week 7 day 1

"Exact testing for model/data fit for log-linear models"
"Part Two"
"Algebraic & Geometric Methods in Statistics"

Sonja Petrović
Created for Math/Stat 561

Feb 20, 2023.

Agenda

- Chapter 9 from our textbook: Fisher's exact test
- Part of chapter 8, as we may need the cone of sufficient statistics.

Goals

- LAST LECTURE:
  - Understand hypotheses testing for model/data fit
- THIS LECTURE: we will work towards
  - What is a p-value for a goodness-of-fit test?
  - Asymptotic vs. exact tests
  - Fisher's test and example
- NEXT LECTURE:
  - General goodness of fit test for log-linear models
  - Open problems and relation to projects!

Recap

Exact test (Fisher)

In an **exact** goodness-of-fit test, one uses the exact distribution of the statistic. . .

Recap

Exact test (Fisher)

In an **exact** goodness-of-fit test, one uses the exact distribution of the statistic. . . . . . which is **what**?

```
        gender
range     M  F Nb
  <=135K  8  1  4
  > 135K  2  9  2
```

```
        gender
range     M  F Nb
  <=135K  9  0  4
  > 135K  1 10  2
        gender
range     M  F Nb
  <=135K  9  1  3
  > 135K  1  9  3
```

Conclusion? Evidence in the data? Significance?

Definition [p-value]

Refer to Chapter 5. Discuss in lecture / board.

- Read the beginning of Chapter 9. Section 9.1: Conditional inference.
  - We are *conditioning* on the row and column sums of the table.
  - These are sufficient statistics for the independence model.
  - This is a *general strategy*. . .

The general exact test for contingency tables [board lecture]

- Proposition 9.1.1. [stated without proof, but it's not difficult. . .]
- p.192 "A similar strategy is based on the likelihood ratio test, where we use the G statistic, instead of the X2 statistic."
- Look back to the example from Lecture 10:

Interpret: what are all the possible tables? What is the probability of any given table?

	M	F	T/Nb	totals
$\leq$ 135K	?	?	?	**13**
$>$ 135K	?	?	?	**13**
totals	**10**	**10**	**6**	**26**

Here's a cheat sheet:

Before we proceed with the Fisher test, we first introduce some notations. We represent the cells by the letters *a, b, c* and *d*, call the totals across rows and columns *marginal totals*, and represent the grand total by *n*. So the table now looks like this:

	Men	**Women**	*Row Total*
Studying	***a***	***b***	*a + b*
Non-studying	***c***	***d***	*c + d*
Column Total	*a + c*	*b + d*	*a + b + c + d (=n)*

Fisher showed that conditional on the margins of the table, *a* is distributed as a hypergeometric distribution with *a+c* draws from a population with *a+b* successes and *c+d* failures. The probability of obtaining such set of values is given by:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{n}{b+d}} = \frac{(a+b)!\;(c+d)!\;(a+c)!\;(b+d)!}{a!\;\;b!\;\;c!\;\;d!\;\;n!}$$

where $\binom{n}{k}$ is the binomial coefficient and the symbol ! indicates the factorial operator. This can be seen as follows. If the marginal totals (i.e. $a+b$, $c+d$, $a+c$, and $b+d$) are known, only a single degree of freedom is left: the value e.g. of a suffices to deduce the other values. Now, $p = p(a)$ is the probability that a elements are positive in a random selection (without replacement) of $a+c$ elements from a larger set containing n elements in total out of which $a+b$ are positive, which is precisely the definition of the hypergeometric distribution.

Figure 1: From Wikipedia :)

Resources & License

- Quick summary notes about p-values that I wrote for Stat 514.
- Read about hypothesis tests for context of the model fitting tests in these lecture notes.
- This lesson from Penn State online offers a one-page summary of Fisher's exact test for 2×2 tables, as it was developed by Sir Fisher!
- Believe it or not, there is a great 2×2 example on Wikipedia, a page which actually contains a really good explanation for this one example.

This document is created for Math/Stat 561, Spring 2023.