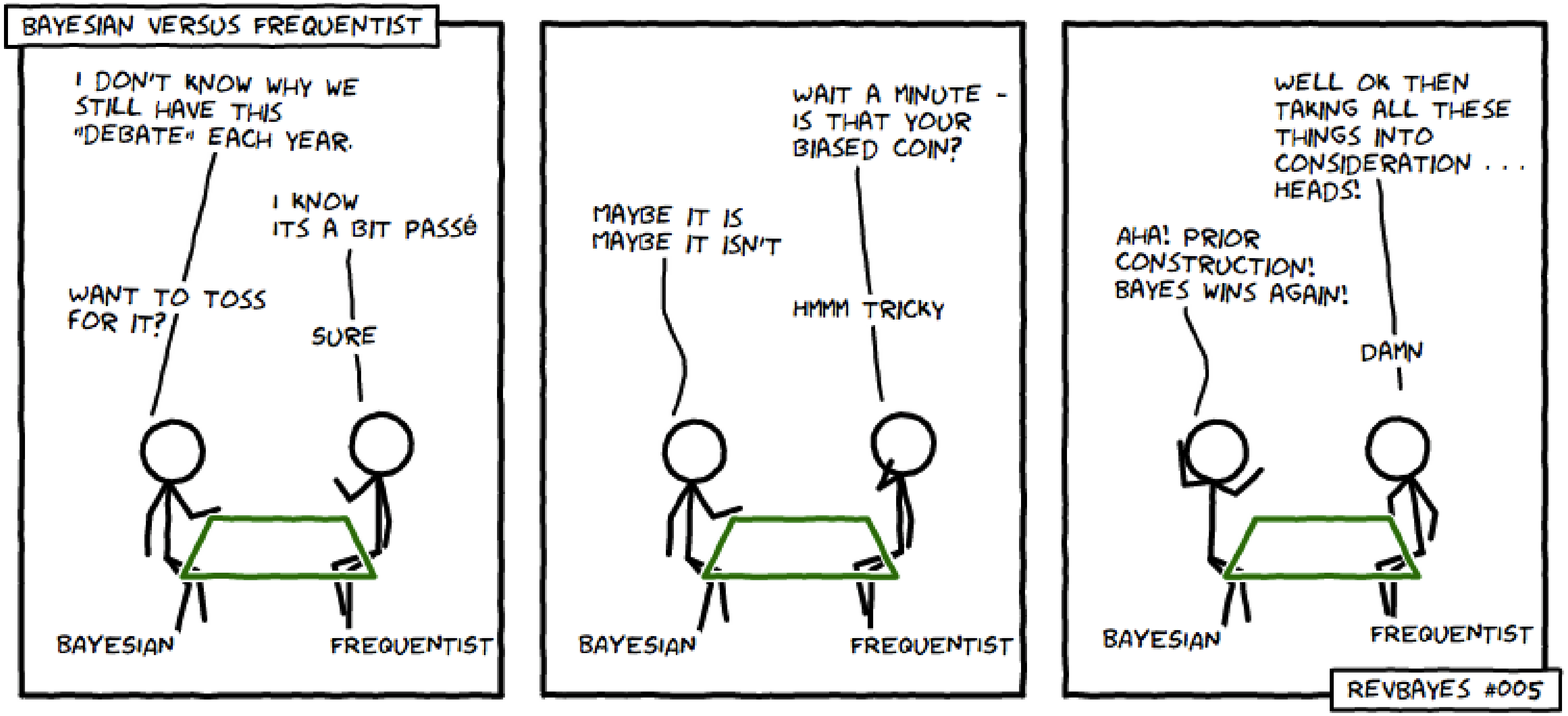
Statistically Speaking: *Contingency Tables*

DAVID L. TABB, PH.D.

OCTOBER 5, 2017

Frequentists work from data, not from expectations.



https://twitter.com/revbayes/

Overview

- Defining contingency tables and Chi-square
- Comparing the binomial and hypergeometric distributions
- Fisher Exact Test

Counts and Contingency

- The Student T-Test expects continuously distributed values, but count data are not!
- Counts often tie, and they have a low bound of zero. A zero may be a measurement.
- A contingency table, also called a cross-tab, shows how samples divide across categories.

USA causes of deaths by gender, ages 25-34 **worldlifeexpectancy.com**

Sex	Poisoning	Suicide	Road Traffic Accidents	Homicide	Other Injuries	Endocrine Disorders	**TOTAL (this table)**
Female	2651	1347	1460	682	230	428	**6798**
Male	6683	5222	4518	3477	1012	590	**21502**
TOTAL	**9334**	**6569**	**5978**	**4159**	**1242**	**1018**	**28300**

Computing expected counts

- Figure proportions for rows and columns, independent of other axis.
- Compute the product of these proportions and multiply by total.

		Poisoning	Suicide	Road Traffic Accidents	Homicide	Other Injuries	Endocrine Disorders
		33%	23%	21%	15%	4%	4%
Female	24%	8%	6%	5%	4%	1%	1%
Male	76%	25%	18%	16%	11%	3%	3%

Sex	Poisoning	Suicide	Road Traffic Accidents	Homicide	Other Injuries	Endocrine Disorders	**TOTAL (this table)**
Female	2242	1578	1436	999	298	245	**6798**
Male	7092	4991	4542	3160	944	773	**21502**
TOTAL	**9334**	**6569**	**5978**	**4159**	**1242**	**1018**	**28300**

Chi-Square test for independence

- Do observed data correspond to expected values?
- $\chi^2 = \sum \frac{(o-e)^2}{e}$
- $df = (rows - 1) * (cols - 1)$

- chisq.test(Mortality)

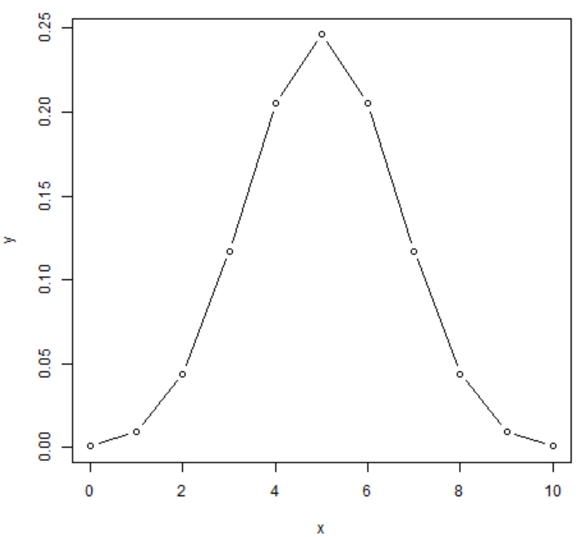
```
Pearson's Chi-squared test

data:  Mortality

X-squared = 477.34,
df = 5,
p-value < 2.2e-16
```

Binomial distribution

- Requirements:
  - experiment is *n* trials
  - only two possibilities
  - each trial has same success probability
  - trials are independent

```
x <- 0:10
y <- dbinom(x,
  size=10,prob=0.5)
```

- What is the probability (*y*) of getting *x* "heads" in ten coin flips?



http://www.stat.yale.edu/Courses/1997-98/101/binom.htm

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Hypergeometric distribution

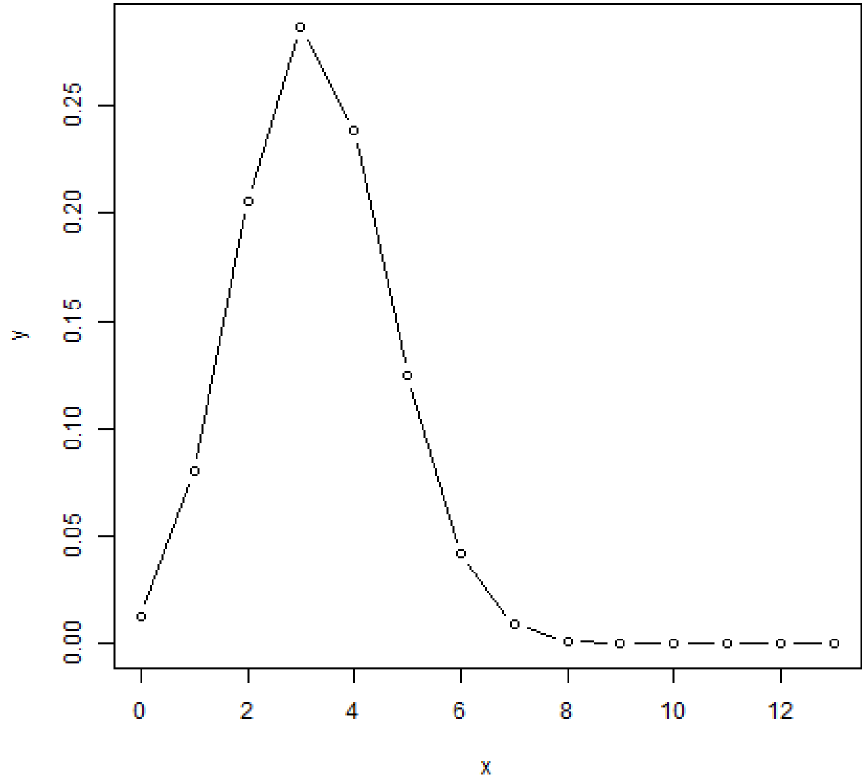
- We need HGD when replacement is not in effect (drawing a hand of cards, handful of marbles from a jar, etc.).
- $P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$

- K: # of successes in population (deck, jar)
- k: # of successes in set we sampled from pop.
- N: # of all items in population (deck, jar)
- n: # of items we sampled from pop.

How many spades am I likely to be dealt in a hand?

```
x <- 0:13
y <- dhyper(x,m=13,
         n=39,k=13)
```

- m="successes" in deck
- n="failures" in deck (52 cards – 13 spades)
- k=cards in hand



Fisher’s Exact Test

- Dr. Muriel Bristol, an algae specialist at Rothamsted, declined a cup of tea because the milk had been added to the tea rather than vice versa.
- RA Fisher created a test to evaluate whether or not she could really tell the difference: Prepare eight cups of tea, where four had tea poured first and four had milk poured first. Could she correctly separate them?

https://io9.gizmodo.com/how-a-tea-party-turned-into-a-scientific-legend-1706697488

A null hypothesis and many combinations

- With four milk-first and four tea-first cups to choose from, Dr. Bristol could pick 70 different combinations of cups (eight-choose-four). Only one answer is correct.
- Null hypothesis: Dr. Bristol has no ability to discriminate between milk-first and tea-first.

https://en.wikipedia.org/wiki/Lady_tasting_tea

Tea-Tasting Distribution Assuming the Null Hypothesis

Success count	Permutations of selection	Number of permutations
0	oooo	1 × 1 = 1
1	ooox, ooxo, oxoo, xooo	4 × 4 = 16
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	6 × 6 = 36
3	oxxx, xoxx, xxox, xxxo	4 × 4 = 16
4	xxxx	1 × 1 = 1
Total		70

Contingency table to p-value

- All correct: p=0.0143

	Truly tea-first	Truly milk-first
Judged tea-first	4	0
Judged milk-first	0	4

- 3 correct: p=0.2429

	Truly tea-first	Truly milk-first
Judged tea-first	3	1
Judged milk-first	1	3

```
Tasting <- matrix(
  c(4,0,0,4), nrow=2,
  dimnames=list(
  Judgment=c(
  "Milk","Tea"),
  Truth=c("Milk","Tea")))

fisher.test(Tasting,
  alternative="greater")
```

Takeaways

- We like replicates for estimating variance, but statistics in the absence of replicates are still possible.
- Binomial and hypergeometric distributions are key to interpreting two-outcome trials.
- Fisher's Exact Test is a useful tool to decide whether or not decision-making is random.