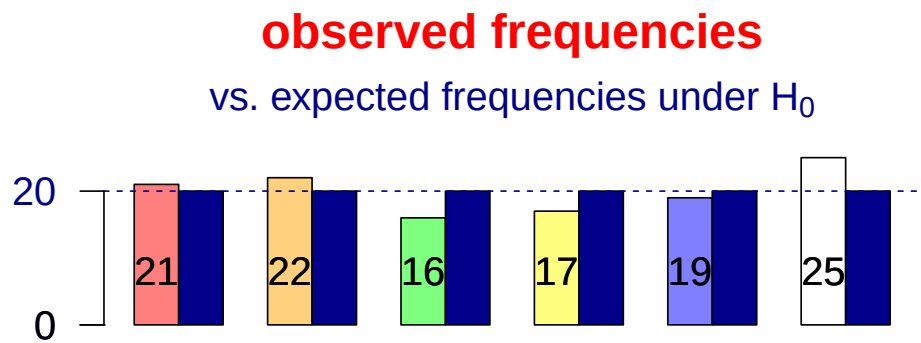
The χ^2-test (goodness of fit)



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.

Motivation

- Is the die fair?

Motivation

- Is the die fair?
- Given a six-sided die
  with sides red, orange, green, yellow, bue and white

Motivation

- Is the die fair?
- Given a six-sided die
  with sides red, orange, green, yellow, bue and white
- Somebody: 'the die is fair!'

Motivation

- Is the die fair?
- Given a six-sided die
  with sides red, orange, green, yellow, bue and white
- Somebody: 'the die is fair!'
- What to do?

Motivation

- Is the die fair?
- Given a six-sided die
  with sides red, orange, green, yellow, bue and white
- Somebody: 'the die is fair!'
- What to do?
- Roll the die! (collect data)

Motivation

- Is the die fair?
- Given a six-sided die
  with sides red, orange, green, yellow, bue and white
- Somebody: 'the die is fair!'
- What to do?
- Roll the die! (collect data)
- roll the die 120 times

Motivation

- Is the die fair?
- Given a six-sided die
  with sides red, orange, green, yellow, bue and white
- Somebody: 'the die is fair!'
- What to do?
- Roll the die! (collect data)
- roll the die 120 times
- the outcome was

  red, blue, blue, white , red, green, orange, green, . . . , orange

Motivation

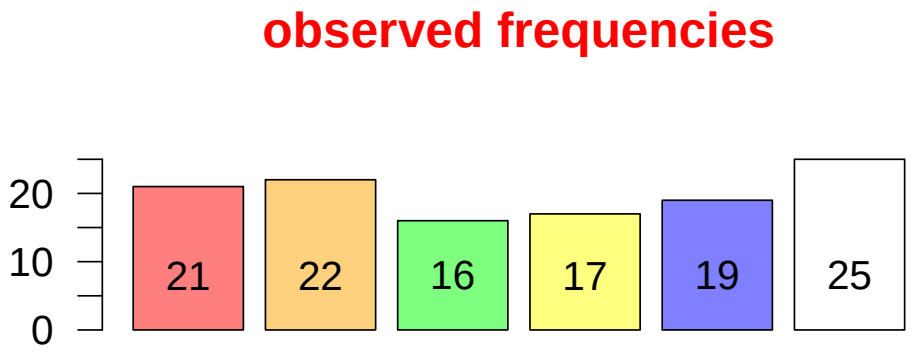
- Is the die fair?
- Given a six-sided die
  with sides red, orange, green, yellow, bue and white
- Somebody: 'the die is fair!'
- What to do?
- Roll the die! (collect data)
- roll the die 120 times
- the outcome was

  red, blue, blue, white , red, green, orange, green, . . . , orange
- Once again: hard to 'understand' anything,
  $\rightarrow$ thus graphical representation, e.g., in the barplot

Motivation

- Is the die fair?
- Given a six-sided die
  with sides red, orange, green, yellow, bue and white
- Somebody: 'the die is fair!'
- What to do?
- Roll the die! (collect data)
- roll the die 120 times
- the outcome was

  red, blue, blue, white , red, green, orange, green, . . . , orange
- Once again: hard to 'understand' anything,
  $\rightarrow$ thus graphical representation, e.g., in the barplot

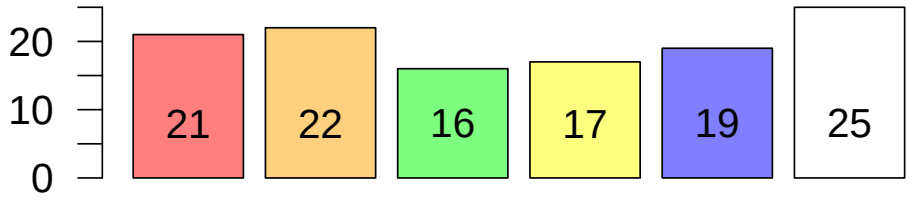


Motivation

- Is the die fair?
- Given a six-sided die
  with sides red, orange, green, yellow, bue and white
- Somebody: 'the die is fair!'
- What to do?
- Roll the die! (collect data)
- roll the die 120 times
- the outcome was

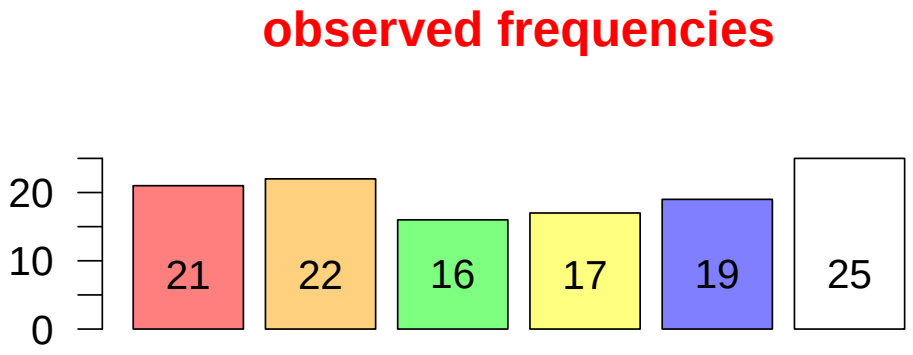
  red, blue, blue, white , red, green, orange, green, . . . , orange
- Once again: hard to 'understand' anything,
  $\rightarrow$ thus graphical representation, e.g., in the barplot
  $\rightarrow$ categorial data
  6 catagories (red, orange, green, yelllow, blue, white)

observed frequencies



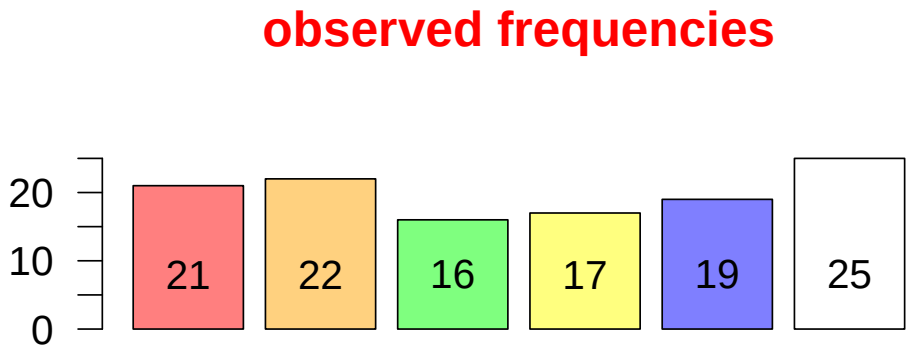
Motivation

- 120 throws



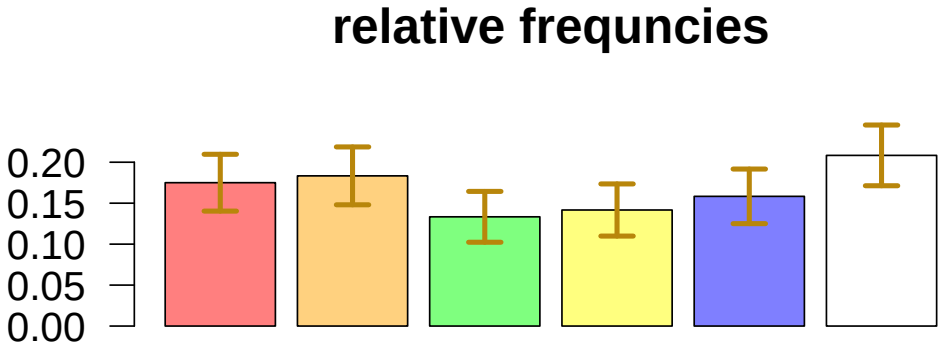
Motivation

- 120 throws
- Question:
  Are the observed frequencies far away from each other?



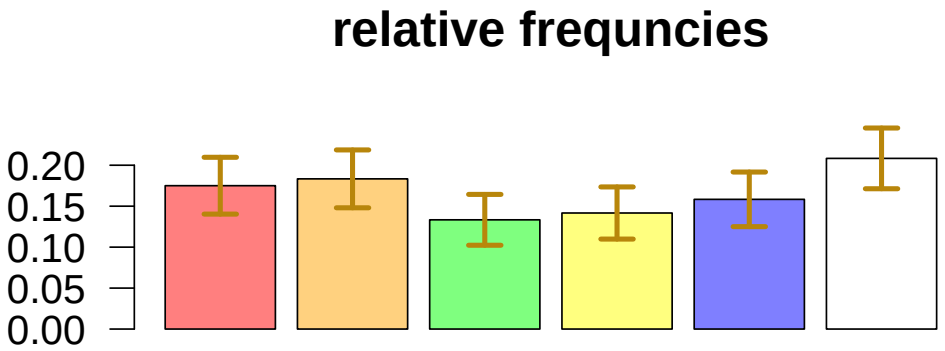
Motivation

- 120 throws
- Question:
  Are the observed frequencies far away from each other?
- 'Answer' 1:
  relative frequencies and standard error $\rightarrow$ rather close?!



Motivation

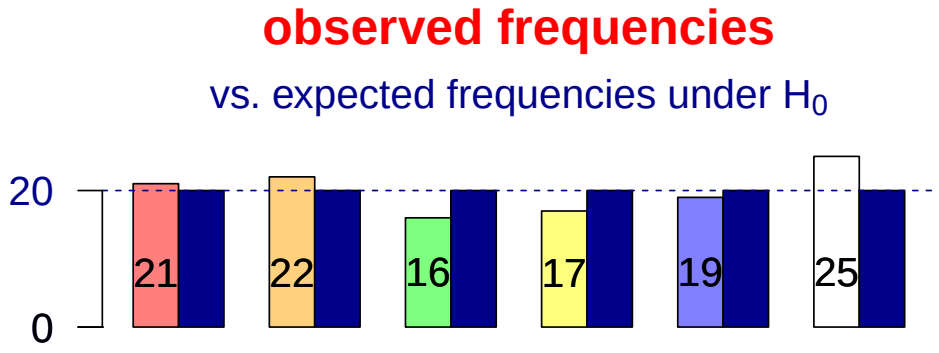
- 120 throws
- Question:
  Are the observed frequencies far away from each other?
- 'Answer' 1:
  relative frequencies and standard error $\rightarrow$ rather close?!



- 'Answer' 2:
  - meaning of 'fair': no side is preferred, probability 1/6

Motivation

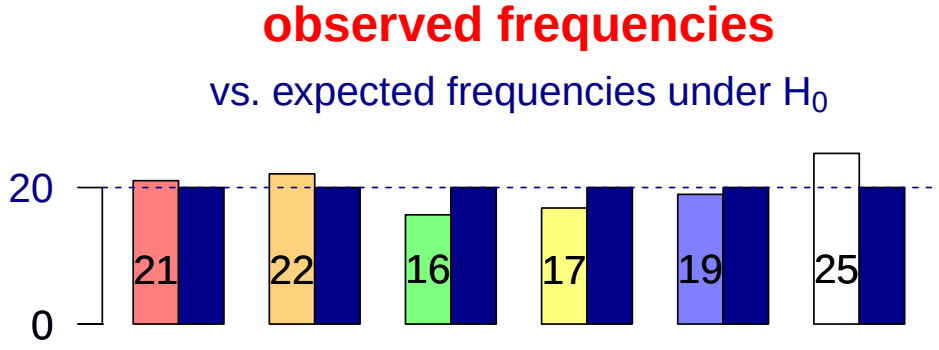
- 120 throws
- Question:
  Are the observed frequencies far away from each other?
- 'Answer' 1:
  relative frequencies and standard error $\rightarrow$ rather close?!



- 'Answer' 2:
  - meaning of 'fair': no side is preferred, probability 1/6
  - per category $120 \cdot 1/6 = 20$ occupations expected, if the die is fair

Motivation

- 120 throws
- Question:
  Are the observed frequencies far away from each other?
- 'Answer' 1:
  relative frequencies and standard error $\rightarrow$ rather close?!

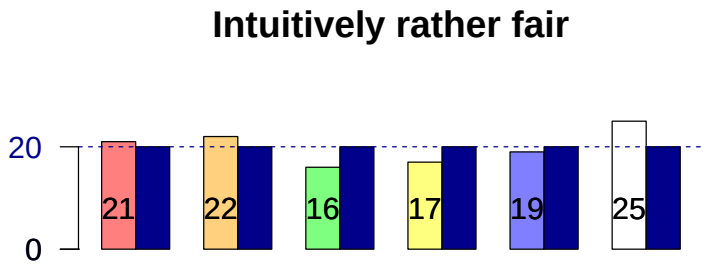


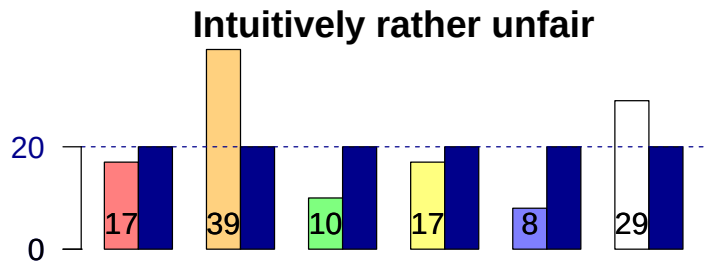
- 'Answer' 2:
  - meaning of 'fair': no side is preferred, probability 1/6
  - per category $120 \cdot 1/6 = 20$ occupations expected, if the die is fair
  - in every category the observed frequencies should then typically be 'close' to the expected frequencies

Motivation

- 120 throws
- Question:
  Are the observed frequencies far away from each other?
- 'Answer' 1:

  relative frequencies and standard error $\rightarrow$ rather close?!



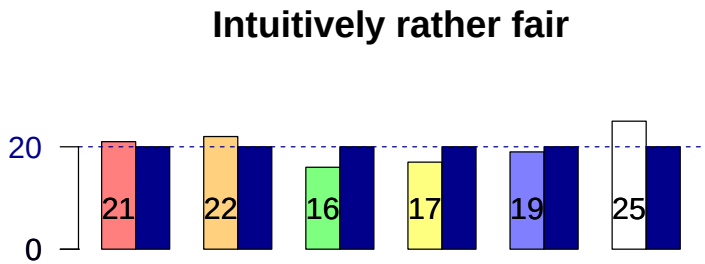


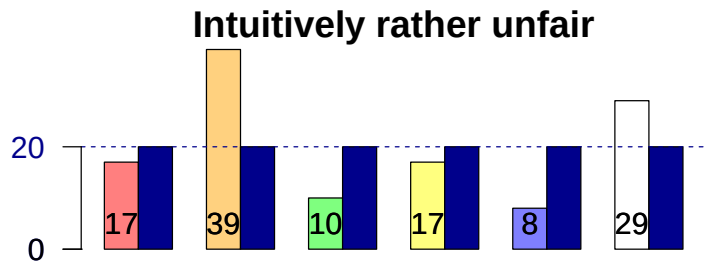
- 'Answer' 2:
  - meaning of 'fair': no side is preferred, probability 1/6
  - per category $120 \cdot 1/6 = 20$ occupations expected, if the die is fair
  - in every category the observed frequencies should then typically be 'close' to the expected frequencies

Motivation

- 120 throws
- Question:
  Are the observed frequencies far away from each other?
- 'Answer' 1:

  relative frequencies and standard error $\rightarrow$ rather close?!



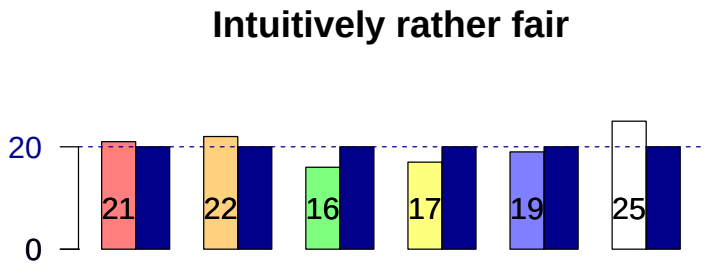


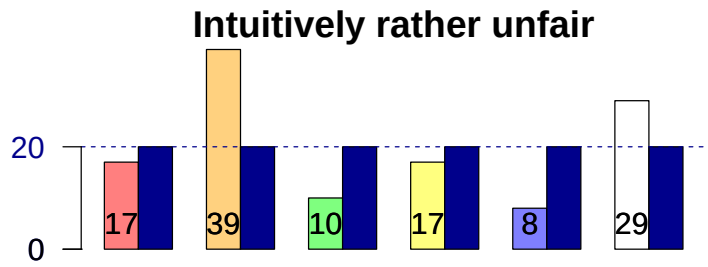
- 'Answer' 2:
  - meaning of 'fair': no side is preferred, probability 1/6
  - per category $120 \cdot 1/6 = 20$ occupations expected, if the die is fair
  - in every category the observed frequencies should then typically be 'close' to the expected frequencies
  - a statistic, that quantifies this discrepancy over all categories, is the χ^2-*statistic*

Motivation

- 120 throws
- Question:
  Are the observed frequencies far away from each other?
- 'Answer' 1:

  relative frequencies and standard error → rather close?!



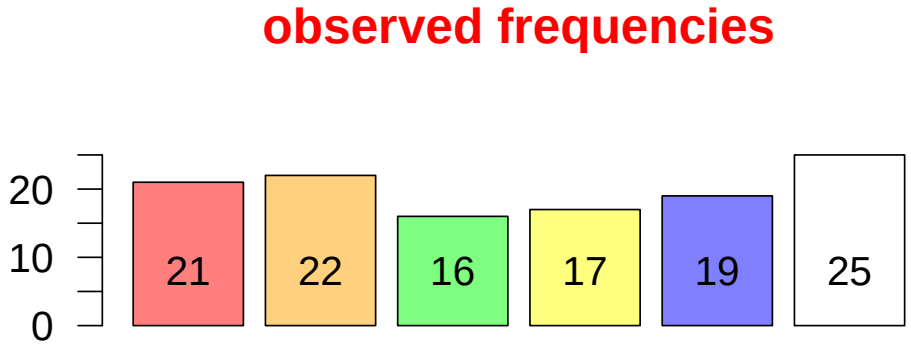


- 'Answer' 2:
  - meaning of 'fair': no side is preferred, probability 1/6
  - per category $120 \cdot 1/6 = 20$ occupations expected, if the die is fair
  - in every category the observed frequencies should then typically be 'close' to the expected frequencies
  - a statistic, that quantifies this discrepancy over all categories, is the χ^2-*statistic*
  - → in the following we construct the so-called χ^2-*test*

Observed and expected frequencies

Notation

- n data (here: $n = 120$)



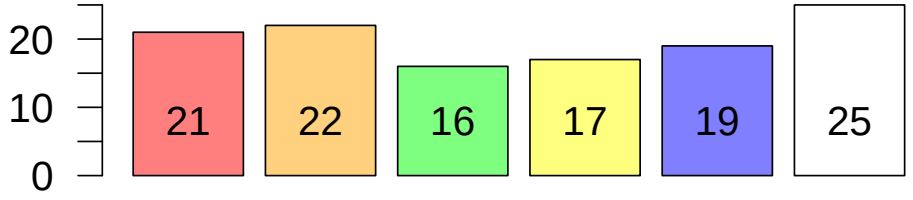
Observed and expected frequencies

Notation

- n data (here: $n = 120$)
- fall in d categories (here: $d = 6$)

k	1	2	3	4	5	6	$\sum$

observed frequencies

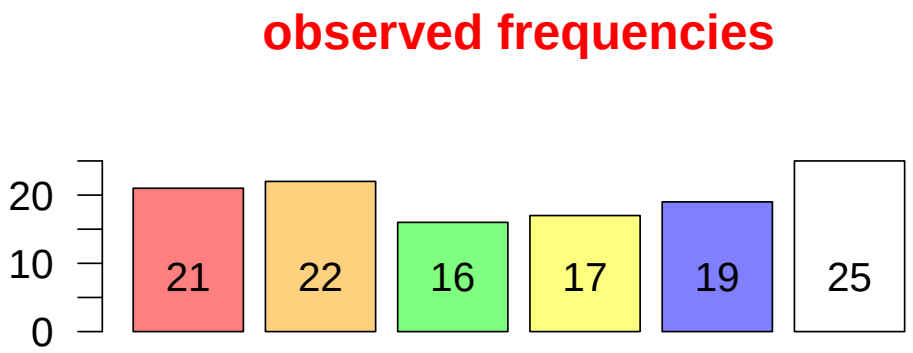


Observed and expected frequencies

Notation

- n data (here: $n = 120$)
- fall in d categories (here: $d = 6$)
- x_k denotes the number of occupations (number of data) in the k-th category $\rightarrow$ observed frequencies

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

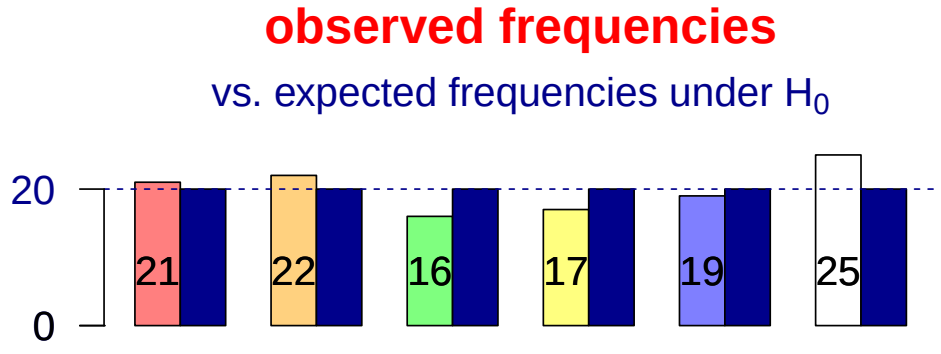


Observed and expected frequencies

Notation

- n data (here: $n = 120$)
- fall in d categories (here: $d = 6$)
- x_k denotes the number of occupations (number of data) in the k-th category $\rightarrow$ observed frequencies
- these are compared to the expected frequencies, assuming that the die is fair

k	1	2	3	4	5	6	$\sum$
x_k	21	22	16	17	19	25	120
expected frequencies, if 'fair'	20	20	20	20	20	20	120

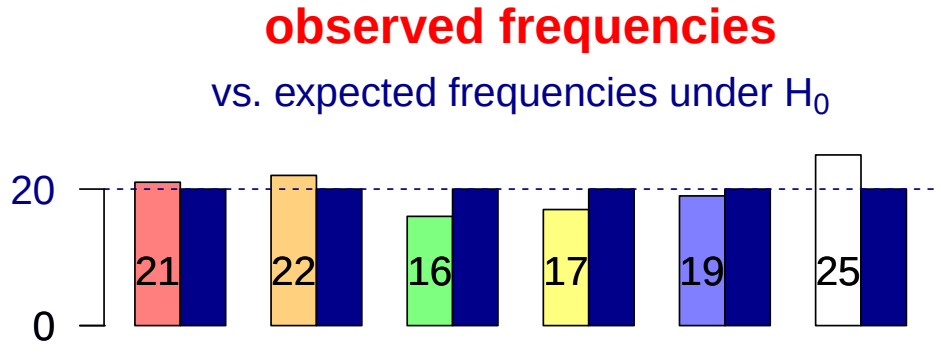


Observed and expected frequencies

Notation

- n data (here: $n = 120$)
- fall in d categories (here: $d = 6$)
- x_k denotes the number of occupations (number of data) in the k-th category $\rightarrow$ observed frequencies
- these are compared to the expected frequencies, assuming that the die is fair
- in order to talk about expectations we need a model

k	1	2	3	4	5	6	$\sum$
x_k	21	22	16	17	19	25	120
expected frequencies, if 'fair'	20	20	20	20	20	20	120



From the binomial to the multinomial distribution

- Which model could we choose with $d = 2$ categories?

From the binomial to the multinomial distribution

- Which model could we choose with $d = 2$ categories?
- categories 'success' and 'failure'

From the binomial to the multinomial distribution

- Which model could we choose with $d = 2$ categories?
- categories 'success' and 'failure'
- Reminder: A random variable X is called *binomial* distributed with parameters n and p, short $X \sim b(n, p)$, if

$$\boxed{\mathbb{P}(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}} \qquad (*)$$

From the binomial to the multinomial distribution

- Which model could we choose with $d = 2$ categories?
- categories 'success' and 'failure'
- Reminder: A random variable X is called *binomial* distributed with parameters n and p, short $X \sim b(n, p)$, if

$$\boxed{\mathbb{P}(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}} \qquad (*)$$

- while $x \in \{0, 1, \ldots, n\}$ (number of successes)

From the binomial to the multinomial distribution

- Which model could we choose with $d = 2$ categories?
- categories 'success' and 'failure'
- Reminder: A random variable X is called *binomial* distributed with parameters n and p, short $X \sim b(n, p)$, if

$$\boxed{\mathbb{P}(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}} \qquad (*)$$

- while $x \in \{0, 1, \ldots, n\}$ (number of successes)
- $p \in (0, 1)$ (success probability)

From the binomial to the multinomial distribution

- Which model could we choose with $d = 2$ categories?
- categories 'success' and 'failure'
- Reminder: A random variable X is called *binomial* distributed with parameters n and p, short $X \sim b(n,p)$, if

$$\boxed{\mathbb{P}(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}} \qquad (*)$$

- while $x \in \{0, 1, \ldots, n\}$ (number of successes)
- $p \in (0,1)$ (success probability)
- and *binomial coefficient*

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

From the binomial to the multinomial distribution

- Which model could we choose with $d = 2$ categories?
- categories 'success' and 'failure'
- Reminder: A random variable X is called *binomial* distributed with parameters n and p, short $X \sim b(n,p)$, if

$$\boxed{\mathbb{P}(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}} \qquad (*)$$

- while $x \in \{0, 1, \ldots, n\}$ (number of successes)
- $p \in (0,1)$ (success probability)
- and *binomial coefficient*

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

- read $(*)$: in n independent 'coin flips' observe x times a success, each with probability p. The binomial coefficient states in how many ways the x successes may have appeared.

From the binomial to the multinomial distribution

- Which model could we choose with $d = 2$ categories?
- categories 'success' and 'failure'
- Reminder: A random variable X is called *binomial* distributed with parameters n and p, short $X \sim b(n,p)$, if

$$\boxed{\mathbb{P}(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}} \qquad (*)$$

- while $x \in \{0, 1, \ldots, n\}$ (number of successes)
- $p \in (0,1)$ (success probability)
- and *binomial coefficient*

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

- read $(*)$: in n independent 'coin flips' observe x times a success, each with probability p. The binomial coefficient states in how many ways the x successes may have appeared.
- it is $\boxed{\mathbb{E}[X] = n \cdot p}$ $\rightarrow$ expected number of successes

From the binomial to the multinomial distribution

- Which model could we choose with $d = 2$ categories?
- categories 'success' and 'failure'
- Reminder: A random variable X is called *binomial* distributed with parameters n and p, short $X \sim b(n,p)$, if

$$\boxed{\mathbb{P}(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}} \qquad (*)$$

- while $x \in \{0, 1, \ldots, n\}$ (number of successes)
- $p \in (0,1)$ (success probability)
- and *binomial coefficient*

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

- read $(*)$: in n independent 'coin flips' observe x times a success, each with probability p. The binomial coefficient states in how many ways the x successes may have appeared.
- it is $\boxed{\mathbb{E}[X] = n \cdot p}$ $\rightarrow$ expected number of successes
- extension to d categories $\rightarrow$ multinomial distribution...

Multinomial distribution

- Definition: A random vector $\mathfrak{X} = (X_1, \ldots, X_d)^t$ is called *multinomial distributed* with parameters n and $p = (p_1, \ldots, p_d)^t$, short $\mathfrak{X} \sim mult(n, p)$, if

$$\boxed{\mathbb{P}(\mathfrak{X} = (x_1, \ldots, x_d)^t) = \binom{n}{x_1, x_2, \ldots, x_d} \prod_{k=1}^{d} p_k^{x_k}} \qquad (*)$$

Multinomial distribution

- Definition: A random vector $\mathfrak{X} = (X_1, \ldots, X_d)^t$ is called *multinomial distributed* with parameters n and $p = (p_1, \ldots, p_d)^t$, short $\mathfrak{X} \sim mult(n, p)$, if

$$\boxed{\mathbb{P}(\mathfrak{X} = (x_1, \ldots, x_d)^t) = \binom{n}{x_1, x_2, \ldots, x_d} \prod_{k=1}^{d} p_k^{x_k}} \qquad (*)$$

- while $(x_1, \ldots, x_d)^t \in \mathbb{N}^d$ with $\sum_{k=1}^{d} x_k = n$ (number of occupations)

Multinomial distribution

- Definition: A random vector $\mathfrak{X} = (X_1, \ldots, X_d)^t$ is called *multinomial distributed* with parameters n and $p = (p_1, \ldots, p_d)^t$, short $\mathfrak{X} \sim mult(n, p)$, if

$$\boxed{\mathbb{P}(\mathfrak{X} = (x_1, \ldots, x_d)^t) = \binom{n}{x_1, x_2, \ldots, x_d} \prod_{k=1}^{d} p_k^{x_k}} \qquad (*)$$

- while $(x_1, \ldots, x_d)^t \in \mathbb{N}^d$ with $\sum_{k=1}^{d} x_k = n$ (number of occupations)
- $p = (p_1, \ldots, p_d)^t \in (0,1)^d$ with $\sum_{k=1}^{d} p_k = 1$ (probabilities for occupations)

Multinomial distribution

- Definition: A random vector $\mathfrak{X} = (X_1, \ldots, X_d)^t$ is called *multinomial distributed* with parameters n and $p = (p_1, \ldots, p_d)^t$, short $\mathfrak{X} \sim mult(n, p)$, if

$$\boxed{\mathbb{P}(\mathfrak{X} = (x_1, \ldots, x_d)^t) = \binom{n}{x_1, x_2, \ldots, x_d} \prod_{k=1}^{d} p_k^{x_k}} \qquad (*)$$

- while $(x_1, \ldots, x_d)^t \in \mathbb{N}^d$ with $\sum_{k=1}^{d} x_k = n$ (number of occupations)
- $p = (p_1, \ldots, p_d)^t \in (0,1)^d$ with $\sum_{k=1}^{d} p_k = 1$ (probabilities for occupations)
- with *multinomial coefficient*

$$\binom{n}{x_1, x_2, \ldots, x_d} := \frac{n!}{x_1! \cdots x_d!}$$

Multinomial distribution

- Definition: A random vector $\mathfrak{X} = (X_1, \ldots, X_d)^t$ is called *multinomial distributed* with parameters n and $p = (p_1, \ldots, p_d)^t$, short $\mathfrak{X} \sim mult(n, p)$, if

$$\boxed{\mathbb{P}(\mathfrak{X} = (x_1, \ldots, x_d)^t) = \binom{n}{x_1, x_2, \ldots, x_d} \prod_{k=1}^{d} p_k^{x_k}} \qquad (*)$$

- while $(x_1, \ldots, x_d)^t \in \mathbb{N}^d$ with $\sum_{k=1}^{d} x_k = n$ (number of occupations)
- $p = (p_1, \ldots, p_d)^t \in (0,1)^d$ with $\sum_{k=1}^{d} p_k = 1$ (probabilities for occupations)
- with *multinomial coefficient*

$$\binom{n}{x_1, x_2, \ldots, x_d} := \frac{n!}{x_1! \cdots x_d!} = \binom{n}{x_1}\binom{n - x_1}{x_2}\binom{n - x_1 - x_2}{x_3} \cdots \binom{n - x_1 - \cdots - x_{d-1}}{x_d}$$

Multinomial distribution

- Definition: A random vector $\mathfrak{X} = (X_1, \ldots, X_d)^t$ is called *multinomial distributed* with parameters n and $p = (p_1, \ldots, p_d)^t$, short $\mathfrak{X} \sim mult(n, p)$, if

$$\boxed{\mathbb{P}(\mathfrak{X} = (x_1, \ldots, x_d)^t) = \binom{n}{x_1, x_2, \ldots, x_d} \prod_{k=1}^{d} p_k^{x_k}} \qquad (*)$$

- while $(x_1, \ldots, x_d)^t \in \mathbb{N}^d$ with $\sum_{k=1}^{d} x_k = n$ (number of occupations)
- $p = (p_1, \ldots, p_d)^t \in (0,1)^d$ with $\sum_{k=1}^{d} p_k = 1$ (probabilities for occupations)
- with *multinomial coefficient*

$$\binom{n}{x_1, x_2, \ldots, x_d} := \frac{n!}{x_1! \cdots x_d!} = \binom{n}{x_1}\binom{n - x_1}{x_2}\binom{n - x_1 - x_2}{x_3} \cdots \binom{n - x_1 - \cdots - x_{d-1}}{x_d}$$

- read $(*)$: in n independent 'occupations' of d categories in which the k-th category is chosen with probability p_k, the k-th category was occupied x_k times. The multinomial coefficient states in how many ways the observed occupations of all categories may have appeared ($\rightarrow$ order)

Multinomial distribution

- Definition: A random vector $\mathfrak{X} = (X_1, \ldots, X_d)^t$ is called *multinomial distributed* with parameters n and $p = (p_1, \ldots, p_d)^t$, short $\mathfrak{X} \sim mult(n, p)$, if

$$\boxed{\mathbb{P}(\mathfrak{X} = (x_1, \ldots, x_d)^t) = \binom{n}{x_1, x_2, \ldots, x_d} \prod_{k=1}^{d} p_k^{x_k}} \qquad (*)$$

- while $(x_1, \ldots, x_d)^t \in \mathbb{N}^d$ with $\sum_{k=1}^{d} x_k = n$ (number of occupations)
- $p = (p_1, \ldots, p_d)^t \in (0,1)^d$ with $\sum_{k=1}^{d} p_k = 1$ (probabilities for occupations)
- with *multinomial coefficient*

$$\binom{n}{x_1, x_2, \ldots, x_d} := \frac{n!}{x_1! \cdots x_d!} = \binom{n}{x_1}\binom{n - x_1}{x_2}\binom{n - x_1 - x_2}{x_3} \cdots \binom{n - x_1 - \cdots - x_{d-1}}{x_d}$$

- read $(*)$: in n independent 'occupations' of d categories in which the k-th category is chosen with probability p_k, the k-th category was occupied x_k times. The multinomial coefficient states in how many ways the observed occupations of all categories may have appeared ($\rightarrow$ order)
- For $d = 2$ the weights equal the binomial weights
  $\rightarrow$ multinomial distribution is 'extension' to d categories.

Multinomial distribution

- Definition: A random vector $\mathfrak{X} = (X_1, \ldots, X_d)^t$ is called *multinomial distributed* with parameters n and $p = (p_1, \ldots, p_d)^t$, short $\mathfrak{X} \sim mult(n, p)$, if

$$\boxed{\mathbb{P}(\mathfrak{X} = (x_1, \ldots, x_d)^t) = \binom{n}{x_1, x_2, \ldots, x_d} \prod_{k=1}^{d} p_k^{x_k}} \qquad (*)$$

- while $(x_1, \ldots, x_d)^t \in \mathbb{N}^d$ with $\sum_{k=1}^{d} x_k = n$ (number of occupations)
- $p = (p_1, \ldots, p_d)^t \in (0,1)^d$ with $\sum_{k=1}^{d} p_k = 1$ (probabilities for occupations)
- with *multinomial coefficient*

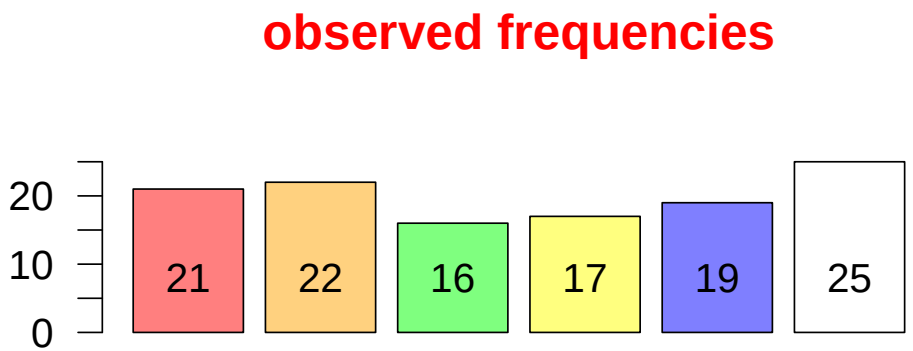
$$\binom{n}{x_1, x_2, \ldots, x_d} := \frac{n!}{x_1! \cdots x_d!} = \binom{n}{x_1}\binom{n-x_1}{x_2}\binom{n-x_1-x_2}{x_3}\cdots\binom{n-x_1-\cdots-x_{d-1}}{x_d}$$

- read $(*)$: in n independent 'occupations' of d categories in which the k-th category is chosen with probability p_k, the k-th category was occupied x_k times. The multinomial coefficient states in how many ways the observed occupations of all categories may have appeared ($\rightarrow$ order)
- For $d = 2$ the weights equal the binomial weights
  $\rightarrow$ multinomial distribution is 'extension' to d categories.
- For the k-th component it holds $X_k \sim b(n, p_k)$,
  thus particularly $\boxed{\mathbb{E}[X_k] = n \cdot p_k}$ $\rightarrow$ 'expected frequencies'

Model and null hypothesis

- n data in d categories (here: $n = 120$, $d = 6$)
- observed frequencies: $x_1, \cdots, x_d$

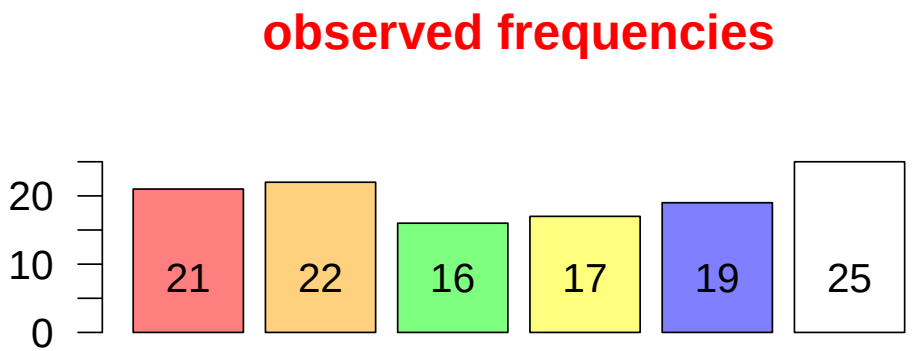
k	1	2	3	4	5	6	$\sum$
x_k	21	22	16	17	19	25	120



Model and null hypothesis

- n data in d categories (here: $n = 120$, $d = 6$)
- observed frequencies: $x_1, \cdots, x_d$
- <u>model:</u> let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$

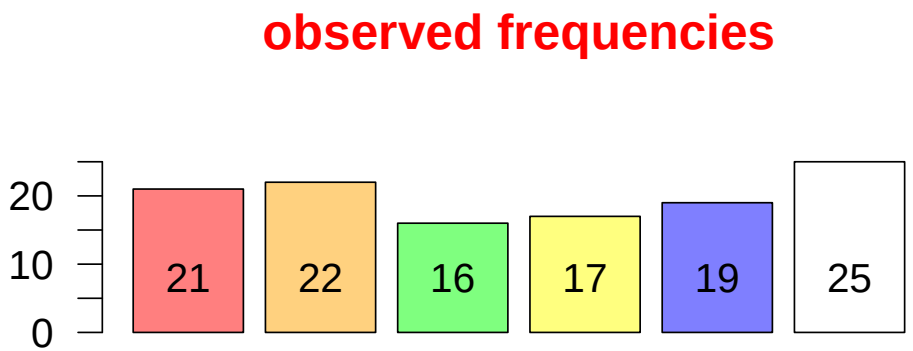
k	1	2	3	4	5	6	$\sum$
x_k	21	22	16	17	19	25	120



Model and null hypothesis

- n data in d categories (here: $n = 120$, $d = 6$)
- 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: $H_0 : p = p_0 := (p_{0,1}, \ldots, p_{0,d})^t$
  claimed occupation probs (here: $p_0 = (1/d, 1/d, \ldots, 1/d)^t \leftrightarrow$ 'fair')

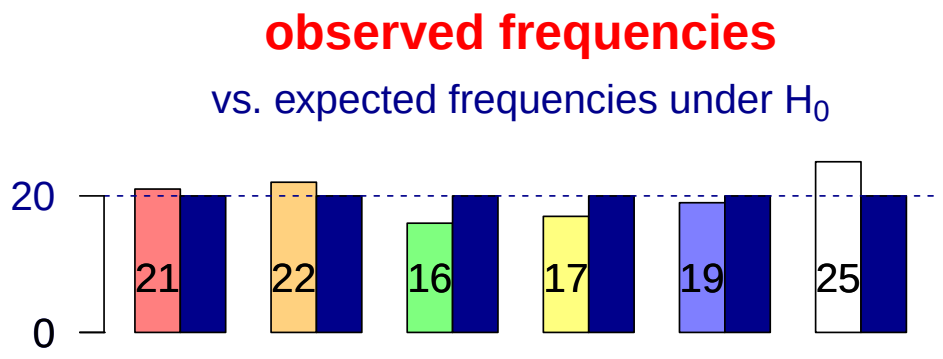
k	1	2	3	4	5	6	$\sum$
x_k	21	22	16	17	19	25	120



Model and null hypothesis

- n data in d categories (here: $n = 120$, $d = 6$)
- 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: $H_0 : p = p_0 := (p_{0,1}, \ldots, p_{0,d})^t$
  claimed occupation probs (here: $p_0 = (1/d, 1/d, \ldots, 1/d)^t \leftrightarrow$ 'fair')
- Under H_0 expected occupations: $\mathbb{E}_{H_0}[X_k] = n \cdot p_{0,k}$
  here $\mathbb{E}_{H_0}[X_k] = 20$, i.e.,. under H_0 there are 20 expected per category

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 χ^2-statistic

- n data in d categories
- observed frequencies $x_1, \cdots, x_d$
- model: $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$
- $\mathbb{E}_{H_0}[X_k] = n \cdot p_{0,k}$

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 χ^2-statistic

- n data in d categories
- observed frequencies $x_1, \cdots, x_d$
- model: $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$
- $\mathbb{E}_{H_0}[X_k] = n \cdot p_{0,k}$
- the χ^2-statistic

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

measures the discrepancy of the observed frequencies from the expected frequencies under the null hypothesis

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 χ^2-statistic

- n data in d categories
- observed frequencies $x_1, \cdots, x_d$
- model: $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$
- $\mathbb{E}_{H_0}[X_k] = n \cdot p_{0,k}$
- the χ^2-statistic

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

measures the discrepancy of the observed frequencies from the expected frequencies under the null hypothesis

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 χ^2-statistic

- n data in d categories
- observed frequencies $x_1, \cdots, x_d$
- model: $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$
- $\mathbb{E}_{H_0}[X_k] = n \cdot p_{0,k}$
- the χ^2-statistic

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

measures the discrepancy of the observed frequencies from the expected frequencies under the null hypothesis

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 χ^2-statistic

- n data in d categories
- observed frequencies $x_1, \cdots, x_d$
- model: $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$
- $\mathbb{E}_{H_0}[X_k] = n \cdot p_{0,k}$
- the χ^2-statistic

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

measures the discrepancy of the observed frequencies from the expected frequencies under the null hypothesis

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 χ^2-statistic

- n data in d categories
- observed frequencies $x_1, \cdots, x_d$
- model: $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$
- $\mathbb{E}_{H_0}[X_k] = n \cdot p_{0,k}$
- the χ^2-statistic

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

measures the discrepancy of the observed frequencies from the expected frequencies under the null hypothesis

- A large positive value of x^2 means a large discrepancy ('positive' due to squares)

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 χ^2-statistic

- n data in d categories
- observed frequencies $x_1, \cdots, x_d$
- model: $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$
- $\mathbb{E}_{H_0}[X_k] = n \cdot p_{0,k}$
- the χ^2-statistic

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

measures the discrepancy of the observed frequencies from the expected frequencies under the null hypothesis

- A large positive value of x^2 means a large discrepancy ('positive' due to squares)
- Is $x^2 = 2.8$ a large value?

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 χ^2-distribution

- Definition: Let $Z_1, \ldots, Z_d$ be i.i.d. RVs, with $Z_1 \sim N(0,1)$.
  A random variable X is called χ^2-*distributed with d degrees of freedom*, short $X \sim \chi^2(d)$, if

$$\boxed{X \sim Z_1^2 + \cdots + Z_d^2}$$

The χ^2-distribution

- Definition: Let $Z_1, \ldots, Z_d$ be i.i.d. RVs, with $Z_1 \sim N(0,1)$.
  A random variable X is called χ^2-*distributed with d degrees of freedom*, short $X \sim \chi^2(d)$, if

$$\boxed{X \sim Z_1^2 + \cdots + Z_d^2}$$

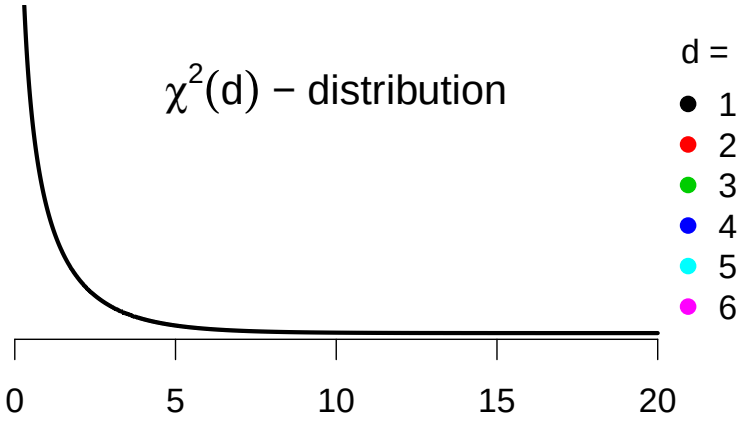
- In words: a $\chi^2(d)$-distributed random variable is distributed like the sum of d squares of independent $N(0,1)$-distributed random variables

The χ^2-distribution

- Definition: Let $Z_1, \ldots, Z_d$ be i.i.d. RVs, with $Z_1 \sim N(0,1)$.
  A random variable X is called χ^2-*distributed with d degrees of freedom*, short $X \sim \chi^2(d)$, if

$$\boxed{X \sim Z_1^2 + \cdots + Z_d^2}$$

- In words: a $\chi^2(d)$-distributed random variable is distributed like the sum of d squares of independent $N(0,1)$-distributed random variables

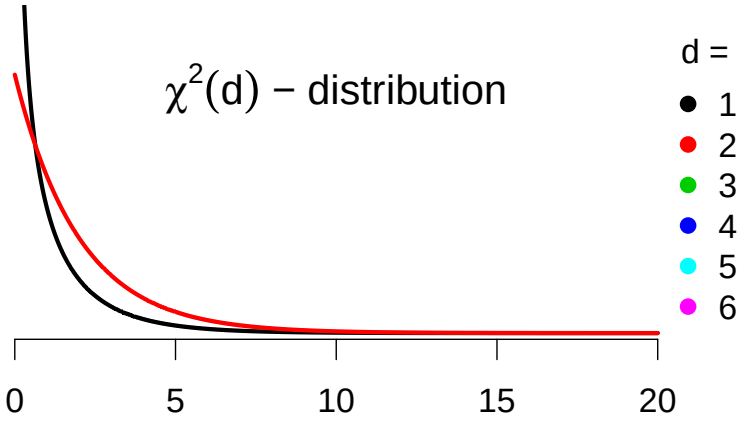


The χ^2-distribution

- Definition: Let $Z_1, \ldots, Z_d$ be i.i.d. RVs, with $Z_1 \sim N(0,1)$.
  A random variable X is called χ^2-*distributed with d degrees of freedom*, short $X \sim \chi^2(d)$, if

$$\boxed{X \sim Z_1^2 + \cdots + Z_d^2}$$

- In words: a $\chi^2(d)$-distributed random variable is distributed like the sum of d squares of independent $N(0,1)$-distributed random variables

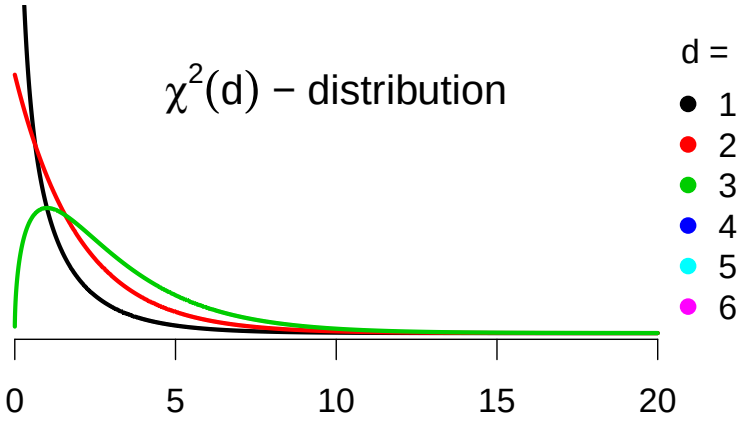


The χ^2-distribution

- Definition: Let $Z_1, \ldots, Z_d$ be i.i.d. RVs, with $Z_1 \sim N(0,1)$.
  A random variable X is called χ^2-*distributed with d degrees of freedom*, short $X \sim \chi^2(d)$, if

$$\boxed{X \sim Z_1^2 + \cdots + Z_d^2}$$

- In words: a $\chi^2(d)$-distributed random variable is distributed like the sum of d squares of independent $N(0,1)$-distributed random variables

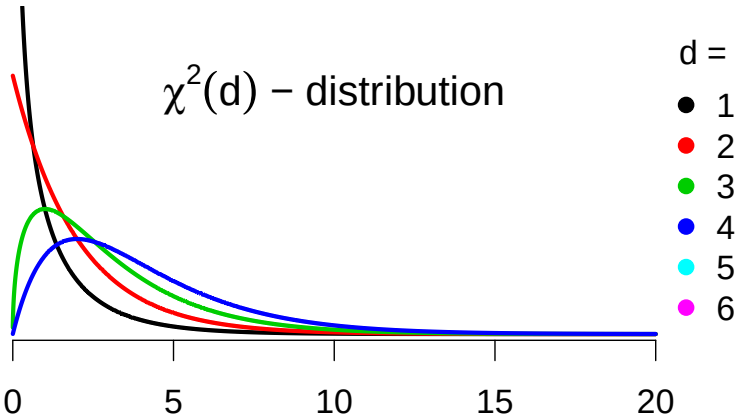


The χ^2-distribution

- Definition: Let $Z_1, \ldots, Z_d$ be i.i.d. RVs, with $Z_1 \sim N(0,1)$.
  A random variable X is called χ^2-*distributed with d degrees of freedom*, short $X \sim \chi^2(d)$, if

$$\boxed{X \sim Z_1^2 + \cdots + Z_d^2}$$

- In words: a $\chi^2(d)$-distributed random variable is distributed like the sum of d squares of independent $N(0,1)$-distributed random variables

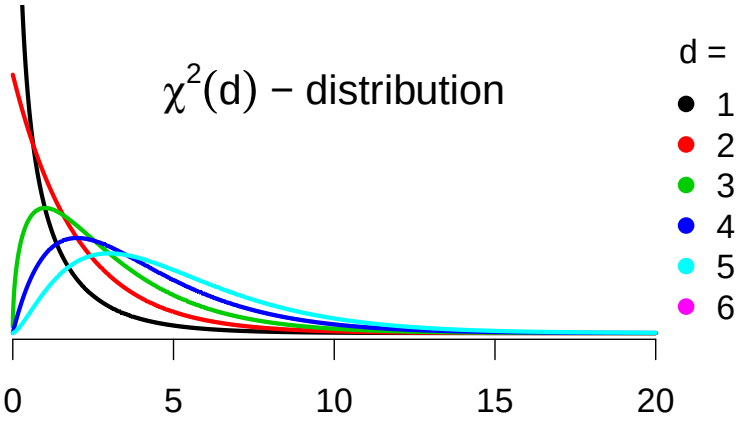


The χ^2-distribution

- Definition: Let $Z_1, \ldots, Z_d$ be i.i.d. RVs, with $Z_1 \sim N(0,1)$.
  A random variable X is called χ^2-*distributed with d degrees of freedom*, short $X \sim \chi^2(d)$, if

$$\boxed{X \sim Z_1^2 + \cdots + Z_d^2}$$

- In words: a $\chi^2(d)$-distributed random variable is distributed like the sum of d squares of independent $N(0,1)$-distributed random variables

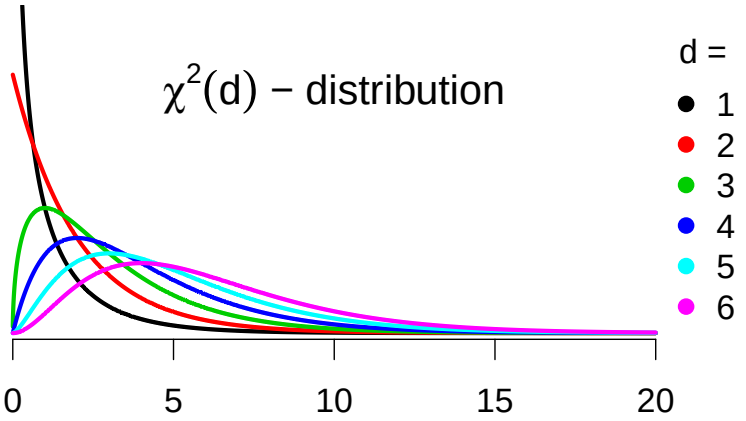


The χ^2-distribution

- Definition: Let $Z_1, \ldots, Z_d$ be i.i.d. RVs, with $Z_1 \sim N(0,1)$.
  A random variable X is called χ^2-*distributed with d degrees of freedom*, short $X \sim \chi^2(d)$, if

$$\boxed{X \sim Z_1^2 + \cdots + Z_d^2}$$

- In words: a $\chi^2(d)$-distributed random variable is distributed like the sum of d squares of independent $N(0,1)$-distributed random variables

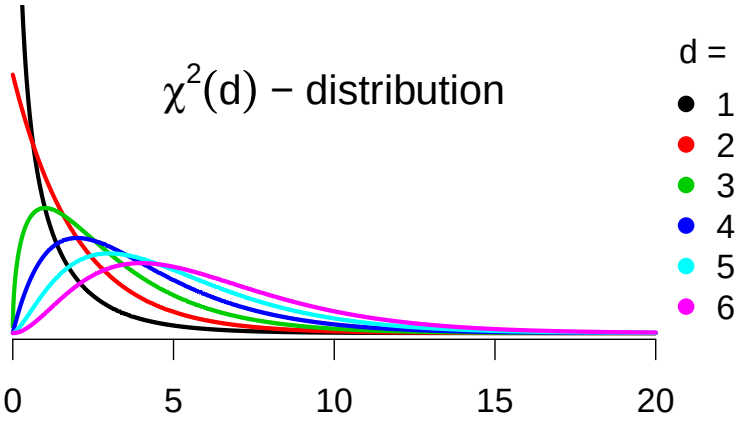


The χ^2-distribution

- Definition: Let $Z_1, \ldots, Z_d$ be i.i.d. RVs, with $Z_1 \sim N(0,1)$.
  A random variable X is called χ^2-*distributed with d degrees of freedom*, short $X \sim \chi^2(d)$, if

$$\boxed{X \sim Z_1^2 + \cdots + Z_d^2}$$

- In words: a $\chi^2(d)$-distributed random variable is distributed like the sum of d squares of independent $N(0,1)$-distributed random variables
- properties: for $X \sim \chi^2(d)$ it holds
  - $X \geqslant 0$

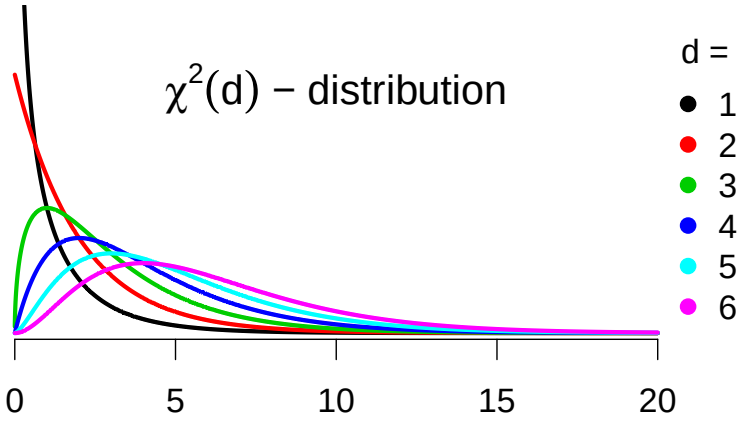


The χ^2-distribution

- Definition: Let $Z_1, \ldots, Z_d$ be i.i.d. RVs, with $Z_1 \sim N(0,1)$.
  A random variable X is called χ^2-*distributed with d degrees of freedom*, short $X \sim \chi^2(d)$, if

$$\boxed{X \sim Z_1^2 + \cdots + Z_d^2}$$

- In words: a $\chi^2(d)$-distributed random variable is distributed like the sum of d squares of independent $N(0,1)$-distributed random variables
- properties: for $X \sim \chi^2(d)$ it holds
  - $X \geqslant 0$
  - $\mathbb{E}[X] = d$ 'linearity of the expectation, and $\mathbb{E}(Z_1^2) = \mathbb{V}ar(Z_1) = 1$'

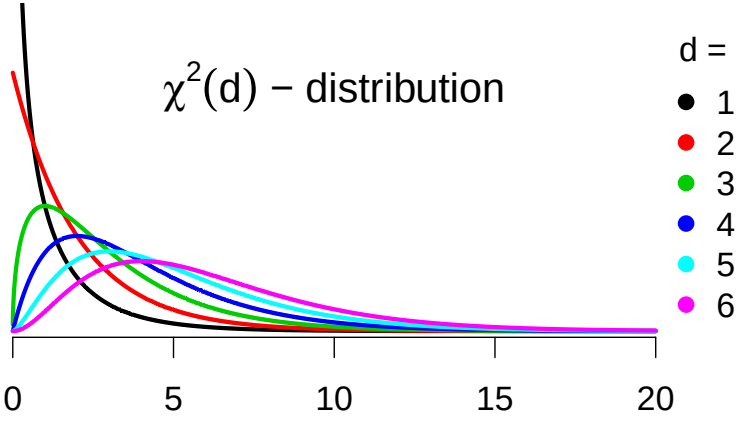


The χ^2-distribution

- Definition: Let $Z_1, \ldots, Z_d$ be i.i.d. RVs, with $Z_1 \sim N(0,1)$.
  A random variable X is called χ^2-*distributed with d degrees of freedom*, short $X \sim \chi^2(d)$, if

$$\boxed{X \sim Z_1^2 + \cdots + Z_d^2}$$

- In words: a $\chi^2(d)$-distributed random variable is distributed like the sum of d squares of independent $N(0,1)$-distributed random variables
- properties: for $X \sim \chi^2(d)$ it holds
  - $X \geqslant 0$
  - $\mathbb{E}[X] = d$ 'linearity of the expectation, and $\mathbb{E}(Z_1^2) = \mathbb{V}ar(Z_1) = 1$'
  - $\mathbb{V}ar(X) = 2d$ 'independence, and $\mathbb{V}ar(Z_1^2) = \mathbb{E}(Z_1^4) - \mathbb{E}(Z_1^2)^2 = 3 - 1 = 2$

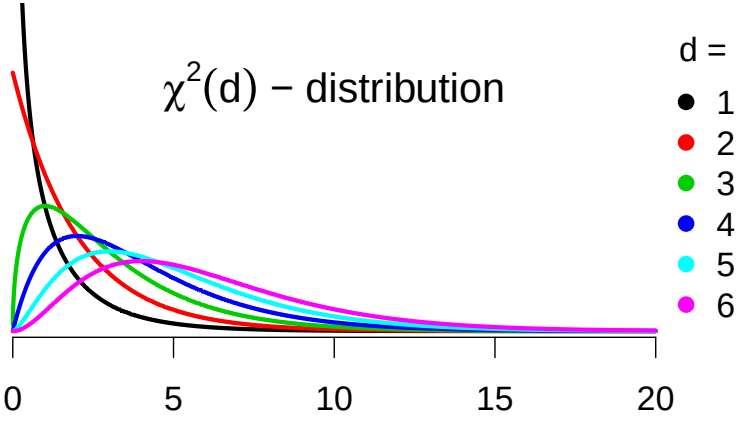


The χ^2-distribution

- Definition: Let $Z_1, \ldots, Z_d$ be i.i.d. RVs, with $Z_1 \sim N(0,1)$.
  A random variable X is called χ^2-*distributed with d degrees of freedom*, short $X \sim \chi^2(d)$, if

$$\boxed{X \sim Z_1^2 + \cdots + Z_d^2}$$

- In words: a $\chi^2(d)$-distributed random variable is distributed like the sum of d squares of independent $N(0,1)$-distributed random variables
- properties: for $X \sim \chi^2(d)$ it holds
  - $X \geqslant 0$
  - $\mathbb{E}[X] = d$ 'linearity of the expectation, and $\mathbb{E}(Z_1^2) = \mathbb{V}ar(Z_1) = 1$'
  - $\mathbb{V}ar(X) = 2d$ 'independence, and $\mathbb{V}ar(Z_1^2) = \mathbb{E}(Z_1^4) - \mathbb{E}(Z_1^2)^2 = 3 - 1 = 2$
  - R knows it well: `rchisq()`, `pchisq()` etc.



The χ^2-test (goodness of fit)

- Let $\mathfrak{x} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^d p_k = 1$

Under $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$ it holds (approximately)

$$X^2 := \sum_{k=1}^{d} \frac{(X_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} \overset{d}{\approx} \chi^2(d-1)$$

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

The χ^2-test (goodness of fit)

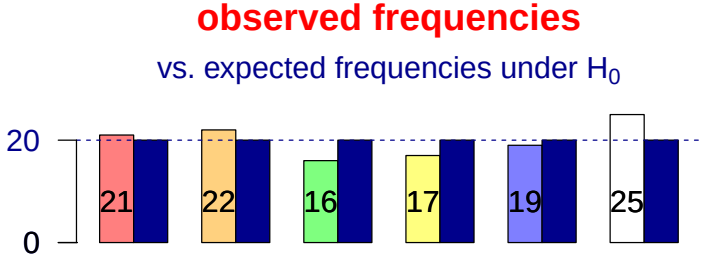
- Let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$

Under $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$ it holds (approximately)

$$X^2 := \sum_{k=1}^{d} \frac{(X_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} \overset{d}{\approx} \chi^2(d-1)$$

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

- here: $n = 120$ and $d = 6$, as well as $p_0 = (1/6, \ldots, 1/6)^t$



The χ^2-test (goodness of fit)

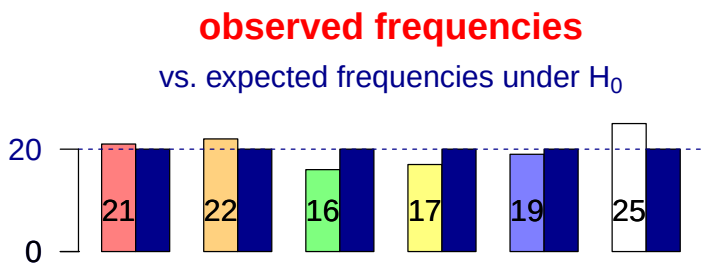
- Let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^d p_k = 1$

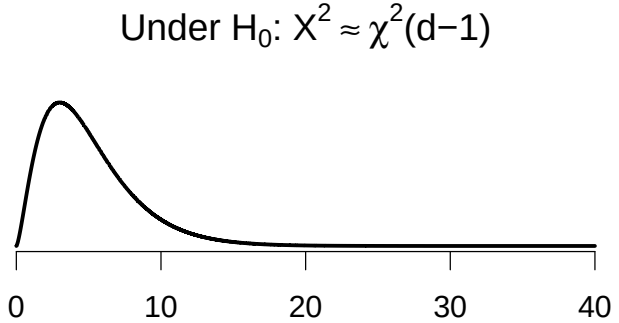
Under $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$ it holds (approximately)

$$X^2 := \sum_{k=1}^{d} \frac{(X_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} \overset{d}{\approx} \chi^2(d-1)$$

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

- here: $n = 120$ and $d = 6$, as well as $p_0 = (1/6, \ldots, 1/6)^t$
- for $\alpha = 5\%$ the $(1-\alpha)$-quantile of the $\chi^2(5)$-distribution is $q_{1-\alpha} \approx 11.1$





The χ^2-test (goodness of fit)

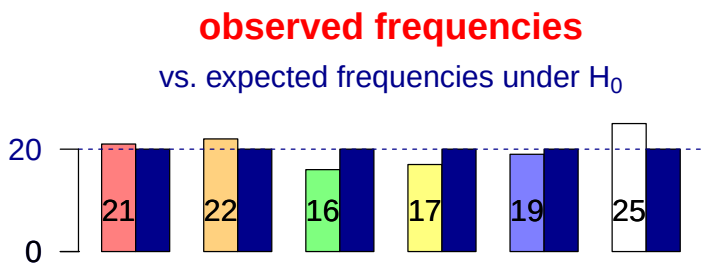
- Let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$

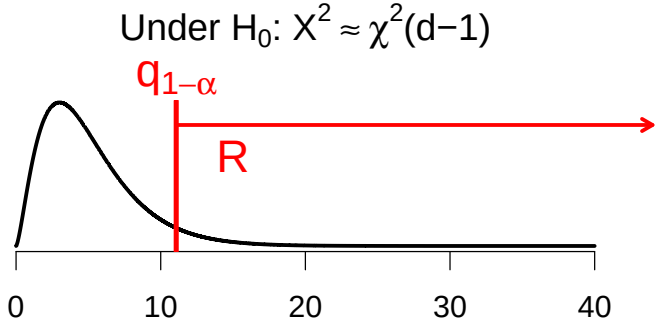
Under $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$ it holds (approximately)

$$X^2 := \sum_{k=1}^{d} \frac{(X_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} \overset{d}{\approx} \chi^2(d-1)$$

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

- here: $n = 120$ and $d = 6$, as well as $p_0 = (1/6, \ldots, 1/6)^t$
- for $\alpha = 5\%$ the $(1-\alpha)$-quantile of the $\chi^2(5)$-distribution is $q_{1-\alpha} \approx 11.1$
- rejection area: $R = [q_{1-\alpha}, \infty)$ (one-sided, x^2 large speaks against H_0)





The χ^2-test (goodness of fit)

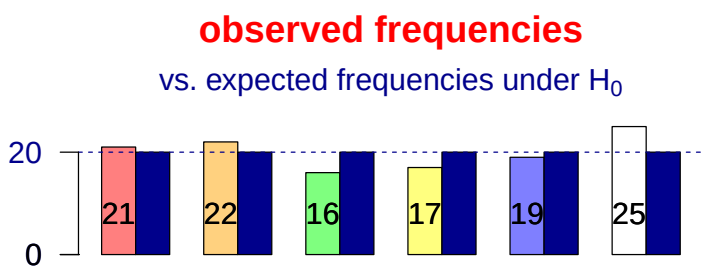
- Let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^d p_k = 1$

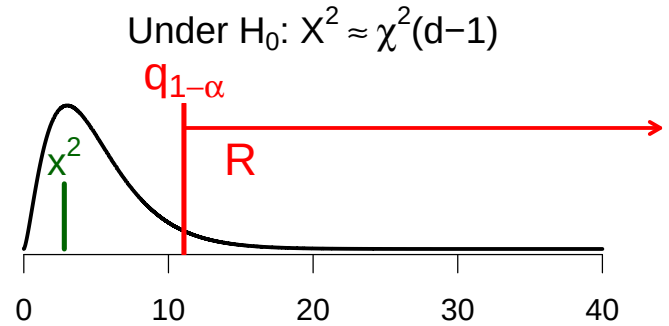
Under $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$ it holds (approximately)

$$X^2 := \sum_{k=1}^{d} \frac{(X_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} \overset{d}{\approx} \chi^2(d-1)$$

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

- here: $n = 120$ and $d = 6$, as well as $p_0 = (1/6, \ldots, 1/6)^t$
- for $\alpha = 5\%$ the $(1-\alpha)$-quantile of the $\chi^2(5)$-distribution is $q_{1-\alpha} \approx 11.1$
- rejection area: $R = [q_{1-\alpha}, \infty)$ (one-sided, x^2 large speaks against H_0)
- data: $x^2 = 2.8 \notin R \to$ can not reject H_0





The χ^2-test (goodness of fit)

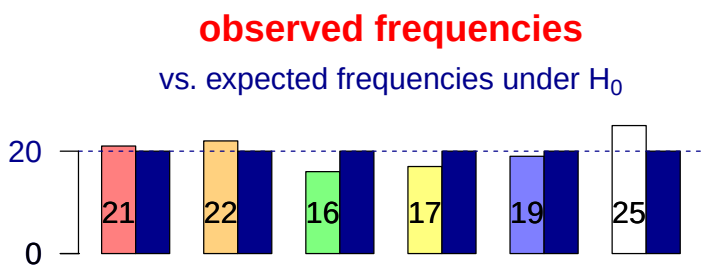
- Let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$

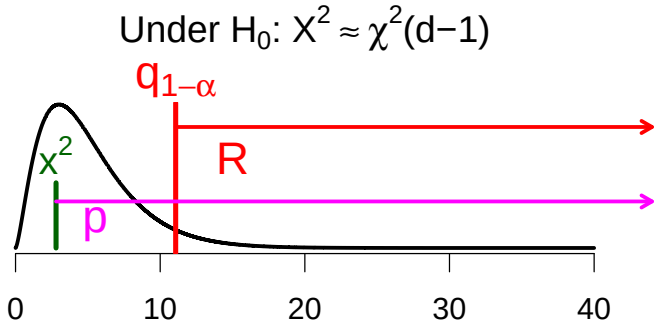
Under $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$ it holds (approximately)

$$X^2 := \sum_{k=1}^{d} \frac{(X_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} \overset{d}{\approx} \chi^2(d-1)$$

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

- here: $n = 120$ and $d = 6$, as well as $p_0 = (1/6, \ldots, 1/6)^t$
- for $\alpha = 5\%$ the $(1-\alpha)$-quantile of the $\chi^2(5)$-distribution is $q_{1-\alpha} \approx 11.1$
- rejection area: $R = [q_{1-\alpha}, \infty)$ (one-sided, x^2 large speaks against H_0)
- data: $x^2 = 2.8 \notin R \to$ can <u>not</u> reject H_0
- $p \approx 0.73$.





The χ^2-test (goodness of fit)

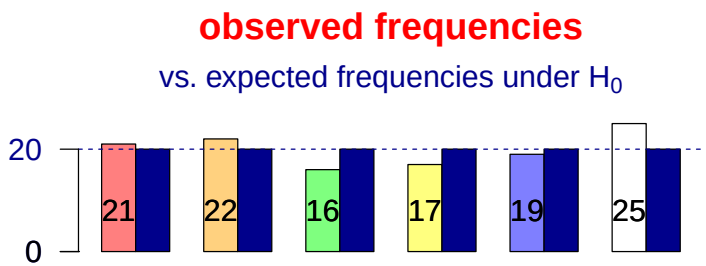
- Let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$

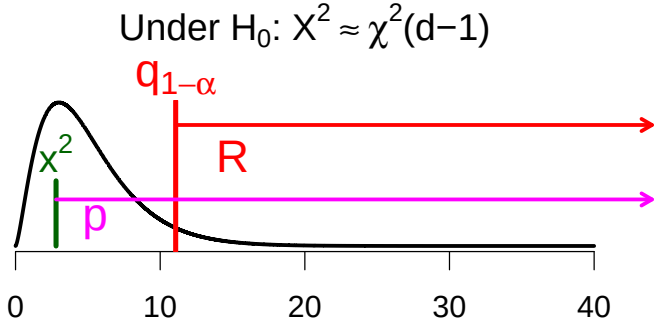
Under $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$ it holds (approximately)

$$X^2 := \sum_{k=1}^{d} \frac{(X_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} \overset{d}{\approx} \chi^2(d-1)$$

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

- here: $n = 120$ and $d = 6$, as well as $p_0 = (1/6, \ldots, 1/6)^t$
- for $\alpha = 5\%$ the $(1-\alpha)$-quantile of the $\chi^2(5)$-distribution is $q_{1-\alpha} \approx 11.1$
- rejection area: $R = [q_{1-\alpha}, \infty)$ (one-sided, x^2 large speaks against H_0)
- data: $x^2 = 2.8 \notin R \to$ can not reject H_0
- $p \approx 0.73$. If the null hypothesis holds true, then we observe in about 7 of 10 cases a discrepancy, which is at least as extreme as in the data.





The χ^2-test (goodness of fit)

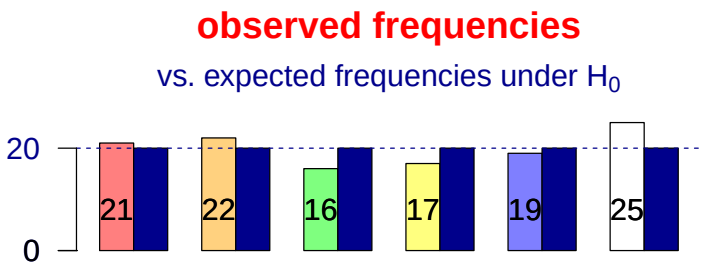
- Let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^d p_k = 1$

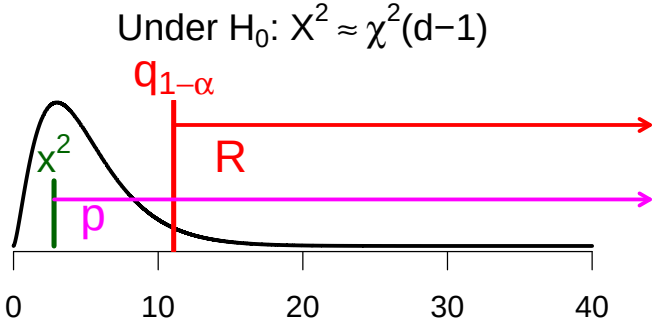
Under $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$ it holds (approximately)

$$X^2 := \sum_{k=1}^{d} \frac{(X_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} \overset{d}{\approx} \chi^2(d-1)$$

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

- here: $n = 120$ and $d = 6$, as well as $p_0 = (1/6, \ldots, 1/6)^t$
- for $\alpha = 5\%$ the $(1-\alpha)$-quantile of the $\chi^2(5)$-distribution is $q_{1-\alpha} \approx 11.1$
- rejection area: $R = [q_{1-\alpha}, \infty)$ (one-sided, x^2 large speaks against H_0)
- data: $x^2 = 2.8 \notin R \to$ can not reject H_0
- $p \approx 0.73$. If the null hypothesis holds true, then we observe in about 7 of 10 cases a discrepancy, which is at least as extreme as in the data. The observed discrepancy is not at all unlikely





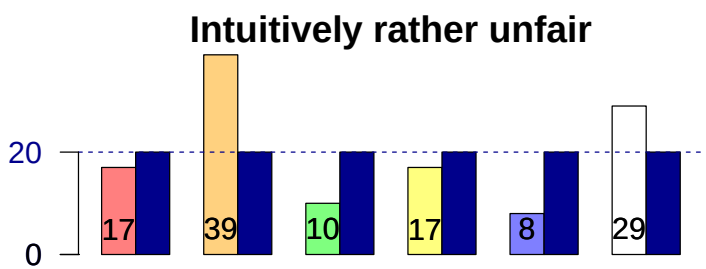
The χ^2-test, goodness of fit (example 2)

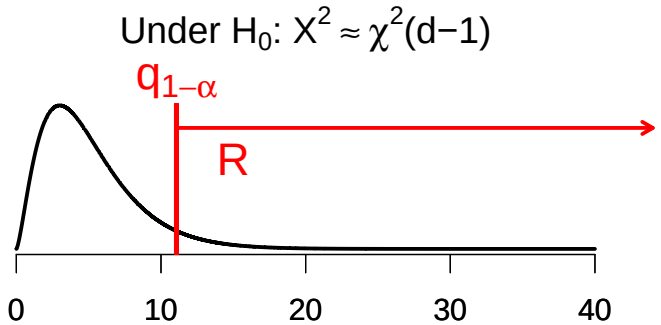
- Let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$

Under $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$ it holds (approximately)

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

- Here: $n = 120$ and $d = 6$, as well as $p_0 = (1/6, \ldots, 1/6)^t$
- For $\alpha = 5\%$ the $(1-\alpha)$-quantile of the $\chi^2(5)$-distribution is $q_{1-\alpha} \approx 11.1$
- Rejection area: $R = [q_{1-\alpha}, \infty)$





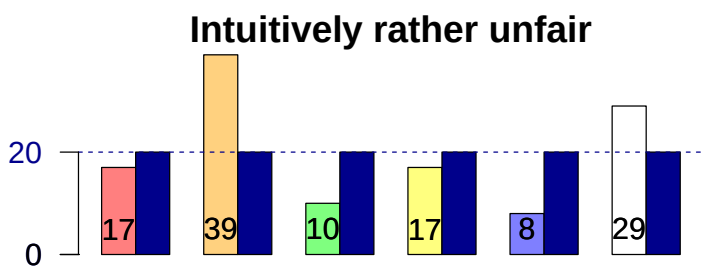
The χ^2-test, goodness of fit (example 2)

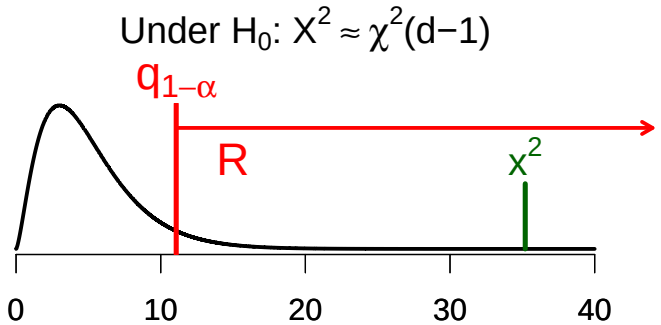
- Let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^d p_k = 1$

Under $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$ it holds (approximately)

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

- Here: $n = 120$ and $d = 6$, as well as $p_0 = (1/6, \ldots, 1/6)^t$
- For $\alpha = 5\%$ the $(1-\alpha)$-quantile of the $\chi^2(5)$-distribution is $q_{1-\alpha} \approx 11.1$
- Rejection area: $R = [q_{1-\alpha}, \infty)$
- data: $x^2 = 35.2 \in R$, $\rightarrow$ we can reject H_0





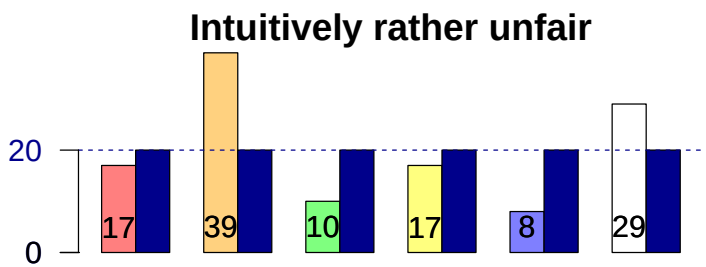
The χ^2-test, goodness of fit (example 2)

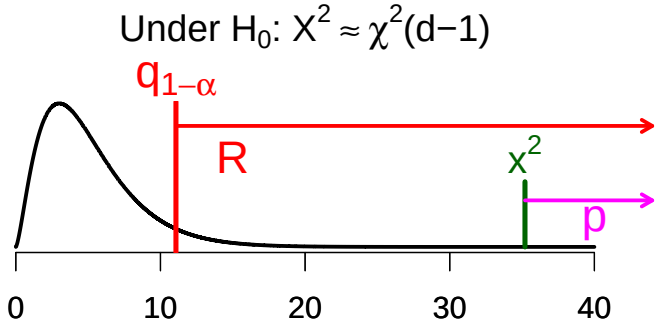
- Let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^d p_k = 1$

Under $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$ it holds (approximately)

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

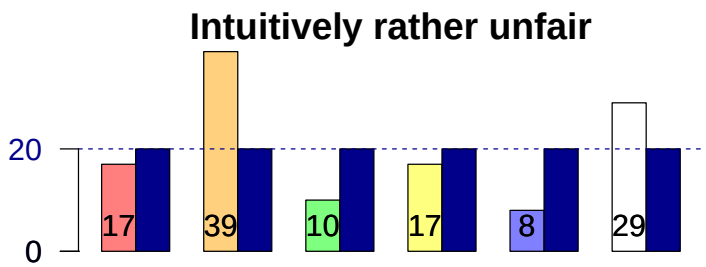
- Here: $n = 120$ and $d = 6$, as well as $p_0 = (1/6, \ldots, 1/6)^t$
- For $\alpha = 5\%$ the $(1-\alpha)$-quantile of the $\chi^2(5)$-distribution is $q_{1-\alpha} \approx 11.1$
- Rejection area: $R = [q_{1-\alpha}, \infty)$
- data: $x^2 = 35.2 \in R$, $\rightarrow$ we can reject H_0
- $p < 10^{-5}$. If H_0 holds true, then we observe in less than 1 of 100000 cases a discrepancy which is at least as extreme as in the data. The data are not at all compatible with the null hypothesis





Loaded die (example 3)

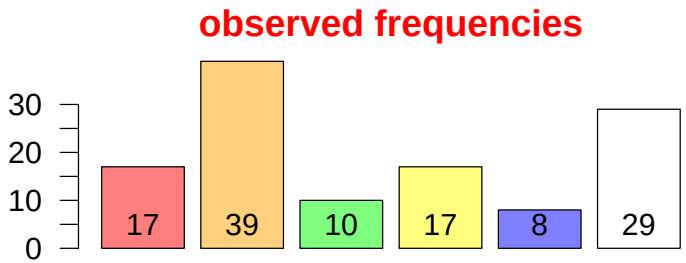
- Somebody claims: 'I loaded the die'



Loaded die (example 3)

- Somebody claims: 'I loaded the die'

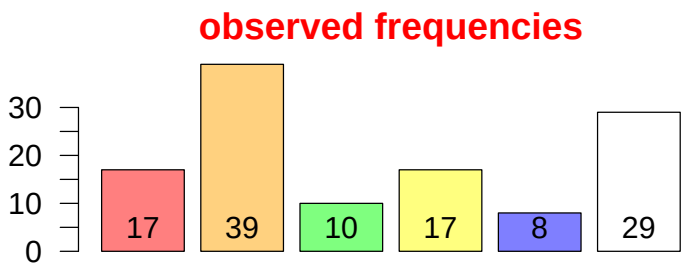
k	1	2	3	4	5	6	$\sum$
x_k	17	39	10	17	8	29	120



Loaded die (example 3)

- Somebody claims: 'I loaded the die' , in a way that
  - 1.: the three sides rot, yellow and white appear with the same frequency

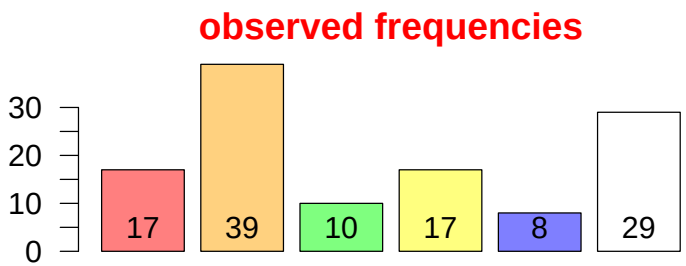
k	1	2	3	4	5	6	$\sum$
x_k	17	39	10	17	8	29	120



Loaded die (example 3)

- Somebody claims: 'I loaded the die' , in a way that
  - 1.: the three sides rot, yellow and white appear with the same frequency
  - 2.: orange appears twice as often as these three

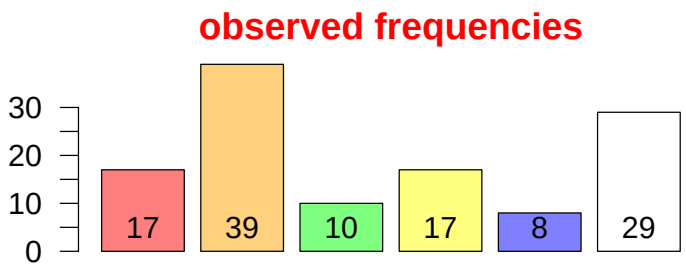
k	1	2	3	4	5	6	$\sum$
x_k	17	39	10	17	8	29	120



Loaded die (example 3)

- Somebody claims: 'I loaded the die' , in a way that
  - 1.: the three sides rot, yellow and white appear with the same frequency
  - 2.: orange appears twice as often as these three
  - 3.: the sides green and blue each half as often as the upper three

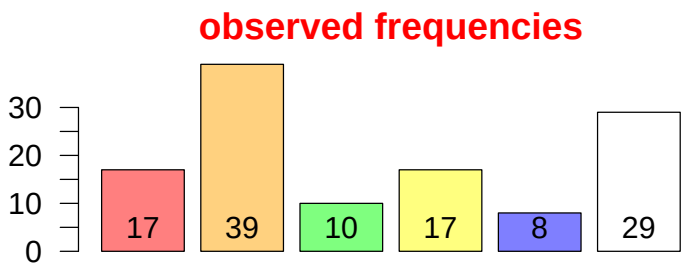
k	1	2	3	4	5	6	$\sum$
x_k	17	39	10	17	8	29	120



Loaded die (example 3)

- Somebody claims: 'I loaded the die' , in a way that
  - 1.: the three sides rot, yellow and white appear with the same frequency
  - 2.: orange appears twice as often as these three
  - 3.: the sides green and blue each half as often as the upper three
- Model: $\mathfrak{x} = (X_1, \ldots, X_6)^t \sim mult(120, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$

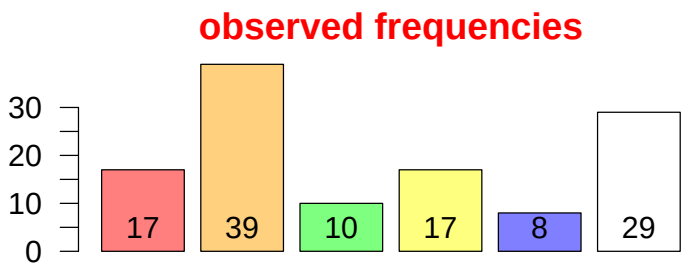
k	1	2	3	4	5	6	$\sum$
x_k	17	39	10	17	8	29	120



Loaded die (example 3)

- Somebody claims: 'I loaded the die' , in a way that
  - 1.: the three sides rot, yellow and white appear with the same frequency
  - 2.: orange appears twice as often as these three
  - 3.: the sides green and blue each half as often as the upper three
- Model: $\mathfrak{x} = (X_1, \ldots, X_6)^t \sim mult(120, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- $H_0 : p = (1/6,\ 1/3,\ 1/12,\ 1/6,\ 1/12,\ 1/6)^t$ ('loaded die')

k	1	2	3	4	5	6	$\sum$
x_k	17	39	10	17	8	29	120

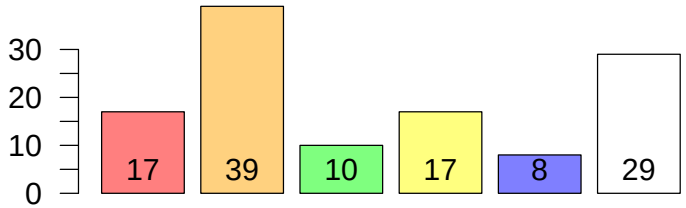


Loaded die (example 3)

- Somebody claims: 'I loaded the die' , in a way that
  - 1.: the three sides rot, yellow and white appear with the same frequency
  - 2.: orange appears twice as often as these three
  - 3.: the sides green and blue each half as often as the upper three
- Model: $\mathfrak{x} = (X_1, \ldots, X_6)^t \sim mult(120, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- $H_0 : p = (1/6,\ 1/3,\ 1/12,\ 1/6,\ 1/12,\ 1/6)^t$ ('loaded die')

k	1	2	3	4	5	6	$\sum$
x_k	17	39	10	17	8	29	120
$\mathbb{E}_{H_0}[X_k]$	20	40	10	20	10	20	120

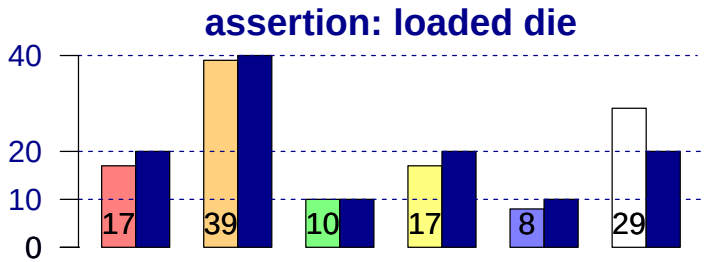
observed frequencies



Loaded die (example 3)

- Somebody claims: 'I loaded the die' , in a way that
  - 1.: the three sides rot, yellow and white appear with the same frequency
  - 2.: orange appears twice as often as these three
  - 3.: the sides green and blue each half as often as the upper three
- Model: $\mathfrak{x} = (X_1, \ldots, X_6)^t \sim mult(120, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- $H_0 : p = (1/6,\ 1/3,\ 1/12,\ 1/6,\ 1/12,\ 1/6)^t$ ('loaded die')

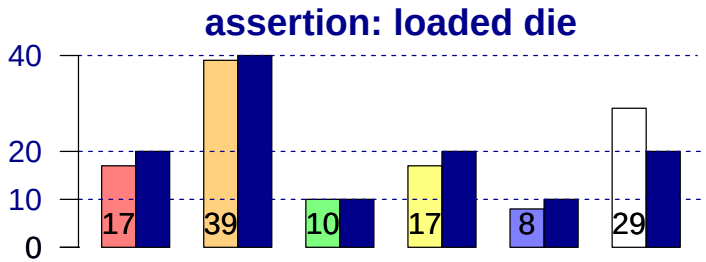
k	1	2	3	4	5	6	$\sum$
x_k	17	39	10	17	8	29	120
$\mathbb{E}_{H_0}[X_k]$	20	40	10	20	10	20	120



Loaded die (example 3)

- Somebody claims: 'I loaded the die' , in a way that
  - 1.: the three sides rot, yellow and white appear with the same frequency
  - 2.: orange appears twice as often as these three
  - 3.: the sides green and blue each half as often as the upper three
- Model: $\mathfrak{x} = (X_1, \ldots, X_6)^t \sim mult(120, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- $H_0 : p = (1/6,\ 1/3,\ 1/12,\ 1/6,\ 1/12,\ 1/6)^t$ ('loaded die')

k	1	2	3	4	5	6	$\sum$
x_k	17	39	10	17	8	29	120
$\mathbb{E}_{H_0}[X_k]$	20	40	10	20	10	20	120

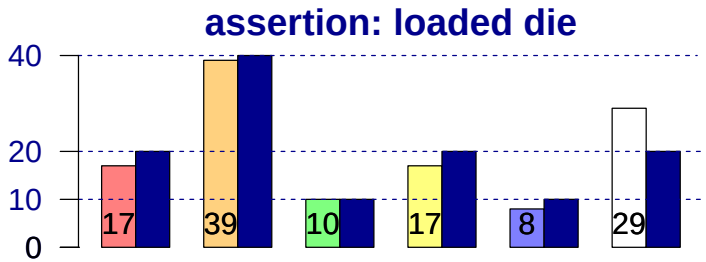


$$x^2 := \sum_{k=1}^{d} \frac{(x_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]}$$

Loaded die (example 3)

- Somebody claims: 'I loaded the die' , in a way that
  - 1.: the three sides rot, yellow and white appear with the same frequency
  - 2.: orange appears twice as often as these three
  - 3.: the sides green and blue each half as often as the upper three
- Model: $\mathfrak{x} = (X_1, \ldots, X_6)^t \sim mult(120, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- $H_0 : p = (1/6,\ 1/3,\ 1/12,\ 1/6,\ 1/12,\ 1/6)^t$ ('loaded die')

k	1	2	3	4	5	6	$\sum$
x_k	17	39	10	17	8	29	120
$\mathbb{E}_{H_0}[X_k]$	20	40	10	20	10	20	120

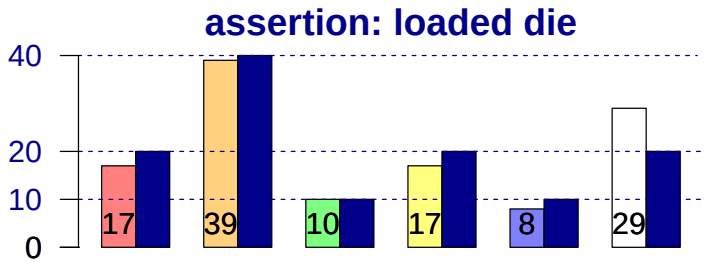


$$x^2 := \sum_{k=1}^{d} \frac{(x_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} = \frac{(17-20)^2}{20} + \frac{(39-40)^2}{40} + \cdots$$

Loaded die (example 3)

- Somebody claims: 'I loaded the die' , in a way that
  - 1.: the three sides rot, yellow and white appear with the same frequency
  - 2.: orange appears twice as often as these three
  - 3.: the sides green and blue each half as often as the upper three
- Model: $\mathfrak{x} = (X_1, \ldots, X_6)^t \sim mult(120, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- $H_0 : p = (1/6,\ 1/3,\ 1/12,\ 1/6,\ 1/12,\ 1/6)^t$ ('loaded die')

k	1	2	3	4	5	6	$\sum$
x_k	17	39	10	17	8	29	120
$\mathbb{E}_{H_0}[X_k]$	20	40	10	20	10	20	120

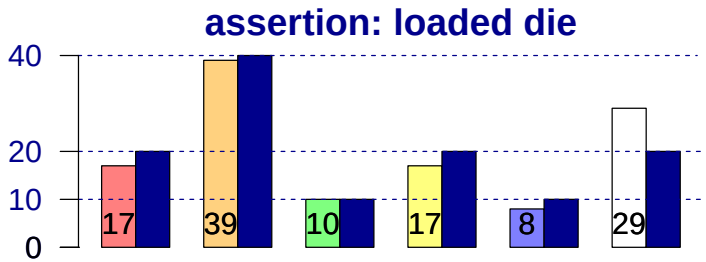


$$x^2 := \sum_{k=1}^{d} \frac{(x_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} = \frac{(17-20)^2}{20} + \frac{(39-40)^2}{40} + \cdots = \frac{9}{20} + \frac{1}{40} + \cdots$$

Loaded die (example 3)

- Somebody claims: 'I loaded the die' , in a way that
  - 1.: the three sides rot, yellow and white appear with the same frequency
  - 2.: orange appears twice as often as these three
  - 3.: the sides green and blue each half as often as the upper three
- Model: $\mathfrak{x} = (X_1, \ldots, X_6)^t \sim mult(120, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- $H_0 : p = (1/6,\ 1/3,\ 1/12,\ 1/6,\ 1/12,\ 1/6)^t$ ('loaded die')

k	1	2	3	4	5	6	$\sum$
x_k	17	39	10	17	8	29	120
$\mathbb{E}_{H_0}[X_k]$	20	40	10	20	10	20	120

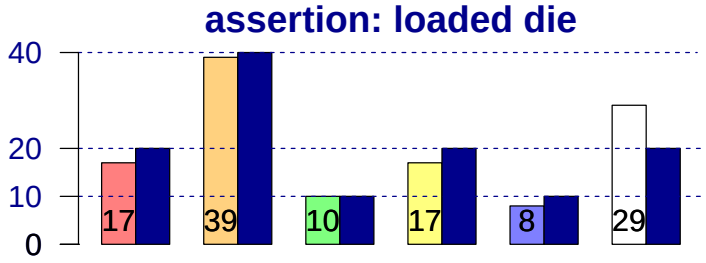


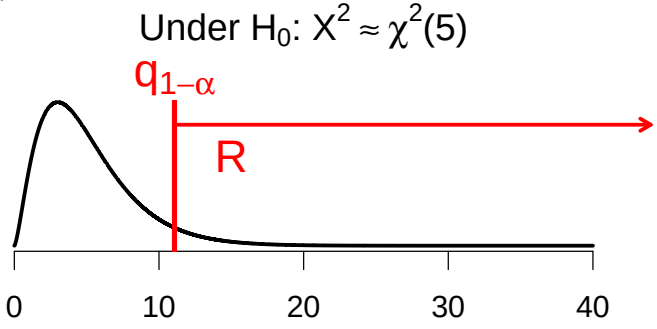
$$x^2 := \sum_{k=1}^{d} \frac{(x_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} = \frac{(17-20)^2}{20} + \frac{(39-40)^2}{40} + \cdots = \frac{9}{20} + \frac{1}{40} + \cdots = 5.375$$

Loaded die (example 3)

- Somebody claims: 'I loaded the die' , in a way that
  - 1.: the three sides rot, yellow and white appear with the same frequency
  - 2.: orange appears twice as often as these three
  - 3.: the sides green and blue each half as often as the upper three
- Model: $\mathfrak{x} = (X_1, \ldots, X_6)^t \sim mult(120, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- $H_0 : p = (1/6,\ 1/3,\ 1/12,\ 1/6,\ 1/12,\ 1/6)^t$ ('loaded die')

k	1	2	3	4	5	6	$\sum$
x_k	17	39	10	17	8	29	120
$\mathbb{E}_{H_0}[X_k]$	20	40	10	20	10	20	120





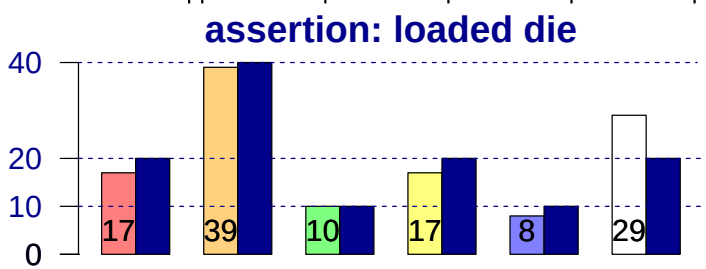
$$x^2 := \sum_{k=1}^{d} \frac{(x_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} = \frac{(17-20)^2}{20} + \frac{(39-40)^2}{40} + \cdots = \frac{9}{20} + \frac{1}{40} + \cdots = 5.375$$

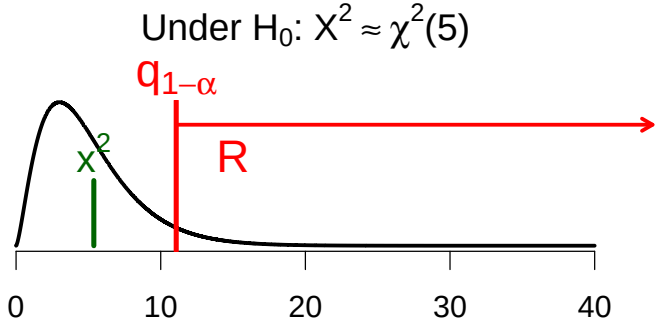
- for $\alpha = 5\%$ we obtain the rejection area $R \approx [11.1, \infty)$

Loaded die (example 3)

- Somebody claims: 'I loaded the die' , in a way that
  - 1.: the three sides rot, yellow and white appear with the same frequency
  - 2.: orange appears twice as often as these three
  - 3.: the sides green and blue each half as often as the upper three
- Model: $\mathfrak{x} = (X_1, \ldots, X_6)^t \sim mult(120, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- $H_0 : p = (1/6,\ 1/3,\ 1/12,\ 1/6,\ 1/12,\ 1/6)^t$ ('loaded die')

k	1	2	3	4	5	6	$\sum$
x_k	17	39	10	17	8	29	120
$\mathbb{E}_{H_0}[X_k]$	20	40	10	20	10	20	120





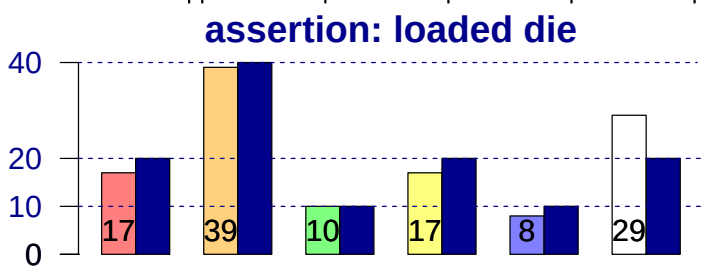
$$x^2 := \sum_{k=1}^{d} \frac{(x_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} = \frac{(17-20)^2}{20} + \frac{(39-40)^2}{40} + \cdots = \frac{9}{20} + \frac{1}{40} + \cdots = 5.375$$

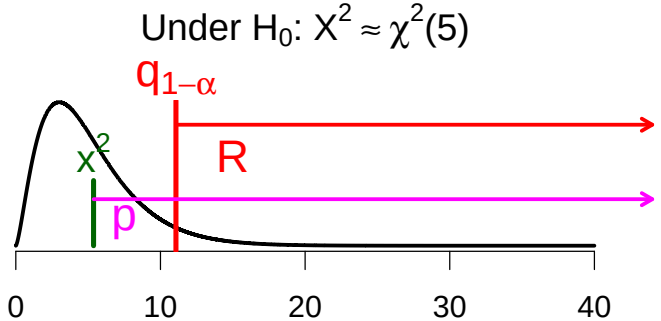
- for $\alpha = 5\%$ we obtain the rejection area $R \approx [11.1, \infty)$
- data: $x^2 = 5.375 \notin R$, $\rightarrow$ can not reject H_0

Loaded die (example 3)

- Somebody claims: 'I loaded the die' , in a way that
  - 1.: the three sides rot, yellow and white appear with the same frequency
  - 2.: orange appears twice as often as these three
  - 3.: the sides green and blue each half as often as the upper three
- Model: $\mathfrak{x} = (X_1, \ldots, X_6)^t \sim mult(120, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^{d} p_k = 1$
- $H_0 : p = (1/6,\ 1/3,\ 1/12,\ 1/6,\ 1/12,\ 1/6)^t$ ('loaded die')

k	1	2	3	4	5	6	$\sum$
x_k	17	39	10	17	8	29	120
$\mathbb{E}_{H_0}[X_k]$	20	40	10	20	10	20	120





$$x^2 := \sum_{k=1}^{d} \frac{(x_k - \mathbb{E}_{H_0}[X_k])^2}{\mathbb{E}_{H_0}[X_k]} = \frac{(17-20)^2}{20} + \frac{(39-40)^2}{40} + \cdots = \frac{9}{20} + \frac{1}{40} + \cdots = 5.375$$

- for $\alpha = 5\%$ we obtain the rejection area $R \approx [11.1, \infty)$
- data: $x^2 = 5.375 \notin R$, $\rightarrow$ can not reject H_0 ($p \approx 0.37$)

Remarks

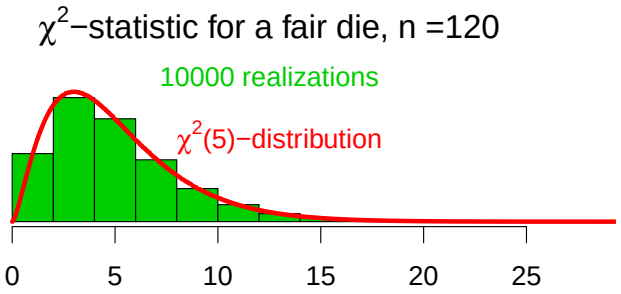
- Initial question:
  How good do the observed frequencies fit to the frequencies expected under the null hypothesis?
  $\rightarrow$ the χ^2-test is also known as goodness of fit test

Remarks

- Initial question:
  How good do the observed frequencies fit to the frequencies expected under the null hypothesis?
  $\rightarrow$ the χ^2-test is also known as goodness of fit test
- the χ^2-statistic is asymptotically $\chi^2(d-1)$-distributed ($n \rightarrow \infty$)

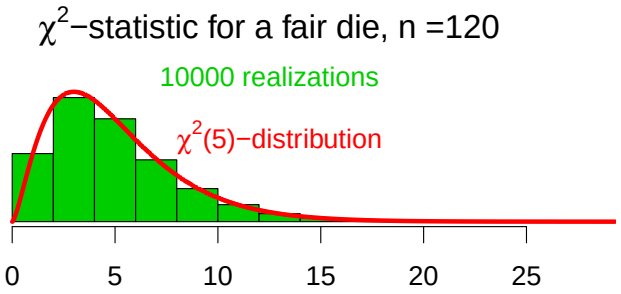
Remarks

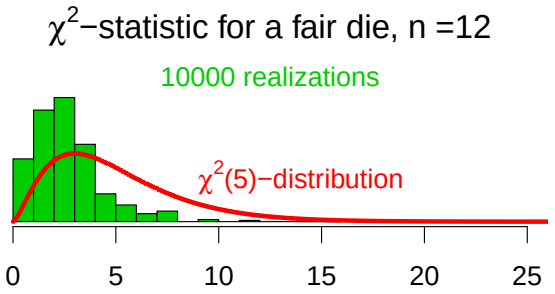
- Initial question:
  How good do the observed frequencies fit to the frequencies expected under the null hypothesis?
  $\rightarrow$ the χ^2-test is also known as goodness of fit test
- the χ^2-statistic is asymptotically $\chi^2(d-1)$-distributed $(n \rightarrow \infty)$
  1. The approximation gets better the more data are found in the categories



Remarks

- Initial question:
  How good do the observed frequencies fit to the frequencies expected under the null hypothesis?
  $\rightarrow$ the χ^2-test is also known as goodness of fit test
- the χ^2-statistic is asymptotically $\chi^2(d-1)$-distributed $(n \rightarrow \infty)$
  1. The approximation gets better the more data are found in the categories





Remarks

- Initial question:
  How good do the observed frequencies fit to the frequencies expected under the null hypothesis?
  $\rightarrow$ the χ^2-test is also known as goodness of fit test
- the χ^2-statistic is asymptotically $\chi^2(d-1)$-distributed ($n \rightarrow \infty$)
  1. The approximation gets better the more data are found in the categories
  2. Why are the degrees of freedom $d-1$ (and not d)?

Remarks

- Initial question:
  How good do the observed frequencies fit to the frequencies expected under the null hypothesis?
  $\rightarrow$ the χ^2-test is also known as goodness of fit test
- the χ^2-statistic is asymptotically $\chi^2(d-1)$-distributed ($n \rightarrow \infty$)
  1. The approximation gets better the more data are found in the categories
  2. Why are the degrees of freedom $d-1$ (and not d)?
     Intuition: if we know that the first $d-1$ categories are occupied with $S = \sum_{k=1}^{d-1} X_i$ data, then it follows that the last category is occupied with $n - S$ data

Remarks

- Initial question:
  How good do the observed frequencies fit to the frequencies expected under the null hypothesis?
  $\rightarrow$ the χ^2-test is also known as goodness of fit test
- the χ^2-statistic is asymptotically $\chi^2(d-1)$-distributed ($n \rightarrow \infty$)
  1. The approximation gets better the more data are found in the categories
  2. Why are the degrees of freedom $d-1$ (and not d)?
     Intuition: if we know that the first $d-1$ categories are occupied with $S = \sum_{k=1}^{d-1} X_i$ data, then it follows that the last category is occupied with $n - S$ data
     $\rightarrow$ only $d-1$ categories are 'free'

Remarks

- Initial question:
  How good do the observed frequencies fit to the frequencies expected under the null hypothesis?
  $\rightarrow$ the χ^2-test is also known as goodness of fit test
- the χ^2-statistic is asymptotically $\chi^2(d-1)$-distributed ($n \rightarrow \infty$)
  1. The approximation gets better the more data are found in the categories
  2. Why are the degrees of freedom $d-1$ (and not d)?
     Intuition: if we know that the first $d-1$ categories are occupied with $S = \sum_{k=1}^{d-1} X_i$ data, then it follows that the last category is occupied with $n - S$ data
     $\rightarrow$ only $d-1$ categories are 'free'
  3. Why is the χ^2-distribution reasonable?

.

Remarks

- Initial question:
  How good do the observed frequencies fit to the frequencies expected under the null hypothesis?
  $\rightarrow$ the χ^2-test is also known as goodness of fit test
- the χ^2-statistic is asymptotically $\chi^2(d-1)$-distributed ($n \rightarrow \infty$)
  1. The approximation gets better the more data are found in the categories
  2. Why are the degrees of freedom $d-1$ (and not d)?
     Intuition: if we know that the first $d-1$ categories are occupied with $S = \sum_{k=1}^{d-1} X_i$ data, then it follows that the last category is occupied with $n-S$ data
     $\rightarrow$ only $d-1$ categories are 'free'
  3. Why is the χ^2-distribution reasonable?
     Intuition: the summands of the χ^2-statistic are squares of rescaled sums (frequencies).

Remarks

- Initial question:
  How good do the observed frequencies fit to the frequencies expected under the null hypothesis?
  $\rightarrow$ the χ^2-test is also known as goodness of fit test
- the χ^2-statistic is asymptotically $\chi^2(d-1)$-distributed ($n \rightarrow \infty$)
  1. The approximation gets better the more data are found in the categories
  2. Why are the degrees of freedom $d-1$ (and not d)?
     Intuition: if we know that the first $d-1$ categories are occupied with $S = \sum_{k=1}^{d-1} X_i$ data, then it follows that the last category is occupied with $n-S$ data
     $\rightarrow$ only $d-1$ categories are 'free'
  3. Why is the χ^2-distribution reasonable?
     Intuition: the summands of the χ^2-statistic are squares of rescaled sums (frequencies). Thus, according to the central limit theorem, each of the d summands is approximately distributed as the square of a $N(0,1)$-distributed random variable.

Remarks

- Initial question:
  How good do the observed frequencies fit to the frequencies expected under the null hypothesis?
  $\rightarrow$ the χ^2-test is also known as goodness of fit test
- the χ^2-statistic is asymptotically $\chi^2(d-1)$-distributed ($n \rightarrow \infty$)
  1. The approximation gets better the more data are found in the categories
  2. Why are the degrees of freedom $d-1$ (and not d)?
     Intuition: if we know that the first $d-1$ categories are occupied with $S = \sum_{k=1}^{d-1} X_i$ data, then it follows that the last category is occupied with $n-S$ data
     $\rightarrow$ only $d-1$ categories are 'free'
  3. Why is the χ^2-distribution reasonable?
     Intuition: the summands of the χ^2-statistic are squares of rescaled sums (frequencies). Thus, according to the central limit theorem, each of the d summands is approximately distributed as the square of a $N(0,1)$-distributed random variable. Under independence we would approximately obtain the $\chi^2(d)$-distribution. But the 'slight dependence' of the d summands (see. 2) results in the reduction of a degree of freedom.

χ^2-test in R

```
# Enter data
die      <- c("red","blue","blue","yellow",...)
# Calculate frequencies, e.g., via
x        <- table(die)
# Enter claimed probabilities
p0       <- c(1/6,1/3,1/12,1/6,1/12,1/6)
# Perform chi^2-test
chisq.test(x,p=p0,...)
# Output

        Chi-squared test for given probabilities

data:  x
X-squared = 5.375, df = 5, p-value = 0.3718
```

- If `p=p0` is not set (default), then equal probabilities are assumed ('fair'), i.e., $p_0 = (1/d, \ldots, 1/d)$
- For few data (n small) a so-called 'continuity correction' (according to Yates) is performed. For that, in the χ^2-statistic the numerator of every summand is (before squaring) replaced by its absolute value, then subtracted by $1/2$ and then squared. Idea: conservative behavior (reject less easily) $\rightarrow$ 'counteract a bad approximation through the χ^2-distribution'.
  The continuity correction can be controlled through the logical argument `correct`.

Thank you!