Goodness of fit - 2 classes

R	W
78	22

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

We could use what we've learned. . .

A couple of lectures ago, we discussed several options for testing $p_R = 0.75$:

- Exact p-value
- Normal approximation
- Randomization test

Goodness of fit - 3 classes

RR	RW	WW
35	43	22

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

The χ^2 test

Back to the first example:

	R	W	total
observed	78	22	100
expected	75	25	100

Say $p_R = \Pr(R)$ and $p_W = \Pr(W) = 1 - p_R$

We want to test $H_0 : (p_R, p_W) = (3/4, 1/4)$ versus $H_a : (p_R, p_W) \neq (3/4, 1/4)$.

Consider the statistic

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

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

Null distribution

Observed counts = (n_R, n_W) with $n_R + n_W = 100$

Under the null hypothesis, $n_R \sim \text{binomial}(n = 100, p = 3/4)$

Possible values of n_R: 0, 1, 2, . . . , 100

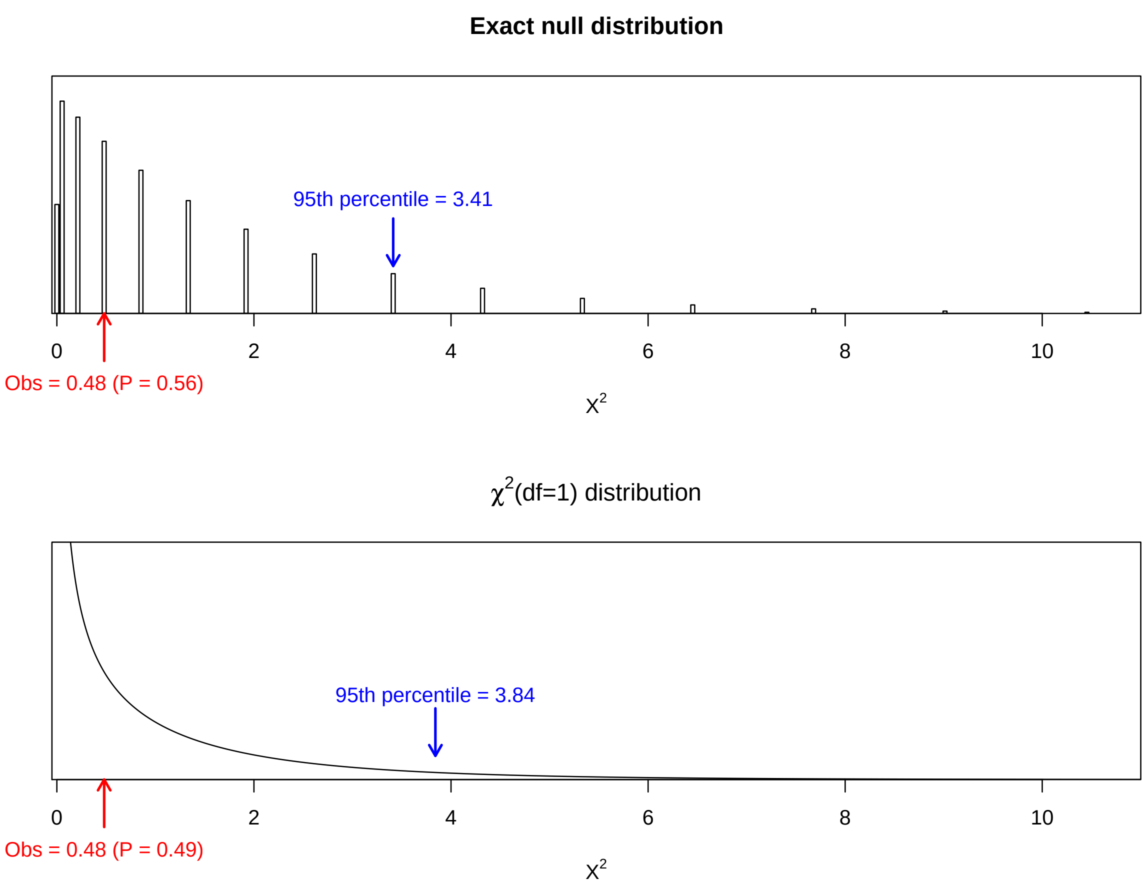
Corresponding probabilities: $\binom{100}{k}\left(\frac{3}{4}\right)^k\left(\frac{1}{4}\right)^{100-k}$

Consider the correponding values of the X^2 statistic

$\longrightarrow$ null distribution of X^2

Alternatively, use computer simulation to estimate the null distribution

Even better: for large samples, the null distribution is approximately $\chi^2(\text{df} = 1)$.



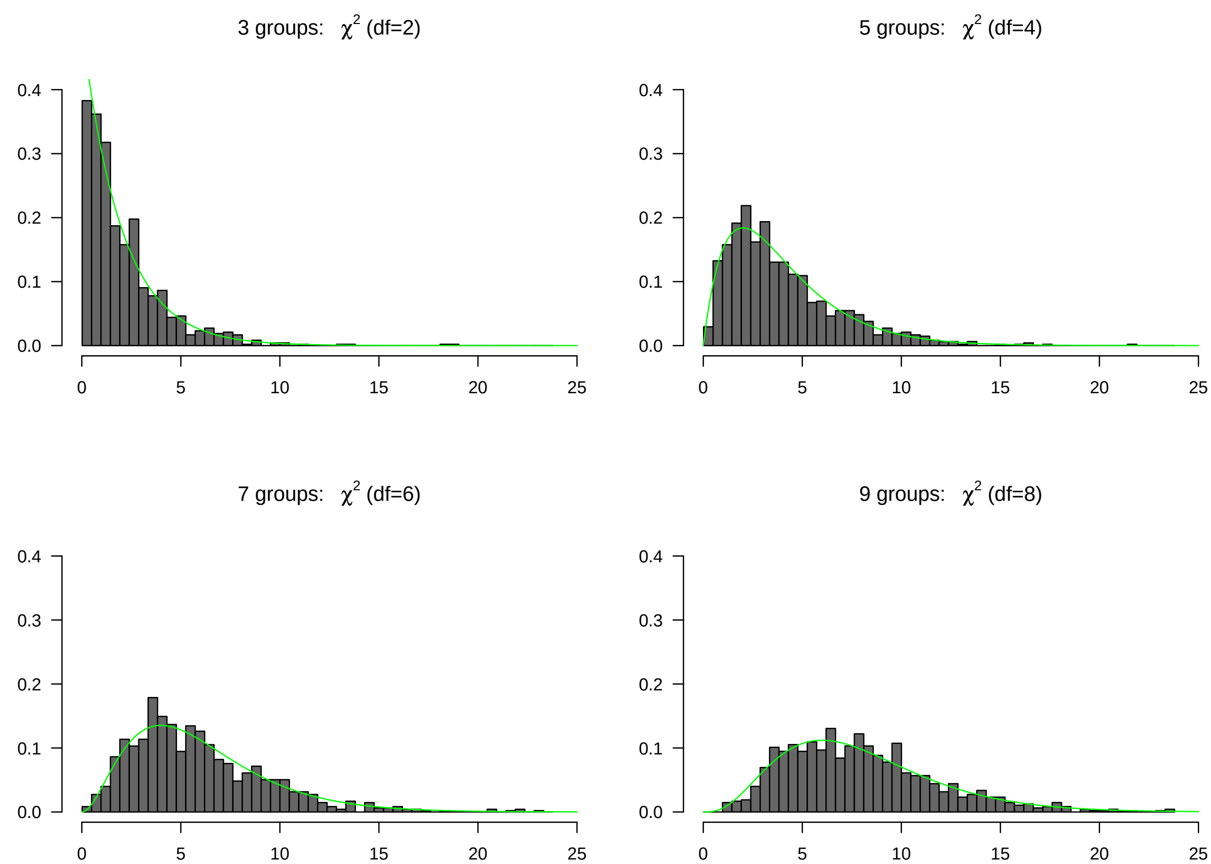
Generalization to more than two groups

If we have k groups, then the χ^2 statistic is still

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

If H_0 is true (and the sample size is large),

$$X^2 \sim \chi^2(\text{df=k}-1).$$



Our 3-group example

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?

Our 3-group example

We observe data like that in the following table:

	RR	RW	WW
observed	35	43	22
expected	25	50	25

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

$$= \frac{(35-25)^2}{25} + \frac{(43-50)^2}{50} + \frac{(22-25)^2}{25}$$

$$= 5.34$$

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

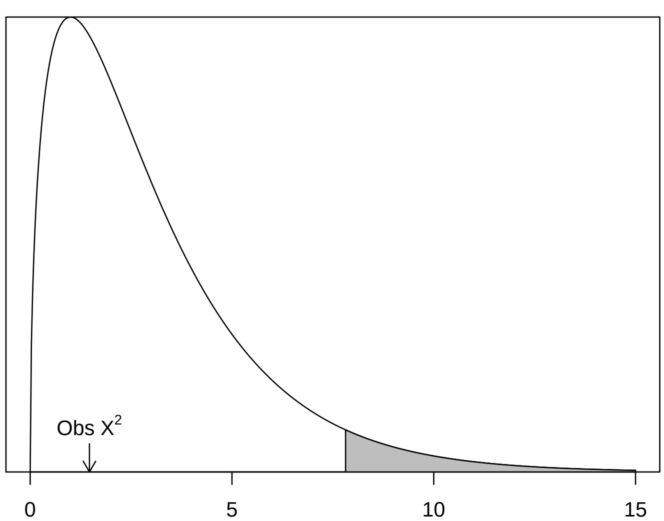
Or: `chisq.test( c(35,43,22), p=c(0.25, 0.5, 0.25) )`

Another 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	Obs	Exp	(Obs-Exp)^2/Exp
Tall, cut-leaf	926	906.2	0.43
Tall, potato-leaf	288	302.1	0.65
Dwarf, cut-leaf	293	302.1	0.27
Dwarf, potato-leaf	104	100.7	0.11
Sum	1611		1.47

Results



The χ^2 statistics is 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.

Stepping back...

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?

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: $(n_1, n_2, \ldots, n_k)$, 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 UW, and classify them according to student/faculty/staff.

Multinomial probabilities

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

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

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. 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 $(n_1, n_2, n_3) \sim$ 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 statistic

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

Null distribution of test statistic

What values of 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

The brute-force method

$$\Pr(X^2 = g \mid H_0) = \sum_{\substack{n_1,n_2,n_3 \\ \text{giving } X^2 = g}} \Pr(n_1, n_2, n_3 \mid H_0)$$

This is usually 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 = 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 mathemathically savy people have shown that,
if the sample size, n, is large,

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

Example

We observe the following data:

RR	RW	WW
35	43	22

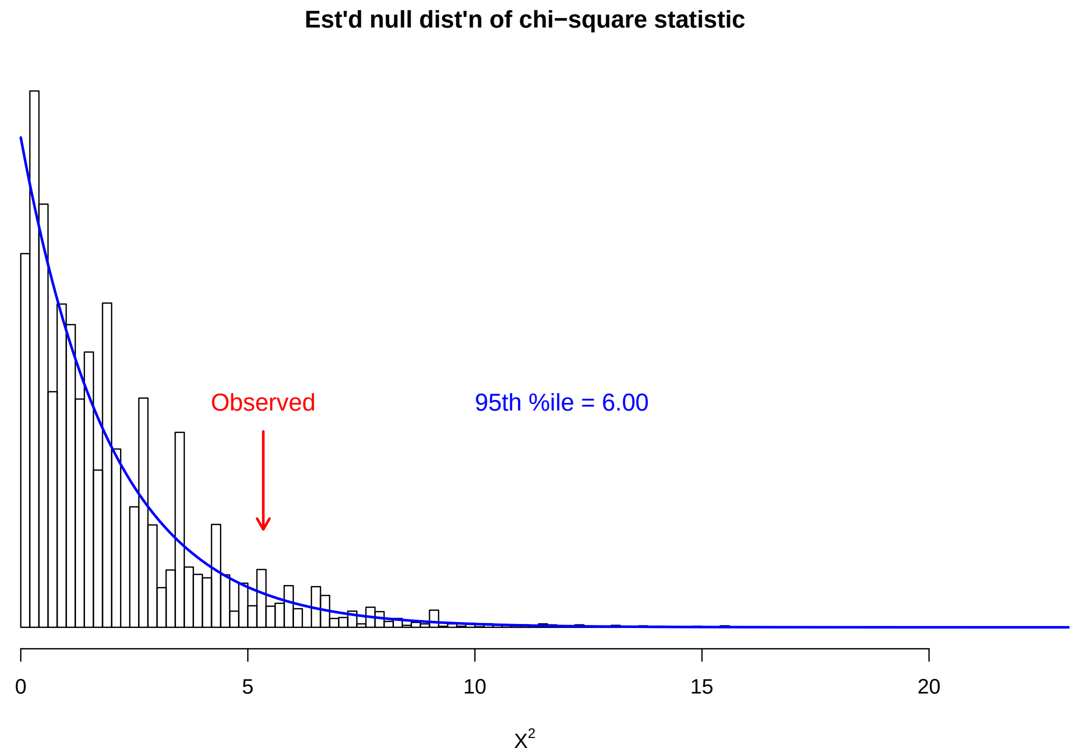
We imagine that these are counts
$(n_1, n_2, n_3) \sim$ multinomial(n=100, (p_1, p_2, p_3)).

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

We calculate $X^2 \approx 5.34$.

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

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



Summary and recommendation

For 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.