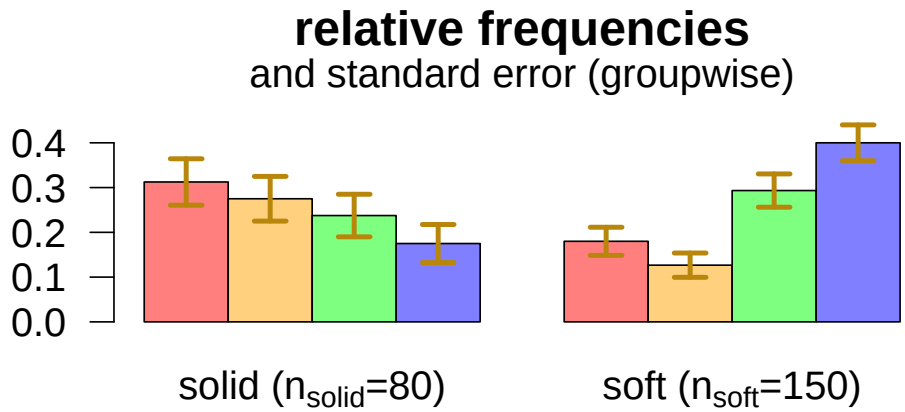
# The $\chi^2$-test (for independence)



All examples are fictitious. All data are simulated and the graphics were created with the statistical program package R.

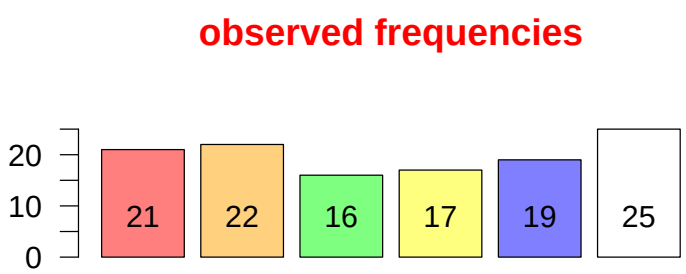


Sämtliche Beispiele sind frei erfunden. Alle Daten sind simuliert und die Grafiken wurden mit statistischen Programmpaket R erstellt.

# Reminder: $\chi^2$-test (goodnes of fit)

- Is the die fair? ($d = 6$ sides, data: $n = 120$ times rolled)
- observed frequencies: $x_1, \cdots, x_d$

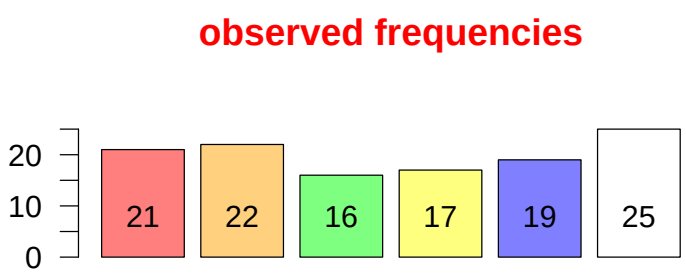
| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | $\sum$ |
|---|---|---|---|---|---|---|---|
| $x_k$ | 21 | 22 | 16 | 17 | 19 | 25 | 120 |



# Reminder: $\chi^2$-test (goodnes of fit)

- Is the die fair? ($d = 6$ sides, data: $n = 120$ times rolled)
- observed frequencies: $x_1, \cdots, x_d$
- Model: let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$

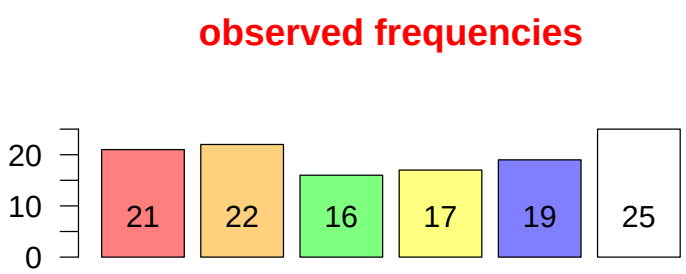
| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | $\sum$ |
|---|---|---|---|---|---|---|---|
| $x_k$ | 21 | 22 | 16 | 17 | 19 | 25 | 120 |



# Reminder: $\chi^2$-test (goodnes of fit)

- Is the die fair? ($d = 6$ sides, data: $n = 120$ times rolled)
- observed frequencies: $x_1, \cdots, x_d$
- Model: let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- null hypothesis: $H_0 : p = (1/6, 1/6, \ldots, 1/6)^t$ ($\leftrightarrow$ 'fair')

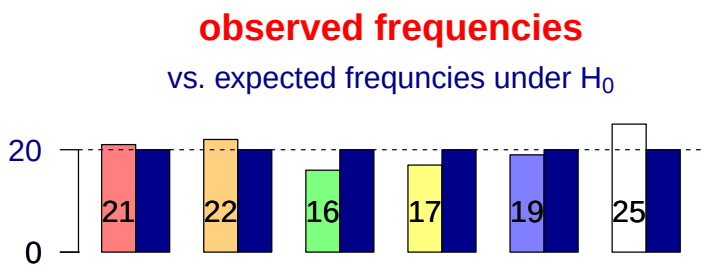
| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | $\sum$ |
|---|---|---|---|---|---|---|---|
| $x_k$ | 21 | 22 | 16 | 17 | 19 | 25 | 120 |



# Reminder: $\chi^2$-test (goodnes of fit)

- Is the die fair? ($d = 6$ sides, data: $n = 120$ times rolled)
- observed frequencies: $x_1, \cdots, x_d$
- Model: let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- null hypothesis: $H_0 : p = (1/6, 1/6, \ldots, 1/6)^t$ ($\leftrightarrow$ 'fair')
- expected occupations under $H_0$ $\mathbb{E}_{H_0}[X_k] = n/6 = 20$

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | $\sum$ |
|---|---|---|---|---|---|---|---|
| $x_k$ | 21 | 22 | 16 | 17 | 19 | 25 | 120 |
| $\mathbb{E}_{H_0}[X_k]$ | 20 | 20 | 20 | 20 | 20 | 20 | 120 |



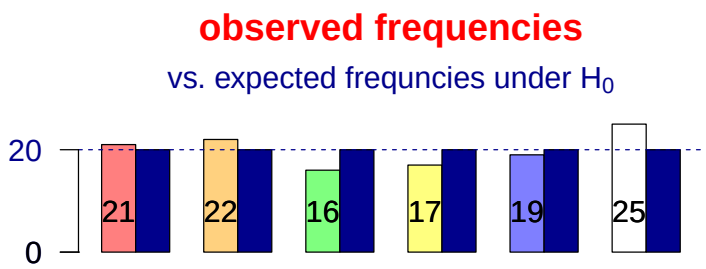
# Reminder: $\chi^2$-test (goodnes of fit)

- Is the die fair? ($d = 6$ sides, data: $n = 120$ times rolled)
- observed frequencies: $x_1, \cdots, x_d$
- Model: let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- null hypothesis: $H_0 : p = (1/6, 1/6, \ldots, 1/6)^t$ ($\leftrightarrow$ 'fair')
- expected occupations under $H_0$ $\mathbb{E}_{H_0}[X_k] = n/6 = 20$

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | $\sum$ |
|---|---|---|---|---|---|---|---|
| $x_k$ | 21 | 22 | 16 | 17 | 19 | 25 | 120 |
| $\mathbb{E}_{H_0}[X_k]$ | 20 | 20 | 20 | 20 | 20 | 20 | 120 |

- The $\chi^2$-statistic (measures discrepancy of 'observed' to 'expected under $H_0$')

$$x^2 := \sum_{k=1}^{d} \frac{(x_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]}$$



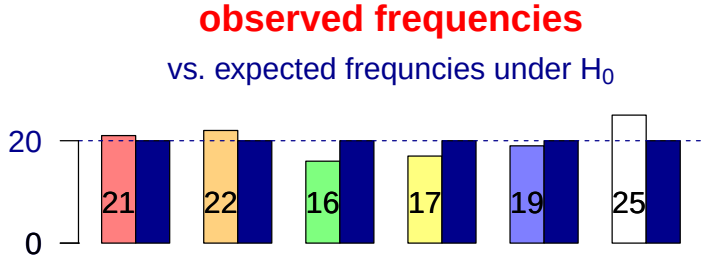
# Reminder: $\chi^2$-test (goodnes of fit)

- Is the die fair? ($d = 6$ sides, data: $n = 120$ times rolled)
- observed frequencies: $x_1, \cdots, x_d$
- Model: let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- null hypothesis: $H_0 : p = (1/6, 1/6, \ldots, 1/6)^t$ ($\leftrightarrow$ 'fair')
- expected occupations under $H_0$ $\mathbb{E}_{H_0}[X_k] = n/6 = 20$

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | $\sum$ |
|---|---|---|---|---|---|---|---|
| $x_k$ | 21 | 22 | 16 | 17 | 19 | 25 | 120 |
| $\mathbb{E}_{H_0}[X_k]$ | 20 | 20 | 20 | 20 | 20 | 20 | 120 |

- The $\chi^2$-statistic (measures discrepancy of 'observed' to 'expected under $H_0$')

$$x^2 := \sum_{k=1}^{d} \frac{(x_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} = \frac{(21-20)^2}{20} + \cdots + \frac{(25-20)^2}{20}$$



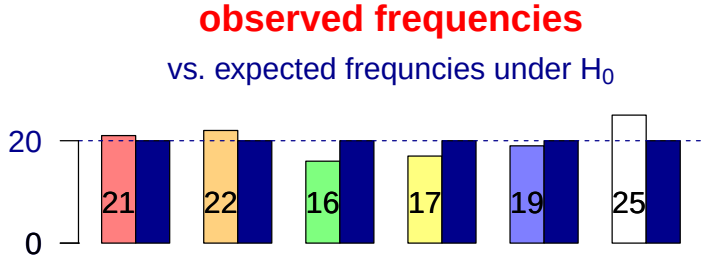
# Reminder: $\chi^2$-test (goodnes of fit)

- Is the die fair? ($d = 6$ sides, data: $n = 120$ times rolled)
- observed frequencies: $x_1, \cdots, x_d$
- Model: let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- null hypothesis: $H_0 : p = (1/6, 1/6, \ldots, 1/6)^t$ ($\leftrightarrow$ 'fair')
- expected occupations under $H_0$ $\mathbb{E}_{H_0}[X_k] = n/6 = 20$

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | $\sum$ |
|---|---|---|---|---|---|---|---|
| $x_k$ | 21 | 22 | 16 | 17 | 19 | 25 | 120 |
| $\mathbb{E}_{H_0}[X_k]$ | 20 | 20 | 20 | 20 | 20 | 20 | 120 |

- The $\chi^2$-statistic (measures discrepancy of 'observed' to 'expected under $H_0$')

$$x^2 := \sum_{k=1}^{d} \frac{(x_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} = \frac{(21-20)^2}{20} + \cdots + \frac{(25-20)^2}{20} = \frac{1}{20} + \cdots + \frac{25}{20}$$



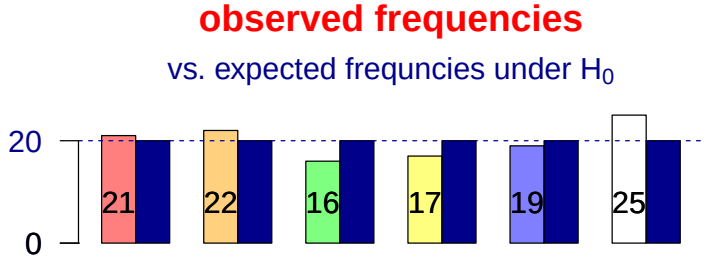
# Reminder: $\chi^2$-test (goodnes of fit)

- Is the die fair? ($d = 6$ sides, data: $n = 120$ times rolled)
- observed frequencies: $x_1, \cdots, x_d$
- Model: let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- null hypothesis: $H_0 : p = (1/6, 1/6, \ldots, 1/6)^t$ ($\leftrightarrow$ 'fair')
- expected occupations under $H_0$ $\mathbb{E}_{H_0}[X_k] = n/6 = 20$

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | $\sum$ |
|---|---|---|---|---|---|---|---|
| $x_k$ | 21 | 22 | 16 | 17 | 19 | 25 | 120 |
| $\mathbb{E}_{H_0}[X_k]$ | 20 | 20 | 20 | 20 | 20 | 20 | 120 |

- The $\chi^2$-statistic (measures discrepancy of 'observed' to 'expected under $H_0$')

$$x^2 := \sum_{k=1}^{d} \frac{(x_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} = \frac{(21-20)^2}{20} + \cdots + \frac{(25-20)^2}{20} = \frac{1}{20} + \cdots + \frac{25}{20} = \frac{56}{20} = 2.8$$



# Reminder: $\chi^2$-test (goodnes of fit)

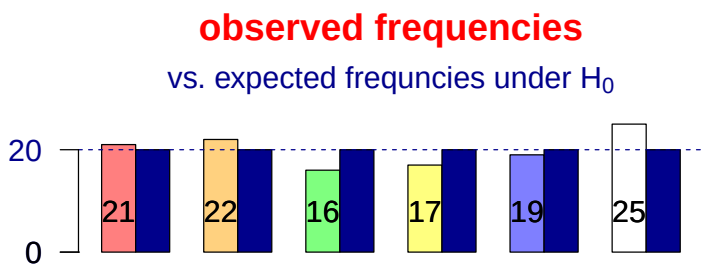
- Is the die fair? ($d = 6$ sides, data: $n = 120$ times rolled)
- observed frequencies: $x_1, \cdots, x_d$
- Model: let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- null hypothesis: $H_0 : p = (1/6, 1/6, \ldots, 1/6)^t$ ($\leftrightarrow$ 'fair')
- expected occupations under $H_0$ $\mathbb{E}_{H_0}[X_k] = n/6 = 20$

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | $\sum$ |
|---|---|---|---|---|---|---|---|
| $x_k$ | 21 | 22 | 16 | 17 | 19 | 25 | 120 |
| $\mathbb{E}_{H_0}[X_k]$ | 20 | 20 | 20 | 20 | 20 | 20 | 120 |

- The $\chi^2$-statistic (measures discrepancy of 'observed' to 'expected under $H_0$')

$$x^2 := \sum_{k=1}^{d} \frac{(x_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} = \frac{(21-20)^2}{20} + \cdots + \frac{(25-20)^2}{20} = \frac{1}{20} + \cdots + \frac{25}{20} = \frac{56}{20} = 2.8$$

- $x^2$ 'large'?



# Reminder: $\chi^2$-test (goodnes of fit)

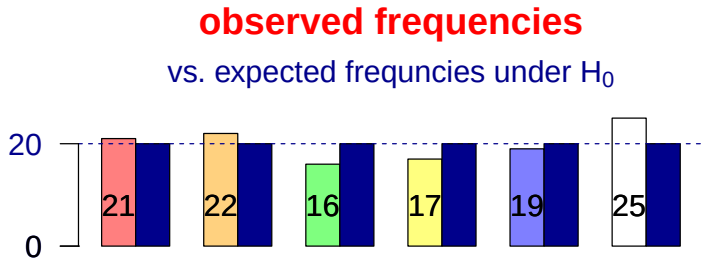
- Is the die fair? ($d = 6$ sides, data: $n = 120$ times rolled)
- observed frequencies: $x_1, \cdots, x_d$
- Model: let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- null hypothesis: $H_0 : p = (1/6, 1/6, \ldots, 1/6)^t$ ($\leftrightarrow$ 'fair')
- expected occupations under $H_0$ $\mathbb{E}_{H_0}[X_k] = n/6 = 20$

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | $\sum$ |
|---|---|---|---|---|---|---|---|
| $x_k$ | 21 | 22 | 16 | 17 | 19 | 25 | 120 |
| $\mathbb{E}_{H_0}[X_k]$ | 20 | 20 | 20 | 20 | 20 | 20 | 120 |

- The $\chi^2$-statistic (measures discrepancy of 'observed' to 'expected under $H_0$')

$$x^2 := \sum_{k=1}^{d} \frac{(x_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} = \frac{(21-20)^2}{20} + \cdots + \frac{(25-20)^2}{20} = \frac{1}{20} + \cdots + \frac{25}{20} = \frac{56}{20} = 2.8$$

- $x^2$ 'large'? No!



# Reminder: $\chi^2$-test (goodnes of fit)

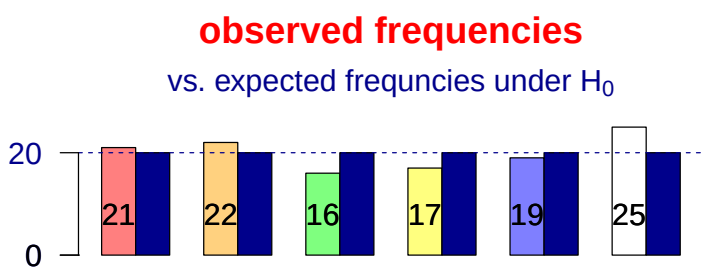
- Is the die fair? ($d = 6$ sides, data: $n = 120$ times rolled)
- observed frequencies: $x_1, \cdots, x_d$
- Model: let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- null hypothesis: $H_0 : p = (1/6, 1/6, \ldots, 1/6)^t$ ($\leftrightarrow$ 'fair')
- expected occupations under $H_0$ $\mathbb{E}_{H_0}[X_k] = n/6 = 20$

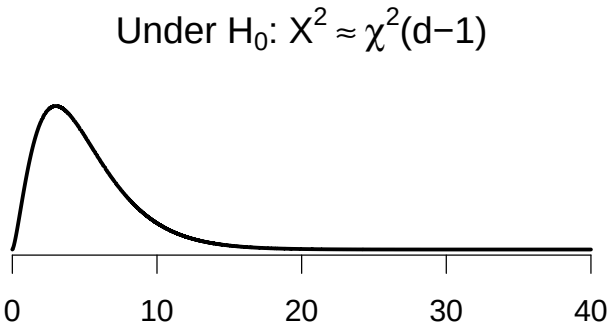
| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | $\sum$ |
|---|---|---|---|---|---|---|---|
| $x_k$ | 21 | 22 | 16 | 17 | 19 | 25 | 120 |
| $\mathbb{E}_{H_0}[X_k]$ | 20 | 20 | 20 | 20 | 20 | 20 | 120 |

- The $\chi^2$-statistic (measures discrepancy of 'observed' to 'expected under $H_0$')

$$x^2 := \sum_{k=1}^{d} \frac{(x_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} = \frac{(21-20)^2}{20} + \cdots + \frac{(25-20)^2}{20} = \frac{1}{20} + \cdots + \frac{25}{20} = \frac{56}{20} = 2.8$$

- $x^2$ 'large'? No! Comparison with $\chi^2(5)$-distribution (as $X^2 \overset{H_0}{\sim} \chi^2(d-1)$ approx)



# Reminder: $\chi^2$-test (goodnes of fit)

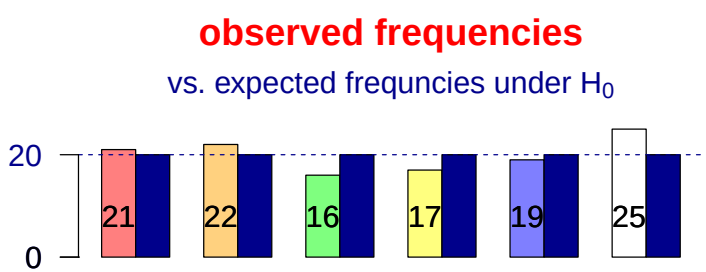
- Is the die fair? ($d = 6$ sides, data: $n = 120$ times rolled)
- observed frequencies: $x_1, \cdots, x_d$
- Model: let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- null hypothesis: $H_0 : p = (1/6, 1/6, \ldots, 1/6)^t$ ($\leftrightarrow$ 'fair')
- expected occupations under $H_0$ $\mathbb{E}_{H_0}[X_k] = n/6 = 20$

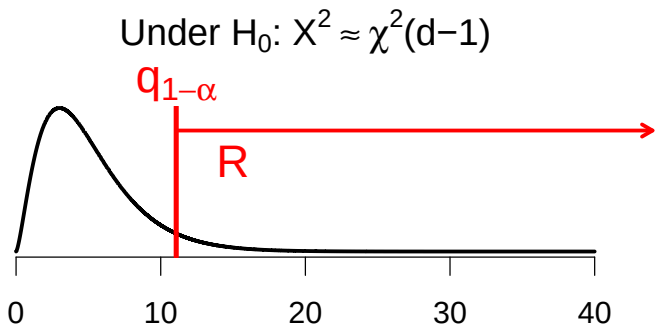
| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | $\sum$ |
|---|---|---|---|---|---|---|---|
| $x_k$ | 21 | 22 | 16 | 17 | 19 | 25 | 120 |
| $\mathbb{E}_{H_0}[X_k]$ | 20 | 20 | 20 | 20 | 20 | 20 | 120 |

- The $\chi^2$-statistic (measures discrepancy of 'observed' to 'expected under $H_0$')

$$x^2 := \sum_{k=1}^{d} \frac{(x_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} = \frac{(21-20)^2}{20} + \cdots + \frac{(25-20)^2}{20} = \frac{1}{20} + \cdots + \frac{25}{20} = \frac{56}{20} = 2.8$$

- $x^2$ 'large'? No! Comparison with $\chi^2(5)$-distribution (as $X^2 \overset{H_0}{\sim} \chi^2(d-1)$ approx)



# Reminder: $\chi^2$-test (goodnes of fit)

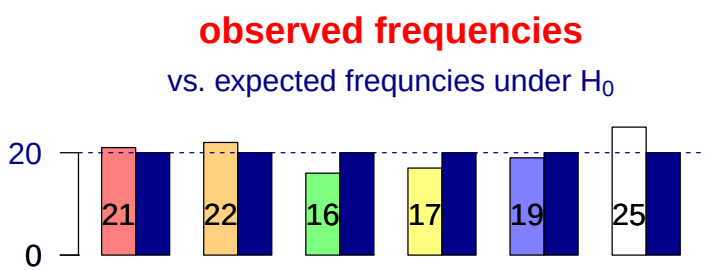
- Is the die fair? ($d = 6$ sides, data: $n = 120$ times rolled)
- observed frequencies: $x_1, \cdots, x_d$
- Model: let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- null hypothesis: $H_0 : p = (1/6, 1/6, \ldots, 1/6)^t$ ($\leftrightarrow$ 'fair')
- expected occupations under $H_0$ $\mathbb{E}_{H_0}[X_k] = n/6 = 20$

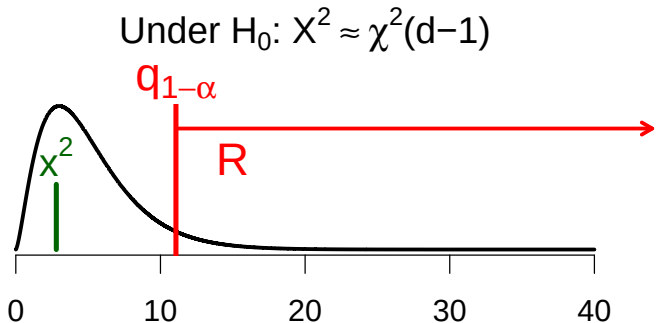
| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | $\sum$ |
|---|---|---|---|---|---|---|---|
| $x_k$ | 21 | 22 | 16 | 17 | 19 | 25 | 120 |
| $\mathbb{E}_{H_0}[X_k]$ | 20 | 20 | 20 | 20 | 20 | 20 | 120 |

- The $\chi^2$-statistic (measures discrepancy of 'observed' to 'expected under $H_0$')

$$x^2 := \sum_{k=1}^{d} \frac{(x_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} = \frac{(21-20)^2}{20} + \cdots + \frac{(25-20)^2}{20} = \frac{1}{20} + \cdots + \frac{25}{20} = \frac{56}{20} = 2.8$$

- $x^2$ 'large'? No! Comparison with $\chi^2(5)$-distribution (as $X^2 \overset{H_0}{\sim} \chi^2(d-1)$ approx)
  $\rightarrow x^2 \notin R$



# Reminder: $\chi^2$-test (goodnes of fit)

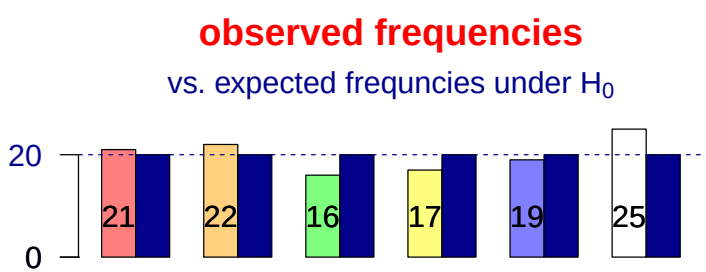
- Is the die fair? ($d = 6$ sides, data: $n = 120$ times rolled)
- observed frequencies: $x_1, \cdots, x_d$
- Model: let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- null hypothesis: $H_0 : p = (1/6, 1/6, \ldots, 1/6)^t$ ($\leftrightarrow$ 'fair')
- expected occupations under $H_0$ $\mathbb{E}_{H_0}[X_k] = n/6 = 20$

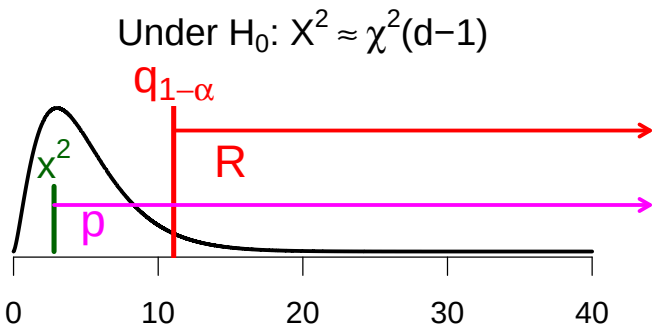
| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | $\sum$ |
|---|---|---|---|---|---|---|---|
| $x_k$ | 21 | 22 | 16 | 17 | 19 | 25 | 120 |
| $\mathbb{E}_{H_0}[X_k]$ | 20 | 20 | 20 | 20 | 20 | 20 | 120 |

- The $\chi^2$-statistic (measures discrepancy of 'observed' to 'expected under $H_0$')

$$x^2 := \sum_{k=1}^{d} \frac{(x_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} = \frac{(21-20)^2}{20} + \cdots + \frac{(25-20)^2}{20} = \frac{1}{20} + \cdots + \frac{25}{20} = \frac{56}{20} = 2.8$$

- $x^2$ 'large'? No! Comparison with $\chi^2(5)$-distribution (as $X^2 \overset{H_0}{\sim} \chi^2(d-1)$ approx)
  $\rightarrow x^2 \notin R \Leftrightarrow p > \alpha = 5\%$



# Reminder: $\chi^2$-test (goodnes of fit)

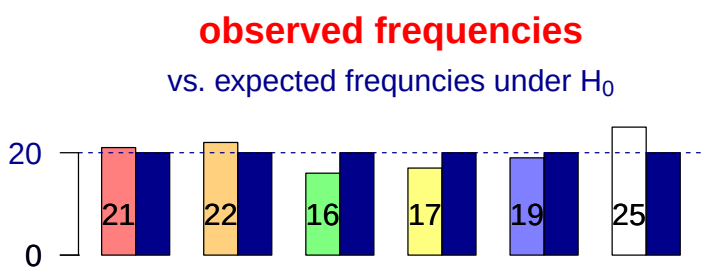
- Is the die fair? ($d = 6$ sides, data: $n = 120$ times rolled)
- observed frequencies: $x_1, \cdots, x_d$
- Model: let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- null hypothesis: $H_0 : p = (1/6, 1/6, \ldots, 1/6)^t$ ($\leftrightarrow$ 'fair')
- expected occupations under $H_0$ $\mathbb{E}_{H_0}[X_k] = n/6 = 20$

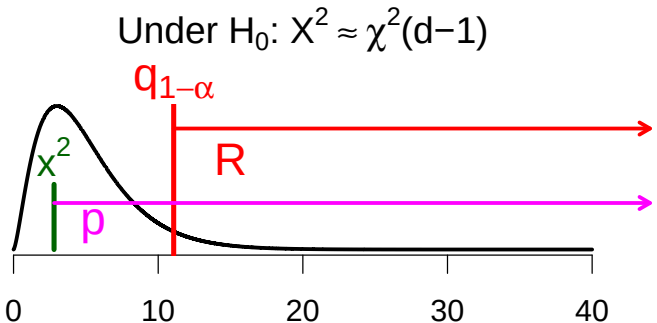
| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | $\sum$ |
|---|---|---|---|---|---|---|---|
| $x_k$ | 21 | 22 | 16 | 17 | 19 | 25 | 120 |
| $\mathbb{E}_{H_0}[X_k]$ | 20 | 20 | 20 | 20 | 20 | 20 | 120 |

- The $\chi^2$-statistic (measures discrepancy of 'observed' to 'expected under $H_0$')

$$x^2 := \sum_{k=1}^{d} \frac{(x_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} = \frac{(21-20)^2}{20} + \cdots + \frac{(25-20)^2}{20} = \frac{1}{20} + \cdots + \frac{25}{20} = \frac{56}{20} = 2.8$$

- $x^2$ 'large'? No! Comparison with $\chi^2(5)$-distribution (as $X^2 \overset{H_0}{\sim} \chi^2(d-1)$ approx)
  $\rightarrow x^2 \notin R \Leftrightarrow p > \alpha = 5\% \rightarrow$ 'If $H_0$ holds true, then the discrepancy is not unlikely'



# Overview

So long:

- $\chi^2$-test, good of fit:

# Overview

So long:

- $\chi^2$-test, good of fit:
- One feature (here: outcome of the rolling die)

# Overview

So long:

- $\chi^2$-test, good of fit:
- One feature (here: outcome of the rolling die)
  $\rightarrow$ in $d$ categories (here: colors)

# Overview

So long:

- $\chi^2$-test, good of fit:
- One feature (here: outcome of the rolling die)
  $\rightarrow$ in $d$ categories (here: colors)
- data: frequencies / occupations

# Overview

So long:

- $\chi^2$-test, good of fit:
- One feature (here: outcome of the rolling die)
  $\rightarrow$ in $d$ categories (here: colors)
- data: frequencies / occupations
- Question: How good do the observed frequencies fit the claimed occupation probabilities?

# Overview

So long:

- $\chi^2$-test, good of fit:
- One feature (here: outcome of the rolling die)
  $\rightarrow$ in $d$ categories (here: colors)
- data: frequencies / occupations
- Question: How good do the observed frequencies fit the claimed occupation probabilities?
- Statistic: $\chi^2$-statistic

# Overview

So long:

- $\chi^2$-test, good of fit:
- One feature (here: outcome of the rolling die)
  $\rightarrow$ in $d$ categories (here: colors)
- data: frequencies / occupations
- Question: How good do the observed frequencies fit the claimed occupation probabilities?
- Statistic: $\chi^2$-statistic

In the following:

- $\chi^2$-test for independence:

# Overview

So long:

- $\chi^2$-test, good of fit:
- One feature (here: outcome of the rolling die)
  $\rightarrow$ in $d$ categories (here: colors)
- data: frequencies / occupations
- Question: How good do the observed frequencies fit the claimed occupation probabilities?
- Statistic: $\chi^2$-statistic

In the following:

- $\chi^2$-test for independence:
- Two features (e.g.,: 1. outcome of the die, and 2. underground used)

# Overview

So long:

- $\chi^2$-test, good of fit:
- One feature (here: outcome of the rolling die)
  $\rightarrow$ in $d$ categories (here: colors)
- data: frequencies / occupations
- Question: How good do the observed frequencies fit the claimed occupation probabilities?
- Statistic: $\chi^2$-statistic

In the following:

- $\chi^2$-test for independence:
- Two features (e.g.,: 1. outcome of the die, and 2. underground used)
- data: frequencies / occupations $\rightarrow$ as above

# Overview

So long:

- $\chi^2$-test, good of fit:
- One feature (here: outcome of the rolling die)
  $\rightarrow$ in $d$ categories (here: colors)
- data: frequencies / occupations
- Question: How good do the observed frequencies fit the claimed occupation probabilities?
- Statistic: $\chi^2$-statistic

In the following:

- $\chi^2$-test for independence:
- Two features (e.g.,: 1. outcome of the die, and 2. underground used)
- data: frequencies / occupations $\rightarrow$ as above
- Question: Is the first feature independent from the second feature?

# Overview

So long:

- $\chi^2$-test, good of fit:
- One feature (here: outcome of the rolling die)
  $\rightarrow$ in $d$ categories (here: colors)
- data: frequencies / occupations
- Question: How good do the observed frequencies fit the claimed occupation probabilities?
- Statistic: $\chi^2$-statistic

In the following:

- $\chi^2$-test for independence:
- Two features (e.g.,: 1. outcome of the die, and 2. underground used)
- data: frequencies / occupations $\rightarrow$ as above
- Question: Is the first feature independent from the second feature?
- Statistic: $\chi^2$-statistic $\rightarrow$ as above

# Overview

So long:

- $\chi^2$-test, good of fit:
- One feature (here: outcome of the rolling die)
  $\rightarrow$ in $d$ categories (here: colors)
- data: frequencies / occupations
- Question: How good do the observed frequencies fit the claimed occupation probabilities?
- Statistic: $\chi^2$-statistic

In the following:

- $\chi^2$-test for independence:
- Two features (e.g.,: 1. outcome of the die, and 2. underground used)
- data: frequencies / occupations $\rightarrow$ as above
- Question: Is the first feature independent from the second feature?
- Statistic: $\chi^2$-statistic $\rightarrow$ as above

Message: On the one hand there is a different question (and setup)...
...on the other hand we will 'technically' work with the same statistics

# Motivation

- A game designer develops a four sided die (sides: red, orange, green and blue)

# Motivation

- A game designer develops a four sided die
  (sides: red, orange, green and blue)
- She presents the cube to a broad audience ($n = 230$ people)...

# Motivation

- A game designer develops a four sided die
  (sides: red, orange, green and blue)
- She presents the cube to a broad audience ($n = 230$ people)...
- ...and claims that the outcome depends on the underground used: soft underground systematically yielded different outcomes than solid underground – a magic cube!
- Are we skeptical? ($\rightarrow$ What do the data say?)

# Motivation

- A game designer develops a four sided die
  (sides: red, orange, green and blue)
- She presents the cube to a broad audience ($n = 230$ people)...
- ...and claims that the outcome depends on the underground used: soft underground systematically yielded different outcomes than solid underground – a magic cube!
- Are we skeptical? ($\rightarrow$ What do the data say?)
- Each person from the audience is allowed to roll the die once:

# Motivation

- A game designer develops a four sided die (sides: red, orange, green and blue)
- She presents the cube to a broad audience ($n = 230$ people)...
- ...and claims that the outcome depends on the underground used: soft underground systematically yielded different outcomes than solid underground – a magic cube!
- Are we skeptical? ($\rightarrow$ What do the data say?)
- Each person from the audience is allowed to roll the die once:
  - first the underground has to be chosen, solid or soft (*feature 1*)...

| | | | | | | |
|---|---|---|---|---|---|---|
| underground | solid | | | | | |
| | soft | | | | | |

# Motivation

- A game designer develops a four sided die (sides: red, orange, green and blue)
- She presents the cube to a broad audience ($n = 230$ people)...
- ...and claims that the outcome depends on the underground used: soft underground systematically yielded different outcomes than solid underground – a magic cube!
- Are we skeptical? ($\rightarrow$ What do the data say?)
- Each person from the audience is allowed to roll the die once:
  - first the underground has to be chosen, solid or soft (*feature 1*)...
  - ...then on this underground the die is rolled and the outcome noted (*feature 2*)

| $x_{j,k}$ | side | red | orange | green | blue | |
|---|---|---|---|---|---|---|
| underground | solid | | | | | |
| | soft | | | | | |

# Motivation

- A game designer develops a four sided die (sides: red, orange, green and blue)
- She presents the cube to a broad audience ($n = 230$ people)...
- ...and claims that the outcome depends on the underground used: soft underground systematically yielded different outcomes than solid underground – a magic cube!
- Are we skeptical? ($\rightarrow$ What do the data say?)
- Each person from the audience is allowed to roll the die once:
  - first the underground has to be chosen, solid or soft (*feature 1*)...
  - ...then on this underground the die is rolled and the outcome noted (*feature 2*)
- The observed frequencies were as follows

| $x_{j,k}$ | side | red | orange | green | blue |
|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 |
| | soft | 27 | 19 | 44 | 60 |

# Motivation

- A game designer develops a four sided die (sides: red, orange, green and blue)
- She presents the cube to a broad audience ($n = 230$ people)...
- ...and claims that the outcome depends on the underground used: soft underground systematically yielded different outcomes than solid underground – a magic cube!
- Are we skeptical? ($\rightarrow$ What do the data say?)
- Each person from the audience is allowed to roll the die once:
  - first the underground has to be chosen, solid or soft (*feature 1*)...
  - ...then on this underground the die is rolled and the outcome noted (*feature 2*)
- The observed frequencies were as follows

| $x_{j,k}$ | side | red | orange | green | blue |
|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 |
| | soft | 27 | 19 | 44 | 60 |

- For example, 22 people chose the solid underground and then the die showed the orange side

# Motivation

- A game designer develops a four sided die (sides: red, orange, green and blue)
- She presents the cube to a broad audience ($n = 230$ people)...
- ...and claims that the outcome depends on the underground used: soft underground systematically yielded different outcomes than solid underground – a magic cube!
- Are we skeptical? ($\rightarrow$ What do the data say?)
- Each person from the audience is allowed to roll the die once:
  - first the underground has to be chosen, solid or soft (*feature 1*)...
  - ...then on this underground the die is rolled and the outcome noted (*feature 2*)
- The observed frequencies were as follows

| $x_{j,k}$ | side | red | orange | green | blue |
|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 |
| | soft | 27 | 19 | 44 | 60 |
| | $x_{\cdot,k}$ | 52 | 41 | 63 | 74 |

- For example, 22 people chose the solid underground and then the die showed the orange side
- Also, we obtain the *column frequencies* $x_{\cdot,k}$

# Motivation

- A game designer develops a four sided die (sides: red, orange, green and blue)
- She presents the cube to a broad audience ($n = 230$ people)...
- ...and claims that the outcome depends on the underground used: soft underground systematically yielded different outcomes than solid underground – a magic cube!
- Are we skeptical? ($\rightarrow$ What do the data say?)
- Each person from the audience is allowed to roll the die once:
  - first the underground has to be chosen, solid or soft (*feature 1*)...
  - ...then on this underground the die is rolled and the outcome noted (*feature 2*)
- The observed frequencies were as follows

| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | |

- For example, 22 people chose the solid underground and then the die showed the orange side
- Also, we obtain the *column frequencies* $x_{\cdot,k}$, the *row frequencies* $x_{j,\cdot}$

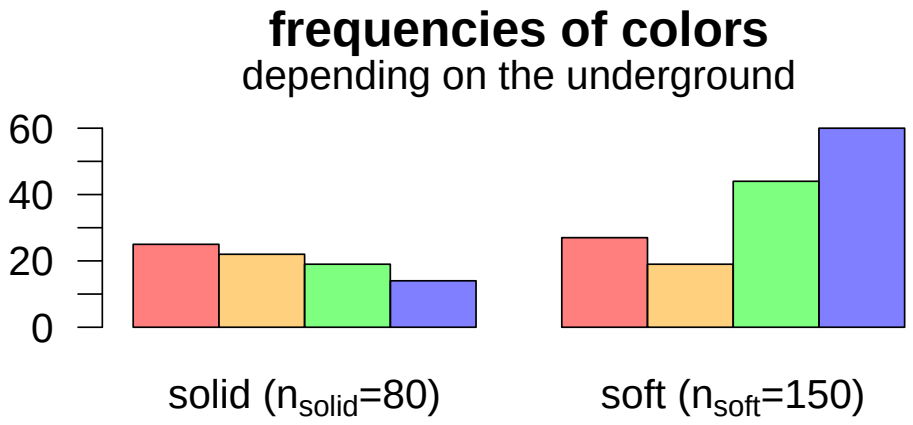
# Motivation

- A game designer develops a four sided die (sides: red, orange, green and blue)
- She presents the cube to a broad audience ($n = 230$ people)...
- ...and claims that the outcome depends on the underground used: soft underground systematically yielded different outcomes than solid underground – a magic cube!
- Are we skeptical? ($\rightarrow$ What do the data say?)
- Each person from the audience is allowed to roll the die once:
  - first the underground has to be chosen, solid or soft (*feature 1*)...
  - ...then on this underground the die is rolled and the outcome noted (*feature 2*)
- The observed frequencies were as follows

| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

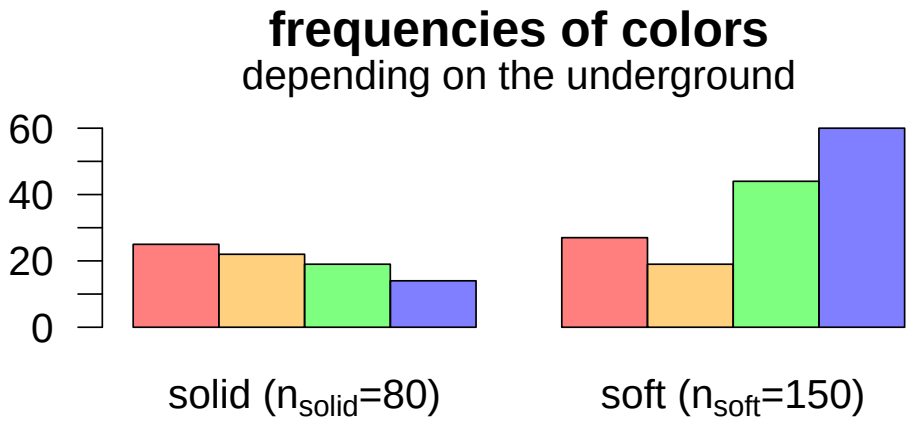
- For example, 22 people chose the solid underground and then the die showed the orange side
- Also, we obtain the *column frequencies* $x_{\cdot,k}$, the *row frequencies* $x_{j,\cdot}$, as well as the *total number* $n = 230$

# Graphically




| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,.}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{.,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

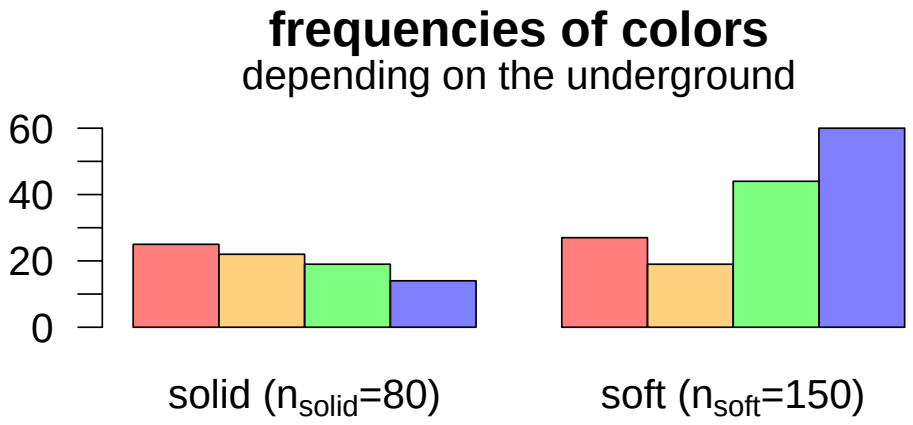
# Graphically




| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,.}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{.,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

- Question: what does 'independence of the features' mean intuitively?

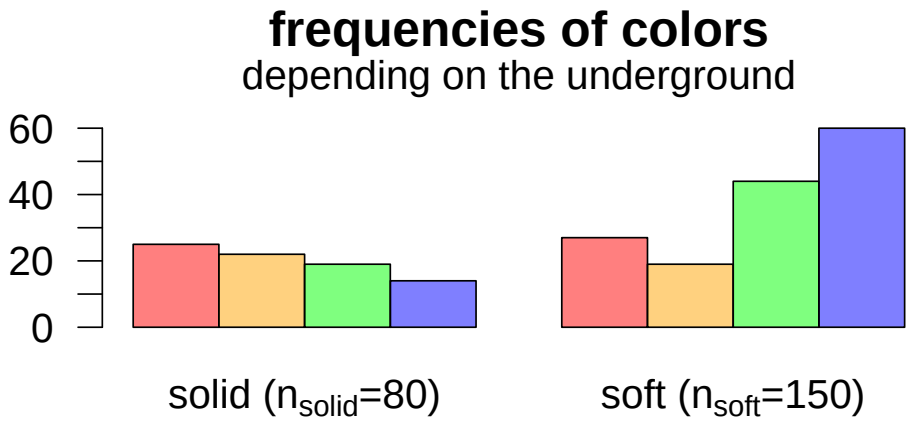
# Graphically




| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,.}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{.,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

- Question: what does 'independence of the features' mean intuitively?
- If the color did not depend on the underground, then the outcomes of both undergrounds should show about the same distribution.

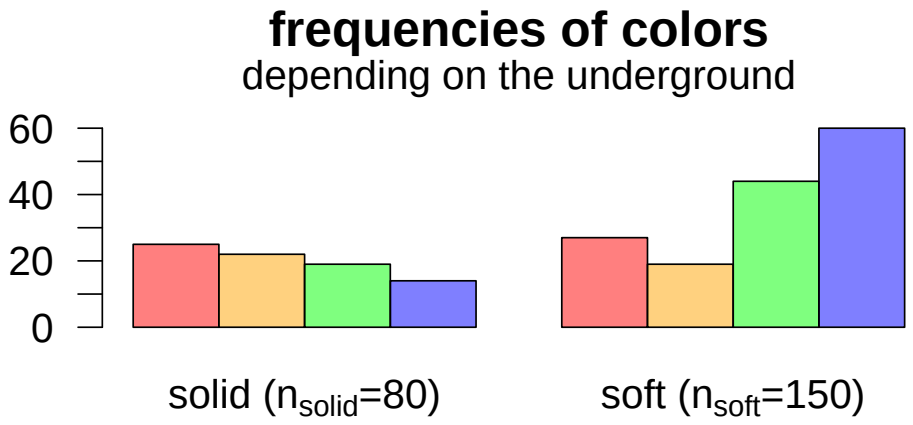
# Graphically




| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,.}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{.,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

- Question: what does 'independence of the features' mean intuitively?
- If the color did not depend on the underground, then the outcomes of both undergrounds should show about the same distribution.
- Here: at 'solid' all colors show about the same frequency, while at 'soft' e.g., the color blue appeared more than thrice as orange

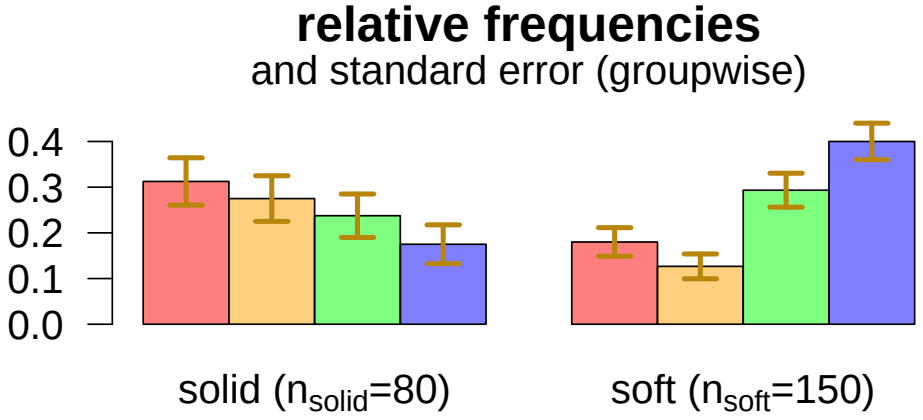
# Graphically




| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,.}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{.,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

- Question: what does 'independence of the features' mean intuitively?
- If the color did not depend on the underground, then the outcomes of both undergrounds should show about the same distribution.
- Here: at 'solid' all colors show about the same frequency, while at 'soft' e.g., the color blue appeared more than thrice as orange
- Can this difference be explained easily by chance under independence?

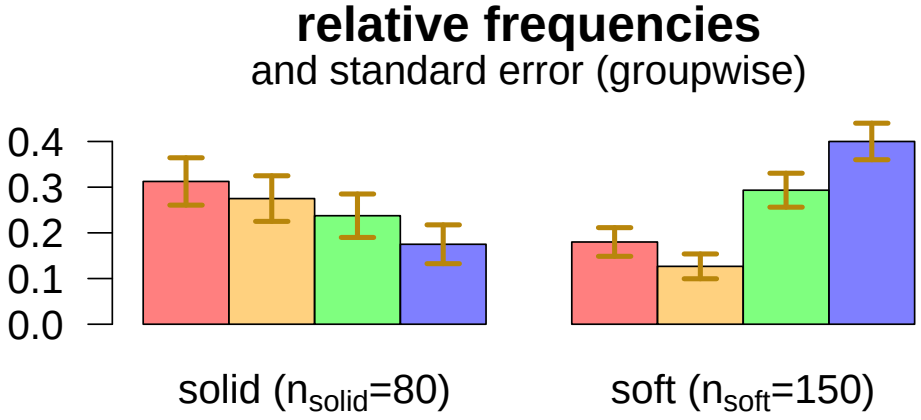
# Graphically




| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

- Question: what does 'independence of the features' mean intuitively?
- If the color did not depend on the underground, then the outcomes of both undergrounds should show about the same distribution.
- Here: at 'solid' all colors show about the same frequency, while at 'soft' e.g., the color blue appeared more than thrice as orange
- Can this difference be explained easily by chance under independence?
- Not really, when considering the standard errors

# Graphically




| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

- Question: what does 'independence of the features' mean intuitively?
- If the color did not depend on the underground, then the outcomes of both undergrounds should show about the same distribution.
- Here: at 'solid' all colors show about the same frequency, while at 'soft' e.g., the color blue appeared more than thrice as orange
- Can this difference be explained easily by chance under independence?
- Not really, when considering the standard errors
  $\rightarrow$ more precisely: $\chi^2$-*test for independence*

# Model

- $n = 230$ data in $d = d_1 \cdot d_2 = 8$ categories

| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

# Model

- $n = 230$ data in $d = d_1 \cdot d_2 = 8$ categories
  feature 1 (underground) has $d_1 = 2$ categories

| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

# Model

- $n = 230$ data in $d = d_1 \cdot d_2 = 8$ categories
  feature 1 (underground) has $d_1 = 2$ categories
  and feature 2 (color) has $d_2 = 4$ categories

| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,.}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{.,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

# Model

- $n = 230$ data in $d = d_1 \cdot d_2 = 8$ categories
  feature 1 (underground) has $d_1 = 2$ categories
  and feature 2 (color) has $d_2 = 4$ categories
- observed frequencies: $x_{1,1}, x_{1,2}, \cdots, x_{2,4}$ ($x_{j,k} \to$ row $j$, column $k$)

| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

# Model

- $n = 230$ data in $d = d_1 \cdot d_2 = 8$ categories
  feature 1 (underground) has $d_1 = 2$ categories
  and feature 2 (color) has $d_2 = 4$ categories
- observed frequencies: $x_{1,1}, x_{1,2}, \cdots, x_{2,4}$ ($x_{j,k} \to$ row $j$, column $k$)
- Model: Let $\mathfrak{X} = (X_{1,1}, \ldots, X_{d_1,d_2})^t \sim mult(n, p)$

| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,.}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{.,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

# Model

- $n = 230$ data in $d = d_1 \cdot d_2 = 8$ categories
  feature 1 (underground) has $d_1 = 2$ categories
  and feature 2 (color) has $d_2 = 4$ categories
- observed frequencies: $x_{1,1}, x_{1,2}, \cdots, x_{2,4}$ ($x_{j,k} \to$ row $j$, column $k$)
- Model: Let $\mathfrak{X} = (X_{1,1}, \ldots, X_{d_1,d_2})^t \sim mult(n, p)$
  with $p = (p_{1,1}, \ldots, p_{d_1,d_2})^t \in (0,1)^{d_1 \cdot d_2}$ and $\sum_{j,k} p_{j,k} = 1$

| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,.}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{.,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| occupation probabilities | red | orange | grün | blue | $p_{j,.}$ |
|---|---|---|---|---|---|
| solid | $p_{1,1}$ | $p_{1,2}$ | $p_{1,3}$ | $p_{1,4}$ | |
| soft | $p_{2,1}$ | $p_{2,2}$ | $p_{2,3}$ | $p_{2,4}$ | |
| $p_{.,k}$ | | | | | |

- while $p_{j,k}$ denotes the probability to fall into row $j$ and column $k$

# Model

- $n = 230$ data in $d = d_1 \cdot d_2 = 8$ categories
  feature 1 (underground) has $d_1 = 2$ categories
  and feature 2 (color) has $d_2 = 4$ categories
- observed frequencies: $x_{1,1}, x_{1,2}, \cdots, x_{2,4}$ ($x_{j,k} \to$ row $j$, column $k$)
- Model: Let $\mathfrak{X} = (X_{1,1}, \ldots, X_{d_1,d_2})^t \sim mult(n, p)$
  with $p = (p_{1,1}, \ldots, p_{d_1,d_2})^t \in (0,1)^{d_1 \cdot d_2}$ and $\sum_{j,k} p_{j,k} = 1$

| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,.}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{.,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| occupation probabilities | red | orange | grün | blue | $p_{j,.}$ |
|---|---|---|---|---|---|
| solid | $p_{1,1}$ | $p_{1,2}$ | $p_{1,3}$ | $p_{1,4}$ | |
| soft | $p_{2,1}$ | $p_{2,2}$ | $p_{2,3}$ | $p_{2,4}$ | |
| $p_{.,k}$ | $p_{.,1}$ | $p_{.,2}$ | $p_{.,3}$ | $p_{.,4}$ | |

- while $p_{j,k}$ denotes the probability to fall into row $j$ and column $k$
- and the *row sums* $p_{j,.}$

# Model

- $n = 230$ data in $d = d_1 \cdot d_2 = 8$ categories
  feature 1 (underground) has $d_1 = 2$ categories
  and feature 2 (color) has $d_2 = 4$ categories
- observed frequencies: $x_{1,1}, x_{1,2}, \cdots, x_{2,4}$ ($x_{j,k} \to$ row $j$, column $k$)
- Model: Let $\mathfrak{X} = (X_{1,1}, \ldots, X_{d_1,d_2})^t \sim mult(n, p)$
  with $p = (p_{1,1}, \ldots, p_{d_1,d_2})^t \in (0,1)^{d_1 \cdot d_2}$ and $\sum_{j,k} p_{j,k} = 1$

| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| occupation probabilities | red | orange | grün | blue | $p_{j,\cdot}$ |
|---|---|---|---|---|---|
| solid | $p_{1,1}$ | $p_{1,2}$ | $p_{1,3}$ | $p_{1,4}$ | $p_{1,\cdot}$ |
| soft | $p_{2,1}$ | $p_{2,2}$ | $p_{2,3}$ | $p_{2,4}$ | $p_{2,\cdot}$ |
| $p_{\cdot,k}$ | $p_{\cdot,1}$ | $p_{\cdot,2}$ | $p_{\cdot,3}$ | $p_{\cdot,4}$ | |

- while $p_{j,k}$ denotes the probability to fall into row $j$ and column $k$
- and the *row sums* $p_{j,\cdot}$, the *column sums* $p_{\cdot,k}$

# Model

- $n = 230$ data in $d = d_1 \cdot d_2 = 8$ categories
  feature 1 (underground) has $d_1 = 2$ categories
  and feature 2 (color) has $d_2 = 4$ categories
- observed frequencies: $x_{1,1}, x_{1,2}, \cdots, x_{2,4}$ ($x_{j,k} \to$ row $j$, column $k$)
- Model: Let $\mathfrak{X} = (X_{1,1}, \ldots, X_{d_1,d_2})^t \sim mult(n, p)$
  with $p = (p_{1,1}, \ldots, p_{d_1,d_2})^t \in (0,1)^{d_1 \cdot d_2}$ and $\sum_{j,k} p_{j,k} = 1$

| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| occupation probabilities | red | orange | grün | blue | $p_{j,\cdot}$ |
|---|---|---|---|---|---|
| solid | $p_{1,1}$ | $p_{1,2}$ | $p_{1,3}$ | $p_{1,4}$ | $p_{1,\cdot}$ |
| soft | $p_{2,1}$ | $p_{2,2}$ | $p_{2,3}$ | $p_{2,4}$ | $p_{2,\cdot}$ |
| $p_{\cdot,k}$ | $p_{\cdot,1}$ | $p_{\cdot,2}$ | $p_{\cdot,3}$ | $p_{\cdot,4}$ | $\sum = 1$ |

- while $p_{j,k}$ denotes the probability to fall into row $j$ and column $k$
- and the *row sums* $p_{j,\cdot}$, the *column sums* $p_{\cdot,k}$, and *total sum* $\sum_{j,k} p_{j,k} = 1$

# Model

- $n = 230$ data in $d = d_1 \cdot d_2 = 8$ categories
  feature 1 (underground) has $d_1 = 2$ categories
  and feature 2 (color) has $d_2 = 4$ categories
- observed frequencies: $x_{1,1}, x_{1,2}, \cdots, x_{2,4}$ ($x_{j,k} \to$ row $j$, column $k$)
- Model: Let $\mathfrak{X} = (X_{1,1}, \ldots, X_{d_1,d_2})^t \sim mult(n, p)$
  with $p = (p_{1,1}, \ldots, p_{d_1,d_2})^t \in (0,1)^{d_1 \cdot d_2}$ and $\sum_{j,k} p_{j,k} = 1$

| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| occupation probabilities | red | orange | grün | blue | $p_{j,\cdot}$ |
|---|---|---|---|---|---|
| solid | $p_{1,1}$ | $p_{1,2}$ | $p_{1,3}$ | $p_{1,4}$ | $p_{1,\cdot}$ |
| soft | $p_{2,1}$ | $p_{2,2}$ | $p_{2,3}$ | $p_{2,4}$ | $p_{2,\cdot}$ |
| $p_{\cdot,k}$ | $p_{\cdot,1}$ | $p_{\cdot,2}$ | $p_{\cdot,3}$ | $p_{\cdot,4}$ | $\sum = 1$ |

- while $p_{j,k}$ denotes the probability to fall into row $j$ and column $k$
- and the *row sums* $p_{j,\cdot}$, the *column sums* $p_{\cdot,k}$, and *total sum* $\sum_{j,k} p_{j,k} = 1$
- Independence means, that

$$p_{j,k} = \mathbb{P}(\text{row } j \text{ and column } k) = \mathbb{P}(\text{row } j) \cdot \mathbb{P}(\text{column } k) = p_{j,\cdot} \cdot p_{\cdot,k}$$

# Model

- $n = 230$ data in $d = d_1 \cdot d_2 = 8$ categories
  feature 1 (underground) has $d_1 = 2$ categories
  and feature 2 (color) has $d_2 = 4$ categories
- observed frequencies: $x_{1,1}, x_{1,2}, \cdots, x_{2,4}$ ($x_{j,k} \to$ row $j$, column $k$)
- Model: Let $\mathfrak{X} = (X_{1,1}, \ldots, X_{d_1,d_2})^t \sim mult(n, p)$
  with $p = (p_{1,1}, \ldots, p_{d_1,d_2})^t \in (0,1)^{d_1 \cdot d_2}$ and $\sum_{j,k} p_{j,k} = 1$

| $x_{j,k}$ | side | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|---|
| underground | solid | 25 | 22 | 19 | 14 | 80 |
| | soft | 27 | 19 | 44 | 60 | 150 |
| | $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| occupation probabilities | red | orange | grün | blue | $p_{j,\cdot}$ |
|---|---|---|---|---|---|
| solid | $p_{1,1}$ | $p_{1,2}$ | $p_{1,3}$ | $p_{1,4}$ | $p_{1,\cdot}$ |
| soft | $p_{2,1}$ | $p_{2,2}$ | $p_{2,3}$ | $p_{2,4}$ | $p_{2,\cdot}$ |
| $p_{\cdot,k}$ | $p_{\cdot,1}$ | $p_{\cdot,2}$ | $p_{\cdot,3}$ | $p_{\cdot,4}$ | $\sum = 1$ |

- while $p_{j,k}$ denotes the probability to fall into row $j$ and column $k$
- and the *row sums* $p_{j,\cdot}$, the *column sums* $p_{\cdot,k}$, and *total sum* $\sum_{j,k} p_{j,k} = 1$
- Independence means, that (e.g., $p_{1,2} = p_{1,\cdot} \cdot p_{\cdot,2}$)

$$p_{j,k} = \mathbb{P}(\text{row } j \text{ and column } k) = \mathbb{P}(\text{row } j) \cdot \mathbb{P}(\text{column } k) = p_{j,\cdot} \cdot p_{\cdot,k}$$

# Model and null hypothesis

- observed frequencies: $x_{1,1}, x_{1,2}, \cdots, x_{2,4}$
- Model: Let $\mathfrak{X} = (X_{1,1}, \ldots, X_{d_1,d_2})^t \sim mult(n, p)$
  with $p = (p_{1,1}, \ldots, p_{d_1,d_2})^t \in (0,1)^{d_1 \cdot d_2}$ and $\sum_{j,k} p_{j,k} = 1$

| $x_{j,k}$ | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|
| solid | 25 | 22 | 19 | 14 | 80 |
| soft | 27 | 19 | 44 | 60 | 150 |
| $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| $\mathbb{P}_{H_0}$ | red | orange | green | blue | $p_{j,\cdot}$ |
|---|---|---|---|---|---|
| solid | $p_{1,\cdot} \cdot p_{\cdot,1}$ | $p_{1,\cdot} \cdot p_{\cdot,2}$ | $p_{1,\cdot} \cdot p_{\cdot,3}$ | $p_{1,\cdot} \cdot p_{\cdot,4}$ | $p_{1,\cdot}$ |
| soft | $p_{2,\cdot} \cdot p_{\cdot,1}$ | $p_{2,\cdot} \cdot p_{\cdot,2}$ | $p_{2,\cdot} \cdot p_{\cdot,3}$ | $p_{2,\cdot} \cdot p_{\cdot,4}$ | $p_{2,\cdot}$ |
| $p_{\cdot,k}$ | $p_{\cdot,1}$ | $p_{\cdot,2}$ | $p_{\cdot,3}$ | $p_{\cdot,4}$ | $\sum = 1$ |

- Independence means that

$$p_{j,k} = \mathbb{P}(\text{row } j \text{ and column } k) = \mathbb{P}(\text{row } j) \cdot \mathbb{P}(\text{column } k) = p_{j,\cdot} \cdot p_{\cdot,k}$$

# Model and null hypothesis

- observed frequencies: $x_{1,1}, x_{1,2}, \cdots, x_{2,4}$
- Model: Let $\mathfrak{X} = (X_{1,1}, \ldots, X_{d_1,d_2})^t \sim mult(n,p)$
  with $p = (p_{1,1}, \ldots, p_{d_1,d_2})^t \in (0,1)^{d_1 \cdot d_2}$ and $\sum_{j,k} p_{j,k} = 1$
- Null hypothesis:
  $H_0 : p = p_0 := (p_{1,\cdot} \cdot p_{\cdot,1}, \ldots, p_{d_1,\cdot} \cdot p_{\cdot,d_2})^t \in (0,1)^{d_1 \cdot d_2}$
  and $\sum_{j=1}^{d_1} p_{j,\cdot} = \sum_{k=1}^{d_2} p_{\cdot,k} = 1$

| $x_{j,k}$ | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|
| solid | 25 | 22 | 19 | 14 | 80 |
| soft | 27 | 19 | 44 | 60 | 150 |
| $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| $\mathbb{P}_{H_0}$ | red | orange | green | blue | $p_{j,\cdot}$ |
|---|---|---|---|---|---|
| solid | $p_{1,\cdot} \cdot p_{\cdot,1}$ | $p_{1,\cdot} \cdot p_{\cdot,2}$ | $p_{1,\cdot} \cdot p_{\cdot,3}$ | $p_{1,\cdot} \cdot p_{\cdot,4}$ | $p_{1,\cdot}$ |
| soft | $p_{2,\cdot} \cdot p_{\cdot,1}$ | $p_{2,\cdot} \cdot p_{\cdot,2}$ | $p_{2,\cdot} \cdot p_{\cdot,3}$ | $p_{2,\cdot} \cdot p_{\cdot,4}$ | $p_{2,\cdot}$ |
| $p_{\cdot,k}$ | $p_{\cdot,1}$ | $p_{\cdot,2}$ | $p_{\cdot,3}$ | $p_{\cdot,4}$ | $\sum = 1$ |

- Independence means that

$$p_{j,k} = \mathbb{P}(\text{row } j \text{ and column } k) = \mathbb{P}(\text{row } j) \cdot \mathbb{P}(\text{column } k) = p_{j,\cdot} \cdot p_{\cdot,k}$$

# Model and null hypothesis

- observed frequencies: $x_{1,1}, x_{1,2}, \cdots, x_{2,4}$
- Model: Let $\mathfrak{X} = (X_{1,1}, \ldots, X_{d_1,d_2})^t \sim mult(n, p)$
  with $p = (p_{1,1}, \ldots, p_{d_1,d_2})^t \in (0,1)^{d_1 \cdot d_2}$ and $\sum_{j,k} p_{j,k} = 1$
- Null hypothesis:
  $H_0 : p = p_0 := (p_{1,\cdot} \cdot p_{\cdot,1}, \ldots, p_{d_1,\cdot} \cdot p_{\cdot,d_2})^t \in (0,1)^{d_1 \cdot d_2}$
  and $\sum_{j=1}^{d_1} p_{j,\cdot} = \sum_{k=1}^{d_2} p_{\cdot,k} = 1$

| $x_{j,k}$ | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|
| solid | 25 | 22 | 19 | 14 | 80 |
| soft | 27 | 19 | 44 | 60 | 150 |
| $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| $\mathbb{P}_{H_0}$ | red | orange | green | blue | $p_{j,\cdot}$ |
|---|---|---|---|---|---|
| solid | $p_{1,\cdot} \cdot p_{\cdot,1}$ | $p_{1,\cdot} \cdot p_{\cdot,2}$ | $p_{1,\cdot} \cdot p_{\cdot,3}$ | $p_{1,\cdot} \cdot p_{\cdot,4}$ | $p_{1,\cdot}$ |
| soft | $p_{2,\cdot} \cdot p_{\cdot,1}$ | $p_{2,\cdot} \cdot p_{\cdot,2}$ | $p_{2,\cdot} \cdot p_{\cdot,3}$ | $p_{2,\cdot} \cdot p_{\cdot,4}$ | $p_{2,\cdot}$ |
| $p_{\cdot,k}$ | $p_{\cdot,1}$ | $p_{\cdot,2}$ | $p_{\cdot,3}$ | $p_{\cdot,4}$ | $\sum = 1$ |

- Independence means that

$$p_{j,k} = \mathbb{P}(\text{row } j \text{ and column } k) = \mathbb{P}(\text{row } j) \cdot \mathbb{P}(\text{column } k) = p_{j,\cdot} \cdot p_{\cdot,k}$$

- Expectations in the categories under $H_0$: $\mathbb{E}_{H_0}[\cdot] = n \cdot p_{j,\cdot} \cdot p_{\cdot,k}$

# Observed and expected frequencies

| $x_{j,k}$ | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|
| solid | 25 | 22 | 19 | 14 | 80 |
| soft | 27 | 19 | 44 | 60 | 150 |
| $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| $\mathbb{E}_{H_0}[\cdot]$ | red | orange | green | blue |
|---|---|---|---|---|
| solid | $n \cdot p_{1,\cdot} \cdot p_{\cdot,1}$ | $n \cdot p_{1,\cdot} \cdot p_{\cdot,2}$ | $n \cdot p_{1,\cdot} \cdot p_{\cdot,3}$ | $n \cdot p_{1,\cdot} \cdot p_{\cdot,4}$ |
| soft | $n \cdot p_{2,\cdot} \cdot p_{\cdot,1}$ | $n \cdot p_{2,\cdot} \cdot p_{\cdot,2}$ | $n \cdot p_{2,\cdot} \cdot p_{\cdot,3}$ | $n \cdot p_{2,\cdot} \cdot p_{\cdot,4}$ |

# Observed and expected frequencies

| $x_{j,k}$ | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|
| solid | 25 | 22 | 19 | 14 | 80 |
| soft | 27 | 19 | 44 | 60 | 150 |
| $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| $\mathbb{E}_{H_0}[\cdot]$ | red | orange | green | blue |
|---|---|---|---|---|
| solid | $n \cdot p_{1,\cdot} \cdot p_{\cdot,1}$ | $n \cdot p_{1,\cdot} \cdot p_{\cdot,2}$ | $n \cdot p_{1,\cdot} \cdot p_{\cdot,3}$ | $n \cdot p_{1,\cdot} \cdot p_{\cdot,4}$ |
| soft | $n \cdot p_{2,\cdot} \cdot p_{\cdot,1}$ | $n \cdot p_{2,\cdot} \cdot p_{\cdot,2}$ | $n \cdot p_{2,\cdot} \cdot p_{\cdot,3}$ | $n \cdot p_{2,\cdot} \cdot p_{\cdot,4}$ |

- Now we can compare 'observed' and 'expected under $H_0$'

# Observed and expected frequencies

| $x_{j,k}$ | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|
| solid | 25 | 22 | 19 | 14 | 80 |
| soft | 27 | 19 | 44 | 60 | 150 |
| $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| $\mathbb{E}_{H_0}[\cdot]$ | red | orange | green | blue |
|---|---|---|---|---|
| solid | $n \cdot p_{1,\cdot} \cdot p_{\cdot,1}$ | $n \cdot p_{1,\cdot} \cdot p_{\cdot,2}$ | $n \cdot p_{1,\cdot} \cdot p_{\cdot,3}$ | $n \cdot p_{1,\cdot} \cdot p_{\cdot,4}$ |
| soft | $n \cdot p_{2,\cdot} \cdot p_{\cdot,1}$ | $n \cdot p_{2,\cdot} \cdot p_{\cdot,2}$ | $n \cdot p_{2,\cdot} \cdot p_{\cdot,3}$ | $n \cdot p_{2,\cdot} \cdot p_{\cdot,4}$ |

- Now we can compare 'observed' and 'expected under $H_0$'

$$\sum_{j,k} \frac{(x_{j,k} - n \cdot p_{j,\cdot} \cdot p_{\cdot,k})^2}{n \cdot p_{j,\cdot} \cdot p_{\cdot,k}} \qquad (*)$$

- Problem: products $p_{j,\cdot} \cdot p_{\cdot,k}$ unknown in practice

# Observed and expected frequencies

| $x_{j,k}$ | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|
| solid | 25 | 22 | 19 | 14 | 80 |
| soft | 27 | 19 | 44 | 60 | 150 |
| $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| $\mathbb{E}_{H_0}[\cdot]$ | red | orange | green | blue |
|---|---|---|---|---|
| solid | $n \cdot p_{1,\cdot} \cdot p_{\cdot,1}$ | $n \cdot p_{1,\cdot} \cdot p_{\cdot,2}$ | $n \cdot p_{1,\cdot} \cdot p_{\cdot,3}$ | $n \cdot p_{1,\cdot} \cdot p_{\cdot,4}$ |
| soft | $n \cdot p_{2,\cdot} \cdot p_{\cdot,1}$ | $n \cdot p_{2,\cdot} \cdot p_{\cdot,2}$ | $n \cdot p_{2,\cdot} \cdot p_{\cdot,3}$ | $n \cdot p_{2,\cdot} \cdot p_{\cdot,4}$ |

- Now we can compare 'observed' and 'expected under $H_0$'

$$\sum_{j,k} \frac{(x_{j,k} - n \cdot p_{j,\cdot} \cdot p_{\cdot,k})^2}{n \cdot p_{j,\cdot} \cdot p_{\cdot,k}} \qquad (*)$$

- Problem: products $p_{j,\cdot} \cdot p_{\cdot,k}$ unknown in practice
- Solution: Estimate marginal probabilities via marginal frequencies

# Observed and expected frequencies

| $x_{j,k}$ | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|
| solid | 25 | 22 | 19 | 14 | 80 |
| soft | 27 | 19 | 44 | 60 | 150 |
| $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| $\mathbb{E}_{H_0}[\cdot]$ | red | orange | green | blue |
|---|---|---|---|---|
| solid | $n \cdot p_{1,\cdot} \cdot p_{\cdot,1}$ | $n \cdot p_{1,\cdot} \cdot p_{\cdot,2}$ | $n \cdot p_{1,\cdot} \cdot p_{\cdot,3}$ | $n \cdot p_{1,\cdot} \cdot p_{\cdot,4}$ |
| soft | $n \cdot p_{2,\cdot} \cdot p_{\cdot,1}$ | $n \cdot p_{2,\cdot} \cdot p_{\cdot,2}$ | $n \cdot p_{2,\cdot} \cdot p_{\cdot,3}$ | $n \cdot p_{2,\cdot} \cdot p_{\cdot,4}$ |

- Now we can compare 'observed' and 'expected under $H_0$'

$$\sum_{j,k} \frac{(x_{j,k} - n \cdot p_{j,\cdot} \cdot p_{\cdot,k})^2}{n \cdot p_{j,\cdot} \cdot p_{\cdot,k}} \qquad (*)$$

- Problem: products $p_{j,\cdot} \cdot p_{\cdot,k}$ unknown in practice
- Solution: Estimate marginal probabilities via marginal frequencies
- More precisely: row proportions $x_{j,\cdot}/n$ estimate row probabilities $p_{j,\cdot}$ and column proportions $x_{\cdot,k}/n$ estimates column probabilities $p_{\cdot,k}$

# Observed and expected frequencies

| $x_{j,k}$ | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|
| solid | 25 | 22 | 19 | 14 | 80 |
| soft | 27 | 19 | 44 | 60 | 150 |
| $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| $\mathbb{E}_{H_0}[\cdot]$ | red | orange | green | blue |
|---|---|---|---|---|
| solid | $n \cdot p_{1,\cdot} \cdot p_{\cdot,1}$ | $n \cdot p_{1,\cdot} \cdot p_{\cdot,2}$ | $n \cdot p_{1,\cdot} \cdot p_{\cdot,3}$ | $n \cdot p_{1,\cdot} \cdot p_{\cdot,4}$ |
| soft | $n \cdot p_{2,\cdot} \cdot p_{\cdot,1}$ | $n \cdot p_{2,\cdot} \cdot p_{\cdot,2}$ | $n \cdot p_{2,\cdot} \cdot p_{\cdot,3}$ | $n \cdot p_{2,\cdot} \cdot p_{\cdot,4}$ |

- Now we can compare 'observed' and 'expected under $H_0$'

$$\sum_{j,k} \frac{(x_{j,k} - n \cdot p_{j,\cdot} \cdot p_{\cdot,k})^2}{n \cdot p_{j,\cdot} \cdot p_{\cdot,k}} \qquad (*)$$

- Problem: products $p_{j,\cdot} \cdot p_{\cdot,k}$ unknown in practice
- Solution: Estimate marginal probabilities via marginal frequencies
- More precisely: row proportions $x_{j,\cdot}/n$ estimate row probabilities $p_{j,\cdot}$ and column proportions $x_{\cdot,k}/n$ estimates column probabilities $p_{\cdot,k}$
  i.e., $(x_{j,\cdot} \cdot x_{\cdot,k})/n$ estimates $n \cdot p_{j,\cdot} \cdot p_{\cdot,k}$

# Observed and expected frequencies

| $x_{j,k}$ | red | orange | green | blue | $x_{j,.}$ |
|---|---|---|---|---|---|
| hard | 25 | 22 | 19 | 14 | 80 |
| soft | 27 | 19 | 44 | 60 | 150 |
| $x_{.,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| $\hat{\mathbb{E}}_{H_0}[\cdot]$ | rot | orange | green | blue |
|---|---|---|---|---|
| hard | $(80 \cdot 52)/n$ | $(80 \cdot 41)/n$ | $(80 \cdot 63)/n$ | $(80 \cdot 74)/n$ |
| soft | $(150 \cdot 52)/n$ | $(150 \cdot 41)/n$ | $(150 \cdot 63)/n$ | $(150 \cdot 74)/n$ |

- Now we can compare 'observed' and 'expected under $H_0$'

$$\sum_{j,k} \frac{(x_{j,k} - n \cdot p_{j,.} \cdot p_{.,k})^2}{n \cdot p_{j,.} \cdot p_{.,k}} \qquad (*)$$

- Estimate marginal probabilities via marginal frequencies
  i.e., $(x_{j,.} \cdot x_{.,k})/n$ estimates $n \cdot p_{j,.} \cdot p_{.,k}$

# Observed and expected frequencies

| $x_{j,k}$ | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|
| hard | 25 | 22 | 19 | 14 | 80 |
| soft | 27 | 19 | 44 | 60 | 150 |
| $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| $\hat{\mathbb{E}}_{H_0}[\cdot]$ | rot | orange | green | blue |
|---|---|---|---|---|
| hard | $(80 \cdot 52)/n$ | $(80 \cdot 41)/n$ | $(80 \cdot 63)/n$ | $(80 \cdot 74)/n$ |
| soft | $(150 \cdot 52)/n$ | $(150 \cdot 41)/n$ | $(150 \cdot 63)/n$ | $(150 \cdot 74)/n$ |

- Now we can compare 'observed' and 'expected under $H_0$'

$$\sum_{j,k} \frac{(x_{j,k} - n \cdot p_{j,\cdot} \cdot p_{\cdot,k})^2}{n \cdot p_{j,\cdot} \cdot p_{\cdot,k}} \qquad (*)$$

- Estimate marginal probabilities via marginal frequencies
  i.e., $(x_{j,\cdot} \cdot x_{\cdot,k})/n$ estimates $n \cdot p_{j,\cdot} \cdot p_{\cdot,k}$
- plugging the estimator into $(*)$ yields the $\chi^2$-statistic

$$x^2 = \sum_{j,k} \frac{\left(x_{j,k} - \frac{x_{j,\cdot} \cdot x_{\cdot,k}}{n}\right)^2}{\frac{x_{j,\cdot} \cdot x_{\cdot,k}}{n}}$$

# Observed and expected frequencies

| $x_{j,k}$ | red | orange | green | blue | $x_{j,.}$ |
|---|---|---|---|---|---|
| hard | 25 | 22 | 19 | 14 | 80 |
| soft | 27 | 19 | 44 | 60 | 150 |
| $x_{.,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| $\hat{\mathbb{E}}_{H_0}[\cdot]$ | rot | orange | green | blue |
|---|---|---|---|---|
| hard | $(80 \cdot 52)/n$ | $(80 \cdot 41)/n$ | $(80 \cdot 63)/n$ | $(80 \cdot 74)/n$ |
| soft | $(150 \cdot 52)/n$ | $(150 \cdot 41)/n$ | $(150 \cdot 63)/n$ | $(150 \cdot 74)/n$ |

- Now we can compare 'observed' and 'expected under $H_0$'

$$\sum_{j,k} \frac{(x_{j,k} - n \cdot p_{j,.} \cdot p_{.,k})^2}{n \cdot p_{j,.} \cdot p_{.,k}} \qquad (*)$$

- Estimate marginal probabilities via marginal frequencies i.e., $(x_{j,.} \cdot x_{.,k})/n$ estimates $n \cdot p_{j,.} \cdot p_{.,k}$
- plugging the estimator into $(*)$ yields the $\chi^2$-statistic

$$x^2 = \sum_{j,k} \frac{\left(x_{j,k} - \frac{x_{j,.} \cdot x_{.,k}}{n}\right)^2}{\frac{x_{j,.} \cdot x_{.,k}}{n}} \approx 19.3$$

# Observed and expected frequencies

| $x_{j,k}$ | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|
| hard | 25 | 22 | 19 | 14 | 80 |
| soft | 27 | 19 | 44 | 60 | 150 |
| $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| $\hat{\mathbb{E}}_{H_0}[\cdot]$ | rot | orange | green | blue |
|---|---|---|---|---|
| hard | $(80 \cdot 52)/n$ | $(80 \cdot 41)/n$ | $(80 \cdot 63)/n$ | $(80 \cdot 74)/n$ |
| soft | $(150 \cdot 52)/n$ | $(150 \cdot 41)/n$ | $(150 \cdot 63)/n$ | $(150 \cdot 74)/n$ |

- Now we can compare 'observed' and 'expected under $H_0$'

$$\sum_{j,k} \frac{(x_{j,k} - n \cdot p_{j,\cdot} \cdot p_{\cdot,k})^2}{n \cdot p_{j,\cdot} \cdot p_{\cdot,k}} \qquad (*)$$

- Estimate marginal probabilities via marginal frequencies
  i.e., $(x_{j,\cdot} \cdot x_{\cdot,k})/n$ estimates $n \cdot p_{j,\cdot} \cdot p_{\cdot,k}$
- plugging the estimator into $(*)$ yields the $\chi^2$-statistic

$$x^2 = \sum_{j,k} \frac{\left(x_{j,k} - \frac{x_{j,\cdot} \cdot x_{\cdot,k}}{n}\right)^2}{\frac{x_{j,\cdot} \cdot x_{\cdot,k}}{n}} \approx 19.3$$

...is this a large value?

# Observed and expected frequencies

| $x_{j,k}$ | red | orange | green | blue | $x_{j,\cdot}$ |
|---|---|---|---|---|---|
| hard | 25 | 22 | 19 | 14 | 80 |
| soft | 27 | 19 | 44 | 60 | 150 |
| $x_{\cdot,k}$ | 52 | 41 | 63 | 74 | $n = 230$ |

| $\hat{\mathbb{E}}_{H_0}[\cdot]$ | rot | orange | green | blue |
|---|---|---|---|---|
| hard | $(80 \cdot 52)/n$ | $(80 \cdot 41)/n$ | $(80 \cdot 63)/n$ | $(80 \cdot 74)/n$ |
| soft | $(150 \cdot 52)/n$ | $(150 \cdot 41)/n$ | $(150 \cdot 63)/n$ | $(150 \cdot 74)/n$ |

- Now we can compare 'observed' and 'expected under $H_0$'

$$\sum_{j,k} \frac{(x_{j,k} - n \cdot p_{j,\cdot} \cdot p_{\cdot,k})^2}{n \cdot p_{j,\cdot} \cdot p_{\cdot,k}} \qquad (*)$$

- Estimate marginal probabilities via marginal frequencies
  i.e., $(x_{j,\cdot} \cdot x_{\cdot,k})/n$ estimates $n \cdot p_{j,\cdot} \cdot p_{\cdot,k}$
- plugging the estimator into $(*)$ yields the $\chi^2$-statistic

$$x^2 = \sum_{j,k} \frac{\left(x_{j,k} - \frac{x_{j,\cdot} \cdot x_{\cdot,k}}{n}\right)^2}{\frac{x_{j,\cdot} \cdot x_{\cdot,k}}{n}} \approx 19.3$$

...is this a large value?

yes, as the comparison with the $\chi^2$-distribution reveals...

# The $\chi^2$-test for independence

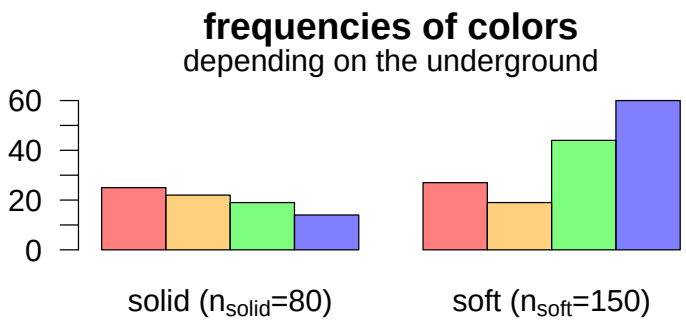
- Model: Let $\mathfrak{x} = (X_{1,1}, \ldots, X_{d_1,d_2})^t \sim mult(n,p)$
  with $p = (p_{1,1}, \ldots, p_{d_1,d_2})^t \in (0,1)^{d_1 \cdot d_2}$ and $\sum_{j,k} p_{j,k} = 1$

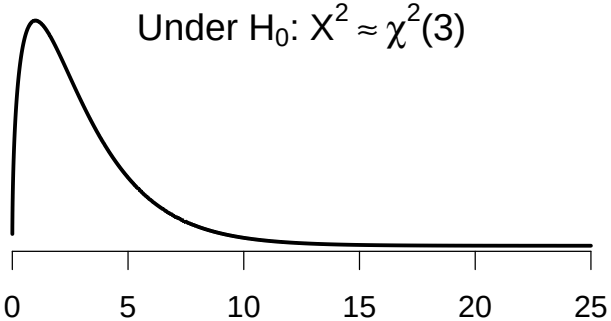
Under $H_0 : p = p_0 := (p_{1,\cdot} \cdot p_{\cdot,1}, \ldots, p_{d_1,\cdot} \cdot p_{\cdot,d_2})^t \in (0,1)^{d_1 \cdot d_2}$
and $\sum_{j=1}^{d_1} p_{j,\cdot} = \sum_{k=1}^{d_2} p_{\cdot,k} = 1$

it holds (approximately)

$$X^2 := \sum_{j,k} \frac{\left(X_{j,k} - \frac{X_{j,\cdot} \cdot X_{\cdot,k}}{n}\right)^2}{\frac{X_{j,\cdot} \cdot X_{\cdot,k}}{n}} \overset{d}{\approx} \chi^2((d_1 - 1) \cdot (d_2 - 1))$$

in fact, it holds that $X^2 \xrightarrow{d} \chi^2((d_1 - 1)(d_2 - 1))$ as $n \to \infty$



# The $\chi^2$-test for independence

- Model: Let $\mathfrak{X} = (X_{1,1}, \ldots, X_{d_1,d_2})^t \sim mult(n,p)$
  with $p = (p_{1,1}, \ldots, p_{d_1,d_2})^t \in (0,1)^{d_1 \cdot d_2}$ and $\sum_{j,k} p_{j,k} = 1$

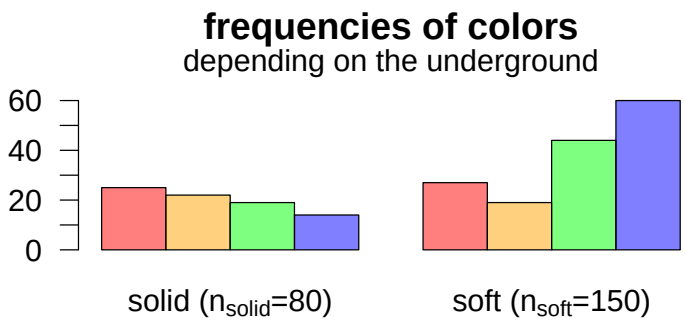
Under $H_0 : p = p_0 := (p_{1,\cdot} \cdot p_{\cdot,1}, \ldots, p_{d_1,\cdot} \cdot p_{\cdot,d_2})^t \in (0,1)^{d_1 \cdot d_2}$
and $\sum_{j=1}^{d_1} p_{j,\cdot} = \sum_{k=1}^{d_2} p_{\cdot,k} = 1$

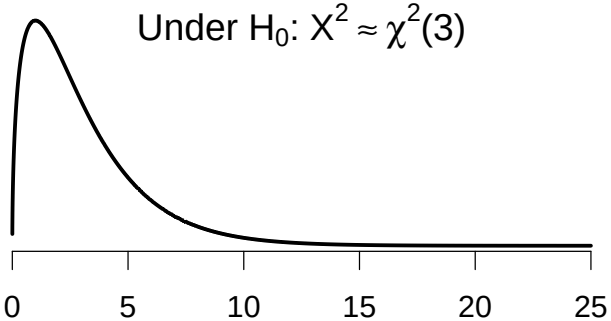
it holds (approximately)

$$X^2 := \sum_{j,k} \frac{\left(X_{j,k} - \frac{X_{j,\cdot} \cdot X_{\cdot,k}}{n}\right)^2}{\frac{X_{j,\cdot} \cdot X_{\cdot,k}}{n}} \overset{d}{\approx} \chi^2((d_1-1)\cdot(d_2-1))$$

in fact, it holds that $X^2 \xrightarrow{d} \chi^2((d_1-1)(d_2-1))$ as $n \to \infty$

- Here: $d_1 = 2, d_2 = 4$, i.e., $X^2 \overset{H_0}{\sim} \chi^2(3)$ (approx)



# The $\chi^2$-test for independence

- Model: Let $\mathfrak{X} = (X_{1,1}, \ldots, X_{d_1,d_2})^t \sim mult(n,p)$
  with $p = (p_{1,1}, \ldots, p_{d_1,d_2})^t \in (0,1)^{d_1 \cdot d_2}$ and $\sum_{j,k} p_{j,k} = 1$

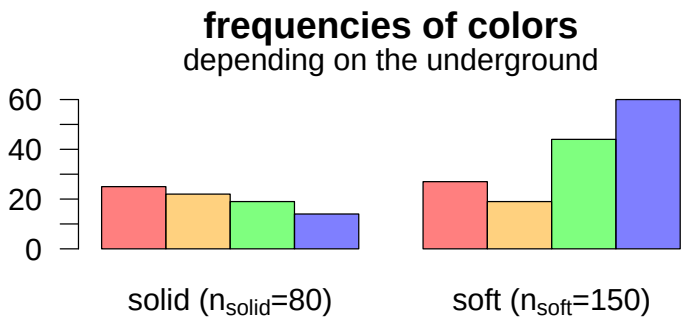
Under $H_0 : p = p_0 := (p_{1,\cdot} \cdot p_{\cdot,1}, \ldots, p_{d_1,\cdot} \cdot p_{\cdot,d_2})^t \in (0,1)^{d_1 \cdot d_2}$
and $\sum_{j=1}^{d_1} p_{j,\cdot} = \sum_{k=1}^{d_2} p_{\cdot,k} = 1$

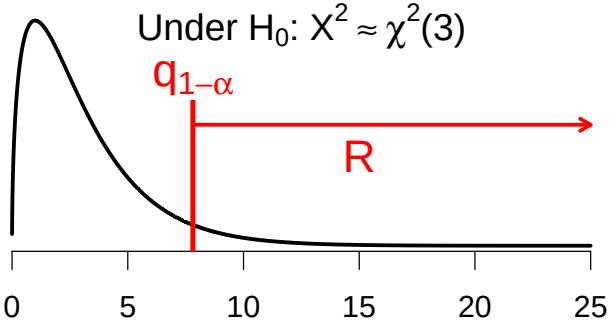
it holds (approximately)

$$X^2 := \sum_{j,k} \frac{\left(X_{j,k} - \frac{X_{j,\cdot} \cdot X_{\cdot,k}}{n}\right)^2}{\frac{X_{j,\cdot} \cdot X_{\cdot,k}}{n}} \overset{d}{\approx} \chi^2((d_1-1)\cdot(d_2-1))$$

in fact, it holds that $X^2 \xrightarrow{d} \chi^2((d_1-1)(d_2-1))$ as $n \to \infty$

- Here: $d_1 = 2, d_2 = 4$, i.e., $X^2 \overset{H_0}{\sim} \chi^2(3)$ (approx)
- for $\alpha = 5\%$ the $(1-\alpha)$-quantile of the $\chi^2(3)$-distribution is $q_{1-\alpha} \approx 7.8$



# The $\chi^2$-test for independence

- Model: Let $\mathfrak{X} = (X_{1,1}, \ldots, X_{d_1,d_2})^t \sim mult(n,p)$
  with $p = (p_{1,1}, \ldots, p_{d_1,d_2})^t \in (0,1)^{d_1 \cdot d_2}$ and $\sum_{j,k} p_{j,k} = 1$

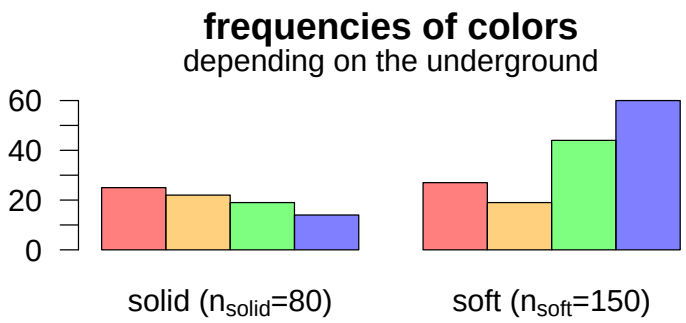
Under $H_0 : p = p_0 := (p_{1,\cdot} \cdot p_{\cdot,1}, \ldots, p_{d_1,\cdot} \cdot p_{\cdot,d_2})^t \in (0,1)^{d_1 \cdot d_2}$
and $\sum_{j=1}^{d_1} p_{j,\cdot} = \sum_{k=1}^{d_2} p_{\cdot,k} = 1$

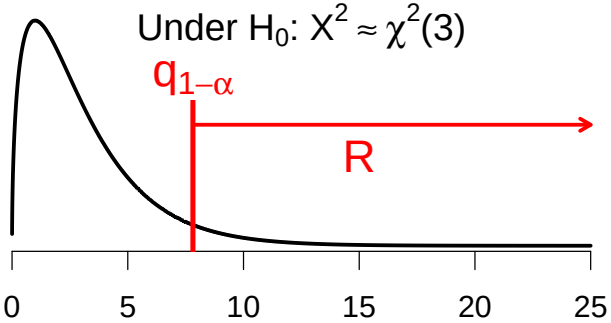
it holds (approximately)

$$X^2 := \sum_{j,k} \frac{\left(X_{j,k} - \frac{X_{j,\cdot} \cdot X_{\cdot,k}}{n}\right)^2}{\frac{X_{j,\cdot} \cdot X_{\cdot,k}}{n}} \overset{d}{\approx} \chi^2((d_1-1)\cdot(d_2-1))$$

in fact, it holds that $X^2 \xrightarrow{d} \chi^2((d_1-1)(d_2-1))$ as $n \to \infty$

- Here: $d_1 = 2, d_2 = 4$, i.e., $X^2 \overset{H_0}{\sim} \chi^2(3)$ (approx)
- for $\alpha = 5\%$ the $(1-\alpha)$-quantile of the $\chi^2(3)$-distribution is $q_{1-\alpha} \approx 7.8$
- rejection area: $R = [q_{1-\alpha}, \infty)$ (one-sided, $x^2$ large speaks against $H_0$)



# The $\chi^2$-test for independence

- Model: Let $\mathfrak{X} = (X_{1,1}, \ldots, X_{d_1,d_2})^t \sim mult(n,p)$
  with $p = (p_{1,1}, \ldots, p_{d_1,d_2})^t \in (0,1)^{d_1 \cdot d_2}$ and $\sum_{j,k} p_{j,k} = 1$

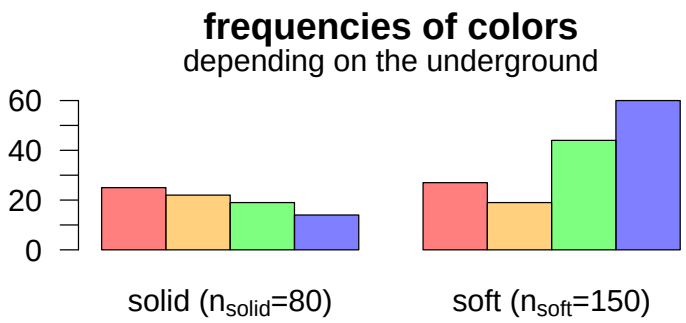
Under $H_0 : p = p_0 := (p_{1,\cdot} \cdot p_{\cdot,1}, \ldots, p_{d_1,\cdot} \cdot p_{\cdot,d_2})^t \in (0,1)^{d_1 \cdot d_2}$
and $\sum_{j=1}^{d_1} p_{j,\cdot} = \sum_{k=1}^{d_2} p_{\cdot,k} = 1$

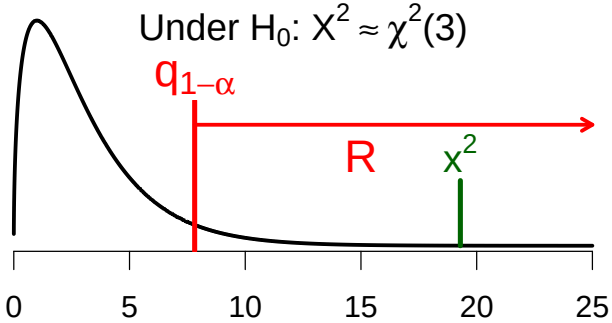
it holds (approximately)

$$X^2 := \sum_{j,k} \frac{\left(X_{j,k} - \frac{X_{j,\cdot} \cdot X_{\cdot,k}}{n}\right)^2}{\frac{X_{j,\cdot} \cdot X_{\cdot,k}}{n}} \overset{d}{\approx} \chi^2((d_1-1)\cdot(d_2-1))$$

in fact, it holds that $X^2 \xrightarrow{d} \chi^2((d_1-1)(d_2-1))$ as $n \to \infty$

- Here: $d_1 = 2, d_2 = 4$, i.e., $X^2 \overset{H_0}{\sim} \chi^2(3)$ (approx)
- for $\alpha = 5\%$ the $(1-\alpha)$-quantile of the $\chi^2(3)$-distribution is $q_{1-\alpha} \approx 7.8$
- rejection area: $R = [q_{1-\alpha}, \infty)$ (one-sided, $x^2$ large speaks against $H_0$)
- data: $x^2 = 19.3 \in R$, $\rightarrow$ reject $H_0$



# The $\chi^2$-test for independence

- Model: Let $\mathfrak{X} = (X_{1,1}, \ldots, X_{d_1,d_2})^t \sim mult(n,p)$
  with $p = (p_{1,1}, \ldots, p_{d_1,d_2})^t \in (0,1)^{d_1 \cdot d_2}$ and $\sum_{j,k} p_{j,k} = 1$

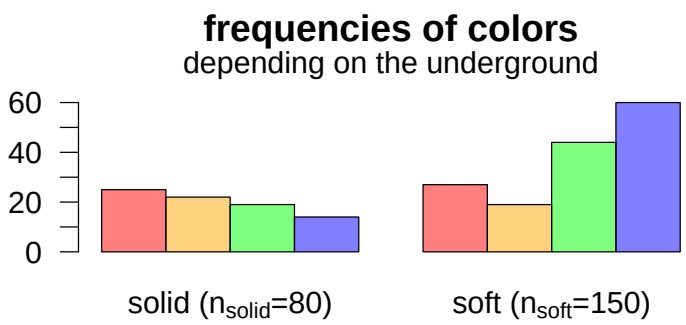
Under $H_0 : p = p_0 := (p_{1,\cdot} \cdot p_{\cdot,1}, \ldots, p_{d_1,\cdot} \cdot p_{\cdot,d_2})^t \in (0,1)^{d_1 \cdot d_2}$
and $\sum_{j=1}^{d_1} p_{j,\cdot} = \sum_{k=1}^{d_2} p_{\cdot,k} = 1$

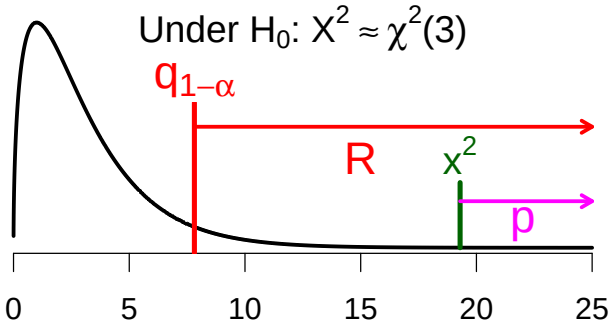
it holds (approximately)

$$X^2 := \sum_{j,k} \frac{\left(X_{j,k} - \frac{X_{j,\cdot} \cdot X_{\cdot,k}}{n}\right)^2}{\frac{X_{j,\cdot} \cdot X_{\cdot,k}}{n}} \overset{d}{\approx} \chi^2((d_1-1)\cdot(d_2-1))$$

in fact, it holds that $X^2 \xrightarrow{d} \chi^2((d_1-1)(d_2-1))$ as $n \to \infty$

- Here: $d_1 = 2, d_2 = 4$, i.e., $X^2 \overset{H_0}{\sim} \chi^2(3)$ (approx)
- for $\alpha = 5\%$ the $(1-\alpha)$-quantile of the $\chi^2(3)$-distribution is $q_{1-\alpha} \approx 7.8$
- rejection area: $R = [q_{1-\alpha}, \infty)$ (one-sided, $x^2$ large speaks against $H_0$)
- data: $x^2 = 19.3 \in R$, $\to$ reject $H_0$
- $p \approx 2.4 \cdot 10^{-4}$.



# The $\chi^2$-test for independence

- Model: Let $\mathfrak{X} = (X_{1,1}, \ldots, X_{d_1,d_2})^t \sim mult(n,p)$
  with $p = (p_{1,1}, \ldots, p_{d_1,d_2})^t \in (0,1)^{d_1 \cdot d_2}$ and $\sum_{j,k} p_{j,k} = 1$

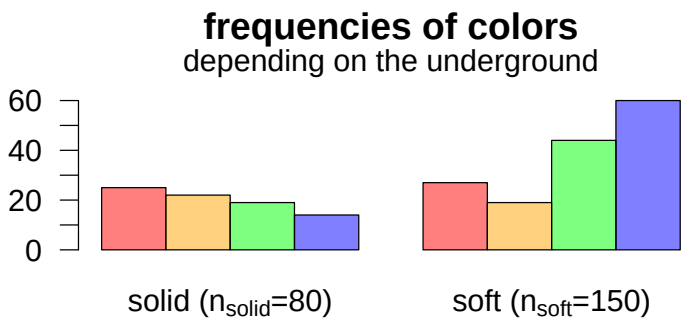
Under $H_0 : p = p_0 := (p_{1,\cdot} \cdot p_{\cdot,1}, \ldots, p_{d_1,\cdot} \cdot p_{\cdot,d_2})^t \in (0,1)^{d_1 \cdot d_2}$
and $\sum_{j=1}^{d_1} p_{j,\cdot} = \sum_{k=1}^{d_2} p_{\cdot,k} = 1$

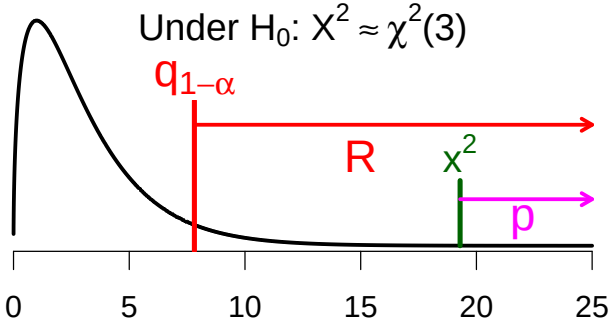
it holds (approximately)

$$X^2 := \sum_{j,k} \frac{\left(X_{j,k} - \frac{X_{j,\cdot} \cdot X_{\cdot,k}}{n}\right)^2}{\frac{X_{j,\cdot} \cdot X_{\cdot,k}}{n}} \overset{d}{\approx} \chi^2((d_1-1)\cdot(d_2-1))$$

in fact, it holds that $X^2 \xrightarrow{d} \chi^2((d_1-1)(d_2-1))$ as $n \to \infty$

- Here: $d_1 = 2, d_2 = 4$, i.e., $X^2 \overset{H_0}{\sim} \chi^2(3)$ (approx)
- for $\alpha = 5\%$ the $(1-\alpha)$-quantile of the $\chi^2(3)$-distribution is $q_{1-\alpha} \approx 7.8$
- rejection area: $R = [q_{1-\alpha}, \infty)$ (one-sided, $x^2$ large speaks against $H_0$)
- data: $x^2 = 19.3 \in R$, $\rightarrow$ reject $H_0$
- $p \approx 2.4 \cdot 10^{-4}$. If the features are independent, then in less than one of 4000 cases we observe a discrepancy, which is at least as extreme as in the data



# Nutshell

| $x_{j,k}$ | | | | |
|---|---|---|---|---|
| | $x_{1,1}$ | $\cdots$ | $x_{1,d_2}$ | |
| | $\vdots$ | $\ddots$ | $\vdots$ | |
| | $x_{d_1,1}$ | $\cdots$ | $x_{d_1,d_2}$ | |
| | | | | $n$ |

- 2 features, with $d_1$ and $d_2$ categories, $\rightarrow d_1 \cdot d_2$ 'cells'

# Nutshell

| $x_{j,k}$ | | | | |
|---|---|---|---|---|
| | $x_{1,1}$ | $\cdots$ | $x_{1,d_2}$ | |
| | $\vdots$ | $\ddots$ | $\vdots$ | |
| | $x_{d_1,1}$ | $\cdots$ | $x_{d_1,d_2}$ | |
| | | | | $n$ |

- 2 features, with $d_1$ and $d_2$ categories, $\rightarrow d_1 \cdot d_2$ 'cells'
- observed frequencies

# Nutshell

| $x_{j,k}$ | | | |
|---|---|---|---|
| | $x_{1,1}$ | $\cdots$ | $x_{1,d_2}$ |
| | $\vdots$ | $\ddots$ | $\vdots$ |
| | $x_{d_1,1}$ | $\cdots$ | $x_{d_1,d_2}$ |

$n$

| $\hat{\mathbb{E}}_{H_0}[\cdot]$ | | | |
|---|---|---|---|
| | $(x_{1,\cdot} \cdot x_{\cdot,1})/n$ | $\cdots$ | $(x_{1,\cdot} \cdot x_{\cdot,d_2})/n$ |
| | $\vdots$ | $\ddots$ | $\vdots$ |
| | $(x_{d_1,\cdot} \cdot x_{\cdot,1})/n$ | $\cdots$ | $(x_{d_1,\cdot} \cdot x_{\cdot,d_2})/n$ |

- 2 features, with $d_1$ and $d_2$ categories, $\rightarrow d_1 \cdot d_2$ 'cells'
- observed frequencies are compared to the expected frequencies under the null hypothesis

# Nutshell

| $x_{j,k}$ | | | | |
|---|---|---|---|---|
| | $x_{1,1}$ | $\cdots$ | $x_{1,d_2}$ | $x_{1,\cdot}$ |
| | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| | $x_{d_1,1}$ | $\cdots$ | $x_{d_1,d_2}$ | $x_{d_1,\cdot}$ |
| | $x_{\cdot,1}$ | $\cdots$ | $x_{\cdot,d_2}$ | $n$ |

| $\hat{\mathbb{E}}_{H_0}[\cdot]$ | | | |
|---|---|---|---|
| | $(x_{1,\cdot} \cdot x_{\cdot,1})/n$ | $\cdots$ | $(x_{1,\cdot} \cdot x_{\cdot,d_2})/n$ |
| | $\vdots$ | $\ddots$ | $\vdots$ |
| | $(x_{d_1,\cdot} \cdot x_{\cdot,1})/n$ | $\cdots$ | $(x_{d_1,\cdot} \cdot x_{\cdot,d_2})/n$ |

- 2 features, with $d_1$ and $d_2$ categories, $\rightarrow d_1 \cdot d_2$ 'cells'
- observed frequencies are compared to the expected frequencies under the null hypothesis (estimated from the marginal frequencies)

# Nutshell

| $x_{j,k}$ | | | | |
|---|---|---|---|---|
| | $x_{1,1}$ | $\cdots$ | $x_{1,d_2}$ | $x_{1,\cdot}$ |
| | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| | $x_{d_1,1}$ | $\cdots$ | $x_{d_1,d_2}$ | $x_{d_1,\cdot}$ |
| | $x_{\cdot,1}$ | $\cdots$ | $x_{\cdot,d_2}$ | $n$ |

| $\hat{\mathbb{E}}_{H_0}[\cdot]$ | | | |
|---|---|---|---|
| | $(x_{1,\cdot} \cdot x_{\cdot,1})/n$ | $\cdots$ | $(x_{1,\cdot} \cdot x_{\cdot,d_2})/n$ |
| | $\vdots$ | $\ddots$ | $\vdots$ |
| | $(x_{d_1,\cdot} \cdot x_{\cdot,1})/n$ | $\cdots$ | $(x_{d_1,\cdot} \cdot x_{\cdot,d_2})/n$ |

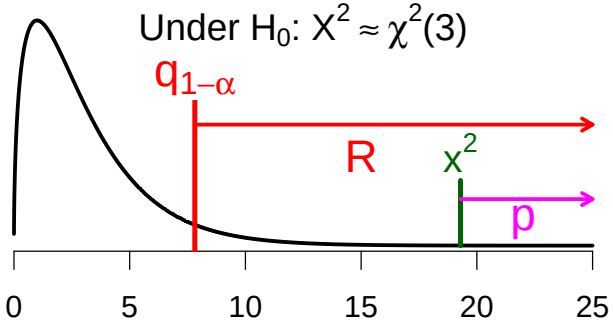
- 2 features, with $d_1$ and $d_2$ categories, $\rightarrow d_1 \cdot d_2$ 'cells'
- observed frequencies are compared to the expected frequencies under the null hypothesis (estimated from the marginal frequencies)
- Comparison ('over all cells') through the $\chi^2$-statistic

$$x^2 = \sum_{j,k} \frac{\left(x_{j,k} - \frac{x_{j,\cdot} \cdot x_{\cdot,k}}{n}\right)^2}{\frac{x_{j,\cdot} \cdot x_{\cdot,k}}{n}}$$

# Nutshell

| $x_{j,k}$ | | | |
|---|---|---|---|
| | $x_{1,1}$ | $\cdots$ | $x_{1,d_2}$ | $x_{1,\cdot}$ |
| | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| | $x_{d_1,1}$ | $\cdots$ | $x_{d_1,d_2}$ | $x_{d_1,\cdot}$ |
| | $x_{\cdot,1}$ | $\cdots$ | $x_{\cdot,d_2}$ | $n$ |

| $\hat{\mathbb{E}}_{H_0}[\cdot]$ | | | |
|---|---|---|---|
| | $(x_{1,\cdot} \cdot x_{\cdot,1})/n$ | $\cdots$ | $(x_{1,\cdot} \cdot x_{\cdot,d_2})/n$ |
| | $\vdots$ | $\ddots$ | $\vdots$ |
| | $(x_{d_1,\cdot} \cdot x_{\cdot,1})/n$ | $\cdots$ | $(x_{d_1,\cdot} \cdot x_{\cdot,d_2})/n$ |



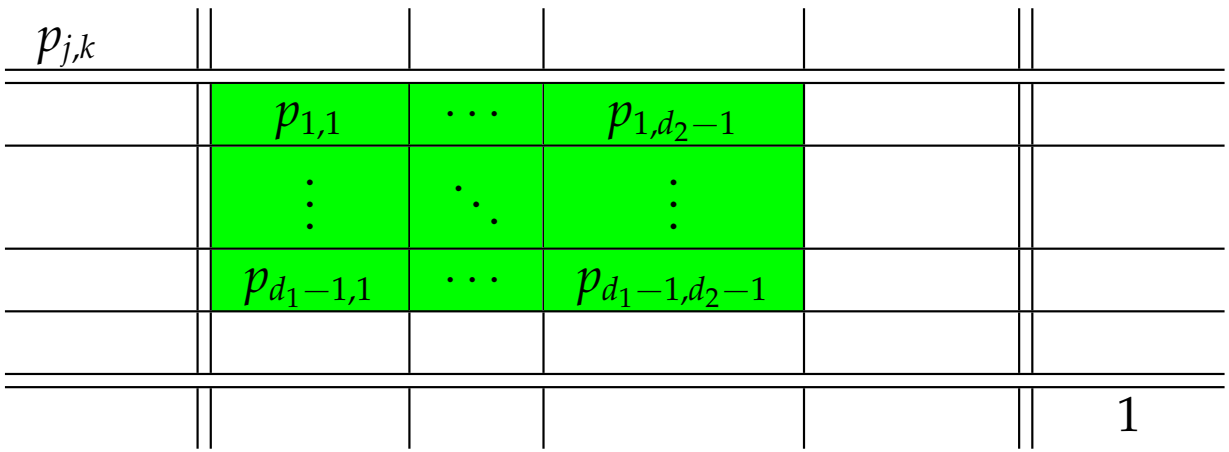

- 2 features, with $d_1$ and $d_2$ categories, $\rightarrow d_1 \cdot d_2$ 'cells'
- observed frequencies are compared to the expected frequencies under the null hypothesis (estimated from the marginal frequencies)
- Comparison ('over all cells') through the $\chi^2$-statistic

$$x^2 = \sum_{j,k} \frac{\left(x_{j,k} - \frac{x_{j,\cdot} \cdot x_{\cdot,k}}{n}\right)^2}{\frac{x_{j,\cdot} \cdot x_{\cdot,k}}{n}}$$

- Judgment of the discrepancy according to the $\chi^2((d_1 - 1) \cdot (d_2 - 1))$-dist.

# Remarks

- It holds: $X^2 \xrightarrow{d} \chi^2((d_1 - 1) \cdot (d_2 - 1))$ under $H_0$ as $n \to \infty$
- Why $(d_1 - 1) \cdot (d_2 - 1)$ degrees of freedom?
- Intuition: only $(d_1 - 1) \cdot (d_2 - 1)$ probabilities can be chosen 'freely'



# Remarks

- It holds: $X^2 \xrightarrow{d} \chi^2((d_1-1)\cdot(d_2-1))$ under $H_0$ as $n \to \infty$
- Why $(d_1-1)\cdot(d_2-1)$ degrees of freedom?
- Intuition: only $(d_1-1)\cdot(d_2-1)$ probabilities can be chosen 'freely'
- The marginal probabilities

| $p_{j,k}$ | | | | | |
|---|---|---|---|---|---|
| | $p_{1,1}$ | $\cdots$ | $p_{1,d_2-1}$ | | $p_{1,\cdot}$ |
| | $\vdots$ | $\ddots$ | $\vdots$ | | $\vdots$ |
| | $p_{d_1-1,1}$ | $\cdots$ | $p_{d_1-1,d_2-1}$ | | $p_{d_1-1,\cdot}$ |
| | | | | | $p_{d_1,\cdot}$ |
| | $p_{\cdot,1}$ | $\cdots$ | $p_{\cdot,d_2-1}$ | $p_{\cdot,d_2}$ | 1 |

# Remarks

- It holds: $X^2 \xrightarrow{d} \chi^2((d_1-1)\cdot(d_2-1))$ under $H_0$ as $n \to \infty$
- Why $(d_1-1)\cdot(d_2-1)$ degrees of freedom?
- Intuition: only $(d_1-1)\cdot(d_2-1)$ probabilities can be chosen 'freely'
- The marginal probabilities already fix the other probabilities e.g.,

$$p_{j,d_2} = p_{j,\cdot} - \sum_{k=1}^{d_2-1} p_{j,k}$$

| $p_{j,k}$ | | | | | |
|---|---|---|---|---|---|
| | $p_{1,1}$ | $\cdots$ | $p_{1,d_2-1}$ | $p_{1,d_2}$ | $p_{1,\cdot}$ |
| | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | $p_{d_1-1,1}$ | $\cdots$ | $p_{d_1-1,d_2-1}$ | $p_{d_1-1,d_2}$ | $p_{d_1-1,\cdot}$ |
| | $p_{d_1,1}$ | $\cdots$ | $p_{d_1,d_2-1}$ | $p_{d_1,d_2}$ | $p_{d_1,\cdot}$ |
| | $p_{\cdot,1}$ | $\cdots$ | $p_{\cdot,d_2-1}$ | $p_{\cdot,d_2}$ | 1 |

# Remarks

- It holds: $X^2 \xrightarrow{d} \chi^2((d_1-1)\cdot(d_2-1))$ under $H_0$ as $n \to \infty$
- Why $(d_1-1)\cdot(d_2-1)$ degrees of freedom?
- Intuition: only $(d_1-1)\cdot(d_2-1)$ probabilities can be chosen 'freely'
- The marginal probabilities already fix the other probabilities e.g.,

$$p_{j,d_2} = p_{j,\cdot} - \sum_{k=1}^{d_2-1} p_{j,k}$$

| $p_{j,k}$ | | | | | |
|---|---|---|---|---|---|
| | $p_{1,1}$ | $\cdots$ | $p_{1,d_2-1}$ | $p_{1,d_2}$ | $p_{1,\cdot}$ |
| | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | $p_{d_1-1,1}$ | $\cdots$ | $p_{d_1-1,d_2-1}$ | $p_{d_1-1,d_2}$ | $p_{d_1-1,\cdot}$ |
| | $p_{d_1,1}$ | $\cdots$ | $p_{d_1,d_2-1}$ | $p_{d_1,d_2}$ | $p_{d_1,\cdot}$ |
| | $p_{\cdot,1}$ | $\cdots$ | $p_{\cdot,d_2-1}$ | $p_{\cdot,d_2}$ | 1 |

- The convergence of $X^2$ to the $\chi^2$-distribution is again reasonable, as according to the central limit theorem we find every summand approx distributed like a square of a $N(0,1)$-distributed random variable (and just $(d_1-1)\cdot(d_2-1)$ summands are 'free').

# $\chi^2$-test in R

```
# Enter data
die_solid       <- c("red","blue","blue",...)
die_soft        <- c("blue","green","blue",...)
# Compute frequencies, e.g., via
x_solid <- table(die_solid)
x_soft  <- table(die_soft)
# Combine frequencies, e.g., as a matrix
x <- rbind(x_solid,x_soft)
x
     1  2  3  4
[1,] 25 22 19 14
[2,] 27 19 44 60
# Perform chi^2-test
chisq.test(x)
# Output

        Pearson's Chi-squared test

data:  x
X-squared = 19.295, df = 3, p-value = 0.0002376
```

# Questions

On which test could you think?

- Does the property 'smoker' (yes/no) depend on the age group (young/old)?

# Questions

On which test could you think?

- Does the property 'smoker' (yes/no) depend on the age group (young/old)?
  $\rightarrow \chi^2$-test for independence (2 features, smoker and age group)

# Questions

On which test could you think?

- Does the property 'smoker' (yes/no) depend on the age group (young/old)?
  $\rightarrow$ $\chi^2$-test for independence (2 features, smoker and age group)
- Is the proportion of students, that regularly visit the cafeteria the same as the proportion of those that do not visit the cafeteria?

# Questions

On which test could you think?

- Does the property 'smoker' (yes/no) depend on the age group (young/old)?
  $\rightarrow \chi^2$-test for independence (2 features, smoker and age group)
- Is the proportion of students, that regularly visit the cafeteria the same as the proportion of those that do not visit the cafeteria?
  $\rightarrow \chi^2$-goodness of fit test (1 feature, regular visit of cafeteria, $p_0 = 50\%$)

# Questions

On which test could you think?

- Does the property 'smoker' (yes/no) depend on the age group (young/old)?
  $\rightarrow \chi^2$-test for independence (2 features, smoker and age group)
- Is the proportion of students, that regularly visit the cafeteria the same as the proportion of those that do not visit the cafeteria?
  $\rightarrow \chi^2$-goodness of fit test (1 feature, regular visit of cafeteria, $p_0 = 50\%$)
- Between the majors math, physics and computer science, is there a difference in the proportions of students that regularly drink coffee?

# Questions

On which test could you think?

- Does the property 'smoker' (yes/no) depend on the age group (young/old)?
  $\rightarrow \chi^2$-test for independence (2 features, smoker and age group)
- Is the proportion of students, that regularly visit the cafeteria the same as the proportion of those that do not visit the cafeteria?
  $\rightarrow \chi^2$-goodness of fit test (1 feature, regular visit of cafeteria, $p_0 = 50\%$)
- Between the majors math, physics and computer science, is there a difference in the proportions of students that regularly drink coffee?
  $\rightarrow \chi^2$-test for independence (2 features, coffee drinker and major)

# Questions

On which test could you think?

- Does the property 'smoker' (yes/no) depend on the age group (young/old)?
  $\rightarrow \chi^2$-test for independence (2 features, smoker and age group)
- Is the proportion of students, that regularly visit the cafeteria the same as the proportion of those that do not visit the cafeteria?
  $\rightarrow \chi^2$-goodness of fit test (1 feature, regular visit of cafeteria, $p_0 = 50\%$)
- Between the majors math, physics and computer science, is there a difference in the proportions of students that regularly drink coffee?
  $\rightarrow \chi^2$-test for independence (2 features, coffee drinker and major)
- Between the sports basketball, table tennis and swimming, is there a difference in the mean body size (in m) of the sportsmen?

# Questions

On which test could you think?

- Does the property 'smoker' (yes/no) depend on the age group (young/old)?
  $\rightarrow \chi^2$-test for independence (2 features, smoker and age group)
- Is the proportion of students, that regularly visit the cafeteria the same as the proportion of those that do not visit the cafeteria?
  $\rightarrow \chi^2$-goodness of fit test (1 feature, regular visit of cafeteria, $p_0 = 50\%$)
- Between the majors math, physics and computer science, is there a difference in the proportions of students that regularly drink coffee?
  $\rightarrow \chi^2$-test for independence (2 features, coffee drinker and major)
- Between the sports basketball, table tennis and swimming, is there a difference in the mean body size (in m) of the sportsmen?
  $\rightarrow$ ANOVA, metric data

Thank you!