Chapter 2

Lin 205C
Santiago Barreda

Chapter 2

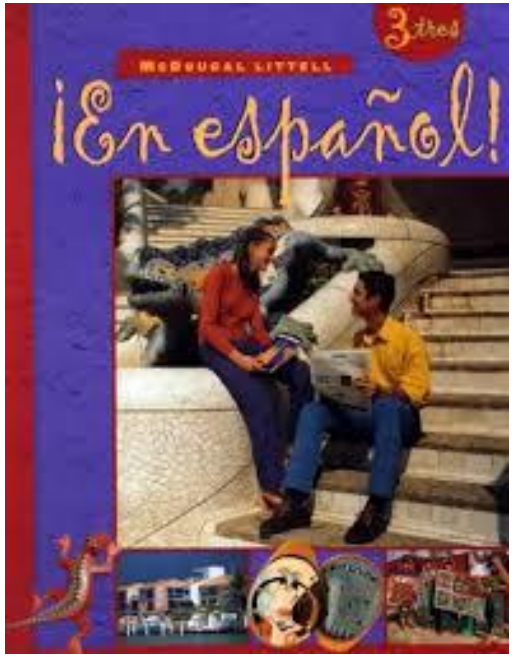
The purpose of this chapter is:

- ‘Basic’ explanation of probability and likelihood.
- Motivate the use of likelihoods in scientific inference.
- Make statistical models less of a ‘black box’.
- Start getting used to mathematical notation.

The purpose of this chapter is *not*:

- Understand all the contents of the chapter

Statistics = Procedural Knowledge



Data and Research Questions

Goal: Use the data in our experiment to answer these questions:

- (Q1) How tall does the average adult male ‘sound’?

- (Q2) Can we set limits on credible average apparent heights based on the data we collected?

Why Estimate Variation?

Interpreting average values relies on understanding variation for at least two reasons:

- Uncertainty in measurements.

- Interpreting Magnitudes (i.e., is the effect large)

Probability

- There are many ways to think of probability mathematically and philosophically.
- Simple Definition: Probability of event is the number of times outcome occurs relative to all the other possible outcomes.
- The probability of each outcomes is a value between 0 and 1.
- The sum of all probabilities for all outcomes is 1.

Probability: Example

- A die has 6 possible outcomes: 1,2,3,4,5, and 6.

$$\text{outcome}_\text{i} \qquad i = \{1,2,3,4,5,6\}$$

- Each outcome is equally likely, a 1/6 (0.167) chance of occurring.

$$P(\text{ outcome}_\text{i}) = 1/6 = 0.16666 \ldots$$

- The sum of the seven outcomes is equal to 1.

$$1 = \sum_{i=1}^{6} P(\text{ outcome}_\text{i}\ )$$

Probability: Not quite so simple…..

- “The probability that the Maple Leafs will win the Stanley Cup is….”
- “The probability of seeing an earthquake larger than any earthquake we’ve ever seen is…”
- “The probability that the die will roll a 6 is…..”

Probability: Quite so simple?

Imagine you have an urn containing 20 red balls and 80 blue balls. You randomly draw two balls from the urn. After each draw, the ball is replaced and the urn is randomized.

What is the probability that both balls are red?

- The probability of drawing one is 20/(20+80) = 20/100 = 0.2

- The probability of drawing two is 0.2 x 0.2 = 0.04

Probability: Not quite so simple!

"In probability theory there is a very clever trick for handling a problem that becomes too difficult. We just solve it anyway by: making it still harder; redefining what we mean by 'solving' it, so that it becomes something we can do; inventing a dignified and technical-sounding word to describe this procedure, which has the psychological effect of concealing the real nature of what we have done, and making it appear respectable.

In the case of sampling with replacement, we apply this strategy as follows. Suppose that, after tossing the ball in, we shake up the urn. However complicated the problem was initially, it now becomes many orders of magnitude more complicated, because the solution now depends on every detail of the precise way we shake it, in addition to all the factors mentioned above.

We now assert that the shaking has somehow made all these details irrelevant, so that the problem reverts back to the simple one where the Bernoulli urn rule applies. We invent the dignified-sounding word randomization to describe what we have done. This term is, evidently, a euphemism, whose real meaning is: deliberately throwing away relevant information when it becomes too complicated for us to handle."

From: "Probability Theory: The Logic of Science", by ET Jaynes

Probability: Not quite so simple!

"We have described this procedure in laconic terms, because an antidote is needed for the impression created by some writers on probability theory, who attach a kind of mystical significance to it. For some, declaring a problem to be 'randomized' is an incantation with the same purpose and effect as those uttered by an exorcist to drive out evil spirits; i.e. it cleanses their subsequent calculations and renders them immune to criticism. We agnostics often envy the True Believer, who thus acquires so easily that sense of security which is forever denied to us.

[…] Shaking does not make the result 'random', because that term is basically meaningless as an attribute of the real world; it has no clear definition applicable in the real world. The belief that 'randomness' is some kind of real property existing in Nature is a form of the mind projection fallacy which says, in effect, 'I don't know the detailed causes – therefore – Nature does not know them.' What shaking accomplishes is very different. It does not affect Nature's workings in any way; it only ensures that no human is able to exert any willful influence on the result. Therefore, nobody can be charged with 'fixing' the outcome."

From: "Probability Theory: The Logic of Science", by ET Jaynes

Probability: Not quite so simple!

- Confused? You're in good company!

- Bertrand Russell: "Probability is the most important concept in modern science, especially as nobody has the slightest notion of what it means."

- John von Neumann: "Young man, in mathematics you don't *understand* things. You just *get used to them*."

Conditional and Marginal Probabilities

- A marginal probability is the general, overall probability of some outcome. It is ‘unconditional’.

- A conditional probability is the probability of some outcome given/if some condition is met.

    - “given” means “assuming that” or “conditional on”.

    - What’s the probability of you being asleep given its 4pm? 4am?

Conditional and Marginal Probabilities

- The marginal probability of some variable proposition is denoted by:

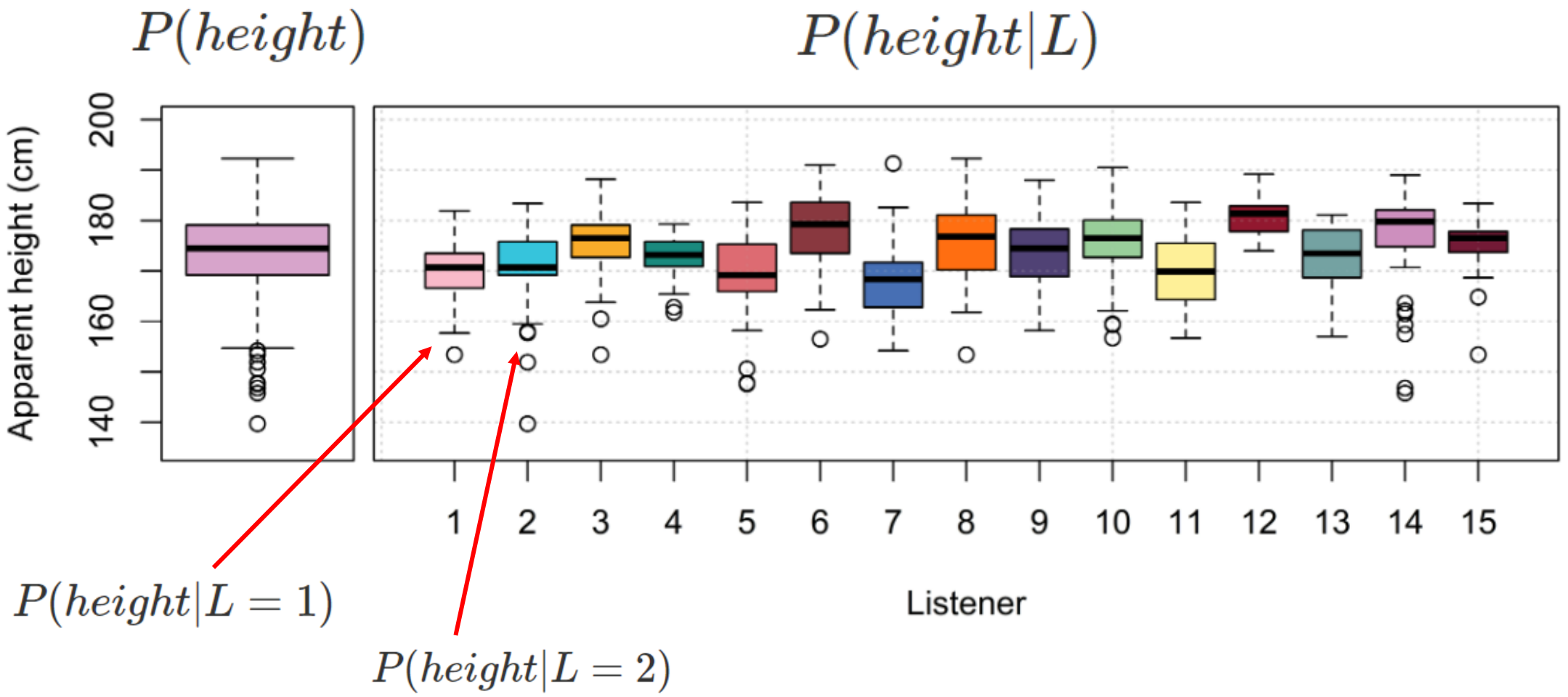
$$P(variable)$$

- The probability of that same variable, conditional on a second is often denoted by:

$$P(outcome\ variable | conditioning\ variable)$$

This means “given”.

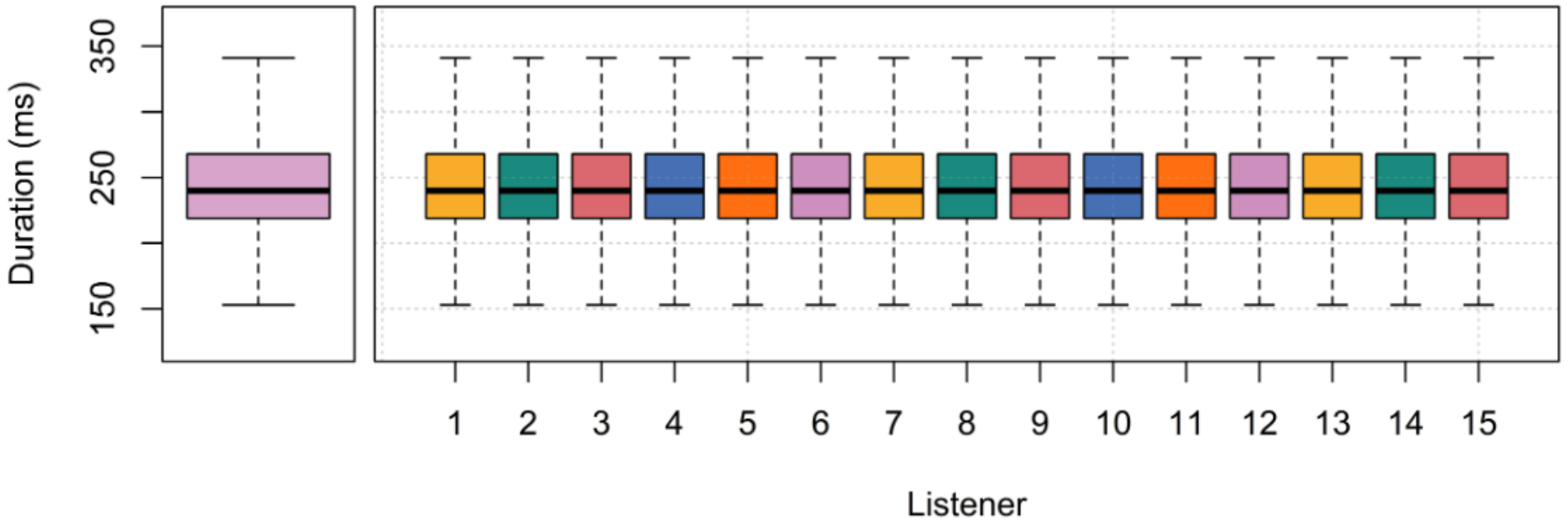
Conditional and Marginal Probabilities



Statistical Independence

- When the marginal distribution of a variable is the same as it's distribution conditional on some variable, it is <u>independent</u> from that variable.

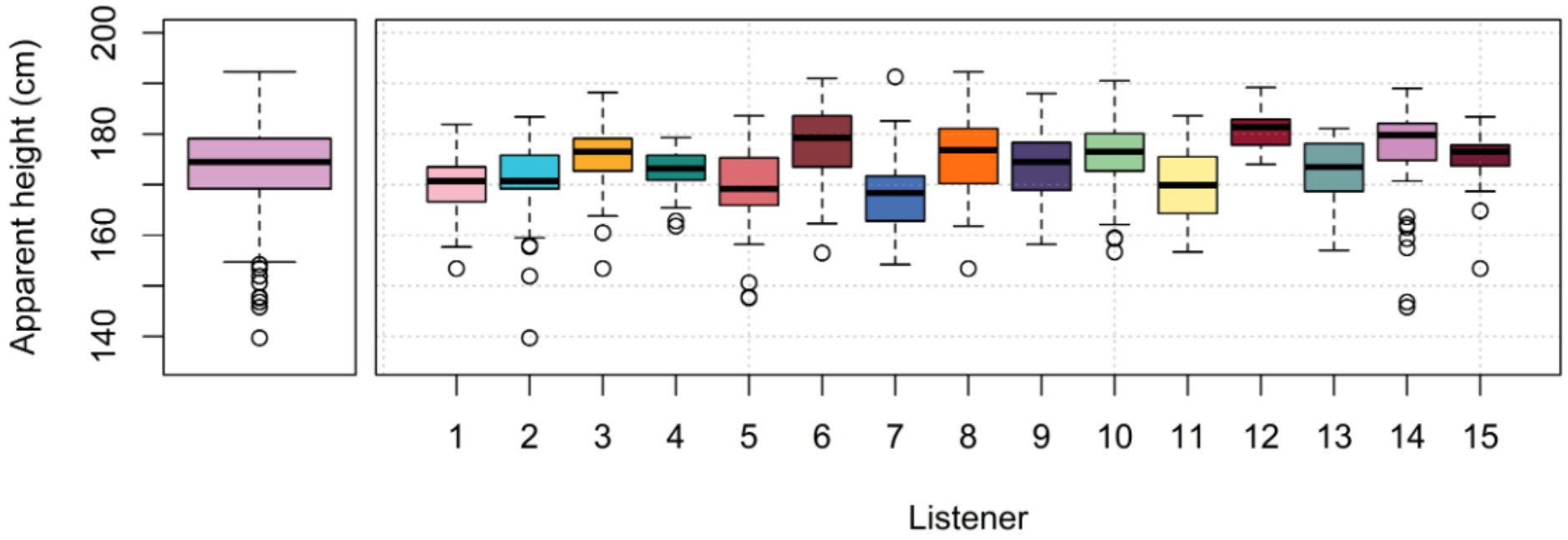
$$P(variable) = P(variable|conditioning\,variable)$$



Statistical Dependence

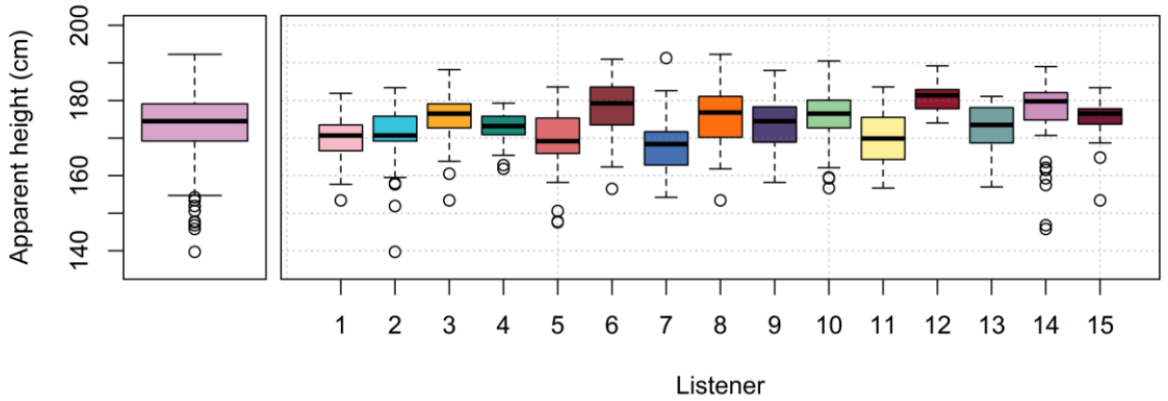
- When the marginal distribution of a variable is not the same as it's distribution conditional on some variable, it is dependent on that variable.

$$P(variable) \neq P(variable|conditioning\,variable)$$

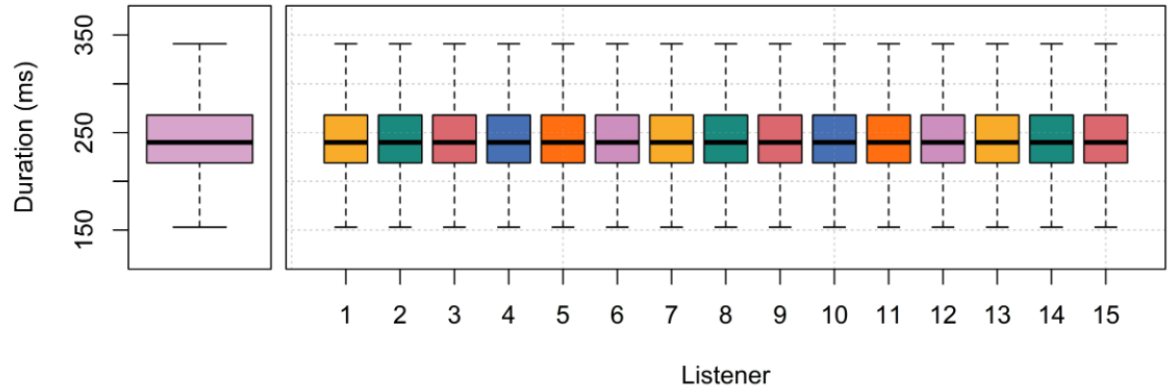


Statistical Dependence and Independence

$$P(variable) \neq P(variable|conditioning\,variable)$$



$$P(variable) = P(variable|conditioning\,variable)$$



Joint Probabilities

- A joint probably is the probability of multiple events happening together.

$$P(A \cap B) \text{ or } P(A\&B)$$

- Calculated by multiplying a conditional and marginal probability.

$$P(A \cap B) = P(A|B) \cdot P(B)$$

Joint Probabilities and Independence

- For independent variables, the conditional probability is equal to its marginal probability.

$$P(A|B) = P(A)$$

- As a result, the joint probability of independent events is the product of their marginal probabilities.

$$P(A|B) \cdot P(B) = P(A) \cdot P(B)$$

Joint Probabilities and Independence

$$\text{Independent events :}$$

$$P(A \,\&\, B \,\&\, C \,\&\, D) = P(A) \cdot P(B) \cdot P(C) \cdot P(D)$$

$$\text{Dependent events :}$$

$$P(A \,\&\, B \,\&\, C \,\&\, D) = P(A|B, C, D) \cdot P(B|C, D) \cdot P(C|D) \cdot P(D)$$

Probability Distributions

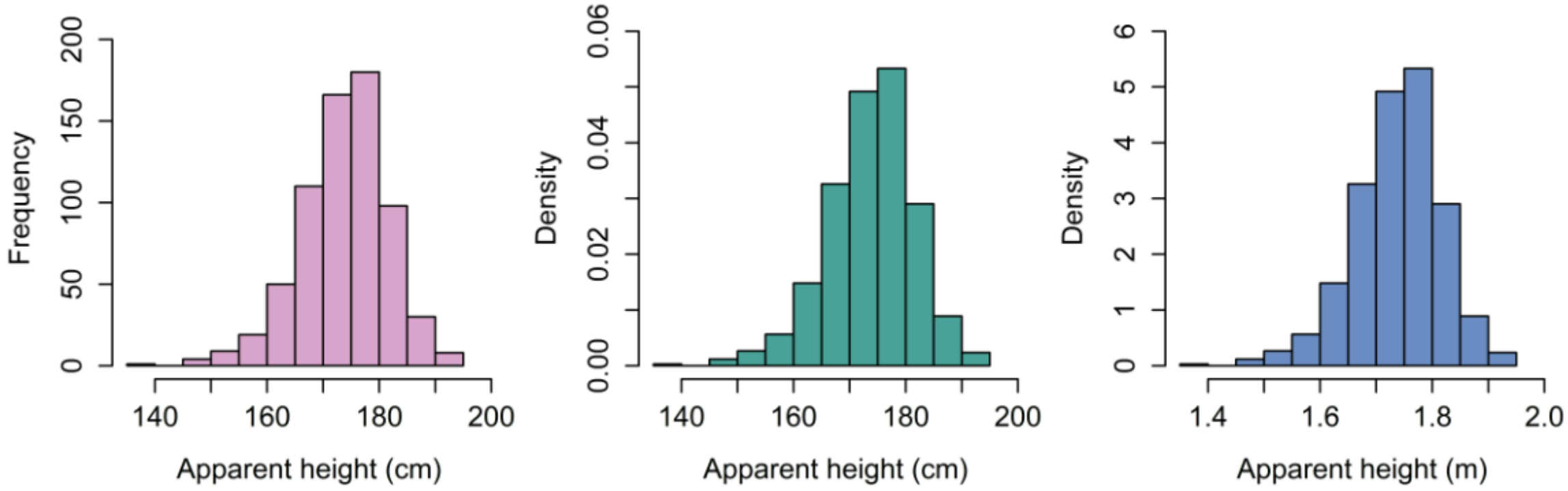
- Formally: A function that assigns probabilities to all values in a sample space.

$$f(x) = p \qquad P(x) = p$$

- Where x is some possible outcome or event.

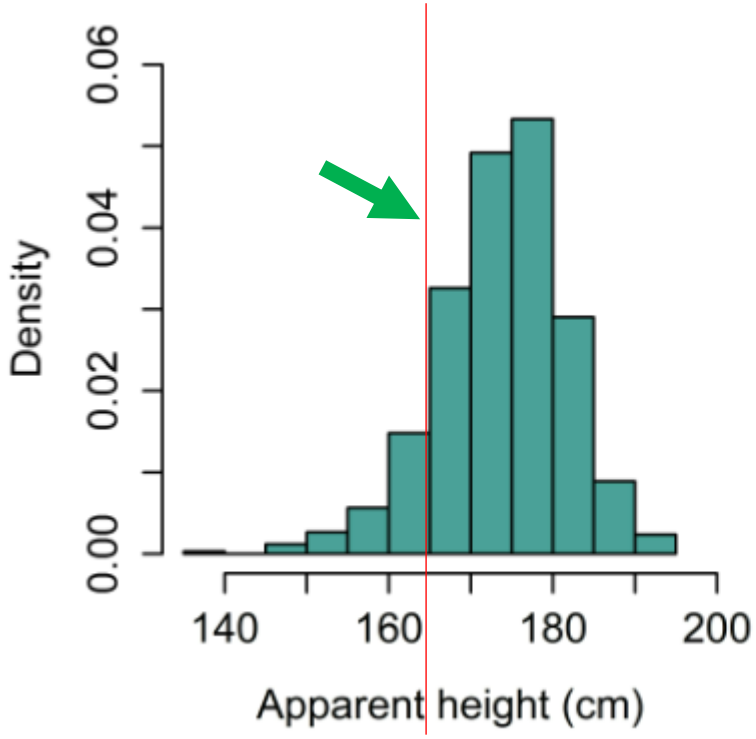
Probability Density

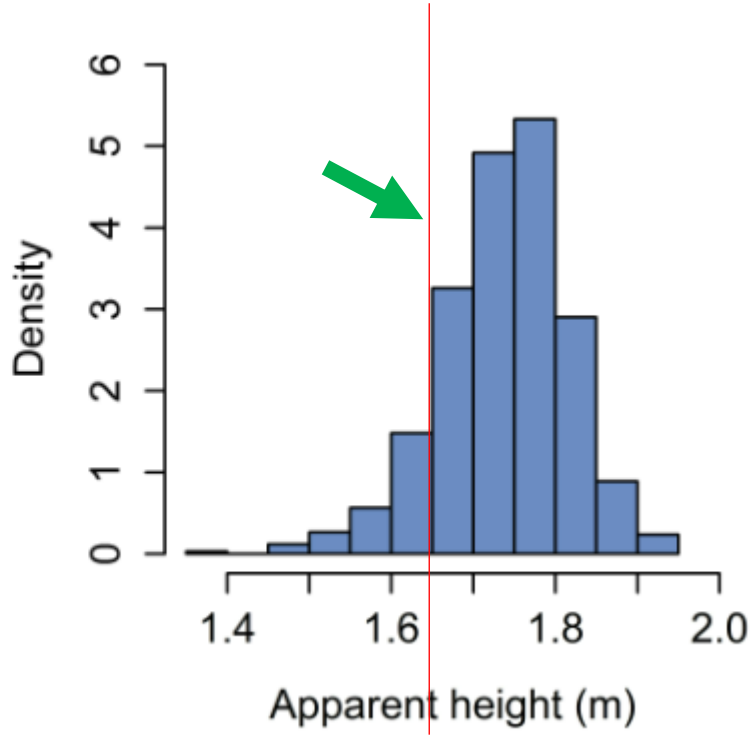
- Used to visualize/understand probability distributions.
- The amount of probability ‘stuff’ per unit of the variable.
- A curve with an area of the curve of 1.



Probability Density: Proportions

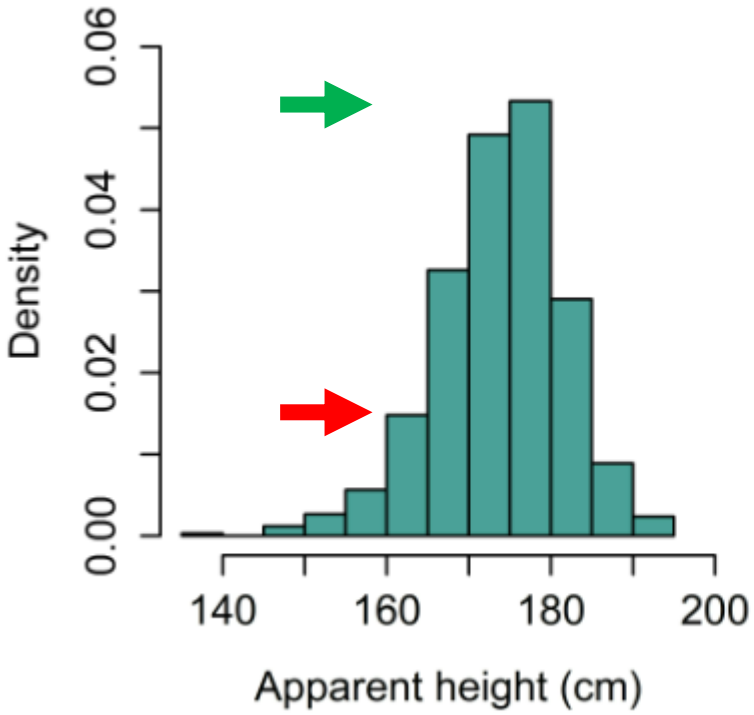
- The area under the curve between two points reflects the probability that the variable will have a value in that range.

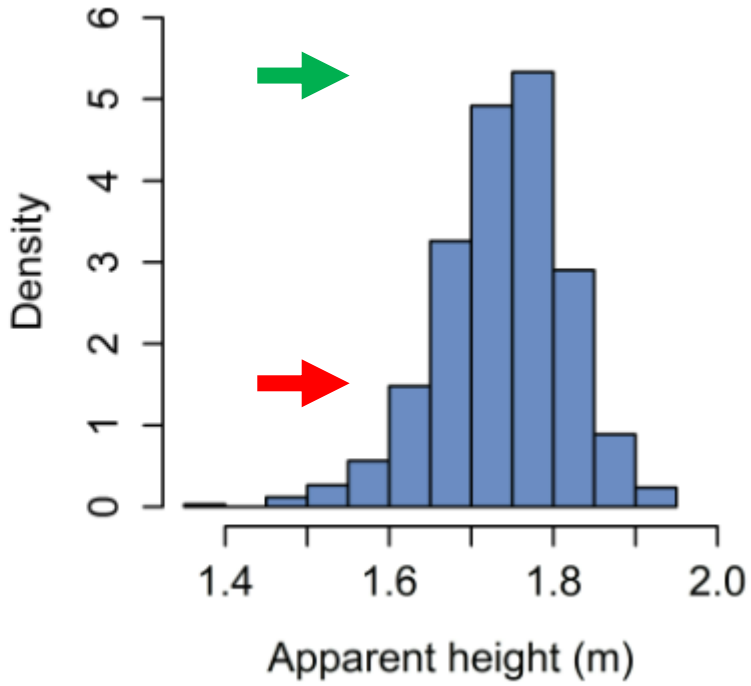




Probability Density: Relative Values

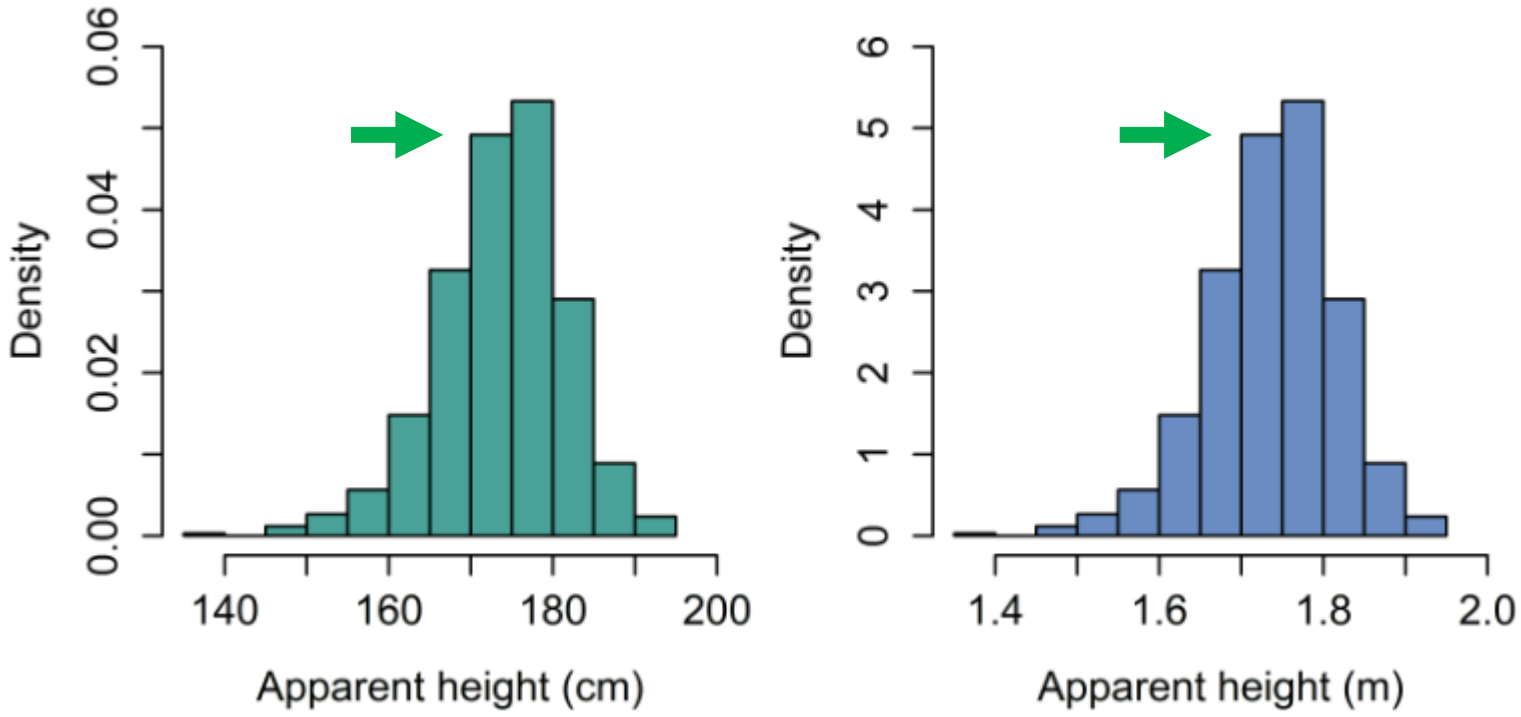
- If the density is x times higher/lower for one values than another, that value is x times more/less likely.





Probability Density: No Absolute Interpretation

- The absolute value of the curve at any given point is not the probability.
- It should not be interpreted in isolation.



Parametric Probability Distributions

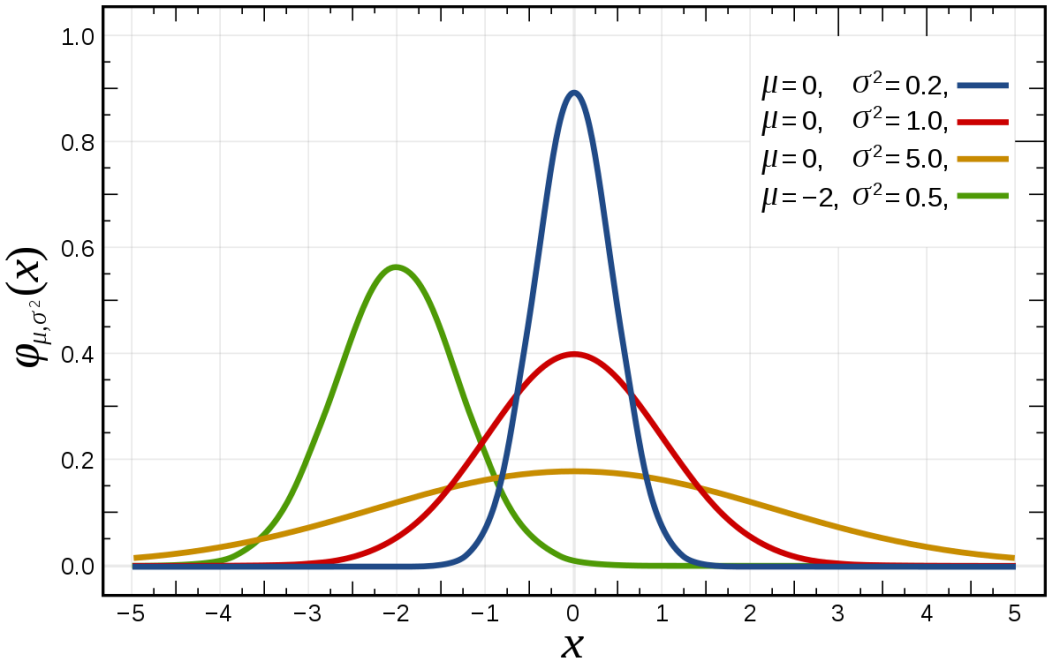
- Some probability distributions have relatively simple behavior.
- This behavior can be predicted using a small set of parameters.
- Think of these like the shapes defined by:

$$y = a + b \cdot x \qquad y = a + b(c - x)^2$$

The Normal Distribution

- Perhaps the most known and used distribution.
- Two parameters: The mean (μ) and the standard deviation (σ).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$



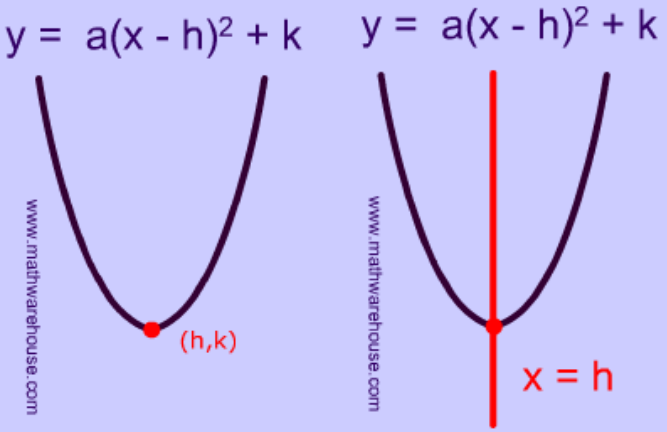
Quadratic Equations

Vertex Form of Equation

The vertex form of a parabola's equation is generally expressed as :

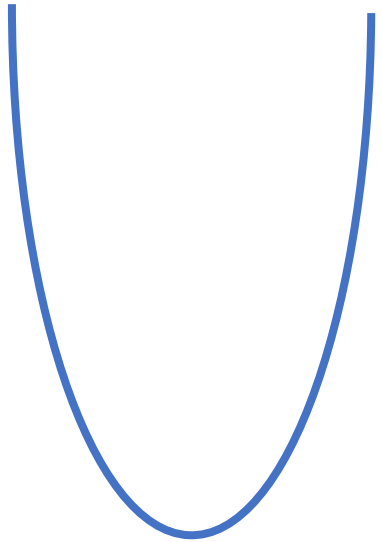
$y = a(x-h)^2+k$

- (h,k) is the vertex as you can see in the picture below



- If a is positive then the parabola opens upwards like a regular "U".
- If a is negative, then the graph opens downwards like an **upside down** "U".

- If |a| < 1, the graph of the parabola widens. This just means that the "U" shape of parabola stretches out sideways . **Explore the way that 'a' works using our interactive parabola grapher** .
- If |a| > 1, the graph of the graph becomes narrower(The effect is the opposite of |a| < 1).



$$-\frac{1}{2\sigma^2}(x-\mu)^2$$

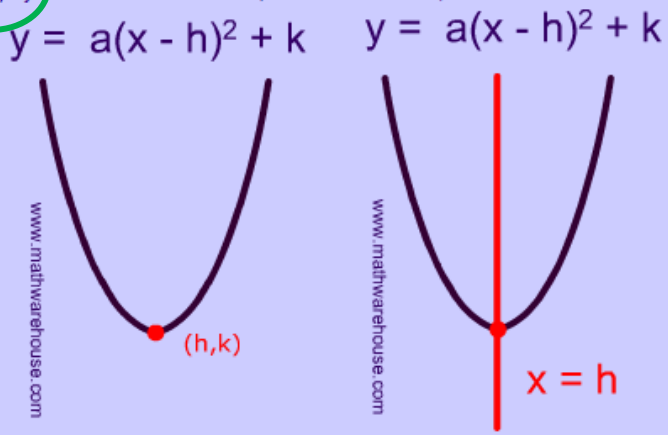
Quadratic Equations

Vertex Form of Equation

The vertex form of a parabola's equation is generally expressed as :

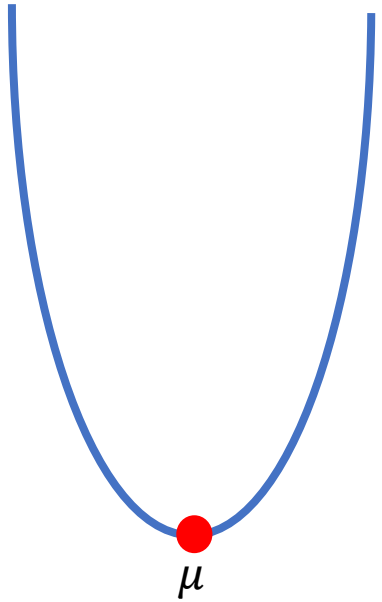
$y = a(x-h)^2+k$

- (h,k) is the vertex as you can see in the picture below



- If a is positive then the parabola opens upwards like a regular "U".
- If a is negative, then the graph opens downwards like an **upside down** "U".

- If |a| < 1, the graph of the parabola widens. This just means that the "U" shape of parabola stretches out sideways . **Explore the way that 'a' works using our interactive parabola grapher** .
- If |a| > 1, the graph of the graph becomes narrower(The effect is the opposite of |a| < 1).



$$-\frac{1}{2\sigma^2}(x-\mu)^2$$

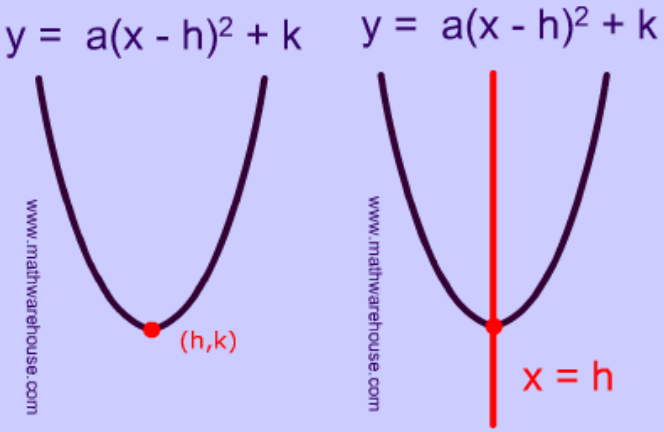
Quadratic Equations

Vertex Form of Equation

The vertex form of a parabola's equation is generally expressed as :

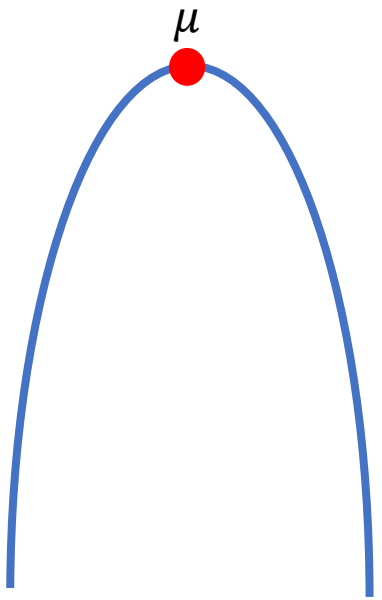
$y = a(x-h)^2+k$

- (h,k) is the vertex as you can see in the picture below



- If a is positive then the parabola opens upwards like a regular "U".
- If a is negative, then the graph opens downwards like an **upside down** "U".

- If |a| < 1, the graph of the parabola widens. This just means that the "U" shape of parabola stretches out sideways . **Explore the way that 'a' works using our interactive parabola grapher** .
- If |a| > 1, the graph of the graph becomes narrower(The effect is the opposite of |a| < 1).



$$-\frac{1}{2\sigma^2}(x-\mu)^2$$

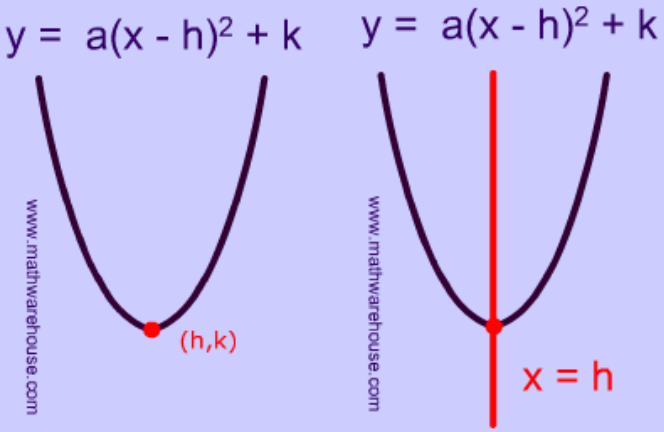
Quadratic Equations

Vertex Form of Equation

The vertex form of a parabola's equation is generally expressed as :

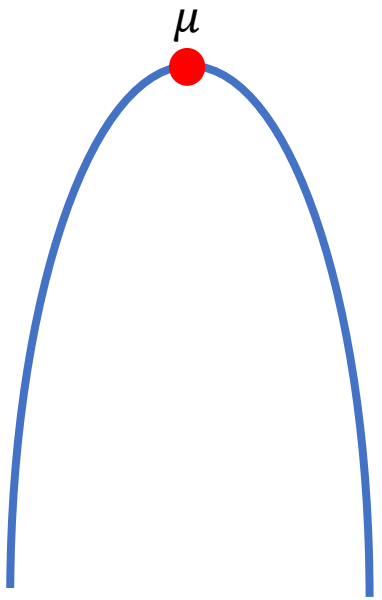
$y = a(x-h)^2+k$

- (h,k) is the vertex as you can see in the picture below



- If a is positive then the parabola opens upwards like a regular "U".
- If a is negative, then the graph opens downwards like an **upside down** "U".

- If |a| < 1, the graph of the parabola widens. This just means that the "U" shape of parabola stretches out sideways . **Explore the way that 'a' works using our interactive parabola grapher** .
- If |a| > 1, the graph of the graph becomes narrower(The effect is the opposite of |a| < 1).



$$-\frac{1}{2\sigma^2}(x-\mu)^2$$

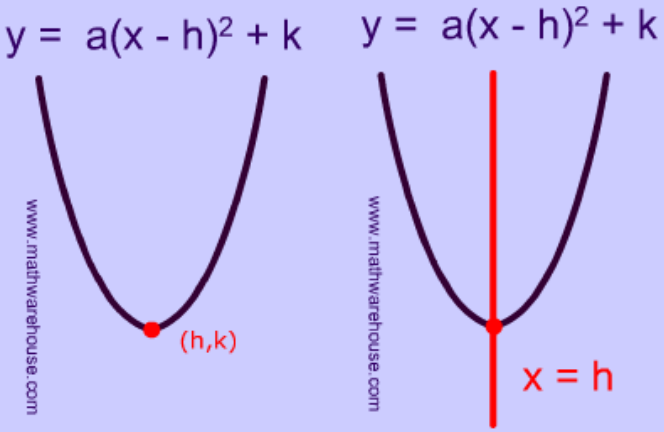
Quadratic Equations

Vertex Form of Equation

The vertex form of a parabola's equation is generally expressed as :

$y = a(x-h)^2+k$

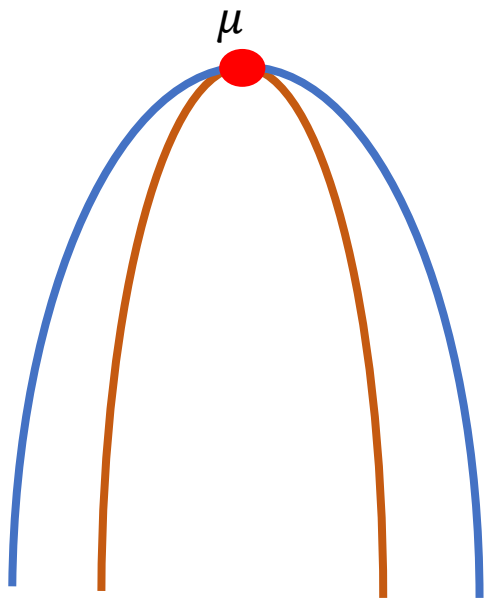
- (h,k) is the vertex as you can see in the picture below



- If a is positive then the parabola opens upwards like a regular "U".
- If a is negative, then the graph opens downwards like an **upside down** "U".

- If |a| < 1, the graph of the parabola widens. This just means that the "U" shape of parabola stretches out sideways . **Explore the way that 'a' works using our interactive parabola grapher** .
- If |a| > 1, the graph of the graph becomes narrower(The effect is the opposite of |a| < 1).

$\sigma < \sigma$

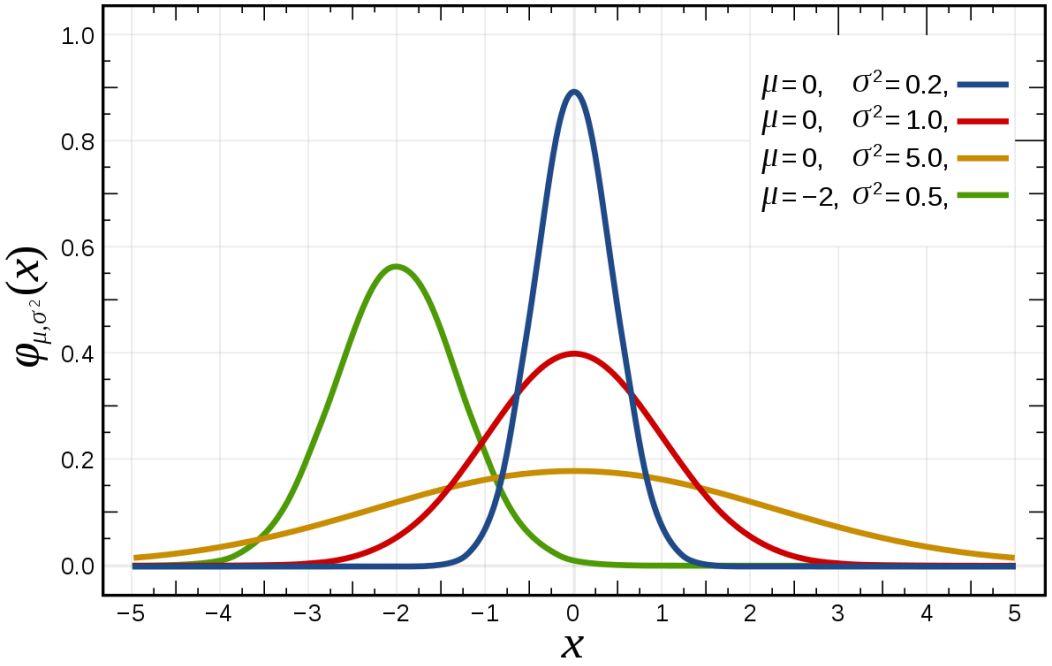


$$-\frac{1}{2\sigma^2}(x-\mu)^2$$

Mean = Location

- The mean of the distribution is the expected/average value.
- We can estimate this using the sample mean below.

$$\hat{\mu}_y = \sum_{i=1}^{n} y_{[i]}/n$$

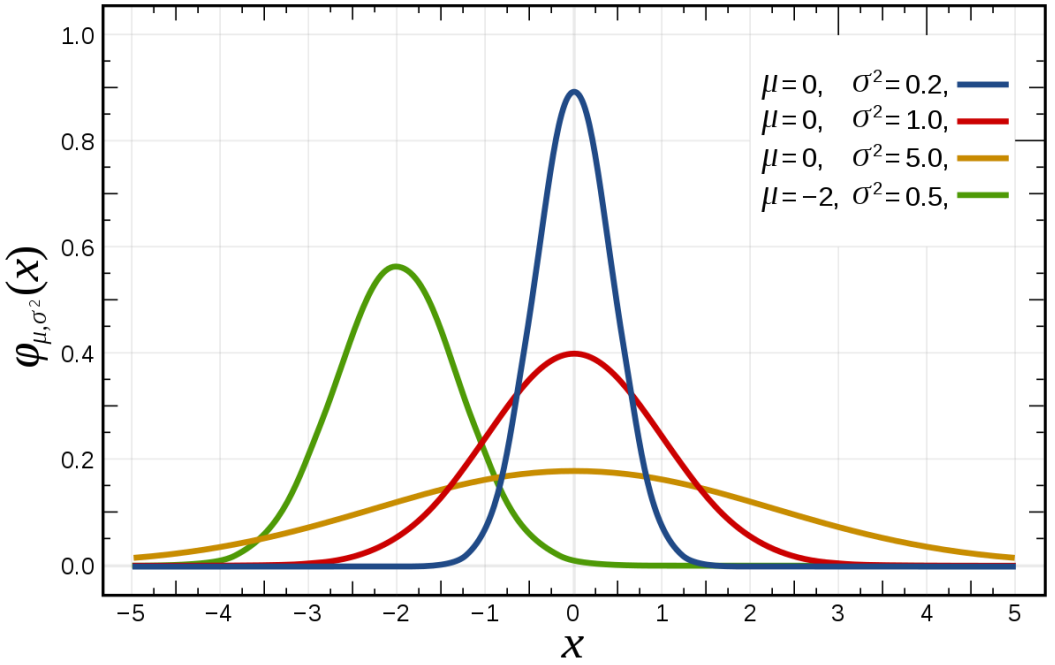


Standard Deviation = Spread

- The variance of the distribution is the expected/average value of squared deviations from the mean.
- We can estimate this using the sample variance below.

$$\hat{\sigma}_y^2 = \sum_{i=1}^{n} (y_{[i]} - \hat{\mu}_y)^2 / (n-1)$$

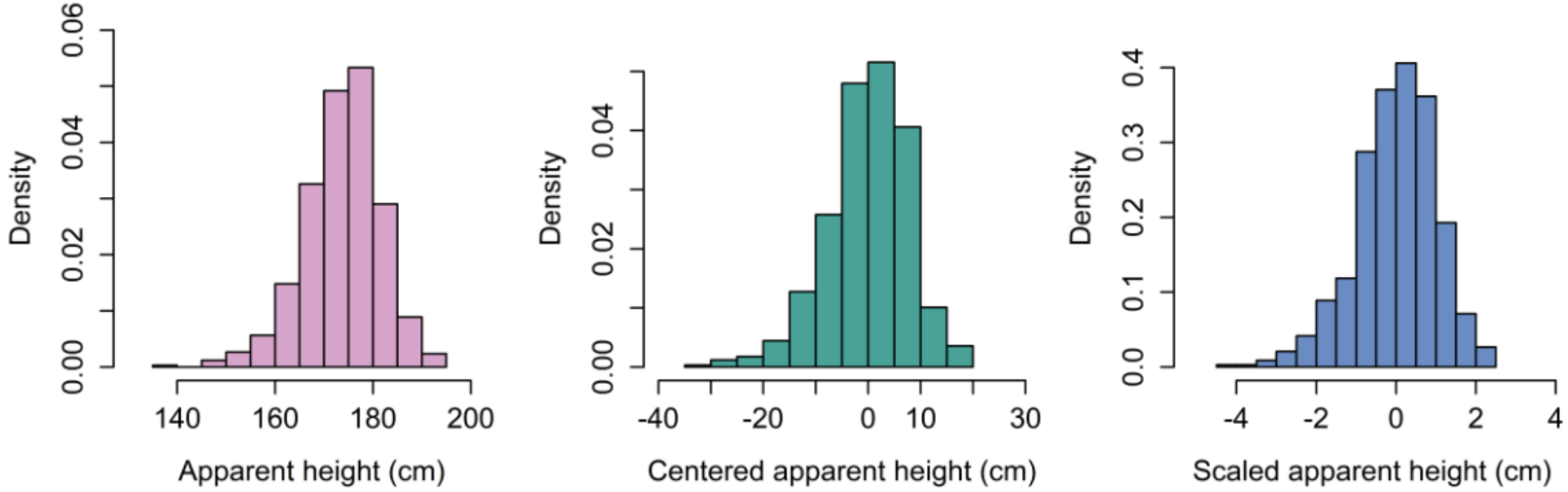
$$\hat{\sigma}_y = \sqrt{\hat{\sigma}_y^2} = \sqrt{\sum_{i=1}^{n} (y_{[i]} - \mu_y)^2 / (n-1)}$$



The Standard Normal Distribution

- A standard normal has $\mu = 0$ and $\sigma = 1$.

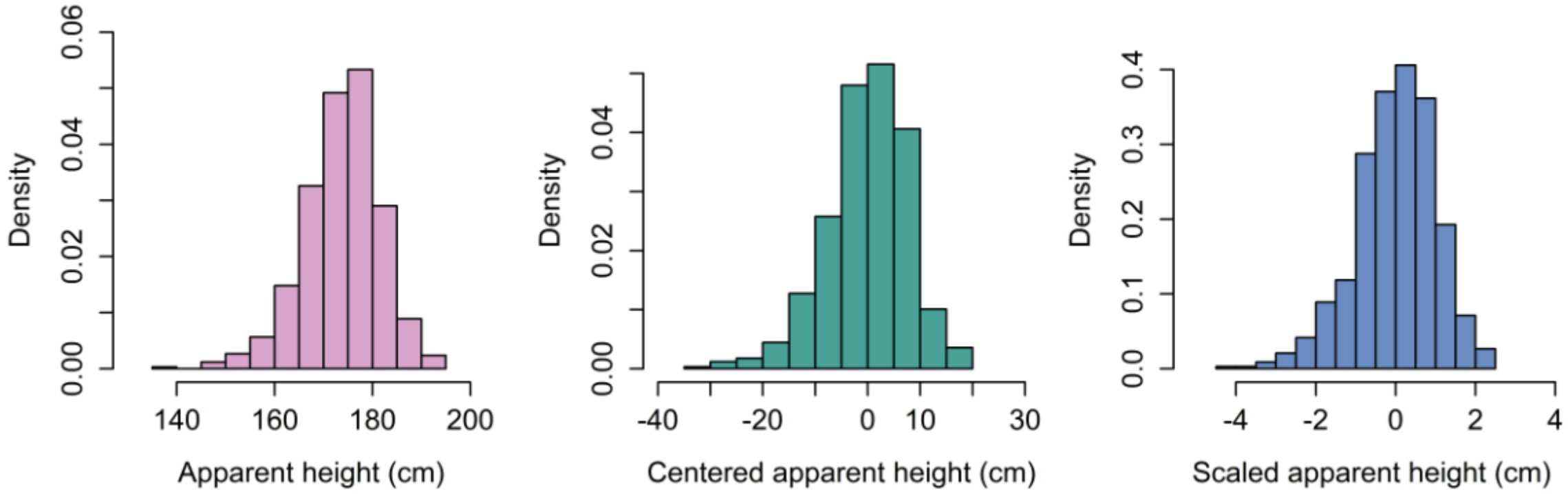
$$z = (x - \mu)/\sigma \qquad x = z \cdot \sigma + \mu$$



Standardization (z scores)

This is useful because:

- All normal data is the same when expressed in standardized units.
- Probabilities for standard quantiles can be learned easily.



Models and Inference

- Exact model: The ‘real’ model that leads to Truth. Relies on actual knowledge of all underlying processes and relations.

- Approximate models: Models good enough for some purpose. May be wrong in unpredictable ways.

We only have access to these.

Theoretical Probabilities

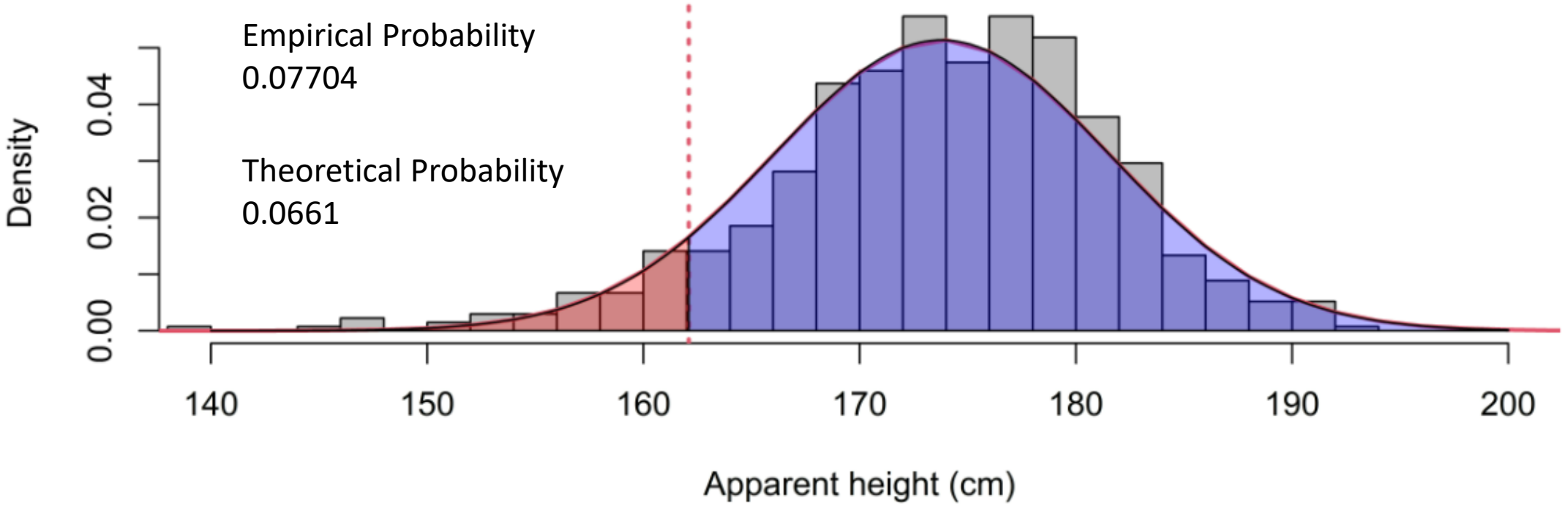
- Theoretical probabilities: based on models about the world and not on observations.

- Statistical models are mostly based on theoretical probabilities.

- Since our models are only approximate and never exact, our theoretical probabilities can also only ever be approximate and never exact.

Theoretical Probabilities

- The theoretical probability is reliable if and if and if and if….



Likelihood

- The joint probability (density) of observing the data you observed, given specific parameter values.

$$\mathcal{L}_{(\mu|x)} = \mathrm{P}(\mathrm{x}|\mu)$$

- For multiple independent data points, we can just multiply the density over each point.

$$\mathcal{L}_{(\mu|x)} = \mathrm{P}(\mathrm{x}_1\,, x_2, \ldots, x_n|\mu)$$

$$\mathcal{L}_{(\mu|x)} = \mathrm{P}(\mathrm{x}_1|\mu) \cdot \mathrm{P}(\mathrm{x}_1|\mu) \cdot \ldots \cdot \mathrm{P}(\mathrm{x}_n|\mu)$$

Likelihood

- Below we have the joint density of x_1 and x_2, assuming these are independent.

$$f(x_1, x_2) = [\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{1}{2\sigma^2}(x_1 - \mu)^2)] \cdot [\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{1}{2\sigma^2}(x_2 - \mu)^2)] \qquad (2.13)$$

- If we consider this as a function of the value of \mu, we have the function for the likelihood of the mean.

$$\mathcal{L}_{(\mu|x)} = [\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{1}{2\sigma^2}(x_1 - \mu)^2)] \cdot [\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{1}{2\sigma^2}(x_2 - \mu)^2)] \qquad (2.14)$$

Likelihood

- If we assume all our data is independent, the likelihood can be calculated by multiplying a bunch of independent probabilities.

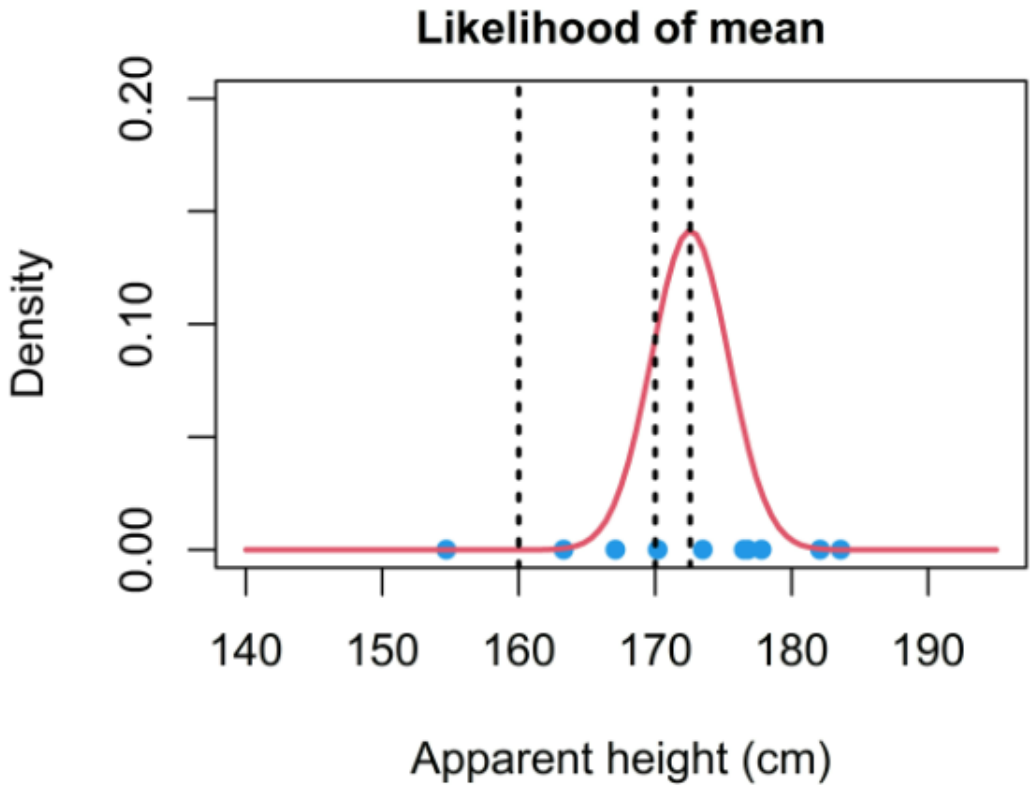
$$\mathcal{L}_{(\mu|x)} = [\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{1}{2\sigma^2}(x_1-\mu)^2)] \cdot [\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{1}{2\sigma^2}(x_2-\mu)^2)] \qquad (2.14)$$

- If our data is not independent, calculating likelihoods can be much more complicated.

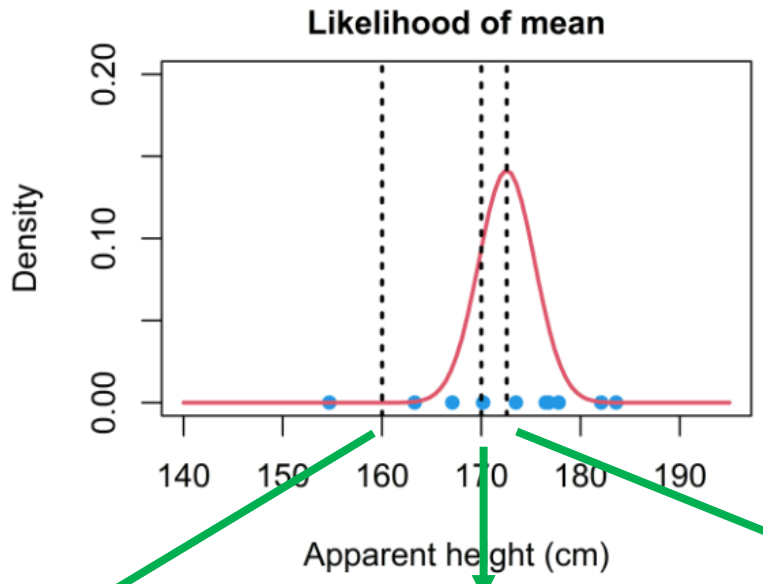
The Likelihood Function

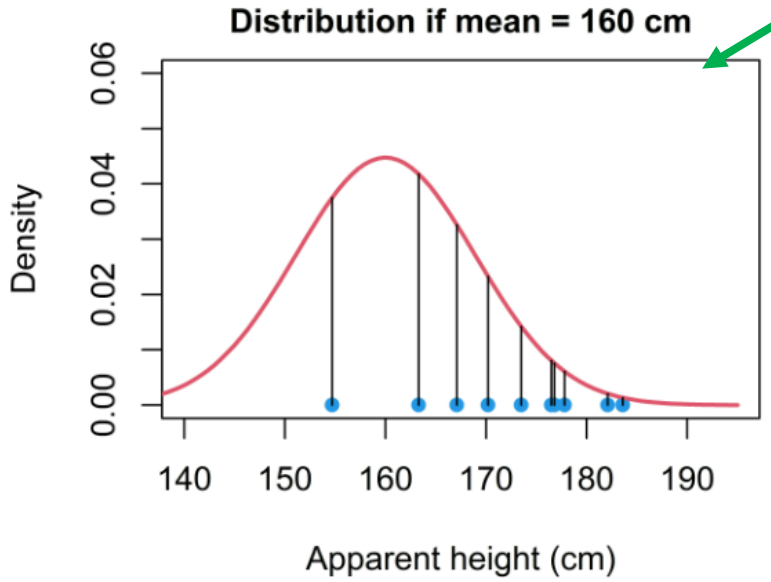
- A function that shows the likelihood for different values of the parameter(s) of interest.

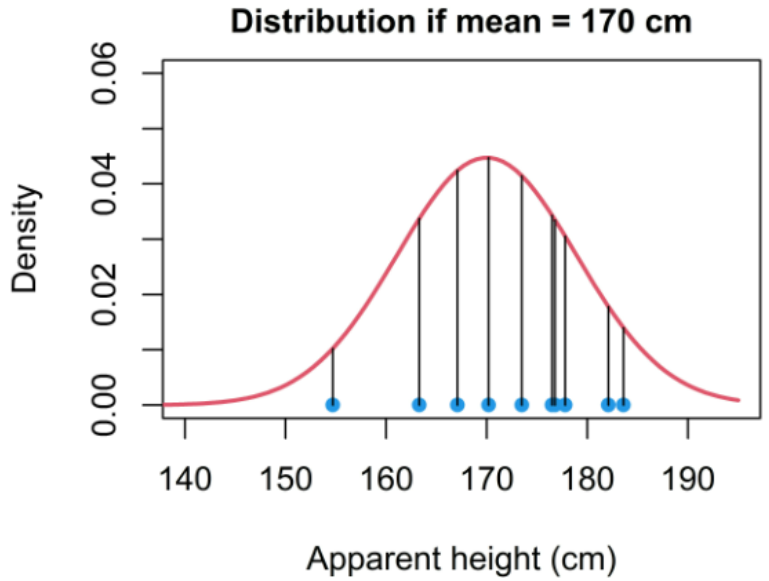
- Absolute values are not terribly useful, relative values matter more.

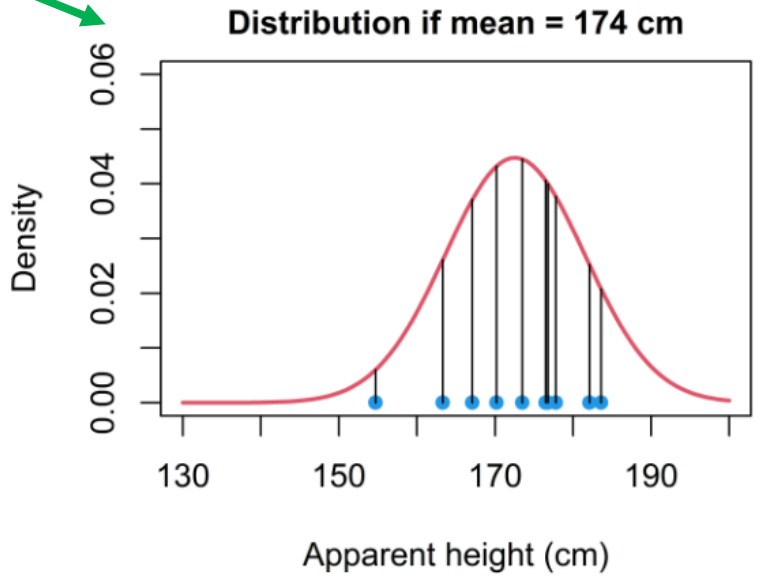


Calculating Likelihoods









Probability and Likelihood

- Probability and likelihood are the inverse of each other.

- Probabilities express the expected relative frequency of observed data, given some parameter values.
  - For probabilities the parameters are fixed and the data is variable.

- Likelihoods express the relative credibility of parameters, given some observed data.
  - For likelihoods the data are fixed and the parameters are variable.

Probability and Likelihood: Example

- You watch your friend shoot 100 three pointers, and they make 12 of them. They claim they usually make about 50%.
- Do you believe them? Probably not. Why?
- If someone really makes 50% of their shots (fixed parameter), it is improbable that they will only make 12% one day (variable data).
- Someone who made only 12% of their shots one day (fixed data) is unlikely to have an underlying ability of 50% (variable parameter).

Characteristics of Likelihoods

- Lots of detail in the book.
- The important part is that likelihoods get narrower under two conditions:
  - The number of observations goes up
  - The error goes down.

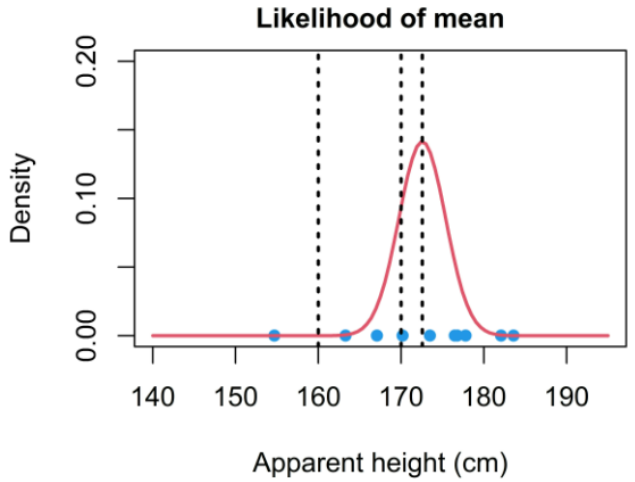
Likelihood and Inference

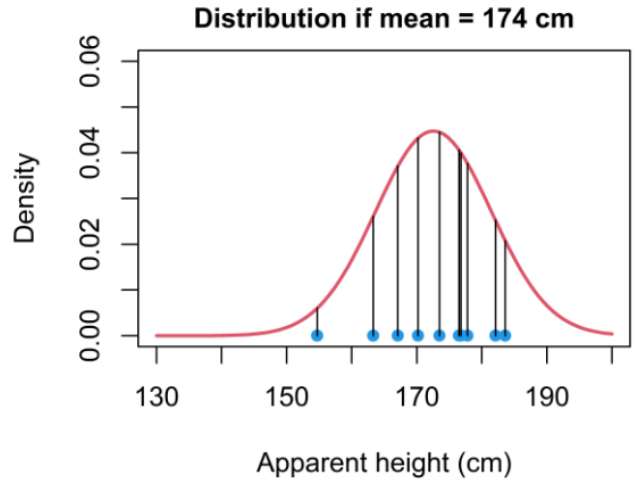
- You usually want to make inferences about parameter values (i.e. *hypotheses*), not about specific results (i.e. *data*).

- Thus, you usually make inferences based on likelihoods rather than probabilities.

- Statistical models center on the interpretation of parameter likelihoods and closely related values.

Calculating the Likelihood for our Data






```r
# make candidates for mean parameter
mus = seq (172.5,175, .01)


# easy way to make zero vector of same length as above
log_likelihood = mus*0


# add the log-density of all observations. Notice only the
# mean changes across iterations of the for loop.
for (i in 1:length(mus)) log_likelihood[i] =
  sum (dnorm (mens_height, mus[i], sd(mens_height), log = TRUE))
```

Logarithms

- Turn multiplication to addition and exponentiation into multiplication.
- The probability of of 675 events with P=0.1 is:

$$0.1^{675}$$

- Or….

$$\log(0.1) = -2.3$$

$$\log(0.1^{675}) = \log(0.1) \cdot 675 = -1554.25$$

$$\log\left(e^{x}\right) = x$$

$$\log(1) = 0$$

$$\text{if } x < 1, \log(x) < 0$$

$$\text{if } x > 1, \log(x) > 0$$

$$\text{if } x < 0 \ \log(x) = \text{undefined}$$

$$\log\left(x^{y}\right) = \log(x) \cdot y$$

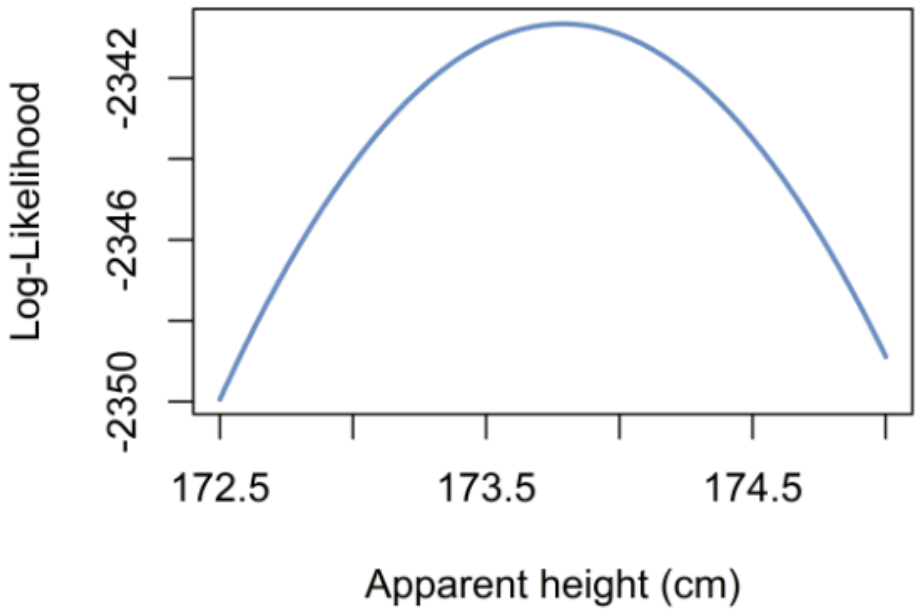
$$\log\left(\sqrt[y]{x}\right) = \log(x) / y$$

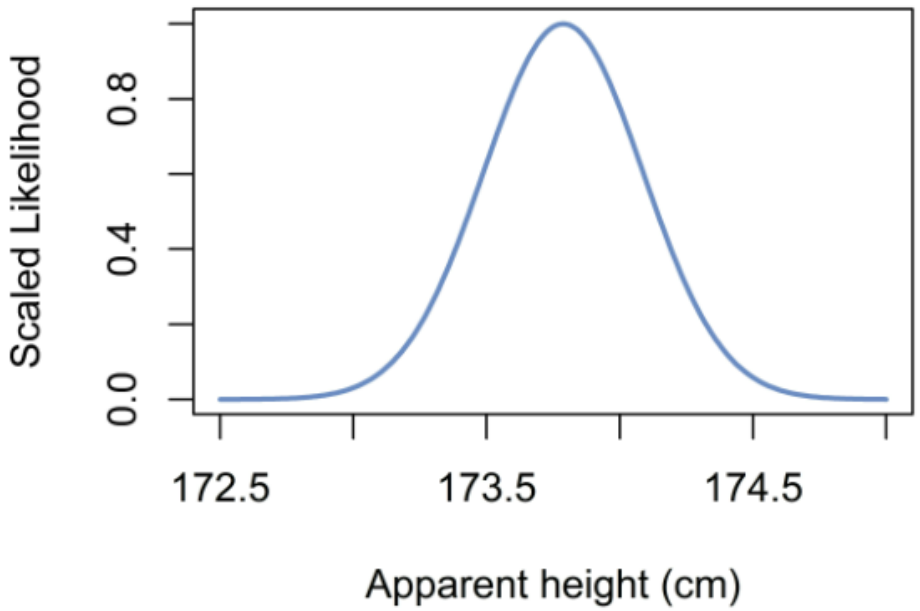
$$\log(x) + \log(y) = \log(x \cdot y)$$

$$\log(x) - \log(y) = \log(x / y)$$

Log-Likelihoods

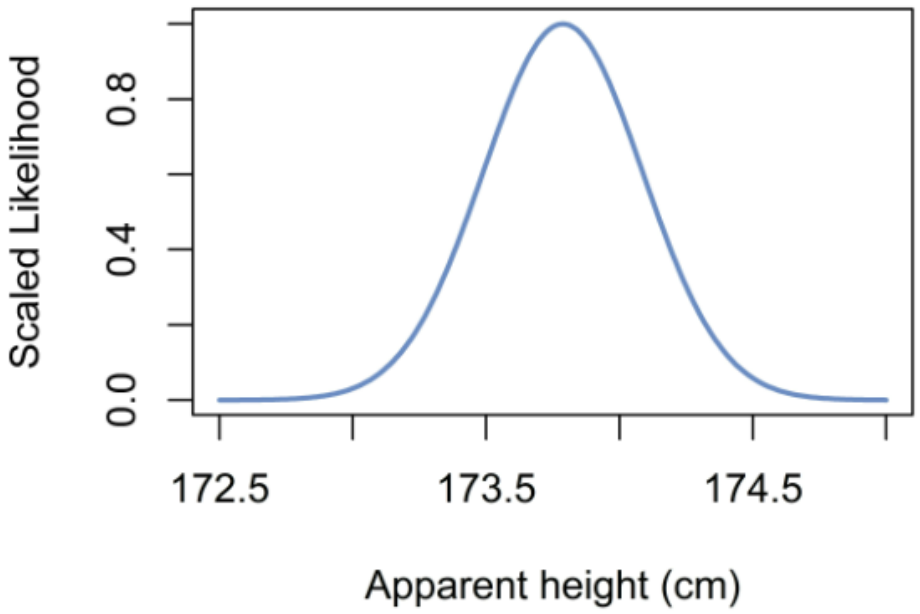
- As a practical matter, likelihoods often need to be calculated as log likelihoods.





Likelihoods and Inference

- We infer likely values of the mean parameter given our data using its likelihood function.




```
# find index number of highest values in log-likelihood
maximum = which.max(scaled_log_likelihood)


# print and compare to sample mean
mus[maximum]
## [1] 173.8
mean (mens_height)
## [1] 173.8
```

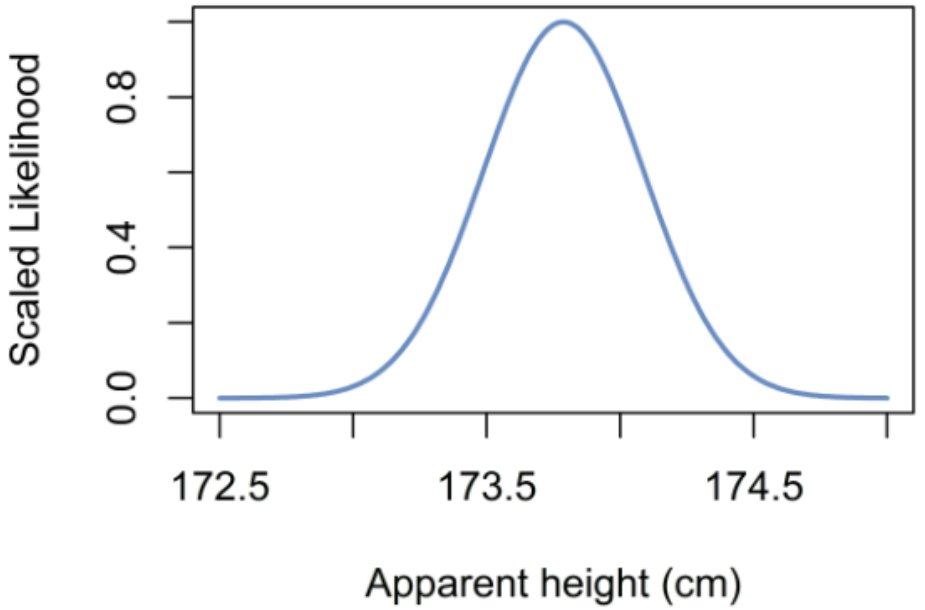
Answering our Research Questions

(Q1) How tall does the average adult male ‘sound’?

A1: About 173.8 cm tall.

(Q2) Can we set limits on credible average apparent heights based on the data we collected?

A2: Yes, between 173 and 174.5 cm.



Exercises

- Use the techniques in chapter 2 to discuss reasonable values for the mean and standard deviation of some variable.

- Write a report using a qmd file. Submit the report and the qmd file.

- Include at least two plots and describe the information they represent.