

Statistics for Data Science: Week 6

Myrto Limnios and Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch, myrto.limnios@epfl.ch



Consider the simplest situation:

$$\Theta_0 = \{\theta_0\} \quad \& \quad \Theta_1 = \{\theta_1\}$$

The Neyman-Pearson Lemma - Continuous Case

Let \mathbf{Y} have joint density/frequency $f \in \{f_0, f_1\}$ and suppose we wish to test

$$H_0 : f = f_0 \quad vs \quad H_1 : f = f_1.$$

If $\Lambda(\mathbf{Y}) = f_1(\mathbf{Y})/f_0(\mathbf{Y})$ is a continuous random variable, then there exists a $k > 0$ such that

$$\mathbb{P}_0[\Lambda(\mathbf{Y}) \geq k] = \alpha$$

and the test whose test function is given by

$$\delta(\mathbf{Y}) = \mathbf{1}\{\Lambda(\mathbf{Y}) \geq k\},$$

is a *most powerful (MP)* test of H_0 versus H_1 at significance level α .

Proof.

Use obvious notation $\mathbb{E}_0, \mathbb{E}_1, \mathbb{P}_0, \mathbb{P}_1$ corresponding to H_0 or H_1 . Let $G_0(t) = \mathbb{P}_0[\Lambda \leq t]$. By assumption, G_0 is a differentiable distribution function, and so is onto $[0, 1]$. Consequently, the set $\mathcal{K}_{1-\alpha} = \{t : G_0(t) = 1 - \alpha\}$ is non-empty for any $\alpha \in (0, 1)$. Setting $k = \inf\{t \in \mathcal{K}_{1-\alpha}\}$ we will have $\mathbb{P}_0[\Lambda \geq k] = \alpha$ and k is simply the $1 - \alpha$ quantile of the distribution G_0 . Consequently,

$$\mathbb{P}_0[\delta = 1] = \alpha \quad (\text{since } \mathbb{P}_0[\delta = 1] = \mathbb{P}_0[\Lambda \geq k])$$

and therefore $\delta \in \mathcal{D}(\{\theta_0\}, \alpha)$ (i.e. δ indeed respects the level α).

To show that δ is also most powerful, it suffices to prove that if ψ is any function with $\psi(\mathbf{y}) \in \{0, 1\}$, then

$$\mathbb{E}_0[\psi(\mathbf{Y})] \leq \underbrace{\mathbb{E}_0[\delta(\mathbf{Y})]}_{=\alpha(\text{by first part of proof})} \stackrel{?}{\Rightarrow} \underbrace{\mathbb{E}_1[\psi(\mathbf{Y})]}_{\beta_1(\psi)} \leq \underbrace{\mathbb{E}_1[\delta(\mathbf{Y})]}_{\beta_1(\delta)}.$$

(recall that $\beta_1(\delta) = 1 - \mathbb{P}_1[\delta = 0] = \mathbb{P}_1[\delta = 1] = \mathbb{E}_1[\delta]$).

$$\frac{f_1}{f_0} > k \Leftrightarrow \Lambda > k$$

WLOG assume that f_0 and f_1 are density functions. Note that

$$\underbrace{f_1(\mathbf{y}) - k \cdot f_0(\mathbf{y}) \geq 0}_{\text{if } \delta(\mathbf{y}) = 1} \quad \& \quad f_1(\mathbf{y}) - k \cdot f_0(\mathbf{y}) < 0 \text{ if } \delta(\mathbf{y}) = 0.$$

Therefore, since ψ can only take the values 0 or 1,

$$\begin{aligned} \underbrace{\psi(\mathbf{y})(f_1(\mathbf{y}) - k \cdot f_0(\mathbf{y}))}_{\text{if } \delta(\mathbf{y}) = 1} &\leq \underbrace{\delta(\mathbf{y})(f_1(\mathbf{y}) - k \cdot f_0(\mathbf{y}))}_{\text{if } \delta(\mathbf{y}) = 0} \\ \int_{\mathbb{R}^n} \psi(\mathbf{y})(f_1(\mathbf{y}) - k \cdot f_0(\mathbf{y})) d\mathbf{y} &\leq \int_{\mathbb{R}^n} \delta(\mathbf{y})(f_1(\mathbf{y}) - k \cdot f_0(\mathbf{y})) d\mathbf{y} \end{aligned}$$

Rearranging the terms yields

$$\begin{aligned} \int_{\mathbb{R}^n} (\psi(\mathbf{y}) - \delta(\mathbf{y})) f_1(\mathbf{y}) d\mathbf{y} &\leq k \int_{\mathbb{R}^n} (\psi(\mathbf{y}) - \delta(\mathbf{y})) f_0(\mathbf{y}) d\mathbf{y} \\ \Rightarrow \mathbb{E}_1[\psi(\mathbf{Y})] - \underbrace{\mathbb{E}_1[\delta(\mathbf{Y})]}_{\text{if } \delta(\mathbf{y}) = 1} &\leq k(\mathbb{E}_0[\psi(\mathbf{y})] - \underbrace{\mathbb{E}_0[\delta(\mathbf{Y})]}_{\text{if } \delta(\mathbf{y}) = 0}) \leq 0 \end{aligned}$$

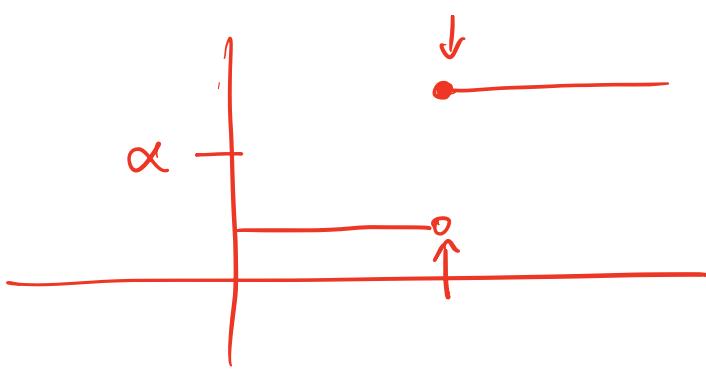
But $k > 0$ by assumption, so when $\mathbb{E}_0[\psi(\mathbf{Y})] \leq \mathbb{E}_0[\delta(\mathbf{Y})]$ the RHS is negative, i.e. δ is an MP test of H_0 vs H_1 at level α . □

The Neyman-Pearson Lemma



$$\Lambda = \frac{f_1}{f_0} > k \Leftrightarrow f_1 > k f_0$$

- Basically we reject if the likelihood of θ_0 is k times higher than the likelihood of θ_1 . This is called a likelihood ratio test, and Λ is the likelihood ratio statistic: *how much more plausible is the alternative than the null?*
- When Λ is a continuous RV, the choice of k is essentially unique. That is, if k' is such that $\delta' = \mathbf{1}\{\Lambda \geq k'\} \in \mathcal{D}(\{\theta_0\}, \alpha)$, then $\delta = \delta'$ almost surely.
- The resulting most powerful test is not necessarily unique.
- Unless Λ is continuous, the most powerful test is not necessarily guaranteed to exist.
type I err $\leq \alpha$
- The problem if Λ is a RV with a discontinuous dist is that there may exist no k for which the equation $\mathbb{P}_0[\Lambda \geq k] = \alpha$ has a solution.
- In any case, typically the distribution of the test statistic converges to a continuous limit with large n , so these problems become inessential.



Example (Poisson Distribution)

Let $Y_1, \dots, Y_n \stackrel{iid}{\sim} \text{Poisson}(\mu)$ and for $\mu_1 > \mu_0$ consider the hypotheses:

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu = \mu_1.$$

These correspond to the Higgs example by setting $\mu_0 = b$ and $\mu_1 = b + s$.

Applying the Neyman-Pearson lemma gives a test statistic

$$\delta(Y_1, \dots, Y_n) = \mathbf{1} \left\{ \sum_{i=1}^n Y_i > q_{1-\alpha} \right\},$$

$$\Lambda = \frac{f_1}{f_0}$$

provided α is such that $\underline{G}_0(q_{1-\alpha}) = \mathbb{P}_{\mu_0}[\tau(Y_1, \dots, Y_n) \leq q_{1-\alpha}] = 1 - \alpha$.

Since the Y_i are independent, one can easily show that

$$\tau(Y_1, \dots, Y_n) \stackrel{H_0}{\sim} \text{Poisson}(n\mu_0).$$



This being a discrete distribution, the only α for which we get an MP test are

$$e^{-n\mu_0}, e^{-n\mu_0}(1 + n\mu_0), e^{-n\mu_0}\left(1 + n\mu_0 + \frac{(n\mu_0)^2}{2}\right), \dots \text{and so on}$$

Nevertheless notice that as $n \rightarrow \infty$, these values become dense near the origin.

When $\{\Theta_0, \Theta_1\}$ are not singletons, choosing a **most powerful test** is a **much stronger requirement**:

- 1 It should respect the level for all $\theta \in \Theta_0$, i.e.

$$\delta \in \mathcal{D}(\Theta_0, \alpha) = \{\delta : \mathcal{Y}^n \rightarrow \{0, 1\} : \mathbb{E}_{\theta}[\delta] \leq \alpha, \forall \theta \in \Theta_0\}$$

- 2 It should be most powerful for all $\theta \in \Theta_1$ (i.e. for all possible simple alternatives),

$$\mathbb{E}_{\theta}[\delta] \geq \mathbb{E}_{\theta}[\delta'] \quad \forall \theta \in \Theta_1 \quad \& \quad \delta' \in \mathcal{D}(\Theta_0, \alpha)$$

Unfortunately UMP tests rarely exist. **Why?**

→ Consider $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$

- A UMP test must be MP test for any $\theta \neq \theta_1$.
- But the form of the MP test typically differs for $\theta_1 > \theta_0$ and $\theta_1 < \theta_0$!
 - e.g. recall exponential mean example

Example (No UMP test exists)

Let $Y_1, \dots, Y_n \sim \text{Bernoulli}(\theta)$ and suppose we want to test:

$$H_0 : \theta = \theta_0$$

vs

$$H_1 : \theta \neq \theta_0$$

$$\Theta_1 = (0, 1)$$

at some level α . To this aim, consider first

$$H'_0 : \theta = \theta_0$$

vs

$$H'_1 : \theta = \theta_1$$

$$\Theta_1 \setminus \{\theta_0\}$$

Neyman-Pearson lemma gives test statistics

$$T = \frac{f(\mathbf{Y}; \theta_1)}{f(\mathbf{Y}; \theta_0)} = \left(\frac{1 - \theta_1}{1 - \theta_0} \right)^n \left(\frac{\theta_1(1 - \theta_0)}{\theta_0(1 - \theta_1)} \right)^{\sum_{i=1}^n Y_i}$$

$$\theta_0 + \varepsilon$$

$$H_1 : (\theta_0, 1)$$

- If $\theta_1 > \theta_0$ then T increasing in $\sum_{i=1}^n Y_i$
 \hookrightarrow MP test would reject for large values of $\sum_{i=1}^n Y_i$
- If $\theta_1 < \theta_0$ then T decreasing in $\sum_{i=1}^n Y_i$
 \hookrightarrow MP test would reject for small values of $\sum_{i=1}^n Y_i$

$$\textcircled{1} \quad \theta_1 > \theta_0 \quad : \quad \theta_1 = 0.5, \quad \theta_0 = 0.25$$

$$T = \left(\frac{0.5}{0.25} \right)^n \left(\frac{0.5(0.75)}{0.25(0.5)} \right)^{\sum y_i}$$
$$= 2^n 8^{\sum y_i} \quad \text{increasing in } \sum y_i$$

$$\textcircled{2} \quad \theta_1 < \theta_0 \quad : \quad \theta_1 = 0.25, \quad \theta_0 = 0.5$$

$$T = \left(\frac{0.25}{0.5} \right)^n \left(\frac{0.25(0.5)}{0.5(0.75)} \right)^{\sum y_i}$$
$$= \left(\frac{1}{2} \right)^n \left(\frac{1}{3} \right)^{\sum y_i} \quad \text{decreasing in } \sum y_i$$

So what can we do for more general $\{\Theta_0, \Theta_1\}$?

$$H_0: (\theta_0, \theta_1] \leftarrow$$
$$H_1: [\theta_0, 1] \leftarrow$$

So what can we do for more general $\{\Theta_0, \Theta_1\}$?

- **One sided hypotheses:** when Θ_0 is an interval of the form $(-\infty, \theta_0]$ or $[\theta_0, +\infty)$ and $\Theta_1 = \Theta_0^c$, there are often uniformly most powerful tests depending on the underlying model.

So what can we do for more general $\{\Theta_0, \Theta_1\}$?

$$H_0 : [\theta_0, 1)$$
$$H_1 : (0, \theta_0)$$

- **One sided hypotheses:** when Θ_0 is an interval of the form $(-\infty, \theta_0]$ or $[\theta_0, +\infty)$ and $\Theta_1 = \Theta_0^c$, there are often uniformly most powerful tests depending on the underlying model.
 - For example, in one-parameter exponential families, one simply uses the Neyman-Pearson lemma, taking the null to be $\theta = \theta_0$ and the alternative $\theta = \theta_1$ for any $\theta_1 \in \Theta_1$ (the form of the test depends only on the direction of the null and the boundary of the null).

So what can we do for more general $\{\Theta_0, \Theta_1\}$?

- **One sided hypotheses:** when Θ_0 is a an interval of the form $(-\infty, \theta_0]$ or $[\theta_0, +\infty)$ and $\Theta_1 = \Theta_0^c$, there are often uniformly most powerful tests depending on the underlying model.
 - For example, in one-parameter exponential families, one simply uses the Neyman-Pearson lemma, taking the null to be $\theta = \theta_0$ and the alternative $\theta = \theta_1$ for any $\theta_1 \in \Theta_1$ (the form of the test depends only on the direction of the null and the boundary of the null).
 - This generalises to families admitting a so-called “monotone likelihood ratio”

So what can we do for more general $\{\Theta_0, \Theta_1\}$?

$$\begin{array}{c} (\theta, \theta_0] \\ \uparrow \rightarrow (\theta_0 - \varepsilon, \theta_0] \end{array}$$

- **One sided hypotheses:** when Θ_0 is an interval of the form $(-\infty, \theta_0]$ or $[\theta_0, +\infty)$ and $\Theta_1 = \Theta_0^c$, there are often uniformly most powerful tests depending on the underlying model.
 - For example, in one-parameter exponential families, one simply uses the Neyman-Pearson lemma, taking the null to be $\theta = \theta_0$ and the alternative $\theta = \theta_1$ for any $\theta_1 \in \Theta_1$ (the form of the test depends only on the direction of the null and the boundary of the null).
 - This generalises to families admitting a so-called “monotone likelihood ratio”
 - In the absence of the “monotone likelihood ratio” property, one can seek **locally most powerfull tests**, near the hypothesis boundary. It can be shown that the score function (derivative of the logliklihood) at the boundary θ_0 can serve as a test statistic to this aim.

So what can we do for more general $\{\Theta_0, \Theta_1\}$?

- **One sided hypotheses:** when Θ_0 is a an interval of the form $(-\infty, \theta_0]$ or $[\theta_0, +\infty)$ and $\Theta_1 = \Theta_0^c$, there are often uniformly most powerful tests depending on the underlying model.
 - For example, in one-parameter exponential families, one simply uses the Neyman-Pearson lemma, taking the null to be $\theta = \theta_0$ and the alternative $\theta = \theta_1$ for any $\theta_1 \in \Theta_1$ (the form of the test depends only on the direction of the null and the boundary of the null).
 - This generalises to families admitting a so-called “monotone likelihood ratio”
 - In the absence of the “monotone likelihood ratio” property, one can seek **locally most powerfull tests**, near the hypothesis boundary. It can be shown that the score function (derivative of the logliklihood) at the boundary θ_0 can serve as a test statistic to this aim.
- **General hypothesis pairs:** we need to abandon optimality, and search for sensible tests. But the **likelihood ratio** idea can serve us well in this pursuit.

Consider now the multiparameter case $\theta \in \mathbb{R}^p$ with general Θ_0, Θ_1

- As noted optimality breaks down.
- But we can still seek general-purpose approaches.

Consider now the multiparameter case $\theta \in \mathbb{R}^p$ with general Θ_0, Θ_1

- As noted optimality breaks down.
- But we can still seek general-purpose approaches.

The idea: Combine Neyman-Pearson paradigm with Maximum Likelihood

$$\Lambda = \frac{f_1(t)}{f_0(t)} \quad \hat{\theta}_{MLE}$$

Consider now the multiparameter case $\theta \in \mathbb{R}^p$ with general Θ_0, Θ_1

- As noted optimality breaks down.
- But we can still seek general-purpose approaches.

The idea: Combine Neyman-Pearson paradigm with Maximum Likelihood

Definition (Likelihood Ratio)

The *likelihood ratio statistic* corresponding to the pair of hypotheses $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ is defined to be

$$\Lambda(\mathbf{Y}) = \frac{\sup_{\theta \in \Theta} f(\mathbf{Y}; \theta)}{\sup_{\theta \in \Theta_0} f(\mathbf{Y}; \theta)} = \frac{\sup_{\theta \in \Theta} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)}$$

Consider now the multiparameter case $\boldsymbol{\theta} \in \mathbb{R}^p$ with general Θ_0, Θ_1

- As noted optimality breaks down.
- But we can still seek general-purpose approaches.

The idea: Combine Neyman-Pearson paradigm with Maximum Likelihood

Definition (Likelihood Ratio)

The *likelihood ratio statistic* corresponding to the pair of hypotheses $H_0 : \boldsymbol{\theta} \in \Theta_0$ vs $H_1 : \boldsymbol{\theta} \in \Theta_1$ is defined to be

$$\Lambda(\mathbf{Y}) = \frac{\sup_{\boldsymbol{\theta} \in \Theta} f(\mathbf{Y}; \boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta_0} f(\mathbf{Y}; \boldsymbol{\theta})} = \frac{\sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta})}$$

- Intuition: choose the “most favourable” $\boldsymbol{\theta} \in \Theta_0$ (in favour of H_0) and compare it against the “most favourable” $\boldsymbol{\theta} \in \Theta_1$ (in favour of H_1) in a simple vs simple setting (applying NP-lemma)

Consider now the multiparameter case $\theta \in \mathbb{R}^p$ with general Θ_0, Θ_1

- As noted optimality breaks down.
- But we can still seek general-purpose approaches.

The idea: Combine Neyman-Pearson paradigm with Maximum Likelihood

Definition (Likelihood Ratio)

The *likelihood ratio statistic* corresponding to the pair of hypotheses $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ is defined to be

$$\Lambda(\mathbf{Y}) = \frac{\sup_{\theta \in \Theta} f(\mathbf{Y}; \theta)}{\sup_{\theta \in \Theta_0} f(\mathbf{Y}; \theta)} = \frac{\sup_{\theta \in \Theta} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)}$$

- Intuition: choose the “most favourable” $\theta \in \Theta_0$ (in favour of H_0) and compare it against the “most favourable” $\theta \in \Theta_1$ (in favour of H_1) in a simple vs simple setting (applying NP-lemma)
- Typically Θ_0 is a lower dimensional subspace of Θ_1 , so taking sup over Θ (rather than Θ_1) incurs no loss. In this case $\Theta_0 \cap \Theta_1 \neq \emptyset$, but $\text{Leb}(\Theta_0 \cap \Theta_1) = 0$, which suffices.

Example

Let $Y_1, \dots, Y_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Consider:

$$H_0: \mu = \mu_0 \quad \text{vs} \quad H_1: \mu \neq \mu_0$$

$$\bar{Y} \rightarrow \mu$$

$$\Lambda(Y) = \frac{\sup_{(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+} f(Y; \mu, \sigma^2)}{\sup_{(\mu, \sigma^2) \in \{\mu_0\} \times \mathbb{R}^+} f(Y; \mu, \sigma^2)} = \left(\frac{\sum_{i=1}^n (Y_i - \mu_0)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \right)^{\frac{n}{2}} = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} \right)^{\frac{n}{2}}$$

So reject when $\Lambda \geq k$, where k is s.t. $\mathbb{P}_0[\Lambda \geq k] = \alpha$. **Distribution of Λ ?** By monotonicity look only at

$$\begin{aligned} \frac{\sum_{i=1}^n (Y_i - \mu_0)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} &= 1 + \frac{n(\bar{Y} - \mu_0)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 + \frac{1}{n-1} \left(\frac{n(\bar{Y} - \mu_0)^2}{S^2} \right) \\ &= 1 + \frac{T^2}{n-1} \end{aligned}$$

With $S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$ and $T = \sqrt{n}(\bar{Y} - \mu_0)/S \stackrel{H_0}{\sim} t_{n-1}$.

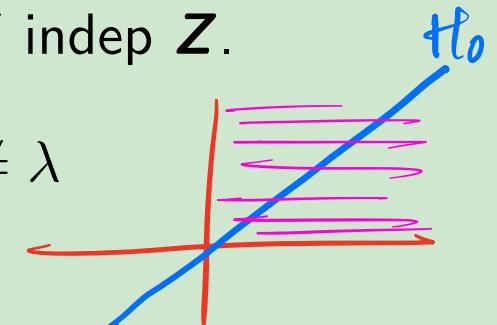
So $T^2 \stackrel{H_0}{\sim} F_{1, n-1}$ and k may be chosen appropriately.

Example

Let $\underbrace{Y_1, \dots, Y_m}_{\text{iid}} \sim \text{Exp}(\lambda)$ and $\underbrace{Z_1, \dots, Z_n}_{\text{iid}} \sim \text{Exp}(\theta)$. Assume \mathbf{Y} indep \mathbf{Z} .

Consider: $H_0 : \theta = \lambda$ vs $H_1 : \theta \neq \lambda$

i.e. $(\theta, \lambda) \in \mathbb{R}_+^2$ against $(\theta, \lambda) \in$ 45 degree line



Unrestricted MLEs: $\hat{\lambda} = \underbrace{1/\bar{Y}}_{\text{sup}_{(\lambda, \theta) \in \mathbb{R}_+^2} f(\mathbf{Y}, \mathbf{Z}; \lambda, \theta)}$ & $\hat{\theta} = \underbrace{1/\bar{Z}}_{= \sum z_i + \sum \ell_i}$

Restricted MLEs: $\hat{\lambda}_0 = \hat{\theta}_0 \stackrel{\downarrow}{=} \left[\frac{m\bar{Y} + n\bar{Z}}{m+n} \right]^{-1}$

$$\Rightarrow \Lambda \stackrel{\downarrow}{=} \left(\frac{m}{m+n} + \frac{n}{n+m} \frac{\bar{Z}}{\bar{Y}} \right)^m \left(\frac{n}{n+m} + \frac{m}{m+n} \frac{\bar{Y}}{\bar{Z}} \right)^n$$

Depends on $T = \bar{Y}/\bar{Z}$ and can make Λ large/small by varying T .

→ But $T \stackrel{H_0}{\sim} F_{2m, 2n}$ so given α we may find the critical value k .

Distribution of Likelihood Ratio?

More often than not, $\text{dist}(\Lambda)$ intractable

→(and no simple dependence on T with tractable distribution either)

Consider asymptotic approximations?

Distribution of Likelihood Ratio?

More often than not, $\text{dist}(\Lambda)$ intractable

→ (and no simple dependence on T with tractable distribution either)

Consider asymptotic approximations?

Setup

- Θ open subset of \mathbb{R}^p
- either $\Theta_0 = \{\theta_0\}$ or Θ_0 open subset of \mathbb{R}^s , where $s < p$
- Concentrate on $\mathbf{Y} = (Y_1, \dots, Y_n)$ has iid components.
- Initially restrict attention to $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$. LR becomes:

$$\Lambda_n(\mathbf{Y}) = \prod_{i=1}^n \frac{f(Y_i; \hat{\theta}_n)}{f(Y_i; \theta_0)}$$

where $\hat{\theta}_n$ is the MLE of θ .

- Impose regularity conditions from MLE asymptotics

Theorem (Wilks' Theorem, case $p = 1$)

Let Y_1, \dots, Y_n be iid random variables with density (frequency) depending on $\theta \in \mathbb{R}$ and satisfying conditions (A1)-(A6), with $\mathcal{I}_1(\theta) = \mathcal{J}_1(\theta)$. If the MLE sequence $\hat{\theta}_n$ is consistent for θ , then the likelihood ratio statistic Λ_n for $H_0 : \theta = \theta_0$ satisfies

from MLE theorem - $2 \log \Lambda_n \xrightarrow{d} V \sim \chi_1^2$

when H_0 is true.

- Obviously, knowing approximate distribution of $2 \log \Lambda_n$ is as good as knowing approximate distribution of Λ_n for the purposes of testing (by monotonicity and rejection method).
- Theorem extends immediately and trivially to the case of general p and for a hypothesis pair $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$.
(i.e. when null hypothesis is simple)

Proof (*).

Under the conditions of the theorem and when H_0 is true,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_1^{-1}(\theta_0))$$

Now take logarithms and expand in a Taylor series around $\hat{\theta}_n$,

$$\begin{aligned} \log \Lambda_n &= \sum_{i=1}^n [\ell(Y_i; \hat{\theta}_n) - \ell(Y_i; \theta_0)] = \sum_{i=1}^n [\ell(Y_i; \hat{\theta}_n) - \ell(Y_i; \hat{\theta}_n)] + \\ &\quad + (\theta_0 - \hat{\theta}_n) \sum_{i=1}^n \ell' (Y_i; \hat{\theta}_n) - \frac{1}{2} (\hat{\theta}_n - \theta_0)^2 \sum_{i=1}^n \ell'' (Y_i; \theta_n^*) \\ &= -\frac{1}{2} n (\hat{\theta}_n - \theta_0)^2 \frac{1}{n} \sum_{i=1}^n \ell'' (Y_i; \theta_n^*) \end{aligned}$$

where θ_n^* lies between $\hat{\theta}_n$ and θ_0 .

If H_0 is true, and since $\hat{\theta}_n$ is a consistent sequence, θ_n^* is sandwiched so

$$\underbrace{\theta_n^* \xrightarrow{p} \theta_0}_{\text{red}}. \quad (\hat{\theta} - \theta) \xrightarrow{p} (n)^{1/2} \sim \mathcal{N}(0, 1)$$

Hence under assumptions (A1)-(A6), and when H_0 is true, a first order Taylor expansion about θ_0 , the continuous mapping theorem and the LLN give

$$\frac{1}{n} \sum_{i=1}^n \ell''(Y_i; \theta_n^*) \xrightarrow{p} -\mathbb{E}_{\theta_0} [\ell''(Y_i; \theta_0)] = \mathcal{I}_1(\theta_0)$$

On the other hand, by the continuous mapping theorem,

$$n(\hat{\theta}_n - \theta_0)^2 \xrightarrow{d} \frac{V}{\mathcal{I}_1(\theta_0)}$$

Applying Slutsky's theorem now yields the result. □

Theorem (Wilk's theorem, general p , general $s \leq p$)

Let Y_1, \dots, Y_n be iid random variables with density (frequency) depending on $\theta \in \mathbb{R}^p$ and satisfying conditions (B1)-(B6), with $\mathcal{I}_1(\theta) = \mathcal{J}_1(\theta)$. If the MLE sequence $\hat{\theta}_n$ is consistent for θ , then the likelihood ratio statistic Λ_n for $H_0 : \{\theta_j = \theta_{j,0}\}_{j=1}^s$ satisfies $2 \log \Lambda_n \xrightarrow{d} V \sim \chi^2_s$ when H_0 is true.

Theorem (Wilks's theorem, general p , general $s \leq p$)

Let Y_1, \dots, Y_n be iid random variables with density (frequency) depending on $\theta \in \mathbb{R}^p$ and satisfying conditions (B1)-(B6), with $\mathcal{I}_1(\theta) = \mathcal{J}_1(\theta)$. If the MLE sequence $\hat{\theta}_n$ is consistent for θ , then the likelihood ratio statistic Λ_n for $H_0 : \{\theta_j = \theta_{j,0}\}_{j=1}^s$ satisfies $2 \log \Lambda_n \xrightarrow{d} V \sim \chi_s^2$ when H_0 is true.

Comments:

- Note that it may potentially be that $s < p$, and this is accommodated by the theorem
- Hypotheses of the form $H_0 : \{g_j(\theta) = a_j\}_{j=1}^s$, for g_j differentiable real functions, can also be handled by Wilks' theorem:
 - Define $(\phi_1, \dots, \phi_p) = g(\theta) = (g_1(\theta), \dots, g_p(\theta))$
 - g_{s+1}, \dots, g_p defined so that $\underline{g(\theta)}$ is 1-1
 - Apply theorem with parameter ϕ

Many other tests possible. For example:

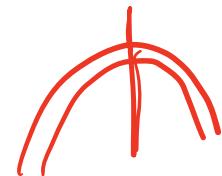
- Wald's test

→ For a simple null, may compare the unrestricted MLE with the MLE under the null. Large deviations indicate evidence against null hypothesis. Distributions are approximated for large n via the asymptotic normality of MLEs.

- Score Test

→ For a simple null, if the null hypothesis is false, then the loglikelihood gradient at the null should not be close to zero, at least when n reasonably large: so measure its deviations from zero. Use asymptotics for distributions (under conditions we end up with a χ^2)

$$\overrightarrow{I(\theta)} = f(\ell')$$



- ...

So far focussed on Neyman-Pearson Framework:

- ➊ Fix a significance level α for the test
- ➋ Consider rules δ respecting this significance level
 - We choose one of those rules, δ^* , based on power considerations
- ➌ We reject at level α if $\delta^*(\mathbf{y}) = 1$.

Useful for attempting to determine optimal test statistics

The infamous p -value (a.k.a. observed significance level)

So far focussed on Neyman-Pearson Framework:

- ➊ Fix a significance level α for the test
- ➋ Consider rules δ respecting this significance level
 - We choose one of those rules, δ^* , based on power considerations
- ➌ We reject at level α if $\delta^*(y) = 1$.

Useful for attempting to determine optimal test statistics

What if we already have a given form of test statistic in mind? (e.g. LRT)

→ A different perspective on testing (used more in practice) says:

Rather then consider a family of test functions respecting level α ...

... consider family of test functions indexed by α



$$P(\Lambda > k) = \alpha$$


So far focussed on Neyman-Pearson Framework:

- ➊ Fix a significance level α for the test
- ➋ Consider rules δ respecting this significance level
 - We choose one of those rules, δ^* , based on power considerations
- ➌ We reject at level α if $\delta^*(\mathbf{y}) = 1$.

Useful for attempting to determine optimal test statistics

What if we already have a given form of test statistic in mind? (e.g. LRT)

→ A different perspective on testing (used more in practice) says:

Rather then consider a family of test functions respecting level α ...

... consider family of test functions indexed by α



- ➊ Fix a family $\{\delta_\alpha\}_{\alpha \in (0,1)}$ of decision rules, with δ_α having level α
 - for a given \mathbf{y} some of these rules reject the null, while others do not
- ➋ Which is the smallest α for which H_0 is rejected given \mathbf{y} ?

$$P(\Lambda > \ell) = \alpha$$

Definition (p -Value)

Let $\{\delta_\alpha\}_{\alpha \in (0,1)}$ be a family of test functions satisfying

$$\underline{\alpha_1} < \underline{\alpha_2} \implies \{\mathbf{y} \in \mathcal{Y}^n : \delta_{\alpha_1}(\mathbf{y}) = 1\} \subseteq \{\mathbf{y} \in \mathcal{Y}^n : \delta_{\alpha_2}(\mathbf{y}) = 1\}.$$

The p -value (or observed significance level) of the family $\{\delta_\alpha\}$ is

$$p(\mathbf{y}) = \inf\{\alpha \downarrow : \delta_\alpha(\mathbf{y}) = 1\}$$

→ The p -value is the smallest value of α for which the null would be rejected at level α , given $\mathbf{Y} = \mathbf{y}$.

Definition (p -Value)

Let $\{\delta_\alpha\}_{\alpha \in (0,1)}$ be a family of test functions satisfying

$$\alpha_1 < \alpha_2 \implies \{\mathbf{y} \in \mathcal{Y}^n : \delta_{\alpha_1}(\mathbf{y}) = 1\} \subseteq \{\mathbf{y} \in \mathcal{Y}^n : \delta_{\alpha_2}(\mathbf{y}) = 1\}.$$

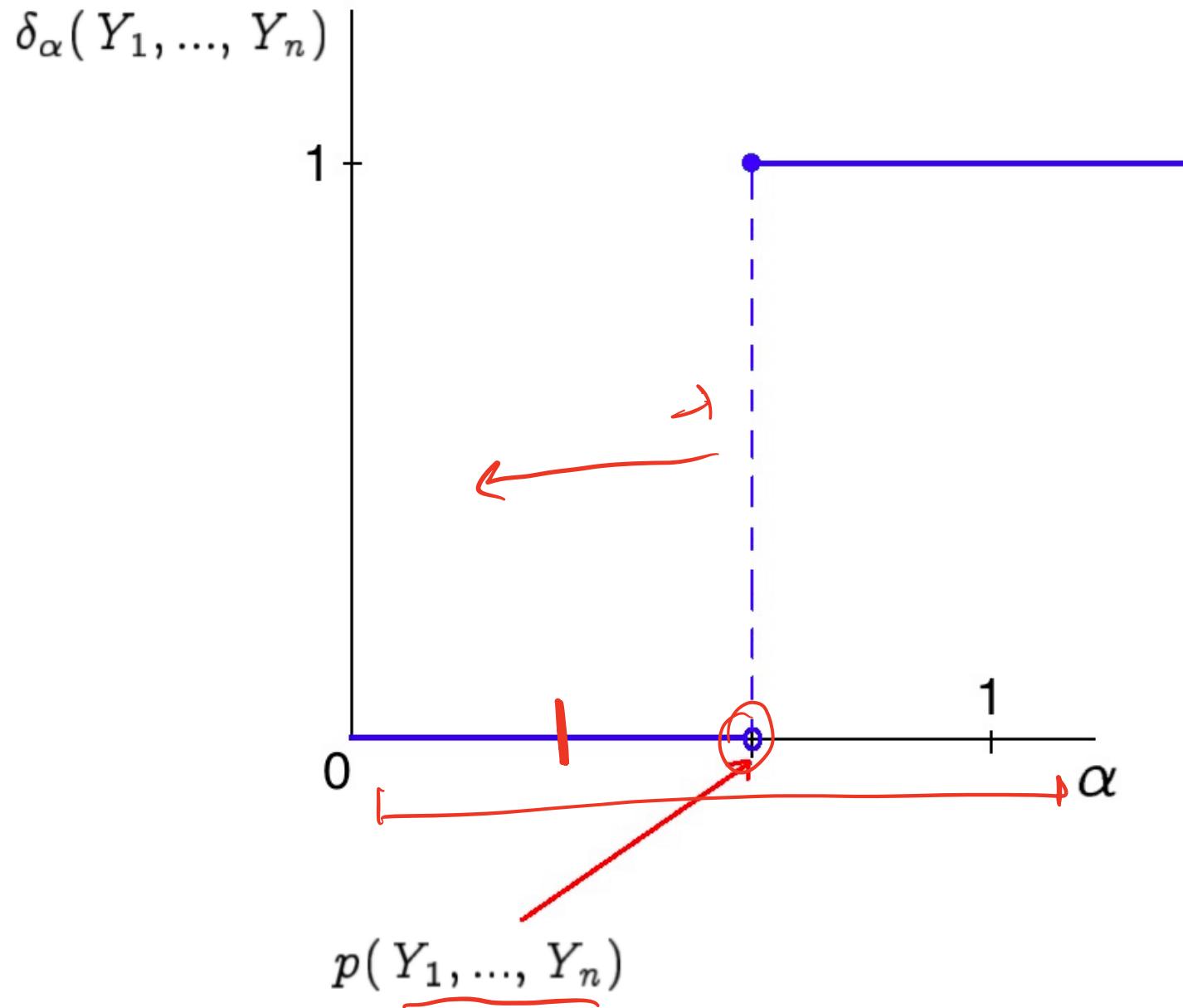
The p -value (or observed significance level) of the family $\{\delta_\alpha\}$ is

$$p(\mathbf{y}) = \inf\{\alpha : \delta_\alpha(\mathbf{y}) = 1\}$$

→ The p -value is the smallest value of α for which the null would be rejected at level α , given $\mathbf{Y} = \mathbf{y}$.

Most usual setup:

- Have a single test statistic T
- Construct family $\delta_\alpha(\mathbf{y}) = \mathbf{1}\{T(\mathbf{y}) > k_\alpha\}$
- If $\mathbb{P}_{H_0}[T \leq t] = G(t)$ then $p(\mathbf{y}) = \mathbb{P}_{H_0}[T(\mathbf{Y}) \geq T(\mathbf{y})] = 1 - G(T(\mathbf{y}))$



Notice: contrary to NP-framework did not make explicit decision!

- We simply reported a p -value
- The p -value is used as a measure of evidence against H_0
 - ↪ Small p -value provides evidence against H_0
 - ↪ Large p -value provides no evidence against H_0
- How small does “small” mean?
 - ↪ Depends on the specific problem...

Notice: contrary to NP-framework did not make explicit decision!

- We simply reported a p -value
- The p -value is used as a measure of evidence against H_0
 - ↪ Small p -value provides evidence against H_0
 - ↪ Large p -value provides no evidence against H_0
- How small does “small” mean?
 - ↪ Depends on the specific problem...

Intuition:

- Recall that extreme values of test statistics are those that are “inconsistent” with null (NP-framework)
- p -value is probability of observing a value of the test statistic as extreme as or more extreme than the one we observed, under the null
- If this probability is [“]small, then we have witnessed something quite unusual under the null hypothesis
- Gives evidence against the null hypothesis

Example (Normal Mean)

Let $Y_1, \dots, Y_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Consider:

$$H_0 : \mu = 0 \quad \text{vs} \quad H_1 : \mu \neq 0$$

Likelihood ratio test: reject when T^2 large, $T = \sqrt{n}\bar{Y}/S \stackrel{H_0}{\sim} t_{n-1}$.

Since $T^2 \stackrel{H_0}{\sim} F_{1,n-1}$, p -value is

$$p(\mathbf{y}) = \mathbb{P}_{H_0}[T^2(\mathbf{Y}) \geq T^2(\mathbf{y})] = 1 - G_{F_{1,n-1}}(T^2(\mathbf{y}))$$

Consider two samples (datasets),

$$\mathbf{y} = (0.66, 0.28, -0.99, 0.007, -0.29, -1.88, -1.24, 0.94, 0.53, -1.2)$$

$$\mathbf{y}' = (1.4, 0.48, 2.86, 1.02, -1.38, 1.42, 2.11, 2.77, 1.02, 1.87)$$

Obtain $p(\mathbf{y}) = \underline{0.32}$ while $p(\mathbf{y}') = \underline{0.006}$.

- Reporting a p -value does not necessarily mean making a decision
- A small p -value can simply reflect our “confidence” in rejecting a null

- Reporting a p -value does not necessarily mean making a decision
- A small p -value can simply reflect our “confidence” in rejecting a null

Recall example: **Statisticians working for Trump** gather iid sample \mathbf{Y} from Florida with $Y_i = \mathbf{1}\{\text{vote Biden}\}$. Trumps team want to test

$$\begin{cases} H_0 : \text{Trump wins Florida} \\ H_1 : \text{Biden wins Florida} \end{cases}$$

- Will statisticians decide for Trump?
- Perhaps better to report p -value to him and let him decide...

- Reporting a p -value does not necessarily mean making a decision
- A small p -value can simply reflect our “confidence” in rejecting a null

Recall example: **Statisticians working for Trump** gather iid sample \mathbf{Y} from Florida with $Y_i = \mathbf{1}\{\text{vote Biden}\}$. Trumps team want to test

$$\begin{cases} H_0 : \text{Trump wins Florida} \\ H_1 : \text{Biden wins Florida} \end{cases}$$

- Will statisticians decide for Trump?
- Perhaps better to report p -value to him and let him decide...

What if statisticians working for newspaper, not Trump?

- Something easier to interpret than test/ p -value?

A Glance Back at Point Estimation

- Let Y_1, \dots, Y_n be iid random variables with density (frequency) $f(\cdot; \theta)$.
- Problem with point estimation: $\mathbb{P}_\theta[\hat{\theta} = \theta]$ typically small (if not zero)
 - ↪ always attach an estimator of variability, e.g. standard error
 - ↪ interpretation?

A Glance Back at Point Estimation

- Let Y_1, \dots, Y_n be iid random variables with density (frequency) $f(\cdot; \theta)$.
- Problem with point estimation: $\mathbb{P}_\theta[\hat{\theta} = \theta]$ typically small (if not zero)
 - ↪ always attach an estimator of variability, e.g. standard error
 - ↪ interpretation?
- Hypothesis tests may provide way to interpret estimator's variability within the setup of a particular problem $\hat{\theta}$
 - ↪ e.g. if observe $\hat{P}[\text{Biden wins}] = 0.52$ can actually see what p -value we get when testing $H_0 : P[\text{Biden wins}] \geq 1/2$.
- Something more directly interpretable?

A Glance Back at Point Estimation

- Let Y_1, \dots, Y_n be iid random variables with density (frequency) $f(\cdot; \theta)$.
- Problem with point estimation: $\mathbb{P}_\theta[\hat{\theta} = \theta]$ typically small (if not zero)
 - ↪ always attach an estimator of variability, e.g. standard error
 - ↪ interpretation?
- Hypothesis tests may provide way to interpret estimator's variability within the setup of a particular problem
 - ↪ e.g. if observe $\hat{P}[\text{Biden wins}] = 0.52$ can actually see what p -value we get when testing $H_0 : P[\text{Biden wins}] \geq 1/2$.
- Something more directly interpretable?

Back to our example: [What do pollsters do in newspapers?](#)

- ↪ They announce their point estimate (e.g. 0.52)
- ↪ They give upper and lower *confidence limits*

What are these and how are they interpreted?

Simple underlying idea:

- Instead of estimating θ by a single value
- Present a whole range of values for θ that are consistent with the data
 - In the sense that they could have produced the data

Simple underlying idea:

- Instead of estimating θ by a single value
- Present a whole range of values for θ that are consistent with the data
 - In the sense that they could have produced the data

Definition (Confidence Interval)

Let $\mathbf{Y} = (Y_1, \dots, Y_n)$ be random variables with joint distribution depending on $\theta \in \mathbb{R}$ and let $L(\mathbf{Y})$ and $U(\mathbf{Y})$ be two statistics with $L(\mathbf{Y}) < U(\mathbf{Y})$ a.s. Then, the random interval $[L(\mathbf{Y}), U(\mathbf{Y})]$ is called a $100(1 - \alpha)\%$ confidence interval for θ if

$$\underbrace{\mathbb{P}_\theta[L(\mathbf{Y}) \leq \theta \leq U(\mathbf{Y})]}_{\text{level} = \alpha} \geq 1 - \alpha$$

for all $\theta \in \Theta$, with equality for at least one value of θ .

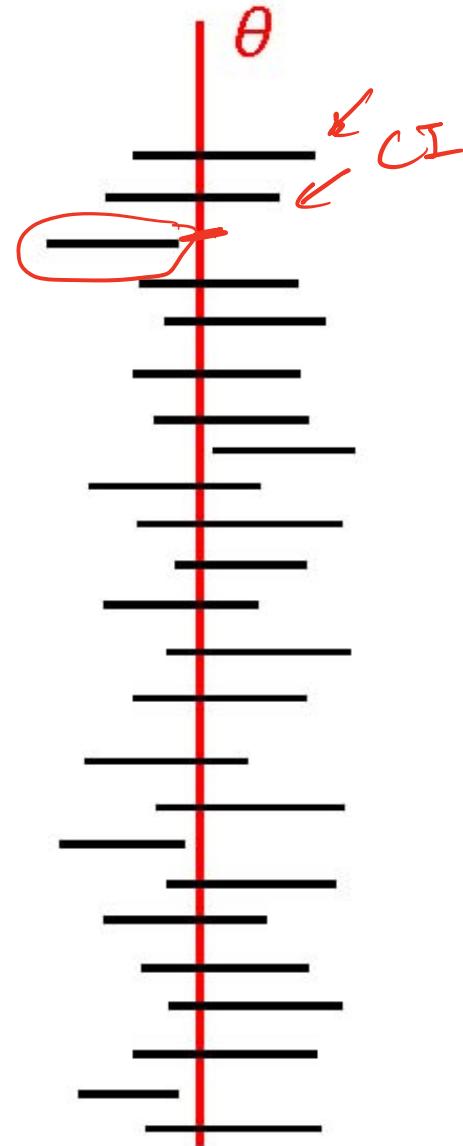
- $1 - \alpha$ is called the coverage probability or confidence level
- Beware of interpretation!

- Probability statement is **NOT** made about θ , which is constant.

- Probability statement is **NOT** made about θ , which is constant.
- Statement is about interval: probability that the interval contains the true value is at least $1 - \alpha$.

- Probability statement is **NOT** made about θ , which is constant.
- Statement is about interval: probability that the interval contains the true value is at least $1 - \alpha$.
- Given any realization $\mathbf{Y} = \mathbf{y}$, the interval $[L(\mathbf{y}), U(\mathbf{y})]$ will either contain or not contain θ .

- Probability statement is **NOT** made about θ , which is constant.
- Statement is about interval: probability that the interval contains the true value is at least $1 - \alpha$.
- Given any realization $\mathbf{Y} = \mathbf{y}$, the interval $[L(\mathbf{y}), U(\mathbf{y})]$ will either contain or not contain θ .
- Interpretation: if we construct intervals with this method, then we expect that $100(1 - \alpha)\%$ of the time our intervals will engulf the true value.



Example (The example that says all) cut

Let $Y_1, \dots, Y_n \stackrel{iid}{\sim} \mathcal{N}(\mu, 1)$. Then $\sqrt{n}(\bar{Y} - \mu) \sim \mathcal{N}(0, 1)$, so that $\alpha = 0.05$

$$\rightarrow \mathbb{P}_\mu[-1.96 \leq \sqrt{n}(\bar{Y} - \mu) \leq 1.96] = 0.95$$

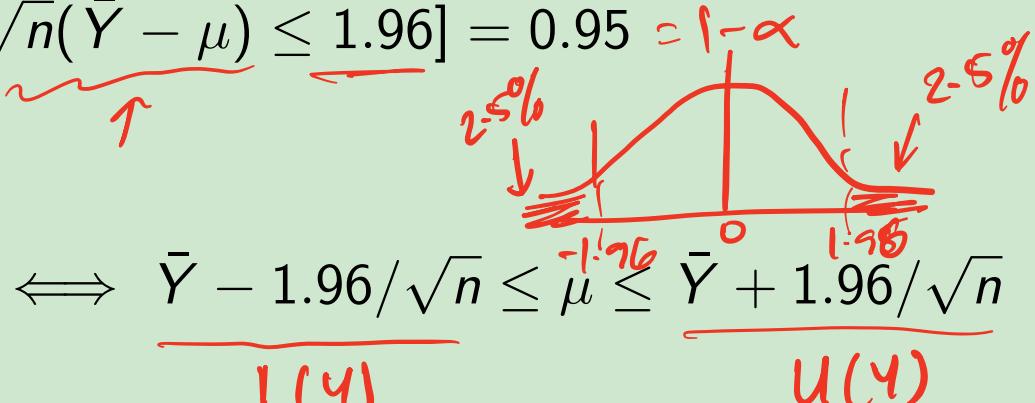
and since

$$-1.96 \leq \sqrt{n}(\bar{Y} - \mu) \leq 1.96 \iff \underbrace{\bar{Y} - 1.96/\sqrt{n} \leq \mu \leq \bar{Y} + 1.96/\sqrt{n}}_{L(\bar{Y}) \quad U(\bar{Y})}$$

we obviously have

$$\mathbb{P}_\mu \left[\bar{Y} - \frac{1.96}{\sqrt{n}} \leq \mu \leq \bar{Y} + \frac{1.96}{\sqrt{n}} \right] = 0.95$$

So that the random interval $[L(\bar{Y}), U(\bar{Y})] = \left[\bar{Y} - \frac{1.96}{\sqrt{n}}, \bar{Y} + \frac{1.96}{\sqrt{n}} \right]$ is a 95% confidence interval for μ .



Example (The example that says all)

Let $Y_1, \dots, Y_n \stackrel{iid}{\sim} \mathcal{N}(\mu, 1)$. Then $\sqrt{n}(\bar{Y} - \mu) \sim \mathcal{N}(0, 1)$, so that

$$\mathbb{P}_\mu[-1.96 \leq \sqrt{n}(\bar{Y} - \mu) \leq 1.96] = 0.95$$

and since

$$-1.96 \leq \sqrt{n}(\bar{Y} - \mu) \leq 1.96 \iff \bar{Y} - 1.96/\sqrt{n} \leq \mu \leq \bar{Y} + 1.96/\sqrt{n}$$

we obviously have

$$\mathbb{P}_\mu \left[\bar{Y} - \frac{1.96}{\sqrt{n}} \leq \mu \leq \bar{Y} + \frac{1.96}{\sqrt{n}} \right] = 0.95$$

So that the random interval $[L(\bar{Y}), U(\bar{Y})] = \left[\bar{Y} - \frac{1.96}{\sqrt{n}}, \bar{Y} + \frac{1.96}{\sqrt{n}} \right]$ is a 95% confidence interval for μ .

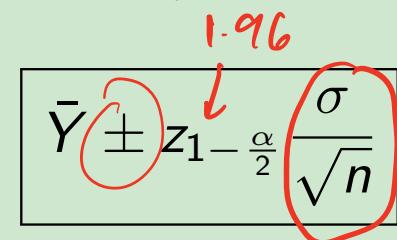
Central Limit Theorem: same argument can yield approximate 95% CI when Y_1, \dots, Y_n are iid, $\mathbb{E}Y_i = \mu$ and $\text{var}(Y_i) = 1$, regardless of their distribution.

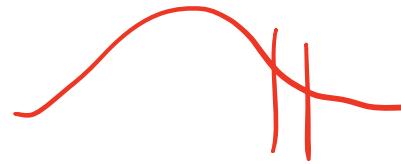
Example (continued)

Notice that the interval is centred at \bar{Y} , the MLE of μ . It's often thus written:

$$\bar{Y} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

1.96





Example (continued)

Notice that the interval is centred at \bar{Y} , the MLE of μ . It's often thus written:

$$\bar{Y} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$\hat{\sigma} \xrightarrow{\rho} \sigma$$

Observations:

- The length of the interval is $2z_{1-\alpha/2}\sigma/\sqrt{n}$, which depends on σ^2 , n and α .
- The parameter σ^2 is beyond our control.
- We can nevertheless control n and $1 - \alpha$. Increasing $\underline{\alpha}$, the length of the interval decays as $1/\sqrt{n}$.
- Reducing α (i.e. increasing $1 - \alpha$) increases the length of the interval (the dependence is quite non-linear, and 5% is chosen as a “sweet spot”).

What can we learn from previous example?

Definition (Pivot)

A random function $\underline{g(Y, \theta)}$ is said to be a *pivotal quantity* (or simply a *pivot*) if it is a function both of Y and θ whose distribution does not depend on θ .

→ $\sqrt{n}(\bar{Y} - \mu) \sim \mathcal{N}(0, 1)$ is a pivot in previous example
 $\underbrace{g(Y, \mu)}$

What can we learn from previous example?

Definition (Pivot)

A random function $g(\mathbf{Y}, \theta)$ is said to be a *pivotal quantity* (or simply a *pivot*) if it is a function both of \mathbf{Y} and θ whose distribution does not depend on θ .

→ $\sqrt{n}(\bar{Y} - \mu) \sim \mathcal{N}(0, 1)$ is a pivot in previous example

Why is a pivot useful?

- $\forall \alpha \in (0, 1)$ we can find constants $a < b$ independent of θ , such that

$$\mathbb{P}_\theta[a \leq g(\mathbf{Y}, \theta) \leq b] = 1 - \alpha \quad \forall \theta \in \Theta$$

- If $g(\mathbf{Y}, \theta)$ can be manipulated then the above yields a CI

Example

Let $Y_1, \dots, Y_n \stackrel{iid}{\sim} \mathcal{U}[0, \theta]$. Recall that MLE $\hat{\theta}$ is $\hat{\theta} = \underline{Y_{(n)}}$, with distribution

$$\mathbb{P}_\theta [Y_{(n)} \leq x] = F_{Y_{(n)}}(x) = \underline{\left(\frac{x}{\theta}\right)^n} \Rightarrow \mathbb{P}_\theta \left[\frac{Y_{(n)}}{\theta} \leq y\right] = y^n \cancel{\downarrow \theta}$$

→ Hence $\underline{Y_{(n)}}/\theta$ is a pivot for θ . Can now choose $a < b$ such that

$$\mathbb{P}_\theta \left[a \leq \frac{Y_{(n)}}{\theta} \leq b\right] = 1 - \alpha$$

→ But there are ∞ -many such choices!

↪ Idea: choose pair (a, b) that minimizes interval's length!

Solution can be seen to be $a = \underline{\alpha^{1/n}}$ and $b = \underline{1}$, yielding

$$\left[Y_{(n)}, \frac{Y_{(n)}}{\alpha^{1/n}}\right]$$

Pivotal method extends to construction of CI for θ_k , when

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_k, \dots, \theta_p) \in \mathbb{R}^p$$

and the remaining coordinates are also unknown. \rightarrow Pivotal quantity should now be function $g(\underline{\mathbf{Y}}; \underline{\theta_k})$ which

- ① Depends on \mathbf{Y} , θ_k , but no other parameters
- ② Has a distribution independent of any of the parameters

→ e.g.: CI for normal mean, when variance unknown

Pivotal method extends to construction of CI for θ_k , when

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_k, \dots, \theta_p) \in \mathbb{R}^p$$

and the remaining coordinates are also unknown. \rightarrow Pivotal quantity should now be function $g(\mathbf{Y}; \theta_k)$ which

- ➊ Depends on \mathbf{Y} , θ_k , but no other parameters
- ➋ Has a distribution independent of any of the parameters

→ e.g.: CI for normal mean, when variance unknown

\rightarrow Main difficulties with pivotal method:

- Hard to find exact pivots in general problems
- Exact distributions may be intractable

Pivotal method extends to construction of CI for θ_k , when

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_k, \dots, \theta_p) \in \mathbb{R}^p$$

and the remaining coordinates are also unknown. \rightarrow Pivotal quantity should now be function $g(\mathbf{Y}; \theta_k)$ which

- ① Depends on \mathbf{Y} , θ_k , but no other parameters
- ② Has a distribution independent of any of the parameters

→ e.g.: CI for normal mean, when variance unknown

\rightarrow Main difficulties with pivotal method:

- Hard to find exact pivots in general problems
- Exact distributions may be intractable

Resort to asymptotic approximations... \downarrow^{MLE}

→ Most classic example when have $a_n(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$.

$$(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$$

What about higher dimensional parameters?

Definition (Confidence Region)

Let $\mathbf{Y} = (Y_1, \dots, Y_n)$ be random variables with joint distribution depending on $\theta \in \Theta \subseteq \mathbb{R}^p$. A random subset $R(\mathbf{Y})$ of Θ depending on \mathbf{Y} is called a $100(1 - \alpha)\%$ confidence region for θ if

$$\mathbb{P}_\theta[R(\mathbf{Y}) \ni \theta] \geq 1 - \alpha$$

for all $\theta \in \Theta$, with equality for at least one value of θ .

- No restriction requiring $R(\mathbf{Y})$ to be convex or even connected
 - ↪ So when $p = 1$ get more general notion than CI
- Nevertheless, many notions extend immediately to CR case
 - ↪ e.g. notion of a pivotal quantity

Let $g : \mathcal{Y}^n \times \Theta \rightarrow \mathbb{R}$ be a function such that $\text{dist}[g(\mathbf{Y}, \theta)]$ independent of θ
 \hookrightarrow Since image space is the real line, can find $a < b$ s.t.

$$\mathbb{P}_{\theta}[a \leq \underbrace{g(\mathbf{Y}, \theta)}_{\text{red wavy line}} \leq b] = 1 - \alpha$$

$$\implies \mathbb{P}_{\theta}[R(\mathbf{Y}) \ni \theta] = 1 - \alpha$$

where $R(\mathbf{y}) = \{\theta \in \Theta : g(\mathbf{y}, \theta) \in [a, b]\}$

Notice that region can be “wild” since it is a random level set of g

Example

Let $\mathbf{Y}_1, \dots, \mathbf{Y}_n \stackrel{iid}{\sim} \mathcal{N}_k(\mu, \Sigma)$. Two unbiased estimators of μ and Σ are

$$\begin{aligned} \mu \in \mathbb{R}^k \\ \Sigma \in \mathbb{R}^{k \times k} \end{aligned}$$

$$\begin{aligned} \hat{\mu} &= \frac{1}{n} \sum_{i=1}^n \mathbf{Y}_i && \text{vector} \\ \hat{\Sigma} &= \frac{1}{n-1} \sum_{i=1}^n (\mathbf{Y}_i - \hat{\mu})(\mathbf{Y}_i - \hat{\mu})^T \end{aligned}$$

Example (cont'd)

Consider the random variable

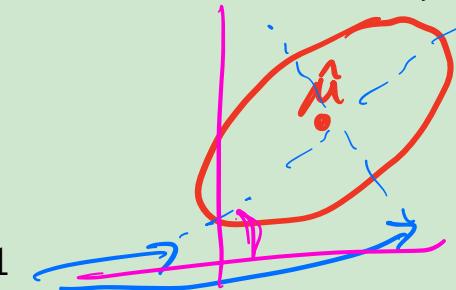
$$g(\{\mathbf{Y}\}_{i=1}^n, \mu) := \frac{n(n-k)}{k(n-1)}(\hat{\mu} - \mu)^T \hat{\Sigma}^{-1}(\hat{\mu} - \mu) \sim F\text{-dist with } k \text{ and } n-k \text{ d.f.}$$

A pivot!

↪ If f_q is q -quantile of this distribution, then get 100 q % CR as

$$R(\{\mathbf{Y}\}_{i=1}^n) = \left\{ \theta \in \mathbb{R}^n : \frac{n(n-k)}{k(n-1)}(\hat{\mu} - \mu)^T \hat{\Sigma}^{-1}(\hat{\mu} - \mu) \leq f_q \right\}$$

- An ellipsoid in \mathbb{R}^n
- Ellipsoid centred at $\hat{\mu}$
- Principle axis lengths given by eigenvalues of $\hat{\Sigma}^{-1}$
- Orientation given by eigenvectors of $\hat{\Sigma}^{-1}$



Visualisation of high-dimensional CR's can be hard

- When these are ellipsoids, spectral decomposition helps
- But more generally?

Visualisation of high-dimensional CR's can be hard

- When these are ellipsoids, spectral decomposition helps
- But more generally?

Things are especially easy when dealing with rectangles - **but they rarely occur!**

→ What if we construct a CR as Cartesian product of CI's?

Let $[L_i(\mathbf{Y}), U_i(\mathbf{Y})]$ be $100q_i\%$ CI's for θ_i , $i = 1, \dots, p$, and define

$$R(\mathbf{Y}) = [L_1(\mathbf{Y}), U_1(\mathbf{Y})] \times \dots \times [L_p(\mathbf{Y}), U_p(\mathbf{Y})]$$

Bonferroni's inequality implies that

$$\underbrace{\mathbb{P}_{\theta}[R(\mathbf{Y}) \ni \theta]}_{\geq 1 - \sum_{i=1}^p \mathbb{P}[\theta_i \notin [L_i(\mathbf{Y}), U_i(\mathbf{Y})]]} \geq 1 - \sum_{i=1}^p (1 - q_i)$$

→ So pick q_i such that $\sum_{i=1}^p (1 - q_i) = \alpha$ (can be conservative...)

Discussion on CR's → provides no guidance on choosing “good” regions

Discussion on CR's → provides no guidance on choosing “good” regions

But: \exists close relationship between CR's and hypothesis tests!

→ exploit this to transform good testing properties into good CR properties

Discussion on CR's \rightarrow provides no guidance on choosing "good" regions

But: \exists close relationship between CR's and hypothesis tests!

\rightarrow exploit this to transform good testing properties into good CR properties

Suppose $R(\mathbf{Y})$ is an exact $100q\% = 100(1 - \alpha)\%$ CR for θ . Consider

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0$$

Define test function:

$$\delta(\mathbf{Y}) = \begin{cases} 1 & \text{if } \theta_0 \notin \underline{R(\mathbf{Y})}, \\ 0 & \text{if } \theta_0 \in R(\mathbf{Y}). \end{cases}$$

Then, $\mathbb{E}_{\theta_0}[\delta(\mathbf{Y})] = 1 - \mathbb{P}_{\theta_0}[\theta_0 \in R(\mathbf{Y})] \leq \alpha$

Can use a CR to construct test with significance level α !

Going the other way around, can *invert* tests to get CR's:

Suppose we have tests at level α for any choice of simple null, $\theta_0 \in \Theta$.

↪ Say that $\delta(\mathbf{Y}; \theta_0)$ is appropriate test function for $H_0 : \theta = \theta_0$

Define

$$R^*(\mathbf{Y}) = \{\theta_0 : \delta(\mathbf{Y}; \theta_0) = 0\}$$

Coverage probability of $R^*(\mathbf{Y})$ is

$$\mathbb{P}_{\theta}[\theta \in R^*(\mathbf{Y})] = \mathbb{P}_{\theta}[\delta(\mathbf{Y}; \theta) = 0] \geq 1 - \alpha$$

Obtain a $100(1 - \alpha)\%$ confidence region by choosing all the θ for which the null would not be rejected given our data \mathbf{Y} .

↪ If test inverted is powerful, then get “small” region for given $1 - \alpha$.

Modern example: looking for signals in noise

- Interested in detecting presence of a signal $\mu(x_t)$, $t = 1, \dots, T$ over a discretised domain, $\{x_1, \dots, x_t\}$, on the basis of noisy measurements
- This is to be detected against some known background, say 0.
- May or may not be specifically interested in detecting the presence of the signal in some particular location x_t , but in detecting whether a signal is present anywhere in the domain.

Formally:

Does there exist a $t \in \{1, \dots, T\}$ such that $\mu(x_t) \neq 0$?

or

for which t 's is $\mu(x_t) \neq 0$?

More generally:

- Observe

$$\underbrace{Y_t = \mu(x_t) + \varepsilon_t}_{\text{---}} \quad t = 1, \dots, T.$$

- Wish to test, at some significance level α :

$$\begin{cases} H_0 : \underbrace{\mu(x_t)}_{\uparrow} = 0 & \text{for all } t \in \{1, \dots, T\}, \\ H_A : \mu(x_t) \neq 0 & \text{for some } t \in \{1, \dots, T\}. \end{cases}$$

- May also be interested in which specific locations signal deviates from zero
- More generally: May have T hypotheses to test simultaneously at level α (they may be related or totally unrelated)
- Suppose we have a test statistic for each individual hypothesis $H_{0,t}$ yielding a p -value p_t .

Bonferroni Method.

If we test each hypothesis individually, we will not maintain the level!

Can we maintain the level α ?

Idea: use the same trick as for confidence regions!

Bonferroni

- 1 Test individual hypotheses separately at level $\alpha_t = \alpha / T$
- 2 Reject $\underline{H_0}$ if at least one of the $\{H_{0,t}\}_{t=1}^T$ is rejected

Global level is bounded as follows:

$$\mathbb{P}[\text{not } H_0 | H_0] = \mathbb{P} \left[\bigcup_{t=1}^T \{\text{not } H_{0,t}\} \middle| H_0 \right] \leq \sum_{t=1}^T \mathbb{P}[\text{not } H_{0,t} | H_0] = T \left(\frac{\alpha}{T} \right) = \alpha$$

Holm-Bonferroni Method.

Bonferroni

$$p_{(t)} \leq \frac{\alpha}{T}$$

- Advantage: Works for any (discrete domain) setup!
- Disadvantage: Too conservative when T large

Holm's modification increases average # of hypotheses rejected at level α (but does not increase power for overall rejection of $H_0 = \cap_{t \in T} H_{0,t}$)

$$p_{(t)} \leq \frac{\alpha}{T-t+1}$$

Holm's Procedure

- ① We reject $\underline{H_{0,t}}$ for small values of a corresponding p -value, $\underline{p_t}$
- ② Order p -values from most to least significant: $\underline{p_{(1)} \leq \dots \leq p_{(T)}}$
- ③ Starting from $t = 1$ and going up, reject all $\underline{H_{0,(t)}}$ such that $\underline{p_{(t)}}$ significant at level $\alpha/(T-t+1)$. Stop rejecting at first insignificant $p_{(t)}$.

Genuine improvement over Bonferroni if want to detect as many signals as possible, not just existence of some signal

Both Holm and Bonferroni reject the global H_0 if and only if $\inf_t p_t$ significant at level α/T .

Taking Advantage of Structure: Independence.

In the (special) case where individual test statistics are independent, one may use Sime's (in)equality,

$$\mathbb{P} \left[p_{(j)} \geq \frac{j\alpha}{T} \text{ for all } j = 1, \dots, T \mid H_0 \right] \geq 1 - \alpha$$

(strict equality requires continuous test statistics, otherwise $\leq \alpha$)

Yields Sime's procedure (assuming independence)

- ① Suppose we reject $H_{0,j}$ for small values of p_j
- ② Order p -values from most to least significant: $p_{(1)} \leq \dots \leq p_{(T)}$
- ③ If, for some $j = 1, \dots, T$ the p -value $p_{(j)}$ is significant at level $\frac{j\alpha}{T}$, then reject the global H_0 .

Provides a test for the global hypothesis H_0 , but does not “localise” the signal at a particular x_t

One can, however, devise a sequential procedure to “localise” Sime’s procedure, at the expense of lower power for the global hypothesis H_0 :

Hochberg’s procedure (assuming independence)

- 1 Suppose we reject $H_{0,j}$ for small values of p_j
- 2 Order p -values from most to least significant: $p_{(1)} \leq \dots \leq p_{(T)}$
- 3 Starting from $j = T, T - 1, \dots$ and down, accept all $H_{0,(j)}$ such that $p_{(j)}$ is insignificant at level $\alpha/(T - j + 1)$.
- 4 Stop accepting for the first j such that $p_{(j)}$ is significant at level α/j , and reject all the remaining ordered hypotheses past that j going down.

Genuine improvement over Holm-Bonferroni both overall (H_0) and in terms of signal localisation:

- 1 Rejects “more” individual hypotheses than Holm-Bonferroni
- 2 Power for overall H_0 “weaker” than Sime’s (for $T > 2$), much “stronger” than Holm (for $T > 1$).

Bonferroni, Hochberg, Simes

