The normal distribution

Learning objectives for today

- Learn about a Normal distribution centered at μ with a standard deviation of σ
- Learn about the standard Normal ($\mu = 0$ and $\sigma = 1$)
- Perform simple calculations by hand using the 68-95-99.7 rule
- Calculate probabilies above, below or within given values in a normal distribution
- Calculate the quantile for a specified cummulative probability for any normal distribution
- Understand and compute Z-scores
- Create and interpret a QQ plot

Statistics is everywhere

Height and success

We often see articles about the correlation of height and success - arguing that taller people are more successfull.

What type of problem is this

Remembering back to our PPDAC framework. . . .

What kind of questions are these studies asking?

Counterfactual

Frayling, one of the authors of a study on height and success said:

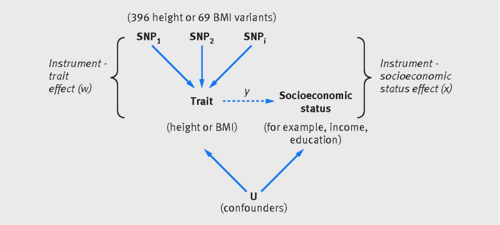
> "if you took the same man - say a 5ft 10in man and make him 5ft 7in - and sent him through life, he would be about £1,500 worse off per year"

Were there important confounders?

This study was published in BMJ (Height, body mass index, and socioeconomic status: Mendelian randomization study in UK Biobank BMJ 2016; 352 doi: https://doi.org/10.1136/bmj.i582 (Published 08 March 2016))

They used mendelian randomization to avoid confounding - and put a DAG in their paper to illustrate

DAG


(396 height or 69 BMI variants)
SNP$_1$
SNP$_2$
SNP$_i$
Instrument - trait effect (w)
Instrument - socioeconomic status effect (x)
y
Trait
Socioeconomic status
(height or BMI)
(for example, income, education)
U
(confounders)

Proposed mechansim

One hypothesized mechanism for the influence of height on success is that those who are perceived as taller than normal experience positive social bias.

But wait. . . . what does "taller than normal" mean?

Three relevant Definitions of Normal (Mirriam-Webster)

1: Conforming to a type, standard , or regular pattern

2: of, relating to, or characterized by average intelligence or development

3: relating to, involving, or being a normal curve or normal distribution

Probability Distributions

Empirical vs Theoretical

- probabilities calculated from a finite amount of observed data are called **Empirical** probabilities
- probabilities based on theoretical functions are called **Theoretical Probability Distributions**

Empirical distributions

What was the empirical distribution we used to visualize a continuous variable in part I of the class?

What did we look at when summarizing continuous variables visually and numerically?

How might we summarize height?

Empirical distribution of height

Height is a continuous variable so we summarize height in a sample with:

- Measure of **central tendancy**
  - mean
  - median
  - mode
- Measure of **variability/spread**
  - standard deviation
  - IQR
- Visually we would look to see if the distribution is
  - Unimodal
  - symmetric

Why are we interested in theoretical distributions?

- Theoretical distributions apply probability theory to describe the behavior of a random variable
- For continuous variables (like height) this allows us to predict the probability associated with a range of values
- They help us answer questions about what is "normal" if by "normal" we essentially mean what is expected

Figure 1: Carl Friedrick Gauss

The Normal Distribution

The Normal distribution

Function for Normal Distribution

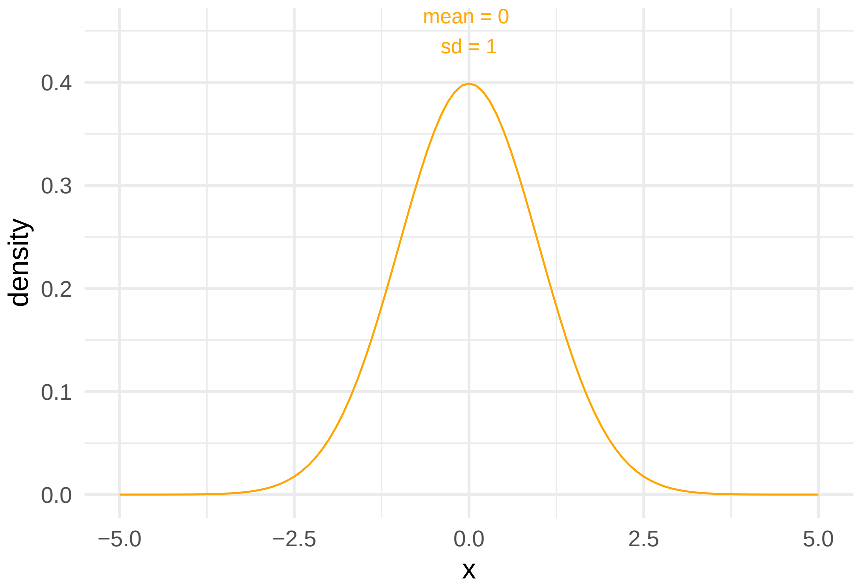
The underlying function which generates a probability distribution for a Normal distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}}^{\left(-\frac{1}{2\sigma^2}\times(x-\mu)^2\right)}$$

You do NOT need to remember this However you should know that the distribution is defined by two parameters, the mean and standard deviation, and that all the values under the curve must total to 1 (the probability space here is the area under the curve)

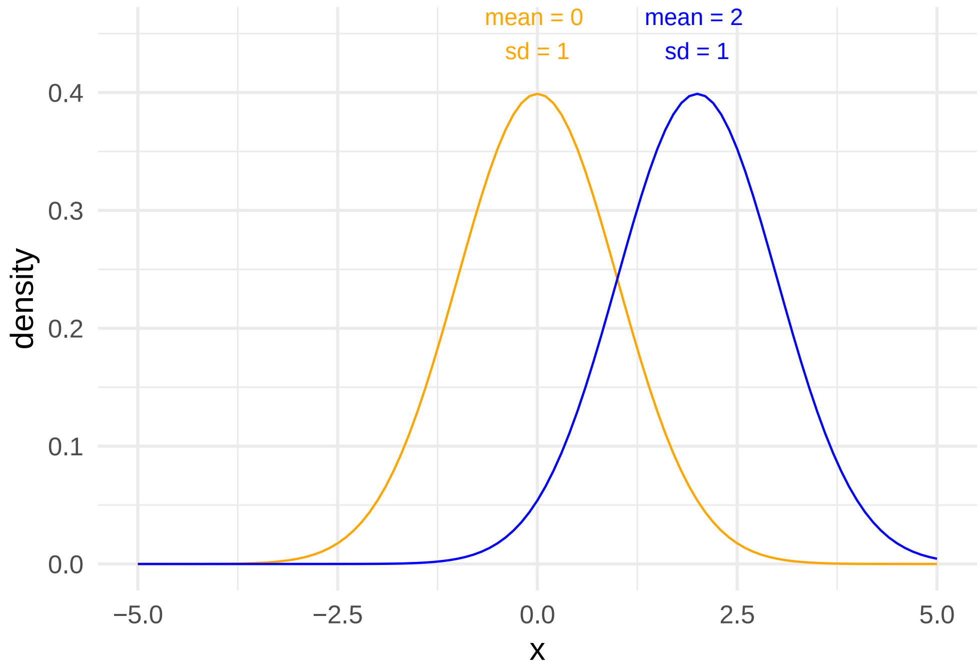
The Normal Distribution N(0,1)

- Here is a Normal distribution with mean (μ) = 0, standard deviation (σ) = 1 .



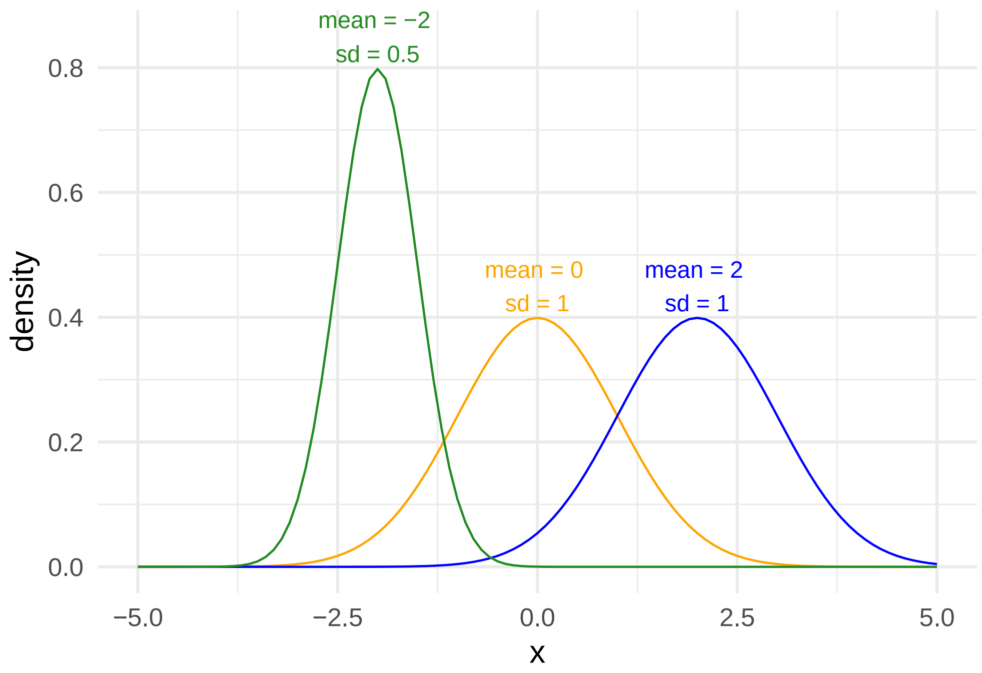
The Normal Distribution N(0,1) and N(2,1)

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



The Normal Distribution N(O,1) and N(2,1) and N(-2,0.5)

- Let's add a third Normal distribution, this one centered at -2, with a standard deviation of 0.5



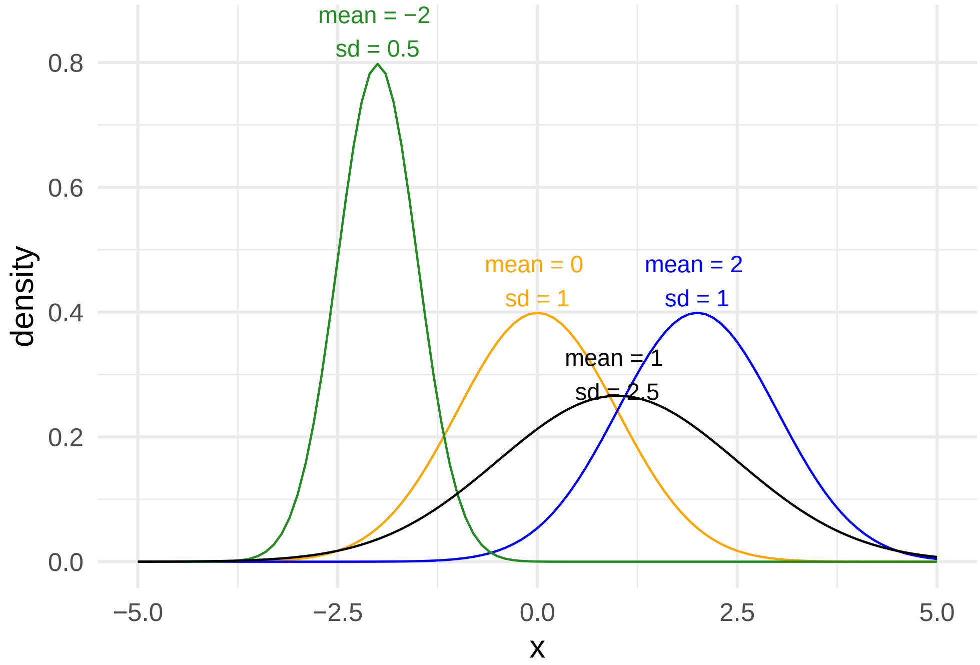
The Normal Distribution

- Notice what happens when we make the standard deviation smaller (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 $\sigma = 2.5$ 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 height of the curve at each point represents the probability of observing that value (however we never calculate probabilities for an exact value, only for ranges…why?)
- 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

The 68-95-99.7 rule for approximation in 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

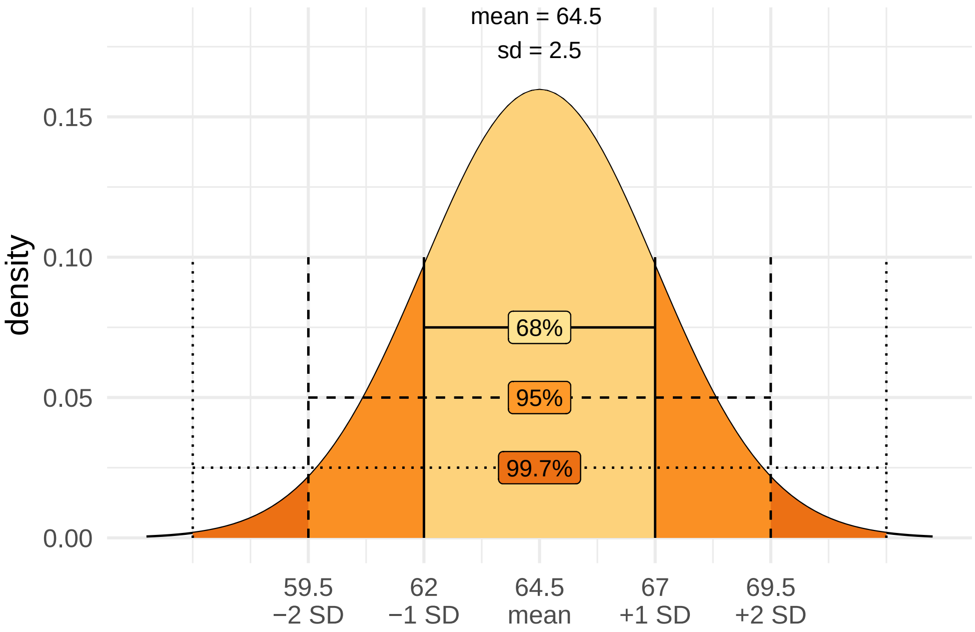
Remembering that the normal curve is symmetric we then also know:

- $P(X < \mu - \sigma) \approx 16\%$
- $P(X > \mu + \sigma) \approx 16\%$
- $P(X < \mu - 2\sigma) \approx 2.5\%$

- $P(X > \mu + 2\sigma) \approx 2.5\%$
- $P(X < \mu - 3\sigma) \approx .15\%$
- $P(X > \mu - 3\sigma) \approx .15\%$

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

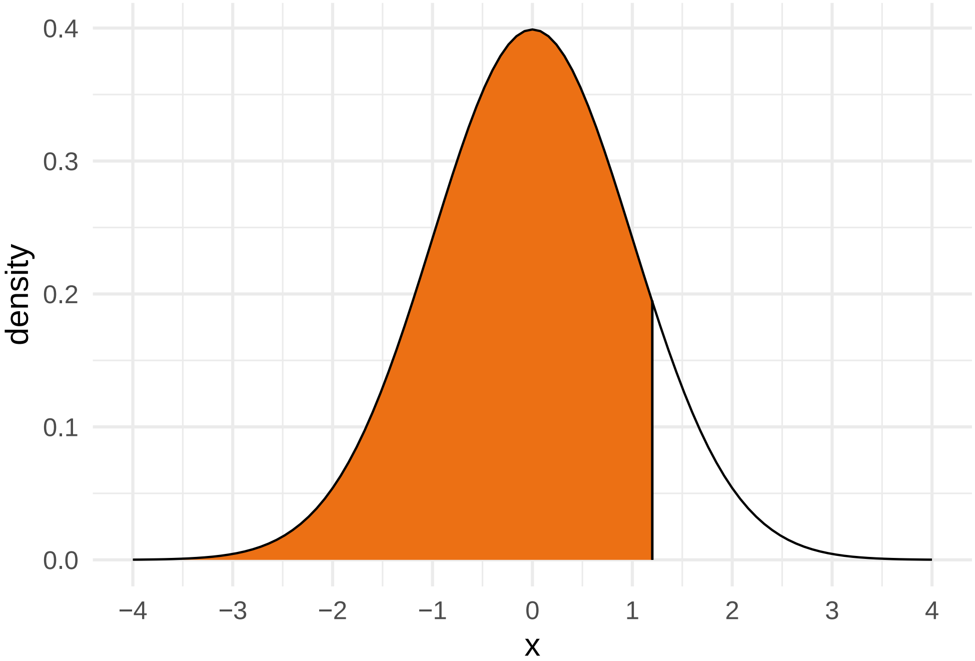
$\mu = 64.5$ inches and standard deviation $\sigma = 2.5$ inches

- What calculations could you do with just the μ and σ values and this rule?
- P(62 < H < 67) = ?
- P(H > 62) = ?
- Thinking back to the article about height - who would we consider “taller than normal” in these data?

Finding Normal probabilities

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

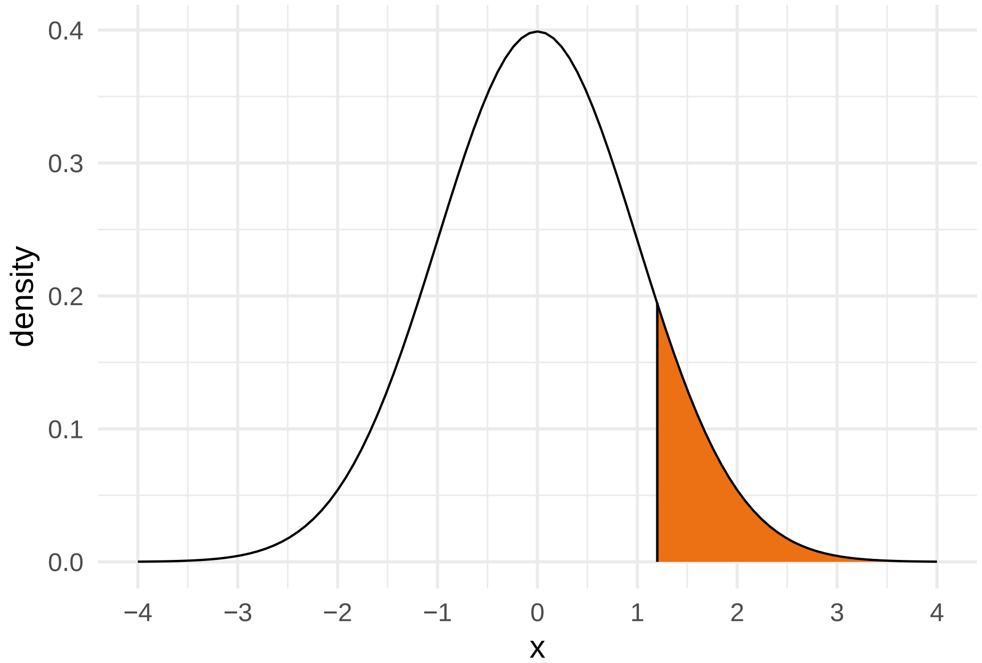
- Recall that 100% of the sample space for the random variable x lies under the probability density function.
- What is the amount of the area that is below x = 1.2?
- To answer this question we use the `pnorm()` function:

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

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

Finding Normal probabilities

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



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

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

Alternatively:

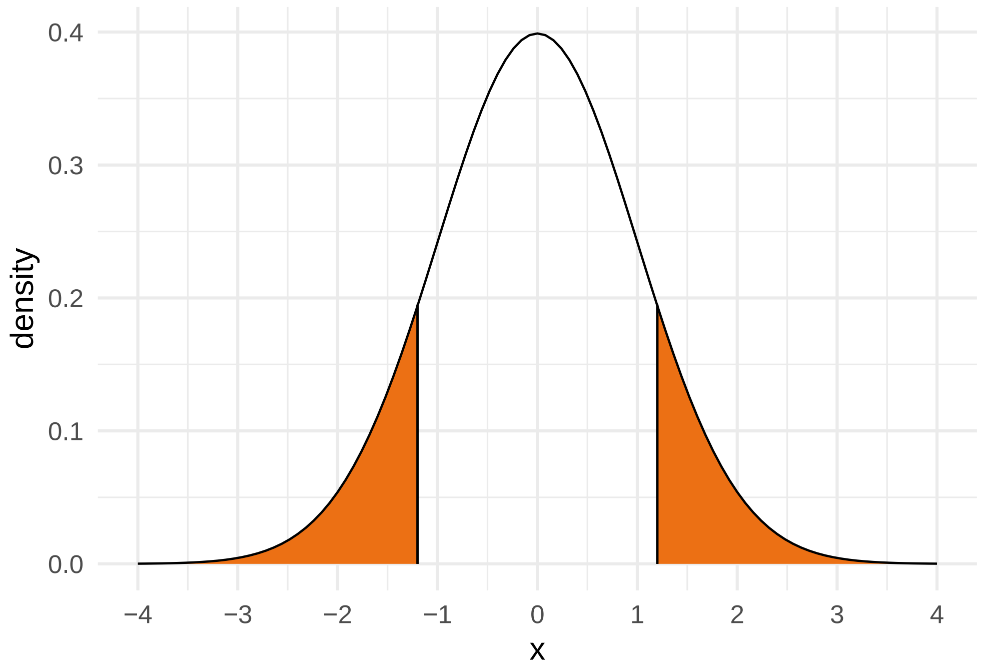
```r
pnorm(q = 1.2, mean = 0, sd = 1, lower.tail = F)
```

```
## [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

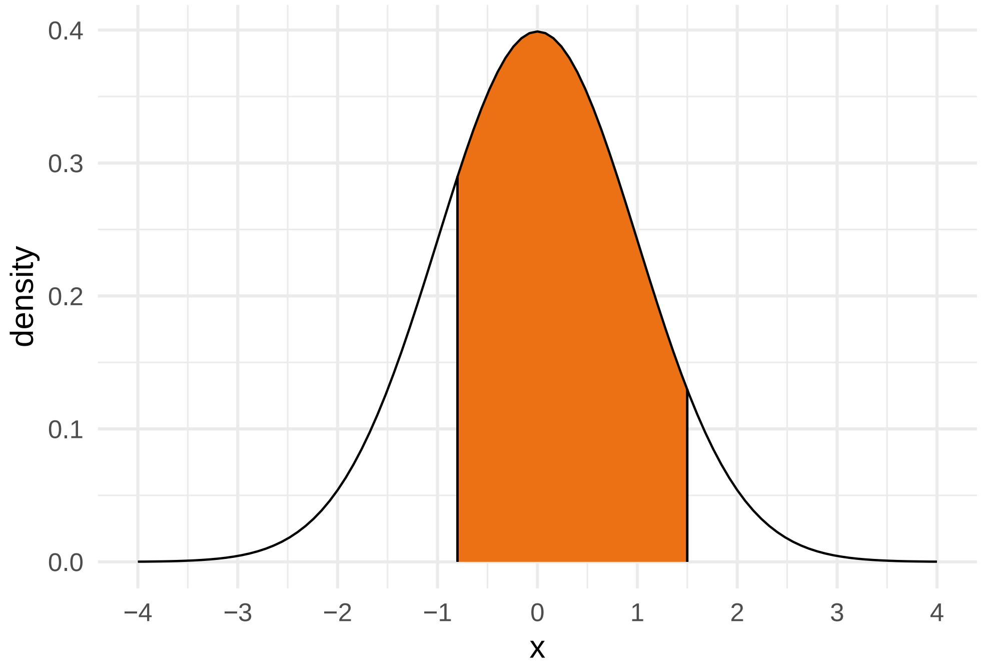
The trick: find one of the tails and then double the area because the distribution is symmetric:

```
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
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?

```
## [1] 0.3085375
```

Finding Normal percentiles

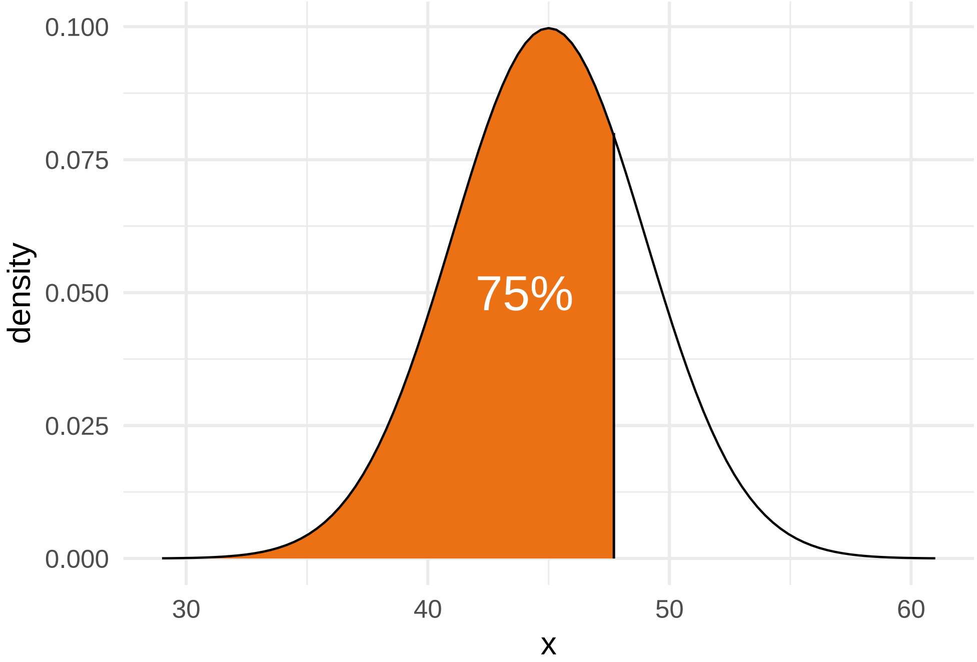
Finding Normal percentiles

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?



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.

The standard normal and Z scores

Convert to a standard

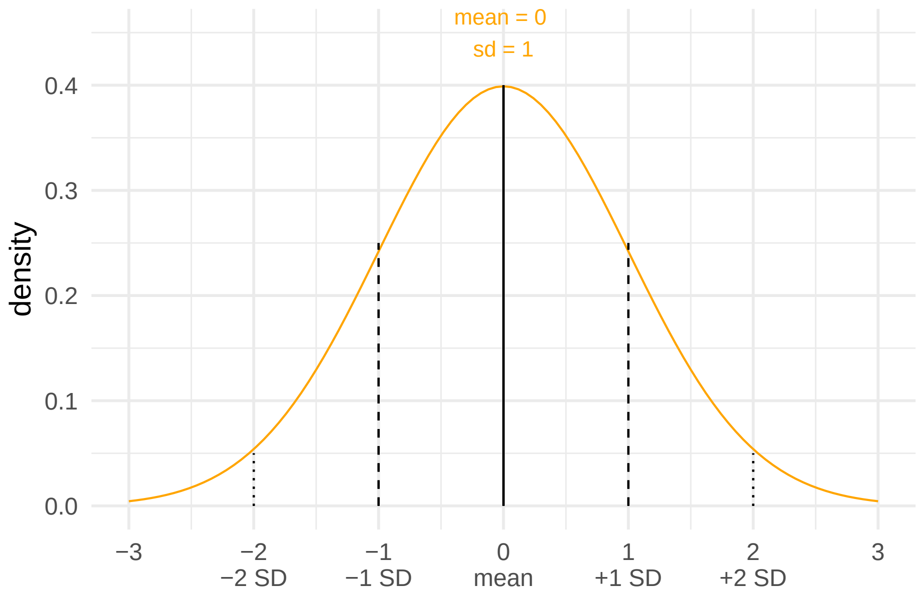
A normal distribution can have an infinite number of possible values for its mean and sd.

It would be impossible to tabulate the area associated with each and every normal curve

So instead of doing the impossible, we convert to the **Standard Normal** Distribution

The standard Normal distribution

- The standard Normal distribution N(0,1) has $\mu = 0$ and $\sigma = 1$.
- $X \sim N(0, 1)$, 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,

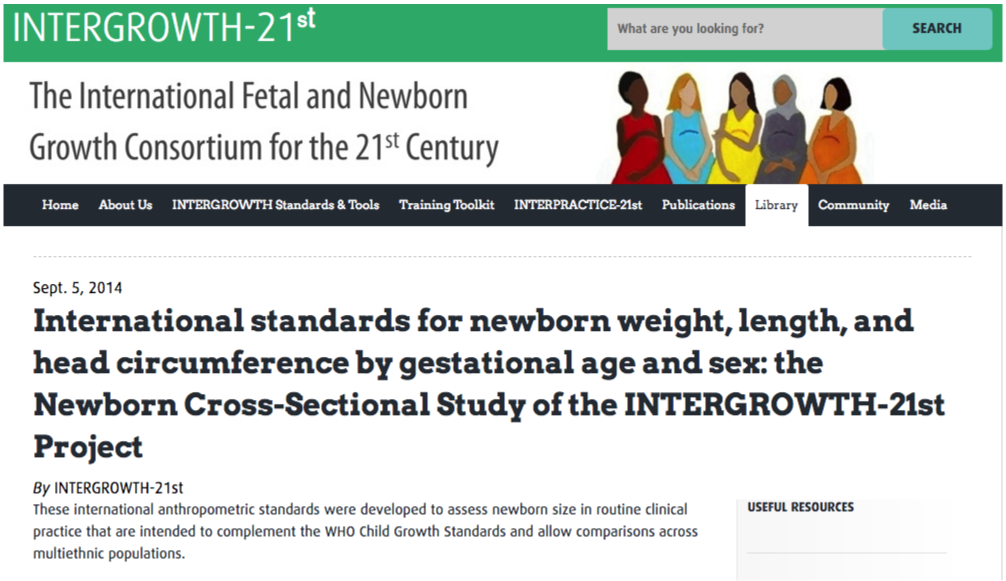
$$z = \frac{x - \mu}{\sigma}$$

What's the Z

By converting our variable of interest X to Z we can use the probabilities of the standard normal probability distribution to estimate the probabilities associated with X.

- 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.

Standardizing Normally distributed data



The International Fetal and Newborn Growth Consortium for the 21st Century

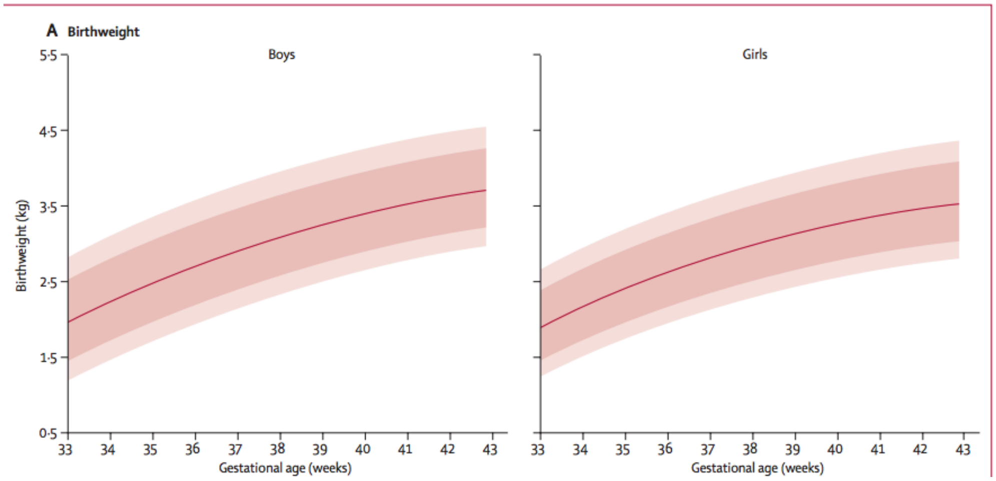
Sept. 5, 2014

International standards for newborn weight, length, and head circumference by gestational age and sex: the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project

By INTERGROWTH-21st

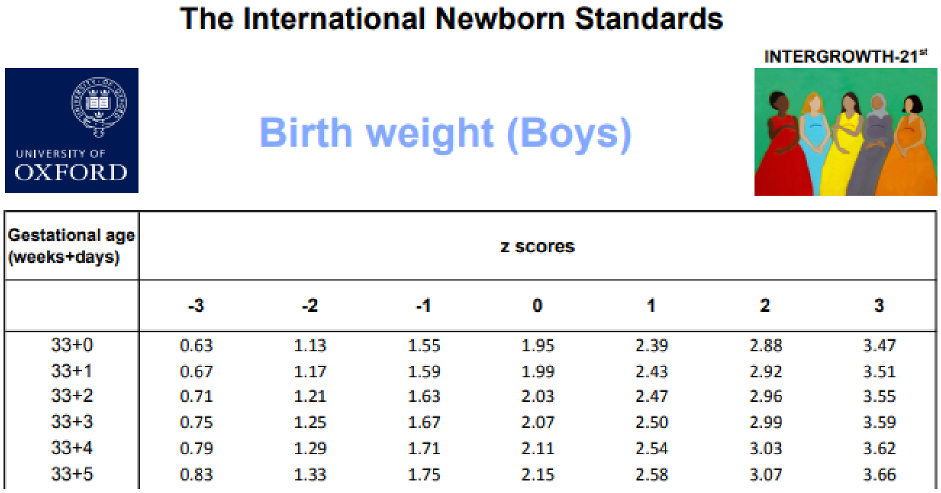
These international anthropometric standards were developed to assess newborn size in routine clinical practice that are intended to complement the WHO Child Growth Standards and allow comparisons across multiethnic populations.

Standardizing Normally distributed data



Reference

Standardizing Normally distributed data



The International Newborn Standards

INTERGROWTH-21st

Birth weight (Boys)

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

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

Simulating Normally distributed data in R

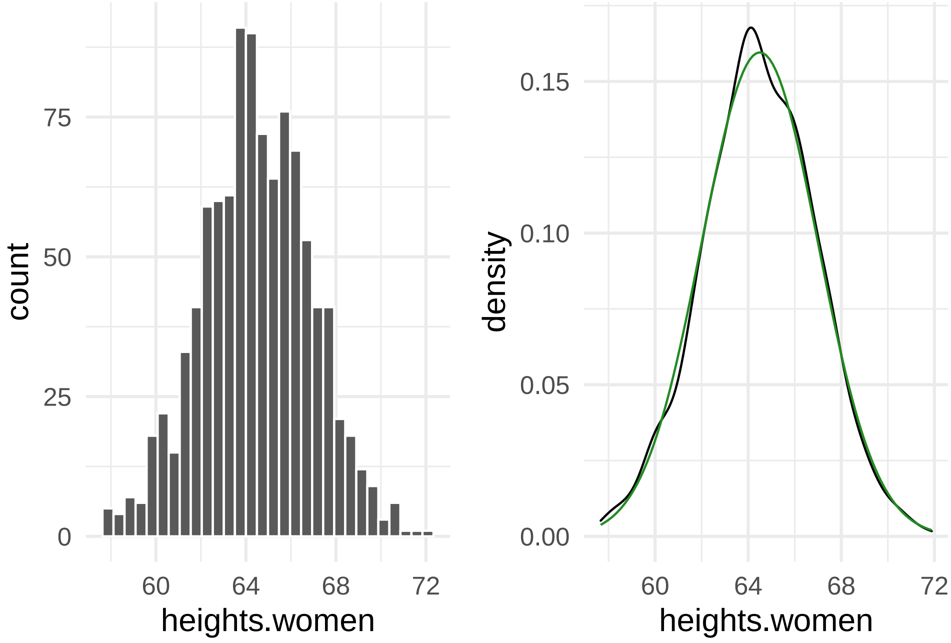
Simulating Normally distributed data in R

Suppose that we measured 1000 heights for young women:

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

Simulating Normally distributed data in R

We can plot the histogram of the heights, and see that they roughly follow from a Normal distribution:



Standardizing Normally distributed data in R

To standard these data, we can apply the formula to compute the z-value:

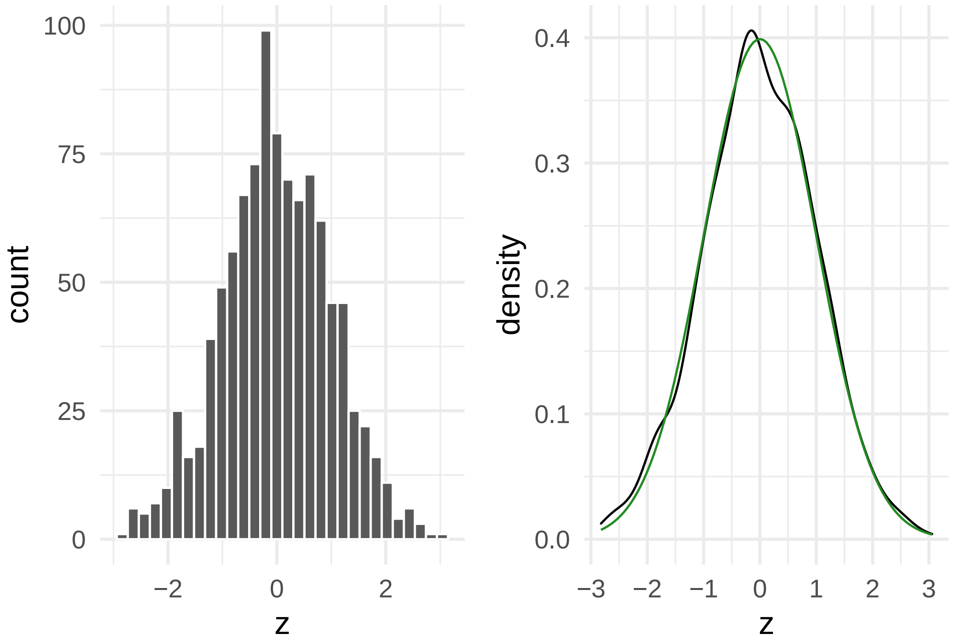
```
heights.women <- heights.women %>% mutate(mean = mean(heights.women),
                                          sd = sd(heights.women),
                                          z = (heights.women - mean)/sd)

head(heights.women)
```

```
##   heights.women     mean       sd          z
## 1      64.72866 64.48788 2.418886  0.0995390
## 2      64.05294 64.48788 2.418886 -0.1798138
## 3      68.01690 64.48788 2.418886  1.4589424
## 4      62.69879 64.48788 2.418886 -0.7396365
## 5      64.93947 64.48788 2.418886  0.1866896
## 6      64.89609 64.48788 2.418886  0.1687589
```

What would the distribution of the standardized heights look like?

Standardizing Normally distributed data in R



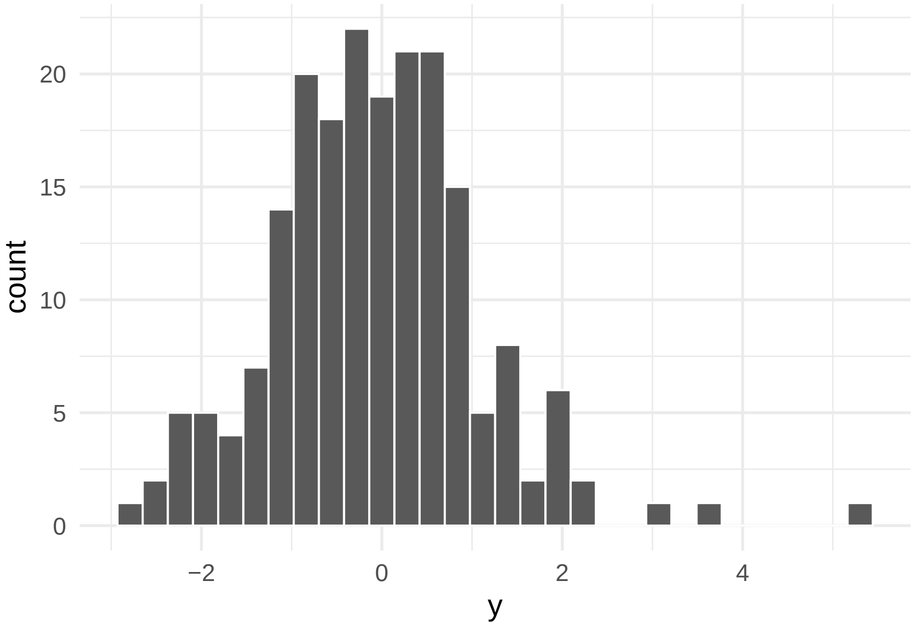
The Normal quantile plot (a.k.a the Q-Q plot)

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 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 you can use a Q-Q plot.

Are these data Normally distributed?

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



Making a QQ plot step by step

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, at x = -1 at the 16th percentile of the N(0, 1) distribution.
3. Make a scatter plot of the calculated x-values on the x-axis and the original variable values on the y-axis.
4. The closer the data is to a straight line, the more closely it approximates a Normal distribution.

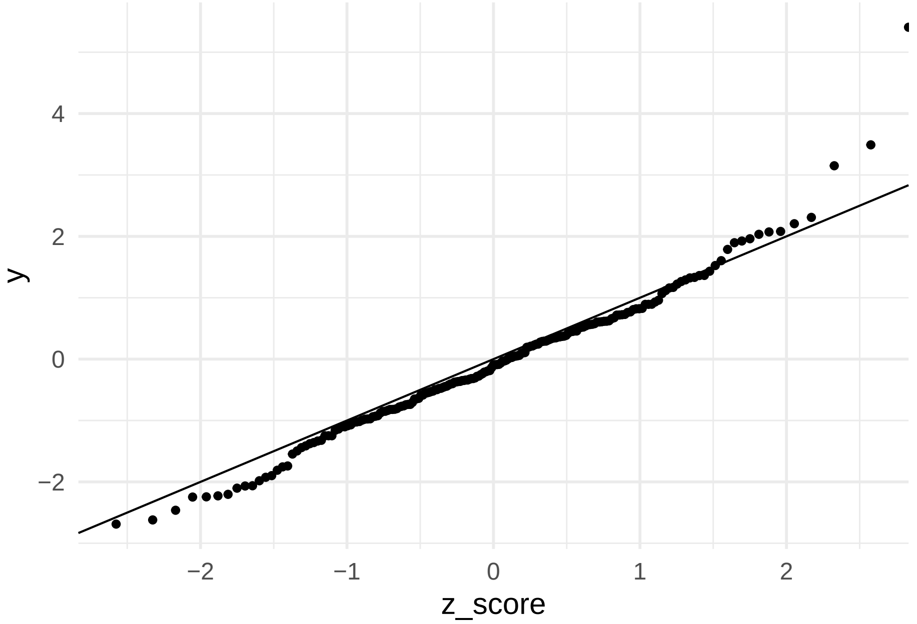
Making a QQ plot step by step

```
#1. calculate the percentile :
example_data <- example_data %>%  arrange(y) %>%
  mutate(quantile = row_number()/n())
# 2. then calculate the x-value at each percentile from the previous step
# 3. 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 -2.687811    0.005 -2.575829
## 2 -2.620008    0.010 -2.326348
## 3 -2.462334    0.015 -2.170090
## 4 -2.247025    0.020 -2.053749
## 5 -2.243396    0.025 -1.959964
## 6 -2.227690    0.030 -1.880794
```

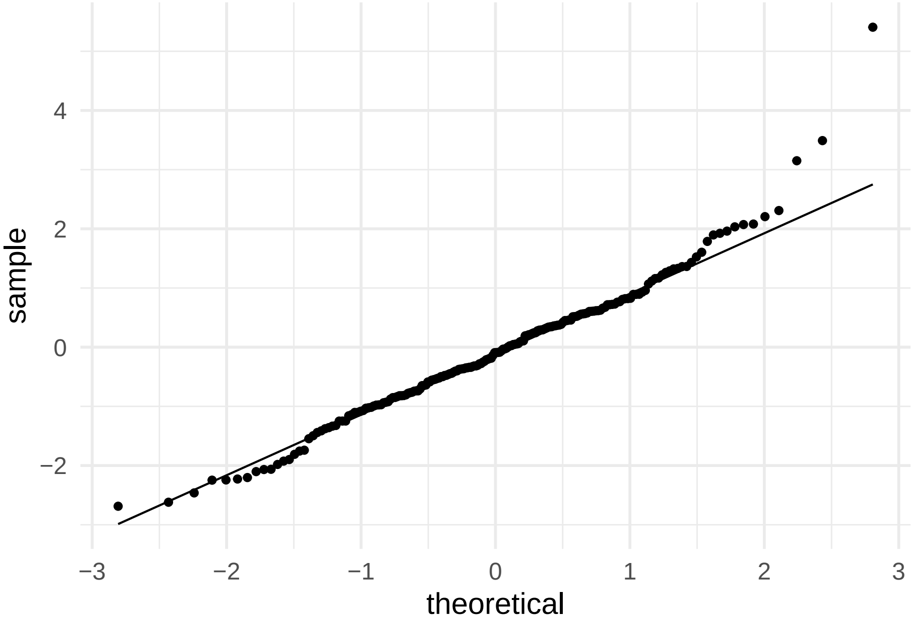
Look at the QQ plot for these data

- 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.



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: Seed Data

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.0.4
```

```
seed_data <- read_csv("Ch04_seed-data")
```

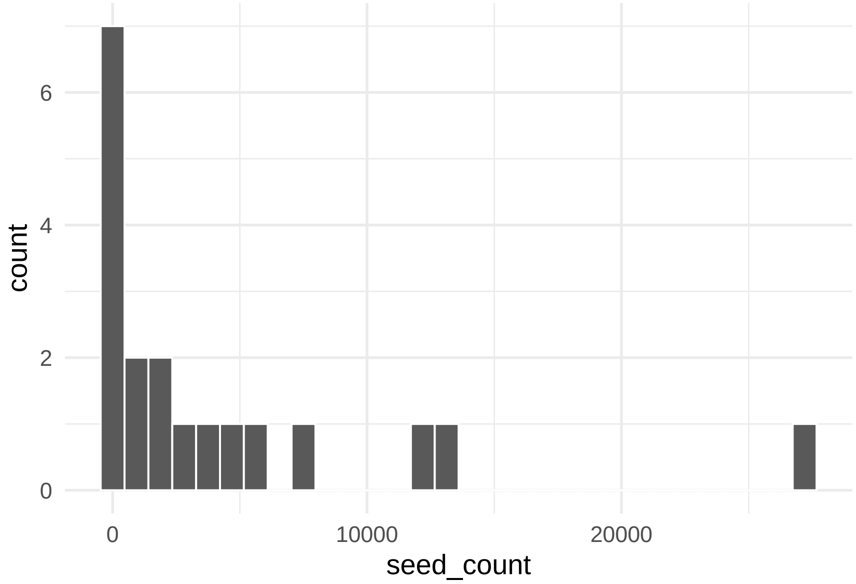
```
##
## -- Column specification --------------------------------------------------------
## cols(
##   species = col_character(),
##   seed_count = col_double(),
##   seed_weight = col_double()
## )
```

```
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
```

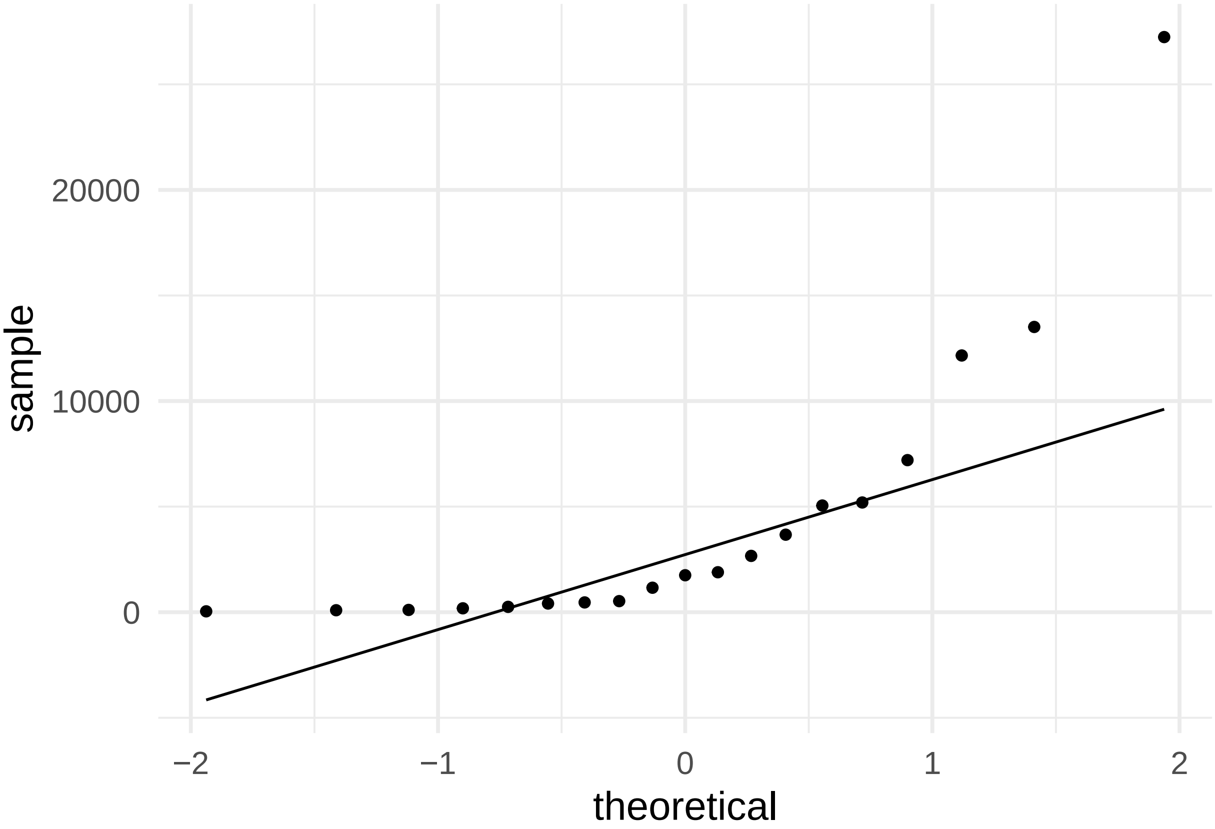
Another example

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



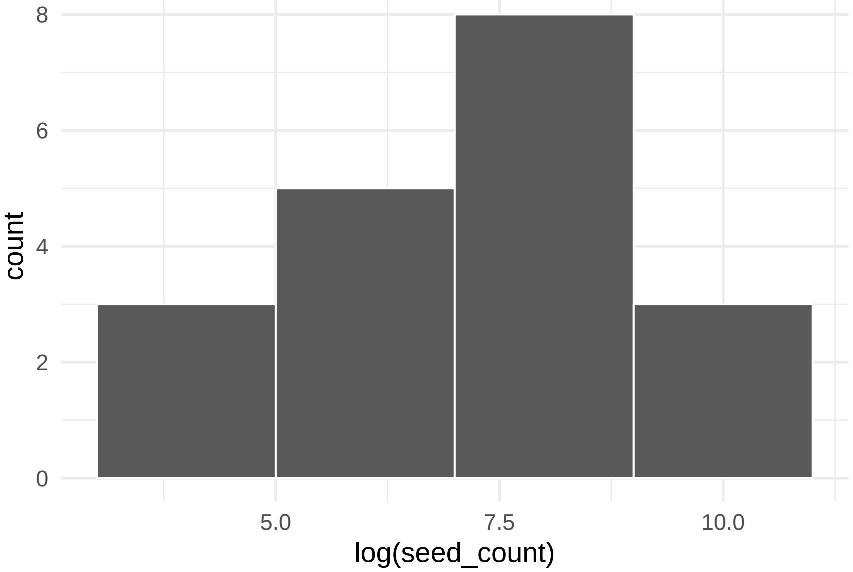
Another example

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



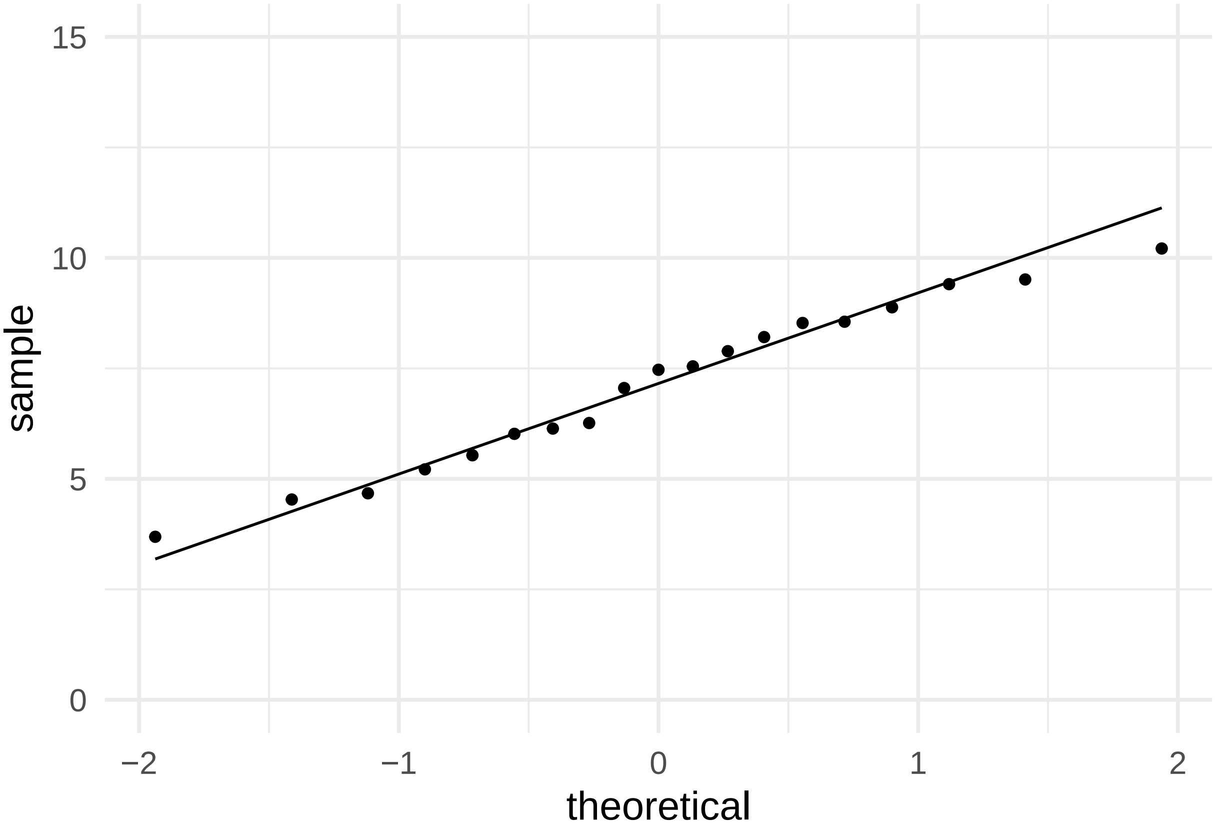
Another example (logged)

You might remember that we took the log of seed_count before we used it in regression. The log values look like this:



Another example (logged)

How does the QQ plot look for the logged variable?



QQ plot summary

- Review the QQ 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 QQplot. Notice that `aes(sample = var1)` is needed

Comic Relief