Testing

Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch

November 10, 2025

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

- 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, $y_{\mathrm{obs}} = 27$ events on the 3-jet, 4-jet, ... channels.
- We compare y_{obs} with its distribution Pr_0 supposing that H_0 is true.
- Here Pr_0 is $\mathrm{Poiss}(\lambda_0 = 6.7)$ and represents the baseline noise under H_0.
- We compute the **P-value**

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

  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 $p_{\mathrm{obs}} \approx 0.001$, maybe they would have decided that H_0 could not (yet) be rejected, and not published their work.

- $n = 92$ 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

- How can we tell if fraud has taken place?

Definition

*For $x \in \mathbb{R}$, let $d(x, j)$ denote the 'j**th significant digit function (base 10)**', so $d(31.4, 1) = 3$, $d(0.314, 2) = 1$ and $d(314, 3) = 4$.*

Definition

If $x \in \mathbb{R}$ and $D_j = d(x, j)$, for $j = 1, 2, \ldots$, then (discarding any leading zeros) the D_j follow **Benford's law** *if*

$$\Pr(D_1 = d_1, D_2 = d_2, \ldots, D_k = d_k) = \log_{10}\left\{1 + \left(\sum_{j=1}^{k} d_j \times 10^{k-j}\right)^{-1}\right\}, \quad d_j \in \{0, \ldots, 9\}.$$

- For example, $\Pr(D_1 = 3, D_2 = 1, D_3 = 4) = \log_{10}\{1 + (314)^{-1}\} \approx 0.0014$.
- This is considered an excellent model for the distribution of all sorts of digits.
- Frequencies (%) of the last digits D_3 for three-digit integers.

Digit	0	1	2	3	4	5	6	7	8	9
Uniform	10	10	10	10	10	10	10	10	10	10
Benford	10.178	10.137	10.097	10.057	10.017	9.978	9.940	9.901	9.864	9.826

- Apparently the last digits of the weighings should be approximately uniform.
- To detect forged deliveries we test the null hypothesis

$$H_0: \text{ the last digits of the weighings are uniformly distributed on } 0, \ldots, 9.$$

Definition

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 $T \dot{\sim} \chi^2_{K-1}$, 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
$$p_{\mathrm{obs}} = \mathrm{Pr}_0(T > t_{\mathrm{obs}}) \doteq \mathrm{Pr}(\chi^2_9 \geq 158.2) \doteq 0,$$
which is essentially impossible for uniformly distributed digits.
- Massive evidence for non-uniformity (and for industrial fraud?)

- 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 $t_{\rm obs}$.
- A **P-value**
$$p_{\rm obs} = {\rm Pr}_0(T \geq t_{\rm obs}),$$
where the **null distribution** ${\rm Pr}_0(\cdot)$ denotes a probability computed under H_0.
- The smaller $p_{\rm obs}$ is, the more we doubt that H_0 is true.
- $p_{\rm obs}$ is a realisation of a **P-variable** P, which is $U(0,1)$ under H_0 (if T is continuous), so
$${\rm Pr}_0(P \leq p_{\rm obs}) = p_{\rm obs}.$$
- If based on $p_{\rm obs}$ I wrongly decide that H_0 is false, then I make an error whose probability under H_0 is exactly $p_{\rm obs}$ — so my uncertainty is quantified, because I know the probability of declaring a "**false positive**".

- $P \sim U(0,1)$ 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

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

  which we estimate by

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

  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

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.

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

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

  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.

- 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**, $p_{\rm obs}$ has at most a countable number of ‘achievable significance levels’. This is only problematic when comparing tests, though randomisation has 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

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

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

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

- 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.

- Fisher regarded a P-value as a **measure of the evidence** against H_0.
- Neyman and Pearson formulated testing as **making a decision** 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_{obs} is lower than some $\alpha \in (0, 1)$.
- For given α we partition the sample space $\mathcal{Y}$ into

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

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

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

  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

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

- 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

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

- 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.,

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

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

		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** $\alpha = \mathrm{Pr}_0(Y \in \mathcal{Y}_1)$, and
  - the **true positive probability** the **power** $\beta = \mathrm{Pr}_1(Y \in \mathcal{Y}_1)$.

Example

If $Y_1, \ldots, Y_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, 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.

- The minimal sufficient statistic 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

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

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

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

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

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

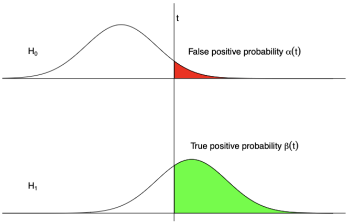
where $\delta = n^{1/2}(\mu_1 - \mu_0)/\sigma$. Hence the Type II error equals $1 - \alpha$ when $\mu_1 = \mu_0$ and decreases as a function of δ. We would expect this, because as μ_1 increases, the distribution of $\overline{Y}$ under H_1 shifts to the right and we are less likely to make a false negative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

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

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

  - reject H_0 correctly (true positive) with probability

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



Definition

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

- As difference in hypotheses 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**,

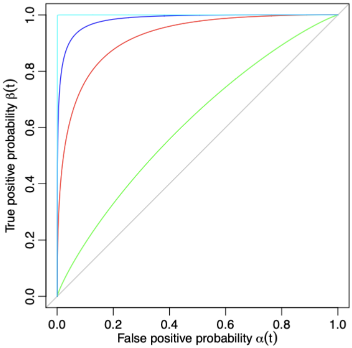
$$\mathrm{AUC} = \int_0^1 \beta(\alpha)\, \mathrm{d}\alpha,$$

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

- In the our example $\alpha(t) = 1 - \Phi\{n^{1/2}(t - \mu_0)/\sigma\}$ 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 $\delta = 2$ (in red). Also shown are curves for $\delta = 0, 0.4, 3, 6$.



Our decision must be based on the sample, so we need to define:

Definition (Test Function)

A test function is a map $\delta : \mathcal{Y}^n \to \{0, 1\}$.

Obtaining 0 or 1 must be decided on whether or not the sample satisfies a certain condition:

$$\delta(Y_1, \ldots, Y_n) = \begin{cases} 1, & \text{if } T(Y_1, \ldots, Y_n) \in C, \\ 0, & \text{if } T(Y_1, \ldots, Y_n) \notin C, \end{cases}$$

where

- T is a statistic called a *test statistic* and
- C is a subset of the range of T, called *critical region*.

In compact form

$$\delta(Y_1, \ldots, Y_n) = \mathbf{1}\{T(Y_1, \ldots, Y_n) \in C\}.$$

- To choose good test functions we need to quantify the performance of a test function.

Remark that, obviously, δ is just a Bernoulli random variable:

$$\delta = \begin{cases} 1, & \text{with probability } \mathbb{P}[T(Y_1, \ldots, Y_n) \in C], \\ 0, & \text{with probability } \mathbb{P}[T(Y_1, \ldots, Y_n) \notin C]. \end{cases}$$

- So a good test function must have a sampling distribution concentrated around the right decision.
- The difference from point estimation is that our action space is discrete.
- Can we get an analogue of mean squared error?

By an abuse of terminology, we could define:

$$\mathrm{MSE}(\delta, H_i) = \mathbb{E}_\theta[(\delta - i)^2], \qquad i \in \{0, 1\}.$$

Since δ is Bernoulli, and i takes values in $\{0, 1\}$, we have

$$\begin{aligned}
\mathrm{MSE}(\delta, H_i) = \mathbb{E}_\theta[(\delta - i)^2] = \mathbb{E}_\theta[|\delta - i|] &= \begin{cases} \mathbb{E}_\theta[\delta], & \text{if } \theta \in \Theta_0, \\ 1 - \mathbb{E}_\theta[\delta], & \text{if } \theta \in \Theta_1. \end{cases} \\
&= \begin{cases} \mathbb{P}_\theta[\delta = 1], & \text{if } \theta \in \Theta_0, \\ 1 - \mathbb{P}_\theta[\delta = 1], & \text{if } \theta \in \Theta_1. \end{cases} \\
&= \begin{cases} \mathbb{P}_\theta[\delta = 1], & \text{if } \theta \in \Theta_0, \\ \mathbb{P}_\theta[\delta = 0], & \text{if } \theta \in \Theta_1. \end{cases}
\end{aligned}$$

In decision theory terms, the action space is $\mathcal{A} = \{0, 1\}$ and the loss function is the so-called "0–1" loss,

$$\mathcal{L}(a, \theta) = \begin{cases} 1 & \text{if } \theta \in \Theta_0 \ \& \ a = 1 \qquad \text{(Type I Error)} \\ 1 & \text{if } \theta \in \Theta_1 \ \& \ a = 0 \qquad \text{(Type II Error)} \\ 0 & \text{otherwise} \qquad\qquad\quad\ \ \text{(No Error)} \end{cases}$$

i.e. we lose 1 unit whenever committing a type I or type II error.

The risk function then becomes

$$R(\delta, \theta) = \begin{cases} \mathbb{E}_\theta[\mathbf{1}\{\delta = 1\}] = \mathbb{P}_\theta[\delta = 1] & \text{if } \theta \in \Theta_0 \quad \text{(prob of type I error)} \\ \mathbb{E}_\theta[\mathbf{1}\{\delta = 0\}] = \mathbb{P}_\theta[\delta = 0] & \text{if } \theta \in \Theta_1 \quad \text{(prob of type II error)} \end{cases}$$

In short,

$$R(\delta, \theta) = \mathbb{P}_\theta[\delta = 1]\mathbf{1}\{\theta \in \Theta_0\} + \mathbb{P}_\theta[\delta = 0]\mathbf{1}\{\theta \in \Theta_1\}$$

Can we hope to simultaneously control both type I and II error probabilities? $\hookrightarrow$ Unfortunately the answer is no.

Here's why let $\delta(Y_1, \ldots, Y_n) = \mathbf{1}\{T(Y_1, \ldots, Y_n) \in C\}$ and suppose we wish to reduce the type I error probability

$$\mathbb{P}_\theta[\delta = 1], \qquad \theta \in \Theta_0,$$

for all $\theta \in \Theta_0$.

To do this, we must replace C by a subset $C_* \subset C$, obtaining

$$\delta_* = \mathbf{1}\{T(Y_1, \ldots, Y_n) \in C_*\}.$$

Observe that, $\forall\, \theta \in \Theta_0$,

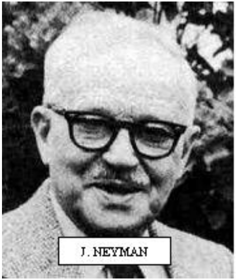
$$\mathbb{P}_\theta[\delta_* = 1] = \mathbb{P}[T(Y_1, \ldots, Y_n) \in C_*] \leq \mathbb{P}[T(Y_1, \ldots, Y_n) \in C] = \mathbb{P}_\theta[\delta = 1]$$

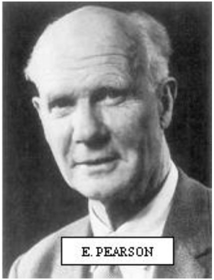
On the other hand $C_* \subset C \implies C_*^c \supset C^c$ and so $\forall\, \theta \in \Theta_1$

$$\mathbb{P}_\theta[\delta_* = 0] = \mathbb{P}[T(Y_1, \ldots, Y_n) \notin C_*] \geq \mathbb{P}[T(Y_1, \ldots, Y_n) \notin C] = \mathbb{P}_\theta[\delta = 0].$$

By reducing the type I error probability we increased the type II error probability

We need to make some concessions...





The fundamental paradigm of *Neyman and Pearson* informally dictates:

1. In applications, one type of error (false positive or negative) is typically more severe.
2. Say this is the type I error, and exploit the asymmetry: fix a tolerance ceiling for the probability of this error.
3. Given this ceiling, consider only test functions that respect it, and focus on minimising type II error (i.e. maximising power).

In mathematical terms:

The Neyman-Pearson Framework

1. We fix an $\alpha \in (0,1)$, usually small (called the significance level)
2. We declare that we only consider test functions $\delta : \mathcal{X} \to \{0,1\}$ such that
$$\delta \in \mathcal{D}(\Theta_0, \alpha) = \left\{\delta : \sup_{\theta \in \Theta_0} \mathbb{P}_\theta[\delta = 1] \le \alpha\right\}$$
i.e. rules for which prob of type I error is bounded above by α

$\hookrightarrow$ *Jargon: we fix a significance level for our test*

3. Within this restricted class of rules, choose δ to minimize prob of type II error:
$$\mathbb{P}_\theta[\delta(\boldsymbol{X}) = 0] = 1 - \mathbb{P}_\theta[\delta(\boldsymbol{X}) = 1], \qquad \theta \in \Theta_1$$
4. Equivalently, maximize the *power*
$$\beta(\theta, \delta) = \mathbb{P}_\theta[\delta(\boldsymbol{X}) = 1] = \mathbb{E}_\theta[\mathbf{1}\{\delta(\boldsymbol{X}) = 1\}] = \mathbb{E}_\theta[\delta(\boldsymbol{X})], \quad \theta \in \Theta_1$$
(since $\delta = 1 \iff \mathbf{1}\{\delta = 1\} = 1$ and $\delta = 0 \iff \mathbf{1}\{\delta = 1\} = 0$)

- Neyman-Pearson setup naturally exploits any asymmetric structure
- But, if natural asymmetry absent, need judicious choice of H_0

Consider the simplest situation:

$$\Theta_0 = \{\theta_0\} \qquad \& \qquad \Theta_1 = \{\theta_1\}$$

The Neyman-Pearson Lemma - Continuous Case

Let $\boldsymbol{Y}$ have joint density/frequency $f \in \{f_0, f_1\}$ and suppose we wish to test

$$H_0 : f = f_0 \qquad vs \qquad H_1 : f = f_1.$$

If $\Lambda(\boldsymbol{Y}) = f_1(\boldsymbol{Y})/f_0(\boldsymbol{Y})$ is a continuous random variable, then there exists a $k > 0$ such that

$$\mathbb{P}_0[\Lambda(\boldsymbol{Y}) \geq k] = \alpha$$

and the test whose test function is given by

$$\delta(\boldsymbol{Y}) = \mathbf{1}\{\Lambda(\boldsymbol{Y}) \geq k\},$$

is a *most powerful (MP)* test of H_0 versus H_1 at significance level α.

Proof.

Use obvious notation $\mathbb{E}_0$, $\mathbb{E}_1$, $\mathbb{P}_0$, $\mathbb{P}_1$ corresponding to H_0 or H_1. Let $G_0(t) = \mathbb{P}_0[\Lambda \leq t]$. By assumption, G_0 is a differentiable distribution function, and so is onto $[0, 1]$. Consequently, the set $\mathcal{K}_{1-\alpha} = \{t : G_0(t) = 1 - \alpha\}$ is non-empty for any $\alpha \in (0, 1)$. Setting $k = \inf\{t \in \mathcal{K}_{1-\alpha}\}$ we will have $\mathbb{P}_0[\Lambda \geq k] = \alpha$ and k is simply the $1 - \alpha$ quantile of the distribution G_0. Consequently,

$$\mathbb{P}_0[\delta = 1] = \alpha \qquad (\text{since } \mathbb{P}_0[\delta = 1] = \mathbb{P}_0[\Lambda \geq k])$$

and therefore δ respects the level α rule for the null hypothesis.
To show that δ is also most powerful, it suffices to prove that if ψ is any function with $\psi(\boldsymbol{y}) \in \{0, 1\}$, then

$$\mathbb{E}_0[\psi(\boldsymbol{Y})] \leq \underbrace{\mathbb{E}_0[\delta(\boldsymbol{Y})]}_{=\alpha\text{(by first part of proof)}} \implies \underbrace{\mathbb{E}_1[\psi(\boldsymbol{Y})]}_{\beta_1(\psi)} \leq \underbrace{\mathbb{E}_1[\delta(\boldsymbol{Y})]}_{\beta_1(\delta)}.$$

(recall that $\beta_1(\delta) = 1 - \mathbb{P}_1[\delta = 0] = \mathbb{P}_1[\delta = 1] = \mathbb{E}_1[\delta]$).

WLOG assume that f_0 and f_1 are density functions. Note that

$$f_1(\boldsymbol{y}) - k \cdot f_0(\boldsymbol{y}) \geq 0 \text{ if } \delta(\boldsymbol{y}) = 1 \quad \& \quad f_1(\boldsymbol{y}) - k \cdot f_0(\boldsymbol{y}) < 0 \text{ if } \delta(\boldsymbol{y}) = 0.$$

Therefore, since ψ can only take the values 0 or 1,

$$\begin{aligned}
\psi(\boldsymbol{y})(f_1(\boldsymbol{y}) - k \cdot f_0(\boldsymbol{y})) &\leq \delta(\boldsymbol{y})(f_1(\boldsymbol{y}) - k \cdot f_0(\boldsymbol{y})) \\
\int_{\mathbb{R}^n} \psi(\boldsymbol{y})(f_1(\boldsymbol{y}) - k \cdot f_0(\boldsymbol{y})) d\boldsymbol{y} &\leq \int_{\mathbb{R}^n} \delta(\boldsymbol{y})(f_1(\boldsymbol{y}) - k \cdot f_0(\boldsymbol{y})) d\boldsymbol{y}
\end{aligned}$$

Rearranging the terms yields

$$\begin{aligned}
\int_{\mathbb{R}^n} (\psi(\boldsymbol{y}) - \delta(\boldsymbol{y})) f_1(\boldsymbol{y}) d\boldsymbol{y} &\leq k \int_{\mathbb{R}^n} (\psi(\boldsymbol{y}) - \delta(\boldsymbol{y})) f_0(\boldsymbol{y}) d\boldsymbol{y} \\
\implies \mathbb{E}_1[\psi(\boldsymbol{Y})] - \mathbb{E}_1[\delta(\boldsymbol{Y})] &\leq k \left(\mathbb{E}_0[\psi(\boldsymbol{y})] - \mathbb{E}_0[\delta(\boldsymbol{Y})]\right)
\end{aligned}$$

But $k > 0$ by assumption, so when $\mathbb{E}_0[\psi(\boldsymbol{Y})] \leq \mathbb{E}_0[\delta(\boldsymbol{Y})]$ the RHS is negative, i.e. δ is an MP test of H_0 vs H_1 at level α. $\square$

- Basically we reject if the likelihood of θ_0 is k times higher than the likelihood of θ_1. This is called a likelihood ratio test, and Λ is the likelihood ratio statistic: *how much more plausible is the alternative than the null?*
- When Λ is a continuous RV, the choice of k is essentially unique. That is, if k' is such that $\delta' = \mathbf{1}\{\Lambda \geq k'\} \in \mathcal{D}(\{\theta_0\}, \alpha)$, then $\delta = \delta'$ almost surely.
- The resulting most powerful test is not necessarily unique.
- Unless Λ is continuous, the most powerful test is not necessarily guaranteed to exist.
- The problem if Λ is a RV with a discontinuous dist is that there may exist no k for which the equation $\mathbb{P}_0[\Lambda \geq k] = \alpha$ has a solution.
- In any case, typically the distribution of the test statistic converges to a continuous limit with large n, so these problems become inessential.

Example (Poisson Distribution)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} Poisson(\mu)$ and for $\mu_1 > \mu_0$ consider the hypotheses:

$$H_0 : \mu = \mu_0 \qquad vs \qquad H_1 : \mu = \mu_1.$$

Applying the Neyman-Pearson lemma gives a test statistic

$$\delta(Y_1, \ldots, Y_n) = \mathbf{1}\left\{\sum_{i=1}^n Y_i > q_{1-\alpha}\right\},$$

provided α is such that $G_0(q_{1-\alpha}) = \mathbb{P}_{\mu_0}[\tau(Y_1, \ldots, Y_n) \leq q_{1-\alpha}] \overset{!}{=} 1 - \alpha$. Since the Y_i are independent, one can easily show that

$$\tau(Y_1, \ldots, Y_n) \overset{H_0}{\sim} Poisson(n\mu_0).$$

This being a discrete distribution, the only α for which we get an MP test are

$$e^{-n\mu_0}, e^{-n\mu_0}\left(1 + n\mu_0\right), e^{-n\mu_0}\left(1 + n\mu_0 + \frac{(n\mu_0)^2}{2}\right), \ldots \text{and so on}$$

Nevertheless notice that as $n \to \infty$, these values become dense near the origin.

When $\{\Theta_0, \Theta_1\}$ are not singletons, choosing a most powerful test is a much stronger requirement:

1. It should respect the level for all $\theta \in \Theta_0$, i.e.

$$\delta \in \mathcal{D}(\Theta_0, \alpha) = \{\delta : \mathcal{Y}^n \to \{0, 1\} : \mathbb{E}_\theta[\delta] \leq \alpha,\ \forall\, \theta \in \Theta_0\}$$

2. It should be most powerful for all $\theta \in \Theta_1$ (i.e. for all possible simple alternatives),

$$\mathbb{E}_\theta[\delta] \geq \mathbb{E}_\theta[\delta'] \qquad \forall \theta \in \Theta_1 \quad \& \quad \delta' \in \mathcal{D}(\Theta_0, \alpha)$$

Unfortunately UMP tests rarely exist. Why?
$\hookrightarrow$ Consider $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$

- A UMP test must be MP test for any $\theta \neq \theta_1$.
- But the form of the MP test typically differs for $\theta_1 > \theta_0$ and $\theta_1 < \theta_0$!
  - $\hookrightarrow$ e.g. recall exponential mean example

Example (No UMP test exists)

Let $Y_1, ..., Y_n \sim$Bernoulli(θ) and suppose we want to test:

$$H_0 : \theta = \theta_0 \qquad vs \qquad H_1 : \theta \neq \theta_0$$

at some level α. To this aim, consider first

$$H_0' : \theta = \theta_0 \qquad vs \qquad H_1' : \theta = \theta_1$$

Neyman-Pearson lemma gives test statistics

$$T = \frac{f(\mathbf{Y};\theta_1)}{f(\mathbf{Y};\theta_0)} = \left(\frac{1-\theta_1}{1-\theta_0}\right)^n \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{\sum_{i=1}^n Y_i}$$

- If $\theta_1 > \theta_0$ then T increasing in $\sum_{i=1}^n Y_i$
  $\hookrightarrow$ MP test would reject for large values of $\sum_{i=1}^n Y_i$
- If $\theta_1 < \theta_0$ then T decreasing in $\sum_{i=1}^n Y_i$
  $\hookrightarrow$ MP test would reject for small values of $\sum_{i=1}^n Y_i$

- The NP lemma applies to simple hypotheses, but sometimes 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 $(\psi_0, \widehat{\lambda}_{\psi_0})$.
- General hypothesis pairs: we need to abandon optimality, and search for sensible tests. But the likelihood ratio idea can serve us well in this pursuit.

Consider now the multiparameter case $\boldsymbol{\theta} \in \mathbb{R}^p$ with general Θ_0, Θ_1

- As noted optimality breaks down.
- But we can still seek general-purpose approaches.

The idea: Combine Neyman-Pearson paradigm with Maximum Likelihood

Definition (Likelihood Ratio)

The *likelihood ratio statistic* corresponding to the pair of hypotheses $H_0 : \boldsymbol{\theta} \in \Theta_0$ vs $H_1 : \boldsymbol{\theta} \in \Theta_1$ is defined to be

$$\Lambda(\boldsymbol{Y}) = \frac{\sup_{\boldsymbol{\theta} \in \Theta} f(\boldsymbol{Y}; \theta)}{\sup_{\boldsymbol{\theta} \in \Theta_0} f(\boldsymbol{Y}; \theta)} = \frac{\sup_{\boldsymbol{\theta} \in \Theta} L(\theta)}{\sup_{\boldsymbol{\theta} \in \Theta_0} L(\theta)}$$

- Intuition: choose the "most favourable" $\theta \in \Theta_0$ (in favour of H_0) and compare it against the "most favourable" $\theta \in \Theta_1$ (in favour of H_1) in a simple vs simple setting (applying NP-lemma)
- Typically Θ_0 is a lower dimensional subspace of Θ_1, so taking sup over Θ (rather than Θ_1) incurs no loss. In this case $\Theta_0 \cap \Theta_1 \neq \emptyset$, but $Leb(\Theta_0 \cap \Theta_1) = 0$, which suffices.

Example

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Consider:

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_1 : \mu \neq \mu_0$$

$$\Lambda(\boldsymbol{Y}) = \frac{\sup_{(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}^+} f(\boldsymbol{Y};\mu,\sigma^2)}{\sup_{(\mu,\sigma^2)\in\{\mu_0\}\times\mathbb{R}^+} f(\boldsymbol{Y};\mu,\sigma^2)} = \left(\frac{\sum_{i=1}^n (Y_i-\mu_0)^2}{\sum_{i=1}^n (Y_i-\overline{Y})^2}\right)^{\frac{n}{2}} = \left(\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}^2}\right)^{\frac{n}{2}}$$

So reject when $\Lambda \geq k$, where k is s.t. $\mathbb{P}_0[\Lambda \geq k] = \alpha$. Distribution of Λ? By monotonicity look only at

$$\begin{aligned} \frac{\sum_{i=1}^n (Y_i-\mu_0)^2}{\sum_{i=1}^n (Y_i-\overline{Y})^2} &= 1 + \frac{n(\overline{Y}-\mu_0)^2}{\sum_{i=1}^n (Y_i-\overline{Y})^2} = 1 + \frac{1}{n-1}\left(\frac{n(\overline{Y}-\mu_0)^2}{S^2}\right) \\ &= 1 + \frac{T^2}{n-1} \end{aligned}$$

With $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2$ and $T = \sqrt{n}(\overline{Y} - \mu_0)/S \overset{H_0}{\sim} t_{n-1}$.

So $T^2 \overset{H_0}{\sim} F_{1,n-1}$ and k may be chosen appropriately.

Example

Let $Y_1, \ldots, Y_m \overset{iid}{\sim} \text{Exp}(\lambda)$ and $Z_1, \ldots, Z_n \overset{iid}{\sim} \text{Exp}(\theta)$. Assume $\boldsymbol{Y}$ indep $\boldsymbol{Z}$.

$$\text{Consider:} \quad H_0 : \theta = \lambda \qquad \text{vs} \qquad H_1 : \theta \neq \lambda$$

i.e. $(\theta, \lambda) \in R_+^2$ against $(\theta, \lambda) \in$ 45 degree line

$$\underset{\sup_{(\lambda,\theta)\in\mathbb{R}_+^2} f(\boldsymbol{Y},\boldsymbol{Z};\lambda,\theta)}{\text{Unrestricted MLEs:}} \quad \widehat{\lambda} = 1/\overline{Y} \qquad \& \qquad \widehat{\theta} = 1/\overline{Z}$$

$$\underset{\sup_{(\lambda,\theta)\in\{(y,z)\in\mathbb{R}_+^2 : y=z\}} f(\boldsymbol{Y},\boldsymbol{Z};\lambda,\theta)}{\text{Restricted MLEs:}} \quad \widehat{\lambda}_0 = \widehat{\theta}_0 = \left[\frac{m\overline{Y}+n\overline{Z}}{m+n}\right]^{-1}$$

$$\implies \Lambda = \left(\frac{m}{m+n} + \frac{n}{n+m}\frac{\overline{Z}}{\overline{Y}}\right)^m \left(\frac{n}{n+m} + \frac{m}{m+n}\frac{\overline{Y}}{\overline{Z}}\right)^n$$

Depends on $T = \overline{Y}/\overline{Z}$ and can make Λ large/small by varying T.

$\hookrightarrow$ But $T \overset{H_0}{\sim} F_{2m,2n}$ so given α we may find the critical value k.

More often than not, dist(Λ) intractable
$\hookrightarrow$(and no simple dependence on T with tractable distribution either)

Consider asymptotic approximations?

Setup

- Θ open subset of $\mathbb{R}^p$
- either $\Theta_0 = \{\boldsymbol{\theta}_0\}$ or Θ_0 open subset of $\mathbb{R}^s$, where $s < p$
- Concentrate on $\boldsymbol{Y} = (Y_1, ..., Y_n)$ has iid components.
- Initially restrict attention to $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. LR becomes:

$$\Lambda_n(\boldsymbol{Y}) = \prod_{i=1}^{n} \frac{f(Y_i; \widehat{\boldsymbol{\theta}}_n)}{f(Y_i; \boldsymbol{\theta}_0)}$$

  where $\widehat{\boldsymbol{\theta}}_n$ is the MLE of $\boldsymbol{\theta}$.
- Impose regularity conditions from MLE asymptotics

Theorem (Wilks' Theorem, case $p = 1$)

Let $Y_1, ..., Y_n$ be iid random variables with density (frequency) depending on $\theta \in \mathbb{R}$ and satisfying conditions (A1)-(A6), with $\imath_1(\theta) = \jmath_1(\theta)$. If the MLE sequence $\widehat{\theta}_n$ is consistent for θ, then the likelihood ratio statistic Λ_n for $H_0 : \theta = \theta_0$ satisfies

$$2 \log \Lambda_n \xrightarrow{d} V \sim \chi_1^2$$

when H_0 is true.

- Obviously, knowing approximate distribution of $2 \log \Lambda_n$ is as good as knowing approximate distribution of Λ_n for the purposes of testing (by monotonicity and rejection method).
- Theorem extends immediately and trivially to the case of general p and for a hypothesis pair $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.
  (i.e. when null hypothesis is simple)

Proof.

Under the conditions of the theorem and when H_0 is true,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_1^{-1}(\theta_0))$$

Now take logarithms and expand in a Taylor series around $\widehat{\theta}_n$,

$$\begin{aligned}
\log \Lambda_n &= \sum_{i=1}^{n}[\ell(Y_i;\widehat{\theta}_n) - \ell(Y_i;\theta_0)] = \sum_{i=1}^{n}[\ell(Y_i;\widehat{\theta}_n) - \ell(Y_i;\widehat{\theta}_n)] + \\
&\quad +(\theta_0 - \widehat{\theta}_n)\sum_{i=1}^{n}\ell'(Y_i;\widehat{\theta}_n) - \tfrac{1}{2}(\widehat{\theta}_n - \theta_0)^2\sum_{i=1}^{n}\ell''(Y_i;\theta_n^*) \\
&= -\tfrac{1}{2}n(\widehat{\theta}_n - \theta_0)^2\,\tfrac{1}{n}\sum_{i=1}^{n}\ell''(Y_i;\theta_n^*)
\end{aligned}$$

where θ_n^* lies between $\widehat{\theta}_n$ and θ_0.

If H_0 is true, and since $\widehat{\theta}_n$ is a consistent sequence, θ_n^* is sandwiched so

$$\theta_n^* \xrightarrow{p} \theta_0.$$

Hence under assumptions (A1)-(A6), and when H_0 is true, a first order Taylor expansion about θ_0, the continuous mapping theorem and the LLN give

$$\frac{1}{n}\sum_{i=1}^{n} \ell''(Y_i;\theta_n^*) \xrightarrow{p} -\mathbb{E}_{\theta_0}[\ell''(Y_i;\theta_0)] = \mathcal{I}_1(\theta_0)$$

On the other hand, by the continuous mapping theorem,

$$n(\widehat{\theta}_n - \theta_0)^2 \xrightarrow{d} \frac{V}{\mathcal{I}_1(\theta_0)}$$

Applying Slutsky's theorem now yields the result. □

Theorem (Wilk's theorem, general p, general $s \leq p$)

Let $Y_1, \ldots, Y_n$ be iid random variables with density (frequency) depending on $\boldsymbol{\theta} \in \mathbb{R}^p$ and satisfying conditions (B1)-(B6), with $\mathcal{I}_1(\boldsymbol{\theta}) = \mathcal{J}_1(\boldsymbol{\theta})$. If the MLE sequence $\widehat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}$, then the likelihood ratio statistic Λ_n for $H_0 : \{\theta_j = \theta_{j,0}\}_{j=1}^s$ satisfies $2\log\Lambda_n \xrightarrow{d} V \sim \chi^2_s$ when H_0 is true.

Comments:

- Note that it may potentially be that $s < p$, and this is accommodated by the theorem
- Hypotheses of the form $H_0 : \{g_j(\boldsymbol{\theta}) = a_j\}_{j=1}^s$, for g_j differentiable real functions, can also be handled by Wilks' theorem:
  - Define $(\phi_1, \ldots, \phi_p) = g(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), \ldots, g_p(\boldsymbol{\theta}))$
  - $g_{s+1}, \ldots, g_p$ defined so that $\boldsymbol{\theta} \mapsto g(\boldsymbol{\theta})$ is 1-1
  - Apply theorem with parameter $\boldsymbol{\phi}$

- 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).

- 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 $t_{\mathrm{obs}} = 3$ 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.
- 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, ...

- 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.

Example

The generalized Pareto distribution, with survival function

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

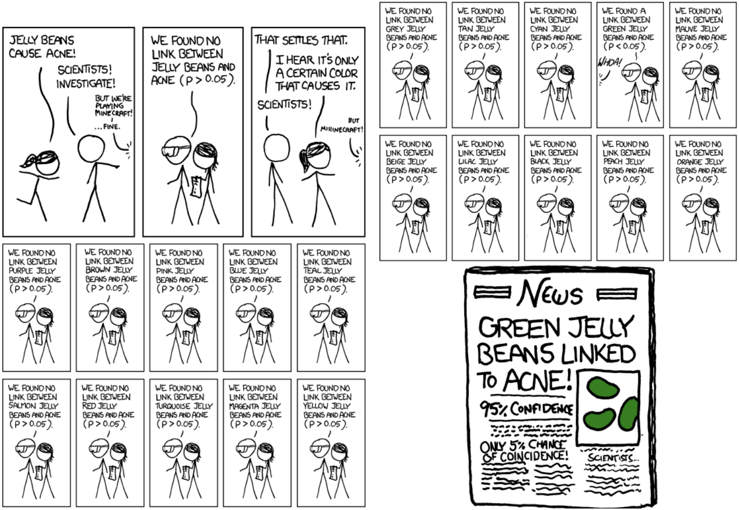
simplifies if $\xi = 0$, and has finite upper support point $x_+ = -\sigma/\xi$ when $\xi < 0$ but $x_+ = \infty$ when $\xi \geq 0$. Here $H_0 : \xi = 0$ is both a simplifying and a dividing hypothesis, of interest (for example) when the distribution is fitted to data on supercentenarians (finite or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

- 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 $H_1, \ldots, H_m$, 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 $\mathrm{E}(FP) = m\alpha = 1$, so this is not a surprise.

https://xkcd.com/882/



- 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 $Z_1, \dots, Z_m \stackrel{\text{iid}}{\sim} \exp(\lambda)$ and the **Rényi representation**

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

applies to their order statistics. Then
  - 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.

- 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

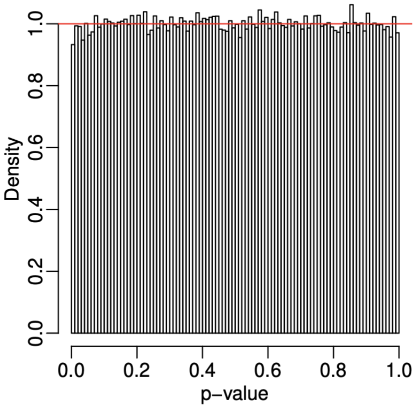
$$H_{0,j} : \text{no association between the expression of the protein and } \mathrm{SNP}_j, \quad j = 1, \ldots, m.$$

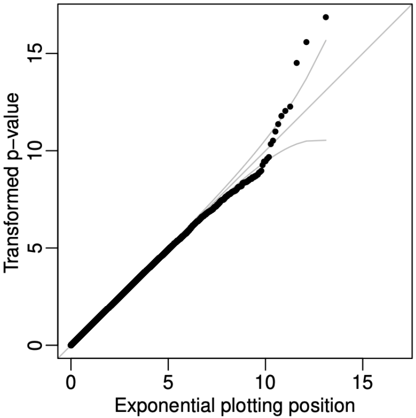
- In a simple model we construct statistics Y_j such that $Y_j \dot{\sim} \mathcal{N}(\theta_j, 1)$, where $\theta_j = 0$ under $H_{0,j}$, and we take $T_j = |Y_j|$, which is likely to be far from zero if $\theta_j \gg 0$ or$\theta_j \ll 0$.
- If $t_{\mathrm{obs},j}$ denotes the observed value of T_j,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

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

  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.

- 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?





- 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,

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

  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 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

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

- 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 FWER$= 0.05$ we need $\alpha \doteq 0.0026$ — the power for individual tests will be tiny (recall ROC curves).

- 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

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

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.

- 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 FP ≥ 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 FP ≥ 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

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

- 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.

- 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)**

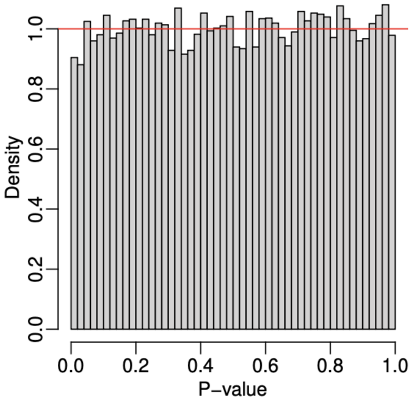
$$\mathrm{FDR} = \mathrm{E}\{I(R > 0)\mathrm{FP}/R\}.$$

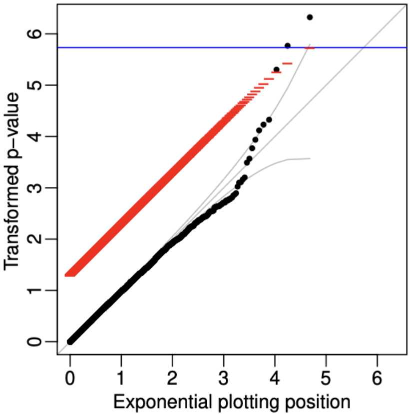
- The **Benjamini–Hochberg procedure** gives strong control for independent tests: specify α, then
  - order the P-values as $P_{(1)} \leq \cdots \leq P_{(m)}$ and the hypotheses as $H_{(1)}, \ldots, H_{(m)}$,
  - reject $H_{(1)}, \ldots, H_{(R)}$, where

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

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, $\ldots$.

- 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.





- 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.

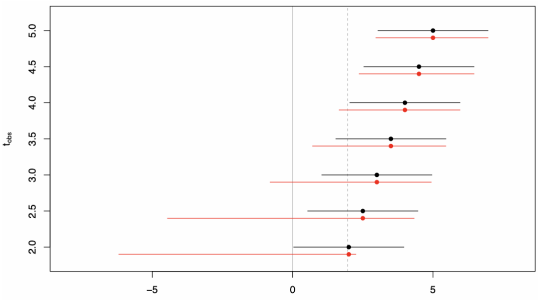
- Contrast
  - **exploratory analysis**, where we study data with no strong prior hypotheses, aiming to find something 'interesting' for future study, and
  - **confirmatory analysis**, where we specify an analysis protocol (hypotheses/tests/...) in advance and stick to it.
- Most statistical procedures assume we are doing the second, but there can be a strong temptation to cheat and treat an exploratory analysis as confirmatory.
- In 'the garden of forking paths' we make a series of choices (which response? transformation? which explanatory variables? ...) but do not then allow for them.
- This leads to non-reproducible results, 'false discoveries', bad science ...
- If we compute a confidence interval $\mathcal{I}$ for θ following a sequence of choices summarised in a selection event $\mathcal{S}$ that is *based on the same data*, and compute

  $$\Pr(\theta \in \mathcal{I}) \quad \text{when we should compute} \quad \Pr(\theta \in \mathcal{I} \mid \mathcal{S}),$$

  we are effectively pretending that $\mathcal{S}$ did not exist.

Example

Suppose $T \sim \mathcal{N}(\theta, 1)$ and we perform a two-sided test of $H_0 : \theta = 0$ at level $\alpha = 5\%$ and then construct a 95% confidence interval $\mathcal{I}_{95}$ around the observed t_{obs} if we reject H_0. Compare the resulting confidence intervals when we do and do not allow for selection. What is the coverage of $\mathcal{I}_{95}$ conditional on $\mathcal{S}$?



95% confidence intervals for θ without (black) and with (red) allowance for selection on event $\mathcal{S} = \{T > z_{0.975}\}$.

- Recall the basis of confidence intervals for θ based on an estimator T satisfying $T \sim \mathcal{N}(\theta, 1)$. We use the fact that $T - \theta \sim \mathcal{N}(0, 1)$ to argue that

$$\Pr(T \le t_{\mathrm{obs}}) = \Pr(T - \theta \le t_{\mathrm{obs}} - \theta) = \Phi(t_{\mathrm{obs}} - \theta)$$

  and then set this equal to α, $1 - \alpha$ to obtain the $(1 - 2\alpha)$ confidence interval $(t_{\mathrm{obs}} - z_{1-\alpha}, t_{\mathrm{obs}} - z_{\alpha})$, which reduces to the 95% confidence interval $\mathcal{I}_{95}$ with limits $t_{\mathrm{obs}} \pm 1.96$ when $\alpha = 0.025$.

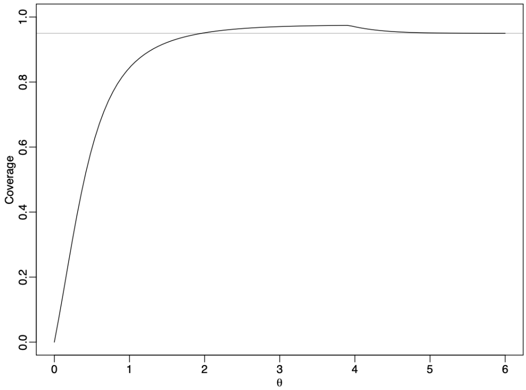
- If we condition on the selection event $\mathcal{S}_R = \{T > z_{1-\beta}\}$ and, compute the 95% confidence interval for θ if this event occurs, we are effectively using the conditional distribution

$$\begin{aligned}\Pr(T \le t_{\mathrm{obs}} \mid T > z_{1-\beta}) &= \Pr(T - \theta \le t_{\mathrm{obs}} - \theta \mid T - \theta > z_{1-\beta} - \theta) \\ &= \frac{\Phi(t_{\mathrm{obs}} - \theta) - \Phi(z_{1-\beta} - \theta)}{1 - \Phi(z_{1-\beta} - \theta)}\end{aligned}$$

  and the $(1 - 2\alpha)$ interval for θ has as endpoints the solutions to

$$\frac{\Phi(t_{\mathrm{obs}} - \theta) - \Phi(z_{1-\beta} - \theta)}{1 - \Phi(z_{1-\beta} - \theta)} = \alpha, 1 - \alpha.$$

- If we set $\beta = \alpha = 0.025$, then we get the limits shown in the graph, which shows that even having $t_{\mathrm{obs}} = 3$ still leads to a 95% CI that contains 0 when we allow for selection. Hence making allowance for selection can radically change inferences, especially when H_0 is only just rejected.



Conditional coverage $\Pr(\theta \in \mathcal{I}_{95} \mid \mathcal{S})$ of $\mathcal{I}_{95}$ as a function of θ.
This graph shows that if we ignore the selection and just use the interval $\mathcal{I}_{95}$ after observing the event $\mathcal{S} = \{|T| > z_{0.975}\}$, then the true coverage varies from 0 when $\theta = 0$ to 0.95 when $\theta \to \infty$, but does not pass its nominal value until $\theta > 2$.

- Lots of work in last decade, in two main categories:
- Methods for specific algorithms (e.g., the lasso) with a selection event $\mathcal{S}$ of a specified form and for which $f(\mathcal{Y} \mid \mathcal{S})$ is tractable;
- More general approaches, including
  - Methods that allow for all possible selection procedures, and hence are hyper-conservative (e.g., so-called universal inference, *e*-values, . . .);
  - Splitting the data into two or more groups (below);
  - Adding noise (less general, since strictly applicable only to certain setttings).
- Garcia Rasines and Young (2023, *Biometrika*) have a good discussion and more references.

- **Sample splitting** is a standard approach to dealing with selection.
- Partition (independent) original data $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$ at random into subsets $\mathcal{Y}_0$ and $\mathcal{Y}_1$, of respective sizes $n_0 = pn$ and $n_1 = (1-p)n$, use $\mathcal{Y}_0$ for selection, and then perform inferences using $\mathcal{Y}_1$.
- As $\mathcal{Y}_1 \perp\!\!\!\perp \mathcal{Y}_0$ and $\mathcal{S}$ depends only on $\mathcal{Y}_0$, we have

  $$f(\mathcal{Y}) = f(\mathcal{Y} \mid \mathcal{S})f(\mathcal{S}) = f(\mathcal{Y}_0, \mathcal{Y}_1 \mid \mathcal{S})f(\mathcal{S}) = f(\mathcal{Y}_1)f(\mathcal{Y}_0 \mid \mathcal{S})f(\mathcal{S}),$$

  so any inference based on $\mathcal{Y}_1$ is unaffected by the selection.
- This approach is simple and widely applicable (at least for random samples), but
  - if the split is random, selections and inferences may be different for different splits;
  - there is a loss of power, both for finding any effects (using $\mathcal{Y}_0$) and for inference for them (using $\mathcal{Y}_1$);
  - if the data are not a random sample (e.g., in a regression setup, (y, x), with x treated as constant), then we should aim for similar information contents in $\mathcal{Y}_0$ and $\mathcal{Y}_1$ (more formally, ancillary statistics should be similar for both parts), and it may be hard to achieve this, particularly in high dimensions.

- Data (Y, X), with X (if present) treated as constant
- Have random variable W, maybe dependent on X, and base selection on $U = u(Y, W)$, e.g., setting selection variable $S = s(U)$ equal to s.
- Then base inference on $Y \mid U$, which is conditionally independent of S.
- If $Y \mapsto (U, V) = (u(Y, W), v(Y, W))$, where (U, V) are jointly sufficient for model and $U \perp\!\!\!\perp V$, then inference from $Y \mid U$ is equivalent to inference from V.

Consider $T \sim \mathcal{N}(\theta, 1)$, and take $U = T + pW$, where $W \sim \mathcal{N}(0, 1)$ independent of T, with p known. Note that if we set $V = T - W/p$, then

$$U \sim \mathcal{N}(\theta, 1 + p^2), \quad V \sim \mathcal{N}(\theta, 1 + 1/p^2), \quad \mathrm{cov}(U, V) = 0,$$

so $U \perp\!\!\!\perp V$, and we can write

$$T = \frac{U + p^2 V}{1 + p^2}.$$

Hence

$$T \mid U = u \overset{D}{=} \frac{u + p^2 V}{1 + p^2},$$

which is equivalent to using the normal distribution of V for inference, as p and u are known.

- if $p \approx 0$, then $U \approx T$ and the selection will be nearly the same as with the original data, but the inference will be poor because $V \not\approx T$;
- if $p \approx 1$, then $V \approx T$ and the inference will be good but $U \not\approx T$ so the selection may be very different from that based on T.
- Implies context-based trade-off between selection and inference.

- Need to be aware of possibility of selection effects and to read the literature critically.
- Must be clear if a study is exploratory or confirmatory:
  - If confirmatory, need to clarify protocol for inference **beforehand**;
  - If exploratory, need to avoid (any?) conclusions that might be due to 'forking paths'.
- At present it looks like randomisation is a good approach in cases with simple sufficient statistics ... and asymptotically when σ^2 can be estimated reasonably well.