36-705: Intermediate Statistics **Fall 2017**

Lecture 23: October 25

Lecturer: Siva Balakrishnan

23.1 False Discovery Rate

Suppose that we tested $d = 1000$ genes for association with some disease, we got a 1000 p-values, and 100 of them were less than 0.01. We'd expect that roughly $0.01d_0 \leq 0.01d = 10$ of these to be falsely rejected nulls, and perhaps this is not a bad tradeoff, i.e. if we rejected the first 100 nulls, we would spend only 10% of our time on falsely rejected nulls, i.e. we would make 90 real discoveries.

The FDR is the expected number of false rejections divided by the number of rejections.

Denote the number of false rejections as V, and the total number of rejections as R. Then the false discovery *proportion* is:

$$\text{FDP} = \begin{cases} \frac{V}{R} & \text{if } \; R > 0 \\ 0 & \text{if } \; R = 0. \end{cases}$$

The FDR is then defined as:

$$\text{FDR} = \mathbb{E}[\text{FDP}].$$

In this notation we can see that the FWER is:

$$\text{FWER} = \mathbb{P}(V \geq 1).$$

We will next consider how one can control the FDR. We will describe a procedure known as the Benjamini-Hochberg (BH) procedure.

23.1.1 The BH procedure

The BH procedure is one that controls the FDR under independence (i.e. the p-values are independent). There is a much weaker form of this procedure that works under dependence (see the Wasserman book). It turns out to be very challenging to tightly control FDR under strong dependence.

The procedure is:

1. Suppose we do d tests. Let us take the p-values $p_1, \ldots, p_d$, and sort them, i.e. we create the list: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(d)}$.

2. Define the thresholds:

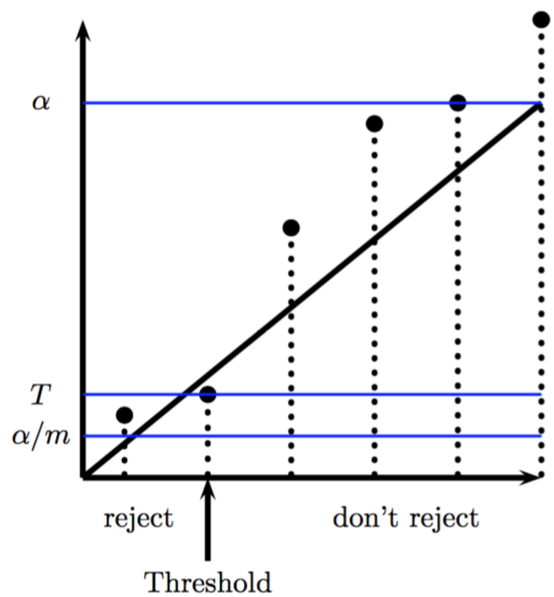
$$t_i = \frac{i\alpha}{d}.$$

3. Find the largest $i_{\max}$ such that

$$i_{\max} = \arg\max_i \{i : p_{(i)} < t_i\}.$$

4. Reject all nulls upto and including $i_{\max}$.

This might seem a bit confusing but here is a simple picture:



23.1.2 Properties of FDR

We have now seen a procedure that controls the FDR under some assumptions. One question of interest is how does FDR control compare to FWER control? Another is just: how do we interpret FDR control?

Interpreting FDR control: The way to think about FDR control is: if we repeat our experiment many times, on average we control the FDP. This is not a statement about the individual experiment we did conduct, and really it does not say much about how likely it is that on a given experiment we have an FDP that is larger than a threshold (think about using Markov's inequality).

FWER on the other hand, does control the error rate for a single experiment. That is, with FWER control, we have managed our false discoveries unless we are very unlucky; with FDR control, on average our test will control FDP, but in our particular experiment we may not have done a very good job. We will see in a second controlling FWER does control the FDR. The way to interpret all of this is that: FDR control is a very weak notion of error control.

Connection to FWER:

1. The first connection is that under the global null (when all the null hypotheses are true) FDR control is equivalent to FWER control.

   Proof: Under the global null, any rejection is a false rejection. There are two possibilities: either we do not reject anything: in this case the FDP $= 0$. If we do reject any null hypothesis then our FDP is 1 (since $V = R$). So we have that:

   $$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}] = \mathbb{P}(V > 0) * 1 + \mathbb{P}(V = 0) * 0 = \mathbb{P}(V > 0) = \mathrm{FWER}.$$

2. The second connection is that the FWER $\geq$ FDR always. This implies that controlling the FWER implies FDR control.

   Proof: We can see that the following is a simple upper bound on the FDP:

   $$\mathrm{FDP} \leq \mathbb{I}(V \geq 1),$$

   since if $V = 0$, FDP $= 0$, and if $V > 0$ then $V/R \leq 1$. Taking expectations of this expression gives:

   $$\mathrm{FDR} \leq \mathbb{P}(V \geq 1) = \mathrm{FWER}.$$

   The flip-side of this is that FDR control is less stringent so if this is the correct measure for you then you will have *more* power by controlling FDR (rather than controlling FWER).

23.2 Proving BH controls FDR

The main result is the following:

Theorem: Suppose that the p-values are independent, the BH procedure controls the FDR at level α. In fact,

$$\mathrm{FDR} \leq \frac{d_0 \alpha}{d} \leq \alpha.$$

Proof Intuition: Suppose that the BH procedure returned a value $i_{\max}$ then we know that,

$$p_{(i_{\max})} < \frac{i_{\max}\alpha}{d}.$$

We have rejected $i_{\max}$ hypotheses. At the cut-off $\frac{i_{\max}\alpha}{d}$ we expect that $\frac{d_0 i_{\max}\alpha}{d}$ nulls to be rejected. This gives us that the FDR should be roughly:

$$\text{FDR} \approx \frac{d_0 i_{\max}\alpha}{d i_{\max}} = \frac{d_0\alpha}{d} \leq \alpha.$$

Formalizing this argument is a bit intricate: notice that $i_{\max}$ is a random variable and furthermore the numerator and denominator in the FDP are not independent random variables so we need to be careful while taking the expectation of the ratio. I have included a formal proof that is identical to one from Emmanuel Candes' Stat 300c notes. These notes are in general a great resource that delve much deeper into theoretical aspects of multiple testing.

Proof: When $d_0 = 0$ there are no false discoveries so there is nothing to prove. We will suppose that $d_0 \geq 1$, and denote the set of nulls as I_0. Let us define:

$$V_i = \mathbb{I}(H_i \text{ is rejected}),$$

then we can write the FDP as:

$$\text{FDP} = \sum_{i \in I_0} \frac{V_i}{\max\{R, 1\}},$$

notice that taking the max in the denominator just avoids the 0/0 problem, and is a shorthand way of writing the FDP. Suppose we could prove that:

$$\mathbb{E}\left[\frac{V_i}{\max\{R, 1\}}\right] = \frac{\alpha}{d},$$

then we are done since,

$$\text{FDR} = \sum_{i \in I_0} \mathbb{E}\left[\frac{V_i}{\max\{R, 1\}}\right] = \frac{d_0\alpha}{d}.$$

To prove the claim we first re-write:

$$\frac{V_i}{\max\{R, 1\}} = \sum_{k=1}^{d} \frac{V_i \mathbb{I}(R = k)}{k},$$

noting that if $R = 0$ both the LHS and RHS are 0. We now need to make some further observations:

1. Suppose that there are k rejections, then we can rewrite:

$$V_i = \mathbb{I}(H_i \text{ is rejected}) = \mathbb{I}(p_i \leq k\alpha/d).$$

2. Suppose that $p_i \leq \alpha k/n$, then we take p_i and set it to 0, and denote the number of rejections as $R(p_i \to 0)$ and note that $R(p_i \to 0)$ is exactly the same as R. On the other hand if $p_i > \alpha k/n$ then $V_i = 0$. So we can write:

$$V_i\mathbb{I}(R = k) = V_i\mathbb{I}(R(p_i \to 0) = k).$$

Now, returning to the main thread suppose we considered the conditional expectation:

$$\begin{aligned}
\mathbb{E}\left[\frac{V_i\mathbb{I}(R = k)}{k}\Big| p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_d\right] &= \frac{\mathbb{E}[\mathbb{I}(p_i \leq k\alpha/d)\mathbb{I}(R(p_i \to 0) = k)|p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_d]}{k} \\
&= \frac{\mathbb{I}(R(p_i \to 0) = k)\alpha}{d},
\end{aligned}$$

where we use the fact that conditional on the other p-values $\mathbb{I}(R(p_i \to 0) = k)$ is deterministic and that the p-values have uniform distribution under the null, and that the nulls are independent so that:

$$\mathbb{E}[\mathbb{I}(p_i \leq k\alpha/d)|p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_d] = \mathbb{E}[\mathbb{I}(p_i \leq k\alpha/d)] = k\alpha/d.$$

Now, by iterated expectations:

$$\begin{aligned}
\mathbb{E}\left[\frac{V_i}{\max\{R, 1\}}\right] &= \sum_{k=1}^{d} \mathbb{E}\left[\mathbb{E}\left[\frac{V_i\mathbb{I}(R = k)}{k}\Big| p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_d\right]\right] \\
&= \sum_{k=1}^{d} \frac{\mathbb{I}(R(p_i \to 0) = k)\alpha}{d} = \frac{\alpha}{d},
\end{aligned}$$

which was the claim we needed to prove.

23.3 Confidence Sets

We have discussed confidence intervals in passing when talking about the central limit theorem. Now, we will begin to discuss them a bit more formally.

The setting is that we have a statistical model (i.e. a collection of distributions) $\mathcal{P}$. Let $C_n(X_1, \ldots, X_n)$ be a set constructed using the observed data $X_1, \ldots, X_n$. This is a random set. C_n is a $1 - \alpha$ confidence set for a parameter θ if:

$$P(\theta \in C_n(X_1, \ldots, X_n)) \geq 1 - \alpha, \quad \text{for all} \;\; P \in \mathcal{P}.$$

This means that no matter which distribution in $\mathcal{P}$ generated the data, the interval guarantees the coverage property described above. Some people would refer to such intervals as *honest* confidence intervals to make explicit the fact that the coverage is uniform over the model.

In cases when $C_n(X_1,\ldots,X_n) = [L(X_1,\ldots,X_n), U(X_1,\ldots,X_n)]$ we refer to the confidence set as a confidence interval.

It is important to note that C_n is random, while θ is fixed.

Example 23.1 *Let $X_1,\ldots,X_n \sim N(\theta,\sigma^2)$. Suppose that σ is known. Let*

$$L = L(X_1,\ldots,X_n) = \overline{X} - c, \qquad U = U(X_1,\ldots,X_n) = \overline{X} + c.$$

Then

$$\begin{aligned}
P_\theta(L \le \theta \le U) &= P_\theta(\overline{X} - c \le \theta \le \overline{X} + c) \\
&= P_\theta(-c < \overline{X} - \theta < c) = P_\theta\left(-\frac{c\sqrt{n}}{\sigma} < \frac{\sqrt{n}(\overline{X}-\theta)}{\sigma} < \frac{c\sqrt{n}}{\sigma}\right) \\
&= P\left(-\frac{c\sqrt{n}}{\sigma} < Z < \frac{c\sqrt{n}}{\sigma}\right) = \Phi(c\sqrt{n}/\sigma) - \Phi(-c\sqrt{n}/\sigma) \\
&= 1 - 2\Phi(-c\sqrt{n}/\sigma) = 1 - \alpha
\end{aligned}$$

if we choose $c = \sigma z_{\alpha/2}/\sqrt{n}$. So, if we define $C_n = \overline{X}_n \pm \sigma z_{\alpha/2}\sqrt{n}$ then

$$P_\theta(\theta \in C_n) = 1 - \alpha$$

for all θ.

Example 23.2 *$X_i \sim N(\theta_i, 1)$ for $i = 1,\ldots,n$. Let*

$$C_n = \{\theta \in \mathbb{R}^n : \ \|X - \theta\|^2 \le \chi^2_{n,\alpha}\}.$$

Then

$$P_\theta(\theta \notin C_n) = P_\theta(\|X-\theta\|^2 > \chi^2_{n,\alpha}) = P(\chi^2_n > \chi^2_{n,\alpha}) = \alpha.$$

Four methods:

1. Probability Inequalities
2. Inverting a test
3. Pivots
4. Large Sample Approximations

NOTE: Optimal confidence intervals are confidence intervals that are as short as possible but we will not discuss optimality.

23.4 Using Probability Inequalities

Intervals that are valid for finite samples can be obtained by probability inequalities.

Example 23.3 *Let* $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. *By Hoeffding's inequality:*

$$\mathbb{P}(|\widehat{p} - p| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

Let

$$\epsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}.$$

Then

$$\mathbb{P}\left(|\widehat{p} - p| > \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}\right) \leq \alpha.$$

Hence, $\mathbb{P}(p \in C) \geq 1 - \alpha$ *where* $C = (\widehat{p} - \epsilon_n, \widehat{p} + \epsilon_n)$.

Example 23.4 *Let* $X_1, \ldots, X_n \sim F$. *Suppose we want a* **confidence band** *for* F. *We can use VC theory. Remember that*

$$\mathbb{P}\left(\sup_x |F_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

Let

$$\epsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}.$$

Then

$$\mathbb{P}\left(\sup_x |F_n(x) - F(x)| > \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}\right) \leq \alpha.$$

Hence,

$$P_F(L(t) \leq F(t) \leq U(t) \ \ \text{for all } t) \geq 1 - \alpha$$

for all F, *where*

$$L(t) = \widehat{F}_n(t) - \epsilon_n, \quad U(t) = \widehat{F}_n(t) + \epsilon_n.$$

We can improve this by taking

$$L(t) = \max\left\{\widehat{F}_n(t) - \epsilon_n,\ 0\right\}, \quad U(t) = \min\left\{\widehat{F}_n(t) + \epsilon_n,\ 1\right\}.$$

23.5 Inverting a Test

For each θ_0, construct a level α test of

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0.$$

Define $\phi_{\theta_0}(x_1, \ldots, x_n) = 1$ if we reject and $\phi_{\theta_0}(x_1, \ldots, x_n) = 0$ if we don't reject. Let $A(\theta_0)$ be the acceptance region, that is,

$$A(\theta_0) = \Big\{ x_1, \ldots, x_n : \; \phi_{\theta_0}(x_1, \ldots, x_n) = 0 \Big\}.$$

Let

$$C_n \equiv C_n(x_1, \ldots, x_n) = \{\theta : \; (x_1, \ldots, x_n) \in A(\theta)\} = \{\theta : \; \phi_\theta(x_1, \ldots, x_n) = 0\}.$$

Theorem 23.5 *For each θ,*

$$P_\theta(\theta \in C(x_1, \ldots, x_n)) = 1 - \alpha.$$

Proof: Note that $1 - P_\theta(\theta \in C(x_1, \ldots, x_n))$ is the probability of rejecting θ when θ is true which is α. ■

The converse is also true:

Lemma 23.6 *If $C(x_1, \ldots, x_n)$ is a $1 - \alpha$ confidence interval then the test:*

$$\text{reject } H_0 \; \text{ if } \theta_0 \notin C(x_1, \ldots, x_n)$$

is a level α test.

Example 23.7 *Suppose we use the LRT. We reject H_0 when*

$$\frac{L(\theta_0)}{L(\widehat{\theta})} \leq c.$$

So

$$C = \left\{ \theta : \; \frac{L(\theta)}{L(\widehat{\theta})} \geq c \right\}.$$

Example 23.8 *Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with σ^2 known. The LRT of $H_0 : \mu = \mu_0$ rejects when*

$$|\overline{X} - \mu_0| \geq \frac{\sigma}{\sqrt{n}} z_{\alpha/2}.$$

So

$$A(\mu) = \left\{ x^n : \ |\overline{X} - \mu_0| < \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right\}$$

and so $\mu \in C(X^n)$ if and only if

$$|\overline{X} - \mu| \leq \frac{\sigma}{\sqrt{n}} z_{\alpha/2}.$$

In other words,

$$C_n = \overline{X}_n \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}.$$

If σ is unknown, then this becomes

$$C_n = \overline{X}_n \pm \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}.$$

23.6 Pivots

A function $Q(X_1, \ldots, X_n, \theta)$ is a *pivot* if the distribution of Q does not depend on θ. For example, if $X_1, \ldots, X_n \sim N(\theta, 1)$ then

$$\overline{X}_n - \theta \sim N(0, 1/n)$$

so $Q = \overline{X}_n - \theta$ is a pivot.

Let a and b be such that

$$P_\theta(a \leq Q(X, \theta) \leq b) \geq 1 - \alpha$$

for all θ. We can find such an a and b because Q is a pivot. It follows immediately that

$$C(x) = \{\theta : \ a \leq Q(x, \theta) \leq b\}$$

has coverage $1 - \alpha$.

Example 23.9 *Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. ($\sigma$ known.) Then*

$$Z = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0, 1).$$

We know that

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

and so

$$P\left(-z_{\alpha/2} \leq \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \leq z_{\alpha/2}\right) = 1-\alpha.$$

Thus

$$C = \overline{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}.$$

If σ is unknown, then this becomes

$$C = \overline{X} \pm \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}$$

because

$$T = \frac{\sqrt{n}(\overline{X}-\mu)}{S} \sim t_{n-1}.$$

Example 23.10 *Let $X_1, \ldots, X_n \sim \text{Uniform}(0,\theta)$. Let $Q = X_{(n)}/\theta$. Then*

$$\mathbb{P}(Q \leq t) = \prod_i \mathbb{P}(X_i \leq t\theta) = t^n$$

so Q is a pivot. Let $c = \alpha^{1/n}$. Then

$$\mathbb{P}(Q \leq c) = \alpha.$$

Also, $\mathbb{P}(Q \leq 1) = 1$. Therefore,

$$\begin{aligned} 1-\alpha &= \mathbb{P}(c \leq Q \leq 1) = \mathbb{P}\left(c \leq \frac{X_{(n)}}{\theta} \leq 1\right) \\ &= \mathbb{P}\left(\frac{1}{c} \geq \frac{\theta}{X_{(n)}} \geq 1\right) \\ &= \mathbb{P}\left(X_{(n)} \leq \theta \leq \frac{X_{(n)}}{c}\right) \end{aligned}$$

so a $1-\alpha$ confidence interval is

$$\left(X_{(n)},\ \frac{X_{(n)}}{\alpha^{1/n}}\right).$$

23.7 Large Sample Confidence Intervals

The Wald Interval. We know that, under regularity conditions,

$$\frac{\widehat{\theta}_n - \theta}{\text{se}} \rightsquigarrow N(0,1)$$

where $\widehat{\theta}_n$ is the mle and $\text{se} = 1/\sqrt{I_n(\widehat{\theta})}$. So this is an asymptotic pivot and an approximate confidence interval is

$$\widehat{\theta}_n \pm z_{\alpha/2}\text{se}.$$

By the delta method, a confidence interval for $\tau(\theta)$ is

$$\tau(\widehat{\theta}_n) \pm z_{\alpha/2}\text{se}(\widehat{\theta})|\tau'(\widehat{\theta}_n)|.$$

The Likelihood-Based Confidence Set. Let's consider inverting the asymptotic LRT. We test

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

Let k be the dimension of θ. We don't reject if

$$-2\log\left(\frac{L(\theta_0)}{L(\widehat{\theta})}\right) \leq \chi^2_{k,\alpha}$$

that is, if

$$\frac{L(\theta_0)}{L(\widehat{\theta})} > e^{-\chi^2_{k,\alpha}/2}.$$

So, the set of non-rejected nulls is

$$C_n = \left\{\theta : \frac{L(\theta)}{L(\widehat{\theta})} > e^{-\frac{\chi^2_{k,\alpha}}{2}}\right\}.$$

This is an upper level set of the likelihood function. Then

$$P_\theta(\theta \in C) \to 1 - \alpha$$

for each θ.

Example 23.11 *Let* $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$*. Using the Wald statistic*

$$\frac{\widehat{p} - p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} \rightsquigarrow N(0,1)$$

so an approximate confidence interval is

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

Using the LRT we get

$$C = \left\{p : \ -2\log\left(\frac{p^Y(1-p)^{n-Y}}{\widehat{p}^Y(1-\widehat{p})^{n-Y}}\right) \leq \chi^2_{1,\alpha}\right\}.$$

These intervals are different but, for large n, they are nearly the same. A finite sample interval can be constructed by inverting a test.

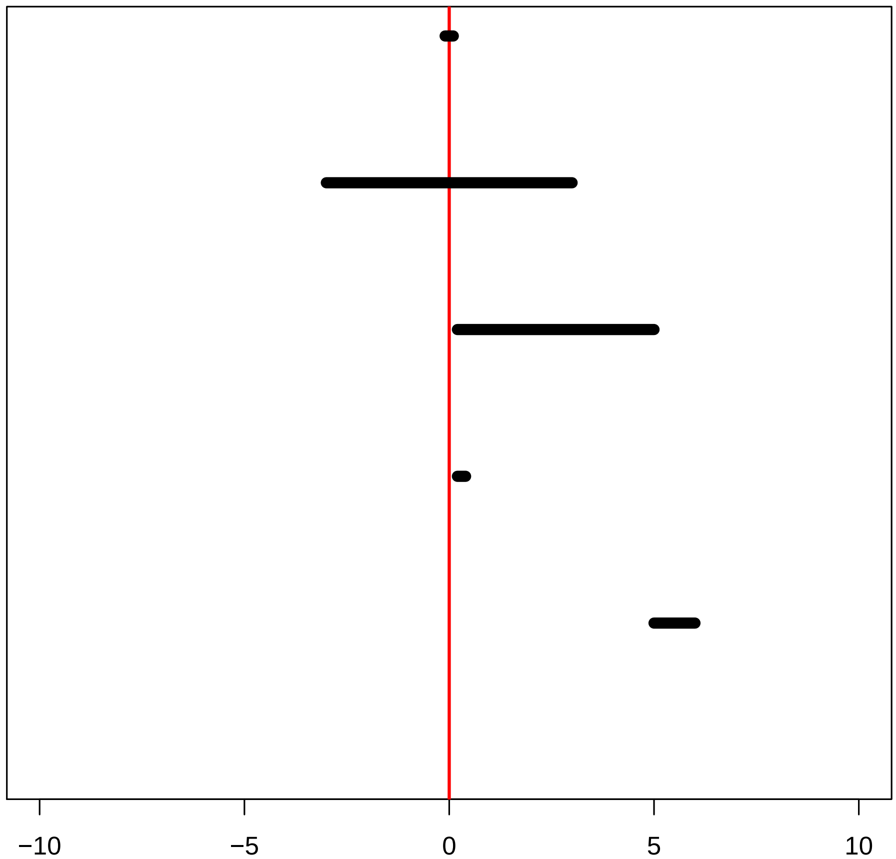


Figure 23.1: Five examples: 1. Not significant, precise. 2. Not significant, imprecise. 3. Barely significant, imprecise. 4. Barely significant, precise. 5. Significant and precise.

23.8 Tests Versus Confidence Intervals

Confidence intervals are more informative than tests. Look at Figure 23.1. Suppose we are testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. We see 5 different confidence intervals. The first two cases (top two) correspond to not rejecting H_0. The other three correspond to rejecting H_0. Reporting the confidence intervals is much more informaive than simply reporting "reject" or "don't reject."