Lec 15: The Normal Distribution continued

Instructors: Tomer Altman and Alan Hubbard

October 4, 2024

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

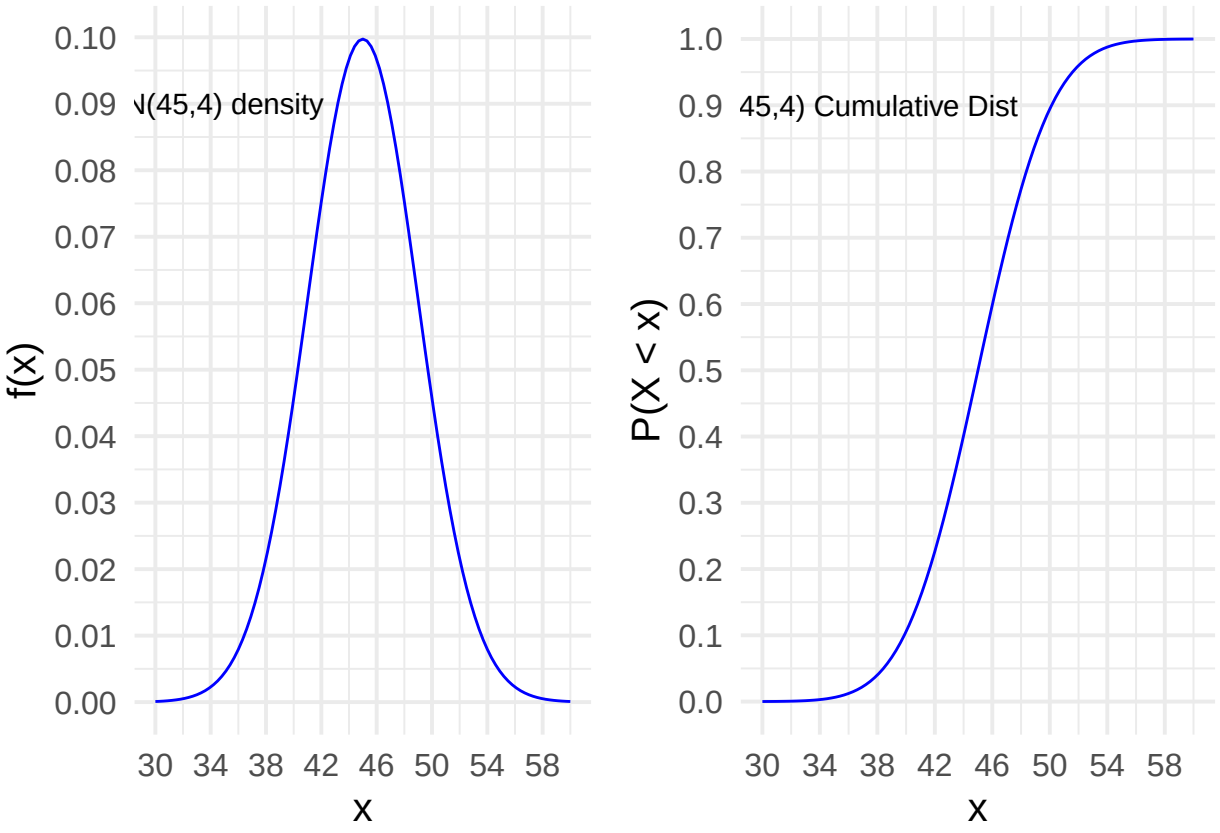
Finding Normal percentiles

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.

```
## Warning in geom_text(aes(x = 35, y = 0.09), label = "N(45,4) density", check_overlap = T, : All aestl
## i Please consider using ‘annotate()‘ or provide this layer with data containing
##   a single row.
```

```
## Warning in geom_text(aes(x = 37, y = 0.9), label = "N(45,4) Cumulative Dist", : All aesthetics have l
## i Please consider using ‘annotate()‘ or provide this layer with data containing
##   a single row.
```

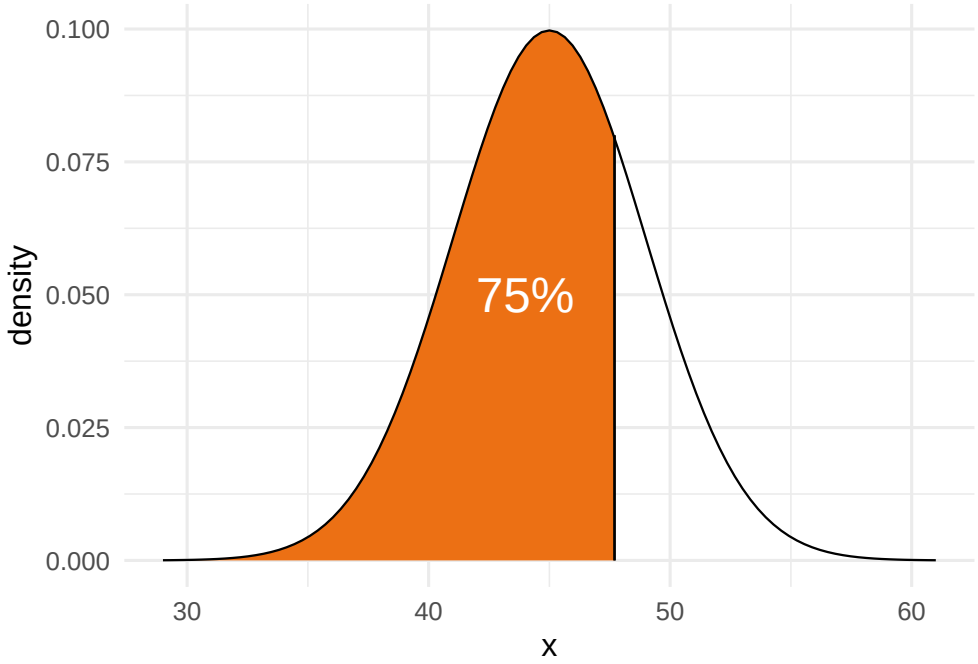


Finding Normal percentiles

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?

```
## Warning in geom_text(aes(x = 44, y = 0.05), check_overlap = T, label = "75%", : All aesthetics have l
## i Please consider using `annotate()` or provide this layer with data containing
##   a single row.
```



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)
```

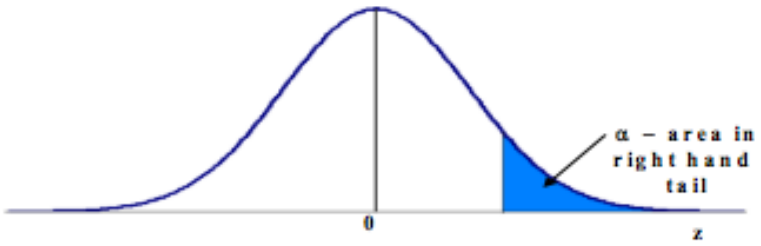
```
## [1] 47.69796
```

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 Inverse Normal Table



Given a probability in the tail this table gives the corresponding z score

Area in Right Hand Tail = α

α	z score	α	z score	α	z score
0.5	0.0000	0.3	0.5244	0.1	1.2816
0.49	0.0251	0.29	0.5534	0.09	1.3408
0.48	0.0502	0.28	0.5828	0.08	1.4051
0.47	0.0753	0.27	0.6128	0.07	1.4758
0.46	0.1004	0.26	0.6433	0.06	1.5548
0.45	0.1257	0.25	0.6745	0.05	1.6449
0.44	0.1510	0.24	0.7063	0.04	1.7507
0.43	0.1764	0.23	0.7388	0.03	1.8808
0.42	0.2019	0.22	0.7722	0.025	1.9600
0.41	0.2275	0.21	0.8064	0.02	2.0537
0.4	0.2533	0.2	0.8416	0.01	2.3263
0.39	0.2793	0.19	0.8779	0.009	2.3656
0.38	0.3055	0.18	0.9154	0.008	2.4089
0.37	0.3319	0.17	0.9542	0.007	2.4573
0.36	0.3585	0.16	0.9945	0.006	2.5121
0.35	0.3853	0.15	1.0364	0.005	2.5758
0.34	0.4125	0.14	1.0803	0.004	2.6521
0.33	0.4399	0.13	1.1264	0.003	2.7478
0.32	0.4677	0.12	1.1750	0.002	2.8782
0.31	0.4958	0.11	1.2265	0.001	3.0902

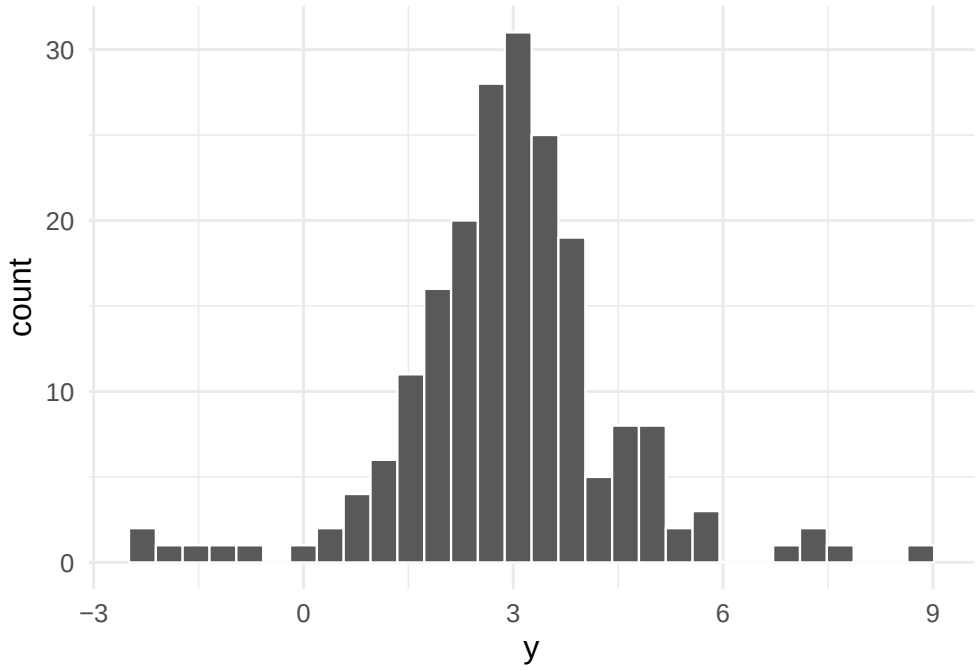
The Normal quantile plot (a.k.a the Q-Q plot)

- The purpose of making a Q-Q plot is to examine whether a continuous variable follows a Normal distribution
- 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`.
```



Basic idea of a Normal quantile plot

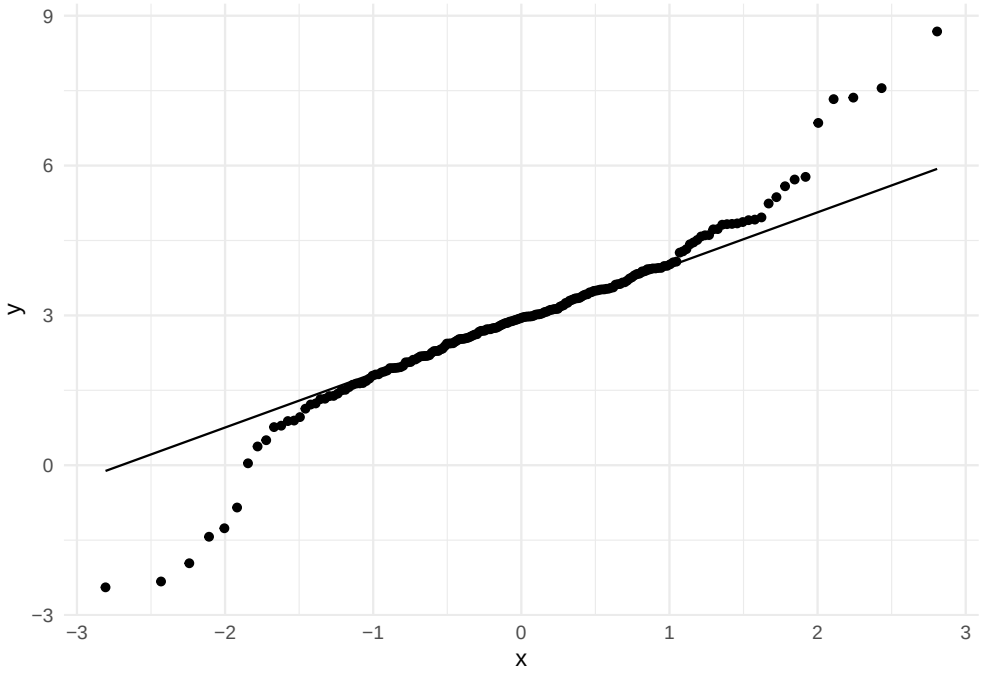
1. Arrange the observed data values from smallest to largest. Record what percentile of the data each value occupies. For example, the smallest observation in a set of 20 is at the 5% point, the second smallest is at the 10% point, and so on.
2. Do Normal distribution calculations to find z-scored at these same percentiles. For example, z = -1.645 is the 5% point of the standard Normal distribution, and z = -1.282 is the 10% point.
3. Plot each data point x against the corresponding z. If the data distribution is close to standard Normal, the plotted points will lie close to the 45-degree line x=z. If the data distribution is close to any normal distribution, the plots points will lie close to some straight line.

Any Normally distributed data set will produce a straight line on a Normal quantile plot, because the data distribution and the z distribution are both Normal and their relationship is thus linear. If the data are not Normally distributed, the data and the z distribution are unrelated, and a Normal quantile plot will not be linear.

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

Looking at this plot, does the pattern in the data lie on a straight line?

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



It does in the middle values of the data, but not at the tails. This means that the original variable is not Normally distributed.

Code template

```
#students, make sure to remove `eval=F` if you copy this code chunk
ggplot(your_data, aes(sample = your_var)) +
  stat_qq() +
  stat_qq_line() +
  theme_minimal()
```

Another example

Recall the seed data:

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

```
## Rows: 19 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): species
## dbl (2): seed_count, seed_weight
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
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
```

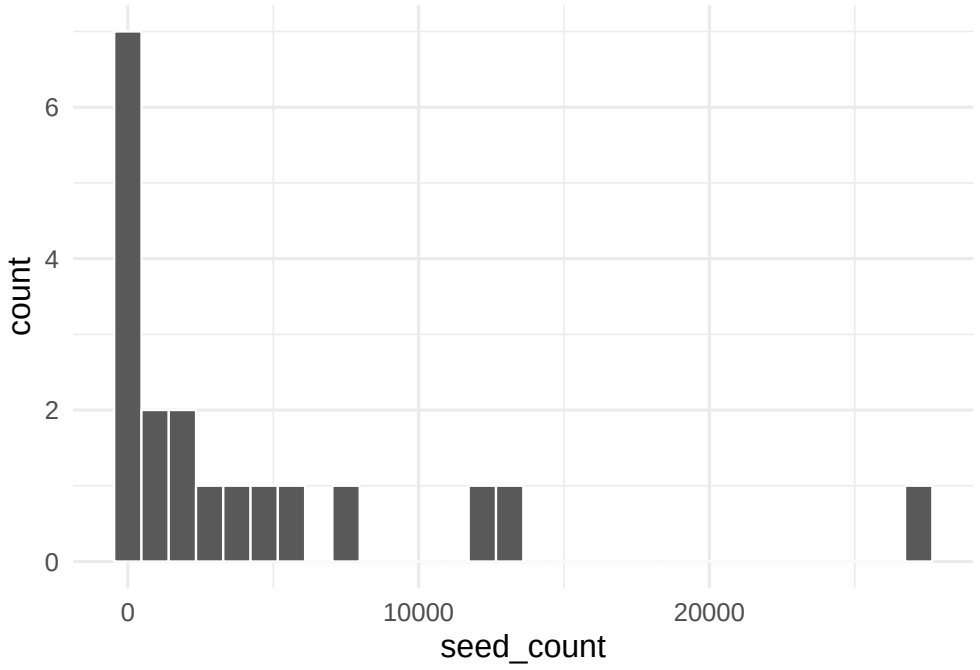
Is the distribution of `seed_count` Normal?

Another example

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

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

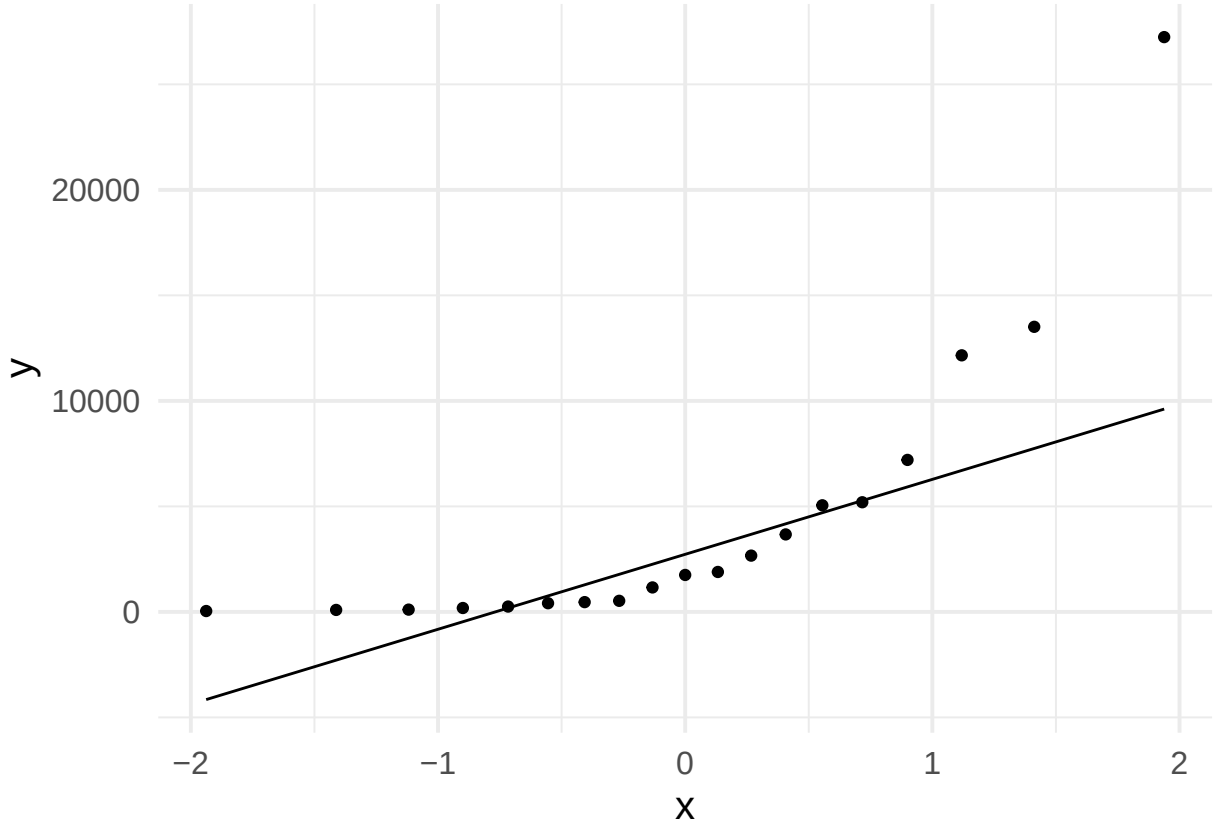
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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)
```

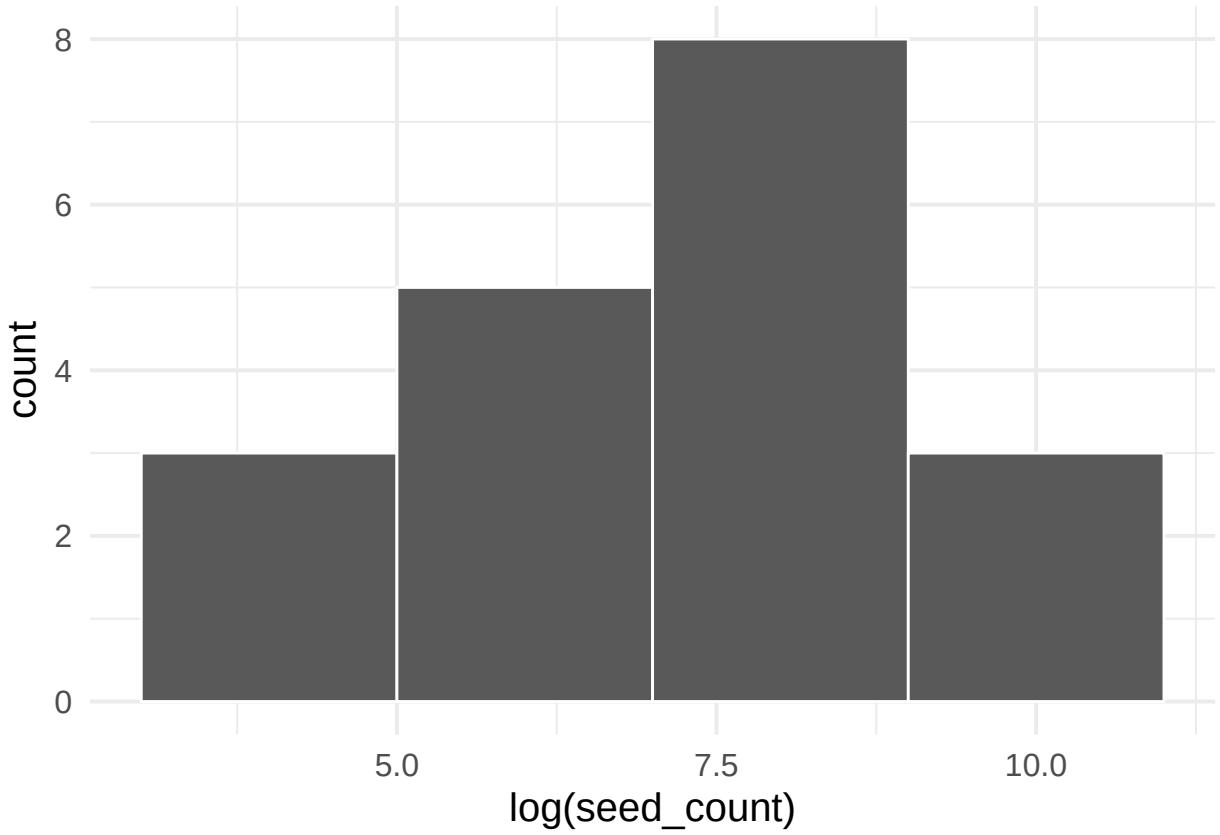


These data definitely do not follow a straight line - there is a curved pattern shown in the plot.

Another example (logged)

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

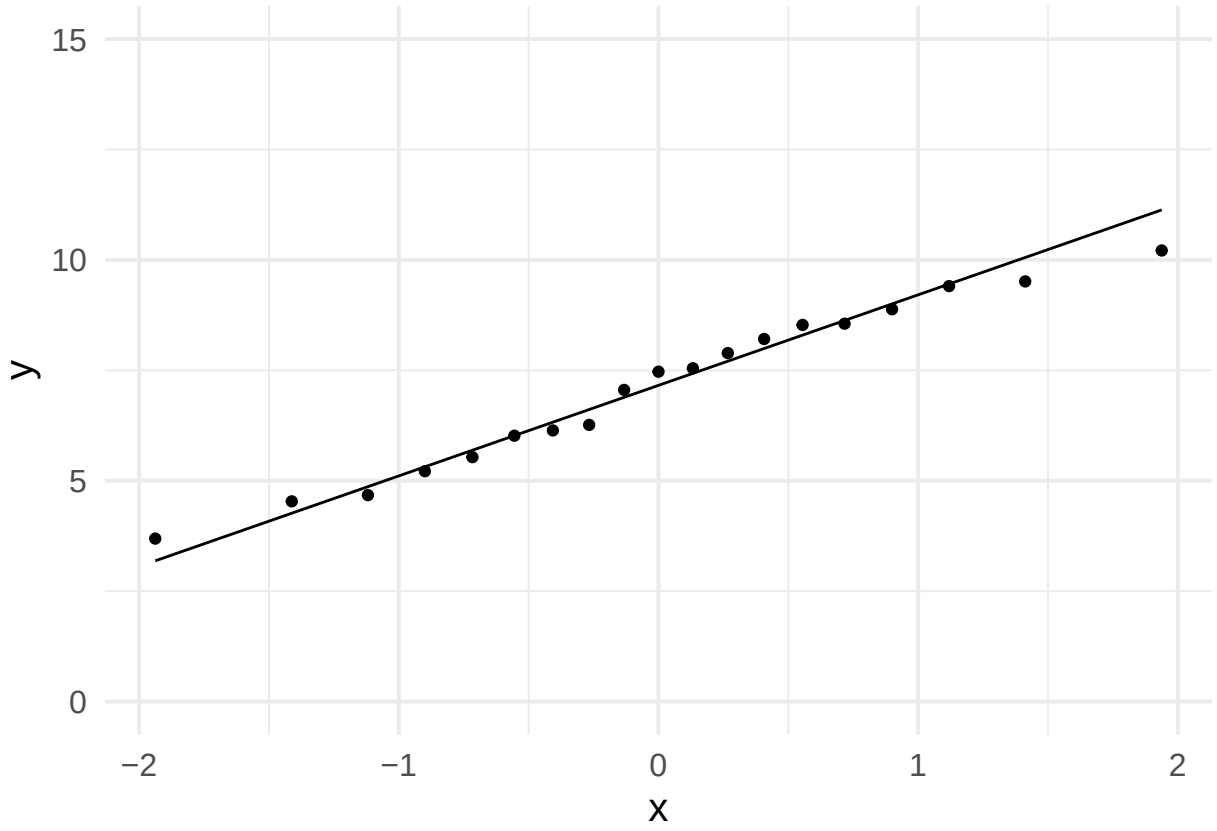
```
ggplot(seed_data, aes(x = log(seed_count))) +
  geom_histogram(col = "white", binwidth = 2) +
  theme_minimal(base_size = 15)
```



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))
```



Once we log transformed the data, the values now follow a Normal distribution

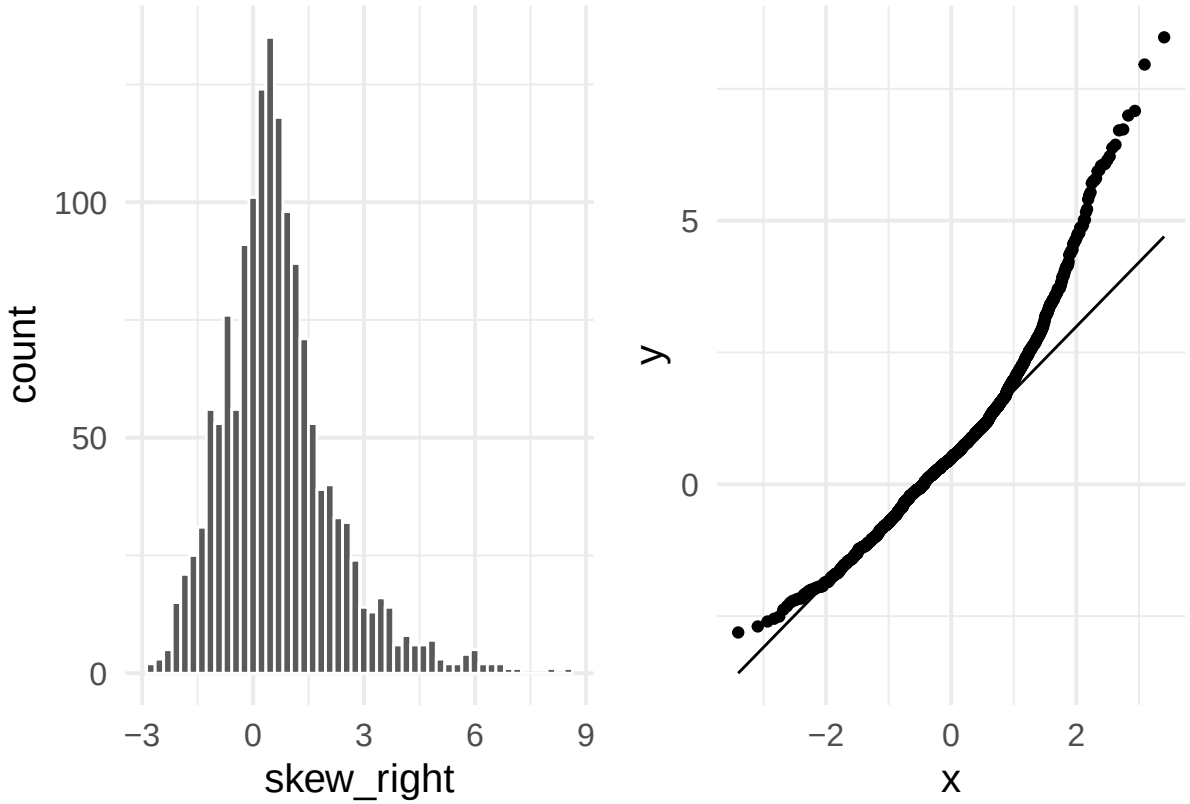
Normal quantile plot when data is skewed right

```
a <- ggplot(sr, aes(x = skew_right)) +
  geom_histogram(col = "white", bins = 50) +
  theme_minimal(base_size = 15)

b <- ggplot(sr, aes(sample = skew_right)) +
  stat_qq() +
  stat_qq_line() +
  theme_minimal(base_size = 15)

library(patchwork)

a + b + plot_layout()
```

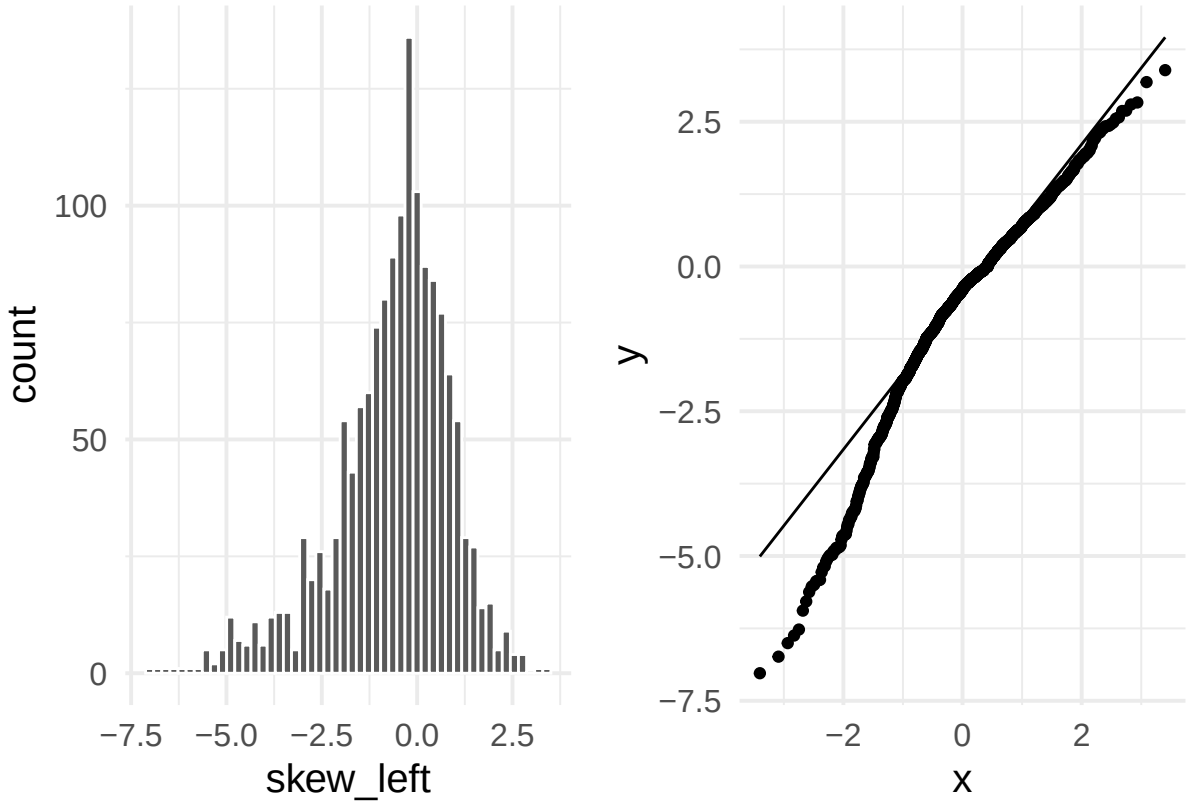


Normal quantile plot when data is skewed left

```
a <- ggplot(sl, aes(x = skew_left)) +
  geom_histogram(col = "white", bins = 50) +
  theme_minimal(base_size = 15)

b <- ggplot(sl, aes(sample = skew_left)) +
  stat_qq() +
  stat_qq_line() +
  theme_minimal(base_size = 15)

a + b + plot_layout()
```

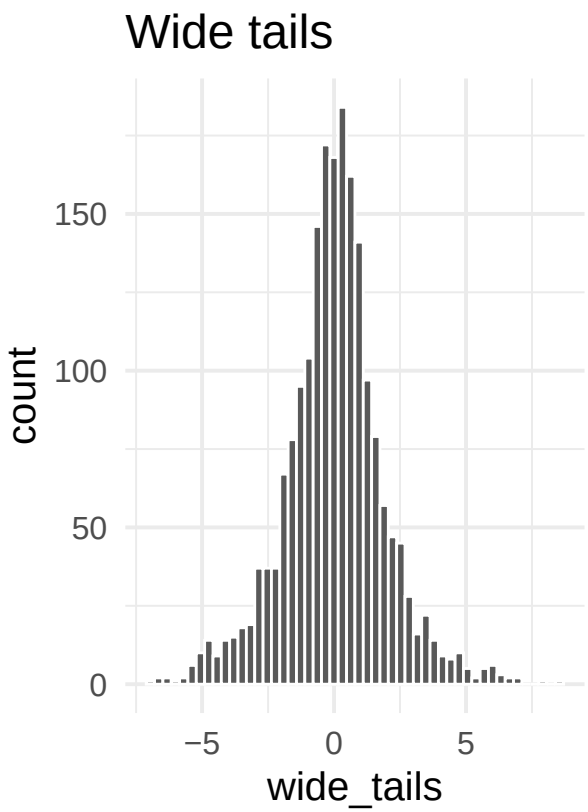


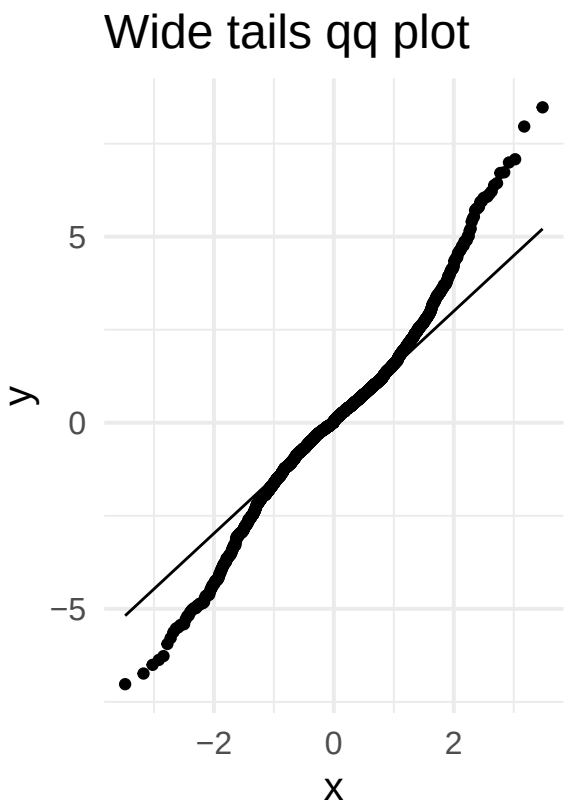
Normal quantile plot when data has "wide tails"

```
e <- ggplot(wt, aes(x = wide_tails)) +
  geom_histogram(col = "white", bins = 50) +
  theme_minimal(base_size = 15) + labs(title = "Wide tails")

f <- ggplot(wt, aes(sample = wide_tails)) +
  stat_qq() +
  stat_qq_line() +
  theme_minimal(base_size = 15) + labs(title = "Wide tails qq plot")

e + f + plot_layout()
```



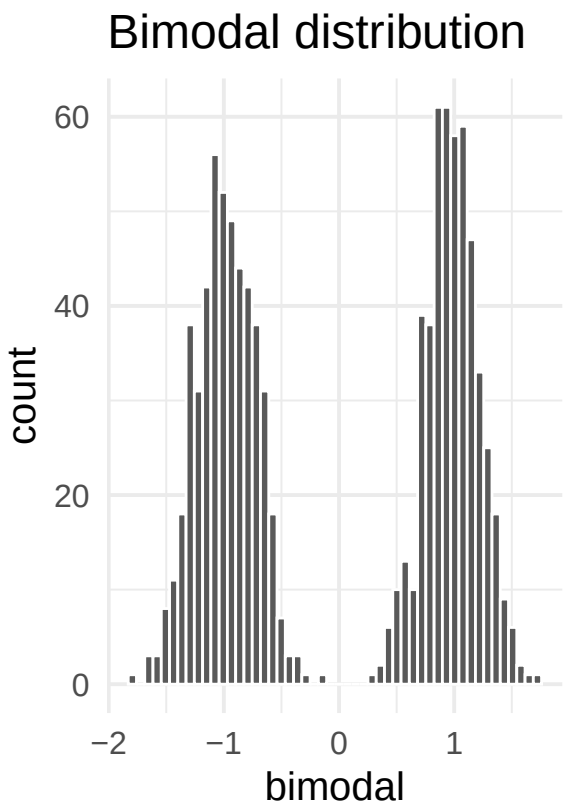


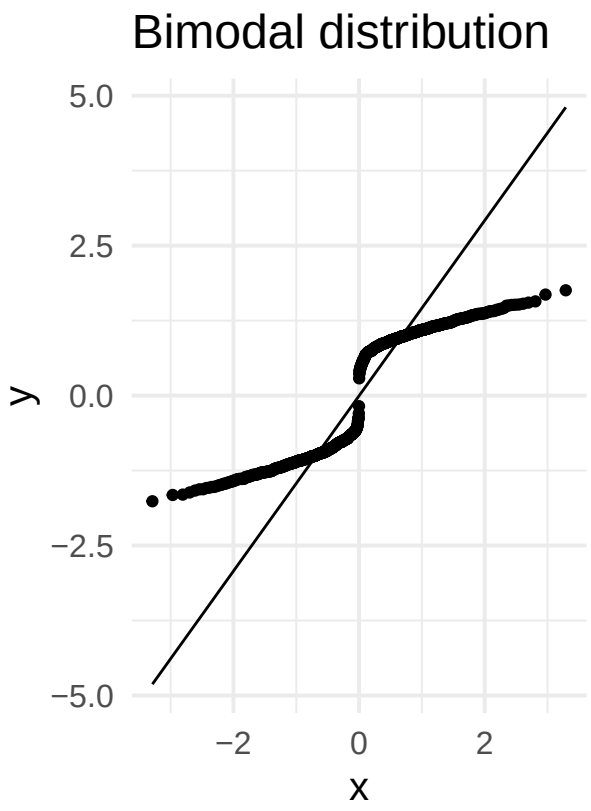
Normal quantile plot when data is bimodal

```
g <- ggplot(bi, aes(x = bimodal)) +
  geom_histogram(col = "white", bins = 50) +
  theme_minimal(base_size = 15) + labs(title = "Bimodal distribution")

h <- ggplot(bi, aes(sample = bimodal)) +
  stat_qq() +
  stat_qq_line() +
  theme_minimal(base_size = 15) + labs(title = "Bimodal distribution")

g + h + plot_layout()
```





Reference

- Read this blog post by Sean Kross (up to and including the Takeaways).
- You can read this if you want, but you don't need to.

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