The Normal Distribution continued

Corinne Riddell

October 1, 2021

Learning objectives for today

- Calculate the quantile for a specified cumulative probability for any specified Normal distribution using R
- Learn about Q-Q plots and how to use them to assess whether a variable is Normally distributed

Finding Normal percentiles

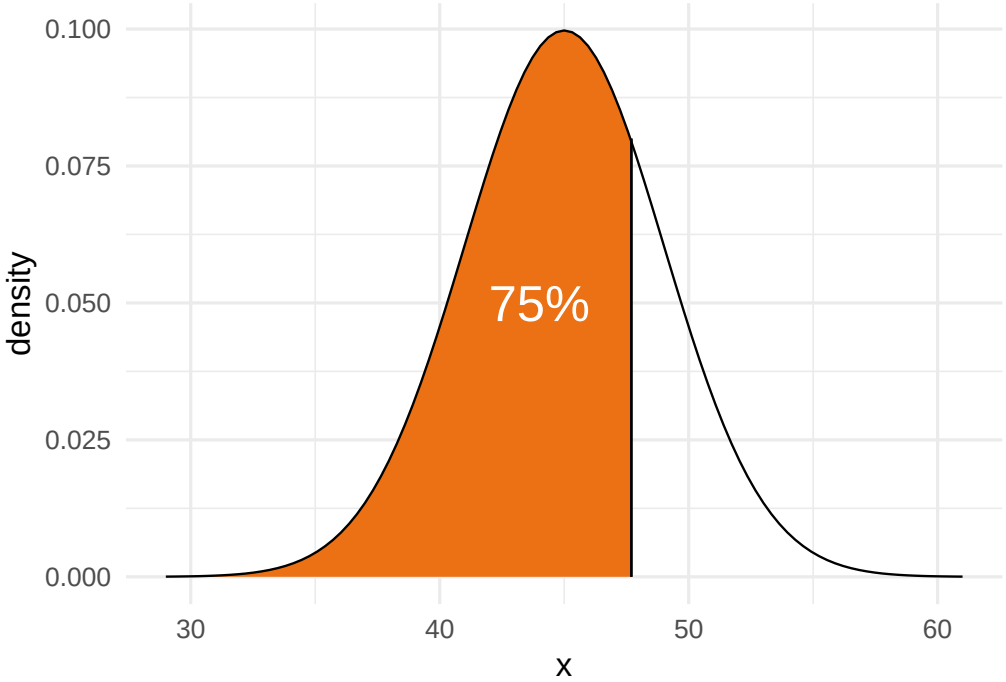
Recap: Last class, we have calculated the *probability* using `pnorm()` given specific values for x.

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

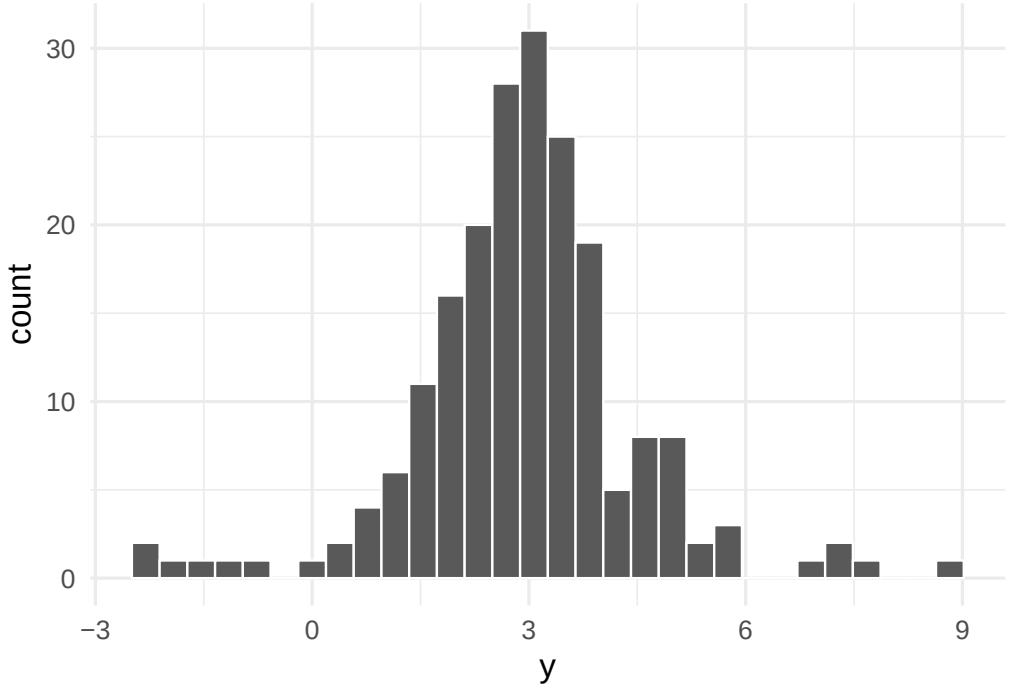
The Normal quantile plot (a.k.a the Q-Q plot)

- 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?

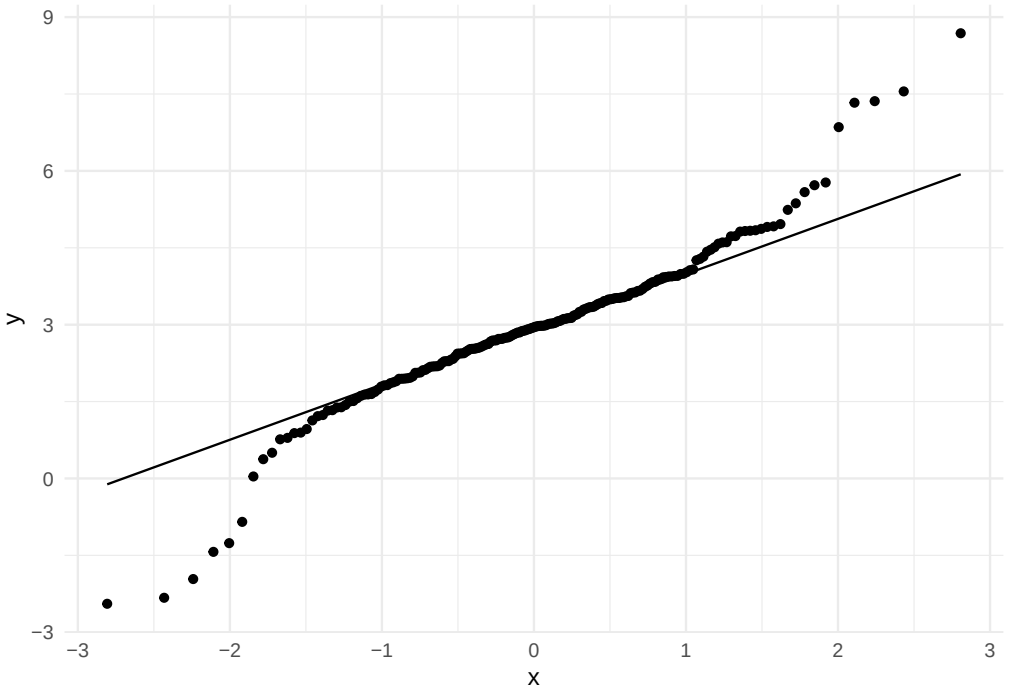
- The data is unimodal and symmetric, but is its distribution Normal?

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



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()
```



Another example

Recall the seed data:

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

```
##
## -- 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
```

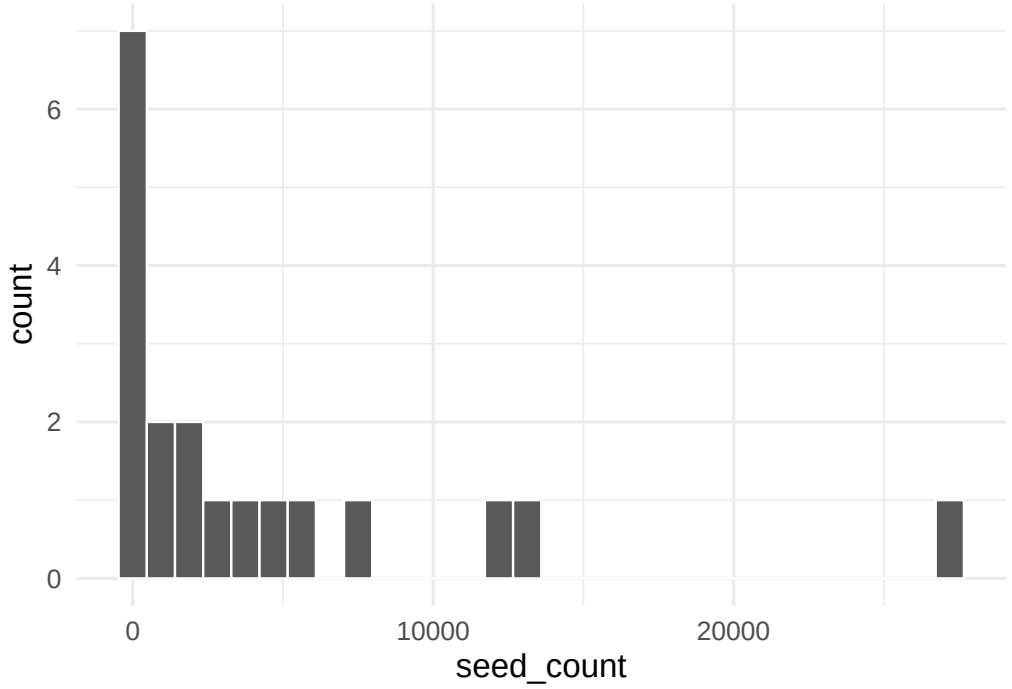
Is the distribution of `seed_count` Normal?

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)
```

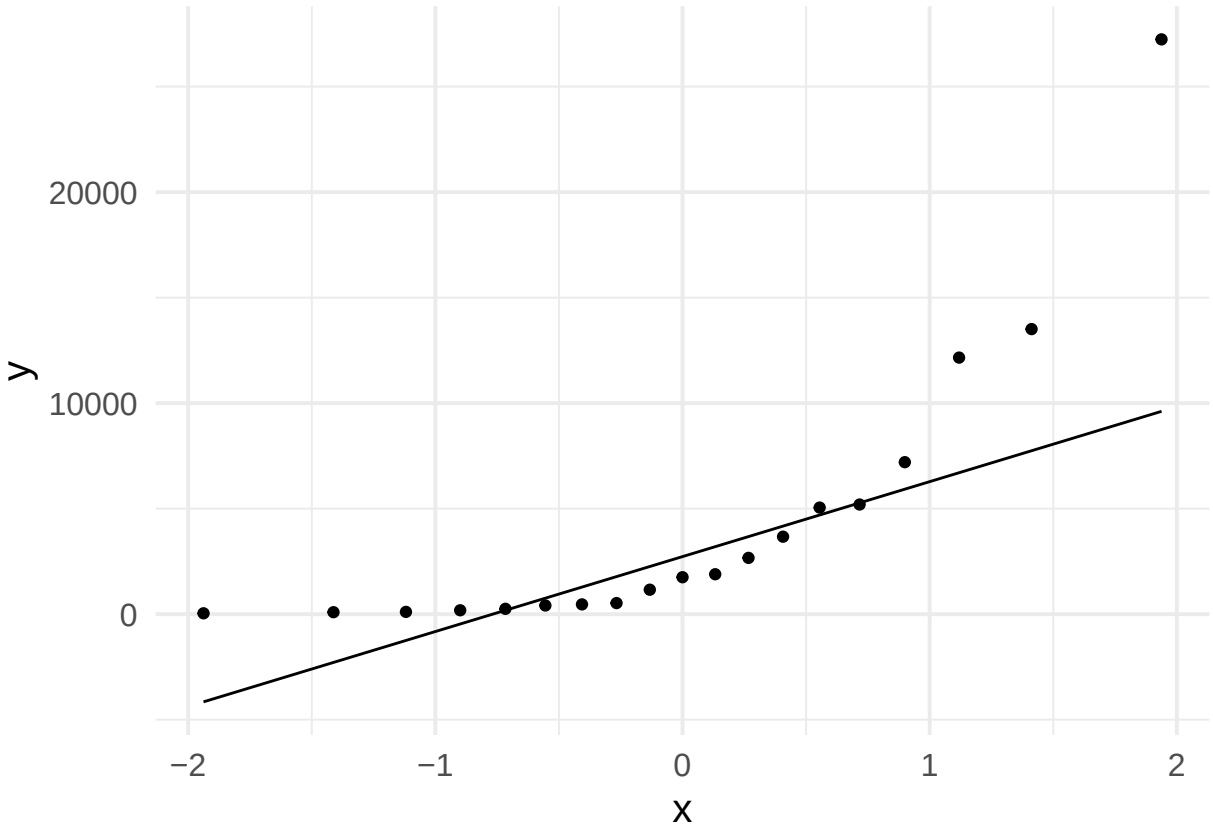
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Another example

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

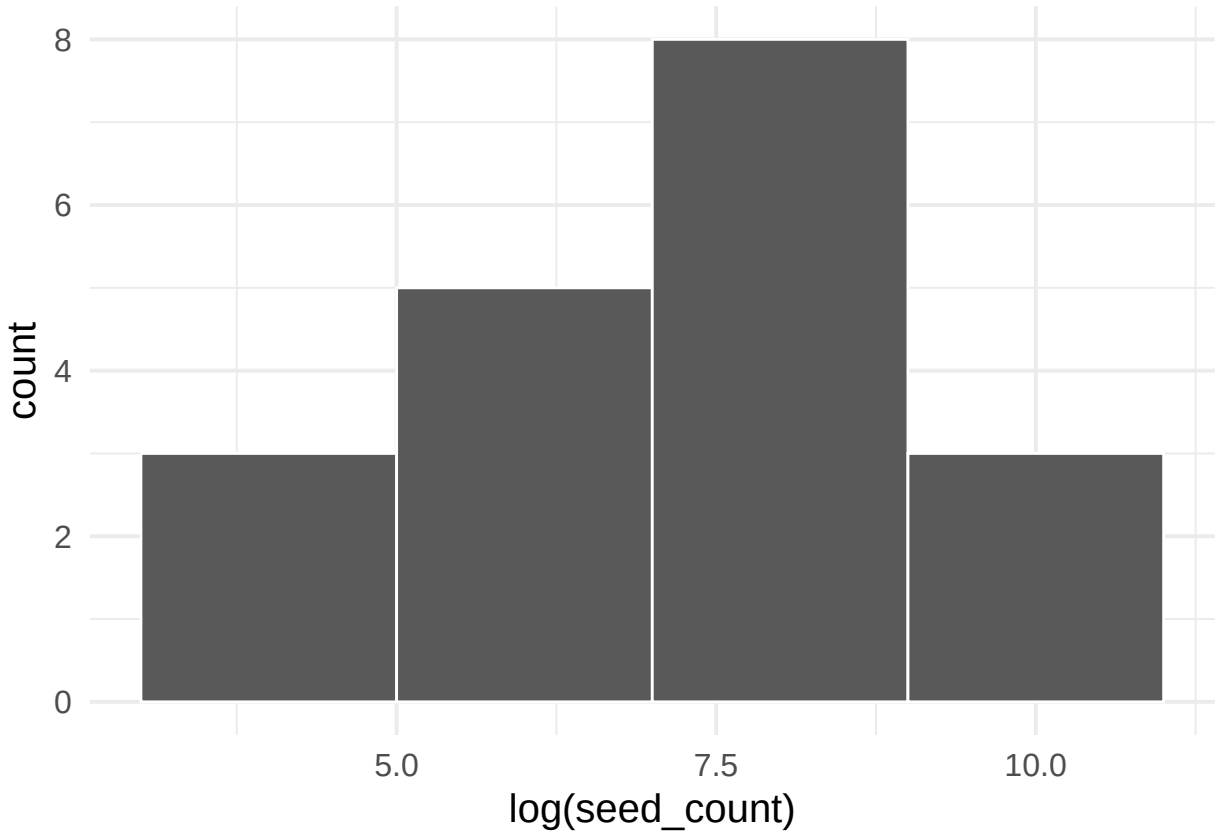
```
ggplot(seed_data, aes(sample = seed_count)) +
  stat_qq() + stat_qq_line() +
  theme_minimal(base_size = 15)
```



Another example (logged)

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

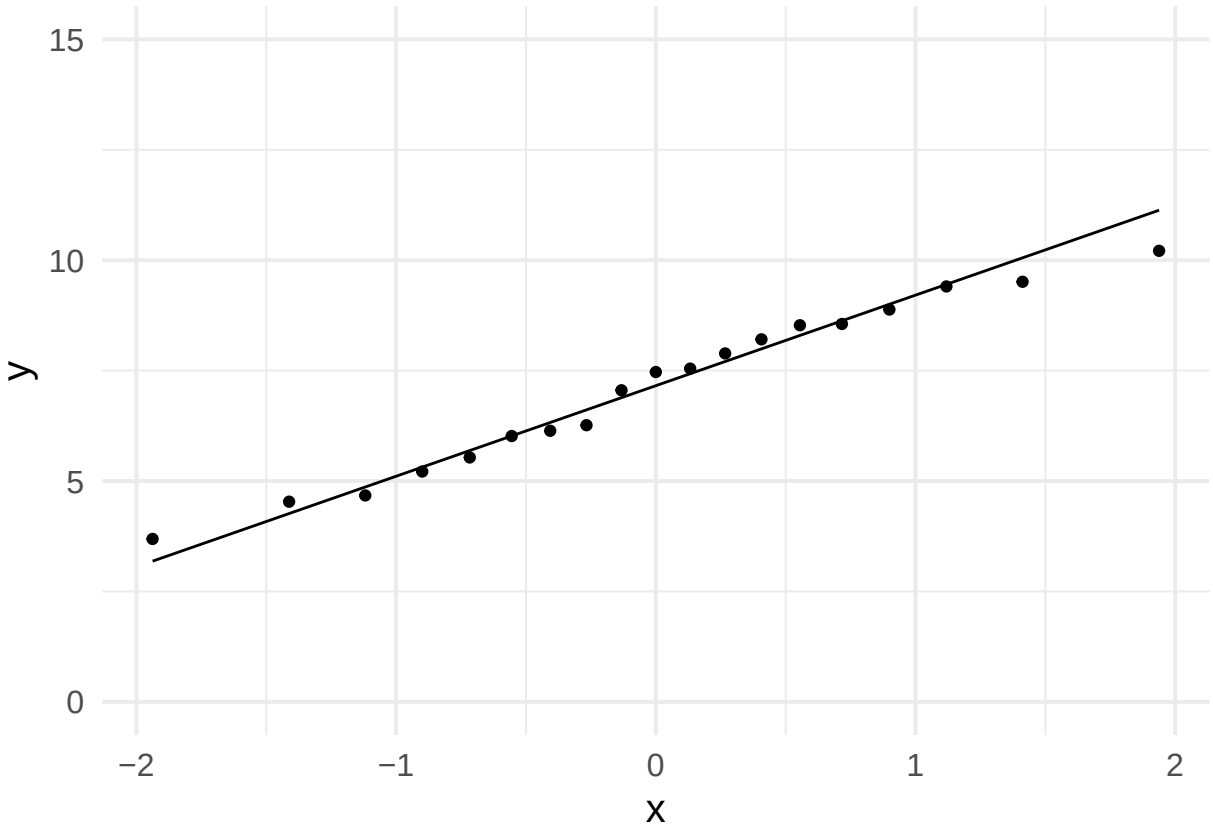
```
ggplot(seed_data, aes(x = log(seed_count))) +
  geom_histogram(col = "white", binwidth = 2) +
  theme_minimal(base_size = 15)
```



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?

Q-Q plots continued

- Read this blog post by Sean Kross (up to and including the Takeaways).
- No need to read the updates, or to understand the code Sean is using–it is different from the code we've been learning in class.
- Pay most attention to the presentation of the Quantile-Quantile plots for all the distributions he presents.
- Important note: Sean is plotting "Q-Q" plots and we've been plotting Normal quantile plots. Q-Q plots are a little different, but the same takeaways apply, meaning that if you understand how to interpret Q-Q plots, you can also apply those interpretations to Normal quantile plots.

Q-Q plot questions

- Look at the charts entitled "Skewed right" and "Skewed left" and the Quantile-Quantile plots beside them. Why does the Quantile-Quantile plot for the skewed right plot curve upwards and to the right (i.e., above the line), while the Quantile-Quantile plot for the skewed left plot curve downwards and to the left?

Q-Q plot answers

"The second graph is "skewed right," meaning that most of the data is distributed on the left side with a long "tail" of data extending out to the right. The third graph is "skewed left" with its tail moving out to the left. Looking at the Q-Q plot for the second graph you can see that the last two theoretical quantiles for this dataset should be around 3, when in fact those quantiles are greater than 8. The points depart upward from the straight blue line as you follow the quantiles from left to right. The blue line shows where the points would fall if the dataset were normally distributed. The point's trend upward shows that the actual quantiles are much greater than the theoretical quantiles, meaning that there is a greater concentration of data beyond the right side of a Gaussian distribution. A similar phenomenon can be seen in the Q-Q plot of the third graph, where there is more data to the left of the Gaussian distribution. The points appear below the blue line because those quantiles occur at much lower values (between -9 and -4) compared to where those quantiles would be in a Gaussian distribution (between -4 and -2)."

Recap of functions used

- `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