Simple linear models:

t-tests and linear regression

What is this model?

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

Outline

- Components of a linear model
  - Stochastic parts of linear models: distributions
  - Deterministic parts of linear models: linear predictor and design matrix
- T-test: equal and unequal variance
- Linear regression

Linear models

- Response = deterministic part + stochastic part
  - Stochastic = random
  - Deterministic = systematic
- Linear models are so called because the expected response can be treated as the results of explanatory variables whose effects are additive

Linear models: stochastic part

Parametric statistical models -> probability distributions

- Types of responses
  - Binary (heads/tails; dead/survived)
  - Categorical (nationality; geographic location)
  - Counts (number of birds in a quadrant)
  - Continuous (body mass; wing length)
- Type of response will dictate the distribution

Linear models: stochastic part

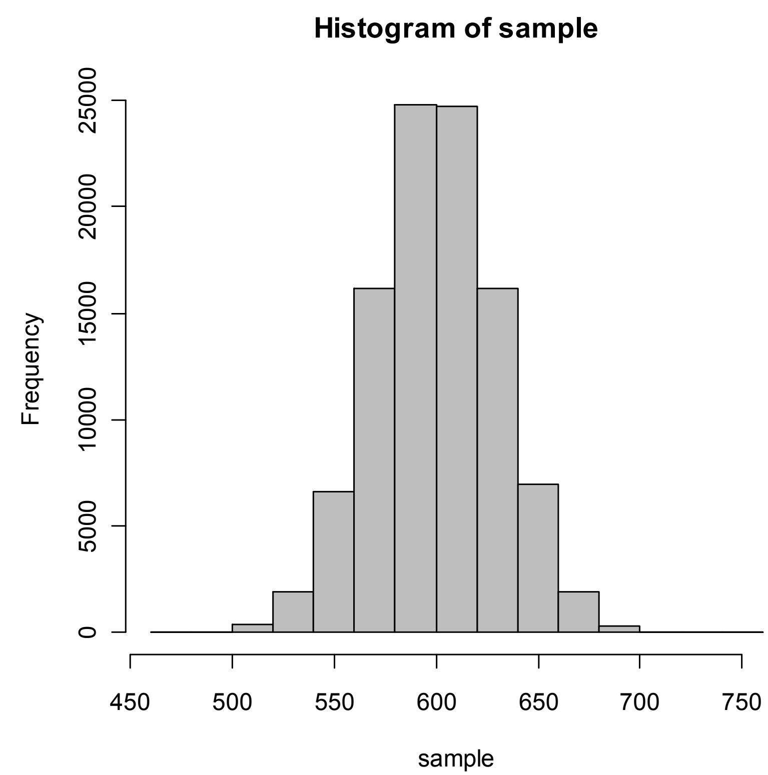
Normal Distribution

One continuous distributions that we need to understand for linear models

- Denoted $N(\mu, \sigma^2)$
- Mean: $\mu \in \boldsymbol{R}$
- Variance: $\sigma^2 > 0$
- Support: $\mathrm{x} \in \boldsymbol{R}$

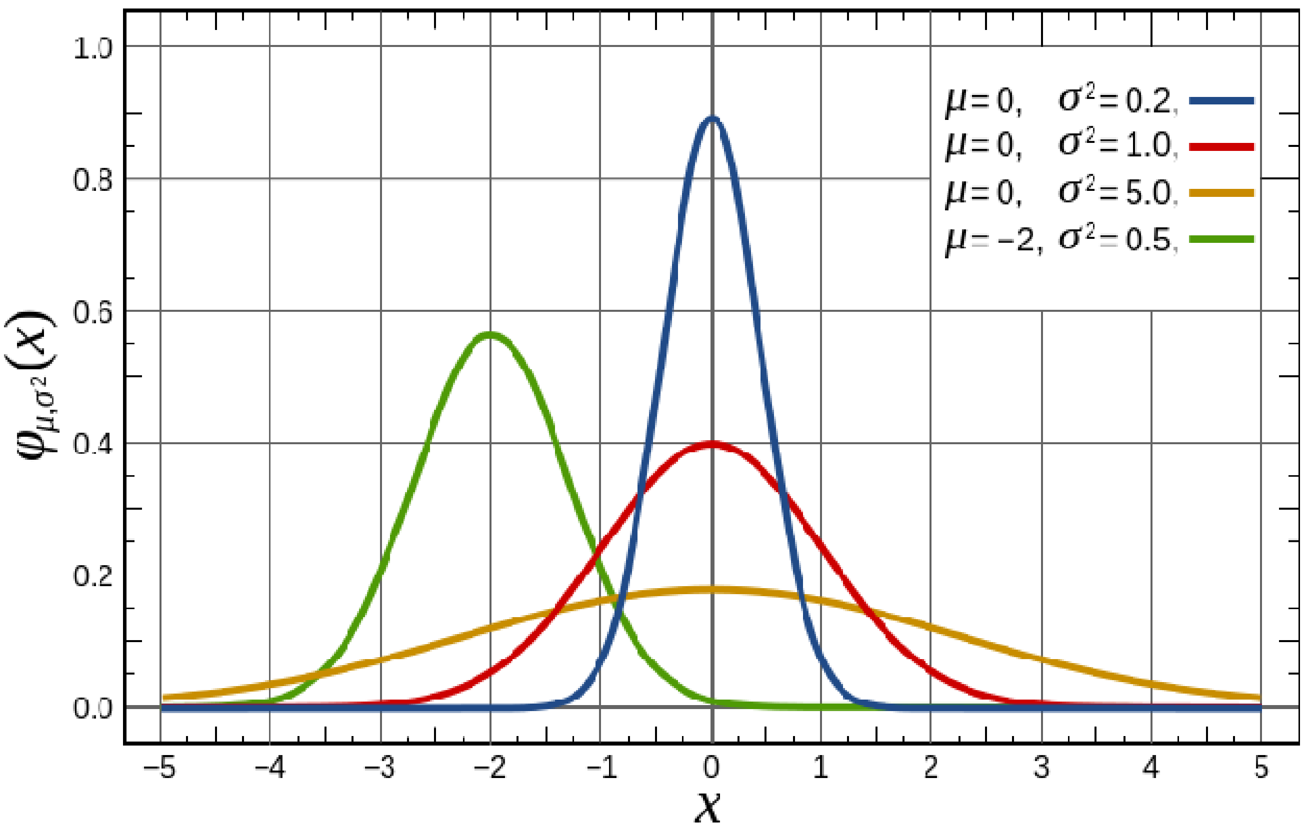
Linear models: stochastic part
Normal Distribution

- *Sampling situation:* Measurements that are affected by a large number of effects that act in an additive way.
- *Classical examples:*
  (1) Body size and other linear measurements on organisms
  (2) Density of species across space
- *Why it's useful:* The Central Limit Theorem – the mean of many RVs independently drawn from some distribution are approximately normal



Linear models: stochastic part
Normal Distribution

Probability density function: $$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Linear models: stochastic part

So called "general linear models" are defined by one feature:

- The assumed error around the deterministic portion of the model is normally distributed

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

Linear models: deterministic part

Design matrix – a matrix of explanatory variables

- For each element of the response vector, the design matrix provides a 0/1 index for which effect is present for categorical (= discrete) explanatory variables and for what "amount" of an effect is present in the case of continuous explanatory variables.

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- The design matrix contains:

  as many columns as the fitted model has parameters

  as many rows as there are data points

- When matrix-multiplied with the parameter vector, yields the *linear predictor*, another vector.

- The linear predictor contains the expected value of the response, on the link scale, given the values of all explanatory variables in the model.

Linear models: deterministic part

Design matrix – example with a t-test

Suppose we have a single, binary explanatory variable (region = North or South) on a continuous response (mass).
Question: Is mass different for organism A in the north than in the south?

Assume 6 data points:

Individual	Location	Mass
1	North	6
2	North	8
3	North	5
4	North	7
5	South	9
6	South	9

Linear models: deterministic part

Design matrix – example with a t-test

- Suppose we have a single, binary explanatory variable (region = North or South) on a continuous response (mass).

$$mass_i = \alpha + \beta * region_i + \varepsilon_i$$
$$\varepsilon_i \sim Norm(0, \sigma^2)$$

- This means that the mass of a snake is made up of the sum of three components: a constant (alpha), the product of another constant (beta) with the value of the indicator for region in which snake was caught plus a third term (ε_i) that is specific to snake.
- Another way to write this model is:

$$mass_i \sim Normal\ (\alpha + \beta * region_i, \sigma^2)$$

Linear models: deterministic part

Design matrix – what does the variable region look like?

$$mass_i \sim Normal\ (\alpha + \beta * region_i, \sigma^2)$$

lm(mass ~ region)
model.matrix(~region)

	(Intercept)	region2
1	1	0
2	1	0
3	1	0
4	1	0
5	1	1
6	1	1

Translates into a set of equations:

$$6 = \alpha * 1 + \beta * 0 + \varepsilon_1$$
$$8 = \alpha * 1 + \beta * 0 + \varepsilon_2$$
$$5 = \alpha * 1 + \beta * 0 + \varepsilon_3$$
$$7 = \alpha * 1 + \beta * 0 + \varepsilon_4$$
$$9 = \alpha * 1 + \beta * 1 + \varepsilon_5$$
$$9 = \alpha * 1 + \beta * 1 + \varepsilon_6$$

Or in matrix notation:

$$\begin{pmatrix} 6 \\ 8 \\ 5 \\ 7 \\ 9 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} * \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

Linear models: deterministic part

Design matrix – what does the variable region look like?

lm(mass ~ region)
model.matrix(~region)

	(Intercept)	region2
1	1	0
2	1	0
3	1	0
4	1	0
5	1	1
6	1	1

Translates into a set of equations:

$$6 = \alpha * 1 + \beta * 0 + \varepsilon_1$$
$$8 = \alpha * 1 + \beta * 0 + \varepsilon_2$$
$$5 = \alpha * 1 + \beta * 0 + \varepsilon_3$$
$$7 = \alpha * 1 + \beta * 0 + \varepsilon_4$$
$$9 = \alpha * 1 + \beta * 1 + \varepsilon_5$$
$$9 = \alpha * 1 + \beta * 1 + \varepsilon_6$$

Or in matrix notation:

$$\begin{pmatrix} 6 \\ 8 \\ 5 \\ 7 \\ 9 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} * \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

What is the interpretation of these parameters?

Linear models: deterministic part

Design matrix – what does the variable region look like?

lm(mass ~ region)
model.matrix(~region)

	(Intercept)	region2
1	1	0
2	1	0
3	1	0
4	1	0
5	1	1
6	1	1

Translates into a set of equations:

$$6 = \alpha * 1 + \beta * 0 + \varepsilon_1$$
$$8 = \alpha * 1 + \beta * 0 + \varepsilon_2$$
$$5 = \alpha * 1 + \beta * 0 + \varepsilon_3$$
$$7 = \alpha * 1 + \beta * 0 + \varepsilon_4$$
$$9 = \alpha * 1 + \beta * 1 + \varepsilon_5$$
$$9 = \alpha * 1 + \beta * 1 + \varepsilon_6$$

Or in matrix notation:

$$\begin{pmatrix} 6 \\ 8 \\ 5 \\ 7 \\ 9 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} * \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

This is an ***effects parameterization*** of the t-test

Linear models: deterministic part

What if we re-parameterize this model?

$$\begin{pmatrix} 6 \\ 8 \\ 5 \\ 7 \\ 9 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} * \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

What is the interpretation of the parameters now?

Linear models: deterministic part

What if we re-parameterize this model?

$$\begin{pmatrix} 6 \\ 8 \\ 5 \\ 7 \\ 9 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} * \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

This is an ***means parameterization*** of the t-test

Linear models: deterministic part

Design matrix – switching to a linear regression

Suppose now we have a continuous explanatory variable (amount of land cover) on a continuous response (mass).
Question: Is there a relationship between the amount of land cover and mass for organism A?

Assume 6 data points:

Individual	Land cover	Mass
1	20	6
2	21	8
3	20	5
4	22	7
5	24	9
6	22	9

Linear models: deterministic part

Design matrix – linear regression

- Continuous explanatory variable (landcover) on a continuous response (mass).

$$mass_i = \alpha + \beta * landcover_i + \varepsilon_i$$
$$\varepsilon_i \sim Norm(0, \sigma^2)$$

- This means that the mass of a snake is made up of the sum of three components: a constant (alpha), the product of another constant (beta) with the value land cover where the snake was caught plus a third term (ε_i) that is specific to snake.
- Another way to write this model is:

$$mass_i \sim Normal\ (\alpha + \beta * landcover_i, \sigma^2)$$

Linear models: deterministic part

Design matrix – what does the variable region look like?

$$mass_i \sim Normal\,(\alpha + \beta * landcover_i, \sigma^2)$$

lm(mass ~ landcover)
model.matrix(~landcover)

	(Intercept)	
1	1	20
2	1	21
3	1	20
4	1	22
5	1	24
6	1	22

Translates into a set of equations:

$$6 = \alpha * 1 + \beta * 20 + \varepsilon_1$$
$$8 = \alpha * 1 + \beta * 21 + \varepsilon_2$$
$$5 = \alpha * 1 + \beta * 20 + \varepsilon_3$$
$$7 = \alpha * 1 + \beta * 22 + \varepsilon_4$$
$$9 = \alpha * 1 + \beta * 24 + \varepsilon_5$$
$$9 = \alpha * 1 + \beta * 22 + \varepsilon_6$$

Or in matrix notation:

$$\begin{pmatrix} 6 \\ 8 \\ 5 \\ 7 \\ 9 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 & 20 \\ 1 & 21 \\ 1 & 20 \\ 1 & 22 \\ 1 & 24 \\ 1 & 22 \end{pmatrix} * \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

Linear models: deterministic part

Design matrix – what does the variable region look like?

lm(mass ~ landcover)
model.matrix(~landcover)

	(Intercept)	
1	1	20
2	1	21
3	1	20
4	1	22
5	1	24
6	1	22

Translates into a set of equations:

$$6 = \alpha * 1 + \beta * 20 + \varepsilon_1$$
$$8 = \alpha * 1 + \beta * 21 + \varepsilon_2$$
$$5 = \alpha * 1 + \beta * 20 + \varepsilon_3$$
$$7 = \alpha * 1 + \beta * 22 + \varepsilon_4$$
$$9 = \alpha * 1 + \beta * 24 + \varepsilon_5$$
$$9 = \alpha * 1 + \beta * 22 + \varepsilon_6$$

Or in matrix notation:

$$\begin{pmatrix} 6 \\ 8 \\ 5 \\ 7 \\ 9 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 & 20 \\ 1 & 21 \\ 1 & 20 \\ 1 & 22 \\ 1 & 24 \\ 1 & 22 \end{pmatrix} * \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

Is there means or effects parameterization with a continuous predictor?

Lab

Linear regression

-Pull up the linear regression R code and we will work through it

T-test: equal and unequal variances

-R script for equal variance

-Homework 3 is to modify the t-test to account for unequal variances in groups (due Oct 13 at midnight)

Note! Quiz next time on the last two lectures. Bring a piece of paper!