Goodness of fit

We observe data like that in the following table:

RR	RW	WW
35	43	22

We want to know:

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

Goodness of fit

	RR	RW	WW
observed	35	43	22
expected	25	50	25

$$
\begin{aligned}
X^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\
&= \frac{(35-25)^2}{25} + \frac{(43-50)^2}{50} + \frac{(22-25)^2}{25} \\
&= 5.34
\end{aligned}
$$

`1-pchisq(5.34, 2)` $\approx$ 6.9%

Or: `chisq.test( c(35,43,22), p=c(0.25, 0.5, 0.25) )`

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:

$$p_A = f_A^2 + 2f_Af_O \qquad p_{AB} = 2f_Af_B$$

$$p_B = f_B^2 + 2f_Bf_O \qquad p_O = f_O^2$$

Example data:

O	A	B	AB
104	91	36	19

χ^2 test for these examples

- Obtain the maximum likelihood estimates (MLE) 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, the χ^2 statistic follows, approximately, a χ^2 distribution with $k - s - 1$ degrees of freedom, where s = 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

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

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

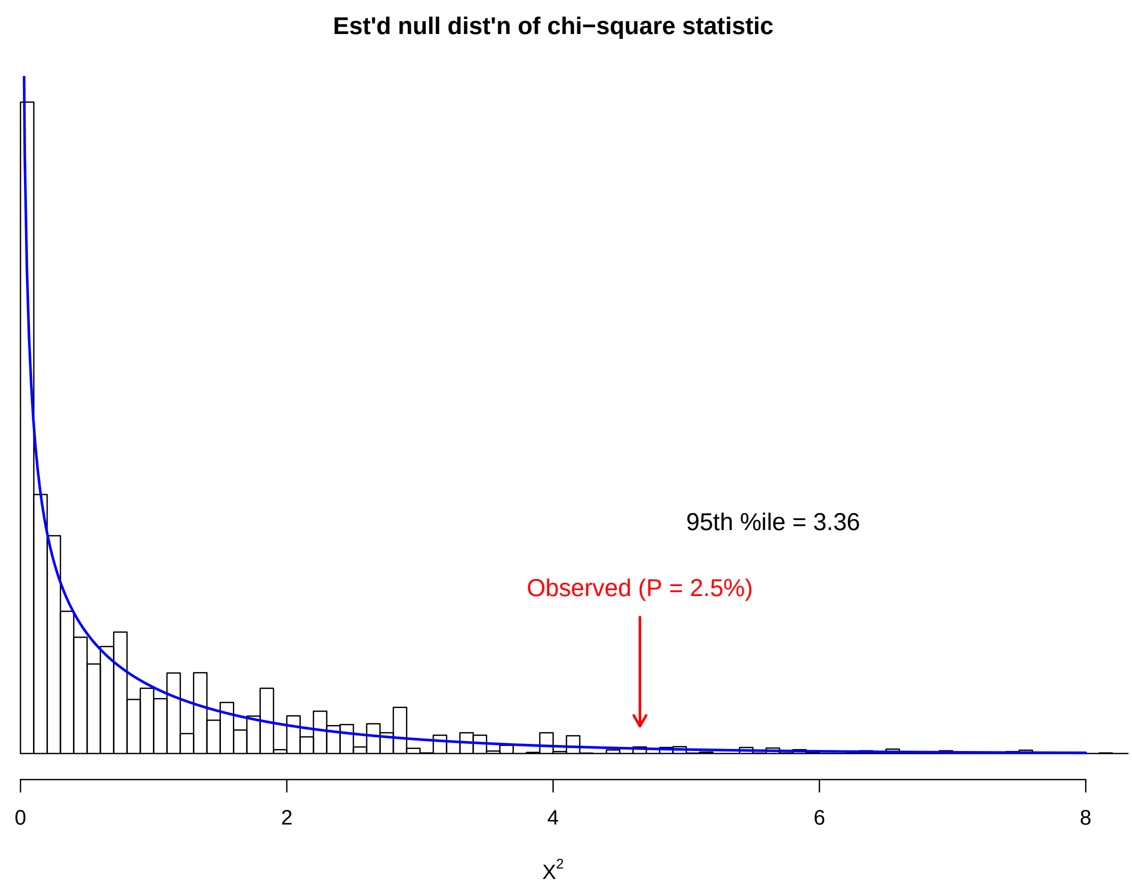
Expected counts:

2.25	25.5	72.25

Test statistics: X^2 = 4.65

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

10,000 computer simulations: $P \approx 2.5\%$



Results, example 2

Example data:

O	A	B	AB
104	91	36	19

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

MLE: $\hat{f}_O \approx 63.4\%, \hat{f}_A \approx 25.0\%, \hat{f}_B \approx 11.6\%.$

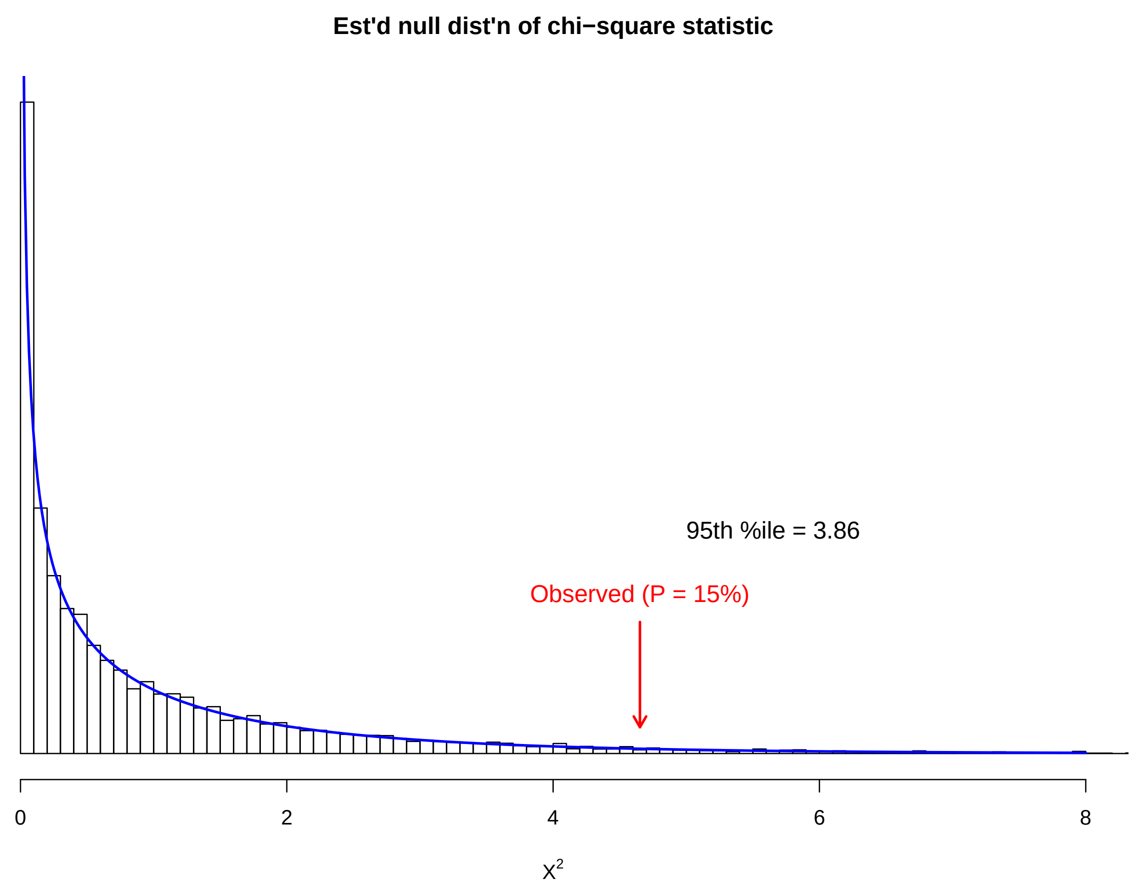
Expected counts:

100.5	94.9	40.1	14.5

Test statistics: $X^2 = 2.10$

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

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



Example 3

A scientist applied a dose of DDT to groups of 10 spider mites and counted the number of mites (out of ten) that survived. A total of 50 groups of mites were considered.

	0	1	2	3	4	5	6	7	8	9	10
count	6	10	15	7	8	1	3	0	0	0	0

Q: Does this look a binomial distribution?

If $X \sim \text{binomial}(n = 10, p)$,

$\Pr(X{=}k) = \binom{10}{k} p^k (1-p)^{10-k}$ for some p.

χ^2 test

MLE, $\hat{p}$ = (0 $\times$ 6 + 1 $\times$ 03 + 2 $\times$ 15 + . . . 10 $\times$ 0) / (50 $\times$ 10) = 0.232

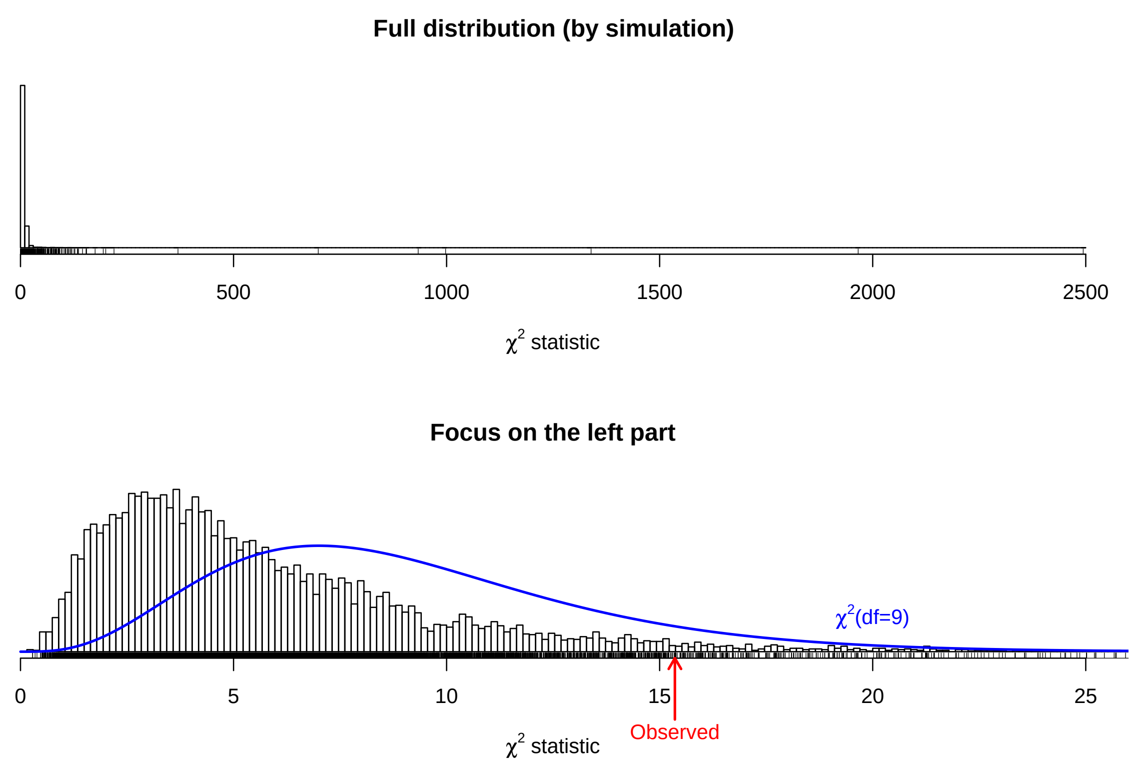
	0	1	2	3	4	5	6	7	8	9	10
observed	6	10	15	7	8	1	3	0	0	0	0
expected	3.6	10.8	14.7	11.8	6.2	2.3	0.6	0.1	~0.0	~0.0	~0.0

$$X^2 = \sum \frac{(\text{obs}-\text{exp})^2}{\text{exp}} = \frac{(6-3.6)^2}{3.6} + \frac{(10-10.8)^2}{10.8} + \frac{(15-14.7)^2}{14.7} + \cdots + \frac{(0-0)^2}{0} = 15.4$$

Compare to χ^2(df = 11 – 1 – 1 = 9) $\longrightarrow$ p-value = 0.082.

By computer simulation: p-value = 0.045

Null simulation results



Combine the rare bins

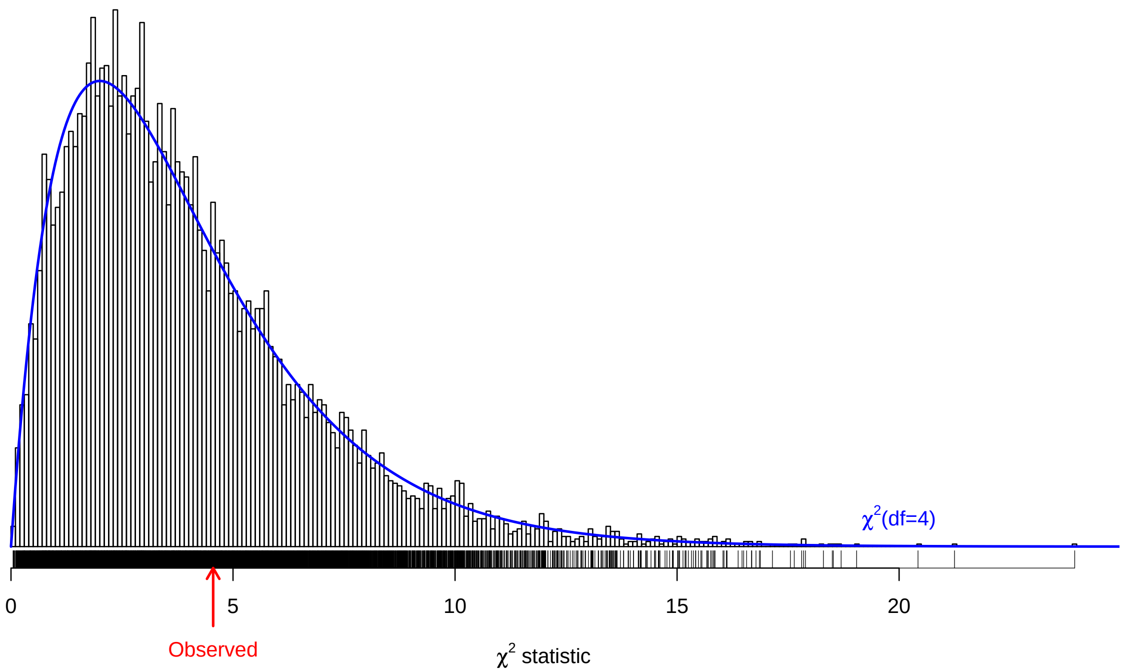
	0	1	2	3	4	≥5
observed	6	10	15	7	8	4
expected	3.6	10.8	14.7	11.8	6.2	2.9

$$X^2 = \sum \frac{(\text{obs}-\text{exp})^2}{\text{exp}} = \frac{(6-3.6)^2}{3.6} + \frac{(10-10.8)^2}{10.8} + \frac{(15-14.7)^2}{14.7} + \cdots + \frac{(4-2.9)^2}{2.9} = 4.55$$

Compare to χ^2(df = 6 – 1 – 1 = 4) $\longrightarrow$ p-value = 0.34.

By computer simulation: p-value = 0.34

Null simulation results (combining rare bins)



Back to the question

A scientist applied a dose of DDT to groups of 10 spider mites and counted the number of mites (out of ten) that survived. A total of 50 groups of mites were considered.

	0	1	2	3	4	5	6	7	8	9	10
count	6	10	15	7	8	1	3	0	0	0	0

Q: Does this look a binomial distribution?

A final note

With these sorts of goodness-of-fit tests, we are often happy when are 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?