Tutorial 6

Biological Data Analysis

Spring 2023

Outline

- χ^2 Chi squared distribution
- Variance estimation
- F distribution
- Comparison of variance
- Categorical data

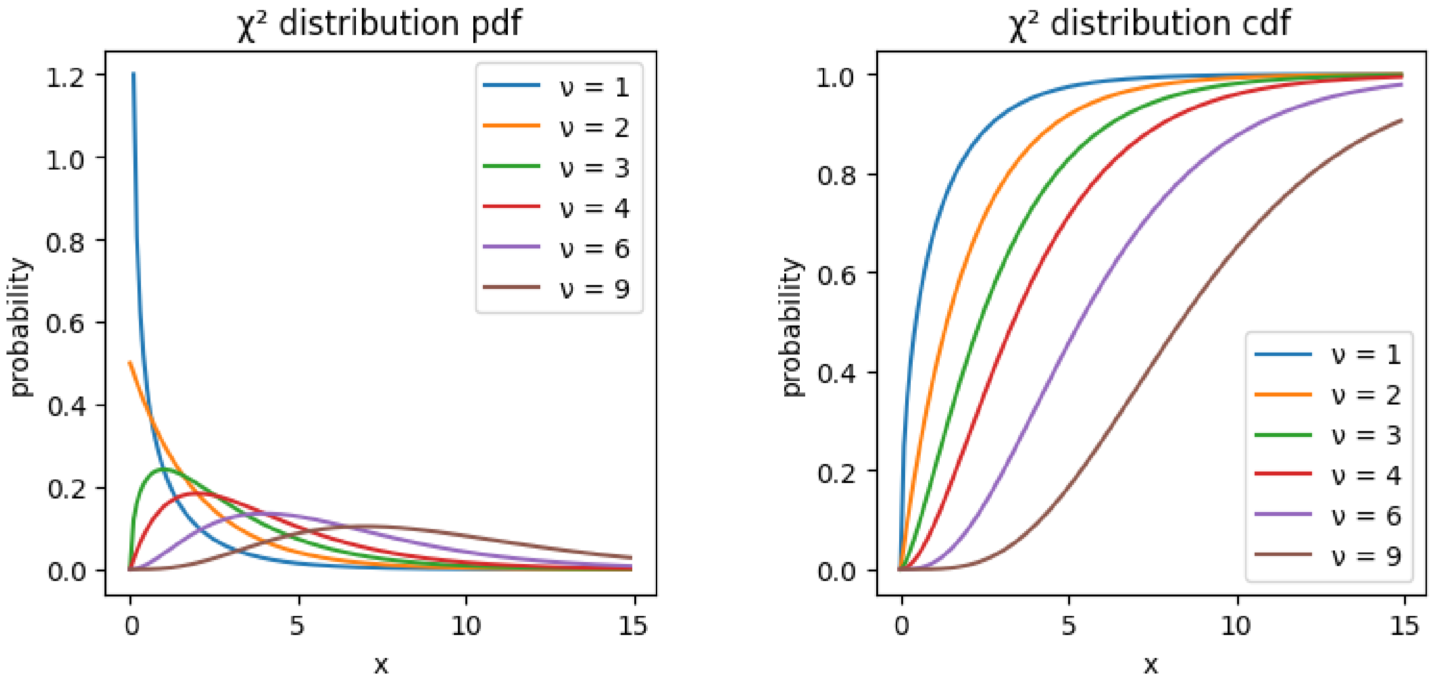
χ^2 Distribution

If x is a random variable from Normal standard distribution $N(0,1)$, than sum of x squared has χ^2_υ distribution

ν - degrees of freedom = N-1
Mean - k, variance - 2ν

$$x_i \sim \mathbb{N}(0,1) \Rightarrow \sum_{i=1}^{\upsilon} x_i^2 \sim \chi^2_\upsilon$$

χ^2 Distribution



χ^2 Distribution

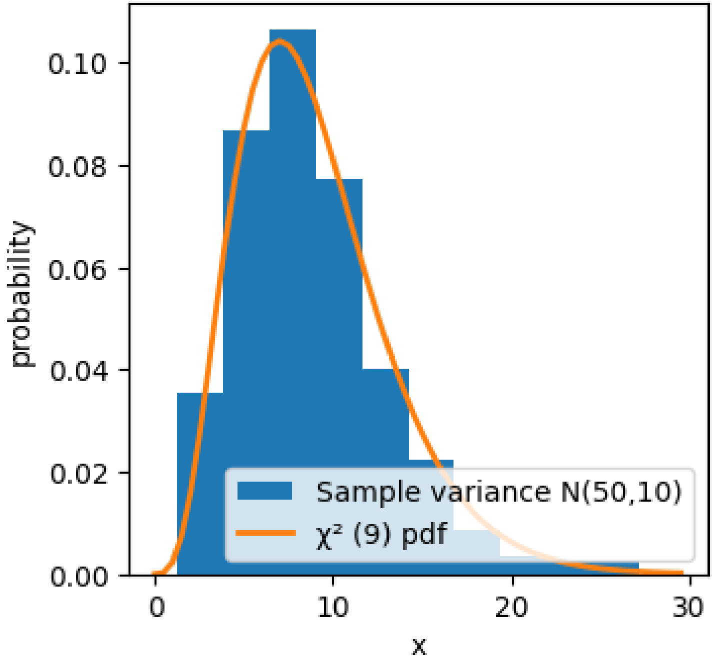
Variance of a sample with size N from Normal Distribution has χ^2 distribution with degrees of freedom ν = N-1

$$\frac{(N-1)*s^2}{\sigma^2} \sim \chi^2_{N-1}$$

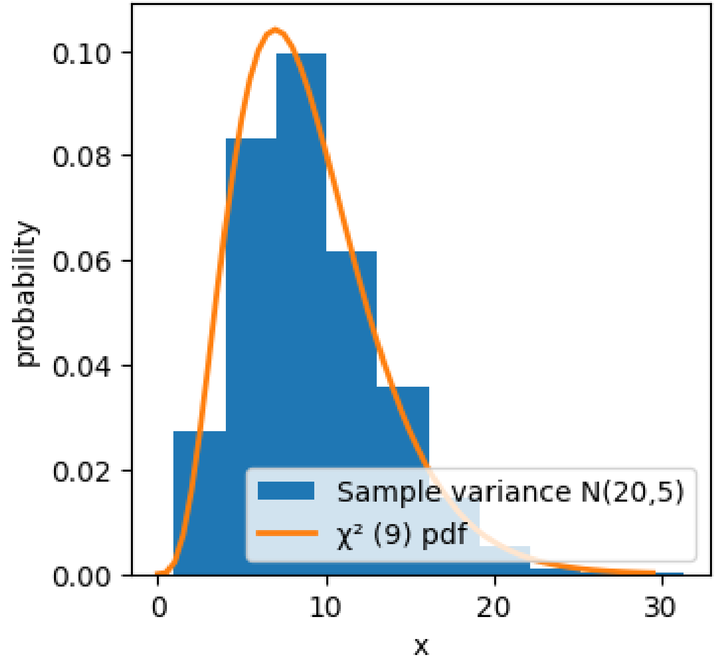
χ^2 Distribution

$$\frac{(N-1)*s^2}{\sigma^2} \sim \chi^2_{N-1}$$

Samples from N(50,10)



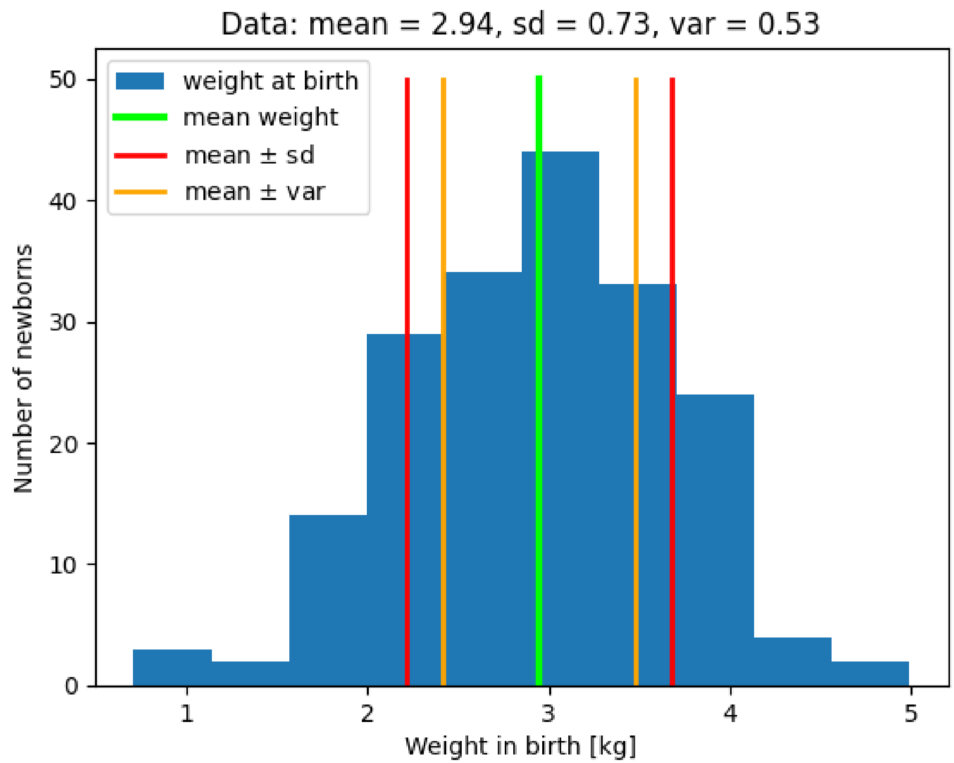
Samples from N(20,5)



Example 1

One hospital gathered data about the weight of the newborns
We want to estimate the spread of the newborns’ weights using variance

```
0       2.523
1       2.551
2       2.557
3       2.594
4       2.600
        ...
184     2.466
185     2.495
186     2.495
187     2.495
188     2.495
Name: bwt, Length: 189
```

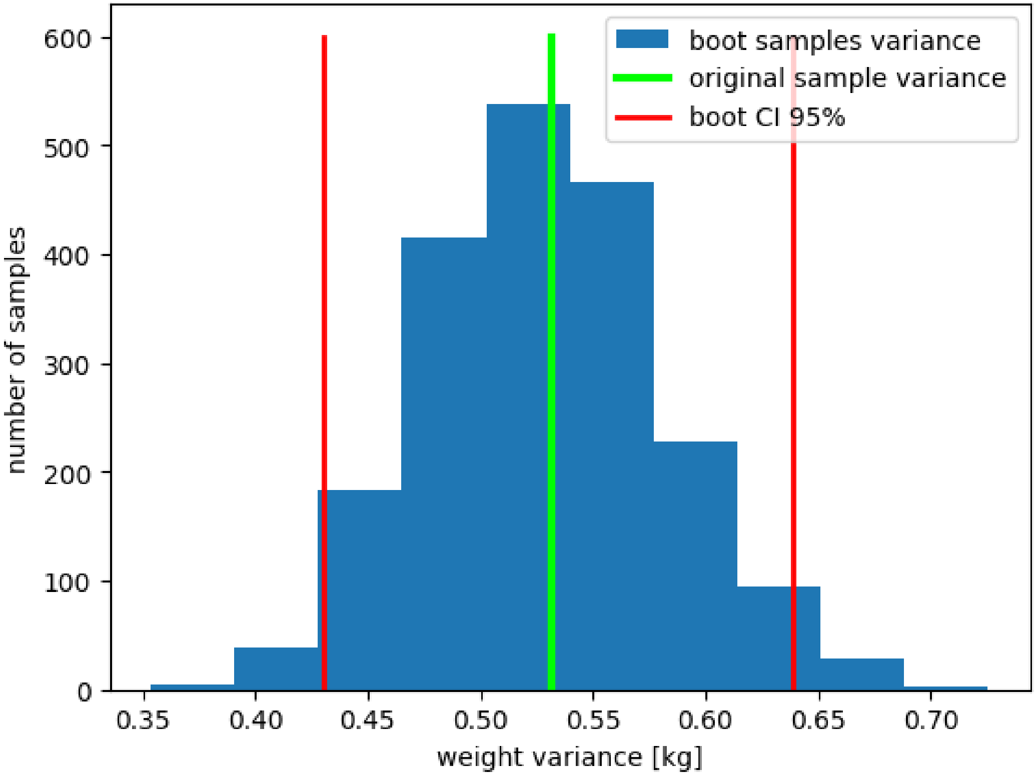


Example 1

Find 95% CI using regular bootstrap

CI 95% for variance

[0.4329 , 0.6344]



Example 1

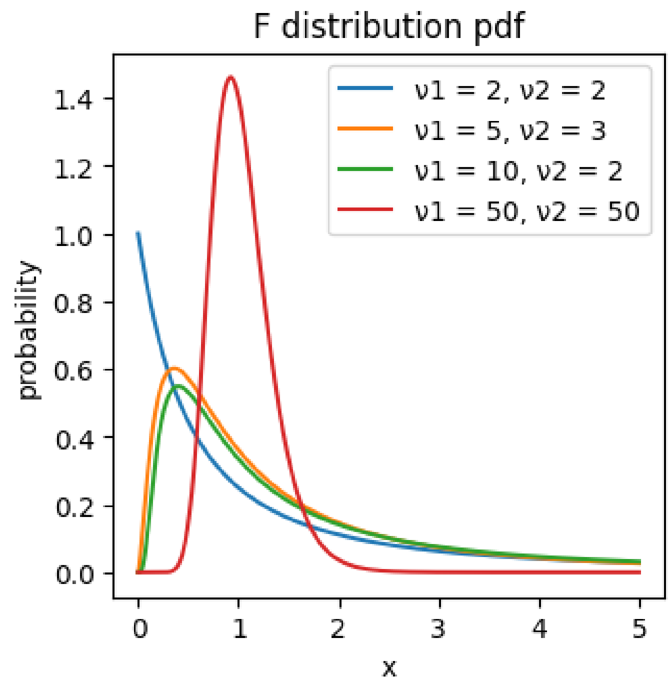
$$\frac{(N-1)*s^2}{\chi_{n-1}^{(1-\alpha/2)}} \le \sigma^2 \le \frac{(N-1)*s^2}{\chi_{n-1}^{(\alpha/2)}}$$

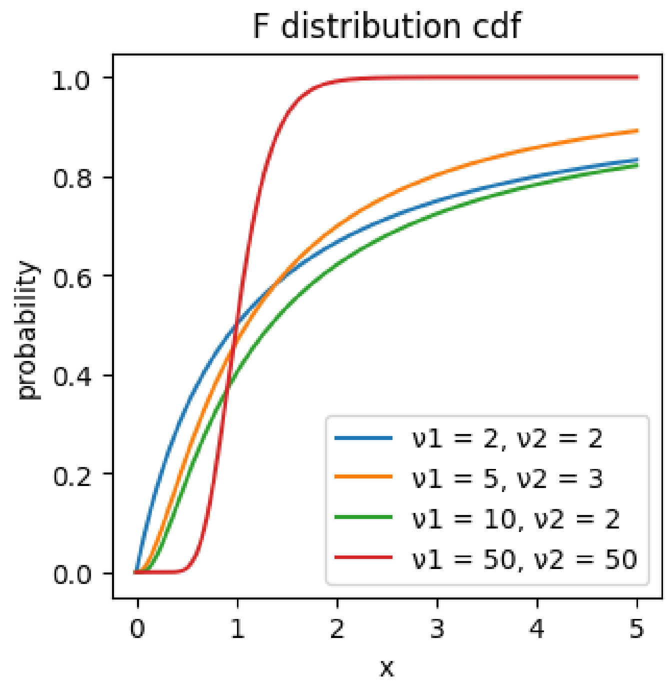
```
CI 95% for variance = [0.4387 , 0.658]
```

F Distribution

Ratio of two estimators for the same variance distributes F defined by 2 parameters: degrees of freedom for the nominator and the denominator

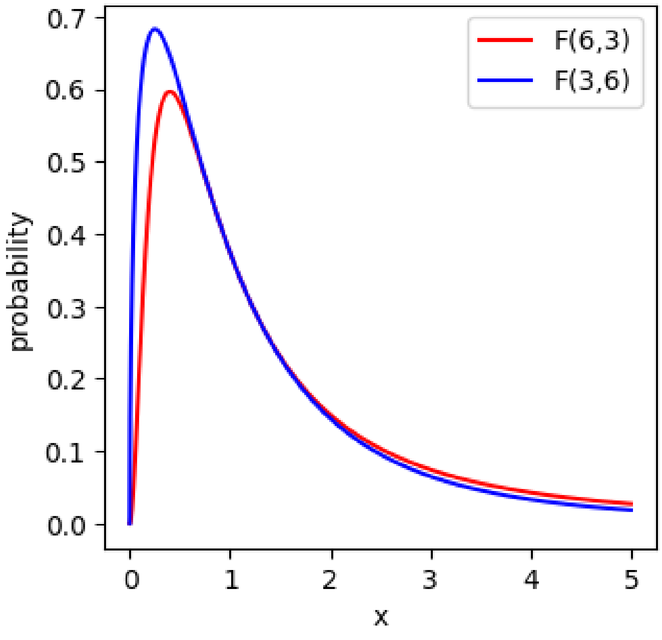
$$F_{\nu_1,\nu_2} = \frac{\dfrac{\chi^2_{\nu_1}}{\nu_1}}{\dfrac{\chi^2_{\nu_2}}{\nu_2}}$$

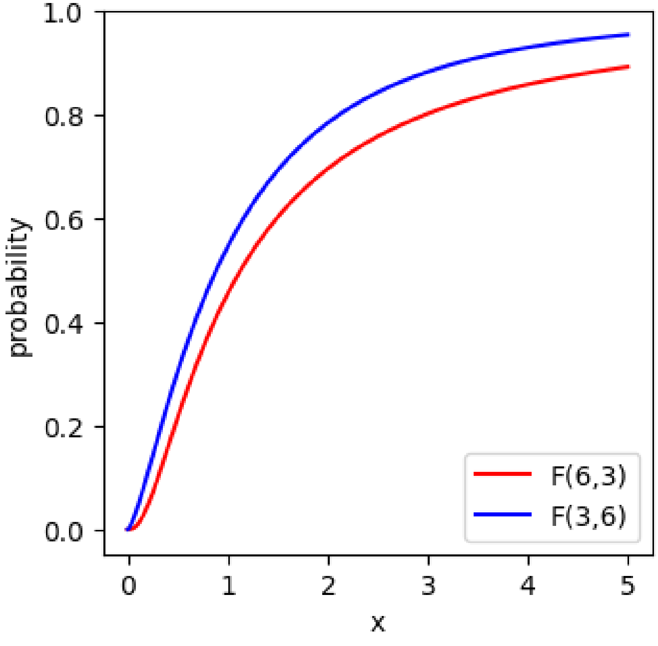




F Distribution

Example of 2 different F distribution:
F(3,6) and F(6,3)



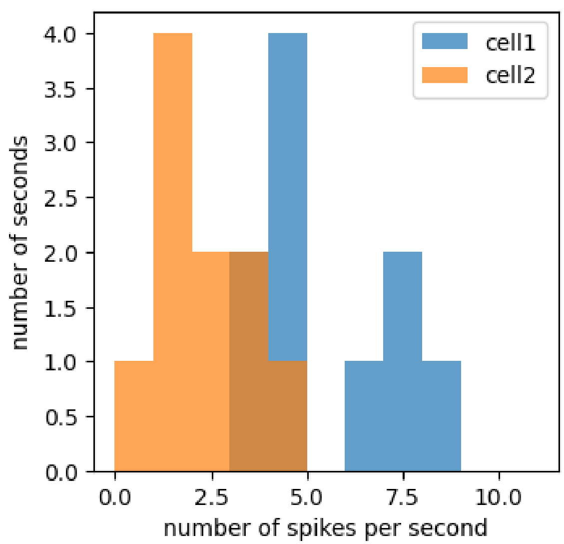


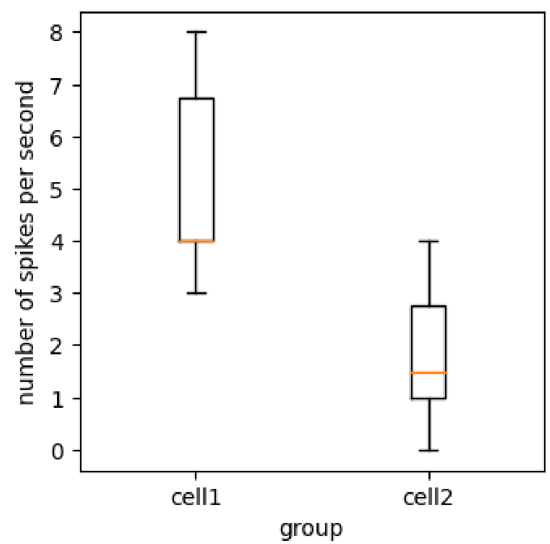
Example 2
compare variances of two groups

Neural activity of two cells was recorded
Data is number of spikes (action potentials) per second

Cell 1 variance = 3.33 ; Cell 2 variance = 1.51

cell1	cell2
8	2
7	0
4	4
6	1
3	3
4	1
4	3
7	1
3	1
4	2



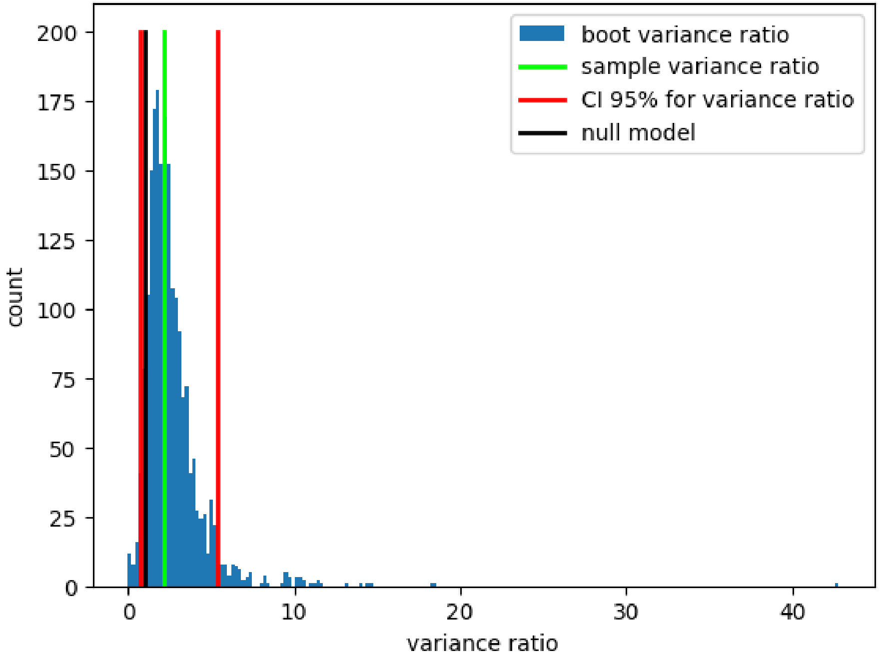


Example 2
regular bootstrap

Confidence Interval 95% for variance ratio and p-value

```
CI 95% for variance ratio =
[0.7405 , 5.3961]
```

```
p-value = 0.0665
```



Null model inside the CI

p=-value >0.05

Null model not rejected

Example 2

Confidence Interval 95% for variance ratio

Nominator - bigger variance, denominator – smaller variance

$$\frac{S_1^2}{S_2^2} * F_{N_1-1, N_2-1, (p=\alpha/2)} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2} * F_{N_1-1, N_2-1, (p=1-\alpha/2)}$$

```
CI 95% for variance ratio =
[0.7166 , 5.6876]
```

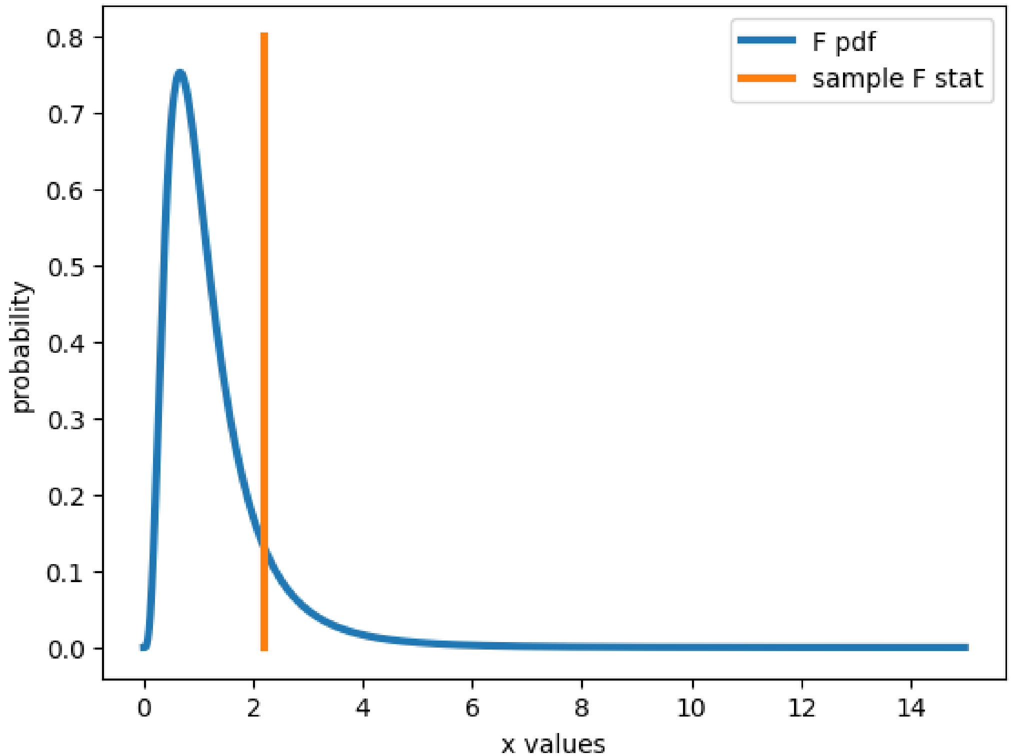
Example 2

Parametric test for variance comparison – F distribution

$$F_{\nu_1,\nu_2} = \frac{S^2_{1(bigger)}}{S^2_{2(smaller)}}$$

F stat = 2.2

P value = 0.1271



Categorical data and Chi2 distribution

- Data is a sample with size N
- Data is divided into k categories:

  Number of observations in category i is Oi

- We check if the data is distributed according to some expected distribution

  Number of observations expected according to the distribution is Ei

- What is the probability that our sample does not come from the expected distribution
- Null Model – sample comes from the expected distribution

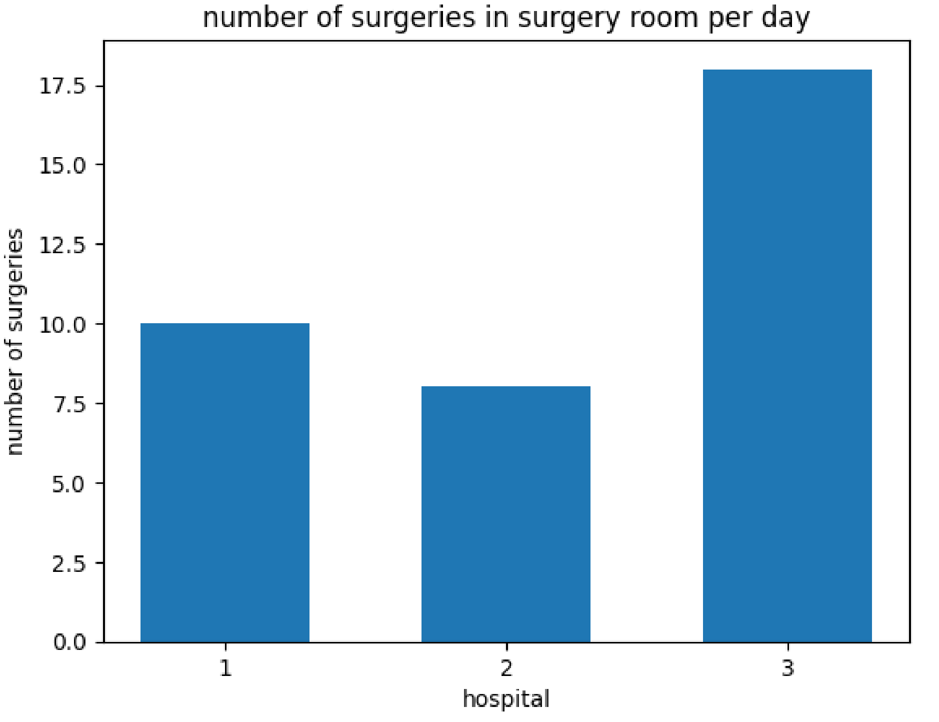
$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1}$$

Example 3

The are three hospitals in one city. A staff of one of the hospitals complained that the burden on the hospitals is distribution equally. To test this claim, an average number of surgeries in surgery room per day was calculated.

```
Number of surgeries

[10 8 18]
```



Example 3

Null model - the burden is equal. Under this assumption, the number of surgeries is 12 per day in each hospital

$$\sum_{i=1}^{k}\frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1}$$

O – observed
E – expected under null model

O = 10   8   18

E = 12 12 12

df = 2

Chi2 = 4.67

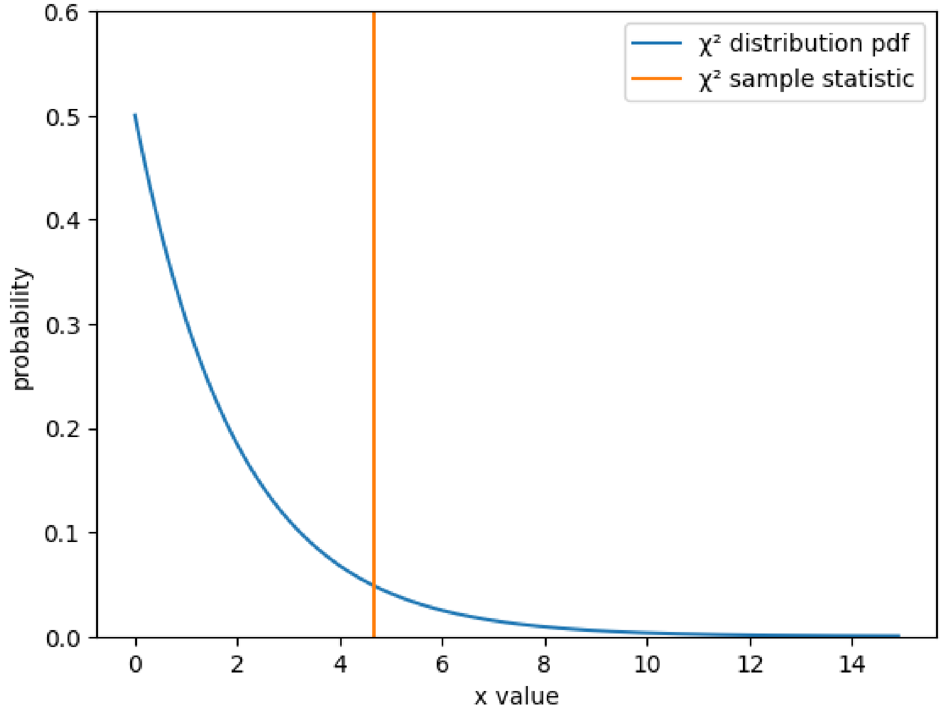
Null model not rejected

P = 0.097

Example 3

Chi2 (df=2) function and chi2 statistic

P (area under the curve to the right of the chi2 statistic) = 0.097

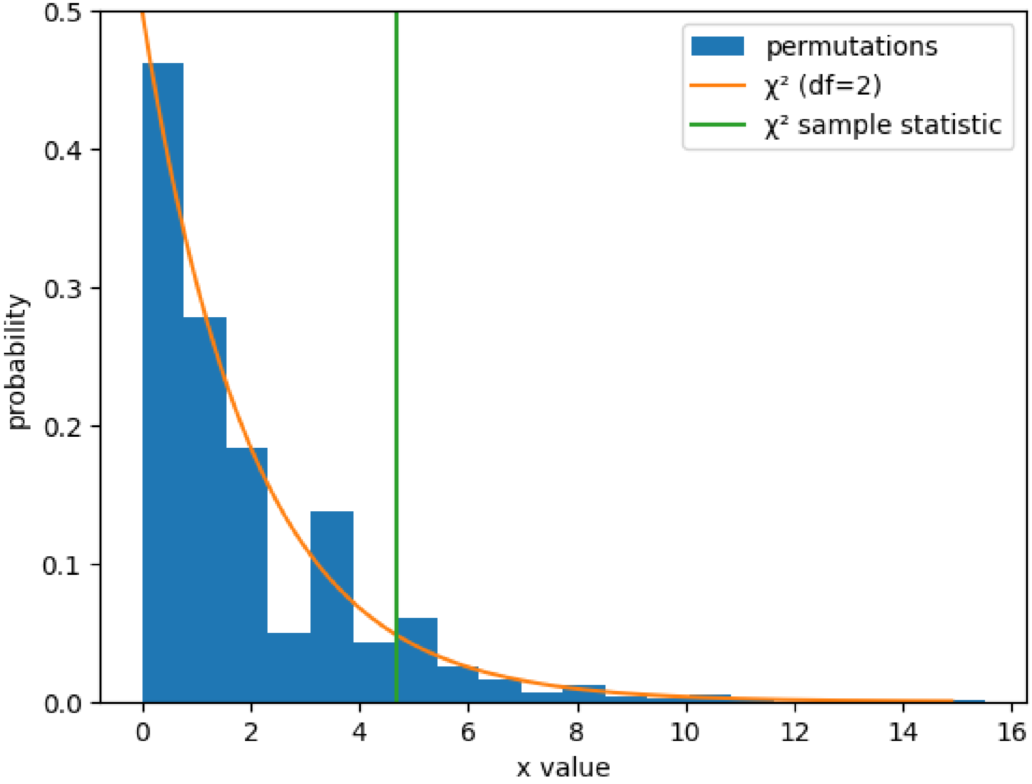


Example 3

Null model distribution using permutations

P (permutations) = 0.1110

P (parametric) = 0.097



Example 3
Effect size

Effect size for categorical data (vector) **Cramer's V**

$$V = \sqrt{\frac{\chi^2}{N * df}}$$

χ^2 - Sample statistic

N - Sample size

df - Degrees of freedom

Cramer's V = 0.2546

V<0.3	- small effect
0.5>V>0.3	- medium effect
V>0.5	- large effect

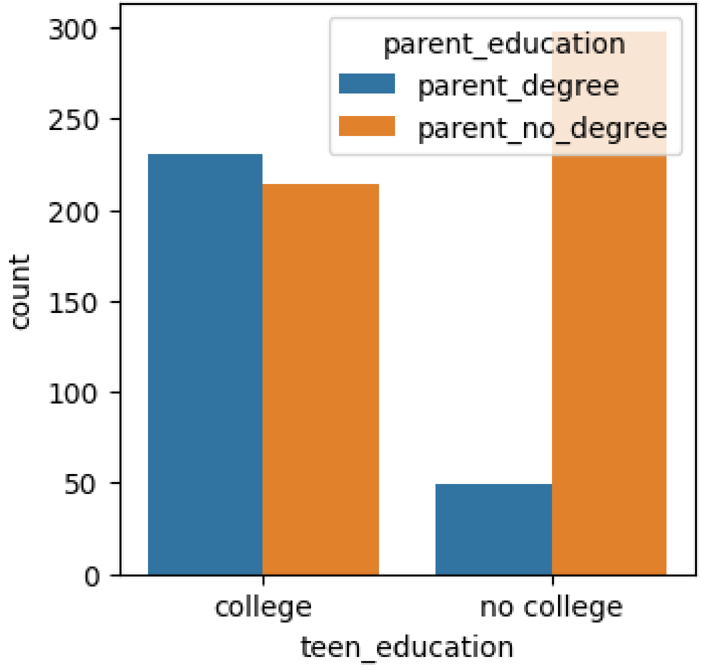
Example 4
Chi square – categorical 2D sample

teen_education	parent_education	count
college	parent_degree	231
no college	parent_degree	49
college	parent_no_degree	214
no college	parent_no_degree	298

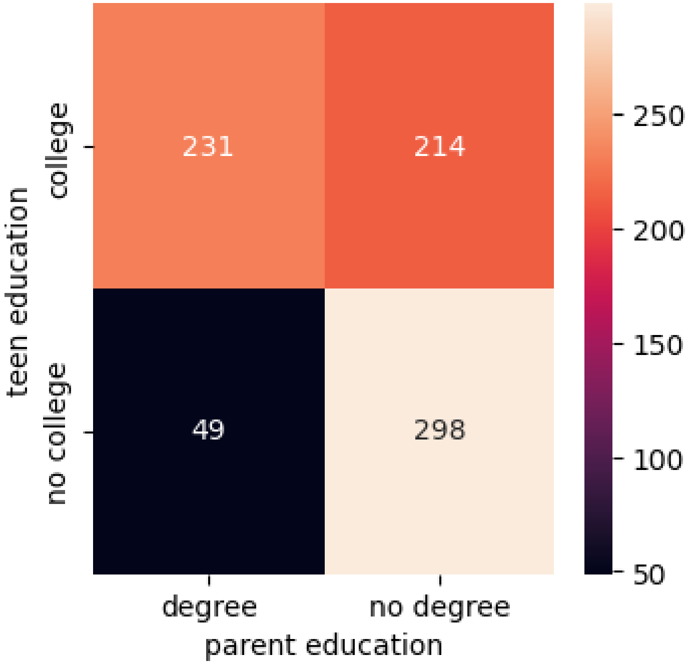
We want to check if parents education affects the education of their children

Example 4

Bar plot



Heatmap



Example 4

```
O = [[231 214] [ 49 298]]
```

N parent degree / no degree [280 512]
N teen college / no college [445 347]
N total 792

	Parent degree	Parent no degree	Total
Teen college	231	214	445
Teen no college	49	298	347
Total	280	512	792

Example 4

E(parent_group & teen_group) = N*P(parent_group)*P(teen_group)=

N*(Nparent_group/N)*(N teen_group s/N)=(Nparent_group*N teen_group)/N

	Parent degree	Parent no degree	Total
Teen college	231	214	445
Teen college expected	*157.32*	*287.68*	
Teen no college	49	298	347
Teen no college expected	*122.68*	*224.32*	
Total	280	512	792

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1}$$

Example 4

chi2(parent_group & teen_group) = (O(parent_group & teen_group) - E(parent_group & teen_group))^2/ E(parent_group & teen_group)

Chi_ParentDegree_TeenCollege	**34.5039**
Chi_ParentDegree_TeenNoCollege	**44.2485**
Chi_ParentNoDegree_TeenCollege	**18.8693**
Chi_ParentNoDegree_TeenNoCollege	**24.1984**

chi2_stat = sum(chi2(parent_group & teen_group))

Chi2 = 121.8202

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1}$$

Example 4

E = N_parent_degree*N_teen_college /N_total

```
[[280],   dot  [[445, 347]] / 792
[512]]
column vector * row vector (vector multiplication)
```

Chi2_stat = sum((O-E)^2)/E

Chi2 = 121.8202

Example 4

Calculating degrees of freedom

For a table with:

number of rows - r

number of columns – c

Degree of freedom of Chi2 distribution:

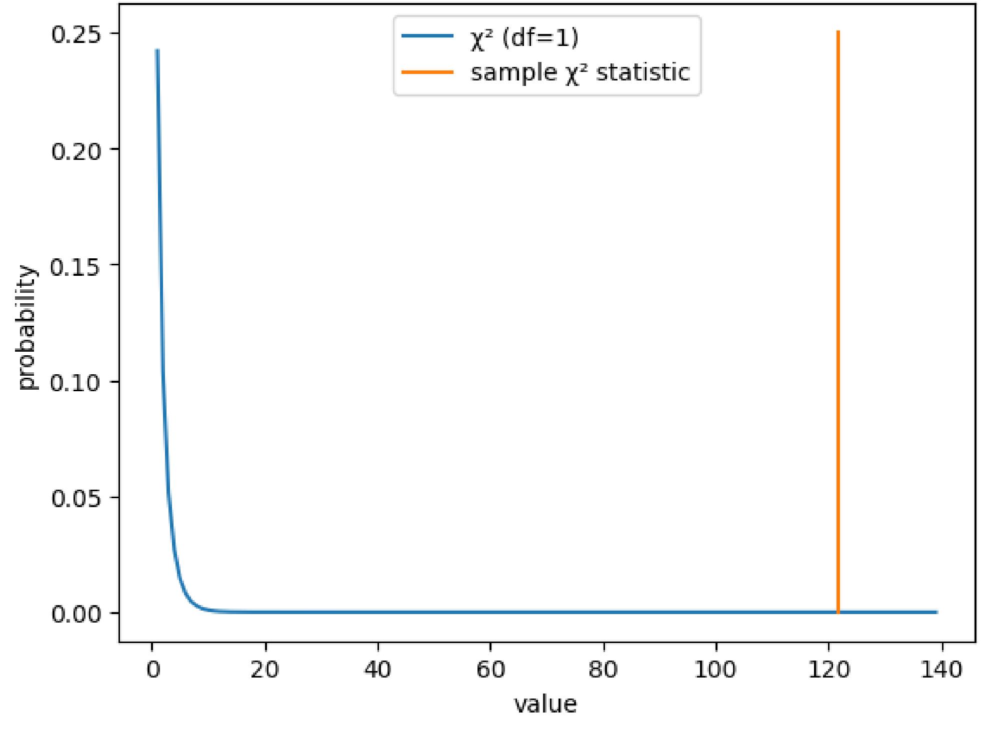
df = (r-1)*(c-1)

Example 4

df = (number of rows-1)*(number of columns-1) = 1

P-value = probability of getting chi2 statistics or larger under chi2(df=1) distribution

P = 0



Null model rejected

parents education affects the education of their children

Effect size

Effect size for categorical data (table) **Cramer's V**

$$V = \sqrt{\frac{\chi^2}{N * df}}$$

χ^2 - Sample statistic

N - Sample size

df - Degrees of freedom = min{r-1 , c-1}

V<0.3	- small effect
0.5>V>0.3	- medium effect
V>0.5	- large effect

V = 0.3922

Effect is medium