Goodness of Fit

Goodness of fit - 2 classes

A	B
78	22

$\longrightarrow$ Do these data correspond reasonably to the proportions 3:1?

We previously discussed options for testing $p_A = 0.75$!

- Exact p-value
- Exact confidence interval
- Normal approximation

Goodness of fit - 3 classes

AA	AB	BB
35	43	22

$\longrightarrow$ Do these data correspond reasonably to the proportions 1:2:1?

Multinomial distribution

Let $(p_1, p_2, p_3) = (0.25, 0.50, 0.25)$ and $n = 100$.

Using the Multinomial distribution function:

$$P(X_1{=}35, X_2{=}43, X_3{=}22) = \frac{100!}{35!\ 43!\ 22!}\ 0.25^{35}\ 0.50^{43}\ 0.25^{22}$$

$$= 7.3 \times 10^{-4}$$

Goodness of fit test

We observe $(n_1, n_2, n_3) \sim$ Multinomial(n,p=$\{p_1, p_2, p_3\}$).

We seek to test $H_0 : p_1 = 0.25, p_2 = 0.5, p_3 = 0.25.$

versus $H_a : H_0$ is false.

We need two things:

$\longrightarrow$ A test statistic.

$\longrightarrow$ The null distribution of the test statistic.

The likelihood-ratio test (LRT)

Back to the first example:

A	B
n_A	n_B

Test $H_0 : (p_A, p_B) = (\pi_A, \pi_B)$ versus $H_a : (p_A, p_B) \neq (\pi_A, \pi_B)$.

$\longrightarrow$ MLE under H_a: $\hat{p}_A = n_A/n$ where $n = n_A + n_B$.

Likelihood under H_a: $L_a = \Pr(n_A | p_A = \hat{p}_A) = \binom{n}{n_A} \times \hat{p}_A^{n_A} \times (1 - \hat{p}_A)^{n-n_A}$

Likelihood under H_0: $L_0 = \Pr(n_A | p_A = \pi_A) = \binom{n}{n_A} \times \pi_A^{n_A} \times (1 - \pi_A)^{n-n_A}$

$\longrightarrow$ Likelihood ratio test statistic: LRT = $2 \times \ln(L_a/L_0)$

$\longrightarrow$ Some clever people have shown that if H_0 is true, then LRT follows a χ^2(df=1) distribution (approximately).

Likelihood-ratio test for the example

We observed $n_A = 78$ and $n_B = 22$.

$H_0 : (p_A, p_B) = (0.75, 0.25)$

$H_a : (p_A, p_B) \neq (0.75, 0.25)$

$L_a = \Pr(n_A{=}78 \mid p_A{=}0.78) = \binom{100}{78} \times 0.78^{78} \times 0.22^{22} = 0.096.$

$L_0 = \Pr(n_A{=}78 \mid p_A{=}0.75) = \binom{100}{78} \times 0.75^{78} \times 0.25^{22} = 0.075.$

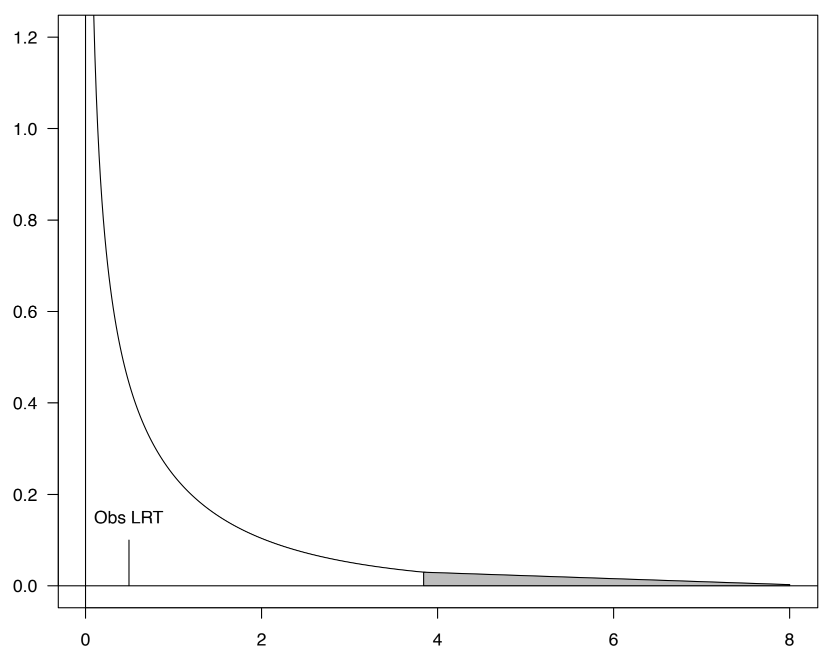
$\longrightarrow$ LRT $= 2 \times \ln(L_a/L_0) = 0.49.$

Using a χ^2(df=1) distribution, we get a p-value of 0.48.

We therefore have no evidence against the null hypothesis.

In R: p-value = `1 - pchisq(0.49,1)`

Null distribution



A little math ...

$n = n_A + n_B, \quad n_A^0 = E[n_A \mid H_0] = n \times \pi_A, \quad n_B^0 = E[n_B \mid H_0] = n \times \pi_B.$

Then $\quad L_a/L_0 = \left(\frac{n_A}{n_A^0}\right)^{n_A} \times \left(\frac{n_B}{n_B^0}\right)^{n_B}$

Or equivalently $\quad LRT = 2 \times n_A \times \ln\left(\frac{n_A}{n_A^0}\right) + 2 \times n_B \times \ln\left(\frac{n_B}{n_B^0}\right).$

$\longrightarrow$ Why do this?

Generalization to more than two groups

If we have k groups, then the likelihood ratio test statistic is

$$LRT = 2 \times \sum_{i=1}^{k} n_i \times \ln\left(\frac{n_i}{n_i^0}\right)$$

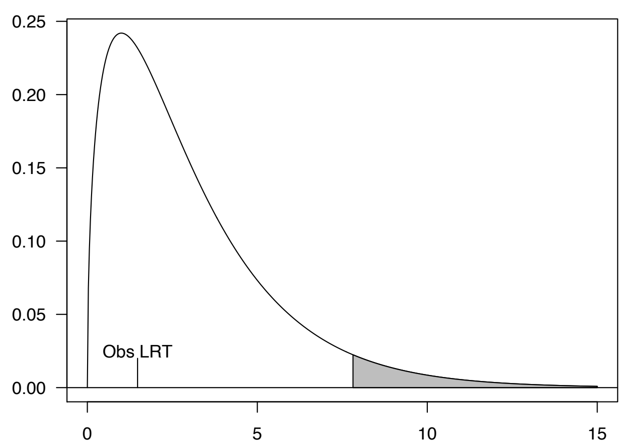
If H_0 is true, $LRT \sim \chi^2(df=k-1)$

Example

In a dihybrid cross of tomatos we expect the ratio of the phenotypes to be 9:3:3:1. In 1611 tomatos, we observe the numbers 926, 288, 293, 104. Do these numbers support our hypothesis?

Phenotype	n_i	n_i^0	n_i/n_i^0	$n_i \times \ln(n_i/n_i^0)$
Tall, cut-leaf	926	906.2	1.02	20.03
Tall, potato-leaf	288	302.1	0.95	-13.73
Dwarf, cut-leaf	293	302.1	0.97	-8.93
Dwarf, potato-leaf	104	100.7	1.03	3.37
Sum	1611			0.74

Results



The test statistics LRT is 1.48. Using a χ^2(df=3) distribution, we get a p-value of 0.69. We therefore have no evidence against the hypothesis that the ratio of the phenotypes is 9:3:3:1.

The chi-square test

There is an alternative technique. The test is called the chi-square test, and has the greater tradition in the literature. For two groups, calculate the following:

$$X^2 = \frac{(n_A - n_A^0)^2}{n_A^0} + \frac{(n_B - n_B^0)^2}{n_B^0}$$

$\longrightarrow$ If H_0 is true, then X^2 is a draw from a χ^2(df=1) distribution (approximately).

Example

In the first example we observed $n_A = 78$ and $n_B = 22$. Under the null hypothesis we have $n_A^0 = 75$ and $n_B^0 = 25$. We therefore get

$$X^2 = \frac{(78\text{-}75)^2}{75} + \frac{(22\text{-}25)^2}{25} = 0.12 + 0.36 = 0.48.$$

This corresponds to a p-value of 0.49. We therefore have no evidence against the hypothesis $(p_A, p_B) = (0.75, 0.25)$.

$\longrightarrow$ Note: using the likelihood ratio test we got a p-value of 0.48.

In R: `chisq.test(c(78,22),p=c(0.75,0.25))`

Generalization to more than two groups

As with the likelihood ratio test, there is a generalization to more than just two groups.

If we have k groups, the chi-square test statistic we use is

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - n_i^0)^2}{n_i^0} \sim \chi^2(\text{df=k-1})$$

Tomato example

For the tomato example we get

$$X^2 = \frac{(926\text{-}906.2)^2}{906.2} + \frac{(288\text{-}302.1)^2}{302.1} + \frac{(293\text{-}302.1)^2}{302.1} + \frac{(104\text{-}100.7)^2}{100.7}$$

$$= 0.43 + 0.65 + 0.27 + 0.11 = 1.47$$

Using a χ^2(df=3) distribution, we get a p-value of 0.69. We therefore have no evidence against the hypothesis that the ratio of the phenotypes is 9:3:3:1.

$\longrightarrow$ Using the likelihood ratio test we also got a p-value of 0.69.

In R: `chisq.test(c(926,288,293,104),p=c(9,3,3,1)/16)`

Test statistics

Let n_i^0 denote the expected count in group i if H_0 is true.

LRT statistic

$$\text{LRT} = 2 \ln \left\{ \frac{\Pr(\text{data} \mid p = \text{MLE})}{\Pr(\text{data} \mid H_0)} \right\} = \ldots = 2 \textstyle\sum_i n_i \ln(n_i/n_i^0)$$

χ^2 test statistic

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_i \frac{(n_i - n_i^0)^2}{n_i^0}$$

Null distribution of test statistic

What values of LRT (or X^2) should we expect, if H_0 were true?

The null distributions of these statistics may be obtained by:

- Brute-force analytic calculations
- Computer simulations
- Asymptotic approximations

$\longrightarrow$ If the sample size n is large, we have

$\text{LRT} \sim \chi^2(k-1)$ and $X^2 \sim \chi^2(k-1)$

The brute-force method

$$\Pr(\text{LRT} \geq g \mid H_0) = \sum_{\substack{n_1, n_2, n_3 \\ \text{giving LRT} \geq g}} \Pr(n_1, n_2, n_3 \mid H_0)$$

This is not feasible.

Computer simulation

1. Simulate a table conforming to the null hypothesis.
   E.g., simulate $(n_1, n_2, n_3) \sim$ Multinomial(n=100, $\{$1/4, 1/2, 1/4$\}$)
2. Calculate your test statistic.
3. Repeat steps (1) and (2) many (e.g., 1000 or 10,000) times.

Estimated critical value $\rightarrow$ the 95th percentile of the results.

Estimated P-value $\rightarrow$ the prop'n of results $\geq$ the observed value.

In R, use `rmultinom(n, size, prob)` to do `n` simulations of a Multinomial(`size`, `prob`).

Example

We observe the following data:

AA	AB	BB
35	43	22

We imagine that these are counts

$$(n_1, n_2, n_3) \sim \text{Multinomial}(n{=}100, \{p_1, p_2, p_3\}).$$

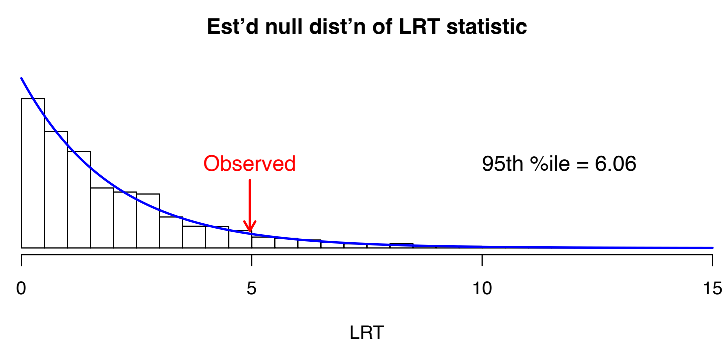
We seek to test $H_0 : p_1 = 1/4,\ p_2 = 1/2,\ p_3 = 1/4$.

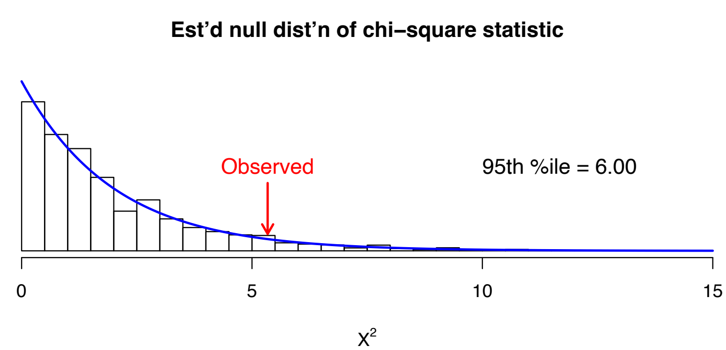
We calculate LRT = 4.96 and X^2 = 5.34.

Referring to the asymptotic approximations (χ^2 dist'n with 2 degrees of freedom), we obtain p = 8.4% and p = 6.9%.

With 10,000 simulations under H_0, we get p = 8.9% and p = 7.4%.

Example





Summary and recommendation

For either the LRT or the χ^2 test:

$\longrightarrow$ The null distribution is approximately $\chi^2(k-1)$ if the sample size is large.

$\longrightarrow$ The null distribution can be approximated by simulating data under the null hypothesis.

If the sample size is sufficiently large that the expected count in each cell is ≥ 5, use the asymptotic approximation without worries.

Otherwise, consider using computer simulations.

Composite hypotheses

Sometimes, we ask not $\quad p_{AA} = 0.25,\ p_{AB} = 0.5,\ p_{BB} = 0.25$

But rather something like:

$$p_{AA} = f^2,\ p_{AB} = 2f(1-f),\ p_{BB} = (1-f)^2 \qquad \text{for some f.}$$

For example: Consider the genotypes, of a random sample of individuals, at a diallelic locus.

$\longrightarrow$ Is the locus in Hardy-Weinberg equilibrium (as expected in the case of random mating)?

Example data:

AA	AB	BB
5	20	75

Another example

ABO blood groups $\longrightarrow$ 3 alleles A, B, O.

Phenotype	A	genotype AA or AO
	B	genotype BB or BO
	AB	genotype AB
	O	genotype O

Allele frequencies: f_A, f_B, f_O (Note that $f_A + f_B + f_O = 1$)

Under Hardy-Weinberg equilibrium, we expect

$$p_A = f_A^2 + 2f_Af_O \quad p_B = f_B^2 + 2f_Bf_O \quad p_{AB} = 2f_Af_B \quad p_O = f_O^2$$

Example data:

O	A	B	AB
104	91	36	19

LRT for example 1

Data: $(n_{AA}, n_{AB}, n_{BB}) \sim \text{Multinomial}(n, \{p_{AA}, p_{AB}, p_{BB}\})$

We seek to test whether the data conform reasonably to

H_0: $p_{AA} = f^2,\ p_{AB} = 2f(1-f),\ p_{BB} = (1-f)^2$ for some f.

General MLEs:

$\hat{p}_{AA} = n_{AA}/n,\ \hat{p}_{AB} = n_{AB}/n,\ \hat{p}_{BB} = n_{BB}/n$

MLE under H_0:

$\hat{f} = (n_{AA} + n_{AB}/2)/n \longrightarrow \tilde{p}_{AA} = \hat{f}^2, \tilde{p}_{AB} = 2\,\hat{f}\,(1-\hat{f}), \tilde{p}_{BB} = (1-\hat{f})^2$

LRT statistic: $$\text{LRT} = 2 \times \ln\left\{\frac{\Pr(n_{AA}, n_{AB}, n_{BB} \mid \hat{p}_{AA}, \hat{p}_{AB}, \hat{p}_{BB})}{\Pr(n_{AA}, n_{AB}, n_{BB} \mid \tilde{p}_{AA}, \tilde{p}_{AB}, \tilde{p}_{BB})}\right\}$$

LRT for example 2

Data: $(n_O, n_A, n_B, n_{AB}) \sim \text{Multinomial}(n, \{p_O, p_A, p_B, p_{AB}\})$

We seek to test whether the data conform reasonably to

H_0: $p_A = f_A^2 + 2f_Af_O$, $p_B = f_B^2 + 2f_Bf_O$, $p_{AB} = 2f_Af_B$, $p_O = f_O^2$

for some f_O, f_A, f_B, where $f_O + f_A + f_B = 1$.

General MLEs: $\hat{p}_O, \hat{p}_A, \hat{p}_B, \hat{p}_{AB}$, like before.

MLE under H_0: Requires numerical optimization

Call them $(\hat{f}_O, \hat{f}_A, \hat{f}_B) \longrightarrow (\tilde{p}_O, \tilde{p}_A, \tilde{p}_B, \tilde{p}_{AB})$

LRT statistic: $\text{LRT} = 2 \times \ln \left\{ \frac{\Pr(n_O, n_A, n_B, n_{AB} \mid \hat{p}_O, \hat{p}_A, \hat{p}_B, \hat{p}_{AB})}{\Pr(n_O, n_A, n_B, n_{AB} \mid \tilde{p}_O, \tilde{p}_A, \tilde{p}_B, \tilde{p}_{AB})} \right\}$

χ^2 test for these examples

- Obtain the MLE(s) under H_0.
- Calculate the corresponding cell probabilities.
- Turn these into (estimated) expected counts under H_0.
- Calculate $X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

Null distribution for these cases

- Computer simulation (with one wrinkle)
  - Simulate data under H_0 (plug in the MLEs for the observed data)
  - Calculate the MLE with the simulated data
  - Calculate the test statistic with the simulated data
  - Repeat many times

- Asymptotic approximation
  - Under H_0, if the sample size, n, is large, both the LRT statistic and the χ^2 statistic follow, approximately, a χ^2 distribution with k – s – 1 degrees of freedom, where s is the number of parameters estimated under H_0.
  - Note that s = 1 for example 1, and s = 2 for example 2, and so df = 1 for both examples.

Example 1

Example data:

AA	AB	BB
5	20	75

MLE: $\hat{f}$ = (5 + 20/2) / 100 = 15%

Expected counts:

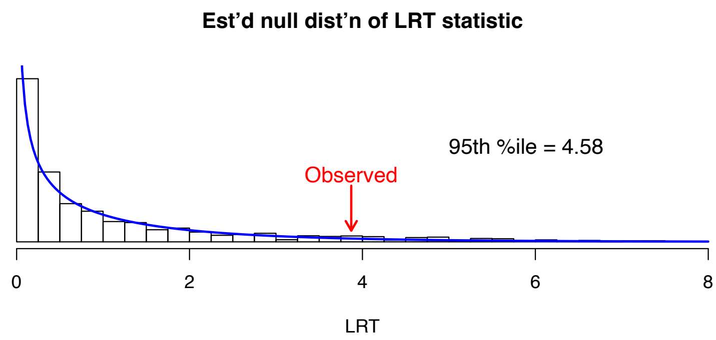
2.25	25.5	72.25

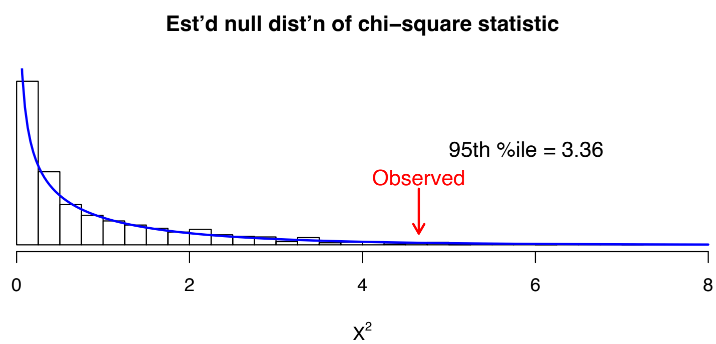
Test statistics: LRT statistic = 3.87 X^2 = 4.65

Asymptotic $\chi^2(\text{df} = 1)$ approx'n: p = 4.9% p = 3.1%

10,000 computer simulations: p = 8.2% p = 2.4%

Example 1





Example 2

Example data:

O	A	B	AB
104	91	36	19

MLE: $\hat{f}_O$ = 62.8%, $\hat{f}_A$ = 25.0%, $\hat{f}_B$ = 12.2%.

Expected counts:

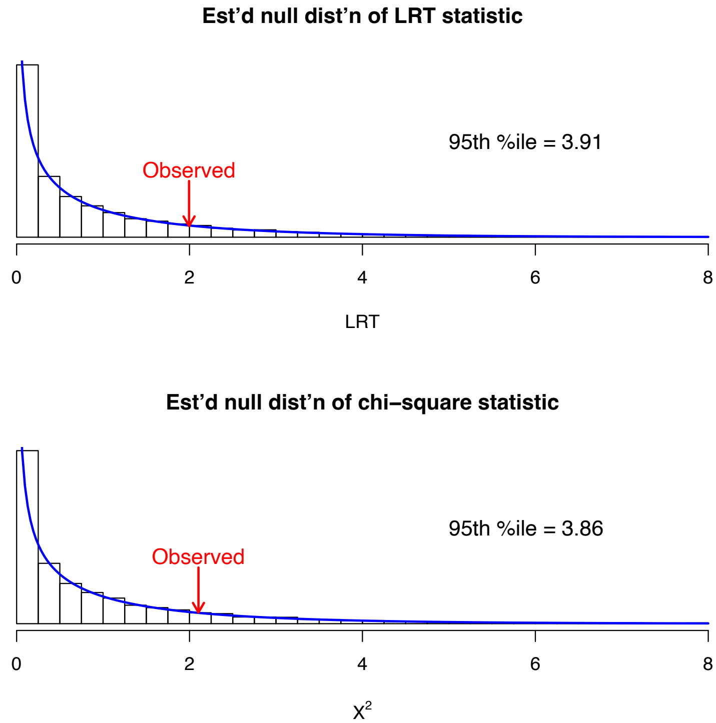
98.5	94.2	42.0	15.3

Test statistics: LRT statistic = 1.99 X^2 = 2.10

Asymptotic $\chi^2(\text{df} = 1)$ approx'n: p = 16% p = 15%

10,000 computer simulations: p = 17% p = 15%

Example 2



Example 3

Data on number of sperm bound to an egg:

	0	1	2	3	4	5
count	26	4	4	2	1	1

$\longrightarrow$ Do these follow a Poisson distribution?

MLE:

$\hat{\lambda}$ = sample average = (0 $\times$ 26 + 1 $\times$ 4 + . . . + 5 $\times$ 1) / 38 = 0.71

Expected counts $\longrightarrow$ $n_i^0 = n \times e^{-\hat{\lambda}}\ \hat{\lambda}^i\ /\ i!$

Example 3

	0	1	2	3	4	5
observed	26	4	4	2	1	1
expected	18.7	13.3	4.7	1.1	0.2	0.0

$X^2 = \sum \frac{(\text{obs}-\text{exp})^2}{\text{exp}} = \ldots = 42.8$

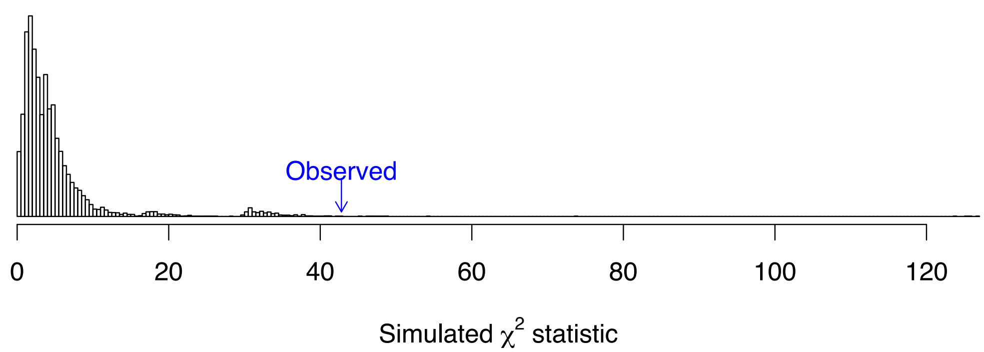
$\text{LRT} = 2 \sum \text{obs} \log(\text{obs}/\text{exp}) = \ldots = 18.8$

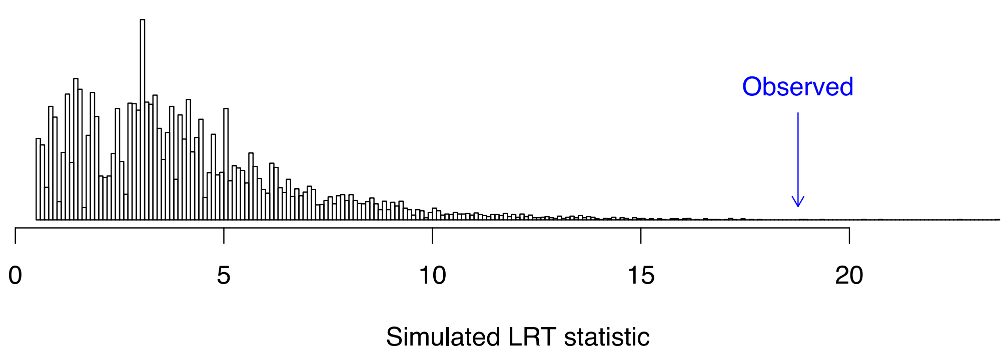
Compare to $\chi^2(\text{df} = 6 - 1 - 1 = 4)$

P-value = 1×10^{-8} (χ^2) and 9×10^{-4} (LRT).

By simulation: p-value = 16/10,000 (χ^2) and 7/10,000 (LRT)

Null simulation results





A final note

With these sorts of goodness-of-fit tests, we are often happy when our model does fit.

In other words, we often prefer to fail to reject H_0.

Such a conclusion, that the data fit the model reasonably well, should be phrased and considered with caution.

We should think: how much power do I have to detect, with these limited data, a reasonable deviation from H_0?