# An introduction to statistical inference

Ben Lambert[1]
ben.c.lambert@gmail.com

[1]University of Oxford

Tuesday 6th July, 2021

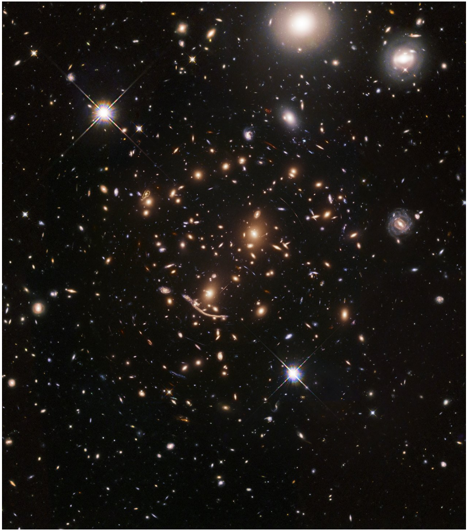
# Course plan

- 9.30am-10.15am: lecture, "An introduction to statistical inference"
- 10.45am-1pm: practical

# Outline

# What is the aim of scientific inquiry?

Understand how the universe works.



# What does it mean to understand the universe?

Understand causality.

# How do we test our understanding?

Make predictions. Compare reality with our predictions.

# What can statistical inference help with?

- Generating understanding from data
- Testing our scientific hypotheses versus data

# But why do we need statistics?

- The universe is complex
- Its mechanisms are not directly observable
- Our data contain information both about the mechanisms and other nuisance factors

Statistics provides a way to separate the signal of the mechanisms from the noise.

# How statistics separates signal from noise?

$$\text{observations} = \text{signal} + \text{noise} \tag{1}$$

- Signal contains our interesting scientific mechanism
- Noise contains a bunch of things not of interest

Since we do not know or observe the exact noise processes, in statistics, it is assumed that the noise is represented as being *random*.

But random does not mean unstructured. In statistics, making assumptions about the nature of the random process allows us to bound its influence on the observed data.

# Example 1: flipping a coin

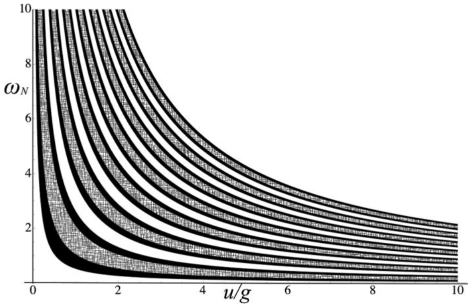
Suppose we flip a coin twice. We could obtain:

- Two tails
- One head; one tail
- Two heads

Why can we get different outcomes each time the coin is flipped?

# Different initial conditions

Precessional frequency: $\omega_N$; Magnitude of upward velocity, $u$.[1]




White indicates heads; hatched indicates lands on sides; hatched indicates tails.

[1] From Probability, geometry, and dynamics in the toss of a thick coin, Yong and Mahadevan (2011)

# Coin flip dynamics: physics approach

Solve complex equations of motion (making assumptions about flipping process). One part of the system:

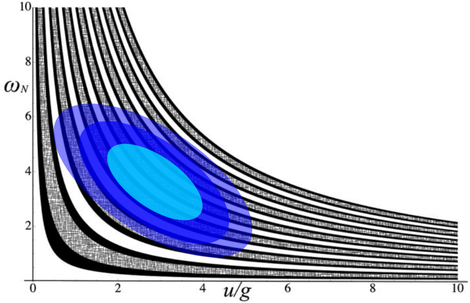
$$\frac{d\boldsymbol{N}}{dt} = \Omega \times \boldsymbol{N} \tag{2}$$

This system determines outcome given a set of initial conditions: precessional frequency and magnitude of upward velocity.

But we still don't know how different people throw a coin, so we'd need to measure this and likely represent this using randomness!

# Coin flip dynamics: statistical approach

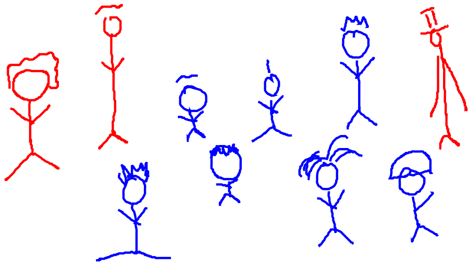
Assume outcome of a coin flip is a random variable with a probability of landing heads up (binomial distribution[2]). Implicitly:




[2] If we forget the landing on side situation.

# Example 2: determining COVID-19 seropositivity

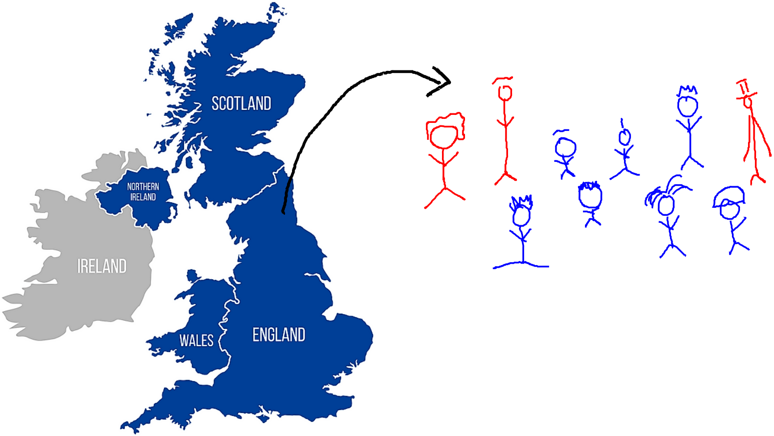
Imagine we want to determine the proportion of the UK population who have COVID-19 antibodies. To do this, we find 10 individuals and test their blood.



Does this mean $3/10 = 30\%$ of the UK population have these antibodies?

# The sampling process yields variation in outputs

No.



# What probability model to use?

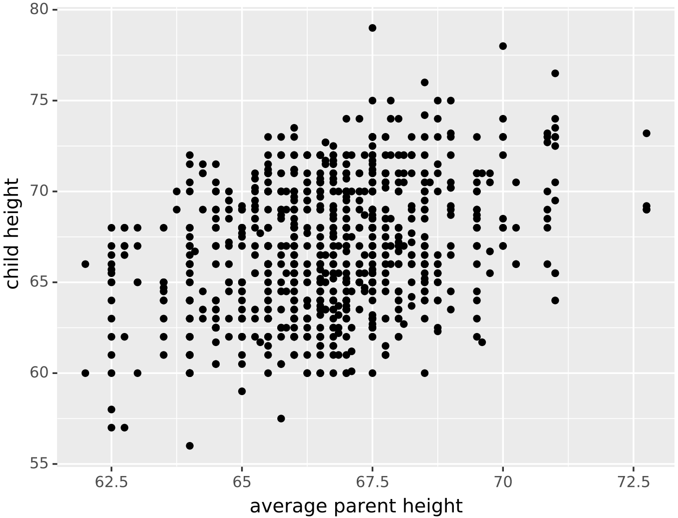
Again, we could use a *binomial* model.

Note, very different problems often can share the same probability model.

Questions?

# Example: Galton's 1885 study of 205 families (898 children)

Question: how inheritable is height from parent to child?



# Potential model

Suppose child height linearly related to parent height:

$$\text{child}_i = a + b * \text{parent}_i. \tag{3}$$

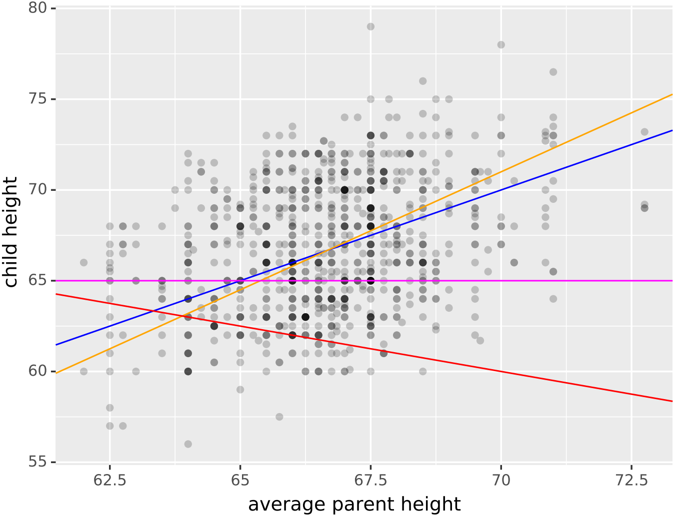
Here, $b$ represents the effect of interest:

- $b = 1 \implies$ a 1 inch increase in average parent height is associated with a 1 inch increase in child height.
- $b = 0 \implies$ no relationship.

This isn't a causal model, so doesn't directly answer our heritability question. But it is still useful.

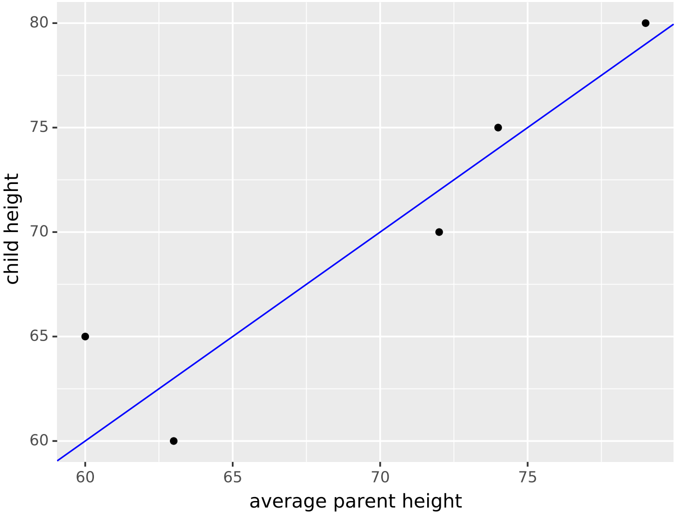
Important: this is a model: a simplified version of reality. It embodies a number of assumptions.

# Example models



# What's the problem with this model?

It doesn't provide a mechanism to exactly hit data.

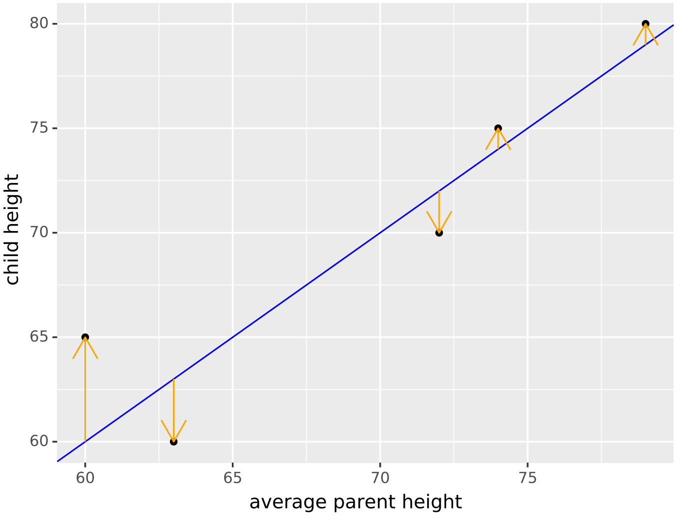


# How to modify the model

$$\text{child}_i = a + b * \text{parent}_i + \epsilon_i. \tag{4}$$

where $\epsilon_i$ is a random error term representing the myriad of other factors not captured by the other parts of the model.

# What does this model look like?



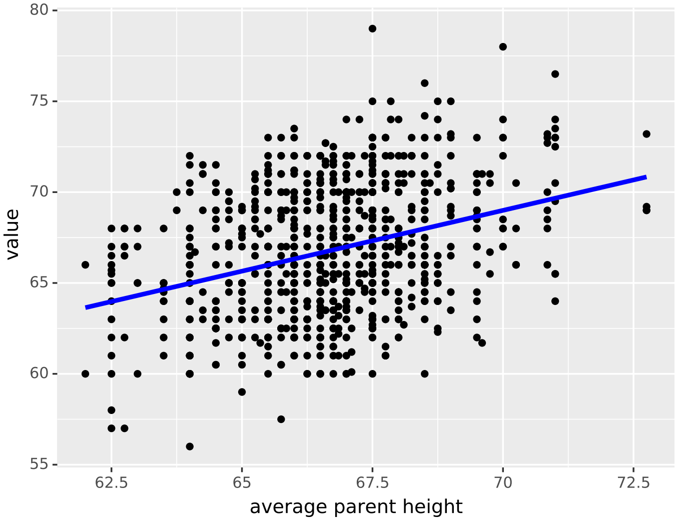
# How to estimate the model's parameters from data?

Want the non-random parts of the model to explain most variation.

So, choose $(a, b)$ so that they minimise some measure of distance between points and line. For example, sum of squared errors:

$$d = \sum_{i=1}^{N} (\text{child}_i - a - b * \text{parent}_i)^2 \tag{5}$$

# Sum of squared errors fit



# Other distance measures

Could have chosen the sum of absolute distances instead:

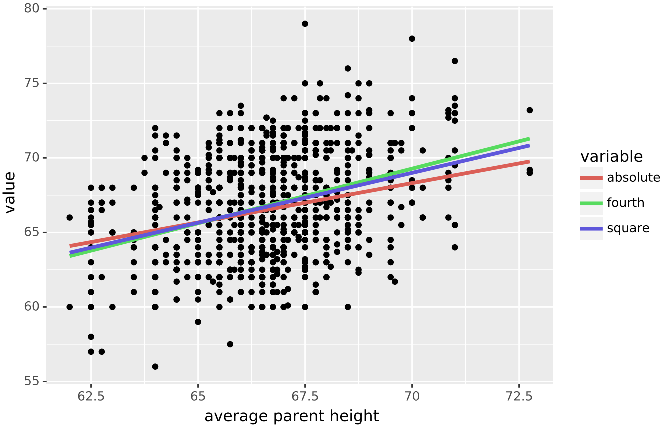
$$d = \sum_{i=1}^{N} \left|\text{child}_i - a - b * \text{parent}_i\right| \tag{6}$$

Or of fourth power:

$$d = \sum_{i=1}^{N} (\text{child}_i - a - b * \text{parent}_i)^4 \tag{7}$$

Question: how would these choices affect the fit?

# Various fits




Conclude: subjective choice of 'distance' affects our estimates.

# A more principled probability-based approach

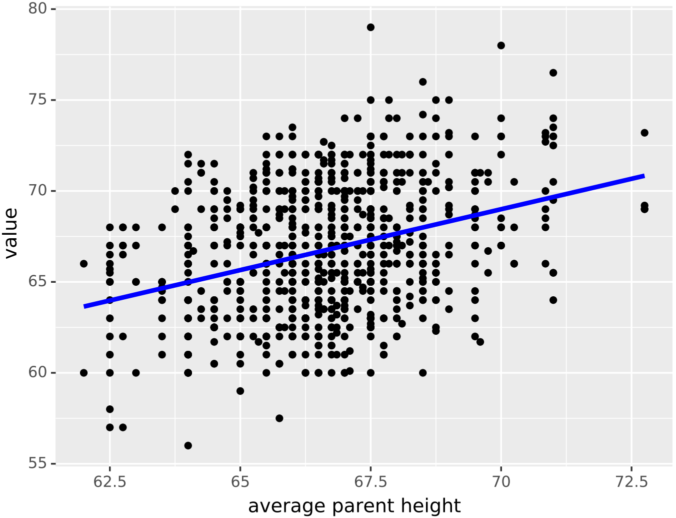
Assume:

$$\text{child}_i = a + b * \text{parent}_i + \epsilon_i, \tag{8}$$

where $\epsilon_i \sim \text{normal}(0, \sigma)$.

# What does this model look like?

Ben to draw:

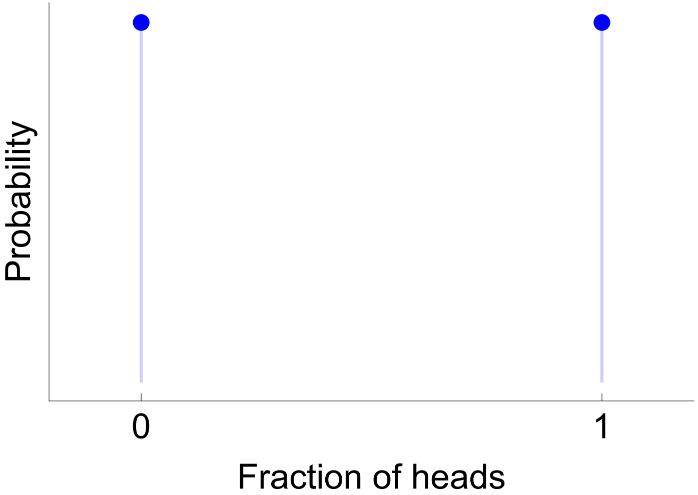


# But isn't a normal distribution also arbitrary?

No! Why? The central limit theorem.

Suppose we flip a fair coin once. What does its probability distribution look like?

# One coin flip

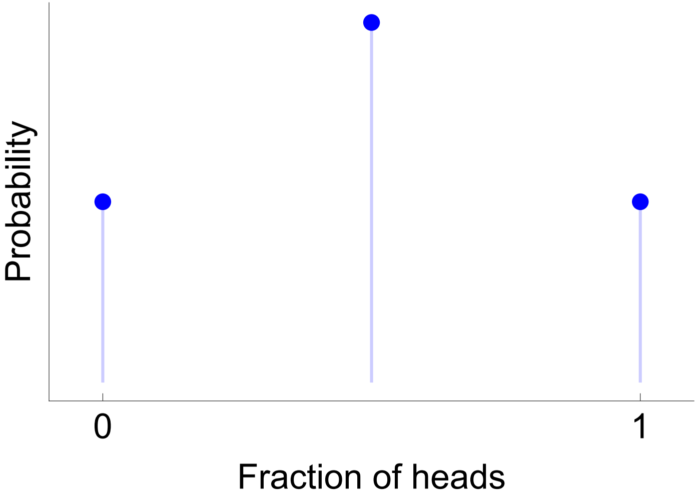


# Two flips

I now flip the coin 2 times.

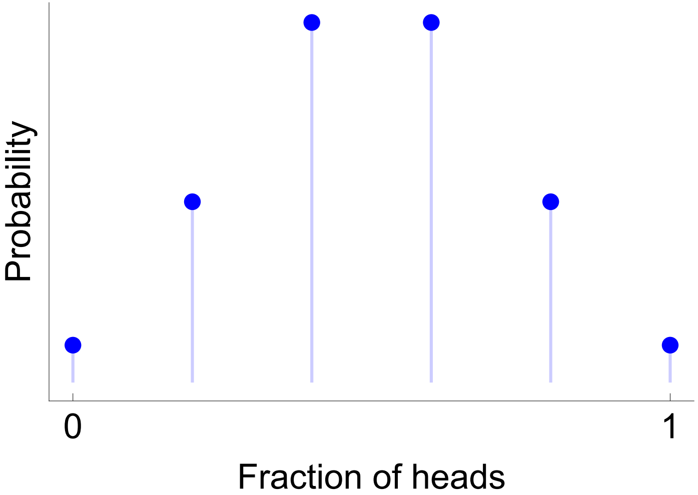
Question: What does the distribution look like now?

# Two flips flips

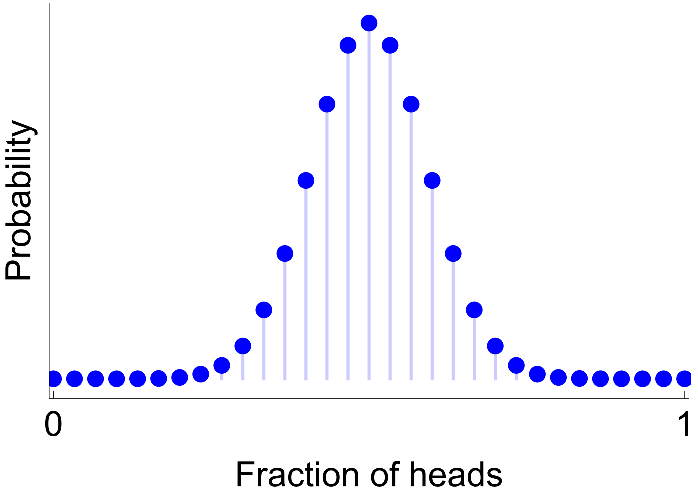


# What about five flips?

# What about five flips?



# What about 30 flips?



# What's going on?

The Central Limit Theorem (CLT) says that under general conditions:

"The distribution of the average of a large number of weakly dependent random variables is approximately normal."

In the coin flipping case, we effectively calculated the average number of independent coin flips landing heads up $\implies$ CLT applies.

# Back to our regression example

$$\text{child}_i = a + b * \text{parent}_i + \epsilon_i, \tag{9}$$

where $\epsilon_i \sim \text{normal}(0, \sigma)$.

Large number of weakly dependent factors – genetic, environmental and so-forth – likely influence a person's height.
$\implies$
CLT applies, so normal distribution may be appropriate.

# Likelihood based inference

$$\text{child}_i = a + b * \text{parent}_i + \epsilon_i, \tag{10}$$

and

$$\epsilon_i \sim \text{normal}(0, \sigma), \tag{11}$$

means that:

$$\text{child}_i - a - b * \text{parent}_i \sim \text{normal}(0, \sigma). \tag{12}$$

This provides us with a way of writing down the overall probability (density) of observations.

# Likelihood

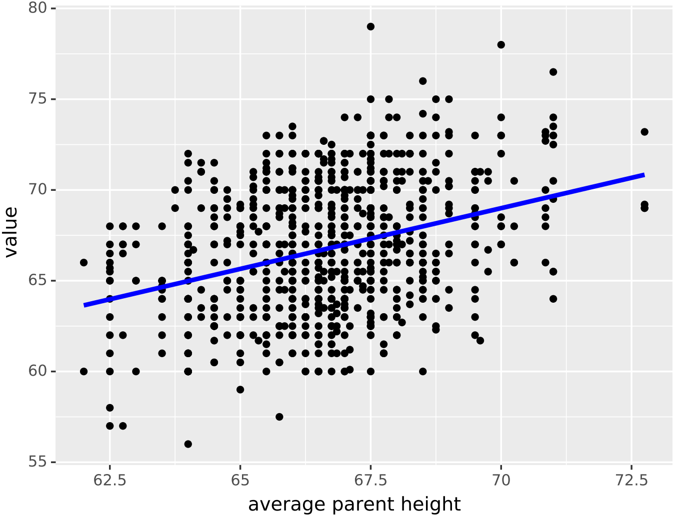
Due to independence of observations:

$$
\begin{aligned}
\mathcal{L} =& \mathbb{P}(\epsilon_1) \times \mathbb{P}(\epsilon_2) \times \ldots \times \mathbb{P}(\epsilon_n) && (13)\\
=& \text{normal}(\text{child}_1 - a - b * \text{parent}_1 | 0, \sigma) \times && (14)\\
& \text{normal}(\text{child}_2 - a - b * \text{parent}_2 | 0, \sigma) && (15)\\
& \ldots && (16)\\
& \text{normal}(\text{child}_n - a - b * \text{parent}_n | 0, \sigma) && (17)\\
& && (18)
\end{aligned}
$$

This object is a function of the parameters of our model: $a$ and $b$. So choose $a$ and $b$ values to maximise $\mathcal{L}$.

This is known as the method of *maximum likelihood*.

# Likelihood estimated model



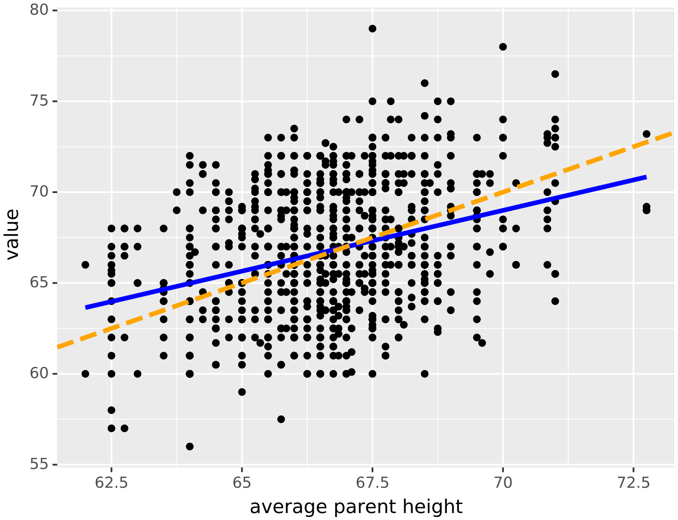
# Strength of association between parent and child heights

Estimates of regression coefficient:

$$\text{child}_i = a + b * \text{parent}_i + \epsilon_i, \tag{19}$$

- Least squares / maximum likelihood: $b = 0.67$
- Absolute deviance: $b = 0.57$
- Fourth power: $b = 0.73$

# Regression to mean



Questions?

# Regression is a model

$$\text{child}_i = a + b * \text{parent}_i + \epsilon_i, \tag{20}$$

and

$$\epsilon_i \sim \text{normal}(0, \sigma), \tag{21}$$

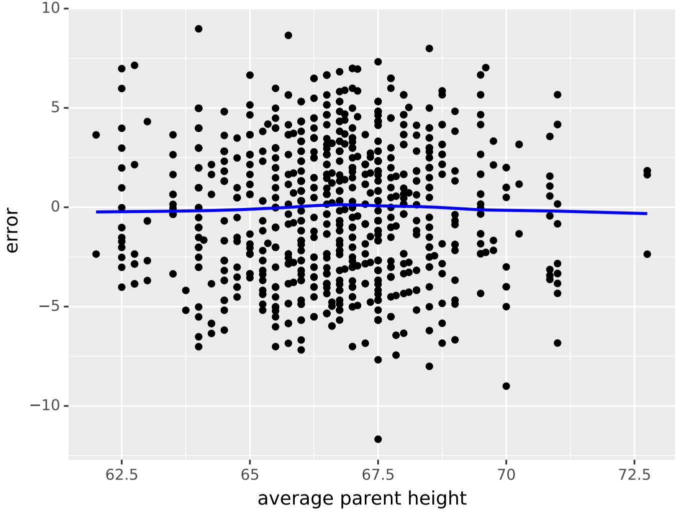
is a model: an idealised version of reality.

# Assumptions

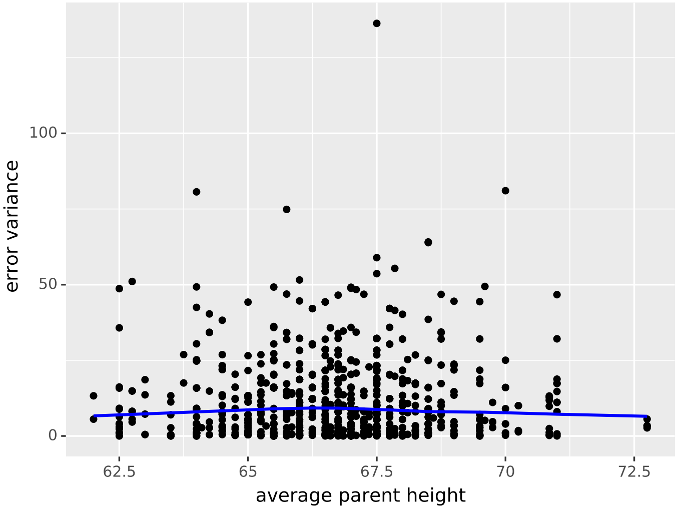
Some of these:

- Linear association of average parental height with child height
- Variance of points around line is constant
- Normality of errors
- Male / female parent heights not separately important
- Sex of child unimportant in relationship

# Checking linearity



# Checking variance homogeneity



Questions?

# How uncertain are we?

Least squares / maximum likelihood estimates:

$$\text{child}_i = 22.15 + 0.67 * \text{parent}_i + \epsilon_i, \tag{22}$$

How representative are these of the population as a whole?

# Gauging uncertainty

Maybe draw this?
Imagine:

- Repeatedly sampling from the population
- Each time calculating an estimate

$\Longrightarrow$ variability in estimates dictates uncertainty.

# Circularity

- If we knew variability in estimates, we'd know how accurate they were
- But to do so, we need exact details of the population

How to resolve this ambiguity?

# Two resolutions

- Make mathematical assumptions about nature of population distribution. e.g. use CLT to determine normality
- Bootstrap sampling

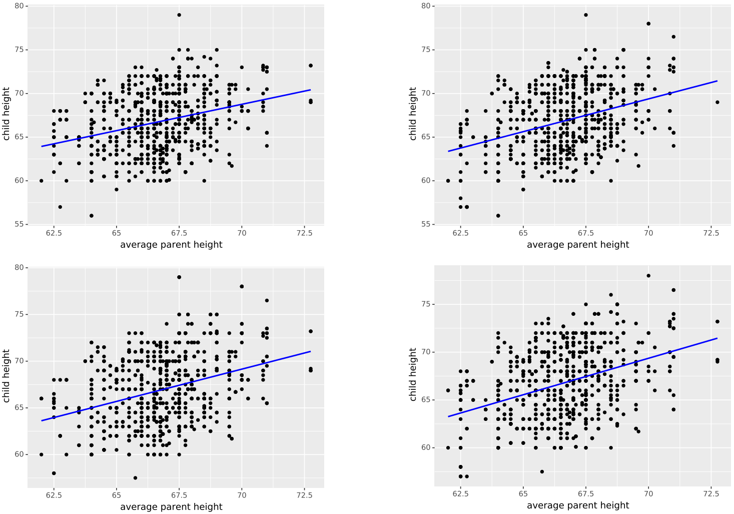
# Bootstrap sampling

Since sample is drawn from the population, it should mirror it.

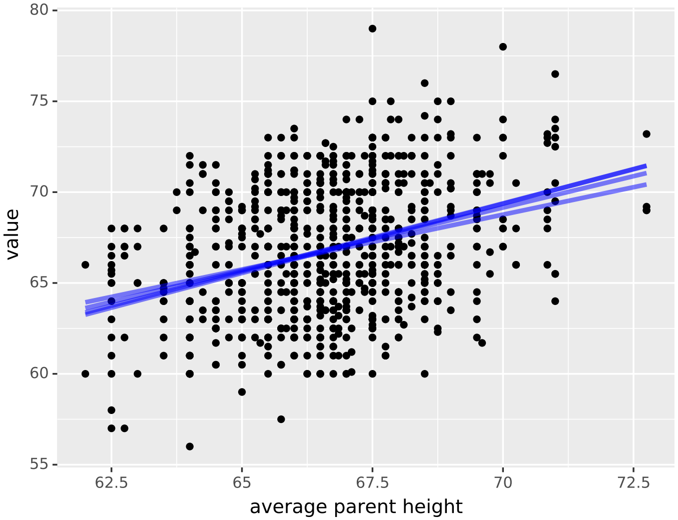
$\implies$ repeatedly draw new samples from our sample! And perform estimation on each.

Note sampling done with replacement.

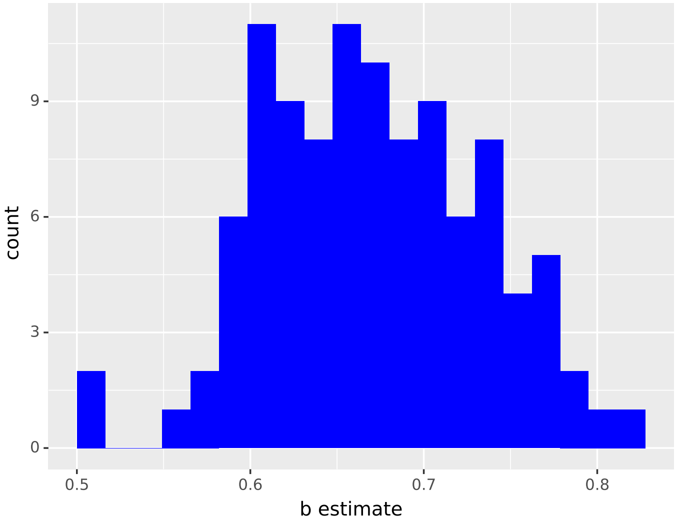
# Bootstrapped Galton samples



# Bootstrapped Galton estimates



# Bootstrapped Galton estimates



# When doesn't bootstrapping work?

- Complex, highly structured data: for example, time series
- Large data $\implies$ too expensive

$\implies$ probability model based approach less cumbersome in these cases.

# That's it!

Questions?