Statistics *

Zhiyuan Bai

Compiled on June 21, 2021

This document serves as a set of revision materials for the Cambridge Mathematical Tripos Part IB course *Statistics* in Lent 2021. However, despite its primary focus, readers should note that it is NOT a verbatim recall of the lectures, since the author might have made further amendments in the content. Therefore, there should always be provisions for errors and typos while this material is being used.

Contents

0 Introduction **2**
0.1 Overview . . . 2
0.2 Probability Review . . . 2

1 Estimation **6**
1.1 Bias and Error . . . 6
1.2 Sufficiency . . . 8
1.3 Maximum Likelihood Estimation . . . 11
1.4 Confidence Interval . . . 12

2 Bayesian Analysis **14**

3 Testing Hypotheses **17**
3.1 Simple Hypotheses . . . 18
3.2 Composite Hypotheses . . . 19
3.3 Goodness-of-fit Tests . . . 21
3.4 Independence . . . 23
3.5 Tests for Homogeneity . . . 25
3.6 Tests and Confidence Intervals . . . 26

4 Multivariate Normals **26**
4.1 Definition and Properties . . . 26
4.2 The (Normal) Linear Model . . . 28
4.3 Two Useful Distributions . . . 31
4.4 Inferences in the Normal Linear Model . . . 31
4.5 Hypothesis Testing . . . 33
4.6 Applications of Normal Linear Model . . . 35

*Based on the lectures under the same name taught by Dr. S. Bacallado in Lent 2021.

0 Introduction

0.1 Overview

What is Statistics? A common modern definition of statistics is to say it is the science of data analysis, but this doesn't really work since it is somewhat circular ("what is data analysis?"). A definitely correct, but useless due to its boardness, definition is that it is the science of learning about the world. Well, it doesn't really tell you anything, does it? A definition that somewhat works well is that it is the science of making informed decisions. It obviously isn't perfect, but it is nice enough.
Statistics can include the design of experiments and studies, graphical exploration of data and informal summaries, formal statistical inference, clear communication of conclusions and uncertainty, and formal decision theory. This course mainly focus on formal statistical inference, more precisely, parametric inference.
Let $X_1, \dots, X_n$ be i.i.d. random variables taking values in some space. In the cases we are interested, we assume the distribution of X_1 belongs to some particular family with parameter $\theta \in \Theta$. For example, $X_1 \sim \text{Poi}(\mu)$ for $\theta = \mu \in \Theta = (0, \infty)$ or $X_1 \sim N(\mu, \sigma^2)$ for $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$.
Let $X = (X_1, \dots, X_n)$, what we want to do is to use the observed phenomena $X = x$ to make inferences about θ. For example, we may attempt to give an estimate $\hat{\theta}(x)$ of the true value of θ, or to produce an interval $(\hat{\theta}_1(x), \hat{\theta}_2(x))$ that contains θ with high probability. Another thing we might do is to test hypotheses about θ, e.g. testing the hypothesis $H : \theta = 0$ so as to determine whether there is probabilistic/statistical evidences against it.
In general, we assume the family of distrbutions is known, but the parameter is unknown. However, some of the results we will explore in this course will not depends on which specific distributions are we using, as we usually only need weaker assumptions (e.g. finite mean and variance).
Of course, statistics is a very highly applicable course. It can be applied in market research, opinion polls, epidemiology, clinical trials, environmental statistics, traffic studies, finance, insurance, science, agriculture, official statistics, sport, machine learning, automation, etc..
What we mainly deal with in this course are developed before 1950. The emergence of computer completely revolutionised the way to approach statistics in recent decades, but of course, we need some basics in order to go there.

0.2 Probability Review

Please, skip this.
Let Ω be the sample space of outcomes of an experiment and $\mathcal{F} \subset 2^\Omega$ the space of events (which consists of subsets that is closed under certain operations that you'd expect – see Probability & Measure for details).

Definition 0.1. A function $\mathbb{P} : \mathcal{F} \to [0, 1]$ is called a probability measure if:
1. $\mathbb{P}(\varnothing) = 0$.
2. $\mathbb{P}(\Omega) = 1$.
3. For a sequence $(A_i)_{i=1}^\infty$ of disjoint events, $\mathbb{P}(\bigcup_i A_i) = \sum_i \mathbb{P}(A_i)$.

Definition 0.2. A random variable is a (measurable – see Probability & Measure) function $X : \Omega \to \mathbb{R}$.

Example 0.1. If we toss two coins, then we can take $\Omega = \{HH, HT, TH, TT\}$ and $\mathcal{F} = 2^\Omega$. We can define a random variable X by assigning it as the number of heads.

Definition 0.3. The distribution function of a random variable X is $F_X(x) = \mathbb{P}(X \leq x)$.
A random variable X is discrete if it takes values in a discrete set. Its probability mass function is $p_X(x) = \mathbb{P}(X = x)$.
A random variable X is continuous if it has a probability density function f_X suhc that

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

for "nice enough" (measurable) A.

Definition 0.4. The expectation $\mathbb{E}X$ of a discrete random variable X is

$$\sum_x x p_X(x)$$

And that of a continuous random variable X is

$$\int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x$$

Easily the expectation is a linear operator.
Note that for a nice enough function g, $g(X) = g \circ X$ is also a random variable, so we can define

Definition 0.5. The variance of X is $\mathrm{var}(X) = \mathbb{E}((X - \mathbb{E}X)^2)$.

Definition 0.6. We say $X_1, \ldots, X_n$ are independent if for all $x_1, \ldots, x_n$,

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

If $X_1, \ldots, X_n$ are continuous and have pdf's $f_{X_1}, \ldots, f_{X_n}$, then this is saying

$$f_X(x) = f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = \prod_{i=1}^n f_{X_i}(x_i)$$

If $Y = \max\{X_1, \ldots, X_n\}$ with $X_1, \ldots, X_n$ independent, then

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X_1, \leq y, \ldots, X_n \leq y) = \prod_{i=1}^n F_{X_i}(y)$$

Definition 0.7. The covariance of random variables X, Y is

$$\mathrm{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y))$$

Then easily $\mathrm{var}(X) = \mathrm{cov}(X, X)$ and for random variables $X_1, \ldots, X_n$ and real $a_1, \ldots, a_n$

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

Many of these rules can be simplified in vector notation, where we have

$$\mathbb{E}(a^\top X) = a^\top \mathbb{E}X, \operatorname{var}(a^\top X) = a^\top (\operatorname{var} X) a$$

where $(\operatorname{var} X)_{i,j} = \operatorname{cov}(X_i, X_j)$.
Let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}X_1 = \mu, \operatorname{var} X_1 = \sigma^2$. The sample mean is defined as

$$\bar{X}_n = \frac{S_n}{n}, S_n = \sum_{i=1}^{n} X_i$$

Then by linearity $\mathbb{E}\bar{X}_n = \mu, \operatorname{var} \bar{X}_n = \sigma^2/n$. The standardised statistic is then defined as

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$$

which has mean 0 and variance 1.

Definition 0.8. The moment generating function of a random variable X is $M_X(t) = \mathbb{E}(e^{tX})$ when it exists.

Mostly, we only need it to exist in a neighbourhood of 0. Then easily we can recover the moments from the mgf:

$$\mathbb{E}(X^n) = \left.\frac{\mathrm{d}^n M_X}{\mathrm{d}t^n}\right|_{t=0}$$

Under very board conditions, the moment generating function, if exists, uniquely determines the random variable. This is equivalent to saying that the Laplace transform is invertible, which is something that will not be covered in this course. Assuming this, then we can use the mgf to find the distribution of sums of some independent random variables.

Example 0.2. Let $X_1, \ldots, X_n \sim \operatorname{Poi}(\mu)$ be i.i.d., then

$$M_{X_i}(t) = \mathbb{E}e^{tX_i} = \sum_{k=0}^{\infty} e^{tk} e^{-\mu} \frac{\mu^k}{k!} = e^{\mu(1-e^t)}$$

and

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

which is the mgf of $\operatorname{Poi}(n\mu)$.

Theorem 0.1 (Weak Law of Large Numbers (WLLN))**.**

$$\forall \epsilon > 0, \lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \to 0$$

Theorem 0.2 (Strong Law of Large Numbers (SLLN))**.**

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

Theorem 0.3 (Central Limit Theorem)**.** *If the moments of X_1 exists up to 4^{th} order, then the standard statistic Z_n has $\mathbb{P}(Z_n \le z) \to \Phi(z)$ as $n \to \infty$ where $\Phi(z)$ is the distribution function of $N(0,1)$.*

Definition 0.9. Let X, Y be discrete random variables, then their joint pmf is $p_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$. The marginal pmf of X is $p_X(x) = \mathbb{P}(X = x) = \sum_y P_{X,Y}(x,y)$. The conditional pmf of X given $Y = y$ is then defined as

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

By convention, this equals 0 in the degenerate case $p_Y(y) = 0$.

We can do the same thing for continuous random variables.

Definition 0.10. Let X, Y be continuous, the joint pdf $f_{X,Y}$ satisfies

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

The marginal pdf of Y is

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

The conditonal pdf of X given Y is

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

which is again 0 if $f_Y(y) = 0$.

Definition 0.11. The conditional expectation of X given Y is

$$\mathbb{E}(X|Y) = \sum_x p_{X|Y}(x|Y)$$

in the discrete case and

$$\int_{-\infty}^{\infty} x f_{X|Y}(x|Y)\,\mathrm{d}x$$

in the continuous case.

Notably, the conditional expectation is a random variable itself.

Proposition 0.4 (Tower Property)**.**

$$\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}X$$

Proof. Exercise. □

Knowing this,

$$\begin{aligned}
\mathrm{var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}X)^2 \\
&= \mathbb{E}(\mathbb{E}(X^2|Y)) - \mathbb{E}(\mathbb{E}(X|Y))^2 \\
&= \mathbb{E}(\mathbb{E}(X^2|Y) - \mathbb{E}(X|Y)^2) + \mathbb{E}(\mathbb{E}(X|Y)^2) - \mathbb{E}(\mathbb{E}(X|Y))^2 \\
&= \mathbb{E}(\mathrm{var}(X|Y)) + \mathrm{var}(\mathbb{E}(X|Y))
\end{aligned}$$

If we do a change in variables, say $(x, y) \to (u, v)$, via smooth bijections, then we have the change-of-variable formula

$$f_{U,V}(u,v) = f_{X,Y}(x(u,v), y(u,v)) \left| \det \begin{pmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{pmatrix} \right|$$

We have a list of common distributions. All of them (especially the discrete ones) has a special interpretation. $\text{Bin}(n, p)$ is the number of success in n independent $\text{Ber}(p)$ trials, and $\text{Multi}(n; p_1, \ldots, p_k)$ is the vector of the number of results for each of the k types with respective probabilities $p_1, \ldots, p_k$ in n independent trials. $\text{NegBin}(k, p)$ is the time where the k^{th} success occurs in i.i.d. $\text{Ber}(p)$ trials (with the special case $\text{NegBin}(1, p) = \text{Geom}(p)$). $\text{Poi}(\lambda)$ is the limit of $\text{Bin}(n, \lambda/n)$ as $n \to \infty$.
For the continuous distributions, it is lengthy to describe the rationale behind some of them. But they usually have good properties. If $X_i \sim \Gamma(\alpha_i, \lambda)$, then $S_n = X_1 + \ldots + X_n$ has distribution $\Gamma(\sum_i \alpha_i, \lambda)$ by looking at the mgf. Also if $X \sim \Gamma(\alpha, \lambda)$, then for any $b > 0$ we have $bX \sim \Gamma(\alpha, \lambda b^{-1})$, so λ is a scaling parameter. The exponential distribution is $\text{Exp}(\lambda) = \Gamma(1, \lambda)$. Notably, the time seperating successive events in a Poisson process with rate λ are i.i.d. $\text{Exp}(\lambda)$. The χ^2 distribution with k degrees of freedom $\chi^2_k = \Gamma(k/2, 1/2)$ is the sum of k independent squared $N(0, 1)$.

1 Estimation

1.1 Bias and Error

Suppose we observe data $X_1, \ldots, X_n$ which are i.i.d. with pdf (or pmf) $f_X(x|\theta)$ with an unknown parameter θ. Let $X = (X_1, \ldots, X_n)$.

Definition 1.1. An estimator is a statistic (a function of the data $\hat{\theta} = T(X)$) which we use to approximate the true parameter θ. The distribution of $T(X)$ is called its sampling distribution.

Example 1.1. Suppose $X_1, \ldots, X_n \sim N(\mu, 1)$ are i.i.d., then we can estimate μ by

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

The sampling distribution of $\hat{\mu}$ is $T(X) \sim N(\mu, 1/n)$.

Definition 1.2. The bias of $\hat{\theta} = T(X)$ is $\text{bias}(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta}) - \theta$ where $\mathbb{E}_\theta$ is the expectation in the model where $X_i \sim f_X(\cdot|\theta)$ are i.i.d..

Remark. Despite the notation, one should bear in mind that the bias is a function of θ.

Definition 1.3. We say $\hat{\theta}$ is unbiased if $\text{bias}(\hat{\theta}) = 0$ for all θ.

Example 1.2. Using the example as before, i.e. $X_1, \ldots, X_n \sim N(\mu, 1)$ are i.i.d. and

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

Then $\mathbb{E}_\mu \hat{\mu} = \mu$, so $\hat{\mu}$ is unbiased.

Definition 1.4. The mean square error of an estimator $\hat{\theta}$ is $\text{mse}(\hat{\theta}) = \mathbb{E}_\theta((\hat{\theta} - \theta)^2)$.

This is, again, a function of θ.
The mse can be related to the bias by the bias-variance decomposition

$$\begin{aligned}\text{mse}(\hat{\theta}) &= \mathbb{E}_\theta((\hat{\theta} - \theta)^2) \\ &= \mathbb{E}_\theta((\hat{\theta} - \mathbb{E}_\theta\hat{\theta})^2) + (\mathbb{E}_\theta\hat{\theta} - \theta)^2 + 2(\mathbb{E}_\theta\hat{\theta} - \theta)\mathbb{E}_\theta(\hat{\theta} - \mathbb{E}_\theta\hat{\theta}) \\ &= \mathbb{E}_\theta((\hat{\theta} - \mathbb{E}_\theta\hat{\theta})^2) + (\mathbb{E}_\theta\hat{\theta} - \theta)^2 \\ &= \text{var}_\theta(\hat{\theta}) + \text{bias}(\hat{\theta})^2\end{aligned}$$

Consequently, there is a tradeoff between bias and variance. But sometimes we can increase the bias in such a way that a greater reduction in variance occurs which gives a smaller mse.

Example 1.3. Suppose $X \sim \text{Bin}(n, \theta)$ with known n. We want to estimate $\theta \in [0, 1]$. One can use the standard estimator $T_U = X/n$ which is unbiased since $\mathbb{E}_\theta T_U = \mathbb{E}_\theta X/n = n\theta/n = \theta$. Then

$$\text{mse}(T_U) = \text{var}_\theta(T_U) + \text{bias}^2(T_U) = \frac{\text{var}_\theta X}{n^2} = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n}$$

We can have an alternative estimator

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

(So 1/2 here is called the "fixed" estimator.) If $X = 8, n = 10$, then $T_U = 0.8$ and $T_B = 0.75$. As you expect from the form of T_B, it is biased

$$\text{bias}(T_B) = \mathbb{E}_\theta T_B - \theta = \mathbb{E}_\theta\left(\frac{X+1}{n+2}\right) - \theta = \frac{n}{n+2}\theta + \frac{1}{n+2} - \theta$$

which is nonzero for all but one value of θ. But if we compute the mse, we obtain

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

So T_B has smaller mse than T_U when θ is close to 1/2. So our prior judgement on the actual value of θ affects which estimator we use.

Unbiasedness is not necessarily a good thing.

Example 1.4. Suppose $X \sim \text{Poi}(\lambda)$ and we want to estimate $\theta = \mathbb{P}(X = 0)^2 = e^{-2\lambda}$. Then if we want $T(X)$ to be an unbiased estimator of θ, then necessarily

$$\sum_{k=0}^{\infty} T(k)\frac{\lambda^k}{k!} = e^{-\lambda} \implies T(X) = (-1)^X$$

which does not even make sense.

We can do simultaneous estimation as well. Suppose $X = (X^{(1)}, X^{(2)}, X^{(3)})$ with $X^{(i)} \sim N(\mu^{(i)}, 1)$ independent. We wish to estimate $(\mu^{(1)}, \mu(2), \mu^{(3)}) = \mu$. If we only have one observation, then the natural estimator is $\hat{\mu} = X = (X^{(1)}, X^{(2)}, X^{(3)})$. How do we decide whether an estimator is good when we are dealing with multiple parameters? One way to generalse the mse is

$$\mathbb{E}_\mu(\|\hat{\mu} - \mu\|^2) = \mathbb{E}_\mu\left(\sum_{i=1}^{3}(\hat{\mu}^{(i)} - \mu^{(i)})^2\right) = \sum_{i=1}^{3}\mathbb{E}_\mu((\hat{\mu}^{(i)} - \mu^{(i)})^2) = \sum_{i=1}^{3}\mathrm{mse}(\hat{\mu}^{(i)})$$

James & Stein (1961) found an alternative estimator

$$\hat{\mu}_{\mathrm{J-S}} = \left(1 - \frac{1}{\|X\|^2}\right)X$$

which takes other parameters into account when estimating one of the parameters. Very surprisingly, as James & Stein has shown in their 1961 paper,

$$\mathbb{E}(\|\hat{\mu}_{\mathrm{J-S}} - \mu\|^2) \leq \mathbb{E}_\mu(\|\hat{\mu} - \mu\|^2)$$

for all values of $\mu \in \mathbb{R}^3$. This means that $\hat{\mu}_{\mathrm{J-S}}$ essentially dominates $\hat{\mu} = X$ in the sense of smaller mse.
This is very counter-intuitive since $X^{(1)}, X^{(2)}, X^{(3)}$ are independent and completely related, so one of them should not give information to the other. But what this result shows is that we can estimate certain parameter better from the information about the other parameters. Fortunately, this result is just about the total mse. For individual mse's, it is not always true that James-Stein estimator dominates the natural estimator.

1.2 Sufficiency

Suppose $X_1, \ldots, X_n$ are i.i.d. random variables with pdf $f_X(\cdot|\theta)$ where θ is the parameter. This notation looks like a conditional. The confusion of notation here is deliberate. The use of it will becomes clear when we introduce Bayesian analysis.
A natural question is if a statistic $T(X)$ in some sense contains alls useful information about the sample X.

Definition 1.5. A statistic T is sufficient for θ if the conditional distribution of X given $T(X)$ does not depend on θ.

Remark. Everything can be vector-valued.

Example 1.5. Suppose $X_1, \ldots, X_n \sim \mathrm{Ber}(\theta)$ are i.i.d. where $\theta \in [0, 1]$. Then $f_X(x|\theta) = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}$. This depends only on $T(x) = \sum_i x_i$, so we guess this is a sufficient statistic.

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

which does not depend on θ. So T is indeed sufficient.

Theorem 1.1 (Factorisation Criterion)**.** *T is sufficient for θ iff $f_X(x|\theta) = g(T(x), \theta)h(x)$ for some suitable functions g, h.*

Proof. We will show the discrete case. The continuous case is analogous. Suppose $f_X(x|\theta) = g(T(x), \theta)h(x)$, then

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

So T is sufficient. Conversely, suppose T is sufficient, then

$$\begin{aligned} \mathbb{P}_\theta(X=x) &= \mathbb{P}_\theta(X=x, T(X)=T(x)) \\ &= \mathbb{P}_\theta(X=x|T(X)=T(x))\mathbb{P}_\theta(T(X)=T(x)) \end{aligned}$$

Writing $h(x) = \mathbb{P}_\theta(X=x|T(X)=T(x)), g(t,\theta) = \mathbb{P}_\theta(T(X)=t)$ finishes the proof. □

Example 1.6. Suppose $X_1, \ldots, X_n \sim \text{Ber}(\theta)$ are i.i.d., then taking $g(t,\theta) = \theta^t(1-\theta)^{n-t}$ and $h(x) = 1$ shows that $T(X) = \sum_i X_i$ is sufficient.

Example 1.7. Suppose $X_1, \ldots, X_n \sim \text{Unif}([0,\theta])$ for some $\theta \in (0,\infty)$. Intuitively, $T(X) = \max_i X_i$ is a sufficient statistic. This can be seen from

$$f_X(x|\theta) = \prod_{i=1}^n \frac{1}{\theta} 1_{x_i \in [0,\theta]} = \frac{1}{\theta^n} 1_{\max_i x_i \leq \theta} 1_{\min_i x_i \geq 0}$$

Quite obviously, sufficient statistics are not unique. If T is any sufficient statistic, so is $g \circ T$ where g is a suitable bijection. Also, we have the trivial sufficien statistic $T(X) = X$.

Definition 1.6. A sufficient statistic $T(X)$ is minimal sufficient if it is a function of every other sufficient statistic. That is, if T' is another sufficient statistic, then $T'(x) = T'(y) \implies T(x) = T(y)$.

Remark. By definition, any two minimal sufficient statistics are in bijection with each other.

Theorem 1.2. *Suppose T is a statistic such that $f_X(x|\theta)/f_X(y|\theta)$ is constant as a function of θ if and only if $T(x) = T(y)$. Then T is minimal sufficient.*

Consider the equivalence relations $x \sim y$ iff $f_X(x|\theta)/f_X(y|\theta)$ is constant in θ and $x \sim' y$ iff $T(x) = T(y)$. The condition in the theorem is basically saying that $\sim$ and $\sim'$ are essentially the same thing. We can always construct T from the quotient map of $\sim$, so the theorem also means that a minimal sufficient statistic always exists.

Proof. For any value t of T, let z_t be a representative from the equivalence class $\{x : T(x) = t\}$. Then,

$$f_X(x|\theta) = f_X(z_{T(x)}|\theta)\frac{f_X(x|\theta)}{f_X(z_{T(x)}|\theta)}$$

So T is sufficient. To see it is minimal, suppose S is any other sufficient statistic, then by factorisation 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

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

which does not depend on θ. So $x \sim y$, therefore $x \sim' y$ which exactly means $T(x) = T(y)$. □

Remark. This is just a sketch of the full proof since there are cases we haven't considered, like when some of the denominators is zero. In this case, what we mean by $f_X(x|\theta)/f_X(y|\theta)$ being constant in θ really means that we can factorise $f_X(x|\theta) = c(x,y)f_X(y|\theta)$ for some nice function c.

Example 1.8. Suppose $X - 1, \ldots, X_n \sim N(\mu, \sigma^2)$ are i.i.d.. Then

$$\frac{f_X(x|\mu,\sigma^2)}{f_X(y|\mu,\sigma^2)} = \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2\right) + \frac{\mu}{\sigma^2}\left(\sum_{i=1}^n x_i - \sum_{i=1}^n y_i\right)\right)$$

So $T(X) = (\sum_i X_i, \sum_i X_i^2)$ is minimal sufficient. Recall that composition with a suitable bijection does not affect a statistic being minimal sufficient. A more common minimal sufficient statistic that is used is

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

Note that in the example above, θ, T both have dimension 2. This does not have to be true in general.

Example 1.9. If we have $X_1, \ldots, X_n \sim N(\mu, \mu^2)$ i.i.d. instead, then the parameter is one-dimensional but the minimal sufficient statistics is still $(\bar{X}, S_{XX})$.

Turns out, we can use sufficient statistic to improve other estimators.

Theorem 1.3 (Rao-Blackwell)**.** *Let T be a sufficient statistic for θ and let $\tilde{\theta}$ be an estimator for θ such that $\mathbb{E}_\theta(\tilde{\theta}^2) < \infty$ for all θ. Define a new estimator by $\hat{\theta} = \mathbb{E}(\tilde{\theta}|T(X))$. Then for all θ,*

$$\mathbb{E}_\theta((\hat{\theta} - \theta)^2) \leq \mathbb{E}_\theta((\tilde{\theta} - \theta)^2)$$

where equality holds iff $\tilde{\theta}$ is a function of $T(X)$.

Remark. $\hat{\theta}$ is always what we'll call a bonafide estimator, i.e. it is a function of X but not of θ since T is sufficient. Also, as we will see in the proof, changing from $\tilde{\theta}$ to $\hat{\theta}$ does not change the bias.

Proof. $\mathbb{E}_\theta\hat{\theta} = \mathbb{E}_\theta(\mathbb{E}_\theta(\tilde{\theta}|T)) = \mathbb{E}_\theta\tilde{\theta}$ by Adam's Law, so $\hat{\theta}$ and $\tilde{\theta}$ have the same bias. Also, Eve's Law gives

$$\mathrm{var}(\tilde{\theta}) = \mathbb{E}_\theta(\mathrm{var}_\theta(\tilde{\theta}|T)) + \mathrm{var}_\theta(\mathbb{E}_\theta(\tilde{\theta}|T)) = \mathbb{E}_\theta(\mathrm{var}_\theta(\tilde{\theta}|T)) + \mathrm{var}(\hat{\theta}) \geq \mathrm{var}(\hat{\theta})$$

which implies the result. Equality holds iff $\mathrm{var}_\theta(\tilde{\theta}|T) = 0$ which requires $\tilde{\theta}$ to be a function of T. □

Example 1.10. Say $X_1, \dots, X_n \sim \text{Poi}(\lambda)$ are i.i.d. and we want to estimate $\theta = \mathbb{P}(X_1 = 0) = e^{-\lambda}$, then

$$f_X(x, \lambda) = \frac{e^{-n\lambda}\lambda^{\sum_i x_i}}{\prod_i (x_i!)} \implies f_X(x|\theta) = \frac{\theta^n(-\log\theta)^{\sum_i x_i}}{\prod_i (x_i!)}$$

By factorisation criterion, $T(X) = \sum_i X_i$ is sufficient. Recall that $\sum_i X_i \sim \text{Poi}(n\lambda)$. To find an unbiased estimator with low mse, we start from the unbiased (but bad) estimator $\tilde{\theta} = 1_{X_1=0}$. Then

$$\begin{aligned}\mathbb{E}(\tilde{\theta}|T = t) &= \mathbb{P}\left(X_1 = 0 \middle| \sum_{i=1}^n X_i = t\right) = \frac{\mathbb{P}(X_i = 0, \sum_{i\geq 2} X_i = t)}{\mathbb{P}(\sum_{i\geq 1} X_i = t)} \\ &= \frac{\mathbb{P}(X_i = 0)\mathbb{P}(\sum_{i\geq 2} X_i = t)}{\mathbb{P}(\sum_{i\geq 1} X_i = t)} = \left(\frac{n-1}{n}\right)^t\end{aligned}$$

So our better estimator is $\hat{\theta} = (1 - n^{-1})^{\sum_i X_i}$. Indeed, $\hat{\theta} \to e^{-\bar{X}}$ as $n \to \infty$, so $\hat{\theta} \approx e^{-\lambda}$ for large n by SLLN.

Example 1.11. Let $X_1, \dots, X_n \sim \text{Unif}[0, \theta]$ be i.i.d. and we want to estimate $\theta \geq 0$. We already know that $T = \max_i X_i$ is sufficient for θ. To get a good unbiased estimator, we start from the unbiased estimator $\tilde{\theta} = 2X_1$, then

$$\begin{aligned}\hat{\theta} &= \mathbb{E}(\tilde{\theta}|T = t) \\ &= 2\mathbb{E}\left(X_1 \middle| \max_i X_i = t\right) \\ &= \frac{2t}{n} + \frac{2(n-1)}{n}\mathbb{E}\left(X_1 \middle| X_1 < t, \max_{i\geq 2} X_i = t\right) \\ &= \frac{2t}{n} + \frac{2(n-1)}{n}\frac{t}{2} = \frac{n+1}{n}\max_i X_i\end{aligned}$$

1.3 Maximum Likelihood Estimation

Let $X = (X_1, \dots, X_n)$ have joint pdf (pmf) $f_X(x|\theta)$.

Definition 1.7. The likelihood of θ is the (random) function $L : \theta \mapsto f_X(X|\theta)$. The maximum likehood estimator (mle) $\hat{\theta}$ of θ is the value of θ maximising L.

If $X_1, \dots, X_n$ are independent with pdf (pmf) $f_i(\cdot|\theta)$ respectively, then $L(\theta) = \prod_i f_i(X_i|\theta)$. Sometimes it is easier to deal with the log-likelihood $\ell(\theta) = \log L(\theta) = \sum_i \log f_i(X_i|\theta)$.

Example 1.12. Let $X_1, \dots, X_n \sim \text{Ber}(p)$ be i.i.d. and

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

which is maximised when $p = \sum_i X_i/n$ which is the mle $\hat{p}$. We have $\mathbb{E}\hat{p} = p$, so it is unbiased.

Example 1.13. Suppose $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ are i.i.d., then

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

is maximised then $\mu = \bar{X}, \sigma^2 = S_{XX}/n$. Again, the mle for μ is unbiased, but we will later see that $S_{XX}/\sigma^2 \sim \chi^2_{n-1}$, so $\mathbb{E}(S_{XX}/n) = \mathbb{E}(\chi^2_{n-1})\sigma^2/n = (n-1)\sigma^2/n$ which means it is biased. But when $n \to \infty$, the bias goes away, so we say it is asymptotically unbiased.

Example 1.14. Take $X_1, \dots, X_n \sim \text{Unif}[0, \theta]$ i.i.d., then $L(\theta) = \theta^{-n} 1_{\max_i X_i \le \theta}$. So the mle is $\hat{\theta} = \max_i X_i$. Recall that we earlier obtained an unbiased estimator in the form $(n+1)\max_i X_i/n$ using Rao-Blackwell, so again although the mle is biased, it is asymptotically unbiased.

If T is a sufficient statistic, then the mle is a function of T because we can factorise $L(\theta) = g(T, \theta)h(X)$ (so the maximiser over θ only depends on X through T).
Also, if $\phi = h(\theta)$ (where h is a bijection) then the mle of ϕ is, of course, $\hat{\phi} = h(\hat{\theta})$ where $\hat{\theta}$ is a mle of θ.
Another important property is that $\sqrt{n}(\hat{\theta} - \theta)$ is asymptotically normal with mean 0 as $n \to \infty$. The covariant matrix of this normal is the "smallest attainable". That is, under suitable conditions, $\mathbb{P}(\sqrt{n}(\hat{\theta} - \theta) \in A) \to \mathbb{P}(z \in A)$ where $Z \sim N(0, \Sigma)$ and Σ is a known function of ℓ.
When the mle is not available in closed form, we can find it numerically for a particular observation $X = x$.

1.4 Confidence Interval

Definition 1.8. A γ confidence interval (with $0 < \gamma < 1$) for a parameter θ is a random interval $(A(X), B(X))$ such that $\mathbb{P}(A(X) \le \theta \le B(X)) = \gamma$ for all values of θ.

What is the right way to interpret this? We can think of this from a frequentist's interpretation: If we repeat the experiment many times, on average γ of the time the interval $(A(X), B(X))$ contains θ. However, this notion is sometimes misinterpreted as "having observed $X = x$, $\theta \in (A(x), B(x))$ with probability γ".

Example 1.15. Suppose $X_1, \dots, X_n \sim N(\theta, 1)$ are i.i.d.. We want to find a 0.95 confidence interval for θ. We know that $Z = \sqrt{n}(\bar{X} - \theta) \sim N(0, 1)$. Let Φ be the cdf of $N(0, 1)$, then for any z_1, z_2 such that $\Phi(z') - \Phi(z) = 0.95$, we have

$$\mathbb{P}\left(\bar{X} - \frac{z'}{\sqrt{n}} \le \theta \le \bar{X} + \frac{z}{\sqrt{n}}\right) = 0.95$$

Hence $(\bar{X} - z'/\sqrt{n}, \bar{X} + z/\sqrt{n})$ is a 0.95 confidence interval.
Of course, there are uncountably many possible choices of z, z'. Typically, we like to center everything (i.e. $-z = z'$) which can be achieved by $-z = z' = z_{0.025} \approx 1.96$ where $z_\alpha = \Phi^{-1}(1-\alpha)$ is called the upper α point of distribution. So the confidence interval is usually $(\bar{X} \pm 1.96/\sqrt{n})$.

In general, to find a confidence interval, we first find a quantity $R(X, \theta)$ (called the pivot) such that the $\mathbb{P}_\theta$-distribution of $R(X, \theta)$ does not depend on θ. In the above example, we took $R(X, \theta) = \sqrt{n}(\bar{X} - \theta)$. Then, we write down $\mathbb{P}(c_1 \leq R(X, \theta) \leq c_2) = \gamma$ and solve for c_1, c_2. This is usually not very hard since the distribution of $R(X, \theta)$ we find is usually common ones (e.g. normal or χ^2). Rearranging everything will give the result if R is nice enough.

Proposition 1.4. *If T s monotone increasing and $(A(X), B(X))$ is a γ confidence interval for θ, then $(T(A(X)), T(B(X)))$ is a γ confidence interval for $T(\theta)$.*

Remark. When θ is a vector, we can, of course, define an analog of confidence intervals by replacing them with "confidence sets" in higher dimensions.

Example 1.16. Suppose $X_1, \ldots, X_n \sim N(0, \sigma^2)$ are i.i.d. and we want to find a 0.95 confidence interval for σ^2. Note that $X_i/\sigma \sim N(0, 1)$, so $\sum_i X_i^2/\sigma^2 \sim \chi_n^2$. Hence we take the pivot $R(X, \sigma^2) = \sum_i X_i^2/\sigma^2$. Then

$$\mathbb{P}\left(\frac{\sum_i X_i^2}{c_2} \leq \sigma^2 \leq \frac{\sum_i X_i^2}{c_1}\right) = 0.95, c_1 = F_{\chi_n^2}^{-1}(0.025), c_2 = F_{\chi_n^2}^{-1}(0.975)$$

where $F_{\chi_n^2}$ is the cdf of χ_n^2. So $(\sum_i X_i^2/c_2, \sum_i X_i^2/c_1)$ is a 0.95 confidence interval for σ^2, which, by the preceding theorem, gives $(\sqrt{\sum_i X_i^2/c_2}, \sqrt{\sum_i X_i^2/c_1})$ as a 0.95 confidence interval for σ.

Example 1.17. Suppose $X_1, \ldots, X_n \sim \text{Ber}(p)$ are i.i.d. with n large. We want to find an approximate 0.95 confidence interval for p. The mle $\hat{p}$ of p has the form $\hat{p} = n^{-1} \sum_i X_i$, which is approximately $N(p, p(1-p)/n)$ for large n. So $\sqrt{n}(\hat{p} - p)/\sqrt{p(1-p)}$ is approximately $N(0, 1)$ and hence

$$\mathbb{P}\left(z_{0.025} \leq \sqrt{n}\frac{(\hat{p} - p)}{\sqrt{p(1-p)}} \leq z_{0.975}\right) \approx 0.95$$

Since we are approximating anyways, we argue that $\sqrt{p(1-p)} \approx \sqrt{\hat{p}(1-\hat{p})}$, which gives (using $z_{0.025} = -z_{0.975}$)

$$\mathbb{P}\left(\hat{p} - z_{0.975}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \leq p \leq \hat{p} + z_{0.975}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right) \approx 0.95$$

Note that $\sqrt{p(1-p)} \leq 1/2$ for all $p \in (0, 1)$, so we can also report a "conservative" 0.95 confidence interval $\hat{p} \pm 1.96\sqrt{1/4n}$.

Let's go back to our original question: How do we interpret confidence intervals? Suppose X_1, X_2 are i.i.d. $\text{Unif}(\theta - 1/2, \theta + 1/2)$, then, as one can easily verify, $(\min(X_1, X_2), \max(X_1, X_2))$ is a 0.5 confidence interval for θ. But if we have $|X_1 - X_2| > 0.5$, then θ has to be contained in $(\min(X_1, X_2), \max(X_1, X_2))$! The frequentist interpretation of the confidence interval is still correct, but it is not sensible anymore to say that having seen a particular X_1, X_2 far apart enough (e.g. $X_1 = 0.1, X_2 = 0.9$) we are "0.5 certain that θ is in the confidence interval" since we are actually absolutely certain.

2 Bayesian Analysis

So far, we have only viewed statistics from a frequentist's point of view. We thought of θ as a fixed true parameter and inferential statements are interpreted in terms of repeated sampling. Frequentist guarantees hold uniformly for all possible true parameters. Bayesian analysis, however, is very very different framework for inference. Which one is better? There is no simple answer.
Bayesians treat θ as a random variable taking values in some space Θ. The prior distribution $\pi(\theta)$ represents the investigator's beliefs or information about θ before observing the data. Conditional on θ, the data X has pdf (pmf) $f(\cdot|\theta)$. Having observed data $X = x$, this information is combined with the prior distribution to form the posterior distribution $\pi(\theta|x)$ which is the conditional distribution of θ given $X = x$. By Bayes' rule,

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

where $f_X(x)$ is the marginal distribution of X, i.e.

$$f_X(x) = \int_\Theta f_X(x|\theta)\pi(\theta)\,\mathrm{d}\theta$$

if θ is continuous and

$$f_X(x) = \sum_{\theta\in\Theta} f_X(x|\theta)\pi(\theta)$$

if θ is discrete. More simply, we can write $\pi(\theta|x) \propto \pi(\theta) f_X(x|\theta)$ up to a constant that does not depend on θ (hence can be recovered by normalising which is usually easy).
By the factorisation criterion of the likelihood, the posterior distribution only depends on x through a sufficient statistic, i.e.

$$\pi(\theta|x) \propto \pi(\theta) \times f_X(x|\theta) = \pi(\theta) \times g(T(x), \theta)h(x) \propto \pi(\theta) \times g(T(x), \theta)$$

Example 2.1. Suppose we have a patient walks into COVID-19 testing centre with no extra factor. We want to know information about

$$\theta = \begin{cases} 1, \text{ if the patient if infected} \\ 0, \text{ otherwise} \end{cases}$$

Let X be the outcome of the test (1 if positive and 0 if negative). Suppose we know about the sensitivity of test $f_X(X = 1|\theta = 1)$ and specificity $f_X(X = 0|\theta = 0)$. To choose a prior, we can set $\pi(\theta = 1)$ to be the proportion of people infected in the country at that time (obtained from e.g. survey). What is the chance of infection given a positive test?

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

which can actually be quite small if $\pi(\theta = 0) >> \pi(\theta = 1)$.

Example 2.2. Suppose $\theta \in (0, 1)$ is the mortality rate for a new surgery in Addenbrokes hospital. In the first 10 operations there are no deaths, In other

hospitals around the country the mortality rate is between 0.03 and 0.2 with an average of 0.1. Before getting data from Addenbrokes, we naturally have in mind a prior distribution based on the country-wide data, say Beta$(3, 27)$ (so that $\pi(\theta)$ has mean 0.1 and $\pi(0.03 < \theta < 0.2) = 0.9$).
Let $X_i \sim \text{Ber}(\theta)$ be the indicator of whether the i^{th} patient at Addenbrookes dies, then the likelihood distribution is $f_X(x|\theta) = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}$. Then the posterior is

$$\begin{aligned}\pi(\theta|x) &\propto \pi(\theta) \times f_X(x|\theta) \propto \theta^{a-1}(1-\theta)^{b-1}\theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i} \\ &\propto \theta^{\sum_i x_i + a - 1}(1-\theta)^{b+n-1-\sum_i x_i}\end{aligned}$$

So $\pi(\theta|x)$ is Beta$(a + \sum_i x_i, b + n - \sum_i x_i)$. Putting in $a = 3, b = 27, n = 10, \sum_i x_i = 0$, the posterior distribution is Beta$(3, 37)$, which is leaning more towards 0 than the prior distribution.
Note that in this example, the prior and the posterior are both beta distributions, which is a phenomenon known as conjugacy.

What to do with the posterior after we computed it? Let $\pi(\theta|x)$ be the information available about θ after doing the experiment. The process of making decisions under uncertainty can then be formalised using it:
Let D be the space of decisions and we aim to pick a decision $\delta \in D$. The loss function $L(\theta, \delta)$ is the loss incurred when making decision δ when the true parameter is θ. For example, we can let $D = \{0, 1\}$ be the decision of whether the patient is required to self-isolate. Then, e.g. $L(\theta = 0, \delta = 1)$ is the loss incurred by requiring self-isolation with no infection. Von Neumann-Morgenstern Theorem states that under axioms of rational behaviour, we must pick δ with minimises the expectation of $L(\theta, \delta)$ in the posterior.
An example of a decision is a "best guess" for θ. The Bayes estimator $\hat{\theta}^{(b)}$ minimises the posterior expected loss

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

Example 2.3. Suppose the loss is quadratic, i.e. $L(\theta, \delta) = (\theta - \delta)^2$, then

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

is minimised when

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

So the Bayes estimator $\hat{\theta}^{(b)}$ under this loss is the posterior mean.

Example 2.4. Suppose we have the absolute error loss $L(\theta, \delta) = |\theta - \delta|$, then

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

So $h'(\delta) = 0$ if δ is the median of the posterior distribution.

Definition 2.1. A γ credible interval $(A(x), B(x))$ is an interval such that $\pi(A(x) \leq \theta \leq B(x)|x) = \gamma$.

Like confidence interval, there can be many choices of (A, B). Unlike confidence interval, credible intervals can be interpreted conditionally, i.e. given the experiment $X = x$, θ has probability γ of being in the credible interval. However, one should note that credible intervals depend on the investigator's choice of prior.

Example 2.5. Suppose $X_1, \ldots, X_n \sim N(\mu, 1)$ are i.i.d. and we choose prior distribution $\pi(\mu)$ to be $\sim N(0, \tau^{-2})$ for a known τ. Then

$$\begin{aligned}\pi(\mu|x) &\propto f_X(x|\mu) \times \pi(\mu) \\ &= \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \times \exp\left(-\frac{\mu^2\tau^2}{2}\right) \\ &\propto \exp\left(-\frac{1}{2}(n+\tau^2)\left(\mu - \frac{\sum_i x_i}{n+\tau^2}\right)^2\right)\end{aligned}$$

which is $N(\sum_i x_i/(n+\tau^2), 1/(n+\tau^2))$. The posterior mean (and median) is $\sum_i x_i/(n+\tau^2)$ which is the Bayes estimator under both quadratic and absolute error loss. The posterior variance decreases as $1/(n+\tau^2) \approx 1/n$.

Example 2.6. Suppose $X_1, \ldots, X_n \sim \text{Poi}(\lambda)$ are i.i.d. and the prior $\pi(\lambda)$ is Exp(1). Then

$$\pi(\lambda|x) \propto f_X(x|\lambda) \times \pi(\lambda) \propto e^{-n\lambda}\lambda^{\sum_i x_i} \times e^{-\lambda} \propto e^{-(n+1)\lambda}\lambda^{\sum_i x_i}$$

which is just $\Gamma(\sum_i x_i + 1, n+1)$. Under quadratic loss, the Bayes estimator is

$$\hat{\lambda} = \frac{\sum_i x_i + 1}{n+1}$$

Under absolute value loss, $\hat{\lambda}$ solves

$$\int_0^{\hat{\lambda}} \frac{(n+1)^{\sum_i x_i + 1}\lambda^{\sum_i x_i}e^{-(n+1)\lambda}}{(\sum_i x_i)!}\,\mathrm{d}\lambda = \frac{1}{2}$$

As we have seen, Bayesian analysis is quite different from a frequentist's approach. So where were they emerged from? In 1763, Thomas Bayes started the theory of statistics based on this Bayesian inference (or "inverse probability"). Before 1920, all statistics were done in the Bayesian idea. In the 1920s, Fisher developed the idea of frequentist procedures due to the fundamentally subjective choice of an prior in Bayesian inference. Between the 1920s and 1950s, many statisticians developed this idea, until a revival of Bayesian statistics between 1950s and 1970s motivated by practical applications. In the 1980s, however, the computer revolution happened which allows statisticians to do posterior inference in very complex situations by Markov Chain Monte Carlo.
What is good about Bayesian methods? Firstly, inferential statements can be easier to interpret (e.g. credible interval versus confidence interval). Also, it provides a natural way to incorporate prior information if that were to be taken into consideration in the model. It also has a relatively more straightforward

mechanics. In addition, the Bayesian estimator provides a natural way to make decisions under uncertainty.
What are the advantages of frequentist statements? We do not have to choose a prior, so it is more "impartial". It is then arguably a better framework for investigators with different prior beliefs to reach agreement.
Prototypically, Bayesian methods are used in public surveys, e.g. how many people are currently infected with COVID-19 in the UK. Frequentist ideas, on the other hand, are mostly used in scientific field like clinical trials for new drug.
In sequential analysis, they can be very very different. Consider two experiments: In Experiment 1, we perform $N = 12$ Ber(p) trials and we saw $X = 3$ successes. In Experiment 2, we perform Ber(p) trials until we obtain $X = 3$ successes, and it turns out it took 12 trials to do it in this occasion. Bayesians would reach the same conclusion for both experiments as the likelihood does not change. But frequentists would reach different conclusions about p (e.g. in terms of confidence intervals). This has important consequences in, say, the design of clinical trials.

3 Testing Hypotheses

A hypothesis is an assumption about the distribution of a random variable X that we want to know about. Scientific questions, for example, are often phrased as a decision between a null hypothesis H_0 (simple model, bare case, no effect) and an alternative hypothesis H_1 (complex model, interesting case, positive/negative effect).

Example 3.1. 1. $X = (X_1, \dots, X_n)$ are i.i.d. Ber(θ). We can take the null hypothesis to be $H_0 : \theta = 1/2$, i.e. these are fair coins, and the alternative hypothesis $H_1 : \theta = 3/4$ saying the coin is biased.
2. Or, as in the previous example, we can take $H_1 : \theta \neq 1/2$.
3. We can take $X = (X_1, \dots, X_n)$ where each X_i takes value in $\mathbb{N}$ and take $H_0 : X_i \sim \text{Poi}(\lambda)$ i.i.d. for some $\lambda > 0$ and $H_1 : X_i \sim f$ i.i.d. for some other distribution f. This the "goodness of fit" test, as we can trying to see whether Poisson is a good fit to the data.
4. Suppose X has pdf $f(\cdot|\theta)$ for some $\theta \in \Theta$. The null hypothesis can be taken as $H_0 : \theta \in \Theta_0 \subsetneq \Theta$ and the altenative hypothesis $H_1 : \theta \notin \Theta_0$.
5. $H_0 : X \sim f_0, H_1 : X \sim f_1$.

Definition 3.1. A simple hypothesis is one that fully specifies the distribution of X, otherwise it is called a composite hypothesis.

One can classify the hypotheses listed above using this criterion (exercise).

Definition 3.2. A test of the null hypothesis H_0 is defined by the critical region C which is a subset of the space of values of X. When $X \in C$, we reject H_0; when $X \notin C$, we "fail to reject" H_0 or "find no evidence against it".

There are, of course, errors that can be made.

Definition 3.3. A type I error occurs when we reject H_0 when H_0 is true. A type II error occurs when we fail to reject H_0 when H_0 is false.

3.1 Simple Hypotheses

Definition 3.4. When H_0, H_1 are both simple, the probability of type I error is

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

and that of type II error is

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

The size of the test is α; The power of the test is $1 - \beta$.

There is, of course, a tradeoff between α and β. Usually we set α at an acceptable level, say 0.01, and minimises β subject to this restriction as best as we can. The role of H_1 is subordinated to that of H_0, as we can design a test of size α without making any reference to H_1. Nonetheless, we want to use tests that have the power to reject H_0 when a given H_1 is true.

Definition 3.5. Let H_0, H_1 be simple with X having pdf f_1, f_2 under H_0, H_1 respectively. The likelihood ratio statistic is

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

A likelihood ratio test (LRT) rejects H_0 when $X \in C = \{x : \Lambda_x(H_0; H_1) > R\}$ for some $R \geq 0$.

Lemma 3.1 (Neyman-Pearson Lemma)**.** *Suppose f_0, f_1 are nonzero on the same sets and suppose there exists $k \geq 0$ such that the LRT with $R = k$ has size α. Then, out of all tests with size at most α, the test with the smallest β is this LRT.*

Remark. We assumed the existence of a LRT of size α, which does not always exist (find a counterexample – exercise). However, we can always define a "randomised" test with exact level α.

Proof. Let C be the critical region for the said LRT with size α and power $1-\beta$ and C^* be the critical region of another test of size $\alpha^* \leq \alpha$ and power $1 - \beta^*$. Denote the complement of a region D as $\bar{D}$, then we have

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

as desired. □

Example 3.2. Take $X_1, \ldots, X_n \sim N(\mu, \sigma_0^2)$ i.i.d. with σ_0 known. We want the best size α test for $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$ for some fixed $\mu_1 > \mu_0$. Then,

$$\Lambda_x(H_0; H_1) = \exp\left(\frac{\mu_1 - \mu_0}{\sigma_0^2} n\bar{x} + \frac{n(\mu_0^2 - \mu_1^2)}{2\sigma_0^2}\right)$$

which is monotone in $\bar{x}$. So for any k there is a $c = c(k)$ such that $\Lambda_x(H_0; H_1) > k$ iff $\bar{x} > c$. The critical region of this LRT is thus $\{x : \bar{x} > c\}$. By way of a linear transformation $z = \sqrt{n}(\bar{x} - \mu_0)/\sigma_0$, we can write the critical region as $\{x : z > c'\}$ for some $c' = c'(k)$. Under the null hypothesis H_0, the random variable Z taking its value as z has $Z \sim N(0, 1)$, so the test $\{x : z > \Phi^{-1}(1-\alpha)\}$ has size α. This is called a z-test.

Definition 3.6. For any test with critical region of the region of the form $\{x : T(x) > k\}$ where T is some statistic, we usually report the p-value (aka observed significance level) defined as $p = \mathbb{P}_{H_0}(T(X) > T(x^*))$ where x^* is the observed data.

So the p-value measures the probability (under the null) of sampling an obervation that is "more extreme" than the observation x^* as measured by T.

Example 3.3. In the previous example with $\mu_0 = 5, \mu_1 = 6, \sigma_0 = 1, \alpha = 0.05$ with observed data $x^* = (5.1, 5.5, 4.9, 5.3)$, then $\bar{x}^* = 5.2$, correspondingly $z^* = 0.4$. $\Phi^{-1}(1 - \alpha) \approx 1.645 > 0.4 = z^*$, so the observation fails to reject H_0. Indeed, $p = 1 - \Phi(z^*) \approx 0.35$.

Proposition 3.2. *Under H_0, the p-value distributes as* Unif$[0, 1]$.

Proof. Let F be the distribution function of the statistic T (which is assumed to be monotone and bijective), then

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

which is the distribution function of Unif$[0, 1]$. □

3.2 Composite Hypotheses

When H_0, H_1 are simple, it is easy to define the error since they fully specified pdf's of data. But of course, we would also want to do statistics over composite hypotheses. The class of composite hypotheses we are considering concerns data distributed in a parametric family $X \sim f_X(\cdot|\theta), \theta \in \Theta$ and $H_0 : \theta \in \Theta_0 \subset \Theta, H_1 : \theta \in \Theta_1 \subset \Theta$. A test of H_0 against H_1 is again constructed in terms of a critical region C.

Definition 3.7. The power function of a test with critical region C is $W(\theta) = \mathbb{P}_\theta(X \in C)$. The size of a test with composite null H_0 is the worst-case Type I error probability, i.e. $\alpha = \sup_{\theta \in \Theta_0} W(\theta)$.
We say the test is uniformly most powerful (UMP) of size α if $\sup_{\theta \in \Theta_0} W(\theta) = \alpha$ and for any other test C^* of size at most α has $W^*(\theta) \leq W(\theta)$ (where W^* is the new power function) for all $\theta \in \Theta_1$.

Note that UMP tests need not exist. However, many likelihood ratio tests are UMP.

Example 3.4. Let $X_1, \ldots, X_n \sim N(\mu, \sigma_0^2)$ be i.i.d. with σ_0^2 known. Fix some μ_0. We wish to test $H_0 : \mu \leq \mu_0$ against $\mu > \mu_0$. We have seen that for the simple hypotheses $H_0' : \mu = \mu_0, H_1' : \mu = \mu_1$, the LRT was $C = \{x : z = \sqrt{n}(\bar{x} - \mu_0)/\sigma_0 > z_\alpha\}$ (where $z_\alpha = \Phi^{-1}(1-\alpha)$ as usual). The claim is that the same test C is UMP of size α for H_0 against H_1. The power function is

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

which is strictly increasing in μ and equals to α at $\mu = \mu_0$. So the test is indeed of size α.
If we have another test C^* of size at most α with power function W^*, we want to show that $W(\mu_1) \geq W(\mu_1)$ for all $\mu_1 > \mu_0$. The main observation is that the critical regions only depend on μ_0 but not μ_1. So C^* also has size at most α for $H_0' : \mu = \mu_0$ against $H_1' : \mu = \mu_1$. Therefore by Neyman-Pearson $W(\mu_1) \geq W^*(\mu_1)$ for any $\mu_1 > \mu_0$, hence C is UMP.

Definition 3.8. The generalised likelihood ratio (GLR) statistic for $H_0 : \theta \in \Theta_0 \subset \Theta, H_1 : \theta \in \Theta_1 \subset \Theta$ is given by

$$\Lambda_X(H_0; H_1) = \frac{\sup_{\theta \in \Theta_1} f_X(X|\theta)}{\sup_{\theta \in \Theta_0} f_X(X|\theta)}$$

Large values of $\Lambda_X(H_0; H_1)$ indicates larger departure from H_0 (or better "fit" with H_1).

Example 3.5 (Two-sided normal mean test)**.** Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma_0^2)$ are i.i.d. wih σ_0^2 known. We wish to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ with μ_0 some constant. Then the GLR statistic is

$$\Lambda_X(H_0; H_1) = \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i - \bar{X})^2\right) \Big/ \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i - \mu_0)^2\right)$$

The GLR test rejects H_0 when $\Lambda_X(H_0; H_1)$ is large. When does this happen? Taking log on both sides gives

$$2\log\Lambda_X(H_0; H_1) = \frac{n}{\sigma_0^2}(\bar{X} - \mu_0)^2$$

So indeed the test rejects H_0 when $|\sqrt{n}(\bar{X} - \mu_0)/\sigma_0|$ is large. How large? Under H_0, we have $\sqrt{n}(\bar{X} - \mu_0)/\sigma_0 \sim N(0, 1)$, so the test that rejects H_0 when $|\sqrt{n}(\bar{X} - \mu_0)/\sigma_0| > z_{\alpha/2}$ has size α.
We can also formulate the critical region in another way. Note that we have $2\log\Lambda_X(H_0; H_1) = n(\bar{X} - \mu_0)^2/\sigma_0^2 \sim \chi_1^2$, so we can write the critical region instead as $\{n(\bar{X} - \mu_0)^2/\sigma_0^2 > \chi_1^2(\alpha)\}$ where $\chi_1^2(\alpha)$ is the upper α point of a χ_1^2 distribution.

Why do we care about writing it in terms of χ^2 distribution? Turns out, in a general setting, the sampling distribution of $2\log\Lambda_X(H_0;H_1)$ can be approximated by a χ^2 distribution.

Definition 3.9. Suppose the parameter θ is k-dimensional. The dimension $\dim\Theta_0$ of a hypothesis $H_0:\theta\in\Theta_0$ is the number of "free parameters" in Θ_0.

Example 3.6. 1. If $\Theta_0=\{\theta\in\mathbb{R}^k:\theta_1=\cdots=\theta_p=0\}$, then $\dim\Theta_0=k-p$.
2. Let $A\in\mathbb{R}^{p\times k}, b\in\mathbb{R}^p, p<k$, then the hyperplane $\Theta_0=\{\theta\in\mathbb{R}^k:A\theta=b\}$ has dimension $k-p$ if the rows of A are linearly independent.
3. Suppose $f:\mathbb{R}^p\to\mathbb{R}^k$ satisfies certain regularity conditions, then the image Θ_0 of it has dimension p.

Theorem 3.3 (Wilk's Theorem)**.** *Suppose $\Theta_0\subset\Theta_1$ (i.e. null is the special case of alternative) and $\dim\Theta_1-\dim\Theta_0=p$. If $X_1,\dots,X_n$ are i.i.d., then under regularity conditions $2\log\Lambda_X(H_0;H_1)\sim\chi^2_p$ as $n\to\infty$. That is, for any $\theta\in\Theta_0, l>0$,*

$$\lim_{n\to\infty}\mathbb{P}_\theta(2\log\Lambda_X(H_0;H_1)\le l)=\mathbb{P}(\Xi\le l),\Xi\sim\chi^2_p$$

Proof. Not in this course, no. □

How do we apply this? Indeed, with this result, we immediately know that the test rejecting H_0 when $2\log\Lambda_X(H_0;H_1)\ge\chi^2_p(\alpha)$ has size α asymptotically.

Example 3.7. In two-sided normal mean test, we saw $\Theta_0=\{\mu_0\},\Theta_1=\mathbb{R}\setminus\{\mu_0\}$. But the completion $\Theta_1=\mathbb{R}$ does not actually change the GLR, so we might as well take that in order to apply Wilk's Theorem. In this case, $\dim\Theta_1-\dim\Theta_0=1$ and we already saw $2\log\Lambda_X(H_0;H_1)\sim\chi^2_1$. The nice thing is that with Wilk's theorem, the same thing is (asymptotically) true in other parametric families.

3.3 Goodness-of-fit Tests

Suppose $X_1,\dots,X_n$ are i.i.d. samples from a distribution on $\{1,\dots,k\}$. Let $p_i=\mathbb{P}(X_1=i)$ and let N_i be the number of observations in $X_1,\dots,X_n$ equal to i. So $\sum_i p_i=1,\sum_i N_i=n$.

Example 3.8. Mendel crossed $n=556$ smooth yellow peas with green wrinkled peas. Each member of progeny can have any combination of the two features. So we end up with 4 types SY, SG, WY, WG with probabilities (p_1,p_2,p_3,p_4). Suppose (N_1,N_2,N_3,N_4) are the numbrs of samples of each type. We are interested in testing the null $H_0:p_i=\tilde{p}_i$ for some fixed known parameter $\tilde{p}$. The alternative H_1 put no constraint on p (other than $\sum_i p_i=1$). Mendel's genetic theory suggests that $p=\tilde{p}=(9/16,3/16,3/16,1/16)$, but how do we test this?

The model can be written in general as $(N_1,\dots,N_k)\sim\text{Multi}(n;p_1,\dots,p_k)$ which has log likelihood $\ell(p)=\sum_i N_i\log p_i+\text{const.}$. We can test H_0 against H_1 using the generalised LRT

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

$\ell(p)$ can be optimised by using Lagrange multipliers. The only constraint here is $\sum_i p_i = 1$, so we consider

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

which, after the standard procedure shows that we we should choose $\hat{p}_i = N_i/n$, so

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

Wilk's theorem says that $2 \log \Lambda \sim \chi^2_p$ asymptotically with $p = \dim \Theta_1 - \dim \Theta_0 = k - 1 - 0 = k - 1$. So we can reject H_0 with size apprximately α if $2 \log \Lambda > \chi^2_{k-1}(\alpha)$.
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 null. Let $\delta = o_i - e_i$, then

$$2 \log \Lambda = 2 \sum_{i} (e_i + \delta_i) \log \left(1 + \frac{\delta_i}{e_i}\right) \approx \sum_{i} \frac{\delta_i^2}{e_i} = \sum_{i} \frac{(o_i - e_i)^2}{e_i}$$

The last expression is known as Pearson's χ^2 statistic, which is again asymptotically χ^2_{k-1}.

Example 3.9. In Mendel's experiment $(n_1, n_2, n_3, n_4) = (315, 108, 102, 31)$ are observed. Then in this case $2 \log \Lambda = 0.618, \sum_i (o_i - e_i)^2/e_i = 0.604$. $\chi^2_3(0.05) = 7.815$ which is bigger than both statistic, so the test does not reject H_0 at 0.05 level. In fact, the p-value is approximately $\mathbb{P}(\chi^2_3 > 0.6) \approx 0.96$, so the fit of Mendel's theorectical distribution $\tilde{p} = (9/16, 3/16, 3/16, 1/16)$ is almost too good!

How about composite hypotheses? In general, this is of course not very easy, but we are often just interested in $H_0 : p_i = p_i(\theta)$ for some parameter θ against H_1 putting no constraint on p.

Example 3.10. Suppose all individuals have one of three genotypes and their respective probabilities in the null hypothesis H_0 are $p_1 = \theta^2, p_2 = 2\theta(1 - \theta), p_3 = (1 - \theta)^2$. We shall test it against H_1 that puts no restriction on p. The generalised LRT has

$$\begin{aligned} 2 \log \Lambda &= 2 \left(\sup_{\sum_i p_i = 1} \ell(p) - \sup_{\theta} \ell(p(\theta)) \right) = 2(\ell(\hat{p}) - \ell(p(\hat{\theta}))) \\ &= 2 \sum_{i} N_i \log(N_i/(np_i(\hat{\theta}))) \end{aligned}$$

where $\hat{p}$ is the mle of p under H_1, i.e. $\hat{p}_i = N_i/n$, and $\hat{\theta}$ is the mle of θ under H_0. So we can still write $e_i = np_i(\hat{\theta}), o_i = N_i$ and the approximation of Pearson

statistic to this still holds.
In the genetic example, we have

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

maximising over θ gives $\hat{\theta} = (2N_1 + N_2)/(2n)$. So we can plug this into the formula for $2\log\Lambda$ and perform test by referring to the χ^2_p distribution with $p = \dim\Theta_1 - \dim\Theta_0 = 2 - 1 = 1$.

3.4 Independence

Suppose we have $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. where X_i's take value in $\{1, \ldots, r\}$ and Y_i in $\{1, \ldots, c\}$. We wish to test H_0 that states that X_i, Y_i are independent.

Definition 3.10. The contingency table is the array with entries $N_{ij} = |\{l : 1 \le l \le n, (X_l, Y_l) = (i, j)\}|$.

If n is fixed and a sample has type (i, j) with probability p_{ij}, then the natural model is $(N_{ij}) \sim \text{Multi}(n; p_{ij})$. It is not very practical in some cases since n is usually not fixed. We will, however, derive a test that is valid in both cases. Let $p_{i*} = \sum_j p_{ij}, p_{*j} = \sum_i p_{ij}$. Then the null hypothesis of X_i, Y_i being independent translates to $H_0 : p_{ij} = p_{i*}p_{*j}$ for $i \in \{1, \ldots, r\}, j \in \{1, \ldots, l\}$. We take the alternative hypothesis H_1 as (p_{ij}) unconstrained as usual. The generalised LRT is

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

where $\hat{p}_{ij}$ is the mle under H_1, i.e. $\hat{p}_{ij} = N_{ij}/n$, and $\hat{p}_{i*}\hat{p}_{*j}$ is the mle under H_0. We have $\hat{p}_{i*} = N_{i*}/n, \hat{p}_{*j} = N_{*j}/n$ by Lagrange multiplier. Writing $o_{ij} = N_{ij}, e_{ij} = n\hat{p}_{i*}\hat{p}_{*j}$, then

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

which can be modelled as a χ^2_p distribution with $p = (r-1)(c-1)$ by Wilk's theorem.
However, there are some problems with this χ^2 test for independence. It usually has low power as the test statistic (e.g. Pearson) can usually detect any deviation from independence. Often we suspect the data deviate in a specific way. So if we are confidence about this, then we can sometimes construct a test with higher power against such.
Another possible problem is small cells in the array might carry too much weight since e_{ij} can be small (in Pearson), which is undesirable. Berrett & Samworth proposed an alternative test using a different statistic namely

$$\hat{U} = \frac{1}{n(n-3)}\sum_{i,j}(o_{ij} - e_{ij})^2 - \frac{4}{n(n-2)(n-3)}\sum_{i,j} o_{ij}e_{ij}$$

that solves the problem.
One last problem is that we need large N_{ij} to use the χ^2 approximation (e.g.

$N_{ij} \geq 5$). The solution to this problem is the use of exact test.
Let π be a permutation of $\{1, \ldots, n\}$ and $Y_\pi = (Y_{\pi(1)}, \ldots, Y_{\pi(n)})$. Suppose $\pi^1, \ldots, \pi^B$ be i.i.d. permutations drawn uniformly from S_n.

Definition 3.11. A tuple of random variables $(Z_1, \ldots, Z_n)$ is exchangable if $(Z_1, \ldots, Z_n)$ and $\pi(Z_1, \ldots, Z_n)$ have the same distribution for any $\pi \in S_n$.

Proposition 3.4. *Under the null hypothesis of independence, the sets of data* $(X, Y), (X, Y_{\pi^1}), \ldots, (X, Y_{\pi^B})$ *are exchangable.*

Proof. Too tedious. Let's skip it. □

Corollary 3.5. *If* T_b *is a test statistic computed from* (X, Y_{π^b}) *for each* $b \in \{1, \ldots, B\}$ *and* T *is the test statistic computed from real data* (X, Y)*, then any ordering of* $T, T_1, \ldots, T_B$ *is equally likely.*
In particular, $\mathbb{P}_{H_0}(T = \max^*\{T, T_1, \ldots, T_B\}) = 1/(B+1)$ *where* $\max^*$ *is just* max *but break ties randomly.*

Proof. Follows pretty much immediately. □

So we can consider a test which rejects H_0 if $T = \max^*\{T, T_1, \ldots, T_B\}$, which has size exactly equal to $1/(B+1)$. This is known as the exact test.

Remark. 1. We do not need to know the null distribution of T analytically.
2. Since we did not restrict T, we can in particular apply it to the Pearson statistic

$$T(X, Y) = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

or the Berrett & Samworth statistic.

Example 3.11. Suppose we have $X = (a, a, b, b, a, b)$ and $Y = (c, c, d, d, d, d)$, then the contingency table N has the form

	a	b
c	2	0
d	1	3

where we cannot use the χ^2 approximation as the counts are too small. So naturally we want to try the exact test (wrt the Pearson statistic) of size $\alpha = 1/2$. The Pearson statistic has value

$$T = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = 3$$

Let $\pi^2 = (1\ 2\ 3)(4)(5)(6)$, then $Y_{\pi^2} = (d, c, c, d, d, d)$ so the modified table N' would be

	a	b
c	1	1
d	2	2

where $T_2 = 0 \implies T = \max\{T, T_2\}$, so we reject H_0 with size $1/2$.

3.5 Tests for Homogeneity

Example 3.12. Suppose we have 150 patients randomly allocated to 3 groups of equal size. Two sets of patients received a drug at 2 different doses while the third group received a placebo. Suppose we have the responses

	Improved	No Difference	Worse
Placebo	18	17	15
Half-Dose	20	10	20
Full-Dose	25	13	12
	63	40	47

How is this different from the model for contingency table? As one can see, the row totals in this table are fixed. What we want here is instead the null hypothesis asserting that the probability of each response (improved/no difference/worse) is the same for each treatment group.

With a table N shaped like in the example above, we have $N_{i1}, \dots, N_{ic} \sim \text{Multi}(n_i; p_{i1}, \dots, p_{ic})$ independently for each $i = 1, \dots, r$, where $n_i = N_{i*}$ is the fixed row total. The null hypothesis we are interested in is $H_0 : p_{1j} = \dots = p_{rj}$ for all $j = 1, \dots, c$. We take the alternative as unconstrained as per usual. Then, under H_1, we have

$$L(p) = \prod_{i=1}^{r} \frac{N_{i*}!}{N_{i1}! \cdots N_{ic}!} p_{i1}^{N_{i1}} \cdots p_{ic}^{N_{ic}} \implies \ell(p) = \sum_{i,j} N_{ij} \log p_{ij} + \text{const.}$$

We have the constraints $\forall i, \sum_j p_{ij} = 1$, so the mle is $\hat{p}_{ij} = N_{ij}/N_{i*}$ by Lagrange multiplier. Under H_0, let $p_j = p_{ij}$ for all i, then

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

Again using Lagrange multiplier with constraint $\sum_j p_j = 1$ we obtain the mle $\hat{p}_j = N_{*j}/N_{**}$. Hence

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

which is exactly the same statistic we found when computing $2 \log \Lambda$ for independence test. Furthermore, if we define again $o_{ij} = N_{ij}, e_{ij} = N_{i*} N_{*j}/N_{**}$, then we can write

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

which is again the same Pearson statistic we used to test independence. Wilk's theorem then tells us that the distribution of $2 \log \Lambda$ is asymptotically χ^2_p with $p = \dim(\Theta_1) - \dim(\Theta_0) = r(c-1) - (c-1) = (r-1)(c-1)$, so the number of degrees of freedom is also the same as in the test of independence. In conclusion, the test of independence and homogeneity always reach the same conclusions (but sadly, they might be different for 3-dimensional tables).

Example 3.13. In the example quoted previously, we have $2 \log \Lambda = 5.129$ and $\sum_{i,j} (o_{ij} - e_{ij})^2 / e_{ij} = 5.173$. As $\chi^2_4(0.05) = 9.488$, H_0 is not rejected at level 0.05.

3.6 Tests and Confidence Intervals

Definition 3.12. The acceptance region A of a test is the complement of the critical region.

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

Theorem 3.6. *1. Suppose for each $\theta_0 \in \Theta$, there is 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 $1-\alpha$ confidence set for θ.*
2. Suppose $I(X)$ is a $1-\alpha$ 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 $\theta_0 \in \Theta$.

Proof. Note that we have $\theta_0 \in I(X) \iff X \in A(\theta_0)$ in both cases. So for 1 we have

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

And for 2 we have

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

which is what we want. □

Example 3.14. Suppose $X_1, \dots, X_n \sim N(\mu, \sigma_0^2)$ are i.i.d. with known σ_0^2, We know that

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}$$

gives a $1-\alpha$ confidence interval for μ. So we can find a size α test for $H_0 : \mu = \mu_0$ by defining the acceptance region

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

which is saying that we reject H_0 when $|\sqrt{n}(\mu_0 - \bar{x})/\sigma_0| > z_{\alpha/2}$ which is the z-test we already found by LRT. We can of course do the opposite direction as well.

4 Multivariate Normals

4.1 Definition and Properties

Recall that for a vector $X = (X_1, \dots, X_n)^\top$ be a vector of random variables. Recall that its expectation is $\mathbb{E}X = (\mathbb{E}X_1, \dots, \mathbb{E}X_n)^\top$ and its variance (or covariance matrix) is $\operatorname{var} X = \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^\top] = (\operatorname{cov}(X_i, X_j))_{i,j}$. Also recall that if A is an $m \times n$ matrix and b is an m-vector, then $\mathbb{E}(AX + b) = A\mathbb{E}X + b, \operatorname{var}(AX + b) = A \operatorname{var}(X)A^\top$.

Definition 4.1. We say X has a multivariate normal (MVN) distribution if for any $t \in \mathbb{R}^n$ the random variable $t^\top X$ has a normal distribution.

Proposition 4.1. *If X is an MVN, so is $AX + b$.*

Proof. $t^\top(Ax + b) = (A^\top t)^\top X + t^\top b$. □

Proposition 4.2. *An MVN distribution is fully determined by its mean and variance.*

Proof. It suffices to show that the mgf of an MVN X depends only on $\mu = \mathbb{E}X, \Sigma = \operatorname{var}(X)$. Indeed,

$$\mathbb{E}e^{t^\top X} = M_{t^\top X}(1) = \exp\left(\mathbb{E}(t^\top X) + \frac{1}{2}\operatorname{var}(t^\top X)\right) = \exp\left(t^\top \mu + \frac{1}{2}t^\top \Sigma t\right)$$

via direct computation. □

Definition 4.2. We say $P \in \mathbb{R}^{n\times n}$ is an orthogonal projection if it is idempotent $P^2 = P$ and symmetric $P^\top = P$. Equivalently, P is an orthogonal projection if $Pv = v$ for all v in the column space of P and $Pw = 0$ for all w in the orthogonal complement of the column space of P.

Proposition 4.3. *The two definitions stated above are indeed equivalent.*

Proof. Simple linear algebra but let's do it.
If P is an orthogonal projection as in the first definition, then for any v in the column space of P we have $v = Pa$ for some a. Then $Pv = P^2a = Pa = v$. Also, if w is in the orthogonal complement of the column space of P, then $Pw = P^\top w = 0$.
Conversely, if P satisfies the properties in the second definition, then note that we can write any $a \in \mathbb{R}^n$ uniquely as $a = v + w$ for some v in the column space of P and w in the orthogonal complement of the column space of P. Then we always have $Pa = v$. In particular, $P^2a = PP(v + w) = Pv = P(v + w) = Pa$ which is idempotence. For symmetry, just observe that $(Pu_1)^\top((I - P)u_2) = 0$ for any u_1, u_2. Expanding it gives the result. □

Corollary 4.4. *If P is an orthogonal projection, so is $I - P$.*

Proof. Obvious. □

Proposition 4.5. *If $P \in \mathbb{R}^{n\times n}$ is an orthogonal projection, then $P = UU^\top$ where $k = \operatorname{rank}(P)$ and the columns of $U \in \mathbb{R}^{n\times k}$ is chosen to be an orthonormal basis for the column space of P.*

Proof. Clearly $UU^\top$ is an orthogonal projection whose column space coincides with the column space of P. □

Note that $k = \operatorname{rank} P = \operatorname{tr}(U^\top U) = \operatorname{tr}(UU^\top) = \operatorname{tr} P$, which is a fact that sometimes comes in handy.

Theorem 4.6. *If $X \sim N(0, \sigma^2 I)$ and P is an orthogonal projection, then:*
1. $PX \sim N(0, \sigma^2 P), (I - P)X \sim N(0, \sigma^2(I - P))$ are independent.
2. $\|PX\|^2/\sigma^2 \sim \chi^2_{\operatorname{rank} P}$

Proof. We know that

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

is MVN and its distribution is fully specified by its mean and variance:

$$\mathbb{E}\begin{pmatrix} PX \\ (I - P)X \end{pmatrix} = \begin{pmatrix} P \\ I - P \end{pmatrix} \mathbb{E}X = 0$$

$$\operatorname{var}\begin{pmatrix} PX \\ (I-P)X \end{pmatrix} = \begin{pmatrix} P \\ (I-P) \end{pmatrix} \sigma^2 I \begin{pmatrix} P \\ (I-P) \end{pmatrix}^\top = \sigma^2 \begin{pmatrix} P & 0 \\ 0 & I-P \end{pmatrix}$$

as P is an orthogonal projection. This exactly means that $PX, (I-P)X$ are independent.
For the second part of the theorem, observe that

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

But $U^\top X \sim N(0, \sigma^2 U^\top U) = N(0, \sigma^2 I)$, so indeed $(U^\top X)_i/\sigma$ are i.i.d. $N(0,1)$ random variables. This shows the result. □

How does this apply to statistics? Suppose we have i.i.d. normal observations $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ with unknown mean and variance. Recall that the mle of μ is $\bar{X} = n^{-1} \sum_i X_i$ and the mle of σ^2 is $\hat{\sigma}^2 = S_{XX}/n$ where $S_{XX} = \sum_i (X_i - \bar{X})^2$.

Theorem 4.7. *1.* $\bar{X} \sim N(\mu, \sigma^2/n)$.
2. $S_{XX}/\sigma^2 \sim \chi^2_{n-1}$.
3. $\bar{X}$ *and* S_{XX} *are independent.*

Proof. Let P be a matrix such that $PX = J\bar{X}$ where $J = (1, \dots, 1)^\top \in \mathbb{R}^n$. Explicitly, $P = n^{-1} JJ^\top$. Easy to check that P is an orthogonal projection. Now we can write $X = \mu J + \epsilon$ with $\epsilon \sim N(0, \sigma^2 I)$. Note that $\bar{X}$ is a function of $PX = \mu J + P\epsilon$. Also $S_{XX} = \sum_i (X_i - \bar{X})^2 = \|X - J\bar{X}\|^2 = \|(I-P)X\|^2 = \|(I-P)\epsilon\|^2$. The independence of $P\epsilon$ and $(I-P)\epsilon$ then shows the third part of the theorem.
The first part is obvious. The second part can be deduced from the previous theorem by writing $S_{XX}/\sigma^2 = \|(I-P)\epsilon\|^2/\sigma^2$. □

4.2 The (Normal) Linear Model

Example 4.1. Suppose you are working in insurance and would like to predict the number Y_i of insurance claims that a car owner i submits in a certain year. We call this the response (or dependent) variable. What we want to know is the statistical relationship between the response and certain predictors (or independent variables), e.g. the age x_{i1} of person i, the number of their insurance claims x_{i2} last year, the number of years x_{i3} they had the licence, the number of miles x_{i4} they've driven in 2019, etc.. Naturally, we will want to know about this relationship from paired data $\{(x_i, Y_i)\}$ with $Y_n \in \mathbb{R}, x_n \in \mathbb{R}^p$ for some p.

One common approach to this problem is the linear model.

Definition 4.3. The linear model is

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

Here, α is called the intercept, $\beta_1, \dots, \beta_p$ are called the coefficients. $\epsilon_1, \dots, \epsilon_n$ are random variables representing the noise, which are assumed to have mean 0.

Remark. 1. We usually elimnate the intercept by adding in the predictor $x_{i1} \equiv 1$ (so that β_1 plays the role of α).
2. We can, of course, model certain non-linear relationships using the linear model by taking individual (possibly nonlinear) terms as predictors. For example, the model $Y_i = \alpha + \beta_1 z_i + \beta_2 z_i^2 + \epsilon_i$ can be rewritten as $Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$ where $x_{i1} = z_i, x_{i2} = z_i^2$.
3. The coefficients β_j can be interprested as the effect on the response by changing x_{ij}. However, in most cases linear model can only reveal correlations between predictors and response instead of causal relationships. Adding predictors to the model can reduce the risk of finding such spurious correlations.

Writing

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

transforms the linear model into the form $Y = X\beta + \epsilon$. Here, X is called the design matrix, which we shall always assume to be fixed. We make the assumption that $\mathbb{E}\epsilon = 0$ and $\operatorname{var}\epsilon = \sigma^2 I$ (known as homoskedasticity, and yes, it is an actual word). Consequently, $\mathbb{E}Y_i = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p$ and $\operatorname{var} Y = \sigma^2 I$. We also assume that X has full rank, in particular, $p \leq n$.

Definition 4.4. The least square estimator $\hat{\beta}$ minimises the residual sum of squares (RSS) $S(\beta) = \|Y - X\beta\|^2 = \sum_i (Y_i - x_i^\top \beta)^2$.

Calculus gives $X^\top X\hat{\beta} = X^\top Y$. As we have assumed that X has full rank, we get the explicit solution $\hat{\beta} = (X^\top X)^{-1} X^\top Y$. In particular, the least squares estimator is linear in Y.
We also know the mean and variance of $\hat{\beta}$:

$$\begin{aligned} \mathbb{E}\hat{\beta} &= \mathbb{E}((X^\top X)^{-1} X^\top Y) = (X^\top X)^{-1} X^\top \mathbb{E}Y = (X^\top X)^{-1} X^\top X\beta = \beta \\ \operatorname{var}\hat{\beta} &= \operatorname{var}((X^\top X)^{-1} X^\top Y) = (X^\top X)^{-1} X^\top \operatorname{var}(Y) X (X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1} \end{aligned}$$

In particular, $\hat{\beta}$ is an unbiased estimator of β. In fact, it is the best one of its kind.

Theorem 4.8 (Gauss-Markov)**.** *Let $\beta^* = CY$ be any other linear estimator of β which is unbiased (i.e. $\mathbb{E}\beta^* = \beta$). Then for any vector $t \in \mathbb{R}^p$, we have $\operatorname{var}(t^\top \hat{\beta}) \leq \operatorname{var}(t^\top \beta^*)$.*

So the least squares estimator is the best linear unbiased estimator (BLUE).

Remark. $t = (t_1, \ldots, t_p)$ can be thought of as the values of the predictors for a new sample. In this case, $t^\top \hat{\beta}, t^\top \beta^*$ are our prediction for the mean response. What the theorem says is then that $\hat{\beta}$ always gives smaller variance and hence smaller mse since the estimators are unbiased.

Proof. Note that $\operatorname{var}(t^\top \beta^*) - \operatorname{var}(t^\top \hat{\beta}) = t^\top (\operatorname{var}\beta^* - \operatorname{var}\hat{\beta}) t$. So it suffices to show that $\operatorname{var}\beta^* - \operatorname{var}\hat{\beta}$ is positive semidefinite. Let $A = C - (X^\top X)^{-1} X^\top$,

then $\mathbb{E}(AY) = 0$ since the estimators are unbiased. Also, for any β we have $AX\beta = A\mathbb{E}Y = \mathbb{E}(AY) = 0$, so essentially $AX = 0$. Therefore

$$\begin{aligned}\operatorname{var}(\beta^*) &= \operatorname{var}((A + (X^\top X)^{-1}X^\top)Y) \\ &= (A + (X^\top X)^{-1}X^\top)\operatorname{var}(Y)(A + (X^\top X)^{-1}X^\top)^\top \\ &= \sigma^2(A + (X^\top X)^{-1}X^\top)(A + (X^\top X)^{-1}X^\top)^\top \\ &= \sigma^2 AA^\top + \operatorname{var}(\hat{\beta})\end{aligned}$$

The result follows. □

Definition 4.5. The fitted values are $\hat{Y} = X\hat{\beta}$ and the residuals are $Y - \hat{Y}$.

If we write $P = X(X^\top X)^{-1}X^\top$ as the "hat matrix", then we simply have $\hat{Y} = PY, Y - \hat{Y} = (I - P)Y$.

Proposition 4.9. *P is the orthogonal projection onto the column space of X.*

Proof. P is symmetric, idempotent, and has the same column space as X. □

So the fitted values $\hat{Y}$ is the orthogonal projection of Y onto the column space of X and the residuals $(I - P)Y$ are the orthogonal projection of Y onto the orthogonal complement of the column space of X.

Definition 4.6. A linear model is called a normal linear model if ϵ is a multivariate normal.

In the case of a normal linear model, we know all the information of the distribution of Y given β, σ^2. Indeed,

$$f_Y(y|\beta,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^\top\beta)^2\right)$$

The log-likelihood is

$$\ell(\beta,\sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}S(\beta)$$

So the least squares estimator $\hat{\beta}$ is, indeed, also the mle of β. We also have

$$\frac{\partial\ell}{\partial\sigma^2}(\hat{\beta},\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}S(\hat{\beta})$$

Therefore the mle of σ^2 is

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

Corollary 4.10. *1. $\hat{\beta} \sim N(\beta, \sigma^2(X^\top X)^{-1})$.*
2. $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$.
3. $\hat{\beta}, \hat{\sigma}^2$ are independent.

Proof. One can simply check 1 from definition. Theorem 4.6 and Proposition 4.9 implies 2 and 3. □

Note that $\hat{\sigma}^2$ is a biased estimator since

$$\mathbb{E}\left(\frac{\hat{\sigma}^2 n}{\sigma^2}\right) = \mathbb{E}(\chi^2_{n-p}) = n - p \implies \mathbb{E}(\hat{\sigma}^2) = \frac{n-p}{n}\sigma^2 < \sigma^2$$

but it is asymptotically unbiased if we fix p and send $n \to \infty$. We can, of couse, consider the unbiased $\tilde{\sigma}^2 = n\hat{\sigma}^2/(n - p)$.

4.3 Two Useful Distributions

Definition 4.7. Let $U \sim N(0,1), V \sim \chi^2_n$ be independent, then $T = U/\sqrt{V/n}$ has a t_n distribution.

Definition 4.8. Let $V \sim \chi^2_n, W \sim \chi^2_m$ be independent, then $F = (V/n)/(W/m)$ has an $F_{n,m}$ distribution.

Remark. $t_n^2 = F_{1,n}$.

Example 4.2. Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ are i.i.d.. When σ^2 is known, we found a $(1-\alpha)$ confidence interval for μ that is $[\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}]$ noting $Z = \sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0,1)$. What if we don't know σ^2? Recall that $\bar{X}$ and $\hat{\sigma}^2 = S_{XX}/n$ are independent with $\bar{X} \sim N(\mu, \sigma^2/n)$ and $S_{XX}/\sigma^2 \sim \chi^2_{n-1}$. Therefore, the random variable

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{S_{XX}/(n-1)}}$$

is t_{n-1}. Consequently we have

$$\mathbb{P}_{\mu,\sigma^2}\left(-t_{n-1}(\alpha/2) \leq \frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{S_{XX}/(n-1)}} \leq t_{n-1}(\alpha/2)\right) = 1 - \alpha$$

which gives the $(1-\alpha)$ confidence interval

$$\left[\bar{X} - \sqrt{\frac{S_{XX}}{n-1}}\frac{t_{n-1}(\alpha/2)}{\sqrt{n}}, \bar{X} + \sqrt{\frac{S_{XX}}{n-1}}\frac{t_{n-1}(\alpha/2)}{\sqrt{n}}\right]$$

for μ. Suppose that we wish to test $H_0 : \mu = \mu_0$ with σ^2 unknown, then under H_0,

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sqrt{S_{XX}/(n-1)}} \sim t_{n-1}$$

Therefore the test that rejects H_0 when $|T| > t_{n-1}(\alpha/2)$ has size α. This is known as the t-test.

Note. Asymptotically, as $n \to \infty$, we have $S_{XX}/(n-1) \approx \sigma^2$ and $t_{n-1}(\alpha/2) \to z_{\alpha/2}$, so the t-test behaves just like the z-test for large n, or if we substitute $S_{XX}/(n-1)$ as an estimator for σ^2. The t-test is useful in the situations where n is small and $X_1, \ldots, X_n$ are approximately normal.

4.4 Inferences in the Normal Linear Model

We wish to find a $1-\alpha$ confidence interval for one of the coefficients in the linear model, WLOG β_1. Note that

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

which is independent of $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$. So, naturally, we want to consider

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{((X^\top X)^{-1})_{11}}}\sqrt{\frac{n-p}{n\hat{\sigma}^2}} = \frac{(\hat{\beta}_1 - \beta_1)/\sqrt{\sigma^2((X^\top X)^{-1})_{11}}}{\sqrt{n\hat{\sigma}^2/((n-p)\sigma^2)}} \sim t_{n-p}$$

The expression is independent of σ (yay), so we have

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

Rearranging gives the $1-\alpha$ confidence interval

$$\left[\hat{\beta}_1 - t_{n-p}\left(\frac{\alpha}{2}\right)\sqrt{\frac{((X^\top X)^{-1})_{11}\hat{\sigma}^2 n}{n-p}}, \hat{\beta}_1 + t_{n-p}\left(\frac{\alpha}{2}\right)\sqrt{\frac{((X^\top X)^{-1})_{11}\hat{\sigma}^2 n}{n-p}}\right]$$

What if we want a nice confidence set for the vector β? We have $\hat{\beta} - \beta \sim N(0, \sigma^2(X^\top X)^{-1})$. As X has full rank, $X^\top X$ is positive definite, so $X^\top X = UDU^\top$ where U is orthogonal and D diagonal with positive diagonal entries. Define $\sqrt{X^\top X} = U\sqrt{D}U^\top$ where $(\sqrt{D})_{ij} = \sqrt{D_{ij}}$. Then $(\sqrt{X^\top X})^2 = X^\top X$ and $\sqrt{X^\top X}(\hat{\beta}-\beta) \sim N(0, \sigma^2 I)$, hence $\|X(\hat{\beta}-\beta)\|/\sigma^2 = \|\sqrt{X^\top X}(\hat{\beta}-\beta)\|^2/\sigma^2 \sim \chi^2_p$ is independent of $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$. So

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

Consequently,

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

This gives the $1-\alpha$ confidence set

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

which is essentially an ellipsoid.
How about confidence in predictions of the linear model? Suppose we have data $(x_1, Y_1), \ldots, (x_n, Y_n)$ and wish to predict the response at a new input point $x^* \in \mathbb{R}^p$.

Example 4.3. Again we imagine we are selling our souls to an insurance company. Suppose we have fitted a linear model to historical data on insurance claims and wish to predict insurance claims for a new customer x^*. If $Y^* = (x^*)^\top\beta + \epsilon^*$ with $\epsilon^* \sim N(0, \sigma^2)$ independent of $\epsilon_1, \ldots, \epsilon_n$, then we may want to obtain a $1-\alpha$ confidence interval for the response $\mathbb{E}Y^* = (x^*)^\top\beta$. Or we might want something even better: A random interval which contains Y^* with probability $1-\alpha$. This is called a prediction interval.

We have $(x^*)^\top\hat{\beta} \sim N((x^*)^\top\beta, \sigma^2(x^*)^\top(X^\top X)^{-1}x^*)$, so

$$\frac{((x^*)^\top\hat{\beta} - (x^*)^\top\beta)/\sqrt{\sigma^2(x^*)^\top(X^\top X)^{-1}x^*}}{\sqrt{n\hat{\sigma}^2/((n-p)\sigma^2)}} \sim t_{n-p}$$

Again those σ cancels out and allows us to give the $1-\alpha$ confidence interval

$$\left[(x^*)^\top\hat{\beta} \pm t_{n-p}\left(\frac{\alpha}{2}\right)\sqrt{\frac{(x^*)^\top(X^\top X)^{-1}x^*\hat{\sigma}^2 n}{n-p}}\right]$$

for $(x^*)^\top \beta$.
How about the prediction interval? Note that we have

$$Y^* - (x^*)^\top \hat{\beta} = (x^*)^\top (\beta - \hat{\beta}) + \epsilon^* \sim N(0, \sigma^2 (x^*)^\top (X^\top X)^{-1} x^* + \sigma^2)$$

which means

$$\frac{(Y^* - (x^*)^\top \hat{\beta})/(\sigma \sqrt{1 + (x^*)^\top (X^\top X)^{-1} x^*})}{\sqrt{n\hat{\sigma}^2/((n-p)\sigma^2)}} \sim t_{n-p}$$

Again cancelling σ gives the $1-\alpha$ prediction interval

$$\left[(x^*)^\top \hat{\beta} \pm t_{n-p}\left(\frac{\alpha}{2}\right) \sqrt{\frac{(1 + (x^*)^\top (X^\top X)^{-1} x^*)\hat{\sigma}^2 n}{n-p}} \right]$$

which, as one shall expect, is wider than the confidence interval.

4.5 Hypothesis Testing

Suppose we want to test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. We already know that

$$I(Y) = \left[\hat{\beta}_1 \pm t_{n-p}\left(\frac{\alpha}{2}\right) \sqrt{\frac{((X^\top X)^{-1})_{11}\hat{\sigma}^2 n}{n-p}} \right]$$

is a $1-\alpha$ confidence interval for β_1, therefore by Theorem 3.6, the acceptance region

$$A = \{y : 0 \in I(y)\}$$

is a test of size α. In other words, we reject H_0 when 0 falls outside the $1-\alpha$ confidence interval. This is equivalent to say that we reject H_0 when

$$|\hat{\beta}_1| > t_{n-p}\left(\frac{\alpha}{2}\right) \sqrt{\frac{n\hat{\sigma}^2((X^\top X)^{-1})_{11}}{n-p}}$$

This is called the t-test.
There is, of course, nothing special about 0. The same procedure can give a test of size α for $H_0 : \beta_1 = \beta_1'$ for any constant β_1'.
Now suppose that we wish to test the null hypothesis that a collection of coefficients has no effect on the response. WLOG we take $H_0 : \beta_1 = \cdots = \beta_{p_0} = 0$ against $H_1 : (\beta_1, \ldots, \beta_{p_0}) \in \mathbb{R}^{p_0}$. We shall make use of the likelihood ratio test. Write

$$X = \begin{pmatrix} X_0 & X_1 \end{pmatrix}, \beta = \begin{pmatrix} \beta^0 \\ \beta^1 \end{pmatrix}$$

where X_0 is an $n \times p_0$ matrix and $\beta^0 \in \mathbb{R}^{p_0}$. The null hypothesis means $\beta^0 = 0$, so under H_0, $Y = X_1 \beta^1 + \epsilon$. Let $P = X(X^\top X)^{-1} X^\top$ be the usual hat matrix and $P_1 = X_1 (X_1^\top X_1)^{-1} X_1^\top$. As X, P have full rank, so do X_1, P_1. Recall that the maximum log-likelihood in the normal linear model is

$$\sup_{\beta \in \mathbb{R}^p, \sigma^2 > 0} \ell(\beta, \sigma^2) = \ell(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2} \log\left(\frac{\|(I-P)Y\|^2}{n} \right) + \text{const.}$$

Hence, the GLR statistic is

$$\begin{aligned} 2\log\Lambda &= n\left(-\log\left(\frac{\|(I-P)Y\|^2}{n}\right)+\log\left(\frac{\|(I-P_1)Y\|^2}{n}\right)\right) \\ &= n\log\left(\frac{\|(I-P_1)Y\|^2}{\|(I-P)Y\|^2}\right) \end{aligned}$$

Lemma 4.11. *1.* $(I-P)(P-P_1)=0$.
2. $P-P_1$ *is an orthogonal projection with rank* p_0.

Proof. Simple verification noting that $P_1P = PP_1 = P_1$. □

So

$$\begin{aligned} \frac{\|(I-P_1)Y\|^2}{\|(I-P)Y\|^2} &= \frac{\|(I-P+P-P_1)Y\|^2}{\|(I-P)Y\|^2} \\ &= \frac{\|(I-P)Y\|^2+\|(P-P_1)Y\|^2+2(I-P)(P-P_1)Y}{\|(I-P)Y\|^2} \\ &= 1+\frac{\|(P-P_1)Y\|^2}{\|(I-P)Y\|^2} \end{aligned}$$

Note also that $(P-P_1)\epsilon$ and $(I-P)\epsilon$ are independent since $(P-P_1)(I-P)=0$, so the statistic is in fact monotone in

$$F = \frac{\|(P-P_1)Y\|^2}{\|(I-P)Y\|^2}\frac{p_0^{-1}}{(n-p)^{-1}} = \frac{\|(P-P_1)\epsilon\|^2/p_0}{\|(I-P)\epsilon\|^2/(n-p)} \sim F_{p_0,n-p}$$

This gives a test (called the F-test) of size α that rejects H_0 when

$$\frac{\|(P-P_1)Y\|^2}{\|(I-P)Y\|^2}\frac{p_0^{-1}}{(n-p)^{-1}} > F_{p_0,n-p}(\alpha)$$

Remark. 1. When $p_0=1$, the F-test reduces to the t-test.
2. One can derive the power function for the test as well (exercise).

What if the predictors are categorical?

Example 4.4. Let Y_i be the level of COVID-19 neutralising antibodies in a subject i. Suppose we want to analyze its relation with the predictors $z_i \in \{\text{control}, \text{vaccine A}, \text{vaccine B}\}$ the treatment received. We can represent the predictors by

$$x_{i,j} = 1_{\text{subject } i \text{ received treatment } j}$$

and test the model $Y_i = \alpha+\beta_1 x_{i,1}+\beta_2 x_{i,2}+\beta_3 x_{i,3}+\epsilon$. However, there is a big problem with this model: The design matrix does not have full rank due to the intercept term α. When there is a single categorical predictor, we can of course solve this by simply removing the intercept term. But we can't do that when there is more than one categorical predictor. The solution to this problem is the following: We apply a corner point constraint, i.e. choose a baseline category (e.g. the control treatment) and set its coefficient to 0. The column space of X remains the same no matter which category we choose as baseline, hence the projection P and the fitted values do not depend on that.

4.6 Applications of Normal Linear Model

Special cases of normal linear model often pose practical importance. THe first one of interest is the analysis of variance (ANOVA) model.

Example 4.5. Consider the usual situation where we'd like to model the level of antibodies in patient receiving one of three treatments: central, vaccine A, vaccine B. We can formulate the model in a new way other than the one we did previously. Let Y_{ij} be the response for subject i in treatment group j and we model $Y_{ij} = \alpha + \mu_j + \epsilon_{ij}$. where $i \in \{1, \dots, N\}$ and $j \in \{1, 2, 3\}$ and $\epsilon_{ij} \sim N(0, \sigma^2)$ are i.i.d.. We use the corner-point constraint $\mu_1 = 0$.
The ANOVA nest does the following: We want to test $\mu_2 = \mu_3 = 0$ against $H_1 : \mu_2, \mu_2 \in \mathbb{R}$. The null hypothesis is saying that $\mathbb{E}Y_{ij} = \alpha$ for all i, j, i.e. each treatment has the same mean response. We can view this as a special case of the F-test by taking $n = 3N$ and concatenating everything. The design matrix then has the form

$$X = \left(\begin{array}{ccc} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ \hline 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ \hline 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{array}\right)$$

Write $X = (X_1, X_0)$ where X_1 is the first column. Recall that the F-test uses the statistic

$$\frac{\|(P - P_1)Y\|^2/p_0}{\|(I - P)Y\|^2/(n - p)}$$

which is referred to an $F_{p,n-p}$ distribution. For the ANOVA test, P projects onto the space of vectors in $\mathbb{R}^{3N}$ which are constant over treatment groups. Take $\bar{Y}_j = N^{-1} \sum_{i=1}^{N} T_{ij}$, then

$$PY = (\bar{Y}_1, \dots, \bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_2, \bar{Y}_3, \dots, \bar{Y}_3)^\top$$

In this case, P_1 projects onto the space of constant vectors in $\mathbb{R}^{3N}$. Precisely $P_1 Y = (\bar{Y}, \dots, \bar{Y})^\top, \bar{Y} = (3N)^{-1} \sum_{i,j} Y_{ij}$. The F statistic is then reduced to the form

$$\left(\frac{1}{2}\sum_{j=1}^{3} N(\bar{Y}_j - \bar{Y})^2\right) \Bigg/ \left(\frac{1}{3N-3}\sum_{i=1}^{N}\sum_{j=1}^{3}(Y_{ij} - \bar{Y}_j)^2\right)$$

Of course, there is nothing special about 3 here. We can easily generalise this to $J \geq 3$ treatment groups where the test statistic has the form

$$\left(\frac{1}{J-1}\sum_{j=1}^{J} N(\bar{Y}_j - \bar{Y})^2\right) \Bigg/ \left(\frac{1}{JN-J}\sum_{i=1}^{N}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_j)^2\right)$$

which can be seens as the fraction of "variance between treatments" and "variance within treatments", hence the name ANOVA.

Remark. This is sometimes called one-way ANOVA. There is also two-way ANOVA where a similar analysis is carried out in an experiment where groups are defined according to 2 variables.

Example 4.6. Suppose we wish to analyse the response which is a student's performance in an exam. The attributes we are interested in are whether or not the student has completed the revision supervisions and whether or not a monetary incentive was given. Denote these attributes as $(j,k) \in \{0,1\}^2$, then we naturally want to model this as $Y_{ijk} = \alpha + \mu_j + \lambda_k + \epsilon_{ijk}$ where ϵ_{ijk} are i.i.d. $N(0,\sigma^2)$ and we use the corner point constraint $\mu_0 = \lambda_0 = 0$. Then the 2-way ANOVA test is the F test applied to $H_0 : \mu_1 = \lambda_1 = 0$ against $H_1 : \mu_1, \lambda_1 \in \mathbb{R}$.

We normally want to centre the predictors

$$Y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \epsilon \sim N(0, \sigma^2 I)$$

which gives rise to what's called a simple linear regression. In this case, the mle actually takes a simple form. Recall that $(\hat{\alpha}, \hat{\beta})$ minimises

$$S(\alpha,\beta) = \sum_{i=1}(Y_i - \alpha - \beta(x_i - \bar{x}))^2$$

Differentiation gives $\hat{\alpha} = \bar{Y}$ and

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}}, S_{XY} = \sum_i (Y_i - \bar{Y})(X_i - \bar{X}), S_{XX} = \sum_i (X_i - \bar{X})^2$$

We now turn to two-sample testing. Suppose that σ^2 is fixed and that the sequences $Z_1, \dots, Z_n \sim N(\mu_1, \sigma^2)$ and $Z'_1, \dots, Z'_m \sim N(\mu_2, \sigma^2)$ are i.i.d. respectively. Suppose also that the sequences (Z_i) and (Z'_i) are independent. We wish to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$. Indeed, we can view this as a special case of an F-test in a linear model with

$$Y = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \\ Z'_1 \\ \vdots \\ Z'_m \end{pmatrix}, X = \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix}, \beta = \begin{pmatrix} \mu_2 \\ \mu_1 - \mu_2 \end{pmatrix}$$

which transforms the hypotheses to $H_0 : \beta_2 = 0, H_1 : \beta_2 \neq 0$. Sadly though, sometimes it is not really reasonable to assume they have normal distributions. Suppose we just assume that $Z_1, \dots, Z_n \sim f_1, Z'_1, \dots, Z'_m \sim f_2$ are independent and attempt to test $H_0 : f_1 = f_2$ against $H_1 : f_1 \neq f_2$. How do we, then, produce a test with exact size α? Let $T(z_1, \dots, z_n, z'_1, \dots, z'_m)$ be a test statistic. For a permutation π of $\{1, \dots, n+m\}$, we can define $T_\pi = T \circ \pi$. Analogous to Proposition 3.4, we have

Lemma 4.12. *If $\pi_1, \ldots, \pi_B$ are i.i.d. permutations drawn uniformly in S_B, then under H_0 the sequence $T, T_{\pi_1}, \ldots, T_{\pi_B}$ is exchangable. Consequently, we have $\mathbb{P}_{H_0}(T = \max^*\{T, T_1, \ldots, T_B\}) = 1/(B+1)$ where $\max^*$ is just $\max$ but break ties randomly.*

So we can build a test with exact size $\alpha = 1/(B+1)$ by rejecting H_0 when $T = \max^*\{T, T_{\pi_1}, \ldots, T_{\pi_B}\}$.