Discrete p-values

Patrick Rubin-Delanchy
patrick.rubin-delanchy@bristol.ac.uk

School of Mathematics, University of Bristol

December 4, 2018

Why is the p-value defined as $p = \mathbb{P}_0(T \geq t)$ and not $p = \mathbb{P}_0(T > t)$?

It obviously only matters if T is discrete (or has discrete components).

Discrete test statistics are ubiquitous, and especially so in cyber-security applications (e.g. count data, presence/absence events).

Why is the p-value defined as $p = \mathbb{P}_0(T \geq t)$ and not $p = \mathbb{P}_0(T > t)$?

It obviously only matters if T is discrete (or has discrete components).

Discrete test statistics are ubiquitous, and especially so in cyber-security applications (e.g. count data, presence/absence events).

Why is the p-value defined as $p = \mathbb{P}_0(T \geq t)$ and not $p = \mathbb{P}_0(T > t)$?

It obviously only matters if T is discrete (or has discrete components).

Discrete test statistics are ubiquitous, and especially so in cyber-security applications (e.g. count data, presence/absence events).

Examples

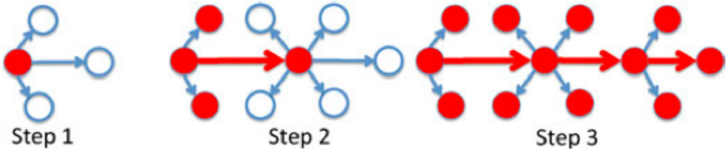
At the beginning of this module we motivated anomaly detection with the following suggestions from the MS-ISAC guide to DDOS attacks (2017):

- "Look for a large number of SYN packets, from multiple sources, over a short duration" (SYN flood)
- "Look for a large number of inbound UDP packets over irregular network ports coming from a large number of source IP addresses" (UDP flood)
- "look for a significant amount of inbound ICMP traffic from a large number of sources" (IMCP flood – e.g. SMURF attack)
- "look for a large number of inbound traffic from a significant number of source IP addresses with a destination port of 80" (HTTP GET flood)

All of these examples would produce discrete-valued anomalies.

Another example

One of the simplest and most direct approaches to intrusion detection is to monitor new connections, e.g. to detect network traversal motifs such as:



(Reproduced from "Scan statistics for the online detection of locally anomalous subgraphs", Joshua Neil *et al.* , 2013).

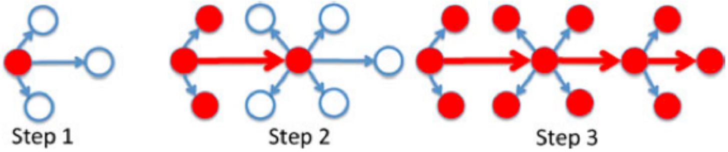
The correct way to formalise this is to consider the test statistic

$$T = \mathbb{I}(\text{edge between } i \text{ and } j),$$

which is binary-valued.

Another example

One of the simplest and most direct approaches to intrusion detection is to monitor new connections, e.g. to detect network traversal motifs such as:



(Reproduced from "Scan statistics for the online detection of locally anomalous subgraphs", Joshua Neil *et al.* , 2013).

The correct way to formalise this is to consider the test statistic

$$T = \mathbb{I}(\text{edge between } i \text{ and } j),$$

which is binary-valued.

Partially discrete p-values occur, for example, with censored data, e.g.

$$T = \min(\text{time to next event, end of observation period/time-out}).$$

The usual stochastic order (Definition 1)

We say that a random variable X is stochastically smaller than (or stochastically dominated by) a random variable Y, denoted $X \leq_{st} Y$, in the usual stochastic order, if their cumulative distribution functions satisfy

$$F_X(a) \geq F_Y(a),$$

for all $a \in \mathbb{R}$.

There are many other stochastic orders, but this is what we mean when we say that X is stochastically smaller than Y.

The usual stochastic order (Definition 1)

We say that a random variable X is stochastically smaller than (or stochastically dominated by) a random variable Y, denoted $X \leq_{st} Y$, in the usual stochastic order, if their cumulative distribution functions satisfy

$$F_X(a) \geq F_Y(a),$$

for all $a \in \mathbb{R}$.

There are many other stochastic orders, but this is what we mean when we say that X is stochastically smaller than Y.

The usual stochastic order (equivalent Definition 2)

We say that a random variable X is stochastically smaller than (or stochastically dominated by) a random variable Y, denoted $X \leq_{st} Y$, in the usual stochastic order, if for any non-decreasing function f we have

$$\mathrm{E}\{f(X)\} \leq \mathrm{E}\{f(Y)\}.$$

It is fairly straightforward to see that Definition 2 implies Definition 1 by considering the family of functions $f_a(x) = \_\_\_$
The reverse implication can also be shown but is out of scope for this course.

The usual stochastic order (equivalent Definition 2)

We say that a random variable X is stochastically smaller than (or stochastically dominated by) a random variable Y, denoted $X \leq_{st} Y$, in the usual stochastic order, if for any non-decreasing function f we have

$$\mathrm{E}\{f(X)\} \leq \mathrm{E}\{f(Y)\}.$$

It is fairly straightforward to see that Definition 2 implies Definition 1 by considering the family of functions $f_a(x) =$ ␣␣␣
The reverse implication can also be shown but is out of scope for this course.

The usual stochastic order (equivalent Definition 2)

We say that a random variable X is stochastically smaller than (or stochastically dominated by) a random variable Y, denoted $X \leq_{st} Y$, in the usual stochastic order, if for any non-decreasing function f we have

$$\mathrm{E}\{f(X)\} \leq \mathrm{E}\{f(Y)\}.$$

It is fairly straightforward to see that Definition 2 implies Definition 1 by considering the family of functions $f_a(x) = -\mathbb{I}(x \leq a)$. The reverse implication can also be shown but is out of scope for this course.

Discrete p-values are super-uniform

Recall the result that (week 1, Fundamentals of hypothesis testing, page 11) under the null hypothesis the p-value is uniform if the test statistic T is continuous.

Theorem (discrete p-values are super-uniform)

Under the null hypothesis, for any distribution of the test statistic T, the p-value is stochastically larger than a uniform random variable on $[0, 1]$.

Discrete p-values are super-uniform

Recall the result that (week 1, Fundamentals of hypothesis testing, page 11) under the null hypothesis the p-value is uniform if the test statistic T is continuous.

Theorem (discrete p-values are super-uniform)

Under the null hypothesis, for any distribution of the test statistic T, the p-value is stochastically larger than a uniform random variable on $[0, 1]$.

Coin-tossing example

Consider testing the hypothesis

$$H_0 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{Bernoulli}(1/2),$$

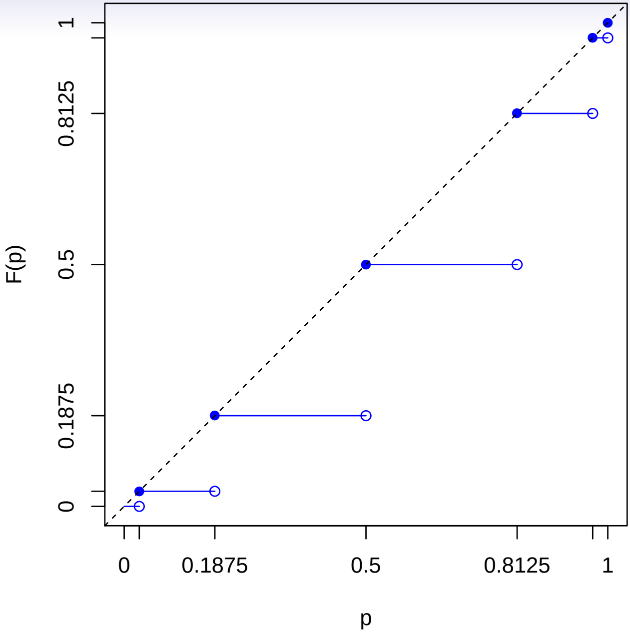
versus,

$$H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{Bernoulli}(\mu) \quad \mu > 0.5,$$

on the basis of the test statistic $T = \sum_{i=1}^{n} X_i$.

Coin-tossing example ($n = 5$)

t	$\mathbb{P}_0(T = t)$	$p = \mathbb{P}_0(T \geq t)$
0	0.03125	1
1	0.15625	0.96875
2	0.3125	0.8125
3	0.3125	0.5
4	0.15625	0.1875
5	0.03125	0.03125


F(p)
1
0.8125
0.5
0.1875
0
0
0.1875
0.5
0.8125
1
p

Proof.

Recall that when viewed as a random variable the p-value is defined as $P = \mathbb{P}_0(T^* \geq T \mid T)$.

Let $S_0(x) = \mathbb{P}_0(T \geq x)$. Then,

$$\begin{aligned}\mathbb{P}_0(P \leq x) &= \mathbb{P}_0\{\mathbb{P}_0(T^* \geq T \mid T) \leq x\} \\ &= \mathbb{P}_0\{S_0(T) \leq x\}.\end{aligned}$$

Proof.

Recall that when viewed as a random variable the p-value is defined as $P = \mathbb{P}_0(T^* \geq T \mid T)$.

Let $S_0(x) = \mathbb{P}_0(T \geq x)$. Then,

$$\begin{aligned}\mathbb{P}_0(P \leq x) &= \mathbb{P}_0\{\mathbb{P}_0(T^* \geq T \mid T) \leq x\} \\ &= \mathbb{P}_0\{S_0(T) \leq x\}.\end{aligned}$$

Let y denote the maximum value of S_0 not exceeding x, and t the supported value at which it occurs. Then,

$$\begin{aligned}\mathbb{P}_0\{S_0(T) \leq x\} &= \mathbb{P}_0\{T \geq t\} \\ &= S_0(t) = y \leq x,\end{aligned}$$

so that $F_U(x) = x \geq \mathbb{P}_0(P \leq x) = F_P(x)$, where F_U, F_P are respectively the distribution functions of a uniform random variable U and the p-value P under H_0. $\square$

Conservative tests

- Remember that in a classical hypothesis testing framework we reject the null hypothesis when the p-value $p \leq \alpha$.
- When the test statistic is continuous this guarantees (week 1, Fundamentals of hypothesis testing, page 12) that:

  the probability of rejecting the null hypothesis if it holds is α.
- More generally, i.e. when the test is discrete or partially discrete, we have shown that

  the probability of rejecting the null hypothesis if it holds does not exceed α.
- A test procedure which rejects the null hypothesis with probability less than α (resp. more than α) is called *conservative* (resp. *liberal*).
- The non-strict inequality in the definition $p = \mathbb{P}_0(T \geq t)$ is therefore well-thought-out, as it guarantees a conservative test procedure.

Conservative tests

- Remember that in a classical hypothesis testing framework we reject the null hypothesis when the p-value $p \leq \alpha$.
- When the test statistic is continuous this guarantees (week 1, Fundamentals of hypothesis testing, page 12) that:

  the probability of rejecting the null hypothesis if it holds is α.
- More generally, i.e. when the test is discrete or partially discrete, we have shown that

  the probability of rejecting the null hypothesis if it holds does not exceed α.
- A test procedure which rejects the null hypothesis with probability less than α (resp. more than α) is called *conservative* (resp. *liberal*).
- The non-strict inequality in the definition $p = \mathbb{P}_0(T \geq t)$ is therefore well-thought-out, as it guarantees a conservative test procedure.

Conservative tests

- Remember that in a classical hypothesis testing framework we reject the null hypothesis when the p-value $p \leq \alpha$.
- When the test statistic is continuous this guarantees (week 1, Fundamentals of hypothesis testing, page 12) that:

  the probability of rejecting the null hypothesis if it holds is α.
- More generally, i.e. when the test is discrete or partially discrete, we have shown that

  the probability of rejecting the null hypothesis if it holds does not exceed α.
- A test procedure which rejects the null hypothesis with probability less than α (resp. more than α) is called *conservative* (resp. *liberal*).
- The non-strict inequality in the definition $p = \mathbb{P}_0(T \geq t)$ is therefore well-thought-out, as it guarantees a conservative test procedure.

Conservative tests

- Remember that in a classical hypothesis testing framework we reject the null hypothesis when the p-value $p \leq \alpha$.
- When the test statistic is continuous this guarantees (week 1, Fundamentals of hypothesis testing, page 12) that:

  the probability of rejecting the null hypothesis if it holds is α.
- More generally, i.e. when the test is discrete or partially discrete, we have shown that

  the probability of rejecting the null hypothesis if it holds does not exceed α.
- A test procedure which rejects the null hypothesis with probability less than α (resp. more than α) is called *conservative* (resp. *liberal*).
- The non-strict inequality in the definition $p = \mathbb{P}_0(T \geq t)$ is therefore well-thought-out, as it guarantees a conservative test procedure.

Conservative tests

- Remember that in a classical hypothesis testing framework we reject the null hypothesis when the p-value $p \leq \alpha$.
- When the test statistic is continuous this guarantees (week 1, Fundamentals of hypothesis testing, page 12) that:

  the probability of rejecting the null hypothesis if it holds is α.

- More generally, i.e. when the test is discrete or partially discrete, we have shown that

  the probability of rejecting the null hypothesis if it holds does not exceed α.

- A test procedure which rejects the null hypothesis with probability less than α (resp. more than α) is called *conservative* (resp. *liberal*).
- The non-strict inequality in the definition $p = \mathbb{P}_0(T \geq t)$ is therefore well-thought-out, as it guarantees a conservative test procedure.

Mid-p-values and randomised p-values

Discrete p-values computed using $\mathbb{P}_0(T \geq t)$ can turn out to be impractically conservative, and there are two proposed solutions:

1. The mid-p-value,

$$q = \frac{1}{2}\mathbb{P}_0(T \geq t) + \frac{1}{2}\mathbb{P}_0(T > t),$$

2. and the randomised p-value,

$$r = X\mathbb{P}_0(T \geq t) + (1 - X)\mathbb{P}_0(T > t),$$

   where X is a randomly generated uniform random variable on $[0, 1]$.

Mid-p-values and randomised p-values

Discrete p-values computed using $\mathbb{P}_0(T \geq t)$ can turn out to be impractically conservative, and there are two proposed solutions:

1. The mid-p-value,

$$q = \frac{1}{2}\mathbb{P}_0(T \geq t) + \frac{1}{2}\mathbb{P}_0(T > t),$$

2. and the randomised p-value,

$$r = X\mathbb{P}_0(T \geq t) + (1 - X)\mathbb{P}_0(T > t),$$

   where X is a randomly generated uniform random variable on $[0, 1]$.

Mid-p-values and randomised p-values

Discrete p-values computed using $\mathbb{P}_0(T \geq t)$ can turn out to be impractically conservative, and there are two proposed solutions:

1. The mid-p-value,

$$q = \frac{1}{2}\mathbb{P}_0(T \geq t) + \frac{1}{2}\mathbb{P}_0(T > t),$$

2. and the randomised p-value,

$$r = X\mathbb{P}_0(T \geq t) + (1 - X)\mathbb{P}_0(T > t),$$

   where X is a randomly generated uniform random variable on $[0, 1]$.

Coin-tossing example ($n = 5$)

t	$\mathbb{P}_0(T = t)$	$p = \mathbb{P}_0(T \geq t)$	$\mathbb{P}_0(T > t)$	q	r
0	0.03125	1	0.96875	0.984375	0.9788779
1	0.15625	0.96875	0.8125	0.890625	0.8154765
2	0.3125	0.8125	0.5	0.65625	0.7774432
3	0.3125	0.5	0.1875	0.34375	0.2327139
4	0.15625	0.1875	0.03125	0.109375	0.0879319
5	0.03125	0.03125	0	0.015625	0.0085281

Consider a test statistic $T \in \{0, 1, \ldots\}$ and let $v_i = \mathbb{P}_0(T \geq i)$, so that $v_0 = 1 \geq v_1 \geq v_2 \geq \cdots > 0$.

If we observe $t = i$:

1. the p-value is $p = v_i$
2. the mid-p-value is $q = \frac{1}{2}v_i + \frac{1}{2}v_{i+1}$
3. the randomised p-value is $r = Xv_i + (1 - X)v_{i+1}$, where $X \sim \text{uniform}[0, 1]$

Since $X \in [0,1]$ and $v_{i+1} < v_i$ we have:

$$v_{i+1} \leq X v_i + (1-X) v_{i+1} \leq v_i.$$

Furthermore, for $x \in [v_{i+1}, v_i]$,

$$\mathbb{P}_0(r \leq x) = \qquad .$$

Since $X \in [0,1]$ and $v_{i+1} < v_i$ we have:

$$v_{i+1} \leq X v_i + (1-X) v_{i+1} \leq v_i.$$

Furthermore, for $x \in [v_{i+1}, v_i]$,

$$\mathbb{P}_0(r \leq x) = \mathbb{P}_0 \left(X \leq \frac{x - v_{i+1}}{v_i - v_{i+1}} \right) = \frac{x - v_{i+1}}{v_i - v_{i+1}}.$$

Since $X \in [0, 1]$ and $v_{i+1} < v_i$ we have:

$$v_{i+1} \leq X v_i + (1 - X) v_{i+1} \leq v_i.$$

Furthermore, for $x \in [v_{i+1}, v_i]$,

$$\mathbb{P}_0(r \leq x) = \mathbb{P}_0\left(X \leq \frac{x - v_{i+1}}{v_i - v_{i+1}}\right) = \frac{x - v_{i+1}}{v_i - v_{i+1}}.$$

Differentiating gives a constant and therefore

$$r \sim \text{uniform}[v_{i+1}, v_i].$$

A quick digression

While the notation's still fresh in your minds, we'll prove this simple result.

Remember that in the coin-tossing example we observed that the distribution function of P under H_0 satisfied $F(v_i) = v_i$. In fact, more generally,

$$F(v_i) = \mathbb{P}_0(P \leq v_i) = \mathbb{P}(T \geq i) = v_i.$$

Define, as with the ordinary p-value, the random counterpart to the randomised p-value:

$$R = X\mathbb{P}_0(T^* \geq T \mid T) + (1 - X)\mathbb{P}_0(T^* > T \mid T).$$

Theorem (the randomised p-value is uniform)

Under the null hypothesis, for any distribution of the test statistic T, the randomised p-value is uniformly distributed on $[0, 1]$.

Proof

For simplicity we will prove the theorem assuming a discrete test statistic $T \in \{0, 1, \ldots\}$.

For $x \in (0, 1]$, let $a = \max(y \in \{0\} \cup \{v_i\} : y < x)$ and $b = \min(y \in \{v_i\} : y \geq x)$.

Then,

$$\begin{aligned}
\mathbb{P}_0(R \leq x) = \mathbb{P}_0(R \leq x \mid R > b)\mathbb{P}_0(R > b) \\
+ \mathbb{P}_0\{R \leq x \mid R \in (a, b]\}\mathbb{P}\{R \in (a, b]\}, \\
+ \mathbb{P}_0(R \leq x \mid R \leq a)\mathbb{P}_0(R \leq a).
\end{aligned}$$

We have:

Proof

For simplicity we will prove the theorem assuming a discrete test statistic $T \in \{0, 1, \ldots\}$.

For $x \in (0, 1]$, let $a = \max(y \in \{0\} \cup \{v_i\} : y < x)$ and $b = \min(y \in \{v_i\} : y \geq x)$.

Then,

$$\begin{aligned}
\mathbb{P}_0(R \leq x) = &\ \mathbb{P}_0(R \leq x \mid R > b)\mathbb{P}_0(R > b) \\
&+ \mathbb{P}_0\{R \leq x \mid R \in (a, b]\}\mathbb{P}\{R \in (a, b]\}, \\
&+ \mathbb{P}_0(R \leq x \mid R \leq a)\mathbb{P}_0(R \leq a).
\end{aligned}$$

We have:

1. $\mathbb{P}_0(R \leq x \mid R > b)\mathbb{P}_0(R > b) = 0$.

Proof

For simplicity we will prove the theorem assuming a discrete test statistic $T \in \{0, 1, \ldots\}$.

For $x \in (0, 1]$, let $a = \max(y \in \{0\} \cup \{v_i\} : y < x)$ and $b = \min(y \in \{v_i\} : y \geq x)$.

Then,

$$\begin{aligned}\mathbb{P}_0(R \leq x) = \mathbb{P}_0(R \leq x \mid R > b)\mathbb{P}_0(R > b) \\ + \mathbb{P}_0\{R \leq x \mid R \in (a, b]\}\mathbb{P}\{R \in (a, b]\}, \\ + \mathbb{P}_0(R \leq x \mid R \leq a)\mathbb{P}_0(R \leq a).\end{aligned}$$

We have:

1. $\mathbb{P}_0(R \leq x \mid R > b)\mathbb{P}_0(R > b) = 0$.
2. $\mathbb{P}\{R \in (a, b]\} = \mathbb{P}(P = b) = F(b) - F(a)$.

Proof

For simplicity we will prove the theorem assuming a discrete test statistic $T \in \{0, 1, \ldots\}$.

For $x \in (0, 1]$, let $a = \max(y \in \{0\} \cup \{v_i\} : y < x)$ and $b = \min(y \in \{v_i\} : y \geq x)$.

Then,

$$\begin{aligned}\mathbb{P}_0(R \leq x) = \mathbb{P}_0(R \leq x \mid R > b)\mathbb{P}_0(R > b) \\ + \mathbb{P}_0\{R \leq x \mid R \in (a, b]\}\mathbb{P}\{R \in (a, b]\}, \\ + \mathbb{P}_0(R \leq x \mid R \leq a)\mathbb{P}_0(R \leq a).\end{aligned}$$

We have:

1. $\mathbb{P}_0(R \leq x \mid R > b)\mathbb{P}_0(R > b) = 0.$
2. $\mathbb{P}\{R \in (a, b]\} = \mathbb{P}(P = b) = F(b) - F(a).$
3. $\mathbb{P}_0(R \leq x \mid R \leq a)\mathbb{P}_0(R \leq a) = 1 \times F(a).$

Proof continued

Finally,

$$\mathbb{P}_0\{R \leq x \mid R \in (a, b]\} = \mathbb{P}_0(Xb + (1 - X)a \leq x)$$

Proof continued

Finally,

$$\begin{aligned}\mathbb{P}_0\{R \le x \mid R \in (a, b]\} &= \mathbb{P}_0(Xb + (1 - X)a \le x) \\ &= \mathbb{P}_0\left(X \le \frac{x - a}{b - a}\right)\end{aligned}$$

Proof continued

Finally,

$$\begin{aligned}
\mathbb{P}_0\{R \le x \mid R \in (a, b]\} &= \mathbb{P}_0(Xb + (1 - X)a \le x) \\
&= \mathbb{P}_0\left(X \le \frac{x-a}{b-a}\right) \\
&= \frac{x-a}{b-a} = \frac{x - F(a)}{F(b) - F(a)}.
\end{aligned}$$

Proof continued

Finally,

$$\begin{aligned}\mathbb{P}_0\{R \leq x \mid R \in (a, b]\} &= \mathbb{P}_0(Xb + (1 - X)a \leq x) \\ &= \mathbb{P}_0\left(X \leq \frac{x - a}{b - a}\right) \\ &= \frac{x - a}{b - a} = \frac{x - F(a)}{F(b) - F(a)}.\end{aligned}$$

Altogether:

$$\mathbb{P}_0(R \leq x) = 0 + \frac{x - F(a)}{F(b) - F(a)}\{F(b) - F(a)\} + F(a) = x.$$

A comparison

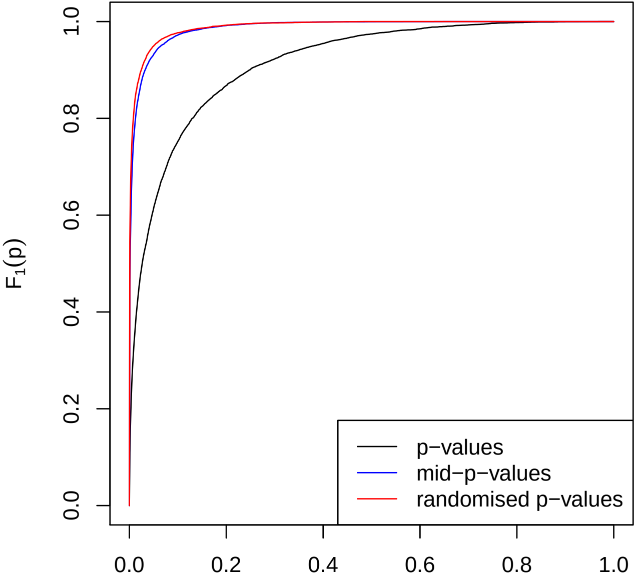
We will generate m "coin-toss" experiments with different, random numbers of tosses.

Each experiment generates m p-values, mid-p-values, and randomised p-values all computed based on a Bernoulli(0.5) null hypothesis. Each of the three sets is combined using Fisher's method.

It will be in your homework to prove that Fisher's method is conservative when ordinary discrete p-values are input. Fisher's method is only approximate with mid-p-values, but is exact with randomised p-values.

In the next example (see code), the coins are actually from a Bernoulli(0.75) distribution and we draw the power curves, $\hat{\mathbb{P}}_1$(combined p-value $\leq x$), for each choice of input.

Comparison



$F_1(p)$
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
p−values
mid−p−values
randomised p−values

A worrying result

"Most people will find repugnant the idea of adding yet another random element to a result which is already subject to the errors of random sampling" (Stevens, 1950).

Yet, randomised p-values often perform best in practice and in theory.

A final result about discrete p-values

Consider testing the hypothesis

$$\begin{aligned} H_0 : X &\sim \text{categorical}\{\mu_1, \ldots, \mu_m\}, \\ &\text{versus,} \\ H_1 : X &\sim \text{categorical}\left\{\frac{1}{m}, \ldots, \frac{1}{m}\right\}, \end{aligned}$$

often known as the "flat alternative".

LRT

The LRT is (replacing probability density functions with probability mass functions)

$$\frac{1/m}{\mu_X},$$

which is equivalent to $1/\mu_X$.

Therefore, upon observing x, we compute the p-value

$$p = \mathbb{P}_0(T \geq 1/\mu_x),$$

LRT

The LRT is (replacing probability density functions with probability mass functions)

$$\frac{1/m}{\mu_X},$$

which is equivalent to $1/\mu_X$.

Therefore, upon observing x, we compute the p-value

$$\begin{aligned} p &= \mathbb{P}_0(T \geq 1/\mu_x), \\ &= \mathbb{P}_0(1/\mu_X \geq 1/\mu_x), \end{aligned}$$

LRT

The LRT is (replacing probability density functions with probability mass functions)

$$\frac{1/m}{\mu_X},$$

which is equivalent to $1/\mu_X$.

Therefore, upon observing x, we compute the p-value

$$\begin{aligned} p &= \mathbb{P}_0(T \geq 1/\mu_x), \\ &= \mathbb{P}_0(1/\mu_X \geq 1/\mu_x), \\ &= \mathbb{P}_0(\mu_X \leq \mu_x), \end{aligned}$$

LRT

The LRT is (replacing probability density functions with probability mass functions)

$$\frac{1/m}{\mu_X},$$

which is equivalent to $1/\mu_X$.

Therefore, upon observing x, we compute the p-value

$$\begin{aligned} p &= \mathbb{P}_0(T \geq 1/\mu_x), \\ &= \mathbb{P}_0(1/\mu_X \geq 1/\mu_x), \\ &= \mathbb{P}_0(\mu_X \leq \mu_x), \\ &= \sum_{i=1}^{m} \mathbb{I}(\mu_i \leq \mu_x), \end{aligned}$$

or "the probability of observing an outcome as unlikely as x."