

3 Hypothesis testing

Often when we observed data we have some ideas about the random processes that are generating the observations. Having collected data it is natural to test whether the observed data are consistent with those expectations. The idea of hypothesis testing is to say if the data provides sufficient evidence to rule out those assumptions. The emphasis is always placed in favour of the assumptions, rather than the alternative. We require strong evidence that the data are inconsistent with the assumptions before we reject them.

Formally, we suppose that we have data $\mathbf{x} = (x_1, \dots, x_n)$ and want to examine whether they are consistent with a hypothesis H_0 (the **null hypothesis** or **hypothesis under test**) about the distribution function $F_{\mathbf{X}}$ of \mathbf{X} .

A hypothesis is **simple** if it defines $P_{\mathbf{X}}$ completely:

$$H_0 : P_{\mathbf{X}} = P_0$$

otherwise, it is **composite**. If $P_{\mathbf{X}}$ is parametric with more than one parameter, a composite hypothesis might specify the values of some or all of them. (e.g. one regression coefficient)

The distribution of \mathbf{X} under H_0 , P_0 , is called null distribution.

Examples of hypotheses

- A significant trigger in a gravitational wave detector is due to instrumental fluctuations. This is a composite hypothesis as the distribution of triggers under the noise assumption is not fully specified.
- The numbers of gravitational wave events x_1, \dots, x_7 observed on Monday, \dots , Sunday. The null hypothesis is that all days are equally likely, i.e., the joint distribution is Multinomial($n; \frac{1}{7}, \dots, \frac{1}{7}$). This is a simple hypothesis.
- The right ascensions x_1, \dots, x_n angles of observed gravitational wave events. The hypothesis that the X_j 's are independently Uniform on $[0, 2\pi)$ is simple.

Suppose we want to test that there is clustering around some angle, then we can assume that the distribution is von Mises with pdf

$$p(x|\theta, \lambda) = \frac{1}{2\pi I_0(\lambda)} e^{\lambda \cos(x-\theta)}, \quad x \in \mathcal{X} = [0, 2\pi); \lambda \geq 0, 0 \leq \theta < 2\pi;$$

for unknown λ . This is a composite hypothesis.

- The hypothesis that the number of gravitational wave events in each month X_1, \dots, X_n are independently Poisson(θ) with unknown θ is composite.

3.1 Definitions and basic concepts

1. A sample of n observations is available to make inference about parameter θ .
2. We wish to decide between two hypotheses: H_0 , the *null hypothesis*, and H_1 , the *alternative hypothesis*.

H_0 is often *simple* (only one value is specified for θ)

$$\text{i.e. } H_0 : \theta = \theta_0 \text{ (e.g. } H_0 : \mu = 100, H_0 : p = \frac{1}{2}\text{)}.$$

H_1 can be *simple*: $H_1 : \theta = \theta_1$ but more commonly it is *composite* (more than one value is allowed for θ). The most common alternatives are

$$\begin{aligned} H_1 : \theta < \theta_0 \quad \text{or} \quad H_1 : \theta > \theta_0 & \text{--- one-sided/one-tailed alternative} \\ \text{or } H_1 : \theta \neq \theta_0 & \text{--- two-sided/two-tailed alternative.} \end{aligned}$$

3. Two possible decisions: *to reject* or *not to reject* H_0 in favour of H_1 .

The decision whether or not to reject H_0 is based on the value of a *test statistic*, which is a function of the observations.

4. Values of the test statistic for which H_0 is not rejected form the *acceptance region*, \bar{C} . Values of the test statistic for which H_0 is rejected form the *rejection region* (or *critical region*), C .

The form of these regions depends on the form of H_1 .

5. There are two possible types of error:

$$\begin{aligned} \text{Reject } H_0 \text{ when } H_0 \text{ is true} & \quad \text{--- Type I error} \\ \text{Fail to reject } H_0 \text{ when } H_0 \text{ is false} & \quad \text{--- Type II error} \end{aligned}$$

The probability of Type I error, denoted by α , is the **significance level** (or **size**) of the test.

The probability of Type II error, denoted by β , is only defined uniquely if H_1 is simple. In which case

$$\eta = 1 - \beta \text{ is the } \mathbf{power} \text{ of the test.}$$

For composite H_1 , $\eta(\theta)$ is the *power function*.

Generally we consider Type-I error (false rejection) to be worse than Type-II (incorrect failure to reject) as usually in the latter case more data will be collected and the test will be re-evaluated. It is therefore usual to specify the **significance level** of the test in order to determine the threshold for rejection, or the quote a **p-value** (see next section) when quoting test results.

We can define a **test function** $\phi(x)$ such that

$$\phi(x) = \begin{cases} 1 & \text{if } t(\mathbf{x}) \in C \\ 0 & \text{if } t(\mathbf{x}) \in \bar{C} \end{cases}$$

and when we observe $\phi(\mathbf{X}) = 1$, we reject H_0 . This function has the property that $\alpha = \mathbb{E}_{H_0}(\phi(\mathbf{X}))$ and $\eta = \mathbb{E}_{H_1}(\phi(\mathbf{X}))$, in which the subscript denotes the hypothesis under which the expectation value is to be calculated.

For discrete distributions, the probability that the test statistic lies on the boundary of the critical region, ∂C , may be non-zero. In that case, it is sometimes necessary to use a **randomized test**, for which the test function is

$$\phi(x) = \begin{cases} 1 & \text{if } t(\mathbf{x}) \in C \\ \gamma(\mathbf{x}) & \text{if } t(\mathbf{x}) \in \partial C \\ 0 & \text{if } t(\mathbf{x}) \in \bar{C} \end{cases}$$

for some function $\gamma(\mathbf{x})$ and we reject H_0 based on observed data \mathbf{x} with probability $\phi(\mathbf{x})$.

3.2 Test statistic

Often to construct a test (i.e. the decision whether to reject H_0 or not based on observed data \mathbf{x}), a *test statistic* is used.

Definition 10. A real-valued function $t(\mathbf{x})$ on \mathcal{X} is a test statistic for testing H_0 iff

- (i) values of t are **ordered** with respect to the evidence for departure from H_0
- (ii) the distribution of $T = t(\mathbf{X})$ under H_0 is known, at least approximately. For composite H_0 the distribution should be (approximately) the same for all simple hypotheses making up H_0 .

For any observation \mathbf{x} , we measure the consistency of \mathbf{x} with H_0 using the *significance probability* or the *p-value*, e.g. if larger values of t correspond to stronger evidence for departure from H_0 , the p-value is defined by

$$p = \mathbb{P}(T \geq t(\mathbf{x}) | H_0),$$

the probability (under H_0) of seeing the observed value of t or any more extreme value. The smaller the value of p the greater the evidence against H_0 .

3.3 Alternative hypothesis

Can be specified or unspecified.

3.3.1 Pure significance tests

In a *pure significance test*, only the null hypothesis H_0 is explicitly specified. The p-value of the observed value under the null distribution is evaluated, and if it is sufficiently small, the null hypothesis would be rejected. Such tests are done if we want to avoid specifying a parametric family of alternative distributions.

There will often be multiple quantities that could be computed under the null hypothesis and we can choose any of them to evaluate the distribution of the test statistic. The best choice can be guided if we have a specific idea of the type of departure from H_0 we are looking for, e.g.,

- Directional data: Might look for a tendency for the observed directions to cluster about a (possibly unknown) direction. But not a specific set of alternatives such as von Mises distributions.
- $\text{Pois}(\theta)$: if the alternative is not a Poisson distribution, we might test whether variance \neq expectation.

An important class of pure significance tests are *goodness of fit* tests where either the sample distribution function $\hat{P}_X(x) = \frac{1}{n} \sum_{i=1}^n I(x \leq x_i)$ or the histogram are compared to those of the null distribution.

Examples

- Event frequency on different days: $H_0 : X_1, \dots, X_7 \sim \text{Mult}(n; \frac{1}{7}, \dots, \frac{1}{7})$.

With no particular alternative we might use Pearson's χ^2 test, comparing

$$X^2 = \sum_{i=1}^7 \frac{(x_i - \frac{n}{7})^2}{\frac{n}{7}} \quad \text{with} \quad \chi_6^2.$$

- Right ascension of GW sources: If alternative to H_0 is clustering about the reference direction (e.g. galactic centre) we could use $\sum \cos x_j$, the projection onto the reference axis of the resultant sum vector ($\sum \cos x_j, \sum \sin x_j$).
- $\text{Pois}(\theta)$: might use index of dispersion,

$$d = \frac{\sum (x_i - \bar{y})^2}{\bar{y}},$$

which is approximately χ^2 with $(n - 1)$ degrees of freedom under H_0 for $\theta \geq 1$.

Note that given $\sum X_j = s$, the distribution of X_1, \dots, X_n is $\text{Mult}(s, \frac{1}{n}, \dots, \frac{1}{n})$ and d is the χ^2 statistic for testing the fit of this distribution.

3.3.2 Specified alternative hypothesis

For a parametrised family of distributions $p(x|\theta)$, $\theta \in \Theta$, say $H_0 : \theta = \theta_0$, then

$$H_1 : \theta \in \Theta_1 \subset \Theta \setminus \{\theta_0\},$$

e.g. $\theta \neq \theta_0$ (two-sided), $\theta > \theta_0$ or $\theta < \theta_0$ (one-sided).

Below we consider two cases: with simple and composite alternative hypotheses (and a simple null hypothesis).

With composite alternative hypotheses, the power of the test becomes the power function defined over $\theta \in \Theta_1$:

$$\eta(\theta) = \mathbb{P}(\text{reject } H_0 | \theta) = \mathbb{P}_\theta(\text{reject } H_0).$$

3.4 Critical regions

In § 3.2 we defined for each $\mathbf{x} \in X$ the significance probability

$$p = \mathbb{P}(T \geq t(\mathbf{x}) | H_0)$$

associated with a test statistic t . A different, but equivalent, approach defines a test using critical regions rather than test statistics. This

- facilitates comparison of different tests of H_0 according to their properties under H_1 ;
- is useful for establishing a connection between tests and confidence regions.

For any α in the interval $(0, 1)$, a subset R_α of X is a **critical region of size α** if

$$\mathbb{P}(\mathbf{X} \in R_\alpha | H_0) = \alpha \tag{54}$$

Interpretations of R_α :

- (i) points in R_α are regarded as not consistent with H_0 at level α ;
- (ii) points in R_α are “significant at level α ”;
- (iii) if $\mathbf{x} \in R_\alpha$, then H_0 is “rejected” in a test of size α .

A significance test is defined by a set of critical regions $\{R_\alpha : 0 < \alpha < 1\}$ satisfying

$$R_{\alpha_1} \subset R_{\alpha_2} \quad \text{if } \alpha_1 < \alpha_2. \quad (55)$$

Thus, for example, if data \mathbf{x} are significant at the 1% level, they are also significant at the 5% level.

The **significance probability** (also called p-value) for data \mathbf{x} is then defined as

$$P = \inf(\alpha; \mathbf{x} \in R_\alpha),$$

i.e. the smallest α for which \mathbf{x} is significant at level α .

The definition of a test in §3.2 corresponds to critical regions of the form

$$R_\alpha^t = \{\mathbf{x} : t(\mathbf{x}) \geq t_\alpha\},$$

where t_α is the upper α point of $T = t(\mathbf{X})$ under H_0 , since

$$\mathbb{P}(\mathbf{X} \in R_\alpha^t | H_0) = \mathbb{P}(t(X) \geq t_\alpha | H_0) = \alpha,$$

by the definition of t_α ; also if $\alpha_1 < \alpha_2$ then $t_{\alpha_1} > t_{\alpha_2}$ and $R_{\alpha_1}^t \subset R_{\alpha_2}^t$ satisfying (55). Finally,

$$\begin{aligned} P &= \mathbb{P}(t(\mathbf{X}) \geq t(\mathbf{x}) : H_0) \\ &= \inf(\alpha; t(\mathbf{x}) \geq t_\alpha) \\ &= \inf(\alpha; \mathbf{x} \in R_\alpha^t), \end{aligned}$$

the smallest α for which \mathbf{x} is significant at level α .

Example

- X_j independent $N(\mu, \sigma^2)$ (σ known and hence =1 without loss of generality) To test $H_0 : \mu = \mu_0$ vs $\mu > \mu_0$, obvious test statistics are \bar{Y} or $(\bar{Y} - \mu_0)\sqrt{n}$. The significance probability is

$$P = \mathbb{P}((\bar{Y} - \mu_0)\sqrt{n} > (\bar{y} - \mu_0)\sqrt{n} | H_0) = 1 - \Phi((\bar{y} - \mu_0)\sqrt{n}).$$

The corresponding critical regions are $R_\alpha = \{\mathbf{x} : (\bar{y} - \mu_0)\sqrt{n} \geq \Phi^{-1}(1 - \alpha)\}$. Thus

$$\mathbb{P}(\mathbf{X} \in R_\alpha | H_0) = \mathbb{P}((\bar{Y} - \mu_0)\sqrt{n} \geq \Phi^{-1}(1 - \alpha)) = \alpha,$$

as required, and if $\alpha_1 < \alpha_2$, then $\Phi^{-1}(1 - \alpha_1) > \Phi^{-1}(1 - \alpha_2)$, so that $R_{\alpha_1} \subset R_{\alpha_2}$. Also

$$\begin{aligned} \inf(\alpha; \mathbf{x} \in R_\alpha) &= \inf(\alpha; (\bar{y} - \mu_0)\sqrt{n} \geq \Phi^{-1}(1 - \alpha)) \\ &= \inf(\alpha; \alpha \geq 1 - \Phi((\bar{y} - \mu_0)\sqrt{n})) \\ &= P. \end{aligned}$$

3.5 Construction of confidence intervals using critical regions

The construction of hypothesis tests leads naturally to the construction of confidence intervals and regions. For any value ψ_0 of ψ , let $R_\alpha(\psi_0)$ be a size- α critical region for testing the null hypothesis $\psi = \psi_0$ against $\psi \neq \psi_0$ (or possibly $\psi < \psi_0$ or $\psi > \psi_0$). For any \mathbf{x} define

$$S_\alpha(\mathbf{x}) = \{\psi_0 : \mathbf{x} \notin R_\alpha(\psi_0)\}.$$

Then $S_\alpha(\mathbf{X})$ is a $(1 - \alpha)$ confidence interval for ψ since

$$\mathbb{P}(S_\alpha(\mathbf{X}) \ni \psi_0; \psi_0, \lambda) = \mathbb{P}(\mathbf{X} \notin R_\alpha(\psi_0) : \psi_0, \lambda) = 1 - \alpha \quad \forall \psi_0, \lambda$$

[$\bar{R}_\alpha(\psi_0)$ comprises \mathbf{x} values judged consistent with ψ_0 (at level α), so $S_\alpha(\mathbf{x})$ comprises ψ values consistent with \mathbf{x} .]

If $\alpha_1 < \alpha_2$, then from (19) $\{\psi_0 : \mathbf{x} \in R_{\alpha_1}(\psi_0)\} \subset \{\psi_0 : \mathbf{x} \in R_{\alpha_2}(\psi_0)\}$, so that (53) holds.

For scalar ψ , critical regions for alternatives $\psi < \psi_0$ lead to upper confidence limits.

Example

- $\text{Exp}(\lambda)$: Find the best size- α critical region for testing $\lambda = \lambda_0$ against $\lambda < \lambda_0$.

The best size- α critical region for testing $\lambda = \lambda_0$ against $\lambda < \lambda_0$ is $R_\alpha(\lambda_0) = \{\mathbf{x} : \sum x_j > \frac{1}{2}\lambda_0^{-1}\chi_{2n}^2(\alpha)\}$. The corresponding $(1 - \alpha)$ confidence region for λ is $\{\lambda_0 : \sum x_j \leq \frac{1}{2}\lambda_0^{-1}\chi_{2n}^2(\alpha)\}$ i.e. $\{\lambda_0 : \lambda_0 \leq \frac{1}{2}(\sum x_j)^{-1}\chi_{2n}^2(\alpha)\}$, so that $\frac{1}{2}(\sum x_j)^{-1}\chi_{2n}^2(\alpha)$ is the $(1 - \alpha)$ upper confidence limit for λ .

3.6 Examples of hypothesis tests

We give three commonly encountered examples of hypothesis tests.

3.6.1 z-test

Suppose that we observe two independent samples

$$X_1, \dots, X_n \sim N(\mu_1, \sigma^2), \quad Y_1, \dots, Y_m \sim N(\mu_2, \sigma^2).$$

We assume additionally that σ^2 is known and we are interested in testing the hypothesis

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_1 : \mu_1 - \mu_2 \neq 0.$$

If the null hypothesis is violated we expect that the magnitude of the difference in sample means, $|\bar{X} - \bar{Y}|$, will be large. The statistic

$$Z = \left(\frac{1}{n} + \frac{1}{m}\right)^{-\frac{1}{2}} \frac{(\bar{X} - \bar{Y})}{\sigma}$$

follows a $N(0, 1)$ distribution under the null hypothesis so we use a critical region of the form

$$|z| > z_{\frac{\alpha}{2}}$$

to define a test with significance α . Here $z_{\frac{\alpha}{2}}$ denotes the upper $\alpha/2$ point in the Normal distribution, i.e., the point such that

$$\mathbb{P}(X \sim N(0, 1) > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}.$$

3.6.2 t-test

We now suppose that we want to test the same hypothesis as in the previous example, but assuming that σ^2 is not known. Once again, we expect the difference in sample means to be large when the null hypothesis is false, but exactly how large now depends on the unknown value of σ^2 . If we use the same test statistic, but with the known variance replaced by the estimated value we have

$$T = \left(\frac{1}{n} + \frac{1}{m} \right)^{-\frac{1}{2}} \frac{(\bar{X} - \bar{Y})}{\hat{\sigma}} \quad \text{where } \hat{\sigma}^2 = \frac{1}{m+n-2} \left(\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_j - \bar{Y})^2 \right)$$

which follows a t_{m+n-2} distribution under the null hypothesis.

The critical region of a size- α test is to reject H_0 when

$$|t| > t_{\frac{\alpha}{2}},$$

where $z_{\frac{\alpha}{2}}$ denotes the upper $\alpha/2$ point in the t-distribution with $m+n-2$ degrees of freedom.

3.6.3 Analysis of variance: F-test

Suppose we have observations of random variables X_{ij} where $j = 1, \dots, n_i$ labels different observations of one particular group, and $i = 1, \dots, k$ labels the different groups. We denote the mean in each group by

$$\bar{X}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$$

and the overall mean by

$$\bar{X}_{\bullet\bullet} = \frac{1}{N} \sum_{ij} X_{ij}, \quad N = \sum_{i=1}^k n_i.$$

We are interested in testing that the means of all the groups are equal. If this is true then we expect that the **between samples sum of squares**

$$SS_b = \sum_i n_i (\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2$$

is comparable to the **within samples sum of squares**

$$SS_w = \sum_{ij} (x_{ij} - x_{i\bullet})^2.$$

If the means are different then we expect the former to be larger than the latter. Therefore, we reject the null hypothesis for large values of SS_b/SS_w . The quantity

$$F = \frac{(N-k)SS_b}{(k-1)SS_w}$$

follows an $F_{k-1, N-k}$ -distribution under the null hypothesis and so our critical regions are of the form to reject H_0 when

$$F > F_{k-1, N-k}(\alpha)$$

the upper α critical point of the $F_{k-1, N-k}$ distribution.

3.7 Calculating thresholds for tests

For the examples above the test statistics followed known distributions under the null hypothesis and so the critical values can be directly calculated. This is not always possible. In other situations it might be possible to compute the mean, μ , and variance, σ^2 , of the test statistic, if not its full distribution. In that case, a Normal approximation can often be used by appealing to the Central Limit Theorem.

Example: $\mathcal{E}(\lambda)$: we saw above that $X = \sum x_j$ can be used for testing $\lambda = \lambda_0$ versus $\lambda < \lambda_0$. While in this case we know the exact distribution of the test statistic, if we did not we can approximate

$$X \sim N\left(\frac{n}{\lambda_0}, \frac{n}{\lambda_0^2}\right)$$

and reject the hypothesis at significance α if

$$\frac{\lambda_0 X - n}{\sqrt{n}} > z_\alpha.$$

The power of the test can be approximated in a similar way, by writing down a Normal approximation to the distribution of the test statistic under the alternative hypothesis.

If the mean and variance cannot be easily calculated, or the form of the test statistic does not lend itself to approximation by the Central Limit Theorem, then usually the best approach is to do a **simulation study**, i.e., generate many realisations of the test statistic under H_0 and determine thresholds numerically. In principle, the power of the test can be evaluated in a similar way although this might not be practical for composite alternative hypotheses.

3.8 Multiple testing

When presented with new data, there is a temptation to keep asking different questions of the same data. When doing this you have to be careful to avoid **multiple testing** (or, in the language of the gravitational wave community **trials factors**). If you keep carrying out independent tests that have a significance of α then you would expect to reject a hypothesis every $1/\alpha$ tests purely by chance. Therefore, if you plan to carry out m independent tests and want the overall significance to be α , the significance levels applied to the individual tests must be lower.

If we carry out m independent tests, each with significance α , then the combined significance is

$$1 - (1 - \alpha)^m = \alpha_c.$$

To reach a target significance of the combined tests requires using individual tests with significance $\alpha = 1 - (1 - \alpha_c)^{1/m} = 1 - \exp(\log(1 - \alpha_c)/m) \approx \alpha_c/m$. The first expression is the *Sidak correction*, while the latter correction is referred to as the *Bonferroni correction*.

It is also possible to not divide the total significance evenly between the different individual tests. The *Holm-Bonferroni method* orders the individual test p -values and then tests the i 'th (starting from the smallest) at a significance level of $\alpha_c/(m - i + 1)$. This approach gives better overall performance.

In practice, multiple tests on the same data will not be independent and so using the corrections based on independence will be conservative and the true significance of any

rejection of the null hypothesis will be greater (i.e., the true p-value will be smaller than that estimated in this way). Understanding the dependency of multiple tests is typically highly non-trivial so it is usually best to assess the true p-value of a testing programme using simulations.

Another issue to be cautious of is changing the question based on the data. Changing the question based on what was observed can lead to results appearing significant when they are not, as the following example illustrates.

Example: LIGO/Virgo operate for 8 months from January to August and sees event counts (1, 0, 0, 0, 0, 1, 1, 4). Are the 4 events in the last month unusual? A total of 7 events have been observed in 8 months, so we have a rate of $\sim 7/8$ per month. Assuming that the events are Poisson distributed with this rate, the probability that a given month would have 4 or more events in it is $\sim 1.2\%$, which would be significant at the 5% level usually used for hypothesis tests. But it is not fair to ask “Is four events in August unusual?”, since we only decided to look at August in particular when we saw the data. The fair question to ask is “Is four events in one of the months unusual”, which means we must multiply by 8 to account for the fact that we have 8 potentially unusual months to choose from. The resulting probability of $\sim 9.8\%$ is much less significant ¹. Note that it is perfectly fine, having made these observations, to ask “Is August unusual in the next observing run?” and specifically target the month that was an outlier in previous data in the next analysis. However, this is less sensitive than doing the test “Is any month unusual?” on all of the data from both observing runs together. Suppose in the next year we also take data from January to August and observe events (0, 1, 0, 1, 1, 0, 0, 2). The probability of observing two or more events in August, given the rate of $5/8$ events per month, is 13%, so this would not be considered significant. However, adding the two observing runs together we have (1, 1, 0, 1, 1, 1, 1, 6) and the rate for binned observations is $4/3$. The probability of seeing 6 or more events in a Poisson distribution with rate $4/3$ is 0.25%, which is significant ².

3.9 Receiver operator characteristic

As mentioned above, Type-I errors are considered to be more serious than Type-II errors and so tests are quoted by the significance level. However, there may be (infinitely) many tests with the same significance, so how do we choose between them? This is done using the power function. Clearly if one test is more powerful than another for the same significance level then it is better and should be used.

In general, one way to compare different tests is by plotting a **receiver operator characteristic** (ROC) curve. This is a plot of the power versus significance of a test, or equivalently the “detection rate” of deviations in the null hypothesis against the “false alarm rate”. For a random test, i.e., we toss a coin and, regardless of the observed data, say that if it is heads we have made a detection, the ROC curve is the diagonal line. Tests that lie above the line are more powerful than random at given significance, and so the further away from the diagonal line the better the test is. ROC curves can be used to compare tests visually, or

¹Another way to tackle this problem is to say that we expect the distribution of events across the 8 months to be Multinomial with equal probability of 0.125 in each month. The distribution of events in a specific month is Binomial with $n = 7$ and $p = 0.125$ and so the probability that a specific event will have four or more events out of the 7 is $\sim 0.6\%$, but this rises to $\sim 5.0\%$ when we compute the probability that one (unspecified) month has four or more events.

²In the multinomial analysis the probabilities are 12% and 0.18% respectively

by computing the area between the curve and the diagonal line. Sometimes the curves can cross, so one test may be better at one significance level and another at another. The best test then depends on what regime you are operating in.

In the following subsections we will present a number of results that describe how to find tests that have the highest power at a given significance, under various assumptions about the hypotheses and the underlying distributions. As we shall see below, it is not always possible to find a test that is the best everywhere.

3.10 Designing the best test: simple null and alternative hypotheses

Consider null and alternative hypotheses H_0 , H_1 corresponding to completely specified p.d.f.'s p_0 , p_1 for \mathbf{X} . For these hypotheses, comparison between the critical regions of different tests is in terms of

$$\mathbb{P}(\mathbf{X} \in R_\alpha | H_1)$$

the **power** of a size- α critical region R_α for alternative H_1 . A **best** critical region of size α is one with maximum power.

In terms of p_0 , p_1 , the power is

$$\begin{aligned} \int_{R_\alpha} p_1(\mathbf{x}) d\mathbf{x} &= \int_{R_\alpha} p_0(\mathbf{x}) r(\mathbf{x}) d\mathbf{x} \quad \left(\text{or } \sum_{R_\alpha} p_0(\mathbf{x}) r(\mathbf{x}) \right) \\ &= \mathbb{E}\{r(\mathbf{X}) | \mathbf{X} \in R_\alpha; H_0\} \end{aligned}$$

where

$$r(\mathbf{x}) = \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} = \frac{L(\theta; H_1)}{L(\theta; H_0)},$$

the **likelihood ratio** (LR) for H_1 vs H_0 . We can **prove** that the power is maximized when R_α has the form $\{\mathbf{x} : r(\mathbf{x}) \geq k_\alpha\}$ or $\{\mathbf{x} : \frac{L(\theta; H_1)}{L(\theta; H_0)} \geq k_\alpha\}$, i.e. when R_α is a LR critical region. Thus we have the Neyman-Pearson lemma.

Theorem 4. (Neyman-Pearson lemma). *For any size α , the LR critical region is the best critical region for testing simple hypotheses H_0 vs H_1 . (It is also better than any critical region of size $< \alpha$.)*

A LR test is a test whose critical regions are LR critical regions for all α for which such a size- α region exists (all α in the continuous case).

Examples

- Angles: If H_0 , H_1 correspond to a Uniform distribution and a von Mises distribution with parameter θ_1 , the LR is

$$r(\mathbf{x}) = \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} = \{2\pi I_0(\theta_1)\}^{-n} \frac{e^{\theta_1 \sum_j \cos x_j}}{(2\pi)^{-n}},$$

which is an increasing function of $t(\mathbf{x}) = \sum \cos x_j$. So the LR critical regions have the form $\{\mathbf{x} : \sum \cos x_j > t_\alpha\}$. For any α , t_α is given by $\mathbb{P}(\sum \cos X_j \geq t_\alpha | H_0) = \alpha$. From §3.3 $\sum \cos X_j$ is approximately $N(0, \frac{1}{2}n)$ under H_0 , so t_α is approximately $(\frac{1}{2}n)^{1/2} \Phi^{-1}(1 - \alpha)$. Note that the critical regions, and hence the test, do not depend on the value of θ_1 .

- $\mathcal{E}(\lambda) : X_1, \dots, X_n$ are i.i.d. with d.f. $1 - e^{-\lambda y}$ ($y > 0$). H_0 is $\lambda = \lambda_0$; H_1 is $\lambda = \lambda_1 < \lambda_0$

$$r(\mathbf{x}) = \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} = \left(\frac{\lambda_1}{\lambda_0}\right)^n \exp\{(\lambda_0 - \lambda_1) \sum x_j\},$$

which is increasing in $\sum x_j$. So the test is based on $\sum x_j$ or $2\lambda_0 \sum X_j$, which is χ_{2n}^2 under H_0 , and the critical regions are $\{\mathbf{x} : \sum x_j > \frac{1}{2}\lambda_0^{-1}\chi_{2n}^2(\alpha)\}$, where $\chi_{2n}^2(\alpha)$ is the upper α point of χ_{2n}^2 . The power is

$$\begin{aligned} \mathbb{P}(2\lambda_0 \sum X_j > \chi_{2n}^2 | H_1) &= \mathbb{P}\left(2\lambda_1 \sum X_j > \frac{\lambda_1}{\lambda_0} \chi_{2n}^2(\alpha) | H_1\right) \\ &= Q_{2n}\left(\frac{\lambda_1}{\lambda_0} \chi_{2n}^2(\alpha)\right) \end{aligned}$$

where Q_{2n} is 1- distribution function for χ_{2n}^2 .

For comparison, we might base a test on $x_{(1)}$, which has distribution function $1 - e^{-n\lambda y}$; size α critical regions are given by $\{\mathbf{x} : x_{(1)} > -(n\lambda_0)^{-1} \ln \alpha\}$, and the power is $\alpha^{\lambda_1/\lambda_0}$, which is $< Q_{2n}\left(\frac{\lambda_1}{\lambda_0} \chi_{2n}^2(\alpha)\right)$ for $n > 1$ and $\lambda_1 < \lambda_0$, and does not depend on n .

3.11 Designing the best test: simple null and composite alternative hypotheses

Suppose now there is a parametric family $\{p(\mathbf{x}|\theta) : \theta \in \Theta_1\}$ of alternative p.d.f.'s for \mathbf{X} . The power of a size- α critical region R_α generalizes to the size- α **power function**

$$\begin{aligned} \text{pow}(\theta; \alpha) &= \mathbb{P}(\mathbf{X} \in R_\alpha | \theta) \\ &= \int_{R_\alpha} p(\mathbf{x}|\theta) dy \quad \left(\text{or } \sum_{R_\alpha} p(\mathbf{x}|\theta) dy\right) \quad (\theta \in \Theta_1). \end{aligned}$$

A size- α critical region R_α is then **uniformly most powerful size α** (UMP size α) if it has maximum power uniformly over Θ_1 . A test is UMP if all its critical regions are UMP. More formally

Definition 11. A **uniformly most powerful** or **UMP test**, $\phi_0(\mathbf{X})$, of size α is a test $t(\mathbf{x})$ for which

$$(i) \quad \mathbb{E}_\theta \phi_0(\mathbf{X}) \leq \alpha \quad \forall \theta \in \Theta_0;$$

$$(ii) \quad \text{given any other test } \phi(\cdot) \text{ for which } \mathbb{E}_\theta \phi(\mathbf{X}) \leq \alpha \quad \forall \theta \in \Theta_0, \text{ we have } \mathbb{E}_\theta \phi_0(\mathbf{X}) \geq \mathbb{E}_\theta \phi(\mathbf{X}) \quad \forall \theta \in \Theta_1.$$

Such tests cannot be found in general, as this requires that the Neyman-Pearson test should be the same for every pair of simple hypotheses. However, for one sided testing problems, i.e., tests of the form $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$, there are a wide class of parametric families for which UMP tests exist. These are distributions that have **monotone likelihood ratio** or MLR.

Definition 12. The family of densities $\{p(\mathbf{x}|\theta), \theta \in \Omega_\theta \subseteq \mathbb{R}\}$ with real scalar parameter θ is said to be of **monotone likelihood ratio** if there exists a function $s(\mathbf{x})$ such that the likelihood ratio

$$\frac{p(\mathbf{x}|\theta_2)}{p(\mathbf{x}|\theta_1)}$$

is a non-decreasing function of $s(\mathbf{x})$ whenever $\theta_1 < \theta_2$.

Note that the same result applies for a non-increasing test statistic, by replacing $t(\mathbf{x})$ by $-t(\mathbf{x})$.

Theorem 5. Suppose \mathbf{X} has a distribution from a family that is monotone likelihood ratio with respect to some continuous test statistic $s(\mathbf{X})$ and we wish to test $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$, then a UMP test exists with critical region of the form $s \geq s_\alpha$.

Proof. For testing $\theta = \theta_0$ against $\theta = \theta_1$ for any specific $\theta_1 \in \Theta_1$, the Neyman-Pearson lemma tells us that the most powerful critical region is given by the likelihood ratio critical region. The LR is a non-decreasing function of $s(\mathbf{y})$ for any $\theta_1 > \theta_0$, and so the critical region is of the form $s \geq s_\alpha$. s_α is determined by the size of the test and depends only on θ_0 . Hence, this critical region is identical for all $\theta_1 \geq \theta_0$ and this test is UMP. \square

Corollary 2. If X_1, \dots, X_n are i.i.d with p.d.f. of the form

$$p(x|\theta) = \exp\{a(x)b(\theta) + c(\theta) + d(x)\}$$

with θ a scalar parameter and $b(\theta)$ strictly increasing, then for testing the null hypothesis that $\theta = \theta_0$ against $\theta > \theta_0$ the LR test has critical regions corresponding to large values of $s = \sum a(x_j)$ and is UMP.

Proof For any $\theta_1 > \theta_0$, the LR is

$$\frac{p_{\mathbf{X}}(\mathbf{x}|\theta_1)}{p_{\mathbf{X}}(\mathbf{x}|\theta_0)} = \exp[\{b(\theta_1) - b(\theta_0)\}s + n\{c(\theta_1) - c(\theta_0)\}].$$

Since $b(\theta_1) > b(\theta_0)$, this is monotone likelihood ratio and so the conditions of Theorem 5 are satisfied. This applies to all one-parameter exponential families, e.g. Normal, Binomial, Poisson. There are similar results for $\theta < \theta_0$, when $b(\theta)$ is a decreasing function.

Example.

- Angles : take H_0 to be that angles X_1, \dots, X_n are i.i.d. and Uniform on $[0, 2\pi)$.

A set of alternatives representing a type of symmetrical clustering about $y = 0$ has the X_j i.i.d. with von Mises p.d.f.

$$\frac{\exp(\theta \cos x)}{2\pi I_0(\theta)} \quad (0 \leq x < 2\pi; \theta > 0).$$

So we test the hypothesis $H_0 : \theta = 0$ against the alternative $\theta > 0$.

3.12 Designing the best test: composite null and alternative hypotheses

3.12.1 One-sided tests

Previously we considered tests of hypotheses where the null hypothesis was simple. Testing composite hypotheses is more complex in general. However, the above result for monotone likelihood ratio distributions also applies to one-sided tests of the form $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$.

Theorem 6. *Suppose \mathbf{X} has a distribution from a family that is monotone likelihood ratio with respect to some continuous test statistic $s(\mathbf{X})$ and we wish to test $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$, then*

(a) *The test*

$$\phi_0(\mathbf{x}) = \begin{cases} 1 & \text{if } s(\mathbf{x}) > s_0, \\ 0 & \text{if } s(\mathbf{x}) \leq s_0, \end{cases} \quad (56)$$

is UMP among all tests of size $\leq \mathbb{E}_{\theta_0} \{\phi_0(\mathbf{X})\}$.

(b) *Given some $0 < \alpha \leq 1$, there exists an s_0 such that the tests in (a) has size exactly equal to α .*

Proof. 1. From Theorem 5, ϕ_0 is UMP for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

2. $\mathbb{E}_\theta \{\phi_0(\mathbf{x})\}$ is a non-decreasing function of θ . If we have $\theta_2 < \theta_1$ and $\mathbb{E}_{\theta_2} \{\phi_0(\mathbf{x})\} = \beta$, then the trivial test $\phi(\mathbf{x}) = \beta$ has $\mathbb{E}_{\theta_1} \{\phi(\mathbf{x})\} = \beta$. The test ϕ_0 is UMP for testing θ_2 against θ_1 and so it must be at least as good as ϕ , i.e., $\mathbb{E}_{\theta_1} \{\phi_0(\mathbf{x})\} \geq \beta$. Hence, if we construct the test with $\mathbb{E}_{\theta_0} \{\phi_0(\mathbf{x})\} = \alpha$, then $\mathbb{E}_\theta \{\phi_0(\mathbf{x})\} \leq \alpha$ for all $\theta \leq \theta_0$, so ϕ_0 is also of size α under the larger hypothesis $H_0 : \theta \leq \theta_0$.

3. For any other test ϕ that is of size α under H_0 , we have $\mathbb{E}_{\theta_0} \{\phi(\mathbf{x})\} \leq \alpha$ and by the Neyman-Pearson lemma $\mathbb{E}_{\theta_1} \{\phi(\mathbf{x})\} \leq \mathbb{E}_{\theta_1} \{\phi_0(\mathbf{x})\}$ for any $\theta_1 > \theta_0$. This shows that this test is UMP among all tests of its size.

4. If α is specified we must show that there exists a s_0 such that $\mathbb{P}_{\theta_0} \{s(\mathbf{X}) > s_0\} = \alpha$, but this follows from the assumption that $s(\mathbf{X})$ is continuous. □

3.12.2 Two-sided tests

In more general situations we will be interested in testing hypotheses of the form $H_0 : \theta \in \Theta_0$, where Θ_0 is either an interval $[\theta_1, \theta_2]$ for $\theta_1 < \theta_2$ or a single point $\Theta_0 = \{\theta_0\}$, against the generic alternative $H_1 : \theta \in \Theta_1$, with $\Theta_1 = \mathbb{R}/\Theta_0$. For a family with monotone likelihood ratio with respect to a statistic $s(\mathbf{X})$, we might expect a good test to have a test function of the form

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } s(\mathbf{x}) > s_2 \text{ or } s(\mathbf{x}) < s_1, \\ \gamma(\mathbf{x}) & \text{if } s(\mathbf{x}) = s_2 \text{ or } s(\mathbf{x}) = s_1, \\ 0 & \text{if } s_1 < s(\mathbf{x}) < s_2. \end{cases}$$

Such a test is called a **two-sided test**. For such two-sided tests, we cannot usually find a UMP test. However, under certain circumstances it is possible to find a **uniformly most powerful unbiased** (UMPU) test.

Definition 13. A test $\phi(\mathbf{y})$ of $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is called **unbiased of size α** if

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta \{ \phi(\mathbf{Y}) \} \leq \alpha$$

and

$$\mathbb{E}_\theta \{ \phi(\mathbf{Y}) \} \geq \alpha \text{ for all } \theta \in \Theta_1.$$

In other words, an unbiased test is one which has higher probability of rejecting H_0 when it is false than when it is true. Note that if the power function is a continuous function of θ then an unbiased test of size α must have size equal to α on the boundary of the critical region (since the size is less than or equal to α within the critical region and greater than or equal to α outside).

Definition 14. A test which is uniformly most powerful among the set of all unbiased tests is called **uniformly most powerful unbiased**.

For a scalar exponential family of the form given in Corollary 2 the following theorem holds

Theorem 7. If X_1, \dots, X_n are i.i.d with p.d.f. of the form

$$p(x|\theta) = \exp\{a(x)b(\theta) + c(\theta) + d(x)\}$$

with θ a scalar parameter and $b(\theta)$ strictly increasing, then there exists a unique UMPU test of size α , ϕ' , for testing the hypothesis $H_0 : \theta \in [\theta_1, \theta_2]$, against the generic alternative $H_1 : \theta \in \mathbb{R} - [\theta_1, \theta_2]$, of the form

$$\phi'(\mathbf{x}) = \begin{cases} 1 & \text{if } s(\mathbf{x}) > s_2 \text{ or } s(\mathbf{x}) < s_1, \\ \gamma_j & \text{if } s(\mathbf{x}) = s_j, \\ 0 & \text{if } s_1 < s(\mathbf{x}) < s_2. \end{cases} \quad (57)$$

where $S = \sum a(x_j)$, for which

$$\mathbb{E}_{\theta_j} \phi'(\mathbf{X}) = \mathbb{E}_{\theta_j} \phi(\mathbf{X}) = \alpha, \quad j = 1, 2.$$

The boundaries of the critical region, s_1, s_2 , and the rejection probabilities on the boundaries, γ_1, γ_2 , are determined from the conditions $\mathbb{E}_{\theta_j} \phi'(\mathbf{X}) = \alpha$.

Example. Suppose a sample Y is drawn from an $\text{Exp}(\lambda)$ distribution, so that $f(y|\lambda) = \lambda \exp(-\lambda y)$. Construct a uniformly most powerful unbiased test of size $\alpha = 0.05$ of the hypothesis $H_0 : \lambda \in [1, 2]$ against the generic alternative $\lambda \in [0, 1) \cup (2, \infty)$.

For a single sample from the exponential distribution, the sufficient statistic is the observed value, y . Using the previous result, the UMPU test is of the form (57). The probability that $s = s_i$ is zero for any single value s_i and therefore the γ_i 's do not need to be determined. The boundaries of the critical region can be found from the constraints

$$\alpha = 0.05 = 1 - \exp(-s_1) + \exp(-s_2) = 1 - \exp(-2s_1) + \exp(-2s_2),$$

from which we find $s_1 = 0.02532$ and $s_2 = 3.6889$. The corresponding power function $\eta(\lambda)$ is shown in Figure 2. This shows that the test is unbiased as the probability of rejecting H_0 is less than or equal to the size α within the region defined by H_0 , it is equal to α on the boundary, and greater than α everywhere outside that region.

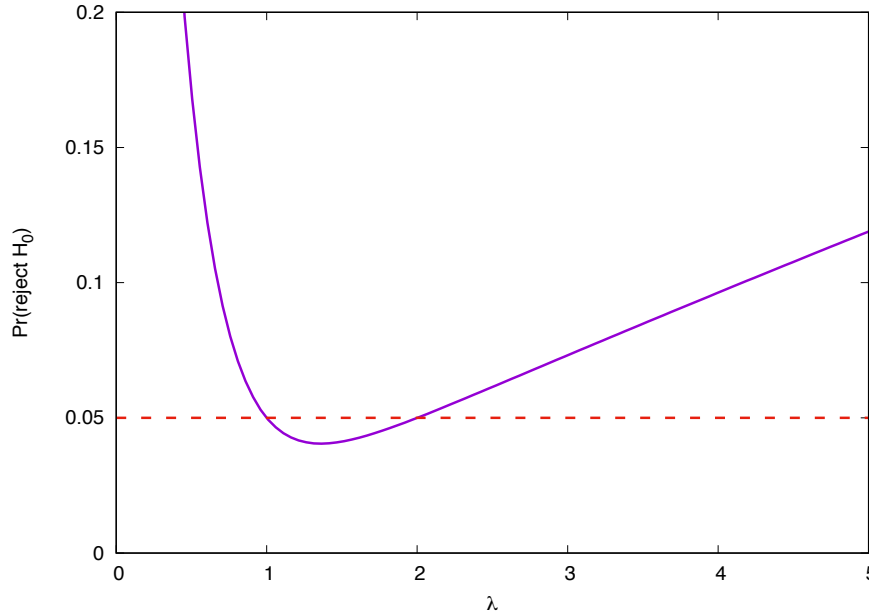


Figure 2: Power of the UMPU test of $\lambda \in [1, 2]$ against a generic alternative for an exponential distribution, as a function of λ , i.e., $\mathbb{P}_\lambda(\text{reject } H_0)$. The horizontal line indicates the size of the test, $\alpha = 0.05$.

3.12.3 Testing a point null hypothesis

A test of the null hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ can be considered as the limit of the preceding two-sided test when $\theta_2 - \theta_1 \rightarrow 0$. Therefore, as a corollary to the previous result, there must exist a unique UMPU test, ϕ' , of this hypothesis of the form (57) for which

$$\mathbb{E}_{\theta_0}\{\phi'(X)\} = \alpha, \quad \frac{d}{d\theta}\mathbb{E}_{\theta}\{\phi'(X)\}|_{\theta=\theta_0} = 0. \quad (58)$$

Differentiability of the power function for any test function is ensured from the assumption that the distribution is in the exponential family.

Example. Returning to the example of the preceding section of a single sample from an $\text{Exp}(\lambda)$ distribution, if we instead want to test the hypothesis that $\lambda = 1$ then we proceed as before, but the constraints on the boundary of the rejection region are now

$$\begin{aligned} \alpha &= 0.05 = 1 - \exp(-t_1) + \exp(-t_2), \\ 0 &= t_1 \exp(-t_1) - t_2 \exp(-t_2), \end{aligned}$$

which can be solved numerically to give $t_1 = 0.0423633$, $t_2 = 4.76517$. The power function is shown in Figure 3. We see that it reaches a minimum of $\alpha = 0.05$ at $\theta = \theta_0$ so it is unbiased and of size α as desired.

3.13 Designing the best test: similar Tests

So far we have focussed on tests of one-parameter distributions. However, often the distribution will depend on more than one parameter. In that case we are interested in tests

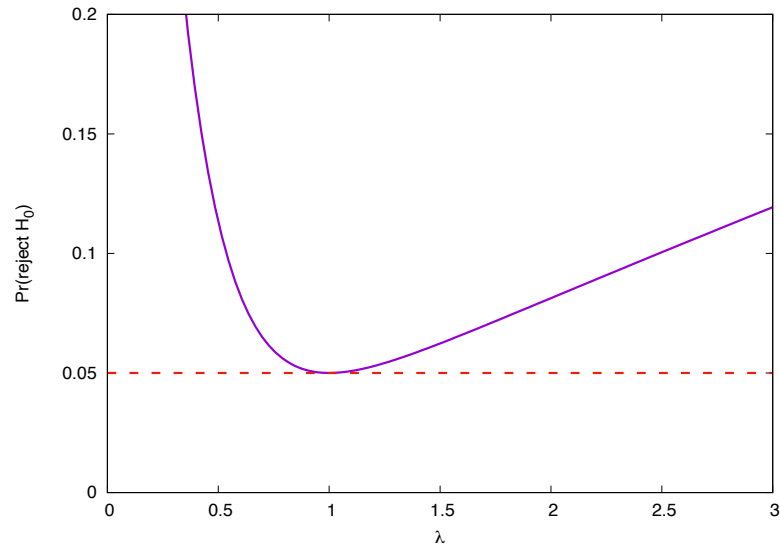


Figure 3: Power of the UMPU test of $\lambda = 1$ against a generic alternative for an exponential distribution, as a function of λ , i.e., $\mathbb{P}_\lambda(\text{reject } H_0)$. The horizontal line indicates the size of the test, $\alpha = 0.05$.

that perform as well as possible in inferring the value of one parameter of the distribution, irrespective of the value of the other parameters of the distribution. This gives rise to the notion of a **similar** test.

Definition 15. Suppose $\theta = (\psi, \lambda)$ and the parameter space is of the form $\Omega_\theta = \Omega_\psi \times \Omega_\lambda$. Suppose we wish to test the null hypothesis $H_0 : \psi = \psi_0$ against the alternative $H_1 : \psi \neq \psi_0$, with λ treated as a nuisance parameter. Suppose $\phi(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$ is a test of size α for which

$$\mathbb{E}_{\psi_0, \lambda} \{ \phi(\mathbf{x}) \} = \alpha \text{ for all } \lambda \in \Omega_\lambda.$$

Then ϕ is called a **similar test of size α** .

This definition can be extended to composite null hypotheses. If the null hypothesis is of the form $\theta \in \Theta_0$, where Θ_0 is a subset of Ω_θ , then a similar test is one for which $\mathbb{E}_\theta \{ \phi(\mathbf{x}) \} = \alpha$ on the boundary of Θ_0 .

If a test is uniformly most powerful among all similar tests then it is called **UMP similar**. There is close connection to UMPU tests. If the power function of a test is continuous then we saw earlier that any unbiased test of size α must have size exactly equal to α on the boundary, i.e., it must be similar. In such cases, if we can find a UMP similar test and it turns out to also be unbiased, then it is necessarily UMPU.

Moreover, in many cases it is possible to demonstrate that a test which is UMP among all tests based on the conditional distribution of a statistic S given the value of an ancillary statistic A , this test is UMP among all similar tests. In particular, this applies if A is a complete sufficient statistic for the variables λ .

One common situation in which this occurs is for multi-parameter exponential families, for which the likelihood can be written

$$p(x|\theta) = \exp \left\{ \sum_{i=1}^p A_i(x) B_i(\theta) + C(\theta) + D(x) \right\}.$$

Consider a test of the form $H_0 : B_1(\theta) \leq \theta_1^*$ against $H_1 : B_1(\theta) > \theta_1^*$. If we take $s(\mathbf{x}) = \sum_j A_1(x_j)$ and $A = (\sum_j A_2(x_j), \dots, \sum_j A_p(x_j))$, then the conditional distribution of S given A is also of the exponential form and doesn't depend on $B_2(\theta), \dots, B_p(\theta)$, so A is both sufficient and complete for $B_2(\theta), \dots, B_p(\theta)$. The Conditionality Principle suggests we should make inference about $B_1(\theta)$ based on the conditional distribution of S given A . Tests constructed in this way are UMPU (Ferguson 1967). The optimal one-sided test is then of the following form. Based on observations $s_1 = \sum_j A_1(x_j)$, $s_2 = \sum_j A_2(x_j), \dots, s_p = \sum_j A_p(x_j)$, we reject H_0 if and only if $s_1 > s_1^*$, where s_1^* is calculated from

$$\mathbb{P}_{B_1(\theta)=\theta_1^*} \{S_1 > s_1^* | S_2 = s_2, \dots, S_p = s_p\} = \alpha.$$

It can be shown this is a UMPU test of size α .

Similarly, to construct a two-sided test of $H_0 : \theta_1^* \leq B_1(\theta) \leq \theta_1^{**}$ against $B_1(\theta) < \theta_1^*$ or $B_1(\theta) > \theta_1^{**}$, we first define the conditional power function

$$w_{\theta_1}(\phi | s_2, \dots, s_p) = \mathbb{E}_{\theta_1} \{ \phi(S_1) | S_2 = s_2, \dots, S_p = s_p \}.$$

Then we can construct a two-sided conditional test of the form

$$\phi'(s_1) = \begin{cases} 1 & \text{if } s_1 < s_1^* \text{ or } s_1 > s_1^{**}, \\ 0 & \text{if } s_1^* \leq s_1 \leq s_1^{**}, \end{cases}$$

where s_1^* and s_1^{**} are chosen such that

$$w_{\theta_1}(\phi' | s_2, \dots, s_p) = \alpha \quad \text{when } B(\theta_1) = \theta_1^* \text{ or } B(\theta_1) = \theta_1^{**}.$$

It can be shown that these tests are also UMPU of size α . If the test is of a simple hypothesis $B(\theta_1) = \theta_1^*$ against the generic alternative $B(\theta_1) \neq \theta_1^*$ then the test is of the same form but the conditions are that the power function is equal to α and its derivative with respect to θ is equal to 0, as in Eq. (58).

3.14 Generalized likelihood ratio tests

In the previous sections we focussed on finding the “best” tests by one metric or another. However, as we have seen this is not always easy and the resulting test statistics are not always straightforward to evaluate. Under many circumstances, in the limit $n \rightarrow \infty$, the likelihood ratio follows a χ^2 distribution and so this can be used to construct a test that is valid asymptotically.

In particular, suppose we are testing $H_0 : \vec{\theta} \in \Theta_0$ versus $H_1 : \vec{\theta} \in \Theta_1$. We define the likelihood ratio

$$L_X(H_0, H_1) = \frac{\sup_{\vec{\theta} \in \Theta_1} p(x|\theta)}{\sup_{\vec{\theta} \in \Theta_0} p(x|\theta)}$$

and denote by $p = |\Theta_1 - \Theta_0|$ the difference in the numbers of degrees of freedom in the unknown parameters between the two hypotheses. Then as $n \rightarrow \infty$

$$2 \log L_X(H_0, H_1) \sim \chi_p^2$$

under H_0 and tends to be larger under H_1 . Therefore critical regions of the form $2 \log L_X > \chi_p^2(\alpha)$ give tests of approximately size α .

The interpretation of p is the number of constraints that have been placed to reduce the, typically more general, alternative hypothesis, to the more restrictive null hypothesis. For example, the null hypothesis might be specified by fixing the values of p of the parameters, or by imposing p linear constraints on the parameters, or by writing the k parameters of Θ_1 as functions of an alternative $k - p$ dimensional parameter space.