

Goodness of fit tests

We observe data like that in the following table:

AA	AB	BB
35	43	22

We want to know:

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

I have neglected to make precise the role of **chance** in this business.

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, \dots, n_k) , with $\sum_i n_i = n$
where n_i = no. balls drawn that were of type i .

Examples

- The binomial distribution: the case $k = 2$.
- Self a heterozygous plant, obtain 50 progeny, and use test crosses to determine the genotypes of each of the progeny.
- Obtain a random sample of 30 people from Hopkins, and classify them according to student/faculty/staff.

Multinomial probabilities

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

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

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

Example

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

$$P(X_1=35, X_2=43, X_3=22) = \frac{100!}{35! 43! 22!} 0.25^{35} 0.50^{43} 0.25^{22}$$
$$\approx 7.3 \times 10^{-4}$$

Rather brutal, numerically speaking.

The solution: take logs (and use a computer).

Goodness of fit test

We observe $(n_1, n_2, n_3) \sim \text{multinomial}(n, (p_1, p_2, p_3))$.

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:

- (a) A test statistic
- (b) The null distribution of the test statistic

Test statistics

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

LRT statistic

$$\begin{aligned} \text{LRT} &= 2 \ln \left\{ \frac{\Pr(\text{data} \mid \hat{p} = \text{MLE})}{\Pr(\text{data} \mid H_0)} \right\} \\ &= \dots = 2 \sum_i n_i \ln(n_i/n_i^0) \end{aligned}$$

χ^2 test statistic

$$\begin{aligned} \chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \sum_i \frac{(n_i - n_i^0)^2}{n_i^0} \end{aligned}$$

Null distribution of test statistic

What values of LRT (or χ^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

The brute-force method

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

This is not feasible.

Computer simulation

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

Estimated critical value = the 95th percentile of the results

Estimated P-value = 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)`.

Asymptotic approximation

Very mathematically savvy people have shown that, if the sample size, n , is large,

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

$$X^2 \sim \chi^2(k - 1)$$

Example

We observe the following data:

AA	AB	BB
35	43	22

We imagine that these are counts

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

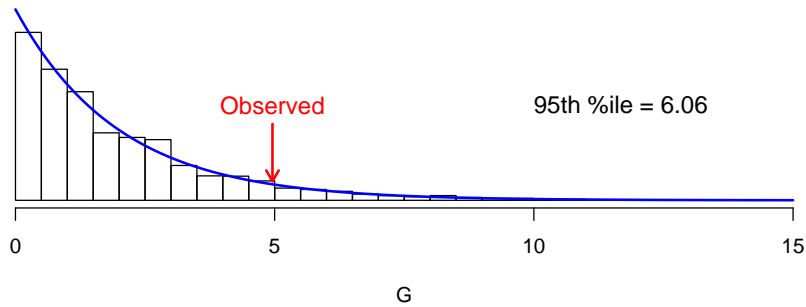
We seek to test $H_0 : p_1 = 1/4, p_2 = 1/2, p_3 = 1/4$.

We calculate $\text{LRT} \approx 4.96$ and $X^2 \approx 5.34$.

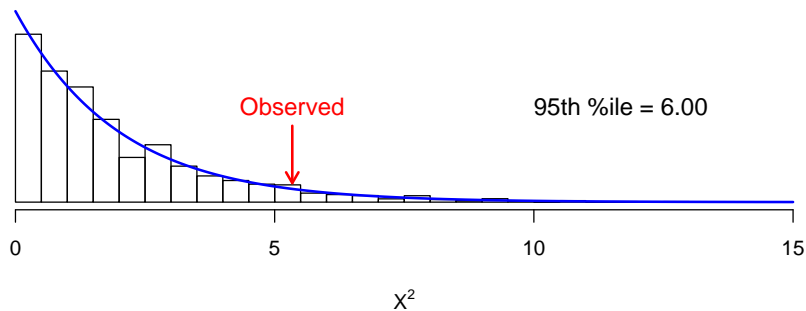
Referring to the asymptotic approximations (χ^2 dist'n with 2 degrees of freedom), we obtain $P \approx 8.4\%$ and $P \approx 6.9\%$.

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

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:

- The null distribution is approximately $\chi^2(k-1)$ if the sample size is large.
- 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 $p_{AA} = 0.25, p_{AB} = 0.5, p_{BB} = 0.25$

But rather something like:

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

For example: Genotypes, of a random sample of individuals, at a diallelic locus.

Question: 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; 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:

$$\begin{aligned} p_A &= f_A^2 + 2f_A f_O & p_{AB} &= 2f_A f_B \\ p_B &= f_B^2 + 2f_B f_O & p_O &= f_O^2 \end{aligned}$$

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

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

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

$$H_0: p_A = f_A^2 + 2f_A f_O, p_B = f_B^2 + 2f_B f_O, p_{AB} = 2f_A f_B, p_O = f_O^2 \\ \text{(for some } f_O, f_A, f_B, \text{ where } f_O + f_A + f_B = 1\text{)}$$

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

MLE under H_0 : Requires numerical optimization.

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

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
$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

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 **$s = \text{no. parameters estimated under } H_0$** .
 - Note that $s = 1$ for example 1, and $s = 2$ for example 2, and so **$df = 1$** for both examples.

Results, example 1

Example data:

AA	AB	BB
5	20	75

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

Expected counts:

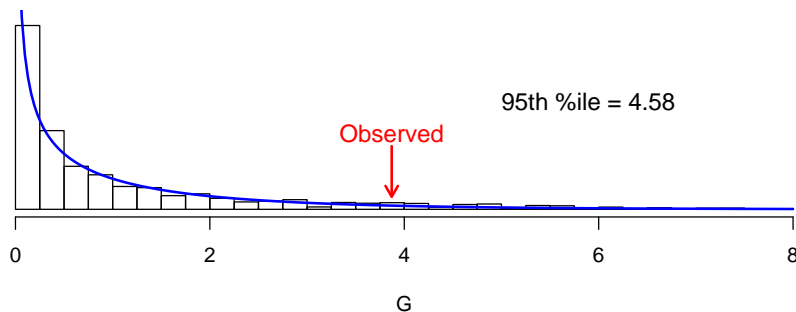
2.25	25.5	72.25
------	------	-------

Test statistics: **LRT statistic = 3.87** $\chi^2 = 4.65$

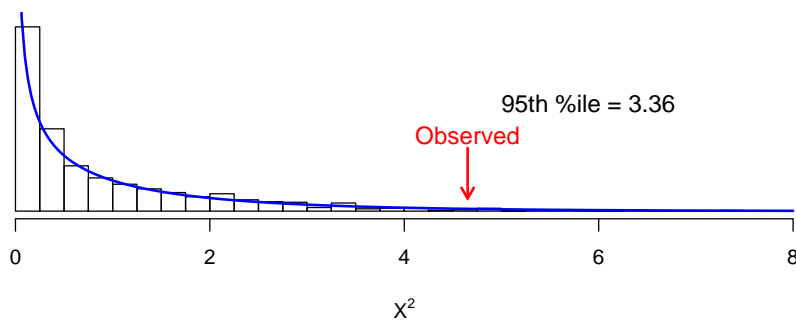
Asymptotic χ^2 (df = 1) approx'n: **P \approx 4.9%** P \approx 3.1%

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

Est'd null dist'n of LRT statistic



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



Results, 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:

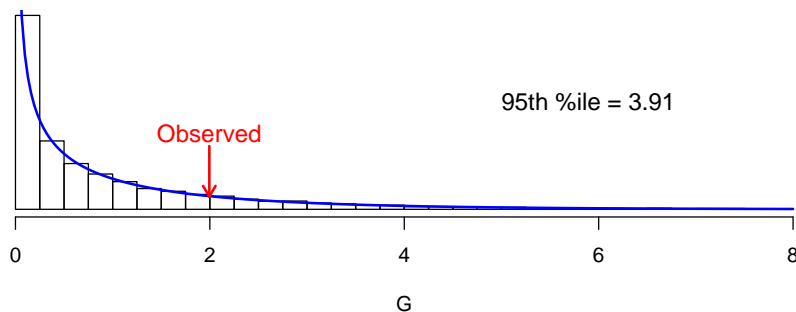
98.5	94.2	42.0	15.3
------	------	------	------

Test statistics: **LRT statistic = 1.99** $\chi^2 = 2.10$

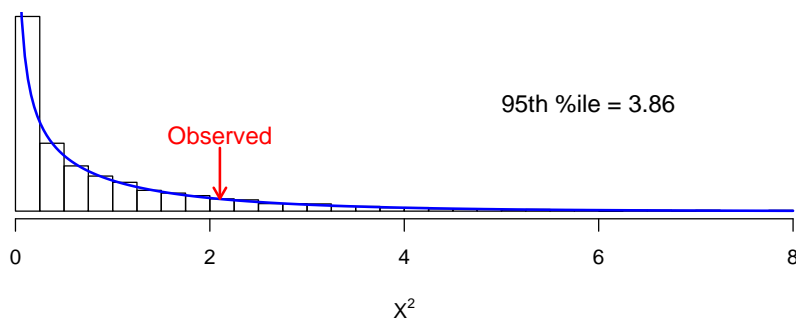
Asymptotic χ^2 (df = 1) approx'n: **P \approx 16%** **P \approx 15%**

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

Est'd null dist'n of LRT statistic



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



Example 3

Data on no. sperm bound to an egg

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

Q: Do these follow a Poisson distribution?

If $X \sim \text{Poisson}(\lambda)$, $\Pr(X=i) = e^{-\lambda} \lambda^i / i!$ where $\lambda = \text{mean}$

χ^2 and likelihood ratio tests

MLE, $\hat{\lambda} = \text{sample average}$

$$= 0 \times 26 + 1 \times 4 + 2 \times 4 + \dots + 5 \times 1 \approx 0.71$$

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

	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

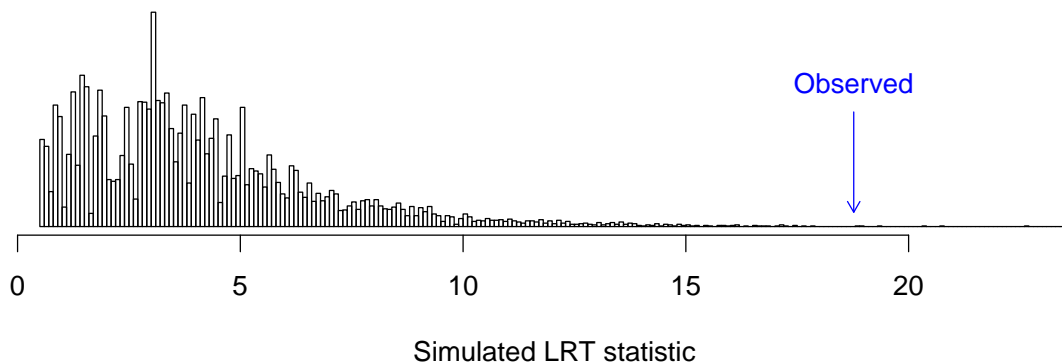
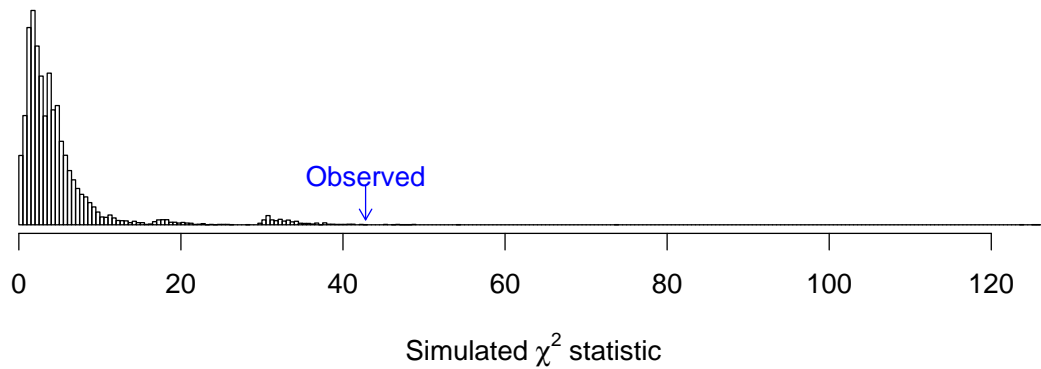
$$\chi^2 = \sum \frac{(\text{obs} - \text{exp})^2}{\text{exp}} = \dots = 42.8$$

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

Compare to $\chi^2(\text{df} = 6 - 1 - 1 = 4) \rightarrow \text{p-value} = 1 \times 10^{-8}$ (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 the 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 ?