error* occurs when we fail to reject H_0 when H_1 is true.

When H_0 and H_1 are simple hypotheses we have the following.

Definition. (Size) We define α as the *size* of the test, defined as

$$\alpha = \mathbb{P}_{H_0}(H_0 \text{ rejected}) = \mathbb{P}_{H_0}(X \in C).$$

Definition. (Power) We define the *power* of the test as $1 - \beta$ where

$$\beta = \mathbb{P}_{H_1}(H_0 \text{ not rejected}) = \mathbb{P}_{H_1}(X \notin C).$$

Remark. Note that α is the probability of a Type I error and β is the probability of a Type II error.

Remark. Type I and Type II errors correspond to a false positive and a false negative respectively.

Usually we set α at an acceptable level for example 1%, and choose a test that minimises β subject to $\alpha \leq 1\%$.

Definition. (Likelihood ratio statistic) Let H_0 and H_1 be simple hypotheses with X having pdf f_i under H_i. The *likelihood ratio statistic* is

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

Definition. (Likelihood ratio test) A *Likelihood ratio test* (LRT) rejects H_0 when $X \in C = \{x \in \Lambda_X(H_0, H_1) > k\}$ for some $k > 0$.

Theorem. (Neyman-Pearson Lemma) Suppose that f_0 and f_1 are nonzero on the same sets and $\exists k$ such that the LRT with critical region $C = \{x : \frac{f_1(x)}{f_0(x)} > k\}$ has size α. Out of all tests with size $\leq \alpha$ the LRT is the test with smallest β.

Proof. Let $\overline{C}$ be the complement of C. Then

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

Let C^* be the critical region of another test of size $\alpha^* \leq \alpha$. We want to show that $\beta \leq \beta^*$.

$$\begin{aligned}
\beta - \beta^* &= \int_{\overline{C}} f_1(x)\mathrm{d}x - \int_{\overline{C}^*} f_1(x)\mathrm{d}x \\
&= \int_{\overline{C}\cap\overline{C}^*} f_1(x)\mathrm{d}x - \int_{\overline{C}^*\cap C} f_1(x)\mathrm{d}x \\
&= \int_{\overline{C}\cap C^*} \frac{f_1(x)}{f_0(x)} f_0(x)\mathrm{d}x - \int_{\overline{C}^*\cap C} \frac{f_1(x)}{f_0(x)} f_0(x)\mathrm{d}x
\end{aligned}$$

and the result follows from algebra manipulation. □

Remark. A LRT with size α for any given α doesn't always exist. However we can always define a "randomised" test with exact level α.

Suppose $X_1, \ldots, X_n \overset{\text{idd}}{\sim} \mathcal{N}(\mu, \sigma_0^2$ where σ_0^2 is known. We want to find the best size α test for

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu = \mu_1$$

for some fixed $\mu_1 > \mu_0$. We have that

$$\begin{aligned}
\Lambda_X(H_0; H_1) &= \frac{(2\pi\sigma_0)^{-n/2} \exp\left(-\frac{1}{2\pi\sigma_0^2}\sum_{i=1}^n (X_i - \mu_1)^2\right)}{(2\pi\sigma_0)^{-n/2} \exp\left(-\frac{1}{2\pi\sigma_0^2}\sum_{i=1}^n (X_i - \mu_0)^2\right)} \\
&= \exp\left(\frac{\mu_1 - \mu_0}{\sigma_0^2} n\overline{X} + \frac{n(\mu_0^2 - \mu_1^2)}{2\sigma_0^2}.\right.
\end{aligned}$$

Since $\Lambda_X(H_0; H_1)$ is monotone in $\overline{X}$. We can depend our critical region for the LRT on $\overline{X}$ equivalently. If we define

$$z = \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma_0}$$

the rejection region is now $\{z > c'\}$ for some c'. Under H_0 we have that $Z \sim \mathcal{N}(0, 1)$ so the test that rejects H_0 when $\{x : z > \Phi^{-1}(1 - \alpha)\}$ has size α. This is called a z-test.

Definition. (p-value) For any test with critical region of the form $\{x : T(x) > k\}$ where T is some statistic, we usually report the *p-value*

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

where x^* is the observed data.

This is the probability of observing "more extreme" data under H_0.

Proposition. Uner H_0 the p-value is Uniform$[0, 1]$.

Proof. Let F be the distribution funciton of T. Then

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

Definition. (Acceptance region) The *acceptance region* of a test is the complement of the critical region.

Let $X \sim f_X(\cdot \mid \theta)$ for some $\theta \in \Theta$.

Theorem. (i) Suppose that for each $\theta_0 \in \Theta$ there exists a test of $H_0 : \theta = \theta_0$ of size α with acceptance region $A(\theta_0)$. Then the set $I(X) = \{\theta : X \in A(\theta)\}$ is a $100(1-\alpha)\%$ confidence set.

(ii) Suppose that $I(X)$ is a $100(1 - \alpha)\%$ confidence set for θ. Then

$$A(\theta_0) = \{x : \theta_0 \in I(X)\}$$

is the acceptance region of a size α test for $H_0 : \theta = \theta_0$ for each $\theta \in \Theta$.

Proof. In both cases $\theta_0 \in I(X) \iff X \in A(\theta_0)$.

(i) We want to show that $\mathbb{P}_{\theta_0}(\theta_0 \in I(X)) = 1 - \alpha$.

$$\mathbb{P}_{\theta_0}(X \in A(\theta_0)) = \mathbb{P}_{\theta_0}(\text{do not reject } H_0) = 1 - \alpha$$

(ii) We want to show that $\mathbb{P}_{\theta_0}(X \notin A(\theta_0)) = \alpha$.

$$\mathbb{P}_{\theta_0}(X \notin A(\theta_0)) = \mathbb{P}_{\theta_0}(\theta_0 \notin I(X)) = \alpha.$$

Hence we're done. □

Suppose that $X = (X_1, \dots, X_n) \overset{\text{idd}}{\sim} \mathcal{N}(\mu, \sigma_0^2)$ where σ_0 is known. We foudn a $100(1-\alpha)\%$ confidence interval for μ, namely, $I(X) = \overline{X} \pm \frac{\sigma_0}{\sqrt{n}}$. Using the second part of the theorem, we can find a size α test for $H_0 : \mu = \mu_0$, by defining the acceptance region as

$$A(\mu_0 \in I(x)_0) = \{x : \mu_0 \in \left[\overline{X} - \frac{z_{\alpha/2}\sigma_0}{\sqrt{n}}, \overline{X} + \frac{z_{\alpha/2}\sigma_0}{\sqrt{n}}\right]\}.$$

Equivalently we reject H_0 when

$$z_{\alpha/2} < \frac{\sqrt{n}|\mu_0 - \overline{X}|}{\sigma_0}.$$

This is a two sided LRT. We could also equivalently go in the opposite direction.

2.2 Composite hypotheses

Previously we considered H_0 and H_1 as *simple* hypotheses with error porbabilities

$$\alpha = \mathbb{P}_{H_0}(X \in C), \qquad \beta = \mathbb{P}_{H_1}(X \notin C).$$

Now we consider $X \sim f_X(\cdot \mid \theta)$, with $\theta \in \Theta$ and

$$H_0 : \theta \in \Theta_0 \subseteq \Theta$$
$$H_1 : \theta \in \Theta_1 \subseteq \Theta$$

Definition. (Power function) The *power function* is $W(\theta) = \mathbb{P}_\theta(X \in C)$.

Definition. (Size) The *size* of a test with composite null H_0 is the worst-case Type I error probability, so

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

Definition. (Uniformly most powerful) We say that a test of H_0 against H_1 is *uniformly most powerful* (UMP) of size α if

(i) $\sup_{\theta \in \Theta_0} W(\theta) \leq \alpha$;

(ii) For any other test of size α, with power function W^* we have that

$$W(\theta) \geq W^*(\theta) \quad \forall \theta \in \Theta_1.$$

Remark. A UMP test might not exist. However many LRTs are UMP.

Let's see an example for a one-sided test for a normal location. Suppose that $X_1, \dots, X_n \overset{\text{idd}}{\sim} \mathcal{N}(\mu, \sigma_0^2)$ with σ_0^2 known. We wish to test that

$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0.$$

Recall that for the simple hypotheses,

$$H_0' : \mu = \mu_0$$
$$H_1' : \mu = \mu_1,$$

the LRT had critical region $C : \left\{x : z = \frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma_0} > z_\alpha\right\}$. We will show that the same test is UMP for H_0 against H_1. The power function is

$$\begin{aligned} W(\mu) &= \mathbb{P}_\mu(\text{reject } H_0) \\ &= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma_0} < z_\alpha\right) \\ &= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma_0} > z_\alpha + \frac{\sqrt{n}(\mu_0-\mu)}{\sigma_0}\right) \\ &= 1 - \Phi\left(z_\alpha + \frac{\sqrt{n}(\mu_0-\mu)}{\sigma_0}\right) \end{aligned}$$

Thus $\sup_{\mu\in\Theta_0} W(\mu) = \alpha$, so the test has size α. Now consider any other size $\leq \alpha$ with power function W^*. We want to show that $W(\mu_1) \geq W^*(\mu_1)$ for all $\mu_1 \in \Theta_1$. Any other test of size $\leq \alpha$ also has size $\leq \alpha$ for H_0' against H_1 since

$$W^*(\mu_0) \leq \sup_{\mu\in\Theta_0} W^*(\mu) \leq \alpha.$$

Thus by Neyman-Pearson, $W(\mu_1) \geq W^*(\mu_1)$. Since this argument works for any $\mu_1 > \mu_0$ we are done.

2.2.1 Generalised likelihood ratio tests

The generalised likelihood ratio (GLR) statistic for

$$H_0 : \theta \in \Theta_0 \subseteq \Theta$$
$$H_1 : \theta \in \Theta_1 \subseteq \Theta$$

as

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

We reject H_0 when Λ_X is large.

Suppose that $X_1, \ldots, X_n\mathcal{N}(\mu, \sigma_0^2)$ with σ_0^2 known. We wish to test

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0.$$

Thus, $\Theta_0 = \{\mu_0\}$ and $\Theta_1 = \mathbb{R} \setminus \{\mu_0\}$, and the GLR statistic is

$$\Lambda_X(H_0; H_1) = \frac{(2\pi\sigma_0^2)^{-n/2} \exp\left(-\frac{1}{2\sigma_0^2}\sum(X_i - \overline{X})^2\right)}{(2\pi\sigma_0^2)^{-n/2} \exp\left(-\frac{1}{2\sigma_0^2}\sum(X_i - \mu_0)^2\right)}$$

and after simplication we get that

$$2\log\Lambda_X(H_0;H_1) = \frac{1}{\sigma_0^2}\left(\sum(X_i-\mu_0)^2 - \sum(X_i-\overline{X})^2\right) = \frac{n}{\sigma_0^2}(\overline{X}-\mu_0)^2.$$

Thus GLRT rejects H_0 if $\frac{\sqrt{n}|\overline{X}-\mu_0|}{\sigma_0}$ is large. Under H_0, $\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma_0} \sim \mathcal{N}(0,1)$ so a test of size α rejects H_0 if $\frac{\sqrt{n}|\overline{X}-\mu|}{\sigma_0} > z_{\alpha/2}$.

Note that in this example $2\log\Lambda_X(H_0;H_1) = \frac{n(\overline{X}-\mu_0)^2}{\sigma_0^2} \sim \chi_1^2$. Thus the critical region of t he GLRT can also be written as $\left\{x : \frac{\sqrt{n}(\overline{X}-\mu_0)^2}{\sigma_0^2} > \chi_1^2(\alpha)\right\}$. In a fact a more general result says that $2\log\Lambda_X(H_0;H_1) \approx \chi^2$ when n is large.

Theorem. (Wilks' theorem) Suppose the parameter θ is k-dimensional so $\theta = (\theta_1,\ldots,\theta_k)$. The *dimension* of a hypothesis $H_0 : \theta \in \Theta_0$ is the number of "free parameters in Θ_0 for example

(i) If $\Theta_0 = \{\theta \in \mathbb{R}^k : \theta_1 = \theta_2 = \cdots = \theta_p = 0\}$ then $\dim(\Theta_0) = k-p$.

(ii) Suppose $\Theta_0 = \{\theta \in \mathbb{R}^k : \theta_i = f_i(\phi),\ \forall 1 \le i \le k$, for some $\phi \in \mathbb{R}^{k-p}\}$. Then $\dim(\Theta_0) = k-p$.

Proof. Not included.

Theorem. Suppose that $\Theta_0 \subseteq \Theta_1$ and $\dim(\Theta_1) - \dim(\Theta_0) = p$. If $(X_1,\ldots,X_n)$ are idd then under regularity conditions, as $n \to \infty$ the limiting distribution of $2\log\Lambda_X(H_0;H_1)$ is χ_p^2. Thus if we reject H_0 when $2\log\Lambda_X(H_0;H_1) \ge \chi_p^2(\alpha)$ and we have a test of size $\approx \alpha$.

Example. In the two-sided normal mean test, we had

$$\Theta_0 = \{\mu_0\}$$
$$\Theta_1 = \mathbb{R} \setminus \{\mu_0\}$$

and we saw that

$$2\log\Lambda_X(H_0;H_1) \sim \chi_1^2$$

exactly. In a different parametric family, with large n this would hold aproximately.

2.3 Goodness-of-fit tests

Example. In one of this experiments, Mendel crossed 556 smooth, yellow, male peas with wrinkled green female peas. He obtained a table of data of the phenotypes of the produced peas. Is there evidence in his data to reject the hypothesis that the genetic theory is correct?

Suppose $X_1,\ldots,X_n$ are iid samples from a distribution on $\{1,\ldots,k$. Let $p_i = \mathbb{P}(X_1 = i)$ and let N_i be the number of observations in $\{X_1,\ldots,X_n\}$ of type i. Hence $\sum p_i = 1$ and $\sum N_i = n$. Then we have that

$$(N_1,\ldots,N_k) \sim \text{Mult}(n;p_1,p_2,\ldots,p_n).$$

The likelihood is $L(p) \propto p_1^{N_1}p_2^{N_2}\ldots p_k^{N_k}$, so

$$\ell(p) = \log(L(p)) = \text{const.} + \sum_{i=1}^{k} N_i \log p_i.$$

We can test H_0 against H_1 using a GLRT

$$2\log\Lambda = 2\left(\sup_{p\in\Theta_1}\ell(p) - \sup_{p\in\Theta_0}\ell(p)\right).$$

Notice that

$$\sup_{p\in\Theta_1}\ell(p) = \sup_{p:\sum p_i=1}\sum_i N_i\log p_i.$$

Using the Lagrangian and calculating we get that the MLE is giving by

$$\hat{p} = \hat{p}_i = \frac{N_i}{n}.$$

So

$$2\log\Lambda = 2\sum_{i=1}^{k} N_i\log\left(\frac{N_i}{n\tilde{p}_i}\right).$$

Wilks' theorem says that $2\log\Lambda \approx \chi^2_p$ with $p = \dim(\Theta_1) - \dim(\Theta_0) = k-1-0 = k-1$. Thus reject H_0 if $2\log\Lambda \geq \chi)k - 1^2(\alpha)$. It is common to write

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

where o_i is N_i and e_i is the expected number of type i equal to $i = n\tilde{p}_i$.

Let $\delta_i = o_i - e_i$. Then

$$2\log\Lambda \approx \sum_i \frac{\delta_i^2}{e_i} = \sum_i \frac{(o_i-e_i)^2}{e_i}$$

which is Pearson's chi-square statistic.

Suppose we wish to test

$$H_0 : p_1 = \theta^2,\ p_2 = 2\theta(1-\theta),\ p_3 = (1-\theta)^2,\ \theta\in[0,1].$$

Then the GLRT is

$$\begin{aligned}
2\log\Lambda &= 2\left(\sup_{p:\sum p_i=1}\ell(p) - \sup_\theta \ell(p(\theta))\right)\\
&= 2(\ell(\hat{p}) - \ell(p(\hat{\theta}))) = 2\sum N_i\log\left(\frac{N_i}{np_i(\hat{\theta})}\right)\\
&= 2\sum_i o_i\log\left(\frac{o_i}{e_i}\right).
\end{aligned}$$

2.4 Contingency tables

2.4.1 Testing independence

Suppose we have $(X_1,Y_1),\ldots,(X_n,Y_n)$ are iid with X_i taking values in $\{1,\ldots,r\}$ and Y_i taking values on $\{1,\ldots,c\}$. We wish to test the null hypothesis that X_i and Y_i are independent. The entries in a contingency table are

$$N_{ij} = \{\ell : 1\leq\ell\leq n; (X_\ell,Y_\ell) = (i,j)\}.$$

Example. Part III admissions statistics from a recent year, by gender/stream

Stream	F	M	Other
MASA	10	9	1
MASTH	11	59	0
MASP	10	56	2
MASS	6	23	0
MMATH	13	87	1

We wish to test the null hypothesiss that gender and stream are independent. Observe n examples (fixed, given). A sample is of type (i, j) with probability p_{ij}. Then $(N_{11}, N_{12}, \ldots, N_{21}, \ldots, N_{rc}) \sim \text{Mult}(n, p_{11}, p_{12}, \ldots, p_{rc})$. Then let $p_{i+} = \sum p_{ij}$ and $p_{+j} = \sum_i p_{ij}$ so

$$H_0 : p_{ij} = p_{i+}p_{+j}$$
$$H_1 : \{p_{ij}\} \text{ unconstrained.}$$

Using Lagrangian methods, the MLEs are $\hat{p}_{ij} = \frac{N_{ij}}{n}$ under H_1 and $(\hat{p}_{i+}, \hat{p}_{+j}) = \left(\frac{N_{i+}}{n}, \frac{N_{+j}}{n}\right)$ under H_0. Then the GLR statistic is

$$2 \log \Lambda = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij} \log \left(\frac{\hat{p}_{ij}}{\hat{p}_{i+}\hat{p}_{+j}} \right).$$

Writing $o_{ij} = N_{ij}$ and $e_{ij} = h\hat{p}_{i+}\hat{p}_{+j}$ we have that

$$2 \log \Lambda = 2 \sum_i \sum_j o_{ij} \log \left(\frac{o_{ij}}{e_{ij}} \right) \approx \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

which converges to χ^2 by Wilks' theorem. The number of degrees of freedom is $p = \dim(\Theta_1) - \dim(\Theta_0) = (rc - 1) - ((r - 1) + (c - 1)) = (r - 1)(c - 1)$.

Applying this to our sample we get that

$$\sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \approx 22.2.$$

Hence our p-value is $\mathbb{P}(\chi^2_8 \geq 22.2) \approx 0.005$, so reject the null hypothesis of indepedence.

Example. MMATH admissions statistic by gender/year

	F	M	Other
Year 1	13	87	1
Year 2	10	81	0
Year 3	15	87	4
Year 4	9	103	0
Year 5	16	126	0

If we have hypotheses

$$H_0 : \text{gender and year and independent}$$
$$H_1 : \text{gender and year are not independent}$$

then our test statistic is

$$\sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \approx 15.5.$$

And our p-value is $\mathbb{P}(\chi^2_8 \geq 15.5) = 0.051$ so we do not reject H_0 at the 5% significance level.

2.4.2 Tests of homogeneity

Example. Suppose that 150 patients are randomly allocated to three groups of equal size. Two sets of patients were given a drug of two different doses. The third group were given a placebo. This is different from the last setting since the row totals are fixed. We want to test the null hypothesis of "homogeneity" which corresponds to the probability of response being unchanged between groups.

Our probability model is $N_{i1}, \dots, N_{ic} \sim \text{Mult}(n_{i+}; p_{i1}, \dots, p_{ic})$ independently for each $1 \le i \le r$ where n_{i+}'s are fixed row totals.

$$H_0 : p_{ij} = p_{2j} = \cdots = p_{rj}$$
$$H_1 : (p_{i1}, \dots, p_{ic}) \text{ is an arbitrary probability distribution}$$

Under H_1,

$$L(p) = \prod_{i=1}^{r} \frac{n_{i+}!}{N_{i1}! \dots N_{ic}!} p_{i1}^{N_{i1}} \cdots p_{ic}^{N_{ic}}$$

and

$$\ell(p) = \log L(p) = \text{const.} + \sum_{i,j} N_{ij} \log p_{ij}.$$

Ynder Lagrangian methods with constraints $\sum_j p_{ij} = 1$ we find the MLE

$$\hat{p}_{ij} = \frac{N_{ij}}{n_{i+}}.$$

Under H_0 let $p_{ij} = p_j$ for all i. Then

$$\ell(p) = \text{const.} + \sum_{i,j} N_{ij} \log p_j = \text{const.} + \sum_{j=1}^{c} N_{+j} \log p_j.$$

Using the Lagrangian method with constraint $\sum_j p_j = 1$ we have the MLE $\hat{p}_j = \frac{N_{+j}}{n_{++}}$. Hence

$$2 \log \Lambda = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij} \log \left(\frac{\hat{p}_{ij}}{\hat{p}_j} \right) = 2 \sum_{i,j} N_{ij} \log \left(\frac{N_{ij}}{n_{i+} n_{+j} / n_{++}} \right)$$

If we define $o_{ij} = N_{ij}$ and $e_{ij} = \frac{n_{i+} N_{+j}}{n_{++}}$, then

$$2 \log \Lambda \approx \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}.$$

And Wilks' theorem states that asympototically the distribution of $2 \log \Lambda$ is χ^2_p with

$$p = \dim(\Theta_1) - \dim(\Theta_0) = r(c-1) - (c-1) = (r-1)(c-1).$$

Remark. The number of degrees of freedom is the same for both tests of homogeneity and independence.

2.5 Multivariate normal theory

Let $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ be a vector of random variables. Recall that

$$\mathbb{E}(X) = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_n) \end{pmatrix} \quad \text{and} \quad \operatorname{Cov}(X) = \mathbb{E}[(X - \mathbb{E}(X))(X - E(X))^T] = (\operatorname{Cov}(X_i, X_j))_{ij}.$$

Furthermore if $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$,

$$\mathbb{E}(AX + b) = A\mathbb{E}(X) + b \quad \text{and} \quad \operatorname{Cov}(AX + b) = A \operatorname{Cov}(X) A^T.$$

Definition. (Multivariate normal) We say that X has a multivariate normal distribution if for any $t \in \mathbb{R}^n$ the random variable $t^T X$ has a normal distribution.

Proposition. If X is multivariate normal then $AX + b$ is multivariate normal.

Proof. Take any $t \in \mathbb{R}^m$. Them $t^T(AX + b) = (A^T t)^T X + t^T b$. Since X is multivariate normal, $(A^T t)^T X \sim \mathcal{N}(\mu, \sigma^2)$ for some (μ, σ^2) then $t^T(AX + b) \sim \mathcal{N}(\mu + t^T b, \sigma^2)$. □