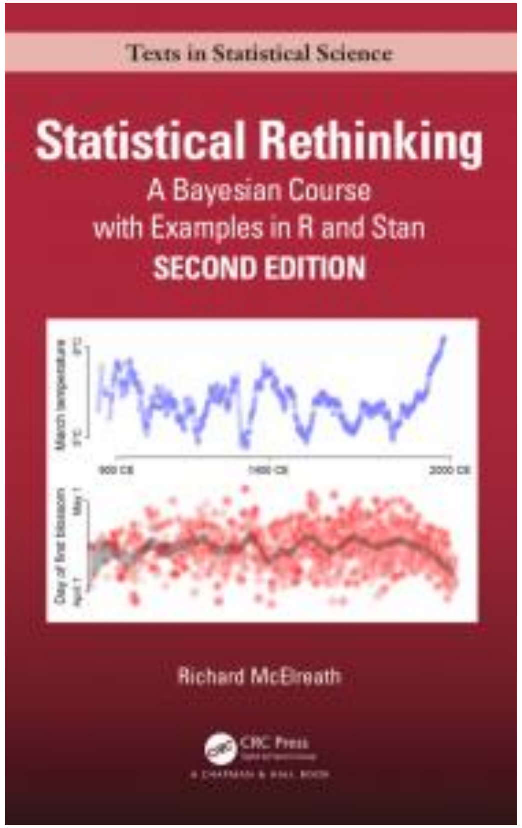
Telling a data story

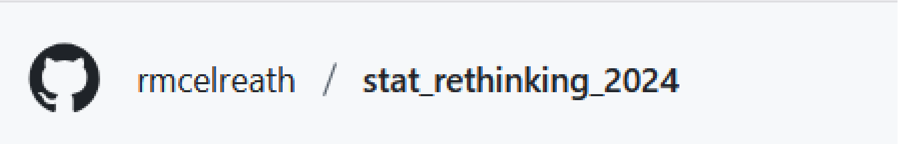
Katherine Muller

What is a data story?

- Story for how the data came to be
- Can be descriptive or causal
- Describes the underlying reality and the sampling process
- Describes how to simulate new data

Credit for “data story” concept





Objectives:

Preview the data storytelling process used in Bayesian data analysis

1) Tell a data story about how data came to be
2) Translate that story into a generative model
3) Use the generative model to simulate new data in R **see code*
4) Predict how a system will respond to change.

Understand what it means for a statistical model to describe a data-generating process, rather than simply describing a dataset

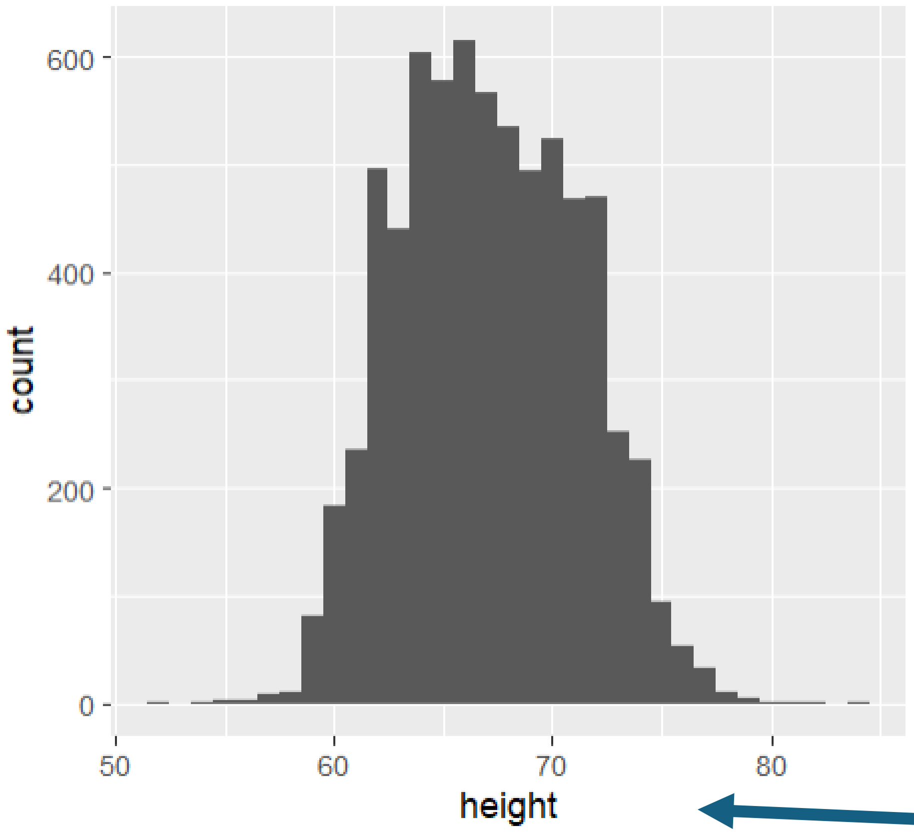
Unlearn bad habits from intro stats

Describing data: Shape

2012 adult height data (n = 7006)
National Longitudinal study
US Bureau of Labor Statistics

Histogram

Count number of people in each bin

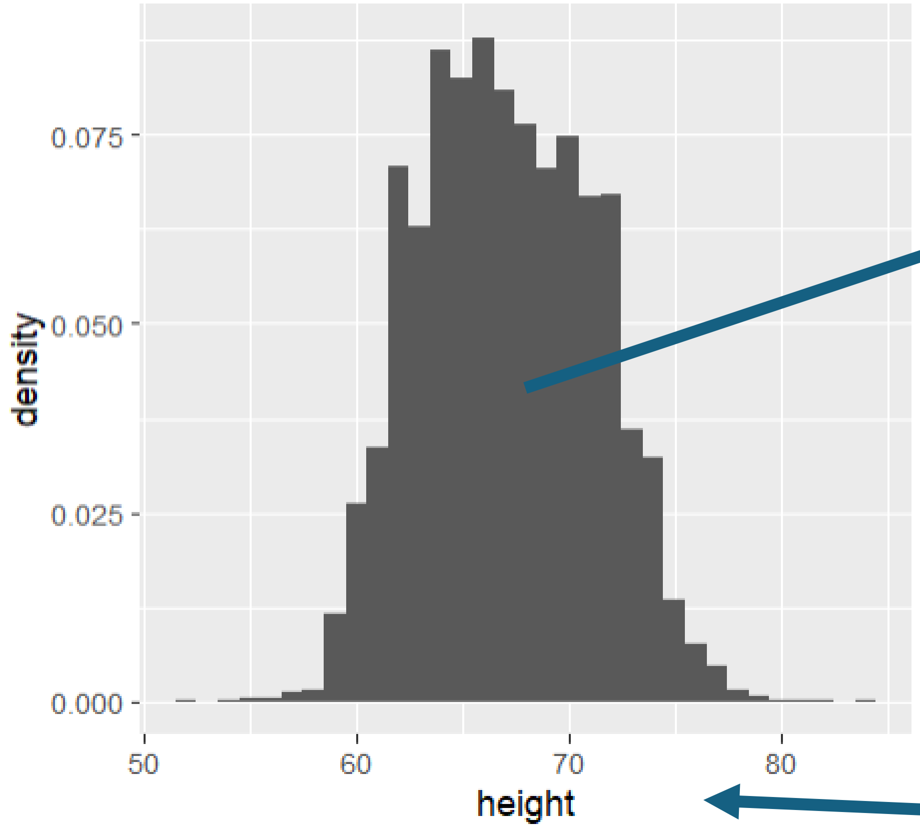


Split height into sensible intervals (bins)—here it's 1"

Describing data: Shape

2012 adult height data (n = 7006)
National Longitudinal study
US Bureau of Labor Statistics

Histogram



Count number of people in each bin

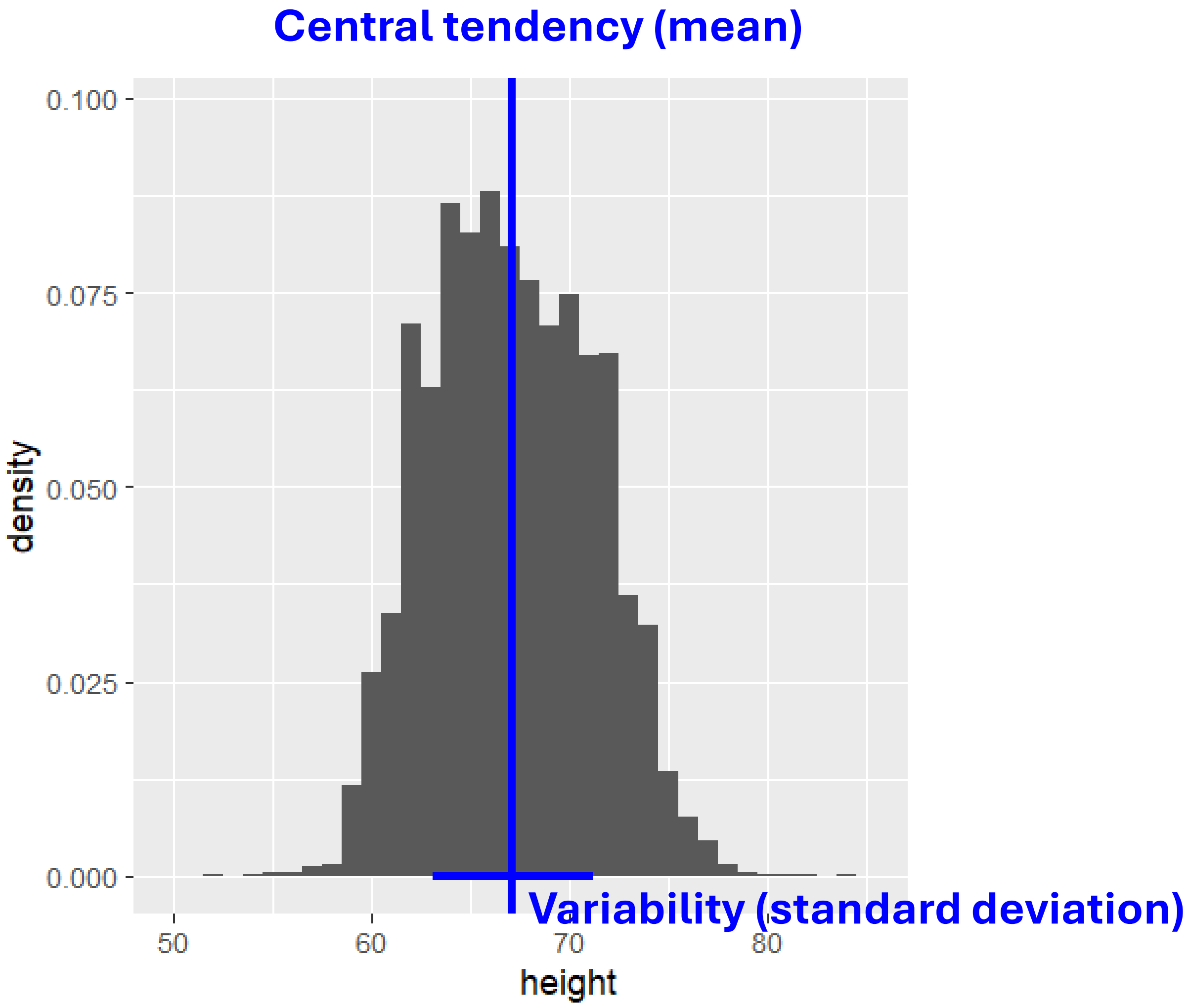
Divide by the total number of people

Sums to 1

Split height into sensible intervals (bins)—here it's 1"

Describing data: Summary stats

2012 adult height data (n = 7006)
National Longitudinal study
US Bureau of Labor Statistics

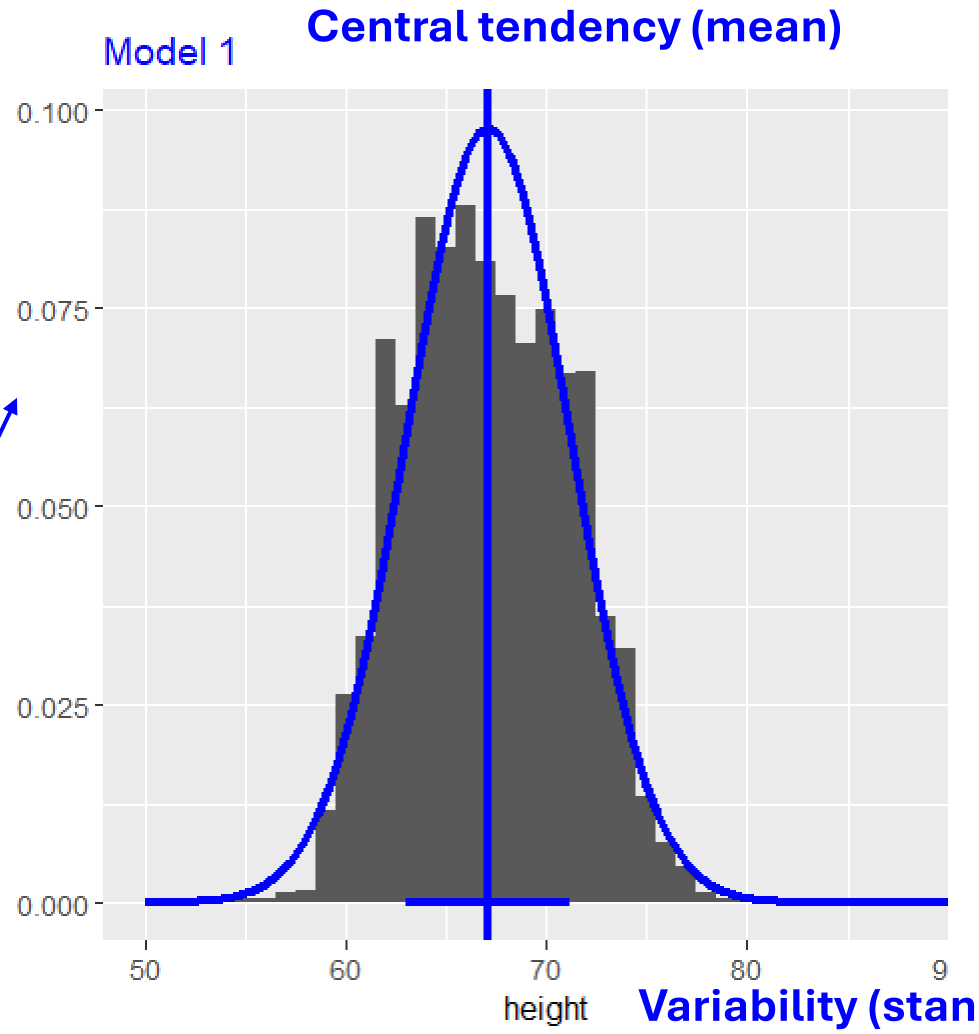


Describing data with models

2012 adult height data (n = 7006)
National Longitudinal study
US Bureau of Labor Statistics

Model 1: Height is a normally distributed random variable with one mean and one standard deviation

Probability density:
Chance of observing height *X* under model 1



$$\boxed{\text{height} \sim \mathcal{N}(\mu, \sigma)}$$

$\mu \approx \text{mean(height)}$

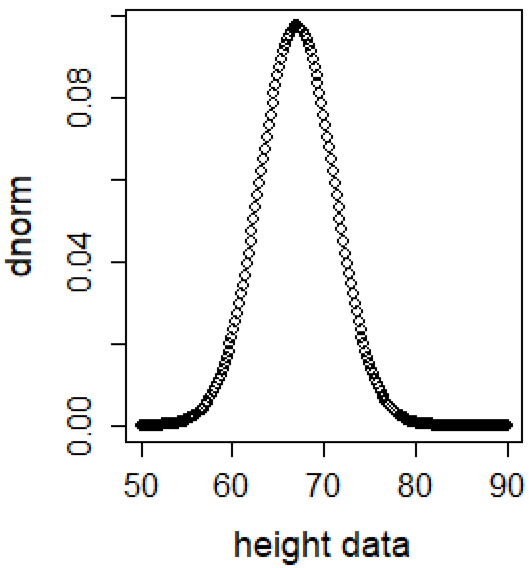
$\sigma \approx \text{sd(height)}$

Theoretical distribution functions in R

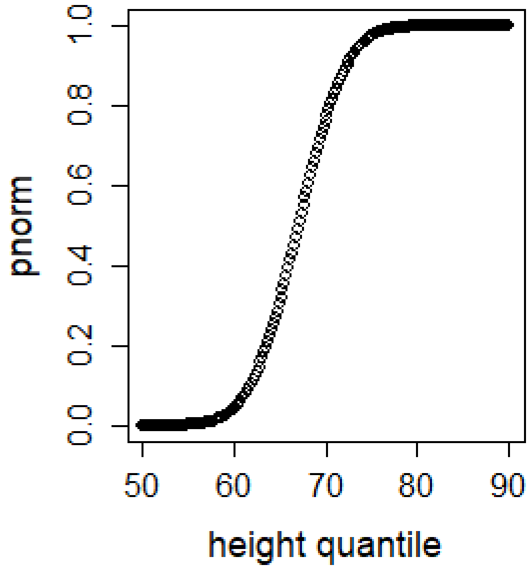
$$\text{height} \sim \mathcal{N}(\mu, \sigma)$$

Theoretical distribution functions in R

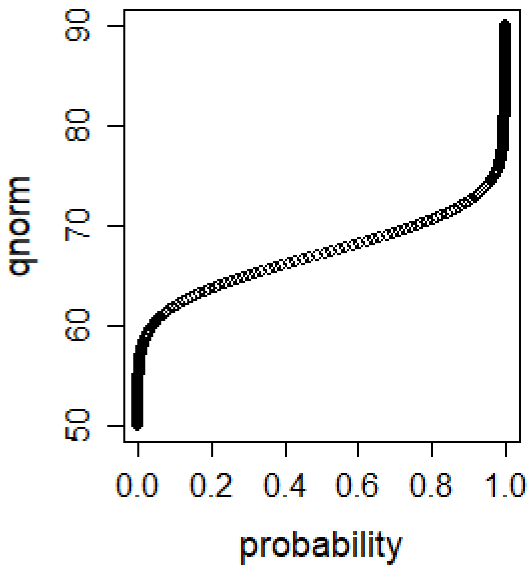
Probability density function (PDF)	**`dnorm(x, mean, sd)`**
Cumulative distribution function (CDF)	**`pnorm(q, mean, sd)`**
Quantile function (inverse CDF)	**`qnorm(p, mean, sd)`**



What is the probability of height = X ?



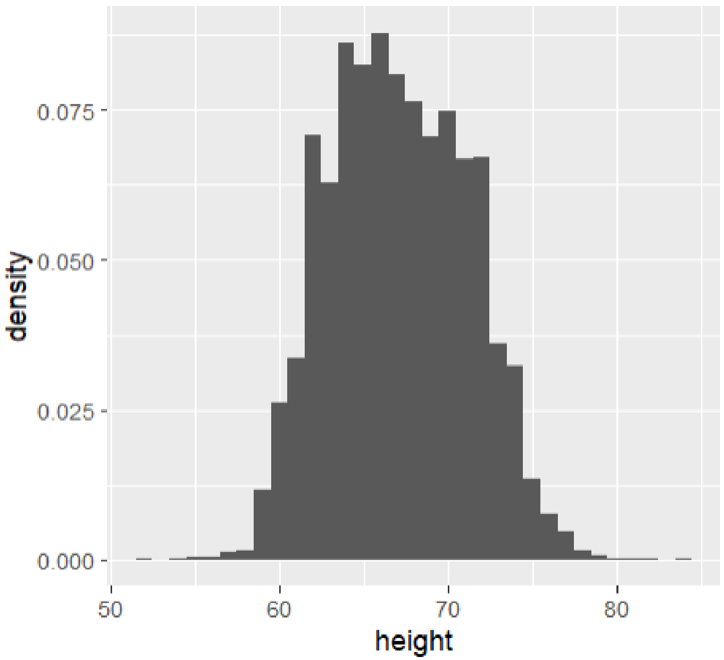
What is the probability of height ≤ X ?



What is the Xth percentile for height? (e.g., median = 50th percentile)

https://rstudio.github.io/r-manuals/r-intro/Probability-distributions.html#r-as-a-set-of-statistical-tables

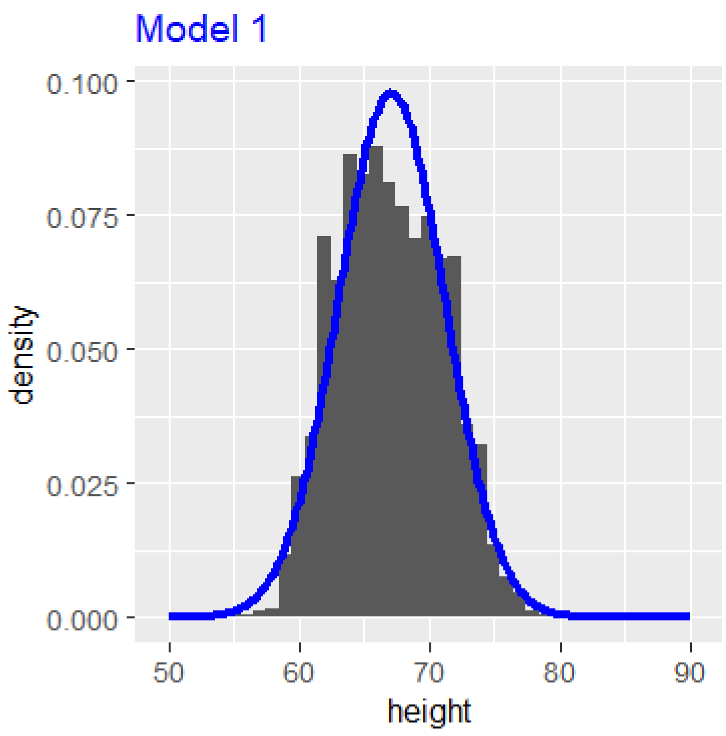
Describing how data are generated



Data story

- People grow up and reach a certain height.
- Some people are taller and some are shorter.
- Most people are somewhere in the middle between very tall and very short.

Describing how data are generated with models

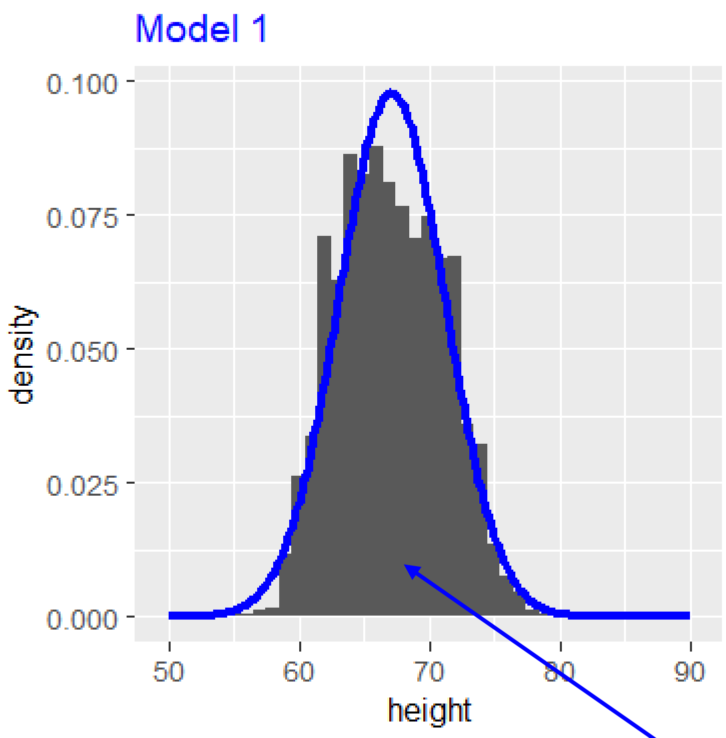


Single gaussian population $\boxed{\text{height} \sim \mathcal{N}(\mu, \sigma)}$

Data story

- People grow up and reach a certain height. → Random variable *height*
- Some people are taller and some are shorter. → Parameter σ
- Most people are somewhere in the middle between very tall and very short. → Parameter μ

Generating data with models



Single gaussian population $\boxed{\text{height} \sim \mathcal{N}(\mu, \sigma)}$

We can use our model to simulate new data!

```
rnorm(n, mean, sd)
```

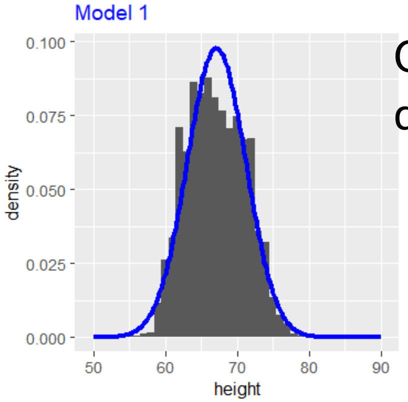
Number of samples — n

μ — mean

σ — sd

- Generate *n* independent random samples
- Values with higher probability density are more likely to be sampled

Generating data with models



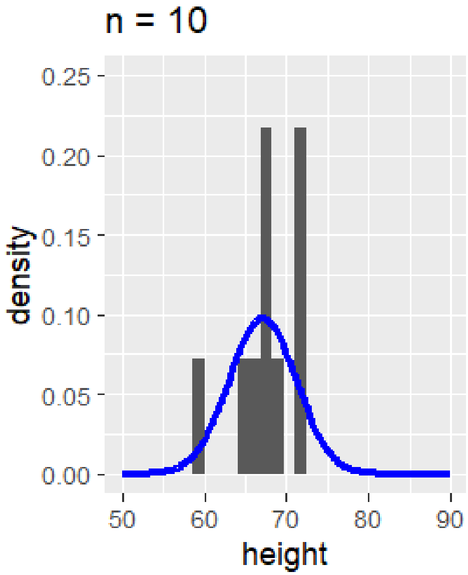
Observed data

Single gaussian population

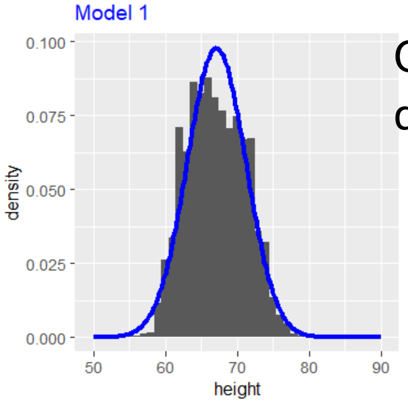
$$\boxed{\text{height} \sim \mathcal{N}(\mu, \sigma)}$$

We can use our model to simulate new data!

```
rnorm(n, mean, sd)
```



Generating data with models



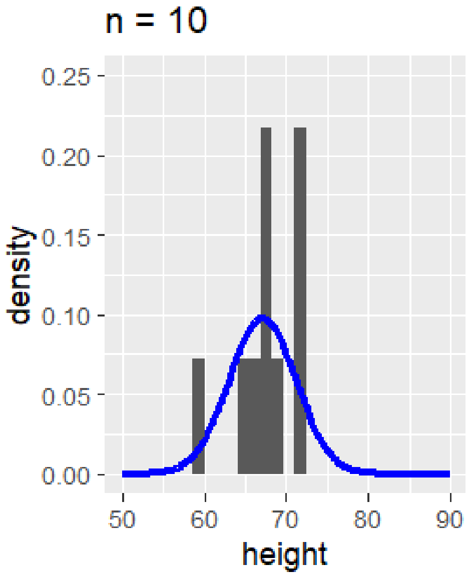
Observed data

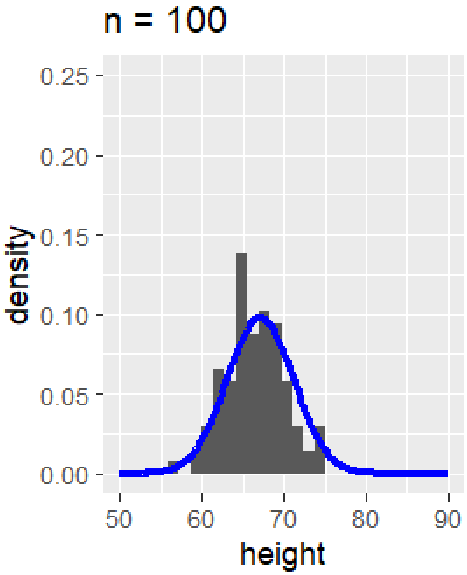
Single gaussian population

$$\boxed{\text{height} \sim \mathcal{N}(\mu, \sigma)}$$

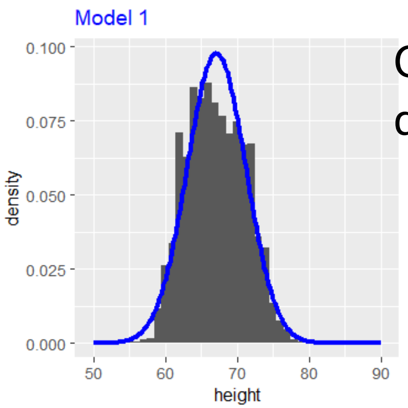
We can use our model to simulate new data!

```
rnorm(n, mean, sd)
```





Generating data with models



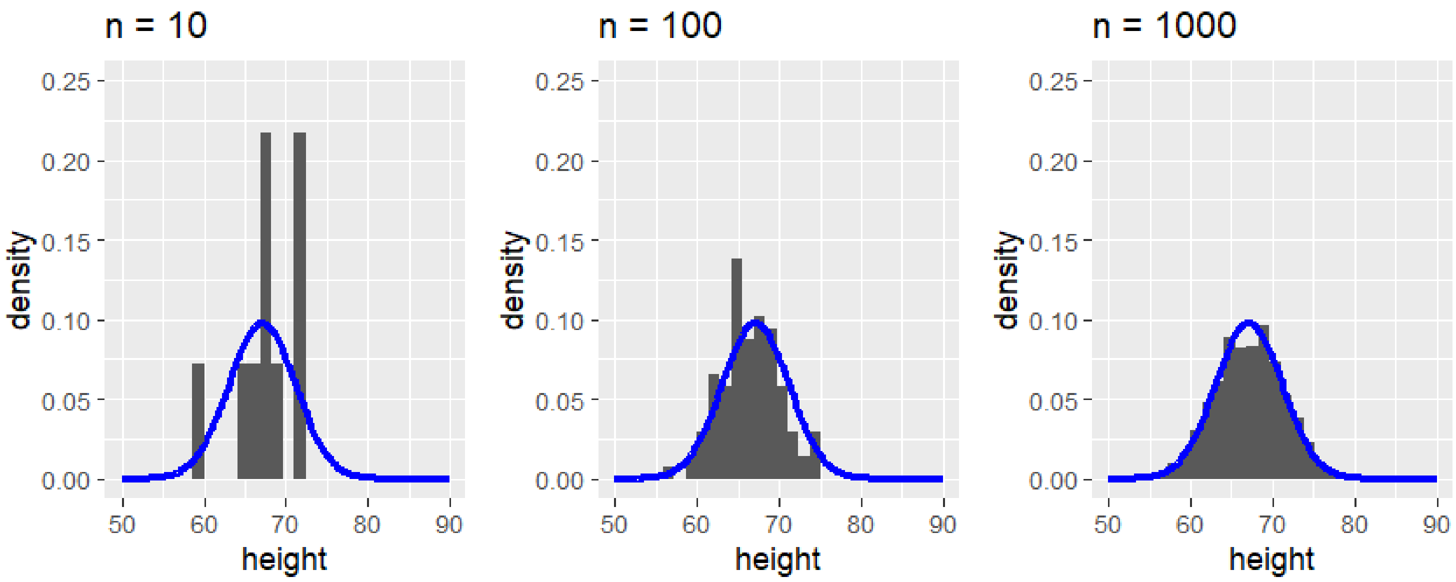
Observed data

Single gaussian population

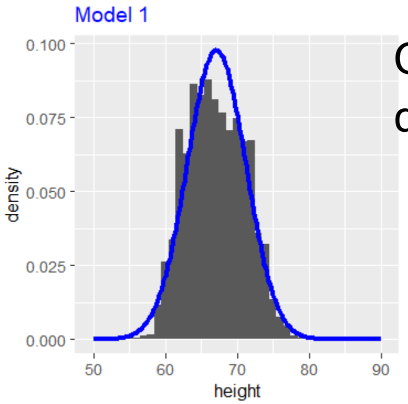
$$\boxed{\text{height} \sim \mathcal{N}(\mu, \sigma)}$$

We can use our model to simulate new data!

```
rnorm(n, mean, sd)
```



Generating data with models



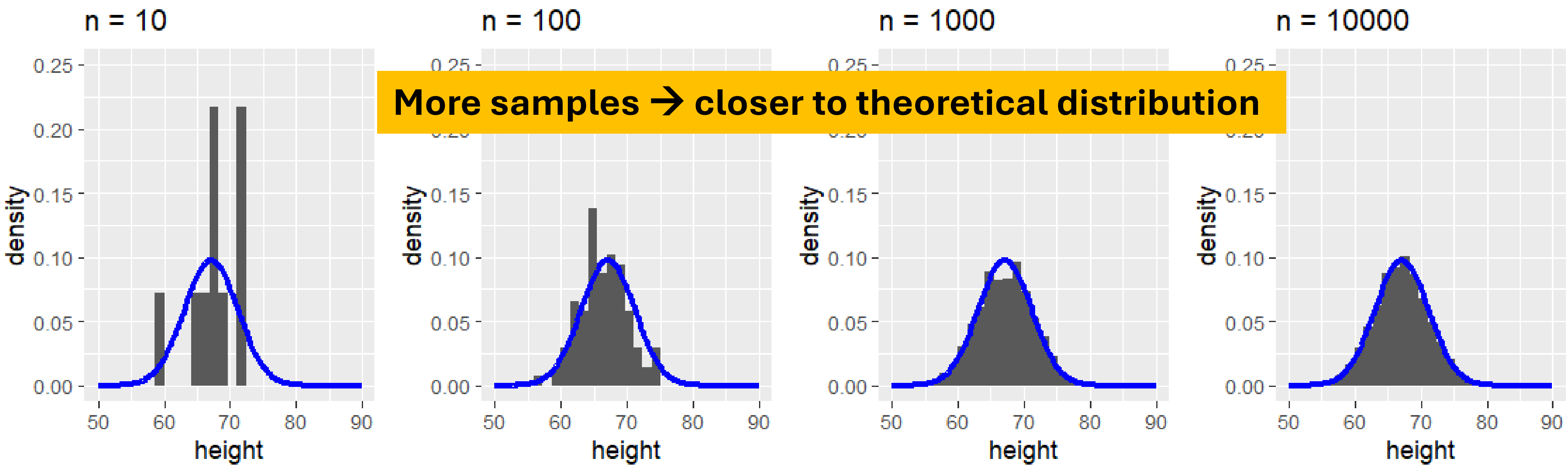
Observed data

Single gaussian population

$$\boxed{\text{height} \sim \mathcal{N}(\mu, \sigma)}$$

We can use our model to simulate new data!

```
rnorm(n, mean, sd)
```



More samples → closer to theoretical distribution

Recap

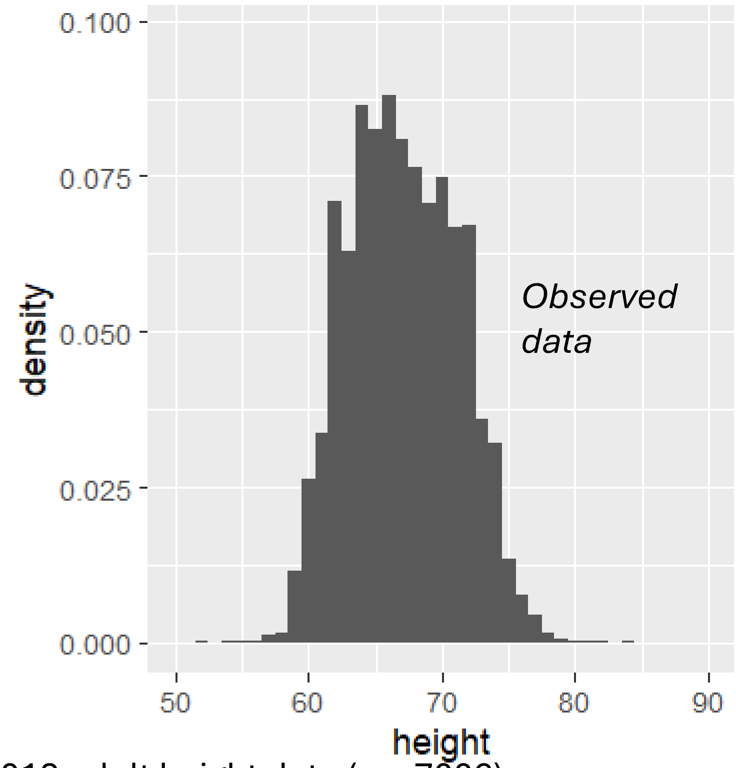
Compare observed and generated data to make inferences about the system

1. Observe some data

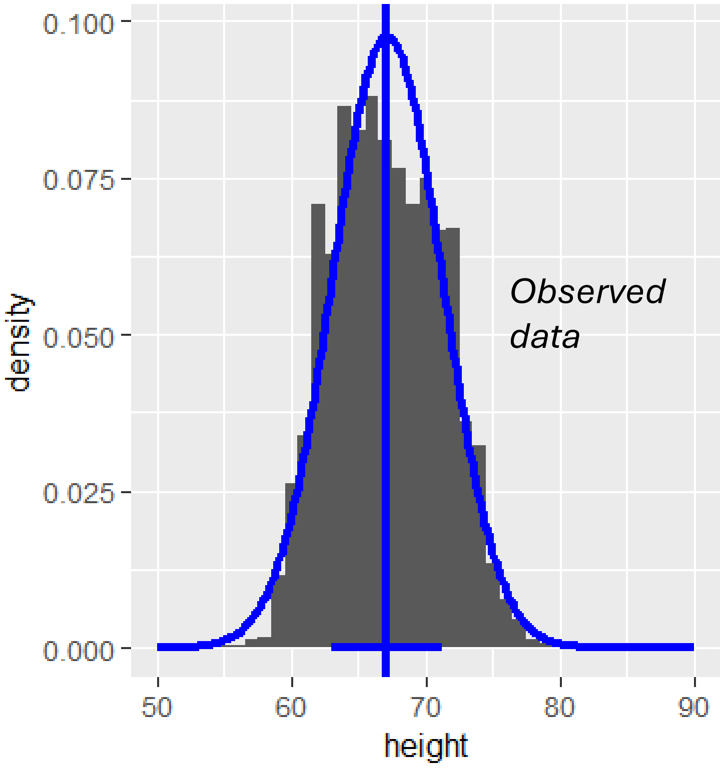
2. Tell a story about how data came to be.

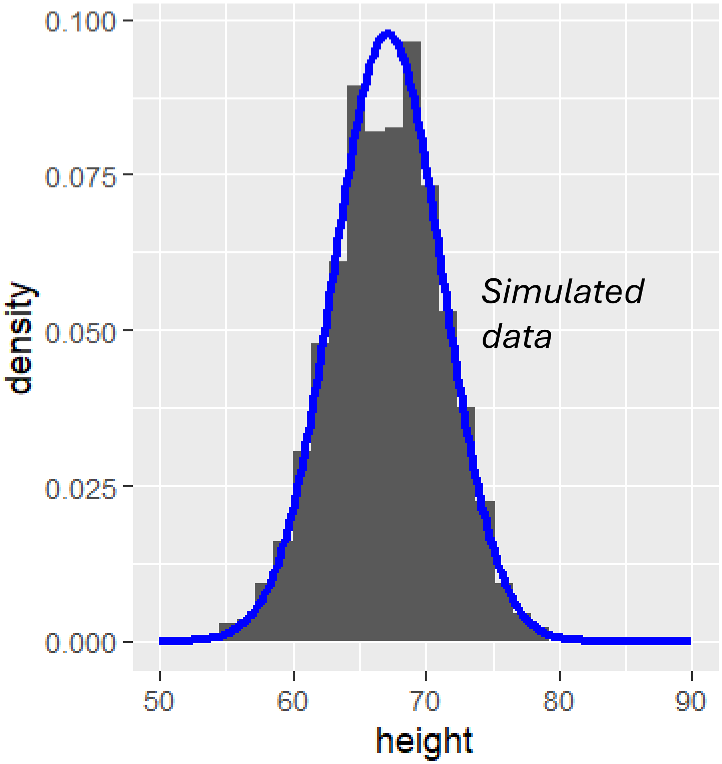
2. Translate that story into a generative model.

3. Simulate data from the model



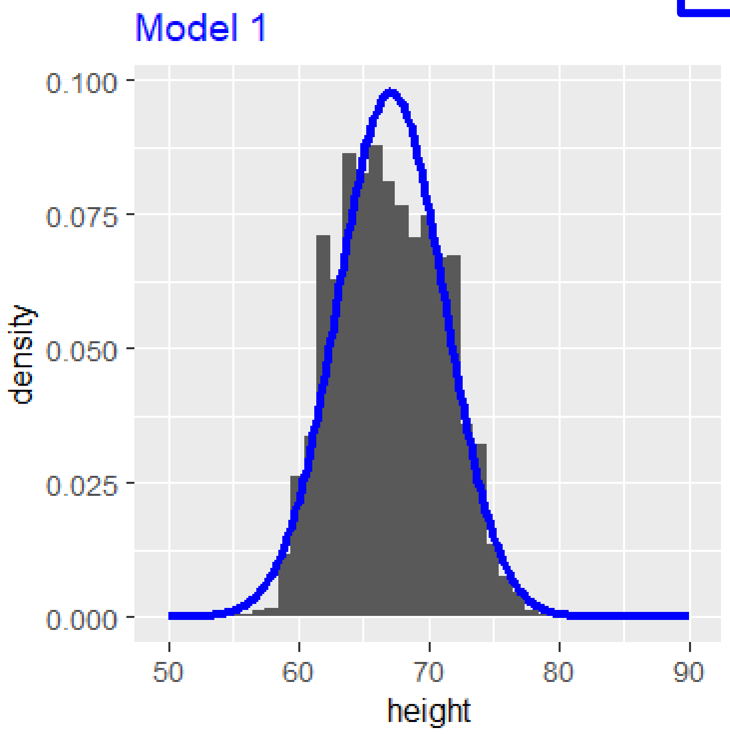
2012 adult height data (n = 7006)
National Longitudinal study
US Bureau of Labor Statistics





Back to the data story

$$\boxed{\text{height} \sim \mathcal{N}(\mu, \sigma)}$$

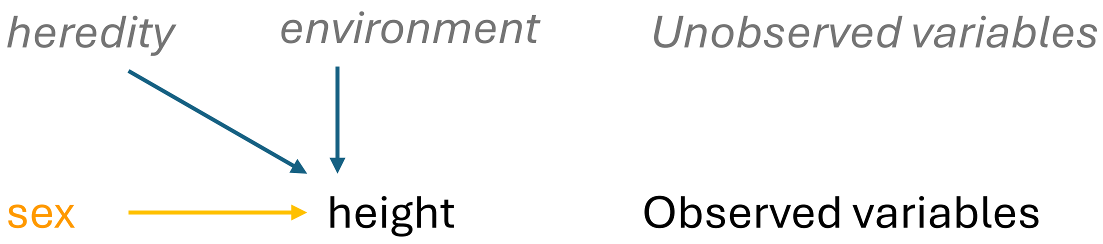


Data story

People grow up and reach a certain height.	Random variable *height*
Some people are taller and some are shorter.	Parameter σ
Most people are somewhere in the middle between very tall and very short.	Parameter μ

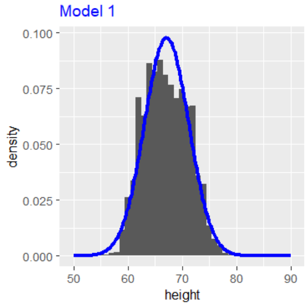
What's missing?

Causal diagram

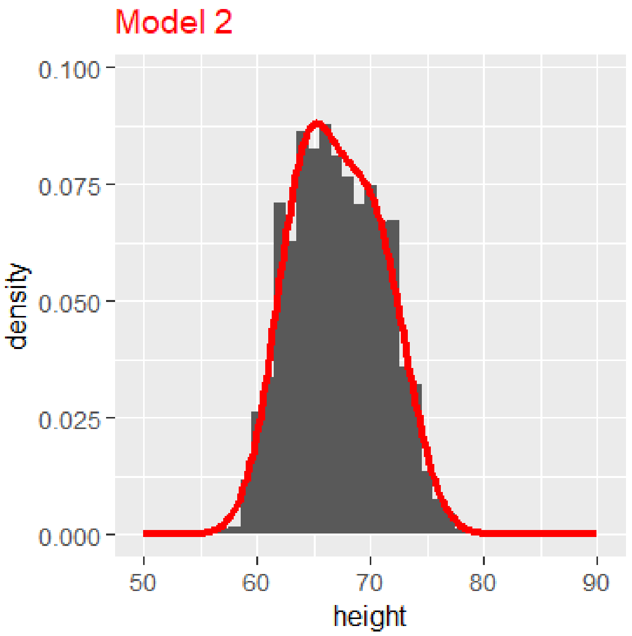


Back to the data story

$$\boxed{\text{height} \sim \phi_W \mathcal{N}(\mu_W, \sigma_W) + (1-\phi_W)\mathcal{N}(\mu_M, \sigma_M)}$$

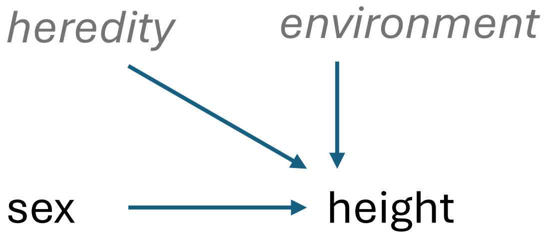


$$\boxed{\text{height} \sim \mathcal{N}(\mu, \sigma)}$$



Data story

People grow up and reach a certain height.	Random variable *height*
Some people are taller and some are shorter.	Parameter σ
Most people are somewhere in the middle between very tall and very short.	Parameter μ
Women and men have a different height distribution.	Separate $\mu_{W, M}$ $\sigma_{W, M}$
The population contains a mixture of men and women.	Parameter ϕ_W



Latent (unobserved) variables

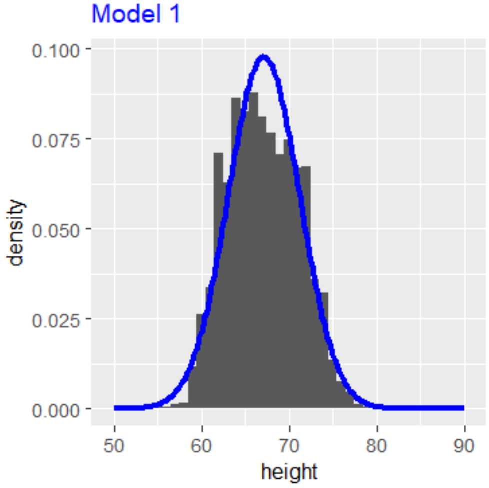
Observed variables

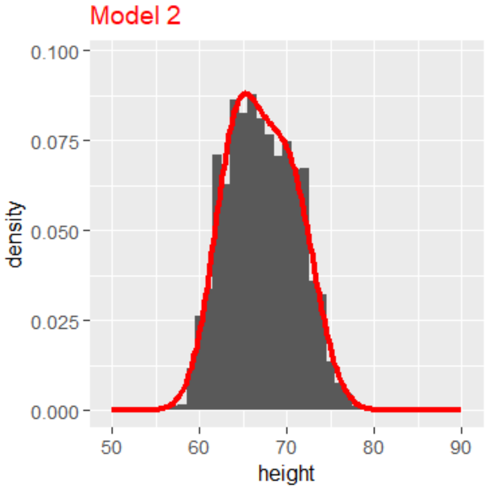
Descriptive framework

Model 1 is rejected in favor of Model 2

Generative framework

Model 1 describes part of the generative process in Model 2





We often need multiple models to tell complex data stories.

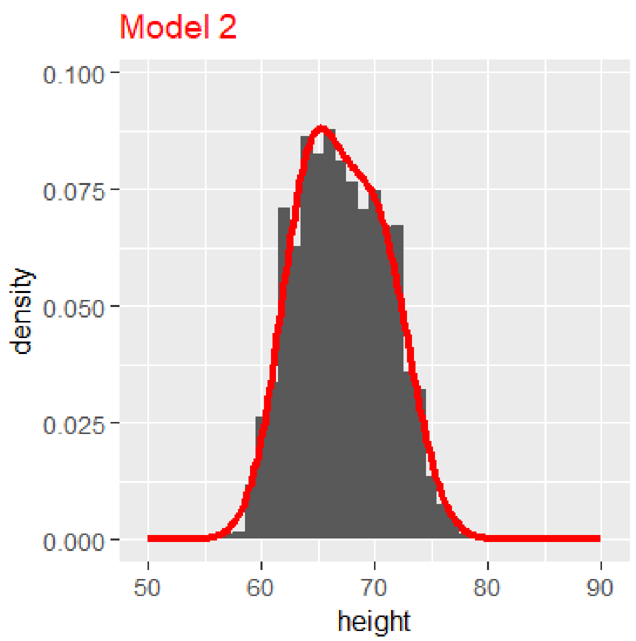
$$\text{height} \sim \mathcal{N}(\mu, \sigma)$$

$$\text{height} \sim \phi_W \mathcal{N}(\mu_W, \sigma_W) + (1 - \phi_W)\mathcal{N}(\mu_M, \sigma_M)$$

The generative model framework allows you to make predictions about how a system will respond to change

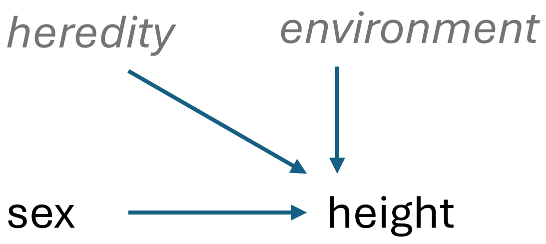
Predicting response to change

$$\text{height} \sim \phi_W \mathcal{N}(\mu_W, \sigma_W) + (1 - \phi_W)\mathcal{N}(\mu_M, \sigma_M)$$



Data story

People grow up and reach a certain height.	Random variable *height*
Some people are taller and some are shorter.	Parameter σ
Most people are somewhere in the middle between very tall and very short.	Parameter μ
Women and men have a different height distribution.	Separate $\mu_{W, M}$ $\sigma_{W, M}$
The population contains a mixture of men and women.	Parameter ϕ_W

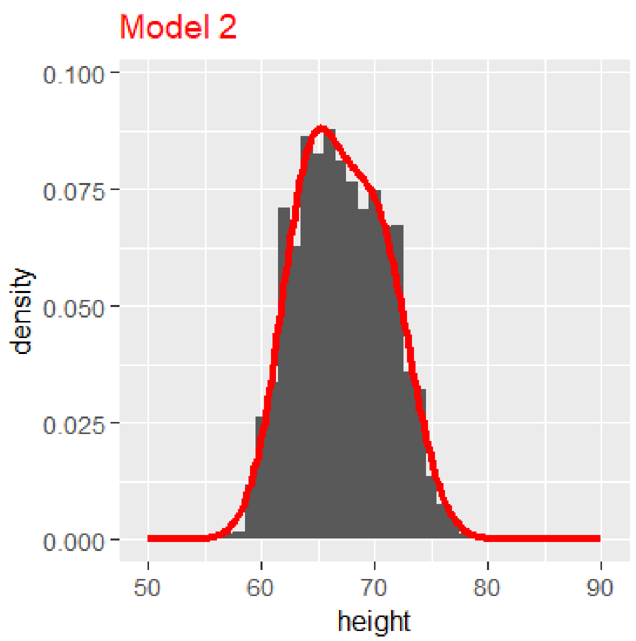


Latent (unobserved) variables

Observed variables

Predicting response to change

$$\boxed{\text{height} \sim \phi_W \mathcal{N}(\mu_W, \sigma_W) + (1-\phi_W)\mathcal{N}(\mu_M, \sigma_M)}$$



Change scenario → *Simulation*

Influx of Amazon warrior women

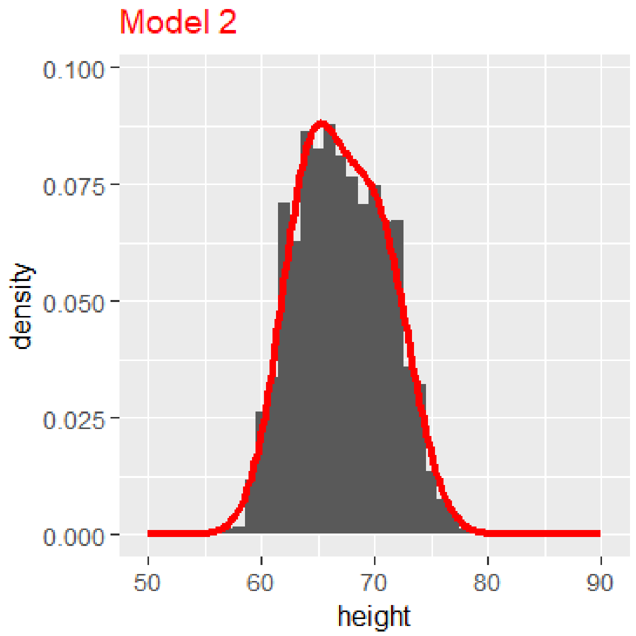
$\uparrow \mu_W$

heredity *environment* *Latent (unobserved) variables*

sex → height Observed variables

Predicting response to change

$$\boxed{\text{height} \sim \phi_W \mathcal{N}(\mu_W, \sigma_W) + (1 - \phi_W)\mathcal{N}(\mu_M, \sigma_M)}$$



Change scenario → *Simulation*

Alice in Wonderland potions

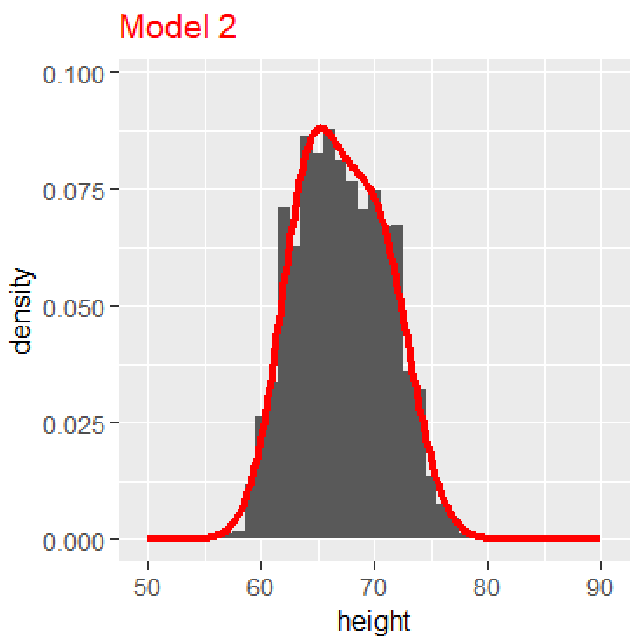
↑σ

heredity *environment* *Latent (unobserved) variables*

sex → height Observed variables

Predicting response to change

$$\boxed{\text{height} \sim \phi_W \mathcal{N}(\mu_W, \sigma_W) + (1 - \phi_W)\mathcal{N}(\mu_M, \sigma_M)}$$



Change scenario → *Simulation*

Male flight

$\uparrow\phi_W$

heredity *environment*

sex → height

Latent (unobserved) variables

Observed variables

Exercises (see lecture code)

1. A terrible disease that kills 10% of women and 30% of men. Simulate a new survey of 10000 people.
2. Thanks to the terrible disease, now the average woman is twice as tall as the average man. Simulate a new survey with 10000 people.
3. Simulate a survey of 100, 1000, and 10000 people. Repeat each survey 3 times and compare the consistency.

Recap

- Tell a story about how your data came to be
- Represent that story with a model
    - Random variables
    - Probability distribution function
    - Parameters
- Generate data with your model
    - Sampling functions in R
- Predict how a system will respond to change
    - Exercises