Binomials, contingency tables, chi-square tests

Joe Felsenstein

Department of Genome Sciences and Department of Biology

Confidence intervals and tests

(A continuation of the last lecture's theme:)

Connected with every confidence interval is a test: a null hypothesis is rejected if that value of the parameter(s) is outside the confidence interval.

(Correspondingly, the confidence interval is all the values that cannot be rejected by the test).

If a null hypothesis is rejected, is it, really?

- If the null hypothesis is rejected, how often does that mean that the alternative hypothesis is true?

If a null hypothesis is rejected, is it, really?

- If the null hypothesis is rejected, how often does that mean that the alternative hypothesis is true?
- *It depends on the fraction of the time we expect the alternative hypothesis to be true.* Here is an example where the alternative hypothesis has a 9% chance of being true:

	Null Hypothesis	Alternative Hypothesis
Null accepted	950	10
Null rejected	50	90
Total	1000	100

If a null hypothesis is rejected, is it, really?

- If the null hypothesis is rejected, how often does that mean that the alternative hypothesis is true?
- *It depends on the fraction of the time we expect the alternative hypothesis to be true.* Here is an example where the alternative hypothesis has a 9% chance of being true:

	Null Hypothesis	Alternative Hypothesis
Null accepted	950	10
Null rejected	50	90
Total	1000	100

- In that case when we reject the null hypothesis, most of the time the alternative hypothesis is true.

If a null hypothesis is rejected, is it, really?

- If the null hypothesis is rejected, how often does that mean that the alternative hypothesis is true?
- *It depends on the fraction of the time we expect the alternative hypothesis to be true.* Here is an example where the alternative hypothesis has a 9% chance of being true:

	Null Hypothesis	Alternative Hypothesis
Null accepted	950	10
Null rejected	50	90
Total	1000	100

- In that case when we reject the null hypothesis, most of the time the alternative hypothesis is true.
- But if the alternative hypothesis is only expected to be true 1% of the time:

	Null Hypothesis	Alternative Hypothesis
Null accepted	950	1
Null rejected	50	9
Total	1000	10

If a null hypothesis is rejected, is it, really?

- If the null hypothesis is rejected, how often does that mean that the alternative hypothesis is true?
- *It depends on the fraction of the time we expect the alternative hypothesis to be true.* Here is an example where the alternative hypothesis has a 9% chance of being true:

	Null Hypothesis	Alternative Hypothesis
Null accepted	950	10
Null rejected	50	90
Total	1000	100

- In that case when we reject the null hypothesis, most of the time the alternative hypothesis is true.
- But if the alternative hypothesis is only expected to be true 1% of the time:

	Null Hypothesis	Alternative Hypothesis
Null accepted	950	1
Null rejected	50	9
Total	1000	10

- ... then when you reject the null hypothesis, it is actually really true in most of those cases. Bummer!

If a null hypothesis is rejected, is it, really?

- If the null hypothesis is rejected, how often does that mean that the alternative hypothesis is true?
- *It depends on the fraction of the time we expect the alternative hypothesis to be true.* Here is an example where the alternative hypothesis has a 9% chance of being true:

	Null Hypothesis	Alternative Hypothesis
Null accepted	950	10
Null rejected	50	90
Total	1000	100

- In that case when we reject the null hypothesis, most of the time the alternative hypothesis is true.
- But if the alternative hypothesis is only expected to be true 1% of the time:

	Null Hypothesis	Alternative Hypothesis
Null accepted	950	1
Null rejected	50	9
Total	1000	10

- ... then when you reject the null hypothesis, it is actually really true in most of those cases. Bummer!

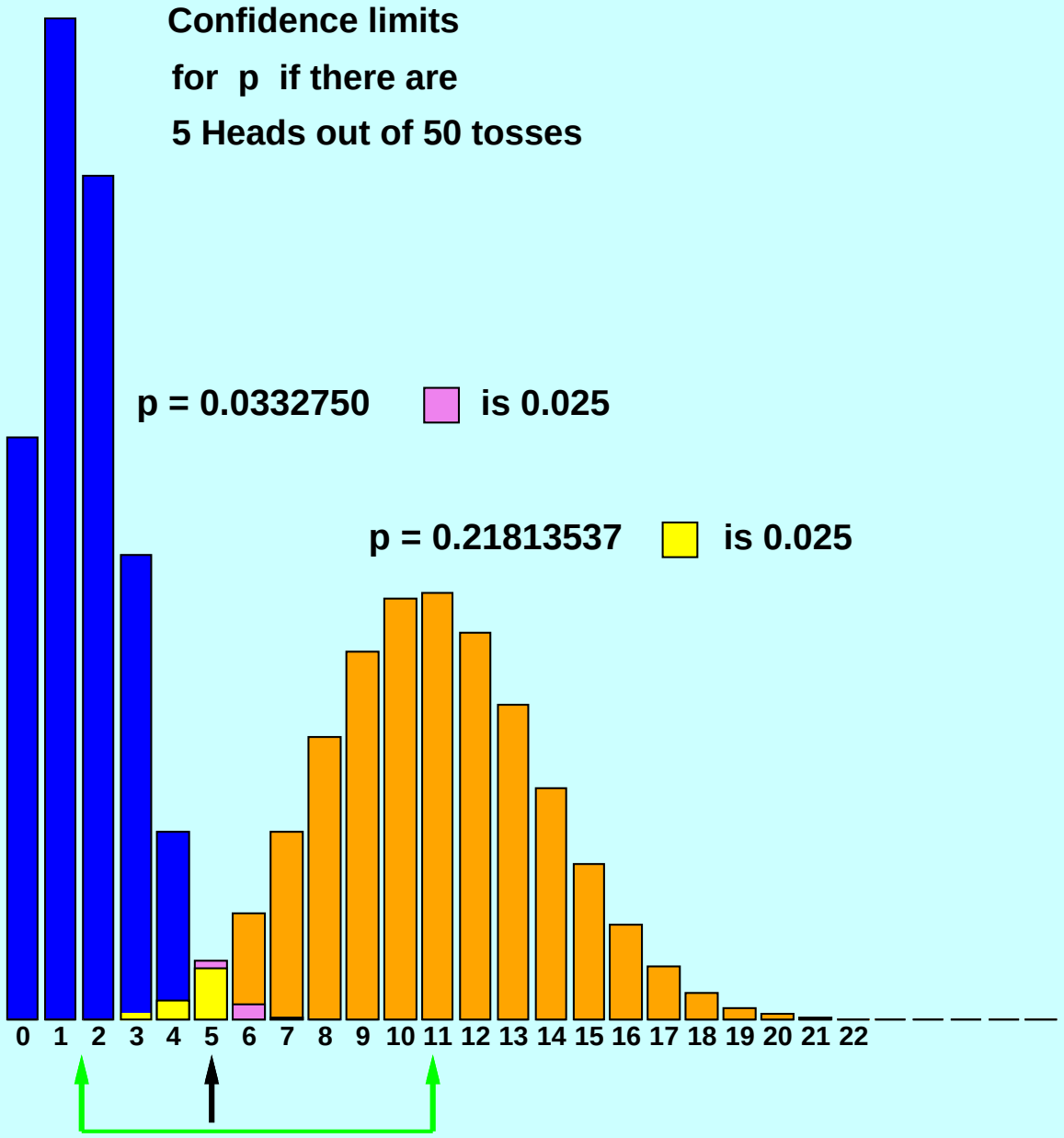
Confidence limits on a proportion

To work out a confidence interval on binomial proportions, use `binom.test`. If we want to see whether 0.20 is too high a probability of Heads when we observe 5 Heads out of 50 tosses, we use `binom.test(5, 50, 0.2)` which gives probability of 5 or fewer Heads as 0.09667, so a test does not exclude 0.20 as the Heads probability.

```
> binom.test(5,50,0.2)

        Exact binomial test
data:  5 and 50
number of successes = 5, number of trials = 50,
p-value = 0.07883
alternative hypothesis:  true probability of success
is not equal to 0.2
95 percent confidence interval:
 0.03327509 0.21813537
sample estimates:
probability of success
                   0.1
```

Confidence intervals and tails of binomials



Testing equality of binomial proportions

How do we test whether two coins have the same Heads probability?

This is hard, but there is a good approximation, the chi-square (χ^2) test. You set up a 2×2 table of numbers of outcomes:

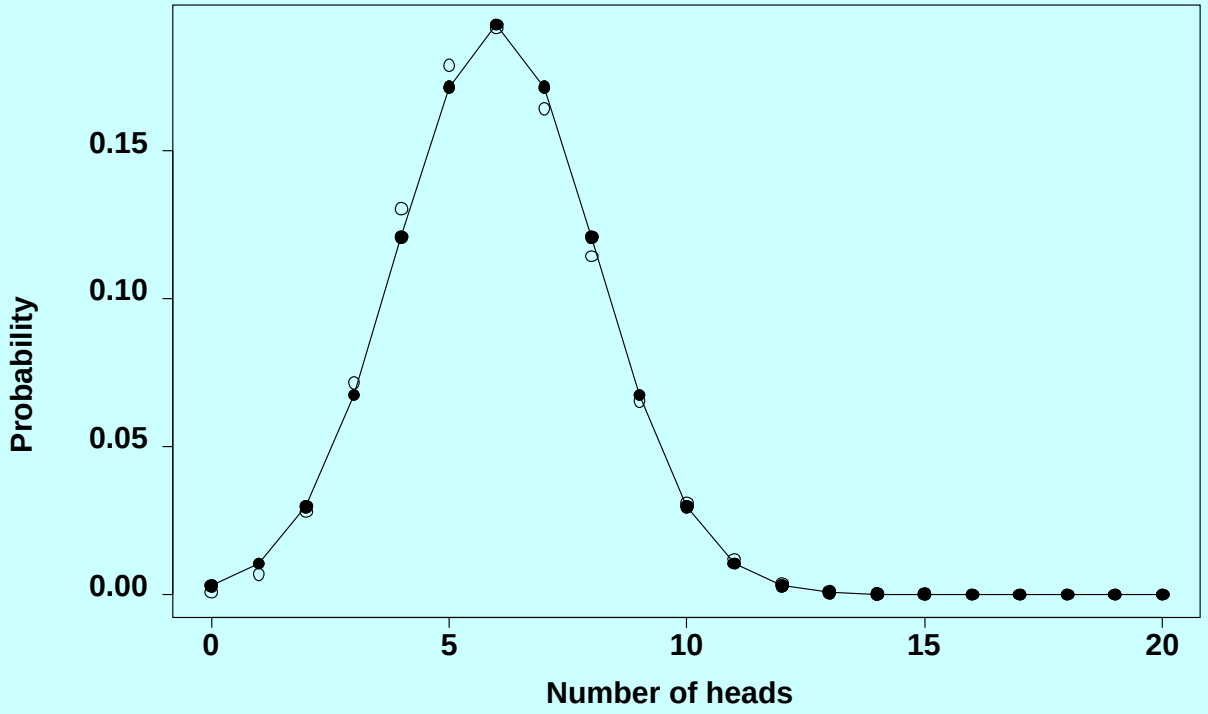
	Heads	Tails
Coin #1	15	25
Coin #2	9	31

In fact the chi-square test can test bigger tables: R rows by C columns.

There is an R function `chisq.test` that takes a matrix as an argument. (We'll see it in use in a moment).

The normal approximation

Actually, the binomial distribution is fairly well-approximated by the Normal distribution:



This shows the Binomial with 20 trials and Heads probability 0.3, the class probabilities are the open circles. For each number of Heads k, we approximate this by the area under a normal distribution with mean 20×0.3 and variance $20 \times 0.3 \times 0.7$, between $k + \frac{1}{2}$ and $k - \frac{1}{2}$ (these are shown as solid circles connected by lines).

The chi-square distribution in effect uses that approximation.

Three cases

There are really three different cases. It turns out that the same chi-square calculation tests them all!

1. We draw individuals from a population and classify them as having or not having one trait, and also as having or not having another. Neither marginal is fixed.

	twitters	does not	Total
flaky	53	28	81
not	69	50	119
Total	122	78	200

```
> tb <- matrix(c(53,28,69,50),c(2,2))
> chisq.test(tb)
   Pearson's Chi-squared test with Yates' continuity correction

data:  tb
X-squared = 0.8328, df = 1, p-value = 0.3615
```

Three cases

2. We draw a given number of sample items (say 100 and 150) in each of two cases. We observe the number of an outcome in each. Thus one marginal (the numbers of the two cases investigated) is fixed.

	Males	Females	Total
went to college	37	52	89
not	63	98	161
Total	100	150	250

```
> tb <- matrix(c(37,52,62,98),c(2,2))
> chisq.test(tb)
        Pearson's Chi-squared test with Yates' continuity
correction

data:  tb
X-squared = 0.0589, df = 1, p-value = 0.8083
```

Three cases

3. “The lady tasting tea” example by R. A. Fisher. A taste test is done where 100 cups of tea have 50 in which the milk was added after the hot water, and 50 in which milk was added before the hot water. The taster tries to decide which cups are the 50 in which the milk was added afterward.

	milk first	milk second	Total
Says milk first	31	19	50
Says milk second	19	31	50
Total	50	50	100

```
 > tb <- c(31,19,19,31)
> a <- matrix(tb,c(2,2))
> chisq.test(a)

        Pearson’s Chi-squared test with Yates’ continuity correction

data:  a
X-squared = 4.84, df = 1, p-value = 0.02781
```

(By the way, this is not a silly example – there are people who claim to be able to taste the difference, and there is some basis in physics for assuming there may be a difference).

How to do a chi-square test

First, figure out the expected numbers in each class. In a contingency table, in each cell). For an $m \times n$ contingency table this is (row sum)$\times$(column sum) / (total for whole table).

Sum over all classes:

$$\sum_{\text{classes}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

The number of degrees of freedom is the total number of classes, less one because the expected frequencies add up to 1, less the number of parameters you had to estimate. For a contingency table you in effect estimated $n-1$ column frequencies and $m-1$ row frequencies so the degrees of freedom are $nm-(n-1)-(m-1)-1$ which is $(n-1)(m-1)$. Look the value up on a chi-square (χ^2) table, which is the distribution of sums of (various numbers of) squares of normally-distributed quantities.

Testing in a one-way table

You can even use chi-square for an $R \times 1$ table, if you have some means of computing expected numbers in each cell. You will do this in your computer exercise.

The chi-square distribution

Here are some critical values of the χ^2 distribution for different numbers of degrees of freedom:

df	Upper 95% point	df	Upper 95% point
1	3.841	15	24.996
2	5.991	20	31.410
3	7.815	25	37.652
4	9.488	30	43.773
5	11.070	35	49.802
6	12.592	40	55.758
7	14.067	45	61.656
8	15.507	50	67.505
9	16.919	60	79.082
10	18.307	70	90.531

Of course you can get the correct P values computed when you use R.

Fisher's exact test

R. A. Fisher invented an exact test that can be used on 2×2 tables if the number of entries is small enough (say less than 100).

We consider only tables that have the same marginals, both the row ones and the column ones. We can rank these by how extreme they are. For example in the Lady Tasting Tea example I gave, this table

$$\begin{array}{cc} 31 & 19 \\ 19 & 31 \end{array}$$

can be shown to have probability of occurring (among all tables that have the same marginal totals, if there is no actual ability of the taster) of

$$\text{Prob}\left(\left[\begin{array}{cc} 31 & 19 \\ 19 & 31 \end{array}\right]\right) = \frac{50!\ 50!\ 50!\ 50!}{100!\ 31!\ 19!\ 19!\ 31!} = 0.00916353$$

and these are summed for all tables as extreme or more extreme (having 31, 32, 33, ... in the upper corner if a one-tailed test).

The Fisher Exact Test in R

```
> a <- matrix(c(31,19,19,31),2,2)
> a
     [,1] [,2]
[1,]   31   19
[2,]   19   31
> fisher.test(a)

        Fisher's Exact Test for Count Data

data:  a
p-value = 0.02731
alternative hypothesis:  true odds ratio is not equal
to 1
95 percent confidence interval:
 1.103007 6.465774
sample estimates:
odds ratio
  2.635069
```

Power and lumping classes

The chi-square test guards against all departures in any pattern. But this can lose power if you are really expecting some kinds of departure and not others. Intelligently collapsing classes can help.

Example: put 2 checkers on each red square of a checkerboard, and none on the black squares.

If we test this as an 8×8 contingency table, $\chi^2 = 32$, with 31 degrees of freedom. Not significant.

If we test this by grouping all red squares into one class, which has 64 checkers, and all black squares into one class, which has 0 checkers, $\chi^2 = 32$ but the degrees of freedom is only 1. Wildly significant!

Although we lose power to detect other patterns of departure by grouping, we gain greatly in ability to detect whether red and black squares have different numbers of checkers.

Goodness-of-fit tests

If the classes represent different outcomes from a distribution (either binned continuous variables or discrete variables), we can use the usual chi-square calculation if we have expectations for each class.

These can be computed by adjusting parameters. One way that is valid is to adjust them to minimize the chi-square. The number of parameters you adjusted has to be subtracted from the degrees of freedom.

Classes with expectations less than 1 should be combined – they may cause inaccuracy of the chi-square approximation. The degrees of freedom is correspondingly reduced.

Nonindependent points

If we collect points that are not independent, this can cause trouble.

Example: surveying proportion of smokers while collecting multiple members of each family, who may all tend to smoke or all not.

If, for example, we count each point twice (for example), the χ^2 value is then twice what it should be. This will exaggerate significance, of courses.

An after-the-fact method

A method (which I will not justify in detail, but works):

We are allowed to collapse rows and columns of a contingency table, *even after the fact, even to maximize our chi-square* until we end up with a 2×2 table.

We can conservatively correct for the collapsing and the after-the-factness by testing not with 1 degrees of freedom, but with $\mathrm{rows} + \mathrm{cols} - 3$.