MS = HYPOTHESIS TESTING

In this chapter, we introduce parametric hypothesis testing. Non-parametric techniques are discussed in the next chapter.

I. SIMPLE HYPOTHESIS TESTING

I.1. The general principle.

× Introductive example: The crop that we wish to grow is known to give the best yield in soils with a pH of 7 (i.e. neutral). The pH of the soil was measured at various locations, giving observations

$\{6.0 \quad 5.7 \quad 6.2 \quad 6.3 \quad 6.5 \quad 6.4 \quad 6.9 \quad 6.6$
$6.8 \quad 6.7 \quad 6.8 \quad 7.1 \quad 6.8 \quad 7.1 \quad 7.1 \quad 7.5$
$7.0\} = \{x_1, \dots, x_n\}, \quad n = 17.$

The sample mean pH of the soil is $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i = 6.676$.
Should we add some chemicals to change the pH of the soil? In other words, is the population mean pH of the soil different from 7?

Thus assuming that our observations are a realization of a random sample $\mathcal{X}_n = \{X_1, \dots, X_n\}$, where the X_i are iid, with mean $\mathbb{E}X$.

↳ We first ask another question: what is the probability of observing a sample mean as small as 6.676, if the population mean is 7?

To answer this question, we need to specify a distribution for the soil pH. Assume that the $X_1, \dots, X_n$ are $\mathcal{N}(\mu, \sigma = 0.5)$. Then $\bar{X} \sim \mathcal{N}(\mu, \frac{0.5}{\sqrt{17}})$, and we find that $P(\bar{X} \leq 6.676) = 0.004$, with $\mu = 7$.

This probability is very small. So...
(i) Either $\mu = 7$, and we have observed something very unlikely, or
(ii) Our assumption about the mean pH is wrong, or
(iii) Our assumption about the model is wrong.

Here, we would most likely conclude that $\mu < 7$, and add some chemicals.

- In simple hypothesis testing, we compare two hypotheses:

↘ The **NULL HYPOTHESIS** (denoted H_0) = the statement whose validity is to be tested. Often the null hypothesis can be expressed in terms of parameters of a model (e.g. $H_0: \mu = 7$).

An example of a simple hypothesis: the parameter of the distribution is specified. As opposed to a composite hypothesis, for which the parameter of the distribution is not completely specified (e.g. $\mu < 7$)

The null hypothesis often expresses an absence of effect, a reference, a "statu quo".

Ex: × amount of savings of customers in a bank is equivalent to the amount of savings of customers of another bank.
× prices of a major sports brand are identical to the prices of its main competitor

The **ALTERNATIVE HYPOTHESIS** (denoted H_1) = it specifies what happens if H_0 is false. H_1 often specifies what we hope, or expect, to be true.

The alternative hypothesis often expresses a difference, a departure from a reference, the presence of an effect.

Ex: ($H_1 : \mu = \mu_1$), with $\mu_1 \neq 0$ for ($H_0 : \mu = 0$.)

Back to our example, we may consider

$H_0 : \mu = 7$ $H_1 : \mu < 7$	or	$H_0 : \mu = 7$ $H_1 : \mu \neq 7$

"two-sided"

if we suspect that the soil is acidic

"one-sided"

if we are unable (or unwilling) to specify the direction in which the mean pH may differ from 7.

× Classical approach to Hypothesis Testing (HT):

- **Rejecting the null**: Rejecting H_0 when it is true is called a type I error. Its probability α is called the SIGNIFICANCE LEVEL. Denoting $H_0 : \theta = \theta_0$, we have

  $$\alpha = \mathbb{P}_{\theta_0}(\text{reject } H_0)$$

  the probability is computed under the null distribution

  (we have not specified yet a rule for rejecting H_0)

- **Failing to reject the null**: Failing to reject H_0 when it is false is called a type II error. Its probability is usually denoted β. The quantity $1-\beta$ is referred to as the POWER of the test, and corresponds to the probability of rejecting H_0 when it is indeed false. To compute β, the

alternative hypothesis must be exactly specified:

e.g. $H_1 : \theta = \theta_1$, so that $\beta = \mathbb{P}_{\theta_1}(\text{fail to reject } H_0)$.

If θ_1 is not specified, β cannot be calculated.

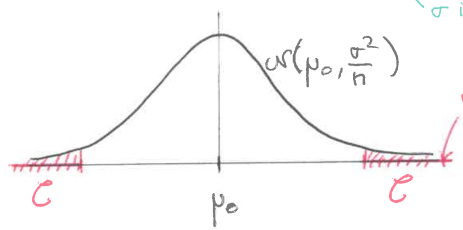
- The decision of a test (reject or not) is often based on determining a critical region $\mathcal{C}$. A statistic $T(X_1, \dots, X_n) \in \mathcal{C}$ is thought to be unlikely to have occured if H_0 is true:

  $$\boxed{\text{reject } H_0 \iff T(X_1, \dots, X_n) \in \mathcal{C}}$$

× Example: Testing for $H_0 : \mu = \mu_0$, under the assumption that $X_1, \dots, X_n$ are $\mathcal{N}(\mu, \sigma^2)$ iid. The sample mean $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = T(X_1, \dots, X_n)$ is $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ distributed.

$\Rightarrow$ Reject H_0 if the observed value $\overline{x}$ of $\overline{X}$ falls in the tails of the $\mathcal{N}(\mu_0, \sigma^2)$ distribution. For example,

σ is assumed to be known.



observing $\overline{x}$ here is highly unlikely if the data $X_1, \dots, X_n$ arised from the $\mathcal{N}(\mu_0, \sigma^2)$ distribution.

$\Rightarrow$ Reject H_0

The significance level α of the test corresponds precisely to the probability that $T(X_1, \dots, X_n) \in \mathcal{C}$. It is usually fixed in advance by the practitioner. [Common values are $\alpha = 0.1$ or $\alpha = 0.05$.]

↳ let $z_{1-\alpha/2}$ be the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution, we

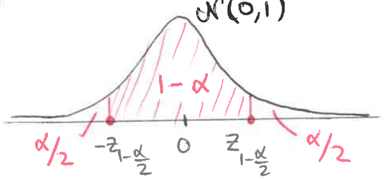
obtain $\mathcal{C} = \left(-\infty,\ \mu_0 - z_{1-\frac{\alpha}{2}}\sigma n^{-1/2}\right] \cup$
$\left[\mu_0 + z_{1-\frac{\alpha}{2}}\sigma n^{-1/2},\ +\infty\right)$,

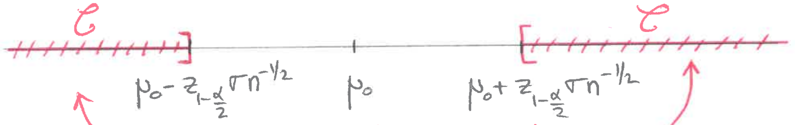
(5)

since

$$\mathbb{P}_{\mu_0}\left(\left| \frac{\bar{X} - \mu_0}{\sigma n^{-1/2}} \right| > z_{1-\frac{\alpha}{2}} \right) = \alpha$$

$Z \sim \mathcal{N}(0,1)$

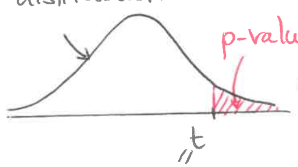


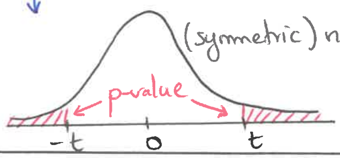


Reject the null hypothesis here

Instead of specifying α, one often reports the p-VALUE = the probability of observing a value of the test statistic $T(X_1, \ldots, X_n)$ as or more extreme than the one actually observed, assuming H_0 is true.

there is some flexibility in what is meant by 'more extreme':
$\mathbb{P}_{H_0}(T \geq t)$ or $\mathbb{P}_{H_0}(T \leq t)$ for a tail event; or
$2 \min\left(\mathbb{P}_{H_0}(T \geq t),\ \mathbb{P}_{H_0}(T \leq t) \right)$ for a double tail event.





H_1 determines what is meant by "more extreme":

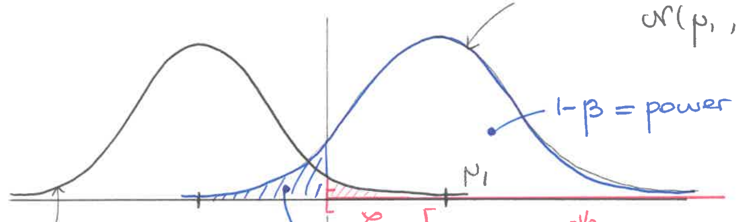
(6)

Tail events correspond to one-sided alternatives, such as $H_1: \mu < \mu_0$ or $H_1: \mu > \mu_0$; while double tail events correspond to two-sided alternatives; $H_1: \mu \neq \mu_0$.

Assuming $H_1: \mu = \mu_1$, we can compute the power $(1-\beta)$ of the test, since

$$\beta = \mathbb{P}_{\mu_1}(\text{fail to reject } H_0)$$
$$= \mathbb{P}_{\mu_1}\left(T(X_1, \ldots, X_n) \in \mathcal{C}\right)$$

↳ With $\mu_1 > \mu_0$,



the threshold is fixed by the critical region constructed from H_0.

The one-sided critical region is obtained from the fact that under H_0,

$$\mathbb{P}_{\mu_0}\left(\frac{\bar{X} - \mu_0}{\sigma n^{-1/2}} > z_{1-\alpha} \right) = \alpha$$

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

<u>In summary</u>, when conducting a HT, one aims at

↘ Minimizing the type I error (aka α)

↘ Maximizing the power (aka $1-\beta$).

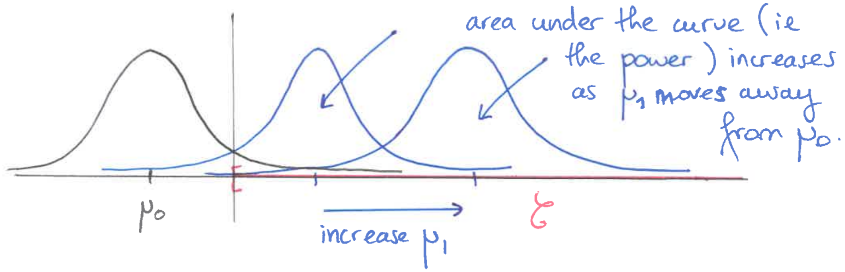
The power of a statistical test is calculated for a particular alternative hypothesis $H_1: \mu = \mu_1$. (7)

$\Rightarrow$ Compute the value of $(1-\beta)$ for a range of simple alternative hypothesis.

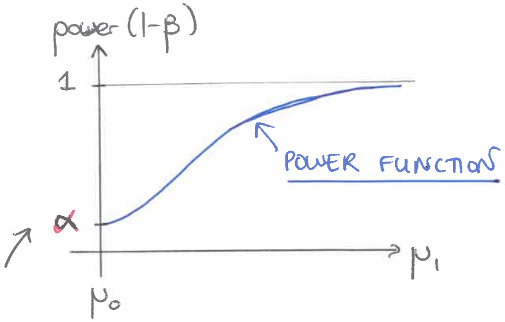
Ex: Compute the value of $(1-\beta)$ under $H_1: \mu = \mu_1$, for all $\mu_1 > \mu_0$.

$\hookrightarrow$ This leads us to the concept of POWER FUNCTIONS.

Note that as μ_1 increases, the power of the test increases as the alternative distribution is shifted to the right.



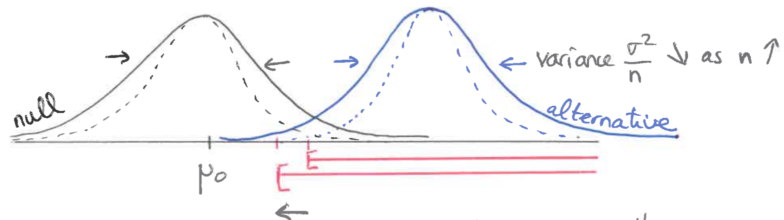
$\Rightarrow$ Plot the value of the power as a function of μ_1, for $\mu_1 \in [\mu_0, +\infty)$:



At $\mu_1 = \mu_0$, the power of the test is equal to α, as the null & alternative distributions coincide

The sample size has an effet on the power of a test: (8)

For a given α, as n increases, the power increases. Indeed, in our previous example, the boundary is shifted to the left as $n \uparrow$, while the variance $\frac{\sigma^2}{n}$ of the normal distributions decreases:



As $n \uparrow$, the end point $\mu_0 + z_{1-\alpha} n^{-1/2} \sigma$ decreases towards μ_0

$\Rightarrow$ One can calculate the minimum sample size required to achive a pre-specified power, for a given alternative hypothesis.

Ex: Looking back at the picture on page 6, the power of the test considered is

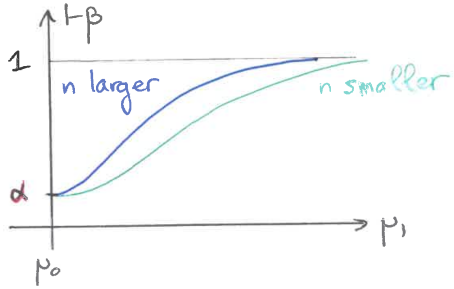
$$1-\beta = \mathbb{P}_{\mu_1}\left(Y \geqslant \mu_0 + z_{1-\alpha} n^{-1/2} \sigma \right), \text{ where } Y \sim \mathcal{N}\left(\mu_1, \frac{\sigma^2}{n}\right)$$

$$= \mathbb{P}_{\mu_1}\left(\underbrace{\frac{Y-\mu_1}{\sigma n^{-1/2}}}_{\sim \mathcal{N}(0,1)} \geqslant \frac{\mu_0 - \mu_1}{\sigma n^{-1/2}} + z_{1-\alpha} \right)$$

Since $\mu_0 - \mu_1 < 0$, as $n \uparrow$, the power of the test tends to 1.

Denoting z_β the β-quantile of the standard normal distribution, we need to solve $\frac{\mu_0-\mu_1}{\sigma n^{-1/2}} + z_{1-\alpha} = z_\beta$

$\Rightarrow$ Choose $n \geqslant \left(\frac{\mu_0-\mu_1}{\sigma(z_\beta - z_{1-\alpha})}\right)^2$ to ensure a power of at least $1-\beta$.



Remark: The test can be conducted without computing its power. However, you have no theoretical guarantee that you are doing something meaningful.

Summary: When designing a HT, you need to

(i) Decide on the null & alternative hypothesis.

(ii) Fix a desired level of type I error (aka α)

(iii) Construct a test statistic $T(X_1,..,X_n)$ from the data $X_1,..,X_n$, and such that the distribution of T is known under H_0.

(iv) Construct a rejection region based on your choices (i)(ii),(iii).

(v) Optional (but recommended): the sample size should be chosen to ensure that the type II error remains small (aka β) / the power $(1-\beta)$ is high.

(vi) Conclude: Reject H_0, or not ($\equiv$ presence of an effect, or not).

Remark: Hypothesis testing was introduced by Ronald Fisher in 1925. His approach however differs from the one presented in this section, in that he did not consider an alternative hypothesis to the null. Instead, given H_0, and a test statistic $T(X_1,..,X_n)$, Fisher suggested calculating the p-value, without the need to fix a desired level of significance α. A result with a low p-value is taken as statistical evidence against the null.

On the top of page 6, we interpreted tail-events and double tail events in terms of a one-sided or two-sided alternative. This association is however superfluous. For Fisher, alternatives to H_0 are implicitely "all what is not H_0". Mathematically, this would translate as $H_1: \mu \neq \mu_0$, if $H_0: \mu = \mu_0$.

Type I errors and type II errors were introduced later by Neyman & Pearson, with the concept of an alternative hypothesis: given a fixed level α, they advocate selecting the test that has the most power.

And indeed, we will see in section I. that to a given problem, you can construct a serie of tests with the same significance level α. The test that should be retained, according to Neyman & Pearson, is the one with maximum power.

I. 2. Further examples.

× Example 1: Testing for the mean of a normal population, when the variance is unknown.

Testing $(H_0 : \mu = \mu_0)$ for $X_1, .., X_n$ iid $\mathcal{N}(\mu, \sigma^2)$, σ unknown.

Then we know from the result on page 25/26 in MS = PARAMETRIC INFERENCE that $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ are independent, and that $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

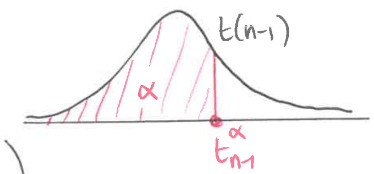
& $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.

It follows that under H_0,

$$T := \frac{(\bar{X}-\mu_0)/\sigma/\sqrt{n}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{n^{1/2}(\bar{X}-\mu_0)}{S} \sim t(n-1).$$

$\Rightarrow$ Consider the test statistic $T(X_1, .., X_n) = \frac{n^{1/2}(\bar{X}-\mu_0)}{S}$, whose distribution under H_0 is $t(n-1)$.

The rejection region C is constructed from the quantiles of the $t(n-1)$ distribution: denoting t_{n-1}^{α} the α-quantile, we obtain



$$C = \left(-\infty, \mu_0 - t_{n-1}^{1-\alpha/2} S n^{-1/2}\right] \cup \left[\mu_0 + t_{n-1}^{1-\alpha/2} S n^{-1/2}, +\infty\right)$$

for a two-sided alternative. The case of a one-sided alternative is treated similarly. ■

× Example 2: Testing for a proportion.

Testing $(H_0 : p = p_0)$ for $X_1, .., X_n$ iid $B(p)$.

The estimator $\hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a consistent estimator of p.

Moreover, under H_0,

$$Z := \frac{\hat{p} - p_0}{\sqrt{np_0(1-p_0)}} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as } n \to +\infty.$$

Therefore, assuming that n and p_0 are such that the normal approximation holds (a rule of thumb: $np_0 \geq 5$, and $n(1-p_0) \geq 5$) we can use Z as our test statistic, and construct a rejection region based on the standard normal distribution. As usual, denoting z_α the α-quantile of $\mathcal{N}(0,1)$, we obtain the (two-sided) critical region

$$C = \left(-\infty, p_0 - z_{1-\frac{\alpha}{2}} n^{-1/2}(p_0(1-p_0))^{1/2}\right] \cup \left[p_0 + z_{1-\frac{\alpha}{2}} n^{-1/2}(p_0(1-p_0))^{1/2}, +\infty\right),$$

with nominal level α.

(and similarly for a one sided region) ■

→ We discuss next common approaches for constructing test statistics:
- Wald tests
- likelihood ratio tests

I.3. The Wald test. (13)

Consider a random sample $X_1, .., X_n$ iid $\sim P_\theta$, for $\theta \in \Theta \in \mathbb{R}^d$.
Test $(H_0 : \theta = \theta_0)$ for some fixed $\theta_0 \in \mathbb{R}^d$.

Let $\hat{\theta}_{ML}$ = maximum likelihood estimate of θ

Under some technical conditions, the MLE is consistent and asymptotically normally distributed, so that we can write:

$$n^{1/2} I_d(\hat{\theta}_{ML})^{1/2}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I_d).$$

$I_d(\hat{\theta}_{ML})$ is the $(d \times d)$ Fisher matrix, evaluated at the MLE.

The multivariate version of the result on pages 15/16 in MS = MAX. LIK. EST.

$$I_d(\theta) = (I_{ij}(\theta)), \text{ with } I_{ij}(\theta) = \mathbb{E}\left\{\frac{\partial \log f(\underline{X};\theta)}{\partial \theta_i} \frac{\partial \log f(\underline{X};\theta)}{\partial \theta_j}\right\}$$

$f(\underline{X};\theta) = \prod_{i=1}^{n} f(X_i;\theta)$ = joint density of $\underline{X} = (X_1, \ldots, X_n)$.

Also, $A^{1/2}$ denotes the square root of the matrix A.

It follows that

$$\underbrace{n(\hat{\theta}_{ML} - \theta_0)^t I_d(\hat{\theta}_{ML})(\hat{\theta}_{ML} - \theta_0)}_{\overset{!!}{T(X_1, .., X_n)} = \text{Wald test statistic}} \xrightarrow{d} \chi^2(d)$$

Test has significance level $\alpha \in (0,1)$ for the rejection region $[q^d_{1-\alpha}, +\infty)$, where q^d_α = α-quantile of the $\chi^2(d)$ distribution.

I.4. Likelihood ratio tests. (14)

Consider a random sample $\underline{X} = (X_1, .., X_n)$, where the X_i are iid and P_θ distributed, for some $\theta \in \Theta$. The probability measure P_θ is assumed to have density $f_\theta(x) = f(x;\theta)$

(likelihood ratio tests are presented in the context of AC RVs, but the results hold true for discrete RVs as well).

Put
- $L(\underline{x};\theta) := \prod_{i=1}^{n} f(x_i;\theta)$
- $\ell(\underline{x};\theta) := \log L(\underline{x};\theta) = \sum_{i=1}^{n} \log f(x_i;\theta)$

$\underline{x} = (x_1, \ldots, x_n)$

Let Θ_0 be a subset of Θ, and consider the MLEs $\hat{\theta}_0$ and $\hat{\theta}$, computed over the parameter spaces Θ_0 and Θ, respectively:

- $\hat{\theta}_0 = \underset{\theta \in \Theta_0}{\operatorname{argmax}} L(\underline{x};\theta)$
- $\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} L(\underline{x};\theta)$.

The ratio $\Lambda(\underline{x}) := \dfrac{L(\underline{x};\hat{\theta}_0)}{L(\underline{x};\hat{\theta})}$ is called the

LIKELIHOOD RATIO STATISTIC.

A few observations:
- $0 \leq \Lambda(\underline{x}) \leq 1$
- If $\hat{\theta}_0$ is far from $\hat{\theta}$, then expect $\Lambda(\underline{x})$ to be small.
- If the true parameter is in Θ_0, then $\Lambda(\underline{x}) \to 1$ as $n \to +\infty$.

⇒ The likelihood ratio statistic may be used as a test statistic to test for $H_0 : \theta \in \Theta_0$, by comparing $\Lambda(x)$ to some threshold c.

It is common to consider as well the quantity

$$\boxed{\begin{aligned}\lambda(x) &:= -2 \log \Lambda(x) \\ &= 2\left(\ell(\underline{x};\hat\theta) - \ell(\underline{x};\hat\theta_0)\right)\end{aligned}}$$

x Example: Let $X_1, \dots, X_n$ iid $\sim B(p)$, for $p \in \Theta = \{p \mid p_0 \leq p \leq 1\}$, for some $p_0 \in (0,1)$.

Consider $(H_0 : p = p_0)$. so that $\Theta_0 = \{p_0\}$

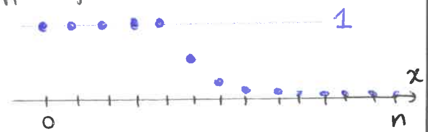
We have :

- $\hat p_0 := \underset{p\in\Theta_0}{\operatorname{argmax}}\; L(\underline{x}; p) = p_0$ since Θ_0 contains a single element
- $\hat p := \underset{p\in\Theta}{\operatorname{argmax}}\; L(\underline{x}; p) = \begin{cases} p_0 & \text{if } \quad \bar x \leq p_0 \\ \bar x := \frac{1}{n}\sum_{i=1}^n x_i & \text{if } \quad \bar x > p_0 \end{cases}$

$$\Rightarrow \Lambda(\underline{x}) = \begin{cases} 1 & \text{if } \frac{x}{n} \leq p_0 \\ \dfrac{p_0^{\,x}(1-p_0)^{x}}{\left(\frac{x}{n}\right)^{x}\left(1-\frac{x}{n}\right)^{n-x}} & \text{if } \frac{x}{n} > p_0 \end{cases}$$

$x := \sum_{i=1}^n x_i$

The LR may have a complicated form...



We see that $\Lambda(\underline{x})$ is a decreasing function of $x = \sum_{i=1}^n x_i$.

⇒ Reject H_0 if $\Lambda(\underline{x})$ is smaller to some c $\iff$ x is larger to some value. Alternatively, we may compute the p-value

$$\begin{aligned} p &:= \mathbb{P}_{p_0}\left(\Lambda(\underline{X}) \leq \Lambda(\underline{x})\right) \\ &= \mathbb{P}_{p_0}\left(X \geq x\right) \end{aligned}$$

where $X = \sum_{i=1}^n X_i$, $x = \sum_{i=1}^n x_i$.

x Example: $X_1, \dots, X_n$ iid $\sim \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \Theta = \mathbb{R}$.
$\sigma^2 > 0$, assumed known.

Consider $(H_0 : \mu = \mu_0)$, for some $\mu_0 \in \mathbb{R}$.
So that $\Theta_0 = \{\mu_0\}$.

Then
- $\hat\mu_0 := \underset{\mu\in\Theta_0}{\operatorname{argmax}}\; L(\underline{x};\mu) = \mu_0$
- $\hat\mu := \underset{\mu\in\Theta}{\operatorname{argmax}}\; L(\underline{x};\mu) = \bar x$

Thus
$$\begin{aligned}\lambda(\underline{x}) &= 2\left(\ell(\underline{x};\hat\mu) - \ell(\underline{x};\hat\mu_0)\right) \\ &= \frac{n}{\sigma^2}(\bar x - \mu_0)^2\end{aligned}$$

Note that $\lambda(\underline{X})$ has a $\chi^2(1)$ distribution, since $\lambda(\underline{X}) = Z^2$; with $Z := \dfrac{\bar X - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.

$z = \dfrac{\bar x - \mu_0}{\sigma/\sqrt{n}}$

Moreover, since $\lambda(\underline{X}) \geq \lambda(\underline{x}) \iff |Z| \geq |z|$, the p-value is

$$\begin{aligned} p &:= \mathbb{P}_{\mu_0}\left(\lambda(\underline{X}) \geq \lambda(\underline{x})\right) \\ &= \mathbb{P}_{\mu_0}\left(|Z| \geq |z|\right) = 2\left(1 - \Phi(|z|)\right) \end{aligned}$$

Theorem: Let θ_0 be an interior point of $\Theta \in \mathbb{R}^d$. Under some technical assumptions (the same as the ones needed to compute the asymptotic properties of the MLE), if the null $(H_0: \theta = \theta_0)$ is true, then the log-likelihood ratio statistic $\lambda(\underline{X}) := -2 \log \Lambda(\underline{X})$ has a limiting χ^2 distribution, with d degrees of freedom.

The limiting $\chi^2(d)$ distribution may be used to decide on a threshold for example.

proof: We have $\lambda(\underline{x}) = 2\left(\ell(\underline{x};\hat{\theta}) - \ell(\underline{x};\theta_0)\right)$.

Expanding $\ell(\underline{x};\theta_0)$ around $\hat{\theta}$, we get

$$\ell(\underline{x};\theta_0) = \ell(\underline{x};\hat{\theta}) + (\theta_0 - \hat{\theta})^t \underbrace{\cancel{\ell'(\underline{x};\hat{\theta})}}_{=0}$$
$$+ \frac{1}{2}(\theta_0 - \hat{\theta})^t \underbrace{\ell''(\underline{x};\tilde{\theta})}_{(d\times d)} (\theta_0 - \hat{\theta}).$$

where $\tilde{\theta}$ is such that $\|\tilde{\theta} - \theta_0\| \leq \|\hat{\theta} - \theta_0\|$

$$\Rightarrow \lambda(\underline{X}) = \sqrt{n}(\theta_0 - \hat{\theta})^t \left\{-\frac{1}{n}\ell''(\underline{X};\tilde{\theta})\right\} \sqrt{n}(\theta_0 - \hat{\theta})^t,$$

$\downarrow$ $\mathcal{N}(0, I_d^{-1}(\theta_0))$

$\downarrow$ $I_d(\theta_0) = \left(I_{ij}(\theta_0)\right)_{k\times k}$ FISHER MATRIX

where

$$I_{jk} = \mathbb{E}\left\{\frac{\partial \log f(X;\theta)}{\partial \theta_i} \frac{\partial \log f(X;\theta)}{\partial \theta_j}\right\}$$

Multivariate version of the asymptotic properties of the MLE established in the previous chapter

and $\lambda(X) \xrightarrow{d} \chi^2(d)$ as required ∎

To test $(H_0: \theta = \theta_0)$, we may consider several test statistics. Consider for example $T(X_1, \dots, X_n)$ and $S(X_1, \dots, X_n)$, both constructed from the same dataset $\underline{X} = (X_1, \dots, X_n)$. There is no guarantee that the two test agree on rejecting or not the null.

$\Rightarrow$ Which test should we consider?

Adopting the paradigm introduced by Neyman & Pearson, we may introduce an alternative simple hypothesis $(H_1: \theta = \theta_1)$ and at a given significance level α, keep the test that has highest power. Neyman & Pearson proved that when testing $(H_0: \theta = \theta_0)$ against $(H_1: \theta = \theta_1)$, the likelihood ratio test is the most powerful. This result is known as the Neyman-Pearson lemma:

NEYMAN-PEARSON LEMMA

Let $X_1, \dots, X_n$ iid with density $f(x;\theta)$, $\theta \in \Theta$.

We test $(H_0: \theta = \theta_0)$ against $(H_1: \theta = \theta_1)$.

The likelihood ratio is $\Lambda(\underline{x}) = \dfrac{f(\underline{x};\theta_1)}{f(\underline{x};\theta_0)}$, and consider

$\phi(\underline{x}) := \mathbb{1}(\Lambda(\underline{x}) \geq k)$, for some $k \geq 0$.

This test is the most powerful test amongst the class of tests with significance level equal to $\alpha := \mathbb{P}_{\theta_0}(\phi(\underline{X}) = 1)$.

In other words, for any other test $\phi'(\underline{X}) \in \{0,1\}$ such that $\mathbb{P}_{\theta_0}(\phi'(X) = 1) \leq \alpha$, we have that

$$\mathbb{P}_{\theta_1}(\phi'(X) = 1) \leq \mathbb{P}_{\theta_1}(\phi(X) = 1)$$

proof: let $C := \{x \mid \phi(x) = 1\} \rightarrow \phi(X) = 1\!\!1(X \in C)$

$C' := \{x \mid \phi'(x) = 1\} \rightarrow \phi'(X) = 1\!\!1(X \in C')$.

Critical regions for each test.

• Take $x \in C$. Then

↘ $\phi'(x) - \phi(x) = 1\!\!1(x \in C') - \underbrace{1\!\!1(x \in C)}_{=1} \leq 0$ —— (*)

↘ $f(x, \theta_1) - k f(x, \theta_0) \geq 0$ —— (**)

by definition of ϕ

Multiplying (*) & (**) together yields

$$(\phi'(x) - \phi(x))(f(x,\theta_1) - k f(x,\theta_0)) \leq 0$$

This expression holds for $x \in C'$ as well, since the signs in (*) & (**) are reversed to ≥ 0 and < 0.

⇒ The expression holds $\forall x$. Integrating it with respect to x gives

$$0 \geq \int (\phi'(x) - \phi(x))(f(x,\theta_1) - k f(x,\theta_0))\,dx$$

$$= \int \phi'(x) f(x,\theta_1)\,dx - \int \phi(x) f(x,\theta_1)\,dx + k\left(\int \phi(x) f(x,\theta_0)\,dx - \int \phi'(x) f(x,\theta_0)\,dx\right)$$

$$= P_{\theta_1}(\phi'(X) = 1) - P_{\theta_1}(\phi(X) = 1) + k\underbrace{\left(\underbrace{P_{\theta_0}(\phi(X)=1)}_{=\alpha} - \underbrace{P_{\theta_0}(\phi'(X)=1)}_{\leq \alpha}\right)}_{\geq 0}$$

$$\geq P_{\theta_1}(\phi'(X) = 1) - P_{\theta_1}(\phi(X) = 1), \text{ as required.} \blacksquare$$

II. TWO-SAMPLE TESTS

II.1. Comparing means.

Let $X_1, \dots, X_{n_1}$ iid $\mathcal{N}(\mu_1, \sigma_1^2)$

$Y_1, \dots, Y_{n_2}$ iid $\mathcal{N}(\mu_2, \sigma_2^2)$ $\qquad \delta := \mu_1 - \mu_2$

We test for equality of the means: $(H_0 : \delta = 0)$.

• σ_1, σ_2 known.

Then $\bar{X} \sim \mathcal{N}(\mu_1, \frac{\sigma_1^2}{n_1})$ and $\bar{Y} \sim \mathcal{N}(\mu_2, \frac{\sigma_2^2}{n_2})$

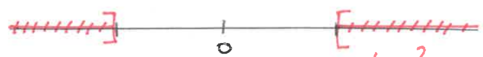
$\Rightarrow \bar{X} - \bar{Y} \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$,

So that $$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1)$$

Under H_0, $\mu_1 = \mu_2$, so $$T(X, Y) = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1)$$

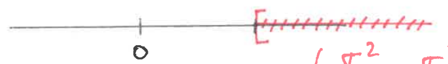
Our test statistic

↳ Critical region is



$-z_{1-\frac{\alpha}{2}}\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{1/2}$ $\qquad$ $z_{1-\frac{\alpha}{2}}\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{1/2}$

For a one-sided test $H_0 : \delta \leq 0$, we get



$z_{1-\alpha}\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{1/2}$

↳ Power of the one-sided test; under $(H_1: \delta = \delta_0 > 0)$

$$1-\beta = \mathbb{P}_{H_1}\left(W \geq z_{1-\alpha} \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)^{1/2} \right),$$

where $W \sim \mathcal{N}\left(\delta_0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$.

$$\Leftrightarrow \beta = \mathbb{P}_{H_1}\left(Z \leq - \frac{\delta_0}{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{1/2}} + z_{1-\alpha} \right),$$

where $Z \sim \mathcal{N}(0,1)$.

Assuming $n_1 = n_2 = n$ (for sample size calculations, we are looking for the minimum value of n to achieve a given power $1-\beta$)

Then $z_\beta = - \dfrac{n^{1/2}\delta_0}{\sqrt{\sigma_1^2+\sigma_2^2}} + z_{1-\alpha}$

Taking $\sigma_1^2 = \sigma_2^2 = \sigma^2 \Rightarrow$ $\boxed{n \geq 2(z_{1-\alpha} - z_\beta)^2 \left(\frac{\sigma}{\delta_0}\right)^2 \quad (*)}$

As we expect, the larger σ, or the smaller δ_0, and the larger n should be.

A function of α, β, σ and δ_0.

- σ_1, σ_2 unknown, assumed equal: $\sigma_1 = \sigma_2 = \sigma$.

Then $\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$

Moreover, $\dfrac{(n_1-1)S_1^2}{\sigma^2} \sim \chi^2(n_1-1)$ and $\dfrac{(n_2-1)S_2^2}{\sigma^2} \sim \chi^2(n_2-1)$

independent

where $S_i^2 = \dfrac{1}{n_i-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

$$\Rightarrow \frac{(n_1-1)S_1^2}{\sigma^2} + \frac{(n_2-1)S_2^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$$

Thus:

our test statistic

$$\frac{\underbrace{\dfrac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}_{\sim \mathcal{N}(0,1)}}{\sqrt{\underbrace{\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2}}_{\sim \chi^2(n_1+n_2-2)} \Big/ (n_1+n_2-2)}} \underset{\text{Under } H_0}{=} \frac{\bar{X}-\bar{Y}}{S_p^2\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$$

indpt

where $S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ is known as the "pooled" estimator of $\sigma^2 \to$ usually, use $(*)$ page 21 for sample size calculations, plugging in S_p^2 for σ^2.

- σ_1, σ_2 unknown.

If we do not assume σ_1 and σ_2 equal, we cannot pool the estimates, and one need to consider estimates of σ_1 and σ_2 separately. We have:

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Inspired by the fact that $\dfrac{(n_i-1)S_i^2}{\sigma_i^2} \sim \chi^2(n_i-1)$,

we are looking for a value of ν such that the quantity

$$U := \frac{\nu\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \text{ is approximately } \chi^2(\nu)$$

Once ν is computed, this will allow us to consider the statistic $T(X_1, .., X_n, Y_1, .., Y_n)$, defined by

$$T(\underline{X},\underline{Y}) = \frac{\dfrac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}}{\sqrt{\dfrac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)\nu}{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\Big/\nu}} \underset{\uparrow \text{Under } H_0}{=} \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \approx t(\nu),$$

whose distribution can be approximated under H_0.

To find ν, we use the method of moments. Let $V \sim \chi^2(\nu)$.
Then $\mathbb{E}V = \nu$, and $\operatorname{Var} V = 2\nu$.

Since $\mathbb{E}S_i^2 = \sigma_i^2$ (unbiased estimator of σ_i), we see that the first moments of U and V coincide.

Next, $\dfrac{(n_i-1)^2 \operatorname{Var} S_i^2}{\sigma_i^4} = 2(n_i-1) =$ Variance of a $\chi^2(n_i-1)$ RV

Thus, $\operatorname{Var}(S_i^2) = \dfrac{2\sigma_i^4}{n_i-1}$, and

$$\operatorname{Var} U = \frac{\nu^2}{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)^2}\left(\frac{\operatorname{Var} S_1^2}{n_1^2}+\frac{\operatorname{Var} S_2^2}{n_2^2}\right)$$

$$= \frac{\nu^2}{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)^2}\left(\frac{2\sigma_1^4}{n_1^2(n_1-1)}+\frac{2\sigma_2^4}{n_2^2(n_2-1)}\right) = \operatorname{Var} V = 2\nu$$

Solving for ν, we obtain

$$\nu = \frac{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)^2}{\frac{\sigma_1^4}{n_1^2(n_1-1)}+\frac{\sigma_2^4}{n_2^2(n_2-1)}}$$

Plug-in S_1^2 and S_2^2 in practice, for an estimate of ν.

the so-called WELCH-SATTERTHWAITE FORMULA

II.2. Comparing two proportions.

Let $X_1, \dots, X_{n_1} \sim B(p_1)$
$Y_1, \dots, Y_{n_2} \sim B(p_2)$. $\qquad \delta = p_1 - p_2$

We test for equality of the two proportions $(H_0: p_1 = p_2)$.

Put $\hat{p}_1 = \dfrac{1}{n_1}\sum_{i=1}^{n_1} X_i$, and $\hat{p}_2 = \dfrac{1}{n_2}\sum_{i=1}^{n_2} Y_i$.

The CLT ensures that $\hat{p}_1 - \hat{p}_2 \approx \mathcal{N}\left(p_1-p_2, \dfrac{p_1(1-p_1)}{n_1}+\dfrac{p_2(1-p_2)}{n_2}\right)$.

Under H_0, $p_1 = p_2 = p$, (equivalently $p = \frac{p_1+p_2}{2}$), and

$$\frac{\hat{p}_1-\hat{p}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \approx \mathcal{N}(0,1).$$

Replace p by a pooled estimate $\hat{p} = \dfrac{\sum(X_i+Y_i)}{n_1+n_2}$,
since X_i and Y_i are iid $B(p)$

Take $T(\underline{X},\underline{Y}) := \dfrac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ as your test statistic,

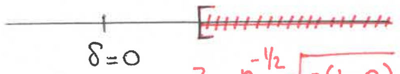
and compare its value with the quantiles of the standard normal distribution.

• <u>Remark</u>: Alternatively, you may consider estimates of p_1 and p_2 separately, and use

$$\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}.$$

• Power & sample size calculations

Assuming equal sample sizes $n_1 = n_2 = n$, the one-sided rejection region is



$z_{1-\alpha} n^{-1/2} \sqrt{p(1-p)}$, $p = \frac{1}{2}(p_1+p_2)$

Under $(H_1 : \delta = \delta_0 > 0)$, the power of the test is

$$1-\beta = \mathbb{P}_{H_1}\left(W \geq z_{1-\alpha} n^{-1/2} \sqrt{p(1-p)} \right),$$

where $W \approx \mathcal{N}\left(\delta_0 , \frac{p_1(1-p_1) + p_2(1-p_2)}{n} \right)$, $p_1 - p_2 = \delta_0 > 0$

$$\Leftrightarrow \quad \beta = \mathbb{P}_{H_1}\left(Z \leq \frac{-\delta_0 n^{1/2}}{\sqrt{p_1(1-p_1)+p_2(1-p_2)}} + z_{1-\alpha} \sqrt{\frac{p(1-p)}{p_1(1-p_1)+p_2(1-p_2)}} \right)$$

where $Z \approx \mathcal{N}(0,1)$

Given β, the minimum value of n achieving the required power must satisfy

$$z_\beta = \frac{-\delta_0 n^{1/2}}{\sqrt{p_1(1-p_1)+p_2(1-p_2)}} + z_{1-\alpha} \sqrt{\frac{p(1-p)}{p_1(1-p_1)+p_2(1-p_2)}} ,$$

so that

$$\boxed{n \geq \frac{1}{\delta_0^2}\left(z_{1-\alpha}\sqrt{p(1-p)} - z_\beta \sqrt{p_1(1-p_1)+p_2(1-p_2)} \right)^2}$$

In practice, plug in the sample estimates $\hat{p}_1, \hat{p}_2$.

where $p_1 - p_2 = \delta_0 > 0$ & $p = \frac{1}{2}(p_1+p_2)$

III - MULTIPLE TESTING

Suppose we want to test multiple null hypothesis $H_{0,1}, \dots, H_{0,d}$. For each $H_{0,i}$ $(i=1,\dots,d)$, we have a p-value p_i associated with the test statistic T_i. It is defined as

$p_i := \mathbb{P}_{0,i}(T_i \geq t_i)$, where t_i is the observed value.

Probability computed under the null $H_{0,i}$.

$= 1 - F_i(t_i)$, denoting $F_i :=$ distribution of T_i under $H_{0,i}$.

Then, under $H_{0,i}$, the random variable $P_i := 1 - F_i(T_i)$ has a $\mathcal{U}(0,1)$ distribution. Indeed,

$$\begin{aligned} \mathbb{P}_{0,i}(P_i \leq p) &= \mathbb{P}_{0,i}(1 - F_i(T_i) \leq p) \\ &= \mathbb{P}_{0,i}(F_i(T_i) \geq 1-p) \\ &= \mathbb{P}_{0,i}(T_i \geq F_i^{-}(1-p)) \\ &= 1 - F_i(F_i^{-}(1-p)) \\ &= 1 - (1-p) \\ &= p . \end{aligned}$$

the generalized inverse of F_i

Now, assume that we decide to reject $H_{0,i}$ if $p_i < \alpha$; so that the probability of a type I error for a single test is α. For d tests, the probability of a type I error associated with the global test $H_0 : \bigcap_{i=1,\dots,d} H_{0,i}$ is $1-(1-\alpha)^d$

Since we reject H_0 as soon as one of the $H_{0,i}$ is rejected.

For $\alpha = 0.05$ and $d = 10$, we get $1-(1-\alpha)^d \approx 0.40$, quite far from the significance level $0.05 \Rightarrow$ Need for correction.

III.1 Bonferroni Correction.

For $H_{0,1}, \dots, H_{0,d}$ = a set of d null hypothesis.

Test for $H_0 := \bigcap_{i=1,\dots,d} H_{0,i}$

Reject each individual test at significance level $\frac{\alpha}{d}$.

$\Rightarrow$ Bonferroni rejects H_0 if $\min_{1 \le i \le d} p_i < \frac{\alpha}{d}$, where

p_i = p-value associated with $H_{0,i}$.

PROCEDURE

The overall significance level is

$$\mathbb{P}\left(\min_{1\le i\le d} p_i < \frac{\alpha}{d}\right) = \mathbb{P}\left(\bigcup_{1\le i\le d} p_i < \frac{\alpha}{d}\right)$$

sub-additivité

Under H_0.

$$\le \sum_{i=1}^{d} \mathbb{P}\left(p_i < \frac{\alpha}{d}\right) \quad (*)$$

Since $p_i \sim \mathcal{U}(0,1)$

$$= \sum_{i=1}^{d} \frac{\alpha}{d} = \alpha$$

The bound may be crude, but if the p-values are independent, then

$$\mathbb{P}\left(\min p_i < \frac{\alpha}{d}\right) = 1 - \mathbb{P}\left(\min p_i \geq \frac{\alpha}{d}\right)$$
$$= 1 - \prod_{i=1}^{d} \mathbb{P}\left(p_i \geq \frac{\alpha}{d}\right)$$
$$= 1 - \left(1 - \frac{\alpha}{d}\right)^d$$

if d is large

$$\approx 1 - e^{-\alpha}$$

if α is small

$$\approx \alpha$$

$\Rightarrow$ For independent tests, the bound $(*)$ is reasonable.

However, Bonferroni correction is known to be too conservative in many practical applications, due to the conservative individual thresholds $\frac{\alpha}{d}$, which can be smaller than needed.

III.2. Fisher's combination test.

Bonferroni is looking at the smallest of the p-values to test for H_0. Alternatively, Fisher suggests to aggregate all the p-values, and considers the quantity

$$X^2 := -2 \sum_{i=1}^{d} \log p_i$$

when the p_i are small, X^2 is large.

$\Rightarrow$ Reject the null $H_0 := \bigcap_{i=1,\dots,d} H_{0,i}$ if X^2 is large. How large?

Under the assumption that $p_1, \dots, p_d$ are independent, $X^2 \sim \chi^2(2d)$. The distribution of X^2 may be used to decide on a threshold.

indeed, if $U \sim \mathcal{U}(0,1)$, then $-2\log U \sim \chi^2(2)$, since

$$\mathbb{P}(-2\log U \le x) = \mathbb{P}\left(U \geq \exp\left(-\frac{x}{2}\right)\right) = 1 - e^{-x/2},$$

with density $\frac{1}{2} e^{-\frac{x}{2}}$.

& the sum of d independent $\chi^2(2)$ RVs is $\chi^2(2d)$.

III.3. Controlling the False Discovery Rate (FDR).

Instead of controlling the intersection $\bigcap H_{0,i}$, we may look at the test separately:

	Not Reject	Reject
$H_{0,i}$ true		FP
$H_{0,i}$ false		TP

FP: total number of false rejections

TP: total # of true rejections.

The Family-Wise Error Rate (FWER) is the probability of making at least one false rejection: $\mathrm{FWER} = \mathbb{P}(FP \geq 1)$. We say that the FWER is controlled at level α if $\mathrm{FWER} \leq \alpha$.

↳ Bonferroni's correction controls the FWER at level α since

$$\mathbb{P}(FP \geq 1) = \mathbb{P}\Big(\bigcup_{i \in I_0} \text{rejecting } H_{0,i}\Big)$$

$$\leq \sum_{i \in I_0} \mathbb{P}(\text{rejecting } H_{0,i})$$

$$= \alpha \frac{d_0}{d} \leq \alpha .$$

I_0 = set of indices in $\{1, .., d\}$ corresponding to the true nulls $H_{0,i}$

Let $d_0 = |I_0|$
$= \#$ elements in I_0

Instead of controlling the FWER, we may wish to control the mean proportion $\left(\frac{FP}{FP+TP}\right)$ of false positives, referred to as the False Discovery Rate (FDR) in the litterature:

$$\mathrm{FDR} = \mathbb{E}\left\{ \frac{FP}{FP+TP} \mathbb{1}(FP+TP \geq 1) \right\} .$$

Goal: Design a procedure to control the FDR, and make sure it remains below a certain level (say α).

The first method we discuss is due to Benjamini & Hochberg (1995) Controlling the False Discovery Rate: A Practical & Powerful Approach to Multiple Testing: JRSS B. Vol 57, No 1, p. 289-300.

Also, see Giraud (2015), Introduction to high dim statistics.

Consider $H_{0,1}, \dots, H_{0,d}$ = d-null hypotheses, with p-values $p_1, \dots, p_d$.
Let $p_{(1)}, \dots, p_{(d)}$ be the ordered p-values, and $H_{0,(i)}$ be the null hypothesis associated with $p_{(i)}$.

Let k be the largest i for which $p_{(i)} \leq \frac{i\alpha}{d}$.

Then, reject all $H_{(i)}$ for $i = 1, -, k$.

PROCEDURE (BENJAMINI & HOCHBERG (1995))

Result: If the $P_1, \dots, P_d$ are independent, then the Benjamini & Hochberg procedure ensures that $\mathrm{FDR} \leq \alpha$.

proof: let $I_0 = \{1 \leq i \leq d \mid H_{0,i} \text{ is true}\}$.
K = (random) number of rejected hypothesis. $(= FP + TP)$

Then

$$\mathrm{FDR} = \mathbb{E}\left\{ \frac{FP}{FP+TP} \mathbb{1}(FP+TP \geq 1) \right\}$$

$$= \mathbb{E}\left\{ \frac{\left|\left\{ i \in I_0 : P_i \leq \frac{\alpha K}{d} \right\}\right|}{K} \mathbb{1}(K \geq 1) \right\}$$

Since all the rejected hypothesis have a p-value smaller than $\frac{\alpha K}{d}$.

$$= \sum_{i \in I_0} \mathbb{E}\left\{ \mathbb{1}\left(P_i \leq \frac{\alpha K}{d}\right) \frac{\mathbb{1}(K \geq 1)}{K} \right\}$$

$$= \sum_{i \in I_0} \mathbb{E}\, \mathbb{E}\Big[\{\dots\} \mid K\Big] ,$$

where $\mathbb{E}\Big[\{\dots\} \mid K = k\Big] = \frac{\mathbb{1}(k \geq 1)}{k} \mathbb{P}\left(P_i \leq \frac{\alpha k}{d} \mid K = k\right)$.

Thus

$$\mathbb{E}\,\mathbb{E}\left[\{\ldots\} \mid K\right] = \sum_{k=1}^{d} \frac{1}{k} \mathbb{P}\left(P_i \leq \frac{\alpha K}{d} \mid K=k\right) \mathbb{P}(K=k),$$

and

$$FDR = \sum_{i\in I_0} \sum_{k=1}^{d} \frac{1}{k} \mathbb{P}\left(K=k,\ P_i \leq \frac{\alpha k}{d}\right)$$

$$= \sum_{i\in I_0} \sum_{k=k_i}^{d} \frac{1}{k} \mathbb{P}\left(K=k \,\Big|\, P_i \leq \frac{\alpha k}{d}\right) \mathbb{P}\left(P_i \leq \frac{\alpha k}{d}\right),$$

where k_i is defined to be the smallest integer ≥ 1 such that $\mathbb{P}\left(P_i \leq \frac{\alpha k_i}{d}\right) > 0$.

Now, we have $P_i \sim \mathcal{U}(0,1)$, so that $\mathbb{P}\left(P_i \leq \frac{\alpha k}{d}\right) = \frac{\alpha k}{d}$, and

$$FDR = \frac{\alpha}{d} \sum_{i\in I_0} \sum_{k=k_i}^{d} \mathbb{P}\left(K=k \,\Big|\, P_i \leq \frac{\alpha k}{d}\right)$$

$$= \frac{\alpha}{d} \sum_{i\in I_0} \sum_{k=k_i}^{d} \left\{ \mathbb{P}\left(K\leq k \,\Big|\, P_i \leq \frac{\alpha k}{d}\right) - \mathbb{P}\left(K\leq k-1 \,\Big|\, P_i \leq \frac{\alpha k}{d}\right)\right\}$$

We claim that $\forall\, k \geq k_i$,

$$\mathbb{P}\left(K\leq k \,\Big|\, P_i \leq \frac{\alpha k}{d}\right) \leq \mathbb{P}\left(K\leq k \,\Big|\, P_i \leq \frac{\alpha (k+1)}{d}\right) \qquad (**)$$

Assuming that $(**)$ holds, the last written sum is a telescoping sum, and

$$FDR \leq \frac{\alpha}{d} \sum_{i\in I_0} \mathbb{1}(k_i \leq d)\, \mathbb{P}\left(K\leq d \,\Big|\, P_i \leq \frac{\alpha(d+1)}{d}\right) \leq \frac{d_0 \alpha}{d} \leq \alpha.$$

It remains to show $(**)$ under the independence assumption of the p-values.

Note that

$$\mathbb{1}(K\leq k) = \mathbb{1}\left(\max\left\{i \mid P_i \leq \frac{\alpha i}{d}\right\} \leq k\right)$$

$$= \text{a function of } P_1, \ldots, P_d$$

$$=: g(P_1, \ldots, P_d).$$

The function

$$g: [0,1]^d \to \{0,1\}$$
$$(p_1,\ldots,p_d) \mapsto g(p_1,\ldots,p_d)$$

is a nondecreasing function of $p_1, \ldots, p_d$. (why?)

To get $(*)$, we are studying the function

$$u \mapsto \mathbb{E}\{\mathbb{1}(K\leq k) \mid P_i \leq u\},$$

and we show that it a non decreasing function of $u\in[0,1]$, provided the $P_1, \ldots, P_d$ are independent. This function can be rewritten

$$u \mapsto \mathbb{E}\{g(P_1,\ldots,P_d) \mid P_i \leq u\}.$$

More generally, a set of distributions satisfying this property $\forall$ positive & bounded g is said to fulfill the Weak Positive Regression Dependency (WPRD) property. As we now show, it holds under the independence assumption, but the WPRD property hold under other assumptions as well, so the result on page 28 is more general than as stated.

We have

$$\mathbb{E}\{g(P_1,\ldots,P_d) \mid P_1 \leq u\}$$ consider wlog $i=1$.

$$= \int_{(x_2,\ldots,x_d)\in[0,1]^{d-1}} \mathbb{E}\{g(P_1,x_2,\ldots,x_d) \mid P_1\leq u\} \times \mathbb{P}(P_2\in dx_2, \ldots, P_d\in dx_d)$$

Under the independence assumption of $P_1, \ldots, P_d$.

$\Rightarrow$ We only need to show that $\forall\, x_2,..,x_d$, the function

$$u \mapsto \mathbb{E}\{\, g(P_1, x_2,..,x_d) \mid P_1 \leq u \,\}$$

is non-decreasing with u.

Since g is nondecreasing, the function $g_1 : x_1 \mapsto g(x_1,..,x_d)$ is also nondecreasing. Thus

$$\mathbb{E}\{ g(P_1, x_2,..,x_d) \mid P_1 \leq u \} = \mathbb{E}\{ \underbrace{g_1(P_1)}_{\text{a non-neg RV}} \mid P_1 \leq u \}$$

$$= \int_0^{+\infty} \mathbb{P}(g_1(P_1) \geq x \mid P_1 \leq u)\, dx$$

$$= \int_0^{+\infty} \mathbb{P}(P_1 \geq g_1^-(x) \mid P_1 \leq u)\, dx$$

generalized inverse of g_1: $g_1^-(x) = \inf_{u\in[0,1]} \{ g_1(u) \geq x \}$

Since $\mathbb{P}(P_1 \geq g_1^-(x) \mid P_1 \leq u) = \left(1 - \dfrac{\mathbb{P}(P_1 < g_1^-(x))}{\mathbb{P}(P_1 \leq u)} \right)_+$ is a nondecreasing function of u for all x, we obtain (*). ■

The Benjamini-Hochberg procedure is a powerful procedure, but theoretical guarantees are obtained under distributional assumptions of the p-values (such as independence). We present next an alternative procedure, introduced by Benjamini & Yekutieli (2001): The Control of the False discovery rate in Multiple Testing under Dependency. Annals of Statistics. Vol 29. p.1165-1188.

The idea is to replace the procedure on page 28 by "Let k be the largest i for which $p_{(i)} \leq \frac{\beta(i)\alpha}{d}$", for some appropriate function $\beta : \{1,..,d\} \to \mathbb{R}_+$, nondecreasing. The Benjamini-Hochberg procedure uses $\beta(i) = i$. The FDR is

$$\mathrm{FDR} = \sum_{i\in I_0} \mathbb{E}\left\{ \mathbb{1}\left(P_i \leq \frac{\alpha\beta(K)}{d} \right) \frac{\mathbb{1}(K\geq 1)}{K} \right\}.$$

See the derivation on page 28.

Noting that on the event $\{K \geq 1\}$, $\frac{1}{K} = \sum_{j\geq 1} \frac{\mathbb{1}(j \geq K)}{j(j+1)}$, we have

$$\mathrm{FDR} = \sum_{i\in I_0} \sum_{j\geq 1} \frac{1}{j(j+1)} \mathbb{E}\left\{ \mathbb{1}\left(P_i \leq \frac{\alpha\beta(K)}{d}\right) \mathbb{1}(j\geq K)\, \mathbb{1}(K\geq 1) \right\}.$$

For $j \geq K$, we have that $\beta(K) \leq \beta(j\wedge d)$, so that

$$\mathbb{E}\left\{ \mathbb{1}\left(P_i \leq \frac{\alpha\beta(K)}{d}\right) \mathbb{1}(j\geq K)\, \mathbb{1}(K\geq 1) \right\} \leq \mathbb{P}\left(P_i \leq \frac{\alpha\beta(j\wedge d)}{d} \right)$$

$$= \frac{\alpha\,\beta(j\wedge d)}{d},$$

and we obtain $\mathrm{FDR} \leq \alpha \frac{d_0}{d} \sum_{j\geq 1} \frac{\beta(j\wedge d)}{j(j+1)}$.

We conclude that as long as $\sum_{j\geq 1} \frac{\beta(j\wedge d)}{j(j+1)} \leq 1$, the FDR of the procedure is less than α. The choice $\beta(i) = i$ yields $\sum_{j\geq 1} \frac{\beta(j\wedge d)}{j(j+1)} = 1 + \frac{1}{2} + \cdots + \frac{1}{d} =: H_d > 1$, and therefore does not guarantee a FDR $\leq \alpha$ (note that Benjamini-Hochberg ensure that FDR $\leq \alpha$ under additional assumptions, such as independence. The bound derived previously is crude, but does not require any distributional assumptions on the p-values $\Rightarrow$ more general, but more conservative).

A popular choice for β ensuring that $\sum_{j \geq 1} \frac{\beta(j \wedge d)}{j(j+1)} \leq 1$ is $\beta(i) = \frac{i}{H_j}$, where $H_i = 1 + \frac{1}{2} + \dots + \frac{1}{i}$ grows as $\log i$.

This procedure is known as the Benjamini-Yekutieli procedure.

> Consider $H_{0,1}, \dots, H_{0,d}$ = d-null hypotheses, with p-values $p_1, \dots, p_d$.
> Let $p_{(1)}, \dots, p_{(d)}$ be the ordered p-values, and $H_{0,(i)}$ the null hypothesis associated with $p_{(i)}$.
> Let k be the largest i for which $p_{(i)} \leq \frac{i\alpha}{d H_d}$.
> Then, reject all $H_{(i)}$ for $i = 1, \dots, k$.

PROCEDURE (BENJAMINI & YEKUTIELI (2001))

The B&Y procedure ensures that FDR $\leq \alpha$.

Conclusion:
- B&H (p.28) more powerful, but theoretical guarantees depend on distributional properties of the p-values
- B&Y (p.33) more conservative, but more general, and theoretical guarantees under dependency.

- Summary: Repeated sampling inflates the type-I error rate. Bonferroni's correction controls the FWER $\mathbb{P}(FP \geq 1)$ at level α by adjusting the individual threshold to $\frac{\alpha}{d}$, where d = number of experiments, since

$$FWER = 1 - \left(1 - \frac{\alpha}{d}\right)^d \approx 1 - e^{-\alpha} \approx \alpha$$

(d large) (α small)

Alternatively, we may look at different quantities, such as the $FDR = \mathbb{E}\left(\frac{FP}{FP+TP} \mathbb{1}(FP+TP \geq 1)\right)$, and control using the Benjamini & Hochberg or the Benjamini & Yekuteli procedure. (Bayesian techniques may be even more appropriate, see MS: BAYESIAN STATISTICS)

- Ex: Run 100 tests with $\alpha = 0.05$

(i) 1 test in 10 is truely effective, $\beta = 0.8$
↘ detect 80% of them; TP = 8
↘ $(100 - 10) \times 0.05 = 4.5 = FP$ } FDR ≈ 0.36

(ii) 1 test in 20 is truely effective, $\beta = 0.8$ [LOW BASE RATE]
↘ TP = 4, FP = 4.75 $\Rightarrow$ FDR ≈ 0.54 ↗

(iii) 1 test in 10 is truely effective, $\beta = 0.3$ [UNDERPOWERED]
↘ TP = 3, FP = 4.75 $\Rightarrow$ FDR ≈ 0.62 ↗

(iv) Tests are all ineffective $\Rightarrow$ FDR = 1
+ "regression to the mean" effect: all tests flagged as improvements will be likely to not reproduce the results in a repeated experiment (cf Kahneman: performance = a bit of talent + a lot of chance).

> ↳ Repeated sampling increase the FWER + low-base Rates & Underpowered tests increase the FDR.

IV. GROUP SEQUENTIAL TESTING

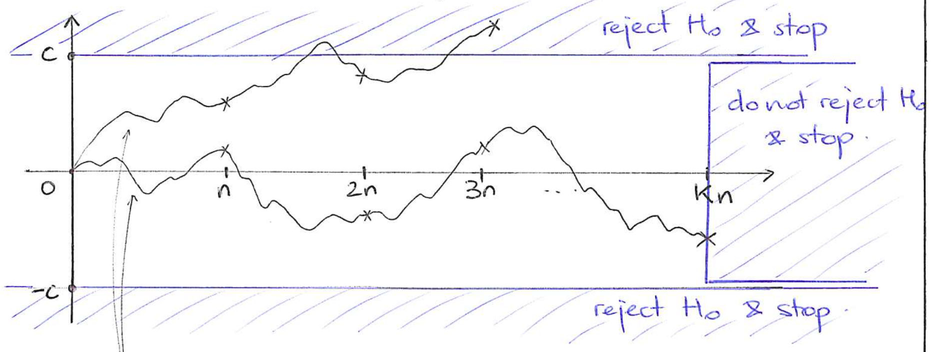
We investigate how the usual testing procedure of Neyman behaves when data are received sequentially, and when an "optimal stopping" of the data collection is performed. The procedure, also known as data peeking, can be described as follows:

(i) Collect n observations $X_1, \dots, X_n$ and compute the test statistic $T(X_1, \dots, X_n)$.
Reject H_0 and stop the experiment if $|T(X_1, \dots, X_n)| > c$.
If not, do not reject and continue.

(ii) Collect another n observations $X_{n+1}, \dots, X_{2n}$ and compute $T(X_1, X_2, \dots, X_n, X_{n+1}, \dots, X_{2n})$.
Reject H_0 & stop if $|T(X_1, \dots, X_{2n})| > c$.
If not, continue.

(iii) .../...

(iv) Until a pre-specified number K of maximum number of iterations is reach, or significance is achieved.



Two possible trajectories. A simple simulation study would show that repeating this many times under the case where H_0 is true would lead to many paths hitting the $\pm c$ boundary before reaching "do not reject H_0 & stop".

In other words, the type I error α is inflated. We artificially detect effects when there is not, at a rate that can well exceed α.
The inflation is know to be not as high as in the case of multiple testing, but can nevertheless easily reach $0.10 / 0.20$ with $\alpha = 0.05$.

To control for the type-I error in sequential testing, the boundaries $\pm c$ need some adjustments. We derive those in the simplest case where observations are iid and normally distribution, leading to Pocock (1977) and O'Brien-Fleming (1979) boundaries. First, we review the procedure with no peeking at the data.

- Case 1: no peeking.

$X_1, \dots, X_{n_0} \sim \mathcal{N}(\mu_X, \sigma^2)$
$Y_1, \dots, Y_{n_0} \sim \mathcal{N}(\mu_Y, \sigma^2)$

Collect n_0 observations in each group, and test for $H_0: \mu_X = \mu_Y$ vs $H_1: \mu_X \neq \mu_Y$.

Assume σ^2 known.

Then $\bar{X} = \frac{1}{n_0} \sum_{i=1}^{n_0} X_i \sim \mathcal{N}(\mu_X, \frac{\sigma^2}{n_0})$, $\bar{Y} \sim \mathcal{N}(\mu_Y, \frac{\sigma^2}{n_0})$

and $\bar{X} - \bar{Y} \sim \mathcal{N}(\mu_X - \mu_Y, \frac{2\sigma^2}{n_0})$.

Put $I_0^{-1}(1) := \frac{2\sigma^2}{n_0}$ = Fisher information.

Then $\sqrt{I_0(1)}\,(\bar{X} - \bar{Y}) \sim \mathcal{N}(\sqrt{I_0(1)}\,\delta_0, 1)$, $\delta_0 := \mu_X - \mu_Y$.

When testing only once, under H_0, $\sqrt{I_0(1)}\,(\bar{X} - \bar{Y}) \sim \mathcal{N}(0,1)$, and $T(1) := \sqrt{I_0(1)}\,(\bar{X} - \bar{Y})$ can be used to construct a rejection region at significance level α: $\mathbb{P}_{H_0}(|T(1)| \geq z_{1-\frac{\alpha}{2}}) = \alpha$.

Under H_1: $\mu_X - \mu_Y = \delta_0 > 0$, the required sample size can be calculated ensuring enough power:

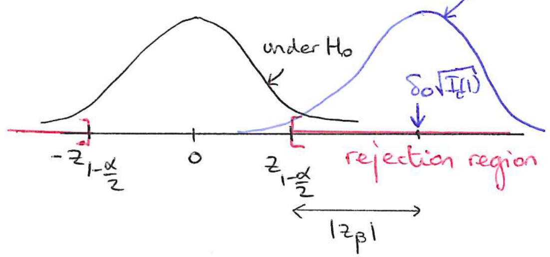
$$\mathbb{P}_{H_1}\left(\sqrt{I(1)}(\bar{X}-\bar{Y}) \geq z_{1-\frac{\alpha}{2}}\right) = 1-\beta \quad \text{, where under}$$

neglecting the left region, which has little mass under H_1

$$H_1,\ \sqrt{I(1)}(\bar{X}-\bar{Y}) \sim \mathcal{N}(\sqrt{I(1)}\,\delta_0, 1)$$

$\Rightarrow$ Take $-z_\beta + z_{1-\frac{\alpha}{2}} = \delta_0\sqrt{I(1)} = \delta_0\sqrt{\dfrac{n_0}{2\sigma^2}}$

$$n_0 \geq 2\sigma^2\left(\frac{z_{1-\beta}+z_{1-\frac{\alpha}{2}}}{\delta_0}\right)^2$$



- Case 2: Optimal stopping.

Suppose now that we allow ourselves to peek at regularly spaced intervals $t_1, .., t_K$, where K is fixed in advance. For simplicity, assume that n observations are collected in each group between two successive times t_k and t_{k+1}, so that at time t_k, each group has collected kn observations.

Reject H_0 at time t_k & stop the experiment if $|T(k)| \geq b(k)$, where $T(k) := \sqrt{I(k)}\,(\bar{X}-\bar{Y}) \sim \mathcal{N}(0,1)$ under H_0, $I^{-1}(k) = \dfrac{2\sigma^2}{kn}$, $k=1,..,K$, for some carefully chosen $b(1), .., b(K)$.

$$\text{Reject } H_0 \Leftrightarrow \bigcup_{k=1}^{K}\{|T(k)| \geq b(k)\}$$

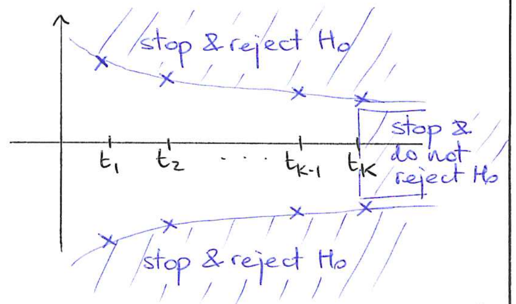
$$\text{Inflated type-I error} = \mathbb{P}_{H_0}\left(\bigcup_{k=1}^{K}\{|T(k)| \geq b(k)\}\right) > 0.05 \quad \text{if } K \geq 2.$$

The goal is to select boundaries $b(1), .., b(K)$ ensuring a

type-I error of α, and a sample size n ensuring a power of $1-\beta$ for all $\delta_0 = \mu_X - \mu_Y$ larger than some value.

$$\mathbb{P}_{H_0}\left(|T(1)| < b(1), \dots, |T(K)| < b(K)\right) = 1-\alpha$$

Reject at time t_k if
$\{|T(1)| < b(1), \dots, |T(k-1)| < b(k-1), |T(k)| \geq b(k)\}$



$\Rightarrow$ We need to derive the joint distribution of the vector $\begin{pmatrix} T(1) \\ \vdots \\ T(K) \end{pmatrix}$.

Recall that $T(k) = \sqrt{I(k)}\,(\bar{X}-\bar{Y})$, where $I(k) = \dfrac{kn}{2\sigma^2}$.

$$\begin{aligned} \text{Put } W(k) &:= \sqrt{I(k)}\,T(k) \\ &= I(k)(\bar{X}-\bar{Y}) \\ &= \frac{\cancel{kn}}{2\sigma^2}\left(\frac{X_1+\dots+X_{kn}}{\cancel{kn}} - \frac{Y_1+\dots+Y_{kn}}{\cancel{kn}}\right) \\ &= \frac{1}{2\sigma^2}\left(X_1+\dots+X_{kn} - Y_1 - \dots - Y_{kn}\right) \end{aligned}$$

$\{W(k)\}_{k=1..K}$ defines an independent increment process since

$$\begin{cases} W(1) = W(1) \\ W(2) = W(1) + [W(2)-W(1)] \\ W(3) = W(1) + [W(2)-W(1)] + [W(3)-W(2)] \\ \dots/\dots \end{cases}$$

independent

$\Rightarrow \operatorname{var} W(k) = I(k)$

(41)

and $\underset{(k<\ell)}{\operatorname{cov}}\big(W(k), W(\ell)\big) = \operatorname{cov}\Big(W(k), [W(\ell)-W(k)] + W(k)\Big)$

$$= \underbrace{\operatorname{cov}\big(W(k), W(\ell)-W(k)\big)}_{=0} + \operatorname{var} W(k)$$

$$= I(k).$$

Thus, since $T(k) = I^{-1/2}(k)\, W(k)$, $\operatorname{var} T(k) = 1$, and

$$\underset{(k<\ell)}{\operatorname{cov}}\big(T(k), T(\ell)\big) = \operatorname{cov}\Big(I^{-1/2}(k)\, W(k),\ I^{-1/2}(\ell)\, W(\ell)\Big)$$

$$= I^{-1/2}(k)\, I^{-1/2}(\ell)\, \underbrace{\operatorname{cov}\big(W(k), W(\ell)\big)}_{=I(k)}$$

$$= \sqrt{\frac{I(k)}{I(\ell)}} = \sqrt{\frac{k}{\ell}}$$

↖ holds as well for $k=\ell$

$\boxed{\text{* Summary: Under } H_0 \text{, the vector } (T(1), \dots, T(K))^t \text{ is multivariate normal with mean vector 0 and covariance matrix } \Sigma = (\Sigma_{k\ell}) \text{, where } \Sigma_{k\ell} = \sqrt{k/\ell},\ k \le \ell.}$

The conditional distributions of a multivariate normal vector with known mean and covariance matrix can be computed using a recursive numerical integration algorithm, see e.g. Armitage, McPherson & Rowe (1969), which can be used to choose $b(1), \dots, b(K)$.

↳ There are ∞-many choices of the $b(1), \dots, b(K)$ that fullfil the requirement

$$\mathbb{P}_{H_0}\big(|T(1)| < b(1), \dots, |T(k)| < b(k)\big) = 1-\alpha .$$

Following Wang & Tsiatis (1987), take the parametric form $b(k) = c\, k^{\gamma - 1/2}$, where $\gamma \in [0, 1/2]$ is referred to as the shape parameter.

(42)

$\gamma = 1/2$ leads to Pocock (1977) boundaries.

$\gamma = 0$ leads to O'Brien-Fleming (1979) boundaries.

We are looking for a value of c such that

$$\mathbb{P}_{H_0}\Big(\bigcap_{k=1}^{K} \{|T(k)| < c\, k^{\gamma - \frac{1}{2}}\}\Big) = 1-\alpha$$

↳ denote the solution $c(\alpha, K, \gamma)$.

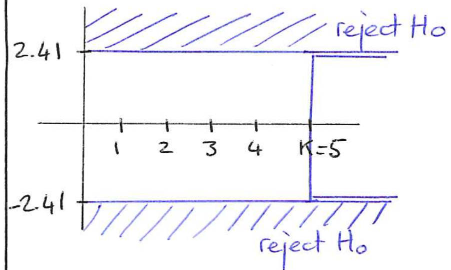
* Ex: take $K=5$, $\alpha = 0.05$.

Then Pocock: $c(\alpha, K, \frac{1}{2}) = 2.4135$

O'Brien-Fleming: $c(\alpha, K, 0) = 4.5618$

Pocock leads to constant boundaries:

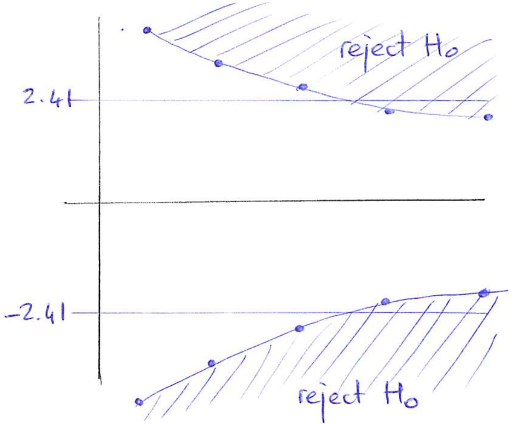
Reject at time t_k if $|T(k)| \geqslant 2.41$.



O'Brien-Flemming rejects H_0 at time t_k if

$$|T(k)| \geqslant \frac{4.5618}{\sqrt{k}}.$$

We see that the Pocock procedure is likely to stop earlier



43

× Remark = generalizations.

The key step in the derivation above is the independence of the increments of the process $\{W(k)\}$. In the case where σ^2 is unknown and must be estimated from the data, this property does not necessarily hold. Another problematic example is when testing for the equality of two binomial proportions $p_X = p_Y$, where the variance σ^2 of the difference of the sample means depends on the unknown parameters themselves. However, although the process does not have independent increments, independence holds asymptotically, as noted in Scharfstein, Tsiatis & Robins (1997).

"Any efficient based test or estimator for [the difference $\mu_X - \mu_Y$], when computed sequentially over time, has, asymptotically, a normal independent increment process whose distribution depends only on the [difference $\mu_X - \mu_Y$] and the statistical information"

$I(1), \ldots, I(K)$.

asymptotic: as $n \to \infty$ in between to times t_k and t_{k+1}.

— The asymptotic covariance $\mathrm{cov}(T(k), T(\ell)) = \sqrt{\dfrac{I(k)}{I(\ell)}}$ —

• Power & sample size.

Under H_1 $T(k) = \sqrt{I(k)}\,(\bar X - \bar Y) \sim \mathcal{N}(\sqrt{I(k)}\,\delta_0, 1)$,

where $I(k) = \dfrac{kn}{2\sigma^2}$. The power under $H_1: \mu_X - \mu_Y = \delta_0 > 0$ is

$$1 - \mathbb{P}_{H_1}\Big(\underbrace{|T(1)| < b(1), \ldots, |T(K)| < b(K)}_{\text{fail to reject } H_0} \Big)$$

44

Information at time t_K is $I(K) = \dfrac{Kn}{2\sigma^2}$.

Put $I(k) = \dfrac{kn}{2\sigma^2} = \dfrac{k}{K} I(K)$.

$\Rightarrow$ Under H_1, $T(k) \sim \mathcal{N}\left(\delta_0 \sqrt{\frac{k}{K} I(K)}, 1 \right)$,

so that $(T(1), \ldots, T(K))^t$ is multivariate normal, with mean vector $\left(\delta_0 \sqrt{\frac{1}{K} I(K)}, \delta_0 \sqrt{\frac{2}{K} I(K)}, \ldots, \delta_0 \sqrt{I(K)} \right)^t$,

and covariance matrix $\Sigma = (\Sigma_{k\ell})$; $\Sigma_{k\ell} = \sqrt{\frac{k}{\ell}}$, $k \le \ell$.

$$\boxed{\begin{array}{l}\text{Power is} \\ 1-\beta = 1 - \mathbb{P}\left(\bigcap_{k=1}^{K} \{ |T(k)| < c(\alpha, K, \gamma)\, k^{\gamma - \frac{1}{2}} \} \right), \\ \text{where } \mathbb{P} = \text{distribution of } (T(1), \ldots, T(K))^t \text{ given above.} \end{array}}$$

For fixed α, K and γ, the power is an increasing function of δ_0. It can be computed numerically using recursive integration.

↳ sample size calculations.

The mean vector can be rewritten

$$\delta_0 \sqrt{I(K)} \left(\sqrt{\tfrac{1}{K}}, \sqrt{\tfrac{2}{K}}, \ldots, 1 \right)^t,$$

where $I(K) = \dfrac{Kn}{2\sigma^2}$ depends on the sample size n between two times t_k and t_{k+1}.

↳ The power is an increasing function of $\delta_0 \sqrt{I(K)}$, and we can solve for $\delta_0 \sqrt{I(K)}$ that gives power $1-\beta$.

Denote this solution $\delta(\alpha, \beta, K, \gamma)$.

$\delta_0 \sqrt{I(K)}$ plays the role of the non-centrality parameter as in the case of no peeking (page 39, it is given by $\delta_0 \sqrt{I_0(1)}$)

no peeking solves for

$$\delta_0 \sqrt{I_0(1)} = z_{1-\beta} + z_{1-\frac{\alpha}{2}}$$

sequential testing solves for

$$\delta_0 \sqrt{I(K)} = \delta(\alpha, \beta, K, \gamma)$$

Take $K=1$, $\gamma = 1/2$,
and $\delta(\alpha, \beta, 1, 1/2) = z_{1-\beta} + z_{1-\frac{\alpha}{2}}$

$$\delta_0^2 \underbrace{\boxed{\frac{n_0}{2\sigma^2}}}_{I_0(1)} = (z_{1-\beta} + z_{1-\frac{\alpha}{2}})^2 \qquad \delta_0^2 \underbrace{\boxed{\frac{nK}{2\sigma^2}}}_{I(K)} = \delta^2(\alpha, \beta, K, \gamma)$$

$$I_0(1) = \frac{(z_{1-\beta} + z_{1-\frac{\alpha}{2}})^2}{\delta_0^2} \qquad I(K) = \frac{\delta^2(\alpha, \beta, K, \gamma)}{\delta_0^2}$$

$$\Rightarrow \boxed{I(K) = I_0(1) \times \left(\frac{\delta(\alpha, \beta, K, \gamma)}{z_{1-\beta} + z_{1-\alpha/2}} \right)^2}$$

Information after K successive peaks at the data = Information required if testing only once × Inflation Factor (IF)

Information after K successive peaks at the data ⇓ nK = total number of obs in each group.

To derive this value (assuming power $1-\beta$), multiply n_0 from a single analysis by the IF to get the total sample size nK.

Inflation Factor (IF) ‖ The relative increase of information necessary for a group-sequential test to have the same power as a single fixed sample test. It depends on α, β, and on the group sequential design parameters K, γ.

Ex: For $\alpha = 0.05$, $1-\beta = 0.8$, $K=5$,
IF = 1.23 (Pocock) and IF = 1.03 (O'Brien)

For $\alpha = 0.05$, $1-\beta = 0.9$, $K=5$,
IF = 1.21 (Pocock) and IF = 1.03 (O'Brien)

In practice, compute the number n_0 of observations required to achieve some power using a usual fixed sample design, and multiply this number by IF to have the same power in a group-sequential test. Interim analyses would be conducted after every $\frac{n_0}{K}$ IF observations in each group.

Which of Pocock or O'Brien should be used?

IF indicate that Pocock require larger sample sizes ($\equiv$ larger information) than O'Brien to do the sequential test, at the same significance level and power.
But Pocock tests have a better chance to stop early because of the shape of the boundary.

Let's formalize this.

We compute the average information required to stop a sequential test when H_1 is true, and the true effect is $\delta_0 > 0$. To ensure a given power of $1-\beta$, recall that

$$I(K) = \frac{\delta^2(\alpha, \beta, K, \gamma)}{\delta_0^2}, \quad \leftarrow I(K) = I_0(1) \times IF$$

while the fixed sample design is such that $I_0(1) = \frac{(z_{1-\beta} + z_{1-\frac{\alpha}{2}})^2}{\delta_0^2}$.

Let V = # interim analyses before a study is stopped.

⇒ The average information before a study is stopped is then

$$\frac{I(K)}{K} \mathbb{E} V = I_0(1) \frac{IF}{K} \mathbb{E} V$$

computed numerically under $H_1: \delta_0 > 0$

For $K=5$, $\alpha=0.05$ and $1-\beta=0.9$, we obtain

$\mathbb{E}V = 2.83$ (Pocock)

$\mathbb{E}V = 3.65$ (O'Brien),

and the average information is $0.68\ I_0(1)$ (Pocock)

$0.75\ I_0(1)$ (O'Brien)

Compare with the maximum info $1.21\ I_0(1)$ (Pocock)

$1.03\ I_0(1)$ (O'Brien)

$\Downarrow$

↘ If we want a design which stops on average with less information when there truly is a treatment difference, then Pocock is preferable over O'Brien.

↘ However, a design with better stopping properties under H_1 need greater maximum information & thus more observations will be needed if H_0 is true.

[REF] B. Tsiatis Lecture Notes
ST 520, Statistical Principles of Clinical Trials (2017).