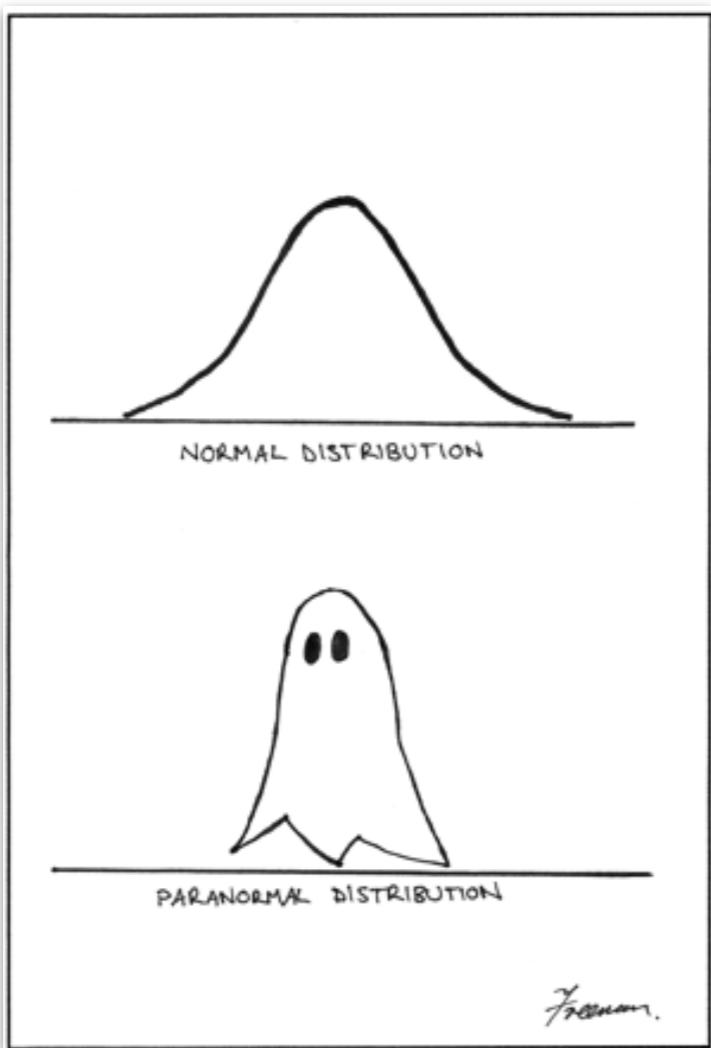


Simulation 1



Chat

What do you think
is the ideal age
(and why)?

To: Everyone ▾ More ▾

Type message here...

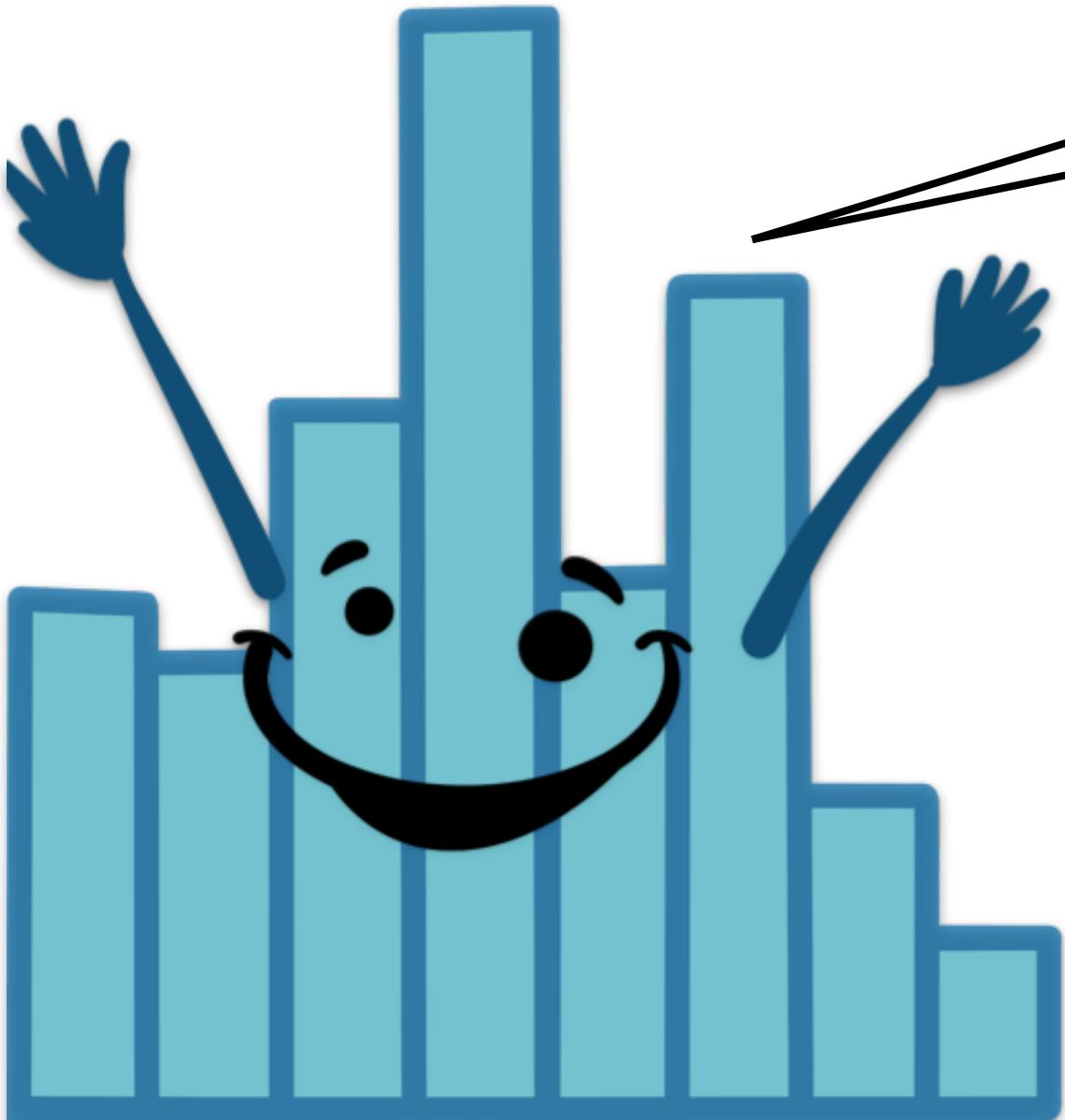
COLLABORATIVE PLAYLIST
psych252

<https://tinyurl.com/psych252spotify21>

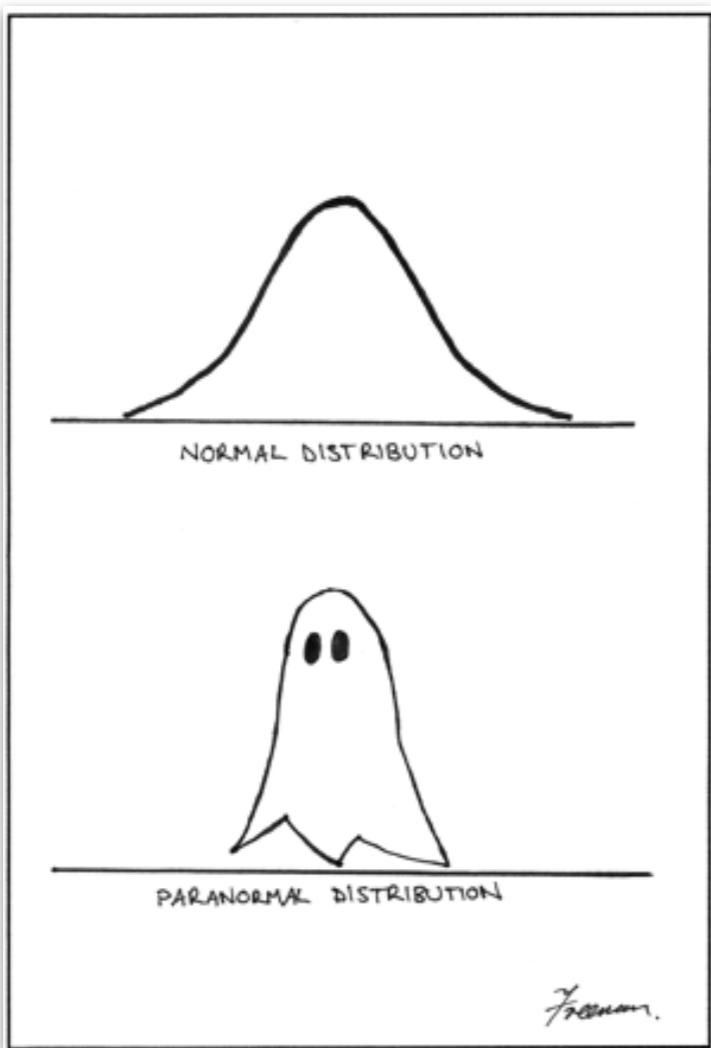
PLAY ...

01/27/2021

Remember to
record the
lecture!



Simulation 1



Chat

What do you think
is the ideal age
(and why)?

To: Everyone ▾ More ▾

Type message here...

O COLLABORATIVE PLAYLIST
psych252

<https://tinyurl.com/psych252spotify21>

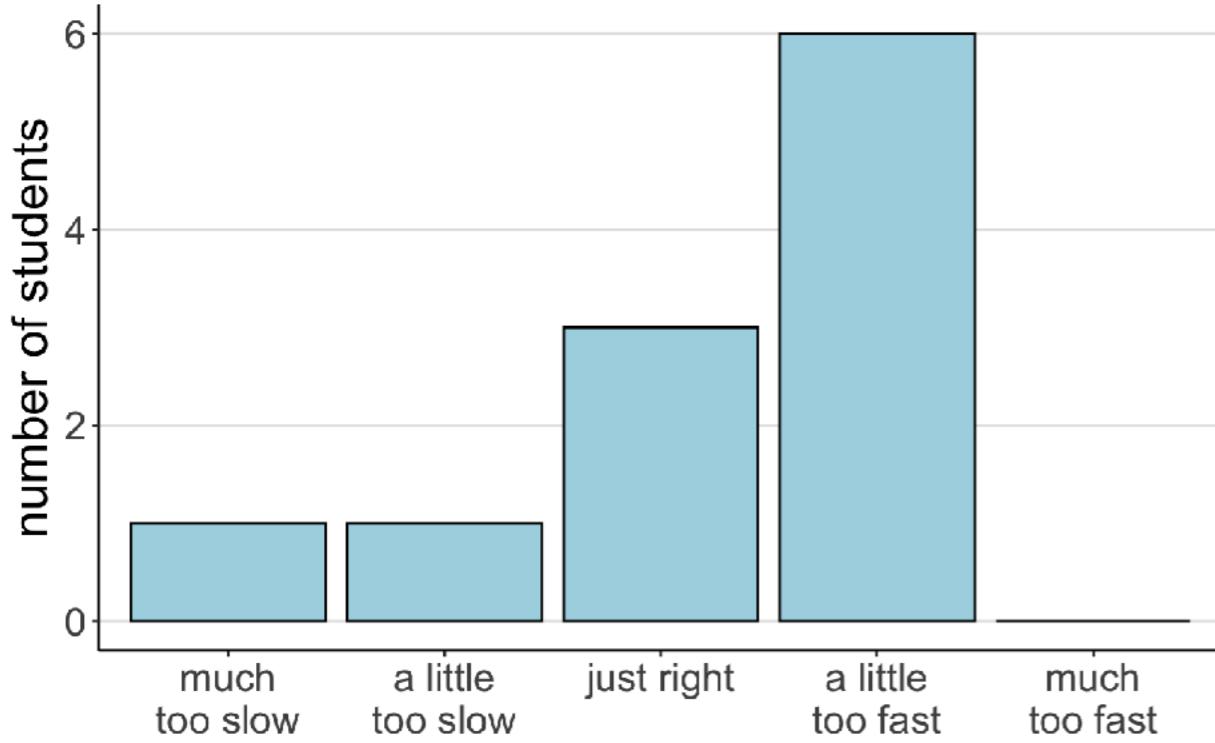
PLAY ...

01/27/2021

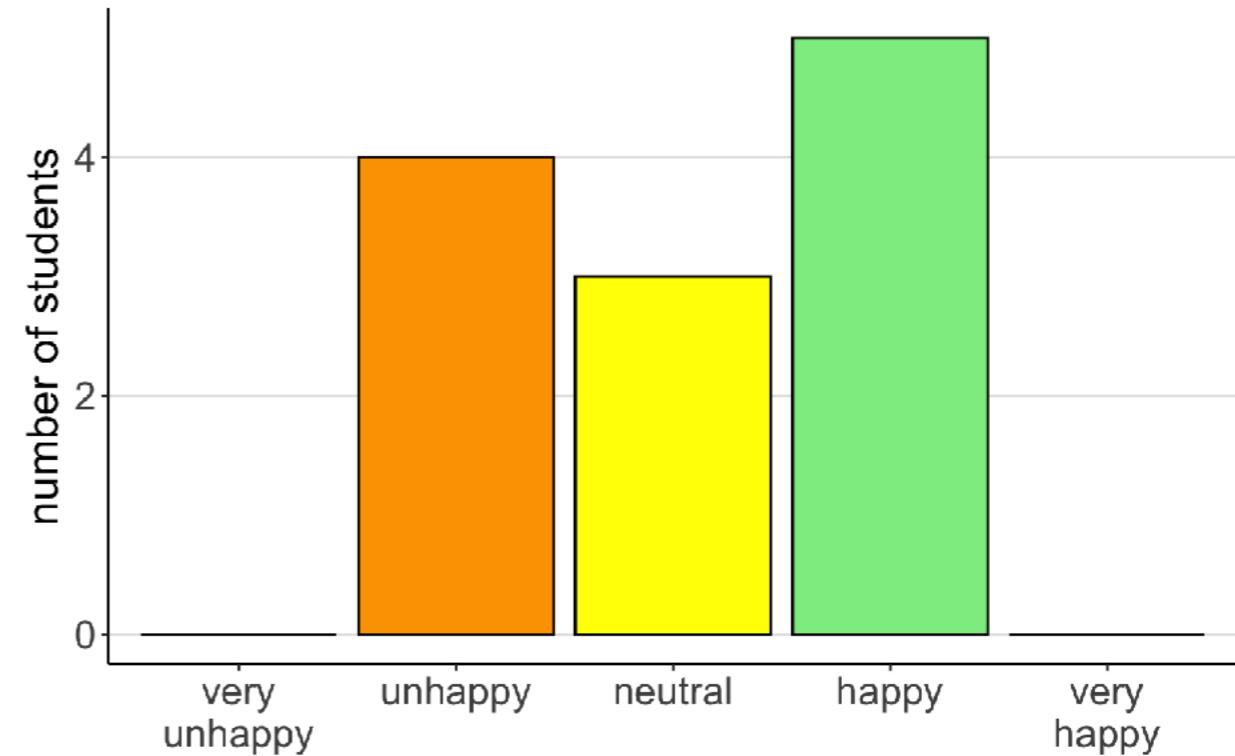
Your feedback

Your feedback

How was the pace of today's class?



How happy were you with today's class overall?

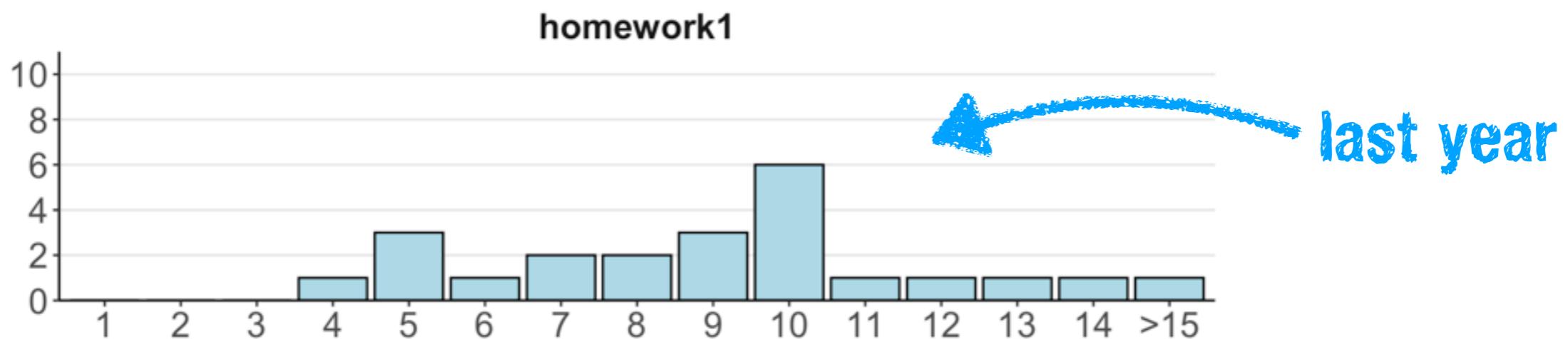
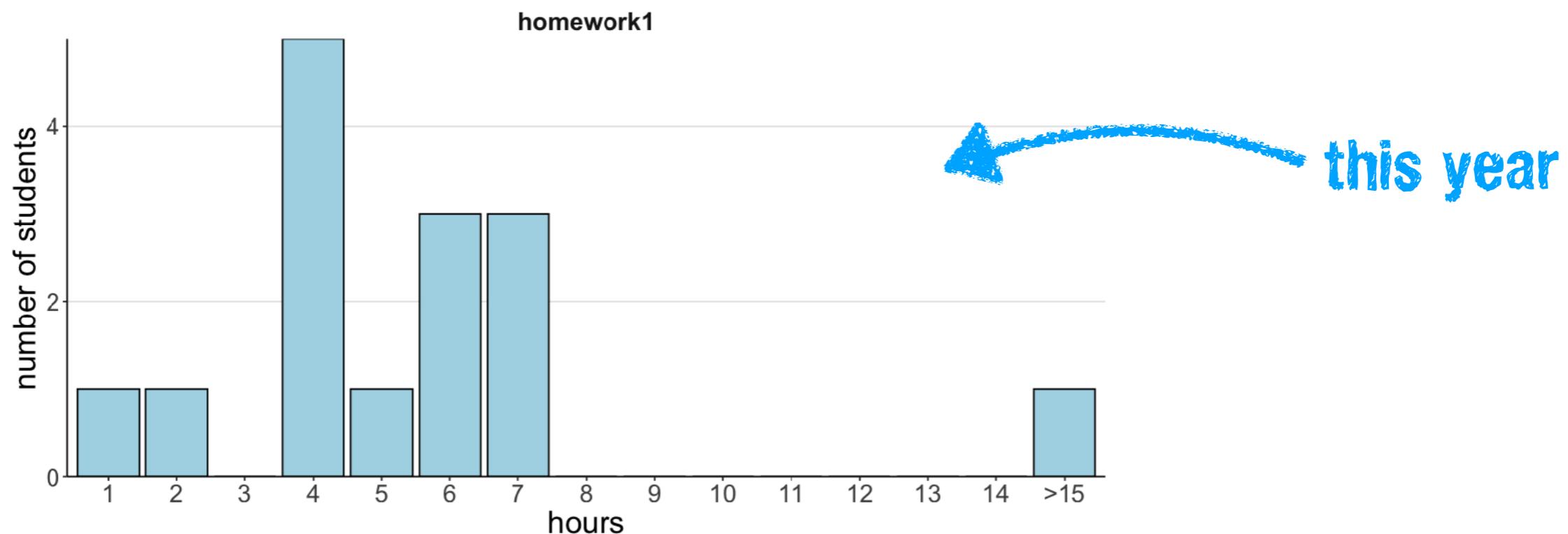


I really liked the fact that you walked us through the bayesian formula using the ipad drawing/writing feature

written comments were mostly positive, please also comment what you didn't like, what was unclear, not helpful, ...

Your feedback

Could we see the results of how long people reported the HW took? Same with the grade distribution for HWs (when that is available). Would be a useful metric



Your feedback

I downloaded the class files for today thinking it would be the lecture slides but it was the R code. Would it be possible to get the slides themselves used for the lecture to notetake on?

I always upload the slides before class

look for
the pdf

Name	Date Created	Date Modified	Modified By	Size	
06_probability_feedback.html	Yesterday	Yesterday	Tobias Gerstenberg	1.3 MB	(c)
06_probability.pdf	Yesterday	Yesterday	Tobias Gerstenberg	14.2 MB	(c)
06_probability.Rproj	Sunday	Sunday		205 bytes	(c)
figures	Sunday			--	
probability.html	Sunday	Sunday		4 MB	(c)
probability.Rmd	Sunday	Sunday		11 KB	(c)

Before class, I upload slides without solutions.

After class, I upload the slides with solutions / and results from polls.

Things that came up

Questions about HW on Piazza

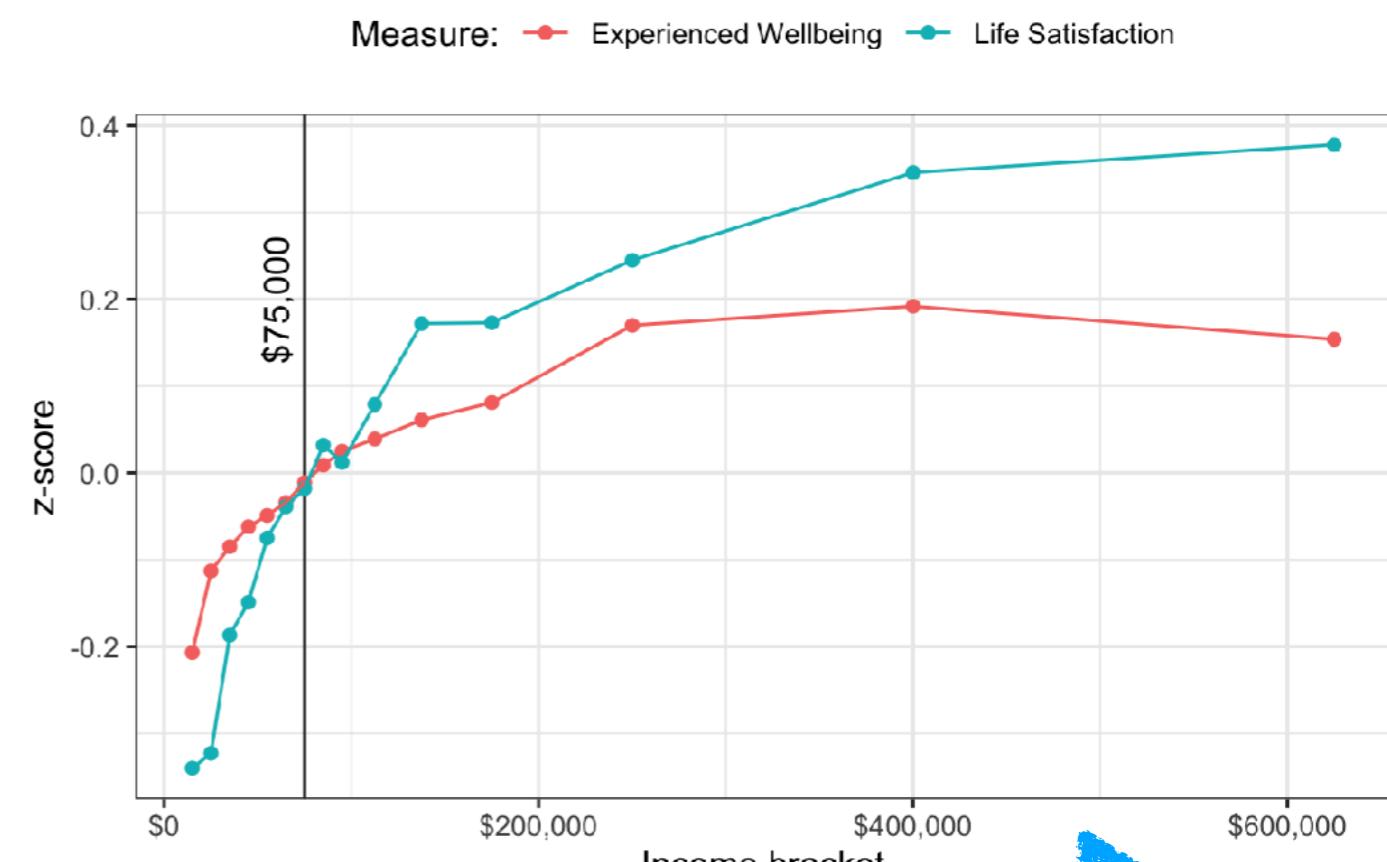
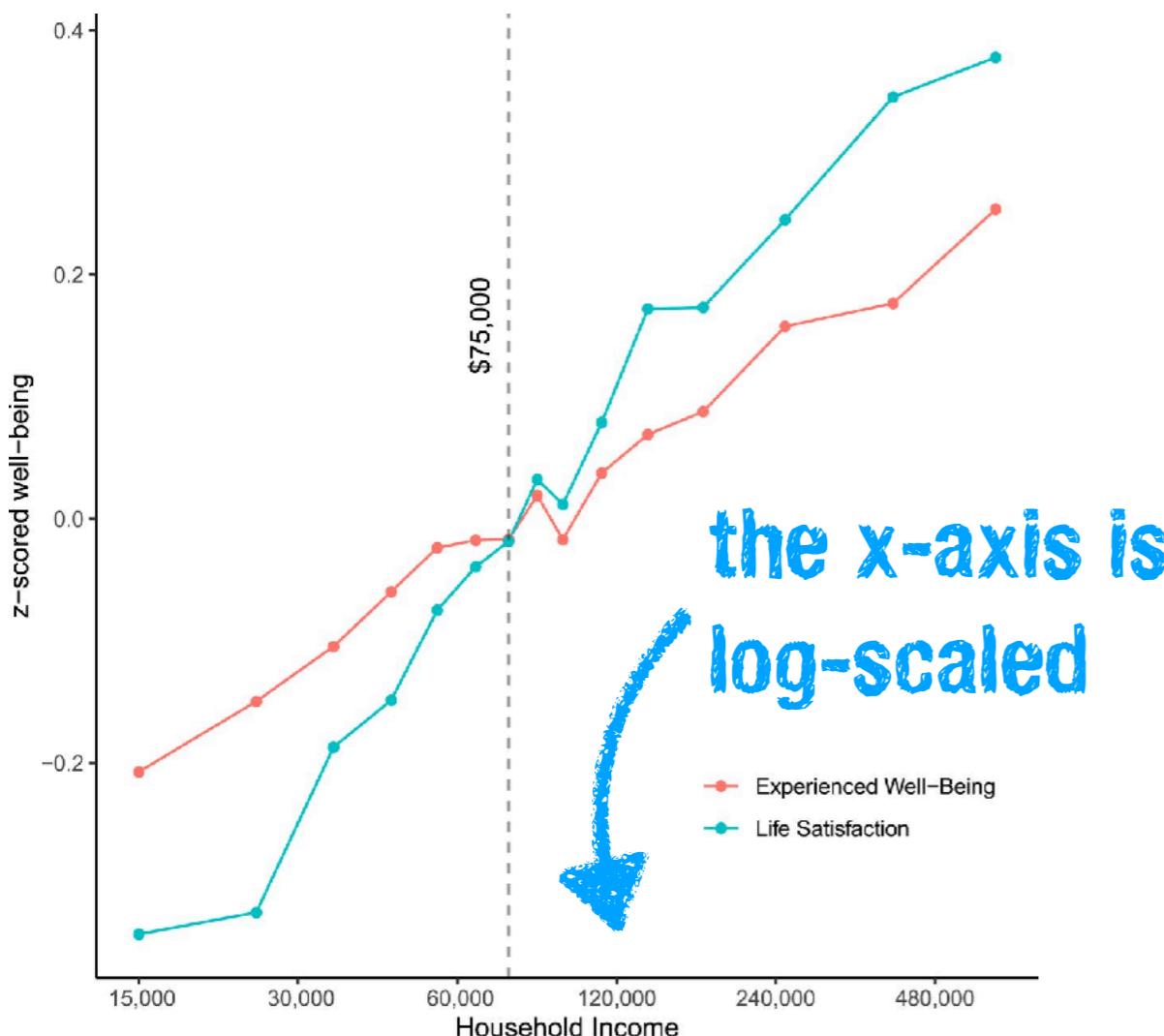
- If in doubt, set the visibility of the question to just the instructor team.
- We will then respond, and will check with you if it's ok for us to change the visibility to public.

Questionable plots, episode #184726

Experienced well-being rises with income, even above \$75,000 per year

Matthew A. Killingsworth^{a,1} 

^aThe Wharton School, University of Pennsylvania, Philadelphia, PA 19104



Plan for today

- Simulating data
 - Drawing samples
 - Working with probability distributions
 - Quick detour: understanding `density()`
 - Asking probability distributions for answers
- Doing Bayesian inference
 - Analytic solution
 - Sampling solution

Simulating data

Drawing samples

Simulating data: Why?



- helps us to:
 - better understand statistical concepts (e.g. p-values, confidence intervals)
 - check how accurately our statistical model can infer the ground truth
 - do power analysis
 - get one step closer to being able to develop our own probabilistic models of an interesting phenomenon

Simulating data: How?



line numbers

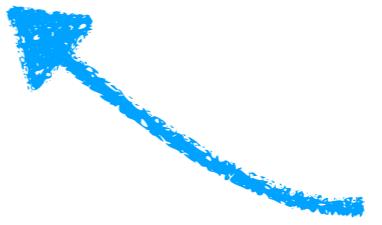
```
1 numbers = 1:3
2
3 numbers %>%
4   sample(size = 10,
5         replace = T)
[1] 3 3 1 2 2 3 2 3 1 2
```

sample 10 times
with replacement please
thank you

Simulating data: How?

```
1 numbers = 1:3
2
3 numbers %>%
4   sample(size = 10,
5         replace = T,
6         prob = c(0.8, 0.1, 0.1))
```

[1] 3 1 1 1 1 2 2 1 1 1



I want
mostly 1s

Simulating data: How?

sets the seed of the random number generator

```
1  `` {r no-seed}
2 numbers = 1:5
3
4 numbers %>%
5   sample(5)
6 ````
```

```
[1] 1 4 5 3 2
[1] 5 3 4 2 1
```

1 `` {r with-seed}
2 set.seed(1)

```
3
4 numbers = 1:5
5
6 numbers %>%
7   sample(5)
8 ````
```

```
[1] 1 4 3 5 2
[1] 1 4 3 5 2
```

every time I run this code chunk, I may get a different outcome

every time I run this code chunk, I get the same outcome 17

set the seed for reproducible code!

sets the seed of the random number generator

```
1 `-- {r no-seed}
2 numbers = 1:5
3
4 numbers %>%
5   sample(5)
6`--
```

[1] 1 4 5 3 2
[1] 5 3 4 2 1

1`-- {r with-seed}
2 set.seed(1)
3
4 numbers = 1:5
5
6 numbers %>%
7 sample(5)
8`--

[1] 1 4 3 5 2
[1] 1 4 3 5 2

every time I run this code chunk, I may get a different outcome

every time I run this code chunk, I get the same outcome 18

Simulating data: How?

Sampling rows from a data frame

```
1 set.seed(1)
2 n = 10
3 df.data = tibble(trial = 1:n,
4                   stimulus = sample(c("flower", "pet"), size = n, replace = T),
5                   rating = sample(1:10, size = n, replace = T))
```

trial	stimulus	rating
1	flower	1
2	flower	4
3	flower	3
4	pet	6
5	flower	10
6	flower	10
7	pet	6
8	pet	4
9	pet	4
10	flower	10

sample 6 rows with replacement

```
1 df.data %>%
2   slice_sample(n = 6,
3                 replace = T)
```

trial	stimulus	rating
9	pet	9
4	flower	5
7	flower	10
1	flower	3
2	pet	1
7	flower	10

sample 50% of the rows

```
1 df.data %>%
2   slice_sample(prop = 0.5)
```

trial	stimulus	rating
9	pet	9
4	flower	5
7	flower	10
1	flower	3
2	pet	1

Working with probability distributions

MOST POPULAR MARVEL MOVIE ACTORS

HEIGHT COMPARISION

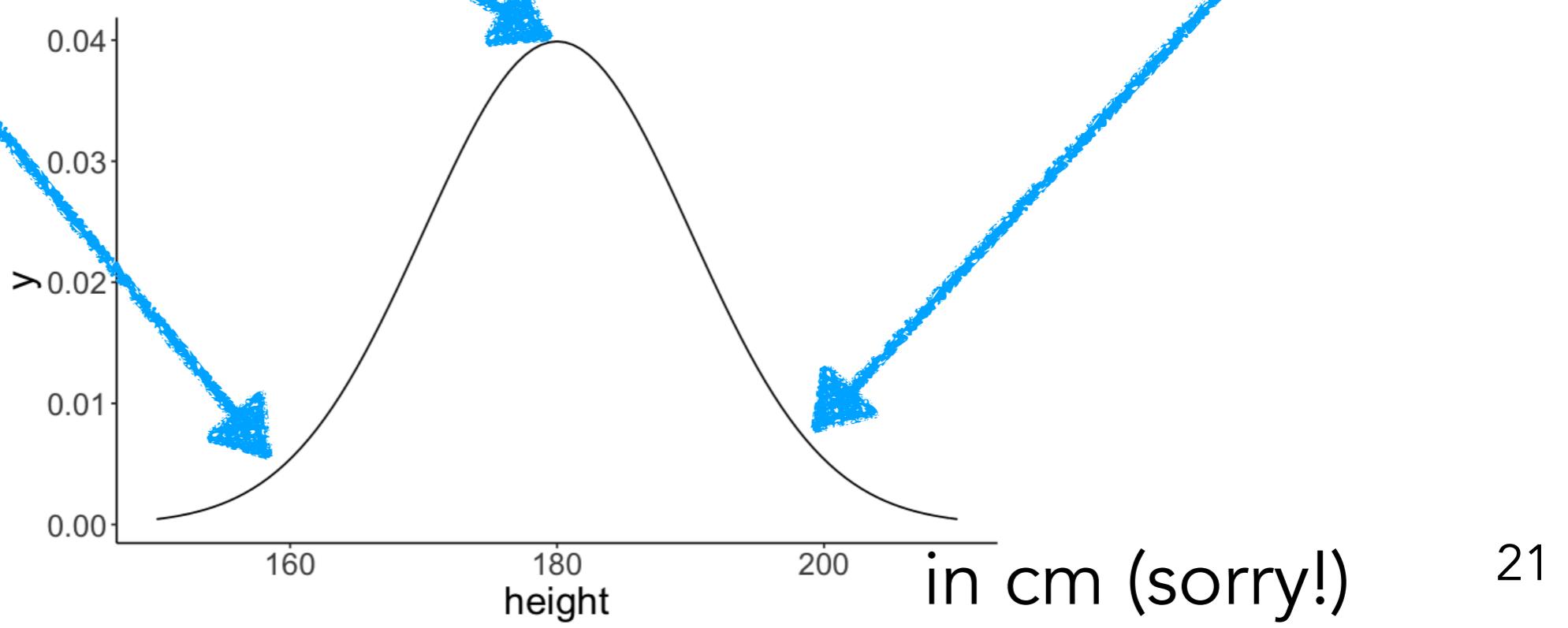
Who's the tallest and shortest actor in the Marvel Cinematic Universe?



Copyright 2018 Maurice Mitchell

TheGeekTwins.com

Not Affiliated With Marvel Studios. All Rights Reserved



Working with probability distributions

letter	description	example
d	for “density”, the density function (probability mass function (for <i>discrete</i> variables) or probability density function (for <i>continuous</i> variables))	<code>dnorm()</code>
p	for “probability”, the cumulative distribution function	<code>pnorm()</code>
q	for “quantile”, the inverse cumulative distribution function	<code>qnorm()</code>
r	for “random”, a random variable having the specified distribution	<code>rnorm()</code>

Normal distribution

make data frame with minimum and maximum x-value

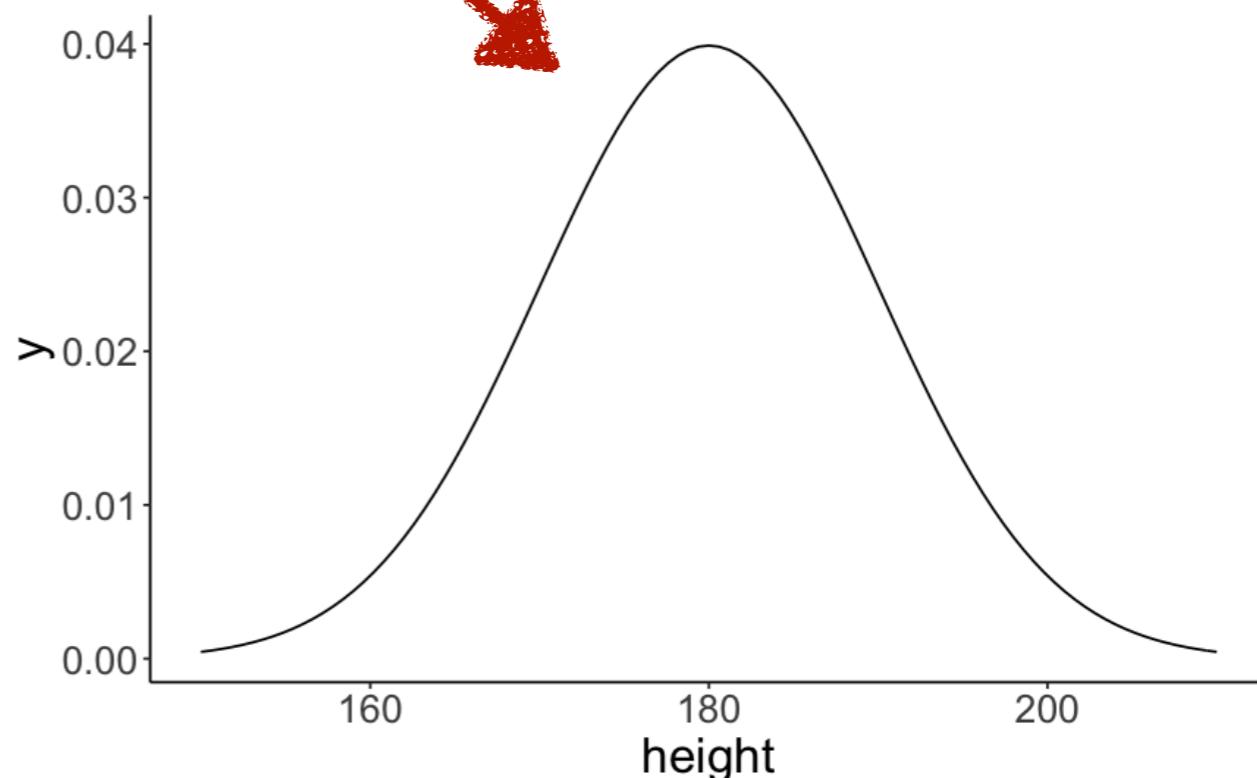
```
1 ggplot(data = tibble(height = c(150, 210)),  
2         mapping = aes(x = height)) +  
3     stat_function(fun = ~ dnorm(x = .,  
4                                     mean = 180,  
5                                     sd = 10))
```

function for plotting
functions

what function
should be plotted?

any parameters for
the function?

the result



dnorm(x, mean = 180, sd = 10)

d = density

norm = normal distribution

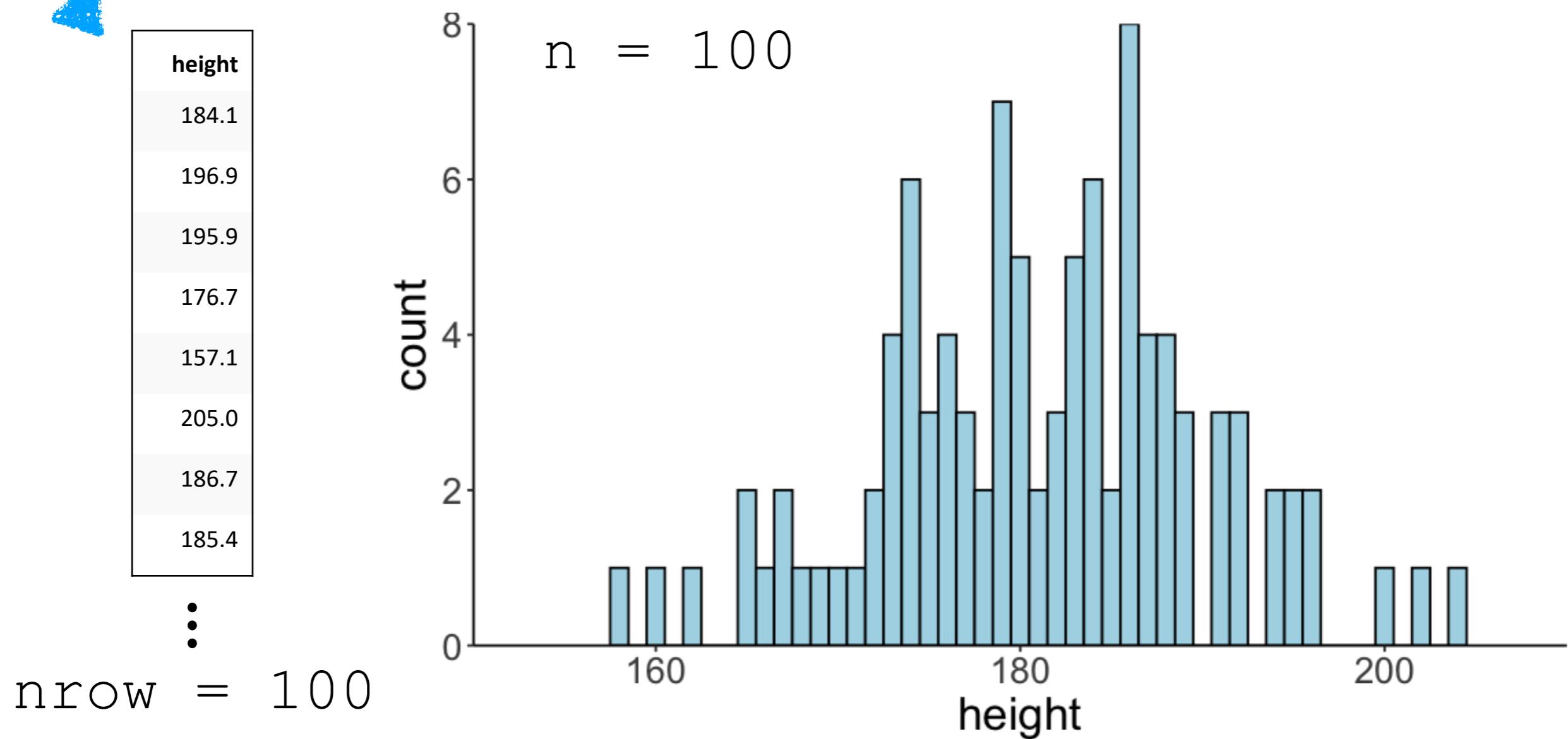
Sampling from distributions

rnorm(n, mean = 180, sd = 10)

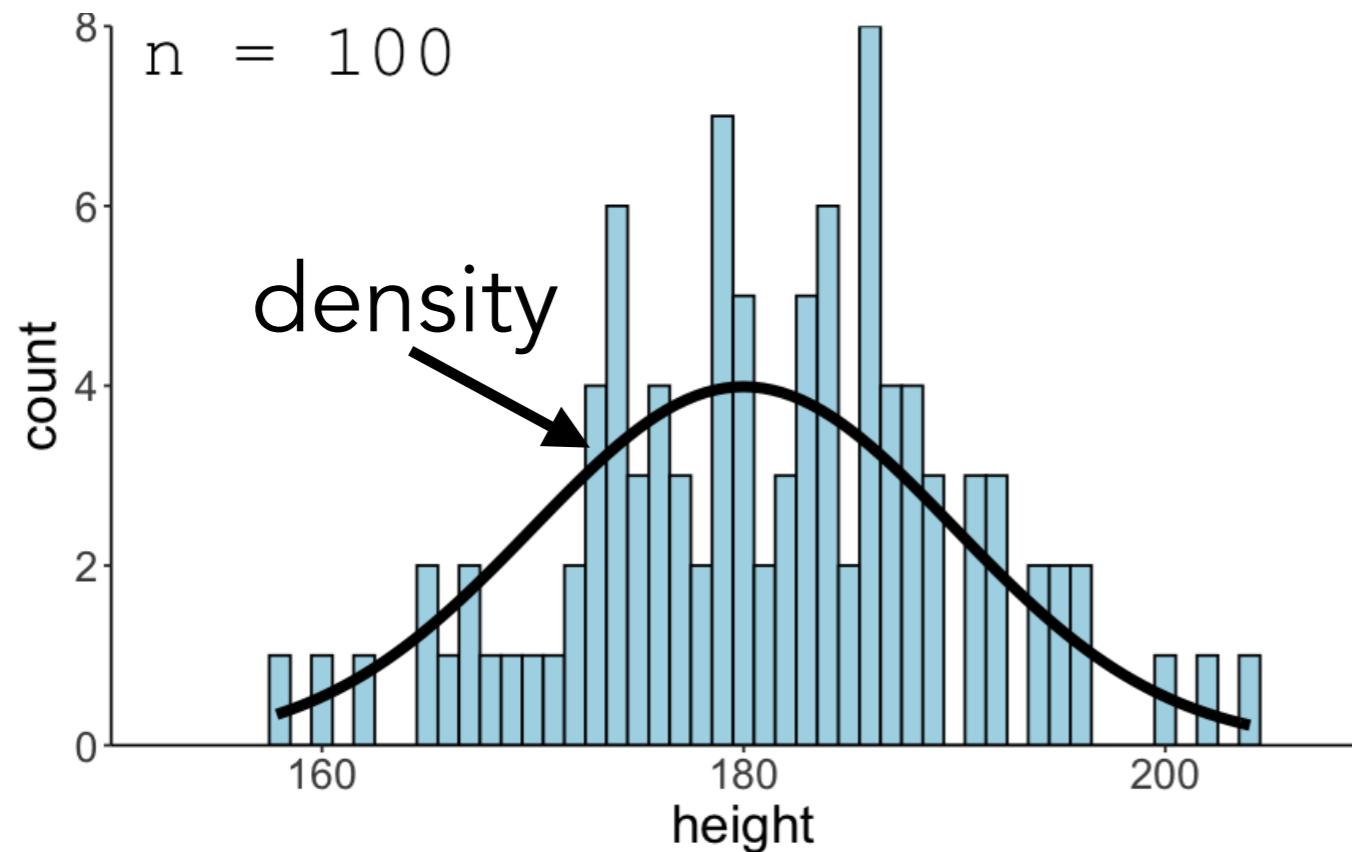
n = number of samples

r = random samples

norm = normal distribution

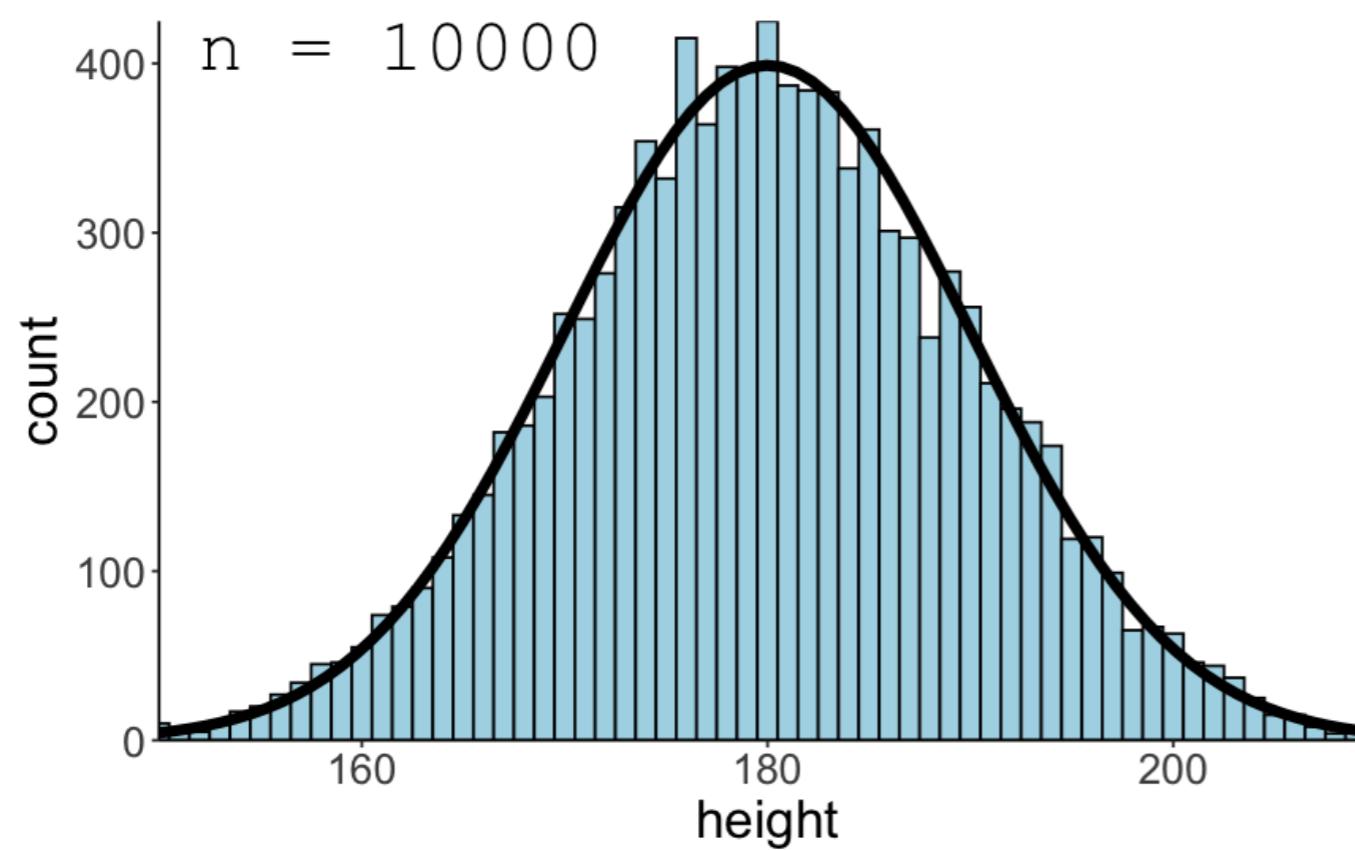


Sampling from distributions

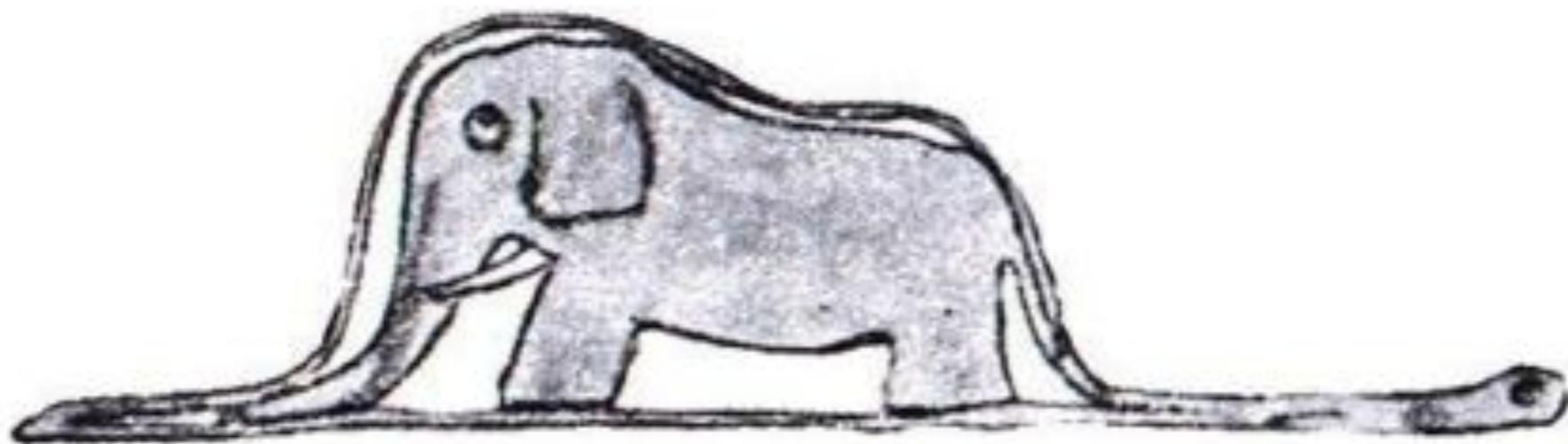
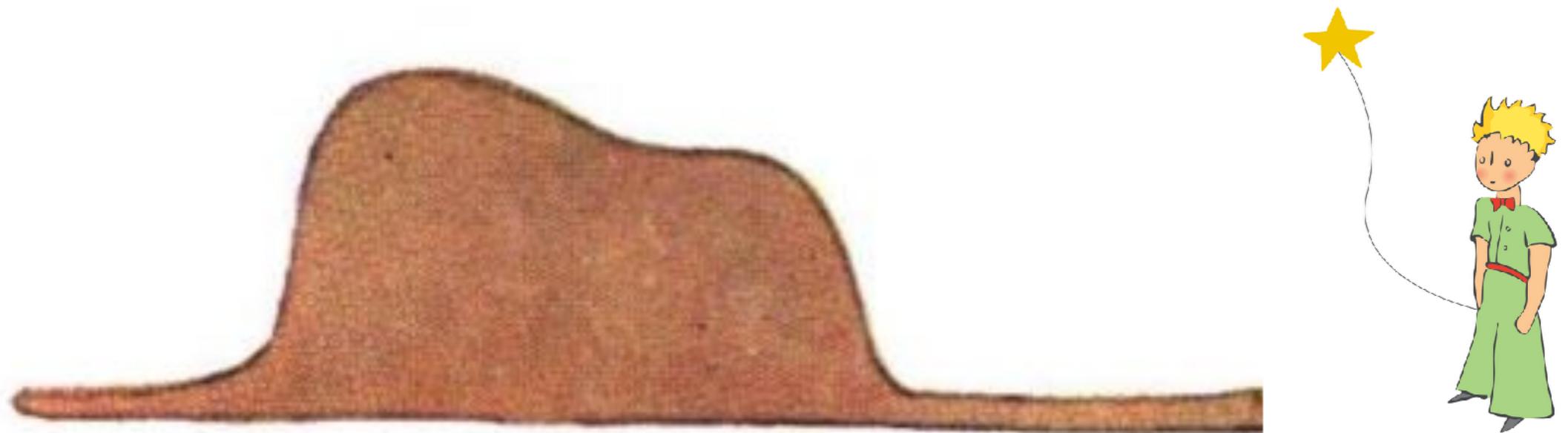


law of large numbers

approximation to true underlying distribution improves with increased sample size



Quick detour: understanding `density()`

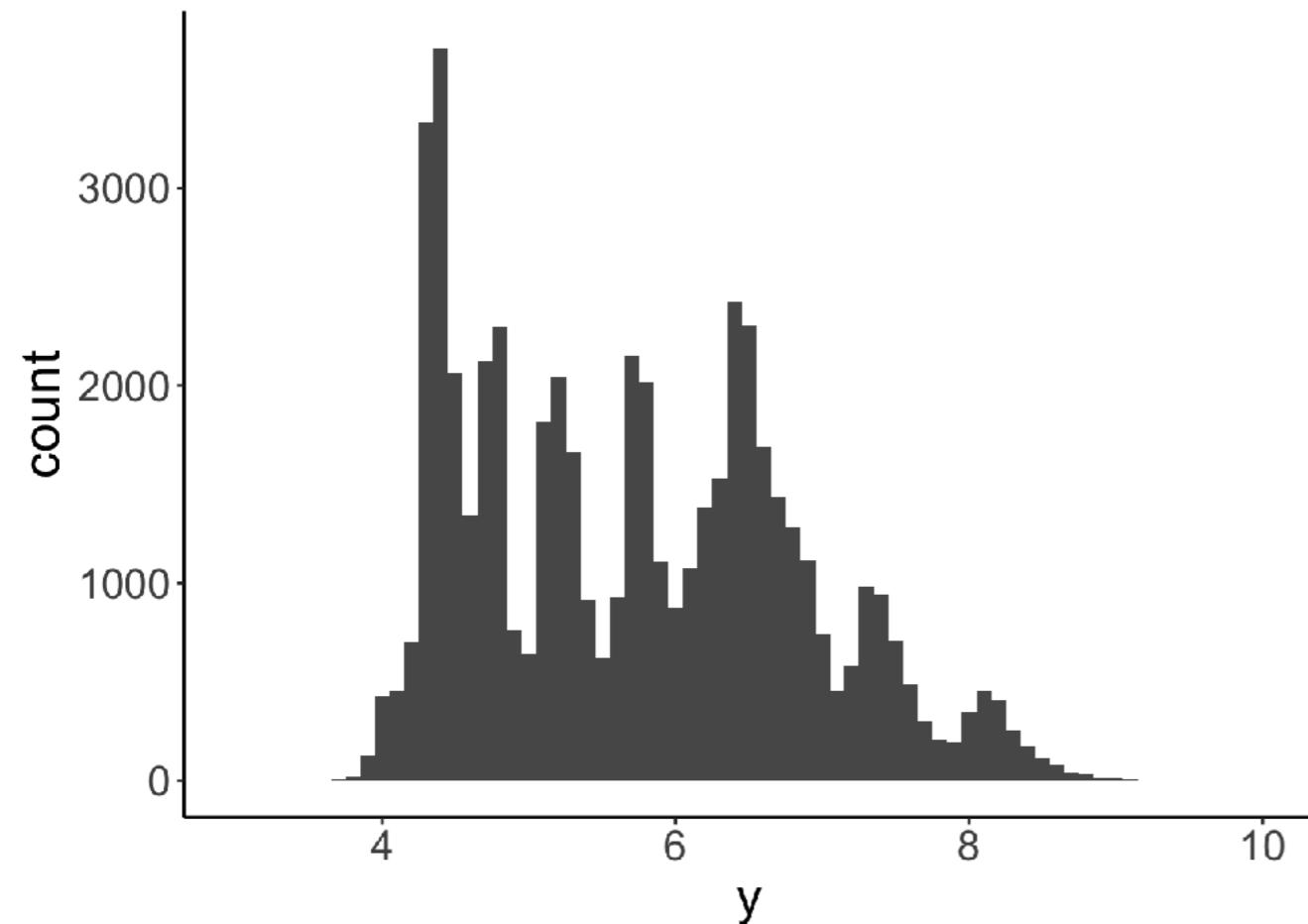


What's underneath the hood (or hat)?

You've seen `density()` before

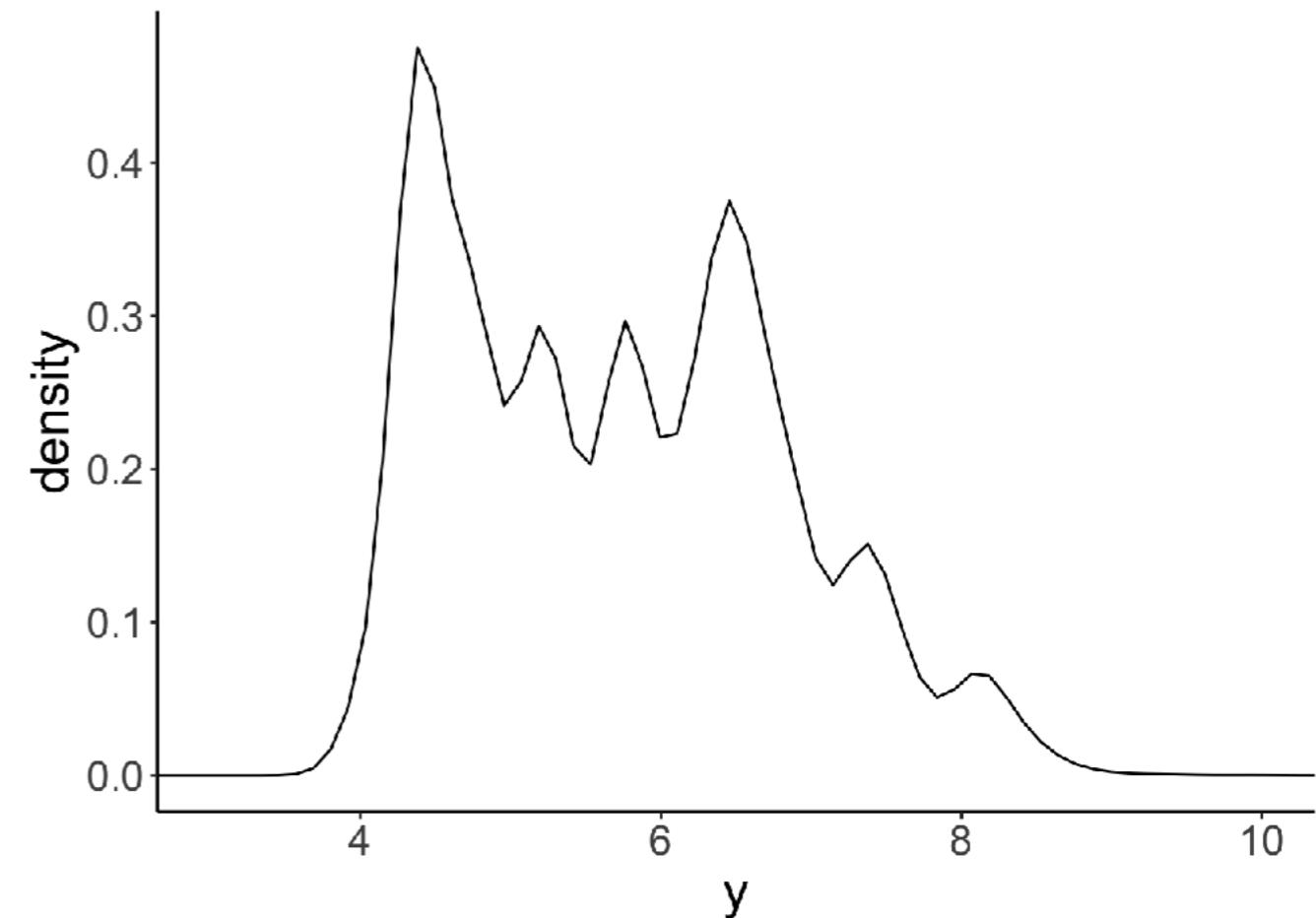
Histogram

```
1 ggplot(data = df.diamonds,  
2         mapping = aes(x = y)) +  
3         geom_histogram(binwidth = 0.1) +  
4         coord_cartesian(xlim = c(3, 10))
```



Density

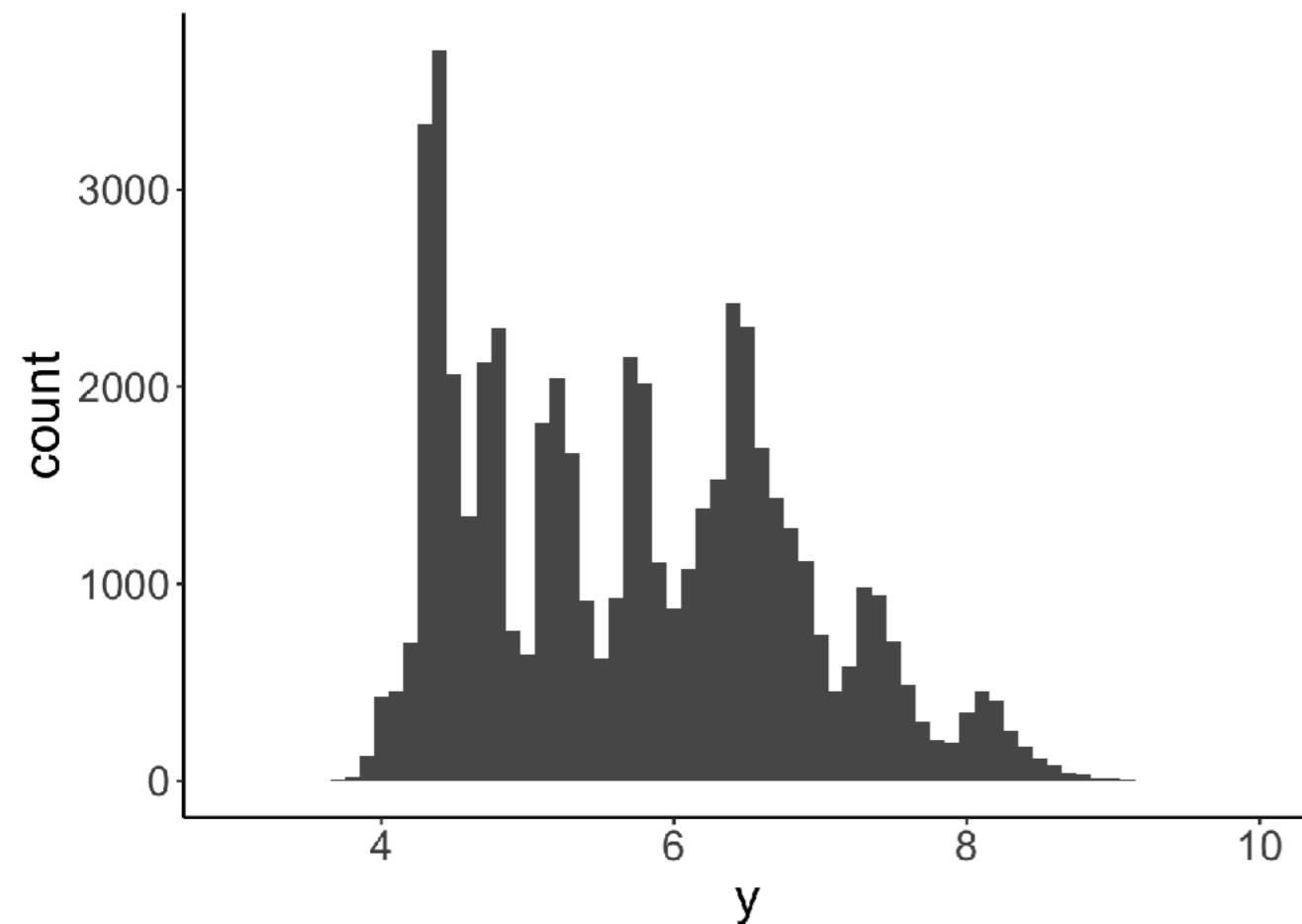
```
1 ggplot(data = df.diamonds,  
2         mapping = aes(x = y)) +  
3         geom_density() +  
4         coord_cartesian(xlim = c(3, 10))
```



You've seen `density()` before

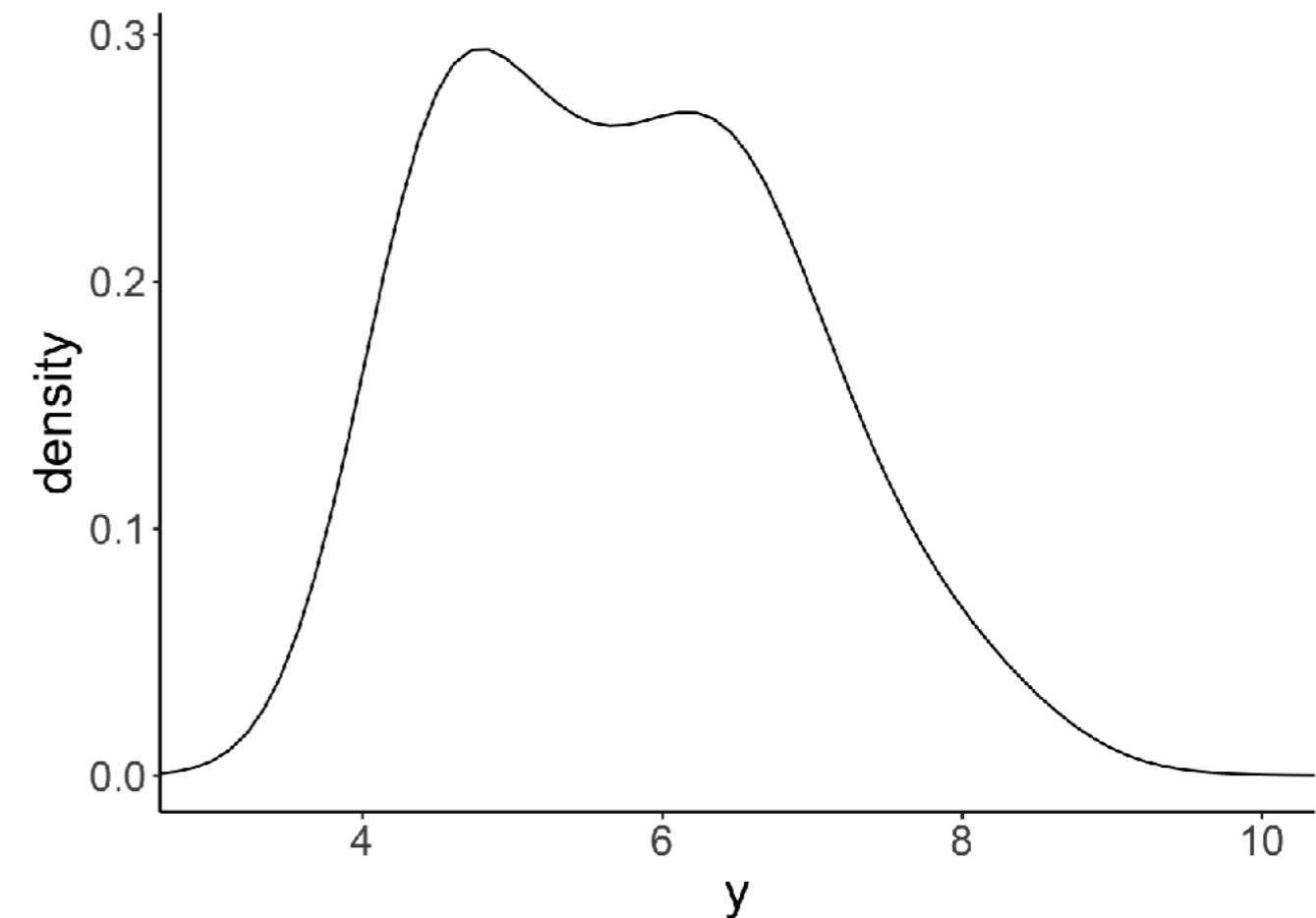
Histogram

```
1 ggplot(data = df.diamonds,  
2         mapping = aes(x = y)) +  
3         geom_histogram(binwidth = 0.1) +  
4         coord_cartesian(xlim = c(3, 10))
```



Density

```
1 ggplot(data = df.diamonds,  
2         mapping = aes(x = y)) +  
3         geom_density(bw = 0.5) +  
4         coord_cartesian(xlim = c(3, 10))
```

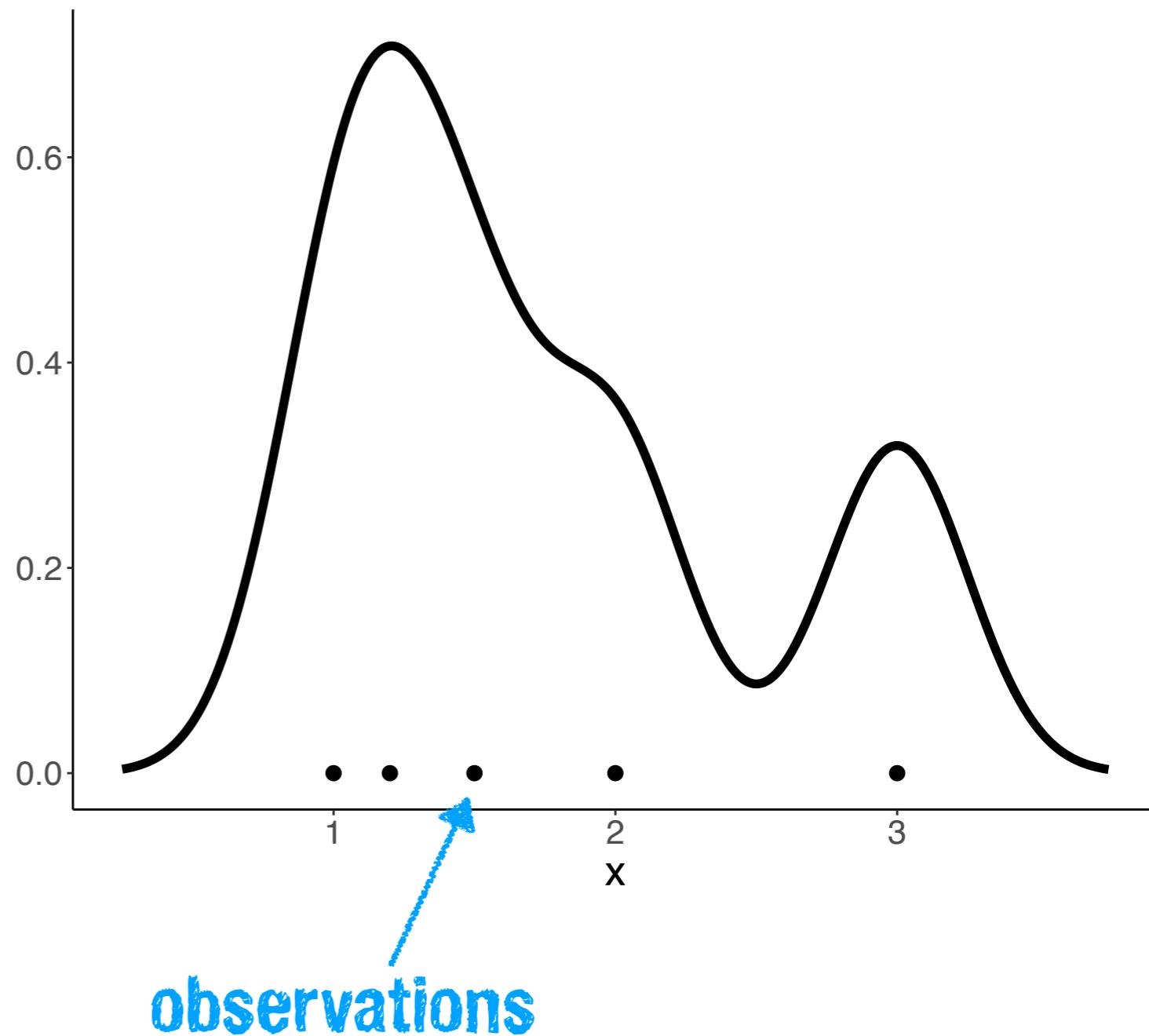


Understanding density()

```
1 # calculate density  
2 observations = c(1, 1.2, 1.5, 2, 3)  
3 bandwidth = 0.25  
4 density = density(observations,  
5   kernel = "gaussian",  
6   bw = bandwidth,  
7   n = 512)
```

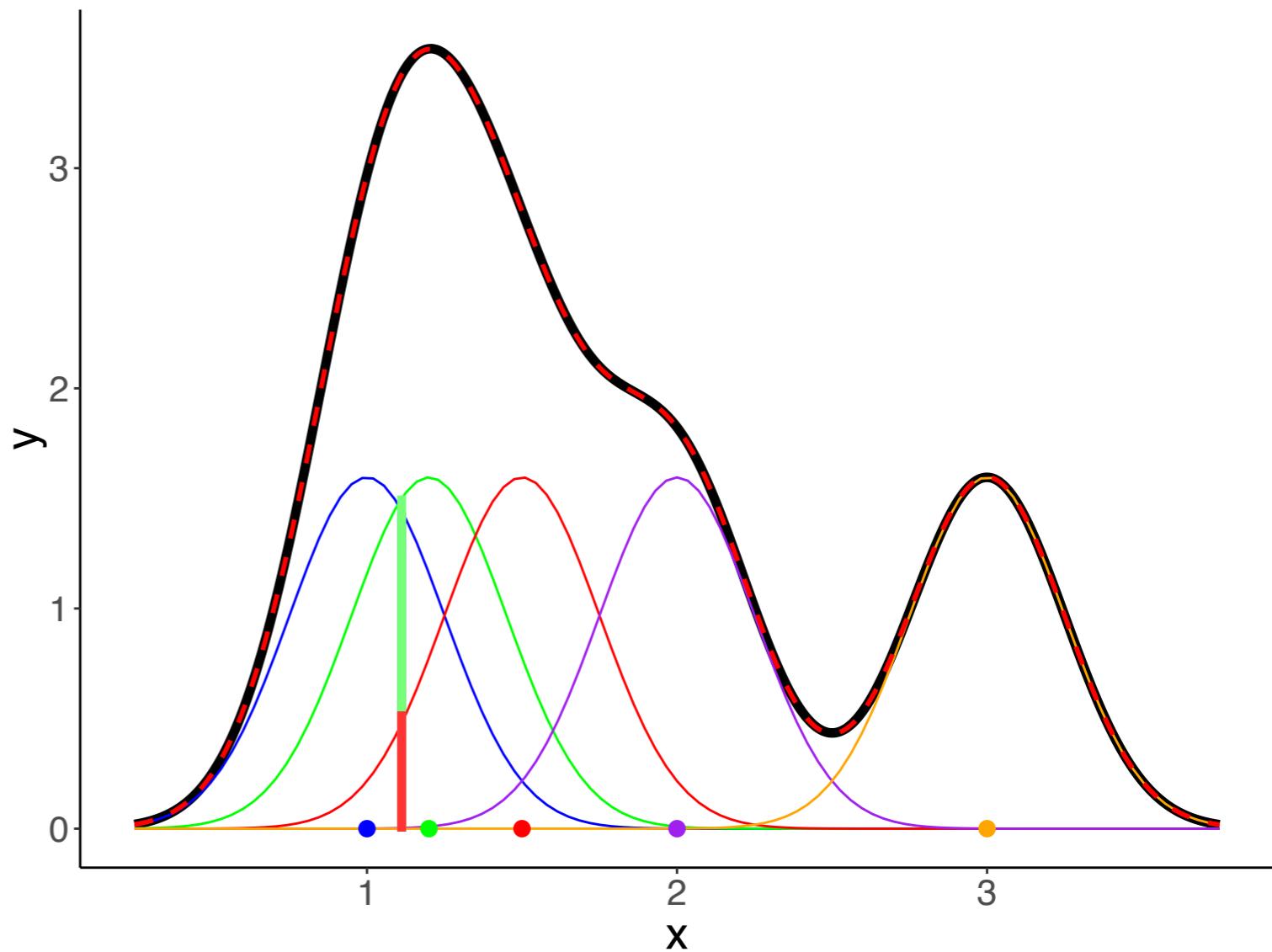
x	y
0.250	0.004
0.257	0.004
0.264	0.005
0.271	0.005
0.277	0.005
0.284	0.006

nrow = 512



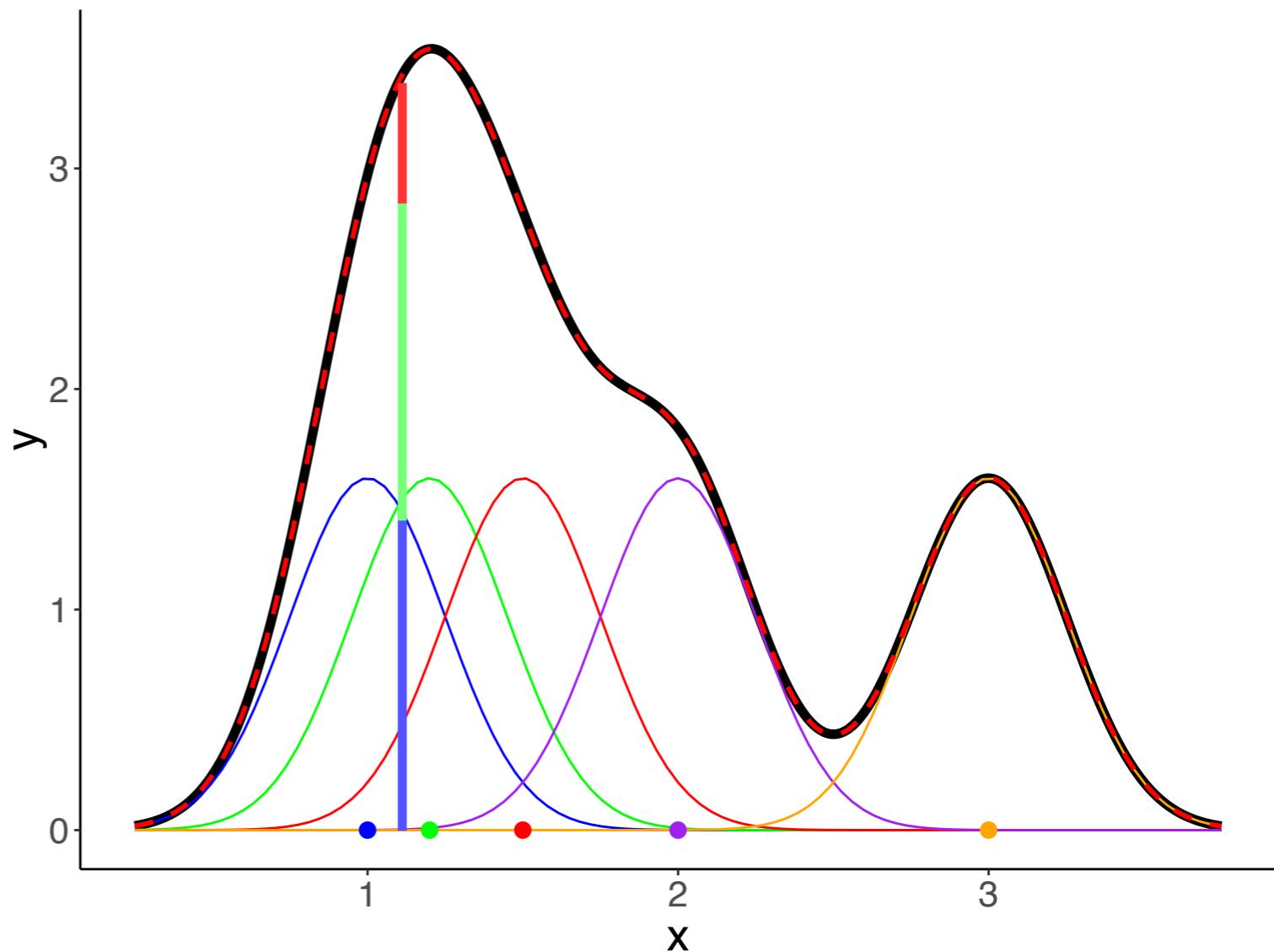
Understanding density()

x	y	observation_1	observation_2	observation_3	observation_4	observation_5	sum_norm
0.250	0.019	0.018	0.001	0	0	0	0.019
0.257	0.021	0.019	0.001	0	0	0	0.021
0.264	0.023	0.021	0.001	0	0	0	0.022
0.271	0.024	0.023	0.002	0	0	0	0.024
0.277	0.027	0.024	0.002	0	0	0	0.026
0.284	0.029	0.026	0.002	0	0	0	0.028

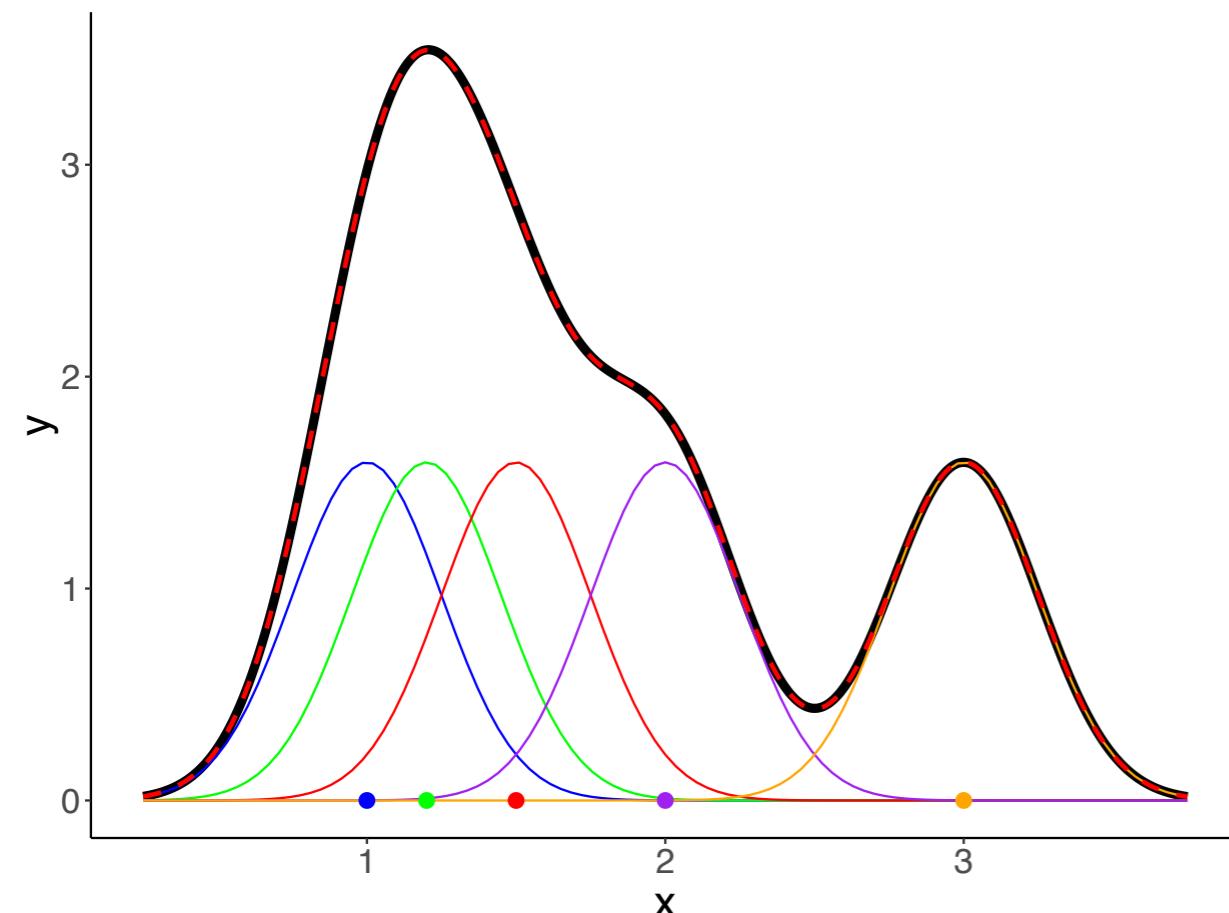


Understanding density()

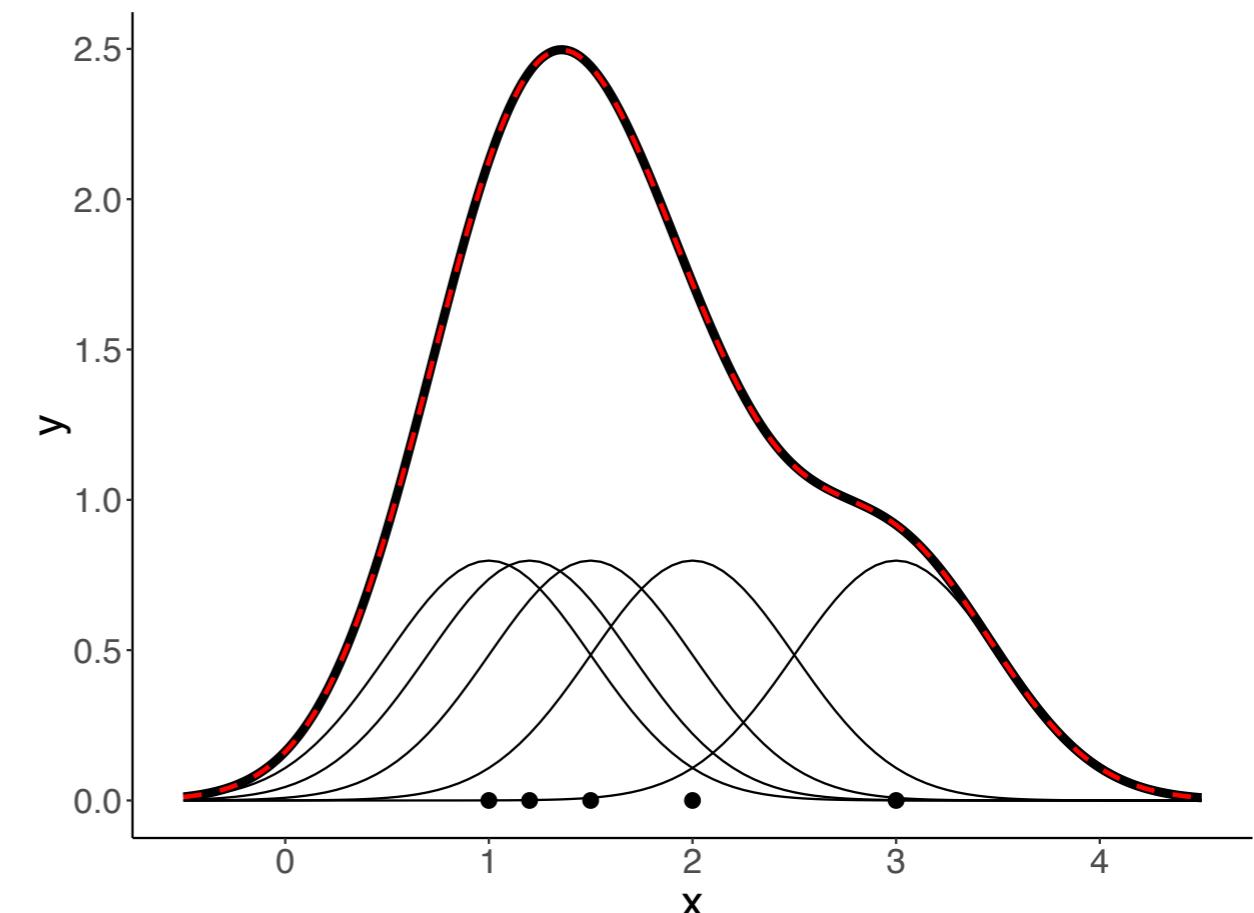
x	y	observation_1	observation_2	observation_3	observation_4	observation_5	sum_norm
0.250	0.019	0.018	0.001	0	0	0	0.019
0.257	0.021	0.019	0.001	0	0	0	0.021
0.264	0.023	0.021	0.001	0	0	0	0.022
0.271	0.024	0.023	0.002	0	0	0	0.024
0.277	0.027	0.024	0.002	0	0	0	0.026
0.284	0.029	0.026	0.002	0	0	0	0.028



Understanding density()



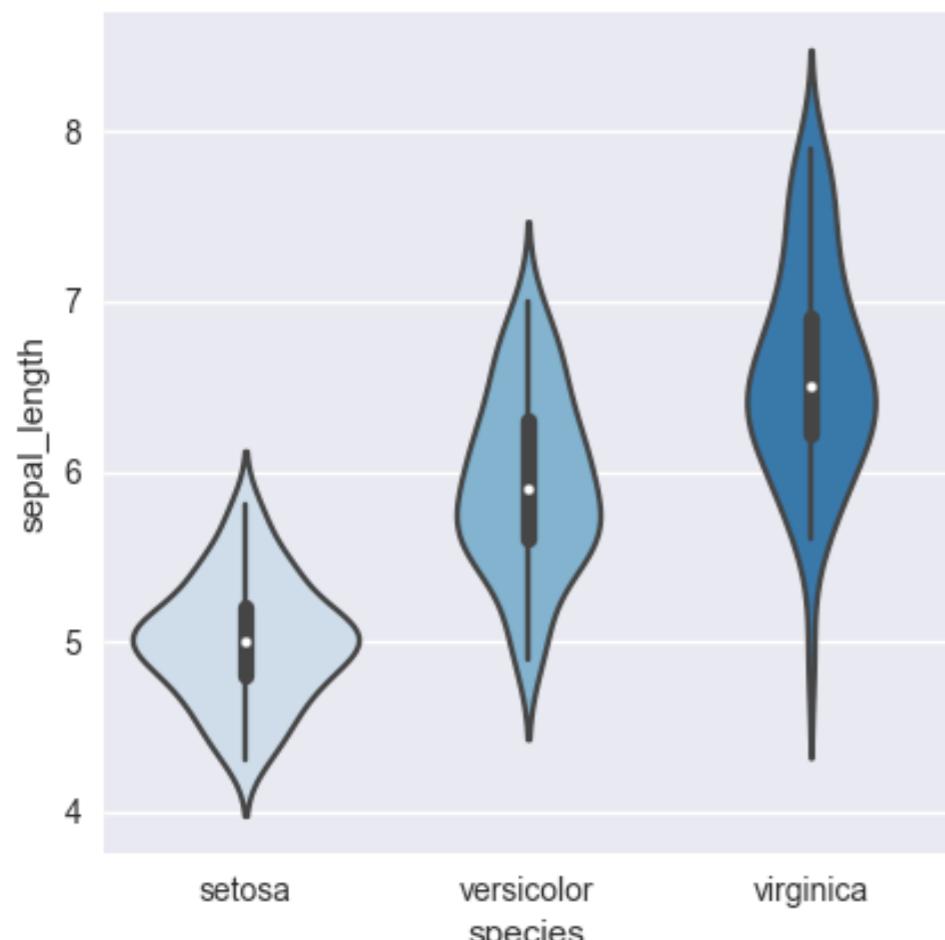
density(bw = 0.25)



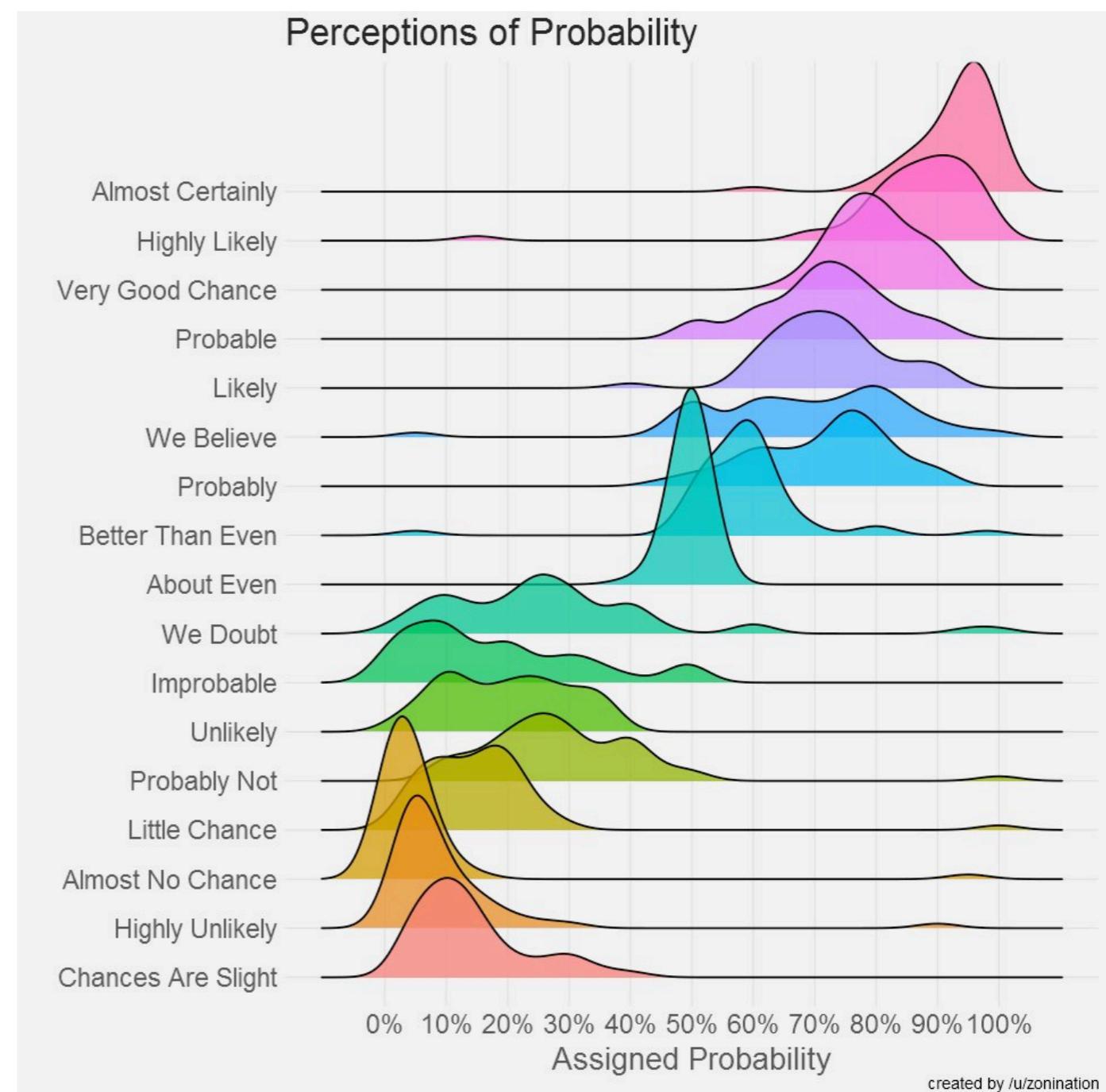
density(bw = 0.5)

Understanding density()

violinplot



joyplot



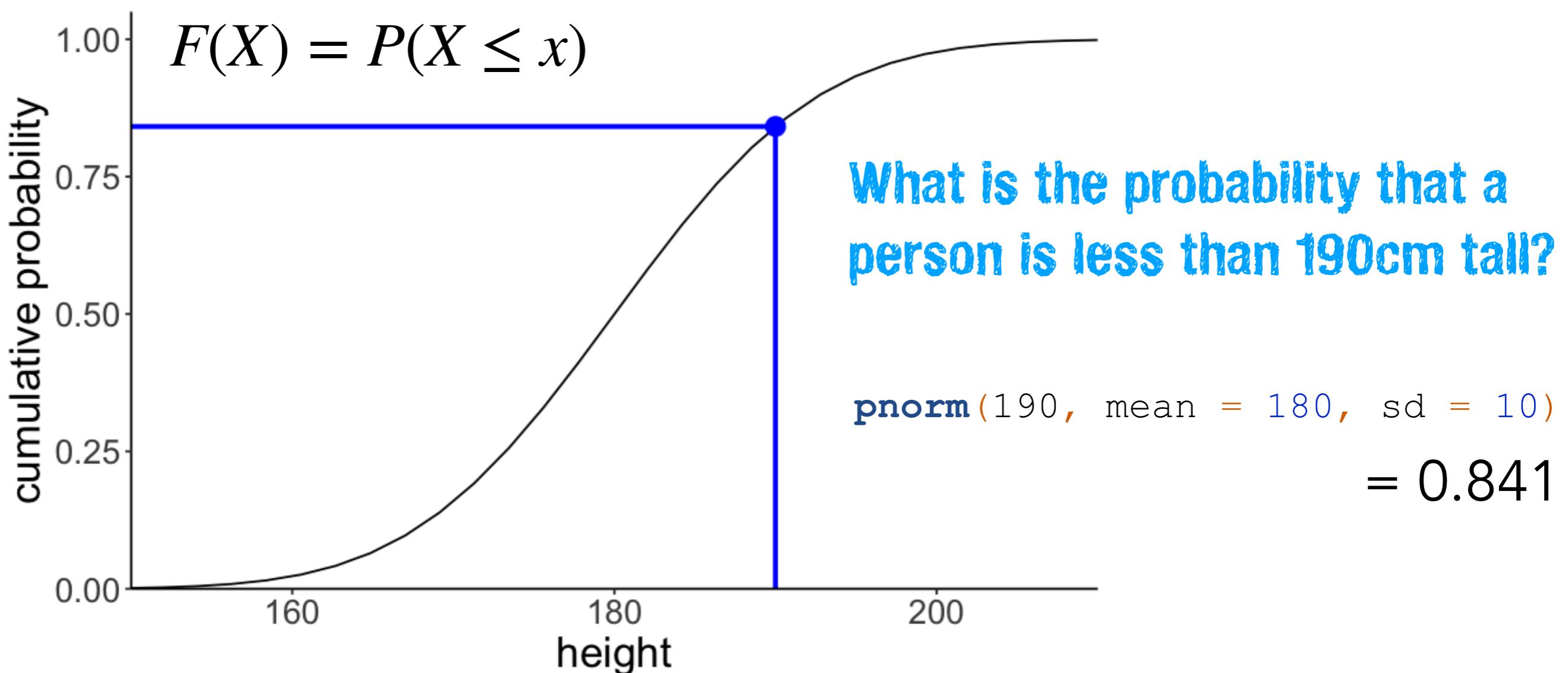
Asking probability distributions for answers



Cumulative probability distribution

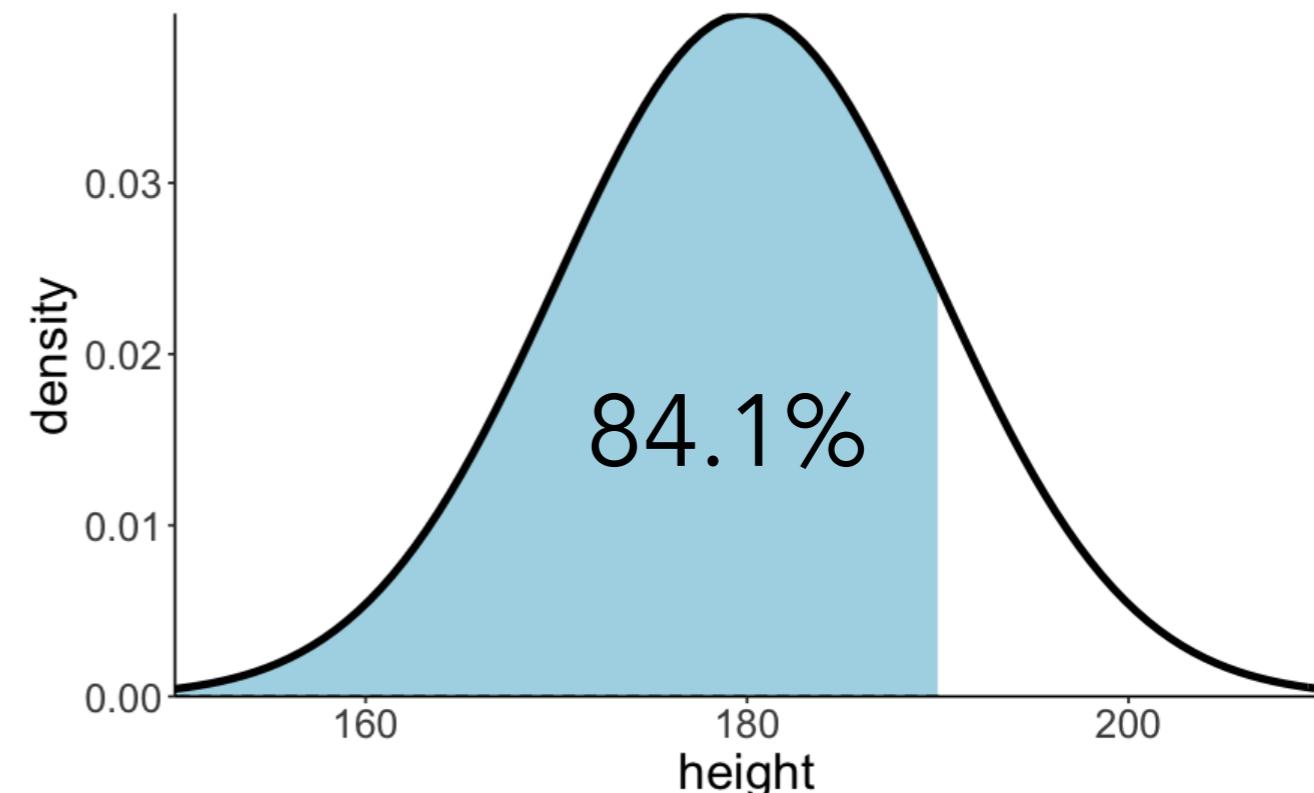
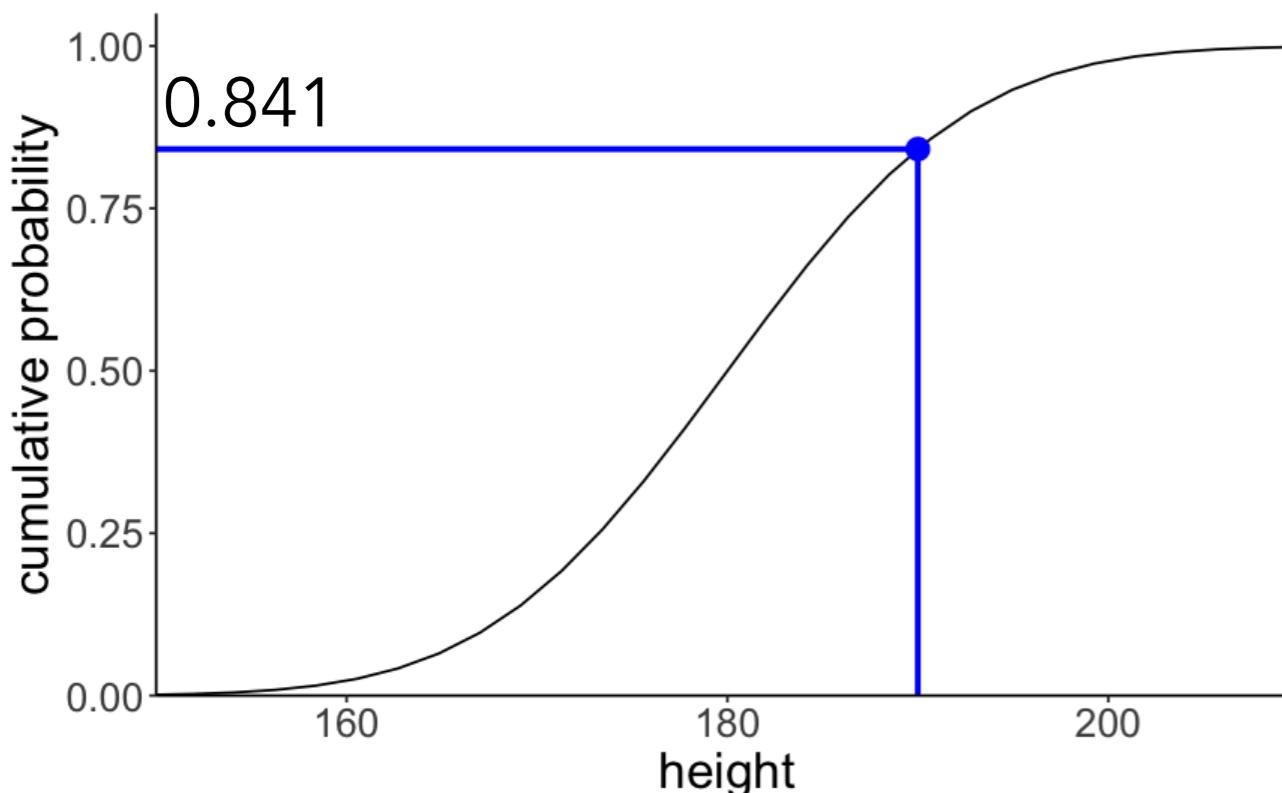
```
1 ggplot(data = tibble(x = c(150, 210)),  
2         mapping = aes(x = x)) +  
3   stat_function(fun = ~ pnorm(q = .,  
4                               mean = 180,  
5                               sd = 10))
```

p = probability
cumulative distribution function



Computing probabilities

`pnorm(190, mean = 180, sd = 10)`

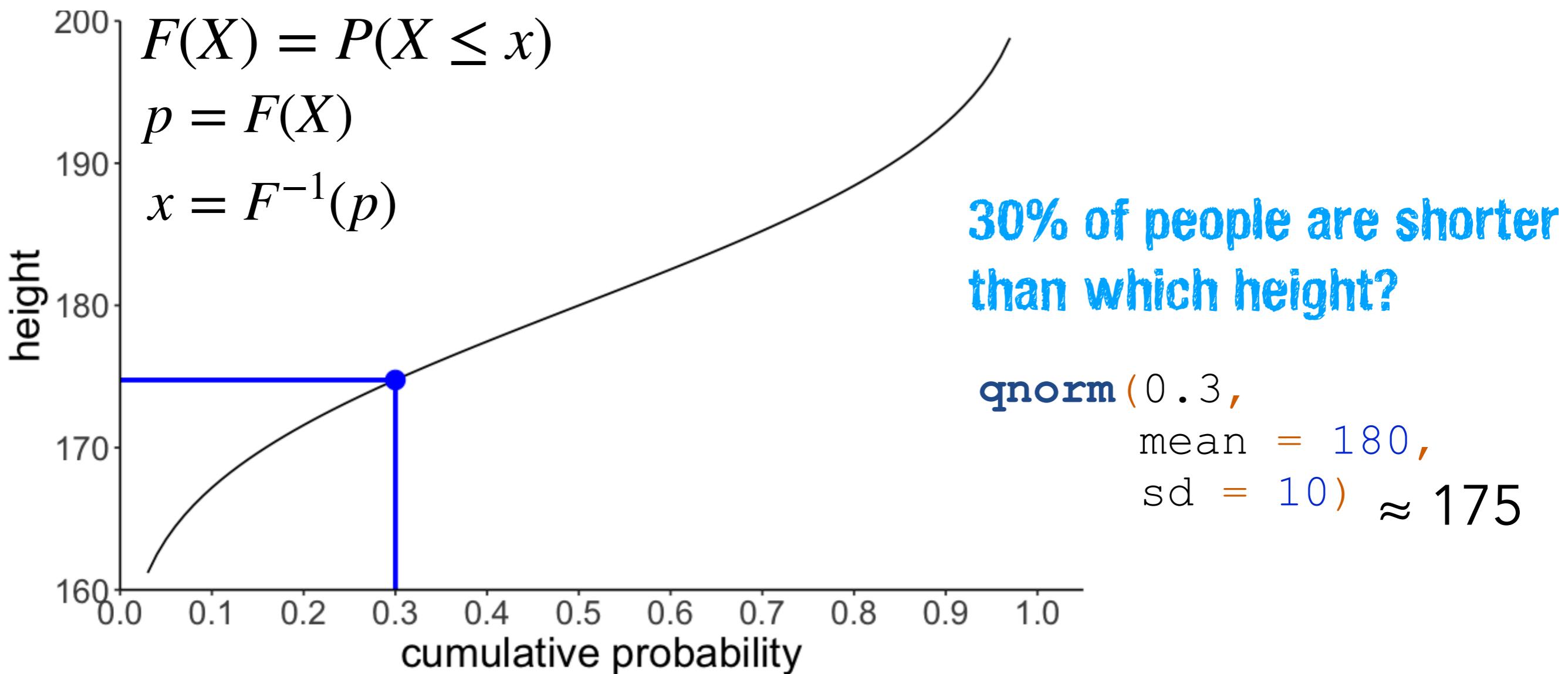


`pnorm(x)` returns the integral from $-\infty$ to x of the probability density function

Inverse cumulative distribution function

```
1 ggplot(data = tibble(x = c(0, 1)),  
2         mapping = aes(x = x)) +  
3   stat_function(fun = ~ qnorm(p = .,  
4                               mean = 180,  
5                               sd = 10))
```

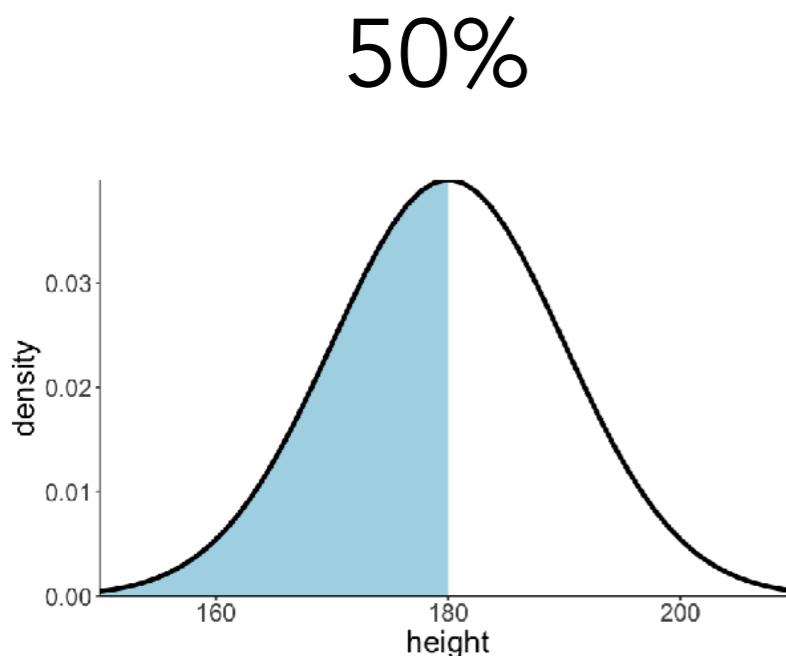
q = quantile
inverse cumulative distribution function



What proportion of people are between 170cm and 180cm?

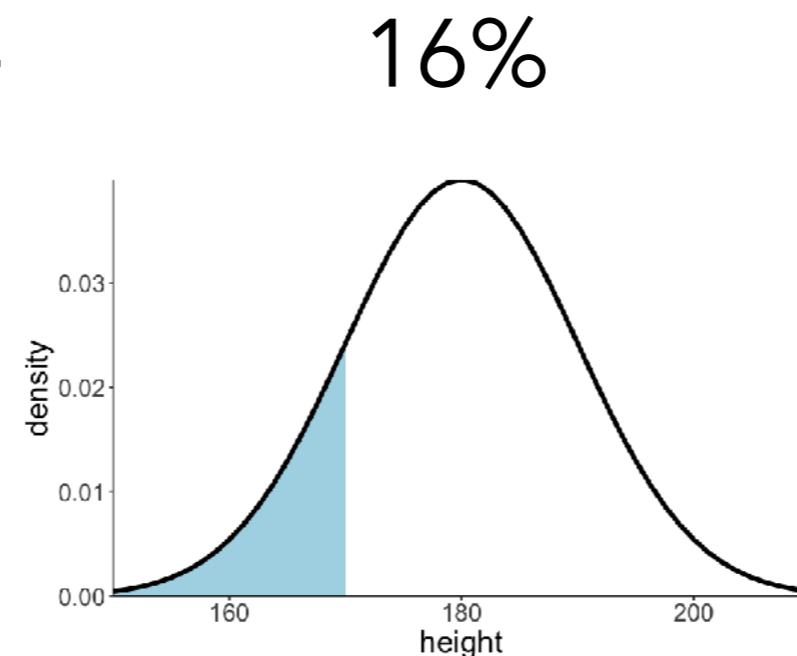
Analytic solution

```
pnorm(180,  
      mean = 180,  
      sd = 10)
```

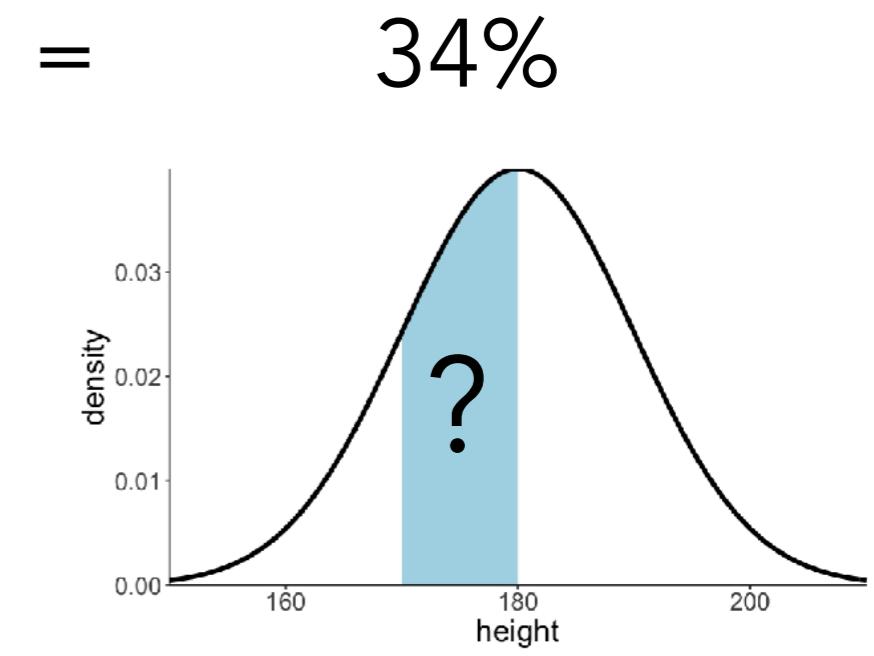


smaller than 180cm

```
pnorm(170,  
      mean = 180,  
      sd = 10)
```



smaller than 170cm

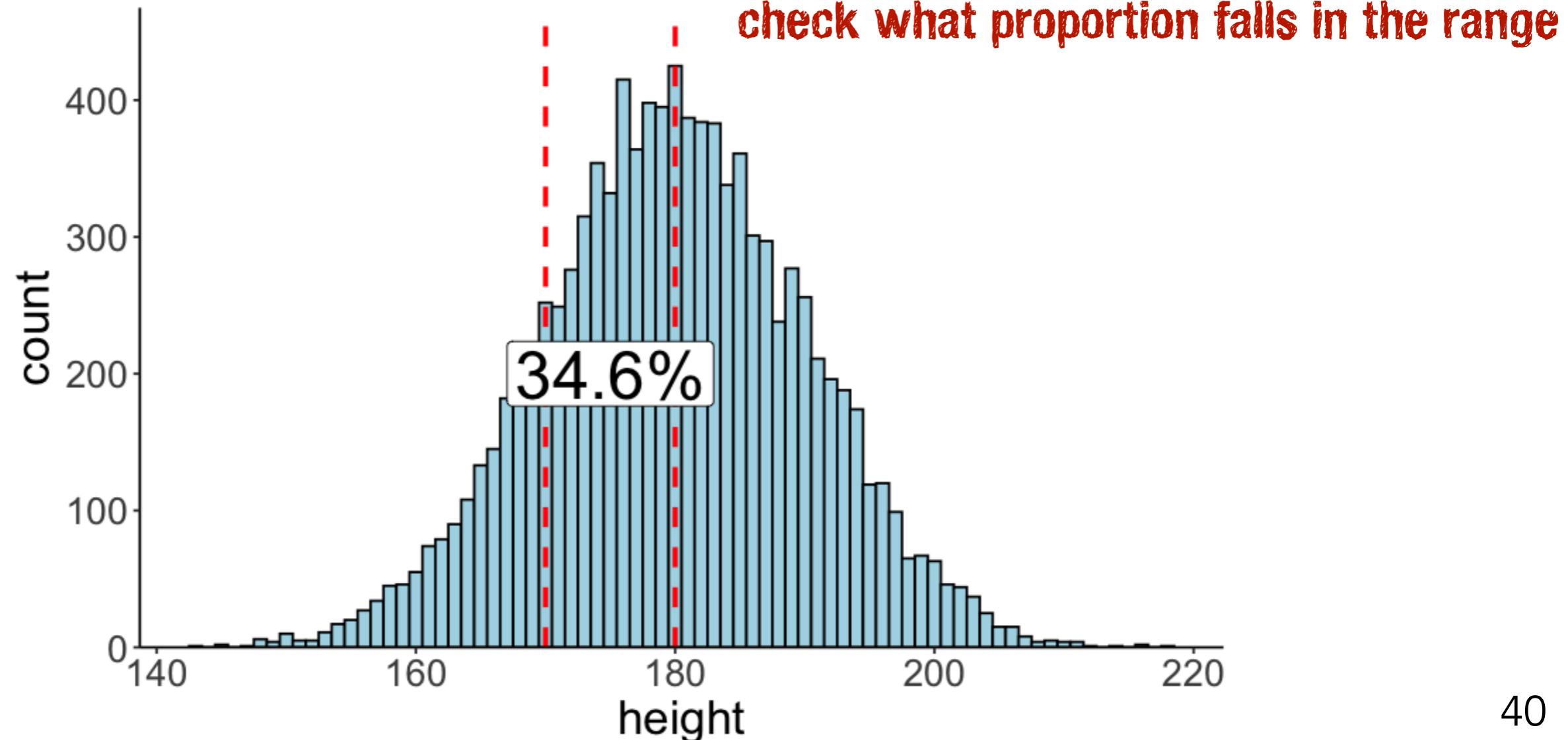


between 170cm
and 180cm

What proportion of people are between 170cm and 180cm?

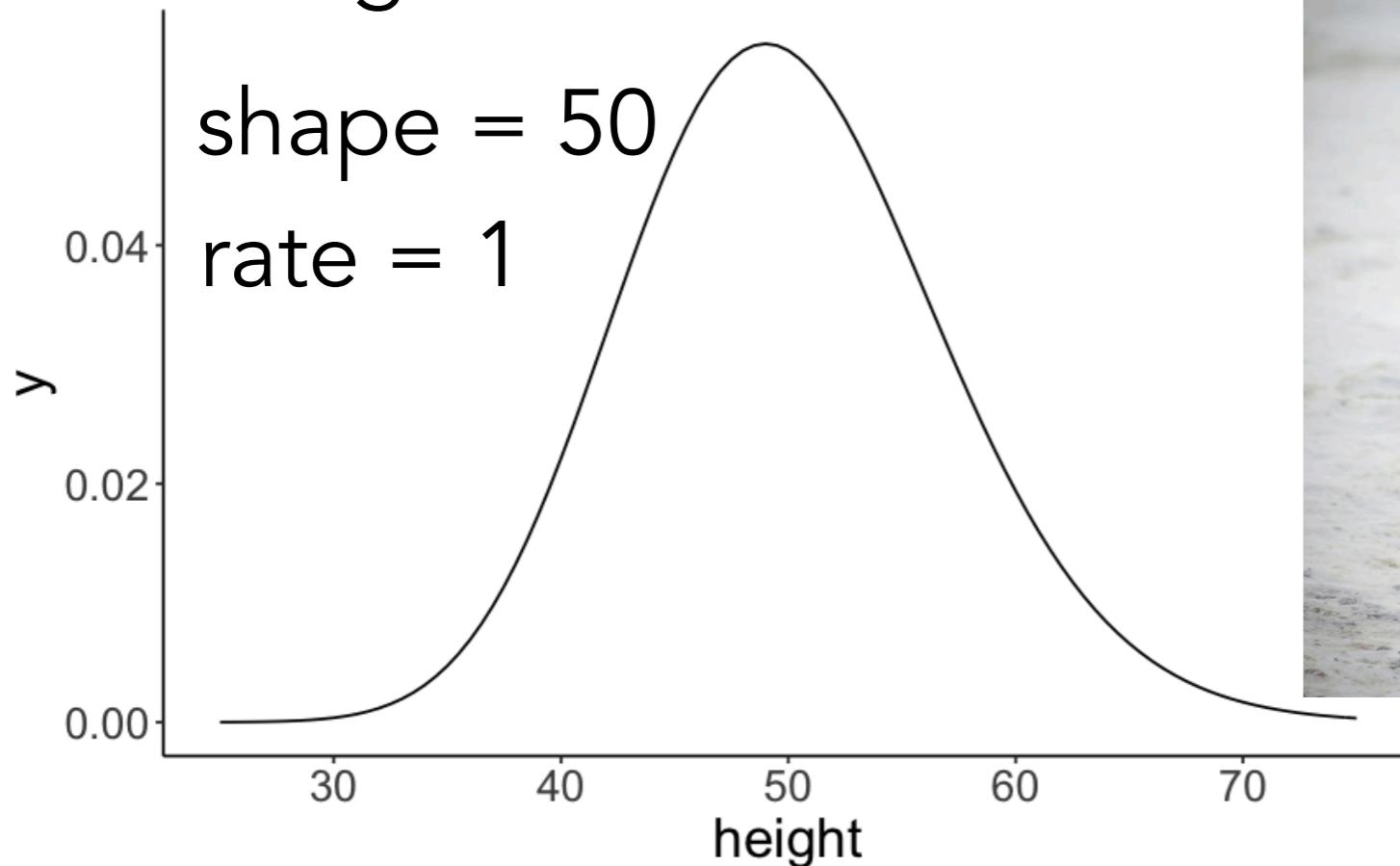
Sampling solution

```
1 tibble(height = rnorm(n = 10000, mean = 180, sd = 10)) %>%
2   summarize(probability = sum(height > 170 & height < 180) / n())
```



Answering questions about Penguins

gamma distribution



1. Make this plot
2. A 60cm tall Penguin claims that no more than 10% are taller than her. Is she correct?
3. Are there more penguins between 50 and 55cm or between 55 and 65cm?
4. What size is a Penguin who is taller than 75% of the rest?

Breakout rooms

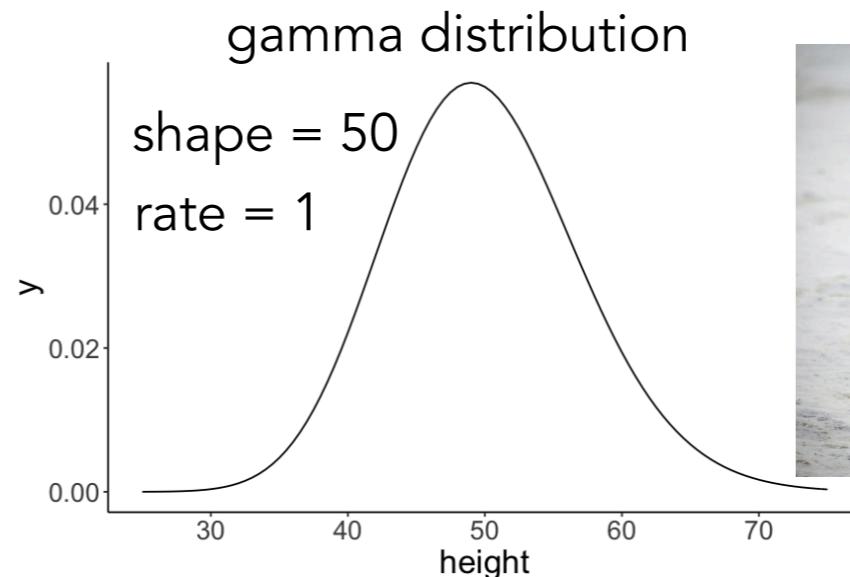
Tasks:

- Answer these questions using R.
- Have one person in your group share their screen.
- Ping us if you need help.

Size: ~4 people

Time: 12 minutes

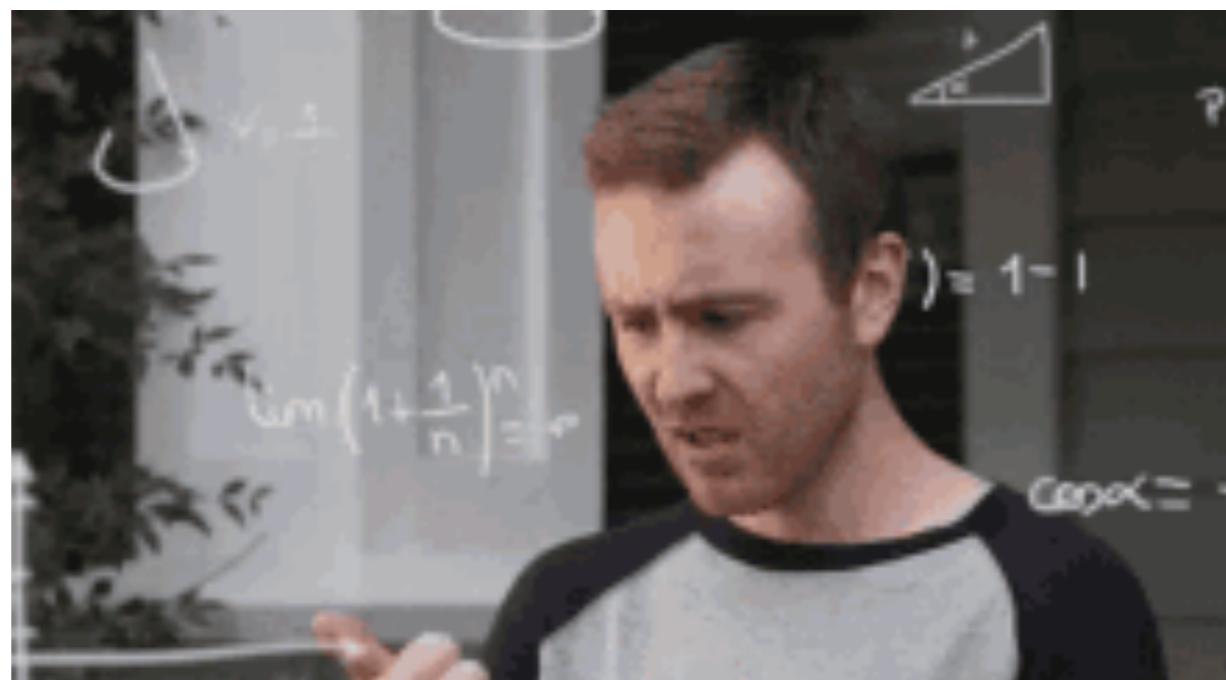
Answering questions about Penguins



1. Make this plot
2. A 60cm tall Penguin claims that no more than 10% are taller than her. Is she correct?
3. Are there more penguins between 50 and 55cm or between 55 and 65cm?
4. What size is a Penguin who is taller than 75% of the rest?



Show solutions in R Studio

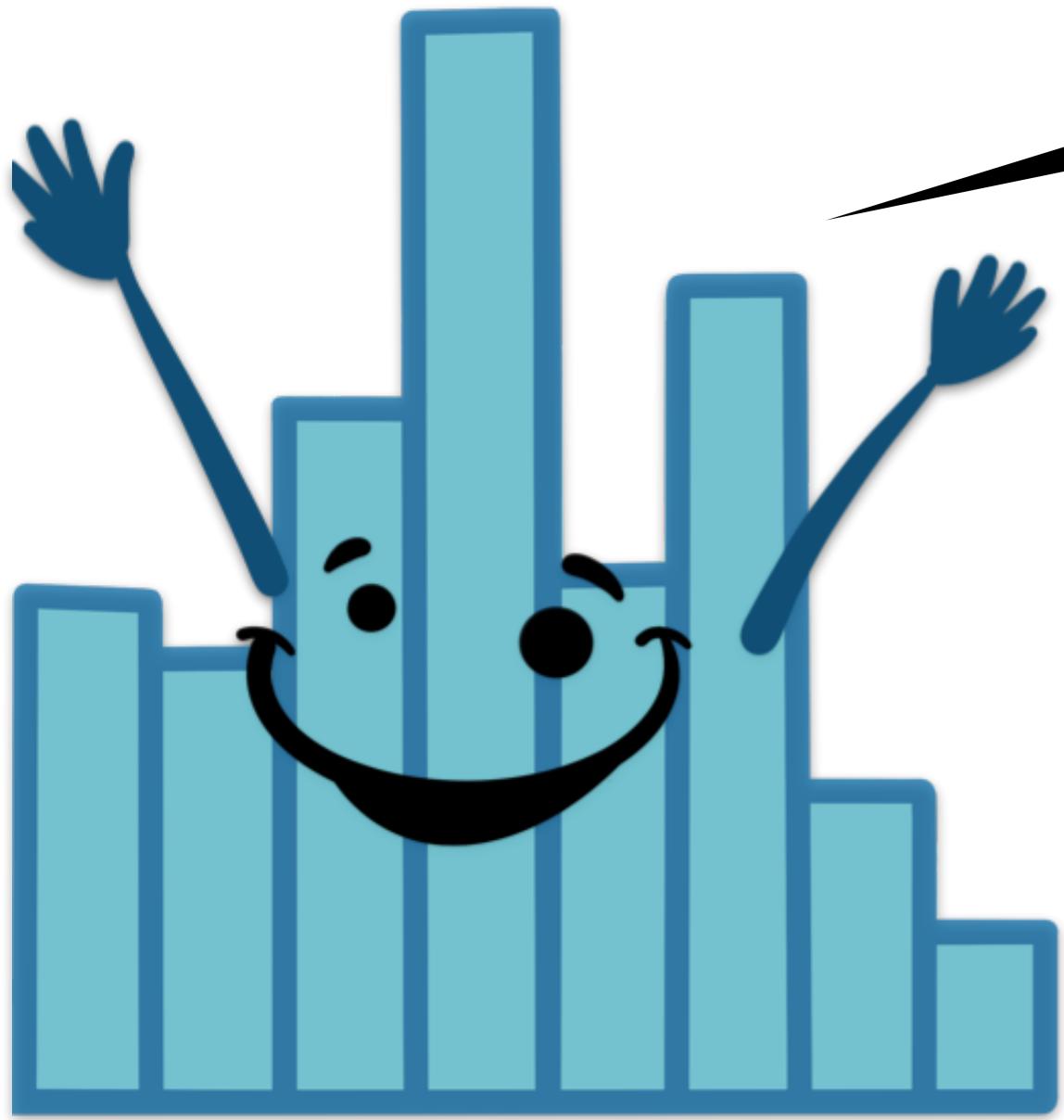


Quick recap

- we can draw random samples in R via
 - `sample()` from a vector
 - `slice_sample()` from a data frame
- generate random samples from a probability distribution via
 - `rnorm()`, `rbinom()`, `rgamma()`, ...
- understand how `density()` works
- answer questions about probabilities via the
 - analytic route: `qnorm()`, `pnorm()`
 - sampling route: `rnorm()` + data wrangling

01:00

stretch break!



Plan for today

- Simulating data
 - Drawing samples
 - Working with probability distributions
 - Quick detour: understanding `density()`
 - Asking probability distributions for answers
- **Doing Bayesian inference**
 - Analytic solution
 - Sampling solution

Summer camp

Register now for Summer Chess Camp!



**think
Move**
CHESS ACADEMY

All skill levels
welcome!

July 23 - July 27
and
August 13 - August 17

www.thinkmovechess.com



twice as many kids go to the basketball camp

$X \sim \text{Normal}(\mu = 170, \sigma = 8)$

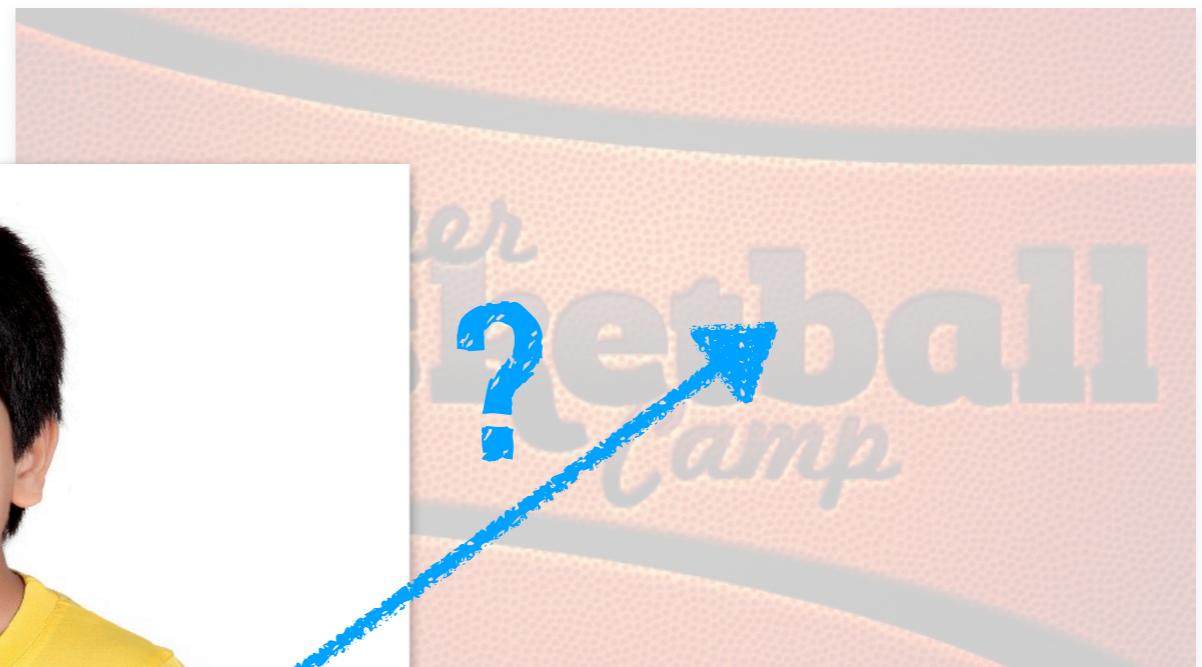
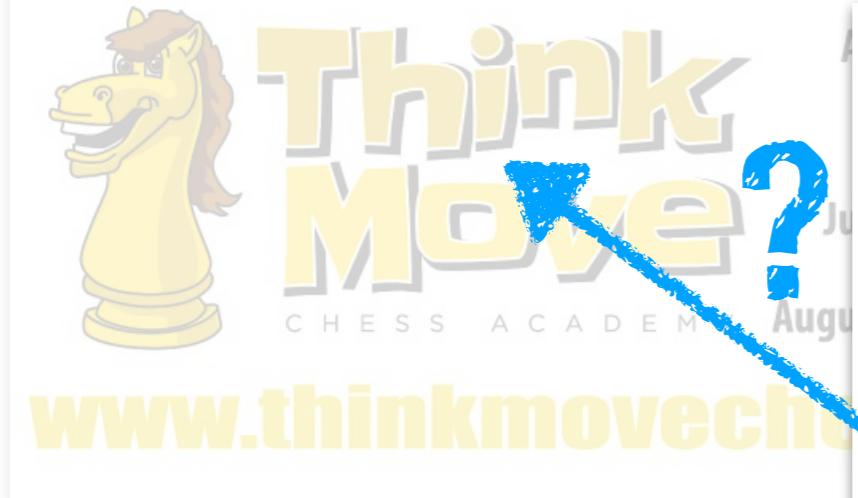


$X \sim \text{Normal}(\mu = 180, \sigma = 10)$



Summer camp

Register now for Summer Chess Camp!



Analytic solution

Can you feel the Bayes?

$H = \{\text{basketball, chess}\}$

$D = 175 \text{ cm}$

$$p(H | D) = \frac{\text{likelihood} \quad \text{prior}}{p(D)} \quad \begin{aligned} H &= \text{Hypothesis} \\ D &= \text{Data} \end{aligned}$$

probability of the data?!

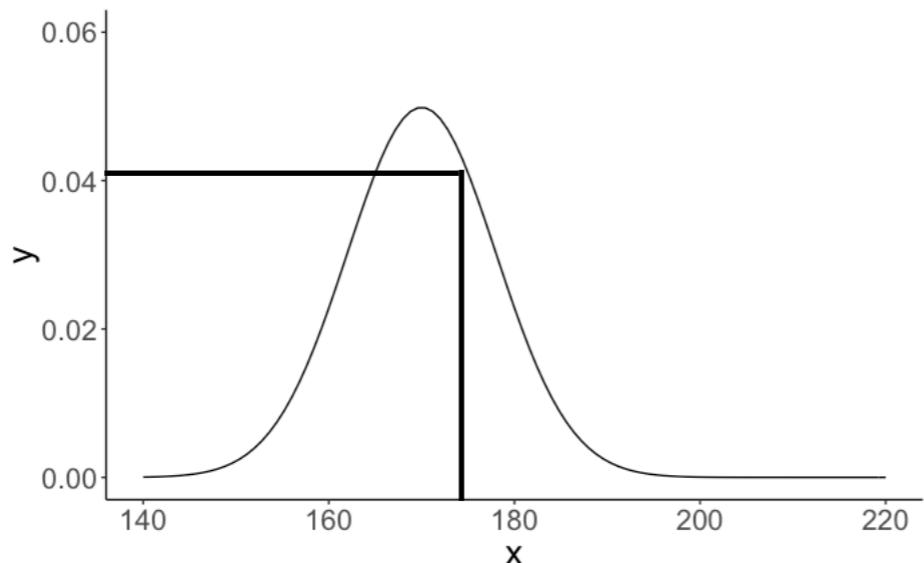
Summer camp

prior

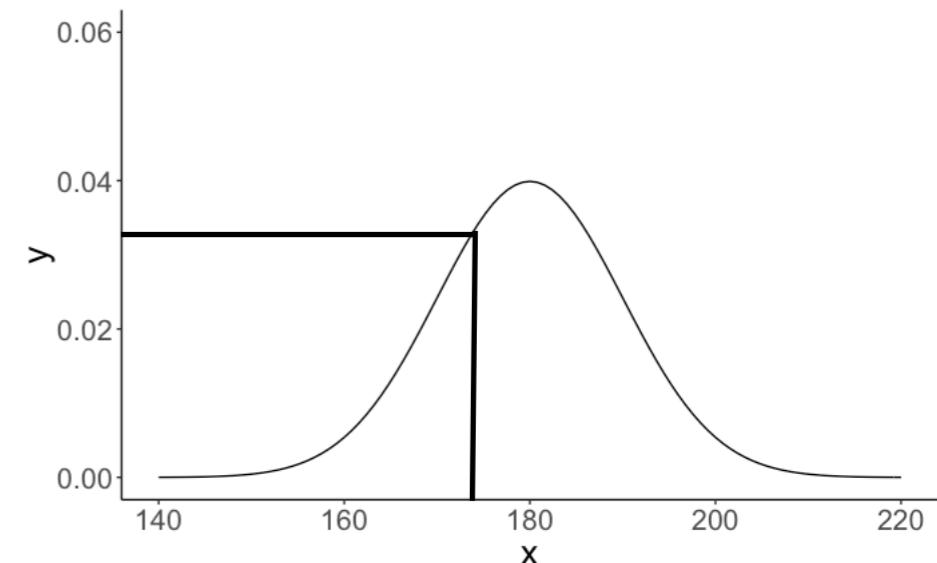
$$p(\text{chess}) = \frac{1}{3}$$

$$p(\text{basketball}) = \frac{2}{3}$$

likelihood



$$\begin{aligned} \text{dnorm}(175, \text{mean} = 170, \text{sd} = 8) \\ = 0.041 \end{aligned}$$



$$\begin{aligned} \text{dnorm}(175, \text{mean} = 180, \text{sd} = 10) \\ = 0.035 \end{aligned}$$

posterior

$$p(\text{sport} = \text{basketball} | \text{height} = 175) = \frac{p(175 | \text{basketball}) \cdot p(\text{basketball})}{p(175)}$$

likelihood **prior**

data

$$p(\text{basketball} | 175) = \frac{p(175 | \text{basketball}) \cdot p(\text{basketball})}{p(175 | \text{basketball}) \cdot p(\text{basketball}) + p(175 | \text{chess}) \cdot p(\text{chess})}$$

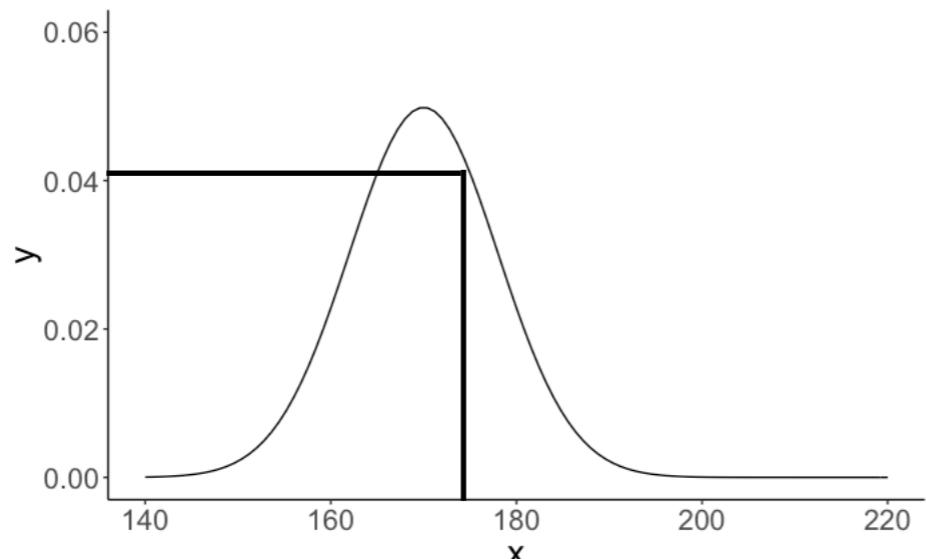
Summer camp

prior

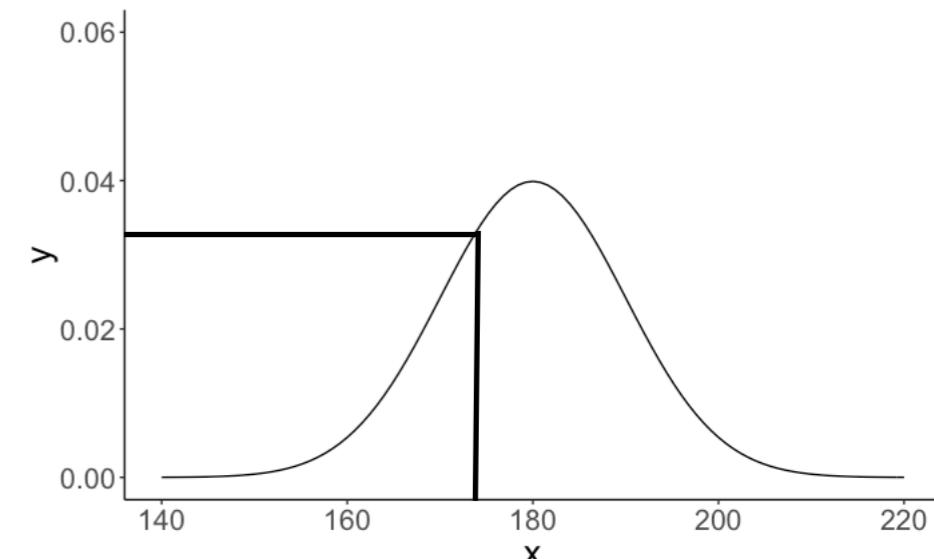
$$p(\text{chess}) = \frac{1}{3}$$

$$p(\text{basketball}) = \frac{2}{3}$$

likelihood



$$\begin{aligned} \text{dnorm}(175, \text{mean} = 170, \text{sd} = 8) \\ = 0.041 \end{aligned}$$



$$\begin{aligned} \text{dnorm}(175, \text{mean} = 180, \text{sd} = 10) \\ = 0.035 \end{aligned}$$

posterior

$$p(\text{basketball} | 175) = \frac{p(175 | \text{basketball}) \cdot p(\text{basketball})}{p(175 | \text{basketball}) \cdot p(\text{basketball}) + p(175 | \text{chess}) \cdot p(\text{chess})}$$

$$p(\text{basketball} | 175) = \frac{0.035 \cdot 2/3}{0.035 \cdot 2/3 + 0.041 \cdot 1/3} \approx 0.63$$

send the kid to
the basketball
gym!

Sampling solution

Summer camp: Via sampling

```
1 df.camp = tibble(  
2   kid = 1:1000,  
3   sport = sample(c("chess", "basketball"),  
4     size = 1000,  
5     replace = T,  
6     prob = c(1/3, 2/3))) %>%  
7   rowwise() %>%  
8   mutate(height = ifelse(test = sport == "chess",  
9     yes = rnorm(., mean = 170, sd = 8),  
10    no = rnorm(., mean = 180, sd = 10))) %>%  
11  ungroup())
```

kid	sport	height
1	basketball	164.84
2	basketball	163.22
3	basketball	191.18
4	chess	160.16
5	basketball	182.99
6	chess	163.54
7	chess	168.56
8	basketball	192.99
9	basketball	171.91
10	basketball	177.12

```
1 df.camp %>%  
2   filter(height == 175) %>%  
3   count(sport)
```

doesn't work!

Summer camp: Via sampling

```
1 df.camp = tibble(  
2   kid = 1:100000,  
3   sport = sample(c("chess", "basketball"),  
4     size = 100000,  
5     replace = T,  
6     prob = c(1/3, 2/3))) %>%  
7   rowwise() %>%  
8   mutate(height = ifelse(test = sport == "chess",  
9     yes = rnorm(., mean = 170, sd = 8),  
10    no = rnorm(., mean = 180, sd = 10))) %>%  
11 ungroup())
```

kid	sport	height
1	basketball	164.84
2	basketball	163.22
3	basketball	191.18
4	chess	160.16
5	basketball	182.99
6	chess	163.54
7	chess	168.56
8	basketball	192.99
9	basketball	171.91
10	basketball	177.12

```
1 df.camp %>%  
2   filter(between(height,  
3     left = 174,  
4     right = 176)) %>%  
5   count(sport)
```

this works!

sport	n
basketball	469
chess	273

$$\frac{\text{basketball}}{\text{basketball} + \text{chess}} \approx 0.63$$

Plan for today

- Simulating data
 - Drawing samples
 - Working with probability distributions
 - Quick detour: understanding `density()`
 - Asking probability distributions for answers
- Doing Bayesian inference
 - Analytic solution
 - Sampling solution

Feedback

How was the pace of today's class?

much a little just a little much
too too right too too
slow slow

How happy were you with today's class overall?



What did you like about today's class? What could be improved next time?

Thank you!