Introduction to R for data analysis

- hypothesis tests -

Carl Herrmann & Carlos Ramirez
R4SC - Freiburg June 2024

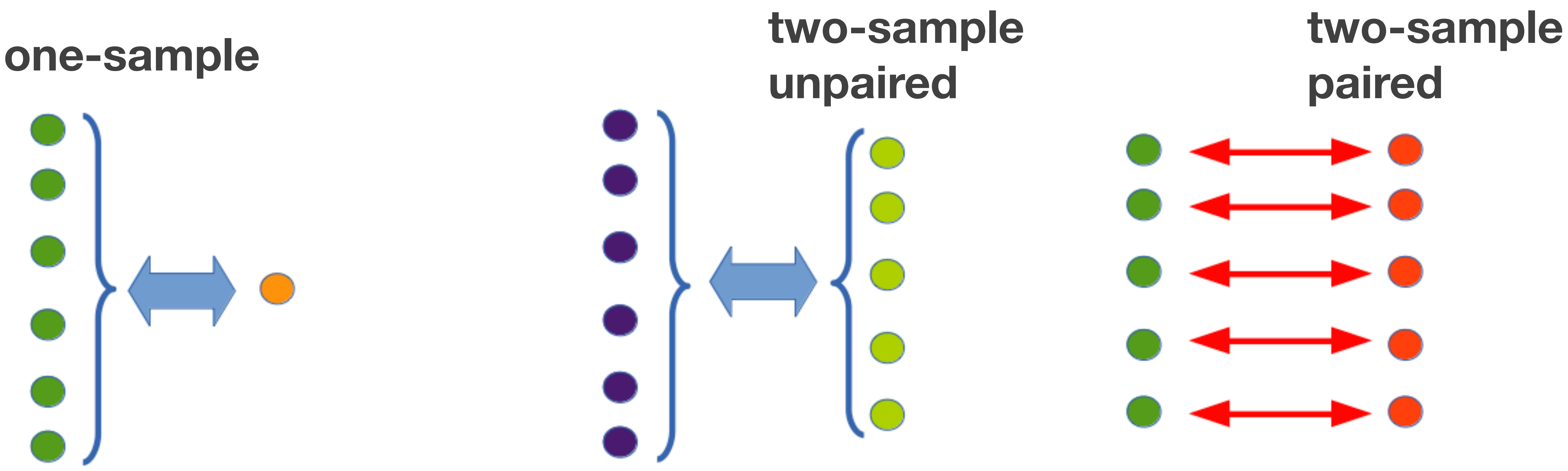
IPMB
Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Testing the means

Test on mean values

- Hypothesis on mean values can be investigated using a ***t-test***
- Family of tests with different version:
  - **one-sample test**: *is the mean body temperature 37.7 C?*
  - **two-sample test, unpaired**: *do men and women have different mean cholesterol levels?*
  - **two-sample test, paired**: *is there a change in cholesterol level after a one-month egg rich diet?*



(do both samples have equal variance?)

Running a t-test in R

two-sample unpaired, two-sided

t = test statistics
df = degrees of freedom

confidence interval differences of the means

```
> t.test(weight.m,weight.f,var.equal=TRUE)

        Two Sample t-test
data:  weight.m and weight.f

t = 1.8265, df = 400, p-value = 0.06852

alternative hypothesis: true difference in
means is not equal to 0

95 percent confidence interval:
 -0.5669448 15.4259192

sample estimates:
mean of x mean of y
 181.9167  174.4872
```

Running a t-test in R

two-sample unpaired, one-sided

t = test statistics
df = degrees of freedom

confidence interval differences of the means

```
>t.test(weight.m,weight.f,alternative="greater",va
r.equal=TRUE)

        Two Sample t-test
data:  weight.m and weight.f

t = 1.8265, df = 400, p-value = 0.03426

alternative hypothesis: true difference in means
is greater than 0

95 percent confidence interval:
 0.723444      Inf
sample estimates:
mean of x mean of y
 181.9167  174.4872
```

Testing proportions

Proportion tests

- This class of tests can be used when searching for
  - **relation between different categorical variables**
    Is there a relation between social background and school grades?
  - comparison of **observed** vs. **expected** counts
    Is there a significant gender bias in the math department if 4 professors out of 10 are women?
- Two tests are generally used
  - **Fisher-Exact test** (FET): gives an exact p-value, used for small samples
  - **chi-square test**: for larger samples (n>5 in each category)
  - both tests are equivalent for large n

Fisher Exact Test

- Tests for a significant relationship between 2 variables
- Starting point: contingency table

	iPhone	other	Total
Men	**4**	**1**	5
Women	**2**	**3**	5
Total	6	4	10

Proportion iPhone/other:
- Men : 4/1 = 4
- Women: 2/3 = 0.66

Odds-Ratio:

OR = (4/1)/(2/3) = 6

If we would <u>randomly</u> distribute 6 iPhone and 4 other smartphones to 5 men and 5 women, how often would we get a larger/smaller*/more extreme

*smaller: < 1/6

**More extreme: > 6 or < 1/6

chi-square test

- The chi-square test compares **observed** and **expected** counts
- Starting point is a **contingency table**
- Test statistics

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

- H_0: expected and observed proportions are equal
- H_0 distribution: chi2 distribution with *n-1* degrees of freedom for *n* observations
- Application possible when $O_i > 2$ and $O_i > 5$ in 80% of observations
- *Note: the chi-square test is always a 1-sided upper tail test!*

Observed

	iPhone	other	Total
Men	**14**	**30**	44
Women	**5**	**20**	25
Total	19	50	69

Observed proportions

	iPhone	other	Total
Men	**31,8 %**	**68,2 %**	100 %
Women	**20 %**	**80 %**	100 %
Total	27,5 %	72,5 %	100 %

H0 proportions

	iPhone	other	Total
Men	**27,5 %**	**72,5 %**	100 %
Women	**27,5 %**	**72,5 %**	100 %
Total	27,5 %	72,5 %	100 %

Expected counts under H0

	iPhone	other	Total
Men	**12,1**	**31,9**	44
Women	**6,9**	**18,1**	25
Total	19	50	69

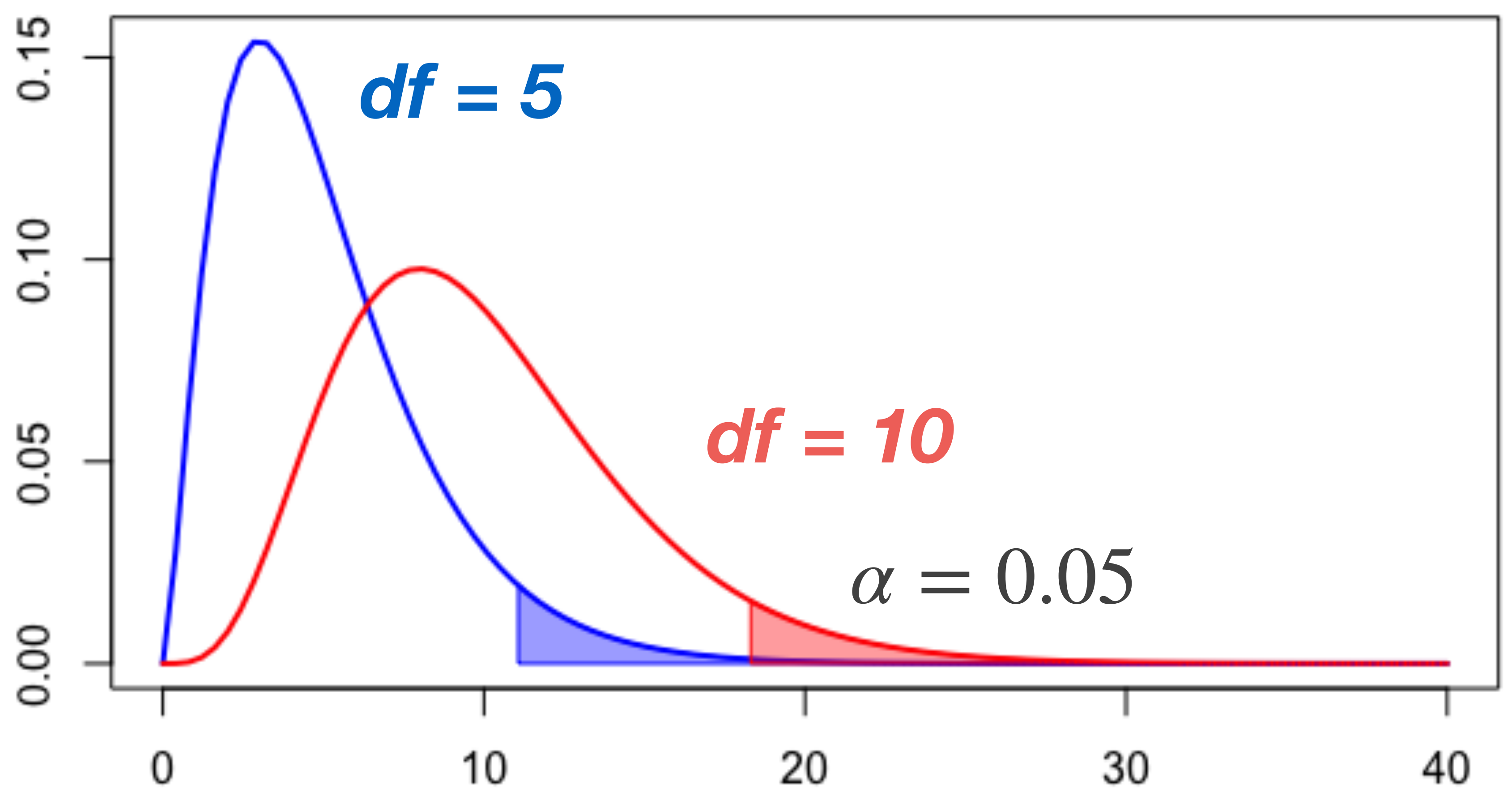
$$\chi^2 = \frac{(14-12.1)^2}{12.1} + \frac{(30-31.9)^2}{31.9} + \frac{(5-6.9)^2}{6.9} + \frac{(20-18.1)^2}{18.1}$$
$$= 0.6022$$

degrees of freedom = (rows-1) x (columns-1)

chi-square distribution

Critical values

	0,025	0,05	0,1
df = 1	5,02	3,84	2,71
df = 2	7,38	5,99	4,61
df = 3	9,35	7,81	6,25
df = 4	11,14	9,49	7,78
df = 5	12,83	11,07	9,24
df = 6	14,45	12,59	10,64
df = 7	16,01	14,07	12,02
df = 8	17,53	15,51	13,36
df = 9	19,02	16,92	14,68
df = 10	20,48	18,31	15,99



$\alpha = 0.05$

$\chi^2 = 0.6022$ **not significant...**

$df = 1$

More than 2 categories

Side effects

	weak	medium	strong	Total
Drug A	**25**	**11**	**13**	49
Drug B	**9**	**14**	**11**	34
Total	34	25	24	83

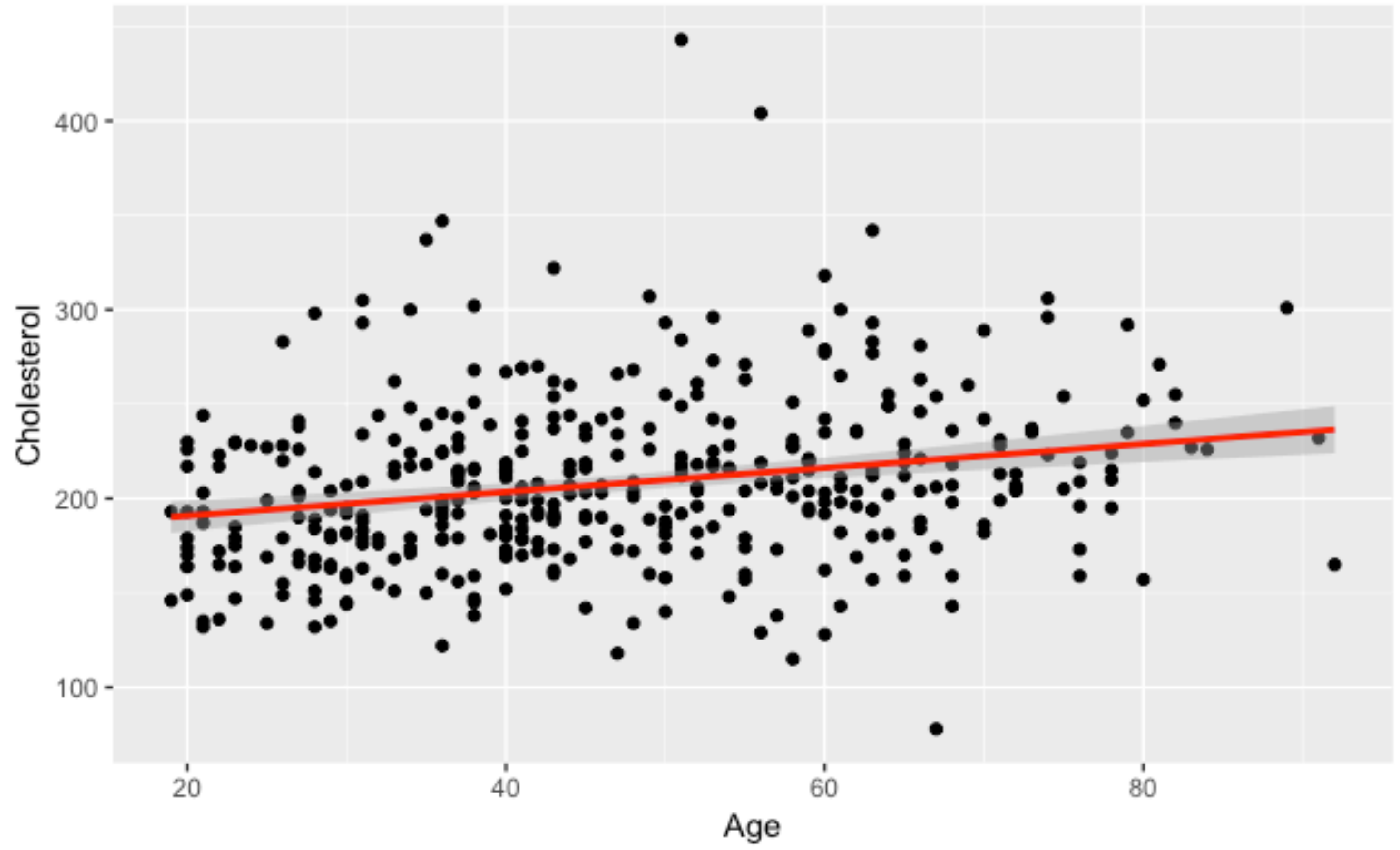
	weak	medium	strong	Total
Drug A	**51 %**	**22,5 %**	**26,5 %**	100 %
Drug B	**26,5 %**	**41,2 %**	**32,3 %**	100 %
Total	41 %	30,1 %	28,9 %	100 %

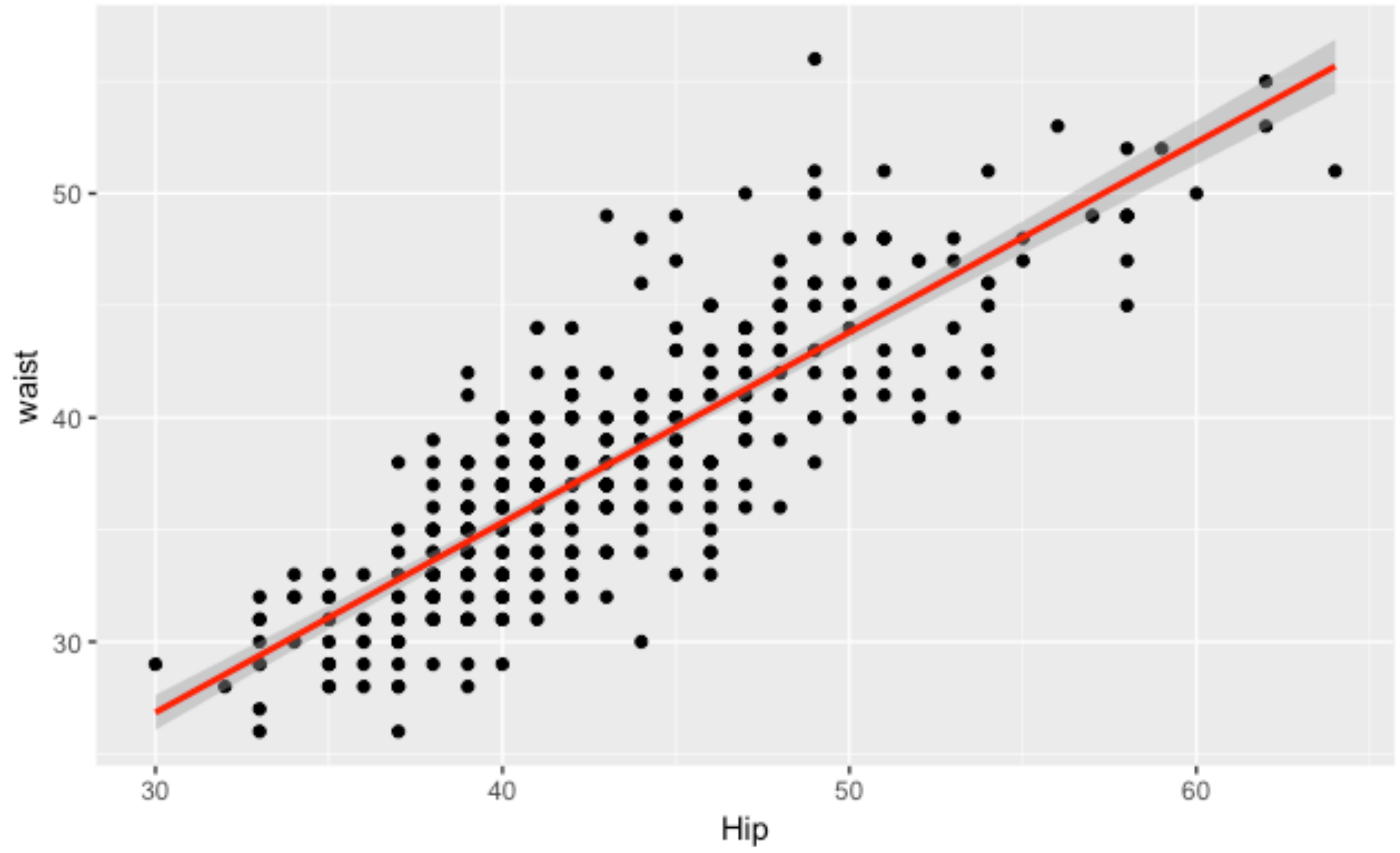
```
> table(sideeffect)
    SideEffect
Drug weak medium strong
   A     25      11      13
   B      9      14      11

> chisq.test(table(sideeffect))
        Pearson's Chi-squared test
data:  table(sideeffect)
X-squared = 5.5257, df = 2, p-value = 0.06311


> fisher.test(table(sideeffect))
        Fisher's Exact Test for Count Data
data:  table(sideeffect)
p-value = 0.06375
alternative hypothesis: two.sided
```

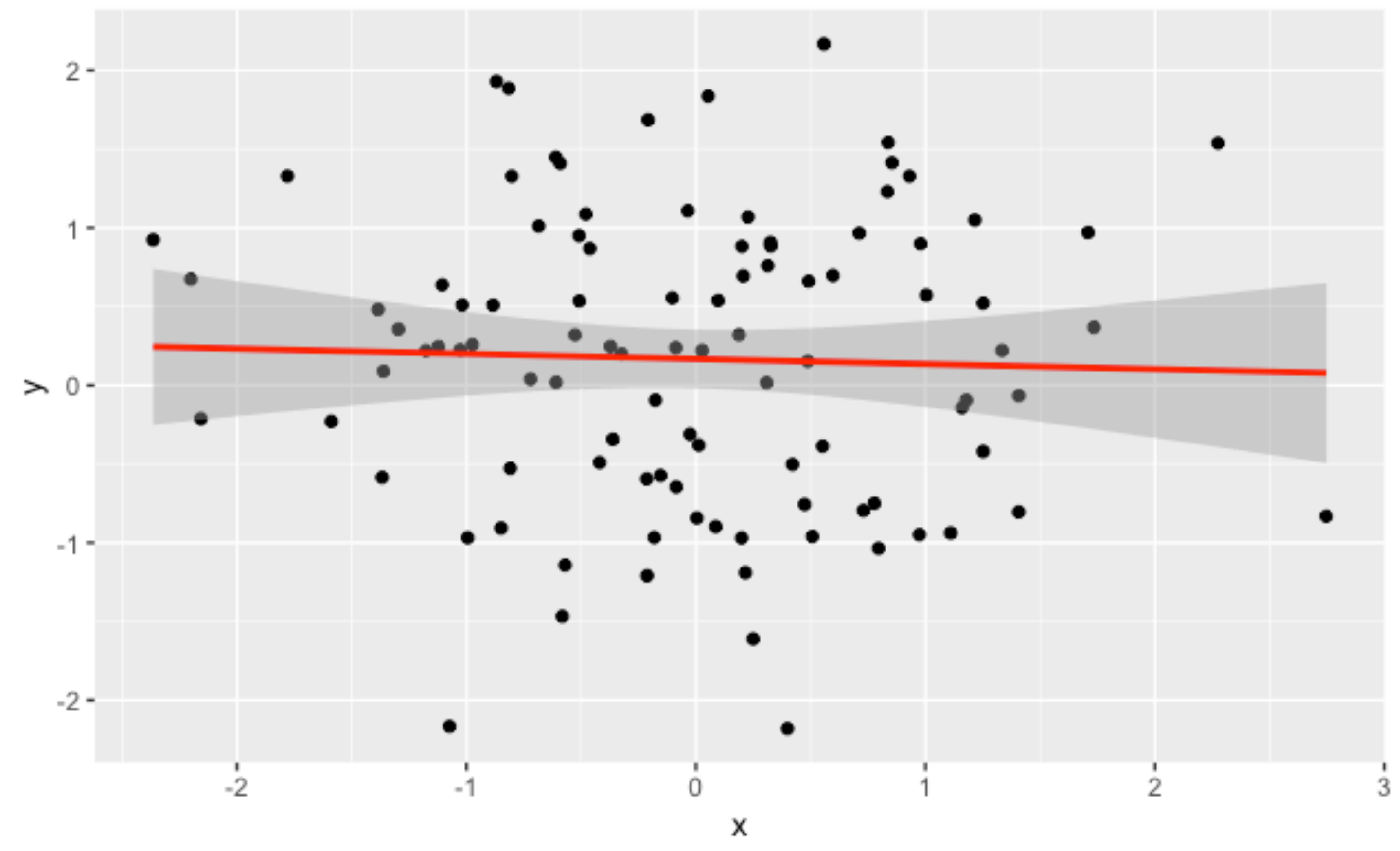
Testing correlations

Relation between numerical variables

- How easy is it to draw a line through a scatter plot?







Relation between numerical variables

- Variance:

$$Var(x) = (s_x)^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

dimension: $[x]^2$

- Covariance :

$$Cov(x,y) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

dimension: [x][y]

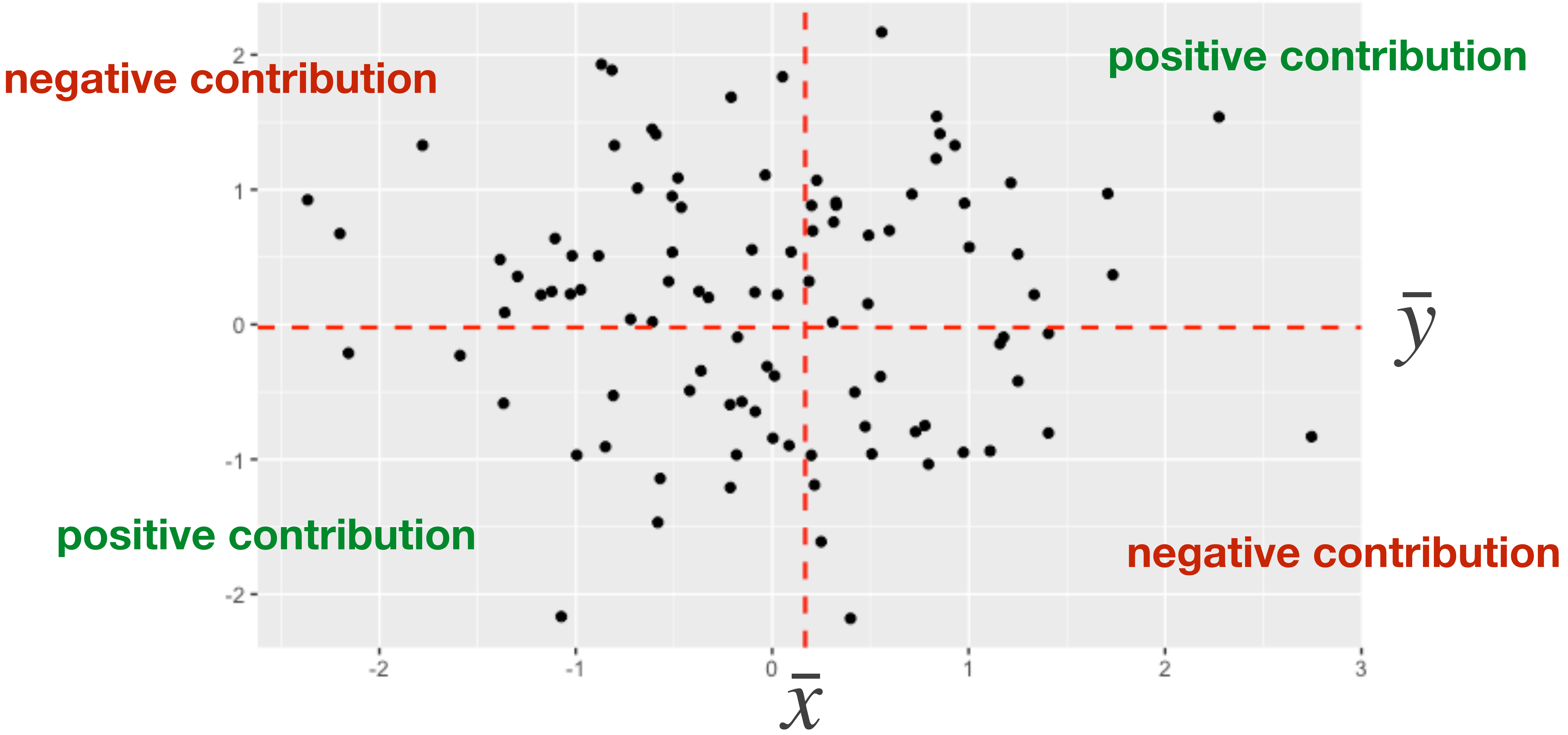
- Pearson Correlation :

$$Corr(x,y) = r = \frac{1}{N-1}\sum_{i=1}^{N}\frac{(x_i - \bar{x})}{s_x}\frac{(y_i - \bar{y})}{s_y}$$

dimension: none

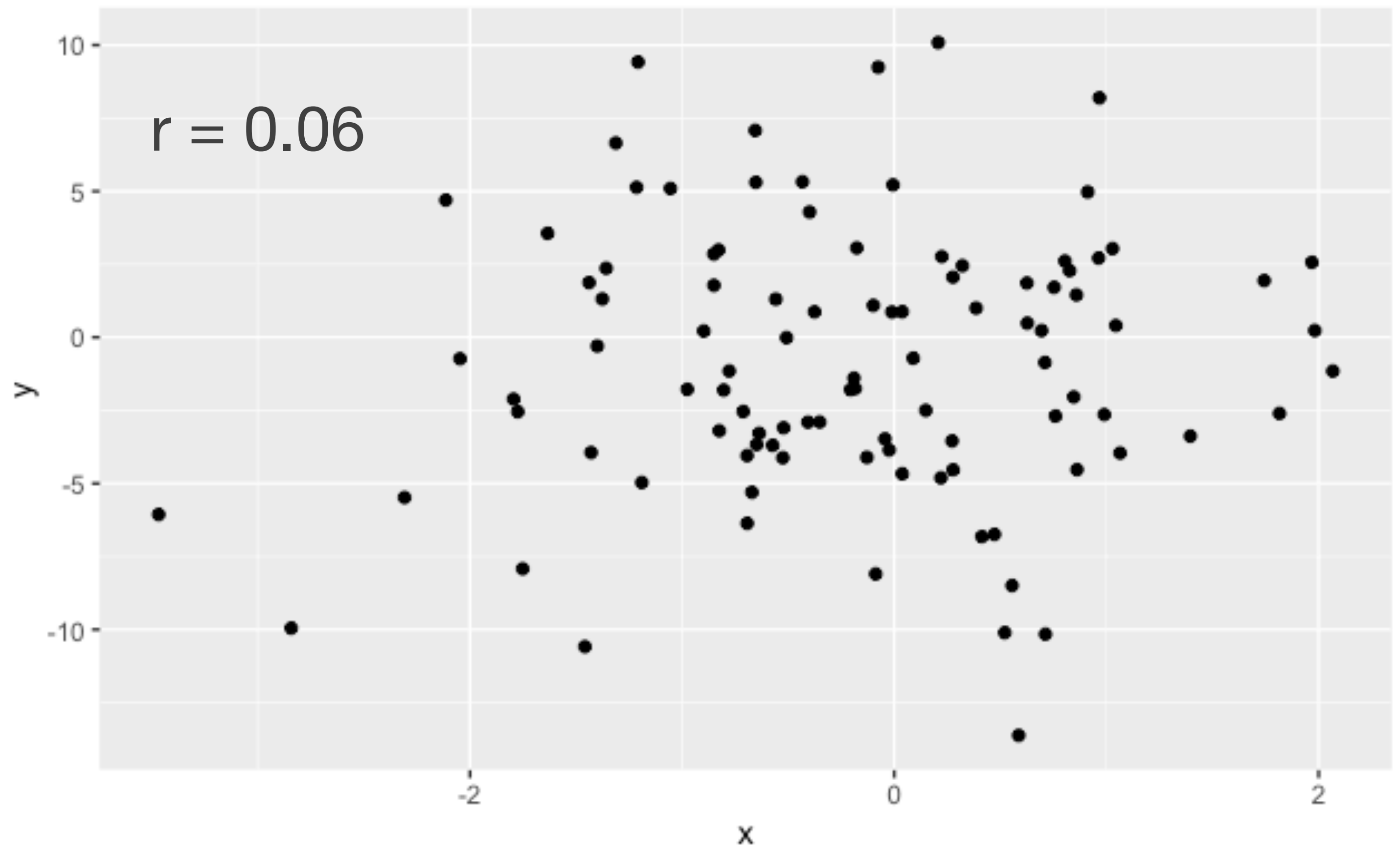
- Properties:
  - correlation is scale invariant, covariance is not!
  - cor(x,x) = 1
  - -1 =< cor(x,y) =< +1

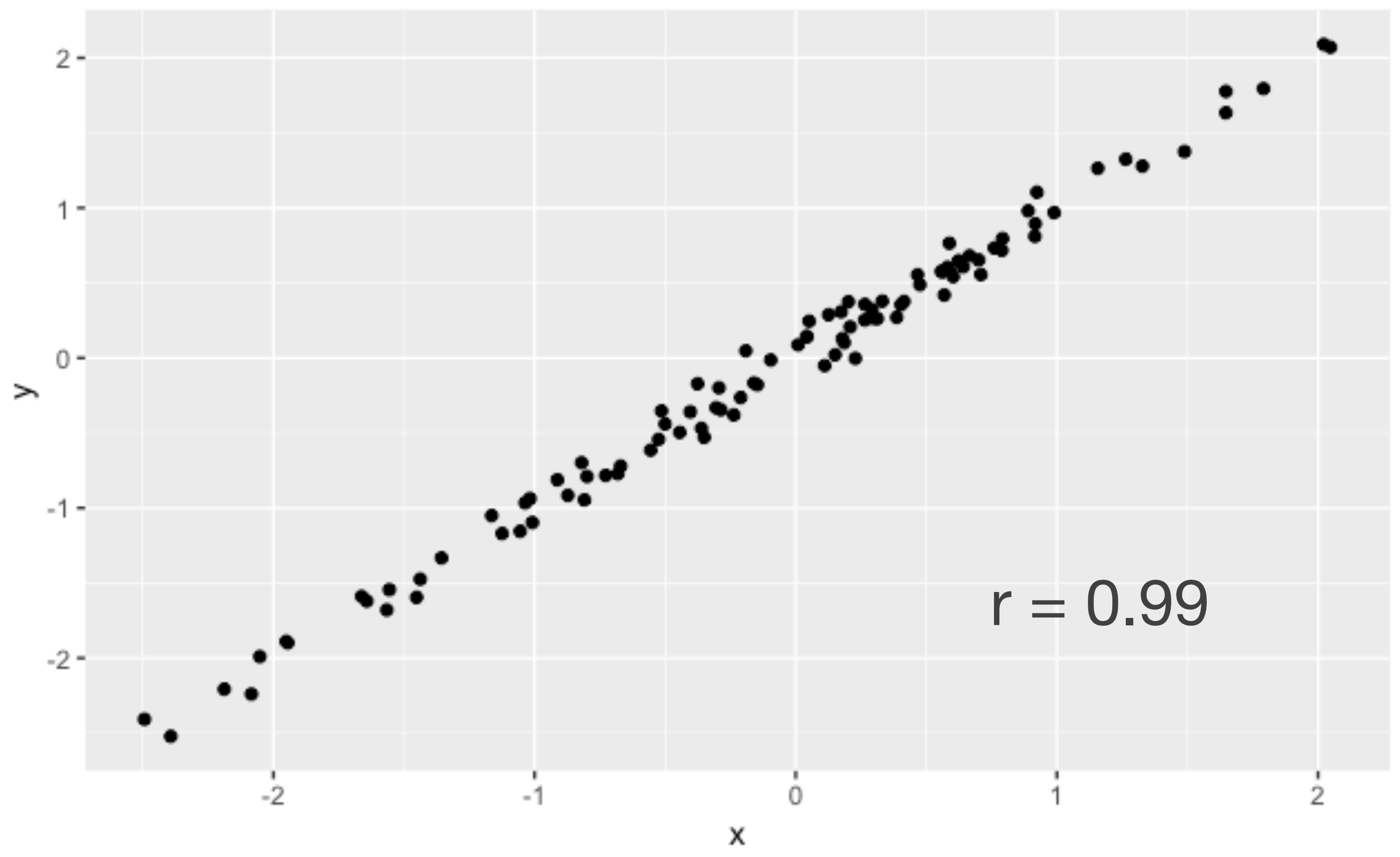
Relation between numerical variables



$$Corr(x,y) = \frac{1}{N-1}\sum_{i=1}^{N}\frac{(x_i-\bar{x})}{s_x}\frac{(y_i-\bar{y})}{s_y}$$

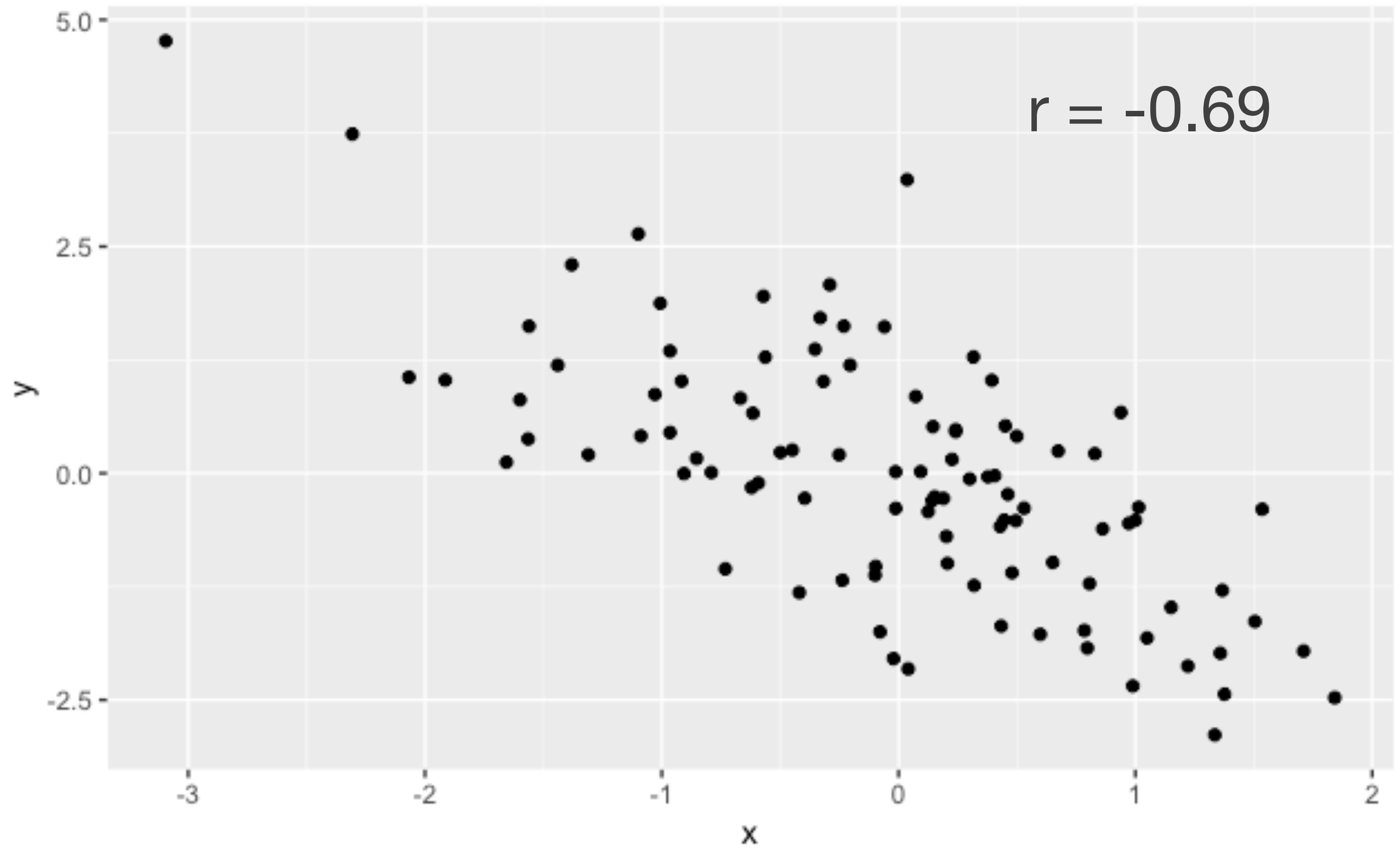
Relation between numerical variables





These are sample-based estimations of the correlation

→ what about the population correlation?



Statistical test on correlation

- the sample correlation coefficient ***r*** is an estimate of the real unknown correlation coefficient **ρ**
- Hypothesis test: ***could ρ actually be zero?***
- t-test with H_0: $\rho = 0$

$$t = \frac{r}{se_r}$$

estimate

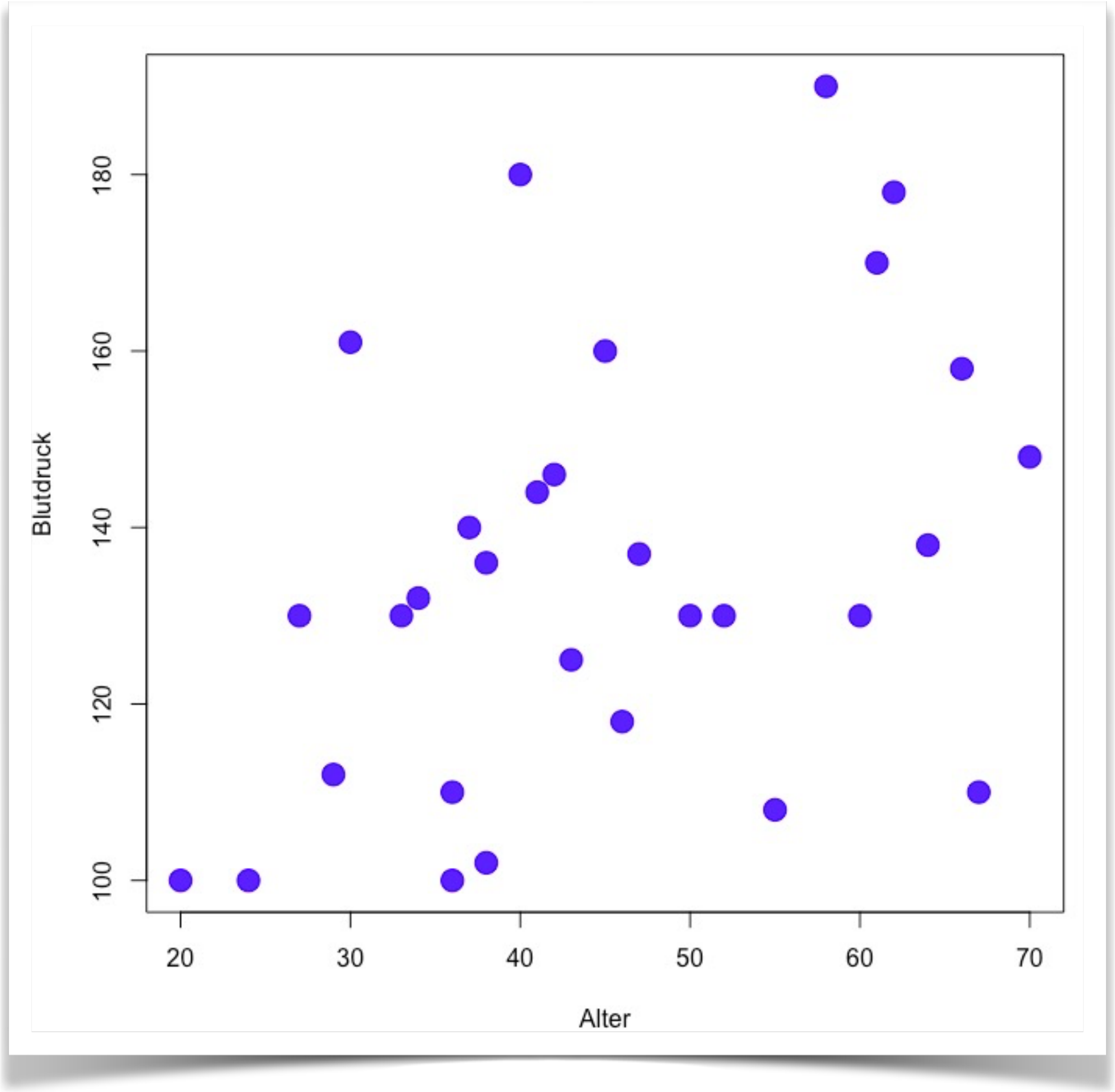
standard error

$$se_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

- H_0 distribution: t-distribution with n-2 degrees of freedom

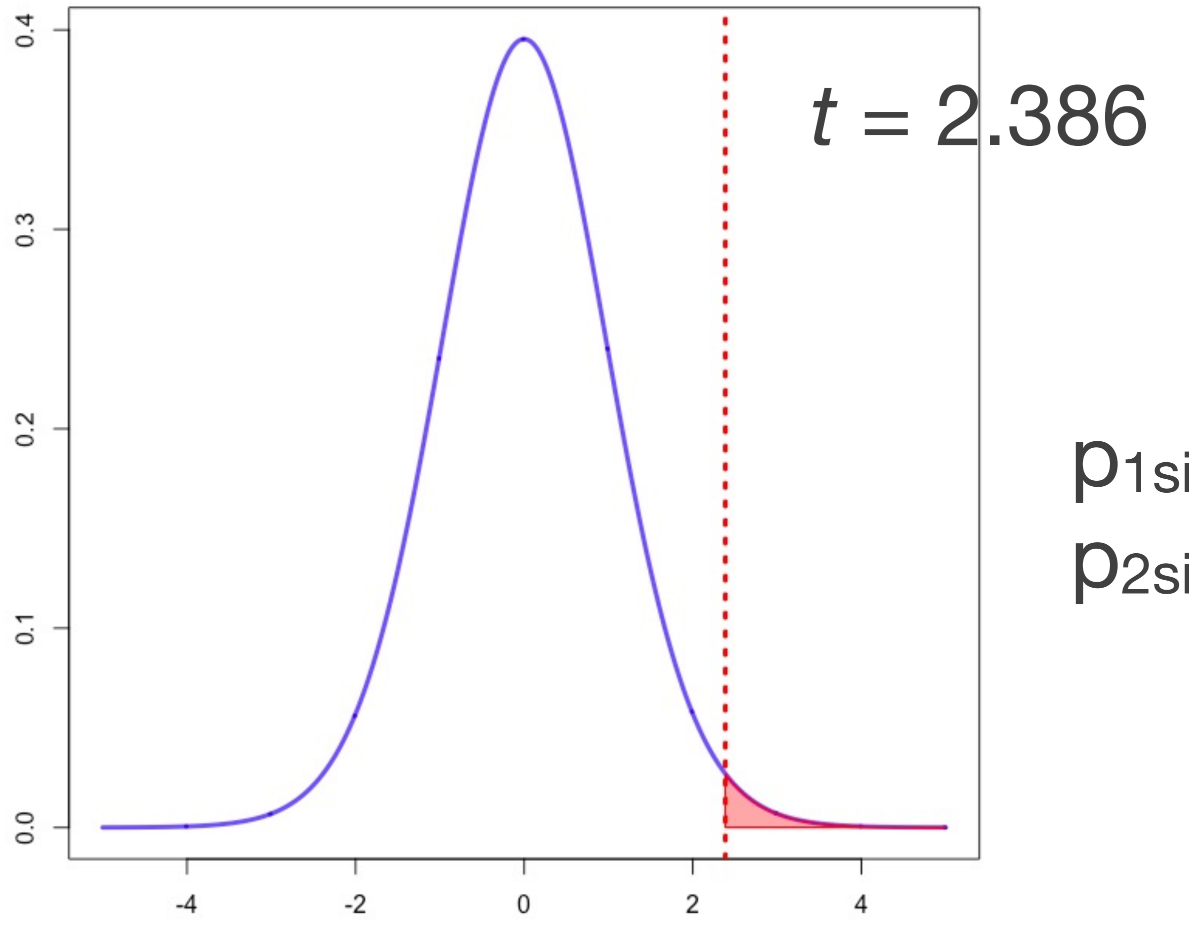
Example

n = 30




```
> cor.test(diab[1:30,7],diab[1:30,12])

    Pearson's product-moment correlation

data:  diab[1:30, 7] and diab[1:30, 12]
t = 2.386, df = 28, p-value = 0.02404
alternative hypothesis: true correlation is
not equal to 0
95 percent confidence interval:
 0.05960801 0.67182894
sample estimates:
    cor
0.41105
```

$$t = \frac{r}{se_r} \qquad se_r = \sqrt{\frac{1-r^2}{n-2}}$$



$p_{1sided} = 0.012$
$p_{2sided} = 0.024$