STAT2004 Content Notes

Ryan White
s4499039

Semester 2, 2022

Contents

1 Week 1

Point Estimation

Suppose that a random variable $\boldsymbol{X}$ is discrete with probability function $f(\boldsymbol{x};\boldsymbol{\theta})$ or is continuous with probability density function $f(x;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a parameter vector belonging to some parameter space Ω. e.g.

$$f(x;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

if $X \sim N(\mu, \sigma^2)$. Here $\boldsymbol{\theta}$ is a vector with $\boldsymbol{\theta} = (\mu, \sigma^2)^T$, where $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathrm{Var}(X)$.
The random variable $\boldsymbol{X}$ can be a vector of p dimensions ($p = 1$ is referred to as the univariate case, while $p > 1$ is the multivariate case).
$\boldsymbol{\theta}$ will be understood to contain all of the unknown feature of the probability function (p.f.) or probability density function (p.d.f) of the random variable $\boldsymbol{X}$. Often, only a part of $\boldsymbol{\theta}$ may be of interest. For example, if $\boldsymbol{\theta} = (\mu, \sigma^2)^T$, interest may be founded exclusively on the estimation of the mean μ, in which the variance would be referred to as a nuisance parameter. If we were to estimate μ by the sample mean $\overline{X}$, we need to estimate σ^2 if we are to provide an estimate of the standard deviation of $\overline{X}$.

$$\begin{aligned}\mathrm{Var}\left(\overline{X}\right) &= \sigma^2/n \\ \mathrm{S.D.}\left(\overline{X}\right) &= \sigma/\sqrt{n} \\ \mathrm{S.E.}\left(\overline{X}\right) &= \hat{\sigma}/\sqrt{n}\end{aligned}$$

where $\hat{\sigma}^2 = \sum_{j=1}^n (x_j - \overline{x})^2/n$.

We concentrate on the case where $\boldsymbol{x_1, ..., x_n}$ denote the observed values of the random variables of $\boldsymbol{X_1, ..., X_n}$. which are taken to be independently distributed with common p.f. or p.d.f., i.e.

$$\boldsymbol{X_1, ..., X_N} \overset{\text{iid}}{\sim} F_\theta$$

where F_θ is the distribution function (F.D.) corresponding to the p.f./p.d.f. In other words, $\boldsymbol{x_1, ..., x_n}$ constitute an observed random sample.

In point estimation, we estimate $\boldsymbol{\theta}$ by a simple point which is taken to be some function of the observed data $\boldsymbol{x_1, ..., x_n}$. That is, we estimate $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$, where

$$\hat{\boldsymbol{\theta}} = \boldsymbol{T}(\boldsymbol{x_1, ..., x_n})$$

and $\boldsymbol{T}$ is some function of $\boldsymbol{x_1, ..., x_n}$. We refer to $\boldsymbol{T}(\boldsymbol{x_1, ..., x_n})$ as an estimate, and to the corresponding random variable $\boldsymbol{T}(\boldsymbol{X_1, ..., X_n})$ as an estimator. We may view $\hat{\boldsymbol{\theta}}$ as our "educated guess" for the unknown value of $\boldsymbol{\theta}$.
If the bias of $\boldsymbol{T}$ is zero for all $\boldsymbol{\theta} \in \boldsymbol{\Omega}$, i.e.

$$\mathbb{E}(\boldsymbol{T}(\boldsymbol{X_1, ..., X_n})) = \boldsymbol{\theta}, \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Omega}$$

then we say that an estimator is unbiased.

It will be seen that if $X \sim N(\mu, \sigma^2)$, then the unbiased estimator of μ with the smallest variance for any value of $\boldsymbol{\theta} = (\mu, \sigma^2)^T$ is the sample mean $\overline{X}$,

$$\overline{X} = \frac{1}{n}\sum_{j=1}^n X_j$$

Note that if $f(x;\boldsymbol{\theta})$ is the p.d.f. of a continuous r.v. X, we define the median $q_{0.5}$ by that value for which

$$0.5 = \int_{-\infty}^{q_{0.5}} f(u;\boldsymbol{\theta})\, du$$

In general, the quantile of order α (or the percentile of order 100α) is defined as that value of q_α such that

$$\alpha = \int_{-\infty}^{q_\alpha} f(u;\boldsymbol{\theta})\, du$$

i.e. the area to the left of q_α under the p.d.f. of X is α.
Let $x_{(1)}, \ldots, x_{(n)}$ denote the order statistics for $x_1, \ldots, x_n$ such that $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$, where $x_{(j)}$ is the j-th smallest value of $x_1, \ldots, x_n$. Them, the sample median $\hat{q}_{0.5}$ is given by

$$\begin{aligned}\hat{q}_{0.5} &= \tfrac{1}{2}\{x_{(n/2)} + x_{(n/2+1)}\} && \text{if } n \text{ is even} \\ &= x_{([n+1]/2)} && \text{if } n \text{ is odd}\end{aligned}$$

Method of Moments

An estimate of a parameter or parameter vector is obtained by equating the population moments to the sample moments and solving for the unknown parameter(s). The latter are calculated using the empirical distribution function in place of the assumed distribution function. The empirical distribution function is formed by placing mass one at each observed data point in the sample.
For example, with an observed random sample $x_1, \ldots, x_n$, if we have one parameter to estimate, we solve the on equation

$$\mathbb{E}(X) = \mu = \sum_{j=1}^n x_j/n$$

or

$$\mathbb{E}\left(X^2\right) = \mu^2 + \sigma^2 = \sum_{j=1}^n x_j^2/n$$

if we have two unknown parameters. In general with k unknown parameters, we have to solve the k equations

$$\mathbb{E}\left(X^i\right) = \sum_{j=1}^n x_j^i/n \quad (i = 1, \ldots, k)$$

to find the moment estimate of $\boldsymbol{\theta}$. The quantity $\sum_{j=1}^n x_j^i/n$ is the moment of the ith order of the empirical distribution F_n which is the nonparametric estimate of the distribution function $F_{\boldsymbol{\theta}}$, putting one mass at each observed point x_j; that is, its probability function p.f. $f_n(x)$ is given by

$$\begin{aligned}f_n(x) &= 1/n && x = x_j \ (j = 1, \ldots, n) \\ &= 0 && \text{otherwise}\end{aligned}$$

Likelihood Function

Let $f_{\boldsymbol{X_1},...,\boldsymbol{X_n}}(\boldsymbol{x_1},...,\boldsymbol{x_n};\boldsymbol{\theta})$ denote the joint p.f. or p.d.f. of the r.v.'s $\boldsymbol{X_1},...,\boldsymbol{X_n}$ corresponding to an observed sample $\boldsymbol{X_1},...,\boldsymbol{X_n}$.
When considered as a function of $\boldsymbol{\theta}$ for the realisations $\boldsymbol{x_1},...,\boldsymbol{x_n}$ (the observed sample), $f_{\boldsymbol{X_1},...,\boldsymbol{X_n}}(\boldsymbol{x_1},...,\boldsymbol{x_n};\boldsymbol{\theta})$ is called the likelihood function, which we shall denote by $L(\boldsymbol{\theta})$ or by $L(\boldsymbol{\theta};\boldsymbol{x_1},...,\boldsymbol{x_n})$ to explicitly reinforce the fact that $L(\boldsymbol{\theta})$ can be considered to be a realisation of the random variable $L(\boldsymbol{\theta};\boldsymbol{X_1},...,\boldsymbol{X_n})$.
In the case of a random sample for which $\boldsymbol{X_1},...,\boldsymbol{X_n}$ are i.i.d. with common p.f. or p.d.f. $f(\boldsymbol{x};\boldsymbol{\theta})$, we have

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{n} f(\boldsymbol{x}_j;\, \boldsymbol{\theta})$$

In the case where $\boldsymbol{\theta}$ is a scalar, the score statistic is defined by

$$S(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n;\theta) = \frac{\partial \log L(\theta)}{\partial \theta}$$

and Fisher's (expected) information is defined by

$$\mathscr{I}(\theta) = \mathbb{E}\left(\frac{\partial \log L(\theta)}{\partial \theta}\right)^2$$

Regularity Conditions

1. The densities $f(x;\theta)$ have common support so that, without loss of generality, the set $\{x; f(x;\theta) > 0\}$ does not depend on θ.
2. Ω is an open interval in $\mathbb{R}$ (finite or not)
3. $\partial f(x;\theta)/\partial\theta$ exists for all $\theta \in \Omega$
4. $\mathscr{I}(\theta) > 0$ for all $\theta \in \Omega$
5. The term $\int\cdots\int L(\theta; x_1,\ldots,x_n)\, dx_1 \ldots dx_n$ may be differentiated under the integral (or summation sign in the case of discrete data
6. The term $\int\cdots\int T(x_1,\ldots,x_n)\, L(\theta; x_1,\ldots,x_n)\, dx_1 \ldots dx_n$ may be differentiated under the integral (or summation sign in the case of discrete data), where T is any unbiased estimator of $g(\theta)$

The conditions above are intended to illustrate the type of restrictions needed, but can be modified and relaxed to some extent depend on the results needed.
Under such regularity conditions as those above, we have that

$$\mathbb{E}\left(\frac{\partial \log L(\theta)}{\partial \theta}\right) = 0$$
$$\mathscr{I} = \mathbb{E}\left(I(\theta)\right)$$
$$\mathrm{Var}\,(T) \geq \frac{\{g'(\theta)\}^2}{\mathscr{I}(\theta)} \qquad (15)$$

where T is any unbiased estimator of the function $g(\theta)$, and $I(\theta)$ is the negative of the Hessian of the log likelihood function:

$$I(\theta) = -\frac{\partial^2 \log L(\theta)}{\partial \theta^2}$$

where the hessian of some function is given by

$$\boldsymbol{H}_f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1\,\partial x_n} \\ \frac{\partial^2 f}{\partial x_2\,\partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2\,\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n\,\partial x_1} & \frac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

The quantity $I(\hat{\theta})$ is known as the observed information, where $\hat{\theta}$ is the ML estimate of θ.
In the case where $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ represent a random sample of size n, we have that the likelihood function for θ is defined by

$$L(\theta;\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) = \prod_{j=1}^{n} f(\boldsymbol{x}_j;\theta)$$

where $f(\boldsymbol{x};\theta)$ denotes the p.f. or p.d.f. of $\boldsymbol{X}$. The (expected) information about θ, $\mathscr{I}(\theta)$, can then be expressed as

$$\mathscr{I}(\theta) = ni(\theta)$$

where

$$i(\theta) = \mathbb{E}\left(\left[\frac{\partial \log f(\boldsymbol{X};\theta)}{\partial \theta}\right]^2\right)$$

is the (expected) information about θ in a single observation.

Equation (15) represents the Cramér-Rao lower bound on the variance of an unbiased estimator of a function $g(\theta)$.

2 Week 2

Information for Parameter Vectors

In the case where $\boldsymbol{\theta}$ is a vector of dimension greater than one, the score statistic is defined by

$$\boldsymbol{S}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n;\boldsymbol{\theta}) = \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

and Fisher's expect information is defined by the matrix

$$\mathscr{I}(\boldsymbol{\theta}) = \mathbb{E}\left(\left\{\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\}\left\{\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\}^T\right)$$

which, under regularity conditions, is equal to

$$\mathscr{I}(\boldsymbol{\theta}) = \mathbb{E}\left(\boldsymbol{I}(\boldsymbol{\theta})\right)$$

where

$$\boldsymbol{I}(\boldsymbol{\theta}) = -\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}$$

In the case where $\boldsymbol{\theta}$ is a d-dimensional vector, the ith element of the score statistic is given by

$$S_i = \frac{\partial \log L(\boldsymbol{\theta})}{\partial \theta_i}$$

and the $(i,\, j)$th element of the information matrix $\mathscr{I}(\boldsymbol{\theta})$ is given by

$$(\mathscr{I}(\boldsymbol{\theta}))_{i,j} = \mathbb{E}\left(\left\{\frac{\partial \log L(\boldsymbol{\theta})}{\partial \theta_i}\right\}\left\{\frac{\partial \log L(\boldsymbol{\theta})}{\partial \theta_j}\right\}\right) \quad (i,\, j = 1,\ldots,d)$$

and finally the $(i,\, j)$th element of $\boldsymbol{I}(\boldsymbol{\theta})$ is given by

$$\boldsymbol{I}(\boldsymbol{\theta}) = -\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \theta_i\, \partial \theta_j} \quad (i,\, j = 1,\ldots,d)$$

Maximum Likelihood

Suppose that the likelihood function $L(\boldsymbol{\theta})$ is bounded in the parameter space $\boldsymbol{\Omega}$. The, the value of $\boldsymbol{\theta}$ that globally maximizes $L(\boldsymbol{\theta})$ over $\boldsymbol{\Omega}$ is called the maximum likelihood (ML) estimate of $\boldsymbol{\theta}$. If we denote the ML estimate of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$, this definition implies that

$$L(\hat{\boldsymbol{\theta}}) \geq L(\boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Omega}$$

In regular situations where the global maximum occur within the interior of the parameter space $\boldsymbol{\Omega}$ (not on the boundary), it follows that $\hat{\boldsymbol{\theta}}$ is an appropriate root of the so called ***likelihood equation***,

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

or equivalently, since log is a monotonic function,

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

Even in situations where the likelihood function is unbounded, there may still exist a root (or a sequence of roots) of the likelihood equation with the desirable asymptotic properties of the ML estimator. We still refer to this root as the ML estimator even though it is not a global maximiser of the likelihood function.

The Regular Exponential Family

In the case of n p-dimensional observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, a d-dimensional parameter $\boldsymbol{\theta}$, and a q-dimensional (sufficient) statistic $\boldsymbol{T}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ $(q \geq d)$, the likelihood function $L(\boldsymbol{\theta})$ for the d-parameter vector $\boldsymbol{\theta}$ or equivalently, the joint p.f. or p.d.f. $f(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n; \boldsymbol{\theta})$, has the following form if it belongs to the d-parameter exponential family:

$$f(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n; \boldsymbol{\theta}) = b(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n) \frac{\exp\left(\boldsymbol{c}(\boldsymbol{\theta})^T \boldsymbol{T}(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n)\right)}{a(\boldsymbol{\theta})}$$

where $\boldsymbol{c}(\boldsymbol{\theta})$ is a $q \times 1$ vector function of the d-dimensional parameter vector $\boldsymbol{\theta}$ and where $a(\boldsymbol{\theta})$ and $b(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n)$ are non-negative scalar functions.
The parameter space $\boldsymbol{\Omega}$ is a d-dimensional convex set such that the above formula defines a p.f. or p.d.f. for all $\boldsymbol{\theta} \in \boldsymbol{\Omega}$; that is,

$$\int \cdots \int b(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n) \exp\left(\boldsymbol{c}(\boldsymbol{\theta})^T \boldsymbol{T}(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n)\right) d\boldsymbol{x}_1, \ldots d\boldsymbol{x}_n < \infty$$

If $q = d$ and the Jacobian of $\boldsymbol{c}(\boldsymbol{\theta})$ is of full rank, then $f(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n; \boldsymbol{\theta})$ is said to be from a regular exponential family. The coefficient $\boldsymbol{c}(\boldsymbol{\theta})$ of the sufficient statistic T is referred to as the natural or canonical parameter (vector).

Derivation of ML Estimates of Parameters in the Regular Exponential Family

In the case of the joint distribution of the data $\boldsymbol{X}_1, \ldots \boldsymbol{X}_n$ belonging to the regular exponential family, the log likelihood function $L(\boldsymbol{\theta})$ has the form

$$\log L(\boldsymbol{\theta}) = \log b(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n) + \boldsymbol{c}(\boldsymbol{\theta})^T \boldsymbol{T}(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n) - \log a(\boldsymbol{\theta})$$

It can be shown that the ML estimate of $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ that satisfies the equation

$$\mathbb{E}\left(\boldsymbol{T}(\boldsymbol{X}_1, \ldots \boldsymbol{X}_n)\right) = \boldsymbol{T}(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n)$$

That is,

$$\left[\mathbb{E}\left(\boldsymbol{T}(\boldsymbol{X}_1, \ldots \boldsymbol{X}_n)\right)\right]_{\theta=\hat{\theta}} = \boldsymbol{T}(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n)$$

Minimum Variance Bound (MVB) Estimators

The unbiased estimator T of $g(\theta)$ attains the Cramér-Rao lower bound, that is, T is the minimum variance bound (MVB), if and only if there is equality in the Cauchy-Schwartz inequality applied to the score statistic and T. Now, equality holds if and only if one is a linear function of the other; that is

$$\frac{\partial \log L(\theta)}{\partial \theta} = k(\theta)T + l(\theta)$$

We can also write this as

$$\frac{\partial \log L(\theta)}{\partial \theta} = k(\theta)\{T - g(\theta)\} \qquad \forall \theta \in \Omega$$

The direct way of establishing whether an estimator T attains the MVB in estimating the function $g(\theta)$ unbiasedly is to calculate its variance and see if it is equal to the expected information $\mathscr{I}(\theta)$.

The variance of T can be expressed as

$$\mathrm{Var}\,(T) = \left|\frac{g'(\theta)}{k(\theta)}\right|$$

where

$$\mathscr{I}(\theta) = |k(\theta)\, g'(\theta)|$$

3 Week 3

Sufficiency

It often turns out thta some part of the data carries no information about the unknown parameter $\boldsymbol{\theta}$ and that $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n$ can be replaced by some statistic $\boldsymbol{T}(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n)$. A statistic is any quantity that is a function of the data alone; that is, it cannot be a function of any unknown parameter.
We let $\boldsymbol{X}_1, \ldots \boldsymbol{X}_n$ denote n replications on a random variable $\boldsymbol{X}$. The joint p.f. or p.d.f. of $\boldsymbol{X}_1, \ldots \boldsymbol{X}_n$ is denoted by $f(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n; \boldsymbol{\theta})$, where $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n$ denote the observed values of $\boldsymbol{X}_1, \ldots \boldsymbol{X}_n$.

A statistic $\boldsymbol{T}$ is said to be a sufficient statistic for $\boldsymbol{\theta}$, or for the family of p.f.'s or p.d.f.'s

$$\{f(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n; \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \boldsymbol{\Omega}\}$$

if the conditional distribution of $\boldsymbol{X}_1, \ldots \boldsymbol{X}_n$ given $\boldsymbol{T}(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n)$ does not depend on $\boldsymbol{\theta}$ for all $\boldsymbol{\theta}$, except perhaps on a set N such that

$$P_\theta(N) = 0, \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Omega}$$

In a sense, a sufficient statistic $\boldsymbol{T}$ contains all the information about $\boldsymbol{\theta}$ contained in the observed sample $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n$. In this sense, it provides a reduction of the data.
It follows from the interpretation of sufficiency that if $\boldsymbol{T}$ is sufficient for $\boldsymbol{\theta}$, and $\boldsymbol{T} = \boldsymbol{H}(\boldsymbol{U})$, then $\boldsymbol{U}$ is also sufficient for $\boldsymbol{\theta}$. Further, $\boldsymbol{T}$ provides a greater reduction of the data than $\boldsymbol{U}$, unless $\boldsymbol{H}$ is 1-1 in which case $\boldsymbol{T} = \boldsymbol{U}$.

A sufficient statistic $\boldsymbol{T}$ is said to be ***minimal sufficient*** if, of all sufficient statistics, it provides the greatest possible reduction of the data; that is, if for any sufficient statistic $\boldsymbol{U}$, there exists a function $\boldsymbol{H}$ such that $\boldsymbol{T} = \boldsymbol{U}$ (in other words, if must be a function of every other sufficient statistic).

A sufficient statistic $\boldsymbol{T}$ is said to be complete if

$$\mathbb{E}\left(w(\boldsymbol{T})\right)_\theta = \boldsymbol{0} \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Omega}$$

which implies

$$w(\boldsymbol{T}) \equiv \boldsymbol{0}$$

The property of completeness guarantees uniqueness of certain statistical procedures based on $\boldsymbol{T}$. A complete sufficient statistic is always minimal.

The definition of sufficiency is not conveniently checked in practice. However, it can be shown that a statistic $\boldsymbol{T}$ is sufficient for $\boldsymbol{\theta}$ if and only if the joint p.f. or p.d.f. $f(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n; \boldsymbol{\theta})$ of $\boldsymbol{X}_1, \ldots \boldsymbol{X}_n$ can be factored as

$$f(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n; \boldsymbol{\theta}) = h_1(\boldsymbol{T}(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n); \boldsymbol{\theta})\ h_2(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n)$$

where h_1 and h_2 are non-negative functions and h_2 is a function of the data alone. This is known as the Fisher-Neyman factorization theorem.

Rao-Blackwell Theorem

Theorem 1 (Rao-Blackwell): Let $\boldsymbol{X}_1, \ldots \boldsymbol{X}_n$ denote a sample of size n and $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n$ the observed data. Suppose $U(\boldsymbol{X}_1, \ldots \boldsymbol{X}_n)$ is an unbiased estimator of the parameter θ; that is,

$$\mathbb{E}\left(U(\boldsymbol{X}_1, \ldots \boldsymbol{X}_n)\right) = \theta \quad \forall \theta \in \Omega$$

Suppose further that $\boldsymbol{T}(\boldsymbol{X}_1, \ldots \boldsymbol{X}_n)$ is a sufficient statistic for θ and set

$$W(\boldsymbol{T}(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n)) = \mathbb{E}\left(U(\boldsymbol{X}_1, \ldots \boldsymbol{X}_n) \mid \boldsymbol{T}(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n)\right)$$

Then,

1. The random variable $W(\boldsymbol{T}(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n))$ is a function of $\boldsymbol{T}$ alone; that is, it does not depend on θ.
2. $W(\boldsymbol{T}(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n))$ is an unbiased estimator of θ; that is,

$$\mathbb{E}\left(W(\boldsymbol{T})\right) = \theta \quad \forall \theta \in \Omega$$

3.

$$\mathrm{Var}\left(W(\boldsymbol{T})\right) < \mathrm{Var}\left(U\right)$$

unless $U = W(\boldsymbol{T})$ with probability one.

This result shows that if we have an unbiased estimator of θ and a sufficient statistic $\boldsymbol{T}$ for θ, we can form another unbiased estimator with smaller variance, provided the original estimator is not a function of $\boldsymbol{T}$ alone.

Theorem 2 (Lehmann-Scheffe): Let $\boldsymbol{X}$ be a random variable with p.f. or p.d.f. $f(\boldsymbol{x}; \theta)$ and suppose that $\boldsymbol{T}$ is a complete statistic for θ. Let $W(\boldsymbol{T})$ be an unbiased estimator of θ, having finite variance. Then $W(\boldsymbol{T})$ is an UMVU estimator of θ and is unique in the sense that if there exists another unbiased estimator V of θ, then $W = V$, except perhaps on a set of probability zero.

4 Week 4

Large Sample Theory

We have seen some optimality results that can be obtained, such as estimators that have minimum risk for a specified loss function; for example, UMVU (uniform minimum variance unbiased) estimators that minimize the variance in the class of unbiased estimators for the estimand under consideration. However, the mathematical precision of these optimality results tend to obscure the fact that statistically they are only rough approximations in view of the approximate nature of both the assumed model and the loss function. As such, introduce a further approximation by assuming that a sample size is "large", which mitigates the effects of the earlier decision-theoretic approximations because the results become much less dependent on the loss function, and, in large samples, it is much easier to study the dependence on the model.

Large sample theory considers a sample $\boldsymbol{X}_S = (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_n^T)^T$ not for fixed n, but as a number of sequence corresponding to $n = 1, 2, \ldots$ (or more generally $n = n_0, n_0 + 1, \ldots$) and assessing the performance of estimator sequences as $n \to \infty$. In applications, the limiting result is used as an approximation to the situation obtaining for the actual finite n.

Consistency

A sequence of estimators T_n of $g(\boldsymbol{\theta})$ is said to be consistent if for every $\boldsymbol{\theta} \in \Omega$,

$$T_n \xrightarrow{P_\theta} g(\theta), \text{ as } n \longrightarrow \infty$$

that is, given any $\varepsilon > 0$, then

$$\mathbb{P}\left(|T_n(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) - g(\boldsymbol{\theta})| \geq \varepsilon\right) \to 0, \text{ as } n \to \infty$$

If $\mathrm{Var}\left(T_n\right) \to 0$ and $\mathrm{bias}(T_n) \to 0$, as $n \to \infty$, then the sequence of estimates T_n (usually abbreviated to just "the estimator T_n") is consistent for estimating $g(\boldsymbol{\theta})$.

Large-sample Comparisons of Estimators

The large sample behaviour of estimators leads to a simple method for computing different estimators. For example,

suppose that T_1 and T_2 are two estimators of $g(\boldsymbol{\theta})$, where as $n \to \infty$

$$\sqrt{n}\{T_i - g(\boldsymbol{\theta})\} \xrightarrow{L} N(0, \tau_i^2) \quad (i = 1, 2)$$

that is, for sufficiently large n,

$$T_i \sim N(g(\boldsymbol{\theta}), \tau_i^2/n) \quad (i = 1, 2)$$

The dependence of T_i on n has been suppressed here. Then, the asymptotic relative efficiency (ARE) of T_2 with respect to T_1 is given by

$$\text{ARE}(T_2) = \frac{\tau_1^2}{\tau_2^2}$$

For example, if $\text{ARE}(T_2) = 1/2$, we say that T_2 is only half as efficient as T_1. To obtain the same limit distribution as T_1, T_2 needs twice as many observations as T_1.

Asymptotic Efficiency

Estimators of interest typically are consistent as the sample sizes tend to infinity and, suitably normalised, are asymptotically normally distributed about the estimated with a variance $v(\boldsymbol{\theta})$ (the asymptotic variance), which provides a reasonable measure of the accuracy of the estimator. Within the class of consistent asymptotically normal estimators, it comes out that under mild additional restrictions, there exists estimators that uniformly minimise $v(\boldsymbol{\theta})$.

It is the nature of of asymptotically optimal solutions not to be unique, since asymptotic results refer to the limiting behaviour of sequences, and the same limiting is shared by many different sequences. More specifically, if

$$\sqrt{n}\{T_n - g(\boldsymbol{\theta})\} \xrightarrow{L} N(0, v(\boldsymbol{\theta})) \quad \text{as } n \to \infty \qquad (74)$$

and T_n is asymptotically optimal in the sense of minimising $v(\boldsymbol{\theta})$, then $T_n + R_n$ is also optimal provided $\sqrt{n}R_n \xrightarrow{P} 0$ as $n \to \infty$.

Suppose that $T(X_1, \ldots, X_n)$ is not necessarily an unbiased estimator of $g(\theta)$ but that it is asymptotically normal where $v(\theta) > 0$. Then it turns out under some additional restrictions in the case of i.i.d. observations $X_1, \ldots, X_n$ that

$$v(\theta) \geq \frac{\{g'(\theta)\}^2}{i(\theta)} \qquad (75)$$

where

$$i(\theta) = \mathbb{E}\left(\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\right)^2\right) = \frac{\mathscr{I}(\theta)}{n}$$

is the expected information about θ in a single observation. A sequence T satisfying (74) with $v(\theta)$ equal to the lower bound in (75) is said to be asymptotically efficient.

Under certain regularity conditions, if T is any estimator satisfying (74), then $v(\theta)$ satisfies (75) (except on a set of probability zero). If $g(\theta)$ is differentiable, it is enough to consider the case $g(\theta) = \theta$. The regularity conditions then in the case of i.i.d. data $X_1, \ldots, X_n$ with common density $f(x;\theta)$ are as follows:

Regularity Conditions (II)

1. Ω is an open interval (not necessarily finite).
2. The distribution of X has support that does not depend on θ, that is, the set $A = \{x : \ f(x;\theta) > 0\}$ does not depend on θ.
3. For every $x \in A$, the density $f(x;\theta)$ is twice differentiable with respect to θ and the 2nd derivative is continuous in θ.
4. The integral $\int f(x;\theta)\ dx$ can be twice differentiated under the integral sign.
5. The Fisher (expected) information $\mathscr{I}(\theta)$ satisfies $0 < \mathscr{I}(\theta) < \infty$.
6. For any given $\theta_0 \in \Omega$, there exists a positive number k and a function $M(X)$, both of which may depend on θ_0 such that
$$\left|\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}\right| \leq M(x), \quad \forall x \in A;\ \theta_0 - k < \theta < \theta_0 + k$$
and $E_{\theta_0}\{M(X)\} < \infty$.

5 Week 5

An Important Theorem on ML Estimation

Suppose that $X_1, \ldots, X_n$ are i.i.d. with p.f. or p.d.f. $f(x;\theta)$. Suppose further that the Regularity Conditions (II) (above) hold, but with (3) and (4) replaced by the corresponding assumptions on the 3rd derivatives, which satisfy

$$\left|\frac{\partial^3 \log f(x;\theta)}{\partial \theta^3}\right| \leq M(x), \quad \forall x \in A$$

in an open interval of θ_0 and that $E_{\theta_0}\{M(X)\} < \infty$.
Then, any consistent sequence $\hat{\theta}_n$ of roots of the likelihood equation is asymptotically normal and efficient. That is,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{L} N(0,\ 1/i(\theta_0))$$

where

$$i(\theta_0) = E_{\theta_0}\left(\left[\frac{\partial \log f(x;\theta)}{\partial \theta}\right]^2\right) = \frac{\mathscr{I}(\theta_0)}{n}$$

and θ_0 denotes the true value of θ.

Further Theorems Concerning Asymptotic Distributions

Theorem 1: If $Y_n \xrightarrow{L} Y$, and a_n and b_n converge in probability to a and b respectively, as n tends to infinity, then

$$a_n Y_n + b_n \xrightarrow{L} aY + b$$

as n tends to infinity.

Theorem 2: If

$$\sqrt{n}(T_n - \theta) \xrightarrow{L} N(0, \xi^2)$$

then,

$$\sqrt{n}(h(T_n) - h(\theta)) \xrightarrow{L} N(0,\ \xi^2(h'(\theta))^2)$$

provided $h'(\theta)$ exists and is not zero.

6 Week 6

P-Value

The P-value is the probability under the null hypothesis H_0 of getting a result as extreme or more extreme as that which has been obtained. It is *not* the probability of the null hypothesis being true given the data.

Positive-Predictive-Value (PPV)

Let T be a test for carrying out a test of a null hypothesis H_0 versus an alternative hypothesis H_1 at a level of significance α. That is, α is the probability of a Type I error given by

$$\alpha = \mathbb{P}\left(\text{reject } H_0 \text{ (i.e. } T \text{ is positive)}; \mid H_0 \text{ is true}\right)$$

The probability of a Type II error is given by

$$\beta = \mathbb{P}\left(\text{reject } H_1 \text{ (i.e. } T \text{ is negative)} \mid H_1 \text{ is true}\right)$$

Here, the sensitivity $(1-\beta)$ corresponds to the power of the test and the specificity $(1-\alpha)$ corresponds to one minus the probability of a Type I error in a hypothesis testing framework.

The Predictive Value of a Positive Test, or Positive Predictive Value (PPV), is defined to be

$$PPV = \mathbb{P}\left(\text{True positive} \mid T \text{ is positive}\right)$$

In order to calculate the PPV, we need to know the prior odds

$$R = \frac{p_1}{p_0}$$

where p_1 is the probability of some condition being true (and hence p_0 the probability of that condition being false). With this knowledge of the prior odds, the PPV can be calculated by

$$PPV = \frac{(1-\beta)}{(1-\beta) + R^{-1}\alpha}$$

Similarly, the Negative Predictive Value (NPV) of the test can be expressed by

$$\begin{aligned} NPV &= \mathbb{P}\left(\text{True negative} \mid T \text{ is negative}\right) \\ &= \frac{(1-\alpha)}{(1-\alpha) + R\beta} \end{aligned}$$

7 Week 7

Review

Consider a large population of units. We are interested in some *numerical property of this population* (i.e. a ***parameter***). This parameter can be an average measurement, μ, in the population, or a population proportion p of a certain characteristic within the population.

Suppose a random sample of n units is taken from our population, and a sample ***statistic*** is computed. This could be a sample mean $\overline{X}$ for quantitative variables, or sample proportions $\hat{p}$ for a categorical variable, etc.
The obvious question to ask next is: how *confident* can we be that a sample statistic (i.e. ***point estimate***) is *close* to our population parameter?

Margin of Error & Level of Confidence

"Confidence" can be formalised as a ***margin of error***. "Confidence" can be formalised as a ***level of confidence***. Recall that an approximate 90%, 95% or 99% confidence interval for p can be constructed via

$$\begin{aligned} 90\%CI &= \hat{p} \pm 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ 95\%CI &= \hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ 99\%CI &= \hat{p} \pm 2.58\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \end{aligned}$$

where the left of the equality is the confidence interval, $\hat{p}$ is the point estimate, and the term after the plus or minus is the margin of error. The values 1.645, 1.96, etc come from the standard normal distribution, $N(0,1)$.

Pivot Quantities

Definition: A ***pivot variable*** (or pivot quantity, or simply pivot) is a theoretical (and generally *not* computable) transformation of a random variable (i.e. data) such that its distribution no longer depends on any unknown parameter.

For example, for some normally distributed data $X_1, X_2, \ldots X_n \sim N(\mu, \sigma^2)$, we know that $\overline{X} \sim N(\mu, \sigma^2/n)$, where n is the sample size. Thus, the theoretical (and non-computable) transformation is

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

which does not depend on any unknown parameters. Hence, $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ is an example of a pivot variable.
Pivot variables can be used to construct confidence intervals and hypothesis tests for population parameters.

In the above example, though, $\overline{X}$ is a random variable and so a confidence interval constructed from it will not always be the same. It is helpful to think of confidence intervals as ***confidence interval procedures***, because we are confident in the procedure and *not* in particular numbers that we happen to get in our sample.

There is nothing inherently special about 90, 95, or 99% levels of confidence. In general, a $(1-\alpha)\times 100\%$ CI for μ when σ^2 is known can be constructed via

$$\overline{X} \pm z^{(1-\alpha/2)}\frac{\sigma}{\sqrt{n}}$$

where $z^{(1-\alpha/2)}$ denotes the upper $\alpha/2$th quantile of the standard normal.
The concept of pivot variables is even more useful for more complex/realistic scenarios, such as when σ^2 is also unknown.

Confidence Intervals

Let $X_1, X_2, \ldots X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, where μ and σ^2 are both unknown. We have already seen that $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. We have seen that a (bias-corrected) MLE for σ^2 is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

which has a χ^2 (chi-squared distribution)

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)} \Longleftrightarrow \frac{S^2}{\sigma^2} \sim \frac{\chi^2_{(n-1)}}{n-1}$$

Moreover, $\overline{X}$ and S^2 are independent.
Consider a transformation given by

$$T = \frac{\overline{X}-\mu}{S/\sqrt{n}} = \frac{(\overline{X}-\mu)/(\sigma/\sqrt{n})}{S/\sigma} \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2_{(n-1)}}{n-1}}} \equiv T_{n-1}$$

Thus, T is a theoretical transformation whose distribution doesn't depend on any unknown parameters, and so T is a *pivot variable*. Like the standard normal, T-distributions are symmetric and somewhat bell curved.
To construct a $(1-\alpha)\times 100\%$ CI for μ when σ^2 is unknown, we can appeal to a known probability statement,

$$\begin{aligned}(1-\alpha)100\% &= \mathbb{P}\left(-t_{n-1}^{(1-\alpha/2)} \le T_{n-1} \le +t_{n-1}^{(1-\alpha/2)}\right)\\ &\overset{\text{pivot}}{=} \mathbb{P}\left(-t_{n-1}^{(1-\alpha/2)} \le \frac{\overline{X}-\mu}{S/\sqrt{n}} \le +t_{n-1}^{(1-\alpha/2)}\right)\\ &= \mathbb{P}\left(\overline{X} + t_{n-1}^{(1-\alpha/2)}\frac{S}{\sqrt{n}} \ge \mu \ge \overline{X} - t_{n-1}^{(1-\alpha/2)}\frac{S}{\sqrt{n}}\right)\end{aligned}$$

and so confidence intervals are formally constructed by

$$(1-\alpha)\times 100\% = \overline{X} \pm t_{n-1}^{(1-\alpha/2)}\frac{S}{\sqrt{n}}$$

How do we find the cutoffs $t^{(1-\alpha/2)}$? We appeal to tables or software. For example, for sample size $n = 42$ (41 degrees of freedom) and a level of confidence of 97%, we would find $t_{41}^{0.985} = 2.248$, and so the CI is $\overline{X} \pm 2.248\ S/\sqrt{n}$.

Confidence Intervals for a Population Variance σ^2

Let $X_1, X_2, \ldots X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ where μ and σ^2 are both unknown. We have already seen that $(n-1)S^2/\sigma^2 \sim \chi^2_{(n-1)}$, which is a pivot quantity. The χ^2 distributions are *not* symmetric. Because not all distributions are symmetric, not all CIs are symmetric about the point estimate. In fact, not all CIs are additive ($\pm$) in their margin of errors, and some confidence intervals have multiplicative (i.e. relative) margin of errors.

Because $(n-1)S^2/\sigma^2$ is a pivot, we can again start with a known probability statement:

$$\begin{aligned}(1-\alpha)100\% &= \mathbb{P}\left(\chi^{2\,(\alpha/2)}_{n-1} \le \chi^2_{n-1} \le \chi^{2\,(1-\alpha/2)}_{n-1}\right)\\ &\overset{\text{pivot}}{=} \mathbb{P}\left(\chi^{2\,(\alpha/2)}_{n-1} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^{2\,(1-\alpha/2)}_{n-1}\right)\\ &= \mathbb{P}\left(\frac{(n-1)S^2}{\chi^{2\,(1-\alpha/2)}_{n-1}} \ge \sigma^2 \ge \frac{(n-1)S^2}{\chi^{2\,(\alpha/2)}_{n-1}}\right)\end{aligned}$$

and so a $(1-\alpha)100\%$ confidence interval for σ^2 can be constructed via the multiplicative bounds,

$$\left(\frac{(n-1)S^2}{\chi^{2\,(1-\alpha/2)}_{n-1}}\ ,\ \frac{(n-1)S^2}{\chi^{2\,(\alpha/2)}_{n-1}}\right)$$

Confidence Intervals for a Population Proportion p

A random sample of size n is taken, and the number X of "successes" is counted. We know that $X \sim \text{Bin}\,(n,p)$, where p is the underlying population proportion of success.
In general, pivot transformations for discrete R.V.'s do *not* exist, because shifting and/or scaling will change the support of the distribution from the integers to something "weird" (that still depends on the underlying parameters). However, we can still construct approximate/***asymptotic pivots***. Recall that the expected value of $\hat{p} = x/n$ is $\mathbb{E}\,(\hat{p}) = p$, and its standard error is $\sqrt{\hat{p}(1-\hat{p})/n}$. We also know that from the CLI, $\hat{p}$ has an asymptotic normal distribution. Combining these three facts gives us the transformation

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

for large n. This is an *asymptotic pivot*, which does not depend on p. Thus, for a 95% CI for the population proportion p, we again start from a known probability statement,

$$\begin{aligned}95\% &= \mathbb{P}\,(-1.96 \le N(0,1) \le +1.96)\\ &\approx \mathbb{P}\left(-1.96 \le \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \le +1.96\right)\\ &\Rightarrow \hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\end{aligned}$$

for large n. This method is only approximate, because the CLI requires n to be increasingly large. A general rule of thumb is that the sample size must be large enough to observe at least 10 "successes: and at least 10 "failures". If n is not large enough, we can use alternative methods for constructing CIs.

General Def$^{\text{n}}$ of a Confidence Interval Procedure

Let $X_1, X_2, \ldots, X_n$ be our data coming from some population that depends on a parameter θ. Then, $(T_1(\underline{X}),\ T_2(\underline{X}))$ is an exact $(1-\alpha)100\%$ CI for θ if $T_1(\underline{X})$ and $T_2(\underline{X})$ are functions of the data $\underline{X}$ alone such that

$$\mathbb{P}\,(T_1(\underline{X}) \le \theta \le T_2(\underline{X})) = 1-\alpha, \quad \forall\theta \in \Omega$$

where θ is fixed, and $T_1(\underline{X}$ and T_2 are random endpoints that depend on data $\underline{X}$.
If $\mathbb{P}\,(T_1(\underline{X}) \le \theta \le T_2(\underline{X})) \ge 1-\alpha$ for all $\theta \in \Omega$, then this is called a ***conservative CI*** with confidence level (at least) $(1-\alpha)100\%$.

Final thoughts on confidence intervals:

- Interval estimates are more informative than a single point estimate because they give us a sense of uncertainty of an estimate along with a level of confidence.
- CIs can be considered as a range/set of parameter values that are constant with the data. In this sense, CIs can be considered as the inverse/complement to ***hypothesis testing***.

Hypothesis Testing

Point and interval estimates are important tools for obtaining a "best guess" of a parameter. However, in applied settings, we need to make decisions between two competing hypotheses H_0 and H_1 about our underlying parameter.
A "hypothesis test" or "hypothesis testing procedure" or simply "test", is just a ***decision rule*** that specifies when to reject H_0 in favour of H_1.

8 Week 8

Some examples of decision rules could be: is the sample mean greater than some critical value c_1? Is the sample median greater than a different critical value c_2? Is $X_7 \geq c_3$?
There is no unique decision rule for any given scenario, but intuitively some rules are better than others.

Null Hypothesis and Alternative Hypothesis

The null and alternative hypothesis are two competing claims about a population parameter. As the name suggests, the null hypothesis is the "status quo" that essentially claims that nothing interesting is happening in the experiment. The alternative hypothesis is the "interesting" hypothesis that claims something is happening.
Mathematically speaking, these two claims are symmetric. However, there are good reasons to treat them asymmetrically. In this course, we mainly focus on a single point hypothesis $H_0 : \theta = \theta_0$, where θ_0 is a specified value. The alternative can be a composite one-sided $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$, or composite two-sided $H_1 : \theta \neq \theta_0$.

Test Statistic: Our decision rule for deciding between H_0 and H_1 needs to be based on the data $X_1, \ldots, X_n$ that was collected. We will consider an approximate summary statistic(s) (i.e. functions of the data alone). Any such function of data can be used as a ***test statistic*** and we write $T = T(X_1, X_2, \ldots, X_n)$ for our test statistic.
Although the choice of test statistic is arbitrary, in many cases we can use tools and principles to suggests some useful test statistics as a starting point. For example, MoM suggests using $\overline{X}$ for testing the population mean μ, etc.
Alternatively, if we have a ***model*** for our population, then we can use the MLE as our starting point. For example, $\overline{X}$ for testing μ if $\underline{X} \sim N(\mu, \sigma^2)$, $1/\overline{X}$ for testing λ if $\underline{X} \sim \exp(\lambda)$, etc. Or, we can use a more general approach by appealing to the likelihood ratio principle.

Decision Rules via Critical Regions

Once we have identified parameter(s) of interest, formulated our two hypotheses, and chosen our test statistic $T = T(X_1, \ldots, X_n)$, then we can propose a ***decision rule*** based on the outcome of T:

> Reject H_0 in favour of H_1 if T falls in some critical region (rejection region) C, i.e. $T \in C$.

Here, our critical region C can also be one-sided or two sided, reflecting whether the alternative hypothesis is one or two-sided.

Right One-Sided Rejection Region: $C = [c, \infty)$, i.e. reject $H_0 : \theta = \theta_0$ in favour of $H_1 : \theta > \theta_0$ "if T is overly large"

Left One-Sided Rejection Region: $C = (-\infty, c]$, i.e. reject $H_0 : \theta = \theta_0$ in favour of $H_1 : \theta < \theta_0$ "if T is overly small"

Two-Sided Rejection Region: $C = \{(-\infty, c_1], [c_2, \infty)\}$, i.e. reject $H_0 : \theta = \theta_0$ in favour of $H_1 : \theta \neq \theta_0$ if "T is overly small or large"

Whenever we make a decision based on some rule, we run the risk of making two types of error:

- Accidentally reject H_0 when it is in fact true (Type I error)
- Accidentally retain H_0 when it is in fact false (Type II error)

Type I Error and Significance Level

The probability of a type I error is denoted by

$$\alpha = \mathbb{P}\left(T(\underline{X}) \in C \mid H_0\right) = \mathbb{P}_{H_0}\left(T(\underline{X}) \in C\right)$$

i.e., the probability that the test statistic is in the critical region, given that the null hypothesis is true. This probability α is also called the "significance level" of a test, or simply "level".
Computing α requires us to know the distribution of $T(\underline{X})$ under H_0. We can either assume an outright model for our data $\underline{X}$, or we can appeal to limit theorems (e.g. Central limit theorem (CLT), Fisher-Tippett) for an approximate distribution.
The weaker the evidence that we require, the higher the Type I error (and vice versa).
Suppose we wanted to control/limit the Type I error to at most 5%, then what critical value should we set in our decision rule? We would want to find $0.05 = \mathbb{P}\left(N(0,1) \geq 1.645\right)$ and then pivot the standard normal term to something involving our data.

It seems tempting to aim for a test with minimal (or zero) Type I error. However, the only rule that achieves this is "always accept/retain H_0". This rule is useless in practice because it lacks ***power*** – the ability to detect an effect where there is in fact one.

Type II Error and Power

The probability of a Type II error is denoted by

$$\begin{aligned}\beta &= \mathbb{P}\left(T(\underline{X}) \notin C \mid H_1\right) = \mathbb{P}_{H_1}\left(T(\underline{X}) \notin C\right) \\ &= 1 - \mathbb{P}\left(T(\underline{X}) \in C \mid H_1\right) = 1 - \mathbb{P}_{H_1}\left(T(\underline{X}) \in C\right)\end{aligned}$$

That is, the probability of accepting H_0 given that H_1 is true.
The power of a test is the probability of rejecting H_0 when it is in fact false. It is related to the Type II error via:

$$\text{Power} = \mathbb{P}_{H_1}\left(T(\underline{X}) \in C\right) = 1 - \beta$$

Since H_1 is typically composite (e.g. $H_1 : \theta > \theta_0$, or $H_1 : \theta < \theta_0$, or $H_1 : \theta \neq \theta_0$), there are many values of the parameter θ under the alternative. For each θ value in the alternative, we can compute its power and then we can graph the power as a function of θ. This graph is called a power curve.

Neyman-Pearson proposed a compromise between Type I error and power that is widely used today.
Neyman-Pearson paradigm: Aim for a test that:

1. Restricts/controls Type I error at a "suitably small" level α
2. Maximise the power under the alternative(s)

Any test that satisfies those two criteria is called a "most powerful test at level α".

Likelihood Ratio Tests

Let $L(\theta \mid \underline{X}) = f_\theta(\underline{X})$ be the likelihood function evaluated at θ for a given dataset $\underline{X}$. Suppose (for now) that we are simply deciding between a single null hypothesis $H_0 : \theta = \theta_0$ vs a single alternative hypothesis $H_1 : \theta = \theta_1$, where θ_0 and θ_1 are two specified values.
Let's consider the likelihood ratio:

$$\frac{L(\theta_0 \mid \underline{X})}{L(\theta_1 \mid \underline{X})}$$

where the numerator is the likelihood of observing the data $\underline{X}$ under θ_0, and the denominator the same under θ_1.
If this ratio is "small", the data would have been more likely observed under H_1 (θ_1) than under H_0 (θ_0), and so may constitute evidence against H_0 in favour of H_1. This leads to the likelihood ratio test (LRT):

$$\begin{array}{l}\text{Reject } H_0 : \theta = \theta_0 \\ \text{in favour of } H_1 : \theta = \theta_1\end{array} \qquad \text{if} \quad \frac{L(\theta_0 \mid \underline{X})}{L(\theta_1 \mid \underline{X})} \leq c$$

for some critical value c.
Neyman-Pearson Lemma: Suppose the LRT that rejects $H_0 : \theta = \theta_0$ in favour of $H_1 : \theta = \theta_1$ if

$$\frac{L(\theta_0 \mid \underline{X})}{L(\theta_1 \mid \underline{X})} \leq c$$

has level α. Then any other test that has level α has power less than or equal to the LRT.

As an example, the LRT for testing $H_0 : \mu = \mu_0$ vs $H_1 : \mu = \mu_1 > \mu_0$ in a $N^*(\mu, \sigma^2)$ distribution with known σ (where the * represents that it only has to be *mostly* normal) is

$$\left\{\frac{L(\mu_0 \mid \underline{X})}{L(\mu_1 \mid \underline{X})} \leq c \text{ for some } c\right\} \Leftrightarrow \left\{\overline{X} \geq d \text{ for some } d\right\}$$

How do we set d? If we want to restrict our Type I error to at most $\alpha(= 0.05)$, then we need to set d to be:

$$\begin{aligned}\alpha &= \mathbb{P}_{H_0}(\overline{X} \geq d) \\ &\overset{CLT}{\approx} \mathbb{P}_{H_0}\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \geq z^{(1-\alpha)}\right) \\ &= \mathbb{P}_{H_0}\left(\overline{X} \geq \mu_0 + z^{(1-\alpha)}\sigma/\sqrt{n}\right) \\ \Rightarrow d &= \mu_0 + z^{(1-\alpha)}\sigma/\sqrt{n}\end{aligned}$$

Because this test does not actually depend on the value of μ_1, it is in fact the uniformly most powerful (UMP) test for $H_0 : \mu = \mu_0$ vs $H_1 : \mu = \mu_1 > \mu_0$.
Similarly, if we are testing $H_0 : \mu = \mu_0$ vs $H_1 : \mu = \mu_1 < \mu_0$, the LRT reduces to $\{\overline{X} \leq d\}$ for some d. If we want to restrict the level of the test to α, we get that $d = \mu_0 - z^{(1-\alpha)}\sigma/\sqrt{n}$ by the same process as before and this test is UMP.

9 Week 9

For a two sided test, we'd need that $\overline{X} \geq d_1$ *or* $\overline{X} \leq d_2$ where d_1 and d_2 are given by each d above respectively. In general, there cannot be a UMP test for two-sided hypotheses, because a correctly chosen one-sided test is always more powerful.

We say that for normal * data, the LRT for testing mean μ reduces simply to large and/or small values of $\overline{X}$. What if we don't know σ^2? We can estimate it using S^2 (sample variance). Then, the LRT reduces to large and/or small values of the pivot

$$\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

To control Type I error of the test to at most α, we can then use t-cutoffs instead of z-cutoffs. Of course, t-tests are *not* the most powerful tests for non-normal data (likelihood function for non-normal data looks different!), but they can still be *asymptotically most powerful!*. That is, for a large sample size n, the t-test has power that approaches the correctly chosen LRT.

LRT for Poisson Data

Let $X_1, \ldots, X_n \sim \text{Poi}(\lambda)$. Suppose we are testing between $H_0 : \lambda = \lambda_0$ vs $H_1 : \lambda = \lambda_1 > \lambda_0$. Construct a LRT for this scenario:

$$\begin{aligned}L(\lambda \mid \underline{X}) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{X_i}}{X_i!} \\ &= e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^n X_i}}{\prod_{i=1}^n X_i!}\end{aligned}$$

So the LR statistic is

$$\frac{L(\lambda_0 \mid \underline{X})}{L(\lambda_1 \mid \underline{X})} = e^{n(\lambda_1-\lambda_0)} \left(\frac{\lambda_0}{\lambda_1}\right)^{\sum_{i=1}^n X_i}$$

Since $\lambda_1 > \lambda_0$, the LRT is small when $\sum_{i=1}^n X_i$ is large. So the LR test for testing $H_0 : \lambda = \lambda_0$ vs $H_1 : \lambda = \lambda_1 > \lambda_0$ reduces to a simple rule:

$$\left\{\sum_{i=1}^n X_i \geq d\right\} \Longleftrightarrow \{\overline{X} \geq c\}$$

for some c and d. Again, our decision rule doesn't depend on the value of λ_1, so it is in fact UMP for testing this.

Knowing the current model (likelihood) for your data leads to optimal decision rules. However, in practice, we never know exactly the data-generating mechanism. Hence, ***asymptotically most powerful tests*** that are robust to model misspecification are important.

Note that in our discussion on hypothesis testing so far, we have not looked at any numerical data yet. We've done this on purpose to make sure that:

1. identifying our parameter of interest
2. formulating our null and alternative hypotheses about this parameter
3. choosing the test statistic $T(\underline{X})$ and region $\{T(\underline{X}) \in C\}$
4. setting our Type I error (by choosing critical values c, d, etc)
5. assessing the power of our test

must be carried out *before* looking at/collecting any numerical data. These are all properties of the ***testing procedure*** and not specific to any sample of data.
We can complement our decision from the data with a *p-value*, which quantifies the *strength of evidence* against our null hypothesis in favour of the alternative.

P-Values: Suppose we have realisations $x_1, x_2, \ldots, x_n$ of $X_1, X_2, \ldots X_n$ from the outcome of some random experiment. We can then compute the observed value of our test statistic: $t_{\text{obs}} = T(\underline{X})$ and make our decision based on this value.

Whatever decision we make, we can supplement our decision by computing a p-value. For a null hypothesis $H_0 : \theta = \theta_0$, we have for one sided tests that $p = \mathbb{P}_{H_0}(T(\underline{X}) \geq t_{\text{obs}})$ for alternative $H_1 : \theta > \theta_0$, and $p = \mathbb{P}_{H_0}(T(\underline{X}) \leq t_{\text{obs}})$ for $H_1 : \theta < \theta_0$. For a two sided test with alternative $H_1 : \theta \neq \theta_0$,

$$p = 2 \times \min\left\{\begin{matrix}\mathbb{P}_{H_0}(T(\underline{X}) \geq t_{\text{obs}}) \\ \mathbb{P}_{H_0}(T(\underline{X}) \leq t_{\text{obs}})\end{matrix}\right\}$$

In all three cases, the p-value is always the probability under H_0 that our test statistic would be "as unusual or more unusual than what we observed".

P-values are often misinterpreted ("the chance that the null is true is <p-value>"), and should be interpreted as something like "the chance that we would see this data given that the null hypothesis is true".
Confidence intervals are equally misinterpreted: "the chance that the true value is between <a> and <b> is 95%..."
There is a duality between CIs and p-values/hypothesis testing.

Duality Between CIs & Hypothesis Testing

Given a dataset, a CI provides a range/set of parameter values under which the data would have reasonably likely to have occurred ("range of plausible parameter values" or "range of parameter values that are consistent with the data").
On the other hand, given a parameter value, a hypothesis test specifies a range of data values that would have been unlikely to be observed.
We can see that CIs are simply the inverse of hypothesis testing (& vice versa). More precisely:

- two-sided CIs are precisely a set of parameter values that would be accepted as the null hypothesis is a two-sided test
- one-sided CIs are precisely a set of parameter values that would be accepted as the null is a one-sided test

Generalized Likelihood Ratio Test (GLRT)

The GLRT can be used to unify all of the methods from chapters 7 and 8 of the supplementary notes. Likelihood ratio tests can be generalised to scenarios in which the null hypothesis and the alternative are *not* single point hypotheses. This is particularly relevant when the parameter is multidimensional.

Suppose that the null hypothesis $H_0 : \theta \in \mathcal{H}_0$, where $\mathcal{H}_0$ is a subset of the full parameter space $\mathcal{H}$, and the alternative $H_1 : \theta \in \mathcal{H}_1$, where $\mathcal{H}_1$ is another disjoint subset of the parameter space.
So, if we are considering a "two-sided" alternative, it really means "outsided" (or "insided") alternative. Typically $\mathcal{H}_1 = {\mathcal{H}_0}^C$.

Consider the generalised likelihood ratio test (GLRT) statistic:

$$\frac{\sup_{\underline{\theta} \in \mathcal{H}_0} L(\underline{\theta} \mid \underline{X})}{\sup_{\underline{\theta} \in \mathcal{H}_1} L(\underline{\theta} \mid \underline{X})} \quad \begin{matrix}\leftarrow \text{ best possible likelihood under } H_0 \\ \leftarrow \text{ best possible likelihood under } H_1\end{matrix}$$

If this ratio is "small", then the data would have more likely occurred under $\mathcal{H}_1$ than $\mathcal{H}_0$. How "small" depends on the Type I error rate that we want to control.

The GLRT ratio Λ is small when the distance between the group sum of squares (SS) is large $\Leftrightarrow$ if $\overline{X}_1$ and $\overline{X}_2$ are far apart (for two distributions in the regular exponential

family). The group sum of square is calculated by

$$SS_{\text{total}} = \sum_{i=1}^{n_1}(X_{1i} - \overline{X}_1)^2 + \sum_{i=1}^{n_2}(X_{2i} - \overline{X}_2)^2$$

The GLRT simplifies to the rule: reject $H_0 : \mu_1 = \mu_2$ in favour of $H_1 : \mu_1 \neq \mu_2$ if

$$\left\{\overline{X}_1 - \overline{X}_2 \leq d_1 \qquad \text{or} \qquad \overline{X}_1 - \overline{X}_2 \geq d_2\right\}$$

for some d_1 and d_2 values.

10 Week 10

How do we choose critical values d_1 and d_2 to control the Type I error rate (at $\alpha = 0.05$, say). Recall that, for two normal distributions with means $\overline{X}_1$ and $\overline{X}_2$ with the same SD σ^2,

$$\overline{X}_1 - \overline{X}_2 \sim \text{N}\left(\mu_1 - \mu_2,\ \tfrac{\sigma^2}{n_1} + \tfrac{\sigma^2}{n_2}\right)$$

and so under $H_0 : \mu_1 - \mu_2 = 0$ (or in general $\mu_1 - \mu_2 = \Delta_0$), we can set

$$d_1 = \Delta_0 - z^{(1-\alpha/2)}\sigma\sqrt{\tfrac{1}{n_1} + \tfrac{1}{n_2}}$$
$$d_2 = \Delta_0 + z^{(1-\alpha/2)}\sigma\sqrt{\tfrac{1}{n_1} + \tfrac{1}{n_2}}$$

(this is the 2-sample Z-test).
What if σ^2 is also unknown? We can estimate σ^2 by

$$S^2_{\text{pooled}} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

(weighted averaged of the two mid-valued sample variances) and we can now use t-quantities instead of z-quantities:

$$d_1 = \Delta_0 - t^{(1-\alpha/2)}_{n_1+n_2-2}\ S_{\text{pooled}}\sqrt{\tfrac{1}{n_1} + \tfrac{1}{n_2}}$$
$$d_2 = \Delta_0 + t^{(1-\alpha/2)}_{n_1+n_2-2}\ S_{\text{pooled}}\sqrt{\tfrac{1}{n_1} + \tfrac{1}{n_2}}$$

This is the two-sample t-test!

Goodness-of-Fit Tests (for Categorical Data)

Suppose each experimental unit can be classified as one of K categories (i.e. of of K levels of a factor).

The underlying population prevalence of each level of a factor is denoted by $p_1, p_2, \ldots, p_K$. Of course, $p_K = 1 - p_1 - p_2 - \cdots - p_{K-1}$, so there are only $K - 1$ free parameters in general.
Now, suppose $X_1, X_2, \ldots, X_K$ are counts in each category from a random sample of n units, so that $X_1 + X_2 + \cdots + X_K = n$, and so $X_K = n - X_1 - \cdots - X_{K-1}$. The pmf of $\underline{X} = (X_1, X_2, \ldots, X_K)^T \sim$ Multinomial $(n; p_1, p_2, \ldots, p_K)$ is

$$\mathbb{P}\left(X_1 = x_1, \ldots, X_K = x_k\right) = \binom{n}{\underline{x}} p_1^{x_1} \cdot \ldots \cdot p_K^{x_K}$$

where

$$\binom{n}{\underline{x}} = \frac{n!}{x_1! x_2! \ldots x_K!}$$

is the ***multinomial coefficient***. As a special case with $K = 2$, we arrive at the binomial distribution.

Suppose for now we are interested in testing a single point null hypothesis

$$H_0 : \underline{p} = (p_1, p_2, \ldots, p_K) = \underline{p}^* = (p_1^*, \ldots, p_K^*)$$

where $\underline{p}^*$ is a given set of proportions, vs $H_1 : \underline{p} \neq \underline{p}^*$. Note that this does *not* mean that all of the elements $\underline{p}_i \neq \underline{p}_i^*$, but rather that *at least* one $\underline{p}_i \neq \underline{p}_i^*$.
Consider the generalised likelihood ratio test:

$$\Lambda = \frac{\sup_{H_0} L(\underline{p} \mid \underline{X})}{\sup_{H_1} L(\underline{p} \mid \underline{X})} = \frac{L(\underline{p}^* \mid \underline{X})}{\sup_{H_1} L(\underline{p} \mid \underline{X})}$$

Denominator: What is the MLE of $\underline{p}$ under H_1?

$$\hat{p}_1 = \frac{X_1}{n}; \qquad \hat{p}_2 = \frac{X_2}{n}; \qquad \hat{p}_K = \frac{X_K}{n}$$

i.e. the *sample proportions*. So,

$$\Lambda = \frac{\binom{n}{\underline{x}} (p_1^*)^{x_1} \cdot \ldots \cdot (p_K^*)^{x_K}}{\binom{n}{\underline{x}} (\hat{p}_1)^{x_1} \cdot \ldots \cdot (\hat{p}_K)^{x_K}}$$
$$= \left(\frac{p_1^*}{\hat{p}_1}\right)^{x_1} \cdot \left(\frac{p_2^*}{\hat{p}_2}\right)^{x_2} \cdot \ldots \cdot \left(\frac{p_K^*}{\hat{p}_K}\right)^{x_K}$$

Small values of $\Lambda \iff$ small values of $\log \Lambda$ (or large values of $-\log \Lambda$).

$$-\log\Lambda = X_1 \log\left(\frac{\hat{p}_1}{p_1^*}\right) + X_2 \log\left(\frac{\hat{p}_2}{p_2^*}\right) + \cdots + X_K \log\left(\frac{\hat{p}_K}{p_K^*}\right)$$
$$= X_1 \log\left(\frac{X_1}{np_1^*}\right) + X_2 \log\left(\frac{X_2}{np_2^*}\right) + \cdots + X_K \log\left(\frac{X_K}{np_K^*}\right)$$
$$= O_1 \log\left(\frac{O_1}{E_1}\right) + O_2 \log\left(\frac{O_2}{E_2}\right) + \cdots + O_K \log\left(\frac{O_K}{E_K}\right)$$
$$= \sum_{\text{cells } i} O_i \log\left(\frac{O_i}{E_i}\right)$$

which is in fact the exact GLRT statistic for categorical data (for some observed data O_i and expected data E_i).
The exact GLRT is often approximated by Pearson's χ^2 statistic:

$$\chi^2 = \sum_{\text{cells } i} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2 \text{ dist.}$$

While a small p-value makes us suspicious about the null hypothesis, a very large p-value can also make us suspicious, but about data fabrication and/or pseudo-replication.

Pseudo-replication occurs when we make multiple measurements from each unit but pretend that these are individual are individual measurements from independent units.

Working with Two (or more) Categorical Factors

Pearson's χ^2 (and the GLRT) *test of independence* and *test of homogeneity* are carried out in exactly the same way, but the underlying hypothesis and the way the data are collected are slightly different.
In a ***test of independence***, units are sampled from a single population and then cross-classified according to two (or more) categorical factors. The null hypothesis is that these two factors are independent/unrelated, i.e.

$$H_0 : p_{ij} = p_{i\cdot} \times p_{\cdot j} \qquad \forall i, j$$

where p_{ij} is the population proportion of being in level i of factor 1, and in level j of factor two; $p_{i\cdot}$ is the overall proportion in level i of factor one and similarly for $p_{\cdot j}$.
The alternative is

$$H_1 : \text{ at least one } p_{ij} \neq p_{i\cdot} \times p_{\cdot j}$$

In a ***test of homogeneity***, instead of sampling units from a single population and cross-classifying, we sample separately from two (or more) (sub) populations and we count the occurrence of a categorical factor in our units. We are interested in testing whether the distribution of the categorical factor are homogeneous (i.e. constant) across our populations.

11 Week 11

Are there any caveats for using a χ^2 distribution to approximate the χ^2 statistic?
Recall that for a binomial, we need at least 10 "successes" and at least 10 "failures" before we can use a Normal approximation. The same is true of Pearson's χ^2 test, although we relax this slightly to:

- at least 80% of cells in the table must have ≥ 10 counts
- all cells must have at least ≥ 5 counts

What happens when this requirement doesn't hold? We have to consider Fisher's exact test instead.

Fisher's exact test involves the following logical argument:

1. Argue why the row and column totals of your table can be considered either fixed or given under the null hypothesis.
2. Under the null hypothesis, any configuration of counts with the same row and column totals would have been equally likely to be observed.
3. We can generate/simulate many different tables with the same fixed margins and determine how often we get something "as unusual or more unusual" than what we observe.

The above 3 steps are Fisher's exact test. This cannot be done by hand, but it is easy to do with computers. The χ^2 distribution is an approximation to Fisher's exact test.

Comparing Groups using Qualitative Data

ANOVA (analysis of variance) is used to compare a qualitative response across levels of one (or more) factor. ANOVA is a special case of *linear models*.

ANOVA setup: (extension of two-sample t-test when there are more than two groups)

$$\begin{aligned} Y_{11}, Y_{12}, \ldots, Y_{1n_1} &\overset{\text{iid}}{\sim} N^*(\mu_1, \sigma^2) \\ \text{indep } Y_{21}, Y_{22}, \ldots, Y_{2n_2} &\overset{\text{iid}}{\sim} N^*(\mu_2, \sigma^2) \\ \vdots \qquad\qquad & \qquad \vdots \\ \text{indep } Y_{J1}, Y_{J2}, \ldots, Y_{2n_J} &\overset{\text{iid}}{\sim} N^*(\mu_J, \sigma^2) \end{aligned}$$

coming from j groups/sub-populations. Say we have

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \cdots = \mu_J \\ \text{vs } H_1 &: \text{not all means are the same} \end{aligned}$$

How do we extend the two-sample t-test to more than 2 groups? We can try our versatile tool, the GLRT:

$$\Lambda = \frac{\sup_{H_0} L(\mu_1, \ldots, \mu_J \mid \underline{Y})}{\sup_{H_1} L(\mu_1, \ldots, \mu_J \mid \underline{Y})}$$

First, let's write down the joint likelihood function (since all Y_{ji} are independent,

$$L(\mu_1, \ldots, \mu_j \mid \underline{Y}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{j} \sum_{i=1}^{n_j} (Y_{ji} - \mu_j)^2\right)$$

where $N = n_1 + \cdots + n_J$ is the total sample size.
Under the null hypothesis, $\mu_1 = \cdots = \mu_J = \mu$, for some common μ, the maximal likelihood achieved would be

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{j} \sum_{i=1}^{n_j} \left(Y_{ji} - \overline{Y_{\cdot\cdot}}\right)^2\right)$$

Under the alternative, it would be

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{j} \sum_{i=1}^{n_j} \left(Y_{ji} - \overline{Y_{j\cdot}}\right)^2\right)$$

where

$$\overline{Y_{\cdot\cdot}} = \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} Y_{ji}}{N}; \qquad \overline{Y_{j\cdot}} = \frac{\sum_{i=1}^{n_j} Y_{ji}}{n_j}$$

It can be shown that the double sum in the exponent (under the null) can be represented as

$$\underbrace{\sum_j \sum_i \left(Y_{ji} - \overline{Y_{j\cdot}}\right)^2}_{SS_{\text{within groups}} = SS_{\text{residuals}}} + \underbrace{\sum_{j=1}^{J} n_j \left(\overline{Y_{j\cdot}} - \overline{Y_{\cdot\cdot}}\right)^2}_{SS_{\text{between groups}} = SS_{\text{groups}}}$$

So the maximum likelihood achieved under the null is

$$\sup_{H_0} L = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2} \left[SS_{\text{groups}} + SS_{\text{residuals}}\right]\right)$$

Source	df	SS	MS	$\mathcal{F}$	$\mathbb{P}(\mathcal{F})$
Groups	$J-1$	$\sum_{j=1}^{J} n_j(\overline{Y_{j\cdot}} - \overline{Y_{\cdot\cdot}})^2$	$SS_{\text{group}}/(J-1)$	$MS_{\text{group}}/MS_{\text{resid}}$	$\mathbb{P}(\mathcal{F}_{J-1,N-J} \geq \mathcal{F})$
Residuals	$N-J$	$\sum_{j=1}^{J}\sum_{i=1}^{n_j} n_j(\overline{Y_{ji}} - \overline{Y_{j\cdot}})^2$	$SS_{\text{resid}}/(N-J)$		
Total	$N-1$	$\sum_{j=1}^{J}\sum_{i=1}^{n_j} n_j(\overline{Y_{ji}} - \overline{Y_{\cdot\cdot}})^2$			

Table 1: Generalised ANOVA Table

for the alternative,

$$\sup_{H_1} L = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2} SS_{\text{residuals}}\right)$$

and so

$$\Lambda = \frac{\sup_{H_0} L}{\sup_{H_1} L} = \exp\left(-\frac{1}{2\sigma^2} SS_{\text{groups}}\right)$$

We see that small values of Λ mean large values of SS_{group}. SS_{group} is a weighted measure of how different the group means are from each other, whereas $SS_{\text{residuals}}$ is a measure of how different data points within a group are from their group mean.

If we know σ^2, then

$$\frac{SS_{\text{group}}}{J-1} \sim \sigma^2 \frac{\chi^2_{J-1}}{J-1}$$

If we don't know σ^2, our best guess is then

$$S^2_{\text{pooled}} = \frac{SS_{\text{residual}}}{N-J} \sim \sigma^2 \frac{\chi^2_{N-J}}{N-J}$$

and

$$\frac{SS_{\text{group}}/(J-1)}{SS_{\text{resid}}/(N-J)} \sim \mathcal{F}_{J-1,N-J}$$

where the first degree of freedom in $\mathcal{F}$ (i.e. $J-1$) is measuring the 'strength of signal', and the $N-J$ DoF is measuring the residual noise.

We can summarise all of these steps in an ANOVA table, shown in Table 1.
We should only look at pairwise comparisons after the overall ANOVA detects a significant result. This protects us from overly large Type I error rates in our conclusions.

Contrasts

Comparing the "best" vs "worst" category, or indeed any pairwise comparison, is an example of a **contrast**.

> In general, a contrast is any linear combination $\sum_{j=1}^{J} a_j\mu_j$ of your group means $(\mu_1, \ldots, \mu_J)^T$ that satisfies $\sum_{j=1}^{J} a_j = 0$.

Contrasts are useful when we want to consider particular comparisons of group means (not necessarily pairwise). Two popular sets of contrasts arise from the so-called "sum parametrisation/sum constraint" and "contrast parametrisation/contrast constraint" of the one-way ANOVA. Note that both of these specify the exact same model, but differ in their parameter interpretation (and some interpretations are clearer for some scenarios than others).

12 Week 12

Mean constraint / mean parametrisation: The one-way ANOVA is usually written via the "mean parametrisation",

$$Y_{ji} = \mu_j + \varepsilon_{ji} \iff Y_{ji} \overset{\text{ind}}{\sim} N^*(\mu_j,\ \sigma^2)$$

with $\varepsilon_{ji} \overset{iid}{\sim} N^*(0, \sigma^2)$, $j = 1, 2, \ldots, J$ (indexes group), $i = 1, 2, \ldots, n_j$ (indexes observations within group). It can also be rewritten using the "sum parametrisation" or "contrast parametrisation". In this case, we would expect the MLE to be $\hat{\mu_j} = \overline{Y_{j\cdot}}$, for $j = 1, 2, \ldots, J$.

Sum constraint / sum parametrisation: The one-way ANOVA can also be rewritten as

$$Y_{ji} = \mu + \alpha_j + \varepsilon_{ji}$$

where μ is the common "intercept", α_j is the main "effect" due to group J, and for identifiability, we set

$$\sum_{j=1}^{J} \alpha_j = 0$$

In this case, μ is interpreted as the overall mean across all J groups, i.e. $\mu = (\mu_1 + \cdots + \mu_J)/J$. Each α_j is then interpreted as the difference between group mean j and the overall mean μ, i.e. $\alpha_j = \mu_j - \mu$.
When is the sum parametrisation useful? All J groups in the study are variations of the same treatment. Or, the J groups exhaust all possible groups in the population.
In this case, we would expect the MLE to be $\hat{\mu} = \overline{Y_{\cdot\cdot}}$, and $\hat{\alpha_j} = \overline{Y_{j\cdot}} - \overline{Y_{\cdot\cdot}}$ for $j = 1, 2, \ldots, J$.

Contrast constraint / contrast parametrisation: The one-way ANOVA can be rewritten as

$$Y_{ji} = \mu + \alpha_j + \varepsilon_{ji}$$

where, for identifiability, $\alpha_1 \equiv 0$. Then, μ is interpreted as the mean of group 1 (i.e. $\mu = \mu_1$) and each α_j represents the difference between group j and group 1: $\alpha_j = \mu_j - \mu_1$.
When is the contrast parametrisation useful? When one group (generically called "group 1") is a baseline group (e.g. placebo/control or current state-of-the-art) to which all others can be compared.
In this case, we would expect the MLE to be $\hat{\mu} = \overline{Y_{1\cdot}}$ and $\hat{\alpha_j} = \overline{Y_{j\cdot}} - \overline{Y_{1\cdot}}$ for $j = 2, 3, \ldots, J$.

Two-Way ANOVA

If the groups in a study are obtained by crossing two (or more) factors, each at two (or more) levels, then we should

respect the structure of the study and consider a two-way ANOVA instead.
The most general two-way ANOVA has the form:

$$Y_{jki} = \mu + \alpha_j + \beta_k + \delta_{jk} + \varepsilon_{jki}$$

where $j = 1, 2, \ldots, J$, $k = 1, 2, \ldots, k$, $i = 1, 2, \ldots, n_{jk}$ and $\varepsilon_{jki} \overset{iid}{\sim} N^*(0, \sigma^2)$. Here, α_js are the "main effects of factor A", β_ks are the "main effects of factor B", δ_{jk} are the "interaction effects of factor A and B".
There are $1 + J + K + JK$ parameters in the setup, and JK groups.
As before, we have constraints.
Sum parametrisation:

$$\sum_{j=1}^{J} \alpha_j = 0; \quad \sum_{k=1}^{K} \beta_k = 0; \quad \sum_{j=1}^{J} \delta_{jk} = 0 = \sum_{k=1}^{K} \delta_{jk}$$

Contrast parametrisation:

$$\alpha_1 \equiv 0; \quad \beta_1 \equiv 0; \quad \delta_{1k} = 0 = \delta_{j1}$$

Under this parametrisation, we interpret each parameter as:

- μ: the mean response of the "first" group (i.e. level 1 of factor A and level 1 of factor B)
- α_j: the expected change in the mean when changing only factor A to level j (but keeping factor B at level 1).
- β_k: the expected change in the mean when changing only factor B to level k (but keeping factor A at level 1).
- δ_{jk}: the additional change in the mean when changing factor A to level j and factor B to level k at the same time.

When we have a two-way ANOVA, our first point of interest is to test for possible interactions. If there are interactions, then it is *not* meaningful to talk about main effects. Only if there are no interactions can we then talk about each main effect in isolation.

13 Week 13

Linear Models

In general, a linear model assumes that the data has the following structure:

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$$

where $\underline{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ is a long vector of responses, and

$$X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \cdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}$$

is a $n \times p$ matrix of covariates / explaining variables. Here also, $\underline{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is a $p \times 1$ vector of corresponding *regression points*. Finally, $\underline{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$ is a $n \times 1$ vector of random deviations away from the "line".

The default ANOVA null hypothesis (i.e. that all of the groups are the same) is

$$H_0 : \alpha_2 = 0 = \alpha_3 = \cdots = \alpha_J$$

This can be written more succinctly in matrix form as

$$H_0 : R\underline{\beta} = 0$$

where $\underline{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$ and

$$R = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

The reason why linear models are so powerful is that we have a universal test for ANY restrictions placed on our model parameters. In general, for testing

$$H_0 : R\underline{\beta} \qquad \text{vs} \qquad H_1 : R\underline{\beta} \neq 0$$

we simply:

1. Fit the null model imposing the restriction
2. Fit the alternative model without restriction

and then we compare the residual sum-of-squares left over from both fits:

$$\frac{(RSS_0 - RSS_1)/r}{RSS_1/df} \sim \mathcal{F}_{r, df_1}$$

where r is the number of restrictions (which is $r = df_0 - df_1$, or $r = \text{rank}(R)$).
This is a generalisation of ANOVA to arbitrary restriction matrices R, and not just the default ANOVA null hypothesis.