Statistics 2 Unit 4

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Kurt Hornik

EQUIS ACCREDITED AACSB ACCREDITED ASSOCIATION OF MBAS ACCREDITED

Outline

- Testing hypotheses and assessing goodness of fit

Outline

- Testing hypotheses and assessing goodness of fit
  - The Neyman-Pearson paradigm
  - Duality of confidence regions and hypothesis tests
  - Generalized likelihood ratio tests
  - Likelihood ratio tests for the multinomial distribution
  - Assessing goodness of fit

The Neyman-Pearson paradigm

Ideally, we would like to make both α and β as small as possible.

But that does not work.

In the Neyman-Pearson paradigm, we thus control for α to be "small enough", and then try to find tests which also have small β (which is not always possible).

What is "small enough"? Social compromise. E.g., if $\alpha = 0.05$, we falsely reject the null "only" in one out of 20 cases.

Technically, we say a test is of **level** α if its size does not exceed the **significance level** α.

(The difference between size and level is usually "ignored".)

Note the asymmetry between H_0 and H_A: we control the probability of falsely rejecting H_0!

The Neyman-Pearson paradigm

Now that looks a bit strange (but then you've seen it before), and it would seem the first idea is preferable.

However:

- The Neyman-Pearson paradigm gives decision rules which have worked rather well for practical decision making
- In quite a few situations we get the same decision rules anyway.

In particular, for simple against simple we get the same decision rules.

The Neyman-Pearson lemma

Theorem (Neyman-Pearson lemma). *Suppose H_0 and H_A are simple hypotheses and that the test that rejects H_0 whenever the likelihood ratio (LR) is less than c has size α. Then any other test whose size does not exceed α has power not exceeding that of the likelihood ratio test.*

The Neyman-Pearson lemma

Theorem (Neyman-Pearson lemma). *Suppose H_0 and H_A are simple hypotheses and that the test that rejects H_0 whenever the likelihood ratio (LR) is less than c has size α. Then any other test whose size does not exceed α has power not exceeding that of the likelihood ratio test.*

Equivalently, any other level α test has a type II error probability β not below that of the LR test.

So for simple against simple, the LR tests are “best” (which agrees with our intuition), but the critical values are obtained by controlling α and not via prior odds!

The Neyman-Pearson lemma

Before we prove the lemma . . .

The Neyman-Pearson lemma

Before we prove the lemma . . .

Note that the likelihood ratio decision rules are now of the form

$$\text{reject } H_0 \Leftrightarrow \text{LR} < c.$$

(So "ties" are broken in favor of H_0.)

The Neyman-Pearson lemma

Before we prove the lemma . . .

The Neyman-Pearson lemma

Before we prove the lemma . . .

Note also the strange formulation: if we use the above decision rule, the type I error probability is α. Why not start with α and choose the critical value c appropriately?

The Neyman-Pearson lemma

Before we prove the lemma . . .

Note also the strange formulation: if we use the above decision rule, the type I error probability is α. Why not start with α and choose the critical value c appropriately?

Well, this does not work in the discrete case. In our introductory example, the only α we can exactly get are

```
R> pbinom(0 : 10, 10, 0.5, lower.tail = FALSE)
```

```
 [1] 0.9990234375 0.9892578125 0.9453125000 0.8281250000 0.6230468750
 [6] 0.3769531250 0.1718750000 0.0546875000 0.0107421875 0.0009765625
[11] 0.0000000000
```

The Neyman-Pearson lemma

So what if we wanted a likelihood ratio test with α exactly 0.05?

The Neyman-Pearson lemma

So what if we wanted a likelihood ratio test with α exactly 0.05?

Well, that's not possible. Unless we take a random decision when $x = 8$.

Hard-core N-P theory thus considers "randomized decision rules" which give the probability of rejecting H_0.

But that's awful, also from a philosophical perspective, so let's only look at non-randomized tests/decisions (and take the lemma as formulated).

The Neyman-Pearson lemma

So what if we wanted a likelihood ratio test with α exactly 0.05?

Well, that's not possible. Unless we take a random decision when $x = 8$.

Hard-core N-P theory thus considers "randomized decision rules" which give the probability of rejecting H_0.

But that's awful, also from a philosophical perspective, so let's only look at non-randomized tests/decisions (and take the lemma as formulated).

A (non-randomized) test is then equivalent to its decision function

$$d(x) = I_{\text{rejection region}}(x) = \begin{cases} 1, & x \in \text{rejection region (i.e., reject } H_0), \\ 0, & x \in \text{acceptance region (i.e., accept } H_0). \end{cases}$$

The Neyman-Pearson lemma

Proof of the Neyman-Pearson lemma. Consider any test with decision function d.

The size is

$$\mathbb{P}(\text{reject } H_0 | H_0) = \mathbb{P}(d(X) = 1 | H_0) = \mathbb{E}_0(d(X)),$$

the power is

$$\mathbb{P}(\text{reject } H_0 | H_A) = \mathbb{P}(d(X) = 1 | H_A) = \mathbb{E}_A(d(X)).$$

Write f_0 and f_A for the densities (or pmfs) under H_0 and H_A, respectively, and d^* for the decision function of the likelihood ratio test, i.e.,

$$d^*(x) = 1 \Leftrightarrow f_0(x)/f_A(x) < c \Leftrightarrow cf_A(x) - f_0(x) > 0.$$

The Neyman-Pearson lemma

For all x, we have

$$d(x)(cf_A(x) - f_0(x)) \leq d^*(x)(cf_A(x) - f_0(x)).$$

Why?

The Neyman-Pearson lemma

For all x, we have

$$d(x)(cf_A(x) - f_0(x)) \leq d^*(x)(cf_A(x) - f_0(x)).$$

Why?

If $d^*(x) = 0$, $cf_A(x) - f_0(x) \leq 0$, so the LHS above is ≤ 0 and the RHS is 0. So o.k.

The Neyman-Pearson lemma

For all x, we have

$$d(x)(cf_A(x) - f_0(x)) \leq d^*(x)(cf_A(x) - f_0(x)).$$

Why?

If $d^*(x) = 0$, $cf_A(x) - f_0(x) \leq 0$, so the LHS above is ≤ 0 and the RHS is 0. So o.k.

If $d^*(x) = 1$, $cf_A(x) - f_0(x) > 0$. So

$$d^*(x)(cf_A(x) - f_0(x)) = (cf_A(x) - f_0(x)) \geq d(x)(cf_A(x) - f_0(x))$$

is also o.k.

The Neyman-Pearson lemma

So for all x, we have

$$d(x)(cf_A(x) - f_0(x)) \le d^*(x)(cf_A(x) - f_0(x)).$$

Now integrate this (with respect to the reference measure for the densities): this gives

$$c\mathbb{E}_A(d(X)) - \mathbb{E}_0(d(X)) \le c\mathbb{E}_A(d^*(X)) - \mathbb{E}_0(d^*(X))$$

or equivalently,

$$\mathbb{E}_A(d^*(X)) - \mathbb{E}_A(d(X)) \ge (\mathbb{E}_0(d^*(X)) - \mathbb{E}_0(d(X)))/c.$$

Thus, $\mathbb{E}_0(d(X)) \le \mathbb{E}_0(d^*(X))$ implies that $\mathbb{E}_A(d^*(X)) \ge \mathbb{E}_A(d(X))$, as asserted.

Example: Normal distribution

Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with known variance σ^2, and consider the simple hypotheses

$$H_0 : \mu = \mu_0 \qquad H_A : \mu = \mu_A.$$

By the Neyman-Pearson lemma, among all level α tests, the one that rejects for small values of the likelihood ratio is most powerful.

What does this test look like?

Example: Normal distribution

We have

$$
\begin{aligned}
\frac{f_0(x)}{f_A(x)} &= \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu_0)^2/2\sigma^2}}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu_A)^2/2\sigma^2}} \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu_0)^2 - \sum_{i=1}^{n}(x_i-\mu_A)^2\right)\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i^2-2x_i\mu_0+\mu_0^2) - \sum_{i=1}^{n}(x_i^2-2x_i\mu_A+\mu_A^2)\right)\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(2n\bar{x}(\mu_A-\mu_0)+n(\mu_0^2-\mu_A^2)\right)\right).
\end{aligned}
$$

Example: Normal distribution

Clearly,

$$
\begin{aligned}
\frac{f_0(x)}{f_A(x)} < c_0 \quad &\Leftrightarrow \quad -\frac{1}{2\sigma^2}\left(2n\bar{x}(\mu_A - \mu_0) + n(\mu_0^2 - \mu_A^2)\right) < \log(c_0) \\
&\Leftrightarrow \quad 2n\bar{x}(\mu_A - \mu_0) + n(\mu_0^2 - \mu_A^2) > -2\sigma^2 \log(c_0) \\
&\Leftrightarrow \quad \bar{x}(\mu_A - \mu_0) > \frac{n(\mu_A^2 - \mu_0^2) - 2\sigma^2 \log(c)}{2n} =: c_1.
\end{aligned}
$$

If $\mu_0 > \mu_A$,

$$
\frac{f_0(x)}{f_A(x)} < c_0 \Leftrightarrow \bar{x} < \frac{c_1}{\mu_A - \mu_0}.
$$

Thus, the LR is small iff $\bar{x}$ is small.

Example: Normal distribution

If $\mu_0 < \mu_A$,

$$\frac{f_0(x)}{f_A(x)} < c_0 \Leftrightarrow \bar{x} > \frac{c_1}{\mu_A - \mu_0}.$$

Thus, the LR is small iff $\bar{x}$ is large.

In this case, the likelihood ratio test (LRT) rejects iff $\bar{x} > c$, where $c = c_\alpha$ is determined by controlling the significance level α:

$$\alpha = \mathbb{P}(\bar{X} > c | H_0) = \mathbb{P}_0\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{c - \mu_0}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{c - \mu_0}{\sigma/\sqrt{n}}\right).$$

Thus,

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

Example: Normal distribution

Note that we get a very nice result.

Intuitively,

- By sufficiency, decision rules should only have to look at $\bar{x}$.

Example: Normal distribution

Note that we get a very nice result.

Intuitively,

- By sufficiency, decision rules should only have to look at $\bar{x}$.
- If $\mu_0 < \mu_A$, then clearly large values of $\bar{x}$ increasingly favor H_A. So decision rules should be of the form “reject H_0 iff $\bar{x} > c$”.
  (The fact that $\mu_0 < \mu_A$ determines the shape of the rejection region.)

The critical value is determined by making the size α.

Example: Normal distribution

Note that we get a very nice result.

Intuitively,

- By sufficiency, decision rules should only have to look at $\bar{x}$.
- If $\mu_0 < \mu_A$, then clearly large values of $\bar{x}$ increasingly favor H_A. So decision rules should be of the form “reject H_0 iff $\bar{x} > c$”.
  (The fact that $\mu_0 < \mu_A$ determines the shape of the rejection region.)

The critical value is determined by making the size α.

By the Neyman-Pearson lemma, this gives the most powerful test with significance level α.

Significance levels and p-values

In the previous example, the decision rules were of the form

$$\text{reject } H_0 \Leftrightarrow \bar{X} > c_\alpha$$

with c_α chosen to make the significance level α, i.e.,

$$\mathbb{P}(\bar{X} > c_\alpha | H_0) = \alpha.$$

(Clearly, the above only makes sense if H_0 is simple. In general, the size is defined as the max/sup of the type I errors.)

Significance levels and p-values

The original N-P idea was

- fix α in advance (e.g., 5 %)
- determine the corresponding critical value c_α
- when observing $x_1, \dots, x_n$, report whether $\bar{x} > c_\alpha$ (reject) or not (accept).

Significance levels and p-values

The original N-P idea was

- fix α in advance (e.g., 5 %)
- determine the corresponding critical value c_α
- when observing $x_1, \ldots, x_n$, report whether $\bar{x} > c_\alpha$ (reject) or not (accept).

Clearly, this loses some information (e.g., was $\bar{x}$ close to the critical value of not?).

Significance levels and p-values

The original N-P idea was

- fix α in advance (e.g., 5 %)
- determine the corresponding critical value c_α
- when observing $x_1, \ldots, x_n$, report whether $\bar{x} > c_\alpha$ (reject) or not (accept).

Clearly, this loses some information (e.g., was $\bar{x}$ close to the critical value of not?).

Do we really need to fix α in advance?

Alternatively, we could report the smallest significance level at which the null hypothesis would be rejected: the so-called ***p*-value** of the test.

Significance levels and p-values

In our case,

$$\bar{x} > c_\alpha \quad \Rightarrow \quad \mathbb{P}(\bar{X} > \bar{x} | H_0) \leq \mathbb{P}(\bar{X} > c_\alpha | H_0) = \alpha.$$

So clearly,

$$\inf\{\alpha : \bar{x} > c_\alpha\} = \mathbb{P}(\bar{X} > \bar{x} | H_0).$$

I.e., the p-value is the probability (under the null) of observing something “more extreme” than we observed.

This recovers Fisher’s approach to testing (older than N-P) by reporting “fiducial values”, interpreted as the null probability of observing something more extreme (in a sense, less fitting with the null) than what was actually observed.

Significance levels and p-values

Can easily be generalized: if T is the test statistic and we reject for large values of T (“large values are significant”), i.e., use decision rules of the form

$$\text{reject } H_0 \Leftrightarrow T > t_\alpha$$

then when observing $t_{\text{obs}} = t(x_1, \ldots, x_n)$,

$$p = \mathbb{P}(T > t_{\text{obs}} | H_0).$$

(Similarly when rejecting for $\geq$ instead of $>$.)

Significance levels and p-values

But note:

$$p = \mathbb{P}(T > t_{\mathrm{obs}} | H_0) = \mathbb{P}(T > t(x_1, \ldots, x_n) | H_0) = p(x_1, \ldots, x_n).$$

So p-values depend on the observations!

They are observations of random variables and **not** probabilities ("the probability that H_0 is true").

In fact, under the null the p-value has a standard uniform distribution (see the homeworks).

Significance levels and p-values

Modern statistical software always reports p-values (i.e., the smallest α for which the null would be rejected) and not the binary decision for a fixed α.

One can easily recover the binary decision by comparing p and a target signficance level α, as

$$\text{reject } H_0 \text{ at level } \alpha \Leftrightarrow p \leq \alpha.$$

Often, software “helps” to see the binary decision by adding “significance stars” (e.g., in R for linear regression modeling).

Roles of H_0 and H_A

By prescribing a significance level to control the size (probability of type I error), one introduces a fundamental asymmetry between H_0 and H_A.

One only rejects H_0 when the data provides significant evidence against it.

In some sense, one can only (significantly) “falsify” H_0, but not “verify” it.

To have the data provide significant evidence **for** something, one has to put that into H_A!

(This is what typically happens in statistical modeling.)

Uniformly most powerful tests

Consider again the situation where $X_1, \ldots, X_n$ are i.i.d. normal with unknown mean μ and known variance σ^2, and we want to test

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu = \mu_A$$

where $\mu_0 < \mu_A$.

We found that the most powerful level α test is of the form

$$\text{reject } H_0 \Leftrightarrow \bar{X} > c_\alpha$$

where c_α is determined by

$$\mathbb{P}(\bar{X} > c_\alpha | H_0) = \mathbb{P}(N(\mu_0, \sigma^2/n) > c_\alpha) = \alpha \quad \Rightarrow \quad c_\alpha = \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}.$$

Uniformly most powerful tests

Consider again the situation where $X_1, \ldots, X_n$ are i.i.d. normal with unknown mean μ and known variance σ^2, and we want to test

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu = \mu_A$$

where $\mu_0 < \mu_A$.

We found that the most powerful level α test is of the form

$$\text{reject } H_0 \Leftrightarrow \bar{X} > c_\alpha$$

where c_α is determined by

$$\mathbb{P}(\bar{X} > c_\alpha | H_0) = \mathbb{P}(N(\mu_0, \sigma^2/n) > c_\alpha) = \alpha \quad \Rightarrow \quad c_\alpha = \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}.$$

I.e., c_α depends on μ_0, but **not** μ_A! (Only $\mu_0 < \mu_A$ matters).

Uniformly most powerful tests

Thus, for all $\mu_A > \mu_0$ we get the same test!

The test is thus **uniformly most powerful** (UMP) for

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu > \mu_0$$

where H_A is now a **one-sided** composite hypothesis.

So also in this case, there is a “best” test. This is good.

Uniformly most powerful tests

Thus, for all $\mu_A > \mu_0$ we get the same test!

The test is thus **uniformly most powerful** (UMP) for

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu > \mu_0$$

where H_A is now a **one-sided** composite hypothesis.

So also in this case, there is a “best” test. This is good.

One can argue that the test is also UMP for $H_0 : \mu \leq \mu_0$ against $H_A : \mu > \mu_0$ (among all tests with size $\sup_{\mu \in H_0} \alpha(\mu)$).

But it is **not** UMP for testing $H_0 : \mu = \mu_0$ against $H_A : \mu \neq \mu_0$ (as for alternatives $> \mu_0$ and $< \mu_0$, the UMP tests reject for large and small values of $\bar{X}$, respectively).

In fact, there clearly cannot be a UMP test for this problem. This is bad.

Outline

- Testing hypotheses and assessing goodness of fit
  - The Neyman-Pearson paradigm
  - Duality of confidence regions and hypothesis tests
  - Generalized likelihood ratio tests
  - Likelihood ratio tests for the multinomial distribution
  - Assessing goodness of fit

Example: Normal distribution

Consider again the situation where $X_1, \ldots, X_n$ are i.i.d. normal with unknown mean μ and known variance σ^2. Suppose we want to test

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu \neq \mu_0$$

The "obvious" level α test is

$$\text{reject } H_0 \Leftrightarrow |\bar{X} - \mu_0| > c_\alpha$$

with c_α determined from

$$\alpha = \mathbb{P}(|\bar{X} - \mu_0| > c_\alpha | \mu_0) = \mathbb{P}(-c_\alpha < \bar{X} - \mu_0 < c_\alpha | \mu_0)$$

from which

$$c_\alpha = z_{1-\alpha/2} \sigma_{\bar{X}} = z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Example: Normal distribution

The test accepts if

$$|\bar{X} - \mu_0| \le c_\alpha \quad \Leftrightarrow \quad -c_\alpha \le \bar{X} - \mu_0 \le c_\alpha \quad \Leftrightarrow \quad \bar{X} - c_\alpha \le \mu_0 \le \bar{X} + c_\alpha.$$

As the acceptance probability is $1 - \alpha$, the above gives the (already known) $100(1 - \alpha)\%$ confidence interval for μ.

I.e.,

$$\mu_0 \text{ is in the confidence interval for } \mu \quad \Leftrightarrow \quad \text{hypothesis test accepts.}$$

This duality between confidence regions (for parameter estimation) and acceptance regions (for hypothesis testing) holds quite generally.

Duality of confidence regions and hypothesis tests

Let θ be a (real-valued) parameter of a family of probability distributions, and Θ be the set of all possible values of θ.

Theorem. *Suppose for every θ_0 in Θ there is a level α test of the hypothesis $H_0 : \theta = \theta_0$. Denote the acceptance region of this test by $A(\theta_0)$. Then the set*

$$C(X) = \{\theta : X \in A(\theta)\}$$

is a $100(1-\alpha)\%$ confidence region for θ.

(Note “confidence region”, as we don’t necessarily get intervals.)

Duality of confidence regions and hypothesis tests

Proof. We have

$$\theta_0 \in C(X) \Leftrightarrow X \in A(\theta_0).$$

Hence, for every $\theta_0 \in \Theta$,

$$\mathbb{P}(\theta_0 \in C(X) | \theta_0) = \mathbb{P}(X \in A(\theta_0) | \theta_0) = 1 - \alpha.$$

Duality of confidence regions and hypothesis tests

Theorem. *Suppose that* $C(X)$ *is a* $100(1-\alpha)\%$ *confidence region for* θ*, i.e., for every* θ_0*,*

$$\mathbb{P}(\theta_0 \in C(X) | \theta_0) = 1 - \alpha.$$

Then

$$A(\theta_0) = \{X : \theta_0 \in C(X)\}$$

defines an acceptance region for a level α *test of the null hypothesis* $H_0 : \theta = \theta_0$*.*

Duality of confidence regions and hypothesis tests

Proof.

$$\mathbb{P}(X \in A(\theta_0)|\theta_0) = \mathbb{P}(\theta_0 \in C(X)|\theta_0) = 1 - \alpha.$$

Outline

- Testing hypotheses and assessing goodness of fit
  - The Neyman-Pearson paradigm
  - Duality of confidence regions and hypothesis tests
  - Generalized likelihood ratio tests
  - Likelihood ratio tests for the multinomial distribution
  - Assessing goodness of fit

Introduction

Consider a general hypothesis testing problem of the form

$$H_0 : \theta \in \Theta_0 \quad \text{against} \quad H_A : \theta \in \Theta_A.$$

If both Θ_0 and Θ_A were simple, we know (N-P lemma!) that the likelihood ratio test is optimal (UMP).

In the general case, we do not know (and UMP tests usually do not exist), but it still seems a good idea to base a test on the ratio of the (maximal) likelihoods under the null and alternative.

Introduction

This gives the **generalized likelihood ratio** test statistic

$$\Lambda^* = \frac{\sup_{\theta \in \Theta_0} \text{lik}(\theta)}{\sup_{\theta \in \Theta_A} \text{lik}(\theta)}$$

or the usually preferred

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} \text{lik}(\theta)}{\sup_{\theta \in \Theta_0 \cup \Theta_A} \text{lik}(\theta)}$$

Both Likelihood ratio tests (LRTs) reject for small values of the test statistic. I.e.,

$$\text{reject } H_0 \Leftrightarrow \Lambda^* < c, \qquad \text{reject } H_0 \Leftrightarrow \Lambda < c.$$

Introduction

Note: finding

$$\sup_{\theta \in \Theta_0} \text{lik}(\theta)$$

gives the **constrained** MLE under the null (i.e., the MLE under the constraint that $\theta \in \Theta_0$).

Finding

$$\sup_{\theta \in \Theta_A} \text{lik}(\theta)$$

gives the constrained MLE under the alternative.

Introduction

Finding

$$\sup_{\theta \in \Theta_0 \cup \Theta_A} \text{lik}(\theta)$$

gives the “usual” (unconstrained) MLE, provided that $\Theta_0 \cup \Theta_A$ gives the whole parameter space Θ.

Example: Normal distribution

Consider again the situation where $X_1, \ldots, X_n$ are i.i.d. normal with unknown mean μ and known variance σ^2. Suppose we want to test

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu \neq \mu_0$$

The likelihood for μ (remember σ is known) is

$$\text{lik}(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/2\sigma^2} = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

Under H_0, $\mu = \mu_0$ (constrained MLE under the null).

The unconstrained MLE is $\hat{\mu} = \bar{x}$.

Example: Normal distribution

Thus, the LRT statistic is

$$\Lambda = \frac{\text{lik}(\mu_0)}{\text{lik}(\hat{\mu})} = \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i - \mu_0)^2 - \sum_{i=1}^{n}(x_i - \bar{x})^2\right)\right)$$

and the LRT does

$$\text{reject } H_0 \quad \Leftrightarrow \quad \Lambda < c_0 \quad \Leftrightarrow \quad -2\log(\Lambda) > -2\log(c_0) := c_1.$$

Now remember that

$$\sum_{i=1}^{n}(x_i - \mu_0)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2.$$

Example: Normal distribution

Therefore,

$$
\begin{aligned}
-2\log(\Lambda) &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu_0)^2-\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\\
&= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu_0)^2-\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\\
&= \frac{1}{\sigma^2}n(\bar{x}-\mu_0)^2\\
&= \left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)^2.
\end{aligned}
$$

Example: Normal distribution

The LRT is thus

$$\text{reject } H_0 \quad \Leftrightarrow \quad \left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)^2 > c_1 \quad \Leftrightarrow \quad \left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right| > c_2$$

with c_1 and c_2 determined to give a level α test.

Note that we again get the “obvious” test. This is nice.

Example: Normal distribution

The LRT is thus

$$\text{reject } H_0 \quad \Leftrightarrow \quad \left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)^2 > c_1 \quad \Leftrightarrow \quad \left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right| > c_2$$

with c_1 and c_2 determined to give a level α test.

Note that we again get the “obvious” test. This is nice.

Under H_0, clearly $\bar{X} \sim N(\mu_0, \sigma^2/n)$, and thus $c_2 = z_{1-\alpha/2}$. Alternatively,

$$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \Rightarrow \left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right)^2 \sim \chi_1^2,$$

from which $c_1 = Q_{\chi_1^2}(1-\alpha)$.

The latter holds more generally, approximately for large samples.

Asymptotic null distribution of the LRT statistic

Theorem. *Under smoothness conditions on the density (or pmf) functions involved, the null distribution of* $-2\log(\Lambda)$ *tends to a chi-squared distribution with* $\dim(\Theta_0 \cup \Theta_A) - \dim(\Theta_0)$ *degrees of freedom as the sample size tends to infinity.*

Asymptotic null distribution of the LRT statistic

Theorem. *Under smoothness conditions on the density (or pmf) functions involved, the null distribution of* $-2\log(\Lambda)$ *tends to a chi-squared distribution with* $\dim(\Theta_0 \cup \Theta_A) - \dim(\Theta_0)$ *degrees of freedom as the sample size tends to infinity.*

In the above, the dimensions are the numbers of free parameters.

(We won’t prove the theorem, sorry.)

Example: Normal distribution

In the previous example,

- Θ_0 is simple and hence has no free parameters,
- Θ_A specifies σ^2 but has μ as free parameter.

So

$$\dim(\Theta_0 \cup \Theta_A) - \dim(\Theta_0) = \dim(\mathbb{R}) - \dim(\{\mu_0\}) = 1 - 0 = 1$$

and the theorem says that under the null,

$$-2\log(\Lambda) \xrightarrow{d} \chi_1^2.$$

In fact, we showed that under the null,

$$-2\log(\Lambda) \stackrel{d}{=} \chi_1^2$$

Outline

- Testing hypotheses and assessing goodness of fit
  - The Neyman-Pearson paradigm
  - Duality of confidence regions and hypothesis tests
  - Generalized likelihood ratio tests
  - Likelihood ratio tests for the multinomial distribution
  - Assessing goodness of fit

Introduction

This is always very confusing at first encounter.

Suppose we have observations from a discrete distribution which attains possible values $v_1, \ldots, v_m$ with (unknown) probabilities $p_1, \ldots, p_m$.

With $p = (p_1, \ldots, p_m)$,

$$\mathbb{P}(X = v_j | p) = p_j.$$

We can also write this as

$$\mathbb{P}(X = x | p) = p_1^{I(x=v_1)} \times \cdots \times p_m^{I(x=v_m)} = \prod_{j=1}^{m} p_j^{I(x=v_j)}, \qquad x \in \{v_1, \ldots, v_m\}.$$

(We already encountered this in the Bernoulli experiment example in the section on sufficiency.)

Introduction

Thus, if $X_1, \ldots, X_n$ are i.i.d. from this discrete distribution, the pmf is

$$
\begin{aligned}
\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n | p) &= \prod_{i=1}^{n} \mathbb{P}(X_i = x_i | p) \\
&= \prod_{i=1}^{n} \left(\prod_{j=1}^{m} p_j^{I(x_i = v_j)} \right) \\
&= \prod_{j=1}^{m} p_j^{\sum_{i=1}^{n} I(x_i = v_j)} .
\end{aligned}
$$

Introduction

If we write

$$n_j = n_j(x_1, \dots, x_n) = \sum_{i=1}^{n} I(x_i = v_j)$$

for the absolute frequency of v_j in the observations $x_1, \dots, x_n$,

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n | p) = \prod_{j=1}^{m} p_j^{n_j}.$$

Thus, the frequencies are sufficient for p, and it makes sense to base inference on the sufficient statistics.

Introduction

Now as we know, the corresponding random variables

$$(N_1, \dots, N_m) = (n_1(X_1, \dots, X_n), \dots, n_m(X_1, \dots, X_n))$$

have a **multinomial distribution** with parameters n and $p_1, \dots, p_m$:

$$\mathbb{P}(N_1 = n_1, \dots, N_m = n_m | p) = \frac{n!}{n_1! \cdots n_m!} p_1^{n_1} \cdots p_m^{n_m}.$$

One thus typically presents inference for the params of discrete distributions (with finite support of size m) as inference for the params of the corresponding m-dimensional multinomial distributions.

Which can be confusing at first encounter, in particular because as the m observed counts (which sum to n) correspond to a sample of size n!

LRT for the multinomial distribution

Now suppose we have a parametric model

$$p_1 = p_1(\theta), \ldots, p_m = p_m(\theta)$$

and want to test whether this model "works".

I.e., we want to perform a **goodness-of-fit test** for the appropriateness of the model.

E.g., test whether the binomial distribution is appropriate.

We can write this as

$$H_0 : p_1 = p_1(\theta), \ldots, p_m = p_m(\theta) \text{ for some } \theta \in \Theta$$

against

$$H_A : \text{there is no } \theta \text{ such that } p_1 = p_1(\theta), \ldots, p_m = p_m(\theta).$$

LRT for the multinomial distribution

Using the multinomial for the observed frequencies,

$$\text{lik}(p) = \frac{n!}{n_1! \cdots n_m!} p_1^{n_1} \cdots p_m^{n_m}.$$

LRT for the multinomial distribution

Using the multinomial for the observed frequencies,

$$\mathrm{lik}(p) = \frac{n!}{n_1! \cdots n_m!} p_1^{n_1} \cdots p_m^{n_m}.$$

Under H_0, we need to find

$$\max_{\theta \in \Theta} \mathrm{lik}(p(\theta)).$$

Write $\hat{\theta}$ for the maximizer (which gives the restricted MLE).

LRT for the multinomial distribution

Using the multinomial for the observed frequencies,

$$\operatorname{lik}(p) = \frac{n!}{n_1! \cdots n_m!} p_1^{n_1} \cdots p_m^{n_m}.$$

Under H_0, we need to find

$$\max_{\theta \in \Theta} \operatorname{lik}(p(\theta)).$$

Write $\hat{\theta}$ for the maximizer (which gives the restricted MLE).

Under H_0 or H_A, p is "unconstrained", i.e., the only constraints are

$$p_1 \geq 0, \ldots, p_m \geq 0, \quad p_1 + \cdots + p_m = 1.$$

LRT for the multinomial distribution

To maximize lik(p) over the set of all probability vectors p, we can use the Lagrangian method. The Lagrangian for the log-likelihood is

$$
\begin{aligned}
L(p) &= \log\left(\frac{n!}{n_1!\cdots n_m!}p_1^{n_1}\cdots p_m^{n_m}\right) + \lambda\left(\sum_{j=1}^m p_j - 1\right) \\
&= \log(n!) - \sum_{j=1}^m \log(n_j!) + \sum_{j=1}^m n_j \log(p_j) + \lambda\left(\sum_{j=1}^m p_j - 1\right).
\end{aligned}
$$

Setting the partials with respect to $p_1, \ldots, p_m$ and λ to zero gives

$$
\frac{n_1}{p_1} + \lambda = 0, \ldots, \frac{n_m}{p_m} + \lambda = 0, \sum_{j=1}^m p_j = 1.
$$

LRT for the multinomial distribution

So

$$p_1 = -\frac{n_1}{\lambda}, \ldots, p_m = -\frac{n_m}{\lambda}$$

where λ can be determined from

$$1 = \sum p_j = \sum_{j=1}^{m} \left(-\frac{n_j}{\lambda} \right) = -\frac{1}{\lambda} \sum_{j=1}^{m} n_j = -\frac{n}{\lambda}$$

from which $\lambda = -n$ and

$$\hat{p}_1 = \frac{n_1}{n}, \ldots, \hat{p}_m = \frac{n_m}{n}$$

("as expected").

LRT for the multinomial distribution

The LRT statistic is thus

$$
\begin{aligned}
\Lambda &= \frac{\mathrm{lik}(p(\hat{\theta}))}{\mathrm{lik}(\hat{p})} \\
&= \frac{\dfrac{n!}{n_1! \cdots n_m!} p_1(\hat{\theta})^{n_1} \cdots p_m(\hat{\theta})^{n_m}}{\dfrac{n!}{n_1! \cdots n_m!} \hat{p}_1^{n_1} \cdots \hat{p}_m^{n_m}} \\
&= \prod_{j=1}^{m} \left(\frac{p_j(\hat{\theta})}{\hat{p}_j} \right)^{n_j}.
\end{aligned}
$$

LRT for the multinomial distribution

Therefore, using $\hat{p}_j = n_j/n$,

$$-2\log(\Lambda) = -2\sum_{j=1}^{m} n_j \log\left(\frac{p_j(\hat{\theta})}{\hat{p}_j}\right) = 2\sum_{j=1}^{m} n_j \log\left(\frac{n_j}{np_j(\hat{\theta})}\right)$$

which is commonly written as

$$-2\log(\Lambda) = 2\sum_{j=1}^{m} O_j \log\left(\frac{O_j}{E_j}\right)$$

with

$$O_j = n_j \ldots \text{observed count}, \qquad E_j = n_j p_j(\hat{\theta}) \ldots \text{expected count.}$$

LRT for the multinomial distribution

Under H_0 or H_A, there are $m-1$ free parameters (as $p_1, \ldots, p_m$ sum to one).

Thus if Θ has k free parameters, our theorem yields that under H_0,

$$-2\log(\Lambda) = 2\sum_{j=1}^{m} O_j \log\left(\frac{O_j}{E_j}\right) \xrightarrow{d} \chi^2_{m-k-1}.$$

LRT for the multinomial distribution

Under H_0 or H_A, there are $m-1$ free parameters (as $p_1, \ldots, p_m$ sum to one).

Thus if Θ has k free parameters, our theorem yields that under H_0,

$$-2\log(\Lambda) = 2\sum_{j=1}^{m} O_j \log\left(\frac{O_j}{E_j}\right) \xrightarrow{d} \chi^2_{m-k-1}.$$

Very confusingly, statistical software typically does not use/report $-2\log(\Lambda)$ but instead an asymptotically equivalent chi-squared statistic.

In R, `chisq.test()`.

LRT for the multinomial distribution

To see why/how, note that when n is large and H_0 is true, $\hat{p}_j \approx p_j(\hat{\theta})$.

Consider the function

$$h(x) = x \log\left(\frac{x}{x_0}\right) = x\log(x) - x\log(x_0)$$

for $x \approx x_0$. We have

$$h'(x) = \log(x) + 1 - \log(x_0), \qquad h''(x) = 1/x$$

so that

$$h(x_0) = 0, \qquad h'(x_0) = 1, \qquad h''(x_0) = 1/x_0$$

for a Taylor series expansion of

$$x \log\left(\frac{x}{x_0}\right) = (x - x_0) + \frac{1}{2x_0}(x - x_0)^2 + \cdots$$

LRT for the multinomial distribution

Therefore, taking $x = \hat{p}_j$ and $x_0 = p_j(\hat{\theta})$,

$$\begin{aligned} -2\log(\Lambda) &= 2n\sum_{j=1}^{m}\hat{p}_j \log\left(\frac{\hat{p}_j}{p_j(\hat{\theta})}\right) \\ &\approx 2n\sum_{j=1}^{m}\left((\hat{p}_j - p_j(\hat{\theta})) + \frac{1}{2p_j(\hat{\theta})}(\hat{p}_j - p_j(\hat{\theta}))^2\right). \end{aligned}$$

Since probabilities sum to one,

$$-2\log(\Lambda) \approx n\sum_{j=1}^{m}\frac{(\hat{p}_j - p_j(\hat{\theta}))^2}{p_j(\hat{\theta})} = \sum_{j=1}^{m}\frac{(n\hat{p}_j - np_j(\hat{\theta}))^2}{np_j(\hat{\theta})} = \sum_{j=1}^{m}\frac{(O_j - E_j)^2}{E_j}.$$

LRT for the multinomial distribution

This is the typically encountered chi-squared approximation:

$$-2\log(\Lambda) \approx X^2 = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j} \xrightarrow{d} \chi^2_{m-k-1}.$$

Note: the X is an upper-case χ.

LRT for the multinomial distribution

If H_0 completely specifies the probabilities, i.e.,

$$H_0 : p_1 = p_{1,0}, \dots, p_m = p_{m,0},$$

clearly $k = 0$ and $E_j = np_{j,0}$, and (under H_0),

$$-2\log(\Lambda) = 2\sum_{j=1}^{m} O_j \log\left(\frac{O_j}{np_{j,0}}\right) \approx X^2 = \sum_{j=1}^{m} \frac{(O_j - np_{j,0})^2}{np_{j,0}} \xrightarrow{d} \chi^2_{m-1}$$

("chi-squared goodness of fit test for given probabilities").

LRT for the multinomial distribution

Note: clearly, if $(N_1, \ldots, N_m)$ has a multinomial distribution with parameters n and $p_1, \ldots, p_m$, each N_j has a binomial distribution with parameters n and p_j.

By the CLT,

$$\frac{N_j - np_j}{\sqrt{np_j(1-p_j)}} \xrightarrow{d} N(0,1) \Rightarrow \frac{(N_j - np_j)^2}{np_j(1-p_j)} \xrightarrow{d} \chi^2_1.$$

But the N_j are not independent (as they sum to n), and interestingly, their asymptotic covariance turns out to be such that

$$\sum_{j=1}^{m} \frac{(N_j - np_j)^2}{np_j} \xrightarrow{d} \chi^2_{m-1}.$$

Outline

- Testing hypotheses and assessing goodness of fit
  - The Neyman-Pearson paradigm
  - Duality of confidence regions and hypothesis tests
  - Generalized likelihood ratio tests
  - Likelihood ratio tests for the multinomial distribution
  - Assessing goodness of fit

Assessing goodness of fit

We already know from Statistics 1 that goodness of fit can be judged via probability or preferably quantile plots, which graphically illustrate the goodness of fit of data to suitable families of probability distributions.

There is also a huge variety of goodness of fit hypothesis tests for nulls that the probability distribution comes from a family of distributions against, e.g., the alternative that it does not.

For discrete distributions (with finite support), these can be based on the likelihood ratio or chi-squared tests for the multinomial distribution discussed above.

Assessing goodness of fit

A very popular problem is testing for normality, either against the general alternative of non-normality, or against departures which take the form of asymmetry (skewness) or non-normal kurtosis, or jointly (Jarque-Bera test, implemented in package tseries).

For departures against symmetry, goodness-of-fit tests can be based on the sample coefficient of skewness

$$b_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}$$

which rejects for large values of $|b_1|$. Under the null of normality, this is asymptotically normal with mean 0 and variance $6/n$.