The Normal Distribution

Corinne Riddell

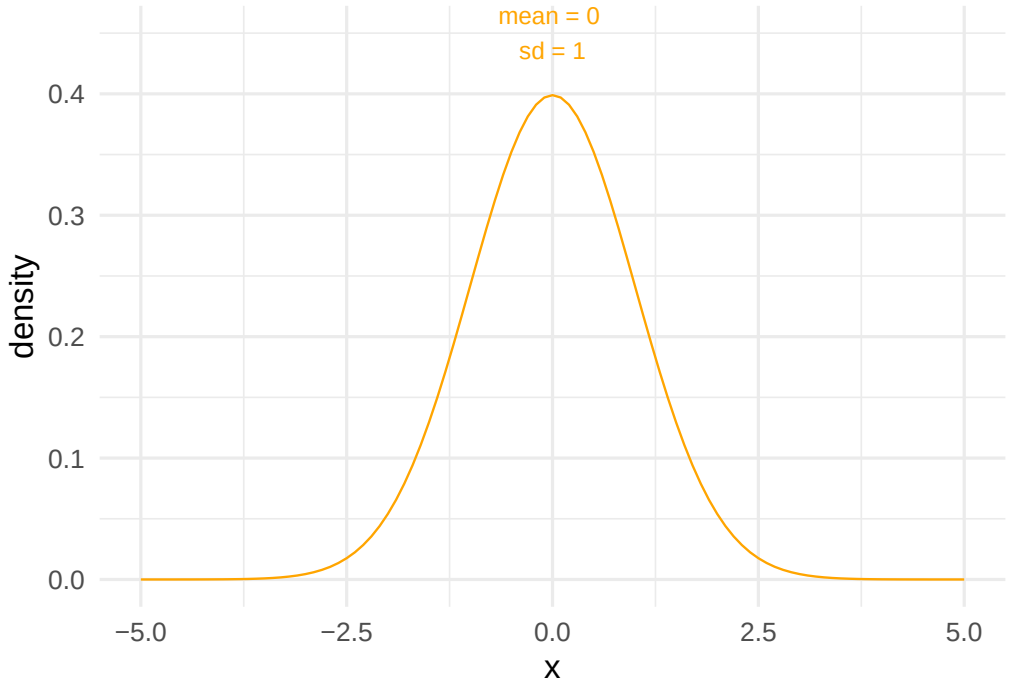
September 30, 2020

Learning objectives for today

- Learn about the Normal distribution centered at μ with a standard deviation of σ
- Learn about the standard Normal distribution where $\mu = 0$ and $\sigma = 1$ and compute z-scores
- Calculate cumulative probabilities below or above a given value for any specified Normal distribution using R
- Calculate the quantile for a specified cumulative probability for any specified Normal distribution using R
- Perform simple calculations by hand (using the 68-95-99.7 rule)
- Learn about Q-Q plots and how to use them to assess whether a variable is Normally distributed

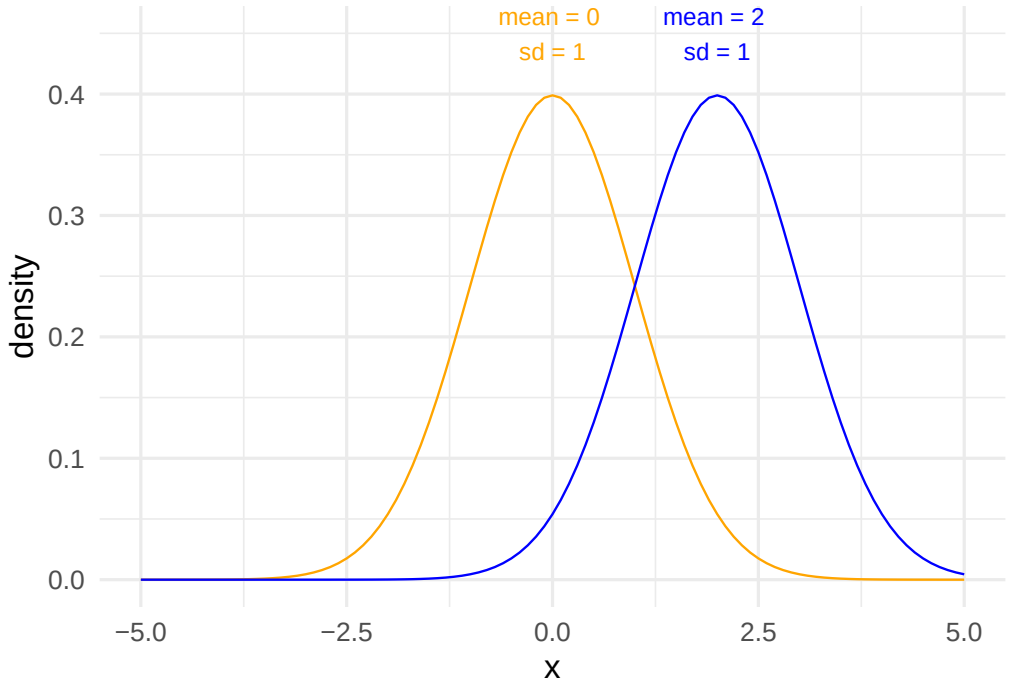
The Normal Distribution

- Here is the Normal distribution with mean of 0 (μ) and standard deviation of 1 (σ).
- It is:
  - symmetric
  - centered at μ



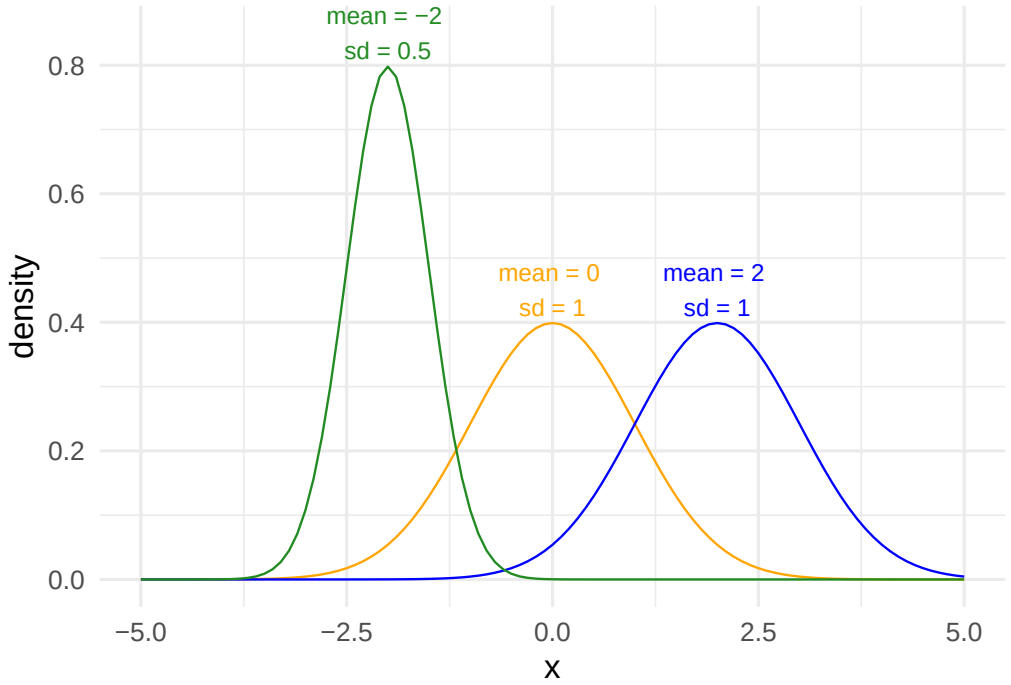
The Normal Distribution

- Let's add another Normal distribution, this one centered at 2, with the same standard deviation



The Normal Distribution

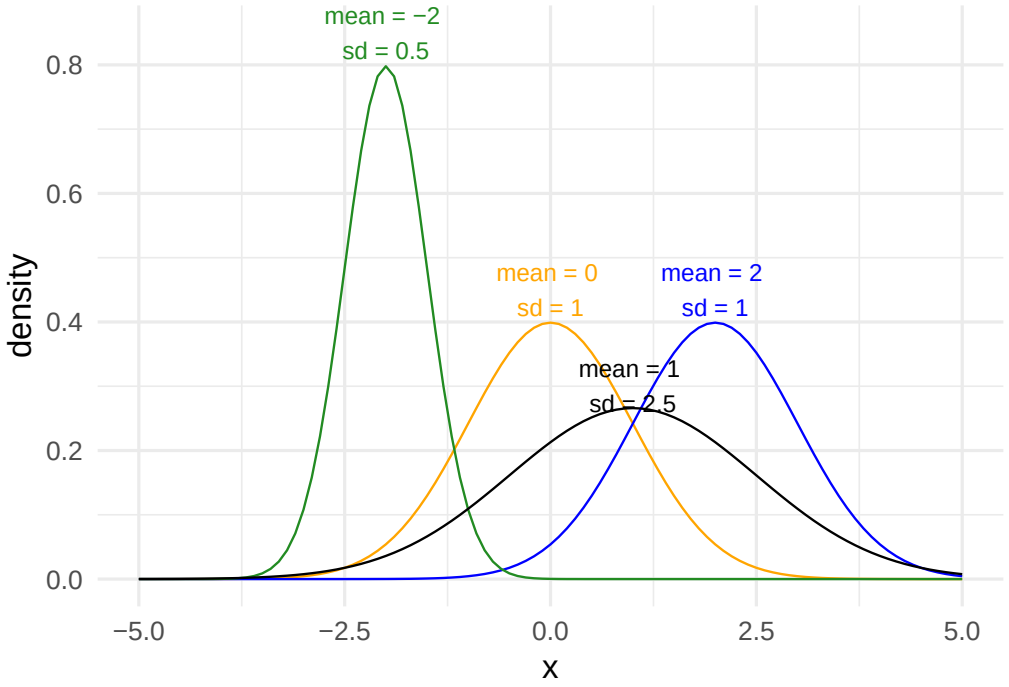
- Let's add a third Normal distribution, this one centered at -2, with a standard deviation of 0.5
- Notice how the distribution is narrowed (i.e., the spread is reduced)
- Why is the distribution "taller"?



The Normal Distribution

- Can you guess what a Normal distribution with $\mu = 1$ and $\sigma = 1.5$ would look like compared to the others?

The Normal Distribution



Properties of the Normal distribution

- the mean μ can be any value, positive or negative
- the standard deviation σ must be a positive number
- the mean is equal to the median (both $= \mu$)
- the standard deviation captures the spread of the distribution
- the area under the Normal distribution is equal to 1 (i.e., it is a density function)
- a Normal distribution is completely determined by its μ and σ

The 68-95-99.7 rule for all Normal distributions

- Approximately 68% of the data fall within one standard deviation of the mean
- Approximately 95% of the data fall within two standard deviations of the mean
- Approximately 99.7% of the data fall within three standard deviations of the mean

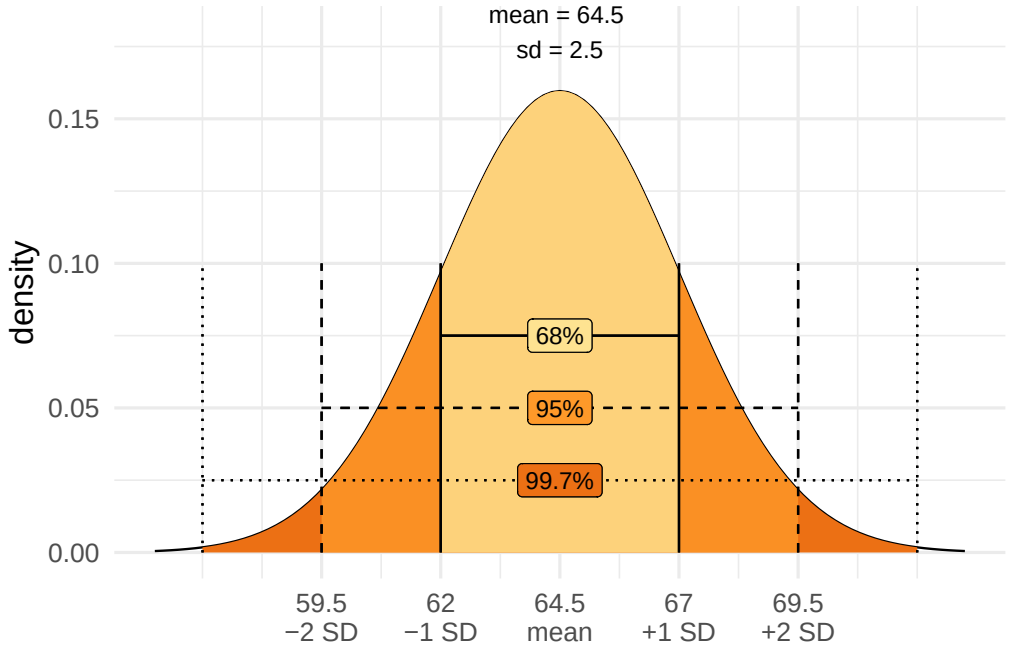
Written probabilistically:

- $P(\mu - \sigma < X < \mu + \sigma) \approx 68\%$
- $P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 95\%$
- $P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 99.7\%$

Calculations using the 68-95-99.7 rule

Example 11.1 from Baldi & Moore on the heights of young women. The distribution of heights of young women is approximately Normal, with mean $\mu = 64.5$ inches and standard deviation $\sigma = 2.5$ inches.

- i.e., $H \sim N(64.5, 2.5)$, where H is defined as the height of a young woman

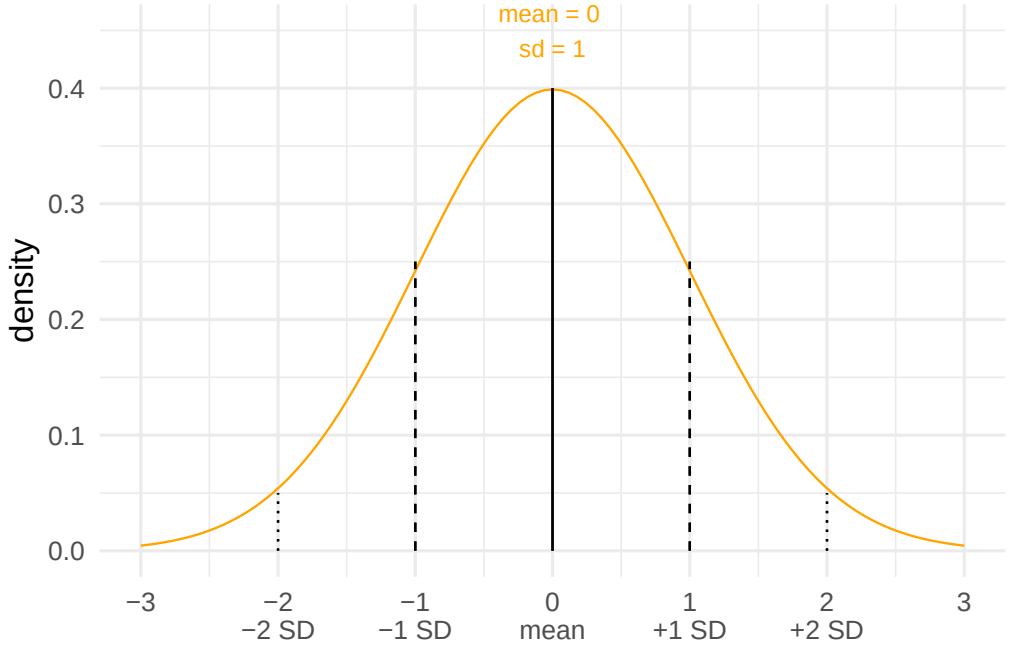


Calculations using the 68-95-99.7 rule

- What calculations could you do with these data alone?
- $P(62 < H < 67) =?$
- $P(H > 62) =?$

The standard Normal distribution

- The standard Normal distribution is the Normal distribution with $\mu = 0$ and $\sigma = 1$.
- We write: $N(0, 1)$ to denote this distribution
- $X \sim N(0, 1)$, implies that the random variable X is Normally distributed.



Standardizing Normally distributed data

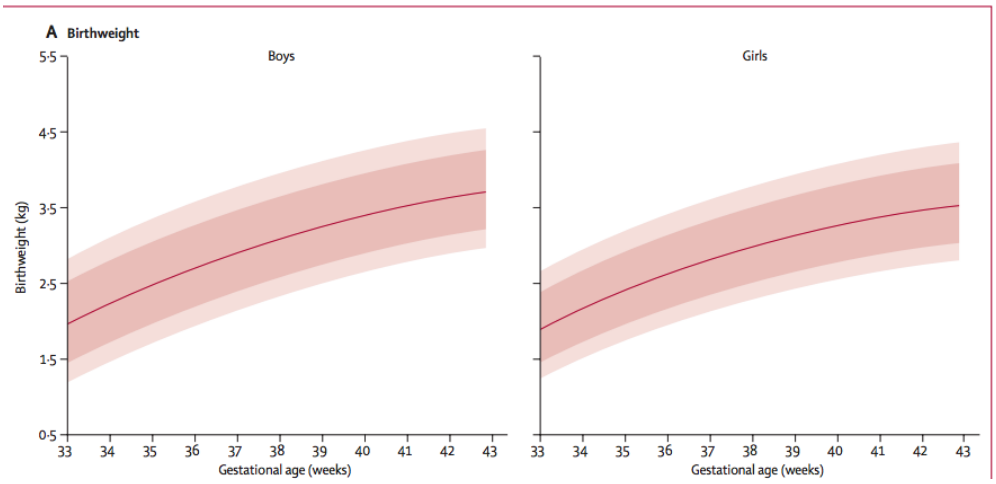
- Any random variable that follows a Normal distribution can be standardized
- If x is an observation from a distribution that has a mean μ and a standard deviation σ, the standardized value of x is:

$$z = \frac{x - \mu}{\sigma}$$

- A standardized value is often called a **z-score**
- Interpretation: z is the number of standard deviations that x is above or below the mean of the data.
- We standardize values so that we can have this interpretation, which is agnostic to the underlying mean, standard deviation, and units of measure. Standardizing Normally-distributed data is a quick way to determine if a specific value is much higher or lower than the average value.

Standardizing Normally distributed data

Standardizing Normally distributed data



Reference

Standardizing Normally distributed data

The International Newborn Standards

UNIVERSITY OF OXFORD

Birth weight (Boys)

INTERGROWTH-21st

Gestational age (weeks+days)	z scores						
	-3	-2	-1	0	1	2	3
33+0	0.63	1.13	1.55	1.95	2.39	2.88	3.47
33+1	0.67	1.17	1.59	1.99	2.43	2.92	3.51
33+2	0.71	1.21	1.63	2.03	2.47	2.96	3.55
33+3	0.75	1.25	1.67	2.07	2.50	2.99	3.59
33+4	0.79	1.29	1.71	2.11	2.54	3.03	3.62
33+5	0.83	1.33	1.75	2.15	2.58	3.07	3.66
33+6	0.87	1.37	1.79	2.18	2.62	3.11	3.70
34+0	0.91	1.40	1.82	2.22	2.65	3.14	3.73
34+1	0.95	1.44	1.86	2.26	2.69	3.18	3.77
34+2	0.98	1.48	1.90	2.29	2.73	3.21	3.80
34+3	1.02	1.51	1.93	2.33	2.76	3.25	3.84
34+4	1.05	1.55	1.97	2.36	2.80	3.28	3.87
34+5	1.09	1.58	2.00	2.40	2.83	3.32	3.91
34+6	1.12	1.62	2.04	2.43	2.86	3.35	3.94
35+0	1.16	1.65	2.07	2.47	2.90	3.38	3.97

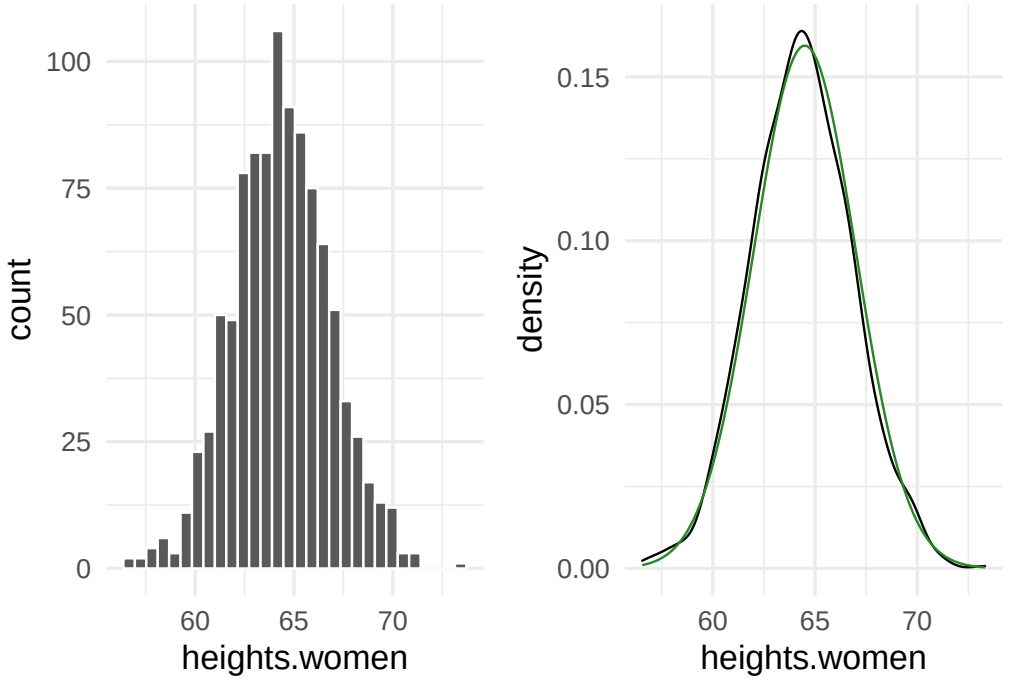
- Birthweight z-scores for boys
- How does this relate to what you see on the previous slide?

Simulating Normally distributed data in R

Suppose that we measured 1000 heights for young women:

```
#students, rnorm() is important to know!
heights.women <- rnorm(n = 1000, mean = 64.5, sd = 2.5)
heights.women <- data.frame(heights.women)
```

We can plot the histogram of the heights, and see that they roughly follow from a Normal distribution. The green curve is a Normal distribution, and the black curve is the density plot based on the actual data:



Standardizing Normally distributed data in R

To standardize these data, we can apply the formula to compute the z-score:

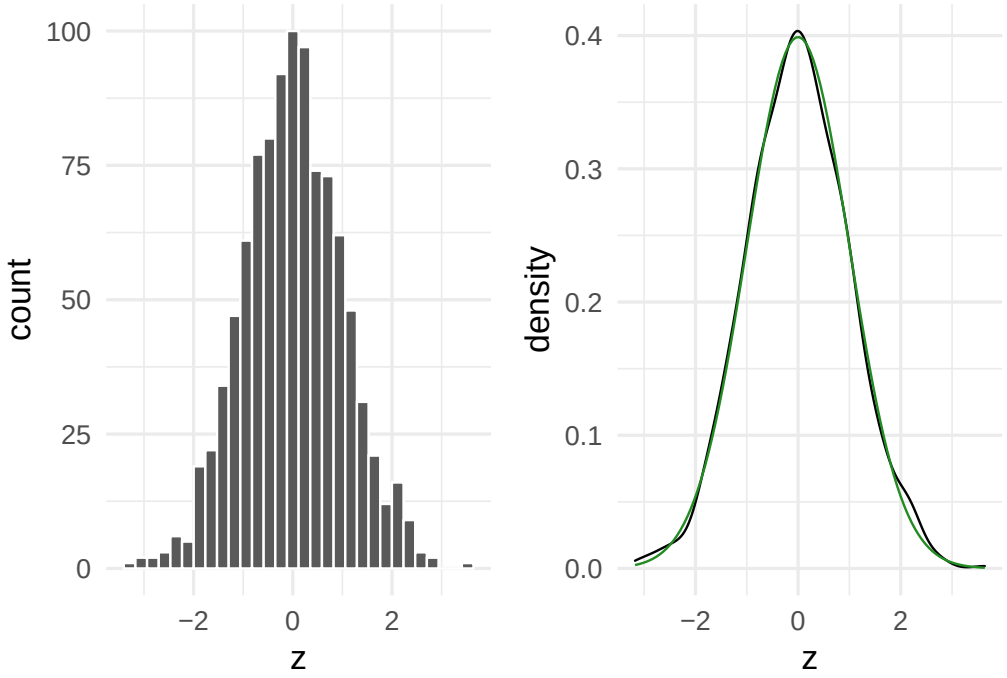
```
heights.women <- heights.women %>% mutate(mean = mean(heights.women),
                                          sd = sd(heights.women),
                                          z = (heights.women - mean)/sd)

head(heights.women)
```

```
##   heights.women     mean       sd          z
## 1      68.09674 64.36835 2.460119  1.5155316
## 2      69.86720 64.36835 2.460119  2.2351975
## 3      61.07293 64.36835 2.460119 -1.3395357
## 4      59.80148 64.36835 2.460119 -1.8563603
## 5      60.71829 64.36835 2.460119 -1.4836926
## 6      63.70202 64.36835 2.460119 -0.2708544
```

What would the distribution of the standardized heights look like?

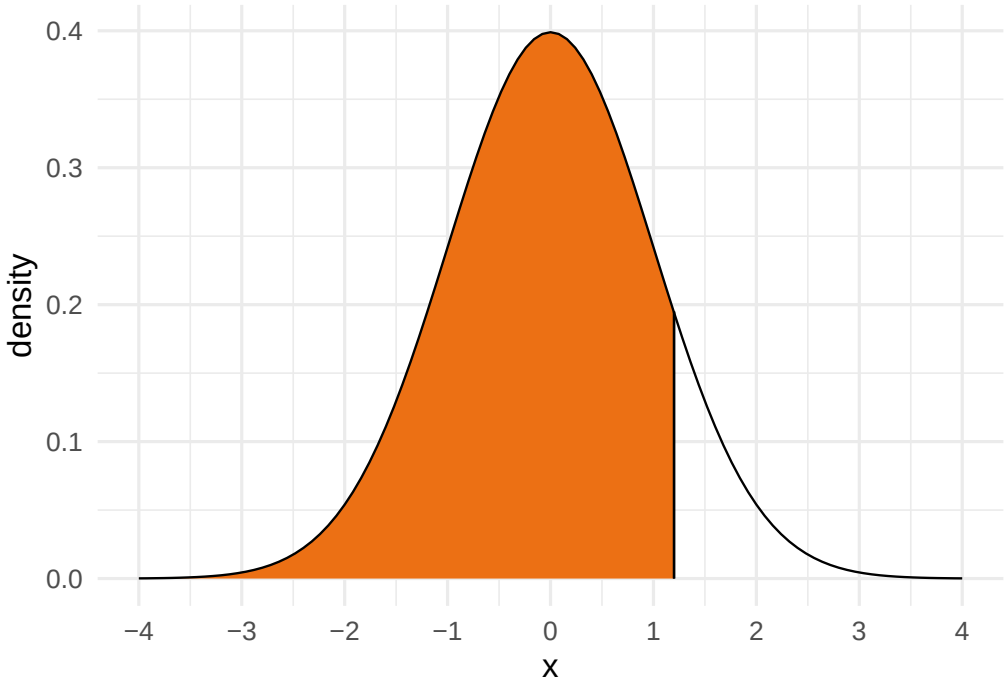
Standardizing Normally distributed data in R

How are these plots different from the previous ones? Hint: look at the x axis.



Finding Normal probabilities

- A cumulative probability for a value x in a distribution is the proportion of observations in the distribution that lie at or below x.
- Here is the cumulative probability for x=1.2



Finding Normal probabilities

- Recall that 100% of the sample space for the random variable X lies under its probability density function.
- What is the amount of the area that is below x = 1.2?

- To answer this question we use the `pnorm()` function. (Think: the **p** in `pnorm` stands for probability):

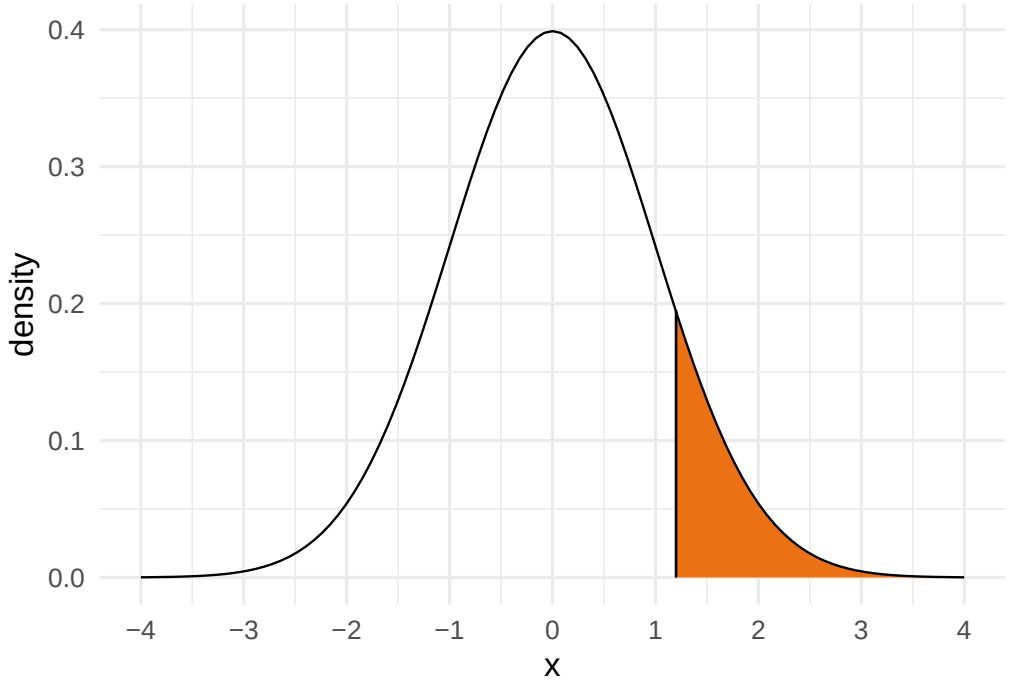
```
pnorm(q = 1.2, mean = 0, sd = 1)
```

```
## [1] 0.8849303
```

This says that approximately 88% of the probability lies below 1.2.

Finding Normal probabilities

What if we wanted the reverse: P(x>1.2)?




```
1 - pnorm(q = 1.2, mean = 0, sd = 1)
```

```
## [1] 0.1150697
```

Alternatively:

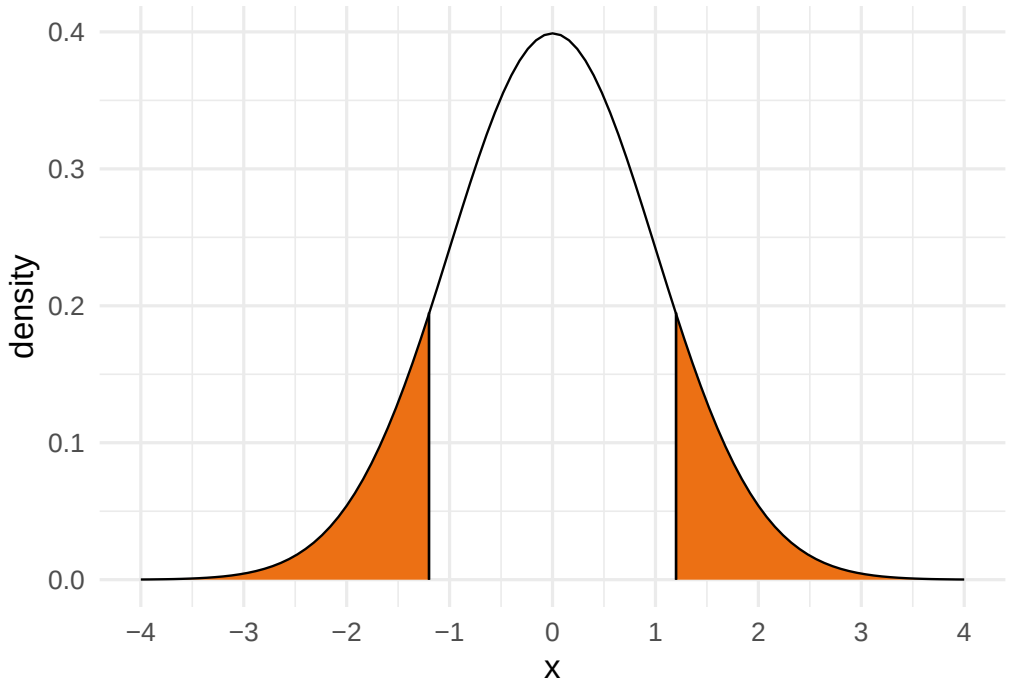
```
 pnorm(q = 1.2, mean = 0, sd = 1, lower.tail = F)
```

```
## [1] 0.1150697
```

So, 11.51% of the data is above x=1.2.

Finding Normal probabilities

What if we wanted two "tail" probabilities?: $P(x < -1.2 \text{ or } x > 1.2)$?



Finding Normal probabilities

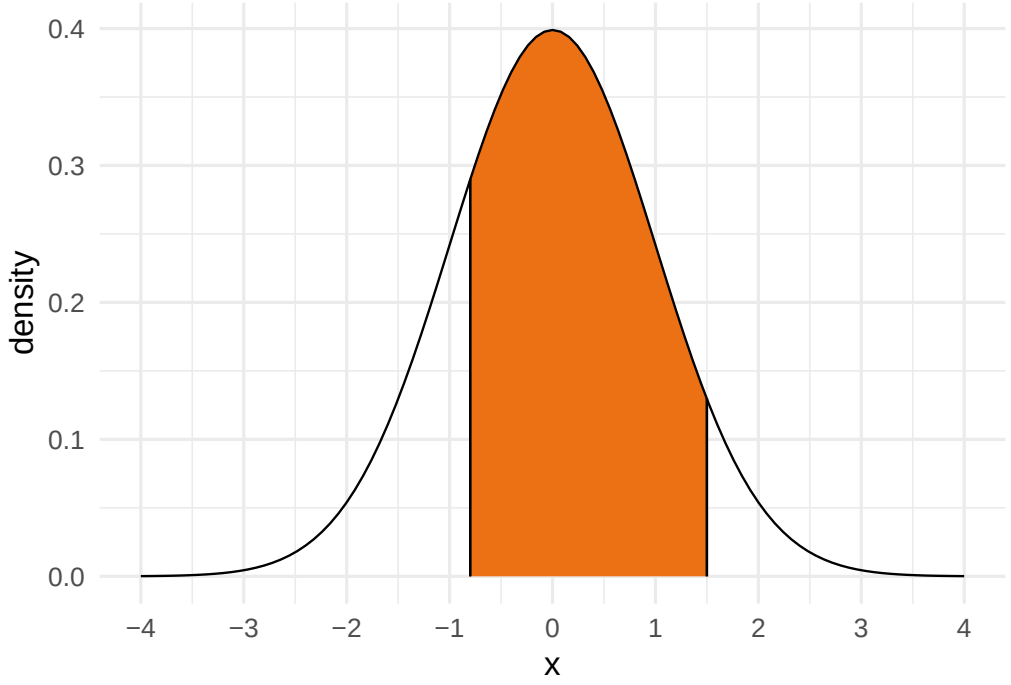
The trick: find one of the tails and then double the area because the distribution is symmetric:

```
pnorm(q = -1.2, mean = 0, sd = 1)*2
```

```
## [1] 0.2301393
```

Finding Normal probabilities

What if we wanted a range in the middle?: $P(-0.8 < x < 1.5)$?



Finding Normal probabilities

```
# step 1: calculate the probability *below* the upper bound (x=1.5)
pnorm(q = 1.5, mean = 0, sd = 1)
```

```
## [1] 0.9331928
```

```
# step 2: calculate the probability *below* the lower bound (x = -0.8)
pnorm(q = -0.8, mean = 0, sd = 1)
```

```
## [1] 0.2118554
```

```
# step 3: take the difference between these probabilities to get what's left in
# the middle
pnorm(q = 1.5, mean = 0, sd = 1) - pnorm(q = -0.8, mean = 0, sd = 1)
```

```
## [1] 0.7213374
```

Thus, 72.13% of the data is in the range -0.8 $<$ x $<$ 1.5.

Your turn

To diagnose osteoporosis, bone mineral density is measured. The WHO criterion for osteoporosis is a BMD score below -2.5. Women in their 70s have a much lower BMD than younger women. Their BMD ~ N(-2, 1). What proportion of these women have a BMD below the WHO cutoff?

Hint: you do not need to find a z-score!

```
#to fill in during class
```

Finding Normal percentiles

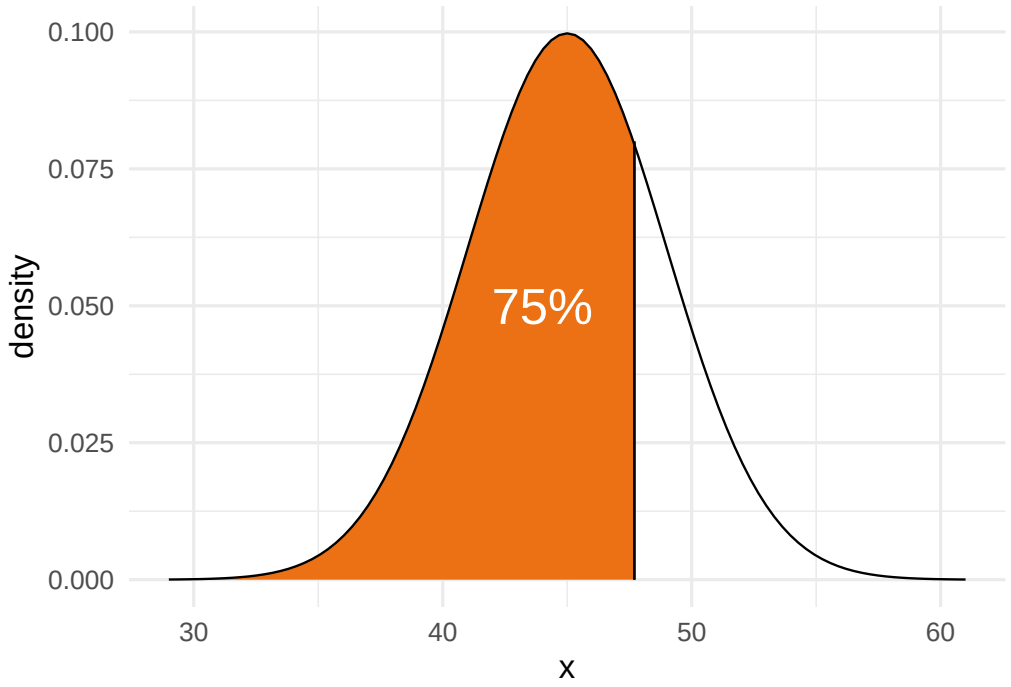
Recap: so far, we have calculated the *probability* using `pnorm()` given specific values for x.

Sometimes we want to go in the opposite direction: We might be given the probability within some range and tasked with finding the corresponding x-values.

Finding Normal percentiles

Example: The hatching weights of commercial chickens can be modeled accurately using a Normal distribution with mean $\mu = 45$ grams and standard deviation $\sigma = 4$ grams. What is the third quartile of the distribution of hatching weights?

That is, what is the x such that 75% of the probability is below it?



Finding Normal percentiles using the `qnorm()` function

Example: The hatching weights of commercial chickens can be modeled accurately using a Normal distribution with mean $\mu = 45$ grams and standard deviation $\sigma = 4$ grams. What is the third quartile of the distribution of hatching weights?

```
qnorm(p = 0.75, mean = 45, sd = 4)
```

```
## [1] 47.69796
```

Thus, 75% of the data is below 47.7 for this distribution.

Using the standard Normal table

- Before we had easy access to computers and software people would use printed out tables to compute probabilities
- We can ignore this section of the textbook because we will always have R to do the calculations for us

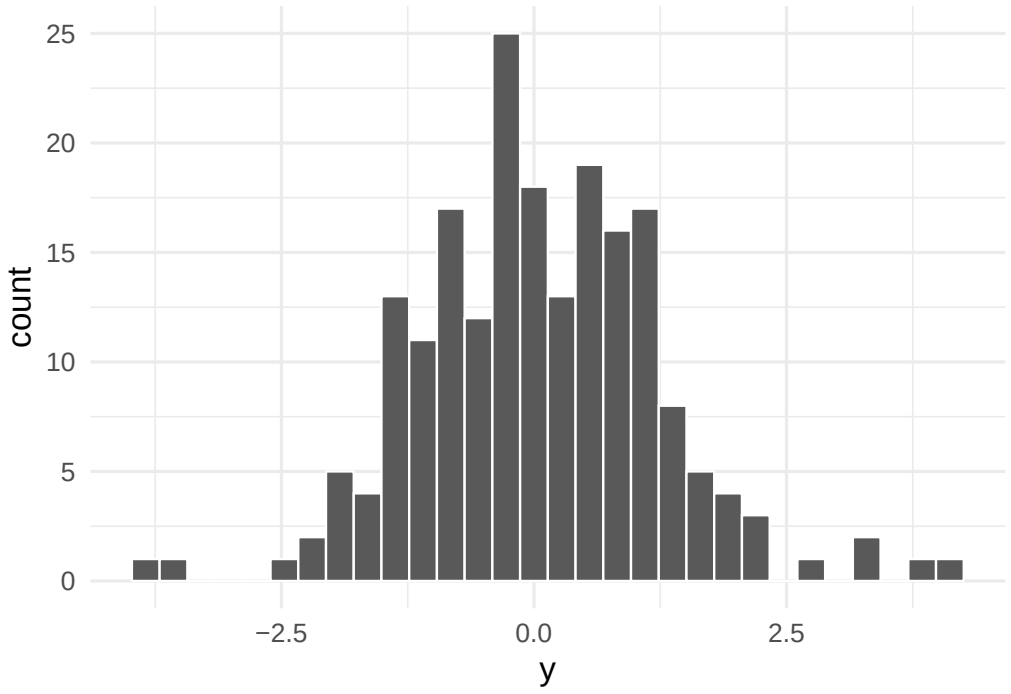
The Normal quantile plot (a.k.a the Q-Q plot)

- The purpose of making a Q-Q plot is to examine the Normality of a distribution of a variable.
- If you want to know whether a variable is Normally distributed you could examine its histogram to see if it is unimodal and symmetric. However, it is still sometimes hard to say if it is truly Normal. To do so we use a Q-Q plot.

Are these data Normally distributed?

- The data is unimodal and symmetric, but is its distribution Normal?

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Making a Q-Q plot by hand

These are the steps to make a Q-Q plot by hand. I want you to understand these steps, because they will help you interpret the QQ plot that we will use R to make (thanks, ggplot!)

1. First, arrange the variable in ascending order. Calculate the percentile for each measurement. For example if you had ten measurements in ascending order, the first measurement is at the 10th percentile because 10% of the data is at or below its value. The second measurement is at the 20% because 20% of the data is at or below its value, and so forth.
2. Then, for each of the percentiles you calculated, use that percentile to calculate the corresponding x-value of the Normal distribution that occurs at that percentile. For example, x = -1 it the 16th percentile of the N(0, 1) distribution. (We can see this using `pnorm(-1, 0, 1)`.)
3. Make a scatter plot of the calculated x-values (from step 1) on the x-axis and the values from step 2 on the y axis.
4. The closer the data is to a straight line, the more closely it approximates a Normal distribution.

Making a Q-Q plot by hand

```
#calculate the percentile by hand:
example_data <- example_data %>%
  arrange(y) %>%
  mutate(quantile = row_number()/n())

# then calculate the x-value at each percentile from the previous step
# this x-value can be called a z-score because it is from the standard Normal distribution
example_data <- example_data %>%
  mutate(z_score = qnorm(quantile, mean = 0, sd = 1))

head(example_data)
```

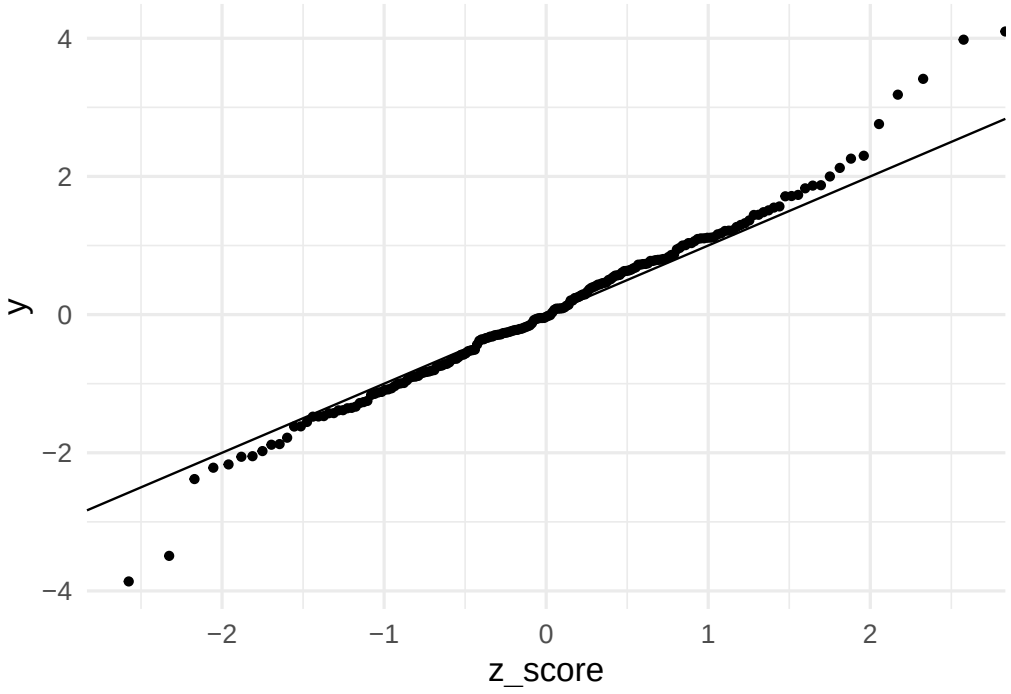
```
##           y quantile   z_score
## 1 -3.862491    0.005 -2.575829
```

```
## 2 -3.492259    0.010 -2.326348
## 3 -2.380508    0.015 -2.170090
## 4 -2.216277    0.020 -2.053749
## 5 -2.168630    0.025 -1.959964
## 6 -2.057844    0.030 -1.880794
```

Look at the Q-Q plot for these data

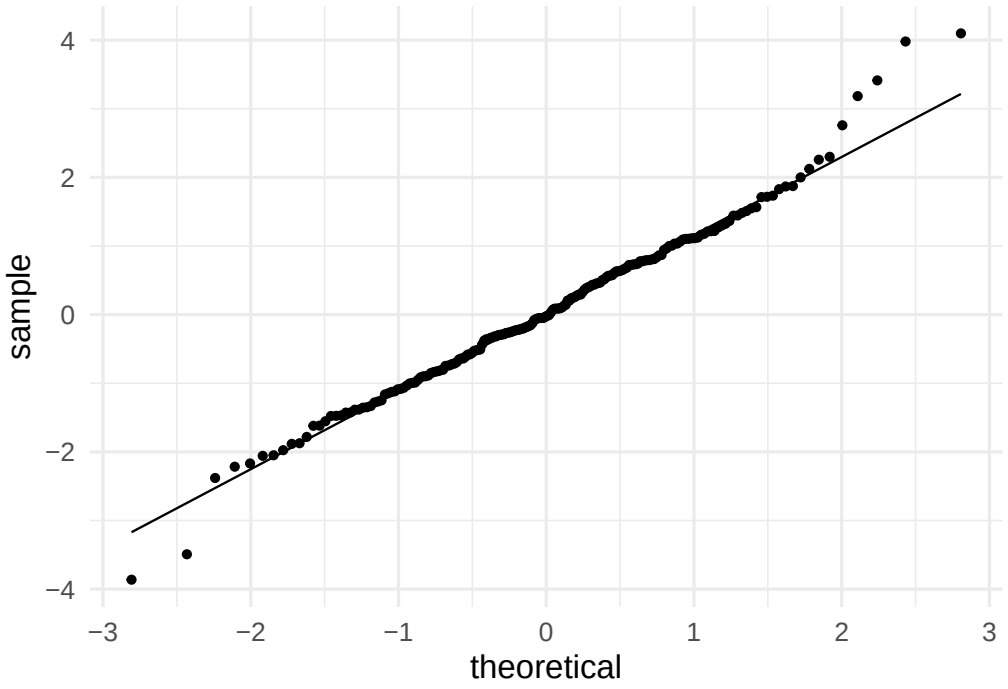
- Notice that the data overlays the 45 degree line in the middle but not in the tails of the distribution. This sort of pattern shows that these data are "wider" (have larger standard deviation) that a Normally distributed variable.

```
ggplot(example_data, aes(x = z_score, y = y)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1) +
  theme_minimal(base_size = 15)
```



Easy way to make a qqplot() where R does all the calculating for you

```
ggplot(example_data, aes(sample = y)) +
  stat_qq() +
  stat_qq_line() +
  theme_minimal(base_size = 15)
```



Another example

Recall the seed data:

```
library(readr)
seed_data <- read_csv("./data/Ch04_seed-data")
```

```
## Parsed with column specification:
## cols(
##   species = col_character(),
##   seed_count = col_double(),
##   seed_weight = col_double()
## )
```

```
head(seed_data)
```

```
## # A tibble: 6 x 3
##   species         seed_count seed_weight
##   <chr>                <dbl>       <dbl>
## 1 Paper birch          27239         0.6
## 2 Yellow birch         12158         1.6
## 3 White spruce          7202         2
## 4 Engelman spruce       3671         3.3
## 5 Red spruce            5051         3.4
## 6 Tulip tree           13509         9.1
```

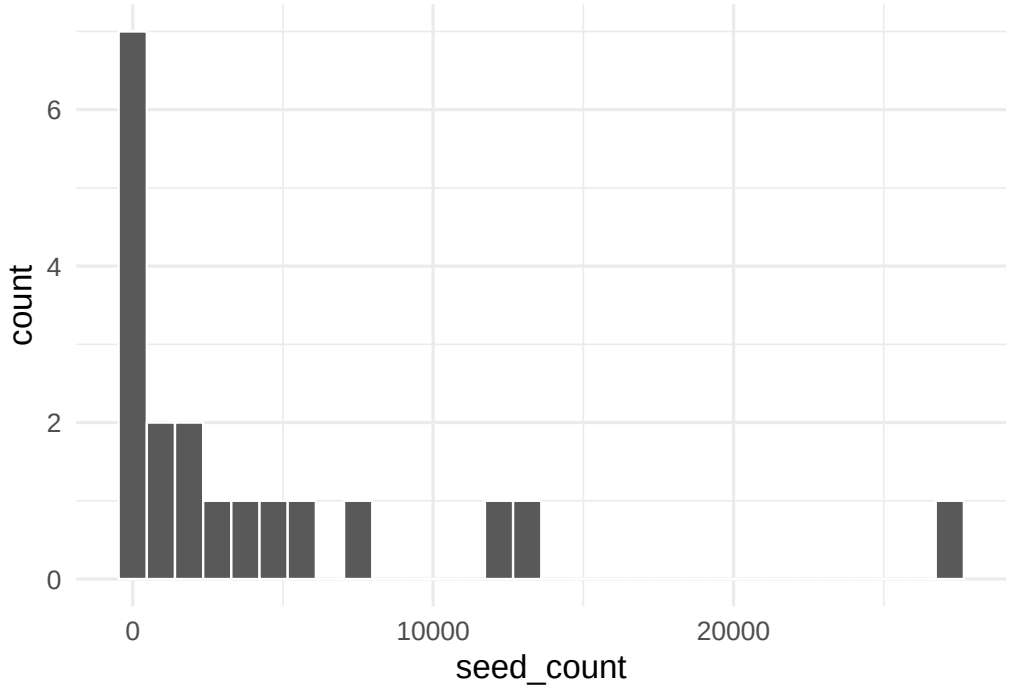
Is the distribution of `seed_count` Normal?

Another example

Check out its distribution. It definitely does not look normal:

```
ggplot(seed_data, aes(x = seed_count)) +
  geom_histogram(col = "white") +
  theme_minimal(base_size = 15)
```

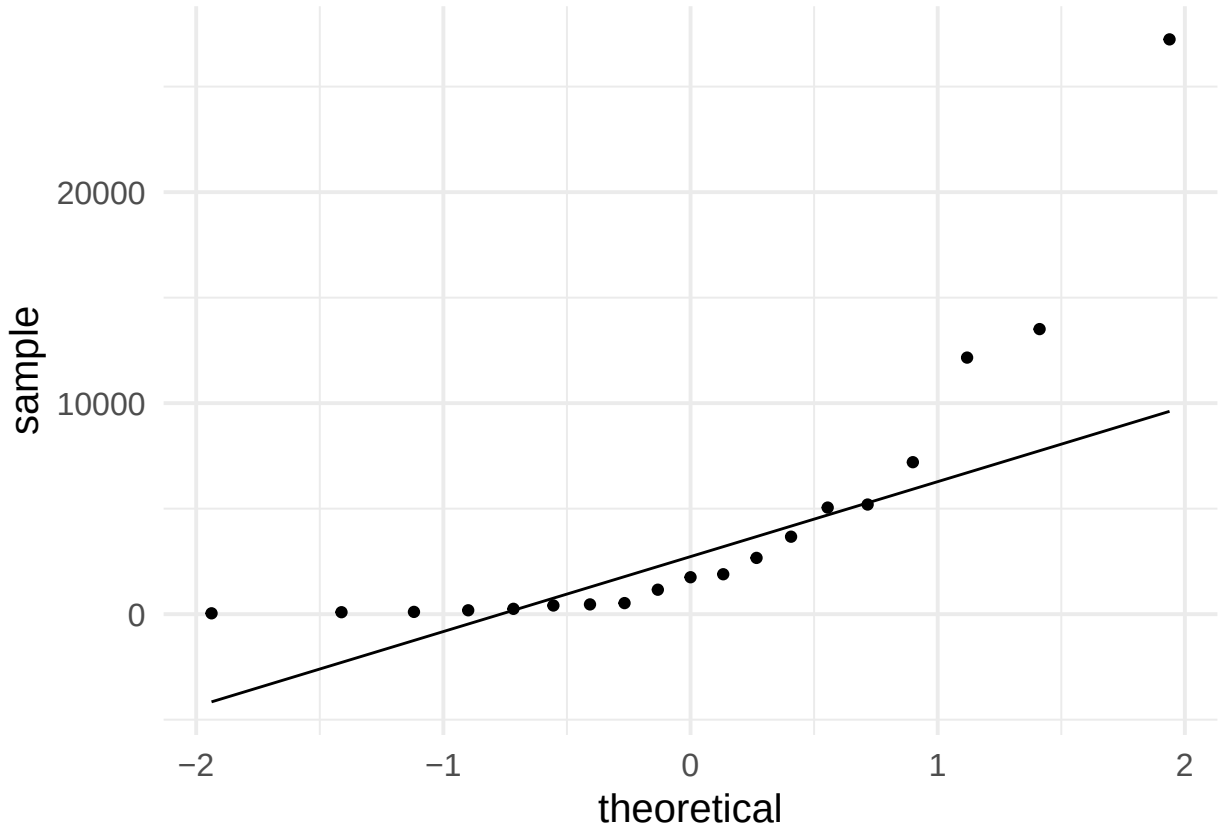
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Another example

And look at its Q-Q plot. Does the data appear to follow a Normal distribution?

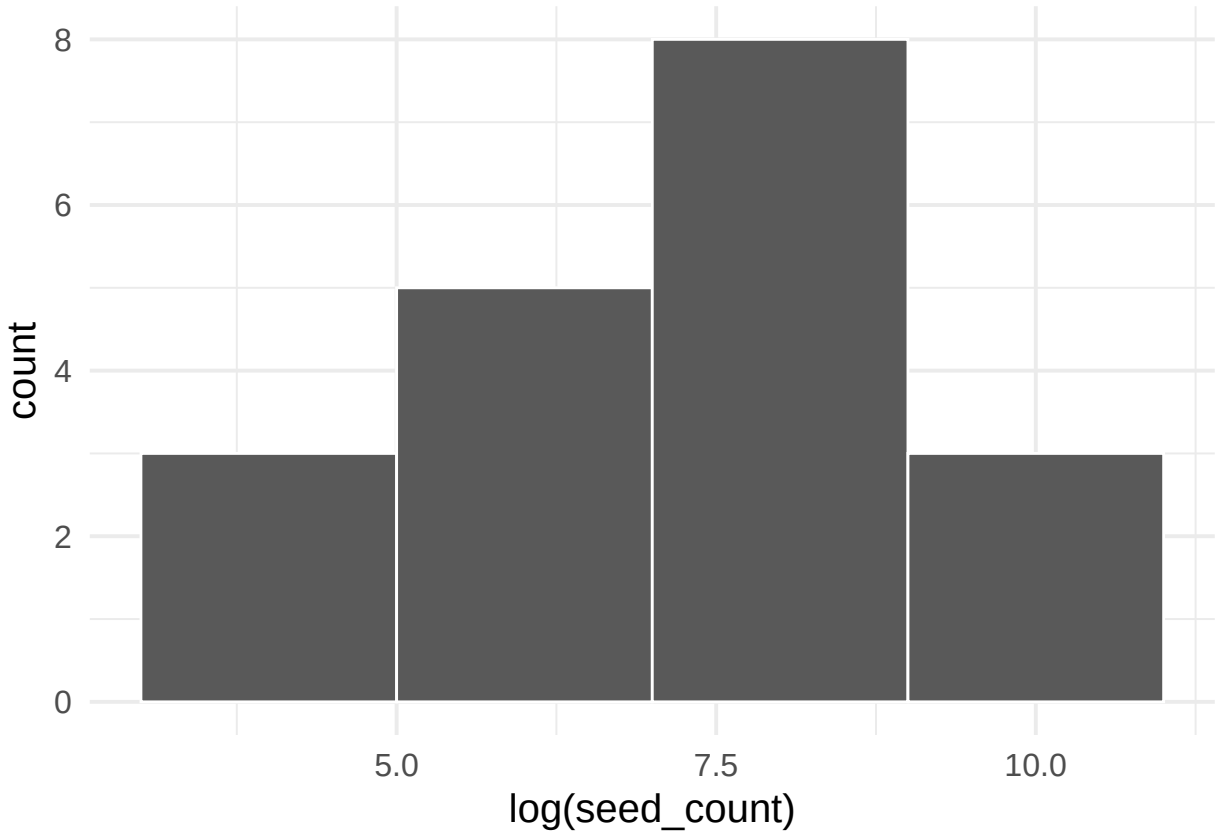
```
ggplot(seed_data, aes(sample = seed_count)) +
  stat_qq() + stat_qq_line() +
  theme_minimal(base_size = 15)
```



Another example (logged)

You might remember that we took the log of seed_count before we used it in regression.

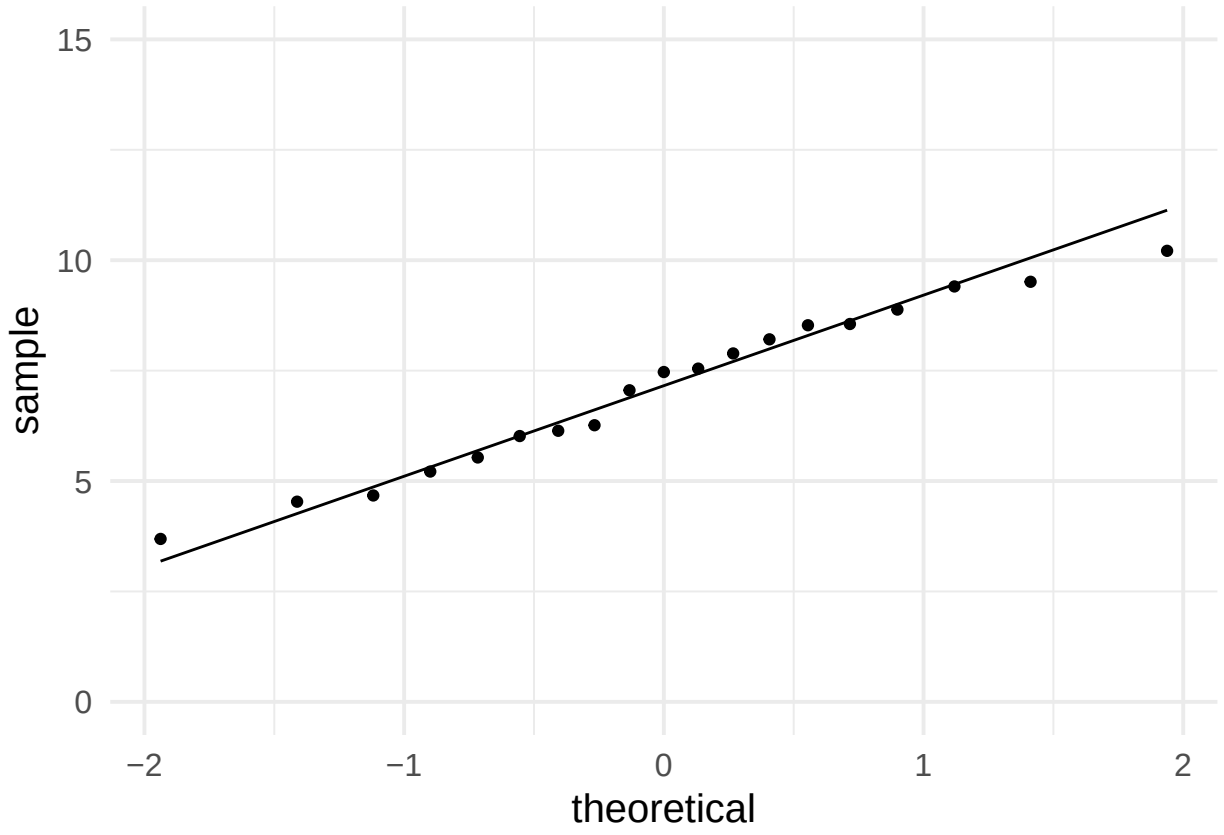
```
ggplot(seed_data, aes(x = log(seed_count))) +
  geom_histogram(col = "white", binwidth = 2) +
  theme_minimal(base_size = 15)
```



Another example (logged)

How does the Q-Q plot look for the logged variable?

```
ggplot(seed_data, aes(sample = log(seed_count))) +
  stat_qq() + stat_qq_line() +
  theme_minimal(base_size = 15) + scale_y_continuous(limits = c(0, 15))
```



Q-Q plot summary

- Review the Q-Q plots from the book on page 290-292 of B&M Edition 4
- Try and gain intuition about when a variable does not appear to fit a Normal distribution
  - Was the distribution skewed?
  - Was there an outlier?
- For each scenario how do these deviations from Normality affect the QQ plot?

Recap of functions used

- `rnorm(n = 100, mean = 2, sd = 0.4)`, to generate Normally distributed data from the specified distribution
- `pnorm(q = 1.2, mean = 0, sd = 2)`, to calculate the cumulative probability below a given value
- `qnorm(p = 0.75, mean = 0, sd = 1)` to calculate the x-value for which some percent of the data lies below it
- `stat_qq()` and `stat_qq_line()` to make a Q-Q plot. Notice also that `aes(sample = var1)` is needed