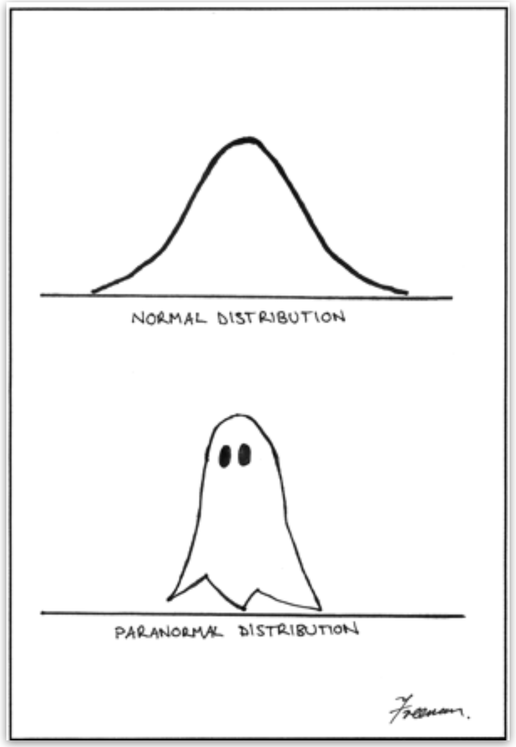
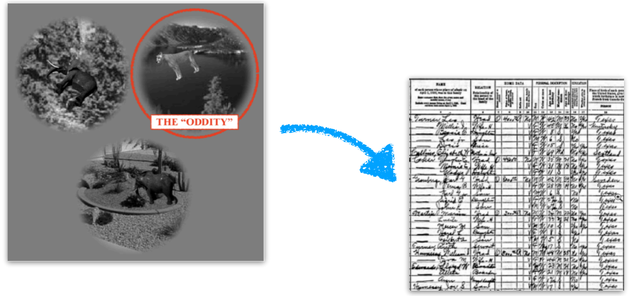
PSYCH 252: Statistical Methods for Behavioral and Social Sciences

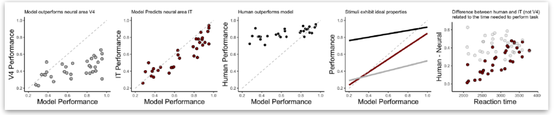
Simulation 1



Logistics

Homework 2





- Due **Thursday Jan 22, at 8pm**
- Don't wait until the very last moment to knit your RMarkdown file into a pdf. It may not compile and debugging takes time ...
- You can upload earlier versions of your homework on Canvas and still update until the deadline.
- Get and give help via Ed Discussion!

Plan for today

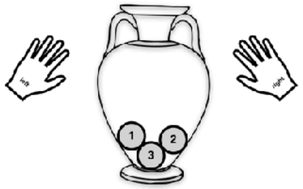
- Quick recap
- Simulating data
  - Drawing samples
  - Working with probability distributions
  - Quick detour: understanding `density()`
  - Asking probability distributions for answers
- Doing Bayesian inference
  - Analytic solution
  - Sampling solution

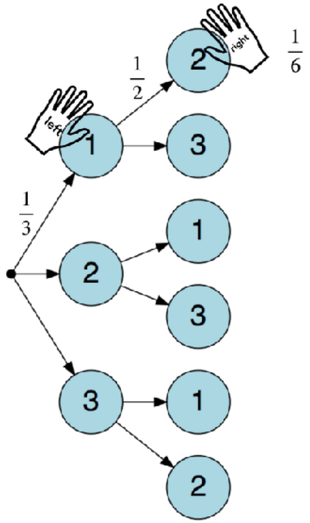
Quick recap

Quick recap

Sampling **without** replacement

$p(\text{left} = 1, \text{right} = 2) = ?$





What is the probability that I first draw the 1 with my left hand, and then, without putting the 1 back into the urn, draw the 2 with my right hand?

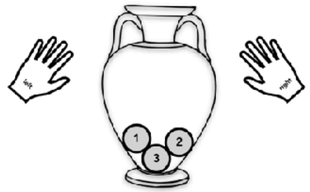
15

Definitions

If $P(X_i)$ is the probability of event X_i

1. Probability cannot be negative.

$$P(X_i) \geq 0$$

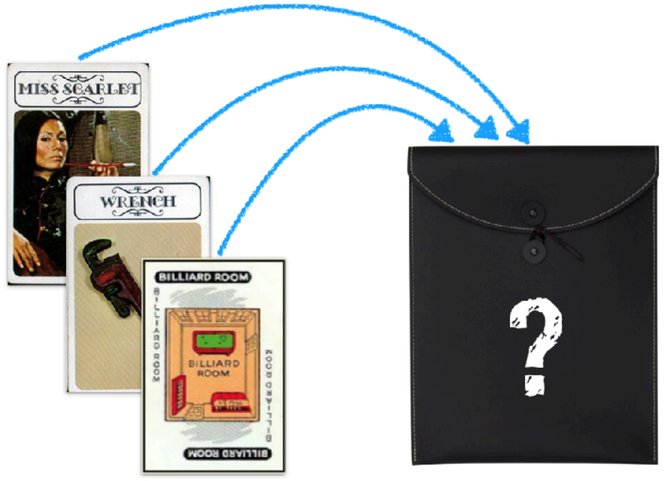


2. Total probability of all outcomes in the sample space is 1.

$$\sum_{i=1}^{N} P(X_i) = P(X_1) + P(X_2) + \ldots + P(X_N) = 1$$

23

Clue guide to probability



28

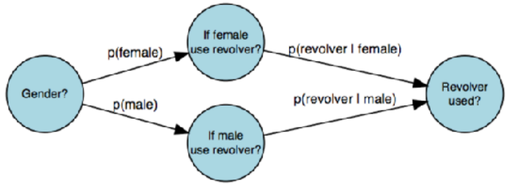
Clue guide to probability

- *law of total probability*
- **Definition**:

$$p(A) = p(A \mid B) \cdot p(B) + p(A \mid \neg B) \cdot p(\neg B)$$

$$p(A) = \sum_{i=1}^{n} p(A \mid B_i) \cdot p(B_i)$$

$p(\text{what} = \text{Revolver}) = ?$



Who?

What?

39

Quick recap

Clue guide to probability

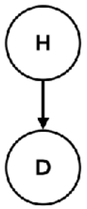
$$p(B \mid A) = \frac{p(A \mid B) \cdot p(B)}{p(A)}$$

$$\text{posterior}\quad p(H \mid D) = \frac{\overbrace{p(D \mid H)}^{\text{likelihood}} \cdot \overbrace{p(H)}^{\text{prior}}}{p(D)}$$

subjective probability interpretation

H = Hypothesis

D = Data



formal framework for learning from data

updating our prior belief $p(H)$, to a posterior belief $p(H \mid D)$ given some data

Clue guide to probability

what we know

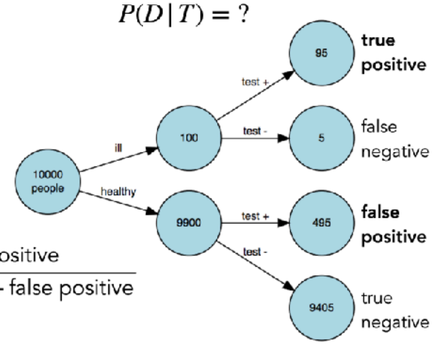
$P(D) = 0.01$

$P(T \mid D) = 0.95$

$P(T \mid \neg D) = 0.05$

what we want to know

$P(D \mid T) = ?$



$$P(D \mid T) = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} = \frac{95}{95 + 495} \approx 0.16$$

Getting Bayes right matters!

Claim:
"White officers are not more likely to shoot minority civilians than non-White officers"

$$\Pr(\text{shot} \mid \text{minority civilian, white officer}, X) - \Pr(\text{shot} \mid \text{minority civilian, minority officer}, X) \leq 0, \quad [1]$$

What the statistic says:
"whether a person fatally shot was more likely to be Black (or Hispanic) than White"

Original claim:
Requires Bayes' rule

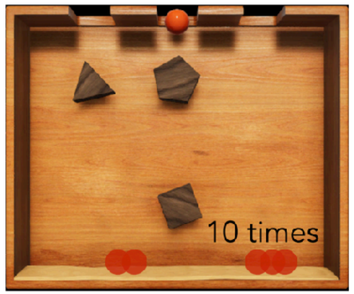
$$\Pr(\text{shot} \mid \text{minority civilian, white officer}, X) - \Pr(\text{shot} \mid \text{minority civilian, minority officer}, X) = \frac{\Pr(\text{min. civ.} \mid \text{shot, white off.}, X) \times \Pr(\text{shot} \mid \text{white off.}, X)}{\Pr(\text{minority civilian} \mid \text{white officer}, X)} - \frac{\Pr(\text{min. civ.} \mid \text{shot, min. off.}, X) \times \Pr(\text{shot} \mid \text{min. off.}, X)}{\Pr(\text{minority civilian} \mid \text{minority officer}, X)} \quad [2]$$

authors didn't have the relevant data to support their claim!

paper was retracted

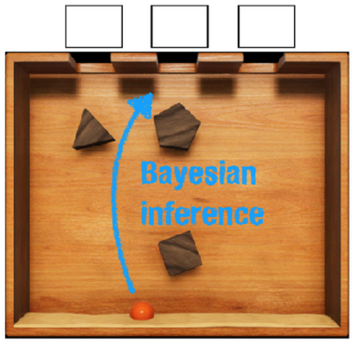
Johnson, D. J., Tress, T., Burkel, N., Taylor, C., & Cesario, J. (2019). Officer characteristics and racial disparities in fatal officer-involved shootings. *Proceedings of the National Academy of Sciences*, 116(32), 15877–15882.
Knox, D., & Mummolo, J. (2020). Making inferences about racial disparities in police violence. *Proceedings of the National Academy of Sciences*, 117(3), 1261–1262.

Prediction



Where will the ball land?

Inference



In which hole was the ball dropped?

Gerstenberg, T., Siegel, M. H., & Tenenbaum, J. B. (2021). What happened? Reconstructing the past from vision and sound. PsyArXiv. https://psyarxiv.com/tfjdk

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)
```

sample 10 times

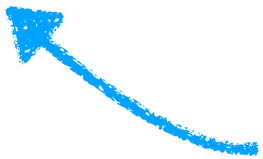
with replacement please

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

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))
```



I want mostly 1s

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

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
```

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

```
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 get the same outcome

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
```

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

```
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 get the same outcome

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

sample 6 rows with replacement

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

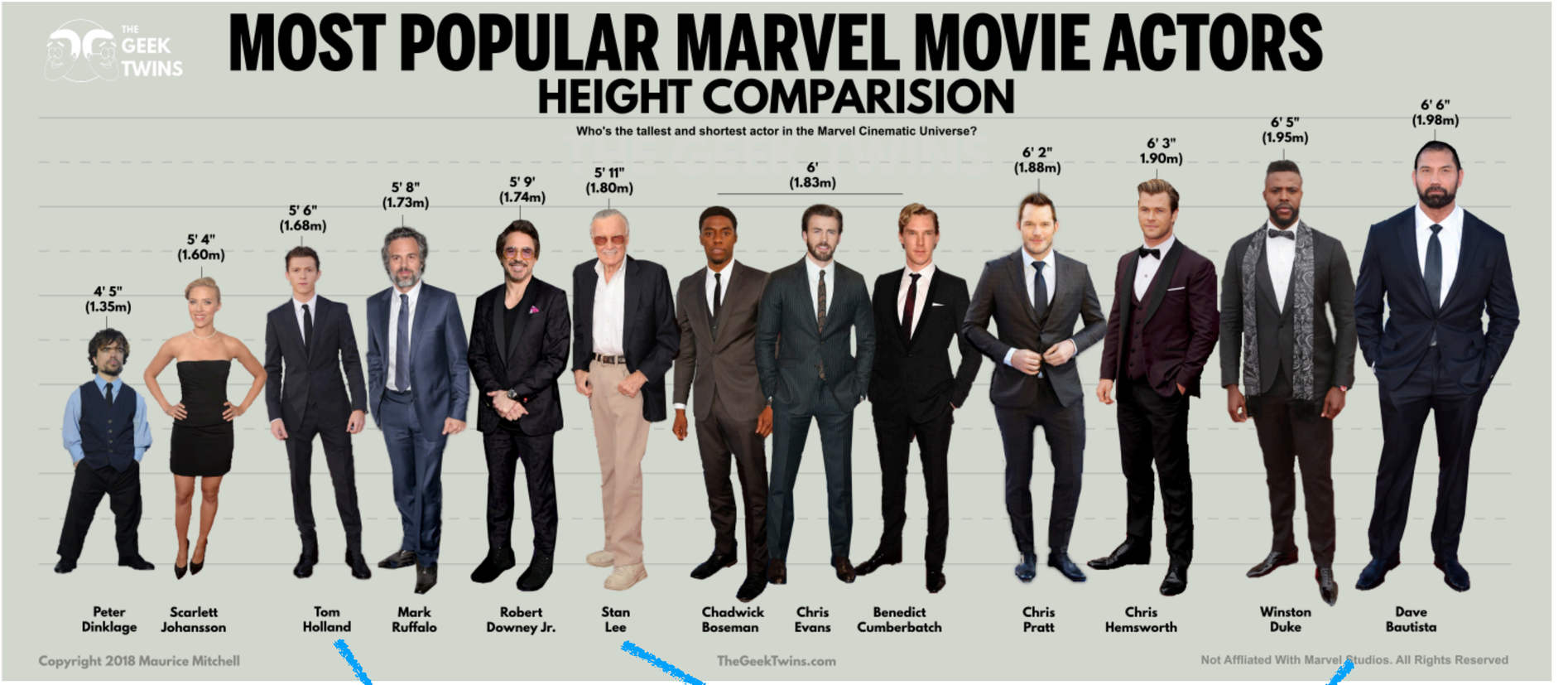
trial	stimulus	rating
9	pet	9
4	flower	5
7	flower	10
1	flower	3
2	pet	1
7	flower	10

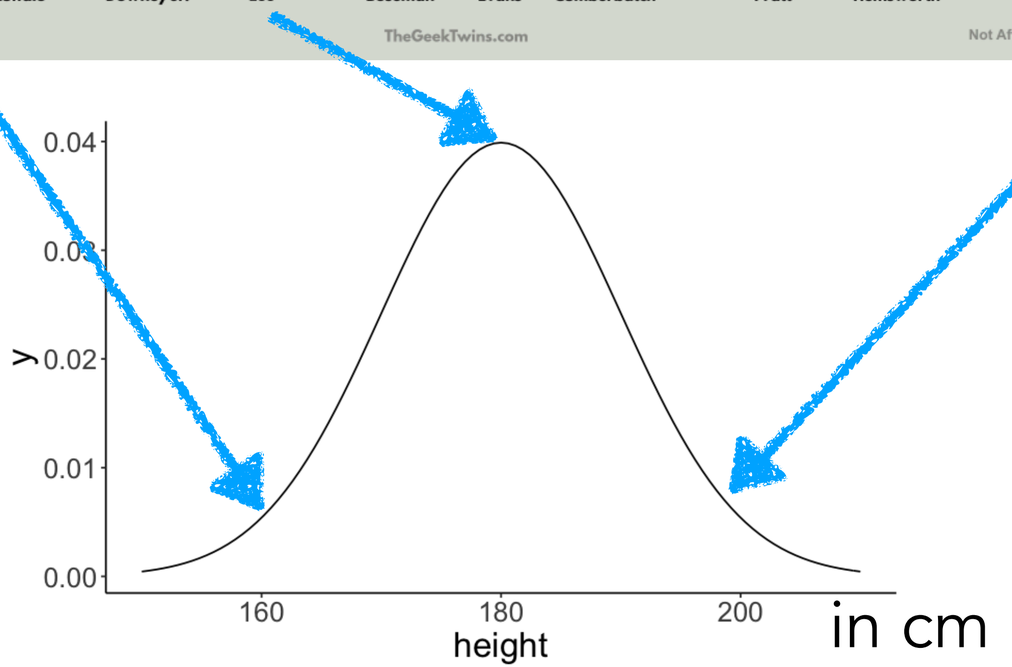
sample 50% of the rows

```
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


THE GEEK TWINS
MOST POPULAR MARVEL MOVIE ACTORS
HEIGHT COMPARISION
Who's the tallest and shortest actor in the Marvel Cinematic Universe?
4' 5" (1.35m)
5' 4" (1.60m)
5' 6" (1.68m)
5' 8" (1.73m)
5' 9" (1.74m)
5' 11" (1.80m)
6' (1.83m)
6' 2" (1.88m)
6' 3" 1.90m)
6' 5" (1.95m)
6' 6" (1.98m)
Peter Dinklage
Scarlett Johansson
Tom Holland
Mark Ruffalo
Robert Downey Jr.
Stan Lee
Chadwick Boseman
Chris Evans
Benedict Cumberbatch
Chris Pratt
Chris Hemsworth
Winston Duke
Dave Bautista
Copyright 2018 Maurice Mitchell
TheGeekTwins.com
Not Affliated With Marvel Studios. All Rights Reserved


0.04
0.03
0.02
0.01
0.00
y
160
180
200
height

in cm

Working with probability distributions

_norm()

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

Normal distribution

make data frame with minimum and maximum x-value

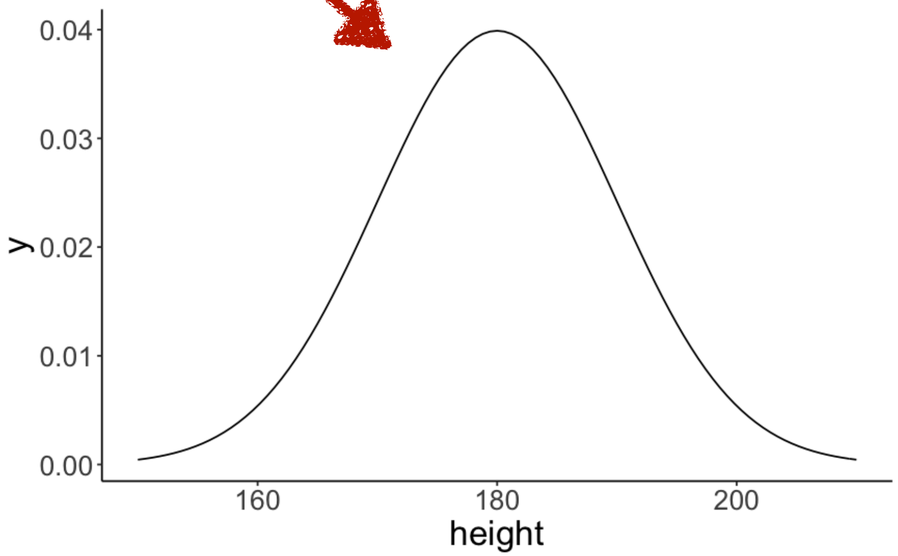
```
1 ggplot(data = tibble(height = c(150, 210)),
2        mapping = aes(x = height)) +
3   stat_function(fun = ~ dnorm(x = .,
4                               mean = 180,
5                               sd = 10))
```

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

n = number of samples

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

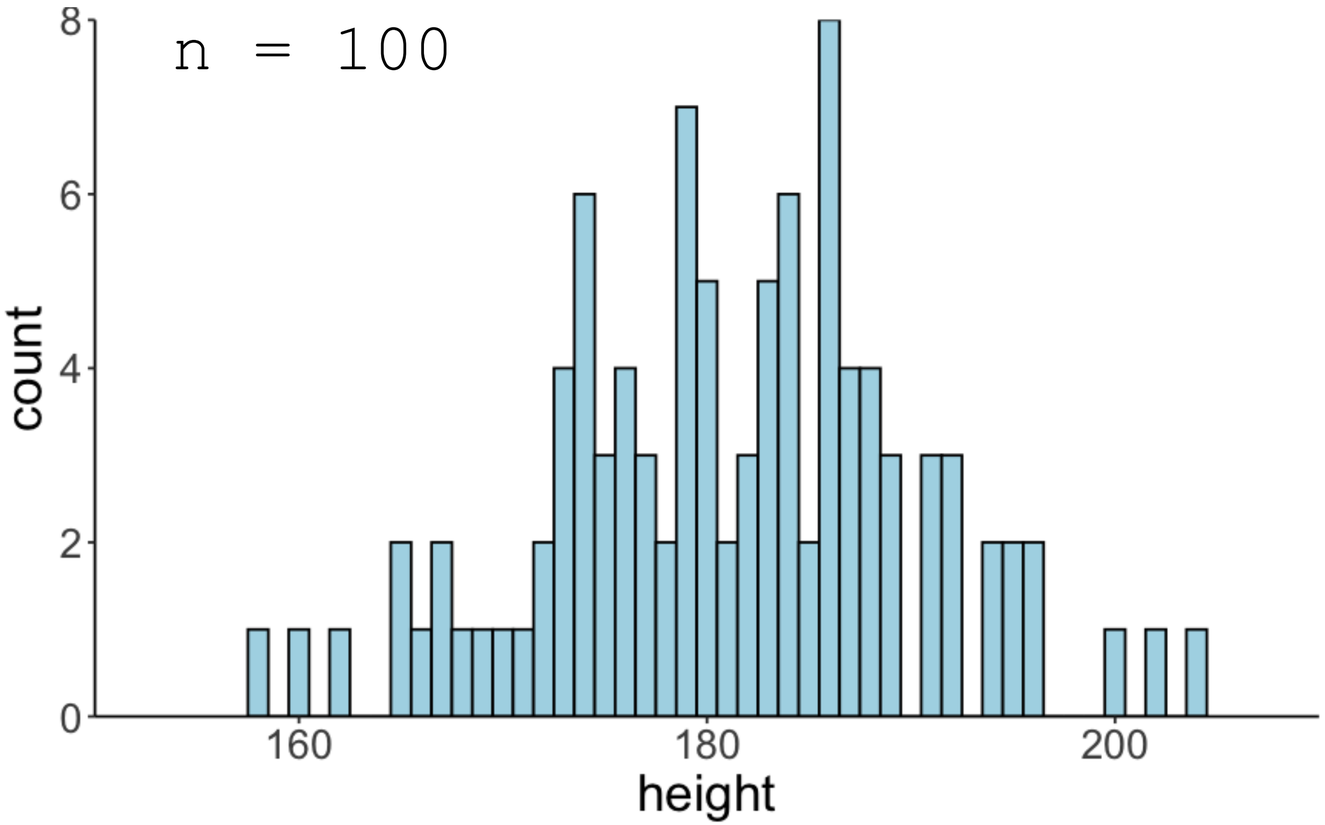
r = random samples **norm** = normal distribution

1	173.7355
2	181.8364
3	171.6437
4	195.9528
5	183.2951
6	171.7953
7	184.8743
8	187.3832
9	185.7578
10	176.9461
11	195.1178
12	183.8984
13	173.7876

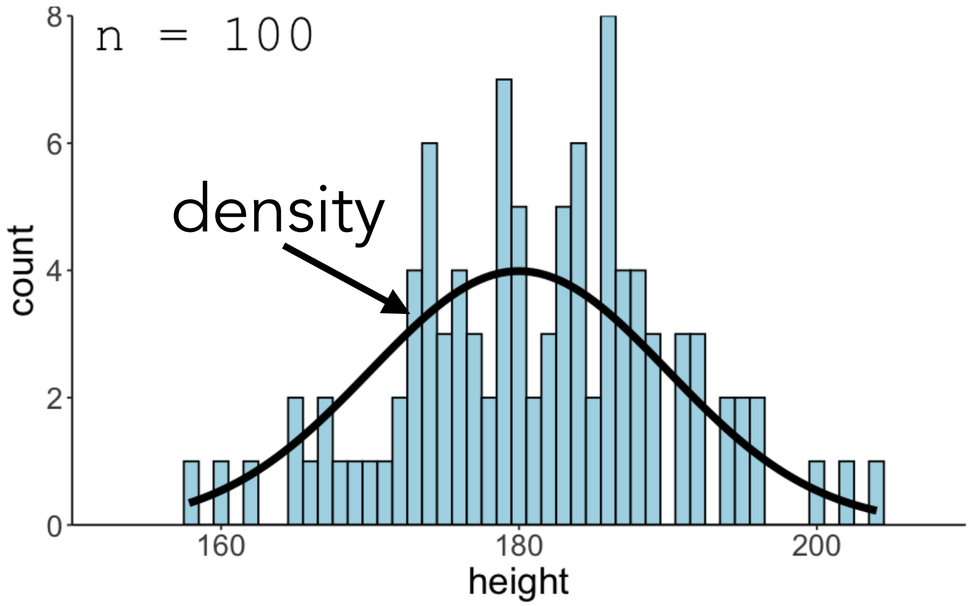
⋮

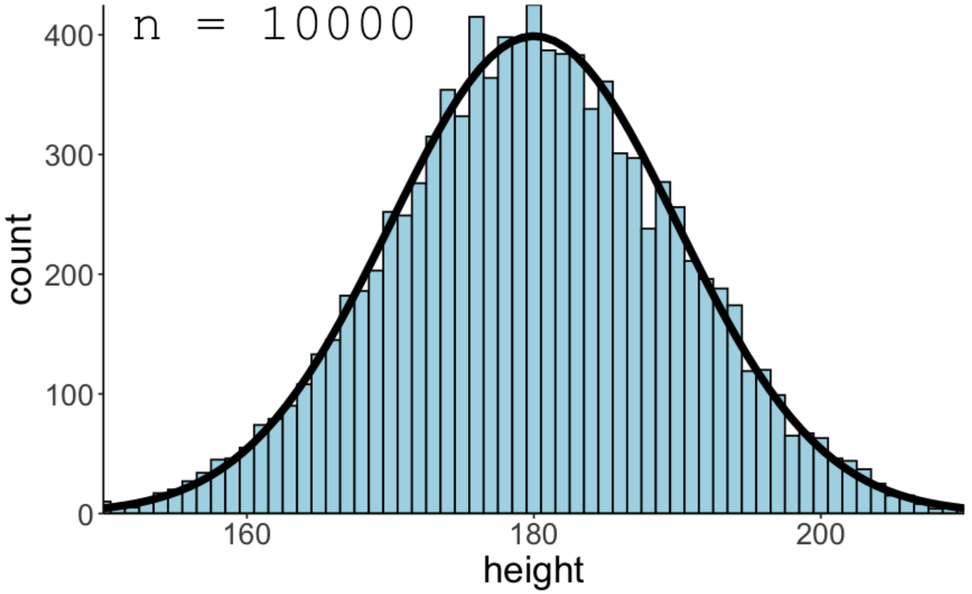
nrow = 100

n = 100



Sampling from distributions

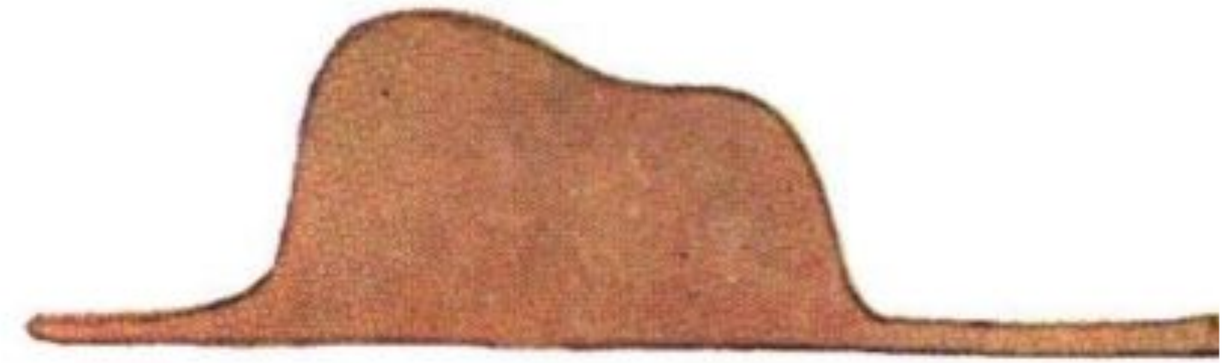


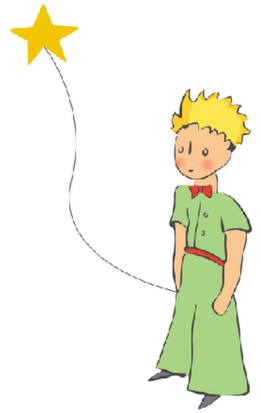


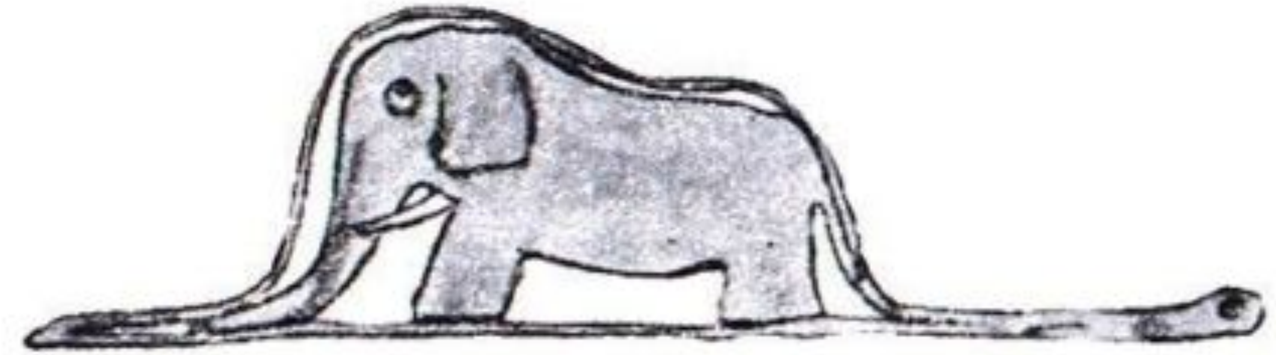
law of large numbers

approximation to true underlying distribution improves with increased sample size

Quick detour: understanding `density()`





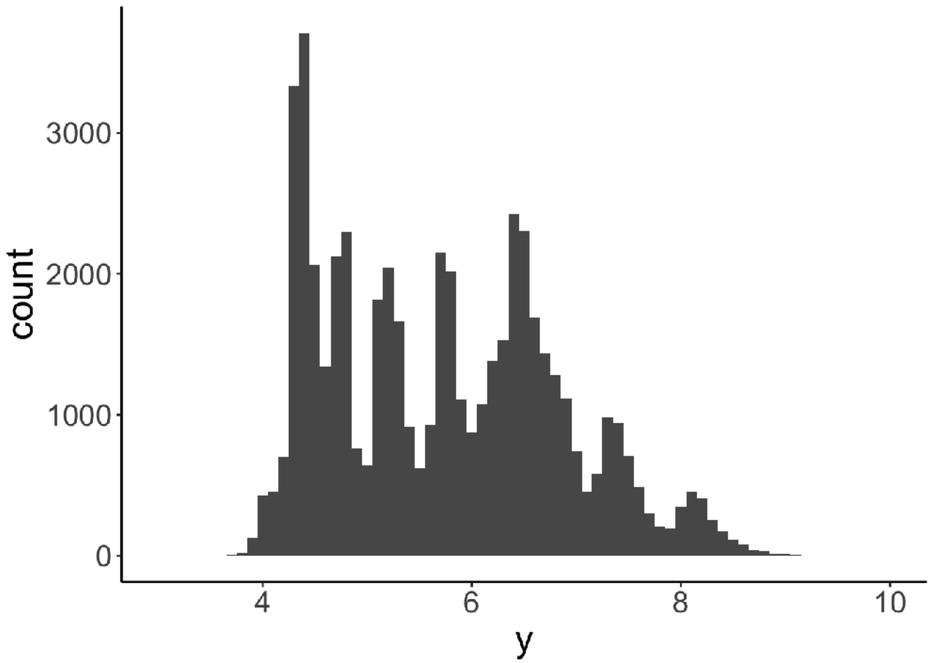


What's underneath the hood (or hat)?

You've seen `density()` before

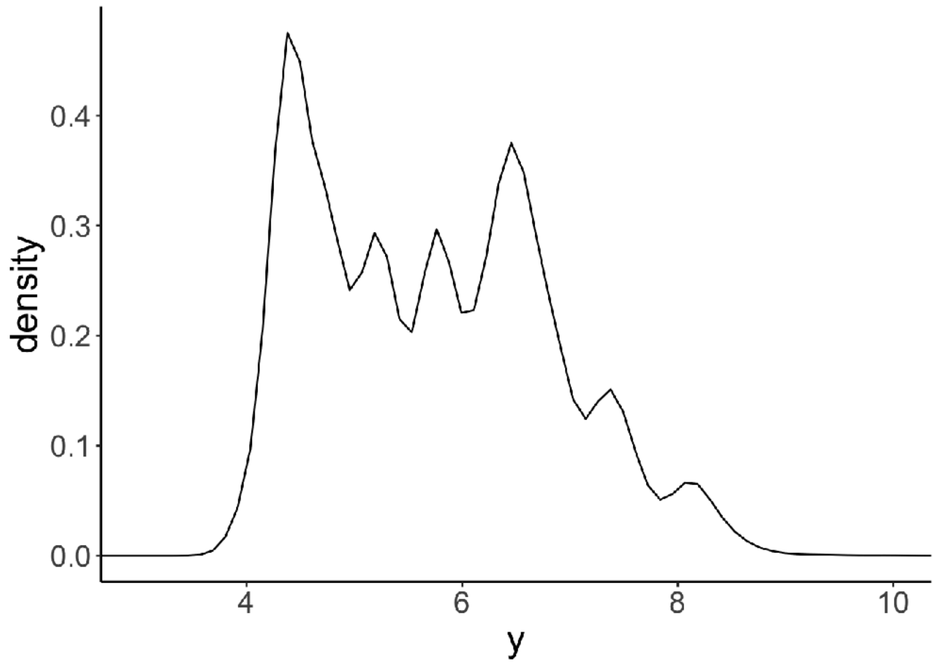
Histogram

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



Density

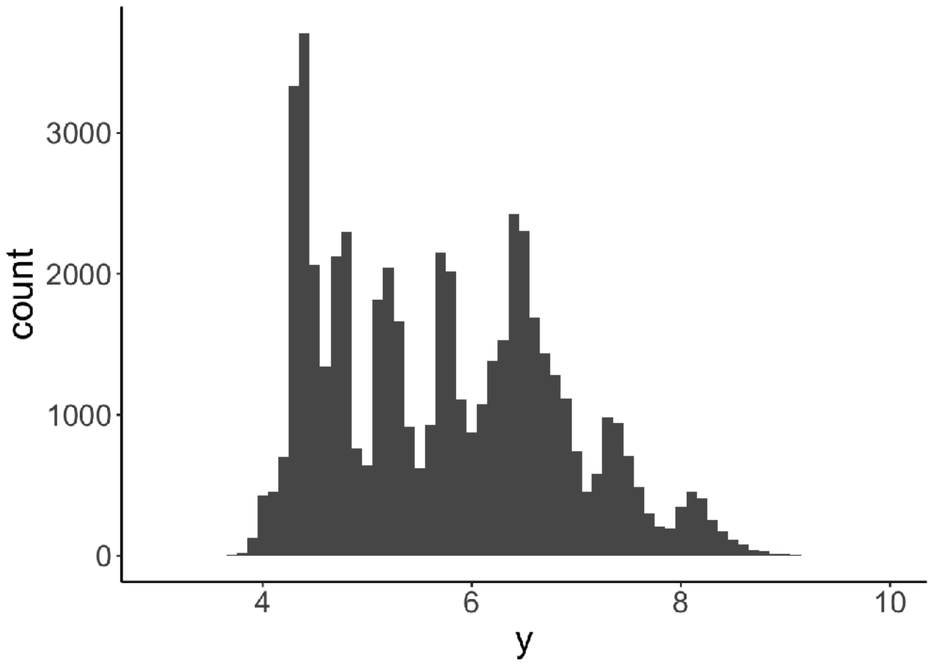
```
1 ggplot(data = df.diamonds,
2        mapping = aes(x = y)) +
3   geom_density() +
4   coord_cartesian(xlim = c(3, 10))
```



You've seen `density()` before

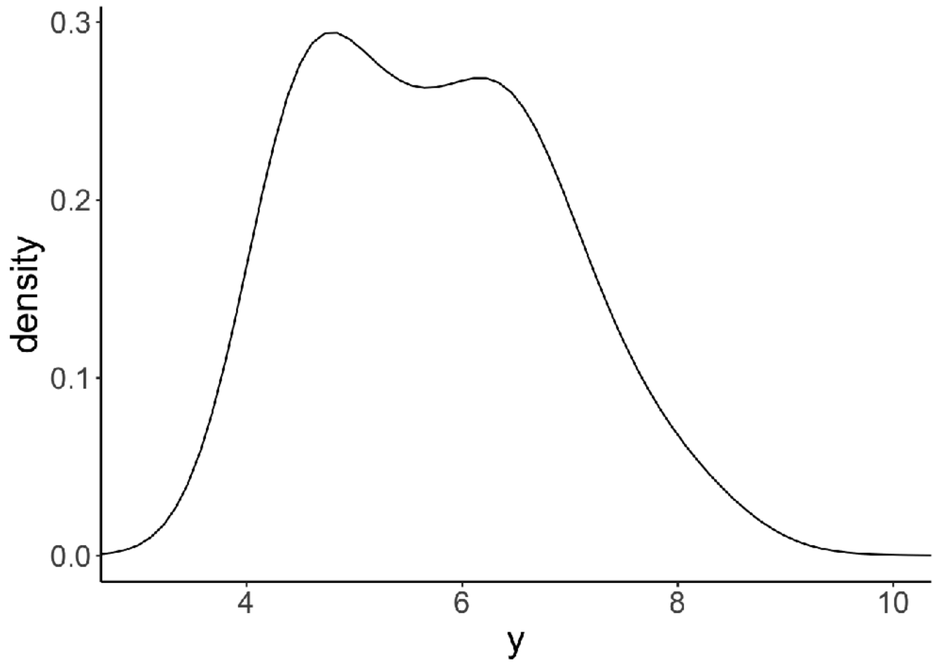
Histogram

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



Density

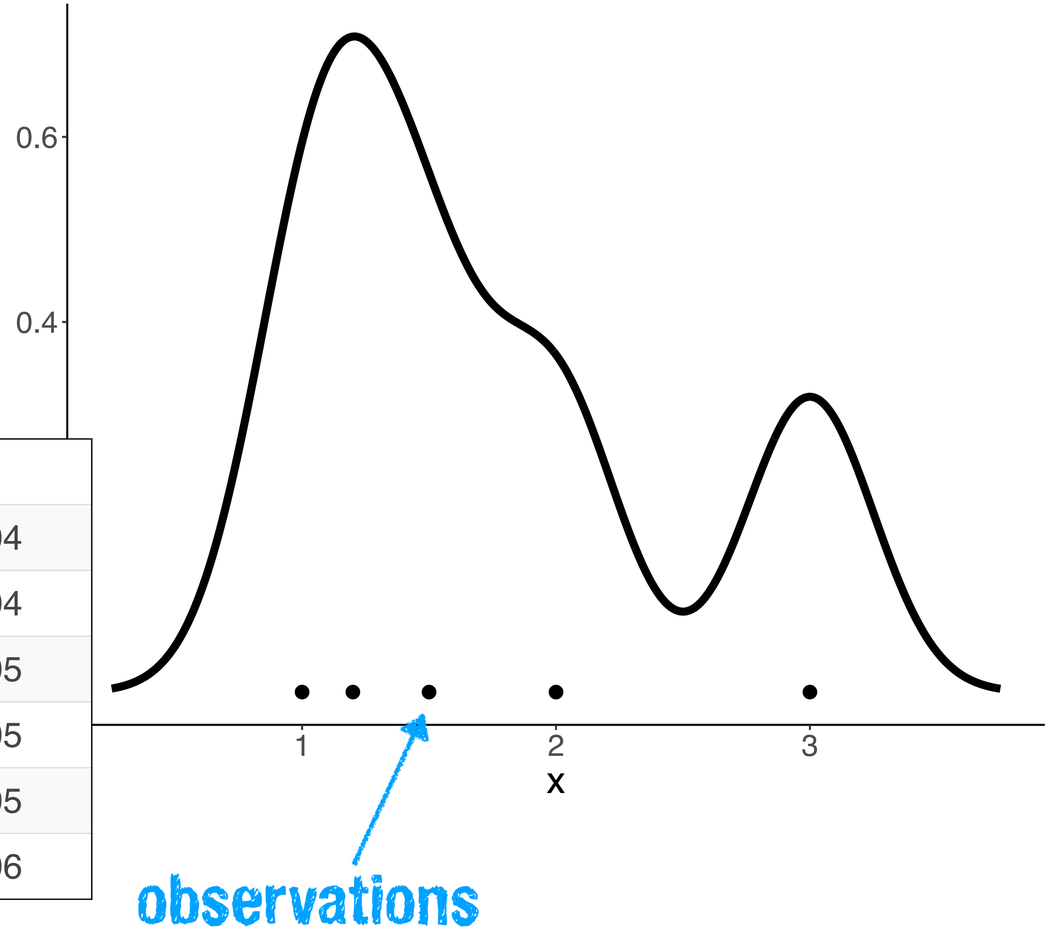
```
1 ggplot(data = df.diamonds,
2        mapping = aes(x = y)) +
3   geom_density(bw = 0.5) +
4   coord_cartesian(xlim = c(3, 10))
```



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)
```

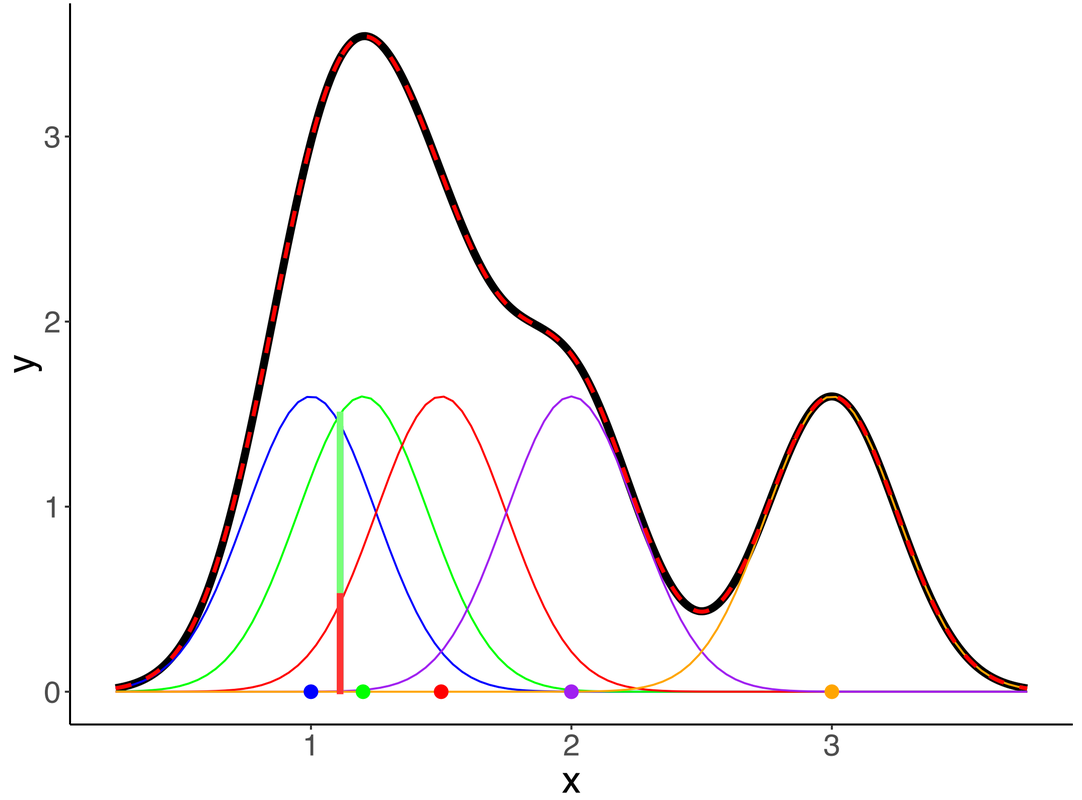
x	y
0.250	0.004
0.257	0.004
0.264	0.005
0.271	0.005
0.277	0.005
0.284	0.006



`nrow = 512`

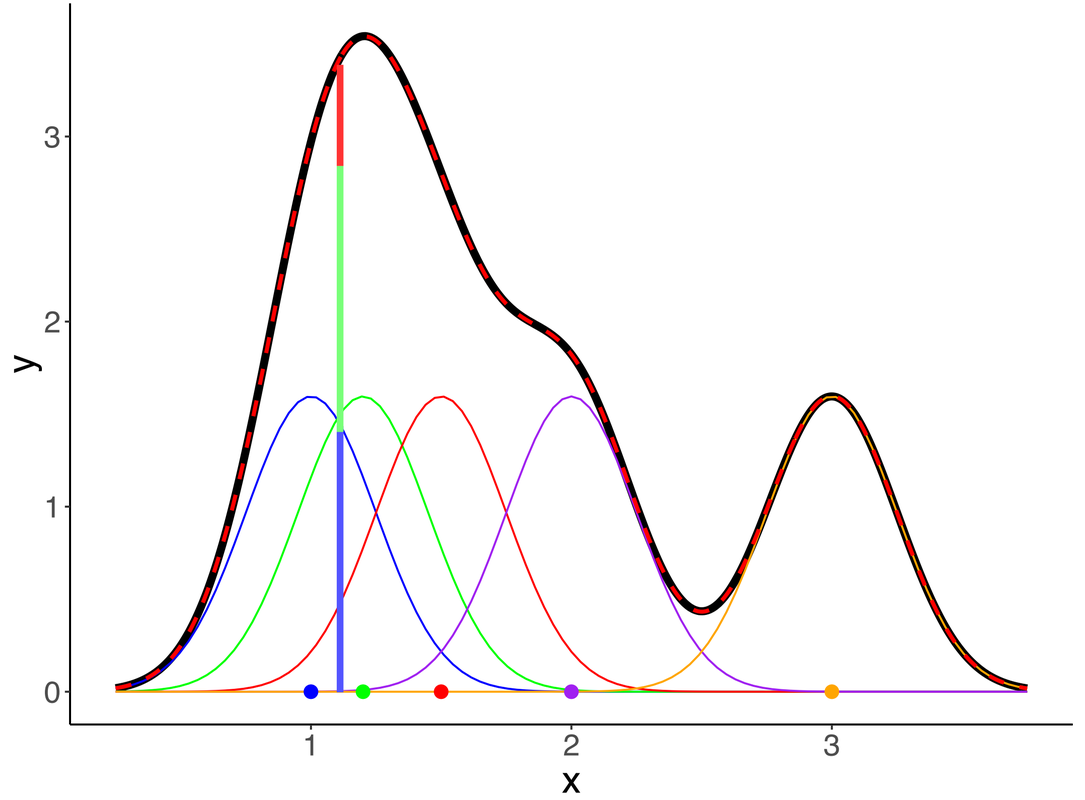
Understanding `density()`

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

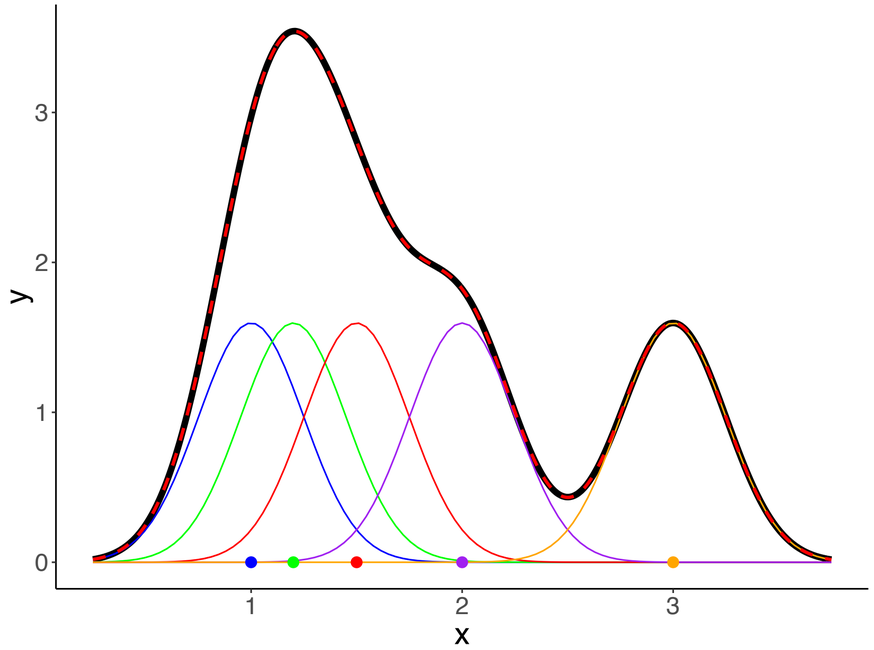


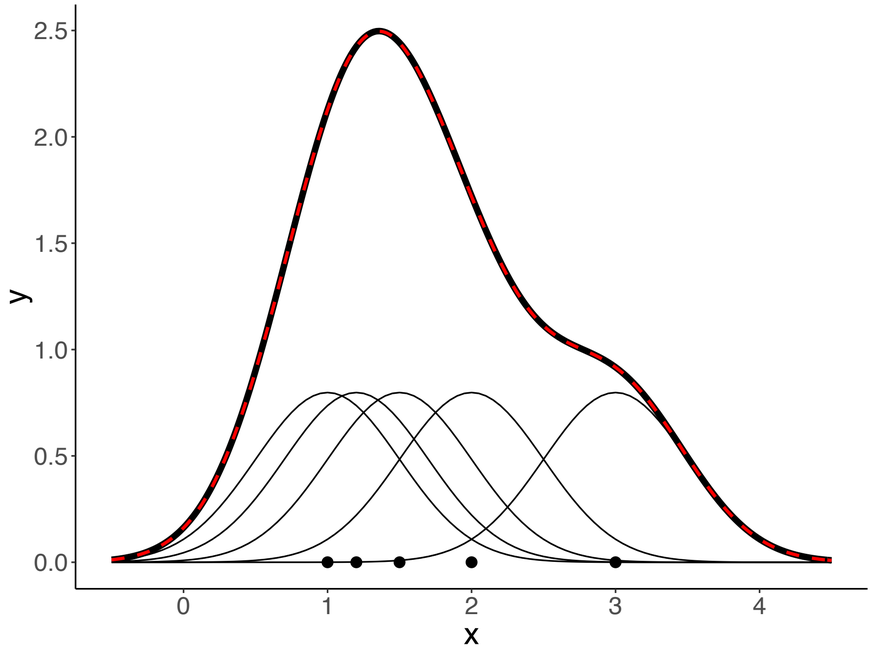
Understanding `density()`

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()`



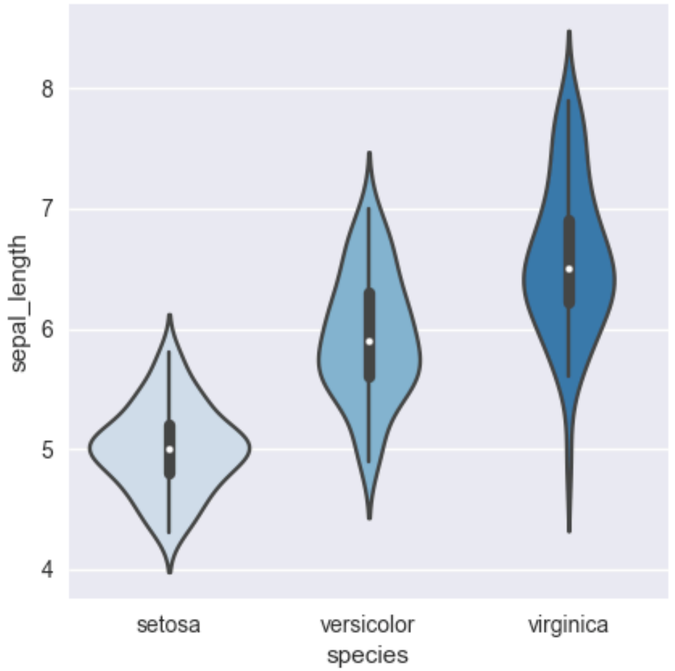


`density``(bw = 0.25)`

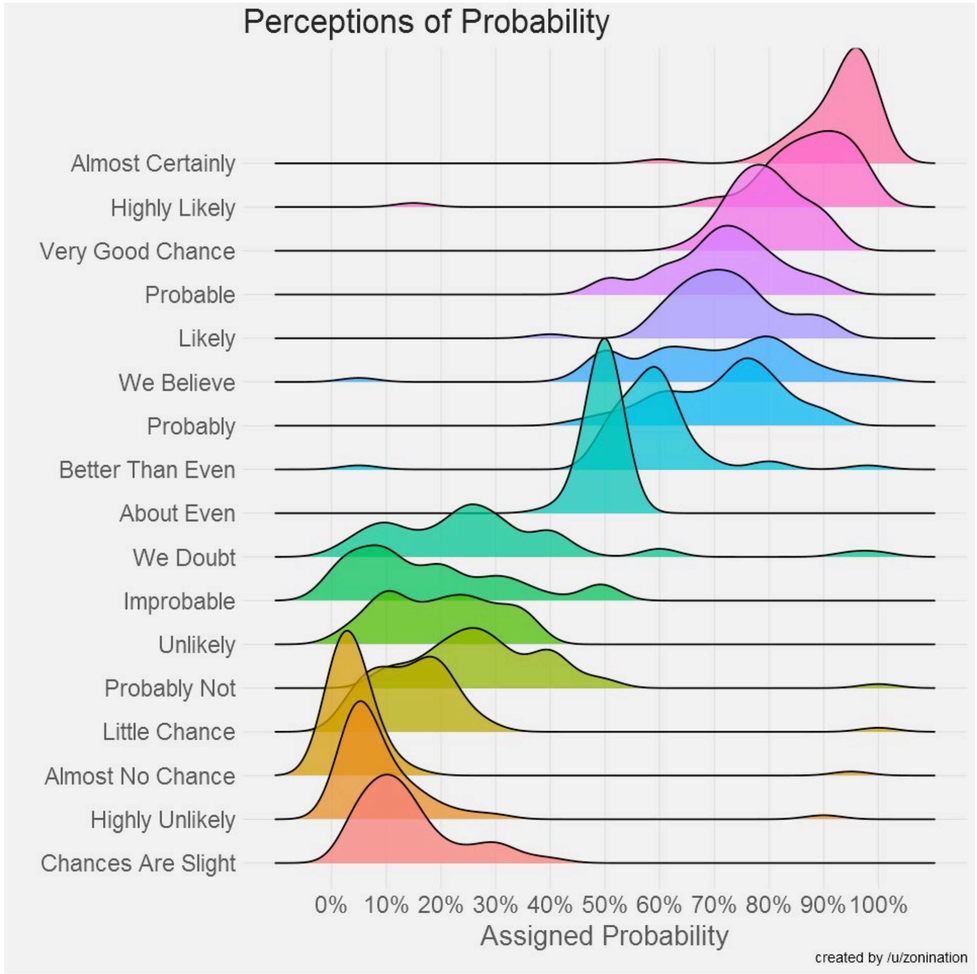
`density``(bw = 0.5)`

Understanding `density()`

violinplot



joyplot



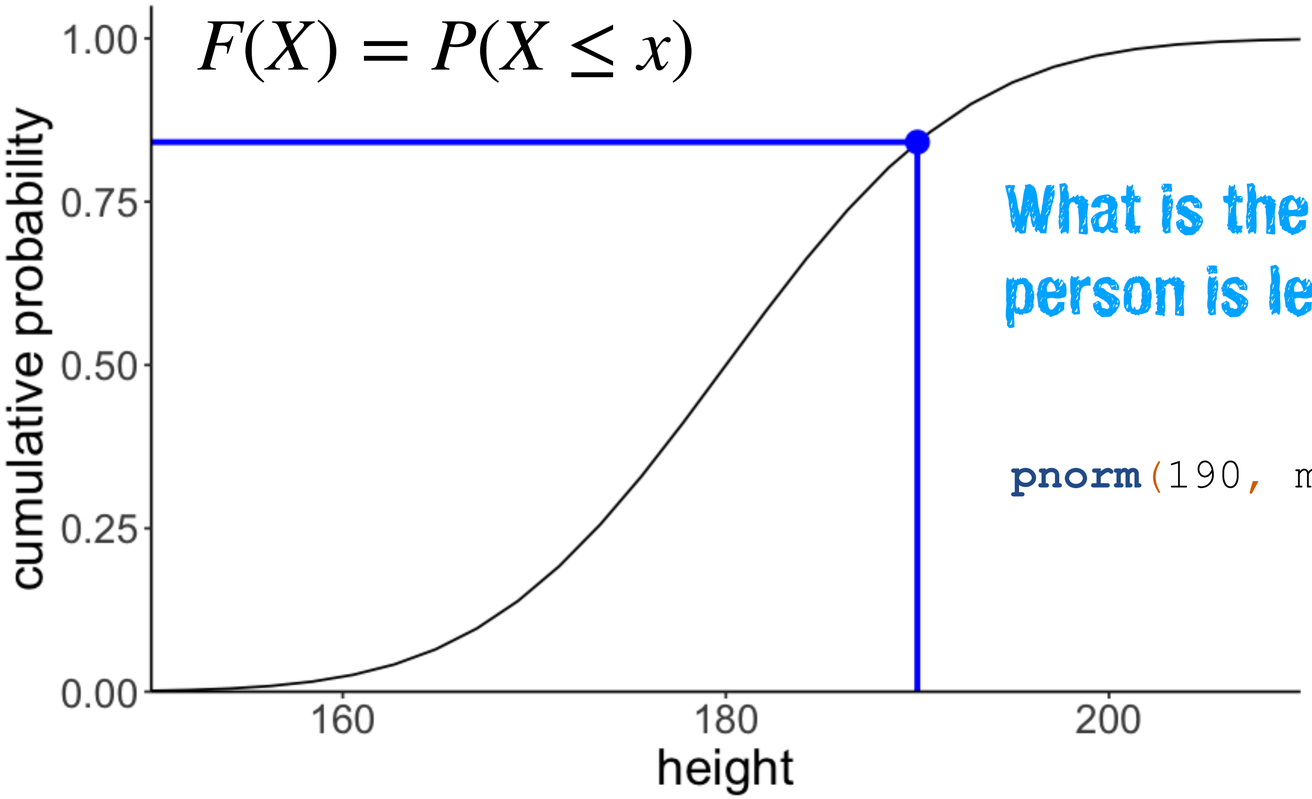
Asking probability distributions for answers

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



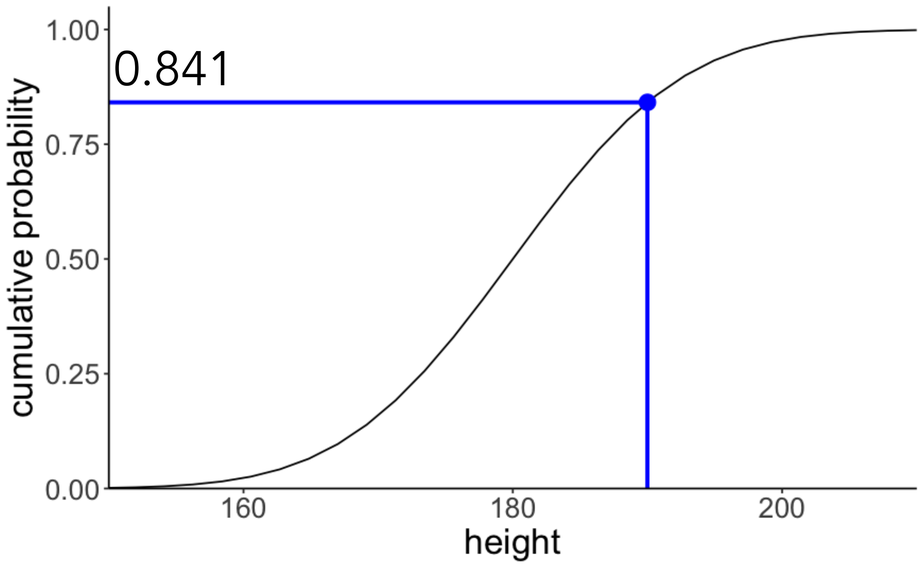
What is the probability that a person is less than 190cm tall?

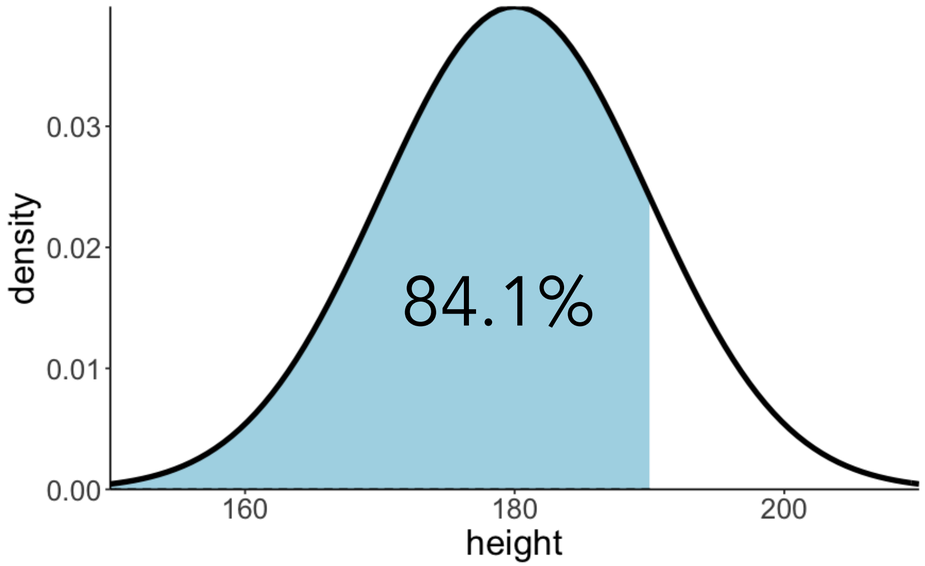
```
pnorm(190, mean = 180, sd = 10)
```

= 0.841

Computing probabilities

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





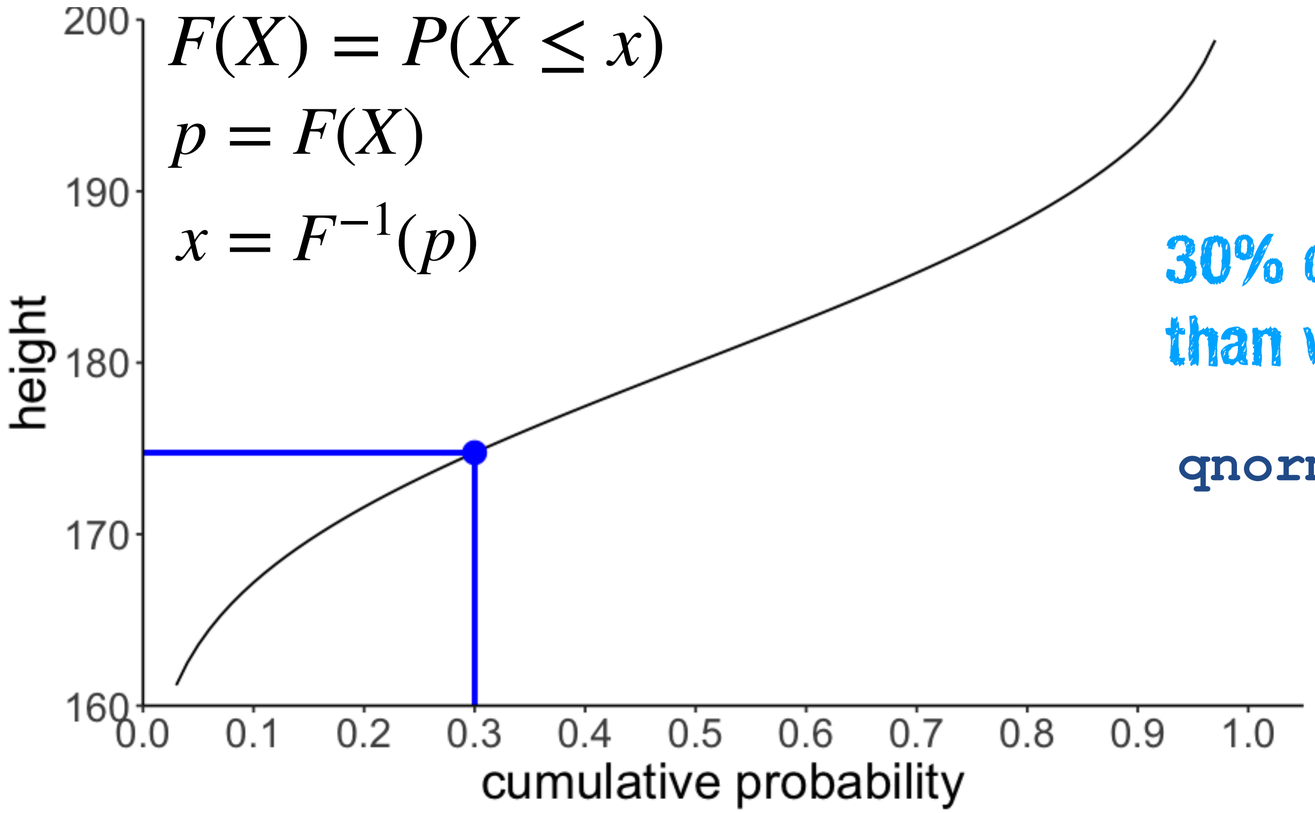
`pnorm(x)` returns the integral from $-\infty$ to `x` of the probability density function

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



30% of people are shorter than which height?

```
qnorm(0.3,
      mean = 180,
      sd = 10)
```

≈ 175

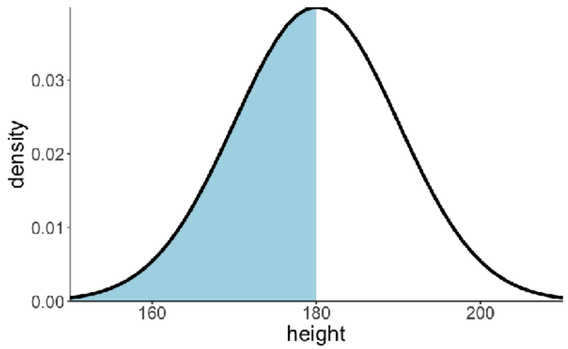
What proportion of people are between 170cm and 180cm?

Analytic solution

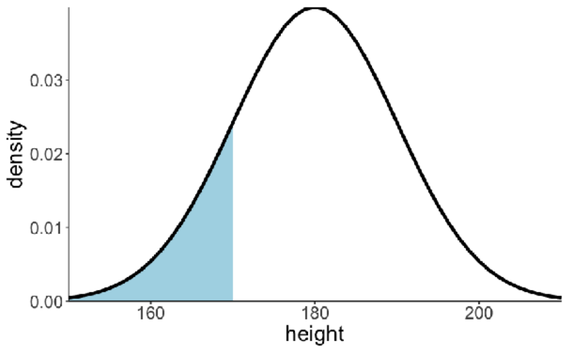
```
pnorm(180,
      mean = 180,
      sd = 10)
```

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

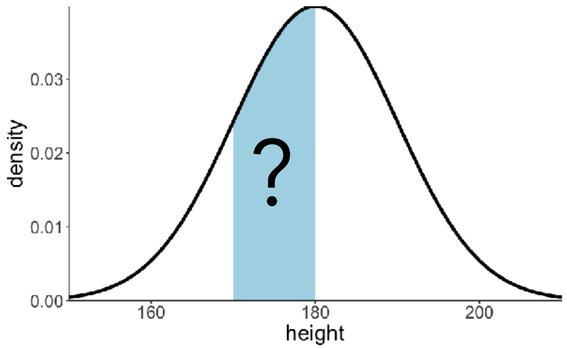
50% - 16% = 34%



smaller than 180cm



smaller than 170cm



between 170cm and 180cm

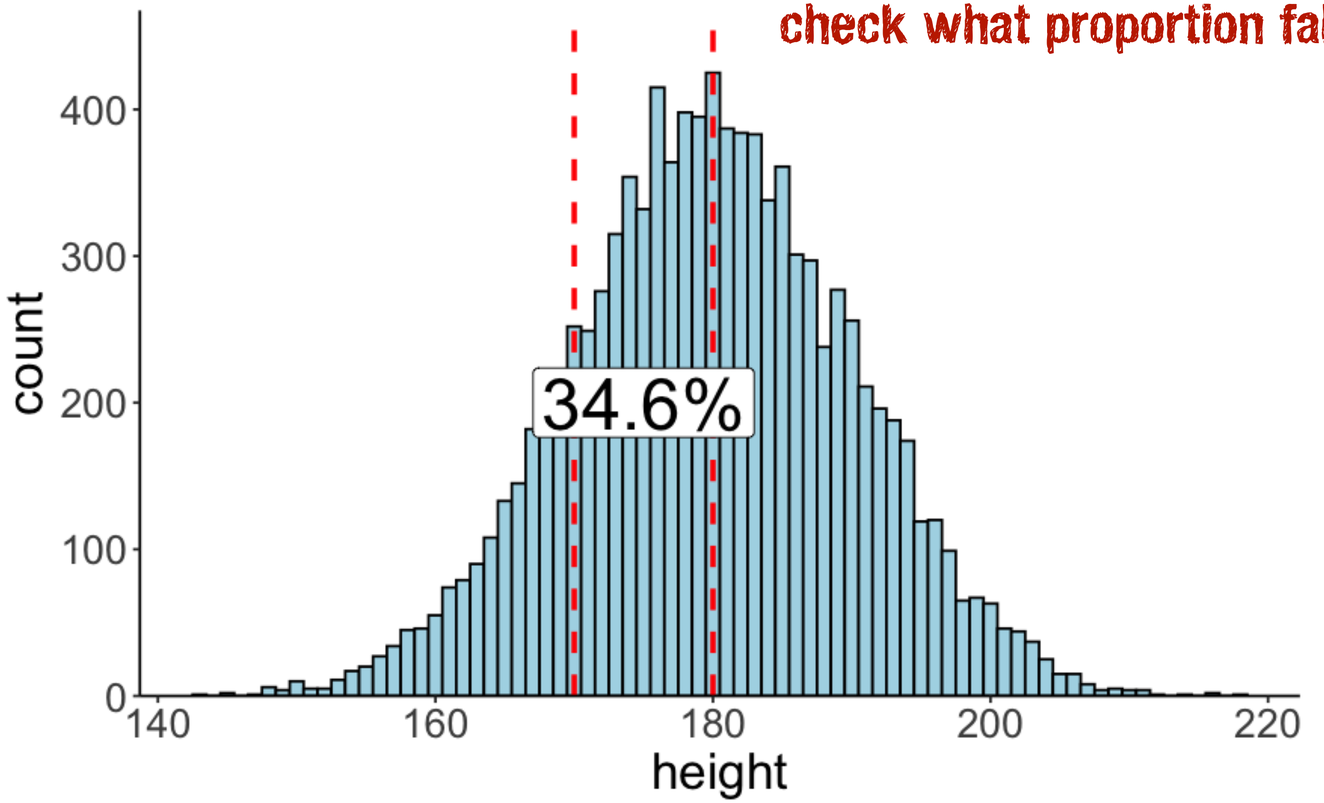
What proportion of people are between 170cm and 180cm?

Sampling solution

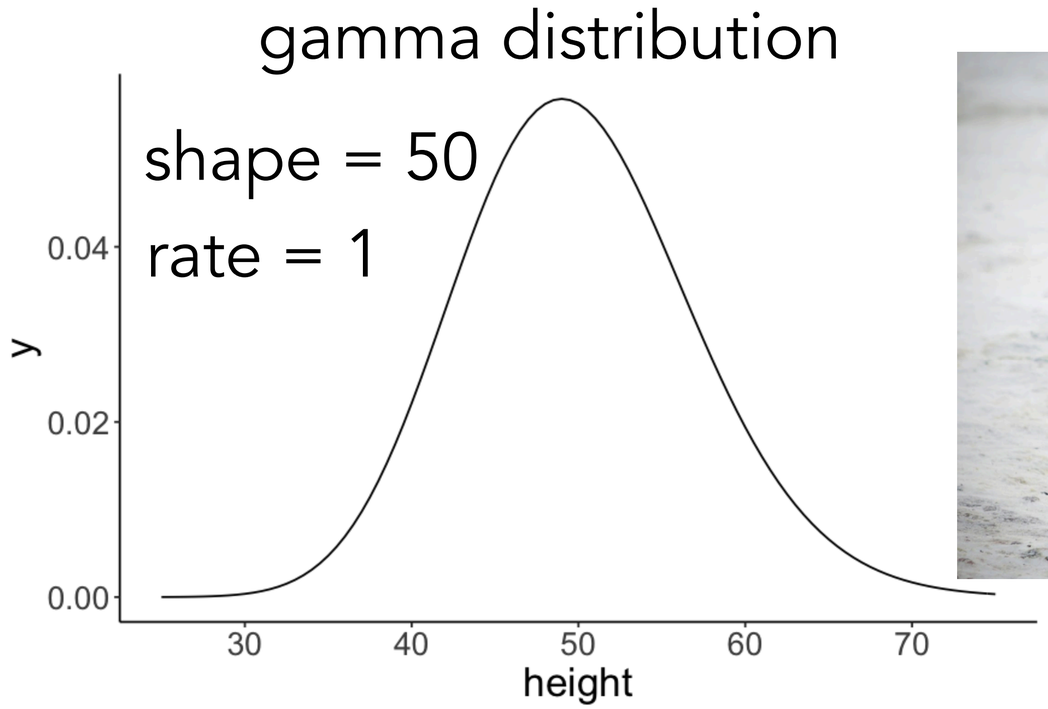
draw a bunch of samples

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

check what proportion falls in the range

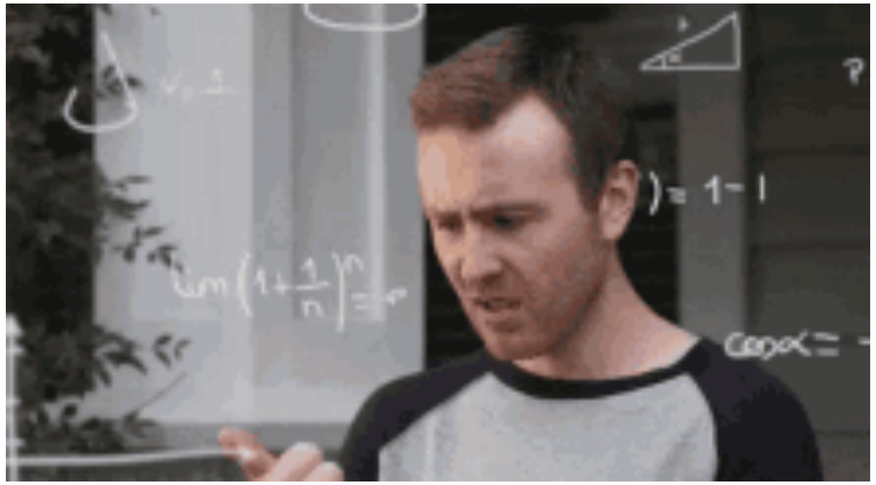


Answering questions about Penguins



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

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

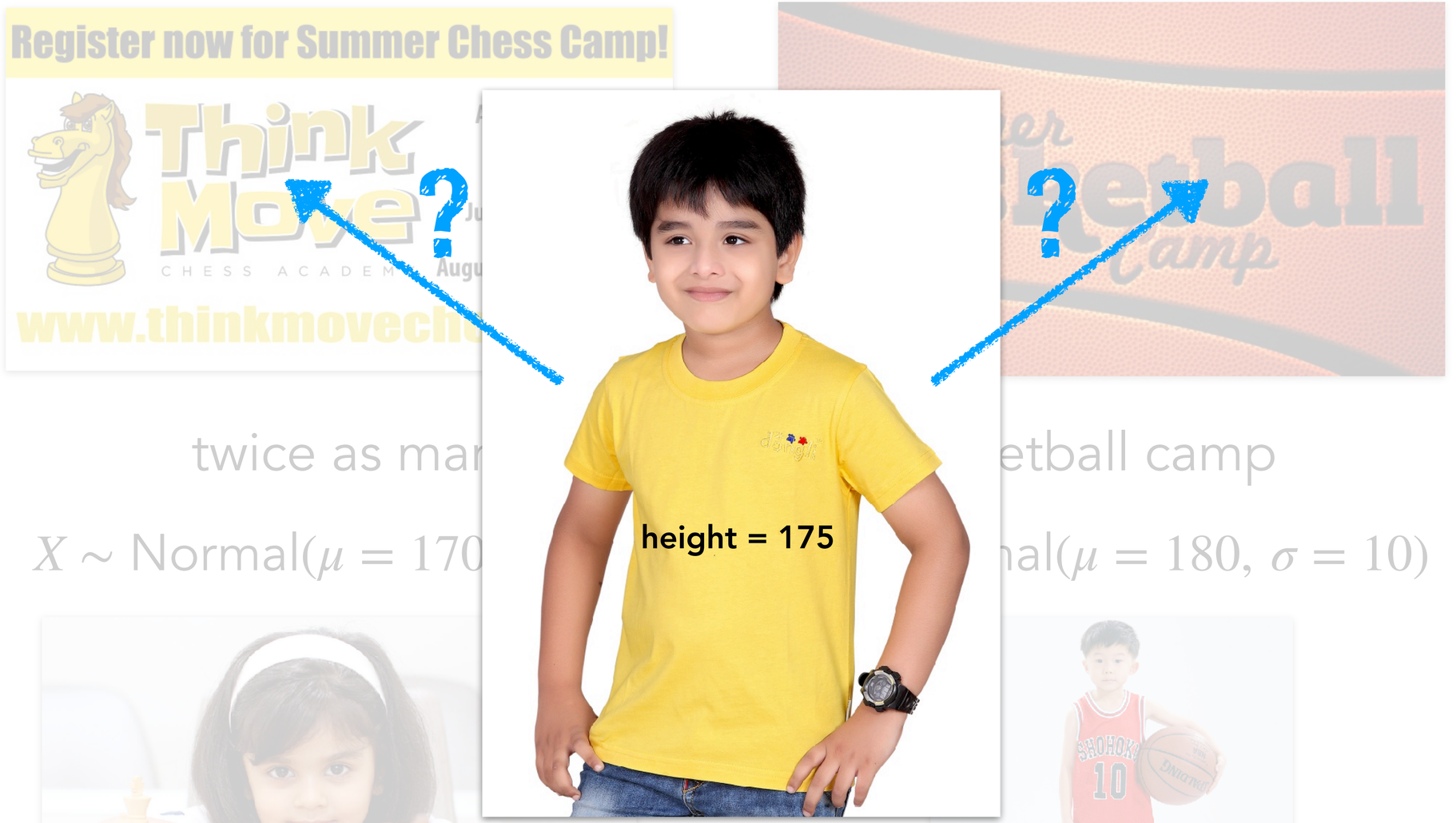
Summer camp

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



twice as mar … etball camp

$X \sim \text{Normal}(\mu = 170$ … $\text{nal}(\mu = 180, \sigma = 10)$

Analytic solution

Can you feel the Bayes?

H = {basketball, chess}

D = 175 cm

$$\underbrace{p(H \mid D)}_{\text{posterior}} = \frac{\overbrace{p(D \mid H)}^{\text{likelihood}} \cdot \overbrace{p(H)}^{\text{prior}}}{\underbrace{p(D)}_{\text{probability of the data?!}}}$$

H = Hypothesis

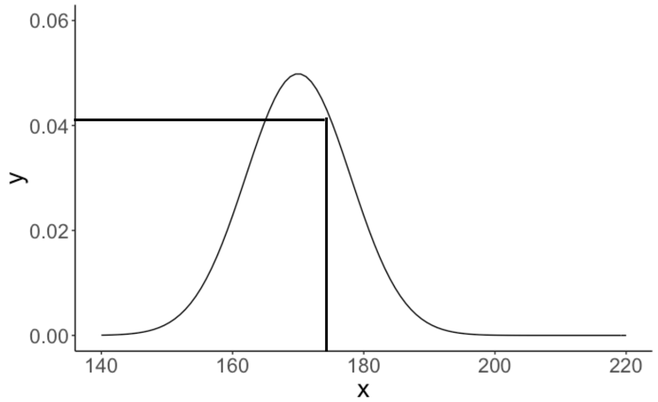
D = Data

Summer camp

prior

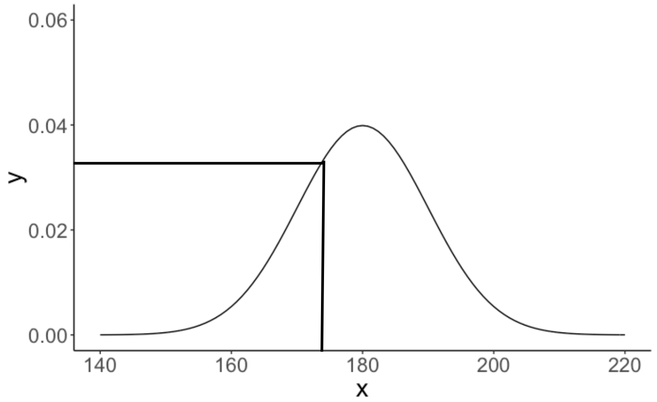
$$p(\text{chess}) = \frac{1}{3} \qquad p(\text{basketball}) = \frac{2}{3}$$

likelihood



`dnorm(175, mean = 170, sd = 8)`

= 0.041



`dnorm(175, mean = 180, sd = 10)`

= 0.035

posterior

$$p(\text{sport} = \text{basketball} \mid \text{height} = 175) = \frac{\overbrace{p(175 \mid \text{basketball})}^{\text{likelihood}} \cdot \overbrace{p(\text{basketball})}^{\text{prior}}}{\underbrace{p(175)}_{\text{data}}}$$

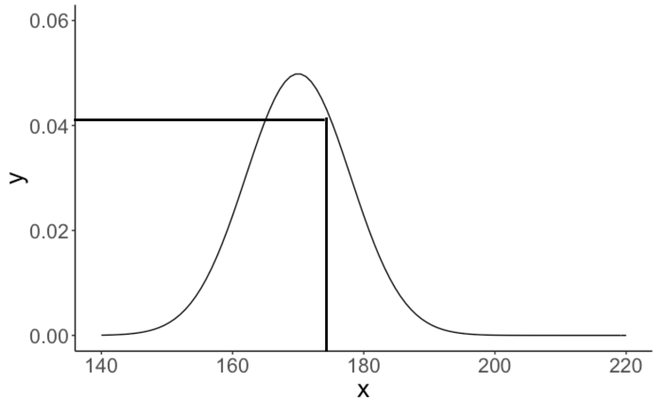
$$p(\text{basketball} \mid 175) = \frac{p(175 \mid \text{basketball}) \cdot p(\text{basketball})}{p(175 \mid \text{basketball}) \cdot p(\text{basketball}) + p(175 \mid \text{chess}) \cdot p(\text{chess})}$$

Summer camp

prior

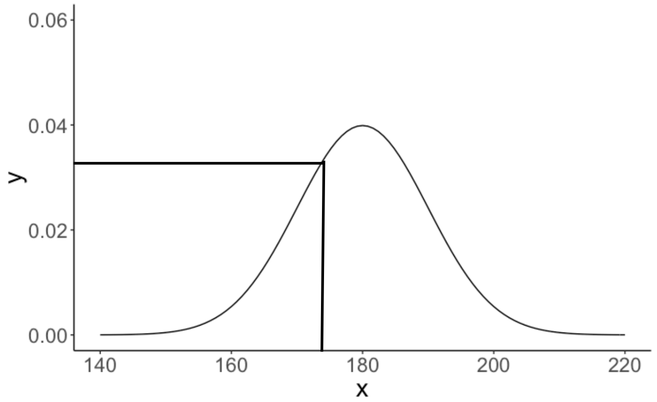
$$p(\text{chess}) = \frac{1}{3} \qquad p(\text{basketball}) = \frac{2}{3}$$

likelihood



`dnorm(175, mean = 170, sd = 8)`

$= 0.041$



`dnorm(175, mean = 180, sd = 10)`

$= 0.035$

posterior

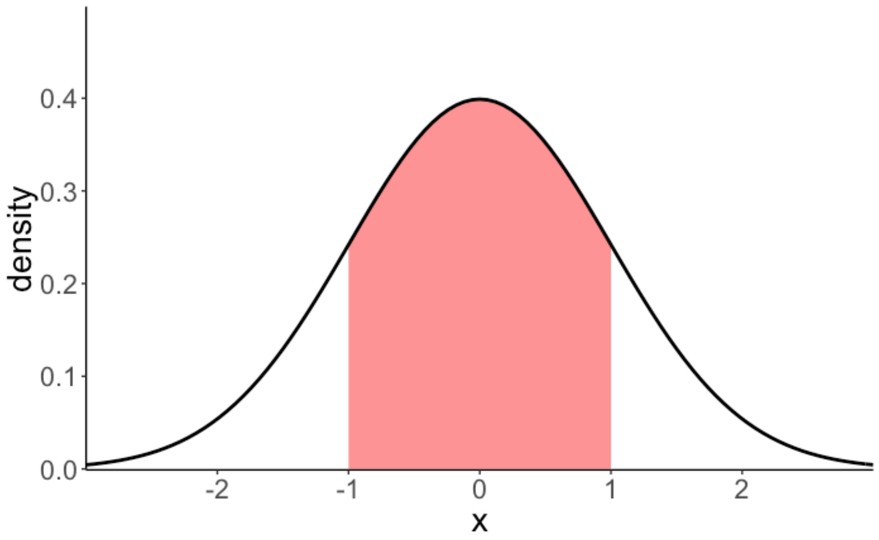
$$p(\text{basketball} \mid 175) = \frac{p(175 \mid \text{basketball}) \cdot p(\text{basketball})}{p(175 \mid \text{basketball}) \cdot p(\text{basketball}) + p(175 \mid \text{chess}) \cdot p(\text{chess})}$$

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

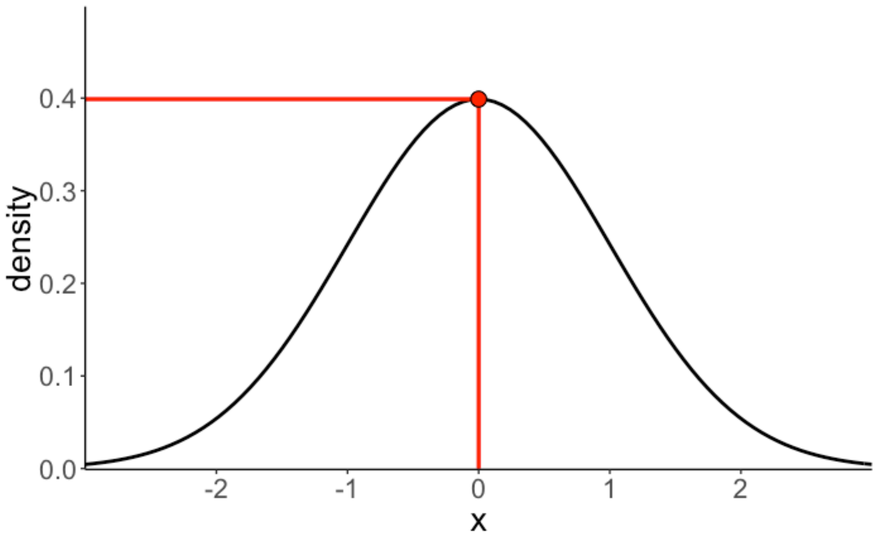
send the kid to the basketball gym!

Probability vs. likelihood

Probability

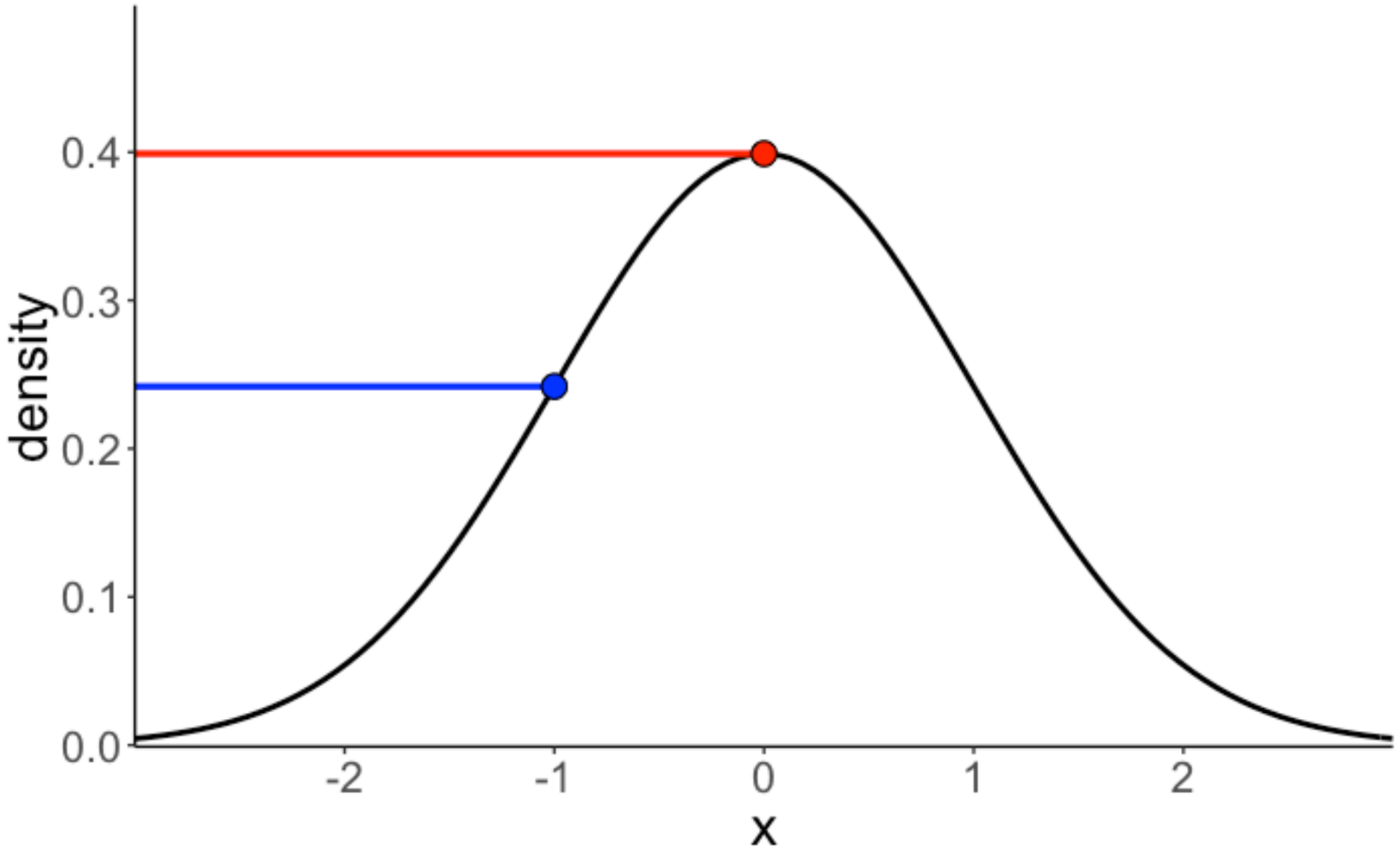


Likelihood



Probability vs. likelihood

```
dnorm(0)/dnorm(-1) = 1.6487
```

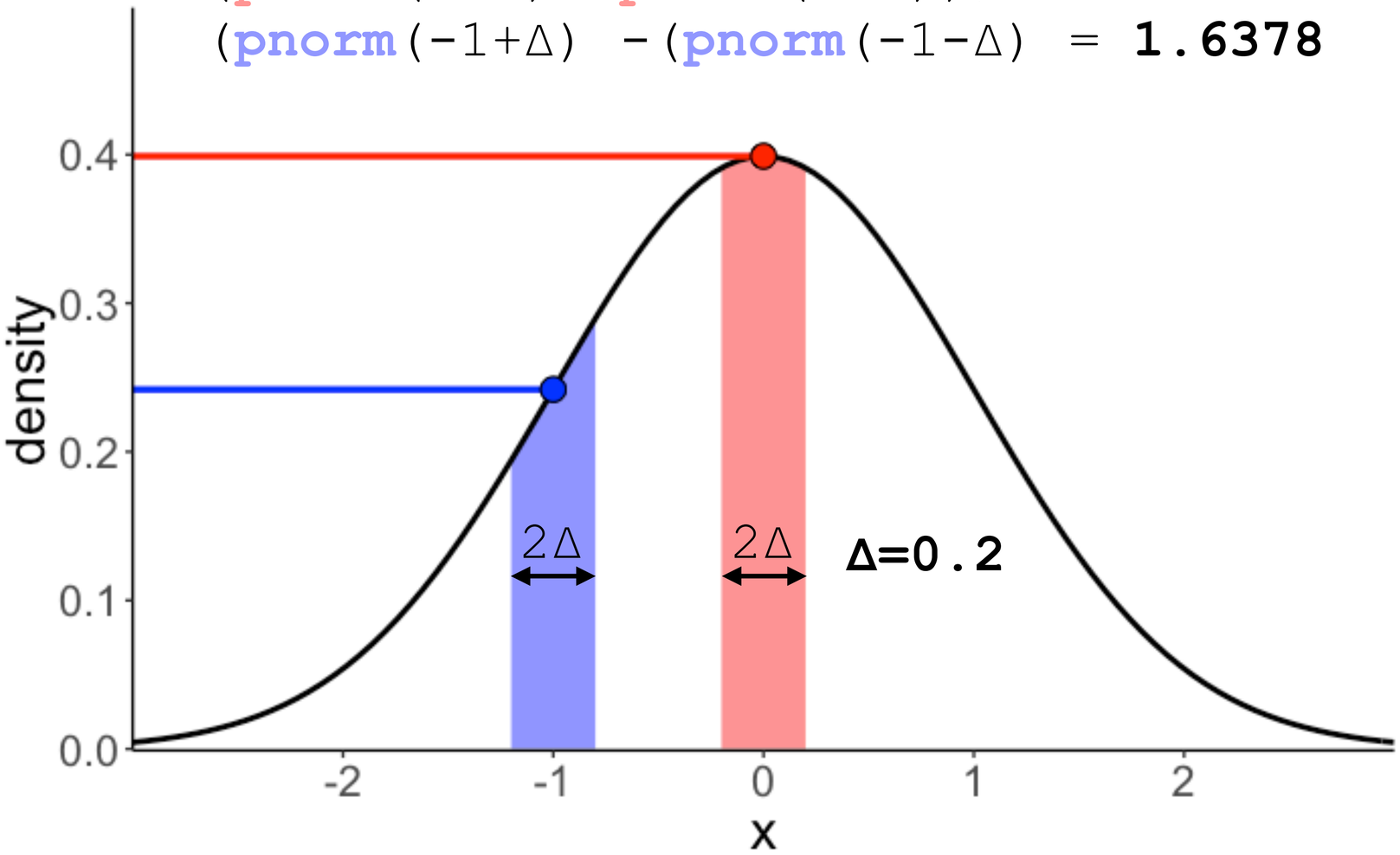


relative probability of one value vs. another

Probability vs. likelihood

```
dnorm(0)/dnorm(-1) = 1.6487
```

```
(pnorm(0+Δ)- pnorm(0-Δ))/
(pnorm(-1+Δ) -(pnorm(-1-Δ) = 1.6378
```

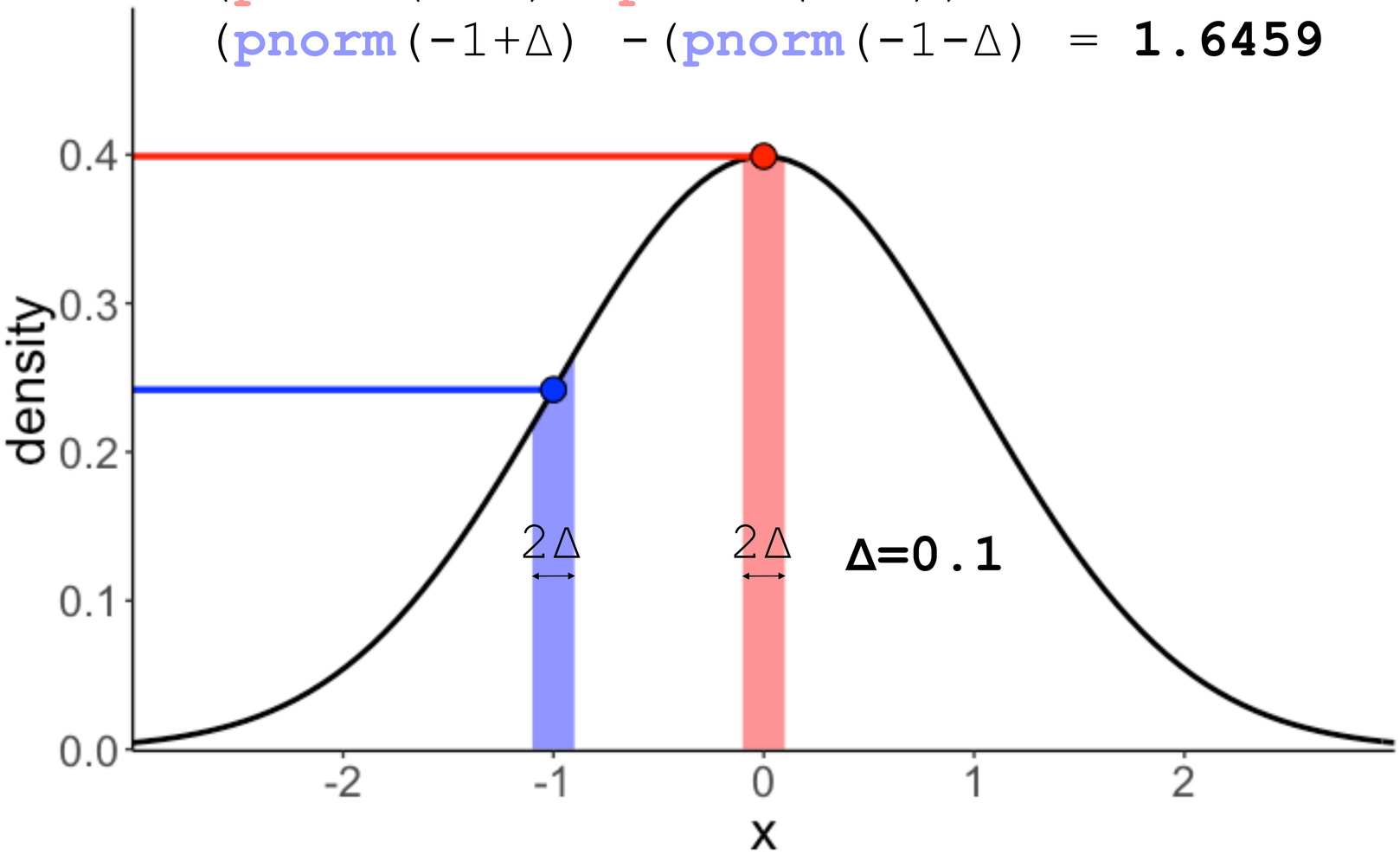


relative probability of one value vs. another

Probability vs. likelihood

```
dnorm(0)/dnorm(-1) = 1.6487
```

```
(pnorm(0+Δ)- pnorm(0-Δ))/
(pnorm(-1+Δ) -(pnorm(-1-Δ) = 1.6459
```

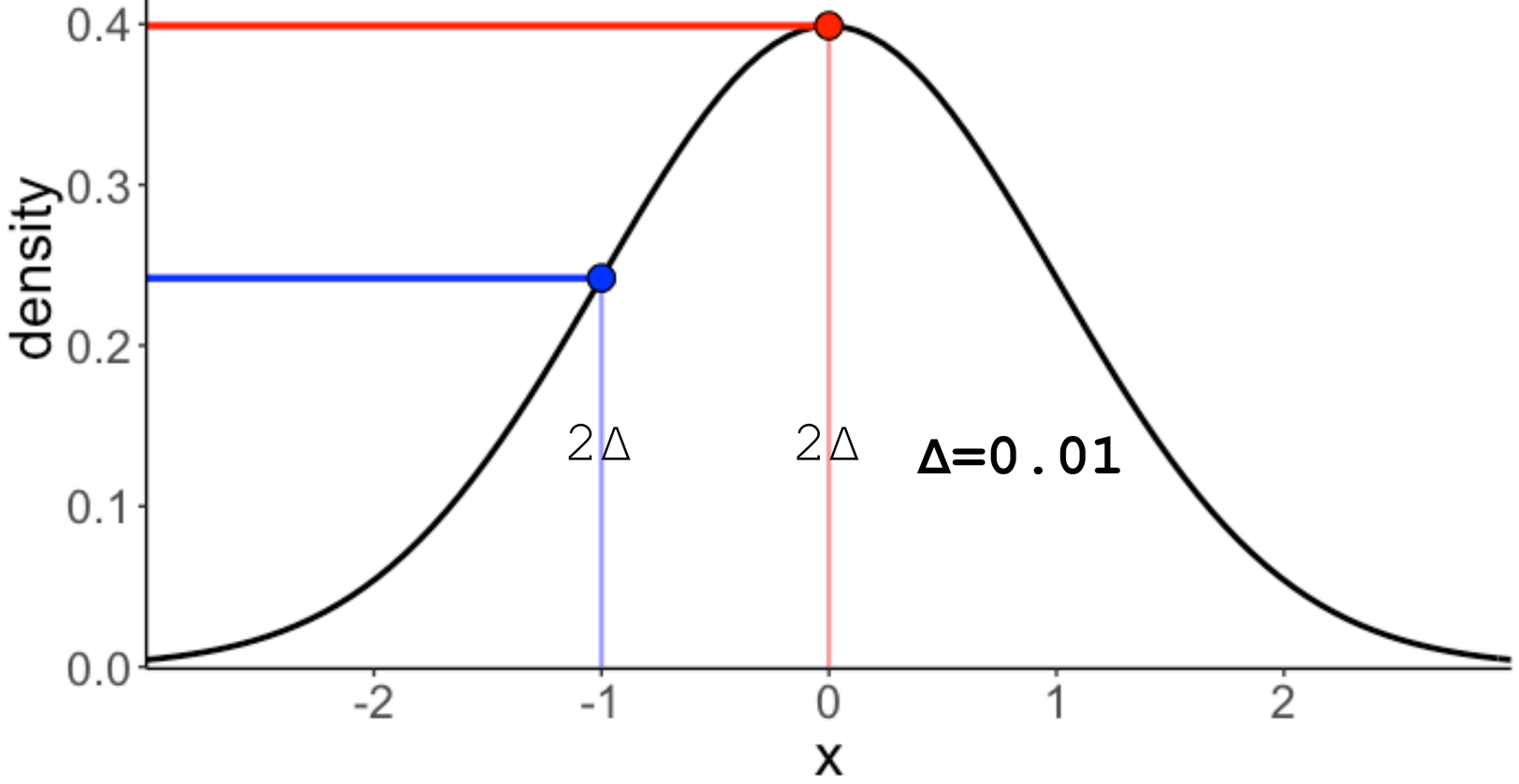


relative probability of one value vs. another

Probability vs. likelihood

```
dnorm(0)/dnorm(-1) = 1.6487
```

```
(pnorm(0+Δ)- pnorm(0-Δ))/
(pnorm(-1+Δ) -(pnorm(-1-Δ) = 1.6486
```



relative probability of one value vs. another

Sampling solution

Summer camp: Via sampling

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

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

doesn't work!

Summer camp: Via sampling

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

```
df.camp %>%
  filter(between(height,
                 left = 174,
                 right = 176)) %>%
  count(sport)
```

this works!

sport	n
basketball	469
chess	273

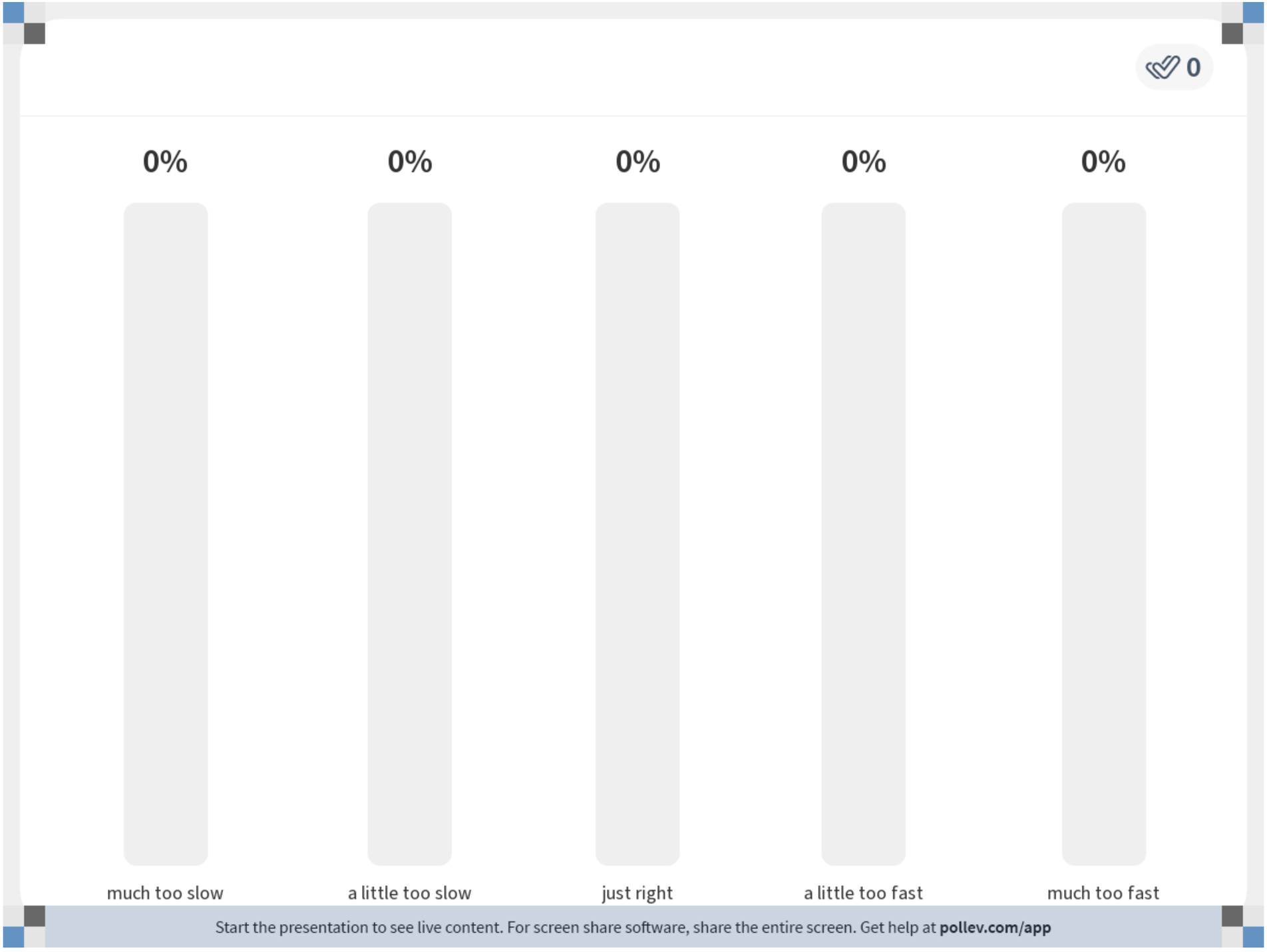
$$\frac{\text{basketball}}{\text{basketball} + \text{chess}} \approx 0.63$$

Plan for today

- Quick recap
- 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

1


0
0%
0%
0%
0%
0%
much too slow
a little too slow
just right
a little too fast
much too fast
Start the presentation to see live content. For screen share software, share the entire screen. Get help at pollev.com/app

0

Nobody has responded yet.

Hang tight! Responses are coming in.

Thank you!