Testing Statistical Hypotheses

Statistical Theory

Guillaume Dehaene
Ecole Polytechnique Fédérale de Lausanne

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Contrasting Theories With Experimental Evidence

Using Data to Evaluate Theories/Assertions

- Scientific theories lead to assertions that are testable using empirical data.
- Data may discredit the theory (or *hypothesis*) or not
  - $\hookrightarrow$ i.e. are empirical findings reasonable under hypothesis?
- Example: Large Hadron Collider in CERN, Genève. Does the Higgs Boson exist? Study if particle trajectories are consistent with what theory predicts.
- Example: Theory of "luminoferous aether" in late 19th century to explain light travelling in vacuum. Discredited by Michelson-Morley experiment.
- Similarities with the logical/mathematical concept of a necessary condition

What would be the appropriate formal statistical framework?

Hypothesis Testing Setup

Statistical Framework for Testing Hypotheses

The Problem of Hypothesis Testing

- $\mathbf{X} = (X_1, ..., X_n)$ random variables with joint density/frequency $f(\mathbf{x}; \theta)$
- $\theta \in \Theta$ where $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \emptyset$ (or $\Lambda(\Theta_0 \cap \Theta_1) = 0$)
- Observe realization $\mathbf{x} = (x_1, ..., x_n)$ of $\mathbf{X} \sim f_\theta$
- Decide on the basis of $\mathbf{x}$ whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$

$\hookrightarrow$ Often $\dim(\Theta_0) < \dim(\Theta)$ so $\theta \in \Theta_0$ represents a *simplified model*.

Example

Let $X_1, ..., X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, 1)$ and $Y_1, ..., Y_n \stackrel{iid}{\sim} \mathcal{N}(\nu, 1)$. Have $\theta = (\mu, \nu)$ and

$$\Theta = \{(\mu, \nu) : \mu \in \mathbb{R}, \nu \in \mathbb{R}\} = \mathbb{R}^2$$

May be interested to see if $\mathbf{X}$ and $\mathbf{Y}$ have same distribution, even though they may be measurements on characteristics of different groups. In this case $\Theta_0 = \{(\mu, \nu) \in \mathbb{R}^2 : \mu = \nu\}$

Type I vs Type II Error

Decision Theory Perspective on Hypothesis Testing

Given **X** we need to *decide* between two hypotheses:

H_0: $\theta \in \Theta_0$ (the NULL HYPOTHESIS)

H_1: $\theta \in \Theta_1$ (the ALTERNATIVE HYPOTHESIS)

$\rightarrow$ Want decision rule $\delta : \mathcal{X} \rightarrow \mathcal{A} = \{0, 1\}$ (chooses between H_0 and H_1)

- In hypothesis testing δ is called a *test function*
- Often δ depends on **X** only through some real-valued statistic $T = T(\mathbf{X})$ called a *test statistic*.

Unlikely that a test function is perfect. Possible errors to be made?

Action / Truth	H_0	H_1
0	☺	Type II Error
1	Type I Error	☺

Potential asymmetry of errors in practice: false positive VS false negative (e.g. spam filters for e-mail)

Decision Theory Perspective on Hypothesis Testing

Typically loss function is "0–1" loss, i.e.

$$\mathcal{L}(\theta, a) = \begin{cases} 1 & \text{if } \theta \in \Theta_0 \ \& \ a = 1 \qquad \text{(Type I Error)} \\ 1 & \text{if } \theta \in \Theta_1 \ \& \ a = 0 \qquad \text{(Type II Error)} \\ 0 & \text{otherwise} \qquad \text{(No Error)} \end{cases}$$

i.e. we lose 1 unit whether we commit a type I or type II error.
$\longrightarrow$ Leads to the following risk function:

$$R(\theta, \delta) = \begin{cases} \mathbb{E}_\theta[\mathbf{1}\{\delta = 1\}] = \mathbb{P}_\theta[\delta = 1] & \text{if } \theta \in \Theta_0 \quad \text{(prob of type I error)} \\ \mathbb{E}_\theta[\mathbf{1}\{\delta = 0\}] = \mathbb{P}_\theta[\delta = 0] & \text{if } \theta \in \Theta_1 \quad \text{(prob of type II error)} \end{cases}$$

In short,

$$\begin{aligned} R(\theta, \delta) &= \mathbb{P}_\theta[\delta = 1]\mathbf{1}\{\theta \in \Theta_0\} + \mathbb{P}_\theta[\delta = 0]\mathbf{1}\{\theta \in \Theta_1\} \\ \text{"} = \text{"} &\ \text{"}\mathbb{P}_\theta[\text{choose } H_1 | H_0 \text{ is true}]\text{" OR "}\mathbb{P}_\theta[\text{choose } H_0 | H_1 \text{ is true}]\text{"} \end{aligned}$$

Optimal Testing?

As with point estimation, we may wish to find *optimal* test functions
$\hookrightarrow$ Find test functions that uniformly minimize risk?

- Almost never exists.
- In general there is a trade-off between the two error probabilities
- How to relax problem in this case? Minimize each type I and type II error probabilities separately?

For example consider: $X \sim \mathcal{N}(\mu, 1)$ where $H_0 : \mu = -1$ and $H_1 : \mu = 1$.

Consider the parametric decision rule: $\delta_t(X) = \mathbf{1}(X \geq t)$ (it's optimal).

If we increase t, probability of type 1 error decreases, but probability of type 2 error increases.

The Neyman-Pearson Setup

The Neyman-Pearson Setup

Classical approach: restrict class of test functions by "minimax reasoning"

1. We fix an $\alpha \in (0,1)$, usually small (called the significance level)
2. We declare that we only consider test functions $\delta : \mathcal{X} \to \{0,1\}$ such that

$$\delta \in \mathscr{D}(\Theta_0, \alpha) = \left\{\delta : \sup_{\theta \in \Theta_0} \mathbb{P}_\theta[\delta = 1] \le \alpha\right\}$$

i.e. rules for which prob of type I error is bounded above by α

$\hookrightarrow$ *Jargon: we fix a significance level for our test*

3. Within this restricted class of rules, choose δ to minimize prob of type II error uniformly on Θ_1:

$$\mathbb{P}_\theta[\delta(\mathbf{X}) = 0] = 1 - \mathbb{P}_\theta[\delta(\mathbf{X}) = 1]$$

4. Equivalently, maximize the *power* uniformly over Θ_1

$$\beta(\theta, \delta) = \mathbb{P}_\theta[\delta(\mathbf{X}) = 1] = \mathbb{E}_\theta[\mathbf{1}\{\delta(\mathbf{X}) = 1\}] = \mathbb{E}_\theta[\delta(\mathbf{X})], \quad \theta \in \Theta_1$$

(since $\delta = 1 \iff \mathbf{1}\{\delta = 1\} = 1$ and $\delta = 0 \iff \mathbf{1}\{\delta = 1\} = 0$)

The Neyman-Pearson Setup

Intuitive rationale of the approach:

- Want to test H_0 against H_1 at significance level α
- Suppose we observe $\delta(\mathbf{X}) = 1$ (so we take action 1)
- α is usually small, so that if H_0 is indeed true, we have observed something rare or unusual
  - $\hookrightarrow$ since $\delta = 1$ has probability at most α under H_0
- Evidence that H_0 is false (i.e. in favour of H_1)
- So taking action 1 is a highly reasonable decision

But what if we observe $\delta(\mathbf{X}) = 0$? (so we take action 0)

- Our significance level does not guarantee that our decision is necessarily reasonable
- Our decision would have been reasonable if δ was such that the type II error was also low (given the significance level).
- If we had maximized power β at level α though, then we would be reassured of our decision.

The Neyman-Pearson Setup

- Neyman-Pearson setup naturally exploits any asymmetric structure
- But, if natural asymmetry absent, need judicious choice of H_0

Example: Obama VS Romney 2012. Pollsters gather iid sample $\mathbf{X}$ from Ohio with $X_i = \mathbf{1}\{\text{vote Romney}\}$. Which pair of hypotheses to test?

$$\begin{cases} H_0: & \text{Romney wins Ohio} \\ H_1: & \text{Obama wins Ohio} \end{cases} \quad \text{OR} \quad \begin{cases} H_0: & \text{Obama wins Ohio} \\ H_1: & \text{Romney wins Ohio} \end{cases}$$

- Which pair to choose to make a prediction? (confidence intervals?)
- If Romney is conducting poll to decide whether he'll spend more money to campaign in Ohio, then his possible losses due to errors are:
    - (a) Spend more $'s to campaign in Ohio even though he would win anyway: lose $'s
    - (b) Lose Ohio to Obama because he thought he would win without any extra effort.
- (b) is much worse than (a) (especially since Romney had lots of $'s)
- Hence Romney would pick $H_0 = \{\text{Obama wins Ohio}\}$ as his null

Optimality in the Neyman Pearson Setup

Finding Good Test Functions

Consider simplest situation:

- Have $(X_1, ..., X_n) \sim f(\cdot;\theta)$ with $\Theta = \{\theta_0, \theta_1\}$

The Neyman-Pearson Lemma - Continuous Case

Let $\mathbf{X} = (X_1, ..., X_n)$ have joint density (frequency) function $f \in \{f_0, f_1\}$ and suppose we wish to test

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

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

$$\mathbb{P}_0[\Lambda \geq k] = \alpha$$

and the test whose test function is given by

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

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, taking values over the whole range $[0, 1]$. 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,

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

and therefore $\delta \in \mathscr{D}(\{\theta_0\}, \alpha)$ (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 $\psi(\mathbf{x}) \in \{0, 1\}$, then

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

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

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

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

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

$$\begin{aligned} \psi(\mathbf{x})(f_1(\mathbf{x}) - k \cdot f_0(\mathbf{x})) &\leq \delta(\mathbf{x})(f_1(\mathbf{x}) - k \cdot f_0(\mathbf{x})) \\ \int_{\mathbb{R}^n} \psi(\mathbf{x})(f_1(\mathbf{x}) - k \cdot f_0(\mathbf{x})) d\mathbf{x} &\leq \int_{\mathbb{R}^n} \delta(\mathbf{x})(f_1(\mathbf{x}) - k \cdot f_0(\mathbf{x})) d\mathbf{x} \end{aligned}$$

Rearranging the terms yields

$$\begin{aligned} \int_{\mathbb{R}^n} (\psi(\mathbf{x}) - \delta(\mathbf{x})) f_1(\mathbf{x}) d\mathbf{x} &\leq k \int_{\mathbb{R}^n} (\psi(\mathbf{x}) - \delta(\mathbf{x})) f_0(\mathbf{x}) d\mathbf{x} \\ \Longrightarrow \mathbb{E}_1[\psi(\mathbf{X})] - \mathbb{E}_1[\delta(\mathbf{X})] &\leq k \left(\mathbb{E}_0[\psi(\mathbf{X})] - \mathbb{E}_0[\delta(\mathbf{X})]\right) \end{aligned}$$

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

The Neyman-Pearson Lemma

- 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 $\delta' = \mathbf{1}\{\Lambda \geq k'\} \in \mathscr{D}(\{\theta_0\}, \alpha)$, then $\delta = \delta'$ almost surely.
- The result does not guarantee uniqueness when MP test exists.
- The result does not guarantee existence of an MP test, unless Λ is continuous.
- 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 these cases, we need to consider *randomised decision rules* too, in order to guarantee the existence of a most powerful test.

The Neyman-Pearson Lemma

General version of Neyman-Pearson lemma considers relaxed problem:

$$\text{maximize } \mathbb{E}_1[\delta] \qquad \text{subject to} \qquad \mathbb{E}_0[\delta] = \alpha \quad \& \quad 0 \le \delta(\mathbf{X}) \le 1 \ a.s.$$

$\rightarrow$ The optimum need not be a test function since now $\delta(x) : \mathcal{X} \to [0, 1]$!

Interpretation? Think of relaxation≡randomization:

- I.e., we are also willing to consider <u>randomised</u> decision rules.
- How does a randomised decision rule work?
  1. If $\delta(X) = 1$, reject.
  2. If $\delta(X) = 0$, don't reject.
  3. If $\delta(X) = p \in (0, 1)$, then sample an independent Bernoulli random variable Y with probability of success p.
     - (3a) If Y takes the value 1, then reject.
     - (3b) If Y takes the value 0, don't reject.

The last step is randomisation: we randomise because we inject further randomness, independent of any randomness

The Neyman-Pearson Lemma

Neyman-Pearson Lemma - General Case

Let $\mathbf{X} = (X_1, ..., X_n)$ have joint density (frequency) function $f \in \{f_0, f_1\}$ and suppose we wish to test

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

at level $\alpha \in (0,1)$. Let $\Lambda(X) = f_1(X)/f_0(X)$. Then, there exist $k > 0$ and $p \in [0,1]$ such that the decision rule

$$\delta(\mathbf{X}) = \begin{cases} 1 & \text{if } \Lambda(X) > k, \\ p & \text{if } \Lambda(X) = k, \\ 0 & \text{if } \Lambda(X) < k, \end{cases}$$

satisfies

$$\mathbb{E}_0[\delta(X)] = \alpha \qquad \& \qquad \mathbb{E}_1[\psi(X)] \leq \mathbb{E}_1[\delta(X)]$$

for all $\psi : \mathcal{X} \to [0,1]$ such that $\mathbb{E}_0[\psi(X)] \leq \alpha$.

Proof.

Let $G_0(t) = \mathbb{P}_0[\Lambda \leq t]$ and $k = \inf\{t : G_0(t) \geq 1-\alpha\}$. If $G_0(k) = 1-\alpha$, then set $p = 0$ and proceed as in the continuous version of the NP-lemma. Otherwise, if $G_0(k) > 1-\alpha$, define $\xi := \lim_{\epsilon\to 0} G(k-\epsilon) < (1-\alpha)$ and

$$p = \frac{G_0(k) - (1-\alpha)}{G_0(k) - \xi}.$$

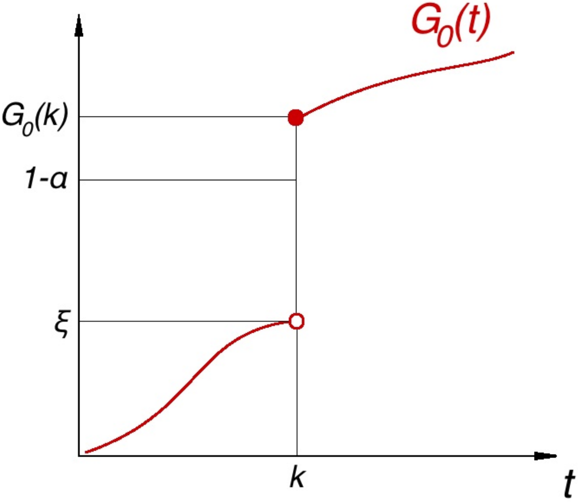
By definition of ξ, it must be that $p \in (0,1)$. Furthermore, note that

$$G_0(k) - \xi = \mathbb{P}_0[\Lambda \leq k] - \lim_{\epsilon\to 0} \mathbb{P}_0[\Lambda \leq k-\epsilon] = \mathbb{P}_0[\Lambda = k]$$

($\lim_{\epsilon\to 0} \mathbb{P}_0[\Lambda \leq k-\epsilon] = \mathbb{P}_0[\Lambda < k]$ by continuity of probability measures from above).

$$\begin{aligned} \text{So:} \quad \mathbb{E}_0[\delta] &= 1 \times \mathbb{P}_0[\Lambda > k] + p \times \mathbb{P}_0[\Lambda = k] + 0 \times \mathbb{P}_0[\Lambda < k] \\ &= 1 - G_0(k) + \frac{G_0(k) - (1-\alpha)}{\mathbb{P}_0[\Lambda = k]} \times \mathbb{P}_0[\Lambda = k] = \alpha. \end{aligned}$$

For the power, repeat the steps in the proof of continuous NP-lemma. □



(recall that G_0 is necessarily *càdlàg*: continue à droite, limite à gauche)

The Neyman-Pearson Setup

Example (Exponential Distribution)

Let $X_1, ..., X_n \overset{iid}{\sim} \text{Exp}(\lambda)$ and $\lambda \in \{\lambda_1, \lambda_2\}$, with $\lambda_1 > \lambda_0$ (say). Consider

$$\begin{cases} H_0 : & \lambda = \lambda_0 \\ H_1 : & \lambda = \lambda_1 \end{cases}$$

Have

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

So Neyman-Pearson say we must base our test on the statistic

$$T = \frac{f(\mathbf{X}; \lambda_1)}{f(\mathbf{X}; \lambda_0)} = \left(\frac{\lambda_1}{\lambda_0}\right)^n \exp\left[(\lambda_0 - \lambda_1) \sum_{i=1}^{n} X_i\right]$$

rejecting the null if $T \geq k$, for k such that the level is α.

Example (cont'd)

To determine k we note that T is a decreasing function of $S = \sum_{i=1}^{n} X_1$ (since $\lambda_0 < \lambda_1$). Therefore

$$T \geq k \iff S \leq K$$

for some K, so that

$$\alpha = \mathbb{P}_{\lambda_0}[T \geq k] \iff \alpha = \mathbb{P}_{\lambda_0}\left[\sum_{i=1}^{n} X_i \leq K\right]$$

For given values of λ_0 and α it is entirely feasible to find the appropriate K: under the null hypothesis, S has a gamma distribution with parameters n and λ_0. Hence we reject H_0 at level α if S exceeds that α-quantile of a gamma(n, λ_0) distribution.

Example (Uniform Distribution)

Let $X_1, ...X_n \overset{iid}{\sim} \mathcal{U}[0,\theta]$ with $\theta \in \{\theta_0, \theta_1\}$ where $\theta_0 > \theta_1$. Consider

$$\begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta = \theta_1 \end{cases}$$

Recall that

$$f(\mathbf{x};\theta) = \frac{1}{\theta^n}\mathbf{1}\left\{\max_{1\le i\le n} X_i \le \theta\right\}$$

so an MP test of H_0 vs H_1 should be based on the discrete test statistic

$$T = \frac{f(\mathbf{X};\theta_1)}{f(\mathbf{X};\theta_0)} = \left(\frac{\theta_0}{\theta_1}\right)^n \mathbf{1}\{X_{(n)} \le \theta_1\}.$$

So if the test rejects H_0 when $X_{(n)} \le \theta_1$ then it is MP for H_0 vs H_1 at

$$\alpha = \mathbb{P}_{\theta_0}[X_{(n)} \le \theta_1] = (\theta_1/\theta_0)^n$$

with power $\mathbb{P}_{\theta_1}[X_{(n)} \le \theta_1] = 1$. What about smaller values of α?

Example (cont'd)

$\hookrightarrow$ What about finding an MP test for $\alpha < (\theta_1/\theta_0)^n$?

An intuitive test statistic is the sufficient statistic $X_{(n)}$, giving the test

$$\text{reject } H_0 \quad \text{iff} \quad X_{(n)} \leq k$$

with k solving the equation:

$$\mathbb{P}_{\theta_0}[X_{(n)} \leq k] = \left(\frac{k}{\theta_0}\right)^n = \alpha,$$

i.e. with $k = \theta_0 \alpha^{1/n}$, with power

$$\mathbb{P}_{\theta_1}[X_{(n)} \leq \theta_0 \alpha^{1/n}] = \left(\frac{\theta_0 \alpha^{1/n}}{\theta_1}\right)^n = \alpha \left(\frac{\theta_0}{\theta_1}\right)^n.$$

Is this the MP test at level $\alpha < (\theta_1/\theta_0)^n$ though?

Example (cont'd)

Use general form of the Neyman-Pearson lemma to solve relaxed problem:

maximize $\mathbb{E}_1[\delta(\mathbf{X})]$ subject to $\mathbb{E}_{\theta_0}[\delta(\mathbf{X})] = \alpha < \left(\frac{\theta_1}{\theta_0}\right)^n$ & $0 \leq \delta(\mathbf{x}) \leq 1$.

One solution to this problem is given by

$$\delta(\mathbf{X}) = \begin{cases} \alpha(\theta_0/\theta_1)^n & \text{if } X_{(n)} \leq \theta_1, \\ 0 & \text{otherwise.} \end{cases}$$

which is not a test function. However, we see that its power is

$$\mathbb{E}_{\theta_1}[\delta(\mathbf{X})] = \alpha\left(\frac{\theta_0}{\theta_1}\right)^n = \mathbb{P}_{\theta_1}[X_{(n)} \leq \theta_0\alpha^{1/n}]$$

which is the power of the test we proposed.
Hence the test that rejects H_0 if $X_{(n)} \leq \theta_0\alpha^{1/n}$ is an MP test for all levels $\alpha < (\theta_1/\theta_0)^n$.

Summary

Hypothesis testing is a key statistical problem.

Key insight: the errors are not symmetric.

Neyman-Pearson setup:

- First, we choose a **significance level** α.
- We seek to maximize (if possible) the **power** of the test while maintaining the significance level.

In a simple vs simple test, there exists an optimal test for any level α. If the data is continuous, this test might be randomized for most values of α.

I personally strongly disagree with randomized rules.