Stat 2004 Second Half

Yutong Ji

Sep 2022

1 Pivotal quantity

Pivotal quantity, pivot variable, or pivot is a theoretical (not computable) transformation of a random variable (collected data) such that its distribution no longer depended on any unknown parameter.

Specifically, let $X = (X_1, X_2, \ldots, X_n)$ be a random sample from a distribution that depends on the unknown parameter θ (could also be a vector). Then g is called a pivotal quantity if $g(X, \theta)$ has the same distribution for any given θ - *which implies that it does not depend on the unknown parameter*
Aside: Pivotal quantity are commonly used for **normalisation** and are fundamental to construct test statistic (an example of Student's t-statistic has shown below)
One of the simplest pivot we have seen before is the *z-score* ($z = \frac{x-\mu}{\sigma}$), in the n-sample case:

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

Where, it has mean 0 and variance 1 - no longer depend on the unknown parameter.

2 Confidence Interval

The confidence interval is a range for an unknown parameter. It is constructed with three components, sample mean (by point estimation), critical value, standard error.
Formula:

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

Margin of error: $z\frac{s}{\sqrt{n}}$ expresses the amount of random sampling error in the result - closeness (how close to the population parameter)
Confident: the level of confidence, for instance 95% of confidence is, $95\% CI = \bar{x} \pm 1.96\frac{s}{\sqrt{n}}$ and we can say that we are 95% confident that the interval $\bar{x} \pm 1.96\frac{s}{\sqrt{n}}$ contains the true population mean (μ)

Interval estimation are more informative than a single point estimation (first half notes), since a CI gives us a sense of the uncertainty of an estimate along with a level of confidence. Moreover, CIs can be considered as a range or set of parameters values that are consistent with the data which CI is the inverse/complement hypothesis testings.

2.1 CI for μ when σ^2 is unknown

Suppose that $X = (X_1, X_2, \ldots, X_n) \sim^{iid} N(\mu, \sigma^2)$, in the pervious notes we see that:

1. $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

2. $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1} \Rightarrow \frac{s^2}{\sigma^2} \sim \frac{\chi^2_{n-1}}{n-1}$

3. $\bar{x}$ and s^2 are independent

Before we go further we need to verify those 3 derivations, the first two has been proven in the first half notes, so let's have a look at the third one:
We are about to use **Basu's theorem**: completed minimal sufficient statistic is independent of any ancillary statistic (pivotal quantity). We know that from the first half notes that, the sample mean is a complete minimal sufficient static and the unbiased sample variance can be written as $\hat{\sigma}^2 = \frac{\sum X_i - \bar{X}}{n-1}$ (moments and the biased-corrected representation) which does not depend on the unknown parameter, so that we can conclude that they are independence.

Now we can make a transformation:

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

Where, n-1 is the degree of freedom
Note that this T transformation is a theoretical transformation known as Student-T distribution (symmetric and bell-shape curve), since to complete this transformation we need to know the true population σ. Moreover, the result of this transformation no longer depend on the unknown parameters so that T is a pivot.

Furthermore, when we know that pivot, it is time to construct the CI.
To construct a (1 - α)% CI:

$$\begin{aligned}(1-\alpha)100\% &= P(-t_{n-1}^{1-\alpha/2} \leq T_{n-1} \leq t_{n-1}^{1-\alpha/2} \\ &= P(\bar{X} + \frac{t_{n-1}^{1-\alpha/2} * s}{\sqrt{n}} \geq \mu \geq \bar{X} - \frac{t_{n-1}^{1-\alpha/2} * s}{\sqrt{n}})\end{aligned}$$

In R: we can use: **qt**($(1-\alpha)$, df) to calculate the result.

2.2 CI for population variance

Suppose that $X = (X_1, X_2, \ldots, X_n) \sim^{iid} N(\mu, \sigma^2)$
We already know that: $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$ and $\frac{(n-1)s^2}{\sigma^2}$ is a pivotal quantity so that we could start to construct CI directly.
To construct a (1 - α)% CI:

$$\begin{aligned}(1-\alpha)100\% &= P(\chi_{n-1}^{2\alpha/2} \leq \chi^2_{n-1} \leq \chi_{n-1}^{21-\alpha/2}) \\ &= P(\frac{(n-1)s^2}{\chi_{n-1}^{2\alpha/2}} \geq \sigma^2 \geq \frac{(n-1)s^2}{\chi_{n-1}^{21-\alpha/2}})\end{aligned}$$

Note that chi-square distribution is **not** symmetric and is always positive, so that it is the reason why we do from $\alpha/2$ to $1-\alpha/2$.
Similar in R, we can do: **qchisq**($\frac{\alpha}{2}$ or $1-\frac{\alpha}{2}$, df)

2.3 CI for population proportion

Suppose that we want to investigate some traits (left hand, colour blind, ...) in a population. The people either have this trait or do not have this trait which implies each individual follows Bernoulli distribution with success probability p. Then a random sample size of n in the population would follow a Binomial distribution (Bin(n, p)). If we want to construct a confidence interval, we need to compute the pivot, but in general, pivot does not exist in discrete random variables.

However, we could construct **approximate/asymptotic pivots**. Suppose that x is the number of people in the sample that have the particular trait, where $X \sim Bin(n,p)$

1. the proportion is $\hat{p} = \frac{x}{n} \Rightarrow E(\hat{p}) = p$, due to the proportion converge to the success probability when sample is large. Mathematically, $E(\hat{p}) = E(\frac{x}{n}) = \frac{E(x)}{n} = \frac{np}{n} = p$
2. standard error of the proportion is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, due to definition of the standard error is $\frac{\text{sample standard deviation}}{\sqrt{\text{sample size}}}$
3. by Central Limit Theorem, when sample size is large, $\hat{p}$ has an asymptotic normal distribution

Then, we standardise the proportion $\hat{p}$: $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim N(0,1)$, so that $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$ is a asymptotic pivot.
Aside: although we can construct a asymptotic pivot, when p is close to 0 or 1, it becomes unreliable.
Since this result is based on the CLT which required sample size to be sufficiently large. A general rule is that the sample size must contain at least 10 successes and at least 10 failures.

2.4 CI procedure

Suppose that $X_1, X_2, \ldots, X_n$ be the random data coming from some population that depends on parameter θ, then the interval we constructed $(T_1(X), T_2(X))$ is an exact $(1-\alpha)100\%$ CI for θ, if ($T_1(X)$ and $T_2(X)$) are functions of the data alone such that these statistic capture the true θ:

$$P(T_1(X) \leq \theta \leq T_2(X)) = 1 - \alpha, \text{ for all } \theta$$

If we let $P(T_1(X) \leq \theta \leq T_2(X)) \geq 1 - \alpha$, for all θ, this is called a constructive CI with confidence level at least $(1-\alpha)100\%$

2.5 CI for two samples

will be added when i have more time

3 Hypothesis Test

Hypothesis testing is about making decisions about certain hypotheses on the basis of the observed data. There is no unique decision rule, however some rules are better than others.

3.1 null & alternative hypothesis

The null & alternative hypothesis are two competing claims about a population parameters (some numeric values on a population). The null hypothesis is denoted as H_0, generally contains the statement that is tested (so-called 'boring' hypothesis); whereas the alternative hypothesis is denoted as H_1 which generally contains the statement we hope or suspect is true (so-called 'interesting' hypothesis). Aside: mathematically, two hypothesis are symmetric, however there are many reasons to treat them as asymmetrically.

We mainly focus on a single point hypothesis, where H_0 is usually: $\theta = \theta_0$ and H_1 can be *one-side*: $\theta > \theta_0$, $\theta < \theta_0$ or *two-side*: $\theta \neq \theta_0$.

3.2 Test Statistic

The decision between H_0 and H_1 needs to be based on the observations $(X_1, X_2, \ldots, X_n)$ so that an appropriate summary statistic (only depends on data alone). So we write it as $T = T(X_1, X_2, \ldots, X_n)$ for our test statistic. Any function only depends on the observed data could be considered as a candidate for test statistic, however some principles can be useful to determine which one is great:

3.2.1 Method of moments

Suggesting we use $\bar{X}$ for testing population mean μ and use s^2 for testing population variance σ^2.

3.2.2 Maximum Likelihood Estimation

Alternatively, if we have a model for the population which falls into a distribution then we can use MLE as a starting point:

For example:

1. $\bar{X}$ for testing μ if $X \sim N(\mu, \sigma^2)$ or $X \sim Poi(\mu)$ and so on
2. $\frac{1}{\bar{X}}$ for testing population rate λ, if $X \sim Exp(\lambda)$
3. $max(X_1, X_2, \ldots, X_n)$ for testing population upper bound θ, if $X \sim U(0, \theta)$

The proof of result 2 is provided below:
Suppose $(X_1, X_2, \ldots, X_n)$ are idd random variables fall into exponential distribution.
The likelihood function is:

$$L(\lambda; x) = f_\lambda(x) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

Following MLE approach, we take the first derivative of the log-likelihood function:

$$\begin{aligned} \frac{d}{d\lambda} log(L(\lambda; x)) &= \frac{d}{d\lambda} log(\lambda^n e^{-\lambda \sum_{i=1}^{n} x_I}) \\ &= \frac{d}{d\lambda} nlog(\lambda) - \lambda \sum_{i=1}^{n} x_i \\ &= \frac{n}{\lambda} - \sum_{i=1}^{n} x_i \end{aligned}$$

then assign it to 0 to solve the function,

$$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

$$\frac{n}{\lambda} = \sum_{i=1}^{n} x_i$$

$$\frac{1}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\lambda = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i}$$

3.3 Critical regions

After specifying the parameters we are interested in and computing the test statistic, we are about to make a decision to reject or retain H_0 on the outcome of T. So we say that reject H_0 in favour of H_1, if T falls in the critical region.
The critical regions C could be one side or two side:
one-side:

- right-one-side: $C = [c, \infty]$, if T falls in this region, we could say we reject $H_0(\theta = \theta_0)$ in favour of $H_1(\theta > \theta_0)$ - overly large

- left-one-side: $C = [-\infty, c]$, if T falls in this region, we could say we reject $H_0(\theta = \theta_0)$ in favour of $H_1(\theta < \theta_0)$ - overly small

two-side:
$C = [-\infty, c_1] \cup [c_2, \infty]$, if T falls in this region, we could say we reject $H_0(\theta = \theta_0)$ in favour of $H_1(\theta \neq \theta_0)$ - overly large or overly small

3.4 Type I & Type II Errors

Whenever we made a decision based on some rules, we run into two type of errors:

- accidentally reject H_0 when H_0 is in fact true - type I error

- accidentally retain H_0 when H_0 is in fact false - type II error

3.4.1 Type I - significance level

The type I error is defined as:

$$\alpha = P(T(X) \in C|H_0)$$

Where it is interpreted as the probability of reject H_0 given that H_0 is true. α is also known as the significance level of a test.
In order to compute α, we need to know the distribution of $T(X)$ given H_0. We usually assume an explicit model for the data or apply the limit theorems (CLT, Fisher-tippett) for an approximate distribution.

A general procedure:
Suppose that we collect data with sample size n, where standard deviation is σ, and the decision

rule states that rejecting H_0 (true mean is μ) in favour of H_1, if sample mean ($\bar{x}$ as the test statistic) is $\geq$ c. Then,

$$\begin{aligned}\alpha &= P_{H_0}(\bar{x} \geq c) \\ &= P_{H_0}\underbrace{(\frac{\bar{x}-\mu}{SE} \geq \frac{c-\mu}{SE})}_{standardise}, SE = \frac{\sigma}{\sqrt{n}} \\ &\approx^{CLT} P_{H_0}(N(0,1) \geq \frac{c-\mu}{\sigma/\sqrt{n}})\end{aligned}$$

We could use R to calculate the result by $1 - pnorm(\frac{c-\mu}{\sigma/\sqrt{n}})$
Notice that only c (the specified critical region in the decision rule) is a variable and the rest are fixed values, which proves that if tight the critical region (by simply increase c), the type I got reduced.

Alternatively, sometimes we want to control type I error, for example, we want α to be 5% Aside:(5% is used in the general science, 10% is used in social science, and 1% normally is used in medicine) we can reverse the computation showed above.

Example: obtain the c with given that we want to achieve 5% significance level
With knowing that 5% has a critical value 1.645

$$\begin{aligned}0.05 &= P_{H_0}(N(0,1) \geq 1.645) \\ &\approx^{CLT} P_{H_0}(\frac{\bar{x}-\mu}{SE} \geq 1.645) \\ &= P_{H_0}(\bar{x} \geq 1.645 \times \frac{\sigma}{\sqrt{n}})\end{aligned}$$

Then, by letting $c = 1.645 \times \frac{\sigma}{\sqrt{n}}$ we can achieve $\alpha = 0.05$

3.4.2 Type II error - power

The type II error is defined as:

$$\beta = P(T(X) \notin C|H_1) = 1 - P(T(X) \in C|H_1)$$

And the power is define as:

$$power = 1 - \beta = P(T(X) \in C|H_1)$$

Where power is interpreted as the probability of rejecting H_0 when it is in fact false so that we often want high power.
If the specified H_1 tights the critical region (increasing c), the power got increase.

3.4.3 compromise two types of errors

There is no such a decision rule can simultaneously minimise type I error and type II error, however Negman-Pearson proposed a compromise between type I error and power. Aiming:

1. restricts and control type I error at suitably small (medicine/engineering has $\alpha = 0.01$, biological science has $\alpha = 0.05$)
2. maximise the power under the alternative

3.5 Likelihood ratio tests

Suppose that $H_0 : \theta \in \theta_0$ and $H_1 : \theta \in \theta_1$, then the likelihood ratio test statistic is defined as:

$$\frac{L(\theta_0|X)}{L(\theta_1|X)}$$

If the ratio is small, the data would have more likely observed under H_1. Hence it leads to the likelihood ratio test (LRT): reject H_0 in favour of H_1 if $\frac{L(\theta_0|X)}{L(\theta_1|X)} \leq c$, where c is some critical value.

3.5.1 Negman Pearson Lemma

One very important lemma derived from LRT is:
Suppose that LRT (rejects H_0 in favour of H_1 if $\frac{L(\theta_0|X)}{L(\theta_1|X)} \leq c$) that has level α, then any other tests that also have level α has power less than or equal to the LRT. Hence in their lemma, they state that LRT is the most powerful test for level α

3.6 LRT for normal mean under variance is known

Suppose $X = X_1, X_2, \ldots, X_n \sim^{iid} N(\mu, \sigma^2)$ and we are interested in testing two point hypothesis about μ:

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

In general, the likelihood function for $N(\mu.\sigma^2)$'s mean is:

$$L(\mu|X) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2}$$

Hence the LRT statistic is:

$$\begin{aligned}\frac{L(\mu_0|X)}{L(\mu_1|X)} &= \frac{(\frac{1}{\sqrt{2\pi\sigma^2}})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu_0)^2}}{(\frac{1}{\sqrt{2\pi\sigma^2}})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu_1)^2}} \\ &= \frac{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu_0)^2}}{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu_1)^2}} \\ &= e^{\frac{1}{2\sigma^2}(\sum_{i=1}^n (x_i-\mu_1)^2 - \sum_{i=1}^n (x_i-\mu_0)^2)}\end{aligned}$$

Since the exponential's range is always positive, smaller the value of the statistic is equivalence as smaller the values of its power, where smaller the value of $\sum_{i=1}^n (x_i - \mu_1)^2 - \sum_{i=1}^n (x_i - \mu_0)^2$.

To slove it:

$$\begin{aligned}\sum_{i=1}^n (x_i - \mu_1)^2 - \sum_{i=1}^n (x_i - \mu_0)^2 &= \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu_1)^2 - \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu_0)^2 \\ &= \sum_{i=1}^n [(x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - \mu_1) + (\bar{x} - \mu_1)^2] \\ &- \sum_{i=1}^n [(x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - \mu_0) + (\bar{x} - \mu_0)^2]\end{aligned}$$

Since $\sum_{i=1}^{n}(x_i - \bar{x})^2$ will cancel out and $\sum_{i=1}^{n} x_i - n\bar{x} = 0$, due to the definition of the sample mean. Hence, the evaluation ends as:

$$\sum_{i=1}^{n}(x_i - \mu_1)^2 - \sum_{i=1}^{n}(x_i - \mu_0)^2 = n(\bar{x} - \mu_1)^2 - n(\bar{x} - \mu_0)^2 = n[(\bar{x} - \mu_1)^2 - (\bar{x} - \mu_0)^2]$$

Notice that, now the LRT only depends on sample mean, so we can say that:

- if $\mu_1 > \mu_0$, $\frac{L(\theta_0|X)}{L(\theta_1|X)} \leq c$ if and only if $\bar{X} \geq d$, for some d.
- if $\mu_1 < \mu_0$, $\frac{L(\theta_0|X)}{L(\theta_1|X)} \leq c$ if and only if $\bar{X} \leq d$, for some d.
- if $\mu_1 \neq \mu_0$, $\frac{L(\theta_0|X)}{L(\theta_1|X)} \leq c$ if and only if $\bar{X} \neq d$, for some d.

How to obtain d? - by restricting the type I error:
if $\bar{X} \geq d$:

$$\begin{aligned}\alpha &= P_{H_0}(\bar{X} \geq d) \\ &\approx^{CLT} P_{H_0}(\frac{\bar{X} - \mu_0}{SE} \geq z^{1-\alpha}) \\ &= P_{H_0}(\bar{X} \geq \underbrace{\mu_0 + z^{1-\alpha} \times \frac{\sigma}{\sqrt{n}}}_{d})\end{aligned}$$

if $\bar{X} \leq d$:

$$P_{H_0}(\bar{X} \leq \underbrace{\mu_0 - z^{1-\alpha} \times \frac{\sigma}{\sqrt{n}}}_{d})$$

if $\bar{X} \neq d$:

$$P_{H_0}(\bar{X} \leq \underbrace{\mu_0 - z^{\frac{1-\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}}_{d}) \text{ or } P_{H_0}(\bar{X} \geq \underbrace{\mu_0 + z^{\frac{1-\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}}_{d})$$

3.7 What if the population variance is unknown?

We can estimate it using the sample variance (s^2):

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1} \text{ ,under } H_0 : \mu = \mu_0$$

So we use t-cutoffs instead of z-cutoffs:
if $\bar{X} \geq d$:

$$P_{H_0}(\bar{X} \geq \underbrace{\mu_0 + t_{n-1}^{1-\alpha} \times \frac{s}{\sqrt{n}}}_{d})$$

if $\bar{X} \leq d$:

$$P_{H_0}(\bar{X} \leq \underbrace{\mu_0 - t_{n-1}^{1-\alpha} \times \frac{s}{\sqrt{n}}}_{d})$$

if $\bar{X} \neq d$:

$$P_{H_0}(\bar{X} \leq \underbrace{\mu_0 - t_{n-1}^{\frac{1-\alpha}{2}} \times \frac{s}{\sqrt{n}}}_{d}) \text{ or } P_{H_0}(\bar{X} \geq \underbrace{\mu_0 + t_{n-1}^{\frac{1-\alpha}{2}} \times \frac{s}{\sqrt{n}}}_{d})$$

3.8 Example of LRT on Poisson distribution

Suppose that $X = (X_1, X_2, \ldots, X_n) \sim^{iid} Poisson(\lambda)$. $H_0 : \lambda = \lambda_0$ & $H_1 : \lambda = \lambda_1 > \lambda_0$:
First of all we need the log-likelihood function:

$$L(\lambda; X) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

Then we can apply them into LRT definition:

$$\frac{L(\lambda_0|X)}{L(\lambda_0|X)} = \frac{e^{-n\lambda_0} \frac{\lambda_0^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}}{e^{-n\lambda_1} \frac{\lambda_1^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}} = e^{-n(\lambda_1 - \lambda_0)} (\underbrace{\frac{\lambda_0}{\lambda_1}}_{<1})^{\sum_{i=1}^n x_i}$$

Since $\lambda_1 > \lambda_0$, the LRT becomes smaller when $\sum_{i=1}^n x_i$ is larger so that the LR test for testing between H_0 and H_1 can be interpreted as:

$$\sum_{i=1}^{n} x_i \geq d$$

Where d is the summation of all the sample data. Note that, the final decision rule does not dependent on the value of λ, it is in fact the uniformly most powerful test for testing null & alternative hypothesis.

Now we just need to decide what d is:
Given that the summation of each iid poisson random variable, we can say that: $X = (X_1, X_2, \ldots, X_n) \sim Poisson(n\lambda)$. Hence the type I error can be construct as:

$$\alpha \geq P_{H_0}(X_1, X_2, \ldots, X_n \geq d) = P_{H_0}(Poisson(n\lambda) \geq d)$$

So that d must be the upper α quantile.
In R, we could do "$qpois(1 - \alpha, n\lambda)$"

Alternatively, we could use t-test to achieve an approximately α compared with LRT, however if the datasets are truly from Poisson distribution, the LRT has more power than t-test. On the other hand, if the data is not from Poisson distribution, the type I error from LRT may not be close to α and power would be much lower as well. Where t-test would always give the approximately correct α and asymmetrically power.

Back up verification for sum of n idd Poison(λ) is Poision(nλ):
Suppose $X\ Poisson(\lambda_1)$ and $Y\ Poisson(\lambda_2)$ and let $Z = X + Y$:

$$
\begin{aligned}
P(Z=n) = P(X+Y=n) &= \underbrace{\sum_{k=0}^{n} P(X=k)P(Y=n-k)}_{\text{convolution}} \\
&= \sum_{k=0}^{n} e^{-\lambda_1}\frac{\lambda_1^k}{k!} e^{-\lambda_2}\frac{\lambda_2^{n-k}}{(n-k)!} \\
&= e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{\lambda_1^k \lambda_2^{n-k}}{k!(n-k)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!}\lambda_1^k \lambda_2^{n-k} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{n!}(\lambda_1+\lambda_2)^n
\end{aligned}
$$

The last step above uses the binomial theorem where: $(a+b)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) a^k b^{n-k}$

3.9 P-values

For everything that discussed before, we have not yet collected any numerical data. This is done on purpose since we have to make sure that:

- identifying population parameters of interest
- formulating null & alternative hypothesis test about parameters
- choosing the test statistic $T(X)$ and critical region c
- setting type I error by choosing critical value d
- accessing the power of the test

All the things above have to be done **before** looking at any numerical data.
So, after all of these, we can carry out experiments and collect data, then make a decision. Moreover, we can also supplement our decision with a p-values, which is an indicator of the strength of evidence against our null hypothesis in favour of the alternative hypothesis.

Suppose that $x = x_1, x_2, \ldots, x_n$ of $X = X_1, X_2, \ldots, X_n$ are the collected data. Then we can compute the observed value of test statistic by

$$t_{obs} = T(x)$$

The p-values can be computed by:

- if $H_0 : \theta = \theta_0$ & $H_1 : \theta > \theta_0$, $p-values = P_{H_0}(T(x) \geq t_{obs})$
- if $H_0 : \theta = \theta_0$ & $H_1 : \theta < \theta_0$, $p-values = P_{H_0}(T(x) \leq t_{obs})$
- if $H_0 : \theta = \theta_0$ & $H_1 : \theta \neq \theta_0$, $p-values = 2 \times min\{P_{H_0}(T(x) \geq t_{obs}), P_{H_0}(T(x) \leq t_{obs})\}$

In all 3 cases, we can observe that p-values is always the probability under null hypothesis that the test statistic would be as unusual or more unusual than the observation.

The general threshold of p-values is:

- very strong evidence against H_0 in favour of H_1: $0 \sim 0.01$
- reasonably strong evidence against H_0 in favour of H_1: $0.01 \sim 0.05$
- some (borderline, marginal, weak) evidence against H_0 in favour of H_1: $0.05 \sim 0.1$
- no evidence against H_0 in favour of H_1: > 0.1

3.10 Example continued from 3.8

Suppose we have done n=15 experiments and collect data as $7, 11, 10, 11, 6, 7, 10, 9, 7, 5, 11, 5, 9, 6, 18$ (sum is 132). And we are testing $H_0 : \lambda = 7.3$ & $H_1 : \lambda > 7.3$ with aiming a 5% significance level.

First of all, computing d:

$$0.05 \geq P(Pois(15 \times 7.3) \geq d) = P(Pois(109.5) \geq d)$$

By using R: $qpois(0.95, 109.5) = 127 \equiv P(Pois(109.5) \geq 128)$, hence d is 128.

$$\sum_{i=1}^{15} x_i \geq 128$$

In conclusion, based on the observed data (sum is 132 - fall in the critical region), we can decided to reject H_0 in favour of H_1.
However, what is the p-values?

$$P - values = P_{H_0}(T(x) \geq t_{obs}) = P_{H_0}(Poisson(109.5) \geq 132)$$

By using R: $1 - ppois(131, 109.5) = 0.02$. This can be interpreted as if the null hypothesis claim $H_0 : \lambda = 7.3$ is true, then the observed total from 15 experiments resulting in 132 or even more would occur 2% of the time in the long run. This is quite rare so that there is reasonably strong evidence against $H_0 : \lambda = 7.3$ in favour of $H_1 : \lambda > 7.3$.

3.11 Duality between CIs & hypothesis testings

Background events: since p-values are often misinterpreted, some journal banned hypothesis test in favour of CIs. However does this really improved?

First of all, CIs are misinterpreted quite often as well.
Secnondly, more importantly, there is a duality between CIs & hypothesis testings. Given a dataset, a CI provides a range/set of parameter values which true value would have reasonably likely have occurred - range of plausible parameter values. On the other hand, given a parameter values, a hypothesis test specifies a range of data values would have been unlikely to be observed.
We can see the CIs are the inverse of hypothesis testing.

For example, a two-sided test, the reject region is: $\bar{X} \leq \mu_0 - t_{n-1}^{\frac{1-\alpha}{2}} \times \frac{s}{\sqrt{n}}$ or $\bar{X} \geq \mu_0 + t_{n-1}^{\frac{1-\alpha}{2}} \times \frac{s}{\sqrt{n}}$
Thus, the accept region is: $\mu_0 - t_{n-1}^{\frac{1-\alpha}{2}} \times \frac{s}{\sqrt{n}} \leq \bar{X} \leq \mu_0 + t_{n-1}^{\frac{1-\alpha}{2}} \times \frac{s}{\sqrt{n}}$
Also equivalence to: $\bar{X} - t_{n-1}^{\frac{1-\alpha}{2}} \times \frac{s}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + t_{n-1}^{\frac{1-\alpha}{2}} \times \frac{s}{\sqrt{n}}$

Therefore, when pivot variable cannot be constructed, we can still construct CIs by inverting hypothesis test.

3.12 Generalised likelihood ratio test - GLRT

Likelihood ratio tests can be generalised to scenarios when the null & alternative hypothesis are not a single point, but is multidimensional.

Suppose now we declare that the null hypothesis as $H_0 : \theta \in \{H_0\}$, where $\{H_0\}$ is a subset of the full parameter space $\{H\}$ and alternative hypothesis as $H_1 : \theta \in \{H_1\}$, where $\{H_1\}$ is another subset of the full parameter space which disjoint with $\{H_0\}$. If we are considering two-sided test: $\{H_1\} = \{H_0\}^c$ (complement).
Mathematically, the GLRT statistic is defined as:

$$\frac{\underset{\theta\in\{H_0\}}{sup}\ L(\theta|X)}{\underset{\theta\in\{H_1\}}{sup}\ L(\theta|X)}$$

Where $\underset{\theta\in\{H_0\}}{sup}\ L(\theta|X)$ is the best possible likelihood under H_0 and $\underset{\theta\in\{H_1\}}{sup}\ L(\theta|X)$ is the best possible likelihood under H_1

If this ratio is small, then the data would be more likely occurred under $\{H_1\}$

3.12.1 Example: comparing two group means under the same variance

Suppose that:

- $X_1 = (X_{11}, X_{12}, \ldots, X_{1n_1}) \overset{iid}{\sim} N^*(\mu_1, \sigma^2)$
- $X_2 = (X_{21}, X_{22}, \ldots, X_{2n_2}) \overset{iid}{\sim} N^*(\mu_2, \sigma^2)$

We want to test that whether the two groups' means are equal: $H_0 : \mu_1 = \mu_2$ & $H_1 : \mu_1 \neq \mu_2$

The likelihood function is:

$$L(\mu_1, \mu_2|X) = (\frac{1}{\sqrt{2\pi\sigma^2}})^{n_1+n_2} exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(X_{1i}-\mu_1)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n_2}(X_{2i}-\mu_2)^2]$$

For numerator: $\underset{\theta\in\{H_0\}}{sup}\ L(\mu_1, \mu_2|X)$
We can use the common mean to maximise it under null hypothesis ($H_0 : \mu_1 = \mu_2$), since the power of the exponential is negative, we want to make it close to zero as much as possible.

$$\hat{\mu_1} = \hat{\mu_2} = \frac{\sum_{i=1}^{n_1} X_{1i} + \sum_{i=1}^{n_2} X_{2i}}{n_1 + n_2} = \bar{X}_{..}$$

Therefore, the maximum likelihood achieved under null is:

$$\begin{aligned}\underset{\theta\in\{H_0\}}{sup}\ L(\mu_1, \mu_2|X) &= L(\bar{X}_{..}, \bar{X}_{..}|X) \\ &= (\frac{1}{\sqrt{2\pi\sigma^2}})^{n_1+n_2} exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_{..})^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n_2}(X_{2i}-\bar{X}_{..})^2]\end{aligned}$$

Again, to maximum it, we need to look at the exponent, since others are just constants:

$$\underbrace{\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_{..})^2+\sum_{i=1}^{n_2}(X_{2i}-\bar{X}_{..})^2}_{\text{total sum of squared deviation - }SS_{total}}=\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_{1.}+\bar{X}_{1.}-\bar{X}_{..})^2\textcircled{1}$$

$$+\sum_{i=1}^{n_2}(X_{2i}-\bar{X}_{2.}+\bar{X}_{2.}-\bar{X}_{..})^2\textcircled{2}$$

Where, $\bar{X_1}.$ is the sample mean of group 1 and $\bar{X_2}.$ is the sample mean of group 2.
For $\textcircled{1}$:

$$\begin{aligned}\sum_{i=1}^{n_1}(\underbrace{X_{1i}-\bar{X}_{1.}}_{a}+\underbrace{\bar{X}_{1.}-\bar{X}_{..}}_{b})^2&=\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_{1.})^2\\&+2(\bar{X}_{1.}-\bar{X}_{..})\underbrace{\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_{1.})}_{=0}\\&+\sum_{i=1}^{n_1}(\bar{X}_{1.}-\bar{X}_{..})^2\\&=\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_{1.})^2+n_1(\bar{X}_{1.}-\bar{X}_{..})^2\end{aligned}$$

Similarly for $\textcircled{2}$:

$$\sum_{i=1}^{n_2}(X_{2i}-\bar{X}_{2.}+\bar{X}_{2.}-\bar{X}_{..})^2=\sum_{i=1}^{n_2}(X_{2i}-\bar{X}_{2.})^2+n_2(\bar{X}_{2.}-\bar{X}_{..})^2$$

Hence overall:

$$\begin{aligned}SS_{total}=\textcircled{1}+\textcircled{2}=\quad&\underbrace{\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_{1.})^2+\sum_{i=1}^{n_2}(X_{2i}-\bar{X}_{2.})^2}_{\text{within groups' sum of squared deviation / residual sum of squares}}\\&+\underbrace{n_1(\bar{X}_{1.}-\bar{X}_{..})^2+n_2(\bar{X}_{2.}-\bar{X}_{..})^2}_{\text{between groups sum of squares}}\end{aligned}$$

For denominator: $\underset{\theta\in\{H_1\}}{sup}\ L(\theta|X)$
Since the alternative hypothesis states that $H_1:\mu_1\neq\mu_2$, we can use $\hat{\mu_1}=\bar{X}_{1.}$ and $\hat{\mu_2}=\bar{X}_{2.}$ to maximise it.

$$\begin{aligned}\underset{\theta\in\{H_1\}}{sup}\ L(\theta|X)&=L(\bar{X}_{1.},\bar{X}_{2.}|X)\\&=(\frac{1}{\sqrt{2\pi\sigma^2}})^{n_1+n_2}exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_{1.})^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n_2}(X_{2i}-\bar{X}_{2.})^2]\\&=(\frac{1}{\sqrt{2\pi\sigma^2}})^{n_1+n_2}exp[-\frac{1}{2\sigma^2}\underbrace{\text{within group }SS_{total}}_{\text{from above}}]\end{aligned}$$

In confusion: For GLRT $\frac{\underset{\theta\in\{H_0\}}{sup}\ L(\theta|X)}{\underset{\theta\in\{H_1\}}{sup}\ L(\theta|X)}$:

The numerator can be expressed as:

$$(\frac{1}{\sqrt{2\pi\sigma^2}})^{n_1+n_2}exp[-\frac{1}{2\sigma^2}(\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_{1.})^2+\sum_{i=1}^{n_2}(X_{2i}-\bar{X}_{2.})^2+n_1(\bar{X}_{1.}-\bar{X}_{..})^2+n_2(\bar{X}_{2.}-\bar{X}_{..})^2] \quad (1)$$

The denominator can be expressed as:

$$(\frac{1}{\sqrt{2\pi\sigma^2}})^{n_1+n_2}exp[-\frac{1}{2\sigma^2}(\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_{1.})^2+\sum_{i=1}^{n_2}(X_{2i}-\bar{X}_{2.})^2)] \quad (2)$$

So the GLRT statistic reduced to:

$$\Lambda = \frac{\underset{\theta\in\{H_0\}}{sup}\ L(\theta|X)}{\underset{\theta\in\{H_1\}}{sup}\ L(\theta|X)} = \frac{(1)}{(2)} = exp[-\frac{1}{2\sigma^2}(n_1(\bar{X}_{1.}-\bar{X}_{..})^2+n_2(\bar{X}_{2.}-\bar{X}_{..})^2)]$$

From the reduced equation, the ratio of Λ is small when the between groups SS_{total} is large, where when the mean of two groups are far apart. Hence, the GLRT simplifies the rule to be reject $H_0 : \mu_1 = \mu_2$ in favour of $H_1 : \mu_1 \neq \mu_2$ if:

$$[\bar{X}_{1.} - \bar{X}_{2.} \leq d_1] \cup [\bar{X}_{1.} - \bar{X}_{2.} \geq d_2]$$

So now, how do we choose the critical value (d_1 and d_2) to control the type 1 error? Recall that (two sample):

$$\bar{X}_{1.} - \bar{X}_{2.} \sim N(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2})$$

Under the null hypothesis, $\mu_1 - \mu_2 = 0$ or in general $\mu_1 - \mu_2 = \Delta_0$ (some specific value) Then we can set the critical value to be:

$$d_1 = \Delta_0 - z^{1-\frac{\alpha}{2}}\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$d_2 = \Delta_0 + z^{1-\frac{\alpha}{2}}\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

This is the 2-sample z-test. (Precondition: we know the population variance)

3.12.2 What if we do not know σ^2 as well?

We could estimate the σ^2 by the **pooled variance** - weighted average of two individual sample variance.

Aside: Suppose we have m samples: $X = (X_1, X_2, \ldots, X_m)$, with sample size $n_1, n_2, \ldots, n_m$ respectively. Then mathematically, the pooled variance can be defined as:

$$s^2_{pooled} = \frac{\sum_{i=1}^m (n_i-1)s_i^2}{\sum_{i=1}^m (\underbrace{n_i-1)}_{\text{biased corrected}}} = \frac{(n_1-1)s_1^2+\ldots+(n_m-1)s_m^2}{n_1+\ldots+n_m-m}$$

So for a two samples:

$$s^2_{pooled} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

And we know that (previously section), we could use t-quantile to replace z-quantile:

$$d_1 = \Delta_0 - t^{1-\frac{\alpha}{2}}_{n_1+n_2-2}\sigma_{pooled}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$d_2 = \Delta_0 + t^{1-\frac{\alpha}{2}}_{n_1+n_2-2}\sigma_{pooled}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

This is how two-sample t-test formed.

3.13 Goodness-of-fit test (categorial data)

Now, we have looked some quantitive data for LRT, surprisedly, LRT can also derive some useful test for categorial data.

Recall that, the test statistic for goodness-of-fit is:

$$\sum_I = \frac{(observed_i - expected_i)^2}{expected_i} \sim \text{ Pearson's } \chi^2$$

How does this statistic come up?

Suppose that $X = (X_1, X_2, \ldots, X_k)$ are the counts in each category from a random sample of n units, where the total counts $X_1 + X_2 + \ldots, +X_k = n$.

We can see that $X = (X_1, X_2, \ldots, X_k) \sim Multinomial(p_1, p_2, \ldots, p_k)$, where p are denoted as the probability of the event. Then we begin with a single point null hypothesis.

- $H_0 : p = (p_1, p_2, \ldots, p_k) = p^* = (p_1^*, p_2^*, \ldots, p_k^*)$, where p^* is the given of proportions.
- $H_1 : p \neq p^*$ - interpret as there is at least one proportion is not correct

Construct the LRT:

$$\Lambda = \frac{L(p^*|X)}{L(p|X)}$$

The MLE of the p under H_1 is the sample proportion.

Verification:

Suppose that $X = (X_1, X_2, \ldots, X_k) \sim Multinomial(p_1, p_2, \ldots, p_k)$.

Then:

$$PMF = \frac{n!}{x_1!, x_2!, \ldots, x_k!}p_1^{x_1} \times p_2^{x_2} \times \cdots \times p_n^{x_k} = n!\prod_{i=1}^{k}\frac{p_i^{x_i}}{x_i!}$$

Where n is total occurrences, k is the number of incidences in the category

The log-likelihood function is:

$$\begin{aligned} log(L(p|X)) &= log(n!\prod_{i=1}^{k}\frac{p_i^{x_i}}{x_i!}) \\ &= log(n!) + \sum_{i=1}^{k} x_i log(p_i) - \sum_{i=1}^{k} log(x_i!) \end{aligned}$$

As we know the constraint is $\sum_{i=1}^{k} p_i = 1$, so we would apply Lagrange multiplier: $L(x, \lambda) = f(x) - \lambda g(x)$, where $g(x)$ is an equality constraint.
Hence:

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

Where we see here, it makes no difference since $1 - \sum_{i=1}^{k} p_i = 0$ always.
Then on differentiating the Lagrange multiplier:

$$\frac{d}{dp_i} log(L(p, \lambda)) = \frac{d}{dp_i} log(L(p)) + \lambda(1 - \sum_{i=1}^{k} p_i)$$

$$since, \frac{d}{dp_i}\lambda(1 - \sum_{i=1}^{k} p_i) = \underbrace{\frac{d}{dp_i}\lambda}_{=0} - \underbrace{\frac{d}{dp_i}\lambda p_1}_{=\lambda \text{ if i is 1, else 0}} - \ldots - \underbrace{\frac{d}{dp_i}\lambda p_k}_{=\lambda \text{ if i is k, else 0}}$$

and we know that each time i would be one of $\{$1, 2, ... ,k$\}$ so,

$$\frac{d}{dp_i}\lambda(1 - \sum_{i=1}^{k} p_i) = \lambda$$

$$= \frac{d}{dp_i}(log(n!) + \sum_{i=1}^{k} x_i log(p_i) - \sum_{i=1}^{k} log(x_i!)) - \lambda$$

$$= \frac{x_i}{p_i} - \lambda$$

So $p_i = \frac{x_i}{\lambda} \Rightarrow$ by applying the constraint $\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} \frac{x_i}{\lambda} \Rightarrow 1 = \sum_{i=1}^{k} \frac{x_i}{\lambda} \Rightarrow \lambda = n$
By substituting back, we got $p_i = \frac{x_i}{n}$

Then we could use sample proportion $(\hat{p}_1 = \frac{X_1}{n}, \ldots, \hat{p}_k = \frac{X_k}{n})$ to apply the LRT under H_1:

$$\Lambda = \frac{\binom{n}{x} p_1^{*x_1} \times p_2^{*x_2} \times \ldots \times p_k^{*x_k}}{\binom{n}{x} \hat{p}_1^{x_1} \times \hat{p}_2^{x_2} \times \ldots \times \hat{p}_k^{x_k}}$$

$$= (\frac{p_1^*}{\hat{p}_1})^{x_1} \times (\frac{p_2^*}{\hat{p}_2})^{x_2} \times \ldots \times (\frac{p_k^*}{\hat{p}_k})^{x_2}$$

The trick we can do here is negative-log-transform. Since the log function is monotone increasing. So small value of $\Lambda \equiv$ small value of $log(\Lambda) \equiv$ large value of $-log(\Lambda)$.

$$-log(\Lambda) = x_1 log(\frac{\hat{p}_1}{p_1^*}) + x_2 log(\frac{\hat{p}_2}{p_2^*}) + \ldots + x_k log(\frac{\hat{p}_k}{p_k^*})$$

$$\text{since } \hat{p}_i = \frac{x_1}{n}$$

$$= x_1 log(\frac{x_1}{np_1^*}) + x_2 log(\frac{x_2}{np_2^*}) + \ldots + x_k log(\frac{x_k}{np_k^*})$$

since p_i^* is the true proportion for the i-th category,
np_i^* would be its expected counts

$$= O_1 log(\frac{O_1}{E_1}) + O_2 log(\frac{O_2}{E_2}) + \ldots + O_k log(\frac{O_k}{E_k})$$

$$= \sum_{cells_i} O_i log(\frac{O_i}{E_i})$$

We end up with the exact form of GLRT statistic for categorial data. However, why do we use the Pearson's $\chi^2 = \frac{(observed_i - expected_i)^2}{expected_i}$?

The reasons are:

1. the distribution of $-log(\Lambda) = \sum_{cells_i} O_i log(\frac{O_I}{E_i})$ is not easy to determine
2. the first-order Taylor expansion of the Pearson's χ^2 is approximate to $-2log(\Lambda)$, since the leading term of Taylor expand of $-2log(\Lambda)$ is the Pearson's χ^2
3. the distribution of Pearson's χ^2 is easier to derive, because Pearson's χ^2 converges to χ^2 distribution for large n
4. Pearson's χ^2 is published in 1900, back the day without calculator, it is easier to compute by hand

3.14 Application of goodness-of-fit test: Mendel's data

In 1865, Mendel publish a paper that illustrate the offsprings (total counts is 556) of crossed produced yellow-smooth-male & green-wrinkly-female peas are:

Type	Frequency	Mendel's counts	Expected counts
Smooth yellow	$\frac{9}{16}$	315	312.75
Smooth green	$\frac{3}{16}$	108	104.25
Wrinkly yellow	$\frac{3}{16}$	102	104.25
Wrinkly green	$\frac{1}{16}$	31	34.75

Now we are interested in, are Mendel's counts consistent with the frequency derived from his theory?

- $H_0 : p = (p_1, p_2, p_3, p_4)^T = (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})^T = p^*$
- $H_1 \neq p^*$ or at least one proportion is wrong

Constructing by the exact GLRT statistic:

$$\begin{aligned}
-2log(\Lambda) &= 2\sum_{cells_i} O_i log(\frac{O_i}{E_i}) \\
&= 2[315 \times log(\frac{315}{312.75}) + 108 \times log(\frac{108}{104.25}) \\
&+ 102 \times log(\frac{102}{104.25}) + 31 \times log(\frac{31}{34.75})] \\
&= 0.618
\end{aligned}$$

Constructing by the Pearson's χ^2:

$$\text{Pearson's } \chi^2 = \frac{(observed_i - expected_i)^2}{expected_i} = 0.604$$

Under the null hypothesis, the statistic has an approximate χ^2 with the degrees of the freedom = k - 1 = 4 - 1 = 3.
Hence, the p-value for the test is:

$$P(\chi^2_3 \geq 0.604/0.618) \approx 0.9$$

With such a large p-value, the conclusion for now is Mendel's counts are consistent with his theory (retain null hypothesis) / there is no evidence against Mendel's theory. However, we cannot say that Mendel's counts support his theory, since the theory might be refined. Moreover, the reason for it is actually, we can never find evidence for a null hypothesis by random experiments (you cannot prove some theories with only one example).

Later, Fisher noticed that in here, the p-value is quite large, where the observed counts are very well alined with the expected counts. Then Fisher went to examine all of Mendel's data in every experiments. Fisher pooled all counts together and calculate the p-value to be 0.99996 for testing the null - something odd here...

While getting a small p-value makes us suspicious about the null (reject), however getting a very large p-value would make us suspicious about the data - could be either fabricated or generated by pseudo-replication.

- data fabrication: the real data should look like the proportion is closed to the expected when the sample size is increase, not for any sample size that they are all very close to the expected proportion. *By adding or subtracting 2 times its standard error won't make it suspicious.*
- pseudo replication: making multiple measurements on each unit but pretend that they are individual from independent units. For example, applying 2 pesticides to two trees and we take 1000 leaves on each tree, then we claim that we collect 2000 individual samples is a typical pseudo replication.

3.15 Working with two or more categorial data

When we have two categorial data, we can know a bit more than just one categorial data, such as the independence and homogeneity (sample from different sample might or might not be identical). The test of independence and homogeneity by using GLRT or Pearson's χ^2 are the same, however the way of collecting data are different.

3.15.1 independence test

Samples are from a **single** population and then cross-clarified based on two or more categorial factors.

- $H_0 : p_{ij} = p_{i.} \times p_{.j}$
- $H_1 : p_{ij} \neq p_{i.} \times p_{.j}$ for at least one pair of i and j.

Where p_{ij} is the population proportion of being in both factor i and j. $p_{i.} \& p_{.j}$ are population proportion in factor i & j respectively.

3.15.2 homogeneity test

Samples are separately collected from **two subpopulations** based on a categorial factor. Then we are interested in whether the distribution of the categorial factor are homogenous across those two subpopulations.

Example: below is a table of computer replacement in UQ's staffs (in Science, Engineering, and Arts faculty).

OS	Science	Engineering	Arts	Total
Windows	12 (12.51)	17 (13.90)	9 (11.58)	38
Mac	8 (10.54)	9 (11.71)	15 (9.75)	32
Linux	7 (3.95)	4 (4.39)	1 (3.66)	12
Total	27	30	25	82

Where the value in the parenthesis is the expected value.
For instance, the expected counts for staff in Science who preferred Windows replacement is: $E = np =$ total of science staff $\times \frac{\text{total of Windows users}}{\text{total of staffs}} = 27 \times \frac{38}{82} = 12.51$

- $H_0 : p_{science} = p_{engineering} = p_{arts} = p_{common}$, where $p_i = (p_w, p_m, p_L)_i^T$
- H_1 : not all vectors of probabilities at the same or equivalently, there are no restrictors on each vector of probability
- The number of free parameters under the alternative: (in general) (#rows - 1) × #columns, in this case, 2 ×3 = 6
- The number of free parameters under the null: (in general) #rows -1, in this case, 3 - 1 = 2
- The degree of freedom is: c(r-1) - (r-1) = (c-1)(r-1), in this case, 2 × 2 = 4

Constructing test statistic:

$$\begin{aligned}\text{Pearson's } \chi^2 &= \sum_{cells} \frac{(observed_i - expected_i)^2}{expected_i} \\ &= \frac{(12-12.51)^2}{12.51} + \frac{(17-13.90)^2}{13.90} + \ldots + \frac{(1-3.66)^2}{3.66} \\ &= 9.66\end{aligned}$$

Finding p-value:

$$P(\chi^2_{df=4} \geq 9.66) = 1 - pchisq(9.66, 4) = 0.047 = 4.7\%$$

So the confusion for now is: there is moderately strong evidence that computer operation system preference may differ across the 3 different facilities.

Is something wrong here?
Recall that, to be able to do Normal approximation on a binomial, we need at least 10 successes and 10 failures. Similarly for Pearson's χ^2 test, we need:

- At least 80% of cells in the table have more than or equal to 10 counts
- all cells must have at least 5 counts

In the example showed above, it does not satisfied that condition, so in this case we have to consider the Fisher exact test.

3.16 Fisher exact test

In general, Fisher exact test uses the definition of p-values (how often we see unusual data?) and works on any sample size, however we only can do it by computers since we need a large random data sets. Steps:

1. argue that why the rows total and columns total can be fixed or given, under the null hypothesis. For instance, in previous example, maybe the computer can be changed in 4 years, now this years total number of staffs who wants to change their computer is depend on how many staff required for a computer 4 years ago - total staffs in each faculty (column). As well as, some staff prefer windows so that no matter which faculty the staff is in, the staff always pick windows - total staffs who use windows/Mac/linux (row).

2. under the null hypothesis, any configuration of counts with same row total and column total would have been equally likely to be observed - assuming the proportion are the same.

3. simulate many different table with the same row total and column total and then determine how often we get something as unusual or more unusual

Referring to pervious example, we can do:

$$
\begin{aligned}
&counts = c(12, 17, 9, 8, 9, 15, 7, 4, 1)\\
&table2 = matrix(counts, nrow = 3, byrow = T)\\
&chisq.test(table2, simulate.p.value = T, B = 1000000)
\end{aligned}
$$

3.17 Comparing groups using quantitative data

Analysis of variance (ANOVA) is used to compare a quantitative factor. Interestingly, ANOVA is also a special case of linear model.

Suppose we have J groups that:

$$
\begin{aligned}
Y_{11}, Y_{12}, \ldots, Y_{1n_1} &\overset{iid}{\sim} N^*(\mu_1, \sigma^2)\\
\text{Independent with, } Y_{21}, Y_{22}, \ldots, Y_{2n_2} &\overset{iid}{\sim} N^*(\mu_2, \sigma^2)\\
\text{Independent with, } Y_{31}, Y_{32}, \ldots, Y_{3n_3} &\overset{iid}{\sim} N^*(\mu_3, \sigma^2)\\
&\cdots\\
\text{Independent with, } Y_{J1}, Y_{J2}, \ldots, Y_{Jn_J} &\overset{iid}{\sim} N^*(\mu_J, \sigma^2)
\end{aligned}
$$

Define our two hypothesis:

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_J$
- H_1 : not all group means are the same or equivalent, there are no restrictors on the groups means

This is really extending the two-sample t-test (pervious notes) to more than two groups: using GLRT:

$$
\Lambda = \frac{\underset{H_0}{sup} L(\mu_1, \mu_2, \mu_3, \ldots, \mu_J | Y)}{\underset{H_1}{sup} L(\mu_1, \mu_2, \mu_3, \ldots, \mu_J | Y)}
$$

The likelihood function is:

$$\begin{aligned}
L(\mu_1,\mu_2,\mu_3,\ldots,\mu_J|Y) &= \prod_{j=1}^{J}(\frac{1}{\sqrt{2\pi\sigma^2}})^{n_j} exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n_j}(Y_{ji}-\mu_j)^2] \\
&= (\frac{1}{\sqrt{2\pi\sigma^2}})^{n_1} exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(Y_{1i}-\mu_1)^2] \\
&\times (\frac{1}{\sqrt{2\pi\sigma^2}})^{n_2} exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n_2}(Y_{2i}-\mu_2)^2] \\
&\ldots \\
&\times (\frac{1}{\sqrt{2\pi\sigma^2}})^{n_J} exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n_J}(Y_{Ji}-\mu_J)^2] \\
&= (\frac{1}{\sqrt{2\pi\sigma^2}})^{N} exp[-\frac{1}{2\sigma^2}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ji}-\mu_j)^2]
\end{aligned}$$

Where $N = n_1 + n_2 + \ldots + n_J$ is the total sample size.

So now we can have a look like the likelihood function under the null and under the alternative:

Under the null: $H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_J$, we can use common mean (MLE):

$$(\frac{1}{\sqrt{2\pi\sigma^2}})^{N} exp[-\frac{1}{2\sigma^2}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ji}-\bar{Y_{..}})^2]$$

Where $\bar{Y_{..}}$ is the common mean, such that $\bar{Y_{..}} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} Y_{ji}}{N}$ - the average in the overall sample

Under the alternative:

$$(\frac{1}{\sqrt{2\pi\sigma^2}})^{N} exp[-\frac{1}{2\sigma^2}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ji}-\bar{Y_{j.}})^2]$$

Where $\bar{Y_{j.}}$ is the mean in each j-th group, such that $\bar{Y_{j.}} = \frac{\sum_{i=1}^{n_j} Y_{ji}}{n_j}$

We can work on a bit of the numerator here, to simplify the expression, we can use the same methodology in before - adding $\bar{Y_{j.}}$ then subtracting $\bar{Y_{j.}}$ to make a pair.

$$\begin{aligned}
\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ji}-\bar{Y_{..}})^2 &= \sum_{j=1}^{J}\sum_{i=1}^{n_j}(\underbrace{Y_{ji}-\bar{Y_{j.}}}_{a}+\underbrace{\bar{Y_{j.}}-\bar{Y_{..}}}_{b})^2 \\
&= \sum_{j=1}^{J}\sum_{i=1}^{n_j}\underbrace{(Y_{ji}-\bar{Y_{j.}})^2}_{a^2}+\sum_{j=1}^{J}\sum_{i=1}^{n_j}\underbrace{2(Y_{ji}-\bar{Y_{j.}})(\bar{Y_{j.}}-\bar{Y_{..}})}_{2ab} \\
&+\sum_{j=1}^{J}\sum_{i=1}^{n_j}\underbrace{(\bar{Y_{j.}}-\bar{Y_{..}})^2}_{b^2} \\
&= \sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ji}-\bar{Y_{j.}})^2+2\sum_{j=1}^{J}(\bar{Y_{j.}}-\bar{Y_{..}})\underbrace{\sum_{i=1}^{n_j}(Y_{ji}-\bar{Y_{j.}})}_{=0} \\
&+\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\bar{Y_{j.}}-\bar{Y_{..}})^2 \\
&= \sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ji}-\bar{Y_{j.}})^2+\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\bar{Y_{j.}}-\bar{Y_{..}})^2
\end{aligned}$$

Where it is similar with two sample t-test, but with more groups.

$$\underbrace{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ji}-\bar{Y_{..}})^2}_{\text{total sum of square}}=\underbrace{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ji}-\bar{Y_{j.}})^2}_{\text{residual sum of squares}}+\underbrace{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\bar{Y_{j.}}-\bar{Y_{..}})^2}_{\text{sum of squares between groups}}$$

Thus, the maximal likelihood function is:

- under the null: $\underset{H_0}{sup}L(\mu_1,\mu_2,\mu_3,\dots,\mu_J|Y)=(\frac{1}{\sqrt{2\pi\sigma^2}})^N exp[-\frac{1}{2\sigma^2}(SS_{groups}+SS_{residuals})]$
- under the alternative: $\underset{H_1}{sup}L(\mu_1,\mu_2,\mu_3,\dots,\mu_J|Y)=(\frac{1}{\sqrt{2\pi\sigma^2}})^N exp[-\frac{1}{2\sigma^2}SS_{groups}]$

Substituting back to the test statistic definition:

$$\Lambda=\frac{\underset{H_0}{sup}L(\mu_1,\mu_2,\mu_3,\dots,\mu_J|Y)}{\underset{H_1}{sup}L(\mu_1,\mu_2,\mu_3,\dots,\mu_J|Y)}=exp[-\frac{1}{2\sigma^2}SS_{groups}]$$

We can see that since the exponent is negative, large values of SS_{groups} resulting small values of Λ
Under the null:

$$\bar{Y_{j.}}\overset{ind}{\sim}N(\mu,\sigma^2/n_j)$$

$$\frac{\sqrt{n_j}}{\sigma}(\bar{Y_{j.}}-\mu)\overset{ind}{\sim}N(0,1)$$

When we square it, we get the chi-square with degree of freedom 1 (chi-square distribution)

$$\frac{n_j}{\sigma^2}(\bar{Y_{j.}}-\mu)^2\overset{ind}{\sim}\chi^2_{df=1}$$

Which can be further extends to:

$$\frac{1}{\sigma^2}\sum_{j=1}^{J} n_j(\bar{Y}_{j.}-\mu)^2 \overset{ind}{\sim} \chi^2_{df=J}$$

Well, we do not know the common mean (μ) for now, however our best guess is the grant mean ($\bar{Y}_{..}$):

$$\frac{1}{\sigma^2}\sum_{j=1}^{J} n_j(\bar{Y}_{j.}-\bar{Y}_{..})^2 \overset{ind}{\sim} \chi^2_{df=J-1}$$

Note: when we use the common mean (grant mean), we lost 1 degree of freedom. With knowing sample size N, if you know n_1 up to n_{J-1}, then without counting, we know that the last one is $N - n_{J-1} - n_{J-2} - \ldots - n_1$ so that the total free parameters is number of groups minus 1. With further rearrangements:

$$\sum_{j=1}^{J} n_j(\bar{Y}_{j.}-\bar{Y}_{..})^2 \overset{ind}{\sim} \sigma^2\chi^2_{df=J-1}$$

$$\frac{1}{J-1}\sum_{j=1}^{J} n_j(\bar{Y}_{j.}-\bar{Y}_{..})^2 \overset{ind}{\sim} \sigma^2\frac{\chi^2_{df=J-1}}{J-1}$$

$$\frac{SS_{groups}}{J-1} \overset{ind}{\sim} \sigma^2\frac{\chi^2_{df=J-1}}{J-1} \qquad \textcircled{1}$$

Moreover, we do not know σ^2 as well, recall that we use pooled variance perviously(3.12.2):

$$\begin{aligned} S^2_{pooled} &= \frac{(n_1-1)s_1^2+(n_2-1)s_2^2+\ldots+(n_J-1)s_J^2}{N-J} \\ &= \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ji}-\bar{Y}_{j.})^2}{N-J} \\ &= \frac{SS_{residuals}}{N-J} \end{aligned}$$

Similarly to the mythology above:

$$\frac{SS_{residuals}}{N-J} \overset{ind}{\sim} \sigma^2\frac{\chi^2_{df=N-J}}{N-J} \qquad \textcircled{2}$$

Another important ingredient is $SS_{residuals}$ is independent with SS_{groups} $\textcircled{3}$.

From $\textcircled{1}$, $\textcircled{2}$, $\textcircled{3}$ we can form the F distribution:

$$\frac{\frac{SS_{groups}}{J-1}}{\frac{SS_{residuals}}{N-J}} \sim \frac{\sigma^2\frac{\chi^2_{df=J-1}}{J-1}}{\sigma^2\frac{\chi^2_{df=N-J}}{N-J}} = F_{J-1,N-J}$$

Where:

- J - 1: measures the differences in complexities (number of parameters) between alternative model and null model - we called it as "signal"
- N-J: determines how accurate we can measure the common variance / residual - we call it as "noise"

All the derivations above can be summarised in a ANOVA table:

Source	df	SS	MS (SS / df)	F statistic	Pr(F)
Groups	J - 1	$\sum_{j=1}^{J} n_j(\bar{Y_{j.}} - \bar{Y_{..}})^2$	$\frac{SS_{group}}{J-1}$	$\frac{MS_{groups}}{MS_{residual}}$	$P(\mathcal{F}_{J-1,N-J} \geq F)$
Residual	N - J	$\sum_{j=1}^{J} \sum_{i=1}^{n_j}(Y_{ji} - \bar{Y_{j.}})^2$	$\frac{SS_{residual}}{N-J}$		
Total	N - 1	$\sum_{j=1}^{J} \sum_{i=1}^{n_j}(Y_{ji} - \bar{Y_{..}})^2$			

3.18 Example of ANOVA - movies genres

suppose we have collected some data about movies rating on different genres as following:

Genre	mean (rating)	standard deviation (rating)	sample size
Action	44.97	25.10	31
Comedy	49.78	26.35	27
Horror	41.29	27.04	17
Romance	47.20	31.57	10
Animation	60.58	26.73	12
Thriller	62.38	28.66	13
Total			110

Now we are interested in are there any difference in average between movies rating across genres?

The overall mean is the sum of each sample in each genre / total sample size:

$$\bar{Y_{..}} = \frac{31 \times 44.97 + 27 \times 49.78 + \ldots + 13 \times 62.38}{110} = 49.55$$

The overall variance is:

$$S_{pooled}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_J - 1)s_J^2}{N - J} = \frac{30 \times 25.10^2 + \ldots + 12 \times 28.66^2}{110 - 6} = 26.91^2$$

The sum of square between groups are:

$$SS_{groups} = 31 \times (44.97 - 49.55(\bar{Y_{..}}))^2 + \ldots + 13 \times (62.38 - 49.55)^2 = 5466.63$$

The ANOVA table (with given $SS_{residual}$ is 75337.42) in this question is:

Source	df	SS	MS (SS / df)	F statistic	Pr(F)
Groups	5	5466.63	1093.3	1093.3 / 724.4 = 1.509	$P(\mathcal{F}_{5,104} \geq 1.509) = 0.193$
Residual	104	75337.42	724.4		
Total	109				

where, 1 - pf(1.509, 5, 104) = 0.193

Conclusion: there is no evidence that the movies rating in average are different across the 6 genres.

3.19 Bonferroni's correction

Continuing from the movie rating example, although the conclusion is no evidence against the movie rating means are different, we could actually see that the mean of "Horror" and "Thriller" are quite different.

A quick two-sample T-test is conducted:

Recall that (3.12.2):

$$\bar{X}_{1.} - \bar{X}_{2.} \sim N(\mu_1 - \mu_2, \frac{\sigma^2_{pooled}}{n_1} + \frac{\sigma^2_{pooled}}{n_2})$$

$$T = \frac{\bar{Y}_{Thriller.} - \bar{Y}_{Horror.}}{\sigma_{pooled}\sqrt{\frac{1}{n_{Thriller}} + \frac{1}{n_{Horror}}}} = \frac{62.38 - 41.29}{26.91 \times \sqrt{\frac{1}{13} + \frac{1}{17}}} = 2.13$$

with a degree of freedom of 104, since pooled variance is involved
So, the p-value is:

$$P(T_{df=104} \geq 2.13) = 0.018 \equiv 1 - pt(2.13, 104)$$

Is this a contradiction compared with our result?
No, this result needs to be corrected, since we compare the best and the worst group, they must have quite larger difference than other pairs. However, that does not interpret the overall (across the whole groups), as maybe next year, it would be another two categories are the best and the worst respectively. So how many ways there can be to form one best and one worst, $\binom{6}{2} = 15$

Above is the intuition of Bonferroni correction, we multiply the p-value with the total number of possible combination of the comparisons we could make. So in this case, the p-value after the correction is: $0.018 \times 15 = 0.27$, hence it is consistent with our conclusion before.

In summary, we should only look pairwise comparisons when there is a significance level result. This protects us from overly large type I error.

3.20 Contrast

In statistics, particularly in analysis of variance and linear regression, a contrast is a linear combination of variables (parameters or statistics) whose coefficients add up to zero, allowing comparison of different treatments. Mathematically, a contrast is any linear combination that:

$$\sum_{j=1}^{J} a_j u_j$$

Where, $u = (u_1, u_2, \ldots, u_J)^T$ and $\sum_{j=1}^{J} a_j = 0$

3.20.1 Sum constraint / Sum parametrisation

Under sum constraint, the one way ANOVA can be represent as

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

Where,

- μ is the overall mean that one has one parameter

- a_j is the main effect (difference between group mean and overall mean) due to a particular group j, where $a_j = \mu_j - \mu$ and it has J(total groups) parameters and it stratifies $\sum_{j=1}^{J} a_j = 0$
- ε_{ji} is the error terms

In the sum parametrisation, each a_j is itself a contrast.
Example: when j = 1,

$$\begin{aligned}
\alpha_1 &= \mu_1 - \mu \\
&= \mu_1 - \frac{\mu_1 + \mu_2 + \ldots + \mu_J}{J} \\
&= (1 - \frac{1}{J})\mu_1 - \frac{1}{J}\mu_2 - \ldots - \frac{1}{J}\mu_J \\
&= (\frac{J-1}{J})\mu_1 - \frac{1}{J}\mu_2 - \ldots - \frac{1}{J}\mu_J \\
&= \sum_{j=1} J\alpha_j\mu_j
\end{aligned}$$

Where, $\alpha = (\frac{J-1}{J}, \underbrace{-\frac{1}{J}, \ldots, -\frac{1}{J}}_{\text{J-1 times}})$
So when is the sum parametrisation useful?

- compare yourself to a particular average
- all J groups are variants of the same treatment or J groups exhaust all possible groups in the population

3.20.2 Contrast constraint / Contrast parametrisation (default in R)

Similar with the sum constraint, under contrast constraint, the one way ANOVA can be represent as

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

However the difference is:

- μ is the mean of group 1, where $\mu = \mu_1$
- each a_j represents the difference between group j and goup 1, where $a_j = \mu_j - \mu_1$

Again, each a_j itself is a contrast
For example:

$$\begin{aligned}
a_2 &= \mu_2 - \mu_1 \\
&= -1 \times \mu_1 + 1 \times \mu_2 + 0 \times \mu_3 + \ldots + 0 \times \mu_j \\
&= \sum_{j=1}^{J} a_j u_j
\end{aligned}$$

Where, $a_j = (-1, +1, 0, \ldots, 0)$
When is the contrast constraint useful? - When we have a control/placebo groups

3.20.3 summary

	Mean parametrisation	Sum parametrisation	Contrast parametrisation
Representation	$Y_{ji} = \mu_j + \varepsilon_{ji}$ j = 1, 2, ..., J	$Y_{ji} = \mu + a_j + \varepsilon_{ji}$ j = 1, 2, ..., J	$Y_{ji} = \mu + a_j + \varepsilon_{ji}$ j = 1, 2, ..., J
Constraints		$\sum_{j=1}^{J} a_j = 0$	$a_1 == 0$
MLEs	$\hat{\mu_j} = \bar{Y}_{j.}$	$\hat{\mu_j} = \bar{Y}_{..}$ $\hat{a_j} = \bar{Y}_{j.} - \bar{Y}_{..}$	$\hat{\mu_j} = \bar{Y}_{1.}$ $\hat{a_j} = \bar{Y}_{j.} - \bar{Y}_{1.}$

3.21 Two way ANOVA

Suppose there are two factors at each level. For example, when we testing different drugs, it also has different dosage. In this case, we need to consider two-way interval.

In general, two-way ANOVA is defined as:

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

Where α_j is the main effect of fact A, βj is the main effect of fact B, and γ_{jki} is the interaction effect of fact A&B
How many parameter it has?
It has $1 + J + K + JK$ parameters and the number of groups is JK.

Under the sum constraint, we have: $\sum_{j=1}^{J} \alpha_j = 0$, $\sum_{k=1}^{K} \beta_k = 0$, $\sum_{j=1}^{J} \gamma_{jk} = \sum_{k=1}^{K} \gamma_k = 0$
Whereas, under the contrast constraint, we have: $\alpha_1 = 0$, $\beta_1 = 0$, and $\gamma_{1k} = \gamma_{j1} = 0$

How do we interpret those parameters under the contrast constraint?

- μ is the mean response to the first group - level 1 of A and level 1 of B
- α_j is the expected change in mean when only change the fact A - keep factor B at level 1, then across J for factor A
- β_k is the expected change in mean when changing only the fact B
- γ_{jk} is the additional change in mean, when change factor A & B at the same time

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

3.22 Linear Model

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

$$Y = \underbrace{X\beta}_{structure} + \underbrace{\varepsilon}_{errors}$$

Where, X is $n\times$ design matrix, β is $p \times 1$ vector of regression parameters, and ε is the deviations away from the linear model.

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} X_{11} & X_{12} & X_{13} & \dots & X_{1p} \\ X_{21} & X_{22} & X_{23} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & X_{n3} & \dots & X_{np} \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \quad \text{and} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Example: suppose we have a contrast constraint one-way ANOVA:

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

Where, $j = 1, 2, \dots, J$ with $\alpha_1 = 0$ and $i = 1, 2, \dots, n_j$

Let's construct our model:

$$Y = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \\ \\ Y_{31} \\ Y_{32} \\ \vdots \\ Y_{3n_3} \\ \vdots \\ \\ Y_{J1} \\ Y_{J2} \\ \vdots \\ Y_{Jn_j} \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} \mu \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_J \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} 1 & 0 & 0 \dots 0 \\ 1 & 0 & 0 \dots 0 \\ \vdots \\ 1 & 0 & 0 \dots 0 \\ \\ 1 & 1 & 0 \dots 0 \\ 1 & 1 & 0 \dots 0 \\ \vdots \\ 1 & 1 & 0 \dots 0 \\ \\ 1 & 0 & 1 \dots 0 \\ 1 & 0 & 1 \dots 0 \\ \vdots \\ 1 & 0 & 1 \dots 0 \\ \vdots \\ \\ 1 & 0 & 0 \dots 1 \\ 1 & 0 & 0 \dots 1 \\ \vdots \\ 1 & 0 & 0 \dots 1 \end{pmatrix}$$

So we can indeed write one-way ANOVA in the form of $Y = \mu + \alpha + \varepsilon$.

In the past, we specified our null hypothesis as: $H_0 : \alpha_2 = 0 = \alpha_3 = \ldots = \alpha_J$. However, now with the matrix representation, we could write the H_0 as:

$$H_0 : R\beta = 0$$

Where,

$$\beta = \begin{pmatrix} \mu \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_J \end{pmatrix} \quad \text{and} \quad R = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

Intuitively, the alternative hypothesis would be:

$$H_1 : R\beta \neq 0$$

Shortly, we fit the null model with the restrictions (imposing); whereas we fit the alternative model without the restriction.

Linear model is powerful, since there is a universal test for any residuals place on our model parameters, then we compare the residual sum of square left over. Specifically,

$$\frac{(RSS_{null} - RSS_{alternative})/r}{RSS_{alternative}/df_{alternative}} \sim \mathcal{F}_{r, df_{alternative}}$$

Where, r is the number of restrictions, it also equals to the rank or $df_{null} - df_{alternative}$, and sometime we call $(RSS_{null} - RSS_{alternative})$ as SS_{model}

So where does this universal test come from?
Suppose Y is a $n \times 1$ vector follows a Normal distribution with common variance σ^2. So,

$$Y \sim N(X\beta, \sigma^2 I)$$

Where, I is an identical matrix. β is a $p \times 1$ vector of parameters
The likelihood function for β with given observed data X is:

$$L(\beta|Y,X) = (\frac{1}{\sqrt{2\pi\sigma^2 I}})^n exp\{-\frac{1}{2\sigma^2 I}(Y - X\beta)^T(Y - X\beta)\} \quad \text{since, } X^T X = |X|^2$$

On differentiating the log-likelihood function, we get:

$$\frac{dL}{d\beta} = \frac{1}{\sigma^2 I} X^T(Y - X\beta) = 0 \Rightarrow X^T Y = X^T X\beta$$

So the MLE is:

$$\hat{\beta}_{MLE} = \underbrace{(X^T X)^{-1} X^T}_{\text{hat matrix}} Y$$

Here, we assume that $X^T X$ is invertible and X^T is full rank.

Aside: $\hat{\beta}_{MLE}$ is unbiased.

$$\begin{aligned}\hat{\beta}_{MLE} &= E((X^T X)^{-1} X^T Y) \\ &= (X^T X)^{-1} X^T E(Y) \\ &\text{since, } Y \sim N(X\beta, \sigma^2 I) \\ &= (X^T X)^{-1} X^T X\beta \\ &= \beta\end{aligned}$$

Under the null:

$$\sup_{H_0: R\beta = 0} L(\beta|Y,X) = (\frac{1}{\sqrt{2\pi\sigma^2 I}})^n exp\{-\frac{1}{2\sigma^2 I} RSS_{null}\}$$

Under the alternative:

$$\sup_{\beta} L(\beta|Y,X) = (\frac{1}{\sqrt{2\pi\sigma^2 I}})^n exp\{-\frac{1}{2\sigma^2 I} RSS_{alternative}\}$$

With using GLRT for testing

- $H_0 : R\beta = 0$
- $H_1 : R\beta \neq 0$

$$\Lambda = \frac{\underset{H_0}{sup} L(\beta|Y,X)}{\underset{H_1}{sup} L(\beta|Y,X)} = exp\{-\frac{1}{2\sigma^2 I}(RSS_{null} - RSS_{alternative})\}$$
$$= exp\{-\frac{1}{2\sigma^2 I} RSS_{model}\}$$

So Λ is large if RSS_{model} is small.

Following the same calculation in ANOVA, we find that:

$$RSS_{model} \sim \sigma^2 \chi^2_r$$

Where r is the number of restrictions.
However, if we do not know about σ^2, the best estimation is the sample pooled variance. Recall the section 3.17:

$$RSS_{alternative} \sim \sigma^2 \chi^2_{df_{alternative}}$$

In this case, since the β is $p \times 1$ vector of parameters, the $df_{alternative} = n - p$, on top of that, because of the independence between $RSS_{alternative}$ and SS_{model}:

$$\frac{SS_{model}/r}{RSS_{alternative}/df_{alternative}} \sim \frac{\chi^2_{df_r}/r}{\chi^2_{df_{alternative}}/df_{alternative}} \equiv \mathcal{F}_{r,n-p}$$

Summary: linear model F-test

Model	df	RSS	F	Pr(F)
Model difference	r	$RSS_{null} - RSS_{alternative}$	$\frac{SS_{model}/r}{RSS_{alternative}/n-p}$	$P(\mathcal{F}_{r,n-p}) \geq F$
Alternative	n-p	$RSS_{alternative}$		
Null	n-(p-r)	RSS_{null}		

Where, r is the number of restrictions, p is the number of rows (parameters) in β, and n is the total sample size.

Note that: the null model is actually smaller than the alternative model, since there is more restrictions on null model. You could calculate r by the difference of length of each model.
For example:

- null: $Y = a + bX + \varepsilon$
- alternative: $Y = a + bX + cY + dZ + \varepsilon$

In this case, r is 2.