Fundamentals of hypothesis testing

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

School of Mathematics, University of Bristol

November 15, 2018

The hypothesis testing framework

In a standard hypothesis testing framework, the data are posited to follow one of two families of distributions, each referred to as a hypothesis.

The null hypothesis, denoted H_0, generally represents a null, default position, e.g. "the drug has no effect".

The alternative hypothesis, denoted H_1, generally represents a rival explanation for the data that would be of interest if true, e.g. "the drug works!"

A running example

The simplest possible example is that we have data $D = (X_1, \ldots, X_n)$ and we are testing:

$$H_0 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(0, 1),$$

versus,

$$H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(0.1, 1).$$

Why is this simple?

- We have fully specified H_0. In practice we would rarely know the null distribution was normal, perhaps only that the mean was 0.
- We have fully specified H_1. In practice we would rarely know the alternative distribution was normal, and in fact rarely know that the mean of interest was 0.1 (rather than simply larger than 0, for example).
- We have picked a simple and continuous distribution, making all computations straightforward. For example, even testing whether a coin is fair is more complicated.

A running example

The simplest possible example is that we have data $D = (X_1, \ldots, X_n)$ and we are testing:

$$H_0 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(0, 1),$$

versus,

$$H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(0.1, 1).$$

Why is this simple?

- We have fully specified H_0. In practice we would rarely know the null distribution was normal, perhaps only that the mean was 0.
- We have fully specified H_1. In practice we would rarely know the alternative distribution was normal, and in fact rarely know that the mean of interest was 0.1 (rather than simply larger than 0, for example).
- We have picked a simple and continuous distribution, making all computations straightforward. For example, even testing whether a coin is fair is more complicated.

A running example

The simplest possible example is that we have data $D = (X_1, \ldots, X_n)$ and we are testing:

$$H_0 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(0, 1),$$

versus,

$$H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(0.1, 1).$$

Why is this simple?

- We have fully specified H_0. In practice we would rarely know the null distribution was normal, perhaps only that the mean was 0.
- We have fully specified H_1. In practice we would rarely know the alternative distribution was normal, and in fact rarely know that the mean of interest was 0.1 (rather than simply larger than 0, for example).
- We have picked a simple and continuous distribution, making all computations straightforward. For example, even testing whether a coin is fair is more complicated.

A running example

The simplest possible example is that we have data $D = (X_1, \ldots, X_n)$ and we are testing:

$$H_0 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(0, 1),$$

versus,

$$H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(0.1, 1).$$

Why is this simple?

- We have fully specified H_0. In practice we would rarely know the null distribution was normal, perhaps only that the mean was 0.
- We have fully specified H_1. In practice we would rarely know the alternative distribution was normal, and in fact rarely know that the mean of interest was 0.1 (rather than simply larger than 0, for example).
- We have picked a simple and continuous distribution, making all computations straightforward. For example, even testing whether a coin is fair is more complicated.

Test statistic

A test statistic T is a real-valued function of the data $T = g(D)$, which is used to decide which hypothesis is true.

We will assume:

1. The larger T, the more evidence there is in favour of H_1.
2. T has a fully-specified distribution under H_0, which we denote $\mathbb{P}_0$.

Note that the second assumption is often much weaker than assuming a fully specified H_0.

Test statistic

In our running example, we might use the test statistic

$$T = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Under H_0,

$T \sim$ ⎵⎵⎵⎵⎵

Test statistic

In our running example, we might use the test statistic

$$T = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Under H_0,

$$T \sim \text{normal} \left\{ 0, \frac{1}{n} \right\}.$$

The p-value

The p-value, p, associated with the observed value of a test statistic, t, is the probability under the null of observing a test statistic *as extreme as* t, that is,

$$p = \mathbb{P}_0(T \geq t) = \mathbb{P}(T \geq t \mid H_0 \text{ holds}).$$

The p-value

In our running example, remember that:

$$T \sim \text{normal}\left\{0, \frac{1}{n}\right\}$$

The p-value associated with t is

$$p = \mathbb{P}_0(T \geq t) = \text{␣␣␣␣␣}$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-y^2/2}\, \mathrm{d}y. \qquad (= \mathbb{P}\{\text{normal}(0,1) \leq x\})$$

T ~ N(0,1/n)
T'~N(0,1)
T'=(n)^0.5*T

The p-value

In our running example, remember that:

$$T \sim \text{normal}\left\{0, \frac{1}{n}\right\}$$

The p-value associated with t is

$$p = \mathbb{P}_0(T \geq t) = 1 - \Phi(\sqrt{n}t),$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-y^2/2}\, \mathrm{d}y. \qquad (= \mathbb{P}\{\text{normal}(0,1) \leq x\})$$

Distribution of the p-value under the null hypothesis

So far we have treated the p-value as an observation p. Now we are going to think about the p-value before any data are observed. In this case the data are random rather than observed, and so the p-value is random too. Denote this random variable by P. We now need "two test statistics":

1. T: the random test statistic that would be observed if we observed D (our old t).
2. T^*: a hypothetical replicate of T under H_0 (our old T).

The p-value is now defined to be the *random variable*:

$$P = \mathbb{P}_0(T^* \geq T \mid T).$$

Distribution of the p-value under the null hypothesis

Assume H_0 holds. Then our example provides a generative model for P:

1. Generate data $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{normal}(0, 1)$
2. Compute $T = \frac{1}{n} \sum_{i=1}^{n} X_i$
3. Compute $P = \mathbb{P}_0(T^* \geq T \mid T) = 1 - \Phi(\sqrt{n}T)$

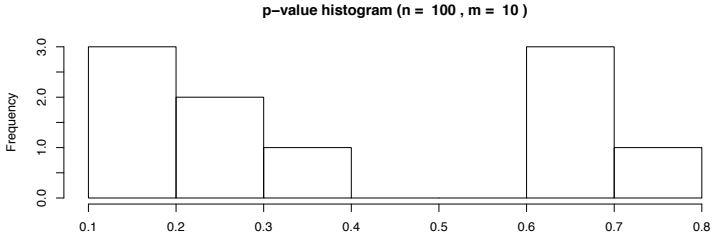
Distribution of the p-value under the null hypothesis

Assume H_0 holds. Then our example provides a generative model for P:

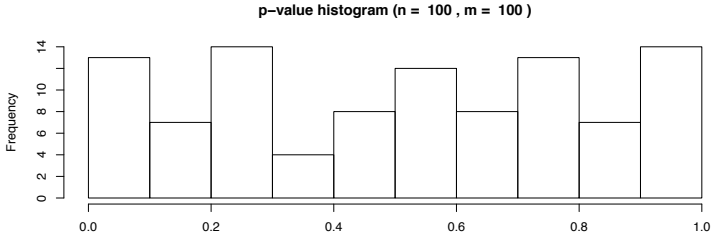
1. Generate data $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{normal}(0, 1)$
2. Compute $T = \frac{1}{n} \sum_{i=1}^{n} X_i$
3. Compute $P = \mathbb{P}_0(T^* \geq T \mid T) = 1 - \Phi(\sqrt{n}T)$

We'll now run this generative model m times to obtain a sample of p-values.

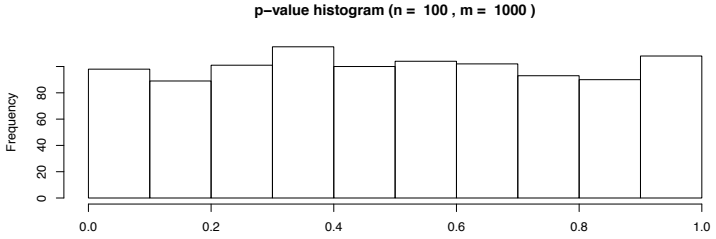
Distribution of the p-value under the null hypothesis



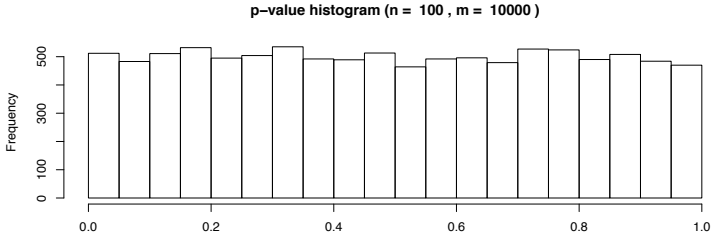
Distribution of the p-value under the null hypothesis



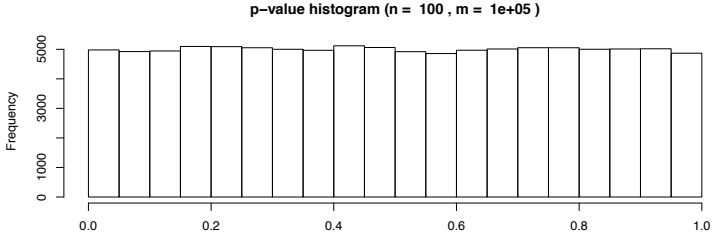
Distribution of the p-value under the null hypothesis



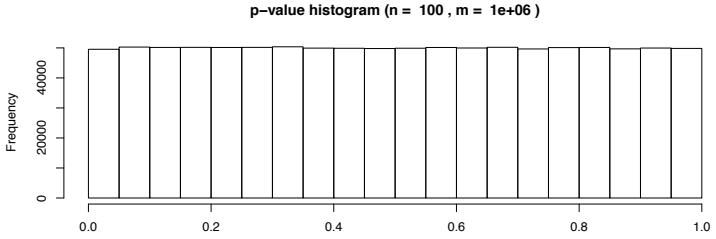
Distribution of the p-value under the null hypothesis



Distribution of the p-value under the null hypothesis



Distribution of the p-value under the null hypothesis



Distribution of the p-value under the null hypothesis

Theorem

Assume the test statistic T has a continuous distribution. Then, under the null hypothesis, the p-value is uniformly distributed on $[0, 1]$.

Proof.

Let $S_0(x) = \mathbb{P}_0(T \geq x)$, which is invertible by continuity of T. Then,

$$\mathbb{P}_0(P \leq x)$$

□

Distribution of the p-value under the null hypothesis

Theorem

Assume the test statistic T has a continuous distribution. Then, under the null hypothesis, the p-value is uniformly distributed on $[0, 1]$.

Proof.

Let $S_0(x) = \mathbb{P}_0(T \geq x)$, which is invertible by continuity of T. Then,

$$\mathbb{P}_0(P \leq x) = \mathbb{P}_0\{\mathbb{P}_0(T^* \geq T \mid T) \leq x\}$$

□

Distribution of the p-value under the null hypothesis

Theorem

Assume the test statistic T has a continuous distribution. Then, under the null hypothesis, the p-value is uniformly distributed on $[0, 1]$.

Proof.

Let $S_0(x) = \mathbb{P}_0(T \geq x)$, which is invertible by continuity of T. Then,

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

□

Distribution of the p-value under the null hypothesis

Theorem

Assume the test statistic T has a continuous distribution. Then, under the null hypothesis, the p-value is uniformly distributed on $[0, 1]$.

Proof.

Let $S_0(x) = \mathbb{P}_0(T \geq x)$, which is invertible by continuity of T. Then,

$$\begin{aligned}\mathbb{P}_0(P \leq x) &= \mathbb{P}_0\{\mathbb{P}_0(T^* \geq T \mid T) \leq x\} \\ &= \mathbb{P}_0\{S_0(T) \leq x\} \\ &= \mathbb{P}_0\{T \geq S_0^{-1}(x)\}\end{aligned}$$

□

Distribution of the p-value under the null hypothesis

Theorem

Assume the test statistic T has a continuous distribution. Then, under the null hypothesis, the p-value is uniformly distributed on $[0, 1]$.

Proof.

Let $S_0(x) = \mathbb{P}_0(T \geq x)$, which is invertible by continuity of T. Then,

$$\begin{aligned}
\mathbb{P}_0(P \leq x) &= \mathbb{P}_0\{\mathbb{P}_0(T^* \geq T \mid T) \leq x\} \\
&= \mathbb{P}_0\{S_0(T) \leq x\} \\
&= \mathbb{P}_0\{T \geq S_0^{-1}(x)\} \\
&= S_0\{S_0^{-1}(x)\}
\end{aligned}$$

□

Distribution of the p-value under the null hypothesis

Theorem

Assume the test statistic T has a continuous distribution. Then, under the null hypothesis, the p-value is uniformly distributed on $[0, 1]$.

Proof.

Let $S_0(x) = \mathbb{P}_0(T \geq x)$, which is invertible by continuity of T. Then,

$$\begin{aligned}
\mathbb{P}_0(P \leq x) &= \mathbb{P}_0\{\mathbb{P}_0(T^* \geq T \mid T) \leq x\} \\
&= \mathbb{P}_0\{S_0(T) \leq x\} \\
&= \mathbb{P}_0\{T \geq S_0^{-1}(x)\} \\
&= S_0\{S_0^{-1}(x)\} \\
&= x
\end{aligned}$$

□

Rejecting the null hypothesis

In classical hypothesis testing, the null hypothesis is rejected when t exceeds a certain threshold (still assuming for simplicity that large test statistics are of interest).

This threshold, τ, is typically chosen according to a *false positive rate* (or significance level) α (e.g. $\alpha = 0.05$) so that

$$\mathbb{P}_0(T \geq \tau) = \alpha.$$

Equivalently, we reject the null hypothesis if $p \leq \alpha$ (i.e., $t \geq \tau \Leftrightarrow p \leq \alpha$).

By the uniformity of P, a rejection event $P \leq \alpha$ occurs with probability α under the null hypothesis, hence:

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

Rejecting the null hypothesis

In classical hypothesis testing, the null hypothesis is rejected when t exceeds a certain threshold (still assuming for simplicity that large test statistics are of interest).

This threshold, τ, is typically chosen according to a *false positive rate* (or significance level) α (e.g. $\alpha = 0.05$) so that

$$\mathbb{P}_0(T \geq \tau) = \alpha.$$

Equivalently, we reject the null hypothesis if $p \leq \alpha$ (i.e., $t \geq \tau \Leftrightarrow p \leq \alpha$).

By the uniformity of P, a rejection event $P \leq \alpha$ occurs with probability α under the null hypothesis, hence:

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

Distribution of P under the alternative

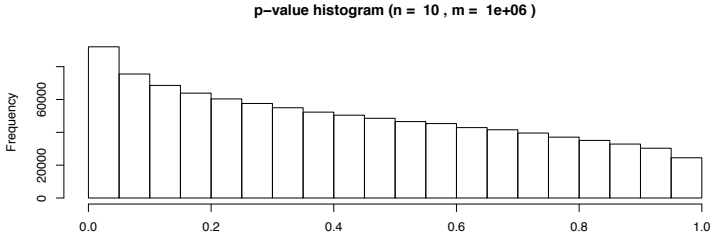
What makes a good test? We can answer that by considering the distribution of P under the alternative hypothesis.

To illustrate this we will:

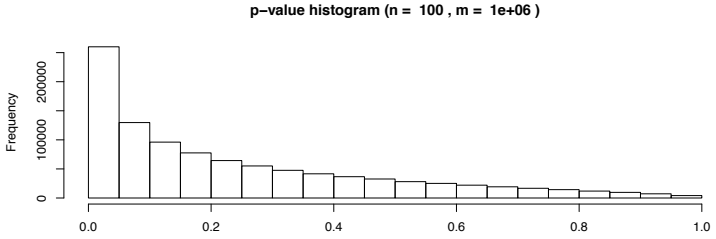
1. Generate data $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{normal}(0.1, 1)$
2. Compute $T = \frac{1}{n} \sum_{i=1}^{n} X_i$
3. Compute $P = \mathbb{P}_0(T^* \geq T \mid T) = 1 - \Phi(\sqrt{n}T)$

and repeat m times.

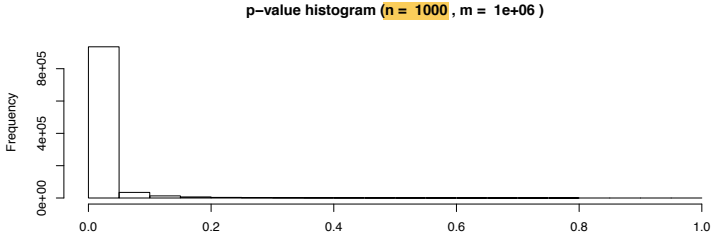
Distribution of P under the alternative



Distribution of P under the alternative



Distribution of P under the alternative



Quick recap

Remember that for a specified false positive rate α we reject the null hypothesis if we observe $p \leq \alpha$, which is equivalent to $t \geq \tau$ if τ satisfies:

$$\mathbb{P}_0(T \geq \tau) = \alpha.$$

In our running example, with $\alpha = 0.05$ and $n = 10$,

$$0.05 = 1 - \Phi(\sqrt{10}\tau) \Leftrightarrow \tau = \frac{\Phi^{-1}(0.95)}{\sqrt{10}} \approx 0.52.$$

We would therefore reject the null hypothesis that the data were independent standard normals if we observed $p = 1 - \Phi(\sqrt{10}\frac{1}{10}\sum_{i=1}^{10} x_i) \leq 0.05$ or equivalently $t = \frac{1}{10}\sum_{i=1}^{10} x_i \geq 0.52$.

This guarantees that there is only a 5% probability of rejecting H_0 if H_0 holds since:

$$\mathbb{P}_0(P \leq 0.05) = \mathbb{P}_0(T \geq \tau) = 0.05.$$

Power

Informally, the power of a hypothesis test is the probability of rejecting the null hypothesis if the alternative holds:

$$\mathbb{P}(\text{Reject } H_0 \mid H_1 \text{ holds}) = \mathbb{P}_1(\text{Reject } H_0).$$

To make this quantity calculable we need:

1. to have a fully specified alternative hypothesis, or at least calculable $\mathbb{P}_1(T \geq \tau)$, for example:
   - with $H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(0.1, 1)$ the calculation is possible, whereas,
   - with $H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(\mu_1, 1), \mu_1 > 0$, the calculation is not.
2. to have a specified false positive rate α.

With these ingredients in place, we define the power of a test to be:

$$\beta = \mathbb{P}_1(P \leq \alpha) = \mathbb{P}_1(T \geq \tau).$$

Power

Informally, the power of a hypothesis test is the probability of rejecting the null hypothesis if the alternative holds:

$$\mathbb{P}(\text{Reject } H_0 \mid H_1 \text{ holds}) = \mathbb{P}_1(\text{Reject } H_0).$$

To make this quantity calculable we need:

1. to have a fully specified alternative hypothesis, or at least calculable $\mathbb{P}_1(T \geq \tau)$, for example:
   - with $H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(0.1, 1)$ the calculation is possible, whereas,
   - with $H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(\mu_1, 1), \mu_1 > 0$, the calculation is not.
2. to have a specified false positive rate α.

With these ingredients in place, we define the power of a test to be:

$$\beta = \mathbb{P}_1(P \leq \alpha) = \mathbb{P}_1(T \geq \tau).$$

Power

Informally, the power of a hypothesis test is the probability of rejecting the null hypothesis if the alternative holds:

$$\mathbb{P}(\text{Reject } H_0 \mid H_1 \text{ holds}) = \mathbb{P}_1(\text{Reject } H_0).$$

To make this quantity calculable we need:

1. to have a fully specified alternative hypothesis, or at least calculable $\mathbb{P}_1(T \geq \tau)$, for example:
   - with $H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(0.1, 1)$ the calculation is possible, whereas,
   - with $H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(\mu_1, 1), \mu_1 > 0$, the calculation is not.
2. to have a specified false positive rate α.

With these ingredients in place, we define the power of a test to be:

$$\beta = \mathbb{P}_1(P \leq \alpha) = \mathbb{P}_1(T \geq \tau).$$

Power

Informally, the power of a hypothesis test is the probability of rejecting the null hypothesis if the alternative holds:

$$\mathbb{P}(\text{Reject } H_0 \mid H_1 \text{ holds}) = \mathbb{P}_1(\text{Reject } H_0).$$

To make this quantity calculable we need:

1. to have a fully specified alternative hypothesis, or at least calculable $\mathbb{P}_1(T \geq \tau)$, for example:
   - with $H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(0.1, 1)$ the calculation is possible, whereas,
   - with $H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(\mu_1, 1), \mu_1 > 0$, the calculation is not.
2. to have a specified false positive rate α.

With these ingredients in place, we define the power of a test to be:

$$\beta = \mathbb{P}_1(P \leq \alpha) = \mathbb{P}_1(T \geq \tau).$$

Power

Informally, the power of a hypothesis test is the probability of rejecting the null hypothesis if the alternative holds:

$$\mathbb{P}(\text{Reject } H_0 \mid H_1 \text{ holds}) = \mathbb{P}_1(\text{Reject } H_0).$$

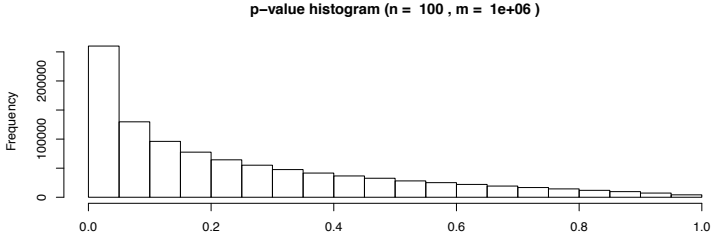
To make this quantity calculable we need:

1. to have a fully specified alternative hypothesis, or at least calculable $\mathbb{P}_1(T \geq \tau)$, for example:
   - with $H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(0.1, 1)$ the calculation is possible, whereas,
   - with $H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(\mu_1, 1), \mu_1 > 0$, the calculation is not.
2. to have a specified false positive rate α.

With these ingredients in place, we define the power of a test to be:

$$\beta = \mathbb{P}_1(P \leq \alpha) = \mathbb{P}_1(T \geq \tau).$$

Running example

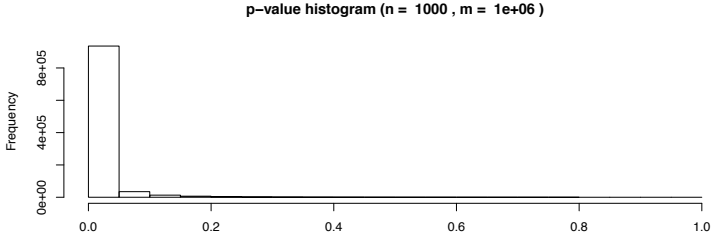


Here $\beta = \mathbb{P}_1(P \leq 0.05) \approx 0.26$.

In words: "if we're prepared to accept a false positive rate 5%, then if the alternative holds, we will detect it with probability 0.26."

For large n

Running example



Here $\beta = \mathbb{P}_1(P \leq 0.05) \approx 0.94$.

In words: "if we're prepared to accept a false positive rate 5%, then if the alternative holds, we will detect it with probability 0.94."

Increasing functions of a test statistic

Lemma

Consider test statistics T_1, T_2 satisfying $T_2 = h(T_1)$ where h is a strictly increasing function. Let t_1, t_2 be the observed values of T_1, T_2. Then the p-values p_1, p_2 associated with t_1, t_2 are equal.

Test statistics that are equal up to an increasing transformation are therefore equivalent. In our running example, there would be no difference in the computed p-value (and therefore the power under the alternative) if we had instead used

$$T = \exp\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\}.$$

Proof.

$$p_1 = \mathbb{P}_0(T_1 \geq t_1) =$$

Increasing functions of a test statistic

Lemma

Consider test statistics T_1, T_2 satisfying $T_2 = h(T_1)$ where h is a strictly increasing function. Let t_1, t_2 be the observed values of T_1, T_2. Then the p-values p_1, p_2 associated with t_1, t_2 are equal.

Test statistics that are equal up to an increasing transformation are therefore equivalent. In our running example, there would be no difference in the computed p-value (and therefore the power under the alternative) if we had instead used

$$T = \exp\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\}.$$

Proof.

$$p_1 = \mathbb{P}_0(T_1 \geq t_1) =$$

Increasing functions of a test statistic

Lemma

Consider test statistics T_1, T_2 satisfying $T_2 = h(T_1)$ where h is a strictly increasing function. Let t_1, t_2 be the observed values of T_1, T_2. Then the p-values p_1, p_2 associated with t_1, t_2 are equal.

Test statistics that are equal up to an increasing transformation are therefore equivalent. In our running example, there would be no difference in the computed p-value (and therefore the power under the alternative) if we had instead used

$$T = \exp\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\}.$$

Proof.

$$p_1 = \mathbb{P}_0(T_1 \geq t_1) =$$

Increasing functions of a test statistic

Lemma

Consider test statistics T_1, T_2 satisfying $T_2 = h(T_1)$ where h is a strictly increasing function. Let t_1, t_2 be the observed values of T_1, T_2. Then the p-values p_1, p_2 associated with t_1, t_2 are equal.

Test statistics that are equal up to an increasing transformation are therefore equivalent. In our running example, there would be no difference in the computed p-value (and therefore the power under the alternative) if we had instead used

$$T = \exp\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\}.$$

T=1/m/X

T'=1/X'

Proof.

$$p_1 = \mathbb{P}_0(T_1 \geq t_1) = \mathbb{P}_0\{h(T_1) \geq h(t_1)\} = \mathbb{P}_0\{T_2 \geq t_2\} = p_2.$$

□

Some terminology and notation

- A simple (resp. composite) hypothesis is a hypothesis in which all (resp. not all) the parameters are specified.
  - The hypothesis $H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(0.1, 1)$ is simple.
  - The hypothesis $H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(\mu, 1), \mu > 0$ is composite.
- Assuming the data $D = (X_1, \ldots, X_n)$ are continuous, consider the simple hypotheses:

$$H_0 : D \sim F_0, \text{with joint density } p_0,$$
$$\text{versus,}$$
$$H_1 : D \sim F_1, \text{with joint density } p_1.$$

- The Likelihood Ratio Test (LRT) is defined as:

$$T = \frac{p_1(D)}{p_0(D)}.$$

Some terminology and notation

- A simple (resp. composite) hypothesis is a hypothesis in which all (resp. not all) the parameters are specified.
  - The hypothesis $H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(0.1, 1)$ is simple.
  - The hypothesis $H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(\mu, 1), \mu > 0$ is composite.
- Assuming the data $D = (X_1, \ldots, X_n)$ are continuous, consider the simple hypotheses:

$$H_0 : D \sim F_0, \text{with joint density } p_0,$$
$$\text{versus,}$$
$$H_1 : D \sim F_1, \text{with joint density } p_1.$$

- The Likelihood Ratio Test (LRT) is defined as:

$$T = \frac{p_1(D)}{p_0(D)}.$$

Equivalent terminology

In other contexts you might see the simple hypothesis testing setup written as:

$$H_0 : \theta = \theta_0,$$
$$\text{versus,}$$
$$H_1 : \theta = \theta_1,$$

and the LRT defined as

$$T = \frac{\ell(D; \theta_1)}{\ell(D; \theta_0)},$$

where ℓ denotes the likelihood of the data. Finally, you will often see the reverse definition

$$T = \frac{\ell(D; \theta_0)}{\ell(D; \theta_1)}.$$

This gives an equivalent test (by the equivalence under increasing transformations lemma) if we instead reject for small values of the test statistic.

The Neyman-Pearson Lemma

Theorem

Consider a hypothesis test between two simple hypotheses $H_0 : D \sim F_0$ versus $H_1 : D \sim F_1$. The test statistic

$$T = \frac{p_1(D)}{p_0(D)},$$

which rejects for large values of T, provides the most powerful test for H_0 versus H_1, at any significance level α.

Proof of the Neyman-Pearson Lemma

Fix α and set τ such that $\mathbb{P}_0(T \geq \tau) = \alpha$. Consider the *rejection region*:

$$R = \left\{ X : \frac{p_1(X)}{p_0(X)} \geq \tau \right\}.$$

Then $\mathbb{P}_0(D \in R) = \mathbb{P}_0(T \geq \tau) = \alpha$.

Now let $T' = g'(D)$ be another test statistic, with associated p-value P', and τ' a threshold satisfying $\mathbb{P}_0(T' \geq \tau') = \alpha$. Define the rejection region:

$$R' = \left\{ X : g(D') \geq \tau' \right\},$$

so that, similarly, $\mathbb{P}_0(D \in R') = \mathbb{P}_0(T' \geq \tau') = \alpha$.

Proof of the Neyman-Pearson Lemma

Fix α and set τ such that $\mathbb{P}_0(T \geq \tau) = \alpha$. Consider the *rejection region*:

$$R = \left\{ X : \frac{p_1(X)}{p_0(X)} \geq \tau \right\}.$$

Then $\mathbb{P}_0(D \in R) = \mathbb{P}_0(T \geq \tau) = \alpha$.

Now let $T' = g'(D)$ be another test statistic, with associated p-value P', and τ' a threshold satisfying $\mathbb{P}_0(T' \geq \tau') = \alpha$. Define the rejection region:

$$R' = \left\{ X : g(D') \geq \tau' \right\},$$

so that, similarly, $\mathbb{P}_0(D \in R') = \mathbb{P}_0(T' \geq \tau') = \alpha$.

Proof of the Neyman-Pearson Lemma

The power of the LRT is

$$\beta = \mathbb{P}_1(P \leq \alpha) = \mathbb{P}_1(T \geq \tau) = \mathbb{P}_1(D \in R),$$

whereas the power of the rival statistic is

$$\beta' = \mathbb{P}_1(P' \leq \alpha) = \mathbb{P}_1(T' \geq \tau') = \mathbb{P}_1(D \in R').$$

We are looking to prove $\beta \geq \beta'$.

Proof of the Neyman-Pearson Lemma

The power of the LRT is

$$\beta = \mathbb{P}_1(P \leq \alpha) = \mathbb{P}_1(T \geq \tau) = \mathbb{P}_1(D \in R),$$

whereas the power of the rival statistic is

$$\beta' = \mathbb{P}_1(P' \leq \alpha) = \mathbb{P}_1(T' \geq \tau') = \mathbb{P}_1(D \in R').$$

We are looking to prove $\beta \geq \beta'$.

Proof of the Neyman-Pearson Lemma

The power of the LRT is

$$\beta = \mathbb{P}_1(P \leq \alpha) = \mathbb{P}_1(T \geq \tau) = \mathbb{P}_1(D \in R),$$

whereas the power of the rival statistic is

$$\beta' = \mathbb{P}_1(P' \leq \alpha) = \mathbb{P}_1(T' \geq \tau') = \mathbb{P}_1(D \in R').$$

We are looking to prove $\beta \geq \beta'$.

Proof of the Neyman-Pearson Lemma

Let $S = R \cap \bar{R}', S' = R' \cap \bar{R}$. Then,

$$\begin{aligned}
\mathbb{P}_0(D \in S) &= \mathbb{P}_0(D \in R) - \mathbb{P}(D \in R \cap R') \\
&= \alpha - \mathbb{P}(D \in R \cap R') \\
&= \mathbb{P}_0(D \in R') - \mathbb{P}(D \in R \cap R') \\
&= \mathbb{P}_0(D \in S') = \alpha_0, \quad \text{say.}
\end{aligned}$$

Proof of the Neyman-Pearson Lemma

Furthermore, because

$$\beta = \mathbb{P}_1(D \in S) + \mathbb{P}_1(D \in R \cap R')$$
$$\beta' = \mathbb{P}_1(D \in S') + \mathbb{P}_1(D \in R \cap R'),$$

then $\beta \geq \beta'$ if and only if $\mathbb{P}_1(D \in S) \geq \mathbb{P}_1(D \in S')$.

Finally,

$$\mathbb{P}_1(D \in S) = \int_S p_1(X)\mathrm{d}X$$

Proof of the Neyman-Pearson Lemma

Finally,

$$\begin{aligned}\mathbb{P}_1(D \in S) &= \int_S p_1(X)\mathrm{d}X \\ &\geq \tau \int_S p_0(X)\mathrm{d}X\end{aligned}$$

Proof of the Neyman-Pearson Lemma

Finally,

$$\begin{aligned}\mathbb{P}_1(D \in S) &= \int_S p_1(X)\mathrm{d}X \\ &\geq \tau \int_S p_0(X)\mathrm{d}X \\ &= \tau\alpha_0\end{aligned}$$

Proof of the Neyman-Pearson Lemma

Finally,

$$\begin{aligned}
\mathbb{P}_1(D \in S) &= \int_S p_1(X)\mathrm{d}X \\
&\geq \tau \int_S p_0(X)\mathrm{d}X \\
&= \tau\alpha_0 \\
&= \tau \int_{S'} p_0(X)\mathrm{d}X
\end{aligned}$$

Proof of the Neyman-Pearson Lemma

Finally,

$$\begin{aligned}
\mathbb{P}_1(D \in S) &= \int_S p_1(X)\mathrm{d}X \\
&\geq \tau \int_S p_0(X)\mathrm{d}X \\
&= \tau\alpha_0 \\
&= \tau \int_{S'} p_0(X)\mathrm{d}X \\
&\geq \int_{S'} p_1(X)\mathrm{d}X
\end{aligned}$$

Proof of the Neyman-Pearson Lemma

Finally,

$$\begin{aligned}
\mathbb{P}_1(D \in S) &= \int_S p_1(X)\mathrm{d}X \\
&\geq \tau \int_S p_0(X)\mathrm{d}X \\
&= \tau\alpha_0 \\
&= \tau \int_{S'} p_0(X)\mathrm{d}X \\
&\geq \int_{S'} p_1(X)\mathrm{d}X \\
&= \mathbb{P}_1(D \in S').
\end{aligned}$$

Running example

Suppose we are testing:

$$H_0 : D = (X_1, \ldots, X_n) \overset{i.i.d}{\sim} \text{normal}(\mu_0, 1),$$

versus,

$$H_1 : D = (X_1, \ldots, X_n) \overset{i.i.d}{\sim} \text{normal}(\mu_1, 1).$$

where $\mu_1 > \mu_0$ (in our earlier example $\mu_0 = 0$, $\mu_1 = 0.1$).

Then the Neyman-Pearson lemma says that the most powerful test statistic is:

$$\frac{p_1(D)}{p_0(D)},$$

where p_0 (resp. p_1) is the joint density of D under H_0 (resp. H_1). These densities have the form:

$$p_k(D) = \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left\{-\frac{1}{2}(X_i - \mu_k)^2\right\}$$

Running example

Suppose we are testing:

$$H_0 : D = (X_1, \ldots, X_n) \overset{i.i.d}{\sim} \text{normal}(\mu_0, 1),$$

versus,

$$H_1 : D = (X_1, \ldots, X_n) \overset{i.i.d}{\sim} \text{normal}(\mu_1, 1).$$

where $\mu_1 > \mu_0$ (in our earlier example $\mu_0 = 0$, $\mu_1 = 0.1$).

Then the Neyman-Pearson lemma says that the most powerful test statistic is:

$$\frac{p_1(D)}{p_0(D)},$$

where p_0 (resp. p_1) is the joint density of D under H_0 (resp. H_1). These densities have the form:

$$\begin{aligned} p_k(D) &= \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left\{-\frac{1}{2}(X_i - \mu_k)^2\right\} \\ &= (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu_k)^2\right\}, \quad k = 0, 1. \end{aligned}$$

Running example

Therefore the LRT is

$$\exp\left[-\frac{1}{2}\left\{\sum(X_i-\mu_1)^2-\sum(X_i-\mu_0)^2\right\}\right],$$

or equivalently (using the increasing transformation lemma)

$$\sum(X_i-\mu_0)^2-\sum(X_i-\mu_1)^2=n\{2\bar{X}(\mu_1-\mu_0)+\mu_0^2-\mu_1^2\},$$

where $\bar{X}=\frac{1}{n}\sum X_i$.

Finally, noting that $\mu_1>\mu_0$, by the increasing transformation lemma this is equivalent to:

$$T=\text{␣␣}$$

Running example

Therefore the LRT is

$$\exp\left[-\frac{1}{2}\left\{\sum(X_i-\mu_1)^2-\sum(X_i-\mu_0)^2\right\}\right],$$

or equivalently (using the increasing transformation lemma)

$$\sum(X_i-\mu_0)^2-\sum(X_i-\mu_1)^2=n\{2\bar{X}(\mu_1-\mu_0)+\mu_0^2-\mu_1^2\},$$

where $\bar{X}=\frac{1}{n}\sum X_i$.

Finally, noting that $\mu_1>\mu_0$, by the increasing transformation lemma this is equivalent to:

$$T=\bar{X}.$$

Lecture outline

Objectives:

1. To make you comfortable with all the notions introduced so far.
   - We'll do this with a real data example.
2. But also to uncover some interesting phenomena that you might already have noticed, but perhaps without fully appreciating their consequences.

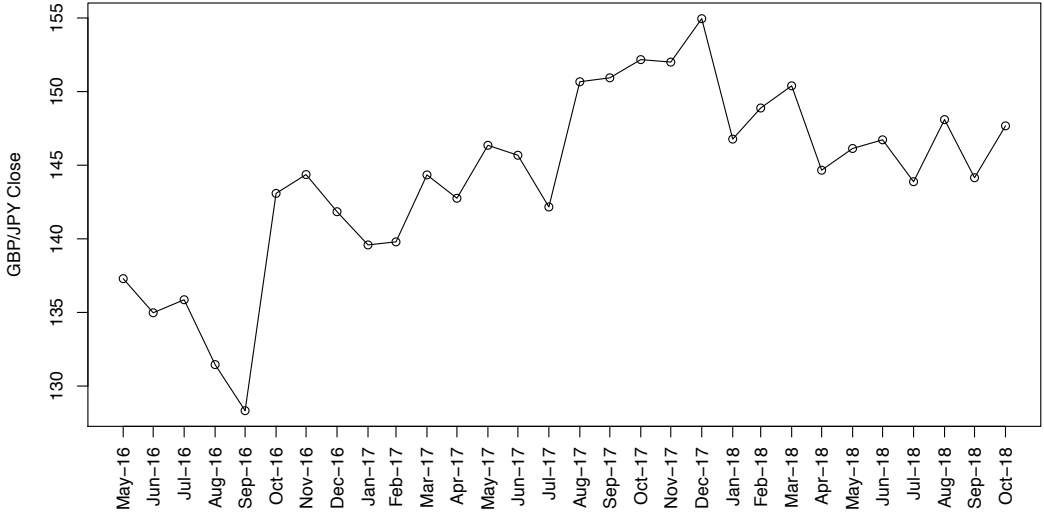
Data

- Monthly GBP/JPY exchange rate from the end of May 2016 until now.
- Because of the Brexit referendum in June 2016, it's a reasonable question whether UK currency has significantly increased or decreased in value.
- JPY taken as a relatively neutral and stable comparison currency.

Data

	Time	Open	High	Low	Close	Volume
1:	31.05.2016 21:00:00.000	160.293	160.654	133.260	137.296	4656524
2:	30.06.2016 21:00:00.000	137.338	143.226	128.788	134.977	6750952
3:	31.07.2016 21:00:00.000	135.188	136.261	129.211	135.863	7768794
4:	31.08.2016 21:00:00.000	135.863	138.830	129.637	131.458	9564064
5:	30.09.2016 21:00:00.000	130.748	132.226	122.383	128.327	6917943
6:	31.10.2016 21:00:00.000	128.333	143.255	126.490	143.087	5347793
7:	30.11.2016 22:00:00.000	143.077	148.456	142.167	144.366	4303096
8:	31.12.2016 22:00:00.000	144.082	145.397	136.455	141.842	5418184
9:	31.01.2017 22:00:00.000	141.838	144.127	138.537	139.587	4240536
10:	28.02.2017 22:00:00.000	139.583	140.750	137.520	139.793	4703494
11:	31.03.2017 21:00:00.000	139.707	144.493	135.597	144.342	3974897
12:	30.04.2017 21:00:00.000	143.958	148.107	141.491	142.756	4380996
13:	31.05.2017 21:00:00.000	142.754	146.543	138.669	146.352	3912584
14:	30.06.2017 21:00:00.000	145.768	147.776	144.026	145.677	4142948
15:	31.07.2017 21:00:00.000	145.688	146.797	139.312	142.163	4845566
16:	31.08.2017 21:00:00.000	142.168	152.856	141.194	150.670	4745183
17:	30.09.2017 21:00:00.000	150.598	151.394	146.944	150.935	3847771
18:	31.10.2017 21:00:00.000	150.916	152.407	146.975	152.173	3787820
19:	30.11.2017 22:00:00.000	152.173	153.408	149.410	152.004	3762345
20:	31.12.2017 22:00:00.000	151.972	156.083	150.193	154.953	5004841
21:	31.01.2018 22:00:00.000	154.936	156.607	146.725	146.774	6531522
22:	28.02.2018 22:00:00.000	146.776	150.589	144.985	148.886	6528444
23:	31.03.2018 21:00:00.000	148.927	153.852	148.384	150.393	4345260
24:	30.04.2018 21:00:00.000	150.419	150.631	143.195	144.662	5277603
25:	31.05.2018 21:00:00.000	144.663	148.117	143.771	146.136	5045401
26:	30.06.2018 21:00:00.000	145.848	149.310	145.183	146.726	4988749
27:	31.07.2018 21:00:00.000	146.726	147.146	139.896	143.880	3752805
28:	31.08.2018 21:00:00.000	143.429	149.712	142.594	148.099	3018616
29:	30.09.2018 21:00:00.000	148.257	149.509	142.768	144.150	5129264
30:	31.10.2018 21:00:00.000	144.175	149.485	144.023	147.673	1987536

Data


GBP/JPY Close
130
135
140
145
150
155
May−16
Jun−16
Jul−16
Aug−16
Sep−16
Oct−16
Nov−16
Dec−16
Jan−17
Feb−17
Mar−17
Apr−17
May−17
Jun−17
Jul−17
Aug−17
Sep−17
Oct−17
Nov−17
Dec−17
Jan−18
Feb−18
Mar−18
Apr−18
May−18
Jun−18
Jul−18
Aug−18
Sep−18
Oct−18

Analysis

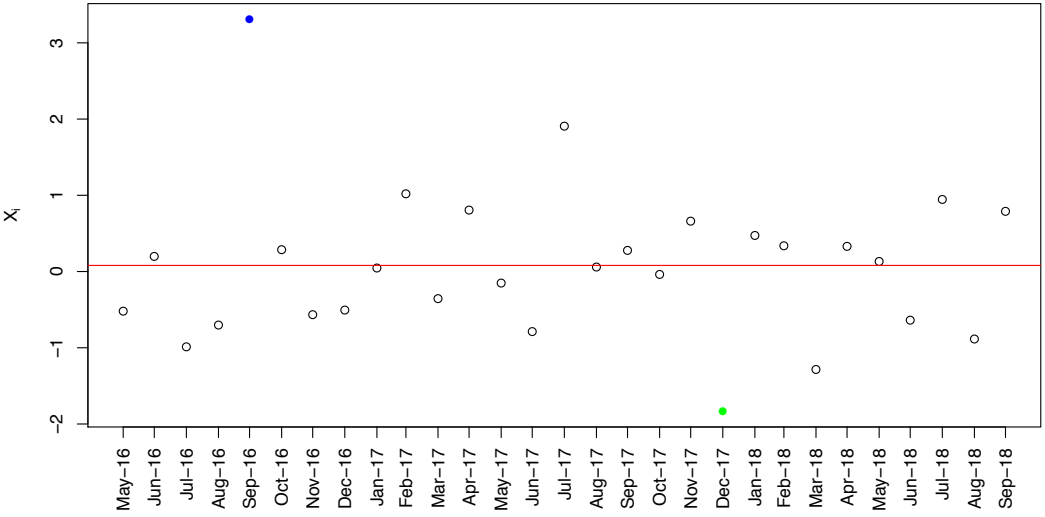
- Let $X_1, \ldots, X_n, n = 29$ denote the difference sequence.
- We'll silently divide by the empirical standard deviation and test the hypothesis:

$$H_0 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(0, 1),$$

versus,

$$H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(\mu_1, 1), \quad \mu_1 > 0.$$

Data


X_i
3
2
1
0
−1
−2
May−16
Jun−16
Jul−16
Aug−16
Sep−16
Oct−16
Nov−16
Dec−16
Jan−17
Feb−17
Mar−17
Apr−17
May−17
Jun−17
Jul−17
Aug−17
Sep−17
Oct−17
Nov−17
Dec−17
Jan−18
Feb−18
Mar−18
Apr−18
May−18
Jun−18
Jul−18
Aug−18
Sep−18

The basic test

- Say we want to reject at a significance level $\alpha = 0.05$.
- Then we will reject H_0 if $p \leq 0.05$ where:

$$p = 1 - \Phi(\sqrt{29}\bar{x}).$$

- Equivalently we will reject H_0 if the test statistic $t = \bar{x} \geq \tau$ where:

$$\tau = \frac{\Phi^{-1}(0.95)}{\sqrt{29}} \approx 0.31.$$

- Why? Because:

$$\begin{aligned} p \leq \alpha &\Leftrightarrow 1 - \Phi(\sqrt{29}\bar{x}) \leq \alpha, \\ &\Leftrightarrow \Phi(\sqrt{29}\bar{x}) \geq 1 - \alpha, \\ &\Leftrightarrow \sqrt{29}\bar{x} \quad \_\_\_ \quad \Phi^{-1}(1-\alpha), \\ &\Leftrightarrow \bar{x} \frac{\Phi^{-1}(1-\alpha)}{\sqrt{29}}. \end{aligned}$$

- We have $\bar{x} = 0.08$
- Equivalently we have $p = 1 - \Phi(\sqrt{29} \times 0.08) = 0.33$

The basic test

- Say we want to reject at a significance level $\alpha = 0.05$.
- Then we will reject H_0 if $p \leq 0.05$ where:

$$p = 1 - \Phi(\sqrt{29}\bar{x}).$$

- Equivalently we will reject H_0 if the test statistic $t = \bar{x} \geq \tau$ where:

$$\tau = \frac{\Phi^{-1}(0.95)}{\sqrt{29}} \approx 0.31.$$

- Why? Because:

$$\begin{aligned} p \leq \alpha &\Leftrightarrow 1 - \Phi(\sqrt{29}\bar{x}) \leq \alpha, \\ &\Leftrightarrow \Phi(\sqrt{29}\bar{x}) \geq 1 - \alpha, \\ &\Leftrightarrow \sqrt{29}\bar{x} \quad \_\_\_ \quad \Phi^{-1}(1-\alpha), \\ &\Leftrightarrow \bar{x} \frac{\Phi^{-1}(1-\alpha)}{\sqrt{29}}. \end{aligned}$$

- We have $\bar{x} = 0.08$
- Equivalently we have $p = 1 - \Phi(\sqrt{29} \times 0.08) = 0.33$

The basic test

- Say we want to reject at a significance level $\alpha = 0.05$.
- Then we will reject H_0 if $p \leq 0.05$ where:

$$p = 1 - \Phi(\sqrt{29}\bar{x}).$$

- Equivalently we will reject H_0 if the test statistic $t = \bar{x} \geq \tau$ where:

$$\tau = \frac{\Phi^{-1}(0.95)}{\sqrt{29}} \approx 0.31.$$

- Why? Because:

$$\begin{aligned} p \leq \alpha &\Leftrightarrow 1 - \Phi(\sqrt{29}\bar{x}) \leq \alpha, \\ &\Leftrightarrow \Phi(\sqrt{29}\bar{x}) \geq 1 - \alpha, \\ &\Leftrightarrow \sqrt{29}\bar{x} \quad \_\_\_ \quad \Phi^{-1}(1-\alpha), \\ &\Leftrightarrow \bar{x} \frac{\Phi^{-1}(1-\alpha)}{\sqrt{29}}. \end{aligned}$$

- We have $\bar{x} = 0.08$
- Equivalently we have $p = 1 - \Phi(\sqrt{29} \times 0.08) = 0.33$

The basic test

- Say we want to reject at a significance level $\alpha = 0.05$.
- Then we will reject H_0 if $p \leq 0.05$ where:

$$p = 1 - \Phi(\sqrt{29}\bar{x}).$$

- Equivalently we will reject H_0 if the test statistic $t = \bar{x} \geq \tau$ where:

$$\tau = \frac{\Phi^{-1}(0.95)}{\sqrt{29}} \approx 0.31.$$

- Why? Because:

$$\begin{aligned} p \leq \alpha &\Leftrightarrow 1 - \Phi(\sqrt{29}\bar{x}) \leq \alpha, \\ &\Leftrightarrow \Phi(\sqrt{29}\bar{x}) \geq 1 - \alpha, \\ &\Leftrightarrow \sqrt{29}\bar{x} \geq \Phi^{-1}(1 - \alpha), \\ &\Leftrightarrow \bar{x} \geq \frac{\Phi^{-1}(1 - \alpha)}{\sqrt{29}}. \end{aligned}$$

- We have $\bar{x} = 0.08$
- Equivalently we have $p = 1 - \Phi(\sqrt{29} \times 0.08) = 0.33$

The basic test

- Say we want to reject at a significance level $\alpha = 0.05$.
- Then we will reject H_0 if $p \leq 0.05$ where:

$$p = 1 - \Phi(\sqrt{29}\bar{x}).$$

- Equivalently we will reject H_0 if the test statistic $t = \bar{x} \geq \tau$ where:

$$\tau = \frac{\Phi^{-1}(0.95)}{\sqrt{29}} \approx 0.31.$$

- Why? Because:

$$\begin{aligned} p \leq \alpha &\Leftrightarrow 1 - \Phi(\sqrt{29}\bar{x}) \leq \alpha, \\ &\Leftrightarrow \Phi(\sqrt{29}\bar{x}) \geq 1 - \alpha, \\ &\Leftrightarrow \sqrt{29}\bar{x} \geq \Phi^{-1}(1 - \alpha), \\ &\Leftrightarrow \bar{x} \geq \frac{\Phi^{-1}(1 - \alpha)}{\sqrt{29}}. \end{aligned}$$

- We have $\bar{x} = 0.08$
- Equivalently we have $p = 1 - \Phi(\sqrt{29} \times 0.08) = 0.33$

The basic test

- Say we want to reject at a significance level $\alpha = 0.05$.
- Then we will reject H_0 if $p \leq 0.05$ where:

$$p = 1 - \Phi(\sqrt{29}\bar{x}).$$

- Equivalently we will reject H_0 if the test statistic $t = \bar{x} \geq \tau$ where:

$$\tau = \frac{\Phi^{-1}(0.95)}{\sqrt{29}} \approx 0.31.$$

- Why? Because:

$$\begin{aligned} p \leq \alpha &\Leftrightarrow 1 - \Phi(\sqrt{29}\bar{x}) \leq \alpha, \\ &\Leftrightarrow \Phi(\sqrt{29}\bar{x}) \geq 1 - \alpha, \\ &\Leftrightarrow \sqrt{29}\bar{x} \geq \Phi^{-1}(1 - \alpha), \\ &\Leftrightarrow \bar{x} \geq \frac{\Phi^{-1}(1 - \alpha)}{\sqrt{29}}. \end{aligned}$$

- We have $\bar{x} = 0.08$
- Equivalently we have $p = 1 - \Phi(\sqrt{29} \times 0.08) = 0.33$

The basic test

- Say we want to reject at a significance level $\alpha = 0.05$.
- Then we will reject H_0 if $p \leq 0.05$ where:

$$p = 1 - \Phi(\sqrt{29}\bar{x}).$$

- Equivalently we will reject H_0 if the test statistic $t = \bar{x} \geq \tau$ where:

$$\tau = \frac{\Phi^{-1}(0.95)}{\sqrt{29}} \approx 0.31.$$

- Why? Because:

$$\begin{aligned} p \leq \alpha &\Leftrightarrow 1 - \Phi(\sqrt{29}\bar{x}) \leq \alpha, \\ &\Leftrightarrow \Phi(\sqrt{29}\bar{x}) \geq 1 - \alpha, \\ &\Leftrightarrow \sqrt{29}\bar{x} \geq \Phi^{-1}(1 - \alpha), \\ &\Leftrightarrow \bar{x} \geq \frac{\Phi^{-1}(1 - \alpha)}{\sqrt{29}}. \end{aligned}$$

- We have $\bar{x} = 0.08 < 0.31$, therefore no rejection.
- Equivalently we have $p = 1 - \Phi(\sqrt{29} \times 0.08) = 0.33$

The basic test

- Say we want to reject at a significance level $\alpha = 0.05$.
- Then we will reject H_0 if $p \leq 0.05$ where:

$$p = 1 - \Phi(\sqrt{29}\bar{x}).$$

- Equivalently we will reject H_0 if the test statistic $t = \bar{x} \geq \tau$ where:

$$\tau = \frac{\Phi^{-1}(0.95)}{\sqrt{29}} \approx 0.31.$$

- Why? Because:

$$\begin{aligned} p \leq \alpha &\Leftrightarrow 1 - \Phi(\sqrt{29}\bar{x}) \leq \alpha, \\ &\Leftrightarrow \Phi(\sqrt{29}\bar{x}) \geq 1 - \alpha, \\ &\Leftrightarrow \sqrt{29}\bar{x} \geq \Phi^{-1}(1 - \alpha), \\ &\Leftrightarrow \bar{x} \geq \frac{\Phi^{-1}(1 - \alpha)}{\sqrt{29}}. \end{aligned}$$

- We have $\bar{x} = 0.08 < 0.31$, therefore no rejection.
- Equivalently we have $p = 1 - \Phi(\sqrt{29} \times 0.08) = 0.33 > 0.05$, therefore no rejection.

A different test

Consider instead the test statistic $T = \max(X_i)$. Under H_0:

$$\mathbb{P}_0(T \geq t) =$$

A different test

Consider instead the test statistic $T = \max(X_i)$. Under H_0:

$$\mathbb{P}_0(T \geq t) = 1 - \mathbb{P}_0\{\max(X_i) \leq t\},$$

A different test

Consider instead the test statistic $T = \max(X_i)$. Under H_0:

$$\begin{aligned}\mathbb{P}_0(T \geq t) &= 1 - \mathbb{P}_0\{\max(X_i) \leq t\}, \\ &= 1 - \mathbb{P}_0(X_1, \ldots, X_n \leq t),\end{aligned}$$

A different test

Consider instead the test statistic $T = \max(X_i)$. Under H_0:

$$\begin{aligned}\mathbb{P}_0(T \geq t) &= 1 - \mathbb{P}_0\{\max(X_i) \leq t\}, \\ &= 1 - \mathbb{P}_0(X_1, \ldots, X_n \leq t), \\ &= 1 - \Phi(t)^n.\end{aligned}$$

A different test

Consider instead the test statistic $T = \max(X_i)$. Under H_0:

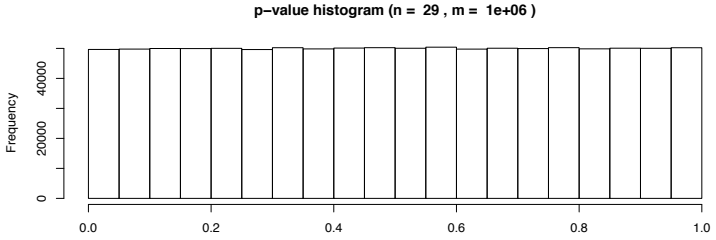
$$\begin{aligned}\mathbb{P}_0(T \geq t) &= 1 - \mathbb{P}_0\{\max(X_i) \leq t\},\\ &= 1 - \mathbb{P}_0(X_1, \ldots, X_n \leq t),\\ &= 1 - \Phi(t)^n.\end{aligned}$$

To simulate a p-value under the null, we will therefore:

1. Generate data $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{normal}(0, 1)$, $n = 29$
2. Compute $T = \max(X_i)$
3. Compute $P = 1 - \Phi(T)^n$

and we'll repeat this a million times to obtain a large sample of p-values.

Distribution of the p-value under the null hypothesis



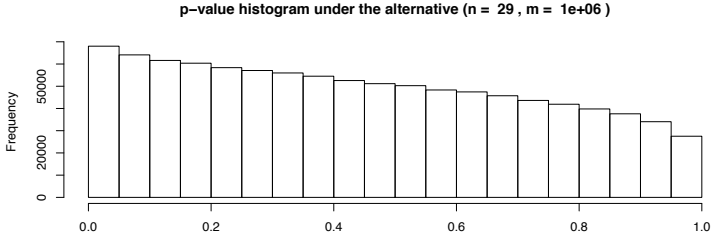
Distribution of the p-value under the alternative

Consider the (usual) alternative hypothesis $H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(0.1, 1)$. To simulate a p-value under this alternative we will:

1. Generate data $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{normal}(0.1, 1)$, $n = 29$
2. Compute $T = \max(X_i)$
3. Compute $P = 1 - \Phi(T)^n$

and we'll repeat this a million times to obtain a large sample of p-values.

Distribution of the p-value under the alternative

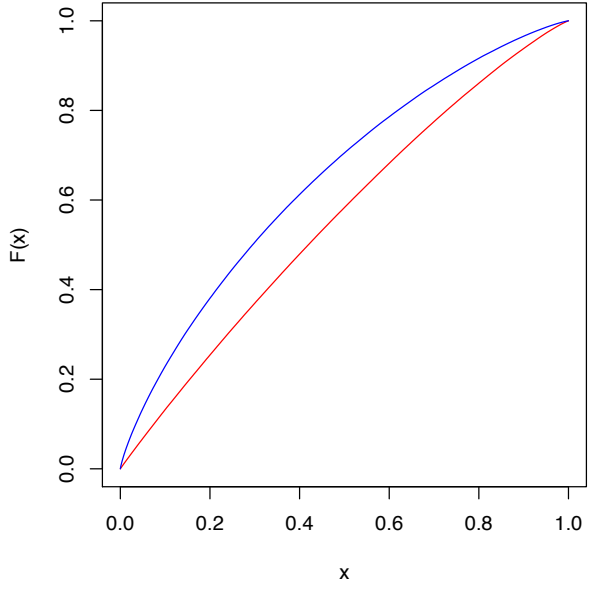


- Remember that the power of the test is the probability of rejecting under the alternative, i.e. $\beta = \mathbb{P}_1(P \leq \alpha)$.
- If we plot this value for different α, we are simply plotting the cumulative distribution of the p-value under the alternative.
- This curve is called the "power curve".

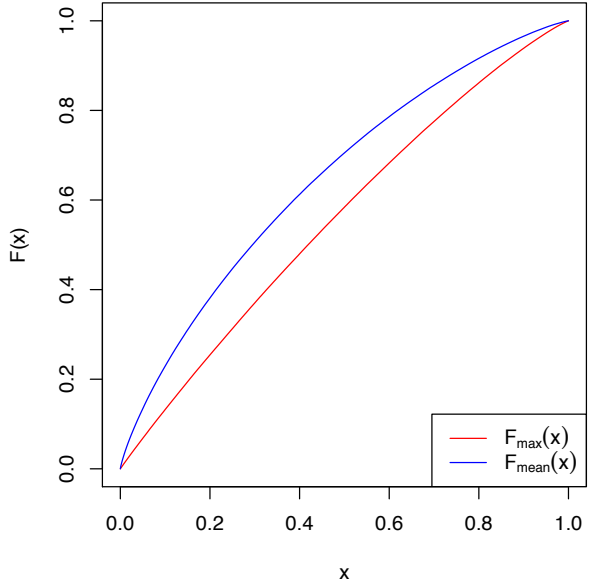
- Remember that the power of the test is the probability of rejecting under the alternative, i.e. $\beta = \mathbb{P}_1(P \leq \alpha)$.
- If we plot this value for different α, we are simply plotting the cumulative distribution of the p-value under the alternative.
- This curve is called the "power curve".

- Remember that the power of the test is the probability of rejecting under the alternative, i.e. $\beta = \mathbb{P}_1(P \leq \alpha)$.
- If we plot this value for different α, we are simply plotting the cumulative distribution of the p-value under the alternative.
- This curve is called the "power curve".

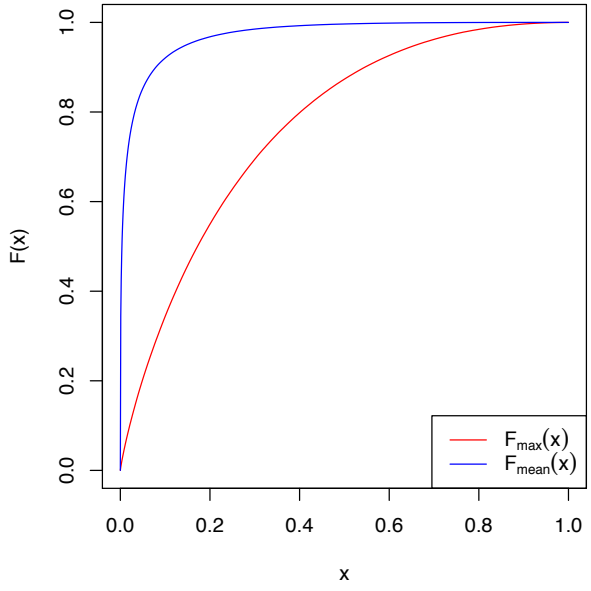
Power comparison under H_1 : normal$(0.1, 1)$ (max versus mean)



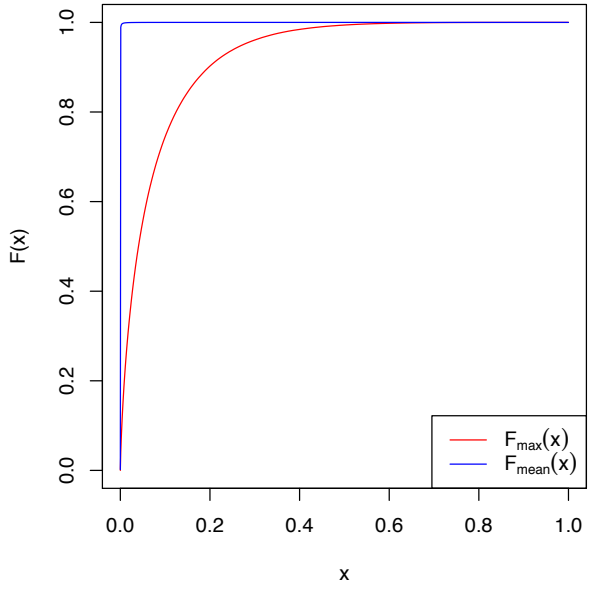
Power comparison under H_1 : normal$(0.1, 1)$ (max versus mean)



Power comparison under H_1 : normal$(0.5, 1)$ (max versus mean)



Power comparison under H_1 : normal$(1, 1)$ (max versus mean)



Uniformly most powerful test

- Note that the Neyman-Pearson lemma gives us the most powerful test for two simple (read: specific) hypotheses. Under a range of alternative hypotheses, as we are considering here with $H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(\mu_1, 1), \mu_1 > 0$, there is no fixed "likelihood ratio" — it depends on μ_1. In general, we will not be able to unilaterally recommend one test over another over a range of alternatives.
- **However**: we found in this instance that the LRT comes down to a test of the mean, which does not depend on the unknown μ_1. No matter what μ_1 is, by testing the mean we are implementing a test that is equivalent to using the LRT based on this unknown value of μ_1. This is the ideal scenario, but it's unfortunately unusual.
- A test is called the uniformly most powerful test (at level α), if it is the most powerful for all alternatives considered (at level α).

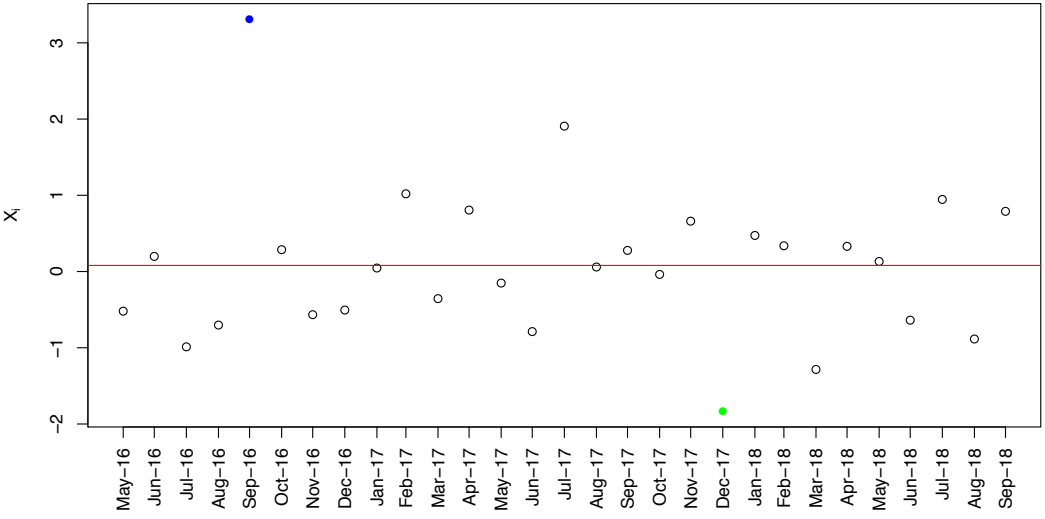
Uniformly most powerful test

- Note that the Neyman-Pearson lemma gives us the most powerful test for two simple (read: specific) hypotheses. Under a range of alternative hypotheses, as we are considering here with $H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(\mu_1, 1), \mu_1 > 0$, there is no fixed "likelihood ratio" — it depends on μ_1. In general, we will not be able to unilaterally recommend one test over another over a range of alternatives.
- **However**: we found in this instance that the LRT comes down to a test of the mean, which does not depend on the unknown μ_1. No matter what μ_1 is, by testing the mean we are implementing a test that is equivalent to using the LRT based on this unknown value of μ_1. This is the ideal scenario, but it's unfortunately unusual.
- A test is called the uniformly most powerful test (at level α), if it is the most powerful for all alternatives considered (at level α).

Uniformly most powerful test

- Note that the Neyman-Pearson lemma gives us the most powerful test for two simple (read: specific) hypotheses. Under a range of alternative hypotheses, as we are considering here with $H_1 : X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{normal}(\mu_1, 1), \mu_1 > 0$, there is no fixed "likelihood ratio" — it depends on μ_1. In general, we will not be able to unilaterally recommend one test over another over a range of alternatives.
- **However**: we found in this instance that the LRT comes down to a test of the mean, which does not depend on the unknown μ_1. No matter what μ_1 is, by testing the mean we are implementing a test that is equivalent to using the LRT based on this unknown value of μ_1. This is the ideal scenario, but it's unfortunately unusual.
- A test is called the uniformly most powerful test (at level α), if it is the most powerful for all alternatives considered (at level α).

Data


X_i
3
2
1
0
−1
−2
May−16
Jun−16
Jul−16
Aug−16
Sep−16
Oct−16
Nov−16
Dec−16
Jan−17
Feb−17
Mar−17
Apr−17
May−17
Jun−17
Jul−17
Aug−17
Sep−17
Oct−17
Nov−17
Dec−17
Jan−18
Feb−18
Mar−18
Apr−18
May−18
Jun−18
Jul−18
Aug−18
Sep−18

Implementing the maximum test

- Say we want to reject at a significance level $\alpha = 0.05$.
- Then we will reject H_0 if $p \leq 0.05$ where:

$$p = 1 - \Phi\{\max(x_i)\}^{29}.$$

- Equivalently we will reject H_0 if the test statistic $t = \max(x_i) \geq \tau$ where:

$$\tau =$$

- We have $\max(x_i) = 3.3$, therefore we reject.
- Equivalently we have $p = 1 - \Phi\{3.3\}^{29} \approx 0.01$, therefore we reject.

Implementing the maximum test

- Say we want to reject at a significance level $\alpha = 0.05$.
- Then we will reject H_0 if $p \leq 0.05$ where:

$$p = 1 - \Phi\{\max(x_i)\}^{29}.$$

- Equivalently we will reject H_0 if the test statistic $t = \max(x_i) \geq \tau$ where:

$$\tau =$$

- We have $\max(x_i) = 3.3$, therefore we reject.
- Equivalently we have $p = 1 - \Phi\{3.3\}^{29} \approx 0.01$, therefore we reject.

Implementing the maximum test

- Say we want to reject at a significance level $\alpha = 0.05$.
- Then we will reject H_0 if $p \leq 0.05$ where:

$$p = 1 - \Phi\{\max(x_i)\}^{29}.$$

- Equivalently we will reject H_0 if the test statistic $t = \max(x_i) \geq \tau$ where:

$$\tau = \quad \sqcup\sqcup\sqcup$$

- We have $\max(x_i) = 3.3$, therefore we reject.
- Equivalently we have $p = 1 - \Phi\{3.3\}^{29} \approx 0.01$, therefore we reject.

Implementing the maximum test

- Say we want to reject at a significance level $\alpha = 0.05$.
- Then we will reject H_0 if $p \leq 0.05$ where:

$$p = 1 - \Phi\{\max(x_i)\}^{29}.$$

- Equivalently we will reject H_0 if the test statistic $t = \max(x_i) \geq \tau$ where:

$$\tau = \Phi^{-1}[\exp\{\log(0.95)/29\}] \approx 2.91.$$

- We have $\max(x_i) = 3.3$, therefore we reject.
- Equivalently we have $p = 1 - \Phi\{3.3\}^{29} \approx 0.01$, therefore we reject.

Other tests

- Let $X_{(1)} \leq \ldots \leq X_{(n)}$ denote the order statistics of $X_1, \ldots, X_n$, that is, $X_{(k)}$ is the kth smallest of $X_1, \ldots, X_n$. We might have instead used any $T_k = X_{(k)}$.
- Denote by $F(x, r)$ the cumulative distribution function of a Binomial variable X with success probability r and number of trials n, that is,

$$F(x, r) = \mathbb{P}(X \leq x).$$

- Then,

$$\mathbb{P}_0(T_k \geq t) =$$

Other tests

- Let $X_{(1)} \leq \ldots \leq X_{(n)}$ denote the order statistics of $X_1, \ldots, X_n$, that is, $X_{(k)}$ is the kth smallest of $X_1, \ldots, X_n$. We might have instead used any $T_k = X_{(k)}$.
- Denote by $F(x, r)$ the cumulative distribution function of a Binomial variable X with success probability r and number of trials n, that is,

$$F(x, r) = \mathbb{P}(X \leq x).$$

- Then,

$$\mathbb{P}_0(T_k \geq t) =$$

Other tests

- Let $X_{(1)} \leq \ldots \leq X_{(n)}$ denote the order statistics of $X_1, \ldots, X_n$, that is, $X_{(k)}$ is the kth smallest of $X_1, \ldots, X_n$. We might have instead used any $T_k = X_{(k)}$.
- Denote by $F(x, r)$ the cumulative distribution function of a Binomial variable X with success probability r and number of trials n, that is,

$$F(x, r) = \mathbb{P}(X \leq x).$$

- Then,

$$\mathbb{P}_0(T_k \geq t) =$$

Other tests

- Let $X_{(1)} \leq \ldots \leq X_{(n)}$ denote the order statistics of $X_1, \ldots, X_n$, that is, $X_{(k)}$ is the kth smallest of $X_1, \ldots, X_n$. We might have instead used any $T_k = X_{(k)}$.
- Denote by $F(x, r)$ the cumulative distribution function of a Binomial variable X with success probability r and number of trials n, that is,

$$F(x, r) = \mathbb{P}(X \leq x).$$

- Then,

$$\mathbb{P}_0(T_k \geq t) = \mathbb{P}_0(\text{at least } n - k + 1 \text{ of } X_i \geq t),$$

Other tests

- Let $X_{(1)} \leq \ldots \leq X_{(n)}$ denote the order statistics of $X_1, \ldots, X_n$, that is, $X_{(k)}$ is the kth smallest of $X_1, \ldots, X_n$. We might have instead used any $T_k = X_{(k)}$.
- Denote by $F(x, r)$ the cumulative distribution function of a Binomial variable X with success probability r and number of trials n, that is,

$$F(x, r) = \mathbb{P}(X \leq x).$$

- Then,

$$\begin{aligned} \mathbb{P}_0(T_k \geq t) &= \mathbb{P}_0(\text{at least } n - k + 1 \text{ of } X_i \geq t), \\ &= \mathbb{P}_0\left\{\sum_{i=1}^{n} \mathbb{I}(X_i \geq t) \geq n - k + 1\right\}, \end{aligned}$$

Other tests

- Let $X_{(1)} \leq \ldots \leq X_{(n)}$ denote the order statistics of $X_1, \ldots, X_n$, that is, $X_{(k)}$ is the kth smallest of $X_1, \ldots, X_n$. We might have instead used any $T_k = X_{(k)}$.
- Denote by $F(x, r)$ the cumulative distribution function of a Binomial variable X with success probability r and number of trials n, that is,

$$F(x, r) = \mathbb{P}(X \leq x).$$

- Then,

$$\begin{aligned}
\mathbb{P}_0(T_k \geq t) &= \mathbb{P}_0(\text{at least } n-k+1 \text{ of } X_i \geq t), \\
&= \mathbb{P}_0\left\{\sum_{i=1}^{n} \mathbb{I}(X_i \geq t) \geq n-k+1\right\}, \\
&= \mathbb{P}_0\left\{\sum_{i=1}^{n} \mathbb{I}(X_i \geq t) > n-k\right\}, \\
&= 1 - \mathbb{P}_0\left\{\sum_{i=1}^{n} \mathbb{I}(X_i \geq t) \leq n-k\right\},
\end{aligned}$$

Other tests

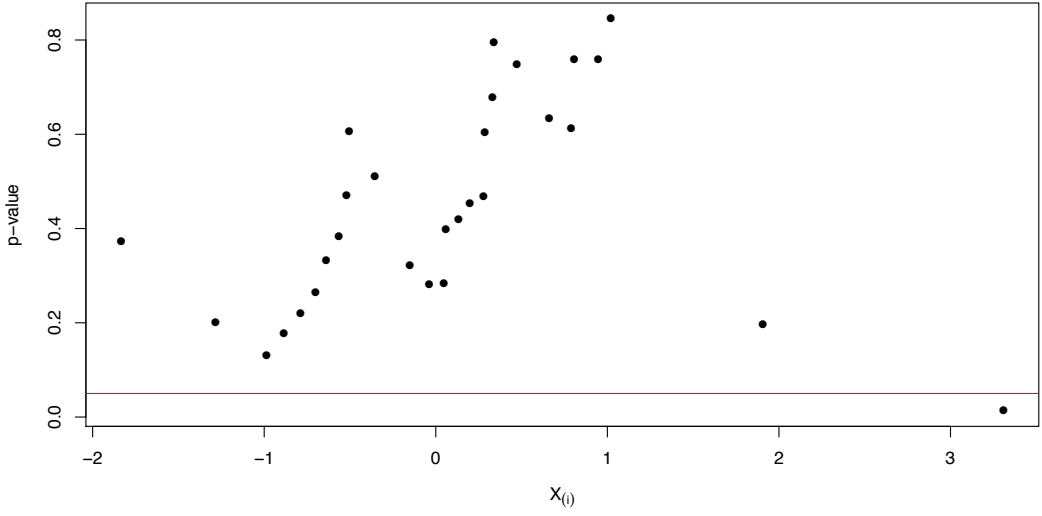
- Let $X_{(1)} \leq \ldots \leq X_{(n)}$ denote the order statistics of $X_1, \ldots, X_n$, that is, $X_{(k)}$ is the kth smallest of $X_1, \ldots, X_n$. We might have instead used any $T_k = X_{(k)}$.
- Denote by $F(x, r)$ the cumulative distribution function of a Binomial variable X with success probability r and number of trials n, that is,

$$F(x, r) = \mathbb{P}(X \leq x).$$

- Then,

$$\begin{aligned}
\mathbb{P}_0(T_k \geq t) &= \mathbb{P}_0(\text{at least } n-k+1 \text{ of } X_i \geq t), \\
&= \mathbb{P}_0\left\{\sum_{i=1}^{n} \mathbb{I}(X_i \geq t) \geq n-k+1\right\}, \\
&= \mathbb{P}_0\left\{\sum_{i=1}^{n} \mathbb{I}(X_i \geq t) > n-k\right\}, \\
&= 1 - \mathbb{P}_0\left\{\sum_{i=1}^{n} \mathbb{I}(X_i \geq t) \leq n-k\right\}, \\
&= 1 - F\{n-k, 1-\Phi(t)\}.
\end{aligned}$$

P-values for all order statistics



GBP/USD Close
1.25
1.30
1.35
1.40
May−16
Jun−16
Jul−16
Aug−16
Sep−16
Oct−16
Nov−16
Dec−16
Jan−17
Feb−17
Mar−17
Apr−17
May−17
Jun−17
Jul−17
Aug−17
Sep−17
Oct−17
Nov−17
Dec−17
Jan−18
Feb−18
Mar−18
Apr−18
May−18
Jun−18
Jul−18
Aug−18
Sep−18
Oct−18

EUR/GBP Close
0.84
0.86
0.88
0.90
0.92
May−16
Jun−16
Jul−16
Aug−16
Sep−16
Oct−16
Nov−16
Dec−16
Jan−17
Feb−17
Mar−17
Apr−17
May−17
Jun−17
Jul−17
Aug−17
Sep−17
Oct−17
Nov−17
Dec−17
Jan−18
Feb−18
Mar−18
Apr−18
May−18
Jun−18
Jul−18
Aug−18
Sep−18
Oct−18