Goodness of fit - 2 classes

A	B
78	22

⟶ 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

⟶ Do these data correspond reasonably to the proportions 1:2:1?

Multinomial distribution

- Imagine an urn with k types of balls.
- Let p_i denote the proportion of type i.
- Draw n balls with replacement.
- Outcome: $(n_1, n_2, \ldots, n_k)$, with $\sum_i n_i = n$, where n_i is the no. balls drawn that were of type i.

$$\longrightarrow \quad P(X_1{=}n_1, \ldots, X_k{=}n_k) = \frac{n!}{n_1! \times \cdots \times n_k!} p_1^{n_1} \times \cdots \times p_k^{n_k}$$

$$\text{if} \quad 0 \le n_i \le n, \quad \sum_i n_i = n$$

$$\text{Otherwise} \quad P(X_1{=}n_1, \ldots, X_k{=}n_k) = 0.$$

Example

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

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

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

Rather brutal, numerically speaking.

⟶ Take logs (and use a computer).

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

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

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

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

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

$\longrightarrow$ 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 $n_B = 22$.

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

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

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

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

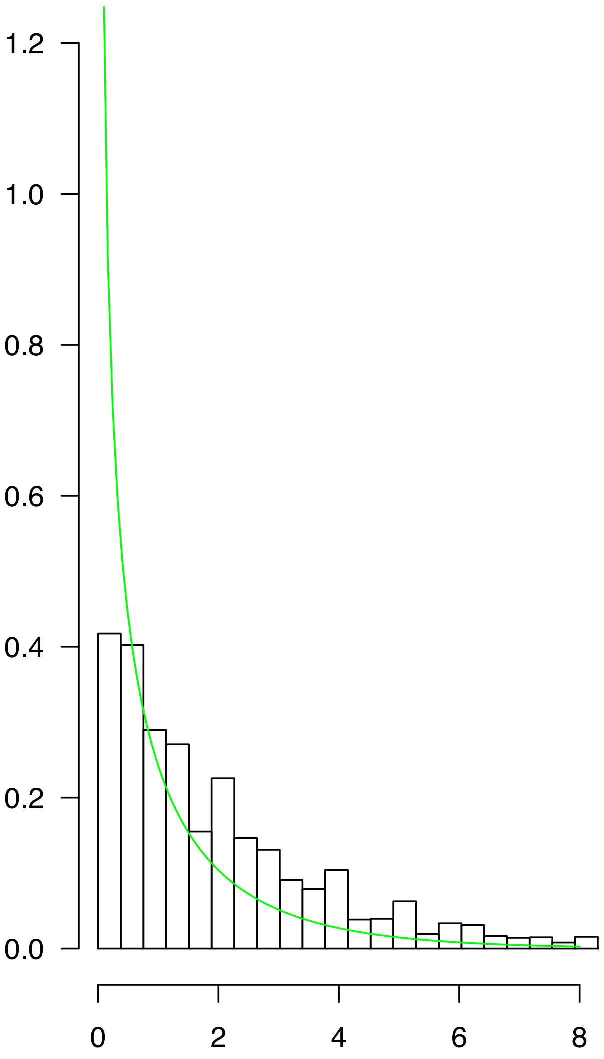
$\longrightarrow \quad \text{LRT} = 2 \times \ln(L_a/L_0) = 0.49.$

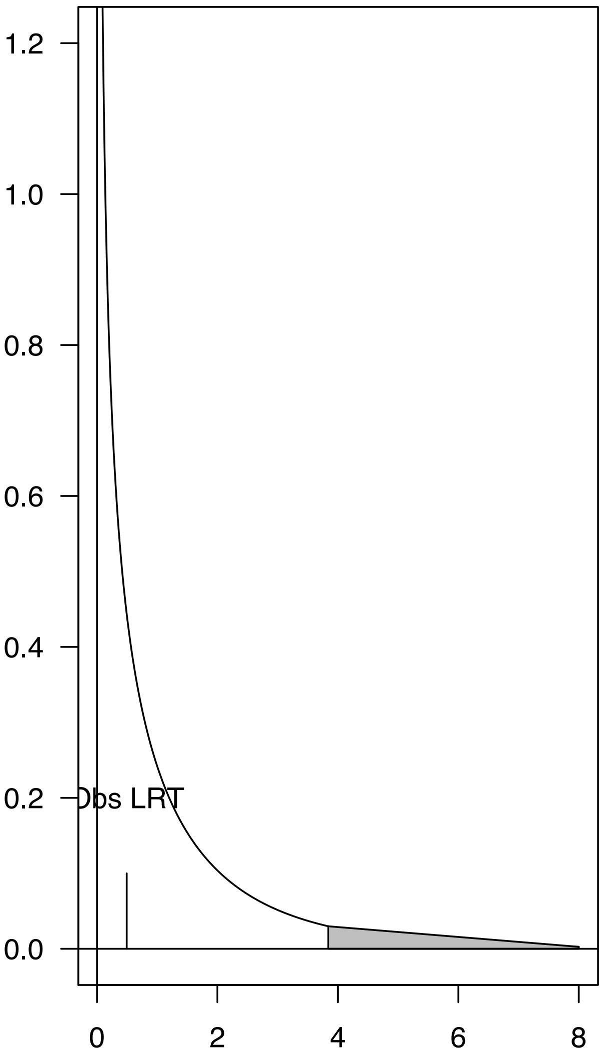
Using a $\chi^2(\text{df}=1)$ distribution, we get a p-value of 0.48.

We therefore have no evidence against the null hypothesis.

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

Null distribution





A little math . . .

$\mathrm{n} = \mathrm{n_A} + \mathrm{n_B}, \quad \mathrm{n_A^0} = \mathrm{E}[\mathrm{n_A} \mid \mathrm{H_0}] = \mathrm{n} \times \pi_\mathrm{A}, \quad \mathrm{n_B^0} = \mathrm{E}[\mathrm{n_B} \mid \mathrm{H_0}] = \mathrm{n} \times \pi_\mathrm{B}.$

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

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

⟶ Why do this?

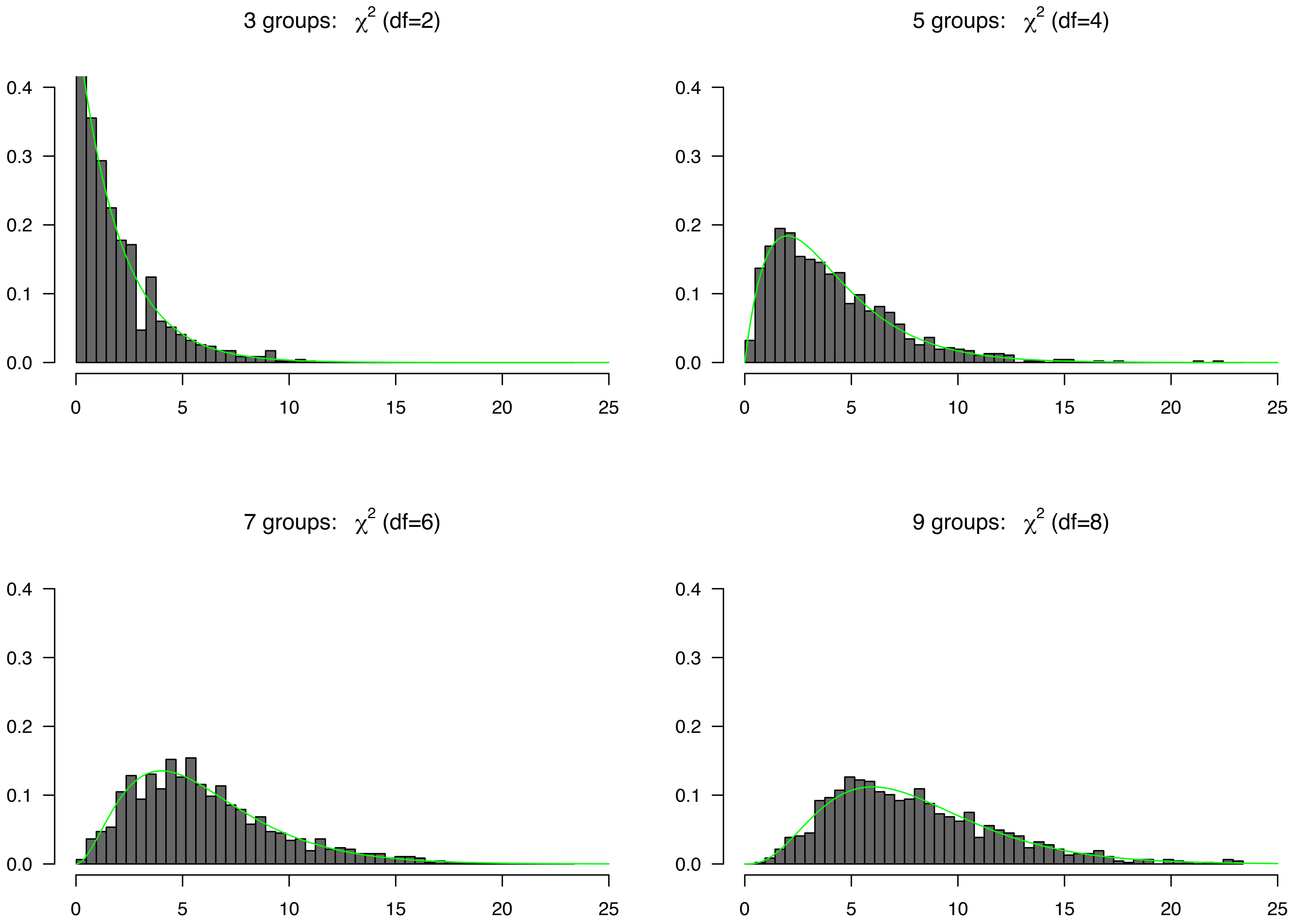
Generalization to more than two groups

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

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

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

Null distributions

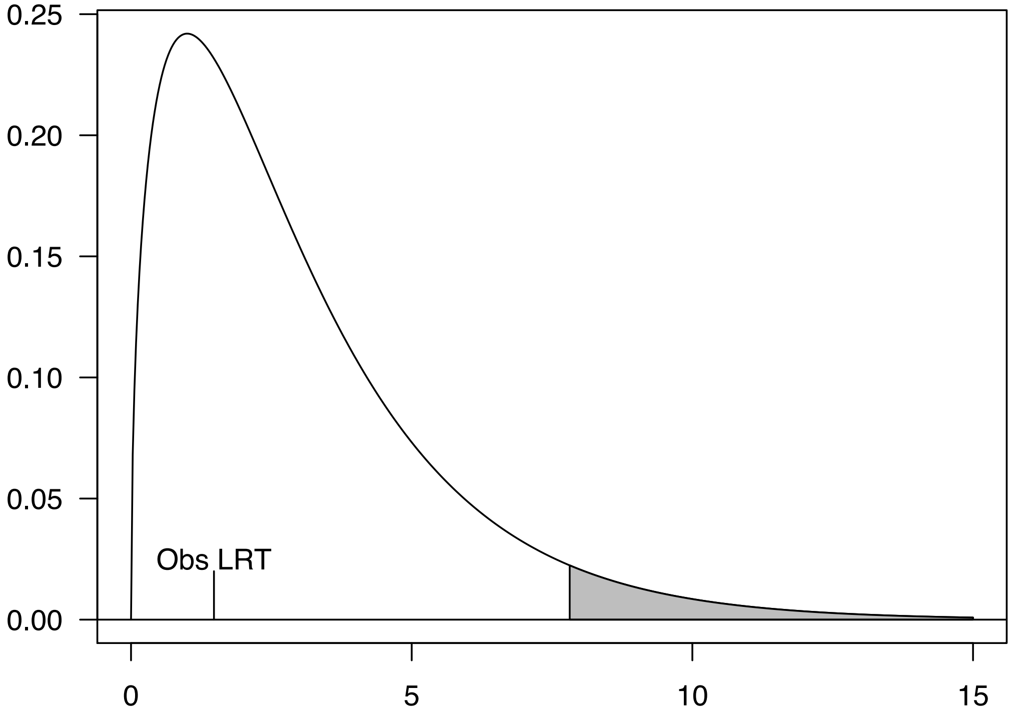


Example

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:

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

$\longrightarrow$ 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

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

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

⟶ 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

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

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

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

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

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

The brute-force method

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

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 $\longrightarrow$ the 95th percentile of the results.

Estimated P-value $\longrightarrow$ 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

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

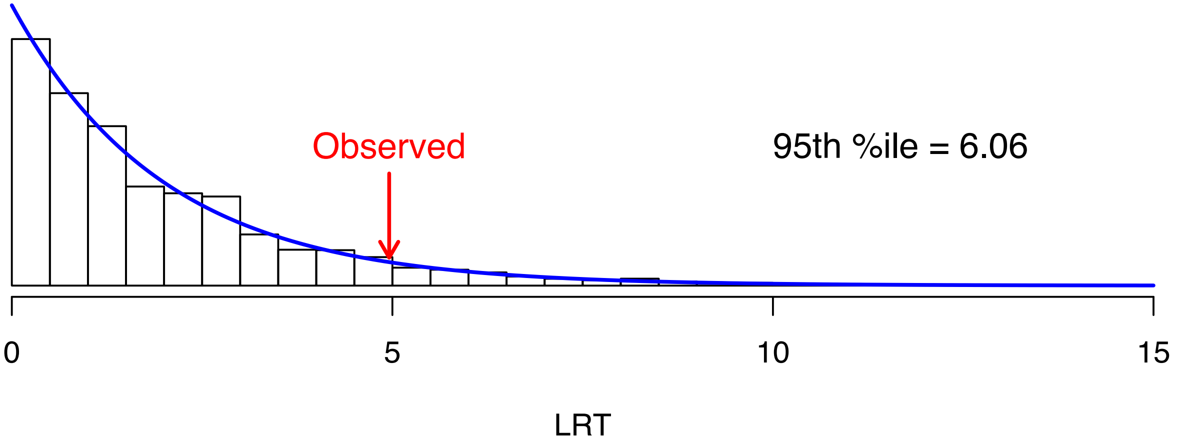
We seek to test $H_0 : p_1 = 1/4,\ p_2 = 1/2,\ p_3 = 1/4$.

We calculate LRT ≈ 4.96 and $X^2 \approx 5.34$.

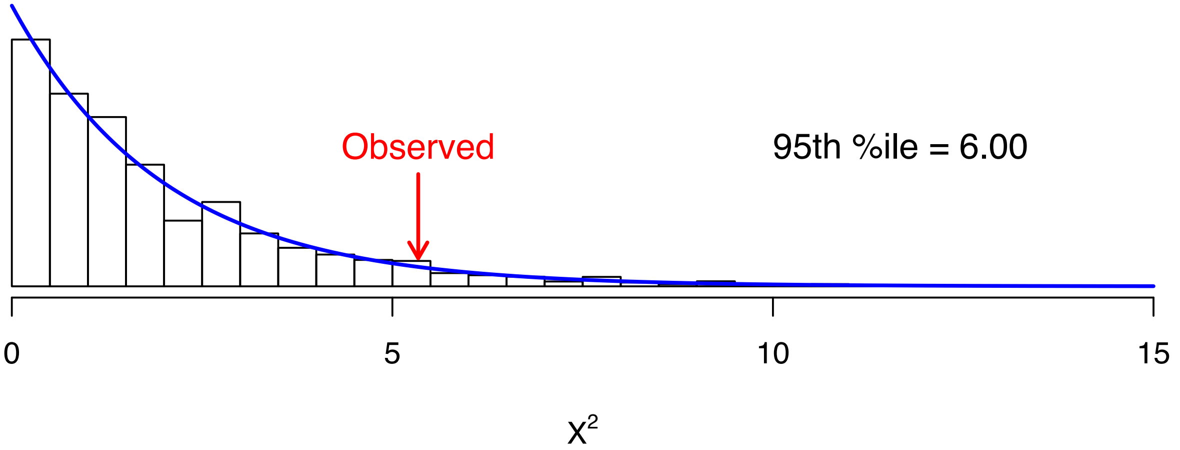
Referring to the asymptotic approximations (χ^2 dist'n with 2 degrees of freedom), we obtain $P \approx 8.4\%$ and $P \approx 6.9\%$.

With 10,000 simulations under H_0, we get $P \approx 8.9\%$ and $P \approx 7.4\%$.

Example



Est’d null dist’n of LRT statistic



Est’d null dist’n of chi–square statistic

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.

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

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:

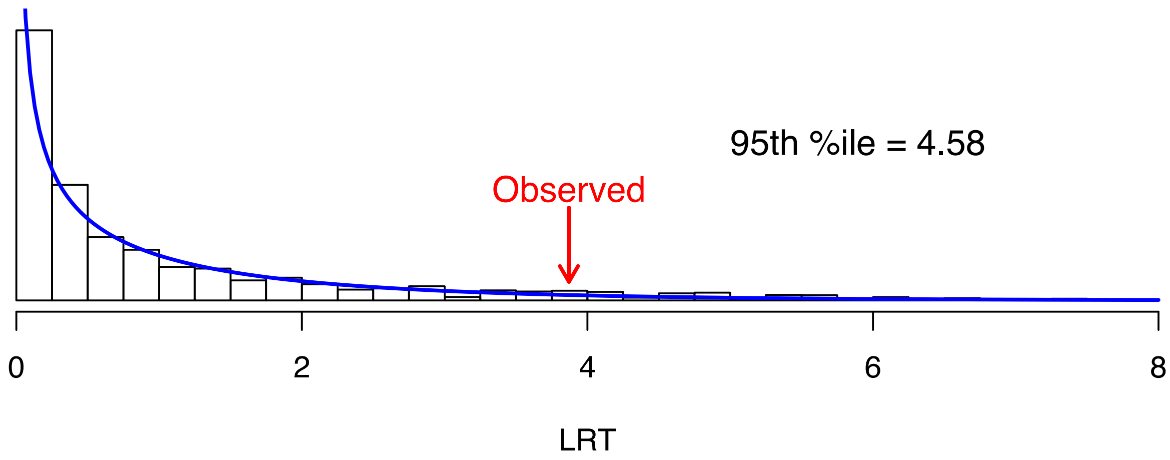
2.25	25.5	72.25

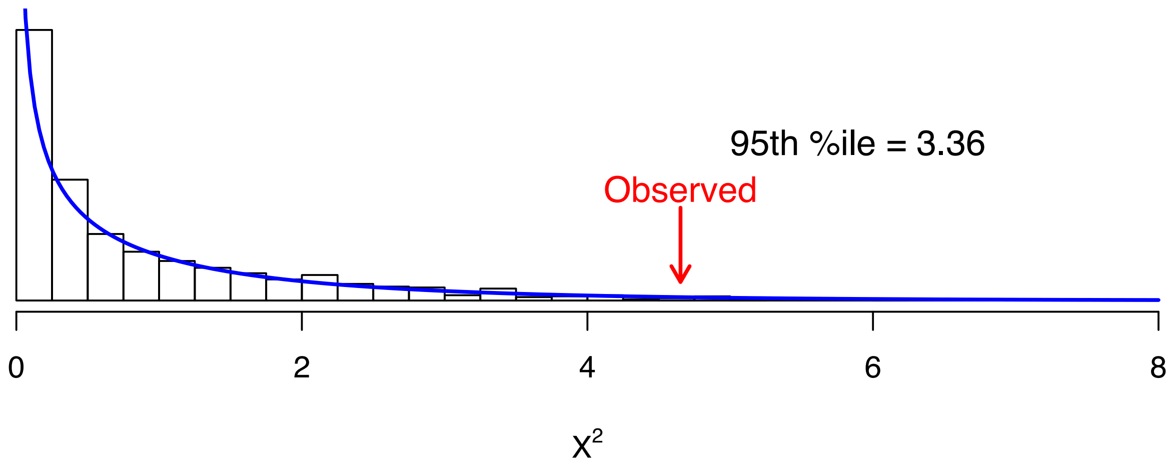
Test statistics: LRT statistic = 3.87 $X^2 = 4.65$

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

10,000 computer simulations: P $\approx$ 8.2% P $\approx$ 2.4%

Example 1


Est'd null dist'n of LRT statistic


Est'd null dist'n of chi–square statistic

Example 2

Example data:

O	A	B	AB
104	91	36	19

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

Expected counts:

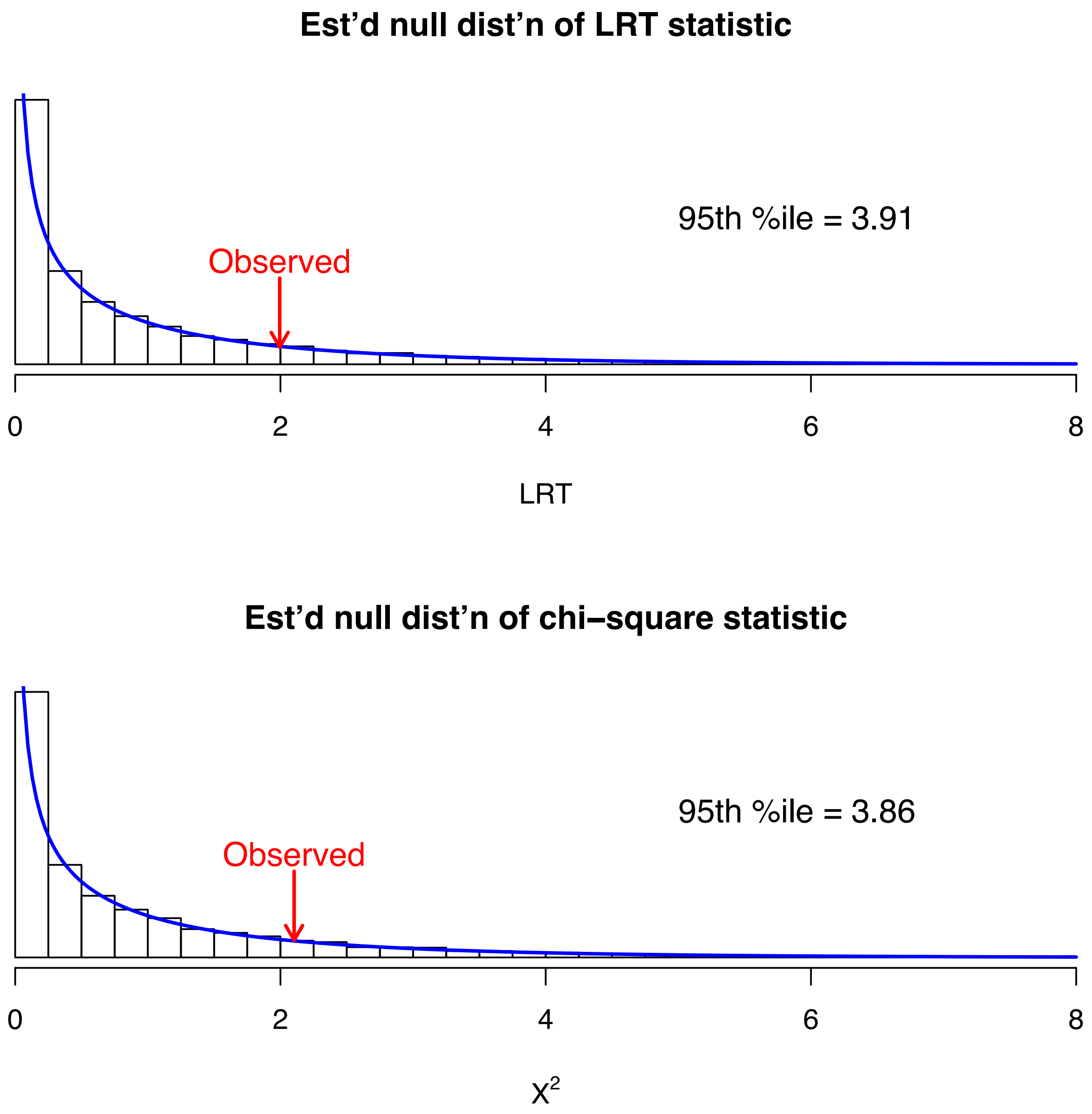
98.5	94.2	42.0	15.3

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

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

10,000 computer simulations: $P \approx 17\%$ $P \approx 15\%$

Example 2



Est'd null dist'n of LRT statistic

Est'd null dist'n of chi–square statistic

Example 3

Data on number of sperm bound to an egg:

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

⟶ Do these follow a Poisson distribution?

MLE:

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

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

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

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

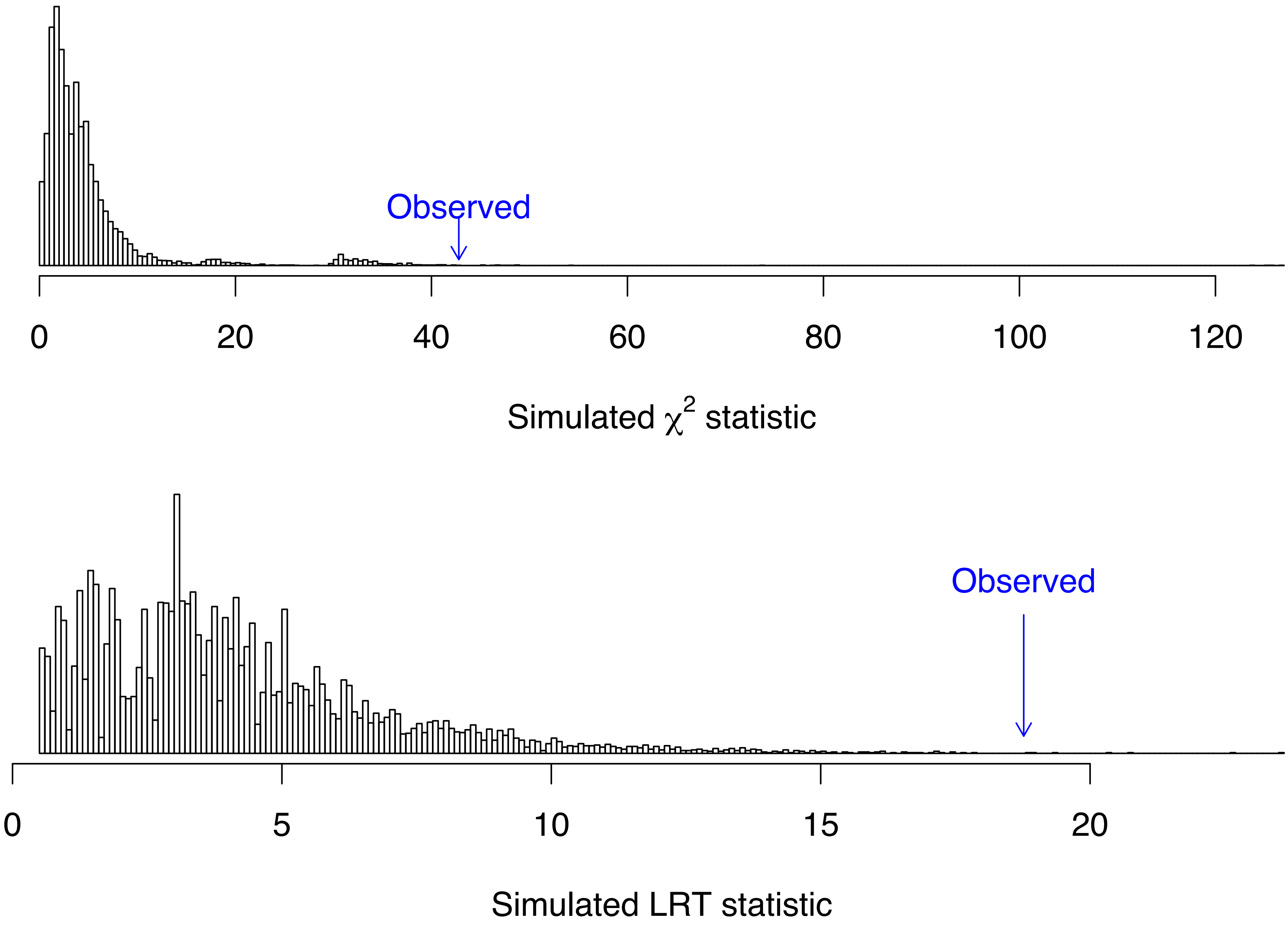
$\text{LRT} = 2 \sum \text{obs} \log(\text{obs}/\text{exp}) = \ldots = 18.8$

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?

2 x 2 tables

Apply a treatment to 20 mice from strains A and B, and observe survival.

	N	Y	
A	18	2	20
B	11	9	20
	29	11	40

Question:

⟶ Are the survival rates in the two strains the same?

Gather 100 rats and determine whether they are infected with viruses A and B.

	I-B	NI-B	
I-A	9	9	18
NI-A	20	62	82
	29	71	100

Question:

⟶ Is infection with virus A independent of infection with virus B?

Underlying probabilities

$\longrightarrow$ Observed data

$$\begin{array}{ccc|cc|c} & & & \multicolumn{2}{c}{\text{B}} & \\ & & & 0 & 1 & \\ \cline{4-5} \text{A} & 0 & & n_{00} & n_{01} & n_{0+} \\ & 1 & & n_{10} & n_{11} & n_{1+} \\ \cline{4-5} & & & n_{+0} & n_{+1} & n \end{array}$$

$\longrightarrow$ Underlying probabilities

$$\begin{array}{ccc|cc|c} & & & \multicolumn{2}{c}{\text{B}} & \\ & & & 0 & 1 & \\ \cline{4-5} \text{A} & 0 & & p_{00} & p_{01} & p_{0+} \\ & 1 & & p_{10} & p_{11} & p_{1+} \\ \cline{4-5} & & & p_{+0} & p_{+1} & 1 \end{array}$$

Model:

$$(n_{00}, n_{01}, n_{10}, n_{11}) \sim \text{Multinomial}(n, \{p_{00}, p_{01}, p_{10}, p_{11}\})$$

or

$$n_{01} \sim \text{Binomial}(n_{0+}, p_{01}/p_{0+}) \quad \text{and} \quad n_{11} \sim \text{Binomial}(n_{1+}, p_{11}/p_{1+})$$

Conditional probabilities

Underlying probabilities

$$\begin{array}{cc|cc|c} & & \multicolumn{2}{c}{B} & \\ & & 0 & 1 & \\ \cline{3-4} A & 0 & p_{00} & p_{01} & p_{0+} \\ & 1 & p_{10} & p_{11} & p_{1+} \\ \cline{3-4} & & p_{+0} & p_{+1} & 1 \end{array}$$

Conditional probabilities

$$\Pr(B = 1 \mid A = 0) = p_{01}/p_{0+}$$

$$\Pr(B = 1 \mid A = 1) = p_{11}/p_{1+}$$

$$\Pr(A = 1 \mid B = 0) = p_{10}/p_{+0}$$

$$\Pr(A = 1 \mid B = 1) = p_{11}/p_{+1}$$

⟶ The questions in the two examples are the same!

They both concern: $p_{01}/p_{0+} = p_{11}/p_{1+}$

Equivalently: $p_{ij} = p_{i+} \times p_{+j}$ for all i,j ⟶ think Pr(A and B) = Pr(A) × Pr(B).

This is a composite hypothesis!

2 x 2 table

$$\begin{array}{cc|cc|c} & & \multicolumn{2}{c}{B} & \\ & & 0 & 1 & \\ \hline A & 0 & p_{00} & p_{01} & p_{0+} \\ & 1 & p_{10} & p_{11} & p_{1+} \\ \hline & & p_{+0} & p_{+1} & 1 \end{array}$$

H_0: $p_{ij} = p_{i+} \times p_{+j}$ for all i,j

A different view

$$\boxed{p_{00} \quad p_{01} \quad p_{10} \quad p_{11}}$$

H_0: $p_{ij} = p_{i+} \times p_{+j}$ for all i,j

Degrees of freedom = 4 - 2 - 1 = 1

Expected counts

Observed data

$$\begin{array}{cc|cc|c} & & \multicolumn{2}{c}{B} & \\ & & 0 & 1 & \\ \hline A & 0 & n_{00} & n_{01} & n_{0+} \\ & 1 & n_{10} & n_{11} & n_{1+} \\ \hline & & n_{+0} & n_{+1} & n \end{array}$$

Expected counts

$$\begin{array}{cc|cc|c} & & \multicolumn{2}{c}{B} & \\ & & 0 & 1 & \\ \hline A & 0 & e_{00} & e_{01} & n_{0+} \\ & 1 & e_{10} & e_{11} & n_{1+} \\ \hline & & n_{+0} & n_{+1} & n \end{array}$$

To get the expected counts under the null hypothesis we:

- Estimate p_{1+} and p_{+1} by n_{1+}/n and n_{+1}/n, respectively.

  These are the MLEs under H_0!
- Turn these into estimates of the p_{ij}.
- Multiply these by the total sample size, n.

The expected counts

The expected count (assuming H_0) for the “11” cell is the following:

$$\begin{aligned} e_{11} &= n \times \hat{p}_{11} \\ &= n \times \hat{p}_{1+} \times \hat{p}_{+1} \\ &= n \times (n_{1+}/n) \times (n_{+1}/n) \\ &= (n_{1+} \times n_{+1})/n \end{aligned}$$

The other cells are similar.

$\longrightarrow$ We can then calculate a χ^2 or LRT statistic as before!

Example 1

Observed data

	N	Y	
A	18	2	20
B	11	9	20
	29	11	40

Expected counts

	N	Y	
A	14.5	5.5	20
B	14.5	5.5	20
	29	11	40

$$X^2 = \frac{(18-14.5)^2}{14.5} + \frac{(11-14.5)^2}{14.5} + \frac{(2-5.5)^2}{5.5} + \frac{(9-5.5)^2}{5.5} = 6.14$$

$$\text{LRT} = 2 \times \left[18 \log\left(\frac{18}{14.5}\right) + \ldots + 9 \log\left(\frac{9}{5.5}\right)\right] = 6.52$$

P-values (based on the asymptotic $\chi^2(\text{df} = 1)$ approximation):

1.3% and 1.1%, respectively.

Example 2

Observed data

	I-B	NI-B	
I-A	9	9	18
NI-A	20	62	82
	29	71	100

Expected counts

	I-B	NI-B	
I-A	5.2	12.8	18
NI-A	23.8	58.2	82
	29	71	100

$$\mathrm{X}^2 = \frac{(9-5.2)^2}{5.2} + \frac{(20-23.8)^2}{23.8} + \frac{(9-12.8)^2}{12.8} + \frac{(62-58.2)^2}{58.2} = 4.70$$

$$\mathrm{LRT} = 2 \times \left[9\log\left(\frac{9}{5.2}\right) + \ldots + 62\log\left(\frac{62}{58.2}\right)\right] = 4.37$$

P-values (based on the asymptotic $\chi^2(\mathrm{df} = 1)$ approximation):

3.0% and 3.7%, respectively.

r x k tables

	Blood type				
Population	A	B	AB	O	
Florida	122	117	19	244	502
Iowa	1781	1351	289	3301	6721
Missouri	353	269	60	713	1395
	2256	1737	367	4258	8618

⟶ Same distribution of blood types in each population?

Underlying probabilities

Observed data

$$\begin{array}{c|cccc|c}
 & 1 & 2 & \cdots & k & \\
\hline
1 & n_{11} & n_{12} & \cdots & n_{1k} & n_{1+} \\
2 & n_{21} & n_{22} & \cdots & n_{2k} & n_{2+} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
r & n_{r1} & n_{r2} & \cdots & n_{rk} & n_{r+} \\
\hline
 & n_{+1} & n_{+2} & \cdots & n_{+k} & n
\end{array}$$

Underlying probabilities

$$\begin{array}{c|cccc|c}
 & 1 & 2 & \cdots & k & \\
\hline
1 & p_{11} & p_{12} & \cdots & p_{1k} & p_{1+} \\
2 & p_{21} & p_{22} & \cdots & p_{2k} & p_{2+} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
r & p_{r1} & p_{r2} & \cdots & p_{rk} & p_{r+} \\
\hline
 & p_{+1} & p_{+2} & \cdots & p_{+k} & 1
\end{array}$$

H_0: $p_{ij} = p_{i+} \times p_{+j}$ for all i,j.

Expected counts

Observed data

	A	B	AB	O	
F	122	117	19	244	502
I	1781	1351	289	3301	6721
M	353	269	60	713	1395
	2256	1737	367	4258	8618

Expected counts

	A	B	AB	O	
F	131	101	21	248	502
I	1759	1355	286	3321	6721
M	365	281	59	689	1395
	2256	1737	367	4258	8618

Expected counts under H_0: $e_{ij} = n_{i+} \times n_{+j}/n$ for all i,j.

χ^2 and LRT statistics

Observed data

	A	B	AB	O	
F	122	117	19	244	502
I	1781	1351	289	3301	6721
M	353	269	60	713	1395
	2256	1737	367	4258	8618

Expected counts

	A	B	AB	O	
F	131	101	21	248	502
I	1759	1355	286	3321	6721
M	365	281	59	689	1395
	2256	1737	367	4258	8618

$$X^2 \text{ statistic} = \sum \frac{(\text{obs}-\text{exp})^2}{\text{exp}} = \cdots = 5.64$$

$$\text{LRT statistic} = 2 \times \sum \text{obs} \ln(\text{obs}/\text{exp}) = \cdots = 5.55$$

Asymptotic approximation

If the sample size is large, the null distribution of the χ^2 and likelihood ratio test statistics will approximately follow a

$$\chi^2 \text{ distribution with } (r - 1) \times (k - 1) \text{ d.f.}$$

In the example, df = $(3 - 1) \times (4 - 1) = 6$

$X^2 = 5.64 \quad \longrightarrow \quad P = 0.46.$

$\text{LRT} = 5.55 \quad \longrightarrow \quad P = 0.48.$

Two-locus linkage in an intercross

	BB	Bb	bb
AA	6	15	3
Aa	9	29	6
aa	3	16	13

Are these two loci linked?

General test of independence

Observed data

	BB	Bb	bb
AA	6	15	3
Aa	9	29	6
aa	3	16	13

Expected counts

	BB	Bb	bb
AA	4.3	14.4	5.3
Aa	7.9	26.4	9.7
aa	5.8	19.2	7.0

χ^2 test: $X^2 = 10.4 \longrightarrow P = 3.5\%$ (df = 4)

LRT test: $\text{LRT} = 9.98 \longrightarrow P = 4.1\%$

Fisher's exact test: P = 4.6%

A more specific test

Observed data

	BB	Bb	bb
AA	6	15	3
Aa	9	29	6
aa	3	16	13

Underlying probabilities

	BB	Bb	bb
AA	$\frac{1}{4}(1-\theta)^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{4}\theta^2$
Aa	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{2}[\theta^2+(1-\theta)^2]$	$\frac{1}{2}\theta(1-\theta)$
aa	$\frac{1}{4}\theta^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{4}(1-\theta)^2$

H_0: θ = 1/2 versus H_a: $\theta < 1/2$

Use a likelihood ratio test!

$\longrightarrow$ Obtain the general MLE of θ.

$\longrightarrow$ Calculate the LRT statistic = 2 ln $\left\{ \frac{\Pr(\text{data} \mid \hat{\theta})}{\Pr(\text{data} \mid \theta{=}1/2)} \right\}$

$\longrightarrow$ Compare this statistic to a $\chi^2(\text{df} = 1)$.

Results

	BB	Bb	bb
AA	6	15	3
Aa	9	29	6
aa	3	16	13

MLE: $\hat{\theta} = 0.359$

LRT statistic: LRT = 7.74 $\longrightarrow$ P = 0.54% (df = 1)

$\longrightarrow$ Here we assume Mendelian segregation, and that deviation from H_0 is “in a particular direction.”

$\longrightarrow$ If these assumptions are correct, we’ll have greater power to detect linkage using this more specific approach.