Modeling categorical relationships

Stats 60/Psych 10
Ismael Lemhadri
Summer 2020

Last time

- Null Hypothesis Statistical Testing (NHST)
- Confidence Intervals (CI)
- The connection between NHST and CI

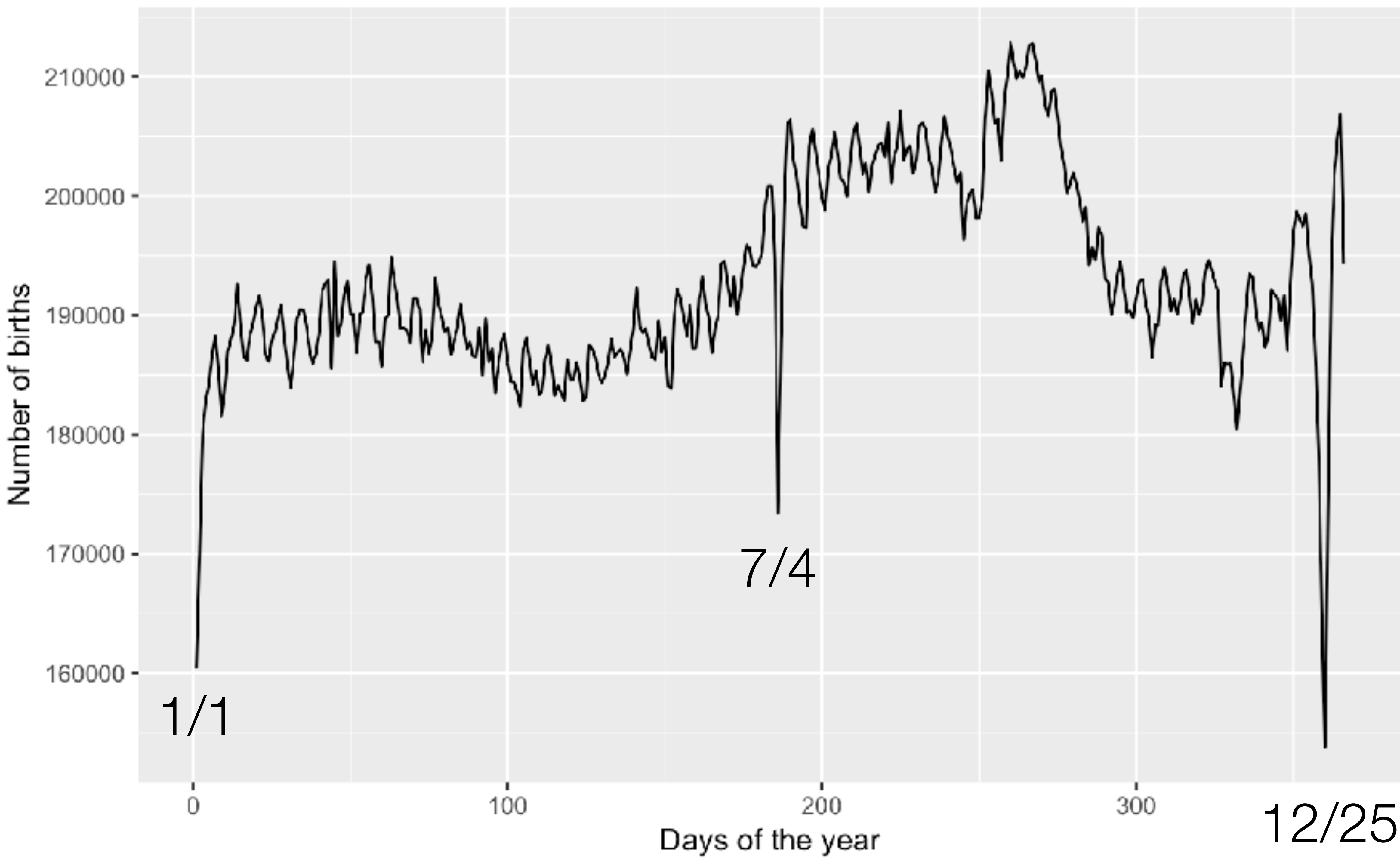
This time

- Modeling categorical relationships
  - contingency tables
  - chi-squared test for goodness of fit
  - Odds ratio

What is a “categorical relationship”?

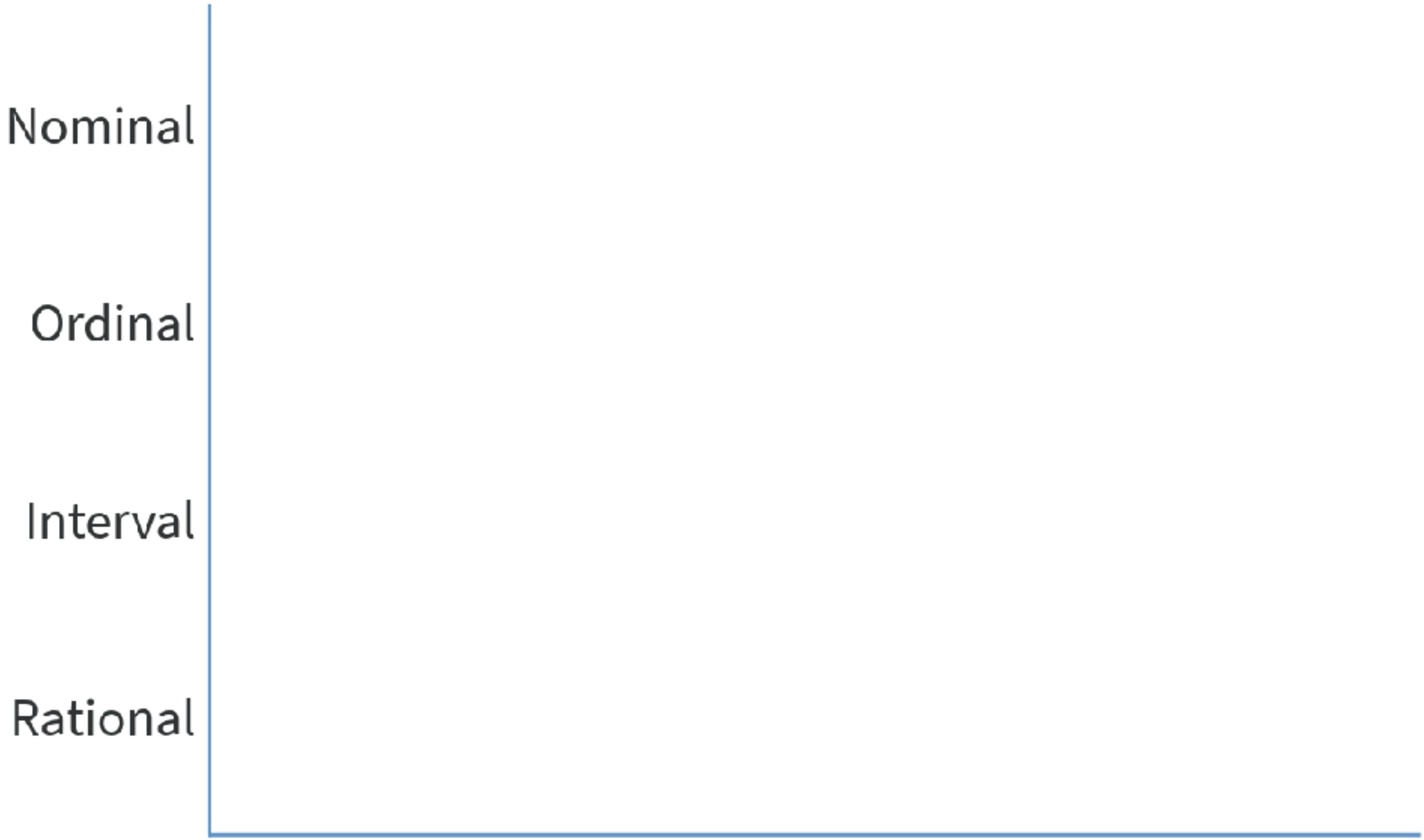
- A relationship between categorical variables
  - Variables on a nominal or (sometimes) ordinal scale
- Usually expressed in terms of counts
  - How many observations fall into each level of the variable?
    - or each combination of levels across variables?

Are births more common on certain days than others?



data from http://chmullig.com/2012/06/births-by-day-of-year/

What kind of variable is the day of the year (recorded as a number, 1-365)?

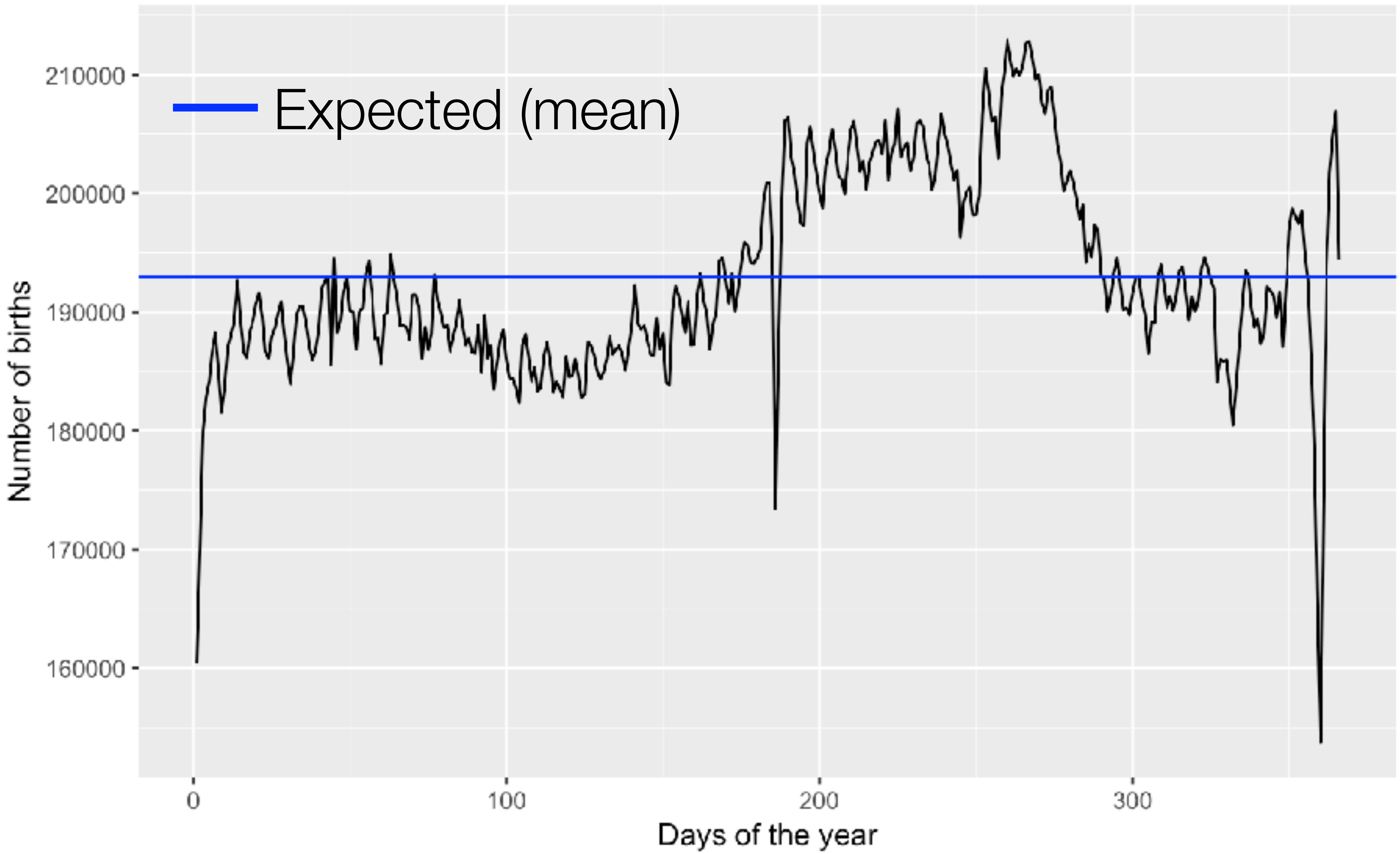


Pearson's chi-squared test for goodness of fit

$$\chi^2 = \sum_{i,j} \frac{(observed_{ij} - expected_{ij})^2}{expected_{ij}}$$

- Compare the observed data to the expected data
  - H_0: birth rates on all days are equal
  - H_A: birth rates differ between days
- If births are equally likely on all days, then the expected value for each day is just the mean number of births per day across the entire year

$$\chi^2 = \sum_{i,j} \frac{(observed_{ij} - expected_{ij})^2}{expected_{ij}} = 132764.2$$



The chi-squared distribution

- Chi-squared distribution describes the distribution of the sum of squares of a set of standard normal random variables
    - with degrees of freedom (df) equal to the number of variables being summed

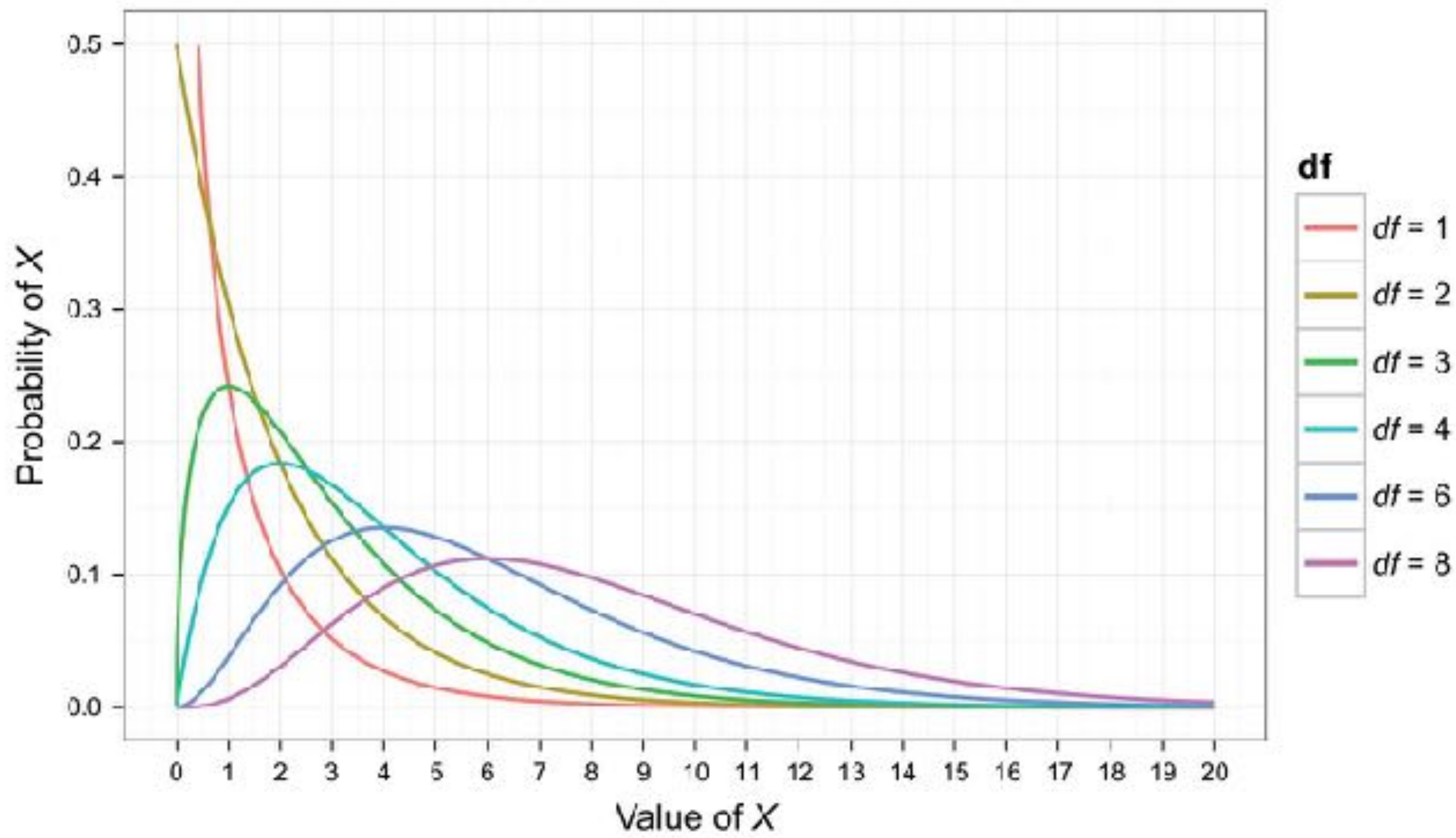
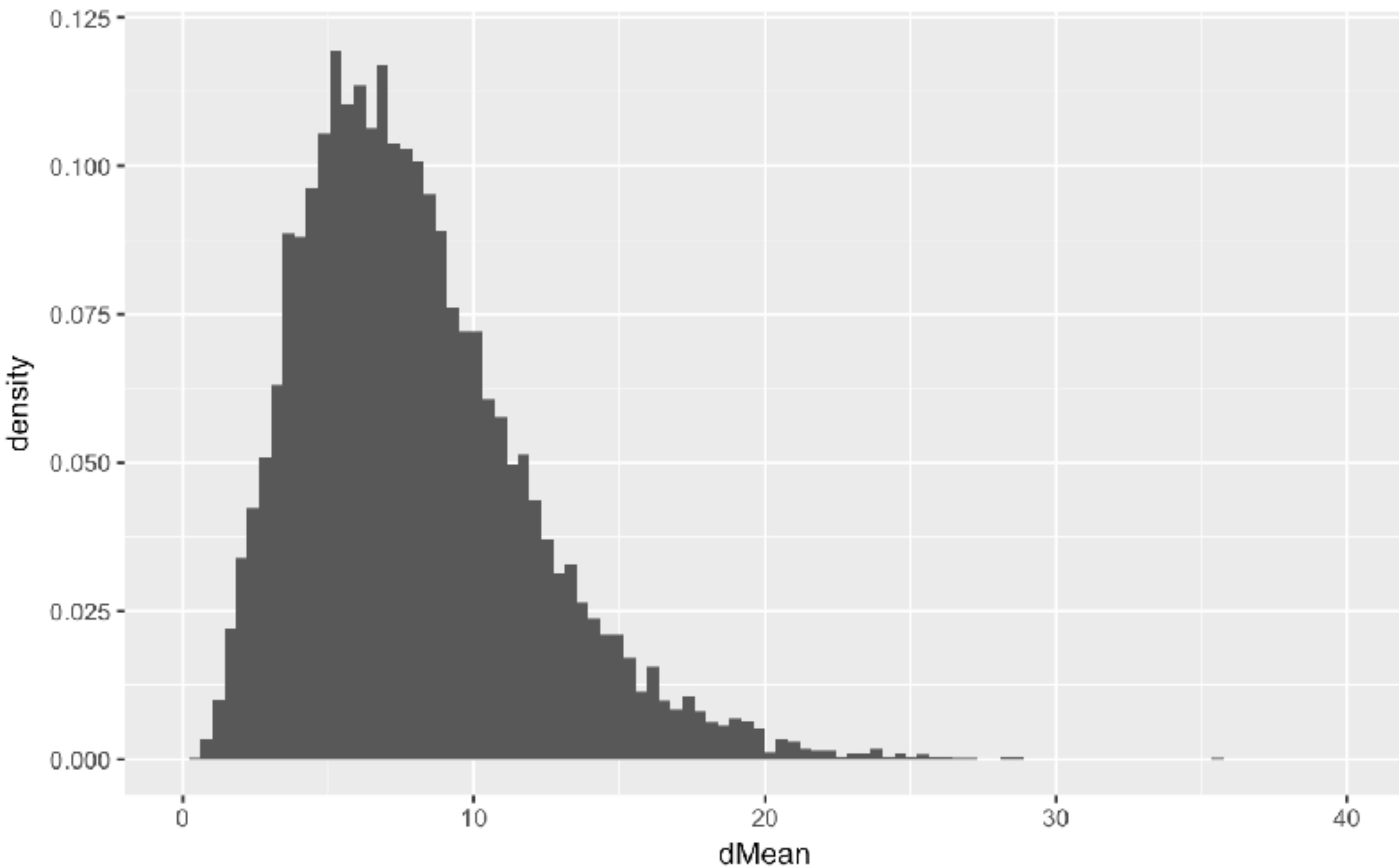


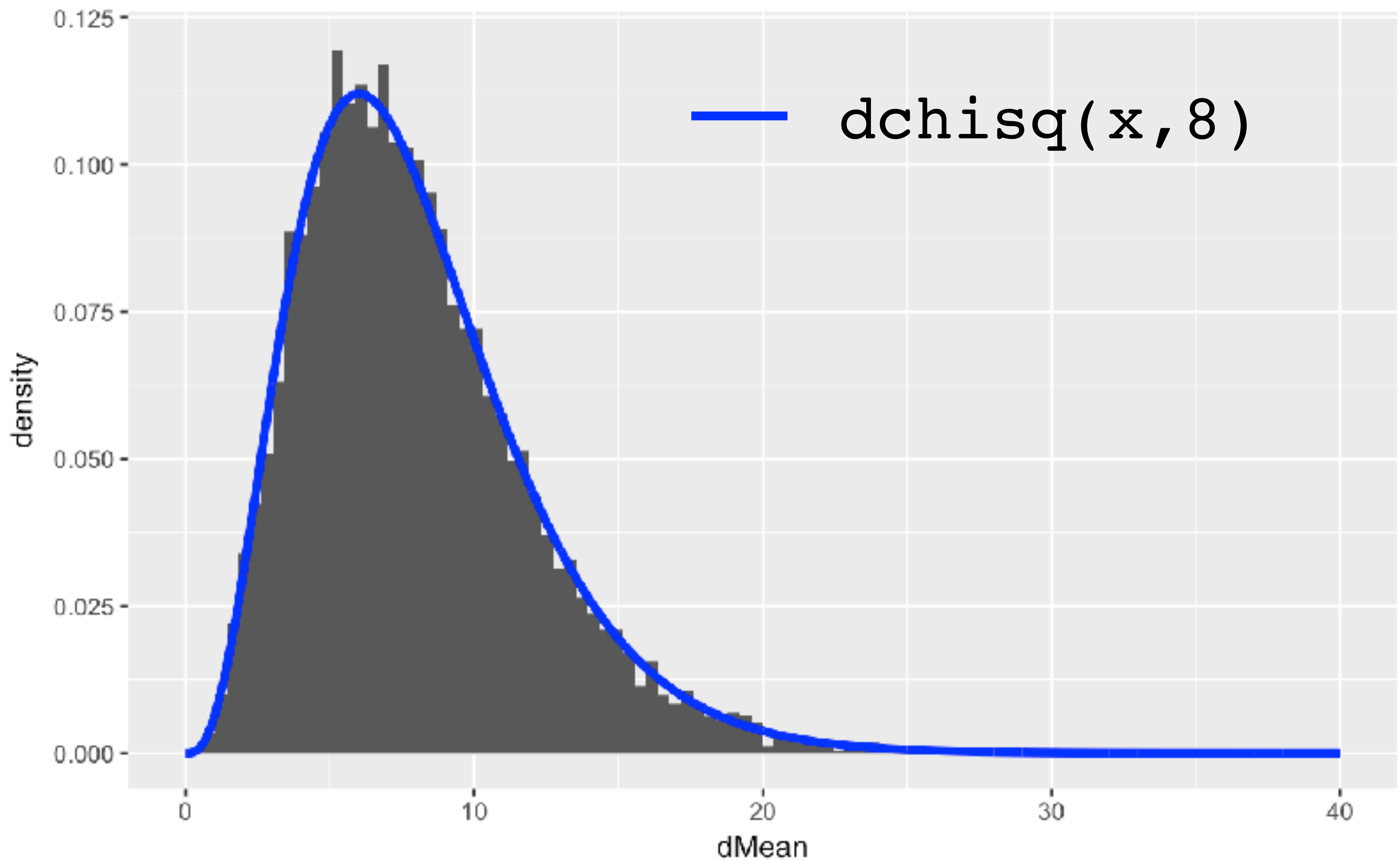
Figure 13.2 The $\chi 2$ distribution changes shape with the degrees of freedom

from Field, An Adventure in Statistics

```
d=replicate(10000,rnorm(8)**2)
dMean=apply(d,2,sum)
```



```
d=replicate(10000,rnorm(8)**2)
dMean=apply(d,2,sum)
```



Chi-squared test in R

```
chisq.test(bdata$smoothbirths)

  Chi-squared test for given probabilities

data:  bdata$smoothbirths
X-squared = 132760, df = 365, p-value < 2.2e-16
```

degrees of freedom = N - 1

```
length(bdata$smoothbirths)
[1] 366
```

Comparing two variables: The contingency table

Counts

Diabetes

TVOver3Hrs	**No** <int>	**Yes** <int>
FALSE	3509	225
TRUE	974	148

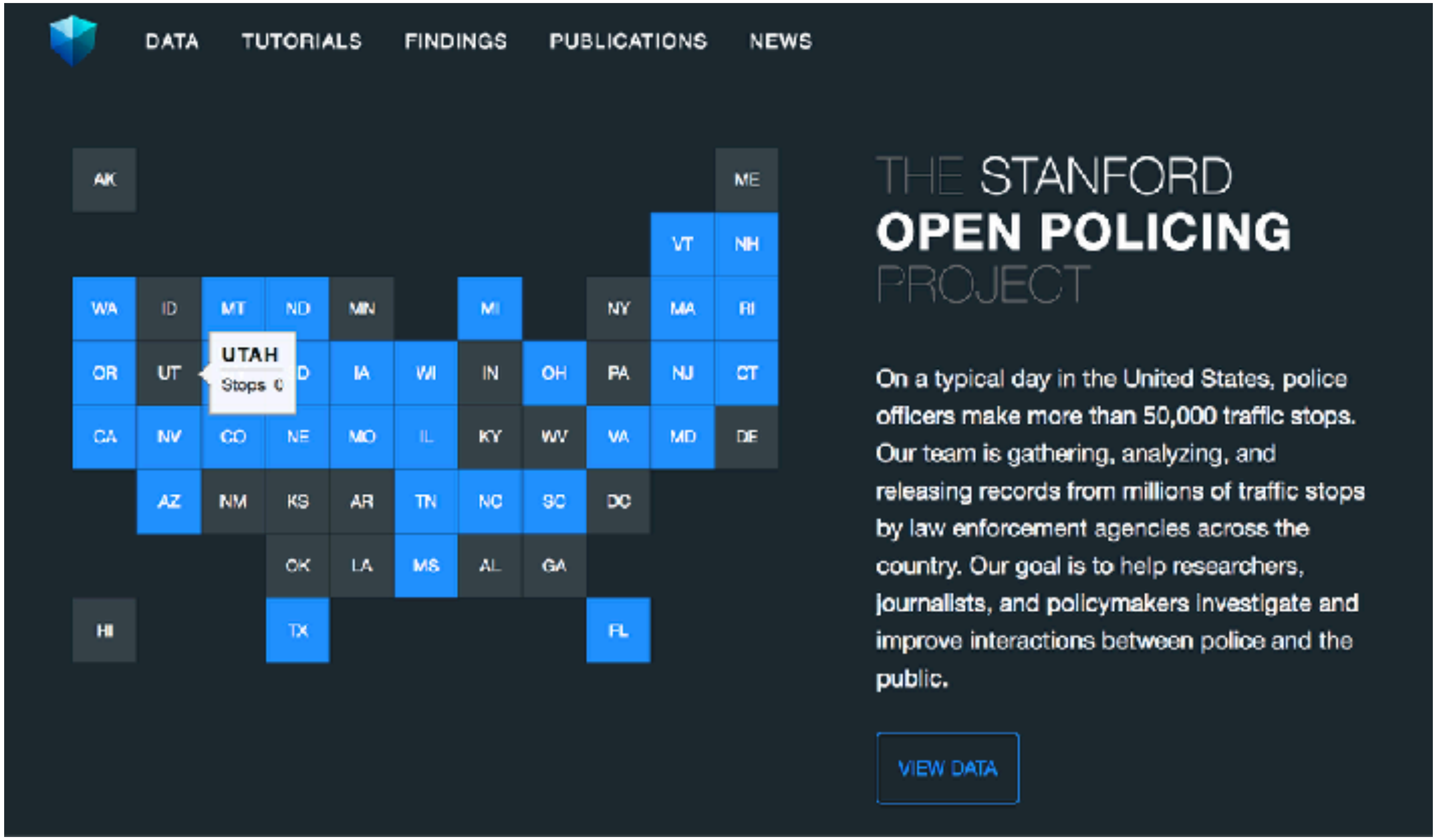
Proportions of total N

Diabetes

TVOver3Hrs	**No** <int>	**Yes** <int>
FALSE	0.722	0.046
TRUE	0.200	0.030

A societally relevant example: Racial disparities in policing

- Are black individuals more likely to be searched when they are stopped by the police, compared to white individuals?



https://openpolicing.stanford.edu/

Representing the data as a contingency table

- State of Connecticut, 318,669 total stops, 2013–2015

Raw counts

	Not searched	Searched
White	239,241	3,108
Black	36,244	1,219

Proportions of total N

	Not searched	Searched
White	0.855	0.011
Black	0.129	0.004

What would we expect if there was no relationship?

Expected probabilities under independence

- Remember that if X and Y are independent, then:

$$P(X \cap Y) = P(X) * P(Y)$$

- So we expect:

“marginal probabilities”

	Not searched	Searched	
White	p(NS)*P(W)	p(S)*P(W)	P(W)
Black	p(NS)*P(B)	p(S)*P(B)	P(B)
	p(NS)	p(S)	

Computing expected probabilities

Observed proportions

	Not searched	Searched
White	0.855	0.011
Black	0.129	0.004

Expected under independence (H_0)

	Not searched	Searched	
White	0.853	0.013	.866
Black	0.132	0.002	.134
	.985	.015	

How can we tell if these are different?

Pearson's chi-squared statistic for goodness of fit

$$\chi^2 = \sum_{i,j} \frac{(observed_{ij} - expected_{ij})^2}{expected_{ij}}$$

Observed

	Not searched	Searched
White	239241	3108
Black	36244	1219

Expected

	Not searched	Searched
White	238601	3748
Black	36884	579

standardized squared difference

	Not searched	Searched
White	1.71	109
Black	11.1	706

$\chi^2 = 828.3$

Degrees of freedom for chi-square on contingency tables

$$df = (r-1)(c-1)$$

where:
r=number of rows
c=number of columns

for a 2 X 2 contingency table:
r=2 rows
c=2 columns
df = (2-1)*(2-1) = 1

Intuition: once we know the marginal sums, then only one number is free to vary

	Not searched	Searched	sum
White	239,241	3,108	242,349
Black	36,244	1,219	37,463
sum	275,485	4327	

Police search example: A parametric test in R

```
summaryDf2wayTable=summaryDf2way %>%
      spread(searched,n) %>%
      select(-driver_race)
```

driver_race	FALSE	TRUE
<fctr>	<int>	<int>
Black	36244	1219
White	239241	3108

Police search example: A parametric test in R

```
chisqTestResult = chisq.test(summaryDf2wayTable,1,
     correct=FALSE)
chisqTestResult
```

```
 Pearson's Chi-squared test

 data:  summaryDf2wayTable
 X-squared = 828.3, df = 1, p-value < 2.2e-16
```

This is a non-directional hypothesis test
H0: searches and race are unrelated
HA: searches and race are related

Another example: diabetes vs. TV watching

- Example from Week 5 PSet:

Counts

	Diabetes	
TVOver3Hrs	No <int>	Yes <int>
FALSE	3509	225
TRUE	974	148

Proportions of total N

	Diabetes	
TVOver3Hrs	No <int>	Yes <int>
FALSE	0.722	0.046
TRUE	0.200	0.030

Chi-squared test on NHANES diabetes/TV data

```
chisq.test(summaryTable[,2:3],correct=FALSE)


Pearson's Chi-squared test

data:  summaryTable[, 2:3]
X-squared = 60, df = 1, p-value = 3e-15
```

Standardized residuals

$$standardized\,residual = \frac{observed_{ij} - expected_{ij}}{\sqrt{expected_{ij}}}$$

can be interpreted as a Z-score

Observed

	Not searched	Searched
White	239241	3108
Black	36244	1219

Expected

	Not searched	Searched
White	238601	3748
Black	36884	579

standardized residual

	Not searched	Searched
White	1.3	-10.4
Black	-3.3	26.5

Odds ratio

- Expresses the relative likelihood of different outcomes
- Odds are the relative likelihood of some event happening versus not happening

Odds ratio

- Expresses the relative likelihood of different outcomes
- Odds are the relative likelihood of some event happening versus not happening

The odds ratio is simply the ratio of two odds

$$\text{odds of } A = P(A)/P(\neg A)$$

Odds ratio

- Expresses the relative likelihood of different outcomes

$$odds_{searched|black} = \frac{N_{searched,black}}{N_{not\ searched,black}} = 0.034$$

$$odds_{searched|white} = \frac{N_{searched,white}}{N_{not\ searched,white}} = 0.013$$

$$odds\ ratio = \frac{odds_{searched|black}}{odds_{searched|white}} = 2.59$$

ODDS RATIO EXAMPLE: SMOKING AND LUNG CANCER

- What is the relationship between smoking and lung cancer?

- $$odds(\text{cancer in smokers}) = \frac{P(\text{cancer in smokers})}{P(\text{no cancer in smokers})}$$

- $$odds(\text{cancer in nonsmokers}) = \frac{P(\text{cancer in nonsmokers})}{P(\text{no cancer in nonsmokers})}$$

- $$oddsratio = \frac{odds(\text{cancer in smokers})}{odds(\text{cancer in nonsmokers})}$$

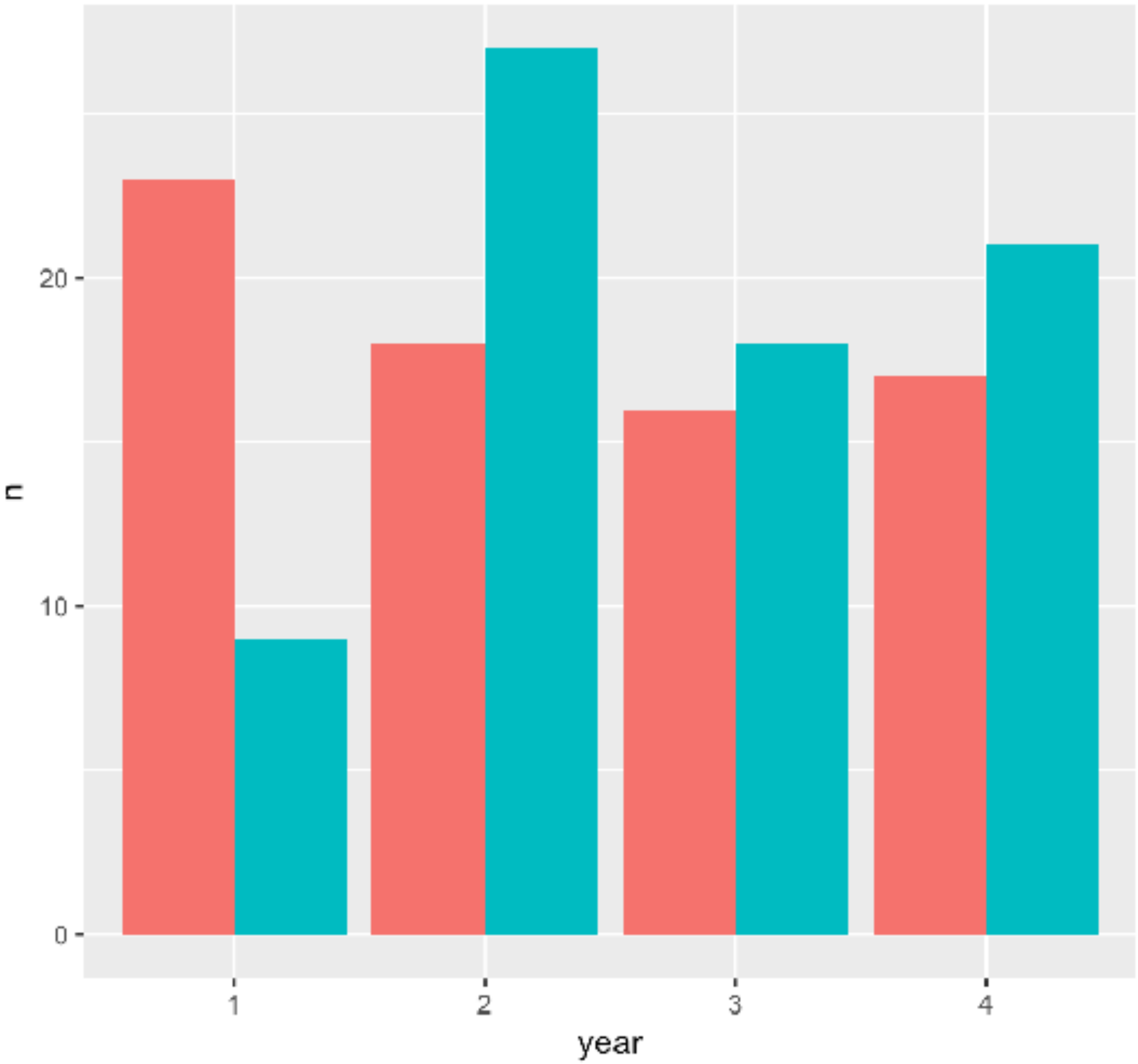
ODDS RATIO EXAMPLE: SMOKING AND LUNG CANCER

Using the data from a published study (Pesch et al., 2012) we can compute these values:

- The odds of someone having lung cancer who has never smoked is 0.08
- the odds of a current smoker having lung cancer is 1.77
- The odds ratio of 23.22 tells us that the odds of cancer in smokers are roughly 23 times higher than never-smokers.

Categorical analysis beyond the 2 X 2 table

- Survey data example: is programming experience related to year?




```
##   programmedBefore   `1`   `2`   `3`   `4`
##   <lgl>            <int> <int> <int> <int>
## 1 F                   23    18    16    17
## 2 T                    9    27    18    21
```

Chi-squared test: year vs. programming experience

H0: Year is unrelated to programming experience
HA: Year and programming experience are related

```
csResult = chisq.test(tableData$year,tableData$programmedBefore)
```

```
csResult

   Pearson's Chi-squared test

data:  tableData$year and tableData$programmedBefore
X-squared = 8, df = 3, p-value = 0.04
```

Group discussion

Seasonal batting averages for Derek Jeter and David Justice, 1995-7

	1995		1996		1997		Combined	
Derek Jeter	12/48	0.250	183/582	0.314	190/654	0.291	385/1284	**0.300**
David Justice	104/411	**0.253**	45/140	**0.321**	163/495	**0.329**	312/1046	0.298

How could this happen?
Which one of them is a better batter?

Sometimes summaries can be misleading: Simpson's paradox

- A pattern that is present in the overall data may be reversed in different subsets of the data
  - Due to a "lurking variable"
    - Different frequencies of at-bats across years
  - Often reflects different frequencies and proportions in subsets of the data

	1995		1996		1997		Combined	
Derek Jeter	12/48	0.250	183/582	0.314	190/654	0.291	385/1284	**0.300**
David Justice	104/411	**0.253**	45/140	**0.321**	163/495	**0.329**	312/1046	0.298

Berkeley graduate admissions example

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

Dept	Men		Women	
	Apps	Admitted	Apps	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

Recap

- We can summarize categorical variables in terms of contingency tables
- We can test for relations between categorical variables using a chi-squared test
- Sometimes combined data can be misleading
  - Always important to think about potentially lurking variables