4 Hypothesis Testing slide 126

4.1 Pure Significance Tests slide 127

Discovery of the top quark (Abe et al., 1995, PRL)

Here are two extracts from the article announcing the discovery:

TABLE I. Number of lepton + jet events in the 67 pb^{-1} data sample along with the numbers of SVX tags observed and the estimated background. Based on the excess number of tags in events with ≥ 3 jets, we expect an additional 0.5 and 5 tags from $t\bar{t}$ decay in the 1- and 2-jet bins, respectively.

N_{jet}	Observed events	Observed SVX tags	Background tags expected
1	6578	40	50 ± 12
2	1026	34	21.2 ± 6.5
3	164	17	5.2 ± 1.7
≥ 4	39	10	1.5 ± 0.4

The numbers of SVX tags in the 1-jet and 2-jet samples are consistent with the expected background plus a small $t\bar{t}$ contribution (Table I and Fig. 1). However, for the $W+ \geq 3$-jet signal region, 27 tags are observed compared to a predicted background of 6.7 ± 2.1 tags [8]. The probability of the background fluctuating to ≥ 27 is calculated to be 2×10^{-5} (see Table II) using the procedure outlined in Ref. [1] (see [9]). The 27 tagged jets are in 21 events; the six events with two tagged jets can be compared with four expected for the top + background hypothesis and ≤ 1 for background alone. Figure 1 also shows the decay lifetime distribution

Performing a test

- ☐ There's a **null hypothesis** to be tested:

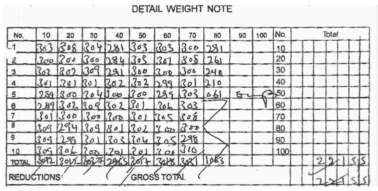
  H_0: the top quark does not exist.

  This seems counter-intuitive, but as one cannot prove a hypothesis, we attempt to refute its opposite — '**proof by (stochastic) contradiction**'.
- ☐ We obtain data, $y_{\text{obs}} = 27$ events on the 3-jet, 4-jet, ... channels.
- ☐ We compare y_{obs} with its distribution P_0 supposing that H_0 is true.
- ☐ Here P_0 is $\mathrm{Poiss}(\lambda_0 = 6.7)$ and represents the baseline noise under H_0.
- ☐ We compute the **P-value**

$$p_{\text{obs}} = \mathrm{P}_0(Y \geq y_{\text{obs}}) = \sum_{y=y_{\text{obs}}}^{\infty} \frac{\lambda_0^y}{y!} e^{-\lambda_0} = 3 \times 10^{-9},$$

  so
  - either H_0 is true but a (very) rare event has occurred,
  - or H_0 is false and the top quark exists.
- ☐ Abe et al. announced a discovery, but if they had found $p_{\text{obs}} \approx 0.001$, maybe they would have decided that H_0 could not (yet) be rejected, and not published their work.

Industrial fraud?



- $n = 92$ weighings of sacks on the 'delivery' (or not?) of a commodity:

261	289	291	265	281	291	285	283	280	261	263	281	291	289	280
292	291	282	280	281	291	282	280	286	291	283	282	291	293	291
300	302	285	281	289	281	282	261	282	291	291	282	280	261	283
291	281	246	249	252	253	241	281	282	280	261	265	281	283	280
242	260	281	261	281	282	280	241	249	251	281	273	281	261	281
282	260	281	282	241	245	253	260	261	281	280	261	265	281	241
260	241													

- Their last digits are

0	1	2	3	4	5	6	7	8	9
14	42	14	9	0	6	2	0	0	5

- How can we tell if fraud has taken place?

Pearson's statistic

Definition 58 *If $O_1, \ldots, O_K$ are the numbers of observations from a random sample of size n falling in categories $1, \ldots, K$, where $\mathrm{E}(O_k) = E_k > 0$ for $k = 1, \ldots, K$ and $\sum_{k=1}^K E_k = n$, then* **Pearson's statistic (aka the 'χ^2 statistic')** *is*

$$T = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k}.$$

- If
$$(O_1, \ldots, O_K) \sim \mathrm{Mult}\{n, (p_1 = E_1/n, \ldots, p_K = E_K/n)\},$$
then $T \dot\sim \chi^2_{K-1}$ (approximation OK if average $E_k \geq 5$), giving a test of whether data $O_1, \ldots, O_K$ agree with specified probabilities $p_1, \ldots, p_K$.
- Here Benford's law suggests all $p_k \doteq 1/10$, so take $E_k = 92/10 = 9.2$.
- For the original dataset we found $t_{\mathrm{obs}} = 158.2$ and hence
$$p_{\mathrm{obs}} = \mathrm{P}_0(T > t_{\mathrm{obs}}) \doteq \mathrm{P}(\chi^2_9 \geq 158.2) \doteq 0,$$
which is essentially impossible for uniformly distributed digits.
- Massive evidence for non-uniformity (and for industrial fraud?)

Elements of a test

- ☐ A **null hypothesis** H_0 to be tested.
- ☐ A **test statistic** T, large values of which will suggest that H_0 is false, and with observed value t_{obs}.
- ☐ A **P-value**

$$p_{\text{obs}} = \mathrm{P}_0(T \geq t_{\text{obs}}),$$

where the **null distribution** $\mathrm{P}_0(\cdot)$ denotes a probability computed under H_0.
- ☐ The smaller p_{obs} is, the more we doubt that H_0 is true.
- ☐ p_{obs} is a realisation of a **P-variable** P, which is $U(0,1)$ under H_0 (if T is continuous), so

$$\mathrm{P}_0(P \leq p_{\text{obs}}) = p_{\text{obs}}.;$$

T is chosen so that P is more likely to be small if H_0 is false.
- ☐ If I decide that H_0 is false, when in fact it is true, then I make an error whose probability under H_0 is exactly p_{obs} — so my uncertainty is quantified, because I know the probability of declaring a "**false positive**".

Note: Why is a P-value uniform?

- ☐ Let T be a test statistic whose distribution is $F_0(t)$ when the null hypothesis is true. Then the corresponding P-value is

$$\mathrm{P}_0(T \geq t_{\text{obs}}) = 1 - F_0(t_{\text{obs}}),$$

and if the value of t_{obs} is a realisation of T_{obs} (because the null hypothesis is true), then we can write the random value of p_{obs} seen in repetitions of the experiment as

$$P_{\text{obs}} = 1 - F_0(T_{\text{obs}}),$$

or equivalently $T_{\text{obs}} = F_0^{-1}(1 - P_{\text{obs}})$. Hence for $x \in [0,1]$,

$$\begin{aligned}
\mathrm{P}_0(P_{\text{obs}} \leq x) &= \mathrm{P}_0\left\{1 - F_0(T_{\text{obs}}) \leq x\right\} \\
&= \mathrm{P}_0\left\{1 - x \leq F_0(T_{\text{obs}})\right\} \\
&= \mathrm{P}_0\left\{T_{\text{obs}} \geq F_0^{-1}(1-x)\right\} \\
&= 1 - F_0\left\{F_0^{-1}(1-x)\right\} \\
&= x,
\end{aligned}$$

which shows that $P_{\text{obs}} \sim U(0,1)$.
- ☐ The above proof works for any continuous T_{obs}, but is only approximate if T_{obs} is discrete (e.g., has a Poisson distribution). In such cases P_{obs} can only take a finite or countable number of values known as the **achievable significance levels**.

Exact and inexact tests

- ☐ $P \sim U(0,1)$ under H_0, exactly in continuous cases and approximately in discrete cases.
- ☐ If the null distribution of the test statistic is estimated, we have $P \dot{\sim} U(0,1)$ only.
- ☐ For example, if the true parameter is $\theta = (\psi_0, \lambda_0)$ and $H_0 : \psi = \psi_0$, then the P-value is

$$p_{\text{obs}} = \mathrm{P}_0(T \geq t_{\text{obs}}) = \mathrm{P}(T \geq t_{\text{obs}}; \psi_0, \lambda_0),$$

which we estimate by

$$\widehat{p}_{\text{obs}} = \mathrm{P}(T \geq t_{\text{obs}}; \psi_0, \widehat{\lambda}_0),$$

where $\widehat{\lambda}_0$ is the estimate of λ under H_0.

- ☐ Exact tests, with $P \sim U(0,1)$, can sometimes be obtained by using a pivot whose distribution is invariant to λ, or by removing λ by conditioning or marginalisation.

Example 59 *If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, show that the distribution of $T = (\overline{Y} - \mu)/\sqrt{S^2/n}$ is invariant to σ^2.*

Example 60 *Find an exact test on a canonical parameter in a logistic regression model.*

Note to Example 60

- ☐ here $\overline{Y}$ and S^2 are minimal sufficient and independent, with $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$ and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, and we can write $\overline{Y} \overset{\text{D}}{=} \mu + \sigma n^{-1/2} Z$ and $S^2 = \overset{\text{D}}{=} \sigma^2 V/(n-1)$, where $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_{n-1}$ are independent. Hence

$$T = \frac{\overline{Y} - \mu}{\sqrt{S^2/n}} \overset{\text{D}}{=} \frac{\mu + \sigma Z/n^{-1/2} - \mu}{[\sigma^2 V/\{n(n-1)\}]^{1/2}} \overset{\text{D}}{=} \frac{Z}{\sqrt{V/(N-1)}} \sim t_{n-1},$$

is pivotal and thus allows tests on μ without reference to σ^2.

- ☐ For a test on σ^2 without regard to μ, we use the marginal distribution of $\S^2$, as $V = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ is a pivot.

Note to Example 59

- In a logistic regression model we have independent binary variables $Y_1, \ldots, Y_n$ each with density

$$\mathrm{P}(Y_j = y_j; \beta) = \pi_j^{y_j}(1-\pi_j)^{1-y_j} = \left(\frac{e^{x_j^{\mathrm{T}}\beta}}{1+e^{x_j^{\mathrm{T}}\beta}}\right)^{y_j}\left(\frac{1}{1+e^{x_j^{\mathrm{T}}\beta}}\right)^{1-y_j} = \frac{e^{y_j x_j^{\mathrm{T}}\beta}}{1+e^{x_j^{\mathrm{T}}\beta}},$$

  for $y_j \in \{0,1\}$, known covariate vectors $X_j \in \mathbb{R}^d$ and parameter $\beta \in \mathbb{R}^d$.
- The corresponding log likelihood is

$$\ell(\beta) = \sum_{j=1}^n \left\{y_j x_j^{\mathrm{T}}\beta - \log\left(1+e^{x_j^{\mathrm{T}}\beta}\right)\right\} = y^{\mathrm{T}}X\beta - \sum_{j=1}^n \log\left(1+e^{x_j^{\mathrm{T}}\beta}\right), \quad \beta \in \mathbb{R}^d.$$

  This is a (d,d) exponential family with canonical statistic $S = X^{\mathrm{T}}y$, canonical parameter $\varphi = \beta$, and cumulant generator $k(\varphi) = \sum_{j=1}^n \log\left(1+e^{x_j^{\mathrm{T}}\varphi}\right)$.
- Hence Lemma 40 implies that if $\varphi = (\psi, \lambda)$ and $S = (T, W) = (X_1^{\mathrm{T}}y, X_2^{\mathrm{T}}y)$, where X_1 is $n \times 1$ and X_2 is $n \times (d-1)$, an exact test on ψ is obtained from the conditional distribution

$$\mathrm{P}(T = t \mid W = w^{\mathrm{o}}; \psi) = \frac{e^{t\psi}}{\sum_{y' \in \mathcal{S}_{w^{\mathrm{o}}}} e^{X_1^{\mathrm{T}}y'\psi}},$$

  where $\mathcal{S}_w = \{(y_1', \ldots, y_n') : X_2^{\mathrm{T}}y' = w^{\mathrm{o}}\}$, with $w^{\mathrm{o}} = X_2^{\mathrm{T}}y^{\mathrm{o}}$ and y^{o} respectively the observed data and the observed value of W.
- Calculation of this conditional density in applications may be awkward, but excellent approximations are available.

Comments

- If we say that a hypothesis is **true**, we mean 'it is reasonable to proceed as if the hypothesis was true' — any model is an idealisation, so a hypothesis cannot be exactly 'true'.
- If we have a **discrete test statistic**, p_{obs} has at most a countable number of 'achievable significance levels'. This is only problematic when comparing tests, though randomisation has (unfortunately) sometimes been proposed to overcome it.
- We may consider a **two-sided test**, with both unusually large and unusually small values of T of interest. We can then define

$$p_+ = \mathrm{P}_0(T \geq t_{\mathrm{obs}}), \quad p_- = \mathrm{P}_0(T \leq t_{\mathrm{obs}}), \quad p_{\mathrm{obs}} = 2\min(p_-, p_+),$$

  so $p_- + p_+ = 1 + \mathrm{P}_0(T = t_{\mathrm{obs}})$, which equals 1 unless T is discrete;
- We sometimes avoid minor problems due to discreteness by computing '**continuity-corrected**' P-values

$$p_+ = \sum_{t > t_{\mathrm{obs}}} \mathrm{P}_0(T = t) + \tfrac{1}{2}\mathrm{P}_0(T = t_{\mathrm{obs}}), \quad p_- = \sum_{t < t_{\mathrm{obs}}} \mathrm{P}_0(T = t) + \tfrac{1}{2}\mathrm{P}_0(T = t_{\mathrm{obs}}).$$

- So far we have described **pure significance tests**, where the situation if H_0 is false is not explicitly considered. We look at the effect of alternatives now.

Testing as decision-making

Neyman and Pearson formulated testing as deciding between two hypotheses:

- the **null hypothesis** H_0, which represents a baseline situation;
- the **alternative hypothesis** H_1, which represents what happens if H_0 is false.
- We choose H_1 and 'reject' H_0 if $p_{\rm obs}$ is lower than some $\alpha \in (0,1)$.
- For given α we partition the sample space $\mathcal{Y}$ into

$$\mathcal{Y}_0 = \{y \in \mathcal{Y} : p_{\rm obs}(y) > \alpha\}, \quad \mathcal{Y}_1 = \{y \in \mathcal{Y} : p_{\rm obs}(y) \le \alpha\},$$

  where the notation $p_{\rm obs}(y)$ indicates that the P-value depends on the data, or equivalently

$$\mathcal{Y}_0 = \{y \in \mathcal{Y} : t(y) < t_{1-\alpha}\}, \quad \mathcal{Y}_1 = \{y \in \mathcal{Y} : t(y) \ge t_{1-\alpha}\},$$

  where t_p denotes the p quantile of the test statistic $T = t(Y)$ under H_0.
- We call $\mathcal{Y}_1$ the **size α critical region** of the test, and we reject H_0 in favour of H_1 if $Y \in \mathcal{Y}_1$, or equivalently if the test statistic exceeds the **size α critical point** $t_{1-\alpha}$.
- Critical regions of different sizes for the same test should be nested, i.e., (in an obvious notation) if $\alpha' > \alpha$, then

$$\mathcal{Y}_1^{\alpha} \subset \mathcal{Y}_1^{\alpha'} \quad \text{and} \quad t_{1-\alpha} > t_{1-\alpha'}.$$

Link to confidence sets

- In a test on a parameter θ, with hypothesis $H_0 : \theta = \theta_0$ and corresponding size α critical region $\mathcal{Y}_1(\theta_0)$, we reject H_0 at level α if

$$p_{\rm obs}(y;\theta_0) < \alpha \quad \Longleftrightarrow \quad y \in \mathcal{Y}_1(\theta_0).$$

- A $(1-\alpha)$ confidence set $\mathcal{C}_{1-\alpha}$ for the 'true value' of θ, i.e., the value that generated the data, is the set of all values of θ_0 for which H_0 is not rejected at significance level α, i.e.,

$$\mathcal{C}_{1-\alpha} = \{\theta : p_{\rm obs}(y;\theta) \ge \alpha\} = \{\theta : y \notin \mathcal{Y}_1(\theta)\}.$$

- This links hypothesis testing and confidence intervals, and enables construction of the latter in general settings, by this process of **test inversion**.

False positives and negatives

		Decision	
		Accept H_0	Reject H_0
State of Nature	H_0 true	Correct choice (True negative)	Type I Error (False positive)
	H_1 true	Type II Error (False negative)	Correct choice (True positive)

- ☐ We can make two sorts of wrong decision:

  Type I error (false positive): H_0 is true, but we wrongly reject it (and choose H_1);

  Type II error (false negative): H_1 is true, but we wrongly choose H_0.
- ☐ Statistics books and papers call
  - the **Type I error/false positive probability** the **size** $\alpha = \mathrm{P}_0(Y \in \mathcal{Y}_1)$, and
  - the **true positive probability** the **power** $\beta = \mathrm{P}_1(Y \in \mathcal{Y}_1)$.
- ☐ Note that losses due to wrong decisions are not taken into account.

Example 61 *If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, with σ^2 known, $H_0 : \mu = \mu_0$ and $H_1 : \mu = \mu_1$, find the Type II error as a function of the Type I error.*

Note to Example 60

- ☐ The minimal sufficient statistic for the normal model with both parameters unknown is $(\overline{Y}, S^2)$, and it is easy to check that if σ^2 is known the minimal sufficient statistic reduces to $\overline{Y}$, which has a $\mathcal{N}(\mu_0, \sigma^2/n)$ distribution under H_0. Hence we take the test statistic T to be $\overline{Y}$, and $\mathcal{Y} = \mathbb{R}^n$.
- ☐ If $\mu_1 > \mu_0$, then clearly we will take

$$\mathcal{Y}_0 = \{y : \overline{y} < t_{1-\alpha}\}, \quad \mathcal{Y}_1 = \{y : \overline{y} \ge t_{1-\alpha}\};$$

  this can be justified using the Neyman–Pearson lemma (below). Now

$$\mathrm{P}_0(Y \in \mathcal{Y}_0) = \mathrm{P}_0(\overline{Y} < t_{1-\alpha}) = \mathrm{P}_0\{\sqrt{n}(\overline{Y}-\mu_0)/\sigma < \sqrt{n}(t_{1-\alpha}-\mu_0)/\sigma\} = \Phi\left\{\sqrt{n}(t_{1-\alpha}-\mu_0)/\sigma\right\},$$

  because $Z = \sqrt{n}(\overline{Y}-\mu_0)/\sigma \sim \mathcal{N}(0,1)$ under H_0, and for this probability to equal $1-\alpha$ we must take $t_{1-\alpha} = \mu_0 + \sigma n^{-1/2} z_{1-\alpha}$; this gives Type I error α.
- ☐ Although the form of $\mathcal{Y}_0$ is determined by H_1, the value of $t_{1-\alpha}$ is given by calculations under H_0.
- ☐ $Z = \sqrt{n}(\overline{Y}-\mu_1)/\sigma \sim \mathcal{N}(0,1)$ under H_1, so the Type II error is

$$\begin{aligned}
\mathrm{P}_1(Y \in \mathcal{Y}_0) &= \mathrm{P}_1(\overline{Y} < t_{1-\alpha}) \\
&= \mathrm{P}_1(\overline{Y} < \mu_0 + \sigma n^{-1/2} z_{1-\alpha}) \\
&= \mathrm{P}_1\{\sqrt{n}(\overline{Y}-\mu_1)/\sigma < \sqrt{n}(\mu_0 + \sigma n^{-1/2} z_{1-\alpha} - \mu_1)/\sigma\} \\
&= \Phi(z_{1-\alpha} - \delta),
\end{aligned}$$

  where $\delta = n^{1/2}(\mu_1 - \mu_0)/\sigma$. Hence the Type II error equals $1-\alpha$ when $\mu_1 = \mu_0$ and decreases as a function of δ. We would expect this, because as μ_1 increases, the distribution of $\overline{Y}$ under H_1 shifts to the right and we are less likely to make a false negative error.

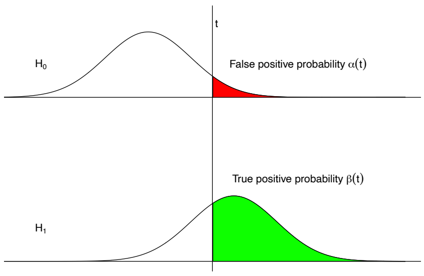
True and false positives: Example

- ☐ It is traditional to fix α and choose T (or equivalently $\mathcal{Y}_1$) to maximise β, but usually more informative to consider $\mathrm{P}_0(T \geq t)$ and $\mathrm{P}_1(T \geq t)$ as functions of t.
- ☐ In Example 60 we would
  - reject H_0 incorrectly (false positive) with probability

$$\alpha(t) = \mathrm{P}_0(T \geq t) = 1 - \Phi\{n^{1/2}(t - \mu_0)/\sigma\},$$

  - reject H_0 correctly (true positive) with probability

$$\beta(t) = \mathrm{P}_1(T \geq t) = 1 - \Phi\{n^{1/2}(t - \mu_0)/\sigma - \delta\}.$$



ROC curve

Definition 62 *The* **receiver operating characteristic (ROC) curve** *of a test plots $\beta(t)$ against $\alpha(t)$ as t varies, i.e., it shows the graph $(x, y) = (\mathrm{P}_0(T \geq t), \mathrm{P}_1(T > t))$, when $t \in \mathbb{R}$.*

- ☐ As μ increases, it becomes easier to detect when H_0 is false, because the densities under H_0 and H_1 become more separated, and the ROC curve moves 'further north-west'.
- ☐ When H_0 and H_1 are the same then the curve lies on the diagonal, and the hypotheses cannot be distinguished.
- ☐ One summary measure of the overall quality of a test is the **area under the curve**,

$$\mathrm{AUC} = \int_0^1 \beta(\alpha)\,\mathrm{d}\alpha,$$

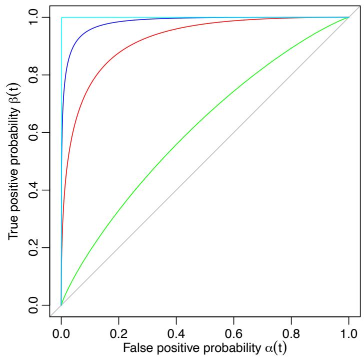
which ranges between 0.5 for a useless test and 1.0 for a perfect test.

Example

- In Example 60 $\alpha(t) = 1 - \Phi\{n^{1/2}(t - \mu_0)/\sigma\}$ and $\beta(t) = 1 - \Phi\{n^{1/2}(t - \mu_0)/\sigma - \delta\}$, so equivalently we graph

$$\beta(t) = 1 - \Phi(-z_{1-\alpha} - \delta) = \Phi(\delta + z_\alpha) \equiv \beta(\alpha) \text{ against } \alpha \in (0,1).$$

- Here is the ROC curve with $\mu = 2$ (in red). Also shown are curves for $\mu = 0, 0.4, 3, 6$. Which is which?



Neyman–Pearson lemma

Definition 63 *A* **simple hypothesis** *entirely fixes the distribution of the data Y, whereas a* **composite hypothesis** *does not fix the distribution of Y.*

Definition 64 *The* **critical region** *of a hypothesis test is the subset $\mathcal{Y}_1$ of the sample space $\mathcal{Y}$ for which $Y \in \mathcal{Y}_1$ implies that the null hypothesis is rejected.*

We aim to choose $\mathcal{Y}_1$ to maximise the power of the test for a given size, i.e., such that $\mathrm{P}_1(Y \in \mathcal{Y}_1)$ is as large as possible provided $\mathrm{P}_0(Y \in \mathcal{Y}_1) \le \alpha$ (with equality in continuous problems).

Lemma 65 (Neyman–Pearson) *Let $f_0(y)$, $f_1(y)$ be the densities of Y under simple null and alternative hypotheses. Then if it exists, the set*

$$\mathcal{Y}_1 = \{y \in \mathcal{Y} : f_1(y)/f_0(y) > t\}$$

such that $\mathrm{P}_0(Y \in \mathcal{Y}_1) = \alpha$ maximises $\mathrm{P}_1(Y \in \mathcal{Y}_1)$ amongst all $\mathcal{Y}_1'$ for which $\mathrm{P}_0(Y \in \mathcal{Y}_1') \le \alpha$. Thus the test of size α with maximal power rejects H_0 when $Y \in \mathcal{Y}_1$.

Example 66 *Construct an optimal test for testing $H_0 : \varphi = \varphi_0$ against $H_1 : \varphi = \varphi_1$ based on a random sample from a canonical exponential family.*

Note to Lemma 64

Suppose that a region $\mathcal{Y}_1$ such that $\mathrm{P}_0(Y \in \mathcal{Y}_1) = \alpha$ exists and let $\mathcal{Y}_1'$ be any other critical region of size α or less. Note that $\mathcal{Y}_0 \cup \mathcal{Y}_1 = \mathcal{Y}_0' \cup \mathcal{Y}_1' = \mathcal{Y}$. If we write $F(\mathcal{C}) = \int_{\mathcal{C}} f(y)\,\mathrm{d}y$ for any density f with corresponding distribution F, then we aim to show that $F_1(\mathcal{Y}_1) \geq F(\mathcal{Y}_1')$. Now

$$\int_{\mathcal{Y}_1} f(y)\,\mathrm{d}y - \int_{\mathcal{Y}_1'} f(y)\,\mathrm{d}y = F(\mathcal{Y}_1) - F(\mathcal{Y}_1') \qquad (5)$$

equals

$$F(\mathcal{Y}_1 \cap \mathcal{Y}_1') + F(\mathcal{Y}_1 \cap \mathcal{Y}_0') - F(\mathcal{Y}_1' \cap \mathcal{Y}_1) - F(\mathcal{Y}_1' \cap \mathcal{Y}_0) = F(\mathcal{Y}_1 \cap \mathcal{Y}_0') - F(\mathcal{Y}_1' \cap \mathcal{Y}_0). \qquad (6)$$

If $F = F_0$, then (**??**) is non-negative, because $\alpha = F_0(\mathcal{Y}_1) \geq F_0(\mathcal{Y}_1')$, so (**??**) is also non-negative, giving

$$tF_0(\mathcal{Y}_1 \cap \mathcal{Y}_0') \geq tF_0(\mathcal{Y}_1' \cap \mathcal{Y}_0), \quad t \geq 0.$$

But $f_1(y) > tf_0(y)$ for $y \in \mathcal{Y}_1$, and $tf_0(y) \geq f_1(y)$ for $y \in \mathcal{Y}_0$, so

$$F_1(\mathcal{Y}_1 \cap \mathcal{Y}_0') \geq tF_0(\mathcal{Y}_1 \cap \mathcal{Y}_0') \geq tF_0(\mathcal{Y}_1' \cap \mathcal{Y}_0) \geq F_1(\mathcal{Y}_1' \cap \mathcal{Y}_0).$$

On adding $F_1(\mathcal{Y}_1 \cap \mathcal{Y}_1')$ to both sides we see that $F_1(\mathcal{Y}_1) \geq F(\mathcal{Y}_1')$, as required.

Note to Example 65

- The likelihood ratio is

$$\frac{f_1(y)}{f_0(y)} = \frac{m^*(y)\exp\{\varphi_1 s^* - nk(\varphi_1)\}}{m^*(y)\exp\{\varphi_0 s^* - nk(\varphi_0)\}} = \exp\{(\varphi_1 - \varphi_0)s^* + nk(\varphi_0) - nk(\varphi_1)\},$$

say, where $s^* = \sum_{j=1}^n s(y_j)$, so

$$\mathcal{Y}_1 = \{y : f_1(y)/f_0(y) > t\} = \{y : (\varphi_1 - \varphi_0)s^* + nk(\varphi_0) - nk(\varphi_1) > \log t\},$$

and if $\varphi_1 > \varphi_0$ then

$$\mathcal{Y}_1 = \{y : s^* > [\log t + nk(\varphi_1) - nk(\varphi_0)]/(\varphi_1 - \varphi_0)\}.$$

This gives the form of $\mathcal{Y}_1$ and we should choose t so that $\mathrm{P}_0(Y \in \mathcal{Y}_1) = \alpha$, or equivalently s_α so that (in the continuous case)

$$\mathrm{P}_0(S^* > s_\alpha) = \int_{s_\alpha}^{\infty} f(s;\varphi_0)\,\mathrm{d}s = \alpha.$$

Example 60 gave such a calculation for normal data with $\varphi_1 = \mu_1/\sigma^2 > \varphi_0 = \mu_0/\sigma^2$ and known σ^2.

- If $\varphi_1 < \varphi_0$, then division by $\varphi_1 - \varphi_0 < 0$ leads to

$$\mathcal{Y}_1^* = \{y : s^* < [\log t + nk(\varphi_1) - nk(\varphi_0)]/(\varphi_1 - \varphi_0)\}.$$

- The Neyman–Pearson lemma tell us that $\mathcal{Y}_1$ gives a most powerful test, but as it does not depend on the value of φ, this test is **uniformly most powerful** for all $\varphi > \varphi_0$, and likewise $\mathcal{Y}_1^*$ is **uniformly most powerful** for $\varphi_1 < \varphi_0$.

Power

- The NP lemma applies to simple hypotheses, but sometimes (e.g., Example 65) gives **uniformly most powerful (UMP) tests** against composite alternatives, i.e., a single critical region $\mathcal{Y}_1$ is most powerful against $\theta = \theta_1$ for all $\theta_1 > \theta_0$ or for all $\theta_1 < \theta_0$.
- If there is no UMP region, we might compare tests of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ by
  - comparing them at some (arbitrary) 'typical' alternative;
  - averaging power over some suitable set of alternatives; or
  - looking at local alternatives, i.e., when $\theta_1 = \theta_0 + \delta$ for small δ.
- For local alternatives, note that with scalar θ and mild regularity of the log likelihood,

$$\log\left\{\frac{f(y;\theta_0+\delta)}{f(y;\theta_0)}\right\} = \ell(\theta_0+\delta) - \ell(\theta_0) = \delta\frac{\mathrm{d}\ell(\theta_0)}{\mathrm{d}\theta} + o(\delta) = \delta\ell_\theta(\theta_0) + o(\delta).$$

- Hence the **locally most powerful critical region** for $\delta > 0$ is obtained from large values of the score statistic, and conversely for $\delta < 0$.
- When $\theta = (\psi, \lambda)$ and we test the composite hypothesis $H_0 : \psi = \psi_0$ against $H_0 : \psi > \psi_0$, without constraints on λ, the optimal local test for each λ will be based on the score $\ell_\psi(\theta) = \partial\ell(\psi,\lambda)/\partial\psi$ evaluated at (ψ_0, λ), which unless λ can somehow be eliminated is often replaced in practice by $(\psi_0, \widehat{\lambda}_{\psi_0})$.

Aside: Score testing

- Score tests can be useful when maximising a full likelihood is difficult or not worthwhile.
- Suppose we want to test $H_0 : \theta = \theta_0$ for scalar θ. Under H_0 and classical asymptotics,

$$\ell_\theta(\theta_0) \dot\sim \mathcal{N}(0, \imath(\theta_0)\} \quad \Longrightarrow \quad \ell_\theta(\theta_0)/\sqrt{\imath(\theta_0)} \dot\sim \mathcal{N}(0,1),$$

  which gives a basis for the test.
- When $\theta = (\psi, \lambda)$ and $H_0 : \psi = \psi_0$, then

$$\ell_\psi(\widehat{\theta}_0) \dot\sim \mathcal{N}(0, \imath^{\psi\psi}(\widehat{\theta}_0)^{-1}\} \quad \Longrightarrow \quad \ell_\psi(\widehat{\theta}_0)^{\mathrm{T}}\imath^{\psi\psi}(\widehat{\theta}_0)\ell_\psi(\widehat{\theta}_0) \dot\sim \chi^2_{\dim\psi},$$

  where $\widehat{\theta}_0 = (\psi_0, \widehat{\lambda}_{\psi_0})$ and

$$\imath^{\psi\psi}(\theta)^{-1} = \imath_{\psi\psi}(\theta) - \imath_{\psi\lambda}(\theta)\imath_{\lambda\lambda}(\theta)^{-1}\imath_{\lambda\psi}(\theta).$$

  If ψ is scalar, then $\ell_\psi(\widehat{\theta}_0)\{\imath^{\psi\psi}(\widehat{\theta}_0)\}^{1/2} \dot\sim \mathcal{N}(0,1)$.
- In both cases
  - any maximisation is needed only on H_0, and
  - if the expected information is difficult to compute, it can be replaced by the corresponding observed information (if this is positive).

Discussion: Interpretation of P-values

- Be careful about interpretation:
  - p_{obs} is a one-number summary of whether data are consistent with H_0;
  - it is NOT the probability that H_0 is true;
  - even a tiny p_{obs} can support H_0 better than an alternative H_1 (consider $t_{\text{obs}} = 3$ when $T \sim \mathcal{N}(\mu, 1)$ with $\mu_0 = 0$, $\mu_1 = 10$);
  - the power depends on analogues of $\delta = n^{1/2}(\mu_1 - \mu_0)/\sigma$, where n is the **sample size**, $\mu_1 - \mu_0$ is the **effect size**, and σ is the **precision**, so
    - ▷ even a tiny (practically irrelevant) effect size can be detected with very large n;
    - ▷ conversely a practically important effect might be undetectable if n is small;
    - ▷ i.e., 'statistical significance' $\neq$ 'subject-matter importance'!
- A confidence interval, or estimate and its standard error, is often more informative.
- Hypothesis testing is often applied by rote — in some medical journals no statement is complete without an accompanying '$(P < 0.05)$' — and is sometimes regarded as controversial, with certain journals now refusing to publish tests and P-values.
- The 'replication crisis' is partly due to abuse of hypothesis testing, e.g., by not correcting for multiple tests, by formulating hypotheses in light of the data, ...

Discussion: Contexts of testing

- It is unwise to be too categorical about testing, because of its different uses:
  - testing a clear hypothesis of scientific interest (e.g., top quark);
  - goodness of fit of a model (e.g., industrial fraud);
  - decision-making with a clearly-specified alternative (e.g., covid testing);
  - model simplification if null hypothesis true (e.g., score test for gamma shape);
  - 'dividing hypothesis' used to partition the parameter space into subsets with sharply different interpretations;
  - as a technical device for generating confidence intervals;
  - to flag which of many similar null hypotheses might be false.
- Hence arguing that testing should be abolished is unreasonable (as well as unrealistic).

Example 67 *The generalized Pareto distribution, with survival function*

$$\mathrm{P}(X > x) = \begin{cases} (1 + \xi x/\sigma)_+^{-1/\xi}, & \xi \neq 0, \\ \exp(-x/\sigma), & \xi = 0, \end{cases}$$

simplifies if $\xi = 0$, and has finite upper support point $x_+ = -\sigma/\xi$ when $\xi < 0$ but $x_+ = \infty$ when $\xi \geq 0$. Here $H_0 : \xi = 0$ is both a simplifying and a dividing hypothesis, of interest (for example) when the distribution is fitted to data on supercentenarians (finite or infinite limit to human life?).

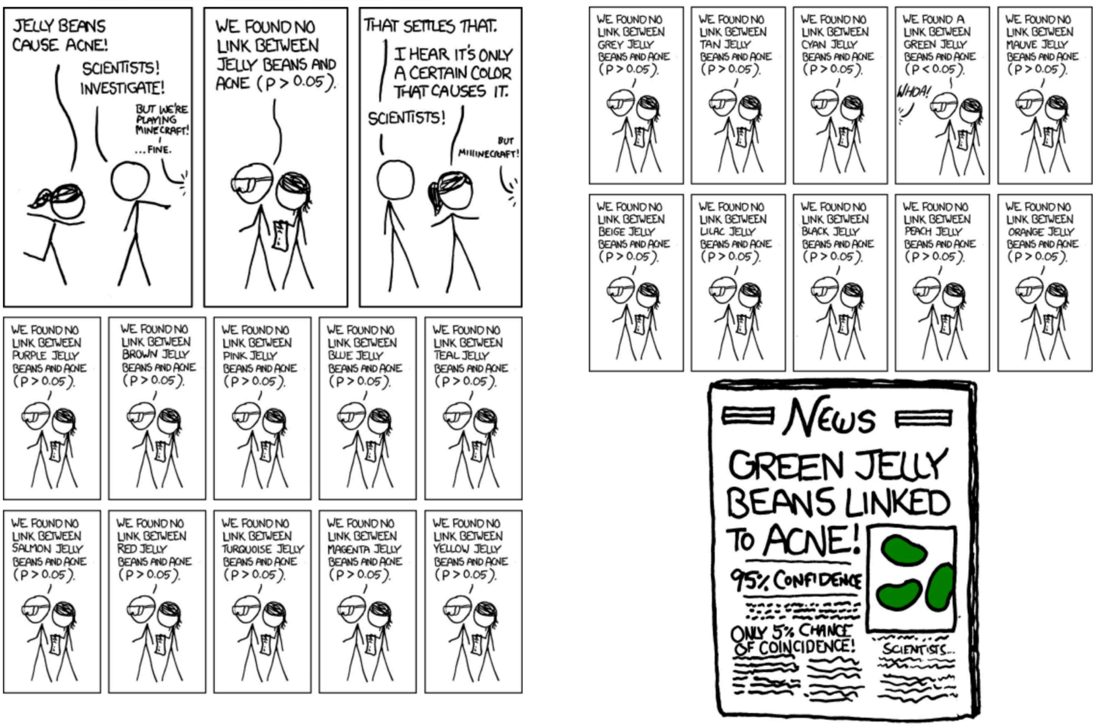
Motivation

- Often require tests of several, even very many, hypotheses:
  - comparison of responses for several treatment groups with the same control group;
  - checking for a change in a series of observations;
  - screening genomic data for effects of many genes on a response.
- There are null hypotheses $H_1, \dots, H_m$, of which
  - m_0 are true, indexed by an unknown set $\mathcal{I}$,
  - $m_1 = m - m_0$ are false, and
  - the **global null hypothesis** is $H_0 = H_1 \cap \cdots \cap H_m$.
- We apply some testing procedure and declare R hypotheses to be significant, of which FP are false positives and TP are true positives. Only R and m are known.

	Non-significant	Significant	
True nulls	TN	FP	m_0
False nulls	FN	TP	$m - m_0$
		R	m

- In the cartoon on the next slide we have $m = 20$ hypotheses individually tested with $\alpha = 0.05$. We observe $R = 1$, but $\mathrm{E}(\mathrm{FP}) = m\alpha = 1$, so this is not a surprise.

The perils of multiple testing



Graphical approach

- Graphs can be helpful in suggesting which hypotheses are most suspect, and can highlight the corresponding (i.e., smallest) P-values.
- $P \sim U(0,1)$ implies $Z = -\log_{10} P \sim \exp(\lambda)$ with $\lambda = \ln 10$.
- With this transformation small P_j become large Z_j; note that $Z_j > a$ iff $P_j < 10^{-a}$.
- If H_0 is true and the tests are independent, then $Z_1, \ldots, Z_m \stackrel{\text{iid}}{\sim} \exp(\lambda)$ and the **Rényi representation**

$$Z_{(r)} \stackrel{\mathrm{D}}{=} \lambda^{-1} \sum_{j=1}^{r} \frac{E_j}{m+1-j}, \quad r = 1, \ldots, m, \quad E_1, \ldots, E_m \stackrel{\text{iid}}{\sim} \exp(1),$$

  applies to their order statistics. Then
  - a plot of the ordered empirical Z_j against their expectations should be straight;
  - outliers, very large Z_j (i.e., very small P_j), cast doubt on the corresponding H_j.
  - For very small P_j (i.e., large Z_j) the uniformity may fail even under H_0, because the null distributions give poor tail approximations; then some form of model-fitting may be needed.
  - Similar ideas apply to z statistics (e.g., in regression): use a normal QQ-plot (excluding the intercept etc.) as a basis for discussion of significant effects.

GWAS, I

- A **genome-wide association study (GWAS)** tests the association between SNPs ('single nucleotide polymorphisms') and a phenotype such as the expression of a protein. The null hypotheses are

$$H_{0,j} : \text{no association between the expression of the protein and SNP}_j, \quad j = 1, \ldots, m.$$

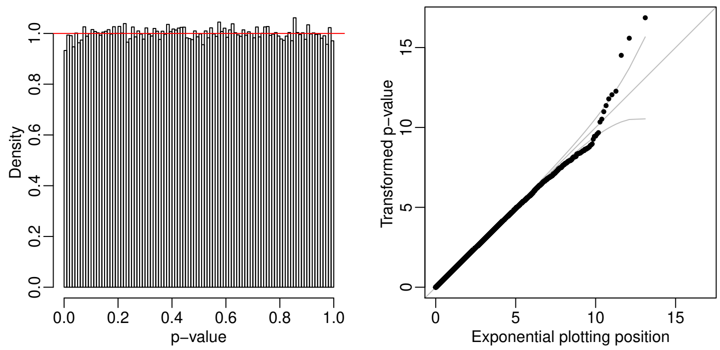
- In a simple model we construct statistics Y_j such that $Y_j \dot\sim \mathcal{N}(\theta_j, 1)$, where $\theta_j = 0$ under $H_{0,j}$, and we take $T_j = |Y_j|$, which is likely to be far from zero if $\theta_j \gg 0$ or$\theta_j \ll 0$.
- If $t_{\text{obs},j}$ denotes the observed value of T_j, then the P-value for association j is

$$p_{\text{obs},j} = \mathrm{P}_0(T_j > t_{\text{obs},j}) = 1 - \mathrm{P}_0(-t_{\text{obs},j} \le Y_j \le t_{\text{obs},j}) \doteq 2\Phi(-t_{\text{obs},j}),$$

  where the approximation comes from the fact that $Y_j \dot\sim \mathcal{N}(0,1)$ under $H_{0,j}$.
- Here it is reasonable to expect that the effects are **sparse**, i.e., most of the $\theta_j = 0$, and we seek a needle in a haystack.
- With many tests it is essential to ensure that the true positives are not drowned in the mass of false positives.

GWAS, II

- Left: a histogram of the P-values for tests of the association between $m = 275297$ SNPs and the expression of the protein CFAB.
- The P-values for SNPs not associated with CFAB are uniformly distributed. Is there an excess of small P-values?
- Right: exponential Q-Q plot of the $Z_j = -\log P_j$. What do you make of it?



Control

- With several tests Type I error generalises to the **familywise error rate (FWER)**, i.e., the probability of at least one false positive when the individual hypotheses are tested,

$$\text{FWER} = \text{P}(\text{FP} \geq 1) = 1 - \text{P}(\text{accept all } H_j, j \in \mathcal{I}),$$

  and we aim to control this by ensuring that $\text{FWER} \leq \alpha$.
- Control of the error rate:
  - **weak control** guarantees $\text{FWER} \leq \alpha$ only under H_0, i.e., $m_0 = m$;
  - **strong control** guarantees $\text{FWER} \leq \alpha$ for any configuration of null and alternative hypotheses.
- If all the tests are independent with individual levels all equal to α, then

$$\text{FWER} = 1 - \text{P}(\text{FP} = 0) = 1 - (1-\alpha)^{m_0} \to 1, \quad m_0 \to \infty.$$

- If conversely we fix FWER and the tests are independent we need

$$\alpha = 1 - (1 - \text{FWER})^{1/m_0},$$

  so with $m_0 = 20$ and $\text{FWER} = 0.05$ we need $\alpha \doteq 0.0026$ — the power for individual tests will be tiny (recall ROC curves).

Bonferroni methods

- If P_j is the P-value for the jth test and we reject H_j if $P_j < \alpha_j$, then **Boole's inequality** (the first **Bonferroni inequality**, aka the **union bound**) gives

$$\mathrm{FWER} = \mathrm{P}(\mathrm{FP} \geq 1) = \mathrm{P}\left(\bigcup_{j=1}^{m_0} \{P_j \leq \alpha_j\}\right) \leq \sum_{j=1}^{m_0} \mathrm{P}\left(P_j \leq \alpha_j\right) = \sum_{j=1}^{m_0} \alpha_j,$$

  so even if the tests are dependent we have strong control of FWER if $\sum_{j=1}^{m} \alpha_j \leq \alpha$.
- Usually we set $\alpha_j \equiv \alpha/m$, so $\sum_{j=1}^{m_0} \alpha_j = m_0\alpha/m \leq \alpha$.
- The resulting **Bonferroni procedure** lacks power when m is large (because α/m is very small), but its assumptions are very weak.
- An improvement is the **Holm–Bonferroni procedure**: for given α,
  - order the P-values as $P_{(1)} \leq \cdots \leq P_{(m)}$ and the hypotheses as $H_{(1)}, \ldots, H_{(m)}$, then
  - reject $H_{(1)}, \ldots, H_{(S-1)}$, where

$$S = \min\left\{s : P_{(s)} > \frac{\alpha}{m+1-s}\right\}.$$

  This still gives strong control but is more powerful than the basic Bonferroni procedure, because it uses higher rejection thresholds. Hence the basic procedure should not be used.

Note: Holm–Bonferroni procedure (HB)

- Recall that there are m hypotheses, of which m_0 are true nulls (for which $j \in \mathcal{I}$) and $m_1 = m - m_0$ are false nulls.
- If we apply HB and $\mathrm{FP} \geq 1$, we must have wrongly rejected some H_j with $j \in \mathcal{I}$. If $H_{(s)}$ is the first such hypothesis to be rejected in the sequential procedure, then the $s-1$ hypotheses rejected before it must have been false null hypotheses, so $s - 1 \leq m_1 = m - m_0$, i.e., $m_0 \leq m + 1 - s$.
- As $H_{(s)}$ was rejected, the corresponding P-value satisfies

$$P_{(s)} \leq \frac{\alpha}{m+1-s} \leq \frac{\alpha}{m_0}.$$

  Thus if $\mathrm{FP} \geq 1$ then the P-value for at least one of the true null hypotheses satisfies $P_j \leq \alpha/m_0$, and Boole's inequality gives

$$\mathrm{FWER} = \mathrm{P}(\mathrm{FP} \geq 1) \leq \mathrm{P}\left(\bigcup_{j \in \mathcal{I}} \{P_j \leq \alpha/m_0\}\right) \leq \sum_{j=1}^{m_0} \mathrm{P}\left(P_j \leq \alpha/m_0\right) = m_0\alpha/m_0 = \alpha.$$

- The only assumption needed above was that the null P-values are $U(0,1)$ (used in Boole's inequality), so HB strongly controls the FWER.

False discovery rate

- When m is large and the goal is exploratory, Bonferroni procedures are unreasonably stringent, and it seems preferable to try and control the **false discovery proportion**

$$I(R > 0)\mathrm{FP}/R,$$

where R is the number of rejected null hypotheses. The aim is to bound the proportion of false positives among the rejections.

- Control of $I(R > 0)\mathrm{FP}/R$ is impossible because the set of true null hypotheses $\mathcal{I}$ is unknown, so instead we try and control the **false discovery rate (FDR)**

$$\mathrm{FDR} = \mathrm{E}\{I(R > 0)\mathrm{FP}/R\}.$$

- The **Benjamini–Hochberg procedure** gives strong control for independent tests: specify α, then
  - order the P-values as $P_{(1)} \leq \cdots \leq P_{(m)}$ and the hypotheses as $H_{(1)}, \ldots, H_{(m)}$,
  - reject $H_{(1)}, \ldots, H_{(R)}$, where

$$R = \max\left\{r : P_{(r)} < \frac{r\alpha}{m}\right\}.$$

This guarantees that $\mathrm{FDR} \leq \alpha$, but does not bound the <u>actual</u> proportion of false positives, just its expectation. Often $\alpha = 0.1$, 0.2, $\ldots$.

Note: Derivation of the Benjamini–Hochberg procedure

- Let the P-values for the false null hypotheses be $P_1', \ldots, P_{m_1}'$, say, independent of the true null P-values $P_1, \ldots, P_{m_0} \stackrel{\text{iid}}{\sim} U(0,1)$. Then the number of rejected hypotheses R satisfies

$$\{R = r\} \cap \{P_1 \le r\alpha/m\} = \{P_1 \le r\alpha/m\} \cap \{R_{-1} = r-1\},$$

where $\{R_{-1} = r-1\}$ is the event that there are exactly $r-1$ rejections among $H_2, \ldots, H_m$. The false discovery proportion is

$$\sum_{r=1}^{m} \frac{\text{FP}}{r} I(R = r) = \sum_{r=1}^{m} \frac{I(R=r)}{r} \sum_{j=1}^{m_0} I(P_j \le r\alpha/m),$$

and by symmetry of the P_j this has the same expectation as

$$m_0 \sum_{r=1}^{m} \frac{I(R=r)}{r} I(P_1 \le r\alpha/m) = m_0 \sum_{r=1}^{m} \frac{I(R_{-1} = r-1)}{r} I(P_1 \le r\alpha/m).$$

Thus the false discovery rate is

$$\begin{aligned}
\text{FDR} &= m_0 \sum_{r=1}^{m} \frac{1}{r} \text{P}(R_{-1} = r-1, P_1 \le r\alpha/m) \\
&= m_0 \sum_{r=1}^{m} \frac{1}{r} \text{P}(R_{-1} = r-1 \mid P_1 \le r\alpha/m) \text{P}(P_1 \le r\alpha/m) \\
&= m_0 \sum_{r=1}^{m} \frac{1}{r} \text{P}(R_{-1} = r-1) \frac{r\alpha}{m} \\
&= \frac{m_0 \alpha}{m} \sum_{r=0}^{m-1} \text{P}(R_{-1} = r) \\
&= \frac{m_0 \alpha}{m} \le \alpha.
\end{aligned}$$

The main steps above successively use the definition of conditional probability, the facts that P_1 and R_{-1} are independent and $P_1 \sim U(0,1)$, and the fact that $R_{-1} \in \{0, 1, \ldots, m-1\}$.

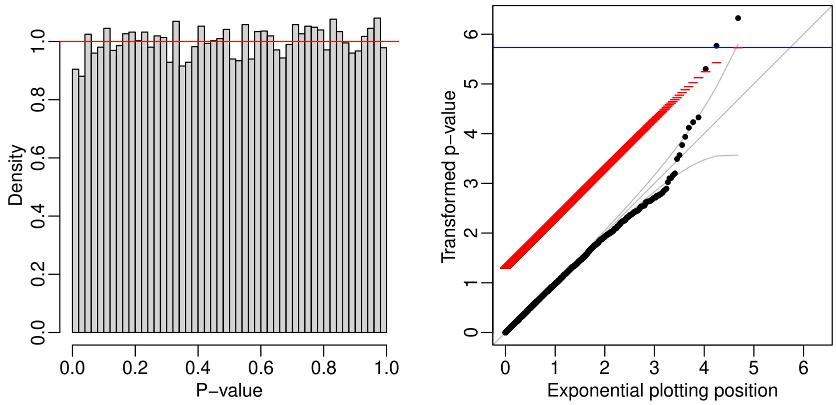
- Hence (under the conditions above) the Benjamini–Hochberg procedure strongly controls the FDR.
- Note that
  - if $m_0 \ll m$, then the last inequality may be very unequal, so possibly FDR $\ll \alpha$.
  - if the P-values are dependent in such a way that

$$\text{P}(R_{-1} = r-1 \mid P_1 \le r\alpha/m) \le \text{P}(R_{-1} = r-1),$$

then the result also holds.

GWAS, II

- Left: histogram of $Q_j = 10P_j$ (when $P_j < 0.1$) for tests of the association between $m = 27530$ SNPs and the expression of the protein CFAB, and the $U(0,1)$ density (red).
- Right: exponential Q-Q plot of $Z_j = -\log_{10} Q_j$, with Bonferroni cutoff (blue) and Benjamini–Hochberg cutoffs (red), both with $\alpha = 0.05$. The grey lines are the target and pointwise 95% confidence sets for the order statistics.



Comments

- The Holm–Bonferroni procedure (HB) compares $P_{(1)}, P_{(2)}, \ldots$ to $\alpha/m, \alpha/(m-1), \ldots$, whereas the ordinary Bonferroni procedure (B) compares all the P_j to α/m.
- The **Simes procedure** (exercises) has exact FWER α for independent tests and then is preferable to the Holm–Bonferroni procedure.
- The Benjamini–Hochberg procedure (BH) strongly controls the false discovery rate, comparing the ordered P-values to $\alpha/m, 2\alpha/m, \ldots, \alpha$.
- HB and B also give strong control when the P-values are dependent. So does BH, taking

$$P_{(j)} \leq \frac{j\alpha}{mc(m)},$$

  with $c(m) = 1$ when the tests are independent or positively dependent, and $c(m) = \sum_{j=1}^{m} 1/j$ under arbitrary dependence.
- Many variants exist, but these versions are simple and widely used.
- Other classical procedures for multiple testing in regression settings are named after
  - Tukey — bounds the maximum of t statistics for different tests;
  - Scheffé — simultaneously bounds all possible linear combinations of estimates $\widehat{\beta}$;
  - Dunnett — compares different treatments with the same control.