

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)

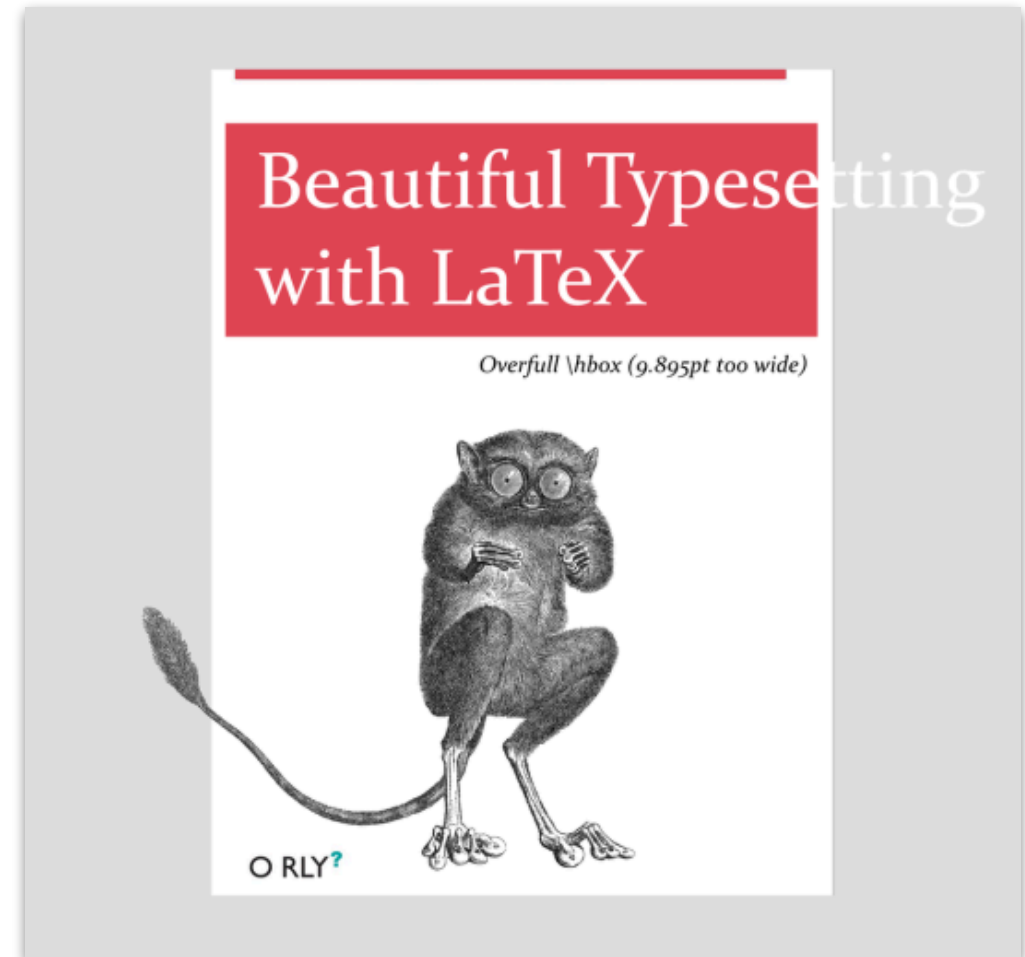
#### Per name across years - filter to names that were given at least 9000 times (overall) and occur at
df.data.35.names <- df.data.cleaned %>%
  group_by(name) %>%
  summarize(occurrences = n_distinct(year),
           overall_total = sum(year_total),
           mse = mean((female - 0.5)^2)) %>%
  filter(occurrences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



visualization2.Rmd

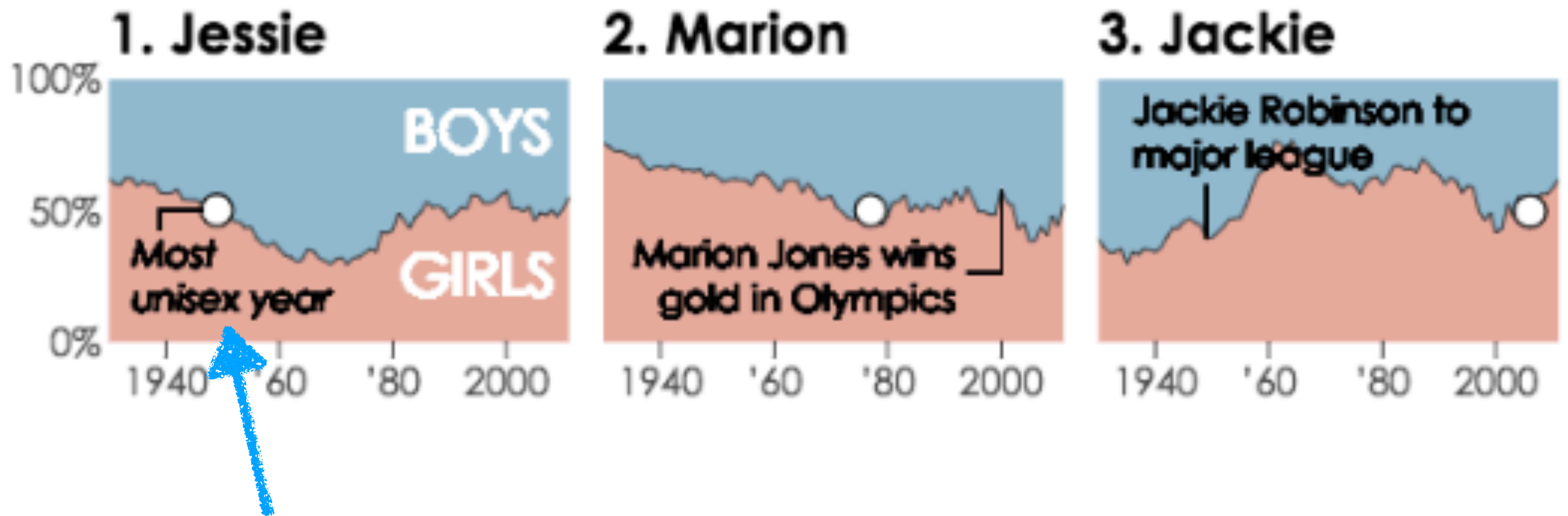
1 ---
2 title: "Class 3"
3 author: "Tobias Gerstenberg"
4 date: "January 11th, 2019"
5 output:
6 bookdown::html_document2:
7 toc: true
8 toc_depth: 4
9 theme: cosmo
10 highlight: tango
11 ---
12
13 {r setup, include=FALSE}
14 # these options here change the formatting of how comments are rendered
15 knitr::opts_chunk\$set(
16 collapse = TRUE,
17 comment = "#>"
18)
19
20 # Visualization 2
21
22 In this lecture, we will lift our 'ggplot2' skills to the next level!
23
24 ## Learning objectives
25
26 - Deciding what plot is appropriate for what kind of data.
27 - Customizing plots: Take a sad plot and make it better.
28 - Saving plots.
29 - Making figure panels.
30 - Debugging.
31 - Making animations.
32 - Defining snippets.
33
34

margin
column

Visualization 2
Learning objectives
Install and load pack...
Overview of different...
Proportions
Stacked bar charts
Pie charts
Comparisons
Boxplots
Violin plots
Joy plots
Practice plot 1
Relationships
Scatter plots
Raster plots
Temporal data
Customizing plots
Changing the order...
Dealing with legends
Choosing good colors
Customizing themes
Saving plots
Creating figure panels
Peeking behind the ...
Making animations
Shiny apps
Defining snippets
Additional resources
Cheatsheets
Data camp courses
Books and chapters
Misc
Session info

12:1 (Top Level) R Markdown

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!

```
1 data = c(1, 3, 4, 2, 5)
2 prediction = c(1, 2, 2, 1, 4)
3
4 # calculate root mean squared error the pipe way
5 rmse = (prediction-data)^2 %>%
6   mean() %>%
7   sqrt() %>%
8   print()
```

can we pipe this even more?



```
1 rmse = prediction %>%
2   subtract(data) %>%
3   raise_to_power(2) %>%
4   mean() %>%
5   sqrt() %>%
6   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

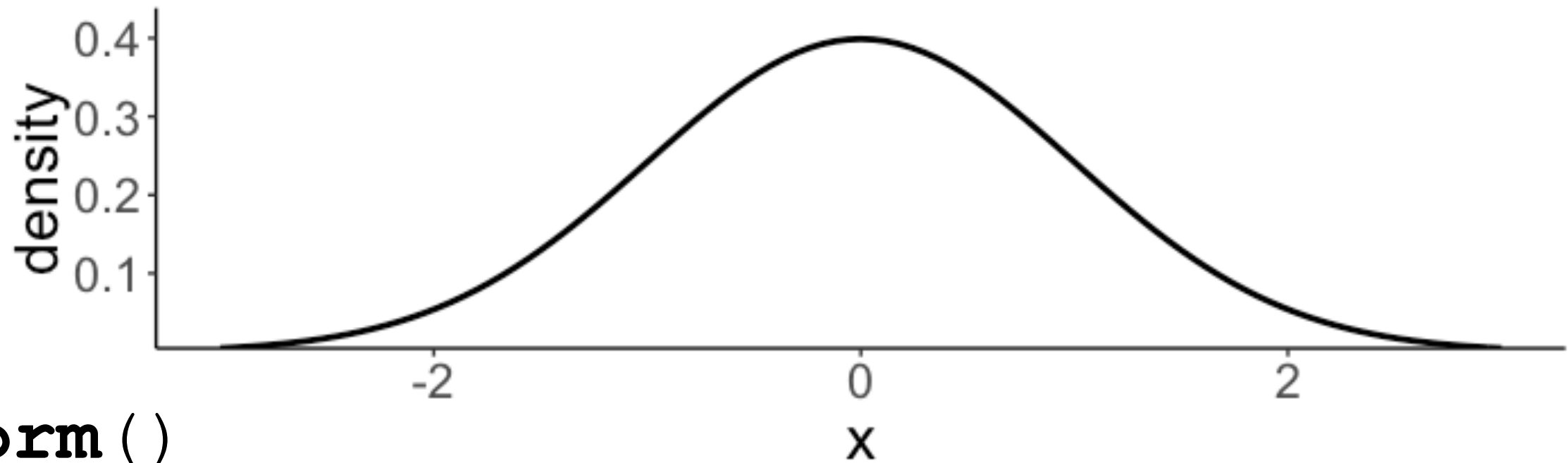
I think you spent too long on CLT which isn't intuitively difficult and would like more time on the harder topics near the end

**sometimes it's tricky
to get it right**

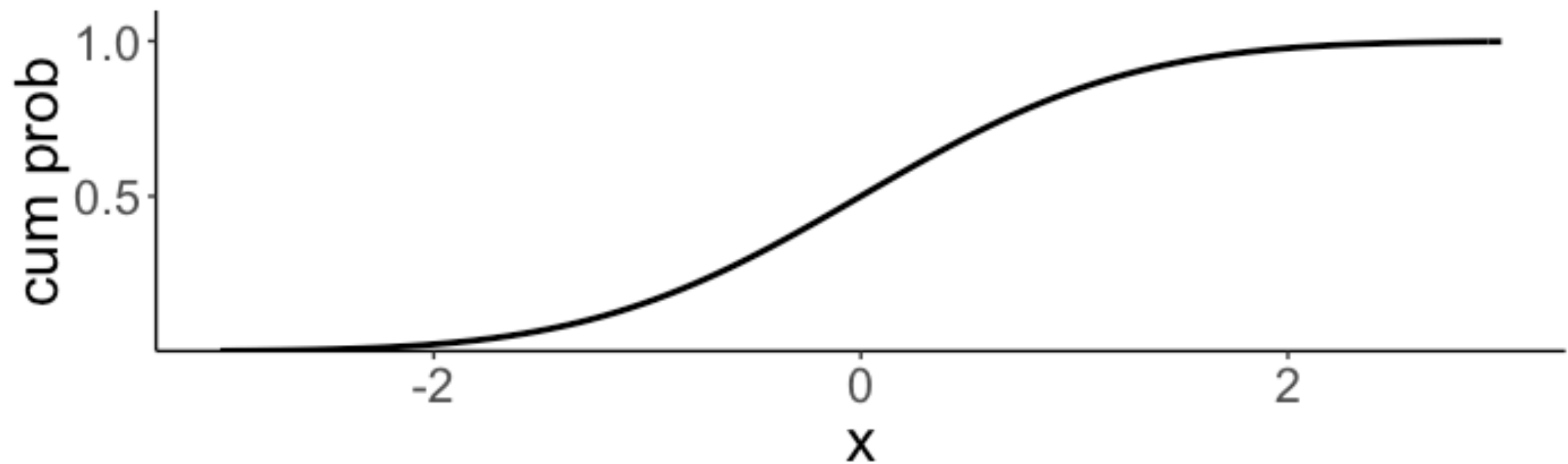
Probability vs. likelihood

Probability vs. likelihood

dnorm ()

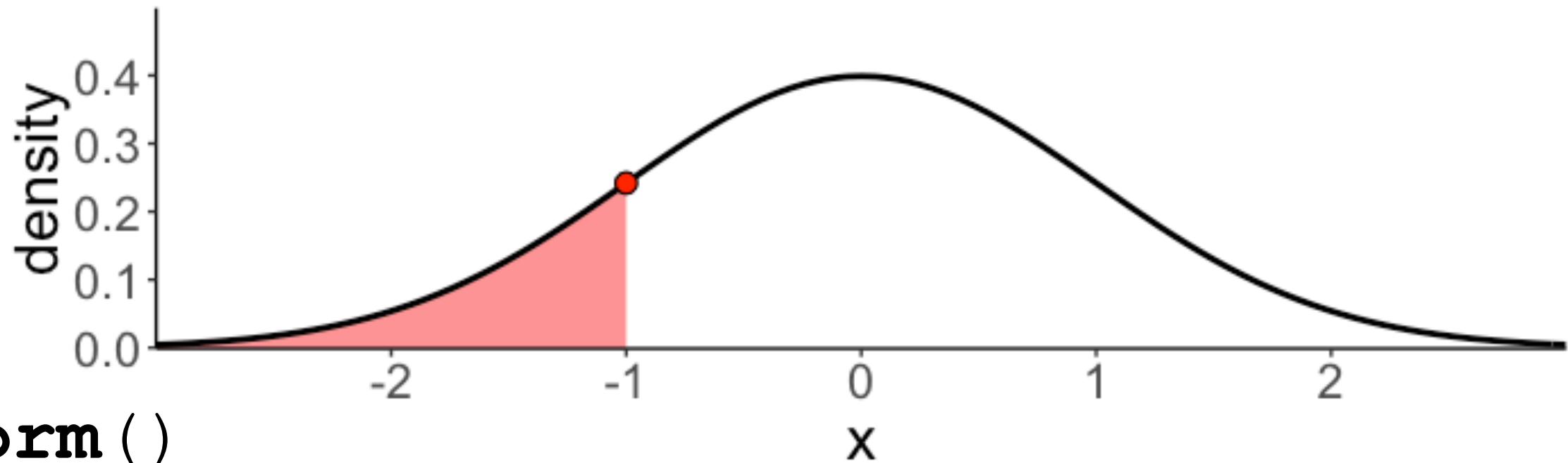


pnorm ()

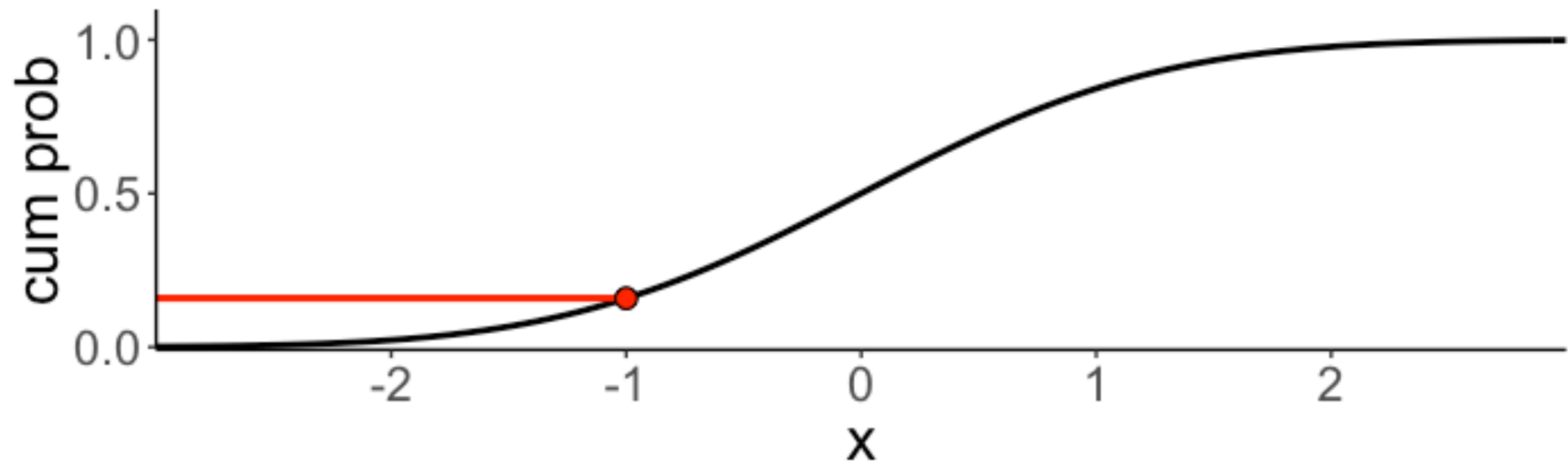


Probability vs. likelihood

dnorm ()

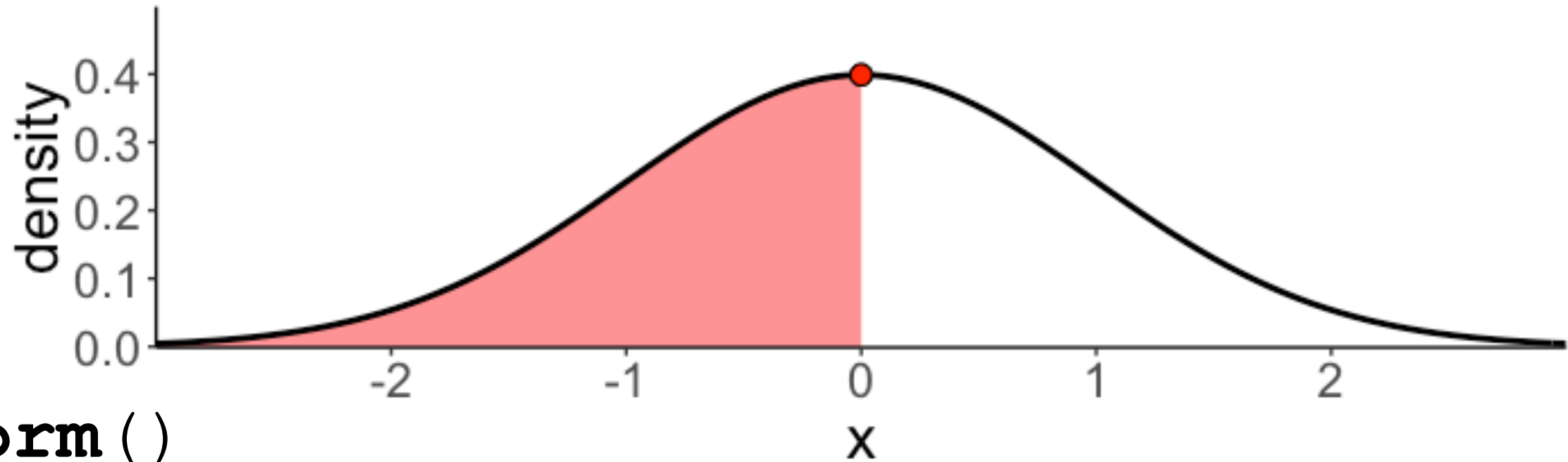


pnorm ()

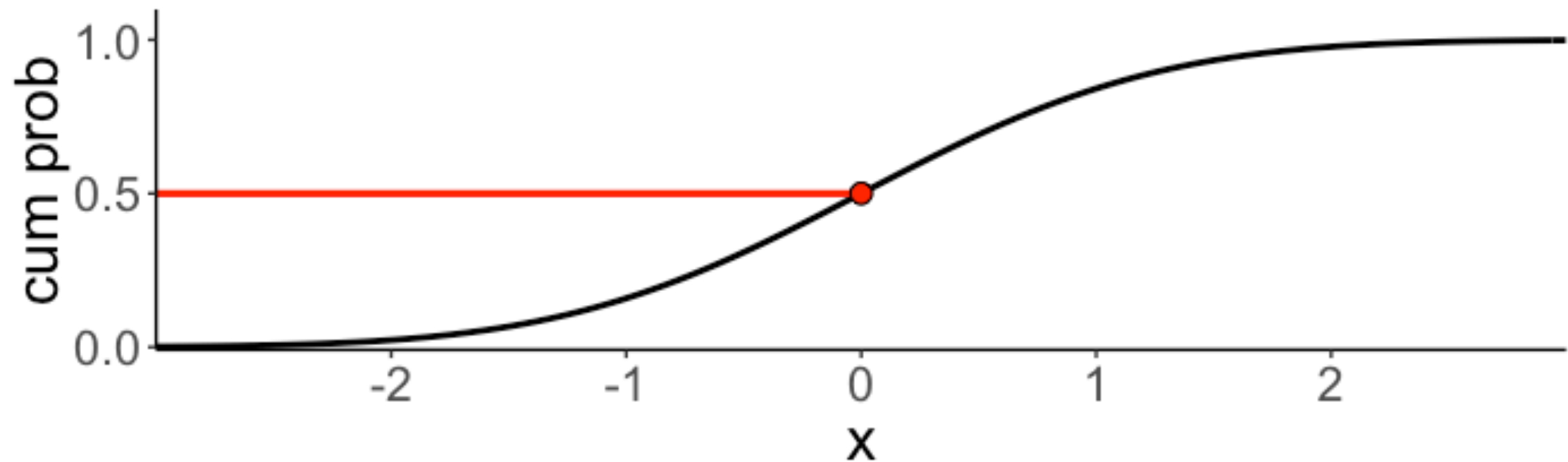


Probability vs. likelihood

dnorm ()

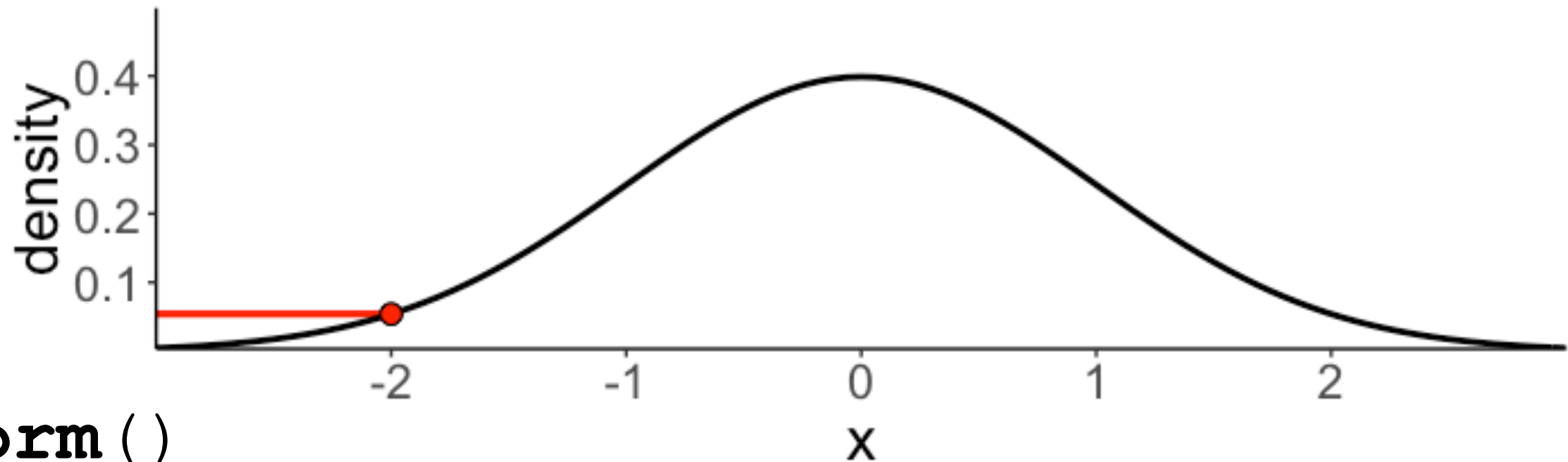


pnorm ()

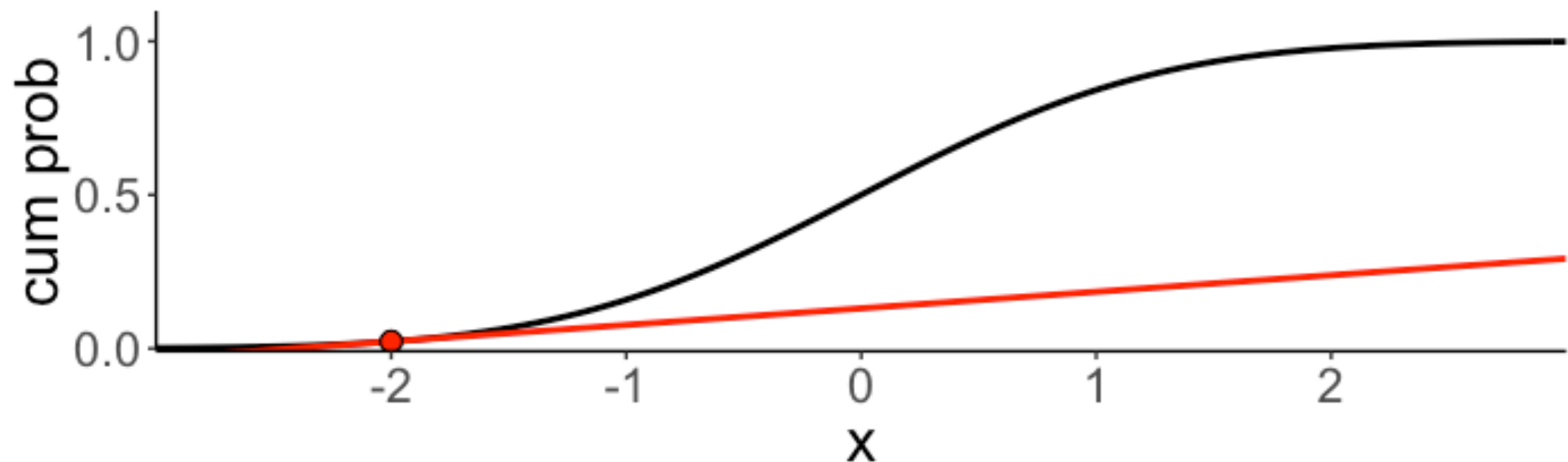


Probability vs. likelihood

dnorm ()



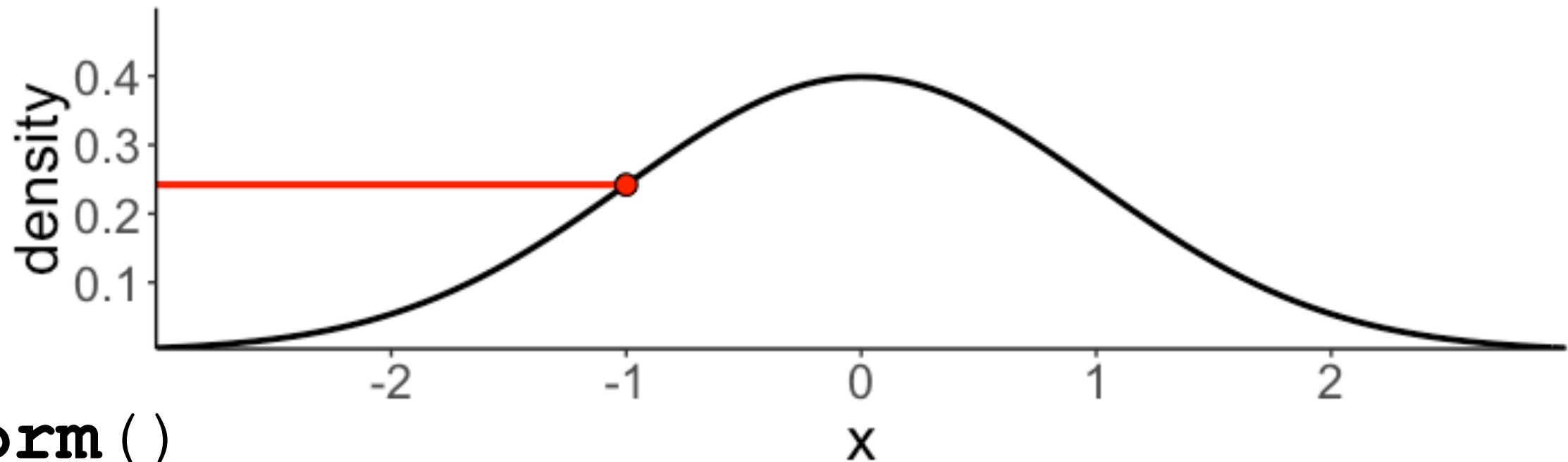
pnorm ()



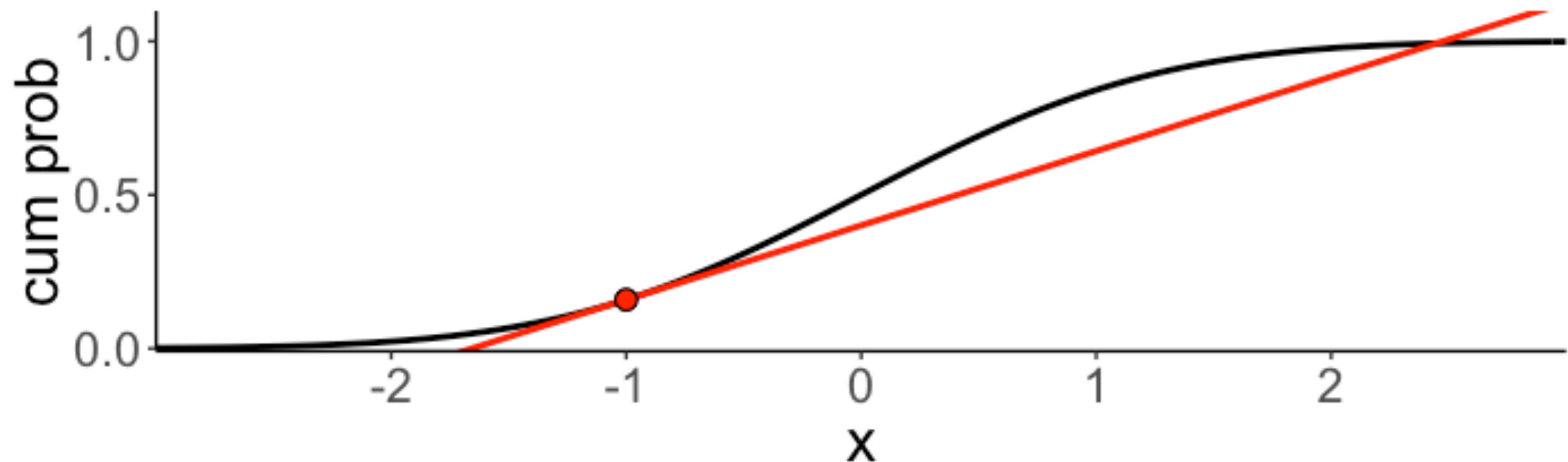
dnorm () is the first derivative of **pnorm ()**

Probability vs. likelihood

dnorm ()



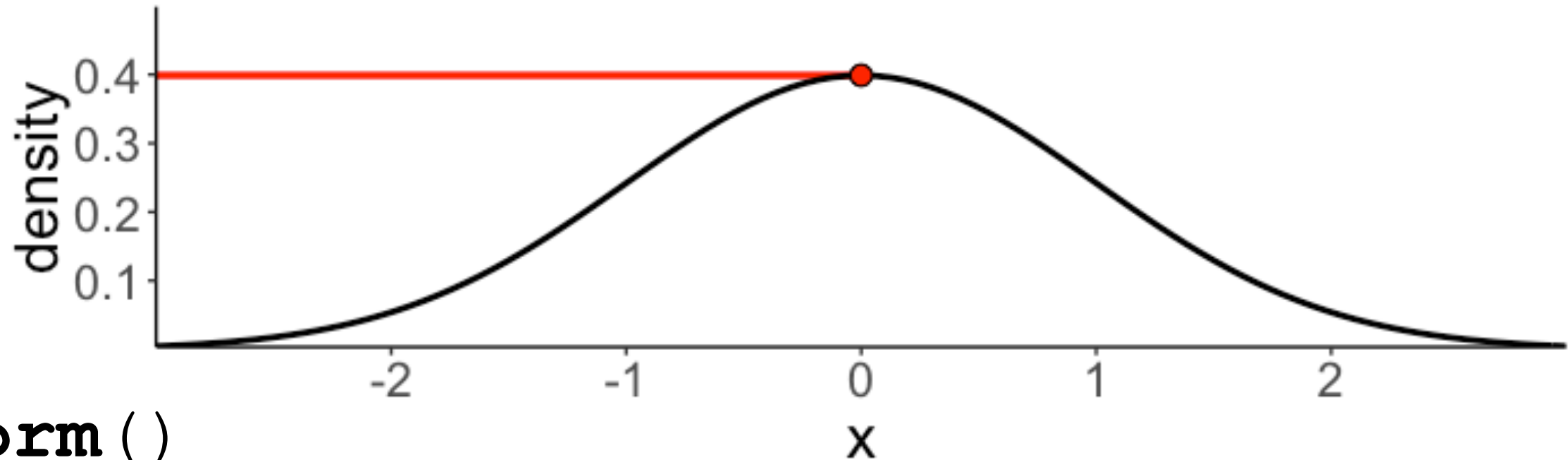
pnorm ()



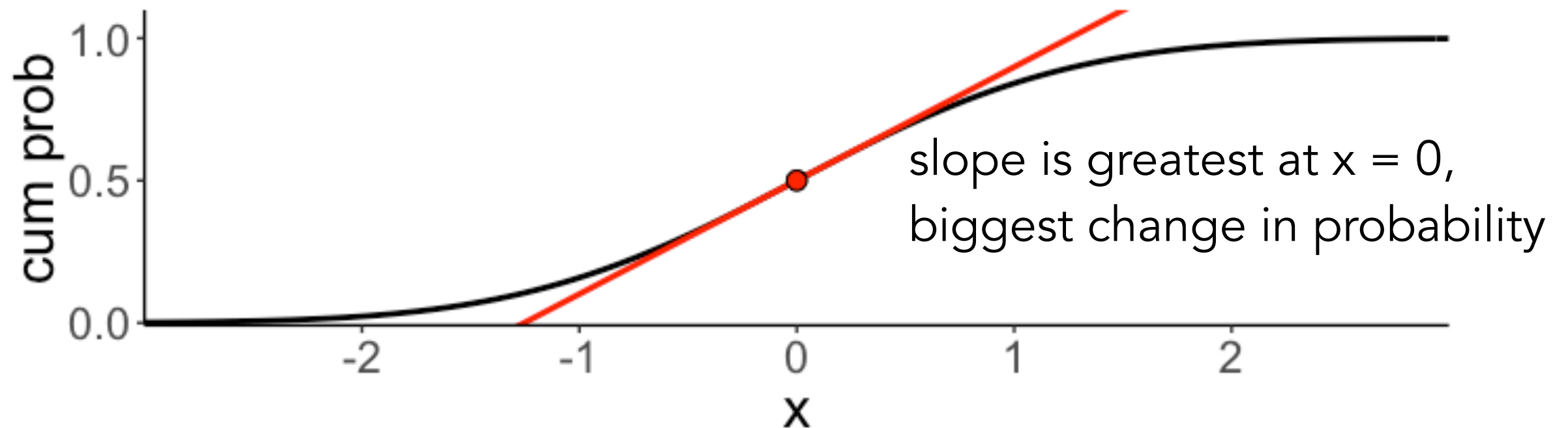
dnorm () is the first derivative of **pnorm ()**

Probability vs. likelihood

dnorm ()



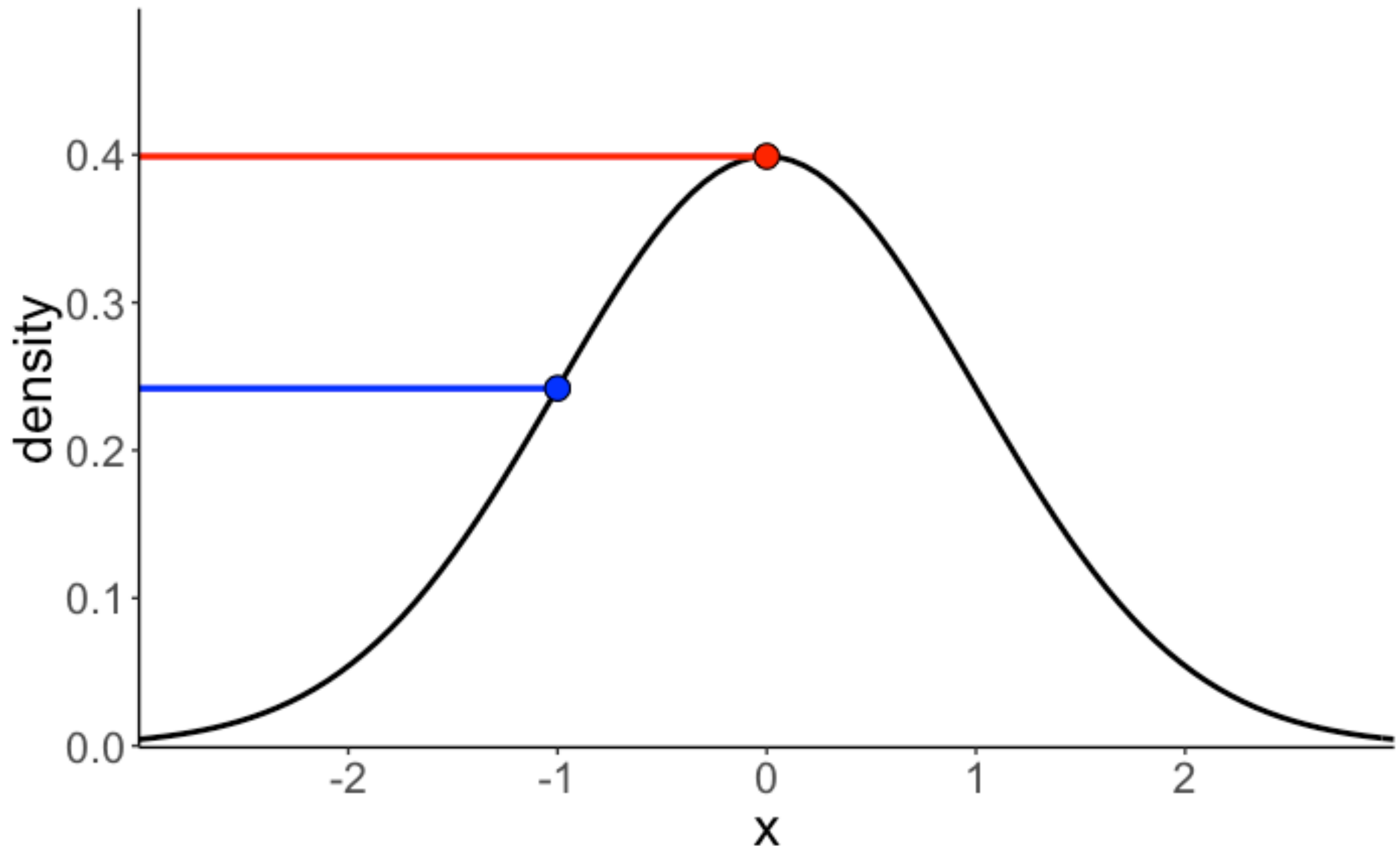
pnorm ()



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

Probability vs. likelihood

$$\text{dnorm}(0) / \text{dnorm}(-1) = 1.6487$$

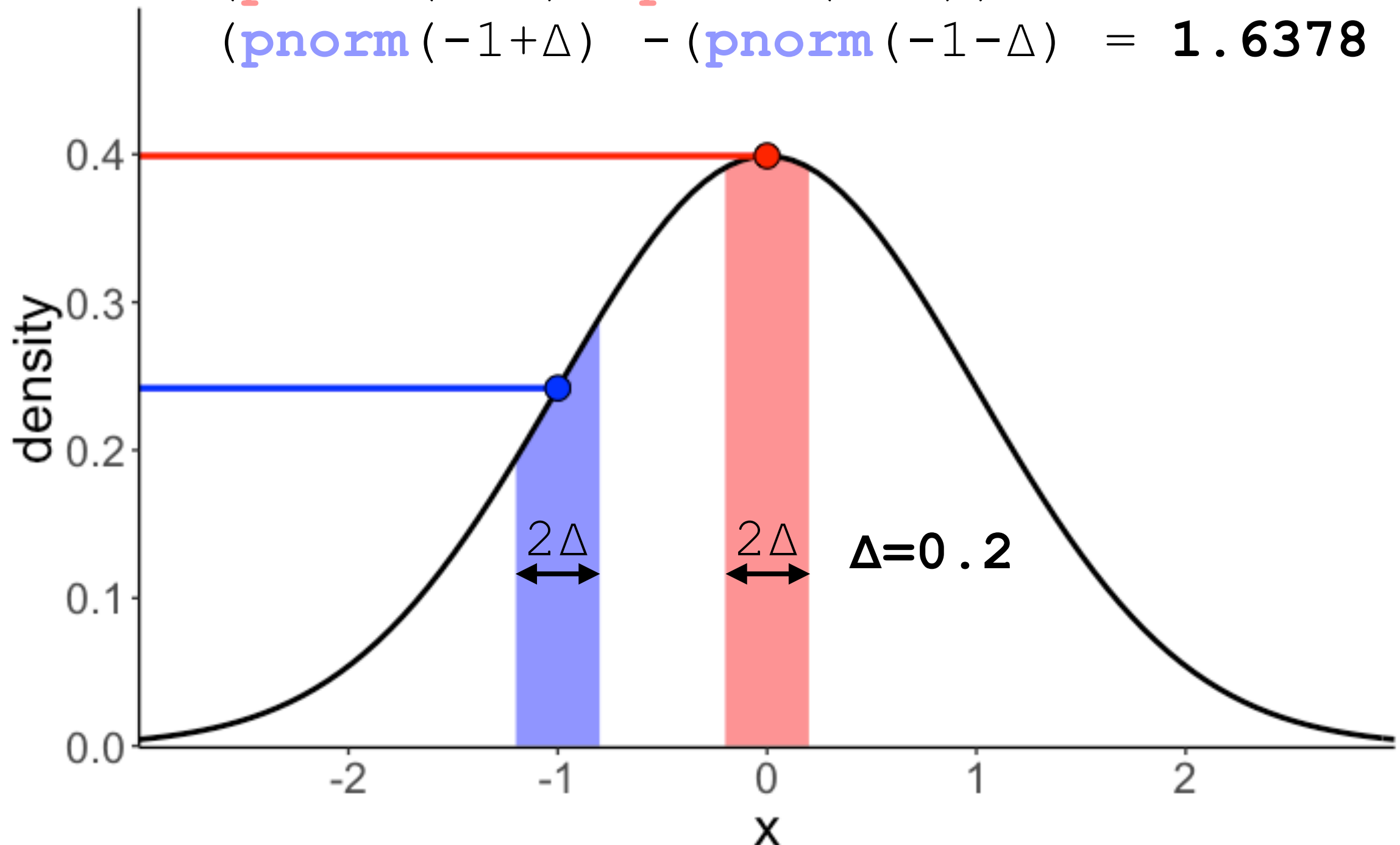


relative probability of one value vs. another

Probability vs. likelihood

$$\text{dnorm}(0) / \text{dnorm}(-1) = 1.6487$$

$$(\text{pnorm}(0+\Delta) - \text{pnorm}(0-\Delta)) / (\text{pnorm}(-1+\Delta) - \text{pnorm}(-1-\Delta)) = \mathbf{1.6378}$$

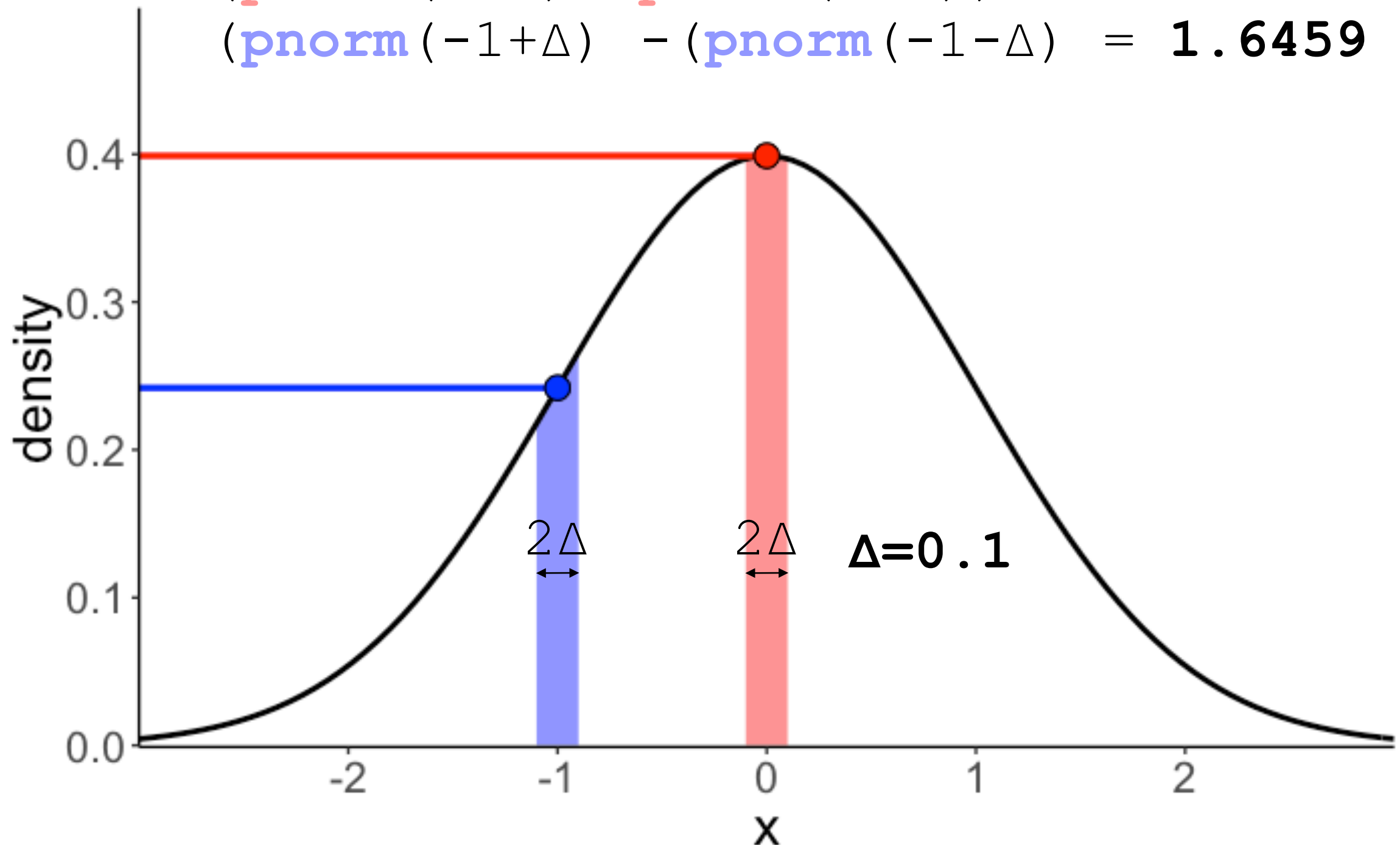


relative probability of one value vs. another

Probability vs. likelihood

$$\text{dnorm}(0) / \text{dnorm}(-1) = 1.6487$$

$$(\text{pnorm}(0+\Delta) - \text{pnorm}(0-\Delta)) / (\text{pnorm}(-1+\Delta) - \text{pnorm}(-1-\Delta)) = 1.6459$$

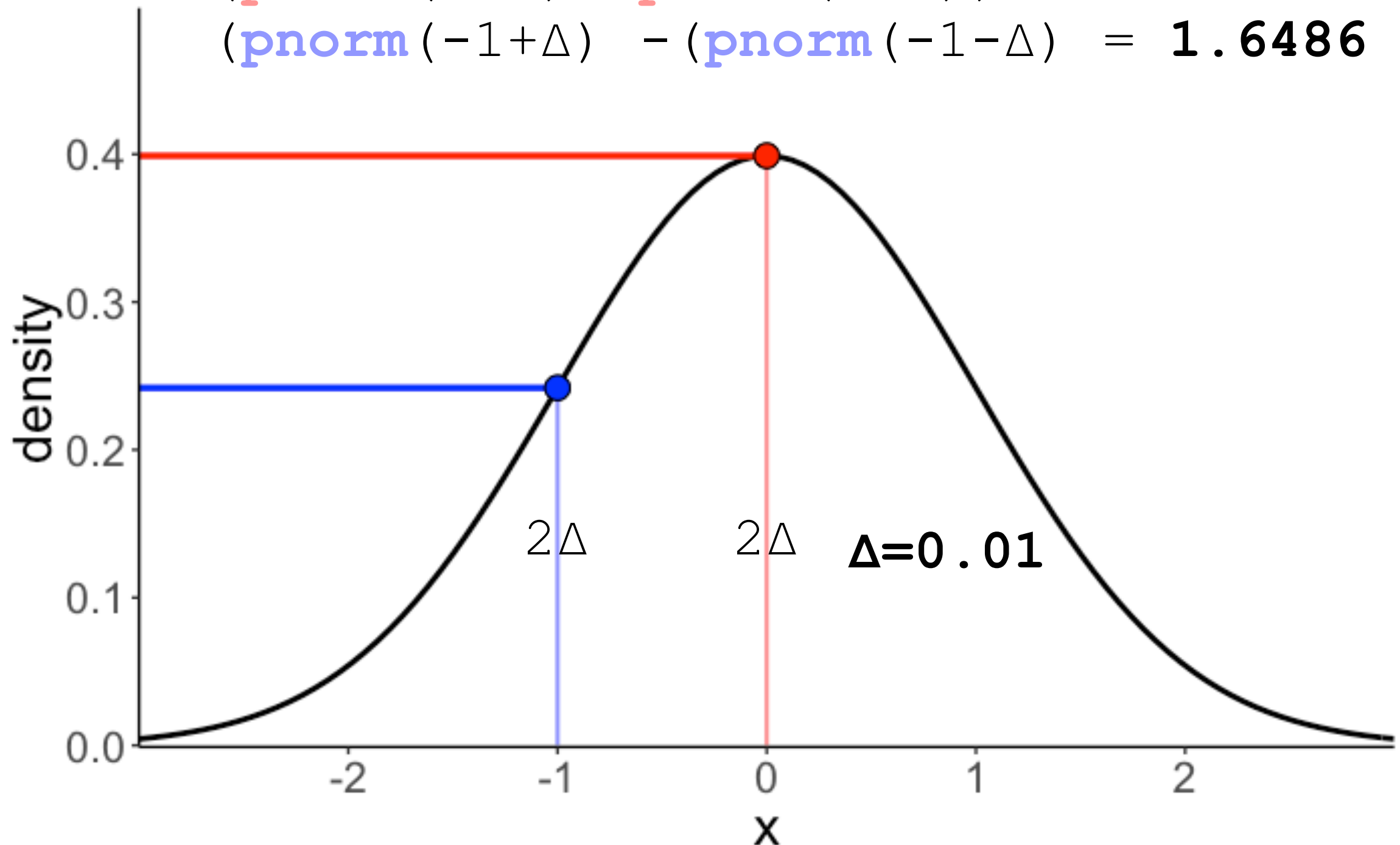


relative probability of one value vs. another

Probability vs. likelihood

$$\text{dnorm}(0) / \text{dnorm}(-1) = 1.6487$$

$$(\text{pnorm}(0+\Delta) - \text{pnorm}(0-\Delta)) / (\text{pnorm}(-1+\Delta) - \text{pnorm}(-1-\Delta)) = 1.6486$$



relative probability of one value vs. another

Probability vs. likelihood

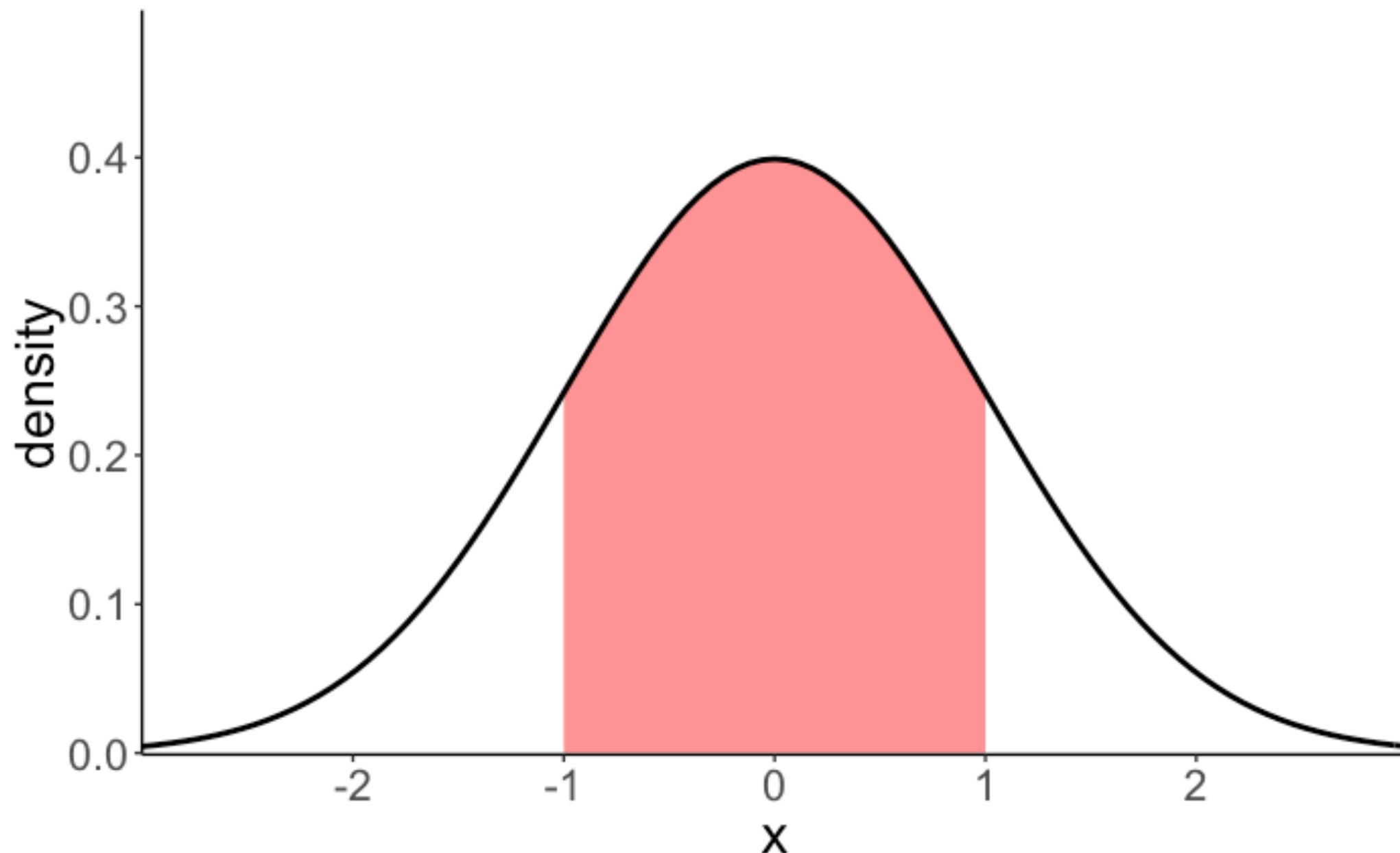
StatQuest makes me feel so happy....

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

Probability vs. likelihood

Probability

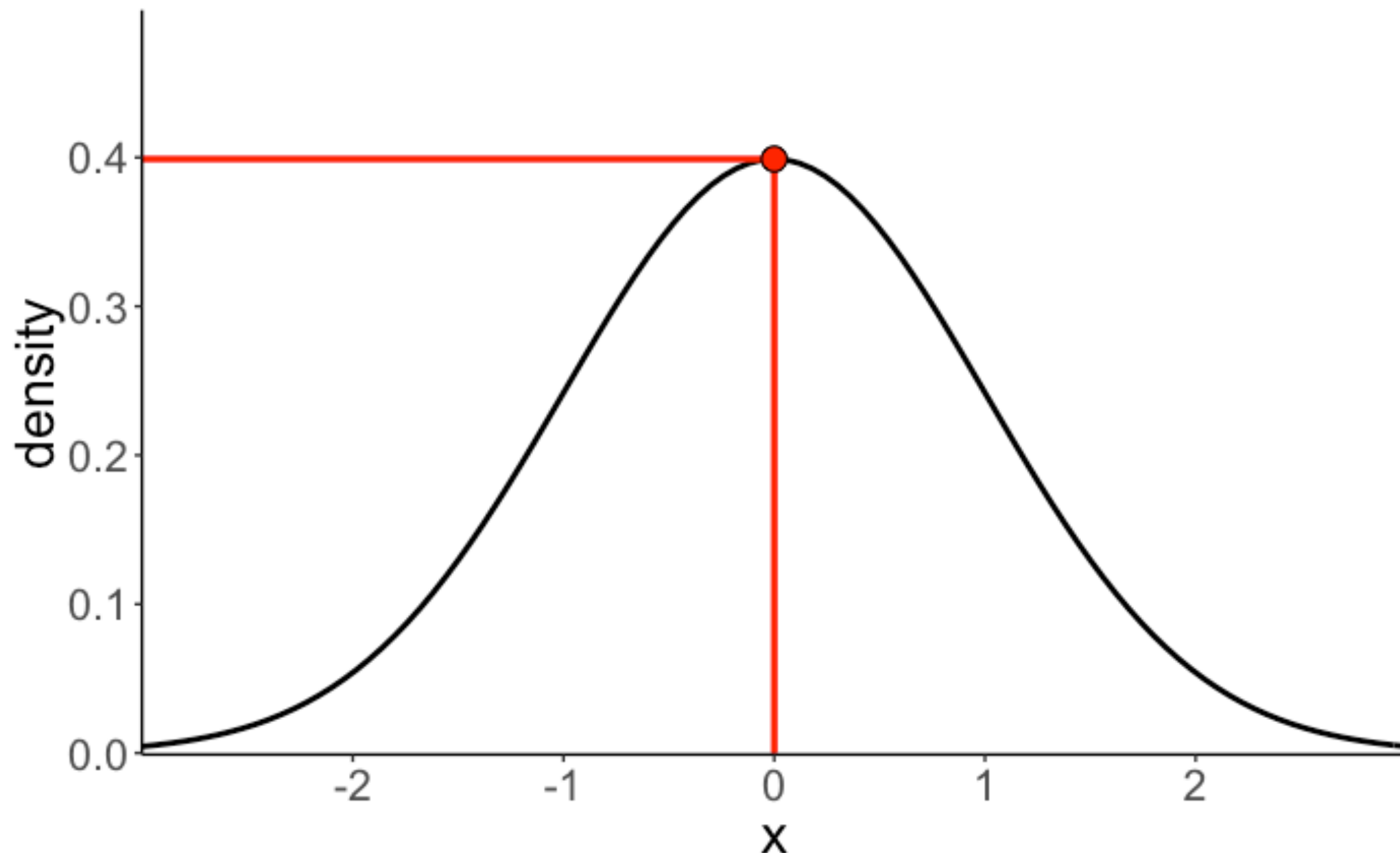
$$p(-1 < x < 1 | \text{mean} = 0, \text{sd} = 1) = 0.68$$



Probability vs. likelihood

Likelihood

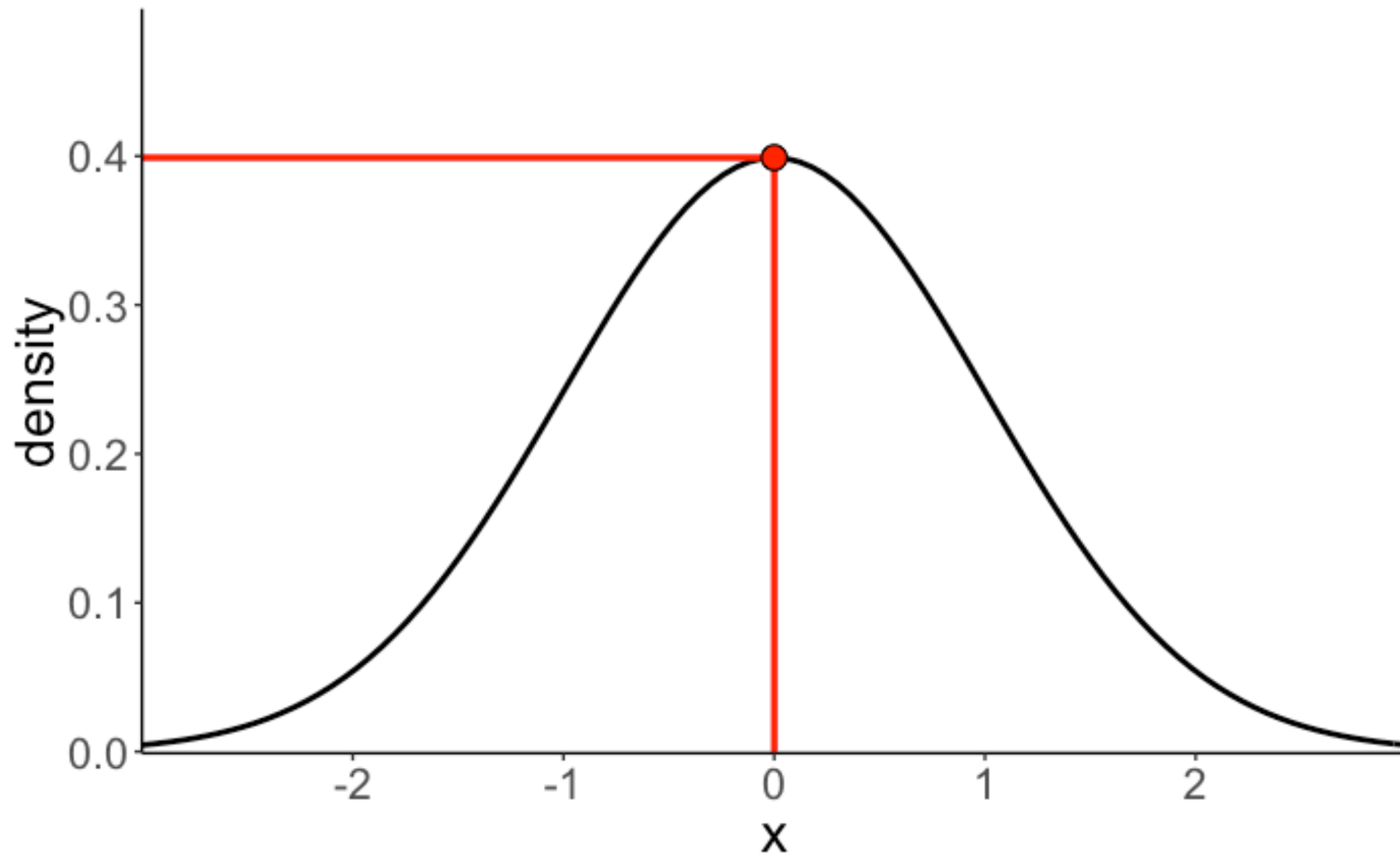
$$L(\text{mean} = 0, \text{sd} = 1 \mid x = 0) = 0.3989$$



Probability vs. likelihood

Likelihood

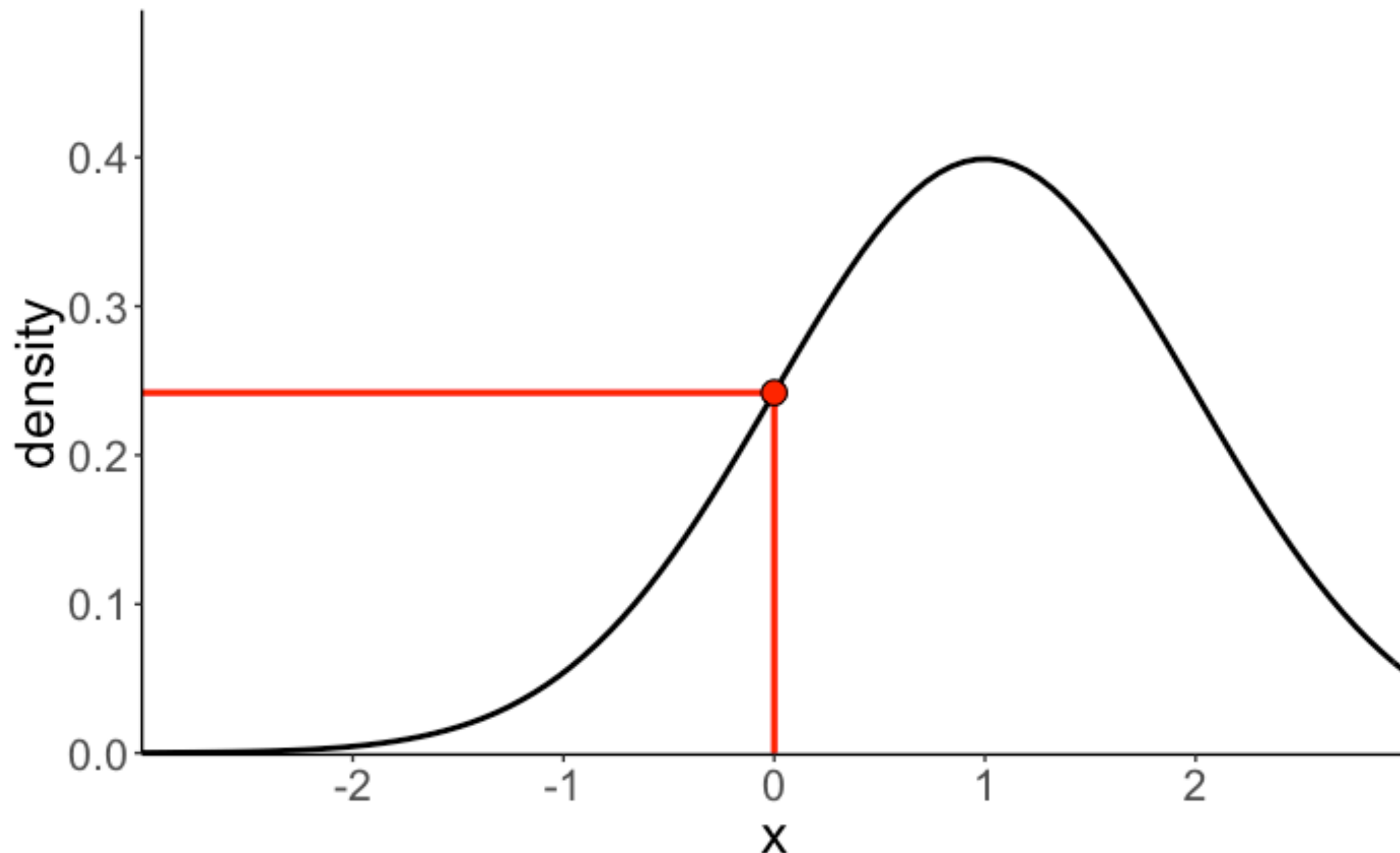
$$L(\text{mean} = 0, \text{sd} = 1 \mid x = 0) = 0.3989$$



Probability vs. likelihood

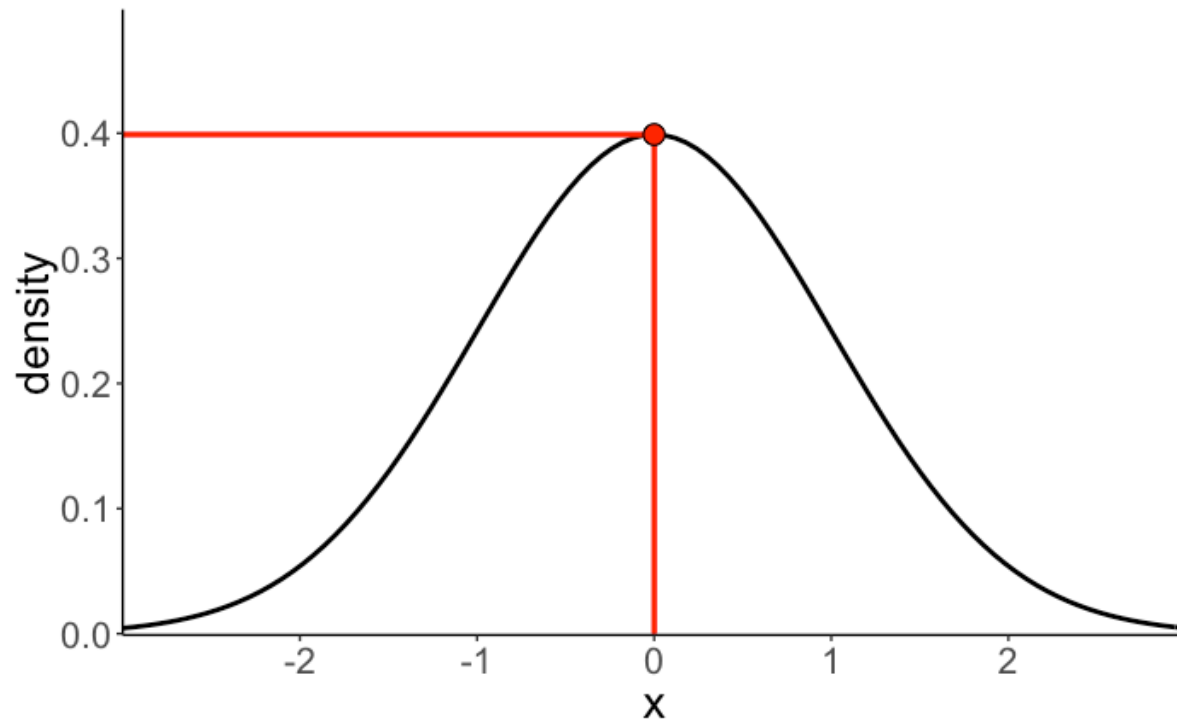
Likelihood

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

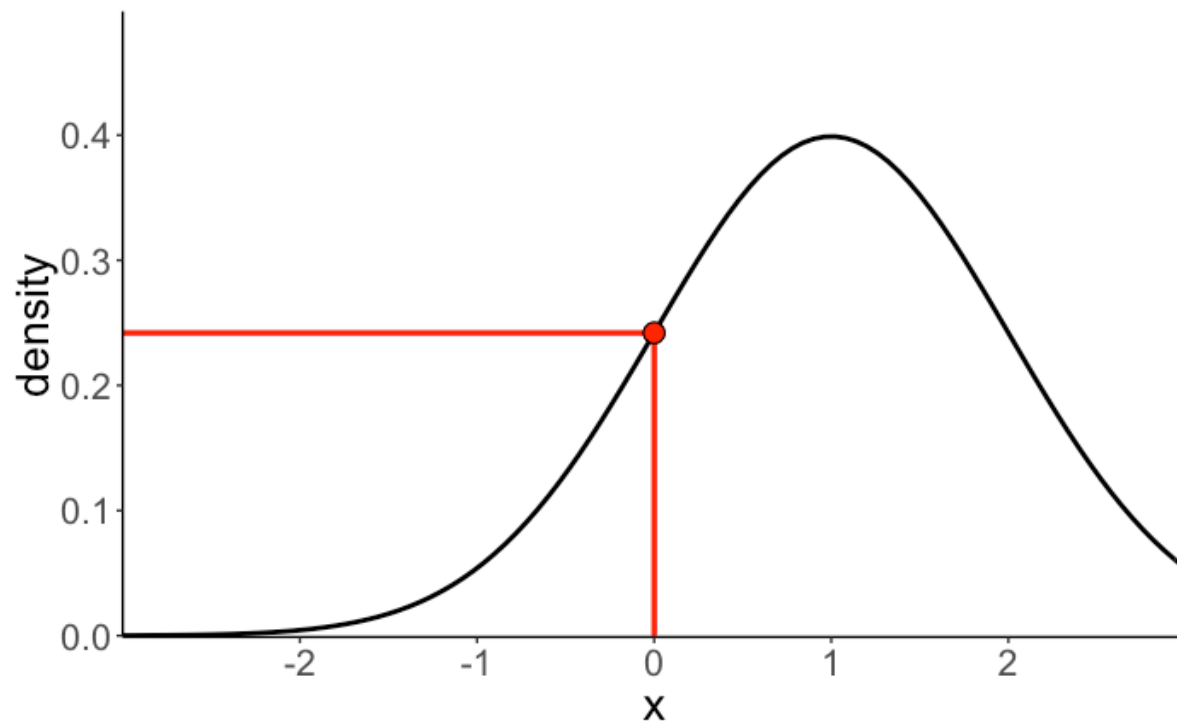


Probability vs. likelihood

Likelihood



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



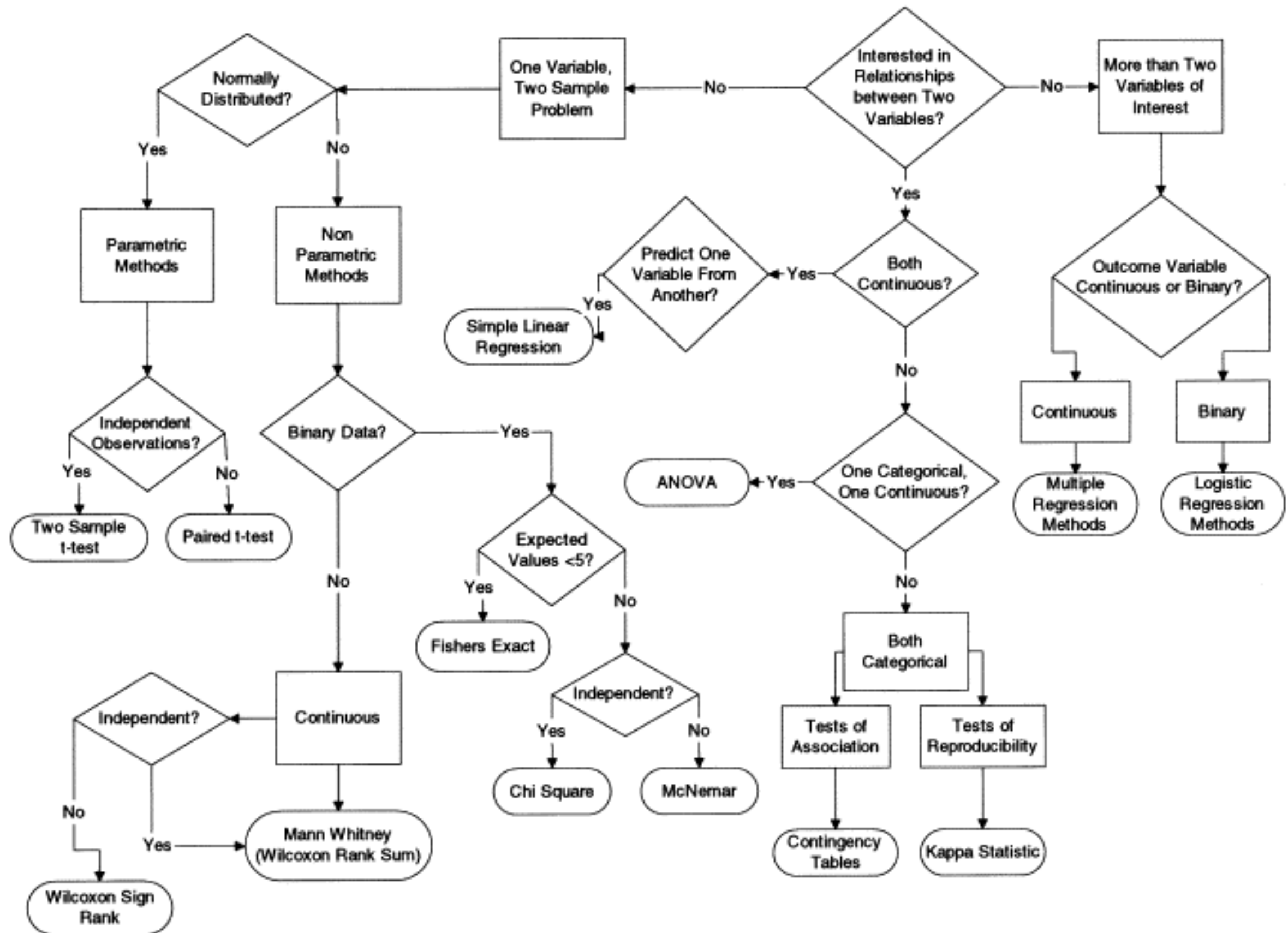
$$L(\text{mean} = 1, \text{sd} = 1 \mid x = 0) = 0.2419$$

Plan for today

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

Cookbook vs. Model Comparison

The cookbook approach

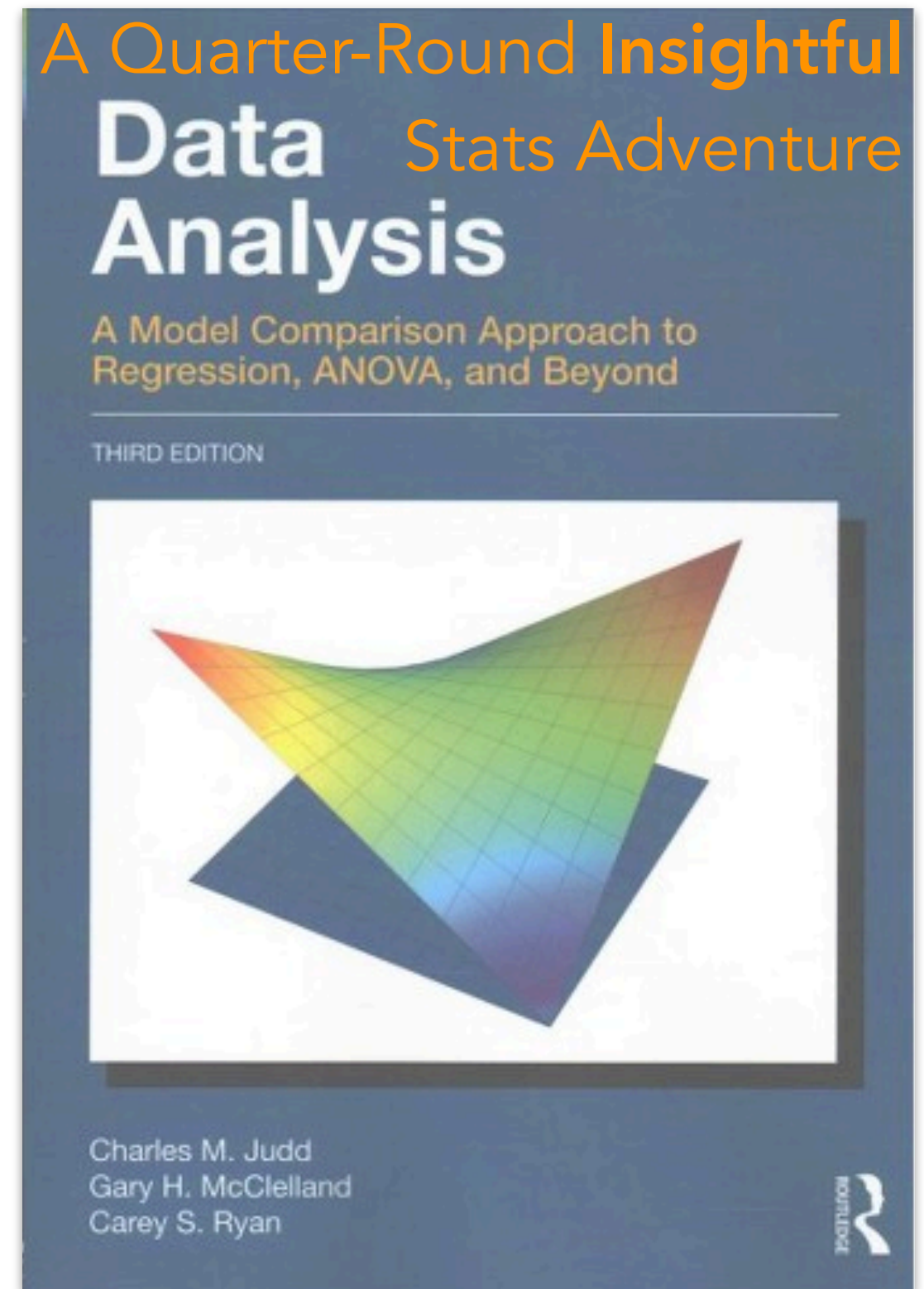


The cookbook approach



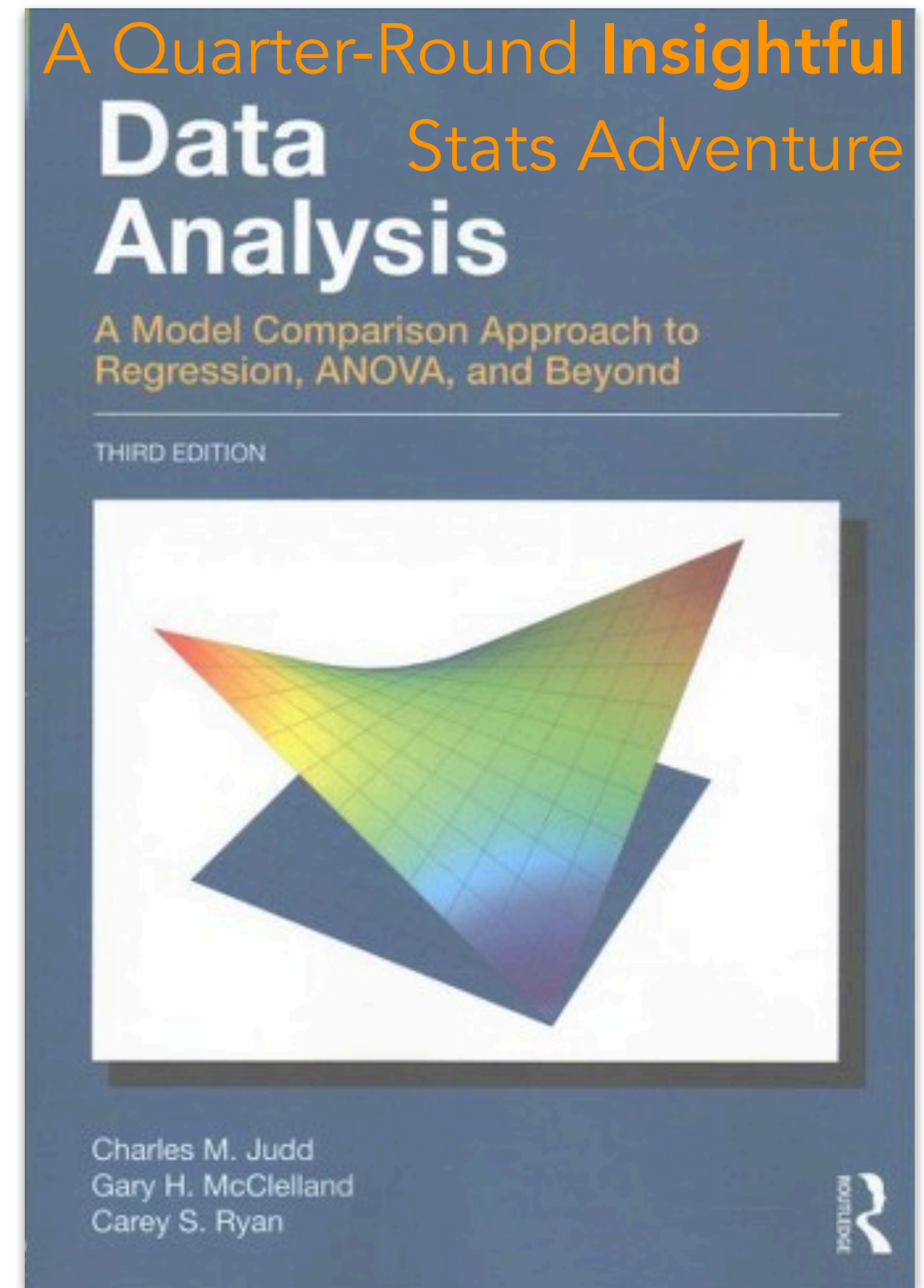
- many statistics textbooks are organized in this way
- works reasonably well if what we want to cook is in the book
- leaves us with no idea what to do if

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

$$\text{Data} = \text{Model} + \text{Error}$$

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?

Thank you!