Chapter 8
Hypothesis Testing
Multiple Testing

Outline

The p-value

In 2016, the American Statistical Association set for itself a task to make a statement on p-values. They note that it is all too easy to set a test, create a test statistic and compute a p-value. Proper statistical practice is much more than this and includes

- appropriately chosen techniques based on a thorough understanding of the phenomena under study,
- adequate visual and numerical summaries of the data,
- properly conducted analyses whose logic and quantitative approaches are clearly explained,
- correct interpretation of statistical results in context, and
- reproducibility of results via a thorough reporting.

Expressing a p-value is one of many approaches to summarize the results of a statistical investigation.

The *p*-value

The statement's six principles, many of which address misconceptions and misuse of the p-value, are the following:

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

The p-value

Let θ_0 be the true state of nature. Assume that the distribution function for the text statistic $T(X)$ has continuous and strictly increasing distribution for all possible values.

Let $F_{T(X)}(t|\theta_0)$ be the distribution function for $T(X)$ under θ_0 and note that the conditions on function $F_{T(X)}(t|\theta_0)$ insure that it is one-to-one and thus has an inverse. The p-value is $1 - F_{T(X)}(T(X)|\theta_0)$. Choose u in the interval $[0, 1]$. Then,

$$\begin{aligned} P_{\theta_0}\{F_{T(X)}(T(X)|\theta_0) \leq u\} &= P_{\theta_0}\{T(X) \leq F^{-1}_{T(X)}(u|\theta_0)\} \\ &= F_{T(X)}(F^{-1}_{T(x)}(u|\theta_0)|\theta_0) = u, \end{aligned}$$

showing that $F_{T(X)}(T(X)|\theta_0)$ is uniformly distributed on the interval $[0, 1]$.

The p-value

The p-value is $P_{\theta_0}\{T(X) > t\} = 1 - F_{T(X)}(t|\theta_0))$. For any value of α, k_α satisfies

$$P_{\theta_0}\{T(X) > k_\alpha\} = 1 - F_{T(X)}(k_\alpha|\theta_0)) = \alpha.$$

The distribution of p-values under θ_1,

$$\begin{aligned} P_{\theta_1}\{F_{T(X)}(T(X)|\theta_0) > 1-\alpha\} &= P_{\theta_1}\{F_{T(X)}(T(X)|\theta_0) > F_{T(X)}(k_\alpha|\theta_0)\} \\ &= P_{\theta_1}\{T(X) > k_\alpha\} = 1 - \beta(\alpha), \end{aligned}$$

the power as a function of the significance level α. This is the receiving operator characteristic.

The p-value

Let's see how this looks. Consider the hypothesis

$$H_0 : \mu \leq 0 \quad \text{versus} \quad H_1 : \mu > 0,$$

and assume that the test statistic is $N(0, 1)$. For $\alpha = 0.05$, the critical value

```
> (zstar<-qnorm(0.95))
[1] 1.644854
```

The power against an alternative $\mu = 3$,

```
> 1-pnorm(zstar,3)
[1] 0.9123145
```

Let's simulate 1000 tests, 90% follow the null and 10% follow the alternative

```
> x <- c(rnorm(900), rnorm(100, 3))
```

and compute p-values.

```
> p <- 1-pnorm(x)
```

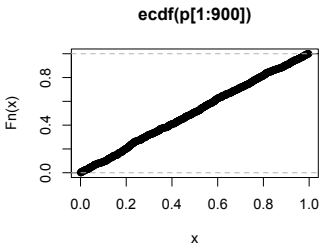
The p-value

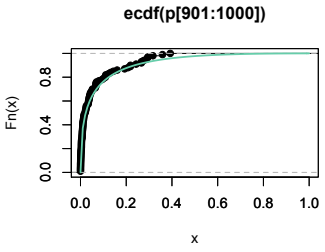
As anticipated, the p-values taken from the first 900 simulations follow a uniform distribution.

p-values for the alternative are close to the receiving operating characteristic whose analytic expression

$$1 - \beta = P_3\{Z > z_\alpha\}, \quad Z \sim N(3,1)$$

```
> par(mfrow=c(2,1))
> plot(ecdf(p[1:900]),ylim=c(0,1),
   xlim=c(0,1))
> plot(ecdf(p[901:1000]),ylim=c(0,1),
   xlim=c(0,1))
> par(new=TRUE)
> curve(1-pnorm(qnorm(1-x),3),ylim=c(0,1),
   col="aquamarine3",lwd=2,xlab="",ylab="")
```





Familywise Error Rate

We now consider testing multiple hypotheses. This is common in the world of "big data" with thousands of hypothesis on many issues in subjects including genomics, internet searches, or financial transactions. For m hypotheses, let $p_1, \ldots, p_m$ be the p-values for m hypothesis tests.

The familywise error rate, (FWER) is the probability of making even one type I error. If we set α_B for the significance level for a single test, then the simplest strategy is to employ the Bonferroni correction. This uses the Bonferroni inequality,

$$P(A_1 \cup \cdots \cup A_m) \leq P(A_1) + \cdots + P(A_m)$$

for events $A_1, \ldots, A_m$.

If A_i is the event of rejecting the null hypothesis when it is true, then $A_1 \cup \cdots \cup A_m$ is the event that at least one of the hypotheses is rejected when it is true.

Familywise Error Rate

For each i, set $P(A_i) = \alpha_B$ and so $\alpha = P(A_1 \cup \cdots \cup A_m) \leq m\alpha_B$. Thus, the Bonferroni correction is to reject if

$$p_i \leq \frac{\alpha}{m} \quad \text{for all } i.$$

For m independent, α_I level hypothesis tests, the familywise error $\alpha = 1 - (1 - \alpha_I)^m$. Thus,

$$(1 - \alpha)^{1/m} = 1 - \alpha_I \quad \text{and} \quad \alpha_I = 1 - (1 - \alpha)^{1/m}$$

is the level necessary to obtain an α familywise error rate.

Example. For $\alpha = 0.05$ and $m = 20$, the Bonferroni correction, $\alpha_B = 0.05/20 = 0.0025$ and the independence correction, $\alpha_I = 1 - (1 - 0.05)^{1/20} = 0.00256$ will guarantee a familywise error rate $\alpha = 0.05$

Familywise Error Rate

For $\alpha = 0.05$, and 1000 tests, so the Bonferroni correction will have us looking for p-values smaller than 0.00005.

Tests 1 through 900 are true under H_0 and

```
> bonftest <- p > 0.00005
> summary(bonftest[1:900])
   Mode    TRUE
logical     900
```

So, there were no type I errors in the simulation. However, for tests 901 through 1000,

```
> summary(bonftest[901:1000])
   Mode   FALSE    TRUE
logical       15      85
```

The stringent reduction in type I error probability has driven the type II error probability, up to 85% in this simulation.

Fisher's Method

We have seen that for continuous test functions, if the null hypothesis is true for all m hypotheses, then

$$p_1, \ldots, p_m \text{ are independent } U(0,1) \text{ random variables.}$$

Recall from the use of the probability transform that

$$-2 \ln p_1, \ldots, -2 \ln p_m \text{ are independent } Exp(1/2) \text{ random variables.}$$

So their sum

$$-2 \ln p_1 - \cdots - 2 \ln p_m \text{ is a } \Gamma(1/2, m) \text{ random variable.}$$

This can serve as a test statistic for the multiple hypothesis that all the null hypotheses are true, rejecting if the sum above is sufficiency large.Traditionally, we use the fact that $\Gamma(1/2, m)$ is also χ^2_{2m}.

Fisher's Method

Returning to our simulation. We will keep 20 values for the null and 5 for the alternative and see how the p-values change.

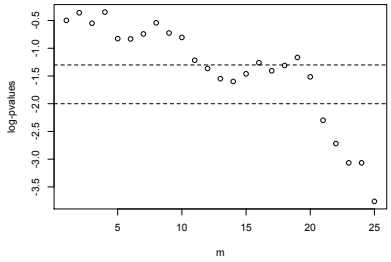
```
> pf<-c(p[881:905])
> sumlnp<- -2*cumsum(log(pf))
```

Next set up the test statistic for $m = 1$ to 25 tests.

```
pvalue<-numeric(25)
> for (m in 1:25){pvalue[m]
  <-1-pchisq(sumlnp[m],2*m)}
```

and plot the base 10 logarithm of the p-values as a function of m

```
> plot(1:25,log(pvalue[1:25],10),
   xlab="m",ylab="log-pvalues")
> abline(h=log(0.05,10),lty=2)
> abline(h=log(0.01,10),lty=2)
```



False Discovery Rate

When the number of tests becomes very large, then having all hypotheses true is an extremely strict criterion. A more relaxed and often more valuable criterion is the false discovery rate.

To set this up we need some notation.

- m - the total number of hypotheses tested
- π_0 - the proportion of true null hypotheses, an unknown parameter
- V - the number of false positives (Type I error) (also called *false discoveries*)
- S - the number of true positives (also called *true discoveries*)
- T - the number of false negatives (Type II error)
- U - the number of true negatives
- $R = V + S$ is the number of rejected null hypotheses (also called *discoveries*, either true or false)

Notation

hypothesis tests			
	H_0 is true	H_1 is true	total
reject H_0	V	S	R
fail to reject H_0	U	T	$m - R$
total	$m\pi_0$	$m(1 - \pi_0)$	m

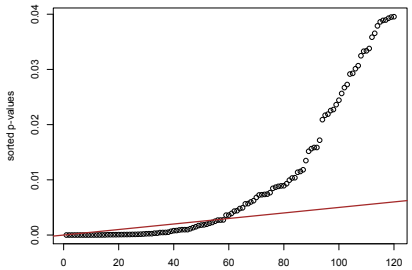
$Q = V/R = V/(V + S)$ as the proportion of false discoveries among the discoveries. The false discovery rate

$$\mathrm{FDR} = E[V/R|R > 0] \cdot P\{R > 0\}.$$

Benjamini-Hochberg Procedure

For m independent tests, the Benjamini-Hochberg procedure follows these steps.

1. Rank the p-values
$$p_{(1)} < p_{(2)} < \cdots < p_{(m)}$$
2. For a given value of α, find the *largest* k such that
$$p_{(k)} \leq \frac{k}{m}\alpha.$$
3. Reject the null hypothesis (i.e., declare discoveries) for all $H_0^{(i)}$ for $i = 1, \ldots, k$.



The Benjamini-Hochberg procedure says to reject the null hypothesis for all tests whose p-value up to the last time the p-value lies below the line with slope α/m

Benjamini-Hochberg Procedure

Returning to the simulations

1. Rank the p-values

```
> psort<-sort(p)
```

2. For a given value of α, find the *largest* k such that $p_{(k)} \leq \frac{k}{m}\alpha$

```
> fdrtest <- NULL
> for (i in 1:1000)fdrtest
    <- c(fdrtest, p[i] > match(p[i],psort)*.05/1000)
```

3. Reject the null hypothesis for all $H_0^{(i)}$ for $i = 1, \ldots, k$.

```
> summary(fdrtest[1:900])
   Mode   FALSE    TRUE
logical       3     897
> summary(fdrtest[901:1000])
   Mode   FALSE    TRUE
logical      55      45
```

False Discovery Rate

We can also model question Is the null hypothesis hypothesis true? as a sequence of Bernoulli trials. Let π_0 be the success parameter for the trials.

- With probability π_0, the null hypothesis is true and the p-values follow F_U, the uniform distribution on the interval $[0, 1]$.
- With probability $1 - \pi_0$, the null hypothesis is false and the p-values follow F_R, the distribution of the receiving operator characteristic

Taken together, we say that the p-values are distributed according to the mixture

$$F(x) = \pi_0 F_U(x) + (1 - \pi_0) F_R(x) = \pi_0 x + (1 - \pi_0) F_R(x).$$

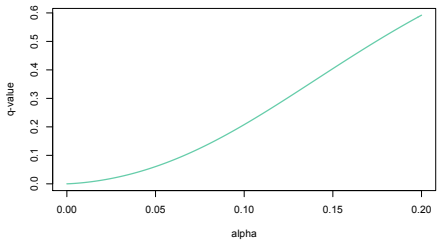
False Discovery Rate

Thus, if we reject whenever the p-value is below a chosen value α, then the type I error probability is α. From this we determine the false discovery rate, here defined as

$$q = P\{H_0 \text{ is true} | \text{reject } H_0\}.$$

Using Bayes formula,

$$\begin{aligned} q &= \frac{P\{\text{reject } H_0 | H_0 \text{ is true}\} P\{H_0 \text{ is true}\}}{P\{\text{reject } H_0\}} \\ &= \frac{\alpha \pi_0}{F(\alpha)}. \end{aligned}$$



q-values versus significance level α. $\pi_0 = 0.10$ and $F_R(\alpha) = P_3\{Z > z_\alpha\}, Z \sim N(3, 1)$.

False Discovery Rate

$F(\alpha)$ can be estimated from the data. π_0 is estimated by smoothing the histogram of the p-values and then estimating when the density strays from uniform.