Analysing Categorical Data

ADS2 - Week 2.5

March 11th, 2024

Chaochen Wang
chaochenwang@intl.zju.edu.cn

Analysing Categorical Data

- Nominal and ordinal data
- Chi-square distribution
- Chi-square goodness-of-fit test
- Chi-square test for homogeneity
- Chi-square test for independence
- Fisher's exact test
- Chi-square 3-way sample
- Tests for ordinal variables

Learning Objectives

After this lecture you will be able to

- Describe and visualise categorical data

- Understand and perform chi-square test of categorical data

Categorical Data

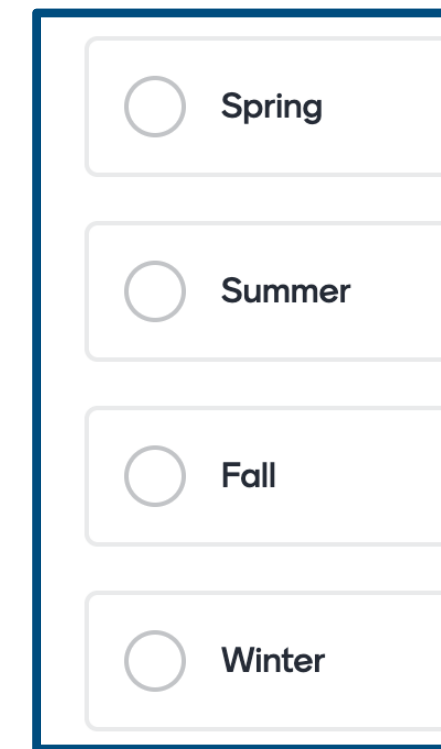Values only chosen from discrete and finite values (categories)

Values are mutually exclusive

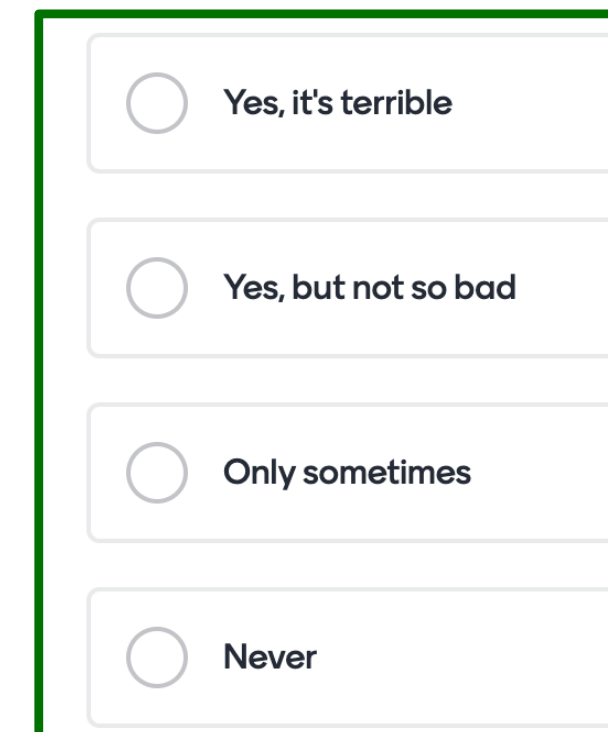Each subject can only choose one category

Nominal and ordinal

Nominal variables have no particular order: Hair color, race,

gender, season preference, ...

Ordinal variables have some inherent ordering: Pain levels,

ratings, ...

Season Preferences: Test for Goodness-of-Fit

	Spring	Summer	Fall	Winter	Total
Observed	13	10	5	3	31
	41 %	33 %	16 %	10 %	100 %
Expected	7.75	7.75	7.75	7.75	31
	25 %	25 %	25 %	25 %	100 %

Question: Is there a difference between the season preferences?

H_0: There is no difference between the observed and expected season preferences

H_1: There is a difference between the observed and expected season preferences

Normal Distribution

μ = 0
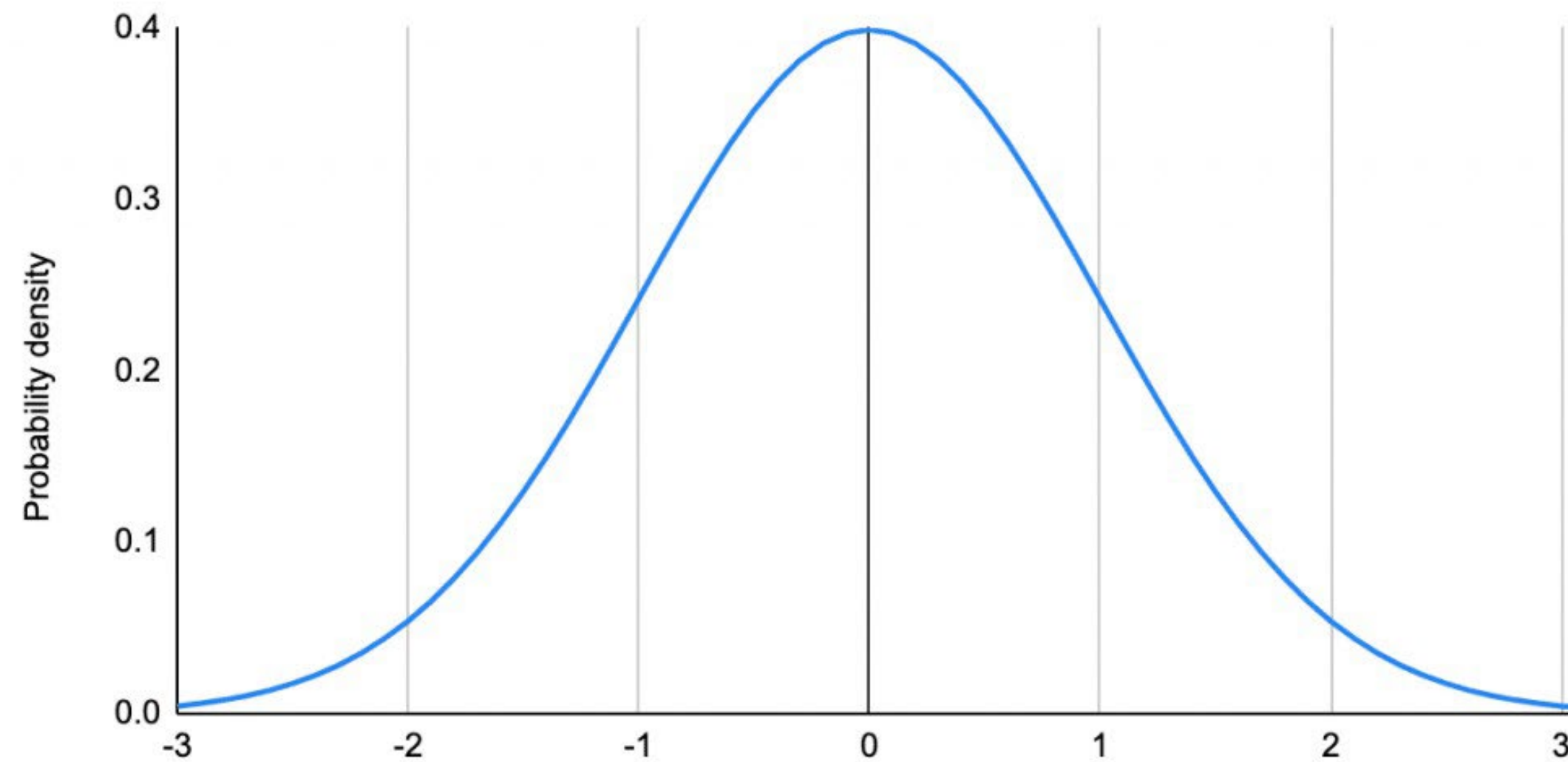
σ = 1

Chi-Square Distribution

Distribution of $Q_1 = X^2$
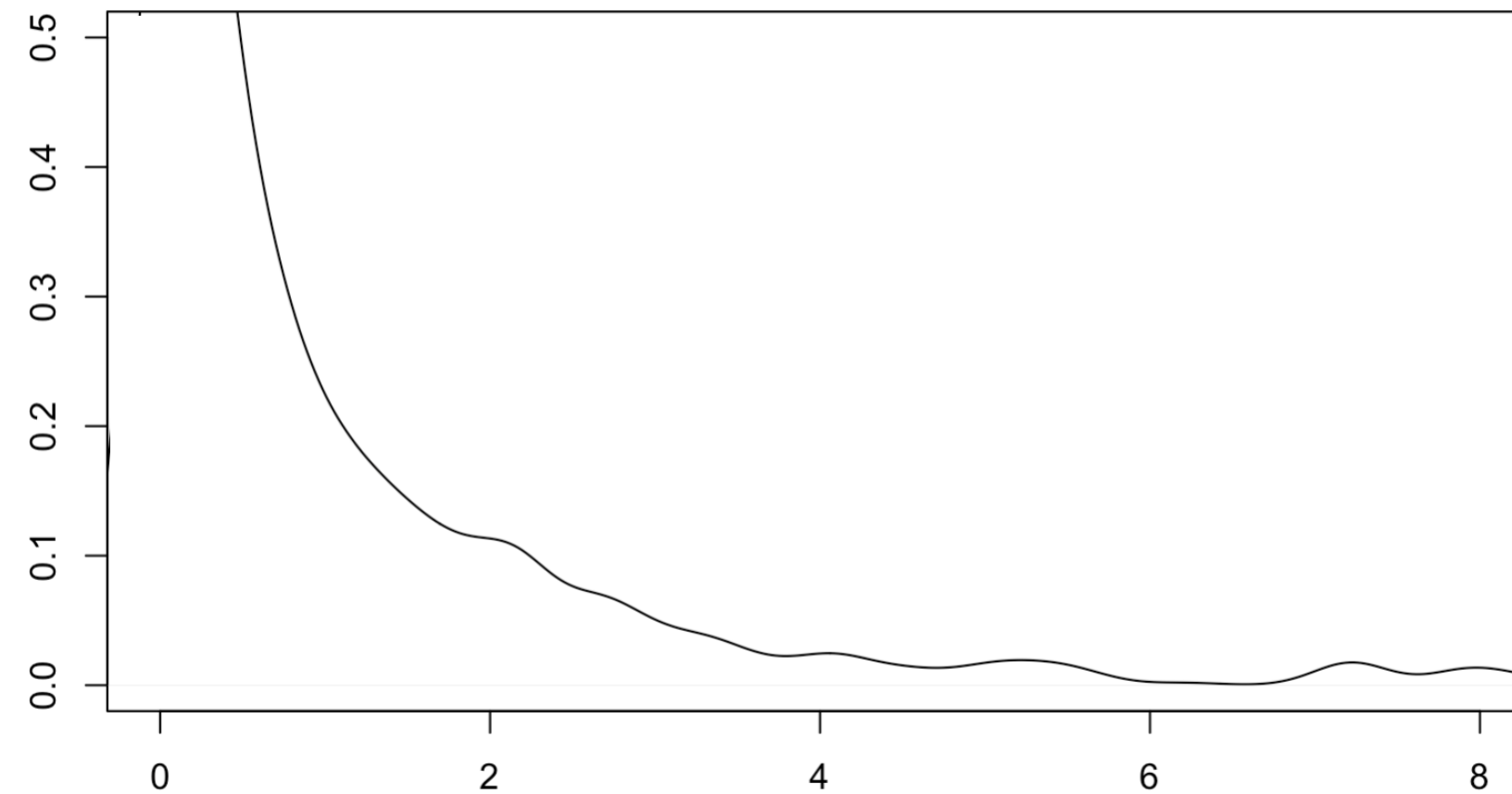
Simulation with 1000 datapoints

Chi-Square Distribution

Distribution of $Q_2 = X_1^2 + X_2^2$

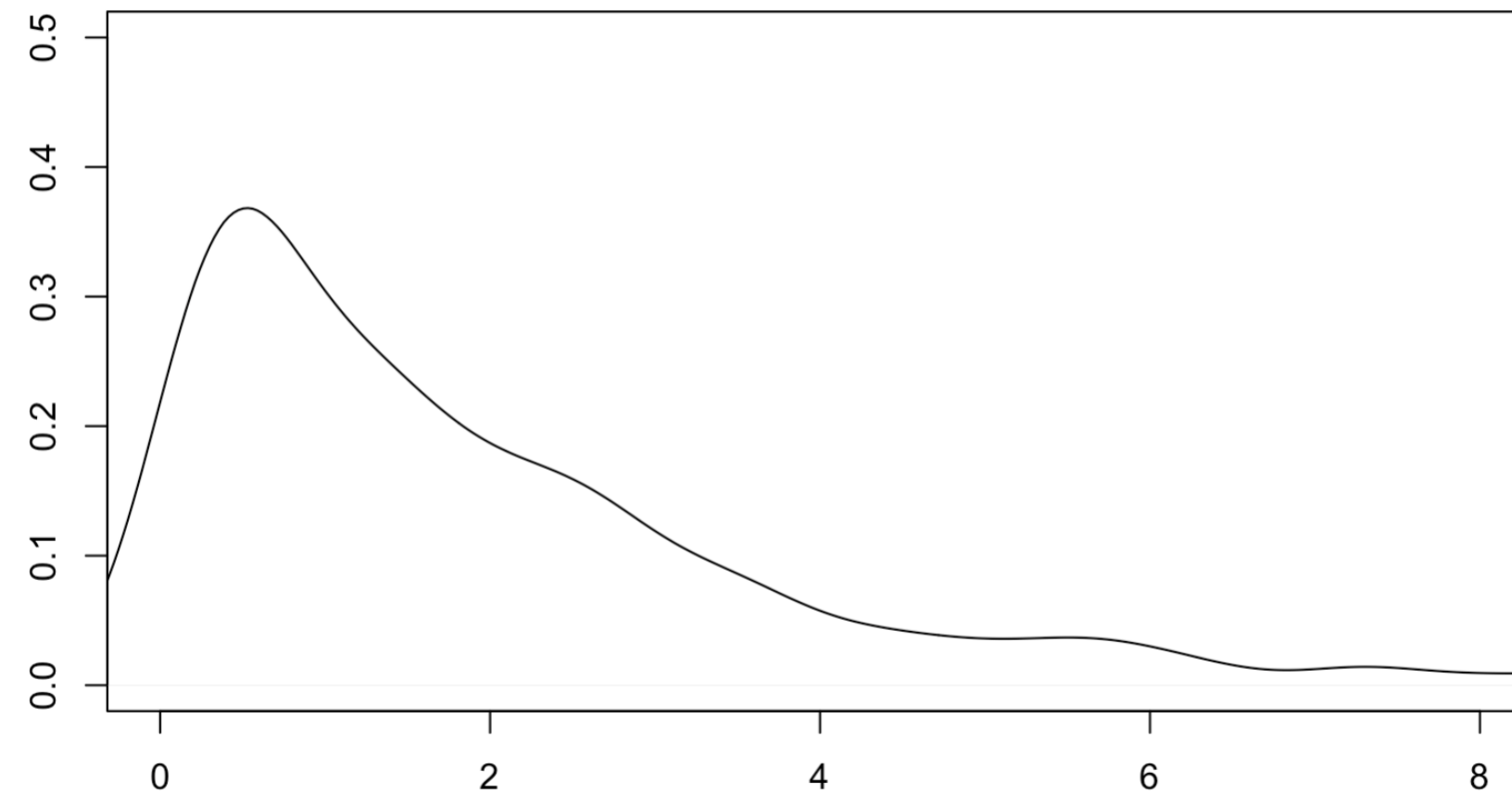

Simulation with 1000 datapoints

Chi-Square Distribution

Distribution of $Q_3 = X_1^2 + X_2^2 + X_3^2$
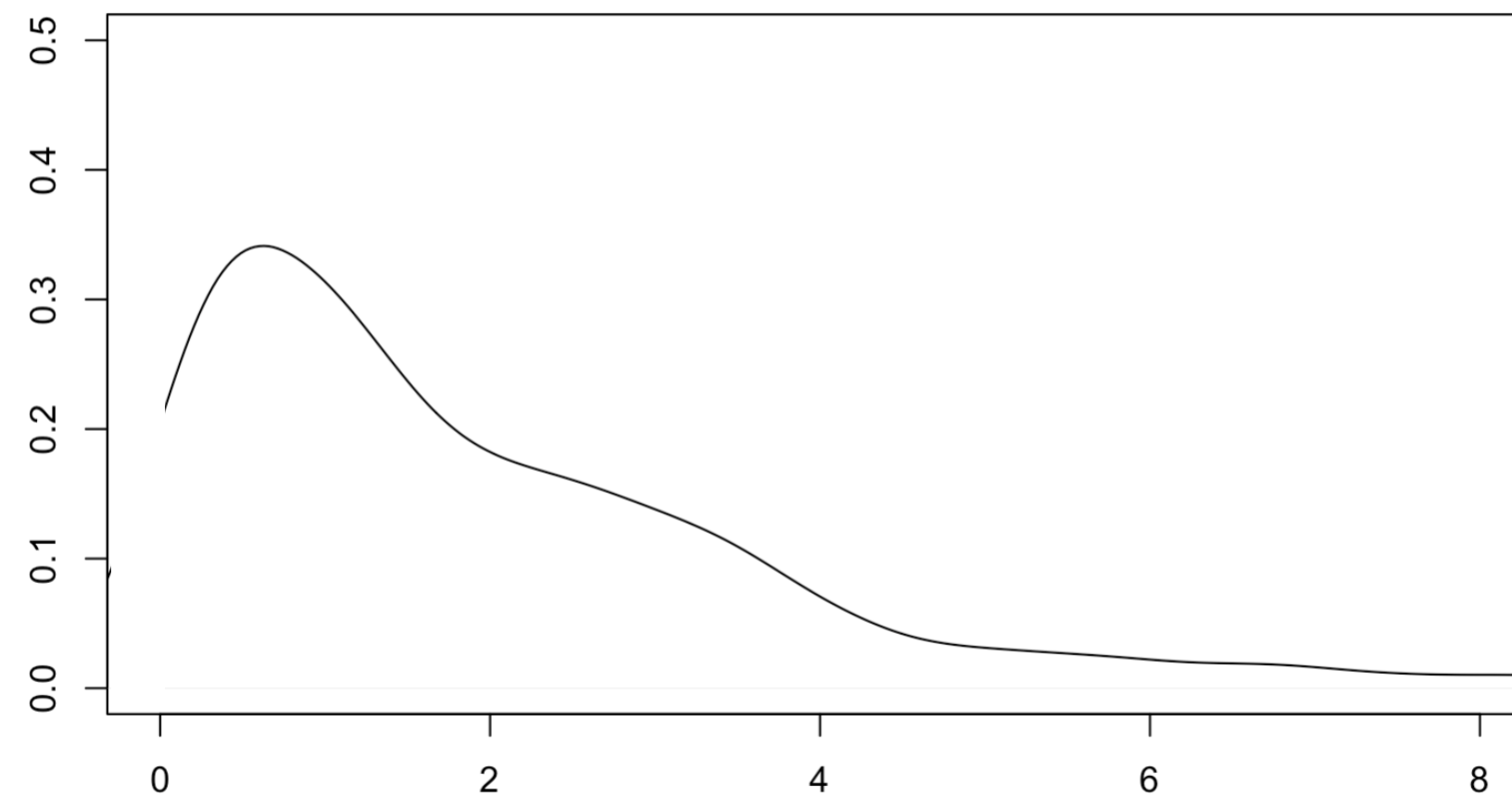
Simulation with 1000 datapoints

Chi-Square Distribution

Distribution of $Q_4 = X_1^2 + X_2^2 + X_3^2 + X_4^2$
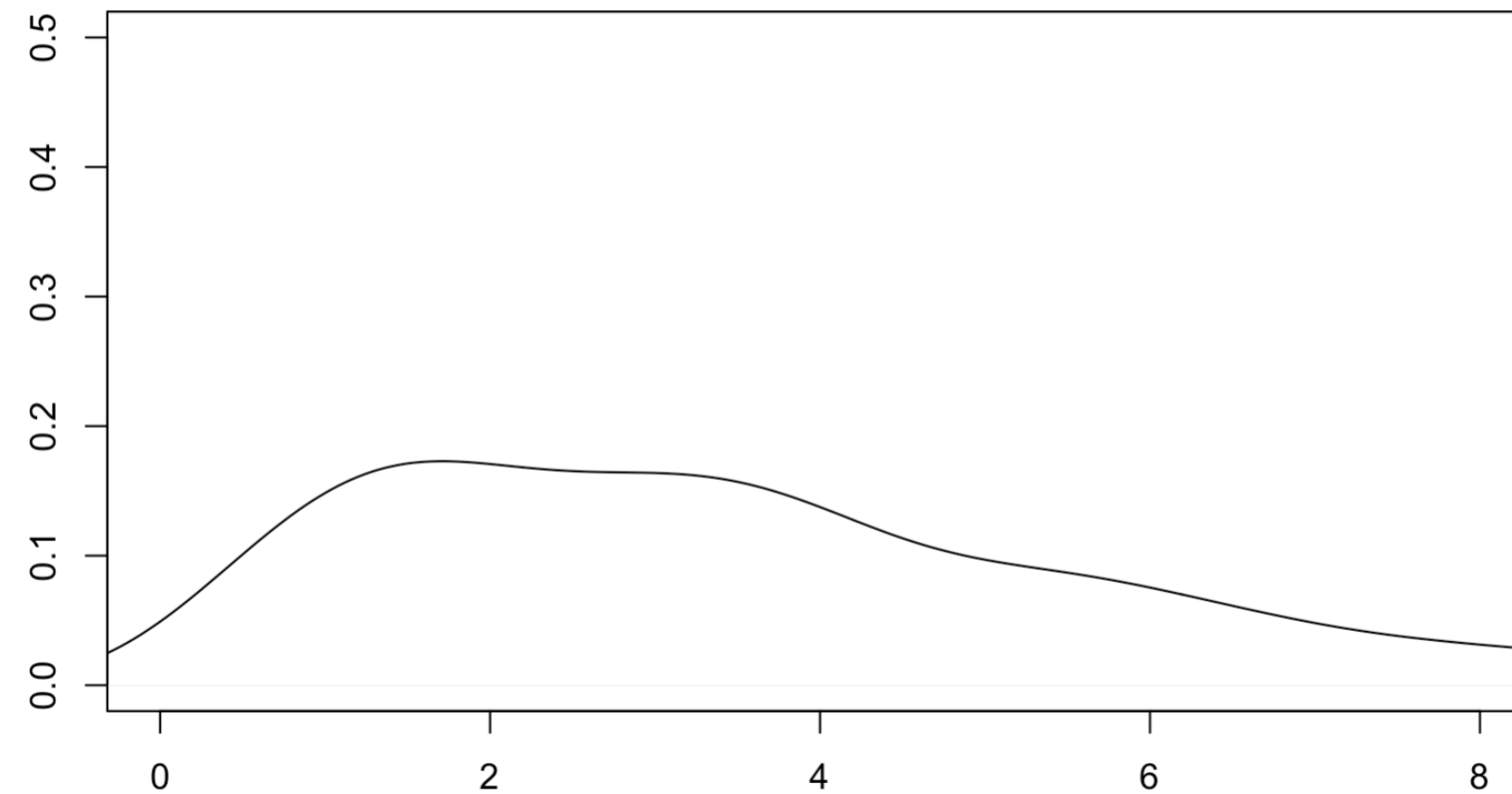
Simulation with 1000 datapoints

Chi-Square Distribution
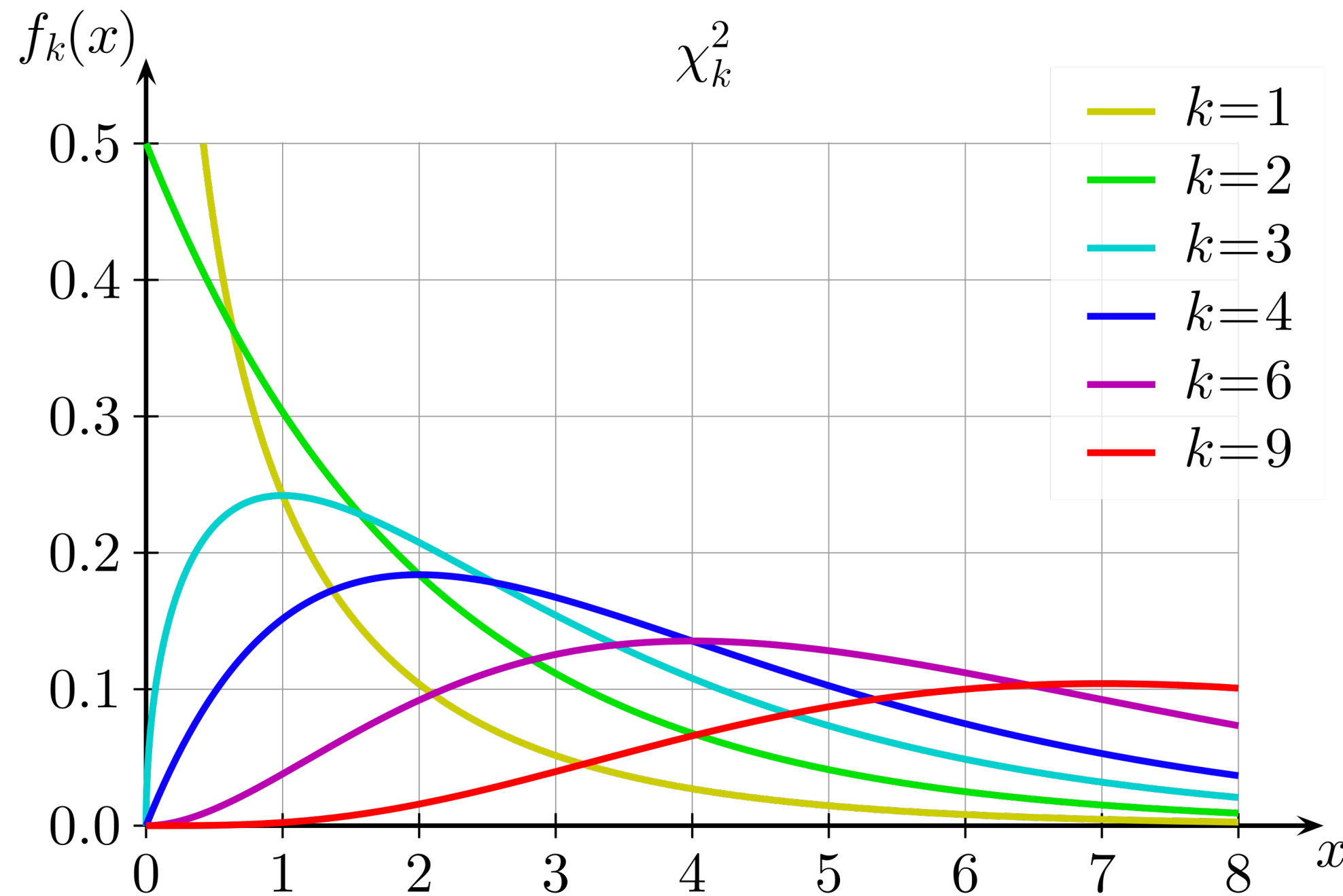
$$Qk = \sum_{i=1}^{k} x_i{}^2$$

x_i :independent standard normal random variables

k: degrees of freedom

Chi-Square Distribution

pdf $f(x)$

$p = \Pr[X \geq \chi^2]$

χ^2

x
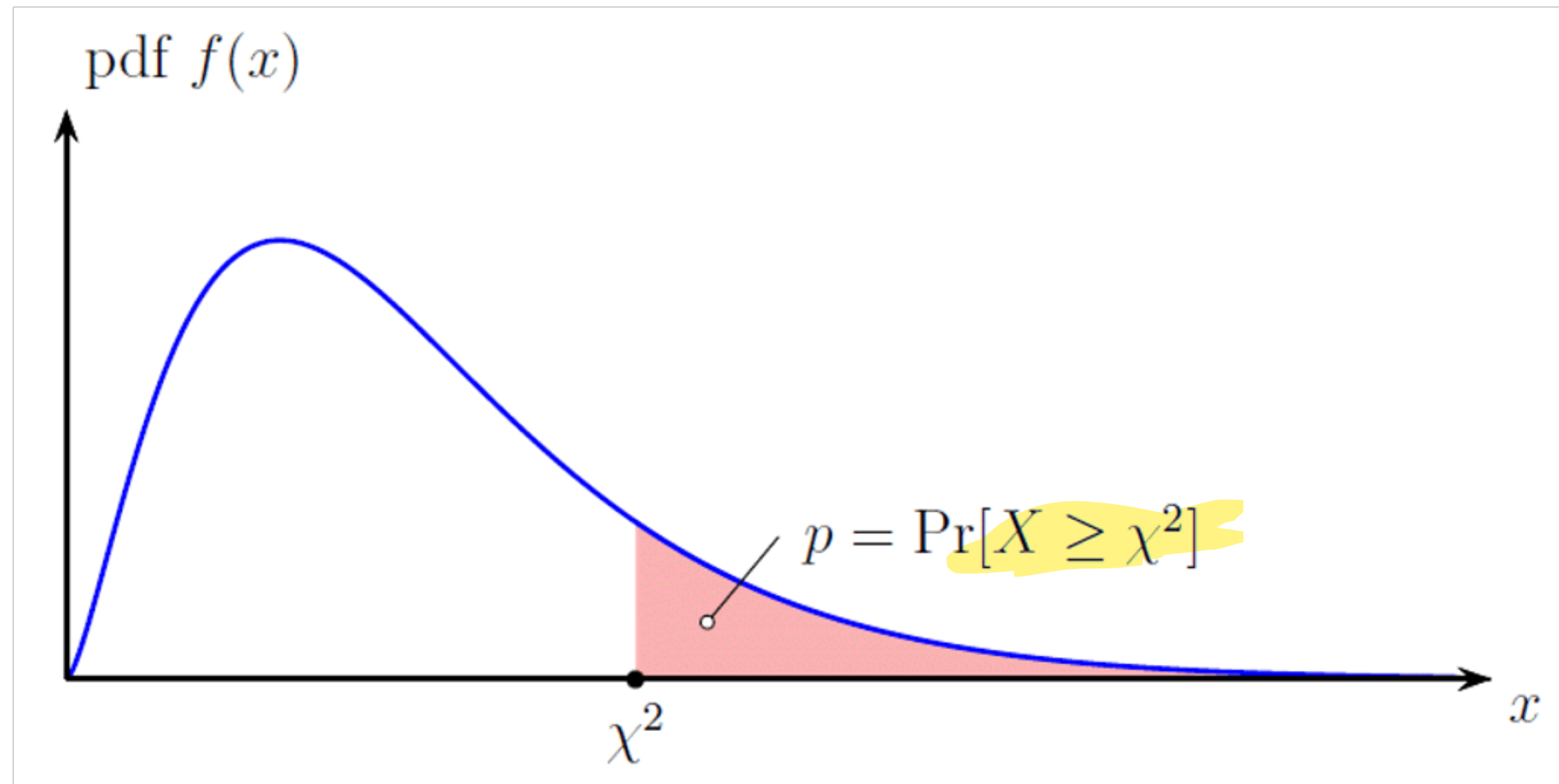
•**Goodness-of-fit tests**: The chi-square goodness-of-fit test is used to determine whether a sample of data comes from a population with a specific distribution. For example, it can test whether observed frequencies differ significantly from expected frequencies.

•**Test for independence**: In a contingency table, the chi-square test for independence can determine whether two categorical variables are independent of each other.

Chi-Square Distribution
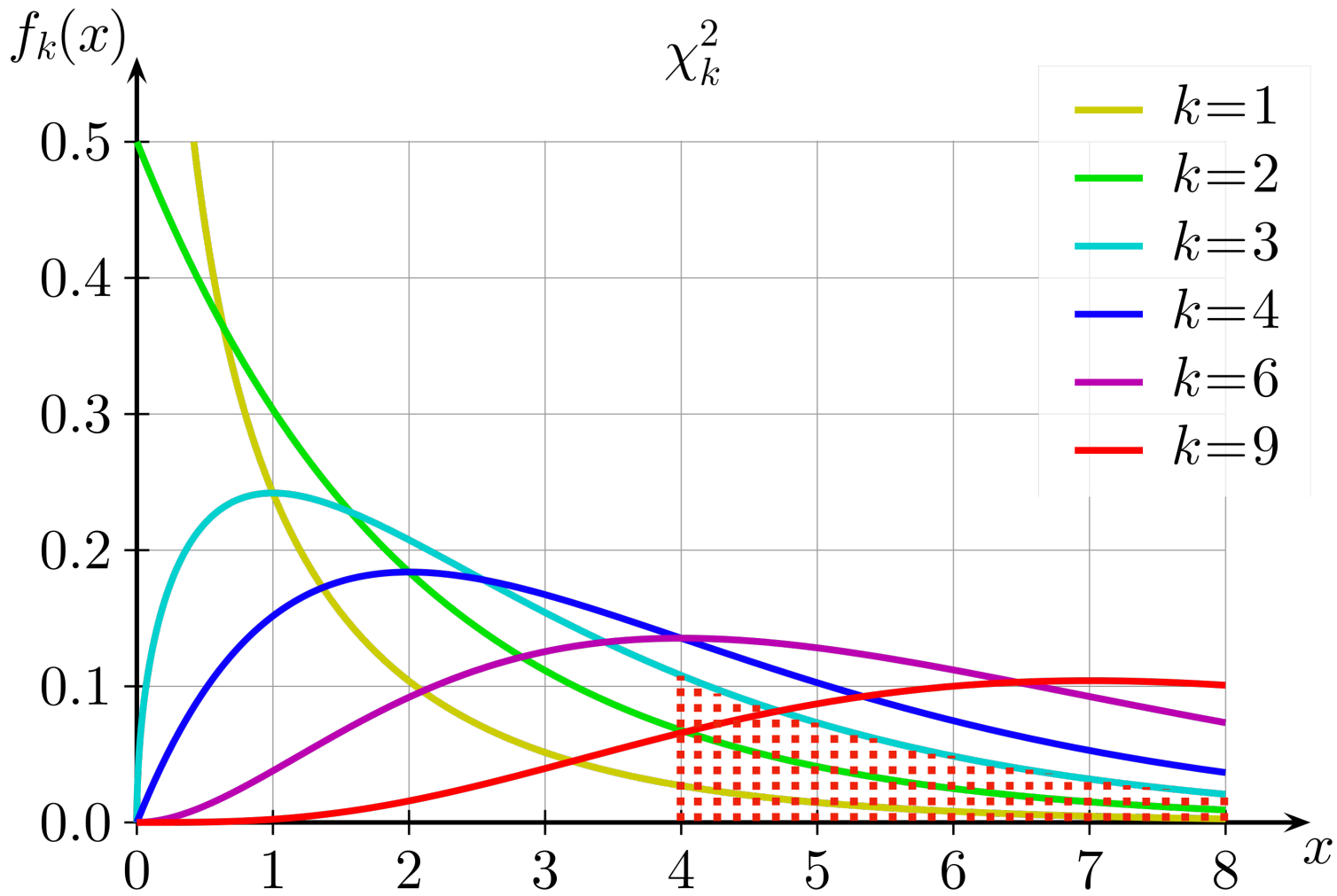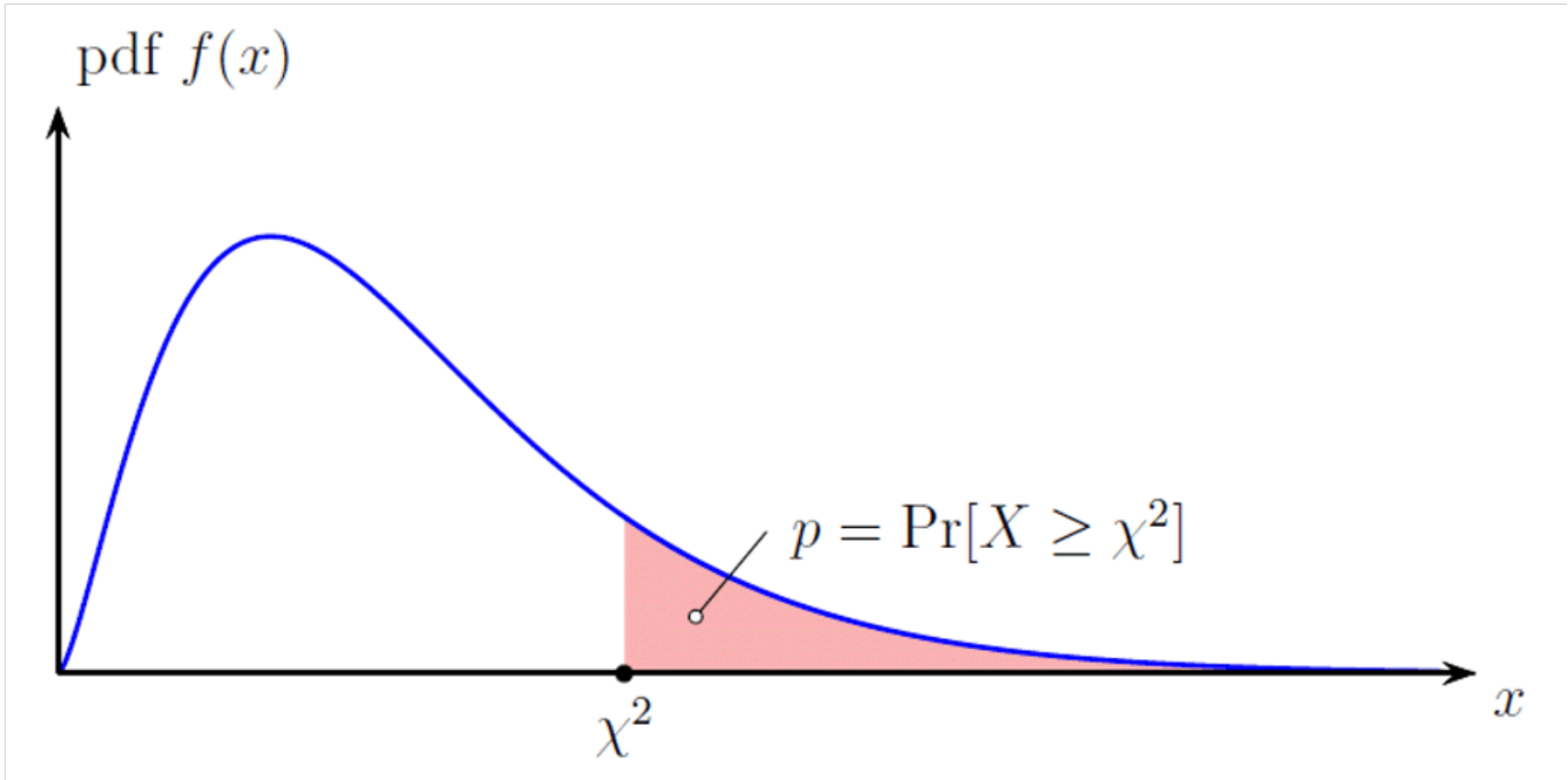
Degrees of freedom (df)	χ^2 value[20]										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.63	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.61	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.81	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
p-value (probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

A χ2 value of 4 corresponds to a probability <0.30

```
> pchisq(4, 3, lower.tail = FALSE)
[1] 0.2614641
```

Season Preferences: Test for Goodness-of-Fit

H_0: There is no difference

H_1: There is a difference

	Spring	Summer	Fall	Winter	Total
Observed	40	30	18	28	116
	41 %	33 %	16 %	10 %	100 %
Expected	29	29	29	29	116
	25 %	25 %	25 %	25 %	100 %

$$\chi^2 = \sum \left(\frac{(O-E)^2}{E} \right) = \left(\frac{(40-29)^2}{29} \right) + \left(\frac{(30-29)^2}{29} \right) + \left(\frac{(18-29)^2}{29} \right) + \left(\frac{(28-29)^2}{29} \right)$$

$= 8.4$
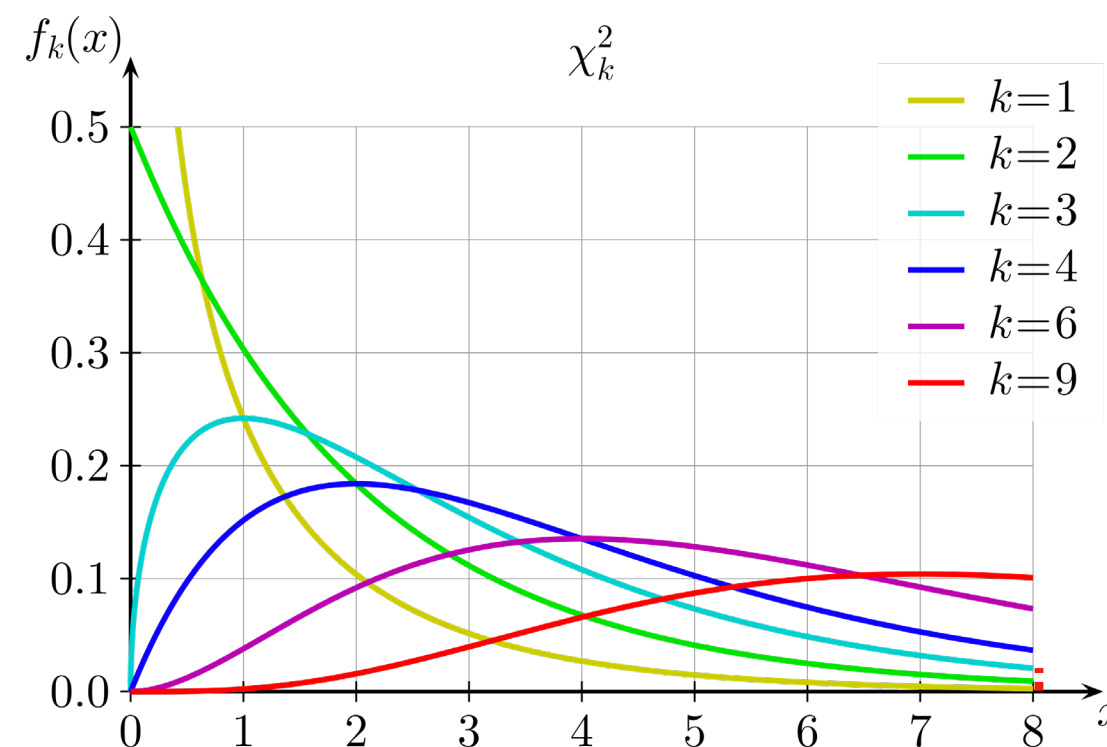

```
> pchisq(8.4, 3, lower.tail = FALSE)
[1] 0.03842932
```

With α = 0.05, with can reject H_0
There is a difference between the observed and the expected season preferences

Season Preferences: Test for Goodness-of-Fit

H_0: There is no difference

H_1: There is a difference

	Spring	Summer	Fall	Winter	Total
Observed	40	30	18	28	116
	41 %	33 %	16 %	10 %	100 %
Expected	29	29	29	29	116
	25 %	25 %	25 %	25 %	100 %

```
> chisq.test(Poll_seasons, correct = FALSE, p = rep(1/4, 4))

        Chi-squared test for given probabilities

data:  Poll_seasons
X-squared = 8.0968, df = 3, p-value = 0.04405
```

One-way table → degrees of freedom = #columns - 1

d.f. = c - 1

Two Categorical Variables

	Spring	Summer	Fall	Winter
Severe allergies	5	1	1	9
Mild allergies	8	5	2	5
Sporadic allergies	9	8	3	9
Never allergic	18	16	12	5
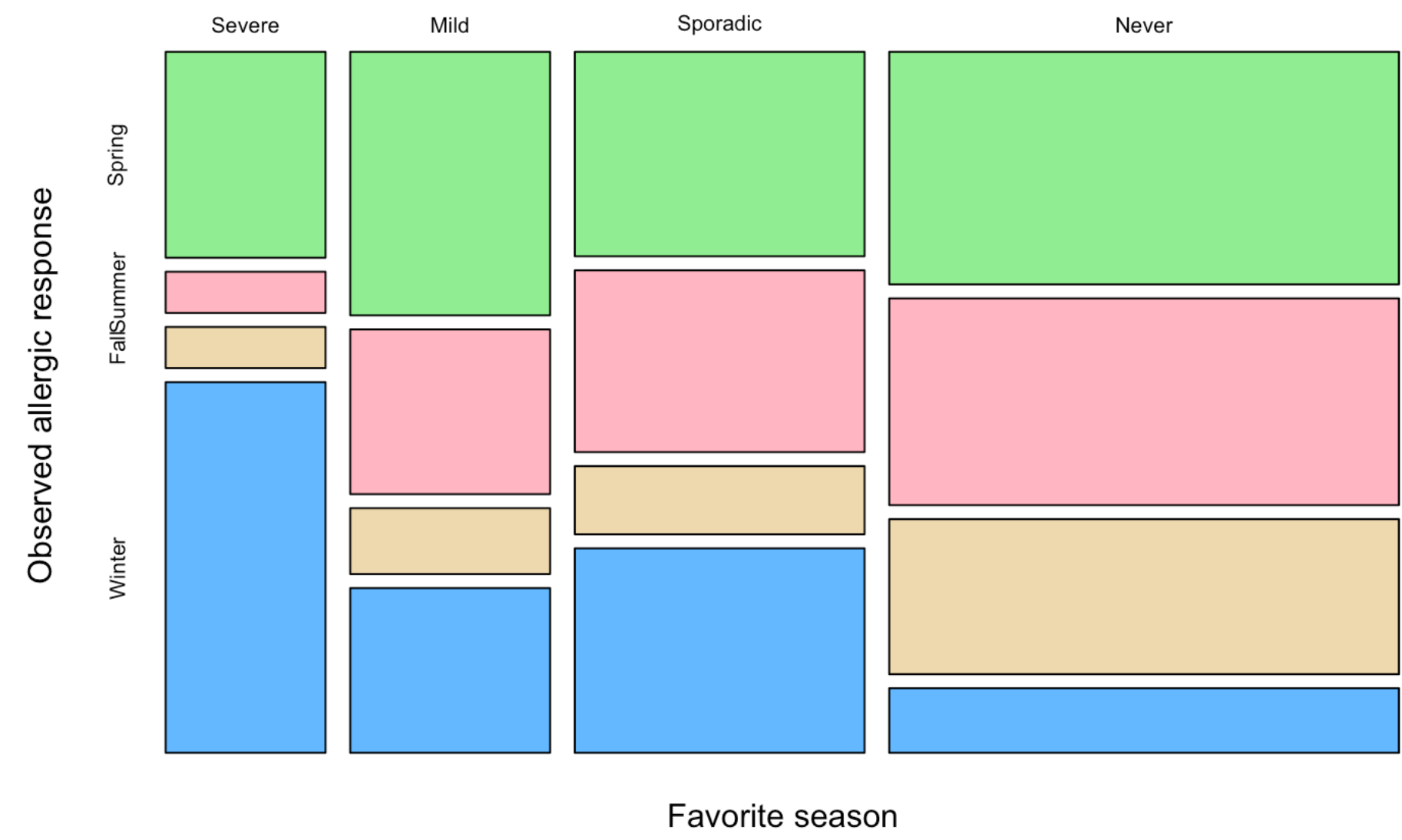
Two Categorical Variables

	Spring	Summer	Fall	Winter
Severe allergies	5	1	1	9
Mild allergies	8	5	2	5
Sporadic allergies	9	8	3	9
Never allergic	18	16	12	5

Test for Homogeneity

Question: Is there a difference between the distribution of allergic reactions in the different seasons?

H_0: The distribution of allergic reactions is the same for the people who preferred different seasons

H_1: The distribution of allergic reactions is **not** the same for the people who preferred different seasons

Two Categorical Variables

	Spring	Summer	Fall	Winter	Total
Severe allergies	5	1	1	9	16
Mild allergies	8	5	2	5	20
Sporadic allergies	9	8	3	9	29
Never allergic	18	16	12	5	51
Total	40	30	18	28	116

= 5.52

Expected Frequencies

$$\chi^2 = \sum \left(\frac{(O-E)^2}{E} \right)$$

	Spring	Summer	Fall	Winter
Severe allergies	5.52(40 x 16 /116)	4.14	2.48	3.86
Mild allergies	6.90	5.17	3.10	4.83
Sporadic allergies	10	7.5	4.5	7
Never allergic	17.59	13.19	7.91	12.31

$$= \left(\frac{(5-5.52)^2}{5.52} \right) + \left(\frac{(1-4.14)^2}{4.14} \right) + \left(\frac{(1-2.48)^2}{2.48} \right) + \left(\frac{(9-3.86)^2}{3.86} \right) + \ldots$$

Two Categorical Variables

	Spring	Summer	Fall	Winter	Total
Severe allergies	5	1	1	9	16
Mild allergies	8	5	2	5	20
Sporadic allergies	9	8	3	9	29
Never allergic	18	16	12	5	51
Total	40	30	18	28	116

Test for Homogeneity

```
> Severe <- data.frame(Spring = 5, Summer = 1, Fall = 1, Winter = 9)
> Mild <- data.frame(Spring = 8, Summer = 5, Fall = 2, Winter = 5)
> Sporadic <- data.frame(Spring = 9, Summer = 8, Fall = 3, Winter = 9)
> Never <- data.frame(Spring = 18, Summer = 16, Fall = 12, Winter = 5)
> Two_categories <- rbind(Severe, Mild, Sporadic, Never)
> chisq.test(Two_categories)
```

Question: Is there a difference between the distribution of allergic reactions and the preferred season?

```
        Pearson's Chi-squared test

data:  Two_categories
X-squared = 18.994, df = 9, p-value = 0.02524
```

With α = 0.05

Includes Yate's correction for continuity

H₀: The distribution of allergic reactions is the same for the people who preferred different seasons

H_1: The distribution of allergic reactions is **not** the same for the people who preferred different seasons

Two Categorical Variables

	Spring	Summer	Fall	Winter	Total
Severe allergies	5	1	1	9	16
Mild allergies	8	5	2	5	20
Sporadic allergies	9	8	3	9	29
Never allergic	18	16	12	5	51
Total	40	30	18	28	116

```
> Severe <- data.frame(Spring = 5, Summer = 1, Fall = 1, Winter = 9)
> Mild <- data.frame(Spring = 8, Summer = 5, Fall = 2, Winter = 5)
> Sporadic <- data.frame(Spring = 9, Summer = 8, Fall = 3, Winter = 9)
> Never <- data.frame(Spring = 18, Summer = 16, Fall = 12, Winter = 5)
> Two_categories <- rbind(Severe, Mild, Sporadic, Never)
> chisq.test(Two_categories)

        Pearson's Chi-squared test

data: Two_categories
X-squared = 18.994, df = 9, p-value = 0.02524
```

Degrees of Freedom

$$(\#rows - 1) \times (\#columns - 1)$$

$$d.f. = r \times c - 1 - (r-1) - (c-1)$$

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

Test for Independency

Question: We need to analyse the survival data of a geneX knockout mice at 1 year. Does geneX affect lifespan of mice?

H_0: The survival of mice is independent on geneX

H_1: The survival of mice is dependent on geneX

The lifespan is dependent on geneX

Explanatory variable

Response variable

	WT	KO	Total
Alive	7	2	9
Dead	3	7	10
Total	10	9	19

Expected frequencies

	WT	KO
Alive	4.7	4.3
Dead	5.3	4.7

$\chi^2 = 4.3372$, $d.f. = 1$, $p = 0.037$

With $\alpha = 0.05$

Use of Chi-Square Analysis

- The appropriate use of the chi-square to approximate the distribution of the good-of-fit test statistic depends on both the sample size and the number of cells

- Pearson's chi-square is an approximation that requires **large** sample sizes

- No expected cell frequencies are less than 1.
- No more than 20% are less than 5.

Fisher's Exact Test

Alternative for contingency tables when sample sizes are **small**

	C1	C2	Row Total
R1	a	b	a+b
R2	c	d	c+d
Column Total	a+c	b+d	a+b+c+d=n

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

The Fisher's exact test can also be used on contingency tables larger than 2x2

Fisher's Exact Test

We need to analyse the survival data of a geneX knockout mice at 1 year. Does geneX affect survival of mice?

	WT	KO	Total
Alive	7	2	9
Dead	3	7	10
Total	10	9	19

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!} \quad = \frac{9!10!10!9!}{7!2!3!7!19!} = 0.069 \quad \text{accept H}_0$$

With α = 0.05

The survival of mice is independent on geneX!

3-Way Sample: Three Categorical Variables (`rxcxl`)

Example: We need to analyse the survival data of a geneX knockout mice at 1 year. Is geneX, sex and lifespan independent of each other?

	WT		KO	
	Male	**Female**	**Male**	**Female**
Alive	40	34	20	25
Dead	9	7	15	20

Chi-square test

H_0: There is no interdependency among geneX, sex and lifespan of mice

H_1: There **is** interdependency among the variables

If A, B and C are independent, then

$$P(A \cap B \cap C) = P(A)P(B)P(C)$$

3-Way Sample: Three Categorical Variables (`rxcxl`)

Example: We need to analyse the survival data of a geneX knockout mice at 1 year. Is geneX, sex and lifespan independent of each other?

	WT		KO	
	Male	**Female**	**Male**	**Female**
Alive	40	34	20	25
Dead	9	7	15	20

Chi-square test

Total mice: 170

Total male vs female: 84:86 (49% vs 51%)

Total alive vs dead: 119:51 (70% vs 30%)

Total WT vs KO: 90:80 (53% vs 47%)

Expected:

Male, alive, WT = `170 x 49% x 70% x 53% = 31`

Male, alive, KO

Male, dead, WT

Male, dead, KO

Female, alive, WT

Female, alive, KO

Female, dead, WT

Female, dead, KO = `170 x 51% x 30% x 47% = 12.1`

$$\chi^2 = \sum \left(\frac{(O-E)^2}{E} \right)$$

3-Way Sample: Three Categorical Variables (`rxcxl`)

Example: We need to analyse the survival data of a geneX knockout mice at 1 year.
  Is geneX, sex and lifespan independent of each other?

	WT		KO	
	Male	**Female**	**Male**	**Female**
Alive	40	34	20	25
Dead	9	7	15	20

Chi-square test

$$\chi^2 = \sum \left(\frac{(O-E)^2}{E} \right) = 15.765, \qquad p = 0.0034, \qquad d.f. = 4$$

With α = 0.05

Degrees of Freedom

```
d.f. = r x c x l - 1 - (r-1) - (c-1) - (l-1)
     = 2 x 2 x 2 - 1 - (2-1) - (2-1) - (2-1)
     = 4
```

There is interdependency among geneX, sex and lifespan in mice at 1 year

Ordinal Variable

Categorical data with an associated set **order** or scale
(ratings, pain levels, age groups, allergy severity)

There is **no** standardised **interval** scale of measurement

	Spring	Summer	Fall	Winter
Severe allergies				
Mild allergies				
Sporadic allergies				
Never allergic				

It can measure **qualitative** traits

Numeric operations cannot be used
Has a **median** (not a mean)

Ordinal Variable

Statistic analysis: compare the **median** values across samples

Wilcoxon test:
  1- or 2-sample test, especially useful for paired samples

Kruskal-Wallis 1-way test:
  3- or more sample test, non-parametric alternative to the 1-way ANOVA

Correlation

Do people who prefer Spring have less allergies?

What is your favourite season?

- ( ) Spring
- ( ) Summer
- ( ) Fall
- ( ) Winter

Submit

Do you have seasonal allergies? (sneezing, red eyes, runny nose, itchy mouth)

- ( ) Yes, it's terrible
- ( ) Yes, but not so bad
- ( ) Only sometimes
- ( ) Never

Submit

Summary

Want to see if there is a difference between an observed variable and expected data?

Large dataset? → **Chi-square goodness-of-fit test**

Small dataset? (integer proportions) → **Fisher's exact test**

Nominal?

Want to see if two or more variables have different distributions of a categorical variable?

Large dataset? → **Chi-square homogeneity test**

Small dataset? → **Fisher's exact test**

Using categorical variables?

Want to see if two or more response variables are independent on an explanatory variable? → **Chi-square independency test**

Ordinal?

1 or 2 samples? → **Wilcoxon test**

3 or more samples? → **Kruscal-Wallis 1-way test**