

3. Hypothesis Testing

Pure significance tests

Data $\mathbf{x} = (x_1, \dots, x_n)$ from $f(\mathbf{x}, \theta)$

Hypothesis H_0 : restricts $f(\mathbf{x}, \theta)$

Are the data consistent with H_0 ?

H_0 is called the *null hypothesis*

simple if it completely specifies the density of $\mathbf{x}$

composite otherwise

pure significance test: a means of examining whether the data are consistent with H_0 where the only distribution of the data that is explicitly formulated is that under H_0

test statistic $T = t(\mathbf{X})$

Suppose: the larger $t(\mathbf{x})$, the more inconsistent the data with H_0

For simple H_0 , the *p-value* of $\mathbf{x}$ is

$$p = \mathbf{P}(T \geq t(\mathbf{x}) \mid H_0).$$

small $p \rightarrow$ more inconsistency with H_0

For composite H_0 :

If S is sufficient for θ then the distribution of $\mathbf{X}$ conditional on S is independent of θ ; when $S = s$, the *p-value* of $\mathbf{x}$ is

$$p = \mathbf{P}(T \geq t(\mathbf{x}) \mid H_0; S = s).$$

Example: Dispersion of Poisson distribution

H_0 : $X_1, \dots, X_n$ i.i.d. $\sim \text{Poisson}(\mu)$, with unknown μ

Under H_0 , $\text{Var}(X_i) = \mathbf{E}(X_i) = \mu$ and so would expect

$T = t(\mathbf{X}) = S^2/\overline{X}$ close to 1. The statistic T is also

called the *dispersion index*.

Suspect that the X_i 's may be over-dispersed, that is,

$\text{Var}(X_i) > \mathbf{E}(X_i)$: discrepancy with H_0 would corre-

spond to large T

Recall that $\overline{X}$ is sufficient for the Poisson distribution; the p -value is then $p = P(S^2/\overline{X} \geq t(\mathbf{x}) | \overline{X} = \bar{x}; H_0)$ under the Poisson hypothesis, which makes p independent of the unknown μ

Given $\overline{X} = \bar{x}$ and H_0 we have that

$$S^2/\overline{X} \approx \chi_{n-1}^2/(n-1)$$

(see Chapter 5 later) and so

$$p \approx \mathbf{P}(\chi_{n-1}^2/(n-1) \geq t(\mathbf{x})).$$

Possible alternatives to H_0 guide the choice and interpretation of T

What is a "best" test?

Simple null and alternative hypotheses

General: random sample $X_1, \dots, X_n$ from $f(x; \theta)$

null hypothesis $H_0 : \theta \in \Theta_0$

alternative hypothesis $H_1 : \theta \in \Theta_1$

where $\Theta_1 = \Theta \setminus \Theta_0$; Θ denoting the whole parameter space

Choose *rejection region* or *critical region* R :

$$\text{reject } H_0 \iff \mathbf{X} \in R$$

Suppose $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$

Type I error: reject H_0 when it is true

$$\alpha = \mathbf{P}(\text{reject } H_0 \mid H_0),$$

also known as *size* of the test

Type II error: accept H_0 when it is false

$$\beta = \mathbf{P}(\text{accept } H_0 \mid H_1).$$

the *power* of the test is $1 - \beta = \mathbf{P}(\text{accept } H_1 \mid H_1)$.

Usually: fix α (e.g., $\alpha = 0.05, 0.01$, etc.), look for a test which minimizes β :

most powerful or *best* test of size α

Intuitively: reject H_0 in favour of H_1 if likelihood of θ_1 is much larger than likelihood of θ_0 , given the data.

Neyman-Pearson Lemma: (see, e.g., Casella and Berger, p.366) The most powerful test of H_0 versus H_1 has rejection region

$$R = \left\{ \mathbf{x} : \frac{L(\theta_1; \mathbf{x})}{L(\theta_0; \mathbf{x})} \geq k_\alpha \right\}$$

where the constant k_α is chosen so that

$$P(\mathbf{X} \in R \mid H_0) = \alpha.$$

This test is called the *likelihood ratio (LR) test*.

Often we simplify the condition

$$L(\theta_1; \mathbf{x})/L(\theta_0; \mathbf{x}) \geq k_\alpha$$

to

$$t(\mathbf{x}) \geq c_\alpha,$$

for some constant c_α and some statistic $t(\mathbf{x})$; determine

c_α from the equation

$$\mathbf{P}(T \geq c_\alpha \mid H_0) = \alpha,$$

where $T = t(\mathbf{X})$; then the test is “reject H_0 if and only if

$T \geq c_\alpha$ ”. For data $\mathbf{x}$ the p -value is $p = \mathbf{P}(T \geq t(\mathbf{x}) \mid H_0)$.

Example: Normal means, one-sided

$$X_1, \dots, X_n \sim N(\mu, \sigma^2), \sigma^2 \text{ known}$$

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu = \mu_1, \text{ with } \mu_1 > \mu_0$$

Then

$$\frac{L(\mu_1; \mathbf{x})}{L(\mu_0; \mathbf{x})} \geq k \Leftrightarrow \ell(\mu_1; \mathbf{x}) - \ell(\mu_0; \mathbf{x}) \geq \log k$$

$$\Leftrightarrow - \sum [(x_i - \mu_1)^2 - (x_i - \mu_0)^2] \geq 2\sigma^2 \log k$$

$$\Leftrightarrow (\mu_1 - \mu_0)\bar{x} \geq k'$$

$$\Leftrightarrow \bar{x} \geq c \quad (\text{since } \mu_1 > \mu_0),$$

where k' , c are constants, indept. of $\mathbf{x}$.

Choose $t(\mathbf{x}) = \bar{x}$, and for size α test choose c so that

$$\mathbf{P}(\bar{X} \geq c \mid H_0) = \alpha;$$

equivalently, such that

$$\mathbf{P} \left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{c - \mu_0}{\sigma/\sqrt{n}} \mid H_0 \right) = \alpha.$$

Hence want

$$(c - \mu_0)/(\sigma/\sqrt{n}) = z_{1-\alpha}$$

(where $\Phi(z_{1-\alpha}) = 1 - \alpha$ with Φ standard normal c.d.f.),

i.e.

$$c = \mu_0 + \sigma z_{1-\alpha}/\sqrt{n}.$$

So our test

“reject H_0 if and only if $\bar{X} \geq c$ ” becomes

“reject H_0 if and only if $\bar{X} \geq \mu_0 + \sigma z_{1-\alpha}/\sqrt{n}$ ”

This is the most powerful test of H_0 versus H_1 at level α .

Recall the notation for standard normal quantiles: If

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

$$\mathbf{P}(Z \leq z_\alpha) = \alpha$$

and

$$\mathbf{P}(Z \geq z(\alpha)) = \alpha,$$

and note that $z(\alpha) = z_{1-\alpha}$. Thus

$$\mathbf{P}(Z \geq z_{1-\alpha}) = 1 - (1 - \alpha) = \alpha.$$

**Example: Bernoulli, probability of success,
one-sided**

$X_1, \dots, X_n$ i.i.d. Bernoulli(θ) then

$$L(\theta) = \theta^r (1 - \theta)^{n-r}$$

where $r = \sum x_i$; test $H_0 : \theta = \theta_0$ against

$H_1 : \theta = \theta_1$, where $\theta_1 > \theta_0$. Now $\theta_1/\theta_0 > 1$, $(1-\theta_1)/(1-$

$\theta_0) < 1$, and

$$\frac{L(\theta_1; \mathbf{x})}{L(\theta_0; \mathbf{x})} = \left(\frac{\theta_1}{\theta_0}\right)^r \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n-r}$$

and so $L(\theta_1; \mathbf{x})/L(\theta_0; \mathbf{x}) \geq k_\alpha \iff r \geq r_\alpha$.

So the best test rejects H_0 for large r . For any given critical value r_c ,

$$\alpha = \sum_{j=r_c}^n \binom{n}{j} \theta_0^j (1 - \theta_0)^{n-j}$$

gives the p -value if we set $r_c = r(\mathbf{x}) = \sum x_i$, the observed value.

Note: The distribution is discrete, so we may not be able to achieve a level α test exactly (unless we use additional randomization).

For example, if $R \sim \text{Binomial}(10, 0.5)$, then $P(R \geq 9) = 0.011$, and $P(R \geq 8) = 0.055$, so there is no c such that $P(R \geq c) = 0.05$. A solution is to randomize: If $R \geq 9$ reject the null hypothesis, if $R \leq 7$ accept the null hypothesis, and if $R = 8$ flip a (biased) coin to achieve the exact level of 0.05.

Composite alternative hypotheses

Suppose θ scalar

$$H_0 : \theta = \theta_0$$

one-sided alternative hypotheses

$$H_1^- : \theta < \theta_0$$

$$H_1^+ : \theta > \theta_0$$

two-sided alternative $H_1 : \theta \neq \theta_0$

The *power function* of a test

$$\text{power}(\theta) = \mathbf{P}(\mathbf{X} \in R \mid \theta)$$

probability of rejecting H_0 as a function of the true value of the parameter θ ; depends on α , the size of the test.

Main uses: comparing alternative tests, choosing sample size.

A test of size α is *uniformly most powerful* (*UMP*) if its power function is such that

$$\text{power}(\theta) \geq \text{power}'(\theta)$$

for all $\theta \in \Theta_1$, where $\text{power}'(\theta)$ is the power function of any other size- α test.

Consider: H_0 against H_1^+

For exponential family problems: usually for any $\theta_1 > \theta_0$

the rejection region of the LR test is independent of θ_1 .

At the same time, the test is most powerful for every

single θ_1 which is larger than θ_0 . Hence the test derived

for one such value of θ_1 is UMP for H_0 versus H_1^+ .

**Example: normal mean, composite one-sided
alternative**

$X_1, \dots, X_n \sim N(\mu, \sigma^2)$ i.i.d., σ^2 known

$$H_0 : \mu = \mu_0$$

$$H_1^+ : \mu > \mu_0$$

First pick arbitrary $\mu_1 > \mu_0$

the most powerful test of $\mu = \mu_0$ versus $\mu = \mu_1$ has a
rejection region of the form

$$\bar{X} \geq \mu_0 + \sigma z_{1-\alpha} / \sqrt{n}$$

for a test of size α .

This rejection region is independent of μ_1 , hence it is

UMP for H_0 versus H_1^+ . The power of the test is

$$\text{power}(\mu)$$

$$= \mathbf{P}(\bar{X} \geq \mu_0 + \sigma z_{1-\alpha}/\sqrt{n} \mid \mu)$$

$$= \mathbf{P}(\bar{X} \geq \mu_0 + \sigma z_{1-\alpha}/\sqrt{n} \mid \bar{X} \sim N(\mu, \sigma^2/n))$$

$$= \mathbf{P}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{1-\alpha} \mid \bar{X} \sim N(\mu, \sigma^2/n)\right)$$

$$= \mathbf{P}\left(Z \geq z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}} \mid Z \sim N(0, 1)\right)$$

$$= 1 - \Phi\left(z_{1-\alpha} - (\mu - \mu_0)\sqrt{n}/\sigma\right).$$

The power increases from 0 up to α at $\mu = \mu_0$ and then

to 1 as μ increases. The power increases as α increases.

Sample size calculation in the Normal example

Suppose want to be near-certain to reject H_0 when $\mu = \mu_0 + \delta$, say, and have size 0.05. Suppose we want to fix n to force $\text{power}(\mu) = 0.99$ at $\mu = \mu_0 + \delta$:

$$0.99 = 1 - \Phi(1.645 - \delta\sqrt{n}/\sigma)$$

so that $0.01 = \Phi(1.645 - \delta\sqrt{n}/\sigma)$. Solving this equation (use tables) gives $-2.326 = 1.645 - \delta\sqrt{n}/\sigma$, i.e.

$$n = \sigma^2(1.645 + 2.326)^2/\delta^2$$

is the required sample size.

UMP tests are not always available. If not, options include:

1. Wald test
2. locally most powerful test (score test)
3. generalized likelihood ratio test.

Wald test

The Wald test is directly based on the asymptotic normality of the m.l.e. $\hat{\theta} = \hat{\theta}_n$, often $\hat{\theta} \approx \mathcal{N}(\theta, I_n^{-1}(\theta))$

if θ is the true parameter. Recall: In random sample,

$$I_n(\theta) = nI_1(\theta).$$

Also it is often true that asymptotically, we may replace θ by $\hat{\theta}$ in the Fisher information,

$$\hat{\theta} \approx \mathcal{N}(\theta, I_n^{-1}(\hat{\theta}))$$

So we can construct a test based on

$$W = \sqrt{I_n(\hat{\theta})}(\hat{\theta} - \theta_0) \approx \mathcal{N}(0, 1).$$

If θ is scalar, squaring gives

$$W^2 \approx \chi_1^2,$$

so equivalently we could use a chi-square test.

For higher-dimensional θ we can base a test on the quadratic form

$$(\hat{\theta} - \theta_0)^T I_n(\hat{\theta})(\hat{\theta} - \theta_0)$$

which is approximately chi-square distributed in large samples.

If we would like to test $H_0 : g(\theta) = 0$, where g is a (scalar) differentiable function, then the delta method gives as test statistic

$$W = g(\hat{\theta})\{G(\hat{\theta})(I_n(\hat{\theta}))^{-1}G(\hat{\theta})^T\}^{-1}g(\hat{\theta}),$$

where $G(\theta) = \frac{\partial g(\theta)}{\partial \theta}^T$.

Advantage of the Wald test:

We do not need to evaluate the likelihood under the null hypothesis, which can be awkward if the null hypothesis contains a number of restrictions on a multidimensional parameter. All we need is (an approximation) of $\hat{\theta}$, the maximum-likelihood-estimator.

Disadvantage: next slide...

Example: Non-invariance of the Wald test

Suppose that $\hat{\theta}$ is scalar and approximately $\mathcal{N}(\theta, I_n(\theta)^{-1})$ -distributed, then for testing $H_0 : \theta = 0$ the Wald statistic becomes

$$\hat{\theta} \sqrt{I_n(\hat{\theta})},$$

which would be approximately standard normal.

If instead we tested $H_0 : \theta^3 = 0$, then the delta method with $g(\theta) = \theta^3$, so that $g'(\theta) = 3\theta^2$, gives

$$\text{Var}(g(\hat{\theta})) \approx 9\hat{\theta}^4(I_n(\hat{\theta}))^{-1}$$

and as Wald statistic

$$\frac{\hat{\theta}}{3}\sqrt{I_n(\hat{\theta})},$$

which would again be approximately standard normal,

but the p -values for a finite sample may be quite different!

Locally most powerful test (Score test)

Test $H_0 : \theta = \theta_0$ against

$H_1 : \theta = \theta_0 + \delta$ for some small $\delta > 0$

The most powerful test has a rejection region of the form

$$\ell(\theta_0 + \delta) - \ell(\theta_0) \geq k.$$

Taylor expansion gives

$$\ell(\theta_0 + \delta) \approx \ell(\theta_0) + \delta \frac{\partial \ell(\theta_0)}{\partial \theta}$$

i.e.

$$\ell(\theta_0 + \delta) - \ell(\theta_0) \approx \delta \frac{\partial \ell(\theta_0)}{\partial \theta}.$$

So a locally most powerful (LMP) test has as rejection region

$$R = \left\{ \mathbf{x} : \frac{\partial \ell(\theta_0)}{\partial \theta} \geq k_\alpha \right\}.$$

This is also called the *(Rao) score test*: $\partial \ell / \partial \theta$ is known as the *score function*.

Under certain regularity conditions,

$$\mathbb{E}_\theta \left[\frac{\partial \ell}{\partial \theta} \right] = 0, \quad \text{Var}_\theta \left[\frac{\partial \ell}{\partial \theta} \right] = I_n(\theta).$$

As ℓ is usually a sum of independent components, so is $\partial \ell(\theta_0) / \partial \theta$, and the CLT (Central Limit Theorem) can be applied.

Example: Cauchy parameter

$X_1, \dots, X_n$ random sample from Cauchy (θ), having density

$$f(x; \theta) = [\pi(1 + (x - \theta)^2)]^{-1} \quad \text{for } -\infty < x < \infty,$$

Test $H_0 : \theta = \theta_0$ against $H_1^+ : \theta > \theta_0$. Then

$$\frac{\partial \ell(\theta_0; \mathbf{x})}{\partial \theta} = 2 \sum \left\{ \frac{x_i - \theta_0}{1 + (x_i - \theta_0)^2} \right\}.$$

Fact: Under H_0 , the expression $\partial \ell(\theta_0; \mathbf{X}) / \partial \theta$ has mean 0, variance $I_n(\theta_0) = n/2$.

The CLT applies, $\partial\ell(\theta_0; \mathbf{X})/\partial\theta \approx \mathcal{N}(0, n/2)$ under H_0 ,

so for the LMP test,

$$P(\mathcal{N}(0, n/2) \geq k_\alpha) = P\left(\mathcal{N}(0, 1) \geq k_\alpha \sqrt{\frac{2}{n}}\right) \approx \alpha$$

gives $k_\alpha \approx z_{1-\alpha} \sqrt{n/2}$, and as rejection region with ap-

proximate size α

$$R = \left\{ \mathbf{x} : 2 \sum \left(\frac{x_i - \theta_0}{1 + (x_i - \theta_0)^2} \right) > \sqrt{\frac{n}{2}} z_{1-\alpha} \right\}.$$

Advantage: We only need the likelihood under the null hypothesis.

Also not generally invariant under reparametrisation.

Multidimensional: Let $U = \partial\ell/\partial\theta$ be the score function, then the score statistic is

$$U^T I_n(\theta_0)^{-1} U.$$

Generalised likelihood ratio (LR) test

Test $H_0 : \theta = \theta_0$ against $H_1^+ : \theta > \theta_0$; use as rejection region

$$R = \left\{ \mathbf{x} : \frac{\max_{\theta \geq \theta_0} L(\theta; \mathbf{x})}{L(\theta_0; \mathbf{x})} \geq k_\alpha \right\}.$$

If L has one mode, at the m.l.e. $\hat{\theta}$, then the likelihood ratio in the definition of R is either 1, if $\hat{\theta} \leq \theta_0$, or $L(\hat{\theta}; \mathbf{x})/L(\theta_0; \mathbf{x})$, if $\hat{\theta} > \theta_0$.

Similar for H_1^- with fairly obvious changes of signs and directions of inequalities

The generalised LRT is invariant to a change in parametrisation.

Two-sided tests

Test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. If the one-sided tests of size α have symmetric rejection regions

$$R^+ = \{\mathbf{x} : t > c\} \quad \text{and} \quad R^- = \{\mathbf{x} : t < -c\},$$

then a two-sided test (of size 2α) is to take the rejection region to

$$R = \{\mathbf{x} : |t| > c\};$$

this test has as p -value $p = \mathbf{P}(|t(\mathbf{X})| \geq t \mid H_0)$.

The two-sided (generalised) LR test uses

$$\begin{aligned} T &= 2 \log \left[\frac{\max_{\theta} L(\theta; \mathbf{X})}{L(\theta_0; \mathbf{X})} \right] \\ &= 2 \log \left[\frac{L(\hat{\theta}; \mathbf{X})}{L(\theta_0; \mathbf{X})} \right] \end{aligned}$$

and rejects H_0 for large T .

Fact: $T \approx \chi_1^2$ under H_0 (later).

Where possible, the exact distribution of T or of a statistic equivalent to T should be used.

If θ is a vector: there is no such thing as a one-sided alternative hypothesis. For the alternative $\theta \neq \theta_0$ we use a LR test based on

$$T = 2 \log \left[\frac{L(\hat{\theta}; \mathbf{X})}{L(\theta_0; \mathbf{X})} \right].$$

Under H_0 , $T \approx \chi_p^2$ where $p = \text{dimension of } \theta$ (see Chapter 5).

For the score test we use as statistic

$$\ell'(\theta_0)^T [I_n(\theta_0)]^{-1} \ell'(\theta_0),$$

where $I_n(\theta)$ is the expected Fisher information matrix:

$$[I_n(\theta)]_{jk} = \mathbf{E}[-\partial^2 \ell / \partial \theta_j \partial \theta_k].$$

If the CLT applies to the score function, then this quadratic form is again approximately χ_p^2 under H_0 (see Chapter 5).

Example: Pearson's Chi-square statistic.

We have a random sample of size n , with p categories;

$P(X_j = i) = \pi_i$, for $i = 1, \dots, p$, $j = 1, \dots, n$. As

$\sum \pi_i = 1$, we take $\theta = (\pi_1, \dots, \pi_{p-1})$. The likelihood

function is then

$$\prod \pi_i^{n_i}$$

where $n_i = \#$ observations in category i (so $\sum n_i = n$).

We think of $n_1, \dots, n_p$ as realisations of random counts

$N_1, \dots, N_p$. The m.l.e. is $\hat{\theta} = n^{-1}(n_1, \dots, n_{p-1})$.

Test $H_0 : \theta = \theta_0$, where $\theta_0 = (\pi_{1,0}, \dots, \pi_{p-1,0})$, against

$H_1 : \theta \neq \theta_0$.

The score vector is vector of partial derivatives of

$$\ell(\theta) = \sum_{i=1}^{p-1} n_i \log \pi_i + n_p \log \left(1 - \sum_{k=1}^{p-1} \pi_k \right)$$

with respect to $\pi_1, \dots, \pi_{p-1}$:

$$\frac{\partial \ell}{\partial \pi_i} = \frac{n_i}{\pi_i} - \frac{n_p}{1 - \sum_{k=1}^{p-1} \pi_k}.$$

The matrix of second derivatives has entries

$$\frac{\partial^2 \ell}{\partial \pi_i \partial \pi_k} = -\frac{n_i \delta_{ik}}{\pi_i^2} - \frac{n_p}{(1 - \sum_{i=1}^{p-1} \pi_i)^2},$$

where $\delta_{ik} = 1$ if $i = k$, and $\delta_{ik} = 0$ if $i \neq k$. Minus the

expectation of this, using $\mathbf{E}[\theta_0 N_i] = n\pi_i$, gives

$$I_n(\theta) = n \text{Diag}(\pi_1^{-1}, \dots, \pi_{p-1}^{-1}) + n \mathbf{1} \mathbf{1}^T \pi_p^{-1},$$

where $\mathbf{1}$ is a $(p-1)$ -dimensional vector of ones.

Compute

$$\ell'(\theta_0)^T [I_n(\theta_0)]^{-1} \ell'(\theta_0) = \sum_{i=1}^p \frac{(n_i - n\pi_{i,0})^2}{n\pi_{i,0}};$$

this statistic is called the *chi-squared statistic*, T say.

The CLT for the score vector gives that $T \approx \chi_{p-1}^2$ under H_0 .

Note: the form of the chi-squared statistic is

$$\sum (O_i - E_i)^2 / E_i$$

where O_i and E_i refer to observed and expected frequencies in category i : This is known as *Pearson's chi-square statistic*.

Composite null hypotheses

$\theta = (\psi, \lambda)$, λ nuisance parameter

want test not depending on the unknown value of λ

Extending two of the previous methods:

Generalised likelihood ratio test: Composite

null hypothesis

$$H_0 : \theta \in \Theta_0$$

$$H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$$

The (generalised) LR test uses the likelihood ratio statistic

$$T = \frac{\max_{\theta \in \Theta} L(\theta; \mathbf{X})}{\max_{\theta \in \Theta_0} L(\theta; \mathbf{X})}$$

and rejects H_0 for large values of T .

Now $\theta = (\psi, \lambda)$, assume that ψ is scalar, test $H_0 : \psi = \psi_0$

against $H_1^+ : \psi > \psi_0$.

The LR statistic T is

$$T = \frac{\max_{\psi \geq \psi_0, \lambda} L(\psi, \lambda)}{\max_{\lambda} L(\psi_0, \lambda)} = \frac{\max_{\psi \geq \psi_0} L_P(\psi)}{L_P(\psi_0)}$$

where $L_P(\psi)$ is the profile likelihood for ψ .

For H_0 against $H_1 : \psi \neq \psi_0$,

$$T = \frac{\max_{\psi, \lambda} L(\psi, \lambda)}{\max_{\lambda} L(\psi_0, \lambda)} = \frac{L(\hat{\psi}, \hat{\lambda})}{L_P(\psi_0)}.$$

Often (see Chapter 5):

$$2 \log T \approx \chi_p^2$$

where p is the dimension of ψ

Important requirement: the dimension of λ does not depend on n .

Example: Normal distribution and Student

t-test

Random sample, size n , $N(\mu, \sigma^2)$, both μ and σ unknown; $H_0 : \mu = \mu_0$. Ignoring an irrelevant additive constant,

$$\ell(\theta) = -n \log \sigma - \frac{n(\bar{x} - \mu)^2 + (n-1)s^2}{2\sigma^2}$$

Maximizing this w.r.t. σ with μ fixed gives

$$\ell_P(\mu) = -\frac{n}{2} \log \left(\frac{(n-1)s^2 + n(\bar{x} - \mu)^2}{n} \right).$$

If $H_1^+ : \mu > \mu_0$: maximise $\ell_P(\mu)$ over $\mu \geq \mu_0$:

if $\bar{x} \leq \mu_0$ then max at $\mu = \mu_0$

if $\bar{x} > \mu_0$ then max at $\mu = \bar{x}$

So $\log T = 0$ when $\bar{x} \leq \mu_0$ and is

$$\begin{aligned} & -\frac{n}{2} \log \left(\frac{(n-1)s^2}{n} \right) \\ & + \frac{n}{2} \log \left(\frac{(n-1)s^2 + n(\bar{x} - \mu_0)^2}{n} \right) \\ & = \frac{n}{2} \log \left(1 + \frac{n(\bar{x} - \mu_0)^2}{(n-1)s^2} \right) \end{aligned}$$

when $\bar{x} > \mu_0$.

Thus the LR rejection region is of the form

$$R = \{\mathbf{x} : t(\mathbf{x}) \geq c_\alpha\},$$

where

$$t(\mathbf{x}) = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s}.$$

This statistic is called *Student-t* statistic. Under H_0 ,

$t(\mathbf{X}) \sim t_{n-1}$, and for a size α test set $c_\alpha = t_{n-1,1-\alpha}$; the

p -value is $p = \mathbf{P}(t_{n-1} \geq t(\mathbf{x}))$. Here we use the quantile

notation $\mathbf{P}(t_{n-1} \geq t_{n-1,1-\alpha}) = \alpha$.

The two-sided test of H_0 against $H_1 : \mu \neq \mu_0$ is easier, as unconstrained maxima are used. The size α test has rejection region

$$R = \{\mathbf{x} : |t(\mathbf{x})| \geq t_{n-1, 1-\alpha/2}\}.$$

Score test: Composite null hypothesis

Now $\theta = (\psi, \lambda)$ with ψ scalar, test $H_0 : \psi = \psi_0$ against

$H_1^+ : \psi > \psi_0$ or $H_1^- : \psi < \psi_0$. The score test statistic is

$$T = \frac{\partial \ell(\psi_0, \hat{\lambda}_0; \mathbf{X})}{\partial \psi},$$

where $\hat{\lambda}_0$ is the MLE for λ when H_0 is true.

Large positive values of T indicate H_1^+ , and large negative values indicate H_1^- . Thus the rejection regions are of

the form $T \geq k_\alpha^+$ when testing against H_1^+ , and $T \leq k_\alpha^-$

when testing against H_1^- .

Recall the derivation of the score test,

$$\ell(\theta_0 + \delta) - \ell(\theta_0) \approx \delta \frac{\partial \ell(\theta_0)}{\partial \theta} = \delta T.$$

If $\delta > 0$, i.e. for H_1^+ , we reject if T is large; if $\delta < 0$, i.e.

for H_1^- , we reject if T is small.

Sometimes the exact null distribution of T is available; more often we use that $T \approx$ normal (by CLT, see Chapter 5), zero mean. To find the approximate variance:

1. compute $I_n(\psi_0, \lambda)$
2. invert to I_n^{-1}
3. take the diagonal element corresponding to ψ
4. invert this element
5. replace λ by the null hypothesis MLE $\hat{\lambda}_0$.

Denote the result by v , then $Z = T/\sqrt{v} \approx \mathcal{N}(0, 1)$ under H_0 .

A considerable advantage is that the unconstrained MLE $\hat{\psi}$ is not required.

Example: linear or non-linear model

We can extend the linear model $Y_j = (x_j^T \beta) + \epsilon_j$, where $\epsilon_1, \dots, \epsilon_n$ i.i.d. $\mathcal{N}(0, \sigma^2)$, to a non-linear model

$$Y_j = (x_j^T \beta)^\psi + \epsilon_j$$

with the same ϵ 's. Test $H_0 : \psi = 1$: usual linear model, against, say, $H_1^- : \psi < 1$. Here our nuisance parameters are $\lambda^T = (\beta^T, \sigma^2)$.

Write $\eta_j = x_j^T \beta$, and denote the usual linear model fitted values by $\hat{\eta}_{j0} = x_j^T \hat{\beta}_0$, where the estimates are obtained under H_0 . As $Y_j \sim \mathcal{N}(\eta_j, \sigma^2)$, we have up to an irrelevant additive constant,

$$\ell(\psi, \beta, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum (y_j - \eta_j^\psi)^2,$$

and so

$$\frac{\partial \ell}{\partial \psi} = \frac{1}{\sigma^2} \sum (y_j - \eta_j^\psi) \eta_j^\psi \log \eta_j,$$

yielding that the null MLE's are the usual LSEs (least-square estimates), which are

$$\hat{\beta}_0 = (X^T X)^{-1} X^T Y,$$

$$\hat{\sigma}^2 = n^{-1} \sum (Y_j - x_j^T \hat{\beta}_0)^2.$$

So the score test statistic becomes

$$T = \frac{1}{\hat{\sigma}^2} \sum (Y_j - \hat{\eta}_{j0})(\hat{\eta}_{j0} \log \hat{\eta}_{j0}).$$

We reject H_0 for large negative values of T .

Compute approximate null variance (see below):

$$I_n(\psi_0, \beta, \sigma) = \frac{1}{\sigma^2} \begin{pmatrix} \sum u_j^2 & \sum u_j x_j^T & 0 \\ \sum u_j x_j & \sum x_j x_j^T & 0 \\ 0 & 0 & 2n \end{pmatrix}$$

where $u_j = \eta_j \log \eta_j$. The $(1, 1)$ element of the inverse of I_n has reciprocal

$$(u^T u - u^T X (X^T X)^{-1} X^T u) / \sigma^2,$$

where $u^T = (u_1, \dots, u_n)$. Substitute $\hat{\eta}_{j0}$ for η_j and $\hat{\sigma}^2$ for σ^2 to get v . For the approximate p -value calculate $z = t/\sqrt{v}$ and set $p = \Phi(z)$.

Calculation trick:

To compute the $(1, 1)$ element of the inverse of I_n above:

if

$$A = \begin{pmatrix} a & x^T \\ x & B \end{pmatrix}$$

where a is a scalar, x is an $(n-1) \times 1$ vector and B is an

$(n-1) \times (n-1)$ matrix, then $(A^{-1})_{11} = 1/(a - x^T B^{-1} x)$.

Recall also:

$$\begin{aligned} \frac{\partial}{\partial \psi} \eta^\psi &= \frac{\partial}{\partial \psi} e^{\psi \ln \eta} \\ &= \ln \eta e^{\psi \ln \eta} \\ &= \eta^\psi \ln \eta. \end{aligned}$$

For the $(1, 1)$ -entry of the information matrix, we calculate

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \psi^2} &= \frac{1}{\sigma^2} \sum \left\{ (-\eta_j^\psi \log \eta_j) \eta_j^\psi \log \eta_j \right. \\ &\quad \left. + (y_j - \eta_j^\psi) \eta_j^\psi (\log \eta_j)^2 \right\}, \end{aligned}$$

and as $Y_j \sim \mathcal{N}(\eta_j, \sigma^2)$ we have

$$\begin{aligned} E \left\{ -\frac{\partial^2 \ell}{\partial \psi^2} \right\} &= \frac{1}{\sigma^2} \sum \eta_j^\psi \log \eta_j \eta_j^\psi \log \eta_j \\ &= \frac{1}{\sigma^2} \sum u_j^2, \end{aligned}$$

as required. The off-diagonal terms in the information matrix can be calculated in a similar way, using that

$$\frac{\partial}{\partial \beta} \eta = x_j^T.$$

Multiple tests

When many tests applied to the same data, there is a tendency for some p -values to be small: Suppose $P_1, \dots, P_m$ are the random P -values for m independent tests at level α (before seeing the data); for each i , suppose that $P(P_i \leq \alpha) = \alpha$ if the null hypothesis is true. But then the probability that at least one of the null hypothesis is rejected if m independent tests are carried out is

$$1 - P(\text{ none rejected}) = 1 - (1 - \alpha)^m.$$

Example: If $\alpha = 0.05$ and $m = 10$, then

$$P(\text{ at least one rejected } | H_0 \text{ true }) = 0.4012.$$

Thus with high probability at least one "significant" result will be found even when all the null hypotheses are true.

Bonferroni: The Bonferroni inequality gives that

$$\mathbf{P}(\min P_i \leq \alpha \mid H_0) \leq \sum_{i=1}^m \mathbf{P}(P_i \leq \alpha \mid H_0) \leq m\alpha.$$

A cautious approach for an overall level α is therefore to declare the most significant of m test results as significant at level p only if $\min p_i \leq p/m$.

Example: If $\alpha = 0.05$ and $m = 10$, then reject only if the p-value is less than 0.005.

Combining independent tests

Suppose we have k independent experiments/studies, for the same null hypothesis. If only the p -values are reported, and if we have continuous distribution, we may use that under H_0 each p -value is $\mathcal{U}[0, 1]$ uniformly distributed (Exercise). This gives that

$$-2 \sum_{i=1}^k \log P_i \sim \chi_{2k}^2$$

(exactly) under H_0 , so

$$p_{\text{comb}} = \mathbf{P}(\chi_{2k}^2 \geq -2 \sum \log p_i).$$

If each test is based on a statistic T such that $T_i \approx \mathcal{N}(0, v_i)$, then the best combination statistic is

$$Z = \sum (T_i/v_i) / \sqrt{\sum v_i^{-1}}.$$

If H_0 is a hypothesis about a common parameter ψ , then the best combination of evidence is

$$\sum \ell_{P,i}(\psi),$$

and the combined test would be derived from this (e.g., an LR or score test).

Advice

Even though a test may initially be focussed on departures in one direction, it is usually a good idea not to totally disregard departures in the other direction, even if they are unexpected.

Warning:

Not rejecting the null hypothesis does not mean that the null hypothesis is true! Rather it means that there is not enough evidence to reject the null hypothesis; the data are consistent with the null hypothesis.

The p -value is **not** the probability that the null hypothesis is true.

Nonparametric tests

Sometimes we do not have a parametric model available, and the null hypothesis is phrased in terms of arbitrary distributions, for example concerning only the median of the underlying distribution. Such tests are called *nonparametric* or *distribution-free*; treating these would go beyond the scope of these lectures.