

G&G Winter School

Distributions, hypothesis tests, variance & covariance, correlation

Kathryn Kemper

Data in statistics

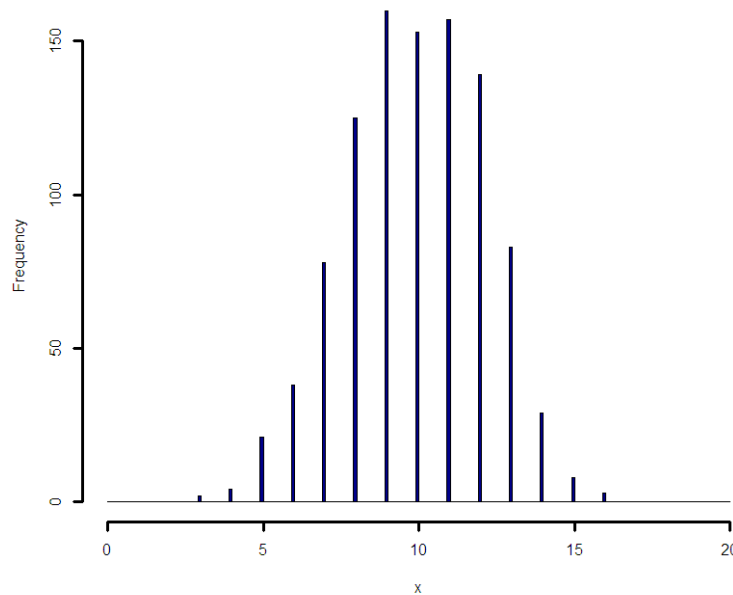
- Statisticians will refer to a vector X (or Y , X_1 etc.) of “random variables”
 - e.g. a coin toss: H, T, H, T, T, T, H
 - e.g. a sample of heights 164.1, 173.2, 150.0
- We normally assume X to be from a distribution. This distribution defines the properties of the *random variable* (e.g the probability that $x < 0$ or $x = 5$)

For example, toss a coin

Toss a coin 20 times, count the number of heads

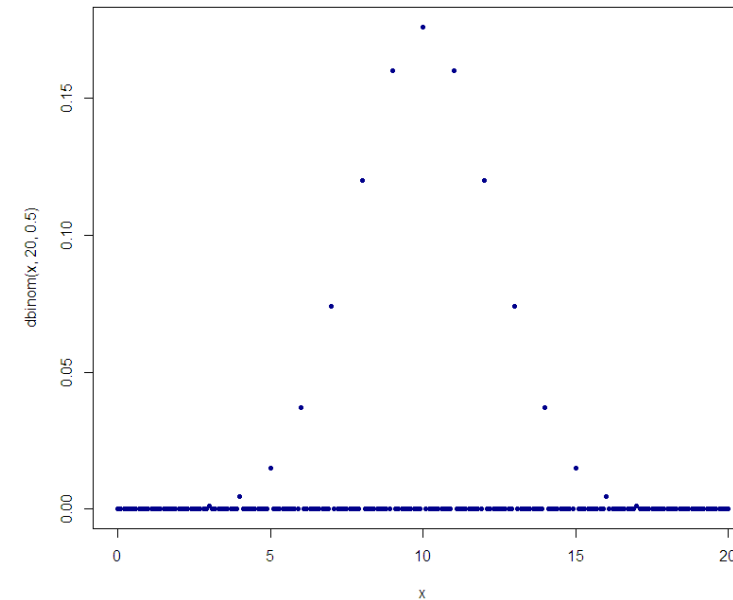
e.g. 7H, 11H, 10H, 19H, 12H, 9H, 6H, 10H, 9H....

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 estimates 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

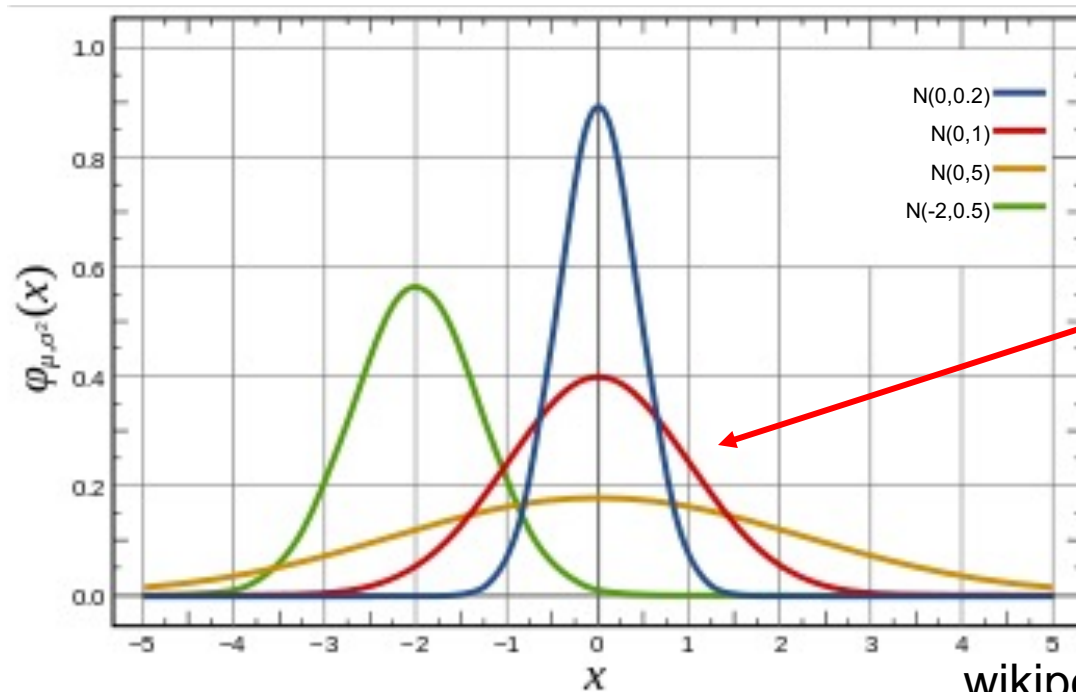
Normal distribution:

$N(\mu, \sigma^2)$: parameters = μ, σ^2

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

centre

$$\left[\sigma^2 = \frac{1}{N} \sum (x_i - \mu)^2 \right] \quad \text{spread}$$



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

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

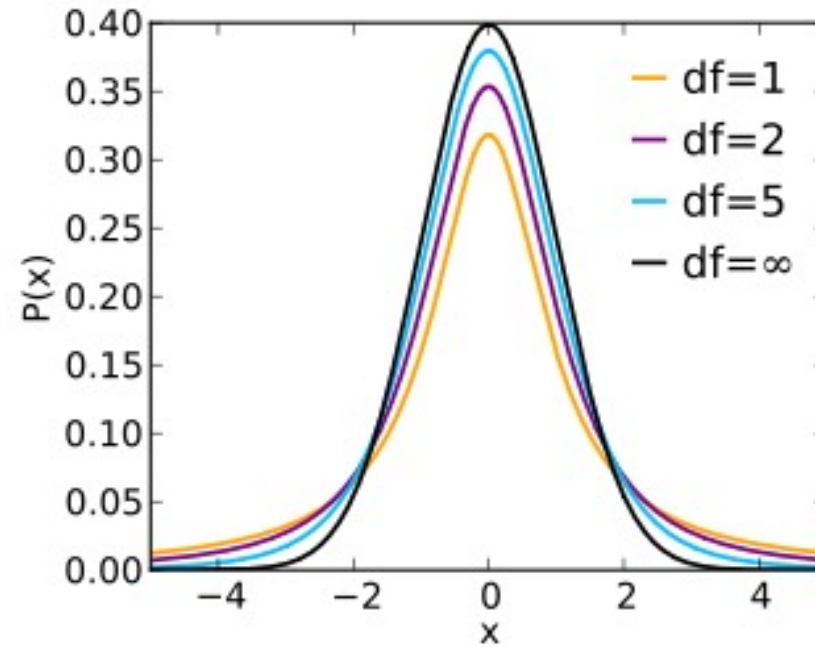
Mean & Variance

- 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'
- $\mu = 0$
- $\sigma^2 = (df/df-2)$
- As $df \rightarrow \infty$, $P(x) \rightarrow N(0,1)$

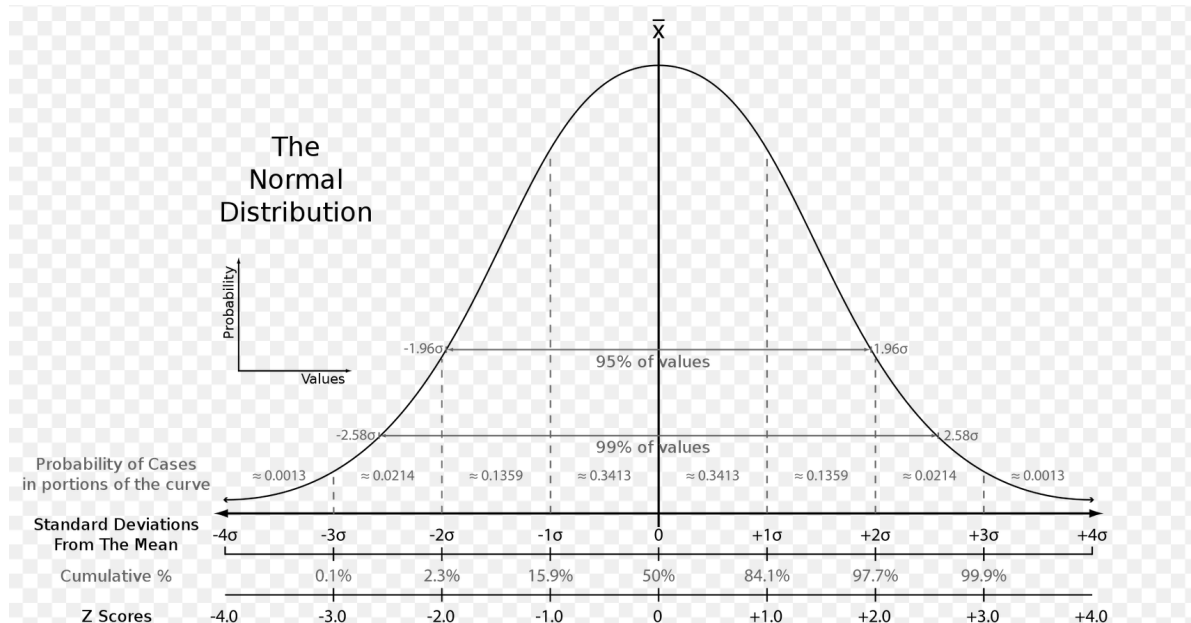


[wikipedia.org/](https://en.wikipedia.org/wiki/Student%27s_t_distribution)

Z-scores & p-values

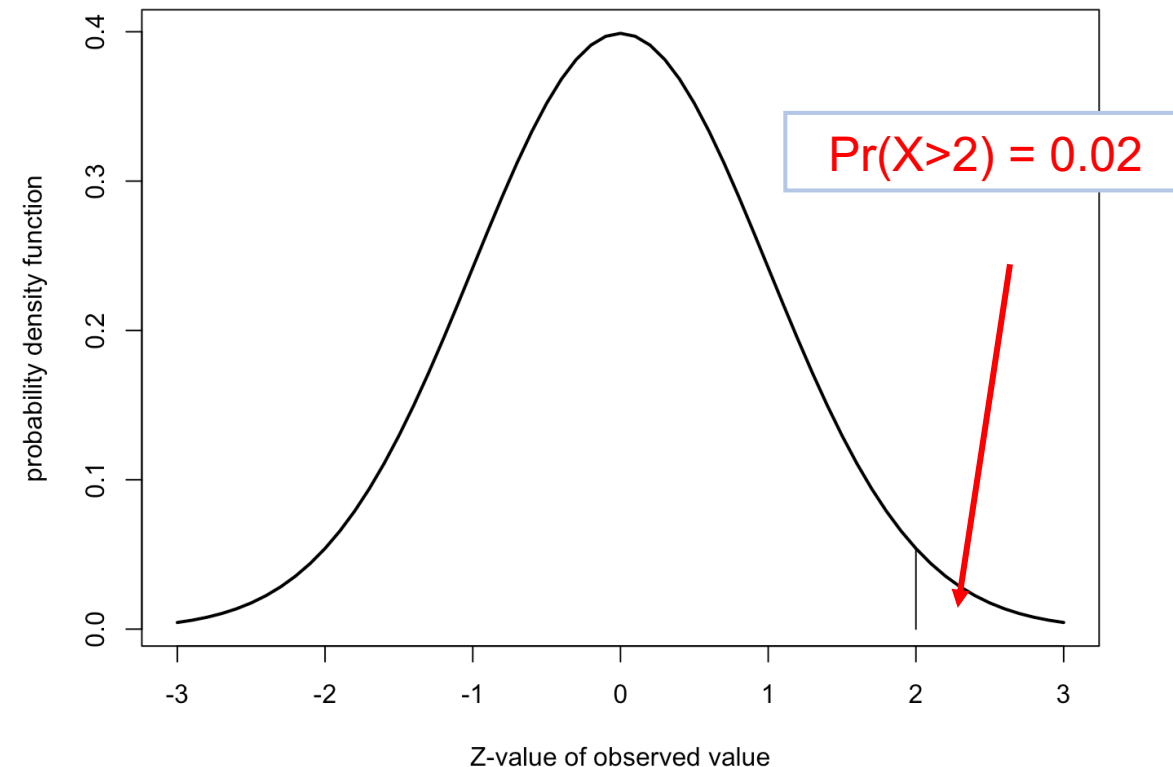
- Z-score

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



P-value

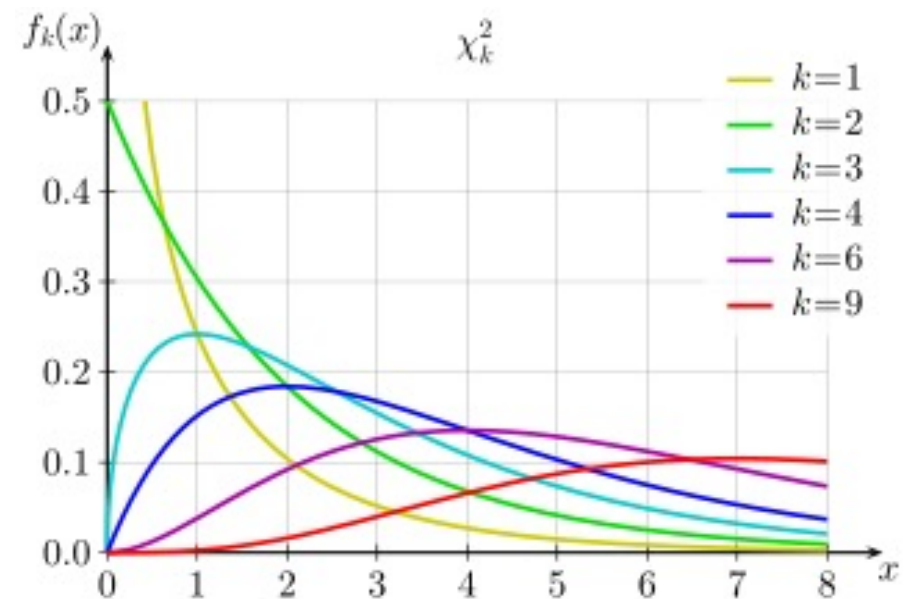
- Probability of observing X (or a more extreme value than X), given an assumed null distribution
- How likely is the observation by chance?



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
- $\mu = df$
- $\sigma^2 = 2df$
- As $df \rightarrow \infty$,
 $P(x) \rightarrow N(\infty, 2df)$

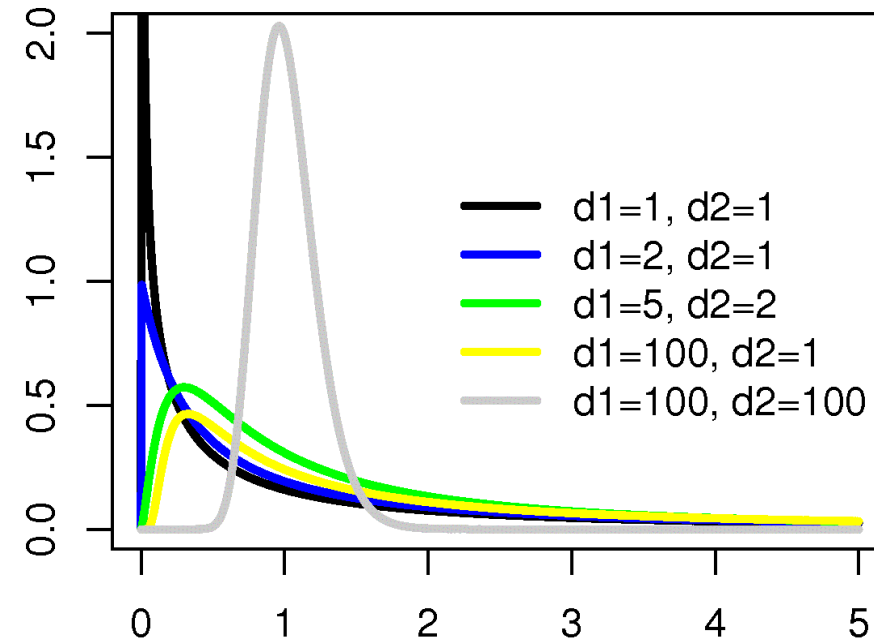


wikipedia.org/

F-distribution

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

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

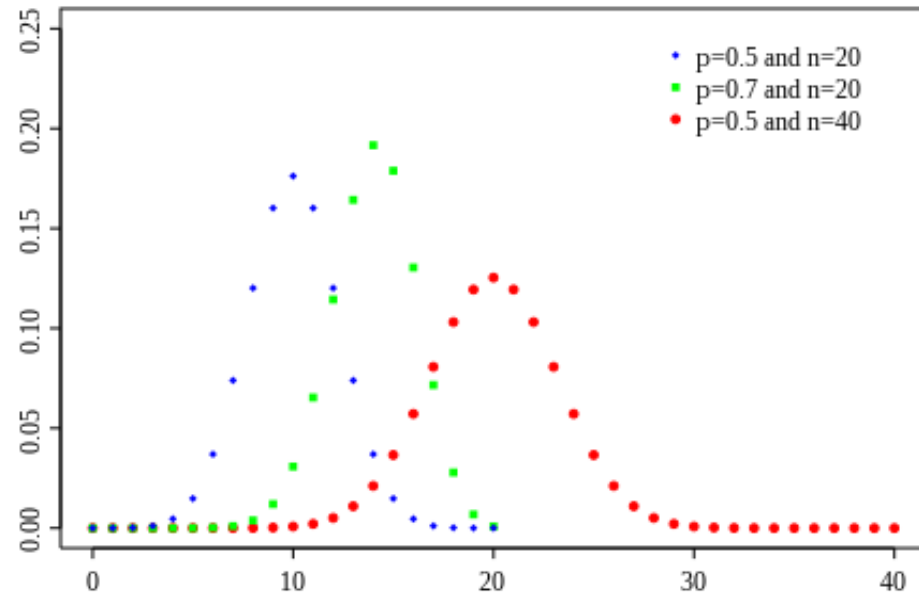


Binomial / multinomial

$B(n,p)$

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.

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

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 $\Rightarrow P(X > 1.96) = 0.162$				

H₀: Is the locus in Hardy-Weinberg Equilibrium?

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

Let freq(M) = p

$$p = (233 \cdot 2 + 385) / (2 \cdot 747) = 0.57$$

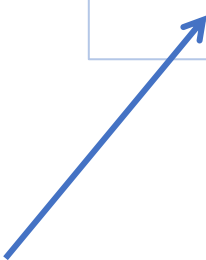
$$\text{e.g. expected MN} = 2 \cdot 0.57 \cdot (1 - 0.57) \cdot 747 = 366.3$$

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

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

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


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

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

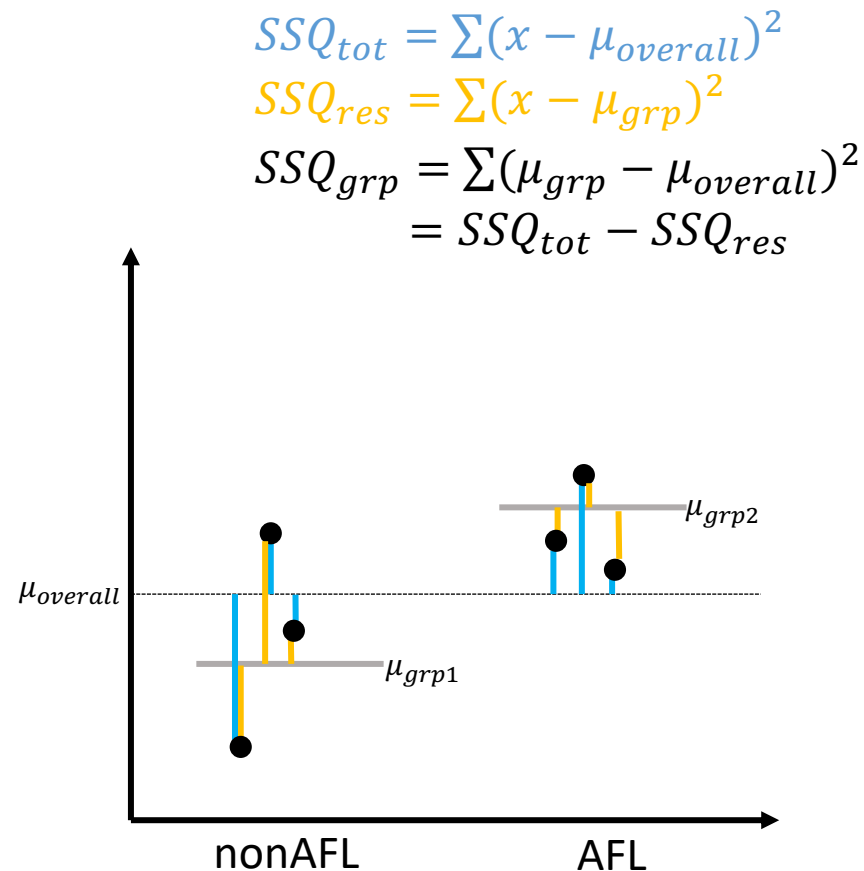
$$= \frac{(192.8 - 184.2)}{\sqrt{\frac{52.6 + 63.8}{6}}}$$

$$= 1.97; \Pr(x > 1.97) = 0.07$$

ANOVA – Analysis of Variance

- Does the mean of the group explain more variance than chance?

	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	



$$SSQ_{tot} = 807.0$$

$$SSQ_{res} = 581.7$$

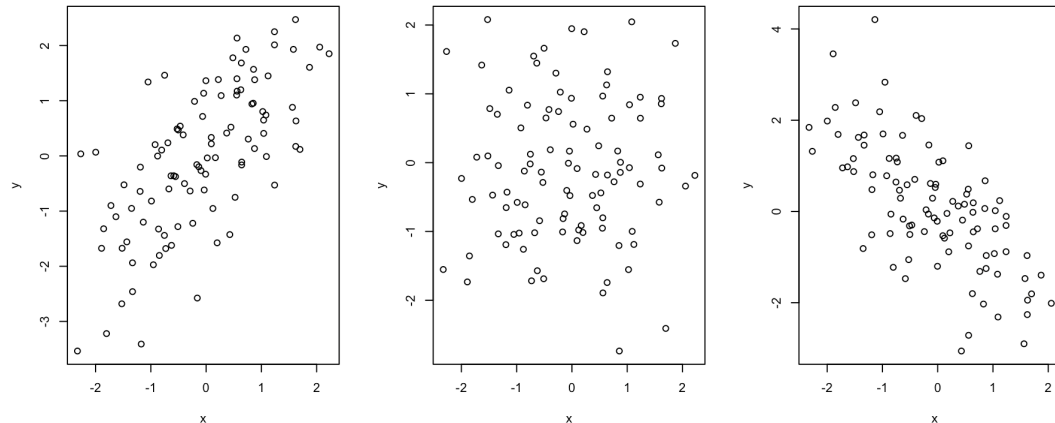
$$SSQ_{grp} = 225.3$$

	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 (cm²)
 - Gives relationship b/w 2 variables: positive, negative or no relationship

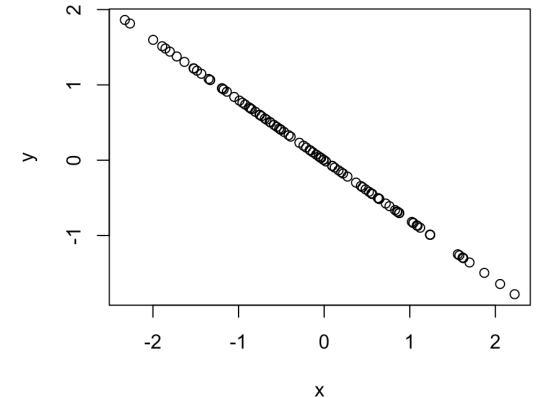
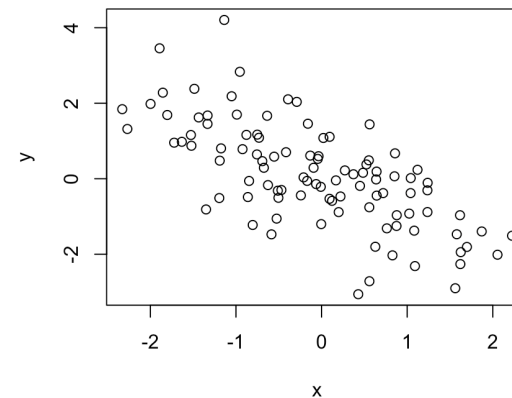
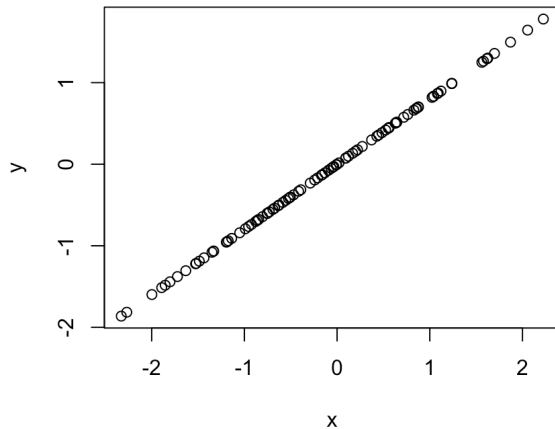
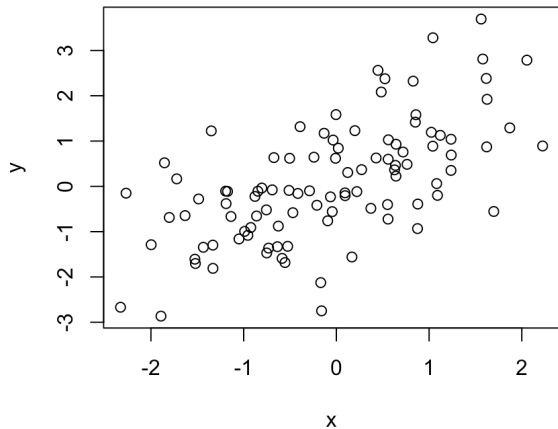


Covariance & correlation

- Covariance does not quantify the *strength* of a relationship, only if the relationship is positive, negative or that there is no relationship
- Correlation is a 'standardized covariance'

$$\rightrightarrows r_{xy} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y},$$

➤ 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 $\mathbf{X}' = [190 \ 195 \ 182 \ 200 \ 201 \ 189]$

➤ $\mathbf{X}'\mathbf{1}/n = 1157/6 = 192.8$

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

➤ $\mathbf{X}_c = \mathbf{X} - \mathbf{1}\mathbf{X}'\mathbf{1}/6 = [-2.8 \ 2.1 \ -10.8 \ 7.1 \ 8.1 \ -3.8]'$

SSQ...

➤ $\mathbf{X}_c' \mathbf{X}_c = [-2.8 \ 2.1 \ -10.8 \ 7.1 \ 8.1 \ -3.8] \begin{bmatrix} -2.8 \\ 2.1 \\ -10.8 \\ 7.1 \\ 8.1 \\ -3.8 \end{bmatrix} = [262.8]$

Variance...

➤ $\mathbf{X}_c' \mathbf{X}_c / 5 = 262.3/5 = 52.6$

Matrix notation: variance/covariance matrix

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

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

$$\begin{aligned}
 \mathbf{X}'\mathbf{X} &= \begin{bmatrix} SSQ_x & SSQ_{xy} \\ SSQ_{xy} & SSQ_y \end{bmatrix} \\
 &= \begin{bmatrix} 1.3 & -2.1 & -0.1 & 1.1 & -0.2 \\ 1.1 & -2.2 & -0.9 & 0.5 & 1.9 \end{bmatrix} \begin{bmatrix} 1.3 & 1.1 \\ -2.1 & -2.2 \\ -0.1 & -0.9 \\ 1.1 & 0.5 \\ -0.2 & 1.9 \end{bmatrix} \\
 &= \begin{bmatrix} 7.4 & 6.3 \\ 6.3 & 10.8 \end{bmatrix}
 \end{aligned}$$

Matrix notation: variance/covariance matrix

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

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

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

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

Matrix notation: variance/covariance matrix

- Let \mathbf{X}_s be a 5 x 2 matrix of data where each column is $\sim N(0,1)$.
 - i.e. data in column j are standardized to a z-score, $\mathbf{x}_{s(ij)} = (\mathbf{x}_{ij} - \boldsymbol{\mu}_j)/\boldsymbol{\sigma}_j$
- What is $\frac{1}{n-1} \mathbf{X}'_s \mathbf{X}_s$?

$$\begin{aligned} \frac{1}{n-1} \mathbf{X}'_s \mathbf{X}_s &= \frac{1}{4} \begin{bmatrix} 0.96 & -1.55 & -0.07 & 0.81 & -0.15 \\ 0.62 & -1.39 & -0.60 & 0.26 & 1.11 \end{bmatrix} \begin{bmatrix} 0.96 & 0.62 \\ -1.55 & -1.39 \\ -0.07 & -0.60 \\ 0.81 & 0.26 \\ -0.15 & 1.11 \end{bmatrix} \\ &= \begin{bmatrix} 1.0 & 0.71 \\ 0.71 & 1.0 \end{bmatrix} \end{aligned}$$