Hypothesis testing: a review

Aldo Solari
Statistical Inference II
PhD in Economics and Statistics
University of Milano-Bicocca

Introduction

Deterministic proof by contradiction

1. Assume a proposition, the opposite of what you think about, i.e. the opposite conclusion of your theorem
2. Write down a sequence of logical steps/math
3. Derive a contradiction
4. Conclude that the proposition is false (which implies that the theorem is true)

Stochastic proof by contradiction

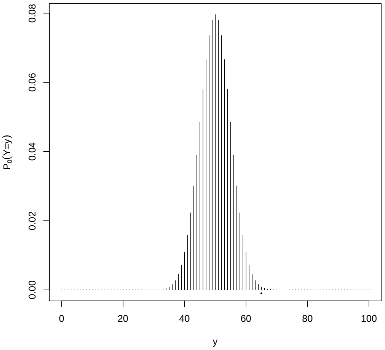
1. Set H_0 (the proposition)
2. Collect data (which is random)
3. Derive an apparent contradiction (i.e. if H_0 is true, then this data is very weird)
4. Hence we reject H_0; this is called a "discovery"

Hypothesis testing is stochastic because we might make errors: *Type I* (false discoveries) and *Type II* (missed discoveries)

Assume we have a coin and we conjecture that it is biased. In this case we can test

$$H_0 : \text{Coin is fair } (\pi = 1/2)$$
$$H_1 : \text{Coin is biased } (\pi \neq 1/2)$$

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



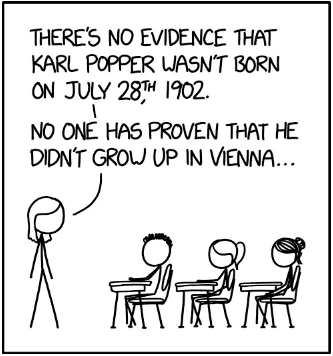
- Is this enough to reject H_0?
- To determine this we calculate a p-**value** associated with our observed data assuming the null hypothesis
- A p-value is "the probability of seeing what you saw - or something more extreme - given that H_0 is true"
- Small p-values imply an unexpected outcome, given that 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

Suppose that in $n = 10000$ trials we get $y = 5001$ heads and $n - y = 4999$ tails. Can we conclude that the coin is fair by testing $H_0 : \pi = 1/2$ against $H_1 : \pi \neq 1/2$?

```
Exact binomial test

data:  5001 and 10000
number of successes = 5001, number of trials
= 10000, p-value = 0.992
alternative hypothesis:
true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4902514 0.5099486
sample estimates:
probability of success
                0.5001
```

Lack of evidence to reject H_0 does not imply that H_0 is true.



source: xkcd

Suppose that we conjecture that the coin is fair.
What about testing

$$
\begin{aligned}
H_0 &: \text{Coin is biased } (\pi \neq 1/2) \\
H_1 &: \text{Coin is fair } (\pi = 1/2)
\end{aligned}
$$

What about this one?

$$
\begin{aligned}
H_0 &: \pi \in [0, 0.49] \cup [0.51, 1] \\
H_1 &: \pi \in (0.49, 0.51)
\end{aligned}
$$

Significance tests

Simple significance test

- Suppose available data y and a **null hypothesis** H_0 that **fully** specifies the distribution of Y
- Choose a **test statistic** $T = t(Y)$, large (or extreme) values of which indicate a departure from H_0
- Then if $t_{\text{obs}} = t(y)$ is the observed value of T we define

$$p_{\text{obs}} = \mathrm{P}_0(T \geq t_{\text{obs}})$$

where P_0 is the probability under H_0

p-value null distribution

- $p_{\text{obs}} = 1 - F_0(t_{\text{obs}})$, where $F_0(t) = \mathrm{P}_0(T \leq t)$ is the null cdf 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 $u \in (0, 1)$

$$
\begin{aligned}
\mathrm{P}_0(P \leq u) &= \mathrm{P}_0(1 - F_0(T) \leq u) \\
&= \mathrm{P}_0(1 - u \leq F_0(T)) \\
&= \mathrm{P}_0(F_0^{-1}(1 - u) \leq T) \\
&= 1 - F_0(F_0^{-1}(1 - u)) = u
\end{aligned}
$$

One- and two-sided tests

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

$$p_{\text{obs}}^{-} = \mathrm{P}_0(T \leq t_{\text{obs}}), \quad p_{\text{obs}}^{+} = \mathrm{P}_0(T \geq t_{\text{obs}})$$

- The p-value is

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

- Note that $P^{-} = 1 - P^{+}$ and $P^{+} \overset{H_0}{\sim} U(0,1)$. Then

$$Q = \min(1 - P^{+}, P^{+}) \overset{H_0}{\sim} U(0, 1/2)$$

thus $P = 2Q \overset{H_0}{\sim} U(0,1)$

Discrete null distribution

- Suppose we want to test $H_0 : \mu = 2$ by $T \sim \text{Poisson}(\mu)$ and we observe $t_{\text{obs}} = 3$
- $$p^{+}_{\text{obs}} = \text{P}_0(T \geq t_{\text{obs}}) = \sum_{t=t_{\text{obs}}}^{\infty} \frac{\mu^t e^{-\mu}}{t!}$$

  $$p^{-}_{\text{obs}} = \text{P}_0(T \leq t_{\text{obs}}) = \sum_{t=0}^{t_{\text{obs}}} \frac{\mu^t e^{-\mu}}{t!}$$
- 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}
- For $t_{\text{obs}} = 3$, $p_{\text{obs}} = 0.458 = \min(0.323, 0.857) + 0.135$

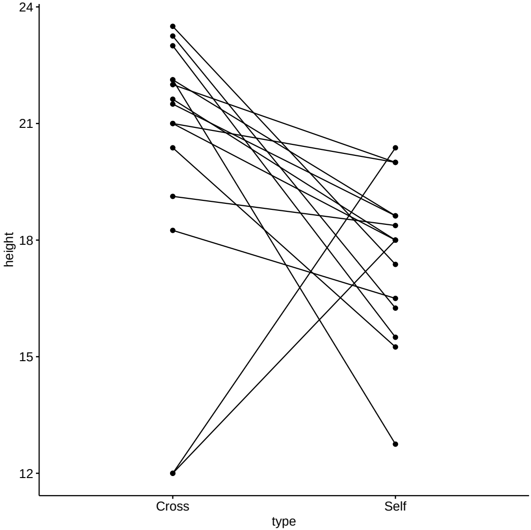
t	0	1	2	3	4	5
$\text{P}_0(T \geq t)$	1	0.865	0.594	0.323	0.143	0.053
$\text{P}_0(T \leq t)$	0.135	0.406	0.677	0.857	0.947	0.983

Example: sign test

- A random sample $Y_1, \ldots, Y_n$ arises from an unknown continuous distribution F
- The null hypothesis H_0 asserts that F is symmetric around 0, i.e. $H_0 : F(-y) + F(y) = 1$
- Under H_0, all points y and $-y$ have equal probability and

$$T = \sum_{i=1}^{n} \mathbb{1}\{Y_i > 0\} \overset{H_0}{\sim} \text{Binomial}(n, 1/2)$$

- Tests where the null hypotheses itself is formulated in terms of arbitrary distributions are called **nonparametric** or **distribution-free** tests



24
21
18
15
12
height
Cross
Self
type

```
binom.test(x=13, n=15, p=0.5, alternative="two.sided")

Exact binomial test

data:  13 and 15
number of successes = 13,
number of trials = 15,
p-value = 0.007385
alternative hypothesis:
true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5953973 0.9834241
sample estimates:
probability of success
             0.8666667
```

Example: adequacy of Poisson model

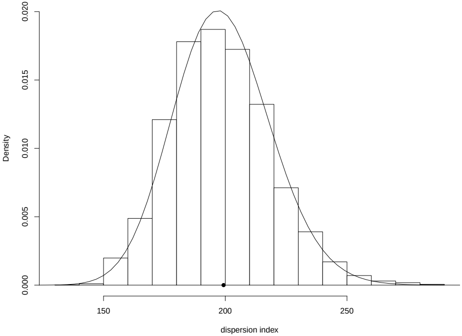
- Null hypothesis H_0: $Y_1, \ldots, Y_n$ i.i.d. Poisson(μ)
- The sufficient statistic is $S = \sum_{i=1}^n Y_i$, so we examine the conditional distribution of the data given $S = s$. This density is zero if $\sum_{i=1}^n y_i \neq s$ and is otherwise

$$\frac{s!}{\prod_{i=1}^n y_i!} \frac{1}{n^s}$$

  i.e., is a multinomial distribution with s trials each giving a response equally likely to fall in one of n cells
- The test statistic may be the dispersion index $\sum_{i=1}^n (Y_i - \bar{Y})^2 / \bar{Y} \overset{H_0}{\approx} \chi^2_{n-1}$ or the number of zeros

Example: von Bortkiewicz's horse-kicks data

Deaths	0	1	2	3	4
Frequency	109	65	22	3	1



Dispersion index = 199.3

exact p-value = 0.505 ($B = 5000$), approximated p-value = 0.48

Example: Kolmogorov-Smirnov test

- The null hypothesis H_0 asserts that the random sample $Y_1, \ldots, Y_n$ is from a known continuous distribution F_0
- We can compare F_0 with the empirical distribution function

$$\hat{F}(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Y_i \leq y\}$$

- A classic test for H_0 is based on the Kolmogorov-Smirnov statistic

$$T = \|\hat{F} - F_0\|_\infty = \sup_y |\hat{F}(y) - F_0(y)|$$

- Kolmogorov (1933, Giornale dell'Istituto Italiano degli Attuari) showed that under H_0 for any $c > 0$

$$\lim_{n \to \infty} \mathrm{P}\left(T > \frac{c}{\sqrt{n}}\right) = 2 \sum_{k=1}^{\infty} (-1)^{k+1} \exp(-2k^2 c^2)$$

- Often referred as **goodness-of-fit** test, but is actually testing for lack-of-fit

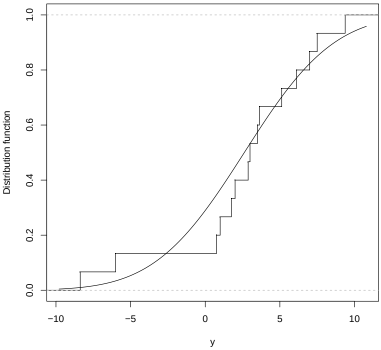
Example: Kolmogorov-Smirnov test (con'd)

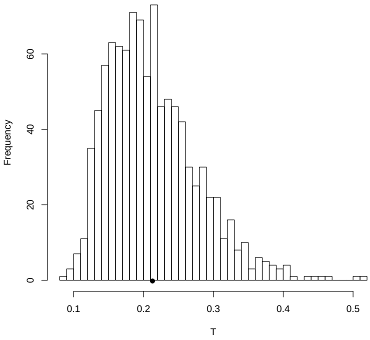
- We can avoid asymptotic approximations by using a Monte Carlo method
- To compute the p-value we can generate B independent sets of data from the null distribution F_0, calculating the corresponding statistics T^b and

$$p_{\text{obs}} = \frac{1 + \sum_{b=1}^{B} \mathbb{1}\{T^b \geq t_{\text{obs}}\}}{1 + B}$$

- If the parameters of F are determined from the data, the resulting test is only approximate

H_0: height differences are $N(\hat{\mu} = 2.6, \hat{\sigma}^2 = 4.7^2)$





p-value = 0.447 ($B = 1000$)

Likelihood-based tests

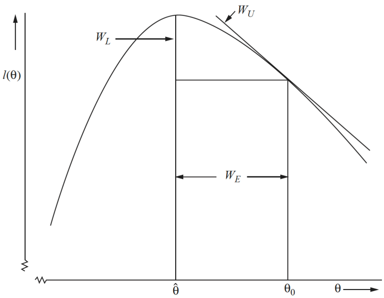


Figure 6.2. Three asymptotically equivalent ways, all based on the log likelihood function of testing null hypothesis $\theta = \theta_0$: W_E, horizontal distance; W_L vertical distance; W_U slope at null point.

$$\text{Wald } W_E = [\hat{\theta} - \theta_0]^2 i(\theta_0)$$

$$\text{Likelihood ratio } W_L = 2\{l(\hat{\theta}) - l(\theta_0)\}$$

$$\text{Score } W_U = [U(\theta_0; Y)]^2 i^{-1}(\theta_0)$$

Example: Student t test

- Let $Y_1, \ldots, Y_n$ be a normal random sample with mean μ and variance σ^2
- Suppose that $H_0 : \mu = \mu_0$
- log likelihood for $y_1, \ldots, y_n$ is

$$l(\mu, \sigma^2) = -\frac{1}{2}\left\{n \log \sigma^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}$$

- The likelihood ratio statistic is

$$W_L = 2\{\max_{\mu,\sigma^2} l(\mu, \sigma^2) - \max_{\sigma^2} l(\mu_0, \sigma^2)\} = n \log\left(1 + \frac{T^2}{n-1}\right)$$

where $T = (\bar{Y} - \mu_0)/(S^2/n)^{1/2} \overset{H_0}{\sim} t_{n-1}$

Example: Permutation two-sample test

- Let $Y_1, \dots, Y_k \overset{i.i.d.}{\sim} F$ and $Y_{k+1}, \dots, Y_n \overset{i.i.d.}{\sim} G$ be independent random samples of size k and $n-k$
- Consider the null hypothesis $H_0 : F = G$
- Under H_0, the sufficient statistic is the set of order statistics of the combined set of observations and all $n!$ permutations of the data are equally likely, i.e.

$$(Y_1, \dots, Y_n) \overset{d}{=} (Y_{\pi(1)}, \dots, Y_{\pi(n)}) \quad \forall \pi$$

- Permutation p-value

$$\mathrm{P}_0(T \geq t_{\text{obs}} | Y_{(1)}, \dots, Y_{(n)}) = \frac{1}{n!} \sum_{\pi} \mathbb{1}\{T^{\pi} \geq t_{\text{obs}}\}$$

- In **randomization tests**, the basis of the procedure is the randomization used in allocating the units to the groups

Relation with two-decision problem

- In the treatment of testing as a two-decision problem, the choice lies between **rejecting** or **not rejecting** the null hypothesis
- In this we fix the probability of rejecting H_0 when it is true (probability of type I error) at **level** α, aiming to maximize the **power**, i.e. the probability of rejecting H_0 when false ($1-$ probability of type II error)
- This amounts to setting in advance a threshold α for p_{obs}
- It demands the explicit formulation of the **alternative hypothesis** H_1

Hypothesis testing

Hypothesis testing

- The decision procedure is called the **test** of H_0 against H_1
- Suppose we have data Y distributed according to P_θ with $\theta \in \Theta$
- About θ we formulate the null hypotheses $H_0 : \theta \in \Theta_0$ with $\Theta_0 \subseteq \Theta$. The alternative hypothesis is $H_1 : \theta \in \Theta_1$ with (usually) $\Theta_1 = \Theta \setminus \Theta_0$.
- A hypothesis that completely determines the distribution of Y is called **simple**; otherwise is **composite**
- A test $\phi = \phi(Y)$ assigns to each possible value y one of these two decisions

$$\phi : \mathcal{Y} \mapsto \{0, 1\}$$

  where 1 denotes the decision of rejecting H_0 and 0 denotes the decision of not rejecting H_0, and thereby partition the sample space $\mathcal{Y}$ into two complementary regions $\mathcal{Y}_0$ and $\mathcal{Y}_1$

Size and power function

- It is required to bound the probability of Type I error at α

$$\mathrm{P}_\theta(\phi = 1) \leq \alpha \quad \forall\ \theta \in \Theta_0$$

  where

$$\sup_{\theta \in \Theta_0} \mathrm{P}_\theta(\phi = 1)$$

  is the **size** of the test
- Subject to this condition, it is desired to maximize the power

$$\mathrm{P}_\theta(\phi = 1) \quad \theta \in \Theta_1$$

- Considered as a function of θ for all $\theta \in \Theta$, this probability is called the **power function** of the test and is denoted by $\beta(\theta)$

p-value

- Usually for varying α, the rejection regions $\mathcal{Y}_1(\alpha)$ and $\mathcal{Y}_1(\tilde{\alpha})$ are nested in the sense that

$$\mathcal{Y}_1(\alpha) \subseteq \mathcal{Y}_1(\tilde{\alpha}) \quad \text{if } \alpha \leq \tilde{\alpha}$$

- When this is the case, the p-value is defined as the smallest significance level at which the hypothesis would be rejected for the given observation:

$$p_{\text{obs}} = \inf\{\alpha \in (0, 1) : y \in \mathcal{Y}_1(\alpha)\}$$

Neyman-Pearson lemma

- Let f_0 and f_1 denote the probability densities of Y specified under H_0 and H_1, respectively, i.e. $H_0 : f = f_0$ vs $H_1 : f = f_1$
- The Neyman-Pearson lemma states that the **most powerful test** of size α has critical region

$$\mathcal{Y}_1 = \left\{ y \in \mathcal{Y} : \frac{f_1(y)}{f_0(y)} \geq t_\alpha \right\}$$

determined by the likelihood ratio

Example: UMP test

- Let $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} N(\mu, 1)$, and suppose that we are testing $\mu \leq \mu_0$ against $\mu > \mu_0$. Suppose we reject the null if $\bar{Y}$ exceed some constant t_α.
- The size of this test is

$$\begin{aligned}
\sup_{\mu \leq \mu_0} \mathrm{P}_\mu(\bar{Y} \geq t_\alpha) &= \mathrm{P}_{\mu_0}(\bar{Y} \geq t_\alpha) \\
&= \mathrm{P}_{\mu_0}\left(\frac{\bar{Y} - \mu_0}{\sqrt{1/n}} \geq \frac{t_\alpha - \mu_0}{\sqrt{1/n}}\right) \\
&= \Phi\left(\frac{\mu_0 - t_\alpha}{\sqrt{1/n}}\right)
\end{aligned}$$

- For a test of size α, we must choose $t_\alpha = \mu_0 + \frac{z_{1-\alpha}}{\sqrt{n}}$ and the critical region is

$$\left\{(y_1, \ldots, y_n) : \bar{y} \geq \mu_0 + \frac{z_{1-\alpha}}{\sqrt{n}}\right\}$$

Example: UMP test (cont'd)

- The power function of the test is

$$\beta(\mu_1) = \mathrm{P}_{\mu_1}(\bar{Y} \geq t_\alpha) = \Phi(z_\alpha + \delta)$$

  where $\delta = \sqrt{n}(\mu_1 - \mu_0)$
- The likelihood ratio for testing $\mu = \mu_0$ against $\mu = \mu_1$ is

$$\frac{f_1(Y)}{f_0(Y)} = \exp\left[\frac{1}{2}(2n\bar{Y}(\mu_1 - \mu_0) - \mu_1^2 + \mu_0^2)\right]$$

- If $\mu_1 > \mu_0$, this is monotone increasing in $\bar{Y}$, and so the critical region rejects H_0 when $\bar{Y} \geq t_\alpha$
- It follows that this test is most powerful for any $\mu_1 > \mu_0$ and so is **uniformly most powerful** (UMP)

Example: UMPU test

- Likewise, the test defined by the critical region

$$\left\{(y_1, \ldots, y_n) : \bar{y} \leq \mu_0 + \frac{z_\alpha}{\sqrt{n}}\right\}$$

  is UMP for testing $\mu \geq \mu_0$ against $\mu \leq \mu_0$
- Suppose now that we wish to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. The critical region

$$\left\{(y_1, \ldots, y_n) : \bar{y} \leq \mu_0 + \frac{z_\alpha}{\sqrt{n}}\right\} \cup \left\{(y_1, \ldots, y_n) : \bar{y} \geq \mu_0 + \frac{z_{1-\alpha}}{\sqrt{n}}\right\}$$

  has size 2α, and no uniformly more powerful test exists for the two-sided alternative. It can be proved that is UMPU
- A test ϕ of $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is **unbiased** of size α if $\sup_{\theta \in \Theta_0} \mathrm{P}_\theta(\phi = 1) = \alpha$ and $\mathrm{P}_\theta(\phi = 1) \geq \alpha$ for all $\theta \in \Theta_1$
- A test which is uniformly most powerful amongst the class of all unbiased tests is **uniformly most powerful unbiased**

Example: Locally most powerful test

- Local alternative where $f_0(y) = f(y;\theta_0)$ and $f_1(y) = f(y;\theta_1)$ with $\theta_1 = \theta_0 + \epsilon$ for small ϵ
-

$$
\begin{aligned}
\frac{f_1(Y)}{f_0(Y)} &= \frac{f(Y;\theta_0+\epsilon)}{f(Y;\theta_0)} \\
&= \frac{1}{f(Y;\theta_0)}\left\{f(Y;\theta_0) + \epsilon\frac{df(Y;\theta_0)}{d\theta_0} + \ldots\right\} \\
&\approx 1 + \epsilon U(\theta_0)
\end{aligned}
$$

- A locally most powerful critical region has form

$$\{(y_1,\ldots,y_n) : u(\theta_0) \geq i(\theta_0)^{1/2} z_{1-\alpha}\}$$

where $i(\theta_0)$ is the Fisher information

Example: location parameter of a Cauchy distribution

- Let $Y_1, \ldots, Y_n$ be i.i.d. in the Cauchy distribution

$$\frac{1}{\pi[1 + (y - \theta)^2]}$$

- For the null hypothesis $H_0 : \theta = \theta_0$ the score from Y_1 is

$$U_1(\theta_0) = \frac{2(Y_1 - \theta_0)}{1 + (Y_1 - \theta_0)^2}$$

  and the information from a single observation is

$$i_1(\theta_0) = \frac{1}{2}$$

- The test statistic is thus

$$U(\theta_0) = 2 \sum_{i=1}^{n} \frac{(Y_i - \theta_0)}{1 + (Y_i - \theta_0)^2}$$

  and under H_0 has zero mean and variance $n/2$

Example: UMPI test

- Let $Y_1, \ldots, Y_n$ be a random sample from the m-variate normal distribution $N_m(\mu, \Sigma)$, and suppose that we are testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.
- If Σ is unknown and $n > m$, we can use the Hotelling T^2 statistic

$$T^2 = n(\bar{Y} - \mu_0)' S^{-1} (\bar{Y} - \mu_0)$$

  where $S = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})(Y_i - \bar{Y})'$
- Under H_0, T^2 follows a Hotelling's T-squared distribution

$$T^2_{m,n-1} = \frac{m(n-1)}{n-m} F_{m,n-m}$$

  where $F_{m,n-m}$ is the F-distribution with parameters m and $n - m$

Example: UMPI test (cont'd)

- No UMP test exists for this problem. It can be proved that the Hotelling T^2 test is the most powerful test in the class of tests that are invariate to full rank linear transformations (UMPI)
- The T^2 statistic is invariant to full rank linear transformations

  $$X = AY + b$$

  with $\underset{m\times m}{A}$ non-singular
- The Hotelling T^2 statistic is a generalization of Student t statistic, i.e. for $m = 1$, $T^2 = (t)^2$

Example: UMP test

- Let $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} N(\mu, 1)$,
- Suppose we want to test $H_0 : \mu \in (-\infty, -\Delta] \cup [\Delta, \infty)$ against $H_1 : \mu \in (-\Delta, \Delta)$ for some pre-specified $\Delta > 0$
- Consider the test statistic

$$T = n\bar{Y}^2 \sim \chi_1^2(n\mu^2)$$

  which rejects for small values, where $\chi_\nu^2(\lambda)$ is a non-central Chi-squared distribution with ν degree of freedom and noncentrality parameter λ

Example: UMP test (cont'd)

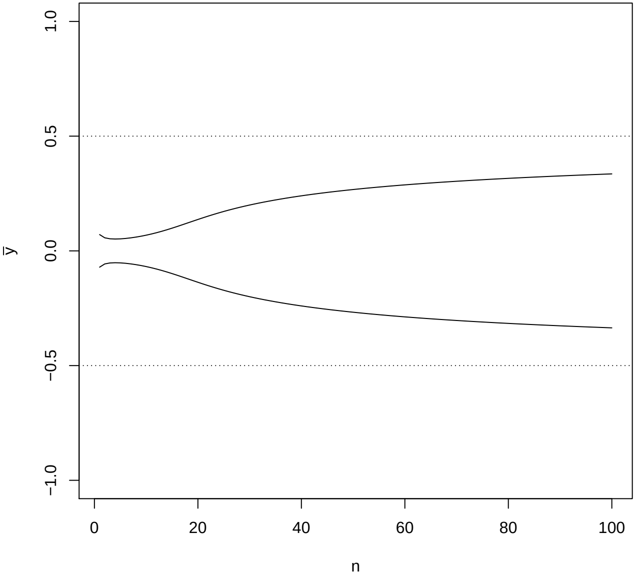
- Since

$$\sup_{\mu \in (-\infty, -\Delta] \cup [\Delta, \infty)} \mathrm{P}_\mu(T \leq t_\alpha) = \mathrm{P}(\chi^2_1(n\Delta^2) \leq t_\alpha)$$

  the critical region of size α is given by

$$\mathcal{Y}_1 = \{(y_1, \ldots, y_n) : -\sqrt{t_\alpha/n} \leq \bar{y} \leq \sqrt{t_\alpha/n}\}$$

  where t_α is the α quantile of $\chi^2_1(n\Delta^2)$
- It can be proved that this test is UMP



1.0
0.5
0.0
−0.5
−1.0
$\overline{y}$
0
20
40
60
80
100
n

Relation with interval estimation

Essentially confidence intervals, or more generally confidence sets, can be produced by testing every possible value θ in Θ and taking all those values not ‘rejected’ at level α, say, to produce a $1 - \alpha$ level interval or region

Confidence intervals

Confidence intervals

- If the density of Y depends on a scalar parameter θ, we define an upper bound for θ at confidence level $1-\alpha$ to be a function $\bar{\theta}_\alpha = \bar{\theta}_\alpha(Y)$ such that

$$\mathrm{P}_\theta(\theta \leq \bar{\theta}_\alpha) \geq 1-\alpha \qquad \forall\, \theta \in \Theta$$

- Lower confidence bounds may be defined analogously
- An equi-tailed $(1-2\alpha)$ confidence interval for θ is $[\underline{\theta}_\alpha, \bar{\theta}_\alpha]$

Duality between tests and confidence intervals

For each $\theta_0 \in \Theta$, let $\mathcal{Y}_0(\theta_0)$ be the acceptance region of a test of size α for testing $\theta = \theta_0$

Theorem

The set of values of θ *not rejected by the test*

$$S(Y) = \{\theta \in \Theta : Y \in \mathcal{Y}_0(\theta)\}$$

contains the true parameter with probability at least $1 - \alpha$

Proof.

By definition of $S(Y)$, $\theta \in S(Y)$ if and only if $Y \in \mathcal{Y}_0(\theta)$, and hence

$$\mathrm{P}_\theta(\theta \in S(Y)) = \mathrm{P}_\theta(Y \in \mathcal{Y}_0(\theta)) \geq 1 - \alpha \qquad \forall\, \theta \in \Theta$$

$\square$

Example: ratio of normal means

- Given two independent sets of random variables from normal distributions of unknown means μ_1 and μ_2 and variance 1
- We first reduce by sufficiency to the sample means $\bar{y}_1$, $\bar{y}_2$
- Suppose that the parameter of interest is $\theta = \mu_2/\mu_1$. Consider the null hypothesis $H_0 : \theta = \theta_0$

$$\frac{\bar{Y}_2 - \theta_0 \bar{Y}_1}{\sqrt{1/n_2 + \theta_0/n_1}} \overset{H_0}{\sim} N(0,1)$$

- We now form a $1-\alpha$ level confidence region by taking all those values of θ_0 that would not be rejected at level α in this test

$$\left\{ \theta \in \mathbb{R} : \frac{(\bar{Y}_2 - \theta \bar{Y}_1)^2}{1/n_2 + \theta/n_1} \leq c_{1-\alpha} \right\}$$

  where $c_{1-\alpha}$ is the $1-\alpha$ quantile of χ^2_1
- Thus we find the limits for θ as the roots of a quadratic equation
- If there are no real roots, all values of θ are consistent with the data at the level in question