Hypothesis Testing

A statistical hypothesis is a statement about a parameter (or a statistical functional in nonparametric models).

A hypothesis test partitions the parameter space Θ into two disjoint sets Θ_0 and Θ_1, and produces a decision rule for choosing between

$$H_0 : \theta \in \Theta_0 \quad \text{and} \quad H_1 : \theta \in \Theta_1$$

H_0 is called the null hypothesis and H_1 is called the alternative hypothesis. The possible choices are

- Reject H_0
- Fail to reject H_0

The decision of whether to reject H_0 is determined by whether the sample $X = (X_1, \ldots, X_n)$ falls into a predefined rejection region R.

Usually, the rejection region R has the form

$$R = \{x_1, \ldots, x_n : T(x_1, \ldots, x_n) > c\}$$

where T is called a test statistic and c is called a critical value.

The idea is to construct R so that the probability of the data falling into it when H_0 is true is small.

Example: Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, and let $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ be the MLEs. If $H_0 : \mu = 0$, one test statistic we might consider is $T = |\hat{\mu}_n / \hat{\sigma}_n|$, reasoning that if H_0 is true, T will tend to be small.

Note: c is just a placeholder. It usually will depend on n and/or our choice of Θ_0 and Θ_1.

We evaluate a test using its power function. This is defined by

$$\beta(\theta) = P_\theta(X \in R)$$

Ideally, we would like $\beta(\theta)$ to be 0 when $\theta \in \Theta_0$ and 1 when $\theta \in \Theta_1$, but that is typically impossible to achieve.

Qualitatively, a good test has small $\beta(\theta)$ when $\theta \in \Theta_0$ and large $\beta(\theta)$ when $\theta \in \Theta_1$.

However, there are typically tradeoffs between the two, so that the researcher must choose between tests based on what kind of error probabilities he/she is willing to accept. More on this shortly.

Example 1: Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, where σ^2 is known. Consider testing $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$, using rejection region

$$R = \{x_1, \ldots, x_n : |\bar{X}_n| > c\}$$

Find and plot $\beta(\mu)$.

Example 2: Let $X \sim Bin(5, p)$. Consider testing $H_0 : p \leq 1/2$ versus $H_1 : p > 1/2$. Consider two different rejection regions:

$$\begin{aligned} R_1 &= \{x : x = 5\} \\ R_2 &= \{x : x \geq 3\} \end{aligned}$$

Plot and compare the corresponding power functions $\beta_1(p)$ and $\beta_2(p)$.

To make the problem of comparing tests better defined, we restrict ourselves to tests of a certain level, and then we try to find a test within that class that has large $\beta(\theta)$ for $\theta \in \Theta_1$.

A test is said to have level α if its size is less than or equal to α. The size of a test is defined to be

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$$

In words, the size of a test is the largest probability of rejecting H_0 when H_0 is true. This is called a Type I error.

	Fail to reject H_0	Reject H_0
H_0 true	Correct	Type I error
H_1 true	Type II error	Correct

Since H_0 usually represents a “default” hypothesis, first guaranteeing that the probability of this is small is a scientifically conservative strategy.

Continuation of Example 1: Find the size of the test as a function of c (and possibly other things). What should c be to produce a size α test?

Continuation of Example 2: Consider a rejection region of the form $R = \{x : x \geq c\}$.

- What values of c do we need to consider?
- For each of these, find the size of the corresponding test.
- What c should we choose if we want a probability of Type I error of no more than 10% (when H_0 is true)?

Correspondence between point-null tests and confidence sets. Practically speaking, this means that if we already have a $1-\alpha$ confidence interval for θ and we want to test $H_0 : \theta = \theta_0$, a level α test is just to reject H_0 if θ_0 falls outside the interval. Often we can get *approximate* $1-\alpha$ confidence intervals using an estimator of θ that is asymptotically normal.

Wald Test: Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Let $\hat{\theta}_n$ be an estimator such that $(\hat{\theta}_n - \theta_0)/\widehat{se}(\hat{\theta}_n) \overset{D}{\to} N(0,1)$. The size α Wald test rejects H_0 when $T > z_{\alpha/2}$, where

$$T = \left| \frac{\hat{\theta}_n - \theta_0}{\widehat{se}(\hat{\theta}_n)} \right|$$

We can show that asymptotically, the Wald test has size α, and that it is obtained by inverting the approximate $1-\alpha$ normal-based CI for θ.

Note that this method is quite general; we just need asymptotic normality.

- Consider a multi-parameter problem in which $\hat{\theta}_n$ is the MLE and g is a continuously differentiable function. Then we can form a Wald test based on
$$\frac{g(\hat{\theta}_n) - g(\theta_0)}{\widehat{se}(g(\hat{\theta}_n))} \overset{D}{\to} N(0,1)$$
where $\widehat{se}(g(\hat{\theta}_n))$ can be found using the Delta method.

- Consider the case that $\theta = T(F)$ for some unknown distribution F. If T is a linear functional, the plug-in estimator is a mean of iid random variables, so we can use the CLT. In the case of a nonlinear functional, if the plug-in estimator is asymptotically normal, we can use the bootstrap to approximate its standard error, $\widehat{se}(\hat{\theta}_n)$, and construct a Wald test statistic accordingly.

Examples

- Consider again $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where σ^2 is known. Show that the size α Wald test for $H_0 : \mu = 0$ produces a rejection region as in Example 1 above. (Actually the size is exactly α in this case).

- Now suppose that σ^2 is unknown. Construct a size α Wald test for $H_0 : \mu = 0$.

- Suppose that $X \sim Bin(m, p_1)$ and $Y \sim Bin(n, p_2)$. Construct a size α Wald test for $H_0 : p_1 = p_2$.

- Let $F(u, v)$ be the joint distribution of two r.v. U and V. Let $\theta = T(F) = \rho(U, V)$, where ρ denotes the correlation. Describe how to construct a size α Wald test for $H_0 : \rho = 0$ using the plug-in estimator and the bootstrap.

Likelihood ratio test (LRT)

Another broadly applicable class of tests is the likelihood ratio test (LRT). Let

$$T(X) = \frac{\sup_{\theta \in \Theta} \mathcal{L}_n(\theta)}{\sup_{\theta \in \Theta_0} \mathcal{L}_n(\theta)}$$

If $T(X)$ is large, it means there are values of θ in Θ_1 which are larger than for any in Θ_0. A likelihood ratio test is a test for which

$$R = \{x : T(x) > c\}$$

If $\hat{\theta}_n$ is the MLE and $\hat{\theta}_{n,0}$ is the MLE restricting $\theta \in \Theta_0$, then

$$T(X) = \frac{\mathcal{L}_n(\hat{\theta}_n)}{\mathcal{L}_n(\hat{\theta}_{n,0})}$$

Sometimes we can calculate the power function for the LRT exactly.

Example: Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Find $T(X)$ and find a simplified expression for the form of the rejection region. Use it to find the size α LRT.

When the power function can not be calculated exactly, and Θ_0 consists of fixing certain elements of θ (e.g., as in a point-null hypothesis), we can use the limiting distribution

$$\lambda(X) = 2 \log T(X) \overset{D}{\rightarrow} \chi^2_{r-q}$$

where r is the dimension of Θ and q is the dimension of Θ_0 .

Aside: The χ^2_k distribution (read "chi squared with k degrees of freedom") is the distribution of the sum of squares of k independent standard normal random variables. That is, if $Z_1, \ldots, Z_k \overset{iid}{\sim} N(0,1)$, then

$$Y = \sum_{i=1}^{k} Z_i^2 \sim \chi^2_k$$

We can use this approximation to find an appropriate critical value.

Example: Suppose $X_1, \ldots, X_n \overset{iid}{\sim} Poisson(\theta)$. Let $\hat{\theta}_n = \sum_{i=1}^n X_i/n$ be the MLE for θ. For testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, we have

$$\begin{aligned}\lambda &= 2\log\frac{\mathcal{L}(\hat{\theta}_n)}{\mathcal{L}(\theta_0)} \\ &= 2\log\frac{e^{-n\hat{\theta}_n}\hat{\theta}_n^{\sum x_i}}{e^{-n\theta_0}\theta_0^{\sum x_i}} \\ &= 2n[(\theta_0 - \hat{\theta}_n) - \hat{\theta}_n\log(\theta_0/\hat{\theta}_n)]\end{aligned}$$

Since for large n, $\lambda \overset{D}{\approx} \chi_1^2$, to construct an approximate size α LRT, we find $\chi_{1,\alpha}^2$ s.t. $P(\chi_1^2 < \chi_{1,\alpha}^2) = 1 - \alpha$ and reject H_0 if $\lambda > \chi_\alpha^2$.

Example: Consider the regression model $Y_i = X_i + \epsilon_i$ for $i = 1, \ldots, n$, with $\epsilon_1, \ldots, \epsilon_n$ iid and $\epsilon_i \sim N(0, \sigma^2)$.

1. Consider $X_1, \ldots, X_n$ as given. Find the MLE for σ.

2. Now we assume that $n = 2m$, $\epsilon_i \sim N(0, \sigma_1^2)$ when $i = 1, ..., m$, and $\epsilon_i \sim N(0, \sigma_2^2)$ when $i = m + 1, ..., 2m$. Carry out a test for testing $H_0 : \sigma_1 = \sigma_2$ vs. $H_1 : \sigma_1 \neq \sigma_2$.

P-value

Suppose that for every $\alpha \in (0, 1)$ we have a size α test with rejection region R_α. When $R_\alpha = \{x : T(x) \geq c_\alpha\}$,

$$\text{p-value} = \sup_{\theta \in \Theta_0} P_\theta(T(X) \geq T(x))$$

where x is the observed data.

Therefore, the p-value is the probability under H_0 of observing a value $T(X)$ the same as or more extreme than what was actually observed.

Equivalently,

$$\text{p-value} = \inf\{\alpha : T(x) \in R_\alpha\}$$

That is, the p-value is the smallest level at which we can reject H_0 with x observed.

In the case of the Wald test, the (approximate) p-value is

$$\text{p-value} = P_{\theta_0}(|W| > |w|) \approx P(|Z| > |w|) = 2\Phi(-|w|)$$

where w is the observed value of the statistic and $Z \sim N(0, 1)$.

In the case of the LRT with point null hypothesis and limiting χ^2_{r-q} distribution, the (approximate) p-value is

$$\text{p-value} = P_{\theta_0}(\lambda(X) > \lambda(x)) \approx P(\chi^2_{r-q} > \lambda(x))$$

Theorem: If the test statistic has a continuous distribution, then under $H_0 : \theta = \theta_0$, the p-value has a $Unif(0, 1)$ distribution. Therefore, if we reject H_0 when the p-value is less than α, the probability of a Type I error is α.

Note! It is very tempting to think that $P(H_0|Data)$, but this is not the case. We have calculated the p-value *assuming H_0 is true.* Moreover, this kind of quantity doesn't make sense in frequentist statistics, in which we think of the parameters (determining H_0) as being fixed. However, we will see soon that this quantity does make sense (and can be calculated) in a Bayesian framework.

Neyman-Pearson Theorem

Suppose we test $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$. Let

$$T(X) = \frac{\mathcal{L}_n(\theta_1)}{\mathcal{L}_n(\theta_0)} = \frac{f(x_1, ..., x_n; \theta_1)}{f(x_1, ..., x_n; \theta_0)}.$$

Suppose we reject H_0 when $T > c$. If we choose c so that $P_{\theta_0}(T > c) = \alpha$, then this test the most powerful, size α test. That is, among all tests with size α, this test maximize the power $\beta(\theta_1)$.

We'll next discuss some tests when the data are multinomial.

Aside: The Multinomial Distribution

Suppose $Z \in \{1, ..., k\}$ and let $p_j = P(Z = j)$. The parameter $p = (p_1, \ldots, p_k)$ is really only $k - 1$ dimensional, since $\sum_{j=1}^{k} p_j = 1$. Suppose we observe an iid sample $Z_1, \ldots, Z_n$. Let $X_j = \#\{Z_i : Z_i = j\}$. Then we say $X = (X_1, \ldots, X_k)$ has $Multinomial(n, p)$ distribution.

The PDF is

$$f(x_1, \ldots, x_k; p) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

Note that the labels $1, \ldots, k$ for the Z's are arbitrary, and that $Binomial(n, p)$ distribution is just a special case.

The MLE is $(\hat{p}_1, \ldots, \hat{p}_k) = (X_1/n, \ldots, X_k/n)$.

Consider testing $H_0 : (p_1, \ldots, p_k) = (p_{01}, \ldots, p_{ok})$ versus the alternative that they are not equal. The LRT rejects when

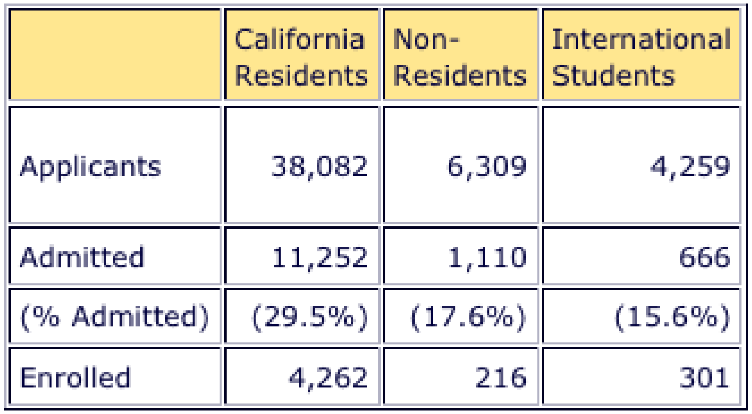
$$T(X) = \frac{\mathcal{L}_n(\hat{p})}{\mathcal{L}(p_0)} = \prod_{j=1}^{k} \left(\frac{\hat{p}_j}{p_{0j}} \right)^{X_j}$$

is large. Since we don't know how to calculate the exact probability of this, we'll use the limiting χ^2. That is,

$$\lambda(X) = 2 \log T(X) = 2 \sum_{j=1}^{k} X_j \log \left(\frac{\hat{p}_j}{p_{0j}} \right) \overset{D}{\to} \chi^2_{k-1}$$

The degrees of freedom is $k-1$ because the dimension of Θ is $k-1$ and the dimension of Θ_0 is zero (a point). The approximate size α LRT rejects H_0 when $\lambda(X) \geq \chi^2_{k-1,\alpha}$.

Example: Consider the following data on 2009 freshman admissions at Berkeley.



	California Residents	Non-Residents	International Students
Applicants	38,082	6,309	4,259
Admitted	11,252	1,110	666
(% Admitted)	(29.5%)	(17.6%)	(15.6%)
Enrolled	4,262	216	301

Treat the enrolled students as a sample from the Multinomial distribution, and test the hypothesis that the proportion of the three groups among the enrolled students is the same as it was for admitted students, i.e., that

$$p = \left(\frac{11252}{13028}, \frac{1110}{13028}, \frac{666}{13028} \right)$$

Another popular test for this situation is called Pearson's χ^2 test. The statistic is defined as

$$T = \sum_{j=1}^{k} \frac{(X_j - np_{0j})^2}{np_{0j}} = \sum_{j=1}^{k} \frac{(X_j - E_j)^2}{E_j}$$

Here $E_j = E[X_j] = np_{0j}$ is the expected value of X_j under H_0.

This statistic also has a limiting χ^2_{k-1} distribution under H_0.

Example: Test the Berkeley data again using Pearson's χ^2 test.

The LRT and Pearson's χ^2 are asymptotically equivalent, so they give similar answers for large n. However, Pearson's χ^2 statistic tends to converge to χ^2_{k-1} in distribution faster, so it is preferable for small n.

The LRT and Pearson's χ^2 also arise in tests of independence. Consider a simple case first, that two r.v.'s Y and Z are binary. The data are shown in the left table and the corresponding probabilities in the right table.

	$Y=0$	$Y=1$	
$Z=0$	X_{00}	X_{01}	$X_{0.}$
$Z=1$	X_{10}	X_{11}	$X_{1.}$
	$X_{.0}$	$X_{.1}$	n

	$Y=0$	$Y=1$	
$Z=0$	p_{00}	p_{01}	$p_{0.}$
$Z=1$	p_{10}	p_{11}	$p_{1.}$
	$p_{.0}$	$p_{.1}$	1

We treat $X = (X_{00}, X_{01}, X_{1,0}, X_{11})$ as a sample from a multinomial distribution. Under the null hypothesis that Y and Z are independent, the cell probabilities are the product of the row and column probabilities:

$$p_{ij} = p_{i.}p_{.j}$$

We can use this to construct either a LRT (doing a constrained maximization) or Pearson's χ^2.

Consider now a table with I rows and J columns. The unconstrained MLEs are $\hat{p}_{ij} = X_{ij}/n$, and under H_0, the constrained MLEs are

$$\hat{p}_{0ij} = \hat{p}_{0i.}\hat{p}_{0.j} = \frac{X_{i.}}{n}\frac{X_{.j}}{n}$$

Therefore for the LRT we have $\lambda = 2\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}\log\left(\frac{nX_{ij}}{X_{i.}X_{.j}}\right)$ and for Pearson's χ^2 we have

$$T = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(X_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(X_{ij}-n\hat{p}_{0ij})^2}{n\hat{p}_{0ij}}$$

Both test statistics have a limiting χ^2_ν distribution, where $\nu = (I-1)(J-1)$.

Finally, we can adapt these ideas to form a test of goodness of fit. Here the null hypothesis is that the data come from an assumed parametric model. The idea is to "discretize" both the data and the model.

First, define k disjoint intervals $I_1, \ldots, I_k$. Define

$$p_j(\theta) = P_\theta(X \in I_j) = \int_{I_j} f(x;\theta)dx$$

Let $N_j = \#\{X_i \in I_j\}$, the number of observations that fall into I_j. Treat $N = (N_1, \ldots, N_k)$ as a sample from a multinomial distribution with $p(\theta) = (p_1(\theta), \ldots, p_k(\theta))$, and maximize the likelihood to get $\tilde{\theta}$.

Then under H_0 that the data are iid draws from $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$, the test statistic $Q = \sum_{j=1}^{k} \frac{(N_j - np_j(\tilde{\theta}))^2}{np_j(\tilde{\theta})} \overset{D}{\to} \chi^2_{k-1-s}$, where s is the dimension of θ.

Examples:

- Given an iid sample $X_1, ..., X_n$, provide a test to determine whether the data follow a Poisson distribution with unknown parameter λ.
  - H_0 : $X_1, ..., X_n \sim Poisson(\lambda)$
  - H_1 : H_0 is not true

- Given an iid sample $X_1, ..., X_n$, provide a test to determine whether the data follow an Exponential distribution with unknown parameter λ.

“Multiple testing” refers to the problem of testing $m > 1$ hypotheses, and wanting to control something more than the error rate for each test.

The Bonferroni Method controls the probability of having at least one false rejection. If α is the upper bound placed on this probability, the method achieves this by using level α/m for each of the tests.

$$\begin{aligned}
P(\text{at least one Type I error}) &= P(\bigcup_{i=1}^{m} \text{Type I error in the } i^{th} \text{ test}) \\
&\leq \sum_{i=1}^{m} P(\text{Type I error in the } i^{th} \text{ test}) \\
&= \sum_{i=1}^{m} \alpha/m = \alpha
\end{aligned}$$

In many cases, Bonferroni is too conservative. Another option is to control the False Discovery Rate (FDR), which is

$$FDR = E\left(\frac{\text{Number of false rejections}}{\text{Total number of rejections}}\right)$$

Benjamini and Hochberg suggested the following procedure, which guarantees $FDR \leq \alpha$:

1. For each test, compute the $p - value$. Let $P_{(1)} < \cdots < P_{(m)}$ denote the ordered p-values.

2. Select $R = \max\{i : P_{(i)} < \frac{i\alpha}{m}\}$, when the p-values are independent.

3. Reject all null hypotheses for which the p-value $\leq P_{(R)}$.