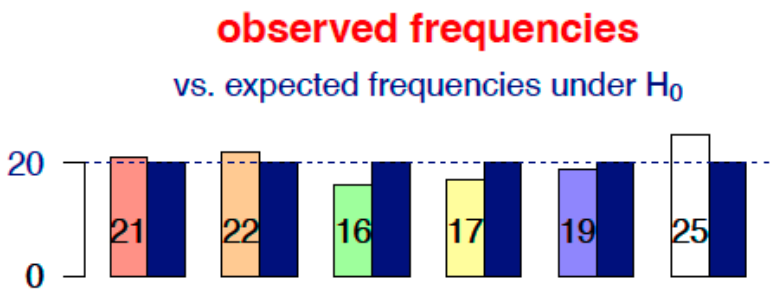
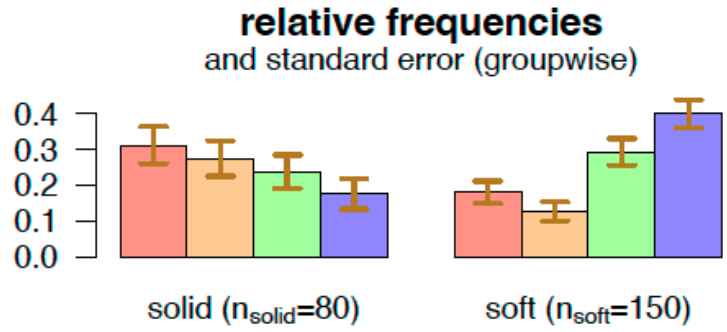
TECHNISCHE UNIVERSITÄT WIEN

The χ^2- tests



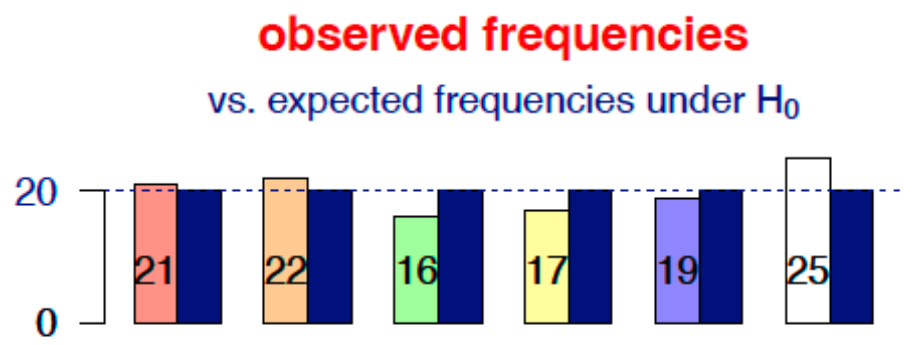


goodness of fit

for independence

Slides were developed together with Dr. Michael Messer, former ASTAT member
Please send your comments to: T. Levajković, tijana.levajkovic@tuwien.ac.at

The χ^2- test (goodness of fit)

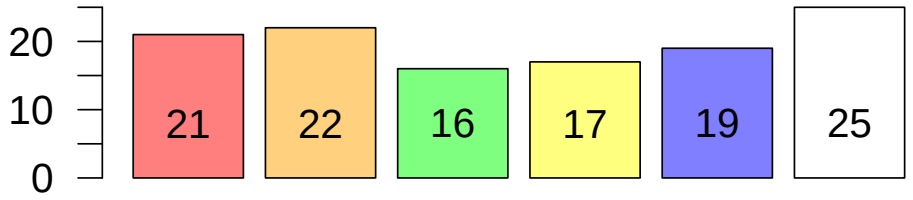


Motivation

- 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)

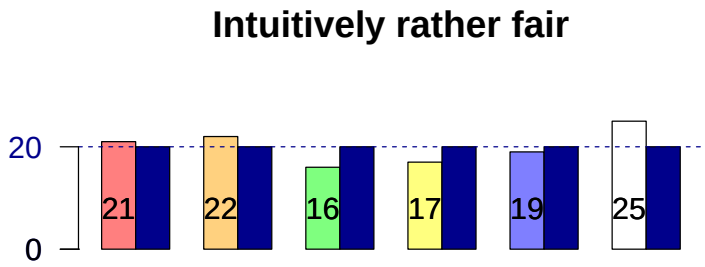
observed frequencies

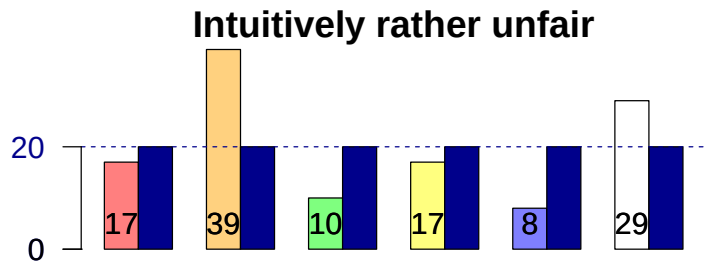


Motivation

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

  relative frequencies and standard error → rather close?!





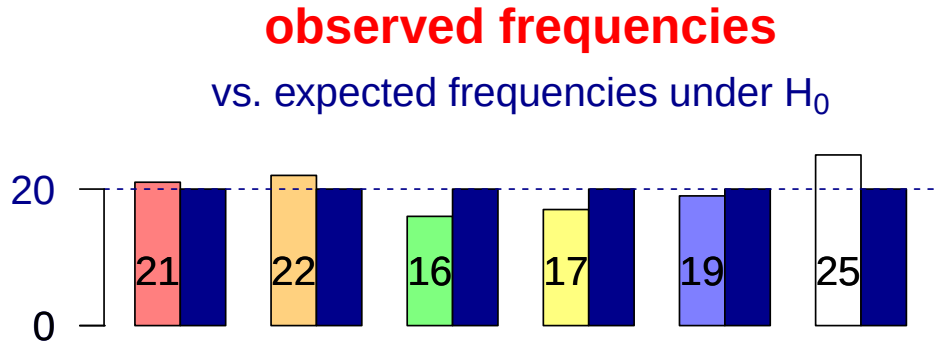
- 'Answer' 2:
  - meaning of 'fair': no side is preferred, probability 1/6
  - per category $120 \cdot 1/6 = 20$ occupations expected, if the die is fair
  - in every category the observed frequencies should then typically be 'close' to the expected frequencies
  - a statistic, that quantifies this discrepancy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

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

- 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*

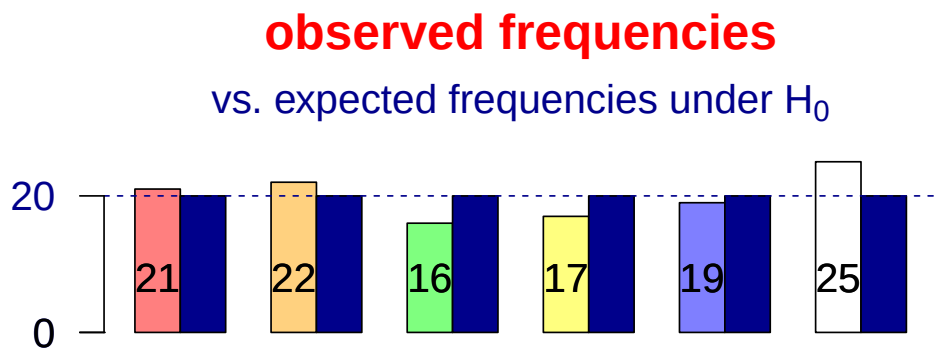
$$\binom{n}{x_1, x_2, \ldots, x_d} := \frac{n!}{x_1! \cdots x_d!} = \binom{n}{x_1}\binom{n-x_1}{x_2}\binom{n-x_1-x_2}{x_3}\cdots\binom{n-x_1-\cdots-x_{d-1}}{x_d}$$

- 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: $x_1, \cdots, x_d$
- 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 $\sum_{k=1}^{d} p_k = 1$
- $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$
- $\mathbb{E}_{H_0}[X_k] = n \cdot p_{0,k}$
- 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?

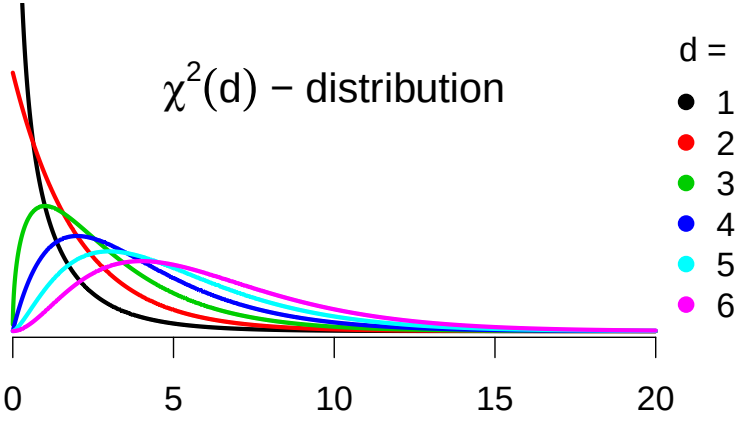
k	1	2	3	4	5	6	$\sum$
x_k	21	22	16	17	19	25	120
$\mathbb{E}_{H_0}[X_k]$	20	20	20	20	20	20	120

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

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

- 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)

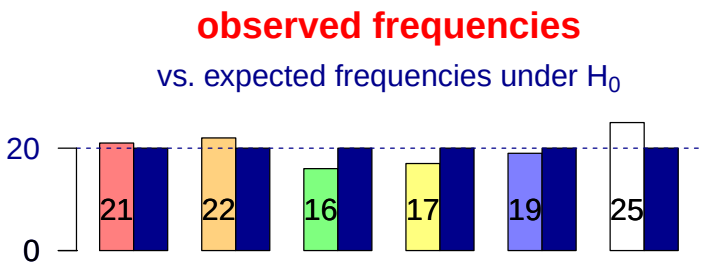
- Let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^d p_k = 1$

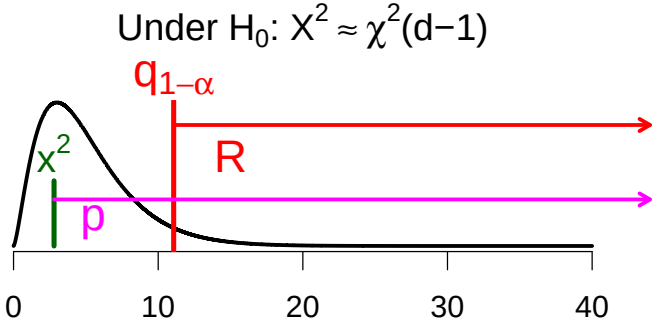
Under $H_0 : p = (p_{0,1}, \ldots, p_{0,d})^t$ it holds (approximately)

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

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

- 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





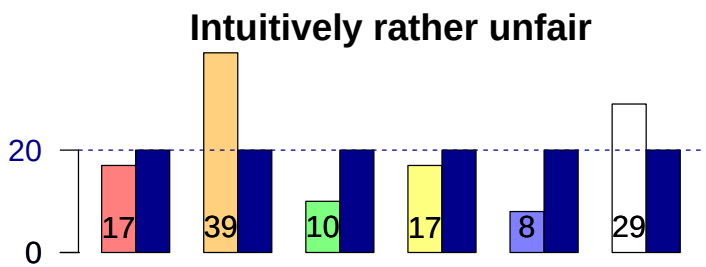
The χ^2-test, goodness of fit (example 2)

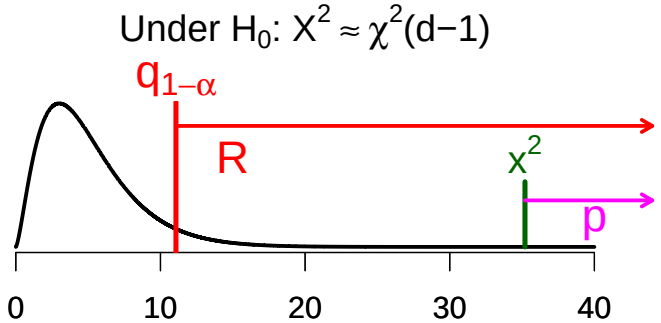
- Let $\mathfrak{X} = (X_1, \ldots, X_d)^t \sim mult(n, p)$, with $p \in (0,1)^d$ and $\sum_{k=1}^d p_k = 1$

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: $R = [q_{1-\alpha}, \infty)$
- 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

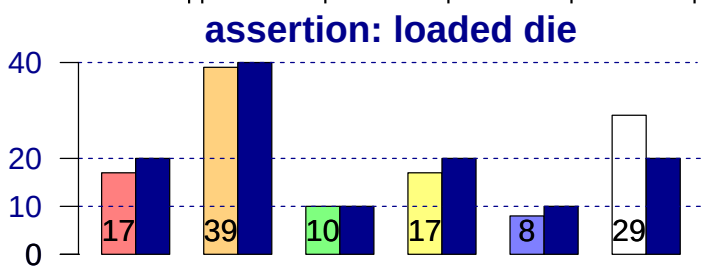


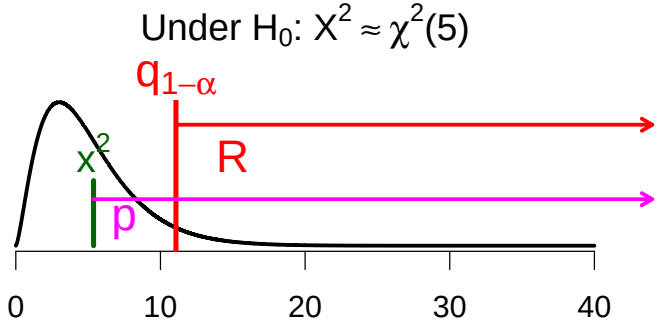


Loaded die (example 3)

- Somebody claims: 'I loaded the die' , 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
- 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`.

Overview

So long:

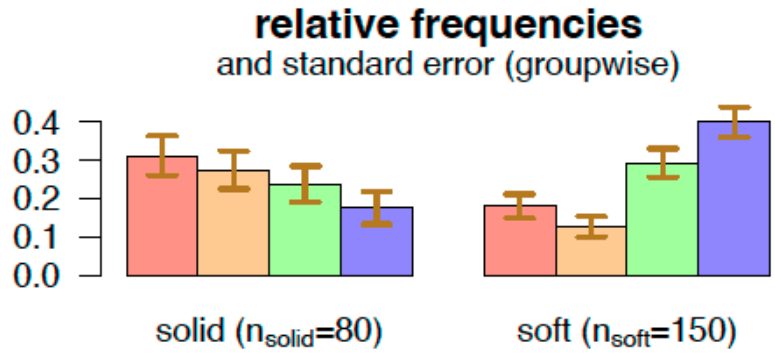
- χ^2-test, good of fit:
- One feature (here: outcome of the rolling die)
  $\rightarrow$ in d categories (here: colors)
- data: frequencies / occupations
- Question: How good do the observed frequencies fit the claimed occupation probabilities?
- Statistic: χ^2-statistic

In the following:

- χ^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: χ^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

The χ^2- test (for independence)



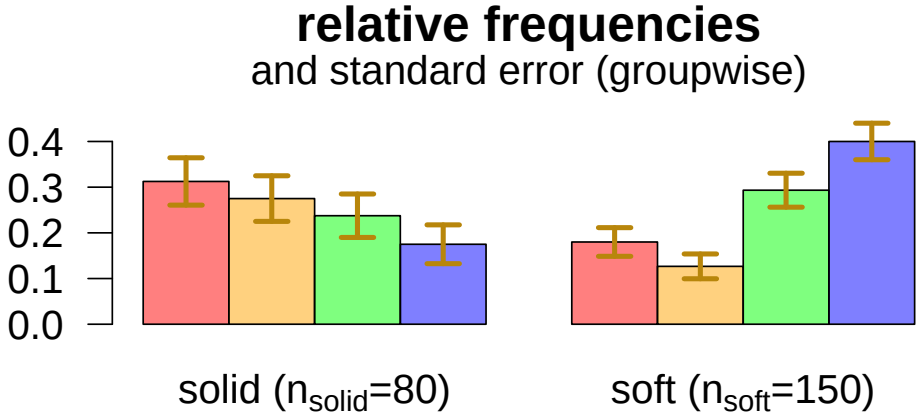
Motivation

- 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	$x_{j,\cdot}$
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: χ^2-*test for independence*

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 $\sum_{j,k} p_{j,k} = 1$
- 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	$n \cdot p_{2,\cdot} \cdot p_{\cdot,1}$	$n \cdot p_{2,\cdot} \cdot p_{\cdot,2}$	$n \cdot p_{2,\cdot} \cdot p_{\cdot,3}$	$n \cdot p_{2,\cdot} \cdot p_{\cdot,4}$

- 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	$x_{j,\cdot}$
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 χ^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 χ^2-distribution reveals...

The χ^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 $\sum_{j,k} p_{j,k} = 1$

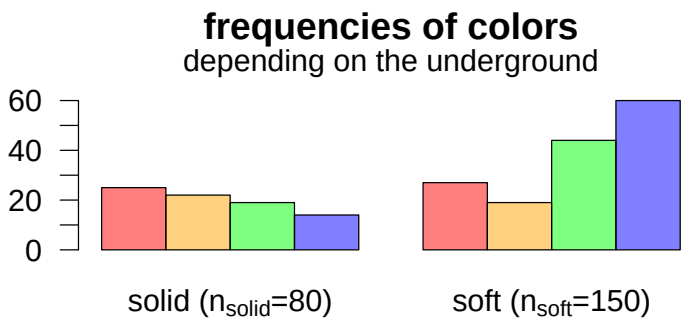
Under $H_0 : p = p_0 := (p_{1,\cdot} \cdot p_{\cdot,1}, \ldots, p_{d_1,\cdot} \cdot p_{\cdot,d_2})^t \in (0,1)^{d_1 \cdot d_2}$
and $\sum_{j=1}^{d_1} p_{j,\cdot} = \sum_{k=1}^{d_2} p_{\cdot,k} = 1$

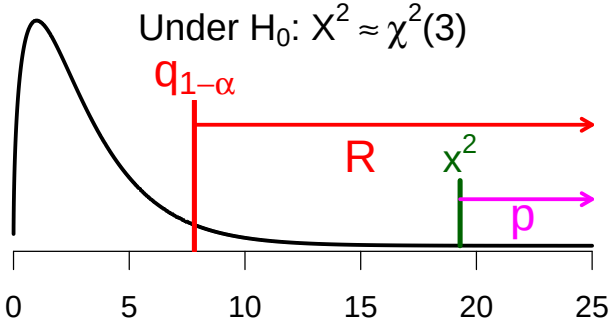
it holds (approximately)

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

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

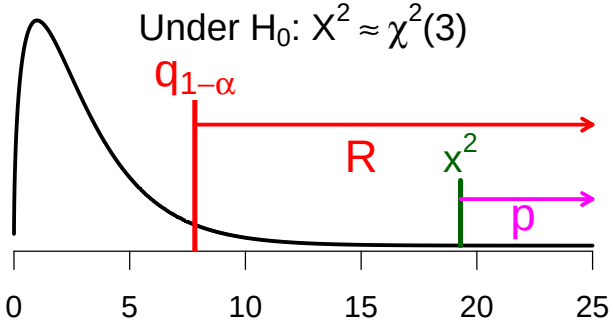




Nutshell

$x_{j,k}$			
	$x_{1,1}$	$\cdots$	x_{1,d_2}	$x_{1,\cdot}$
	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	$x_{d_1,1}$	$\cdots$	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 χ^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'
- 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 χ^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').

χ^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
```

Multiple-choice questions

(1) On which test could you think?
Between the majors math, physics and computer science, is there a difference in the proportions of students that regularly drink coffee?

a. χ^2-goodness of fit test with one category
b. χ^2-goodness of fit test with two categories
c. χ^2-test for independence
d. t-test for proportions

Multiple-choice questions

(2) For a project, Anna randomly picks 100 fellow VO Statistics students to survey on whether each has either a PC or Apple at home (all students have a home computer) and what score (1, 2, 3, 4, 5) each expects to receive on the exam. She applied a χ^2- test of independence. How many degrees of freedom are there?

a. 1
b. 4
c. 7
d. 9

Multiple-choice questions

(3) For a project, Anna randomly picks 150 fellow VO Statistics students to survey on whether each has either a PC or an Apple computer at home (all students have a computer at home) and what score (1, 2, 3, 4, 5) each expects to receive on the VO Statistics exam. A χ^2-test of independence results in a test statistic of 11. What is the p-value of this test?

a. 0.001
b. 0.025
c. 0.15
d. 0.05

Multiple-choice questions

(4) To test the claim that dogs bite more or less depending upon the phase of the moon, a university hospital counts admissions for dog bites and classifies with moon phase.

	New moon	First quarter	Full moon	Last quarter
Dog bite admissions	32	27	47	38

Which of the following is the proper conclusion?

a. The data prove that dog bites occur equally during all moon phases.
b. The data give evidence that dog bites occur equally during all moon phases.
c. The data give evidence that dog bites do not occur equally during all moon phases.
d. The data do not give sufficient evidence to conclude that dog bites are related to moon phases.

Multiple-choice questions

(5) A geneticist claims that four species of fruit flies should appear in the ratio 1 : 3 : 3 : 9. Suppose that a sample of 2000 flies contained 110, 345, 360 and 1185 flies of each species, respectively.

Is there sufficient evidence to reject the geneticist's claim?

a. The data prove the geneticist's claim.
b. The data prove the geneticist's claim is false.
c. The data do not give sufficient evidence to reject the geneticist's claim.
d. The data give sufficient evidence to reject the geneticist's claim.

Multiple-choice questions

(6) Anna performs a χ^2-test for independence in `R` using `chisq.test()`.
A sufficient input is

a. the matrix of absolute cell-frequencies
b. the matrix of relative cell-frequencies
c. the total number of observations
d. the vector of all observations

Multiple-choice questions

(7) In crosses between two types of maize four distinct types of plants were found in the second generation. In a sample of 1301 plants there were 773 green, 231 golden, 238 green-striped and 59 golden-green-striped. According to a simple theory of genetical inheritance the probabilities of obtaining these four plants are $\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}$ respectively. Is the theory acceptable as a model for this experiment?

a. Yes and the test statistic is 9.272.

b. No because the p-value is 0.026.

c. Yes at 5% level of significance.

d. The model is not specified.

Multiple-choice questions

(8) In the context of the goodness of fit χ^2-test for four categories let the observed frequencies be $5, 10, 10$ und 15. Let the null hypothesis be that no category is preferred. Further let the rejection region be $R = [7, \infty)$. Then,

a. we reject the null hypothesis

b. we do not reject the null hypothesis

c. we can not say of whether we reject the null hypothesis

d. due to the data type we should have performed another test

Multiple-choice questions

(9) A dice is tossed 120 times with the following results

number turned up	1	2	3	4	5	6
frequancy	30	25	18	10	22	15

Anna tests the hypothesis that the dice is unbiased. If the rejection region is $R = [11.7, +\infty)$, which one of the following statements is correct?

a. the χ^2-statistic is 11.3 and she should reject the null.
b. the χ^2-statistic is 12.9 and she should reject the null.
c. the χ^2-statistic is 10.9 and she should not reject the null.
d. the χ^2-statistic is 12.3 she should not reject the null.

Multiple-choice questions

(10) The null hypothesis is rejected in a χ^2-test for independence with the level of significance α when

a. the p-value is larger than α.
b. the p-value is larger than $1 - \alpha$.
c. the χ^2-statistic is larger than the critical value for the given α.
d. the χ^2-statistic is smaller than the critical value for the given α.

Thank you for your attention!