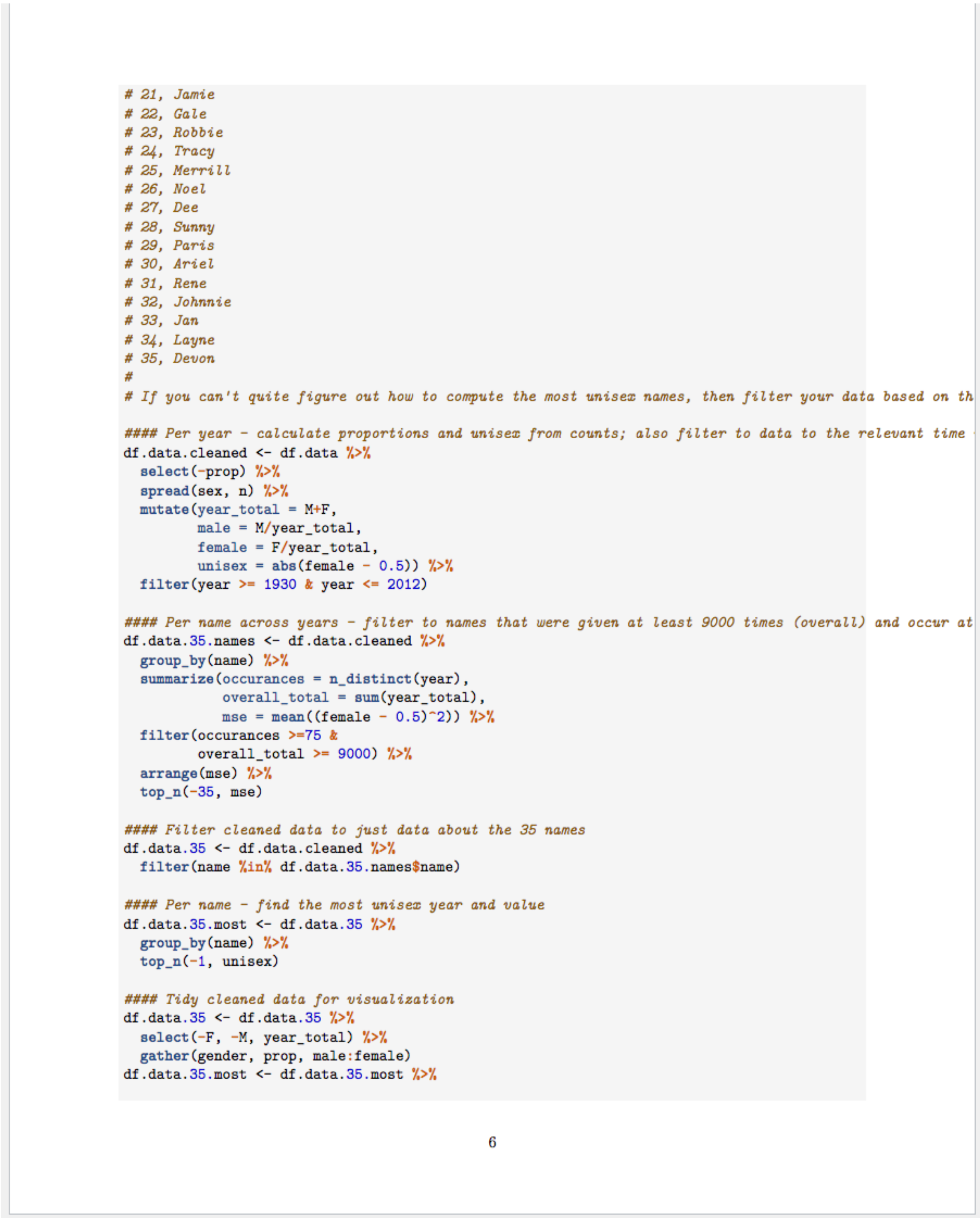
Modeling data

01/28/2019

Logistics

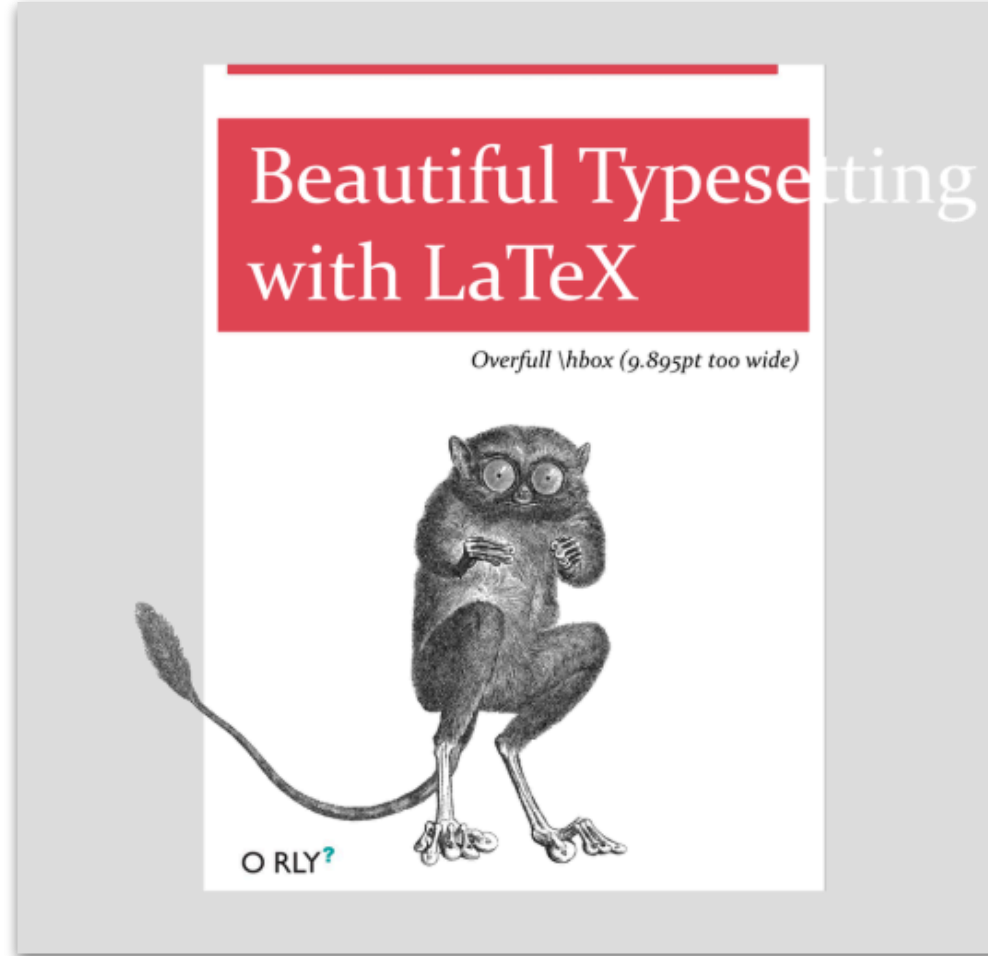
Homework 2

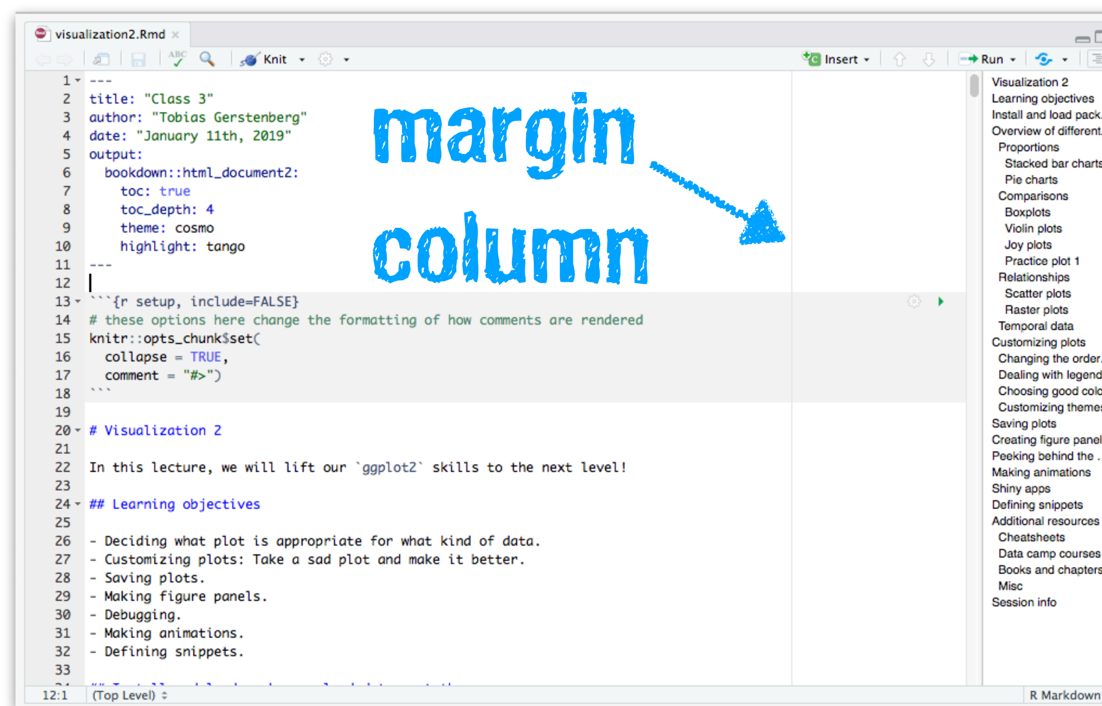
Homework 2



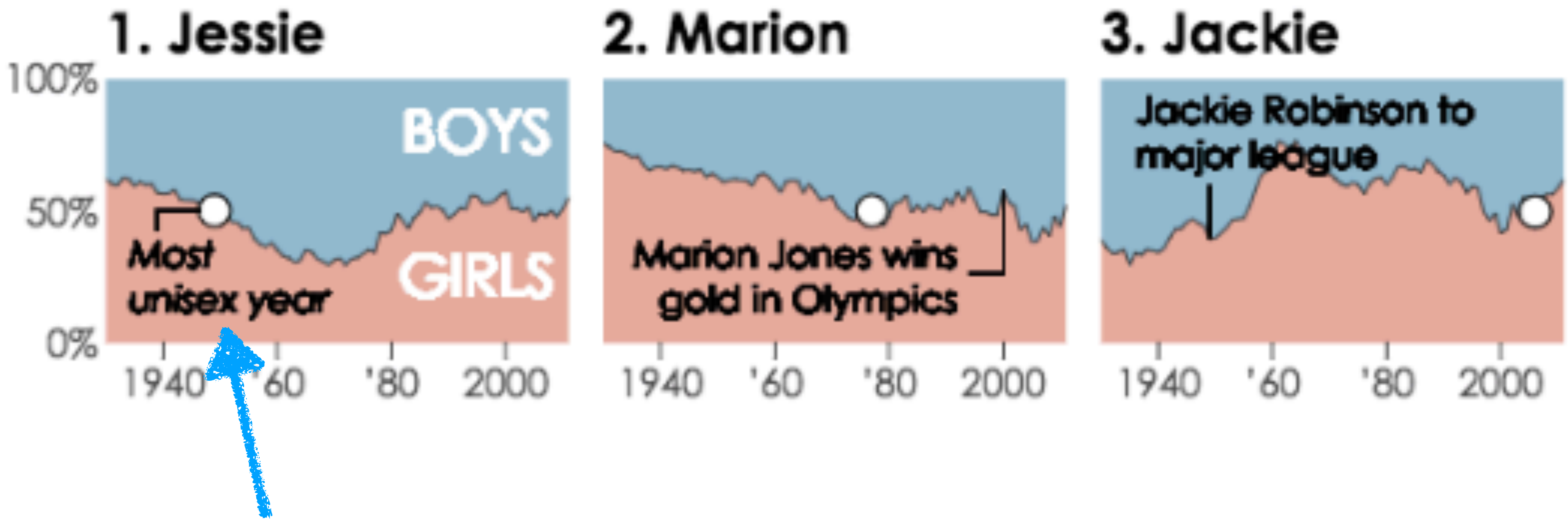
```
# 21, Jamie
# 22, Gale
# 23, Robbie
# 24, Tracy
# 25, Merrill
# 26, Noel
# 27, Dee
# 28, Sunny
# 29, Paris
# 30, Ariel
# 31, Rene
# 32, Johnnie
# 33, Jan
# 34, Layne
# 35, Devon
#
# If you can't quite figure out how to compute the most unisex names, then filter your data based on th

#### Per year - calculate proportions and unisex from counts; also filter to data to the relevant time
df.data.cleaned <- df.data %>%
  select(-prop) %>%
  spread(sex, n) %>%
  mutate(year_total = M+F,
         male = M/year_total,
         female = F/year_total,
         unisex = abs(female - 0.5)) %>%
  filter(year >= 1930 & year <= 2012)

#### Per name across years - filter to names that were given at least 9000 times (overall) and occur at
df.data.35.names <- df.data.cleaned %>%
  group_by(name) %>%
  summarize(occurances = n_distinct(year),
            overall_total = sum(year_total),
            mse = mean((female - 0.5)^2)) %>%
  filter(occurances >=75 &
         overall_total >= 9000) %>%
  arrange(mse) %>%
  top_n(-35, mse)

#### Filter cleaned data to just data about the 35 names
df.data.35 <- df.data.cleaned %>%
  filter(name %in% df.data.35.names$name)

#### Per name - find the most unisex year and value
df.data.35.most <- df.data.35 %>%
  group_by(name) %>%
  top_n(-1, unisex)

#### Tidy cleaned data for visualization
df.data.35 <- df.data.35 %>%
  select(-F, -M, year_total) %>%
  gather(gender, prop, male:female)
df.data.35.most <- df.data.35.most %>%
```

6





Homework 2



if text looks pixelated, it's likely that there are many layers of text on top of each other

`geom_text()` needs a separate data frame with one entry per facet

Homework 2

I learned something new!

```r
data = c(1, 3, 4, 2, 5)
prediction = c(1, 2, 2, 1, 4)

# calculate root mean squared error the pipe way
rmse = (prediction-data)^2 %>%
  mean() %>%
  sqrt() %>%
  print()
```

can we pipe this even more?

```r
rmse = prediction %>%
  subtract(data) %>%
  raise_to_power(2) %>%
  mean() %>%
  sqrt() %>%
  print()
```

Homework 2

I learned something new!

library("magrittr")

extract	\`[\`
extract2	\`[[\`
inset	\`[<-\`
inset2	\`[[<-\`
use_series	\`$\`
add	\`+\`
subtract	\`-\`
multiply_by	\`*\`
raise_to_power	\`^\`
multiply_by_matrix	\`%*%\`
divide_by	\`/\`
divide_by_int	\`%/%\`
mod	\`%%\`
is_in	\`%in%\`
and	\`&\`
or	\`|\`
equals	\`==\`
is_greater_than	\`>\`
is_weakly_greater_than	\`>=\`
is_less_than	\`<\`
is_weakly_less_than	\`<=\`
not (\`n'est pas\`)	\`!\`

Your feedback

Your feedback

Central limit theorem was a little bit confusing……

Good explanation. I haven’t really understood the CLT before

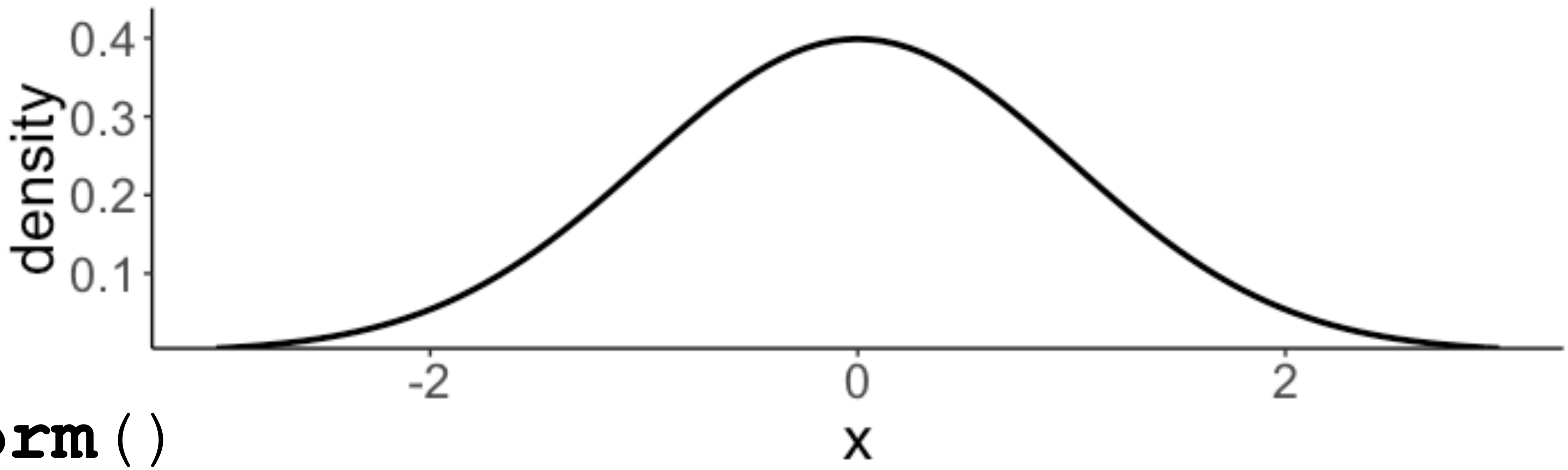
I think you spent too long on CLT which isn’t intuitively difficult and would like more time on the harder topics near the end

sometimes it's tricky to get it right

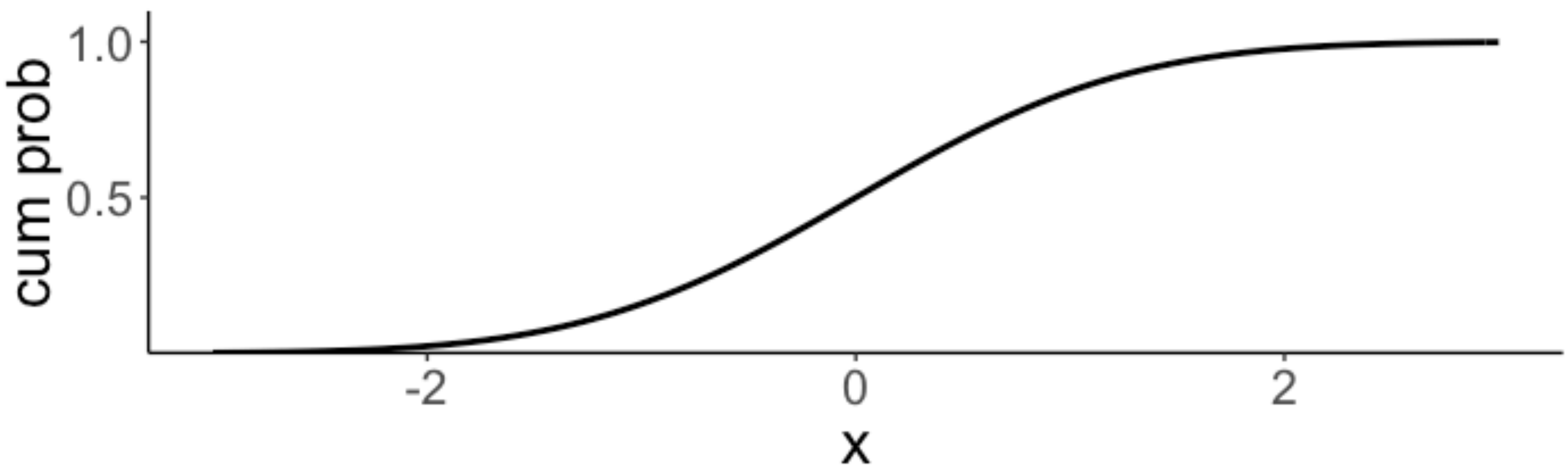
Probability vs. likelihood

Probability vs. likelihood

`dnorm()`

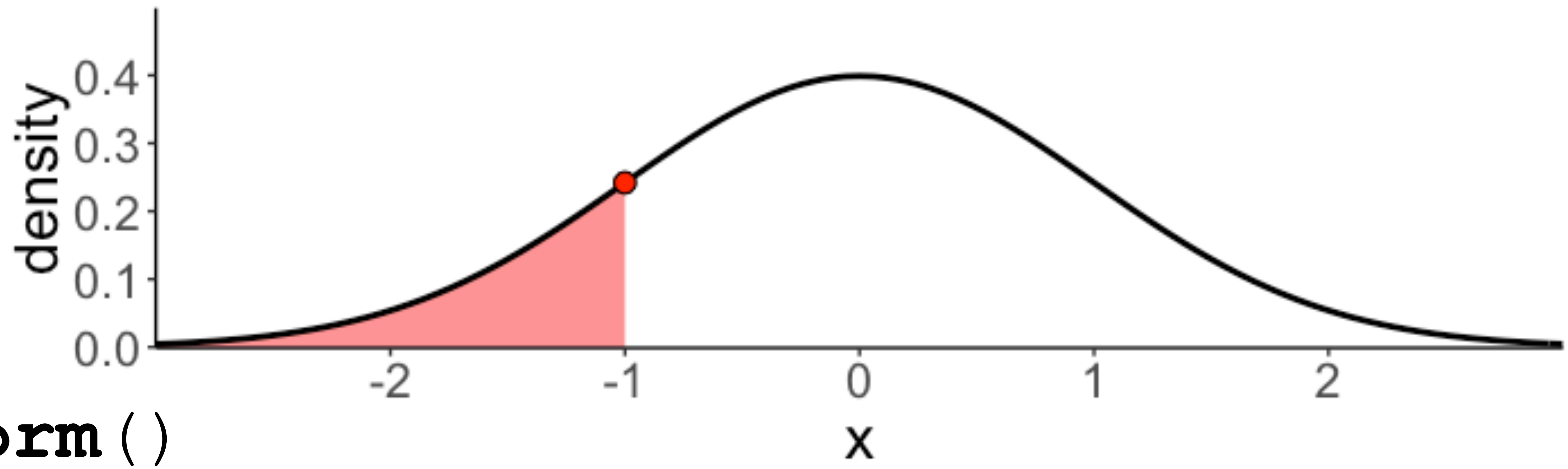


`pnorm()`

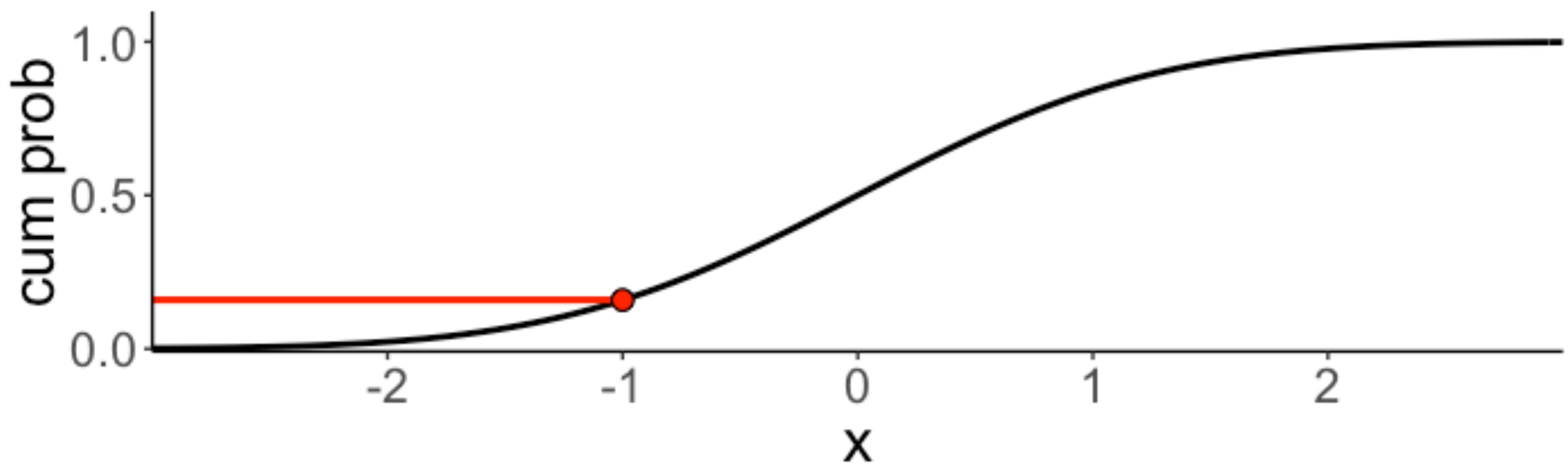


Probability vs. likelihood

`dnorm()`

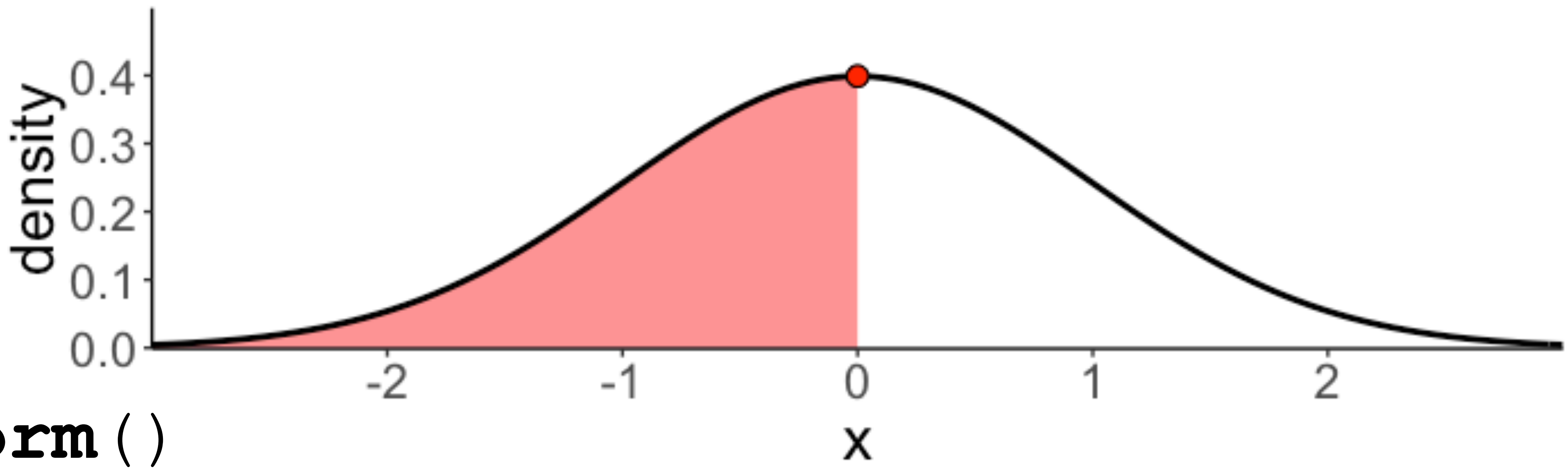


`pnorm()`

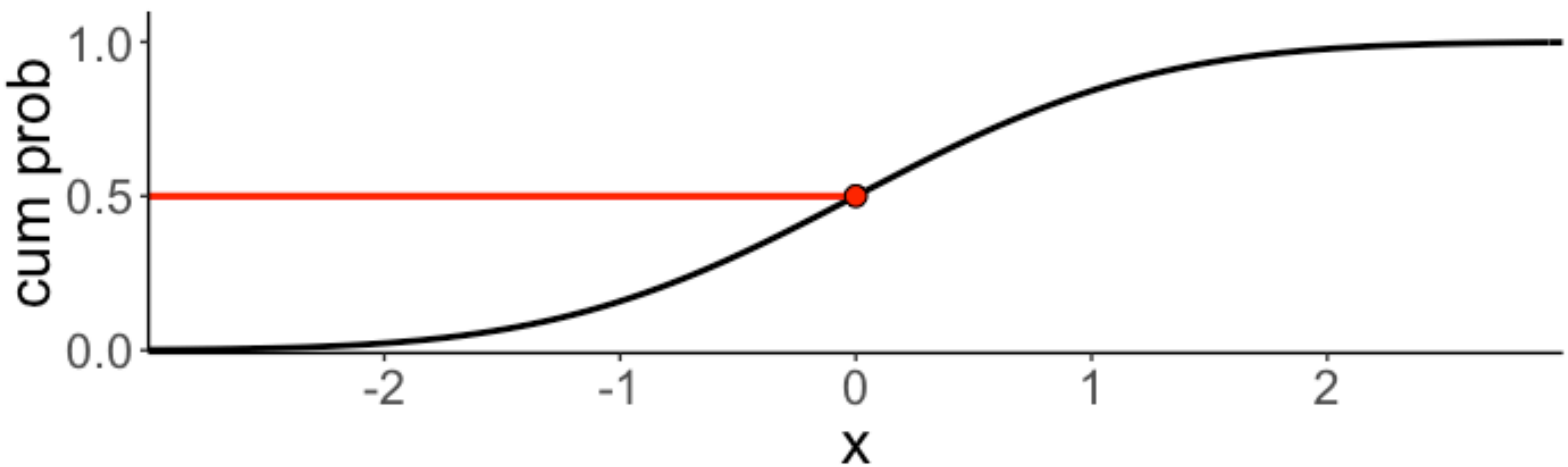


Probability vs. likelihood

`dnorm()`

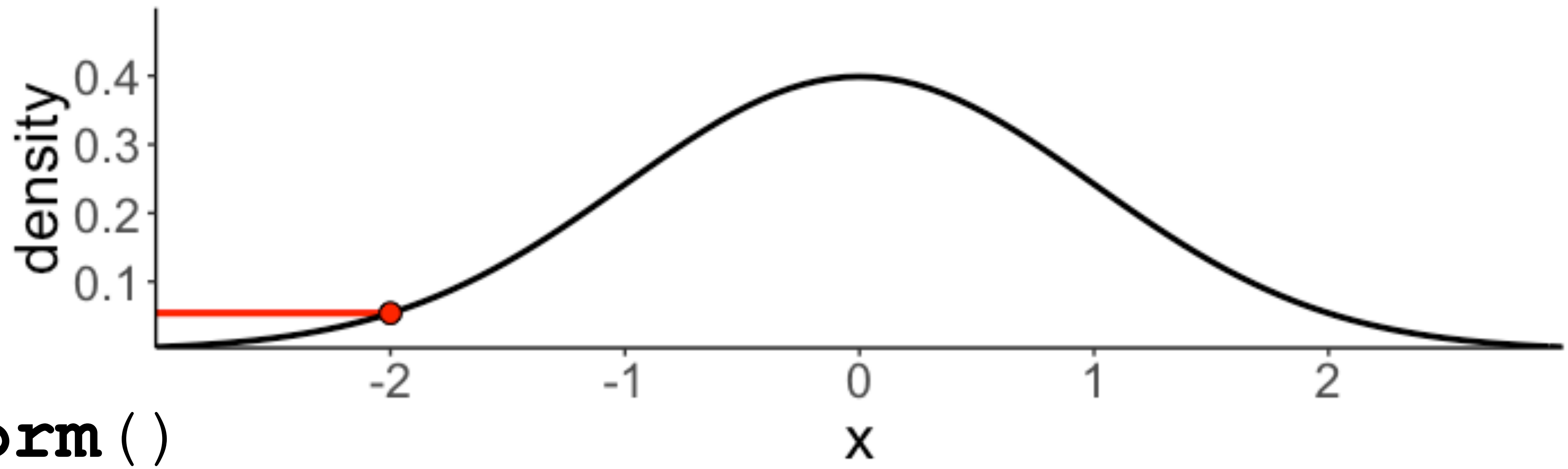


`pnorm()`

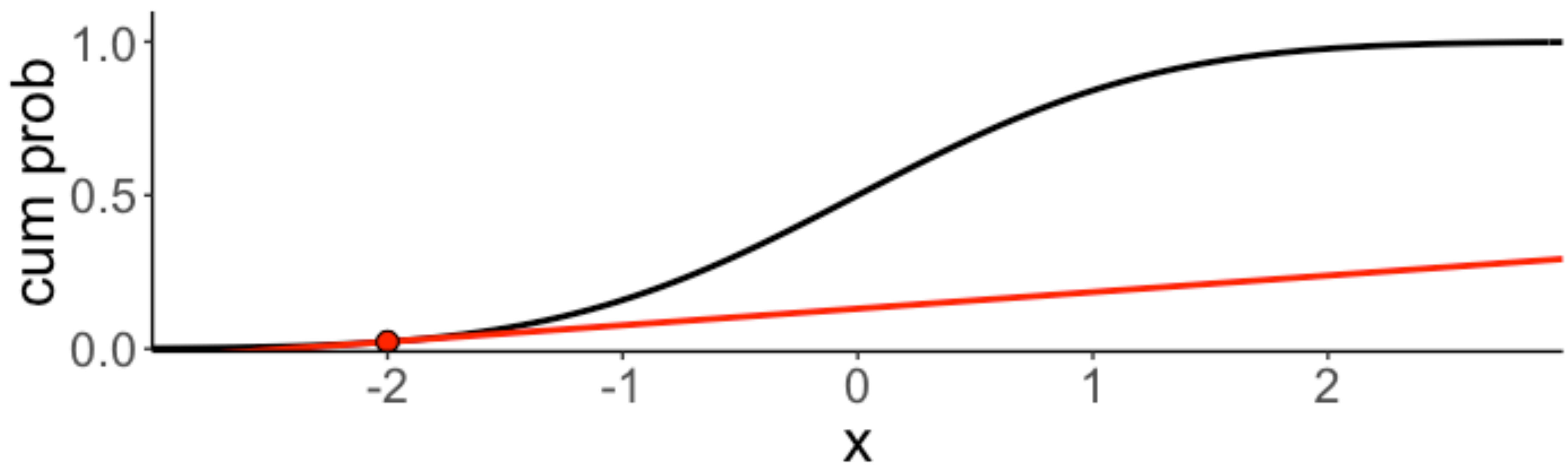


Probability vs. likelihood

dnorm()



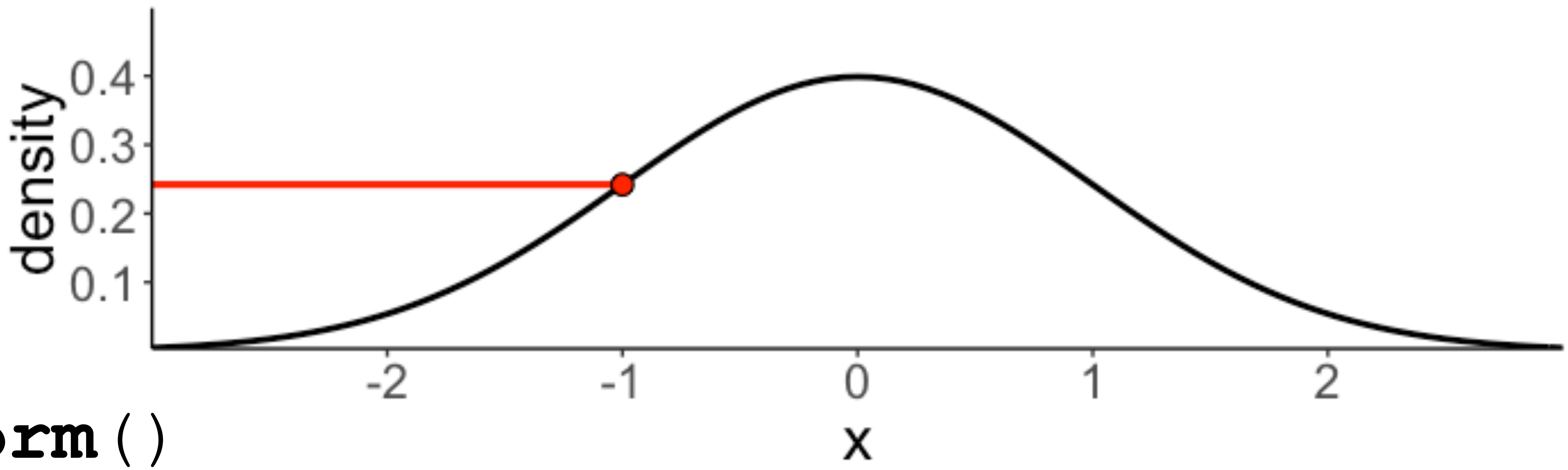
pnorm()



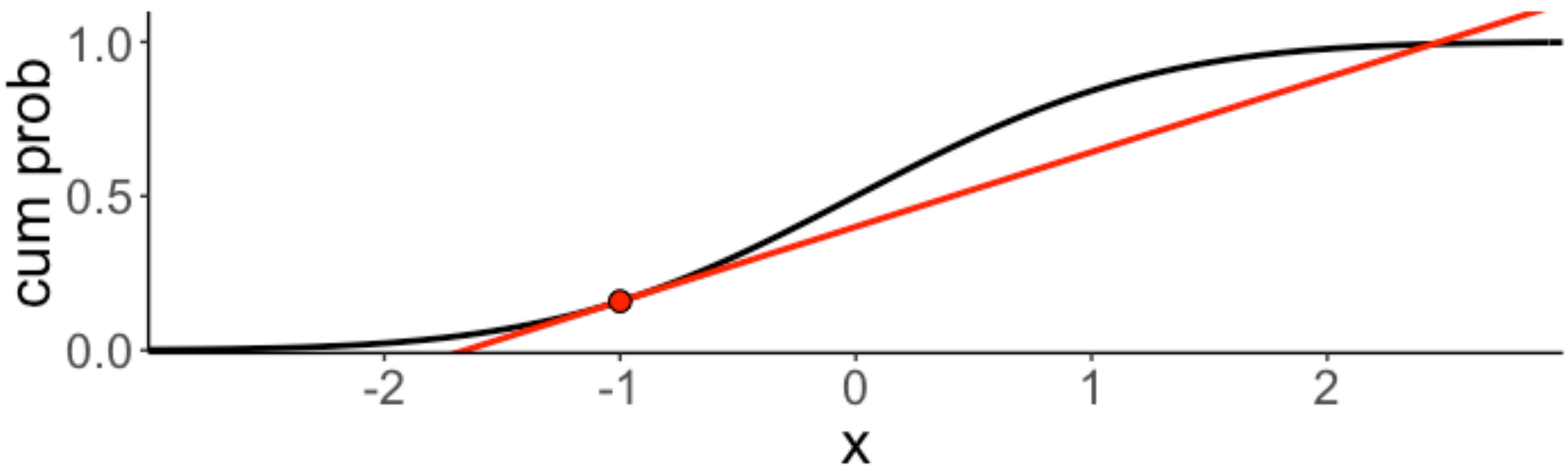
dnorm() is the first derivative of **pnorm()**

Probability vs. likelihood

dnorm()



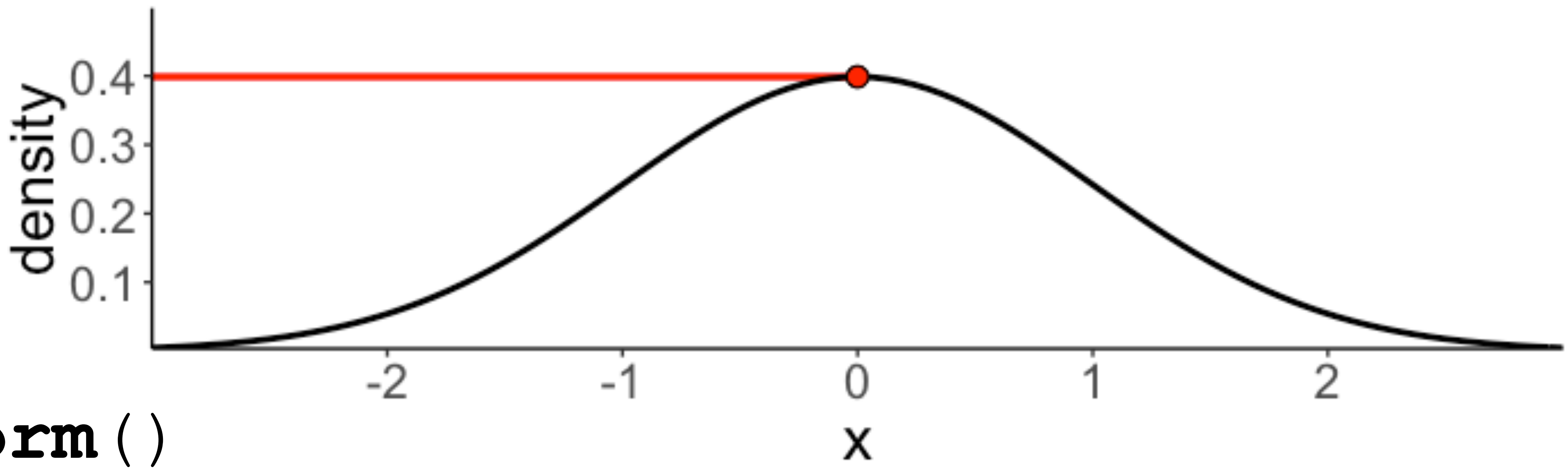
pnorm()



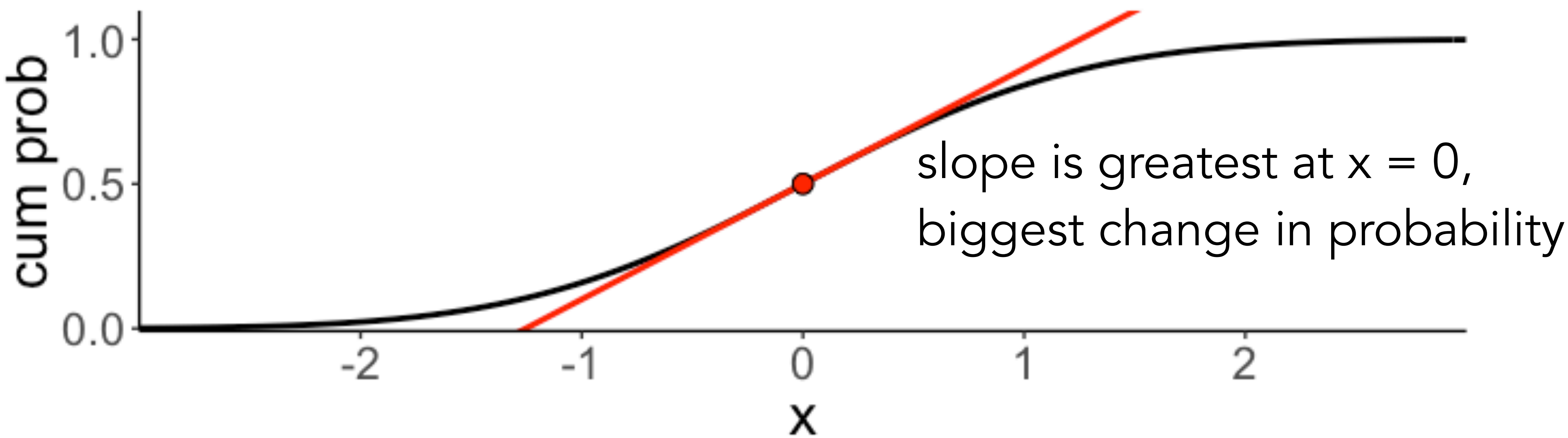
dnorm() is the first derivative of **pnorm**()

Probability vs. likelihood

dnorm()



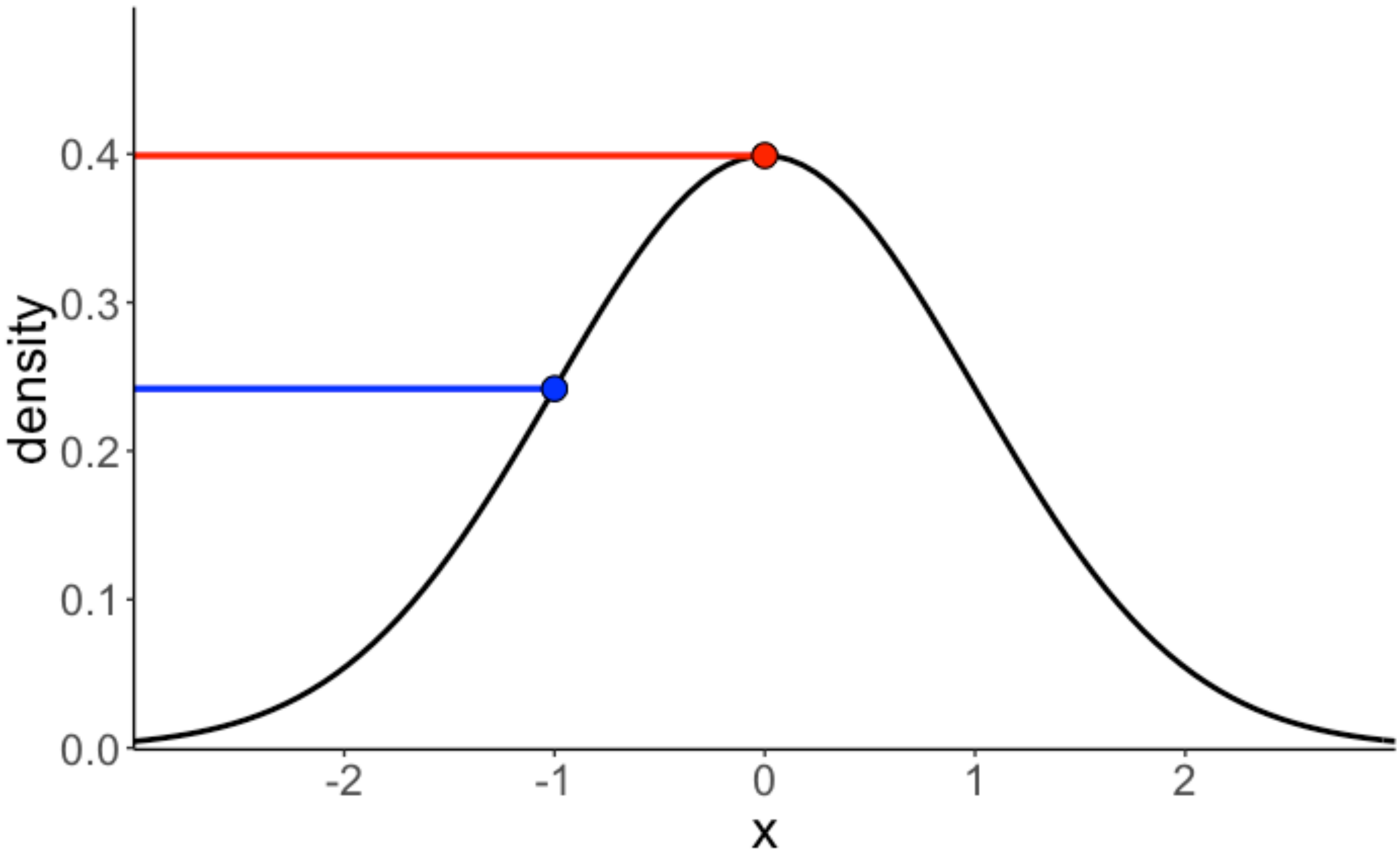
pnorm()



dnorm() is the first derivative of **pnorm**()

Probability vs. likelihood

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

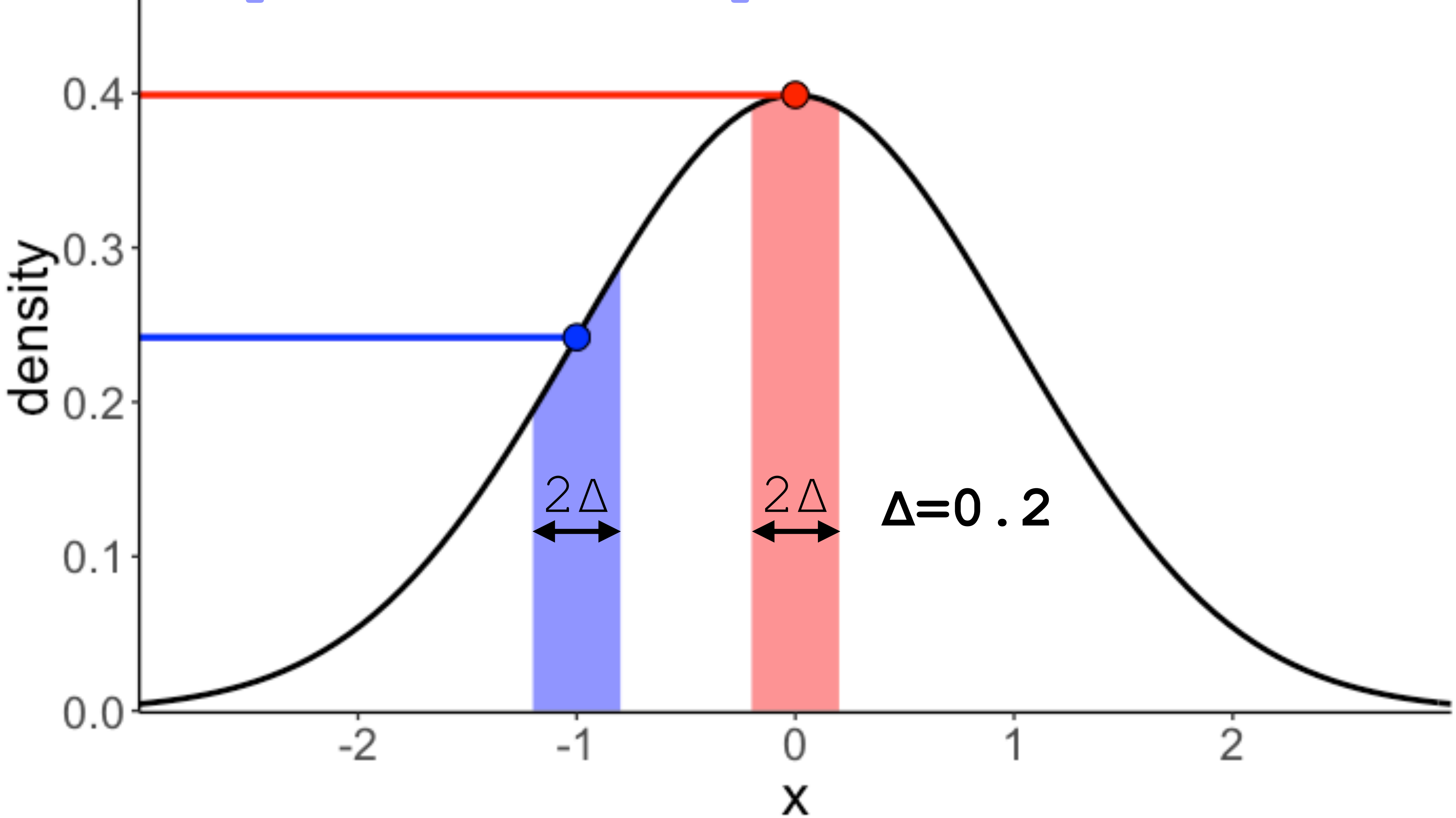


relative probability of one value vs. another

Probability vs. likelihood

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

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

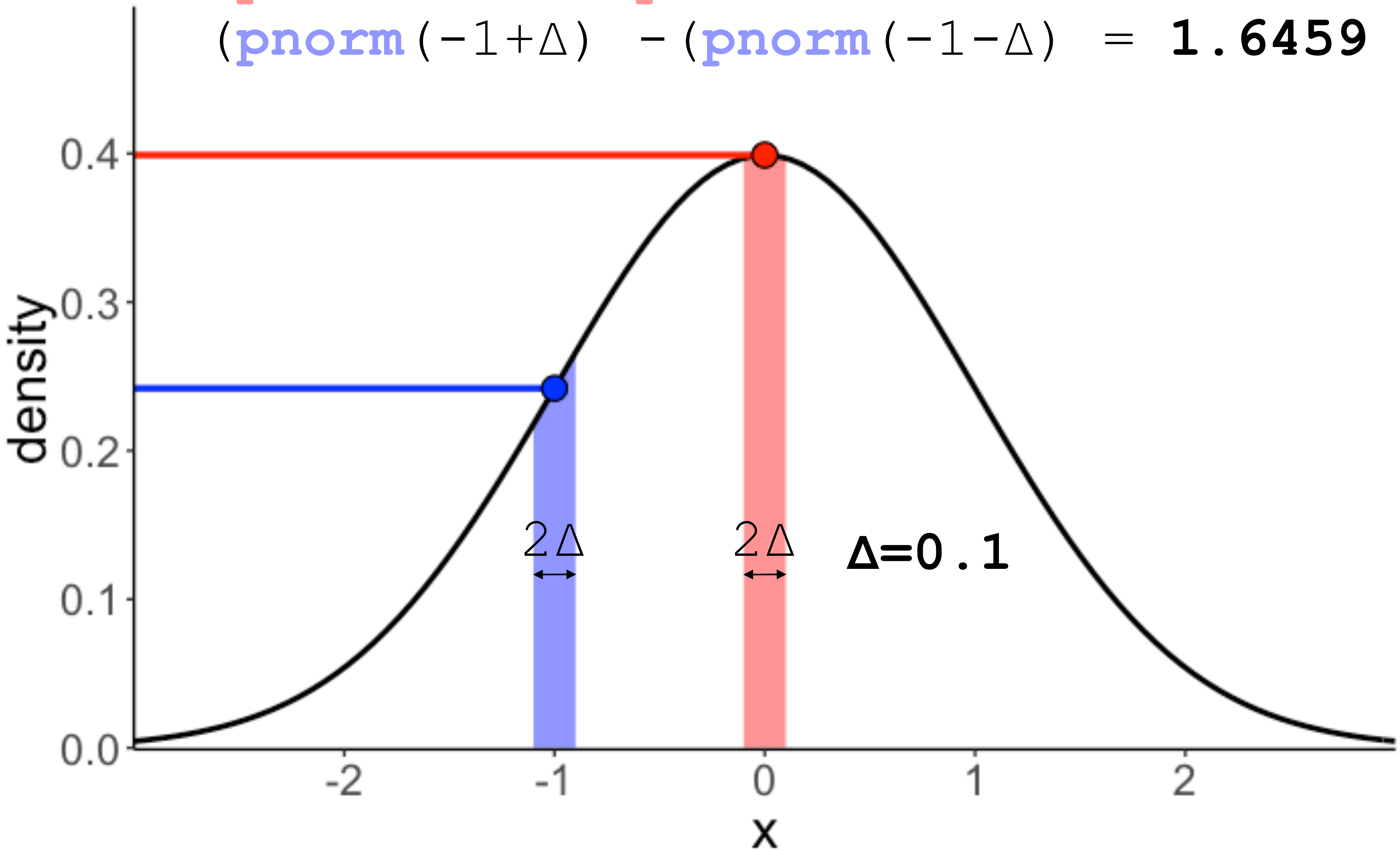


relative probability of one value vs. another

Probability vs. likelihood

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

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

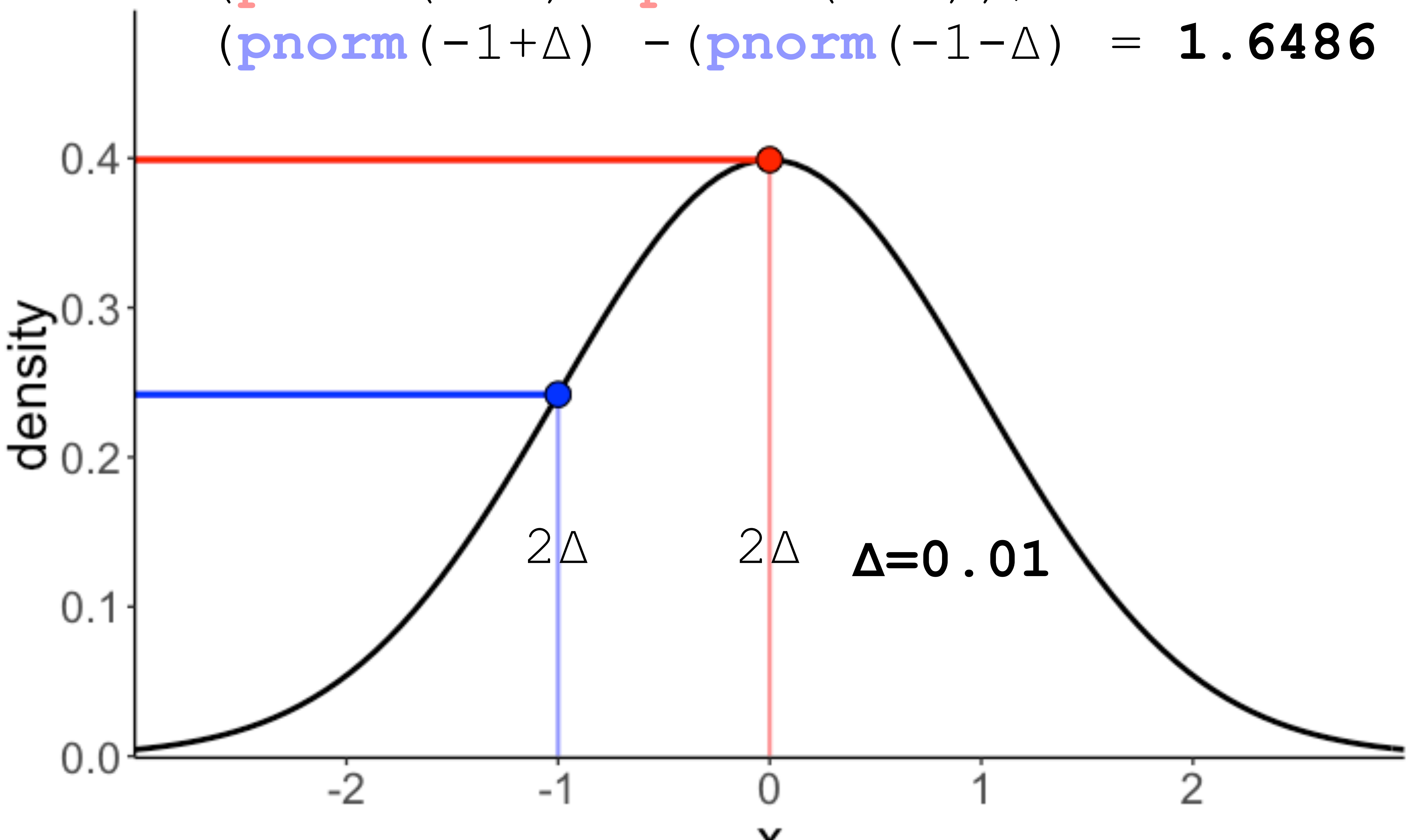


relative probability of one value vs. another

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

Probability vs. likelihood

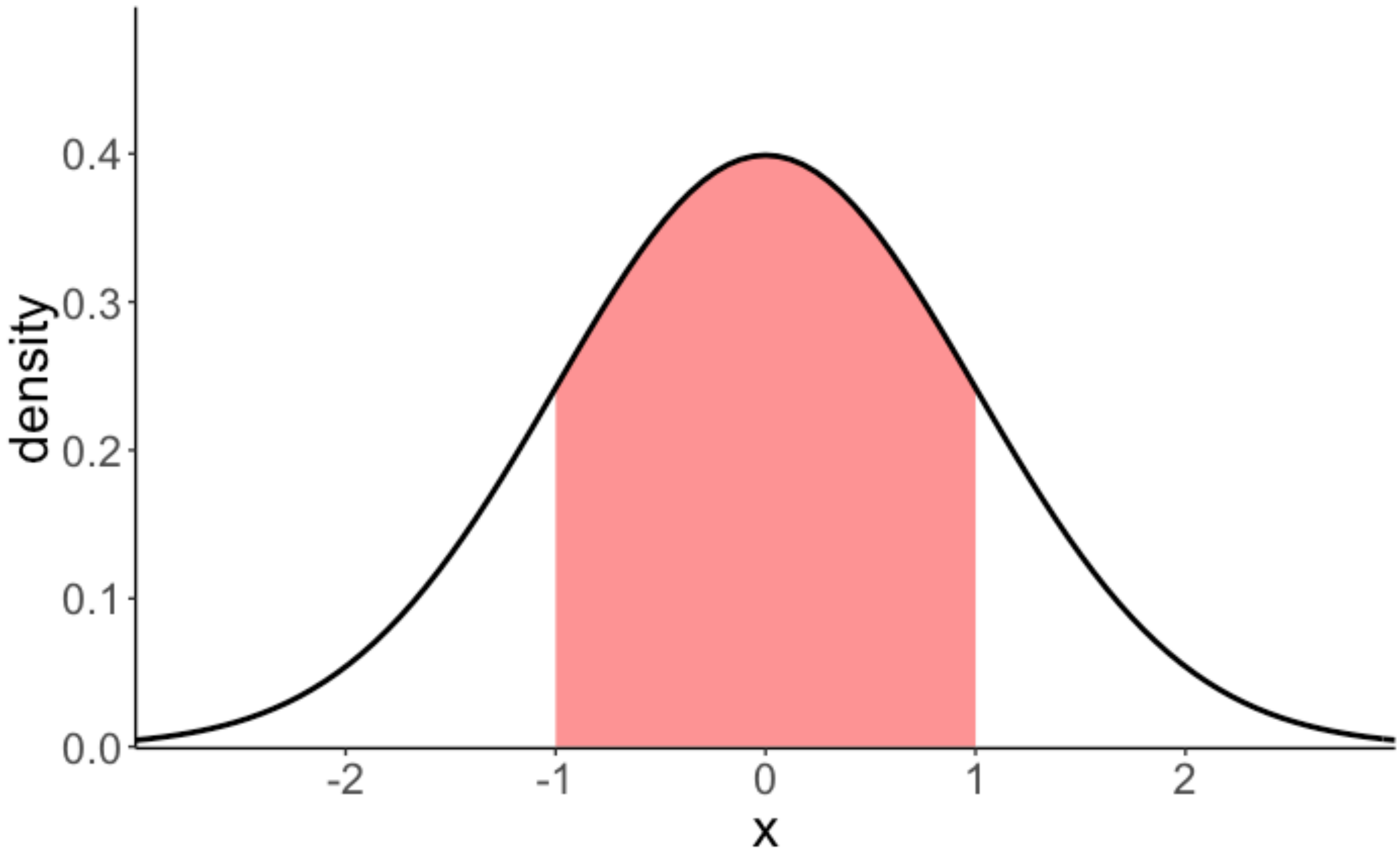
StatQuest makes me feel so happy….

https://www.youtube.com/watch?v=pYxNSUDSFH4

Probability vs. likelihood

Probability

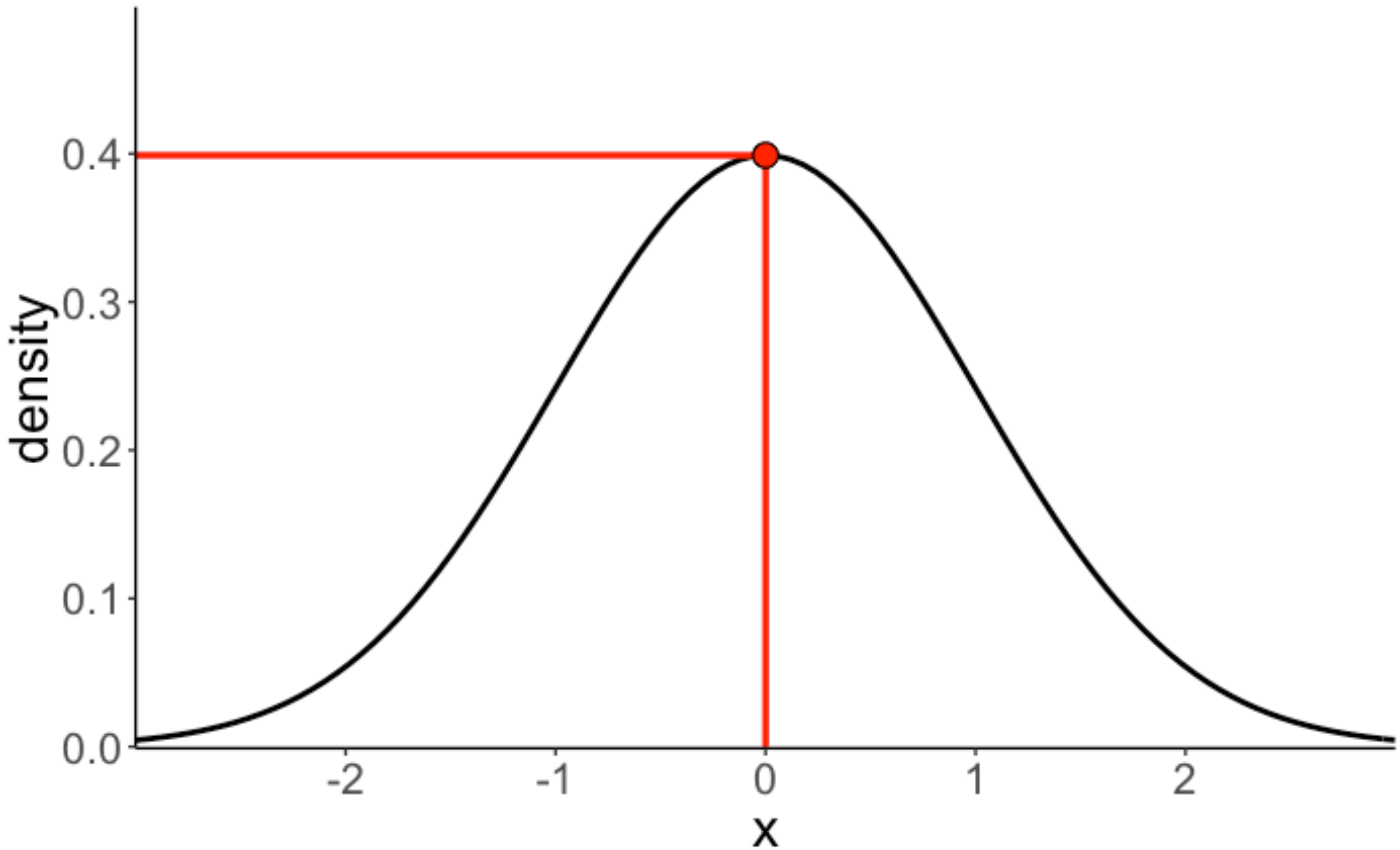
$$p(-1 < x < 1 \,|\, \text{mean} = 0, \text{ sd} = 1) = 0.68$$



Probability vs. likelihood

Likelihood

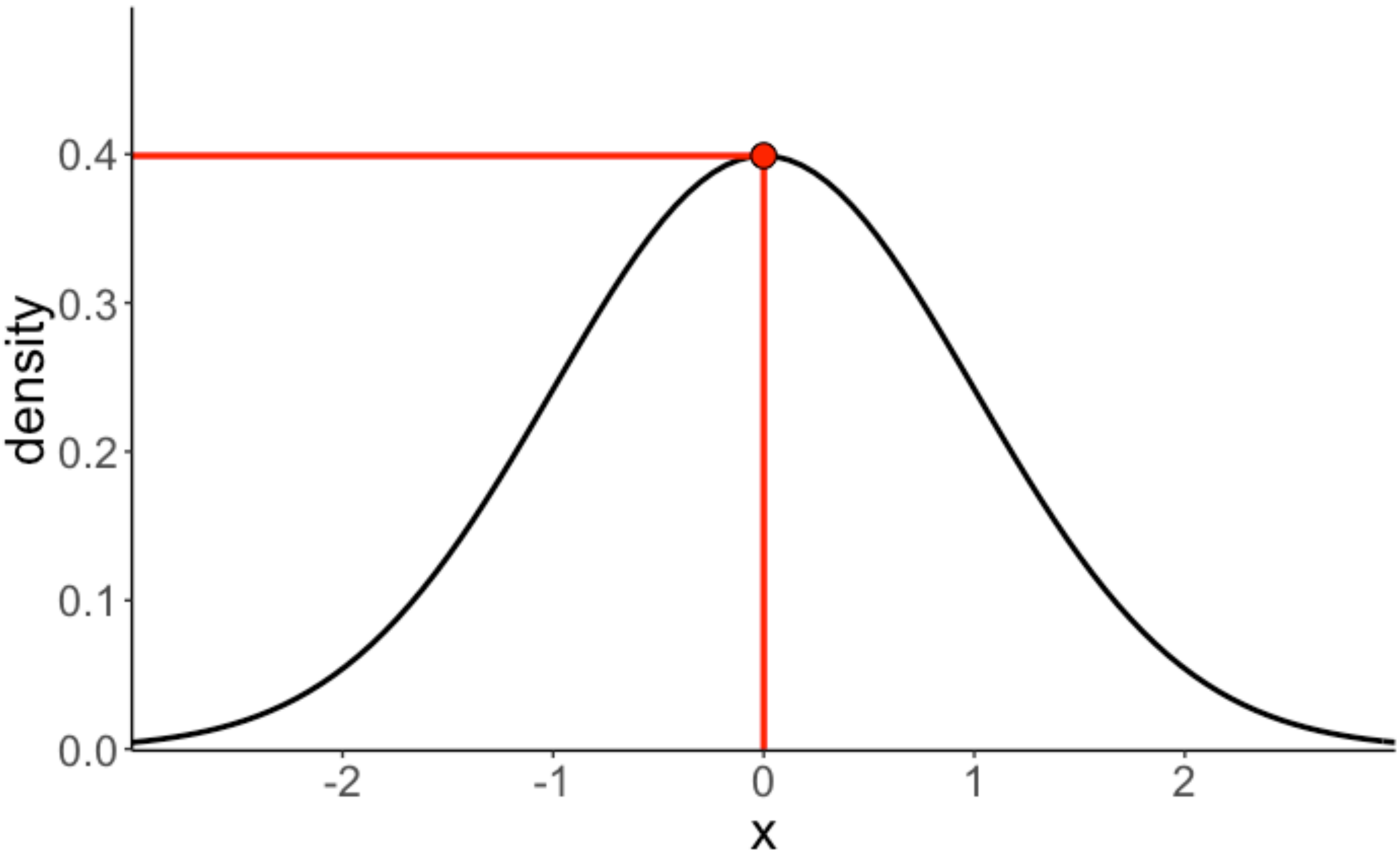
$$L(\text{mean} = 0,\ \text{sd} = 1 \,|\, x = 0) = 0.3989$$



Probability vs. likelihood

Likelihood

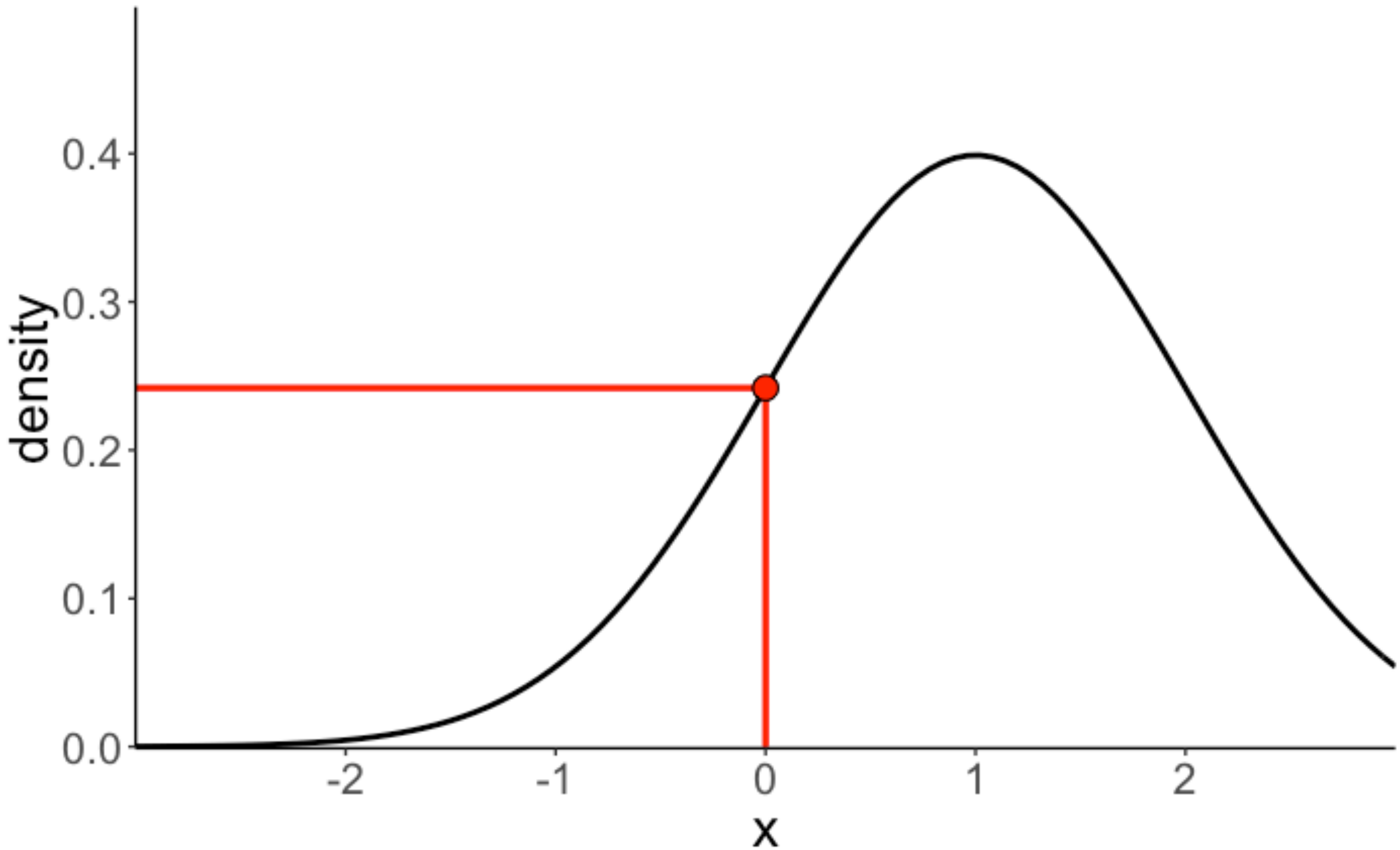
$$L(\text{mean} = 0,\ \text{sd} = 1 \,|\, x = 0) = 0.3989$$



Probability vs. likelihood

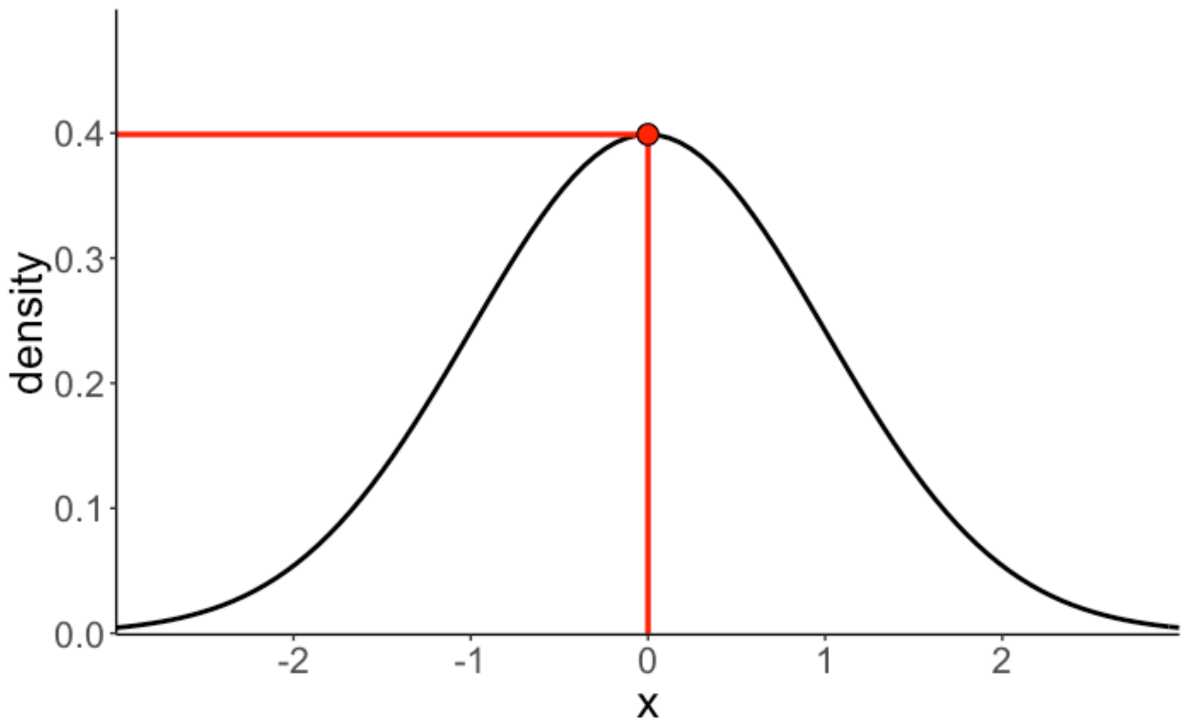
Likelihood

$$L(\text{mean} = 1,\ \text{sd} = 1 \,|\, x = 0) = 0.2419$$

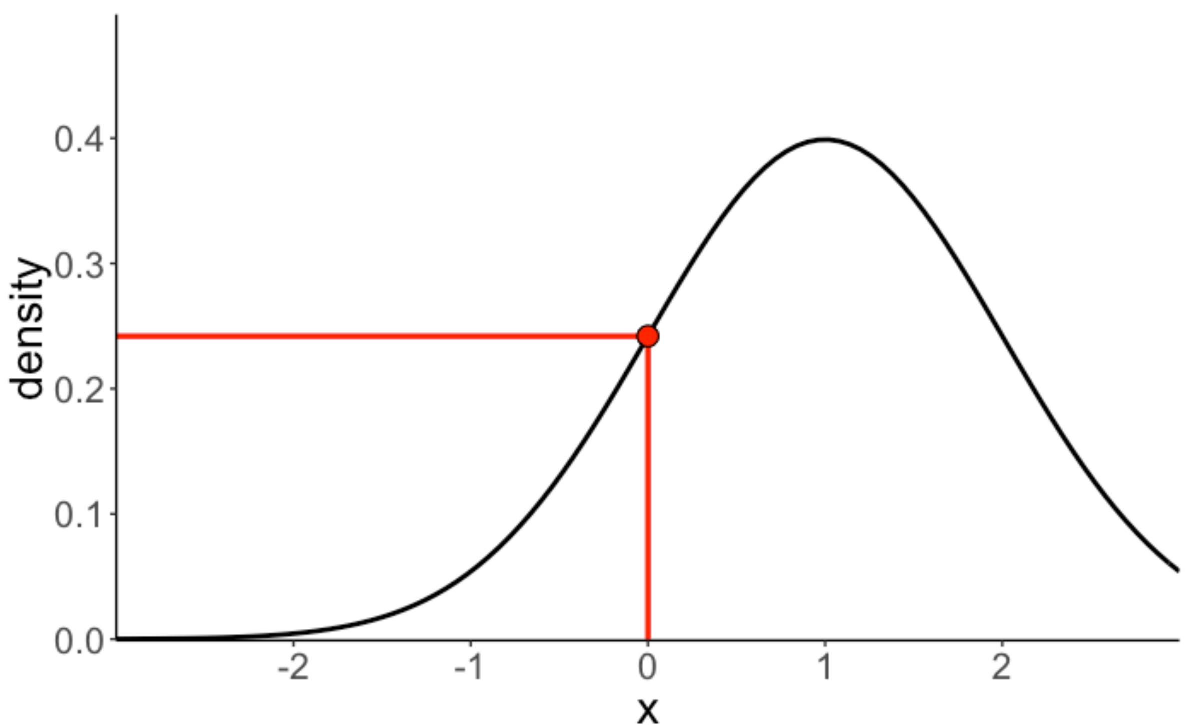


Probability vs. likelihood

Likelihood



$$L(\text{mean} = 0,\ \text{sd} = 1 \,|\, x = 0) = 0.3989$$



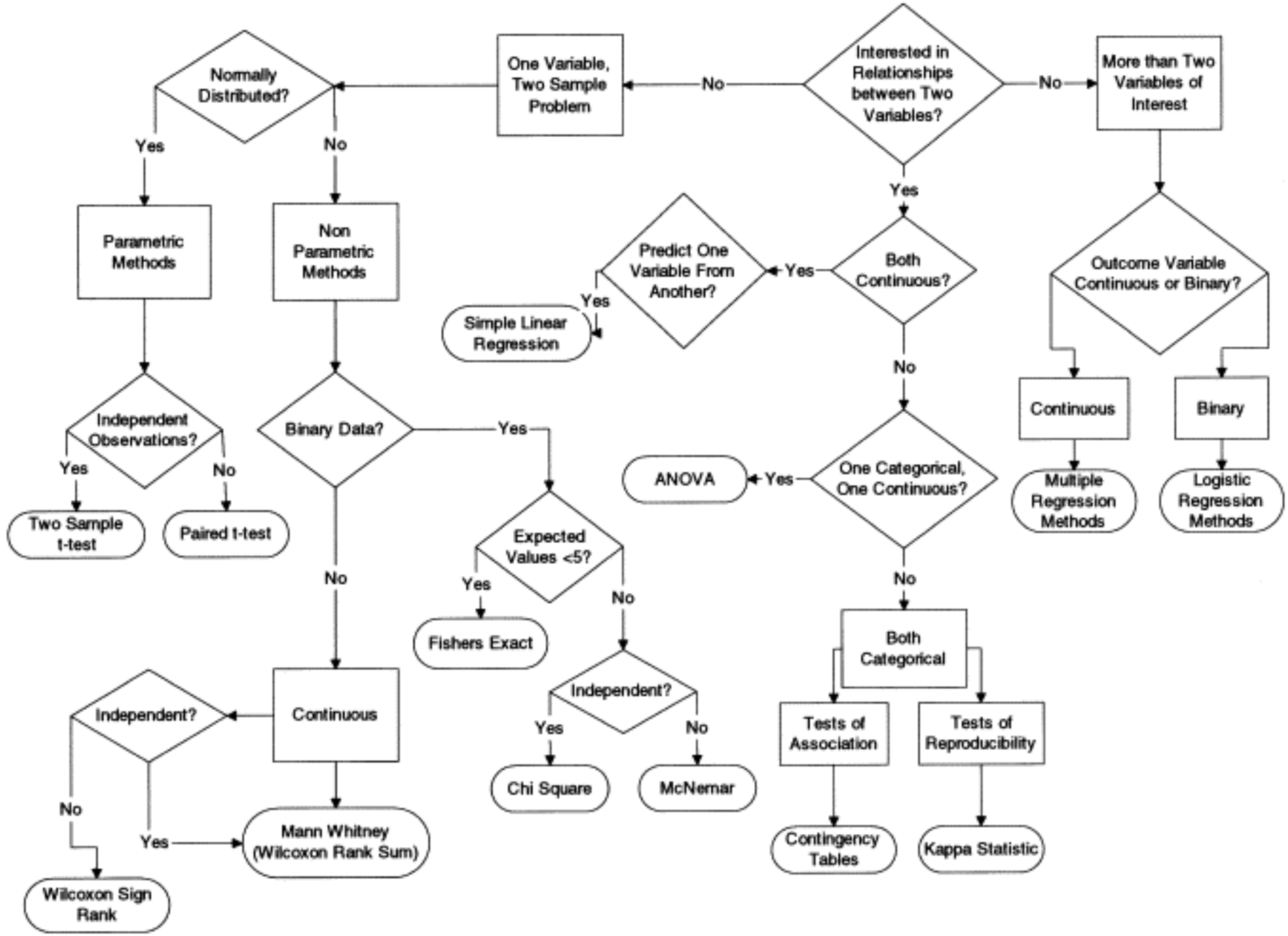
$$L(\text{mean} = 1,\ \text{sd} = 1 \,|\, x = 0) = 0.2419$$

Plan for today

- **Motivation**: Cookbook vs. Model Comparison
- Modeling data: Data = Model + Error
- Model: Choosing a model
- Error: Defining error
- Hypothesis testing as model comparison

Cookbook vs. Model Comparison

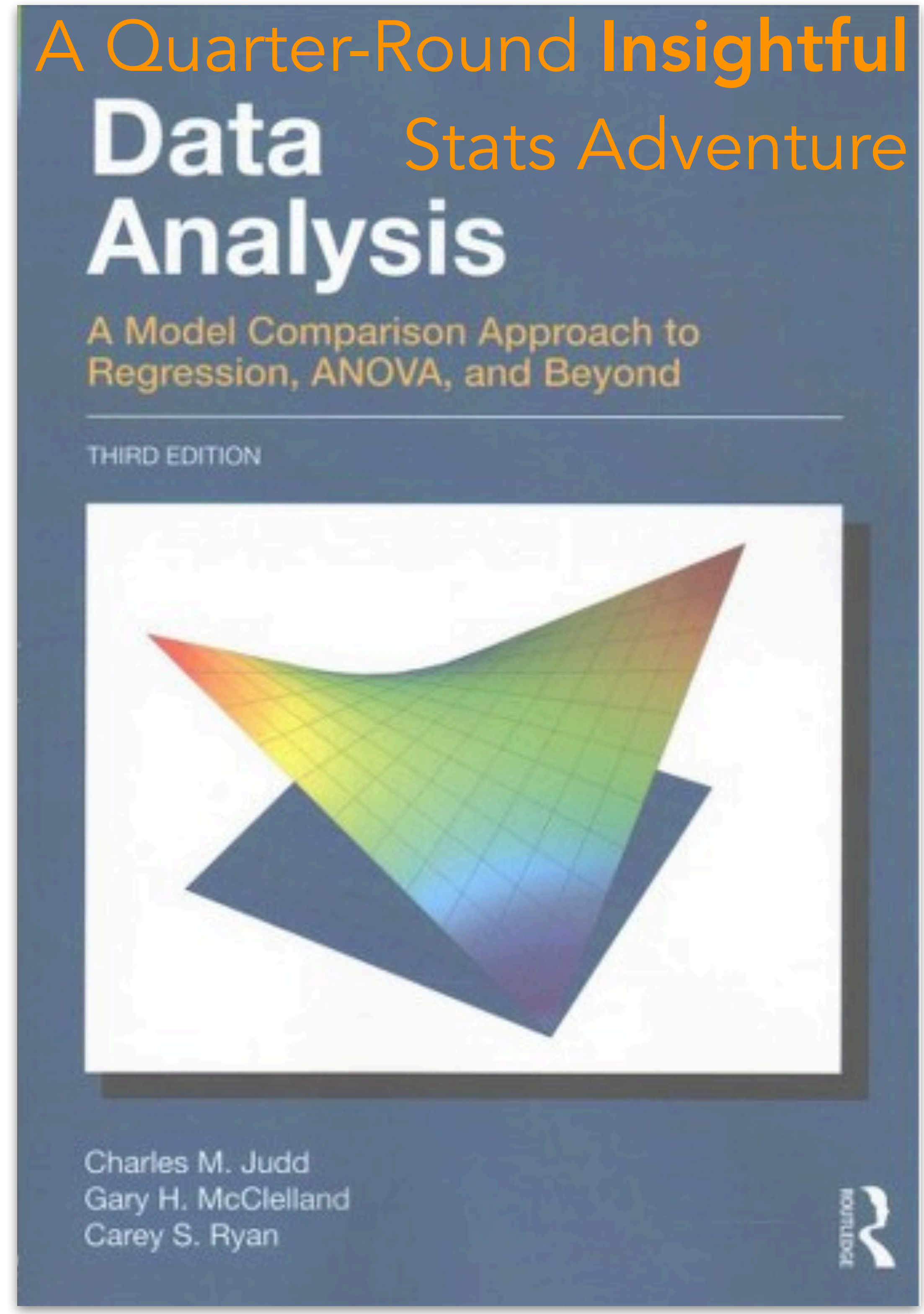
The cookbook approach



The cookbook approach

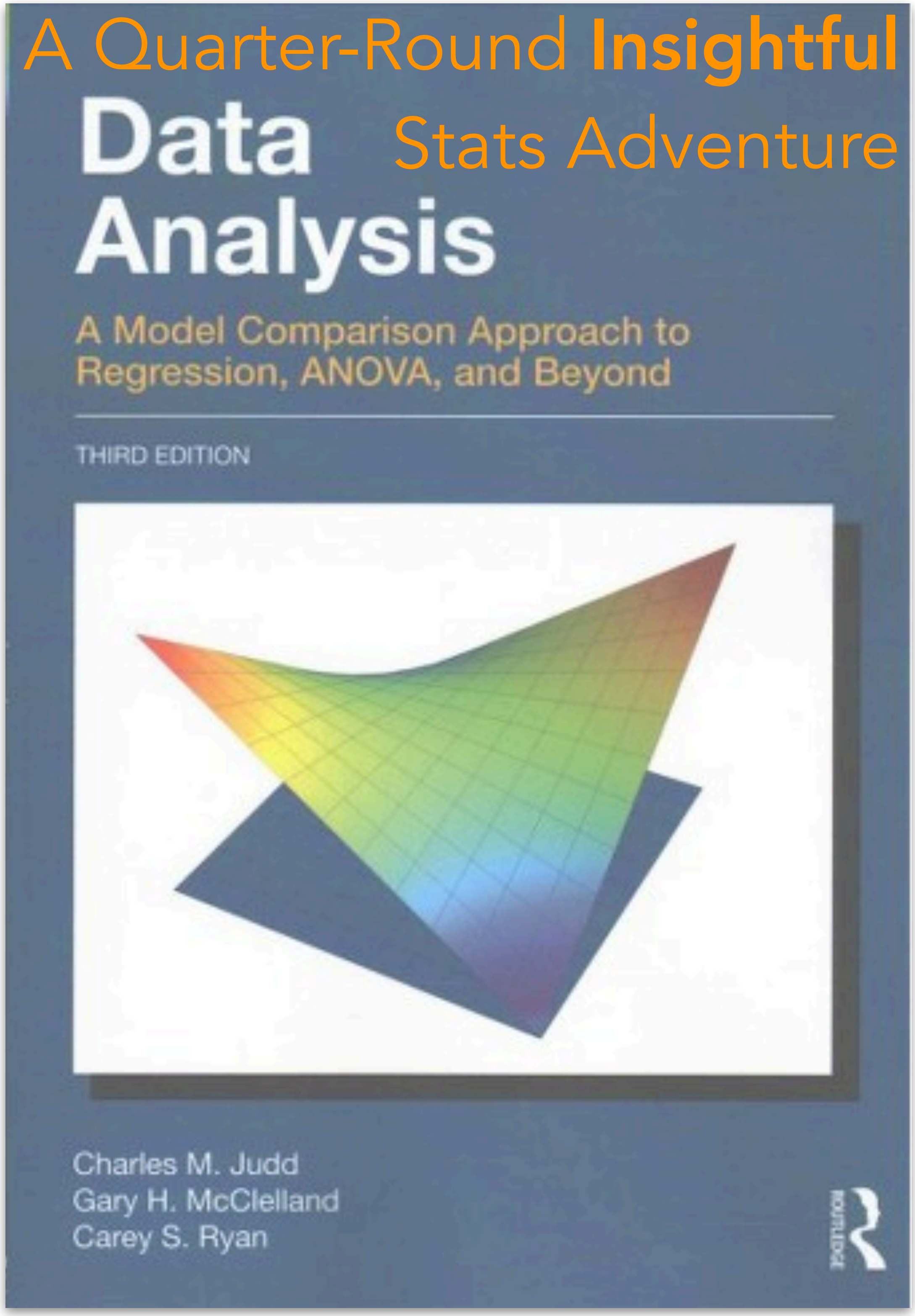
- many statistics textbooks are organized in this way
- works reasonably well if what we want to cook is in the book
- leaves us with no idea what to do if

Model comparison approach



Judd, C. M., McClelland, G. H., & Ryan, C. S. (2011). *Data analysis: A model comparison approach.* Routledge.

Model comparison approach



Judd, C. M., McClelland, G. H., & Ryan, C. S. (2011). *Data analysis: A model comparison approach.* Routledge.

Modeling data

Data = Model + Error

Feedback

How was the pace of today's class?

much too slow

a little too slow

just right

a little too fast

much too fast

How happy were you with today's class overall?

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

Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**

Thank you!