2022 Winter School

Distributions, hypothesis tests, variance & covariance, correlation

Kathryn Kemper

Data in statistics

- Statisticians will refer to a vector X (or Y, X_1 etc.) of “random variables”
  - e.g. a coin toss: H, T, H, T, T, T, H
  - e.g. a sample of heights 164.1, 173.2, 150.0

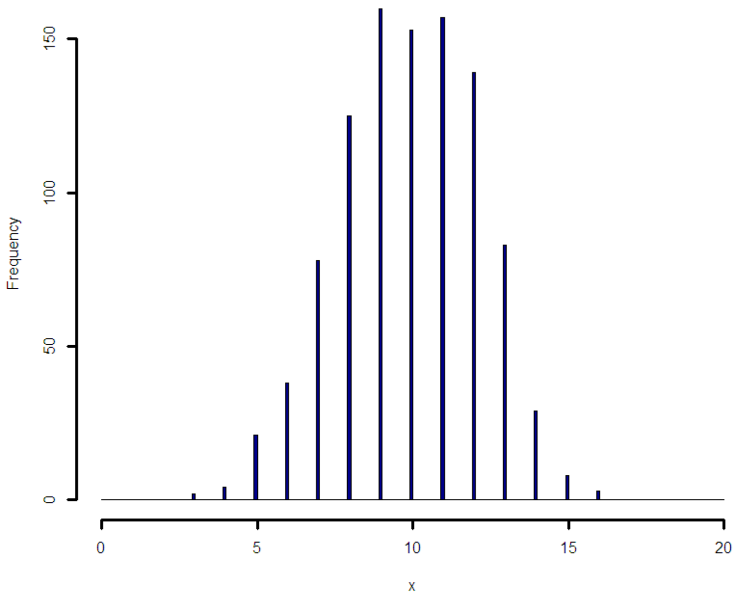
- We normally assume X to be from a distribution. This distribution defines the properties of the *random variable* (e.g the probability that $x < 0$ or $x = 5$)

For example, toss a coin

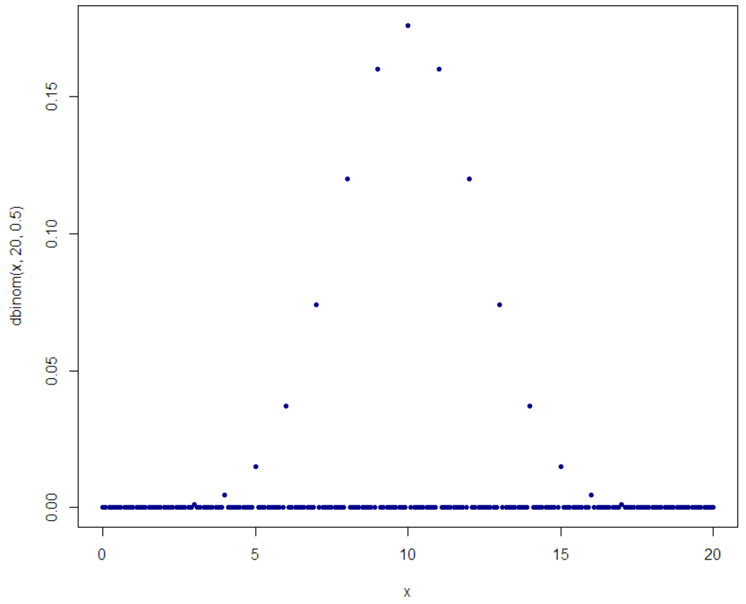
Toss a coin 20 times, count the number of heads

e.g. 7H, 11H, 10H, 19H, 12H, 9H, 6H, 10H, 9H….

Histogram of a 1000 samples



Probability density function (PDF)
P(x) = B(20,0.5)



$$P(X=15)=\frac{20!}{15!(20-15)!}0.5^{15}0.5^{(20-15)}$$

What is a parameter?

- Number that describes a population
- Often unknown and unknowable
  - e.g. how many people suffer from diabetes in Australia
- In statistics, we usually make <u>estimates</u> of parameters
- Parameters (once estimated) define a range of possible outcomes, or the assumed relationship between two variables
  - Fundamental descriptor

Useful distributions

Parameters are often represented by Greek letters, e.g. μ, λ, σ

- Continuous:
  - Normal
  - T-distribution
  - Chi-squared
  - F-distribution

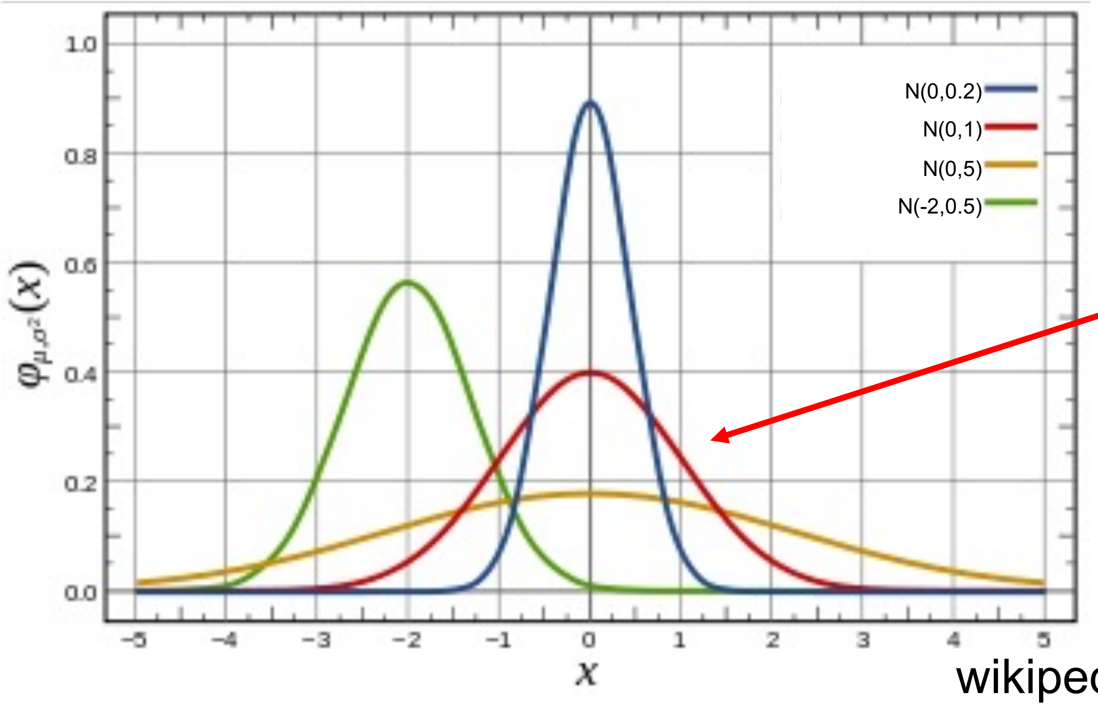
Discrete:

- Binomial

Normal distribution:

N(μ,σ²) : parameters = μ,σ²

- μ ('mu'; the mean) $\left[\mu = \frac{1}{N}\sum x_i\right]$ centre
- σ² ('sigma-squared'; the variance) $\left[\sigma^2 = \frac{1}{N}\sum (x_i - \mu)^2\right]$ spread



a 'standard' normal; N(0,1)

wikipedia.org/

$$SSQ = \sum (y - \mu)^2$$

Mean & Variance

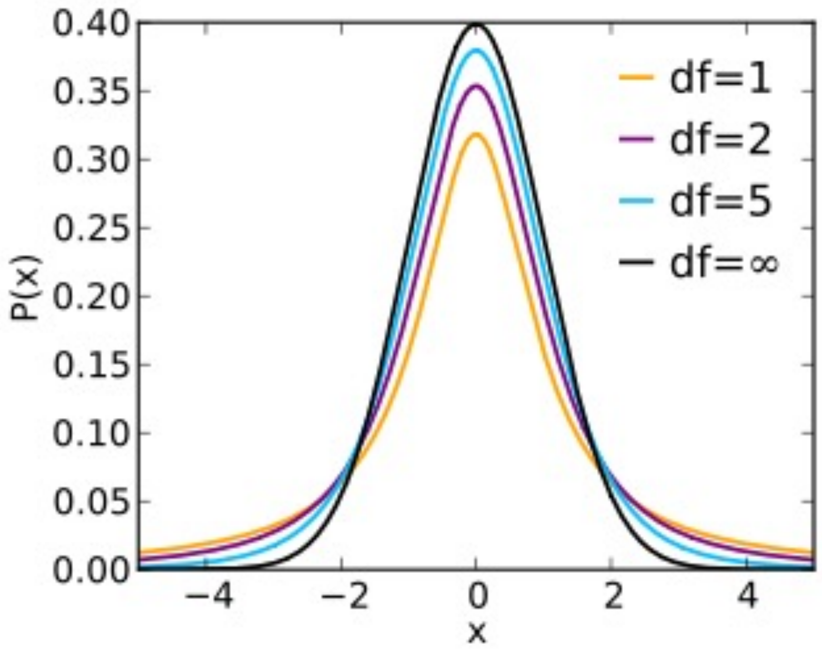
- Mean = a number which describes the center of the data
  - ‘central tendency’ also measured by mode and median
- Variance = a number which describes the data spread around the mean
  - data spread can also measured by range, quartiles & IQR

- Variance (σ^2, sigma squared) = average SSQ = SSQ/’n’
  - SD (standard deviation, σ)

T-distribution

t_{df} : parameter = degrees of freedom

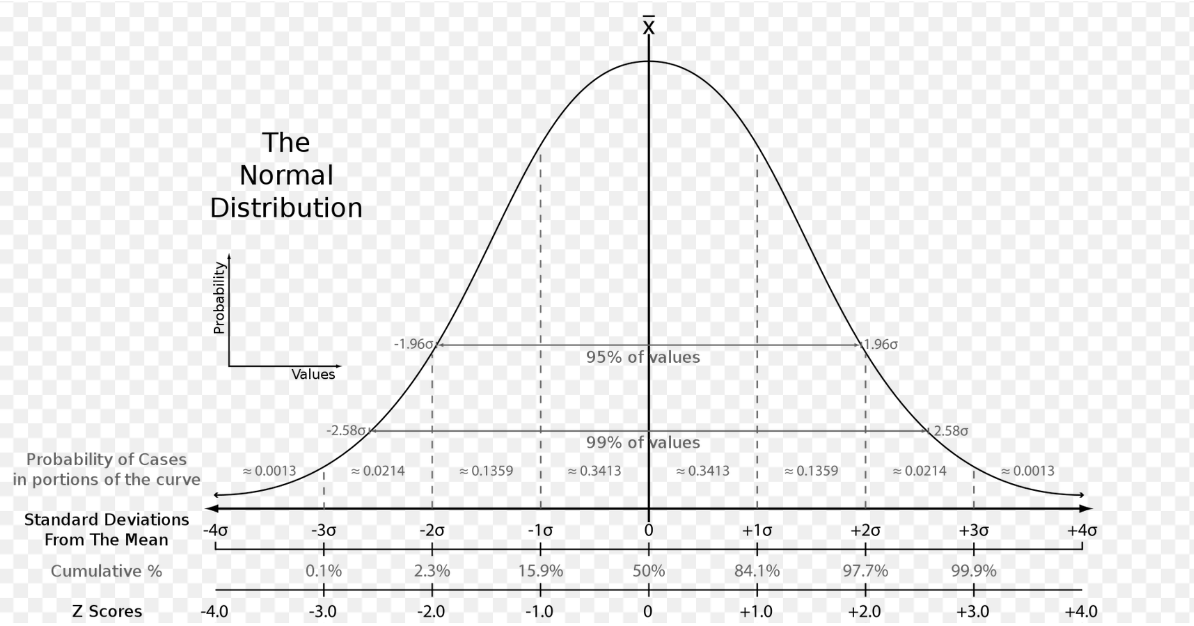
- e.g. T-tests
- A *sampling* distribution
- 'thick tails'
- $\mu = 0$
- $\sigma^2 = (df/df\text{-}2)$
- As df → ∞, P(x) → N(0,1)



wikipedia.org/

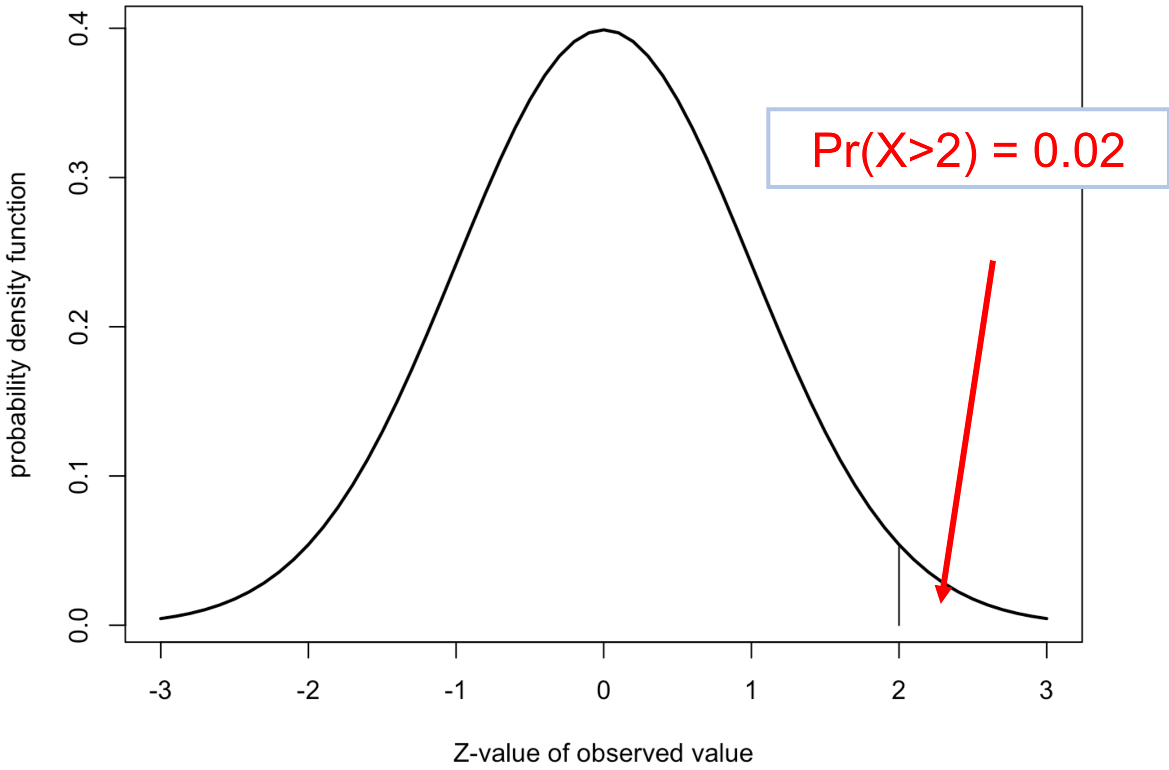
Z-scores & p-values

- Z-score
  - ➢number of standard deviations from the mean, i.e. $Z = \frac{(x-\mu)}{\sigma}$



P-value

- ➢Area under the pdf = 1
- ➢Probability of observing X (or a more extreme value than X), given a distribution is the area under the pdf

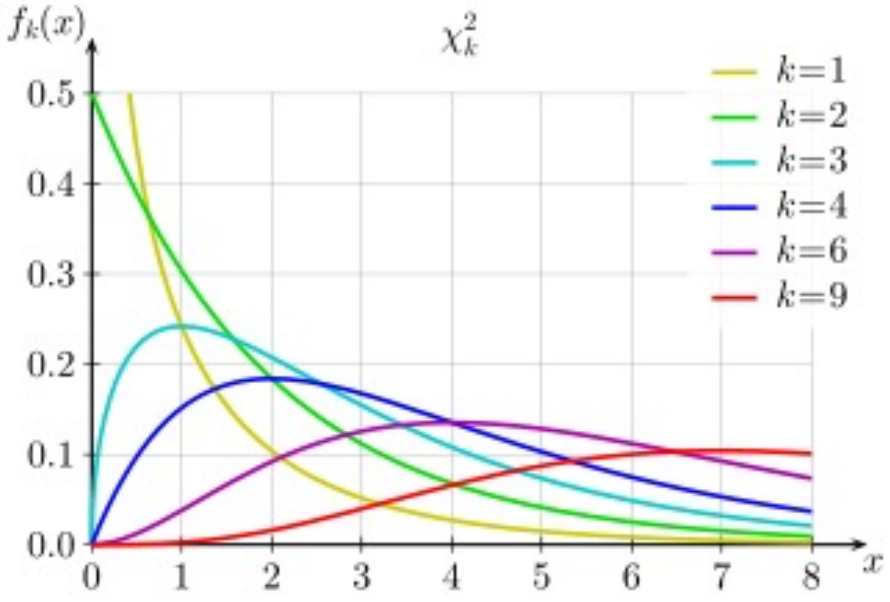


Chi-squared distribution

χ^2_{df} : parameters = degrees of freedom

- If z = N(0,1), then z^2 is χ^2_1
- 'goodness of fit' tests
- μ = df
- σ^2 = 2df
- As df → ∞,

P(x) → N(∞,2df)

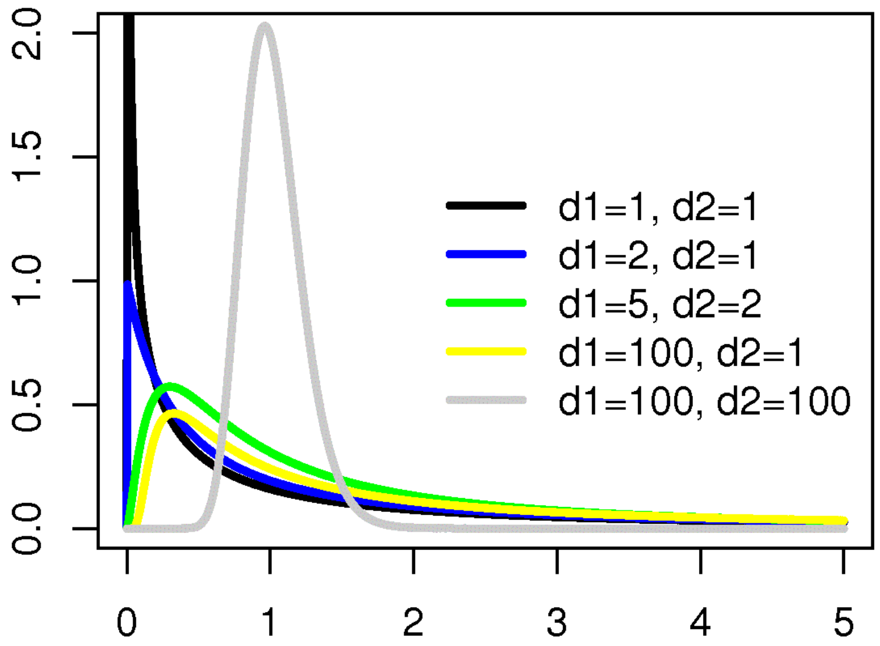


wikipedia.org/

F-distribution

$F_{df1,df1}$: parameters df1, df2

- Ratio of 2 chi-squared distributions
- e.g. used in ANOVA tests



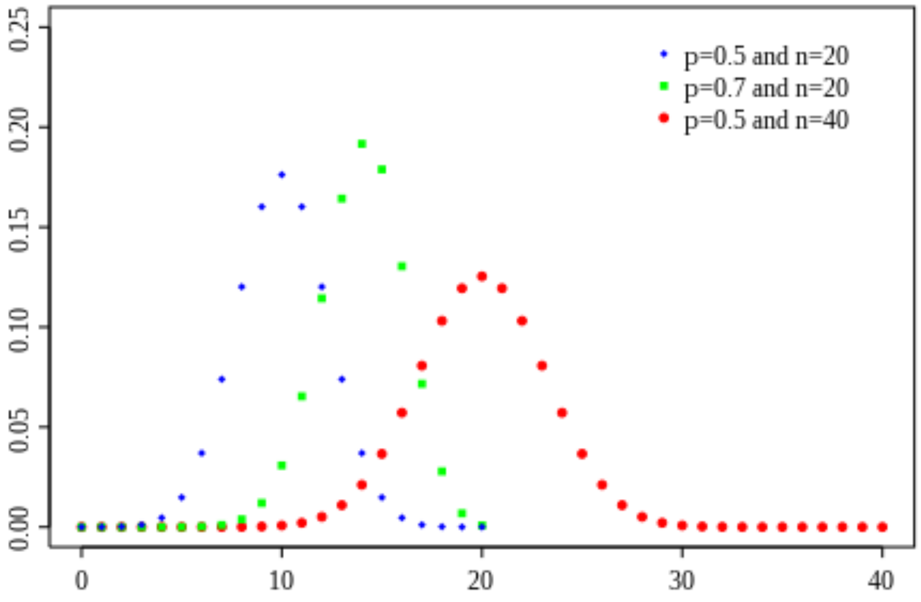
Binomial / mulitnomial

B(n,p)

parameters: n (number of trials) p (probability of success)

- discrete
- $\mu = np$
- $\sigma^2 = np(1-p)$

- Think of 2pq!



Hypothesis testing options for 2 variables

- Hypothesis testing:- statistical inference (ie reasoning) asking does the data collected support a defined scientific question

- The test to use* & the approach depends on type of data, e.g.
  - 2 categorical variables, i.e. count data (blood type & disease status)
    - ➢ Chi-squared test

  - 1 categorical variable & 1 continuous variable, (drug treatment & bp)
    - ➢ Analysis of variance (ANOVA) & t-test

  - 2 continuous variables, (height & weight)
    - ➢ correlation, regression

* Not an exhaustive list

Chi-squared test

- Count data: do we expect to observe the counts given the frequency of each category.

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

Example:

Genotype	M	MN	N	Total
Observed	233	385	129	747
Expected	242.4	366.3	138.4	
$\chi^2 = 1.96$ with 1 df => P(X>1.96) = 0.162				

H_0: Is the locus in Hardy-Weinberg Equilibrium?

$$p^2 + 2pq + q^2 = 1$$

Let freq(M) = p

p = (233*2+385)/(2*747) = 0.57

e.g. expected MN = 2*0.57*(1-0.57)*747 = 366.3

T-test

- Usually testing for mean difference between populations A and B
- Example: Is height of AFL players differ from non-football players?

	AFL	non-AFL
	190	180
	195	172
	182	185
	200	190
	201	183
	189	195
mean	192.8	184.2
SSQ	262.8	318.8
variance	52.6	63.8
n	6	6

$$SSQ = \sum (x_i - \mu_x)^2$$

$$s^2 = \frac{SSQ}{n-1}$$

$$t = \frac{estimate}{s.e.}$$

$$= \frac{estimate}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{estimate}{\sqrt{\frac{SSQ_x + SSQ_y}{n(n-1)}}}$$

$$= \frac{(192.8 - 184.2)}{\sqrt{\frac{52.6 + 63.8}{6}}}$$

= 1.97; Pr(x>1.97) = 0.07

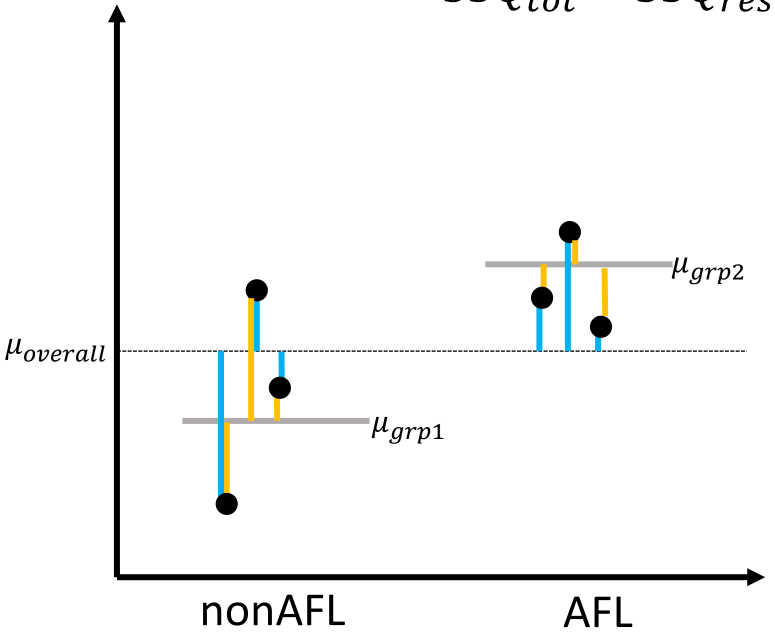
ANOVA – Analysis of Variance

- Generalization of t-test for >2 groups
- Does the variation between groups explain more variation than within groups?

	AFL	non-AFL
	190	180
	195	172
	182	185
	200	190
	201	183
	189	195
mean (μ_{grp})	192.8	184.2
SSQ	262.8	318.8
variance	52.6	63.8
n	6	6
$\mu_{overall}$	188.5	

$$SSQ_{tot} = \sum(x - \mu_{overall})^2$$
$$SSQ_{res} = \sum(x - \mu_{grp})^2$$
$$SSQ_{grp} = \sum(\mu_{grp} - \mu_{overall})^2 = SSQ_{tot} - SSQ_{res}$$

$$SSQ_{tot} = 807.0$$
$$SSQ_{res} = 581.7$$
$$SSQ_{grp} = 225.3$$

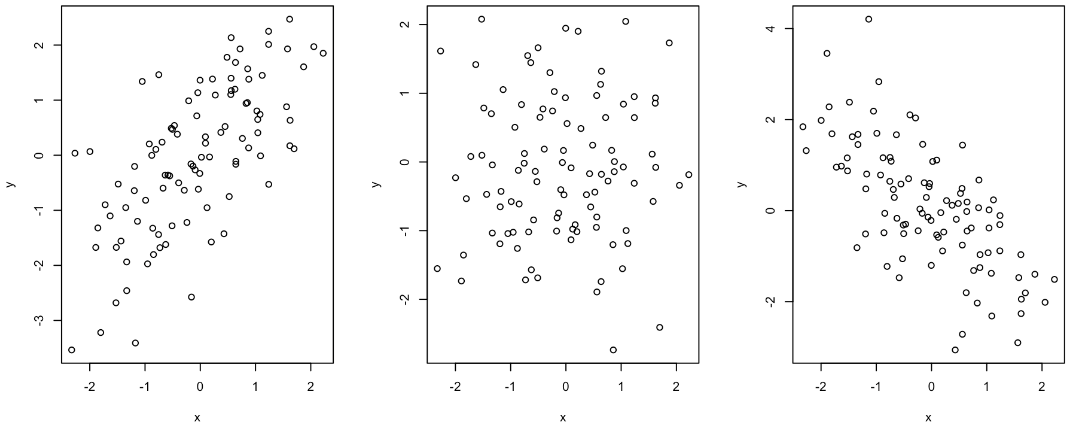


	df	SSQ	MSQ	F	P
group	1	225.3	225.3	3.87	0.07
residual	10	581.7	58.2		
total	11	807.0			

Covariance & correlation

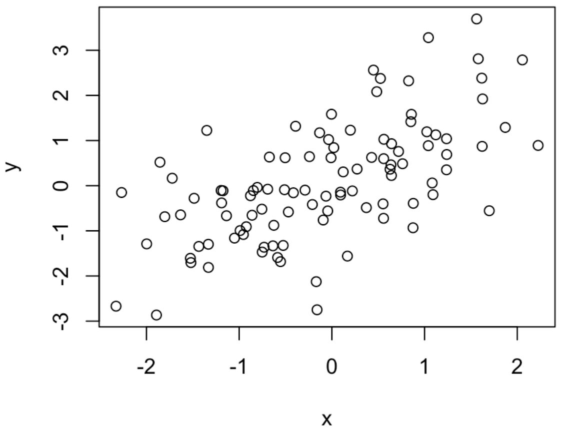
$$\sigma_x^2 = \frac{1}{N}\sum(x-\mu_x)(x-\mu_x) \qquad \sigma_{xy}^2 = \frac{1}{N}\sum(x-\mu_x)(y-\mu_y)$$

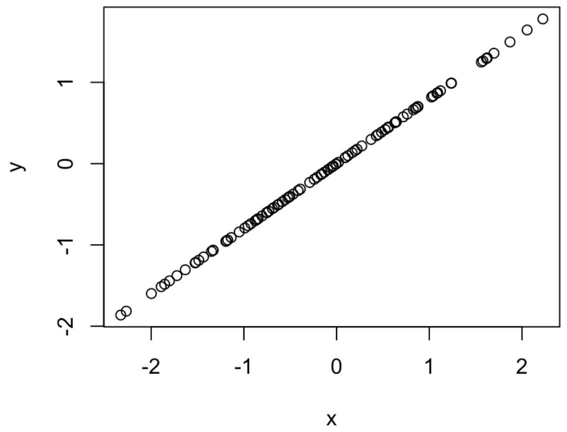
- Variance = SSQ/'n'
- Covariance describes relationship between two different variables
- Covariance = average 'sums of products'
  - ➢Scale dependent, covariance are in squared trait units (e.g. cm^2)
  - ➢Describes the relationship between 2 variables: positive, negative or no relationship

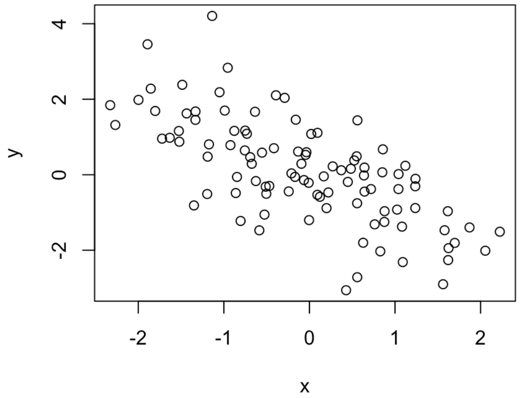


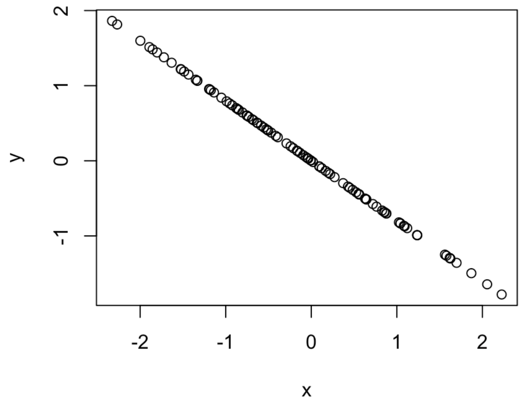
Covariance & correlation

- Covariance does not quantify the *strength* of a relationship, only if the relationship is positive, negative or null
- Correlation is a ‘standardized covariance’
  - $r_{xy} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}$,
  - values from -1 (perfect negative), 0 (no relationship) to 1 (perfect positive)









Matrix notation: mean & variance

	AFL
	190
	195
	182
	200
	201
	189
mean (μ_{grp})	192.8
SSQ	262.8
variance	52.6
n	6

- Let **X’** = [190 195 182 200 201 189]
  - **X’1**/n = 1157/6 = 192.8

Let’s correct the data for the mean...

- $\boldsymbol{X_c} = \boldsymbol{X} - \boldsymbol{1X'1/6}$ = [-2.8 2.1 -10.8 7.1 8.1 -3.8]’

SSQ...

- $\boldsymbol{X_c'}\,\boldsymbol{X_c} = [-2.8\ 2.1\ -10.8\ 7.1\ 8.1\ -3.8]\begin{bmatrix}\mathbf{-2.8}\\ \mathbf{2.1}\\ \mathbf{-10.8}\\ \mathbf{7.1}\\ \mathbf{8.1}\\ \mathbf{-3.8}\end{bmatrix}$ = [262.8]

Variance...

$\boldsymbol{X_c'}\,\boldsymbol{X_c}$/5 = 262.3/5 = 52.6

Matrix notation: variance/covariance matrix

x	y
76.0	61.2
72.6	57.9
74.6	59.2
75.8	60.6
74.5	62.0
74.7	**60.1**

- Let's assume $\boldsymbol{X}$ is the 5 x 2 mean corrected data matrix
- Then,

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} SSQ_x & SSQ_{xy} \\ SSQ_{xy} & SSQ_y \end{bmatrix}$$

$$= \begin{bmatrix} 1.3 & -2.1 & -0.1 & 1.1 & -0.2 \\ 1.1 & -2.2 & -0.9 & 0.5 & 1.9 \end{bmatrix} \begin{bmatrix} 1.3 & 1.1 \\ -2.1 & -2.2 \\ -0.1 & -0.9 \\ 1.1 & 0.5 \\ -0.2 & 1.9 \end{bmatrix}$$

$$= \begin{bmatrix} 7.4 & 6.3 \\ 6.3 & 10.8 \end{bmatrix}$$

Matrix notation: variance/covariance matrix

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} SSQ_x & SSQ_{xy} \\ SSQ_{xy} & SSQ_y \end{bmatrix}$$

$$\frac{\mathbf{1}}{\boldsymbol{n}-\mathbf{1}}\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy}^2 \\ \sigma_{xy}^2 & \sigma_x^2 \end{bmatrix}$$

$$= \frac{\mathbf{1}}{\mathbf{4}} \begin{bmatrix} 7.4 & 6.3 \\ 6.3 & 10.8 \end{bmatrix} = \begin{bmatrix} 1.8 & 1.6 \\ 1.6 & 2.7 \end{bmatrix}$$

- If you encounter $\frac{\mathbf{1}}{\boldsymbol{n}-\mathbf{1}}\boldsymbol{X}'\boldsymbol{X}$ of a mean-corrected matrix then it's a variance-covariance matrix

Matrix notation: variance/covariance matrix

- Let $\boldsymbol{X_s}$ be a 5 x 2 matrix of data where each column is ~ N(0,1).
  - i.e. data in column *j* are standardized to a z-score, $\boldsymbol{x_{s(ij)}} = (\boldsymbol{x_{ij}} - \boldsymbol{\mu_j})/\boldsymbol{\sigma_j}$
- What is $\frac{\mathbf{1}}{\boldsymbol{n}-\mathbf{1}}\boldsymbol{X_s'X_s}$?

$$\frac{\mathbf{1}}{\boldsymbol{n}-\mathbf{1}}\boldsymbol{X_s'X_s} = \frac{\mathbf{1}}{\mathbf{4}}\begin{bmatrix} 0.96 & -1.55 & -0.07 & 0.81 & -0.15 \\ 0.62 & -1.39 & -0.60 & 0.26 & 1.11 \end{bmatrix}\begin{bmatrix} 0.96 & 0.62 \\ -1.55 & -1.39 \\ -0.07 & -0.60 \\ 0.81 & 0.26 \\ -0.15 & 1.11 \end{bmatrix}$$

$$= \begin{bmatrix} 1.0 & 0.71 \\ 0.71 & 1.0 \end{bmatrix}$$

basics – practical 2

- Work through “basicsPrac2.pdf”
  - Q1: z-scores and normal distribution
  - Q2: chi-square test
  - Q3: t-test
  - Q4: ANOVA
  - Q5: normal probability density function

- Software: R