

Global null testing

Modern Inference

Aldo Solari

Outline

- ❶ The p -value
- ❷ Global null testing
- ❸ Maximum statistic
- ❹ Sum of statistics
- ❺ Higher Criticism

Deterministic proof by contradiction

- ➊ Assume a proposition, the opposite of what you think about, i.e. the opposite conclusion of your theorem
- ➋ Write down a sequence of logical steps/math
- ➌ Derive a contradiction
- ➍ Conclude that the proposition is false (which implies that the theorem is true)

Stochastic proof by contradiction

- ➊ Set H_0 (the proposition)
- ➋ Collect data (which is noisy)
- ➌ Derive an apparent contradiction (i.e. if H_0 is true, then this data is very weird)
- ➍ Hence we reject H_0 ; this is called a “discovery”

Hypothesis testing is stochastic because we might make errors:

Type I (false discoveries) and *Type II* (missed discoveries)

Theorem: *There is no greatest even integer.*

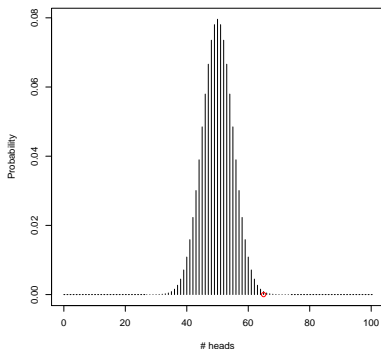
- We take the negation of the theorem and suppose it to be true: Suppose there is greatest even integer N . [We must deduce a contradiction.]
 - Then, for every even integer n , $N \geq n$.
 - Now suppose $M = N + 2$. Then, M is an even integer. [Because it is a sum of even integers.]
 - Also, $M > N$ [since $M = N + 2$].
 - Therefore, M is an integer that is greater than the greatest integer.
 - This contradicts the supposition that $N \geq n$ for every even integer n . [Hence, the supposition is false and the statement is true.]
- And this completes the proof.

Assume we have a coin and want to determine whether or not it is fair. In this case we have

H_0 : Coin is fair ($\theta = 1/2$)

H_1 : Coin is biased ($\theta \neq 1/2$)

The probability distribution of X = "the number of heads in 100 trials" under H_0 is Binomial($n = 100, \theta = 1/2$). After tossing the coin $n = 100$ times we then get $x = 65$ heads and $n - x = 35$ tails.



- Is this then enough to reject H_0 ?
- To determine this we calculate a p -value associated with our observed data assuming the null hypothesis
- A p -value is the probability of seeing what you saw - or something more extreme - given that H_0 is true.
- Small p -values imply an unexpected outcome, given that the null (H_0) is true
- So if $p = 0.0018$ then either H_0 isn't true or we are really unlucky and saw this data

The p -value

- Choose a test statistic $T = t(Y)$, large values of which cast doubts on H_0
- Observe the data y_{obs} , realization of Y
-

$$p_{obs} = P_0(T \geq t_{obs})$$

where $t_{obs} = t(y_{obs})$ and P_0 is the probability under H_0

p-value null distribution

- $p_{\text{obs}} = 1 - F_0(t_{\text{obs}})$, where F_0 is the null distribution function of T , supposed to be continuous and invertible.
- One interpretation of p_{obs} stems from the corresponding random variable $P = 1 - F_0(T)$
- The null distribution of P is *Uniform*(0,1): for any $u \in (0, 1)$,

$$P_0(P \leq u) = P_0(F_0^{-1}(1 - u) \leq T) = 1 - F_0(F_0^{-1}(1 - u)) = u$$

One- and two-sided tests

- Suppose that we have a test statistic T with continuous distribution, small and large values of which indicate a departure from H_0
- Calculate

$$p_{\text{obs}}^- = P_0(T \leq t_{\text{obs}}), \quad p_{\text{obs}}^+ = P_0(T \geq t_{\text{obs}})$$

- The p -value is

$$p_{\text{obs}} = 2 \min(p_{\text{obs}}^-, p_{\text{obs}}^+)$$

This follows because the null distribution of $Q = \min(P^-, P^+)$ is

$$Q = \min(1 - U(0, 1), U(0, 1)) = U(0, 1/2)$$

thus the null distribution of $2Q$ is $U(0, 1)$

Discrete null distribution

- Suppose $T \sim \text{Poisson}(\mu)$ and we observe $t_{\text{obs}} = 3$
- We want to test $H_0 : \mu = 2$ vs $H_1 : \mu \neq 2$
-

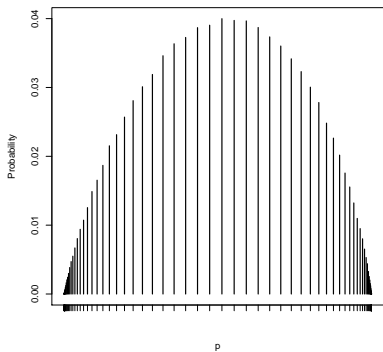
$$p_{\text{obs}}^+ = P_0(T \geq t_{\text{obs}}) = \sum_{t=t_{\text{obs}}}^{\infty} \frac{\mu^t e^{-\mu}}{t!}$$

$$p_{\text{obs}}^- = P_0(T \leq t_{\text{obs}}) = \sum_{t=0}^{t_{\text{obs}}} \frac{\mu^t e^{-\mu}}{t!}$$

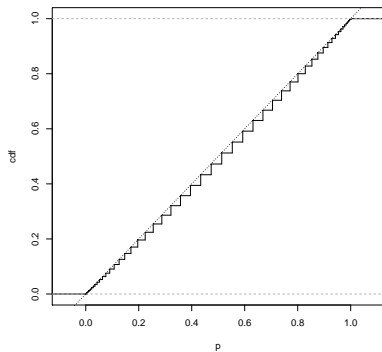
With discrete null distribution, p_{obs} is $q_{\text{obs}} = \min(p_{\text{obs}}^-, p_{\text{obs}}^+)$ plus the achievable p -value from the other tail of the distribution nearest to but not exceeding q_{obs}

- In the example, $p_{\text{obs}} = 0.458 = \min(0.323, 0.857) + 0.135$

t	0	1	2	3	4	5
$P_0(T \geq t)$	1	0.865	0.594	0.323	0.143	0.053
$P_0(T \leq t)$	0.135	0.406	0.677	0.857	0.947	0.983



$$P_0(P = p)$$



$$P_0(P \leq p)$$

Example of null discrete distribution (pmf and cdf) of the p -value

Valid p -values

We have a *valid test* if the p -value is uniformly distributed under H_0 , i.e.

$$P_0(P \leq u) = u \quad \forall u \in (0, 1)$$

or more generally (e.g. for discrete null distributions) if the p -value is *stochastically dominated* by the uniform distribution under H_0 , i.e.

$$P_0(P \leq u) \leq u \quad \forall u \in (0, 1)$$

Outline

- ① The p -value
- ② **Global null testing**
- ③ Maximum statistic
- ④ Sum of statistics
- ⑤ Higher Criticism

Inference on the mean

$$y \sim N_m(\mu, \Sigma)$$

where $y = (y_1, \dots, y_m)^T$ is the response, $\mu = (\mu_1, \dots, \mu_m)^T$ is the mean vector (where $\mu_i = 0$ means “no effect” and $\mu_i \neq 0$ means “effect”) and Σ is the correlation matrix. Marginally, $y_i \sim N(\mu_i, 1)$

Consider the following questions:

- ❶ *Detecting effects*: There is at least one μ_i different from 0?
- ❷ *Counting effects*: How many μ_i are different from 0?
- ❸ *Identifying effects*: Which μ_i are different from 0?

Global null

Testing the global null hypothesis aims at detecting any effect

$H_0 : \mu = 0$, i.e. $\mu_i = 0$ for all $i = 1, \dots, m$

$H_1 : \mu \neq 0$, i.e. $\mu_i \neq 0$ for at least one i

One-sided alternative

$H_0: \mu_i = 0$ for all $i = 1, \dots, m$

$H_1: \mu_i > 0$ for at least one i

We will consider three different tests (one-sided alternative):

- Maximum statistic: $T_{\max} = \max(y_1, \dots, y_m)$
- Sum of statistics $T_{\text{sum}} = \sum_{i=1}^m y_i$
- Higher criticism

For simplicity, assume that the y_i s are independent (i.e. $\Sigma = I_m$)

Outline

- ① The p -value
- ② Global null testing
- ③ Maximum statistic**
- ④ Sum of statistics
- ⑤ Higher Criticism

Maximum statistic

- The critical value $t_{1-\alpha}$ of the test based on the maximum statistic is

$$P_0(T_{\max} \geq t_{1-\alpha}) = \alpha$$

where $t_{1-\alpha}$ is the $1 - \alpha$ quantile of the distribution of the maximum of m independent standard normal variables

$$\int_{t_{1-\alpha}}^{\infty} m\phi(y)\Phi(y)^{m-1}dy = \alpha$$

where ϕ and Φ are the density and cdf of $N(0, 1)$

Minimum p -value

- Equivalently, define $p_i = 1 - \Phi(y_i) \stackrel{H_0}{\sim} U(0, 1)$ and use the minimum p -value

$$p_{\min} = \min(p_1, \dots, p_m) \stackrel{H_0}{\sim} \text{Beta}(1, m)$$

- The MinP test rejects H_0 if $p_{\min} \leq 1 - (1 - \alpha)^{\frac{1}{m}}$
- To see this

$$\begin{aligned} P_0(p_{\min} \leq 1 - (1 - \alpha)^{\frac{1}{m}}) &= 1 - P_0\left(\bigcap_{i=1}^m \{p_i > 1 - (1 - \alpha)^{\frac{1}{m}}\}\right) \\ &= 1 - [(1 - \alpha)^{\frac{1}{m}}]^m = \alpha \end{aligned}$$

Approximated critical value

- Replace $t_{1-\alpha}$ by $z_{1-\alpha/m}$, where z_α is the α quantile of $N(0, 1)$

$$\begin{aligned} P_0(T_{\max} \geq z_{1-\alpha/m}) &= P_0\left(\bigcup_{i=1}^m \{y_i \geq z_{1-\alpha/m}\}\right) \\ &\leq \sum_{i=1}^m P_0(y_i \geq z_{1-\alpha/m}) = m \frac{\alpha}{m} = \alpha \end{aligned}$$

- The union bound might seem crude, but with independent y_i s the size of the test is very near α

$$\begin{aligned} P_0(T_{\max} \geq z_{1-\alpha/m}) &= 1 - \prod_{i=1}^m P_0(y_i < z_{1-\alpha/m}) \\ &= 1 - \left(1 - \frac{\alpha}{m}\right)^m \xrightarrow{m \rightarrow \infty} 1 - e^{-\alpha} \approx \alpha \end{aligned}$$

For $\alpha = 0.05$, the size is 0.0487 (asymptotically)

Magnitude of the critical value

- We have

$$P(Z > t) \leq \frac{\phi(t)}{t}$$

- To see this

$$\int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz < \frac{1}{t\sqrt{2\pi}} \int_t^\infty z e^{-z^2/2} dz = \frac{1}{t\sqrt{2\pi}} e^{-t^2/2}$$

- It follows that

$$P(Z > \sqrt{2 \log m}) \leq \frac{\phi(\sqrt{2 \log m})}{\sqrt{2 \log m}} = \frac{1}{2m\sqrt{\pi \log m}} < \frac{\alpha}{m}$$

as soon as $\sqrt{\log m} > \frac{1}{2\sqrt{\pi\alpha}}$

- Therefore $z_{1-\alpha/m} \leq \sqrt{2 \log m}$ if $\sqrt{\log m} > \frac{1}{2\sqrt{\pi\alpha}}$

Magnitude of the critical value

- It can be proved that $(1 - 1/t^2) \frac{\phi(t)}{t} < P(Z > t)$
- Then, for any fixed α and $\epsilon > 0$, the following inequalities hold for all large enough m :

$$\sqrt{(1 - \epsilon)2 \log(m)} \leq z_{1-\alpha/m} \leq \sqrt{2 \log(m)}$$

Hence, $z_{1-\alpha/m}$ grows like $\sqrt{2 \log m}$

- For large m , the maximum test rejects H_0 when

$$T_{\max} \geq \sqrt{2 \log m}$$

and there is (asymptotically) no dependence on α

Needle in a haystack problem

$H_0 : \mu_i = 0 \text{ for all } i = 1, \dots, m$

$H_1 : \mu_i = c > 0, \mu_j = 0 \text{ for } j \neq i$

- What is the limiting power of the test?

$$\lim_{m \rightarrow \infty} P_1(T_{\max} > z_{1-\alpha/m})$$

- The answer to this question depends on the limiting ratio

$$\lim_{m \rightarrow \infty} \frac{c}{\sqrt{2 \log m}}$$

where $c = c(m)$ is the value of the single non-zero mean, which is a function of m

Needle in a haystack problem

Two cases:

- Suppose $c > (1 + \epsilon)\sqrt{2 \log m}$. Then, assuming without loss of generality that $\mu_1 = c$,

$$P_1(T_{\max} > z_{1-\alpha/m}) \geq P_1(y_1 > z_{1-\alpha/m}) = P(Z > z_{1-\alpha/m} - c) \rightarrow 1$$

- Suppose $c < (1 + \epsilon)\sqrt{2 \log m}$. Then

$$\begin{aligned} P_1(T_{\max} > z_{1-\alpha/m}) &\leq P(y_1 > z_{1-\alpha/m}) + P(\max_{i>1} y_i > z_{1-\alpha/m}) \\ &= P(Z > z_{1-\frac{\alpha}{m}} - c) + P(\max_{i>1} y_i > z_{1-\frac{\alpha}{m}}) \\ &\rightarrow 0 + (1 - e^{-\alpha}) \approx \alpha \end{aligned}$$

- Can we do better than this test? No, it is asymptotically equivalent to optimal test given by Neyman-Pearson lemma

Outline

- ① The p -value
- ② Global null testing
- ③ Maximum statistic
- ④ Sum of statistics**
- ⑤ Higher Criticism

Sum of statistics

$$T_{\text{sum}} = \sum_{i=1}^m y_i \sim N\left(\sum_{i=1}^m \mu_i, m\right)$$

- $Z_{\text{sum}} = \frac{T_{\text{sum}}}{\sqrt{m}} \stackrel{H_0}{\sim} N(0, 1)$
- $Z_{\text{sum}} \stackrel{H_1}{\sim} N(\theta, 1)$ where $\theta = \frac{\sum_{i=1}^m \mu_i}{\sqrt{m}}$
- If $\theta \rightarrow 0$ when $m \rightarrow \infty$, then the test has no power
- By the Neyman-Pearson lemma, T_{sum} is the uniformly most powerful (UMP) test for testing against a dense alternative with constant effect:
 $H_0 : \mu_i = 0$ for all i
 $H_1 : \mu_i = c > 0$ for all i
- Here $\theta = \sqrt{m}c$, but if $c = \frac{1}{m}$ the optimal test has no power

Comparison

The two test are effective in two different regimes:

- **Few strong effects:**

$m^{1/4}$ of the μ_i s are equal to $\sqrt{2 \log m}$, the rest 0.

E.g. when $m = 10^6$, $n^{1/4} \approx 36$ and $\sqrt{2 \log m} \approx 5.3$. In this setting T_{\max} has full power, but T_{sum} has no power because

$$\theta = \frac{m^{1/4} \sqrt{2 \log m}}{\sqrt{m}} \rightarrow 0$$

- **Small, distributed effects:**

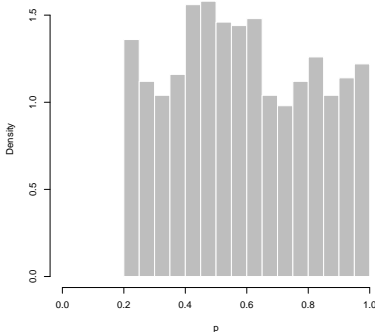
$\sqrt{2m}$ of the μ_i s are equal to 3, the rest 0.

The T_{sum} has (almost) full power, but T_{\max} has no power because when m is large it's very likely that the largest y_i value comes from a null μ_i , not a true signal. An intuitive argument is as follows: among the nulls, the largest y_i has size $\approx \sqrt{2 \log m}$ while among the true signals, the largest y_i has size $\approx 3 + \sqrt{2 \log \sqrt{2m}}$. If m is large, the former is larger

Outline

- ① The p -value
- ② Global null testing
- ③ Maximum statistic
- ④ Sum of statistics
- ⑤ Higher Criticism**

- Suppose that p_1, \dots, p_m are independent and $p_i \stackrel{H_0}{\sim} U(0, 1)$
- For $m = 1000$ you observe the following histogram of p -values:



- Would you reject the global null hypothesis H_0 ?
- Note that in this case $H_0 : p_1, \dots, p_m$ i.i.d. $U(0, 1)$

Kolmogorov-Smirnov test

- Consider the empirical CDF $\hat{F}_m(t) = \frac{\sum_{i=1}^m \mathbb{1}\{p_i \leq t\}}{m}$
- Under H_0 , each p_i has marginal distribution $U(0, 1)$ thus $\mathbb{E}_0(\hat{F}_m(t)) = t$
- Moreover, if we assume that p_i s are independent, i.e. p_1, \dots, p_m i.i.d. $U(0, 1)$ under H_0 , then

$$m\hat{F}_m(t) \sim \text{Binomial}(m, t)$$

- Hence, we measure the distance between what we observe and what we expect and reject if the difference is large, e.g. the Kolmogorov-Smirnov test statistic

$$T_{\text{KS}} = \sup_{t \in [0, 1]} |\hat{F}_m(t) - t|$$

Dvoretzky-Kiefer-Wolfowitz inequality

- Given m , let y_1, \dots, y_m be real-valued i.i.d. r.v. with cdf F and let \hat{F}_m denote the associated ecdf. Then for a given constant $\epsilon > 0$

$$P(\|\hat{F}_m - F\|_\infty > \epsilon) \leq 2e^{-2m\epsilon^2}$$

where $\|\hat{F}_m - F\|_\infty = \sup_{t \in [0,1]} |\hat{F}_m(t) - F(t)|$.

- It can be rephrased as follows:

$$\|\hat{F}_m - F\|_\infty \leq \sqrt{\frac{\log \frac{2}{\alpha}}{2m}}$$

with probability $\geq 1 - \alpha$

Higher criticism

- This test statistic is designed to take account for the variance in the Binomial distribution of the statistic T_{KS}

$$T_{HC} = \sup_{0 \leq t \leq \alpha_0} \frac{\hat{F}_m(t) - t}{\sqrt{t(1-t)/m}}$$

- It can be equivalently written as follows:

$$T_{HC} = \max_{0 \leq i \leq m\alpha_0} T_i, \quad T_i = \sqrt{m} \frac{(i/m) - p_{(i)}}{\sqrt{p_{(i)}(1 - p_{(i)})}}$$

where $p_{(1)} \leq \dots \leq p_{(n)}$ are the sorted p -values

Critical value

- T_{HC} can be connected with the maximum of a standardized empirical process. For $m \rightarrow \infty$, $b_m T_{\text{HC}} - c_m$ converges weakly to the standard Gumbel distribution, where $b_m = \sqrt{2 \log \log m}$ and $c_m = \frac{1}{2}(\log \log \log(m) - 4\pi)$
- For any fixed α and $m \rightarrow \infty$, the $1 - \alpha$ quantile of the null distribution for T_{HC} is

$$t_{1-\alpha} \approx (1 + a)\sqrt{2 \log \log m}$$

for some $a > 0$, e.g. $a = 1.08$ for $m \approx 10^6$, $\alpha_0 = 1$ and $\alpha = 0.05$ by simulations

Mixture distribution

- We assume that our samples follow a mixture of $N(0, 1)$ and $N(\mu, 1)$ distributions with μ fixed, resulting in

$$H_0 : y_i \stackrel{i.i.d}{\sim} N(0, 1)$$

$$H_1 : y_i \stackrel{i.i.d}{\sim} \pi_0 N(0, 1) + \pi_1 N(\mu, 1)$$

where $\pi_1 = 1 - \pi_0$

- To carry out asymptotic analysis, we must specify the dependence scheme of $\pi_1 = \pi_1(m)$ and $\mu = \mu(m)$ on m :

$$\begin{aligned} \pi_1 &= m^{-\beta} & \frac{1}{2} < \beta < 1 \\ \mu &= \sqrt{2r \log m} & 0 < r < 1 \end{aligned}$$

- Note that The needle in a haystack problem corresponds to $\beta = 1$ and $r = 1$; the small distributed effects case corresponds to $\beta = 1/2$

Threshold curve

The following threshold curve for r

$$\rho_{\text{HC}}(\beta) = \begin{cases} \beta - \frac{1}{2} & \text{if } \frac{1}{2} < \beta \leq \frac{3}{4} \\ (1 - \sqrt{1 - \beta})^2 & \text{if } \frac{3}{4} \leq \beta \leq 1 \end{cases}$$

is such that

- If $r > \rho_{\text{HC}}(\beta)$ the Neyman-Pearson optimal test achieves

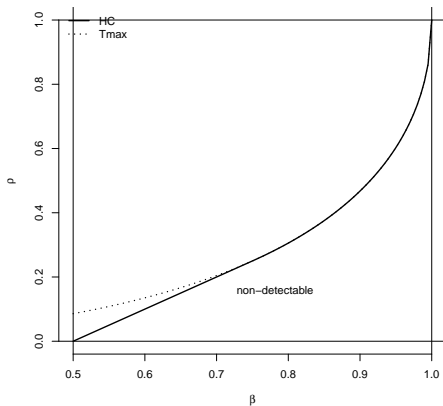
$$P_0(\text{Type I Error}) + P_1(\text{Type II Error}) \rightarrow 0$$

The Higher Criticism is asymptotically equivalent to the optimal test without knowledge of π_1 and/or μ

- If $r < \rho_{\text{HC}}(\beta)$ then for *any* test

$$\liminf_{m \rightarrow \infty} P_0(\text{Type I Error}) + P_1(\text{Type II Error}) \geq 1$$

Detectable region



$$\rho_{T_{\max}}(\beta) = (1 - \sqrt{1 - \beta})^2$$