Comments for MEDB 5501, Week 6

Assessing normality

- Problems caused by non-normality
    - Poor confidence intervals, hypothesis tests
        - Too much imprecision
        - Poor coverage probability
            - Especially for one tailed tests
    - Inability to extrapolate
- What about the Central Limit Theorem?

Speaker notes

Sometimes, I think that researchers obsess too much about non-normality, but in fairness to them, it is an important issue. If your data does not follow a bell shaped curve, then several problems could happen.

First, you might see a greater degree of imprecision, reflected in very wide confidence intervals and loss of statistical power.

Second, you might have poor coverage probability. The 95% confidence interval might only reprent a 92% confidence level. The Type I error rate might be greater than 5%.

Third, and this point is not emphasized enough, is that non-normal data makes it difficult for you to extrapolate to future observations. Prediction of future events is a big part of Statistics. It is difficult even with normal data and becomes much more difficult with non-normal data.

Yes, you might say, but doesn’t the Central Limit Theorem help us out? Well, yes, if the sample size is large, but it only helps with assuring accurate confidence levels and good control of the Type I error rate. Non-normal data will still often produce intervals that are too wide and tests that have too little power.

How to handle non-normality

- Ignore it
  - Central Limit Theorem
- Transform your data
- Use alternatives
  - Nonparametric tests (covered in a later module)
  - Bootstrap (covered in a later module)
  - Randomization tests (not covered in this class)

Speaker notes

Yes, you might say, but doesn’t the Central Limit Theorem help us out? Well, yes, but it only helps with assuring accurate confidence levels and good control of the Type I error rate. Non-normal data will still often produce intervals that are too wide and tests that have too little power.

Still, there often is benefit in not worrying so much about non-normality. If your sample size is large and the deviations from normality are minor, don’t let this keep you up at night.

Even so, there are some benefits to addressing non-normality. Transformations, which I will cover in just a bit, can often help out tremendously. There are also alternatives to the simple tests using a sample mean: nonparametric tests and the bootstrap. I will cover those in a later module. There are also randomization tests, which I don’t think are covered. I’ve given a talk on randomization tests and I have a few webpages that talk about this topic.

Approaches to examine normality

- Histogram, boxplot
- Normal probability plot
  - Normal correlation (not covered, not recommended)
- Skewness, Kurtosis
- Kolmogorov-Smirnow test (not recommended)
- Shapiro-Wilk test (not recommended)

Speaker notes

There are many ways to check normality. The easiest one is a simple histogram. There’s a problem with the histogram in that you might draw different conclusions depending on how many bars you use. But it still is commonly used and easily understood.

The boxplot can also help distinguish normal versus non-normal data. It doesn’t do so well for certain types of non-normality, such as bimodal distributions, but it is also easy to use.

I’ll describe the normal probability plot in a bit. It requires a lot more interpretation, and you need to get some experience with it before you can get the feel of it. But it avoids some of the problems with the histogram and boxplot.

Some researchers advocate the test of a correlation based off of the normal probability plot as a measure of normality. I do not recommend this test.

There are two summary statistics, skewness and kurtosis, that are also commonly used.

There are two tests that I also do not recommend, the Kolmogorov-Smirnow test and the Shapiro-Wilk test

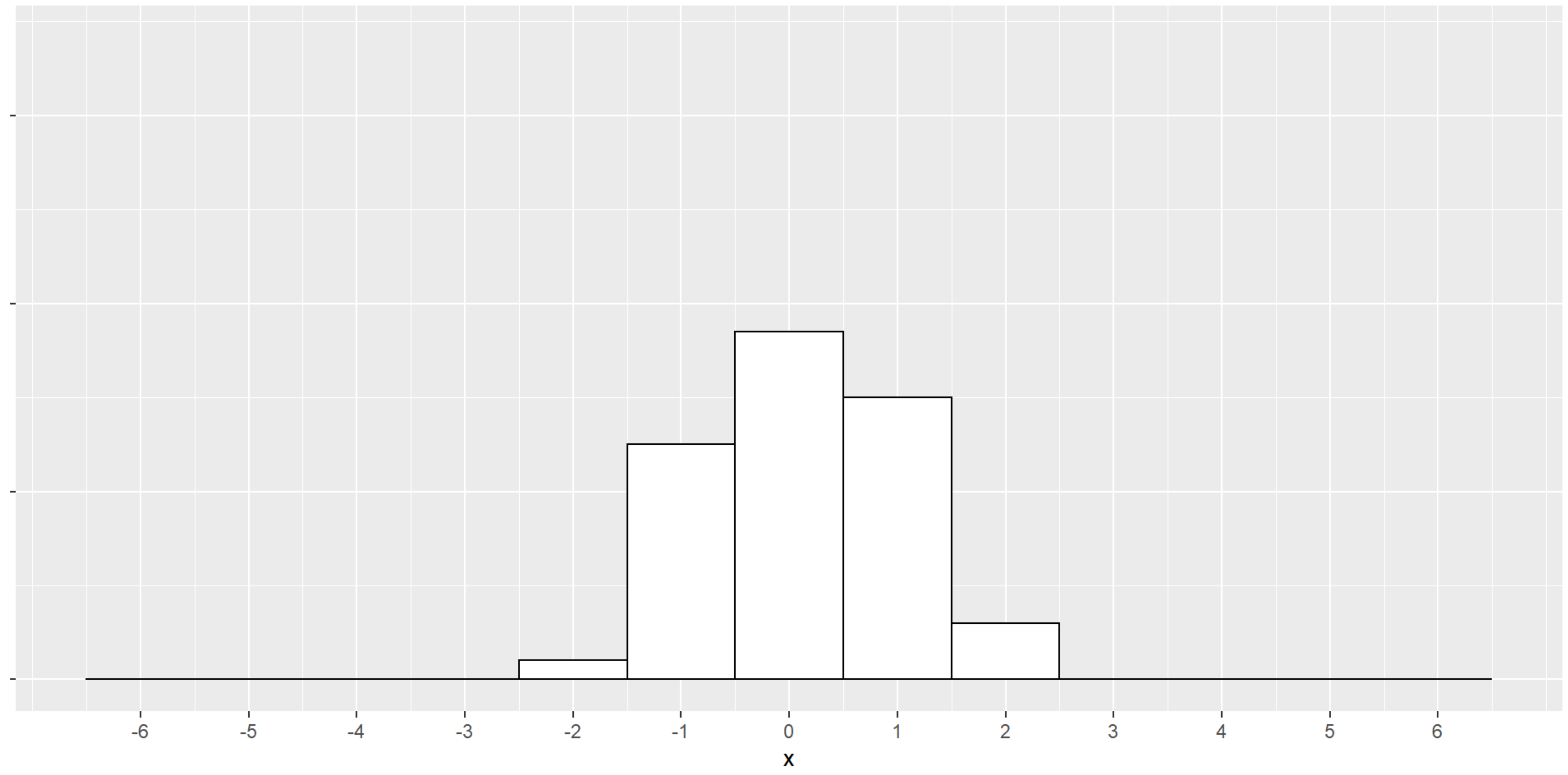
Never use p-values to test normality

- H0: Data comes from a normal population
- Why is this bad?
  - ASA recommendation against the use of p-values
  - Too little power for small sample sizes
  - Too much power for large sample sizes
  - Ignore the type of non-normality

Speaker notes

There has been a lot of criticisms of the excessive use of p-values, including a statement from the American Statistical Association. While I do not feel you should NEVER use p-values, I do think that you can reduce the number of times you use them. And p-values associated with normality tests is one of them.

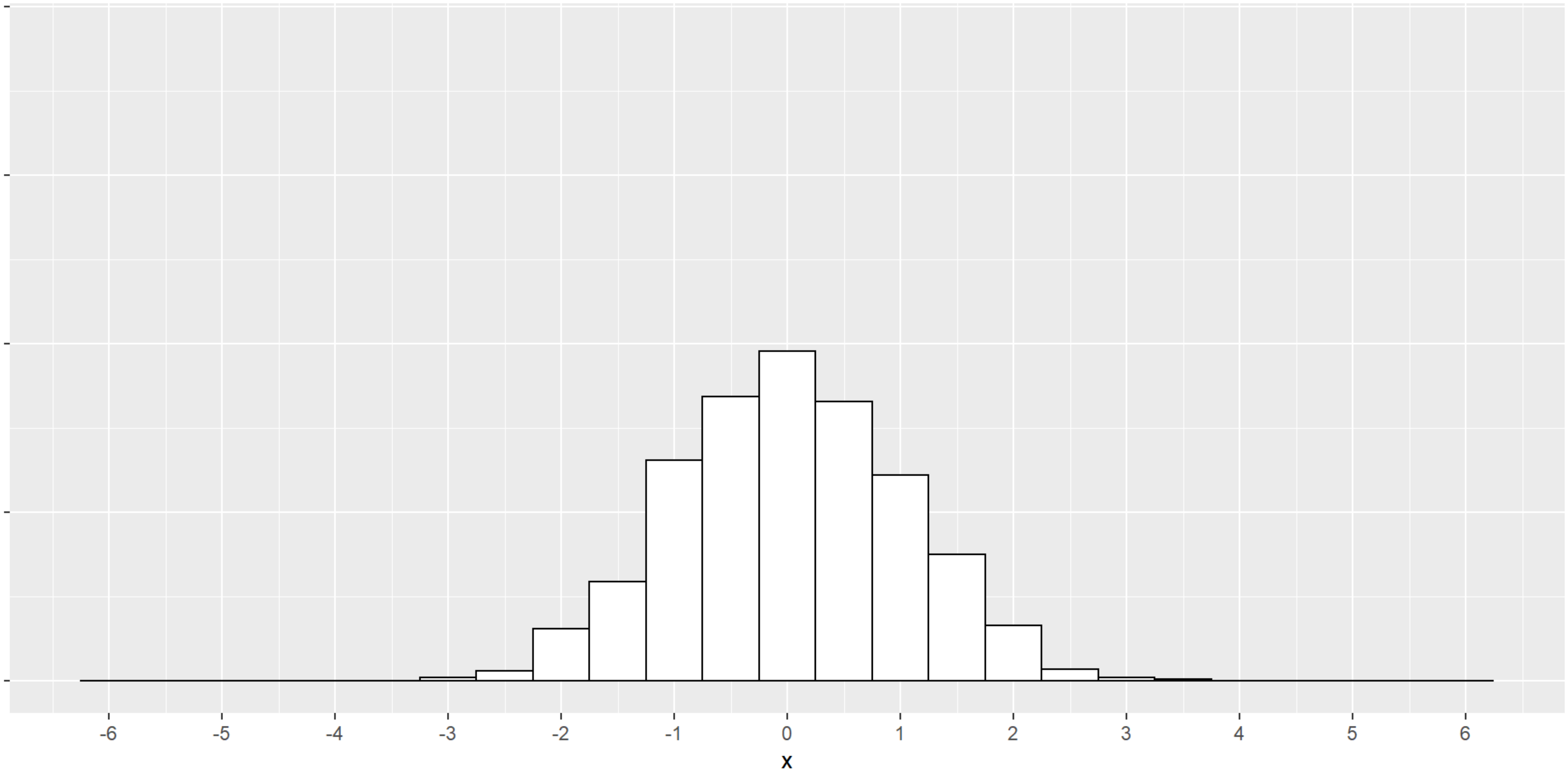
Normal histogram with n=100



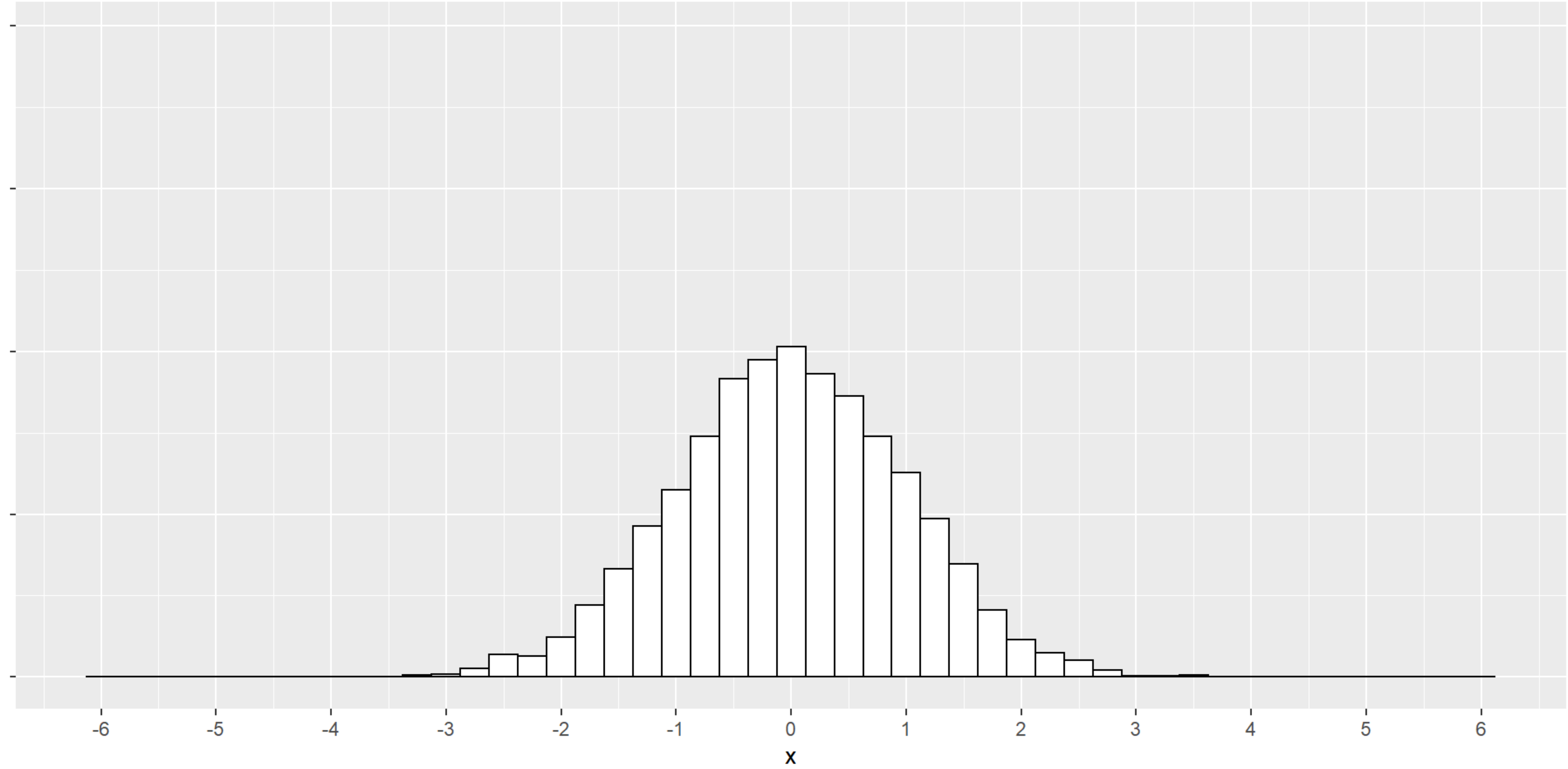
Speaker notes

I’m going to show a sequence of histograms where the sample size increases and the width of the bars simultaneously decreases.

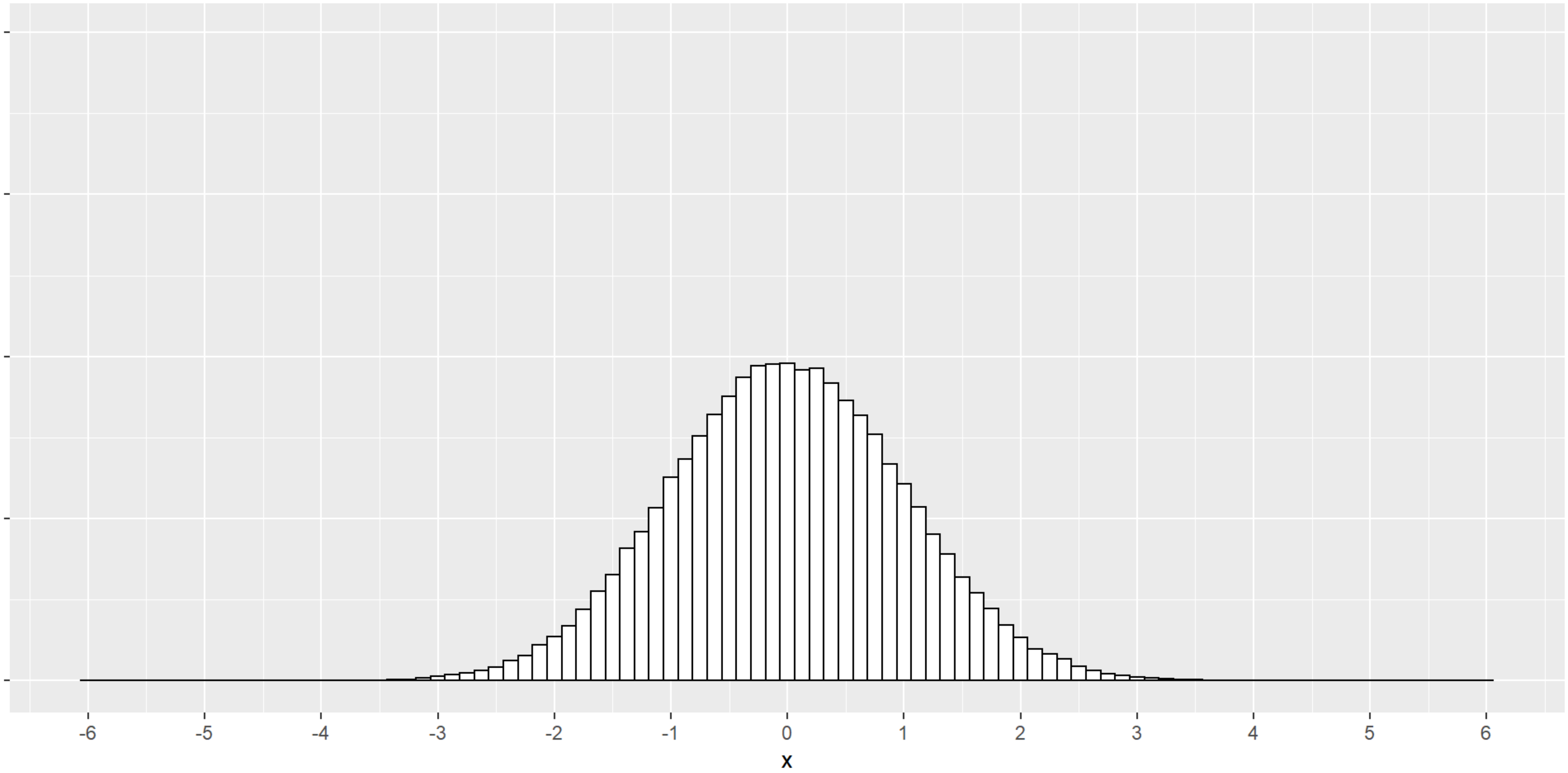
Normal histogram with n=1,000



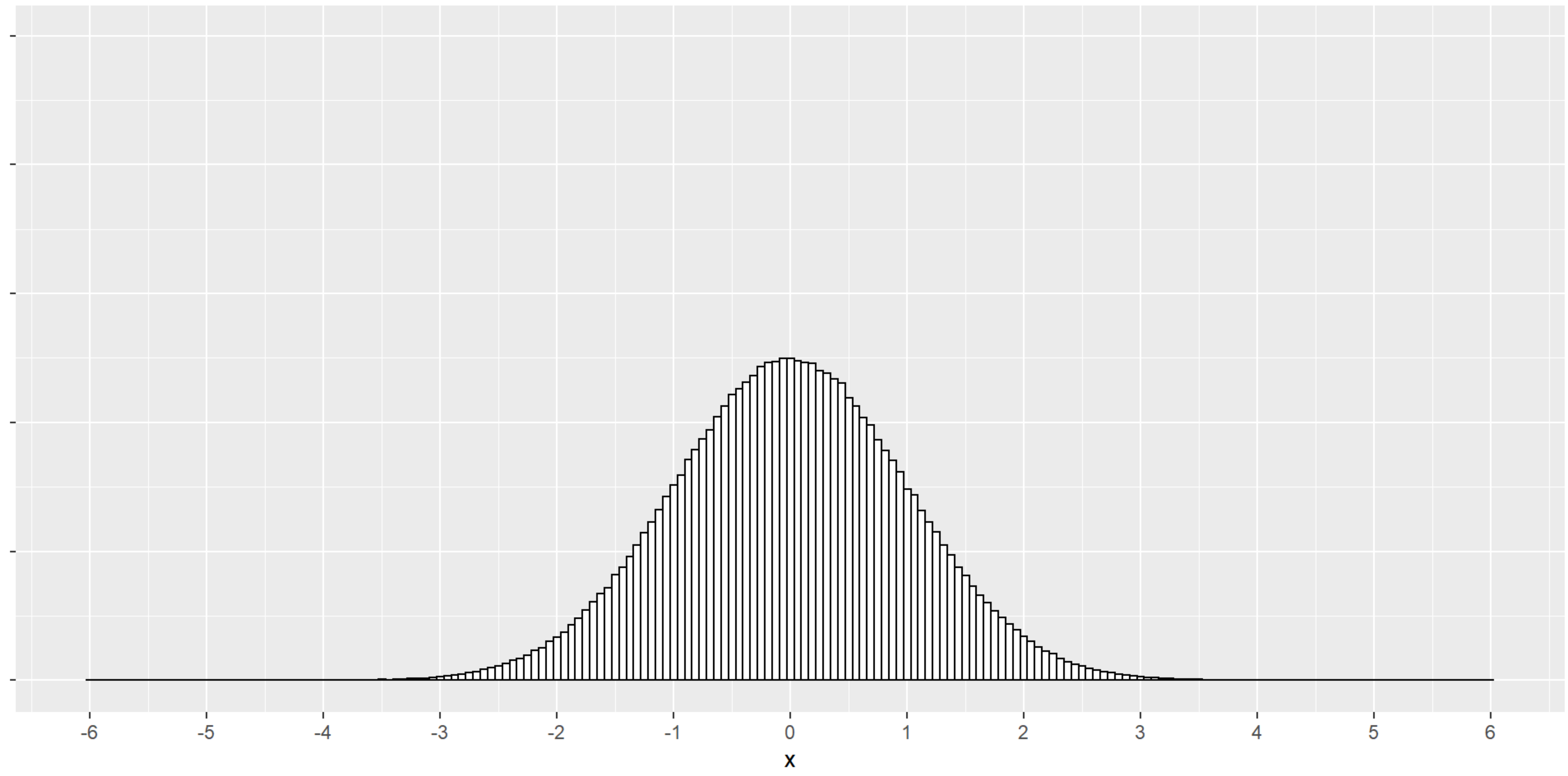
Normal histogram with n=10,000



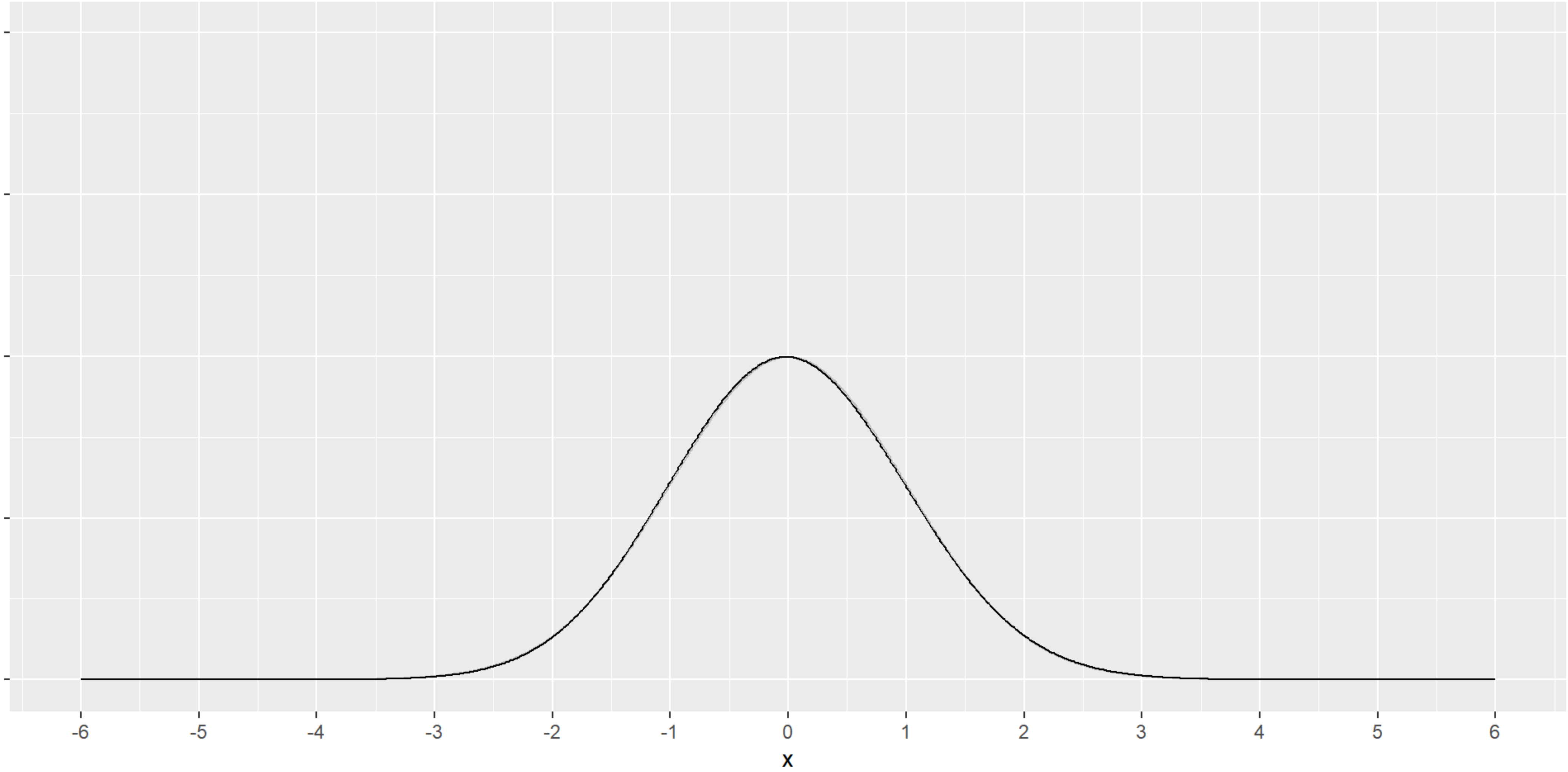
Normal histogram with n=100,000



Normal histogram with n=1,000,000



Normal distribution (n=infinity)



Speaker notes

It’s impossible to show an infinite number of bars with zero width, but if you could, it would look like this.

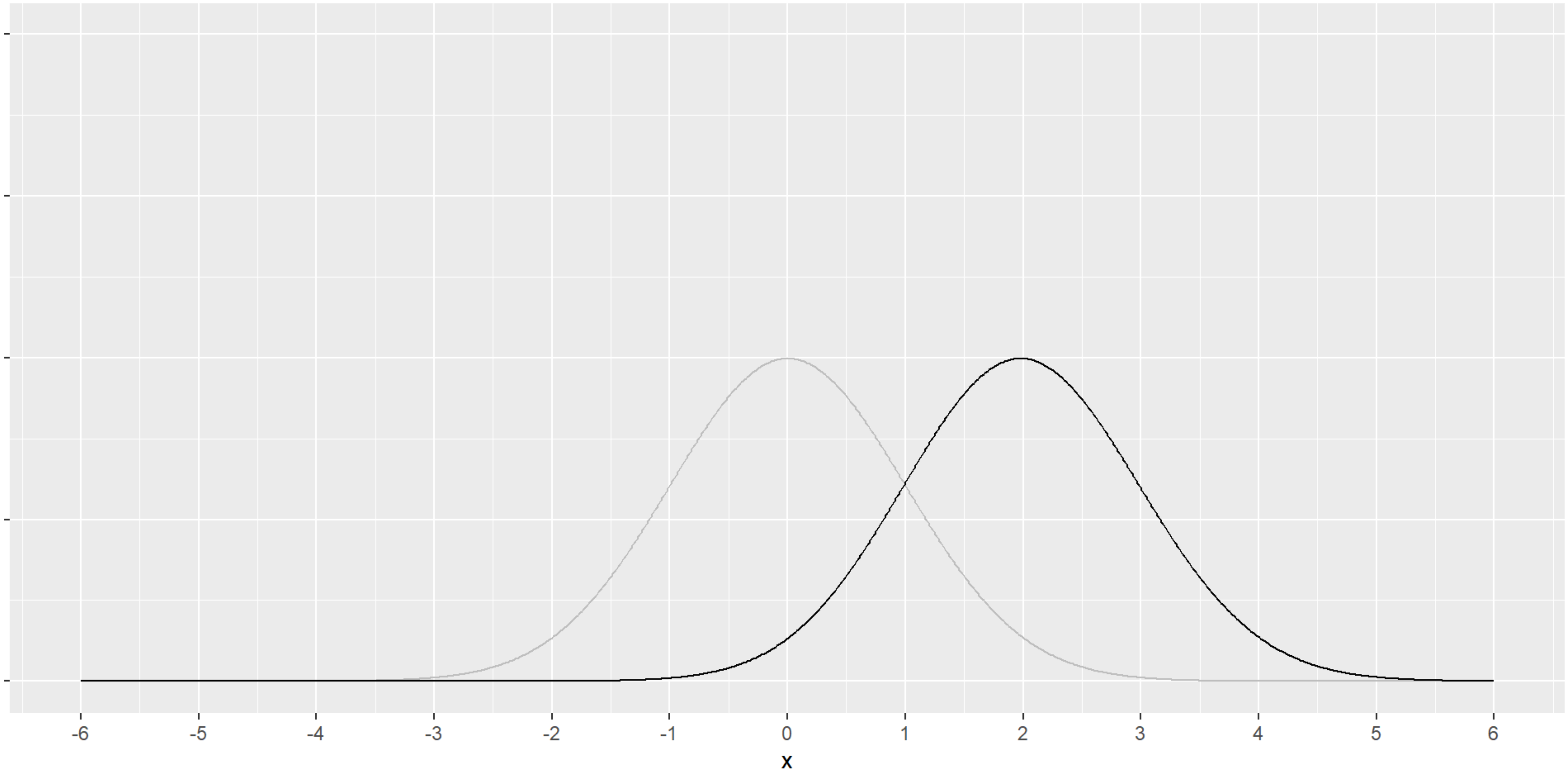
This is the curve that you see in your text box and many other sources. It is a theoretical curve becuase

a. it represents a mathematical limit, and

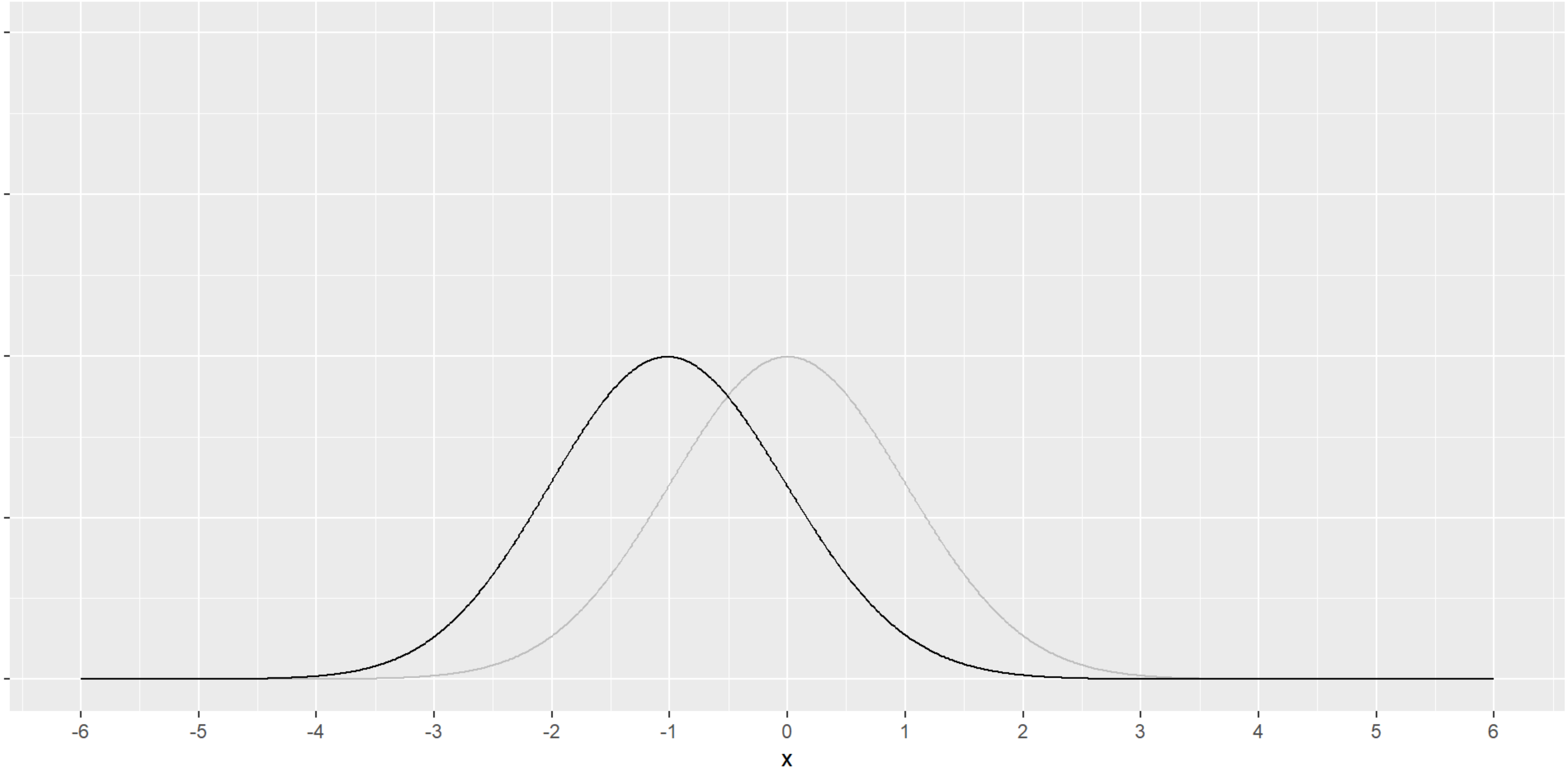
b. no reall world data will ever match the smoothness and symmetry that you see here. Some data sets do come close, though.

This curve is for the standard normal distribution, a distribution with mu=0 and sigma=1. There are other distributions, which I will show in the upcoming slides.

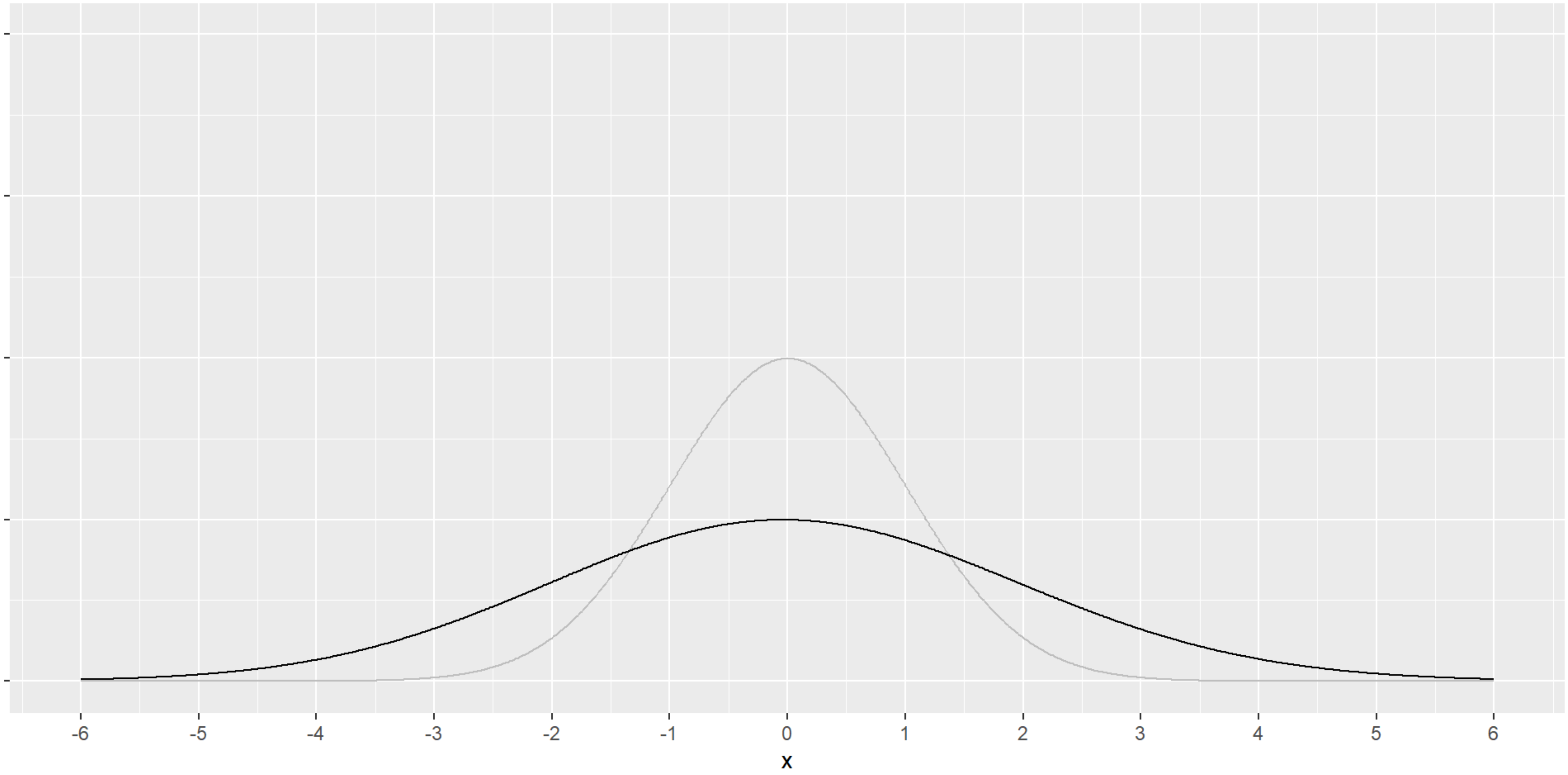
Normal(2, 1)



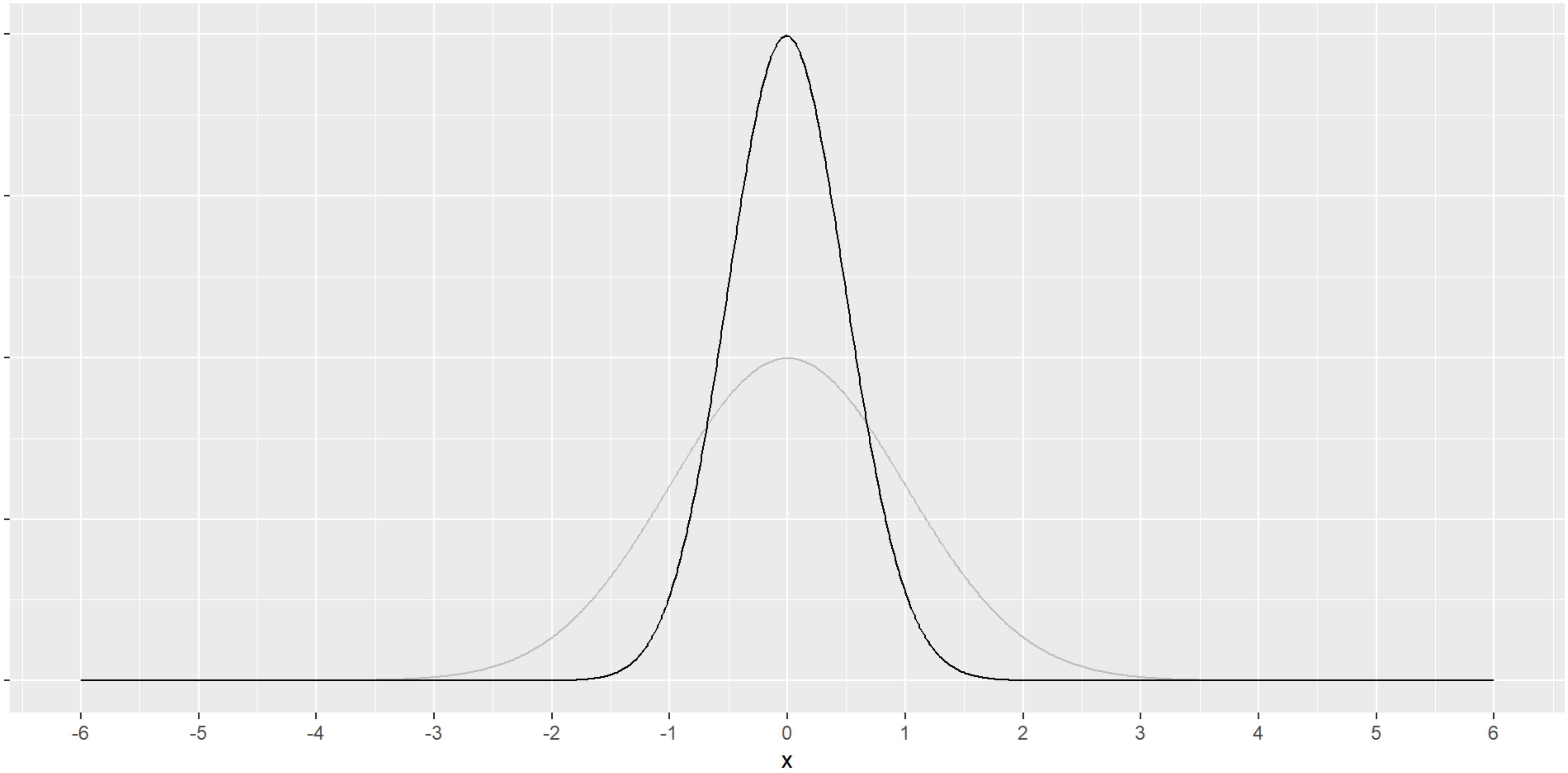
Normal(-1, 1)



Normal(0, 2)



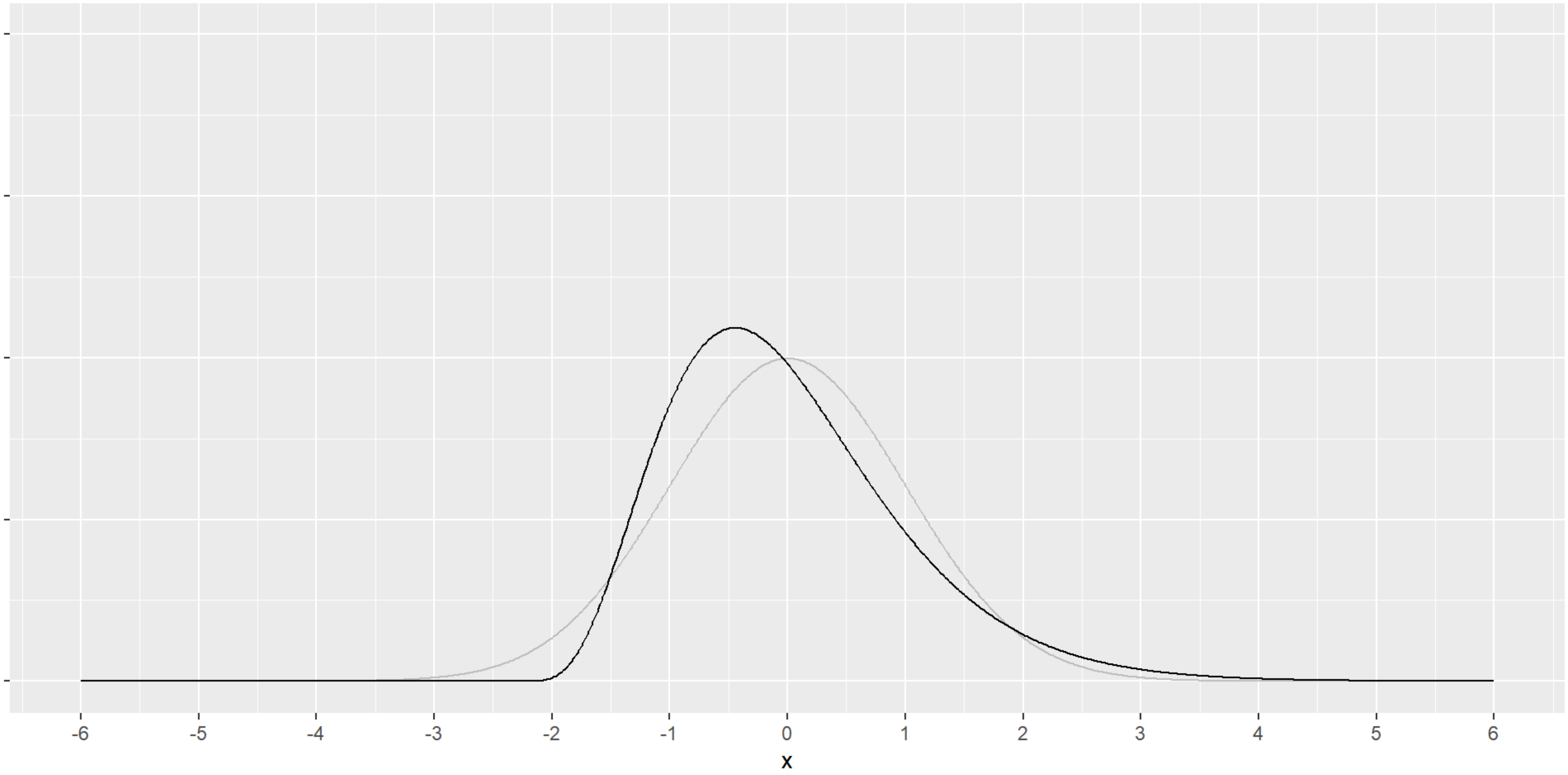
Normal(0, 0.5)



Speaker notes

You can stretch of squeeze the normal curve, you can shift it to the left or to the right, but it still maintains a characteristic shape.

Skewed right



Speaker notes

Here’s an example of a distribution that is skewed to the right. I have scaled this distribution so that the mean is zero and the standard deviation is one.

There are three key features for a skewed right distribution.

First, the chances of seeing an outlier on the low end (more than two standard deviations below the mean) is less for a skewed right distribution. Skewed right distributions almost never produce outliers on the low end.

Second, the chances of seeing an outlier on the high end (more than two standard deviations above the mean) is greater for a skewed right distribution.

Third, the probability of any value being less than the mean is a bit more than 50%. That’s because the outliers on the high end of this distribution tend to have an undue influence on the mean, making it bigger than the bulk of the data.

Right skewness is characterized by the tails of the distribution

- Heavy right tail
  - Greater tendency to produce extreme values on the right
- Light left tail
  - Lesser tendency to produce extreme values on the left
- Right skewness is the most common type of non-normality

Speaker notes

A right skewed distribution means a heavy tail on the right and a light tail on the left. A heavy right tail means lots of probability associated with extreme values on the right (the high end). A light left tail means much less probability associated with extreme values on the left (the low end).

Recognize this pattern right away because it is the most common deviation from normality.

Normal probability plot

- Compare data to evenly spaced percentiles of the normal distribution
- Example with n=4
  - Compare smallest value with $Z_{0.2}$
  - Compare next value with $Z_{0.4}$
  - Compare next value with $Z_{0.6}$
  - Compare largest value with $Z_{0.8}$
- No best definition for evenly spaced
  - 12.5, 37.5, 62.5, 87.5, for example

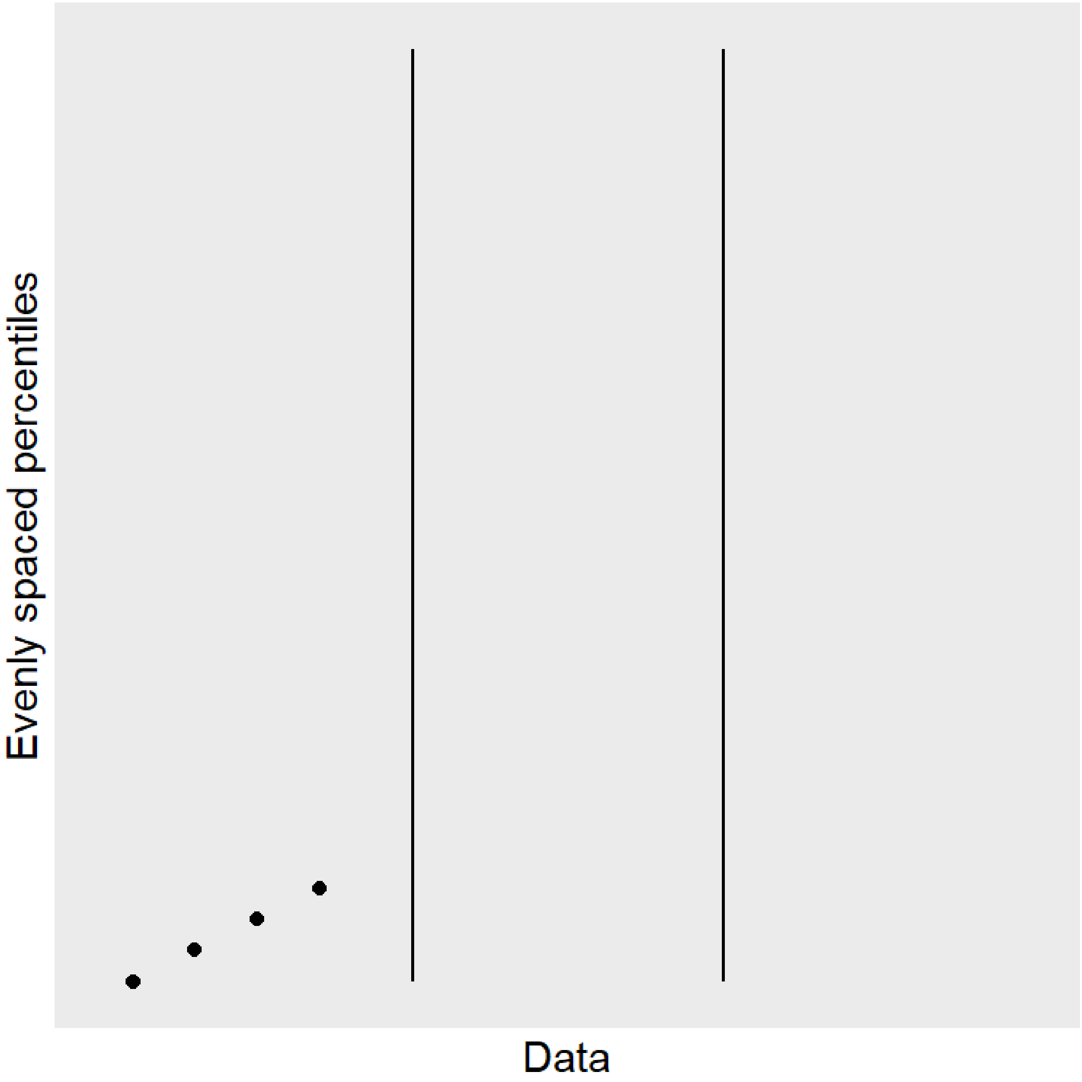
Speaker notes

The normal probability plot is an excellent way to examine whether data is normally distributed. Compare each value to the corresponding percentile of a normal distribution.

For a simple example with four data points, compare your values to the 20th, 40th, 60th, and 80th percentiles.

There are alternative ways to define evenly spaced. They might lead to slightly different plots, but they are not worth fussing over. Any definition of evenly spaced works just fine.

Interpreting, heavy left tail

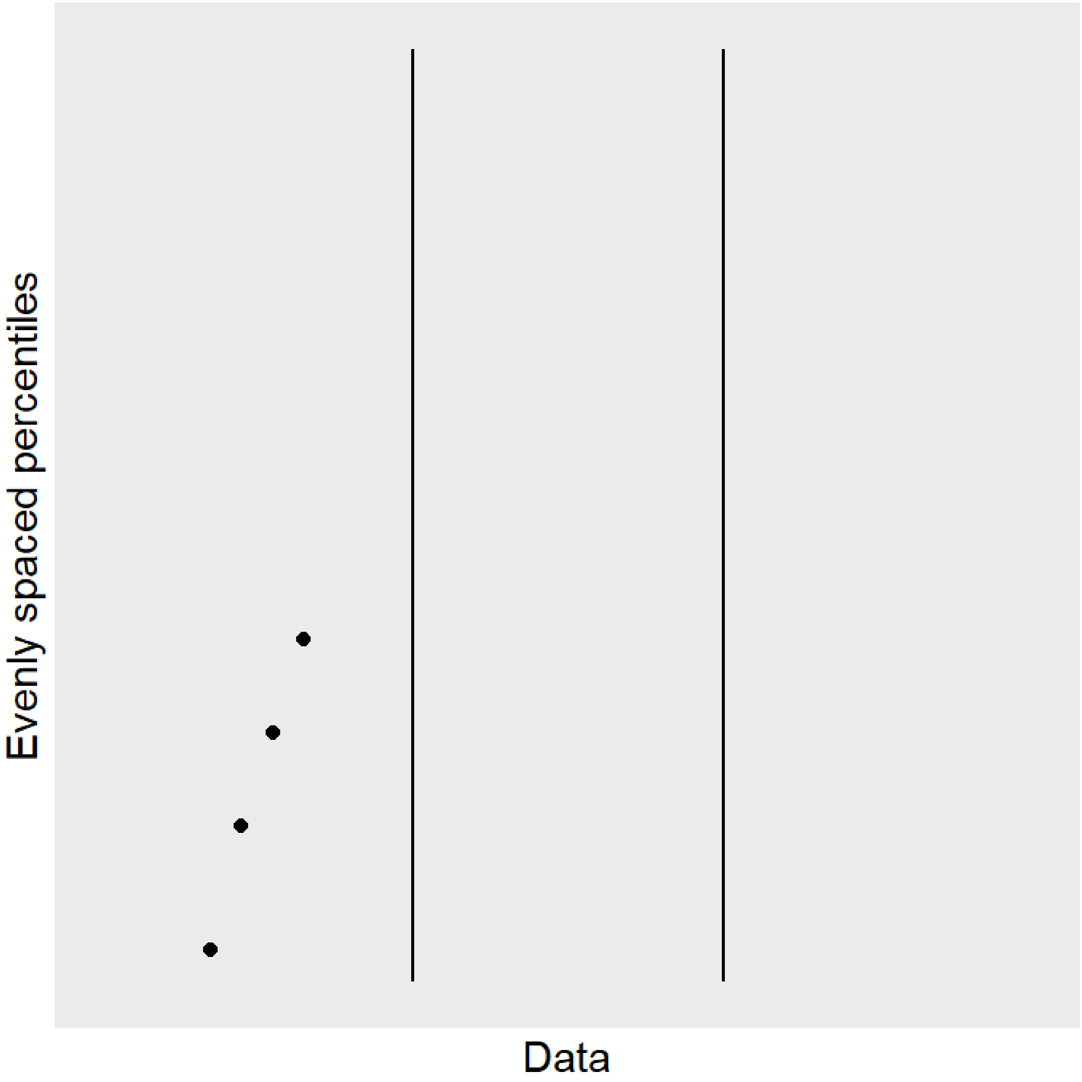


Speaker notes

Interpreting the normal probability plot is a bit tricky. I like to look at the behavior on the left third of the data then the right third of the data.

If the left third of the data is relatively flat (less than 45 degrees), this is an indication that the extreme values on the low end are moving faster than the normal percentiles, indicating a tendency towards extreme values on the left. This is evidence of a heavy left tail.

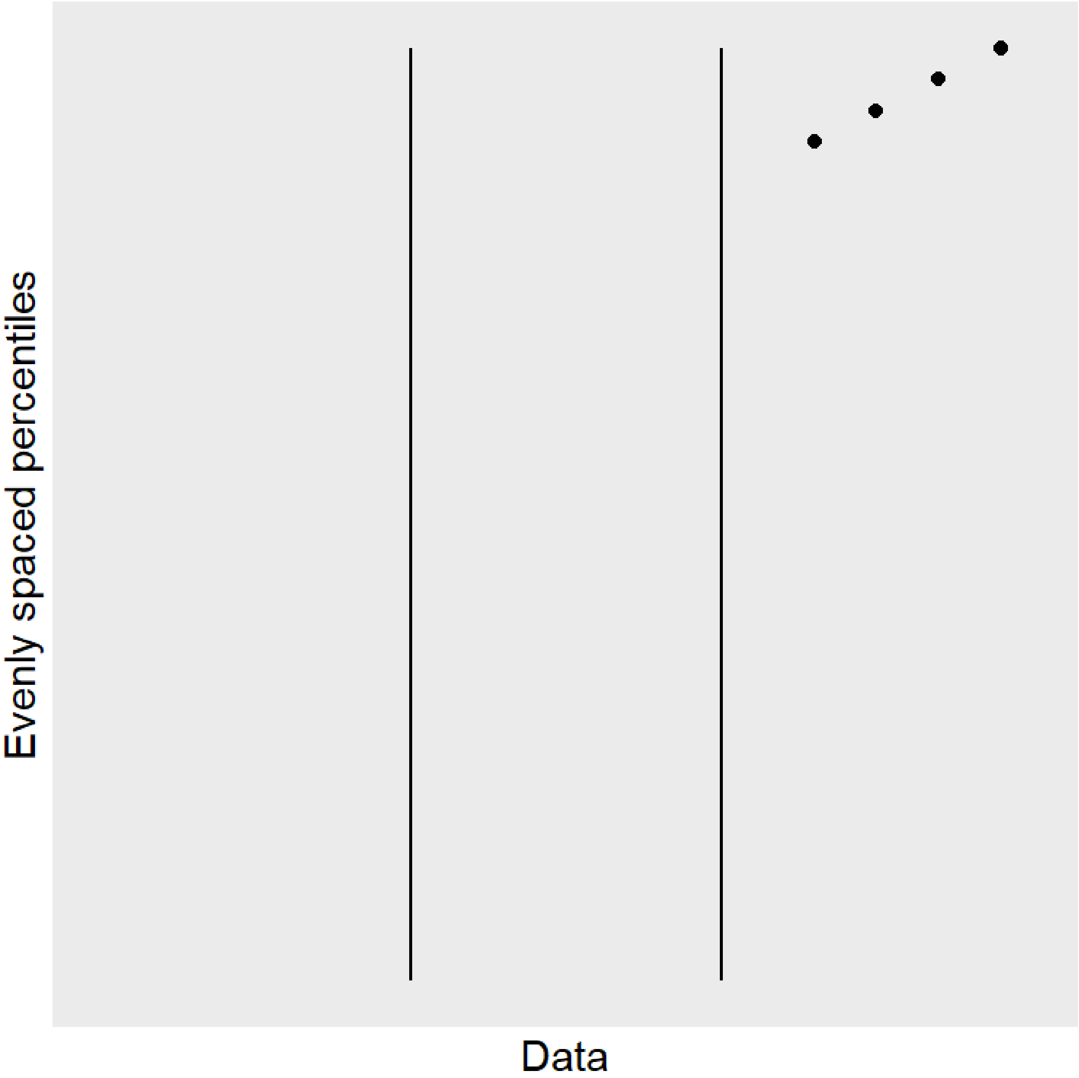
Interpreting, light left tail



Speaker notes

In contrast, if the left third of the data is relatively steep (greater than 45 degrees), this is an indication that the extreme values on the low end are moving slower than the normal percentiles, indicating a tendency towards little or no extreme values on the left. This is evidence of a light left tail.

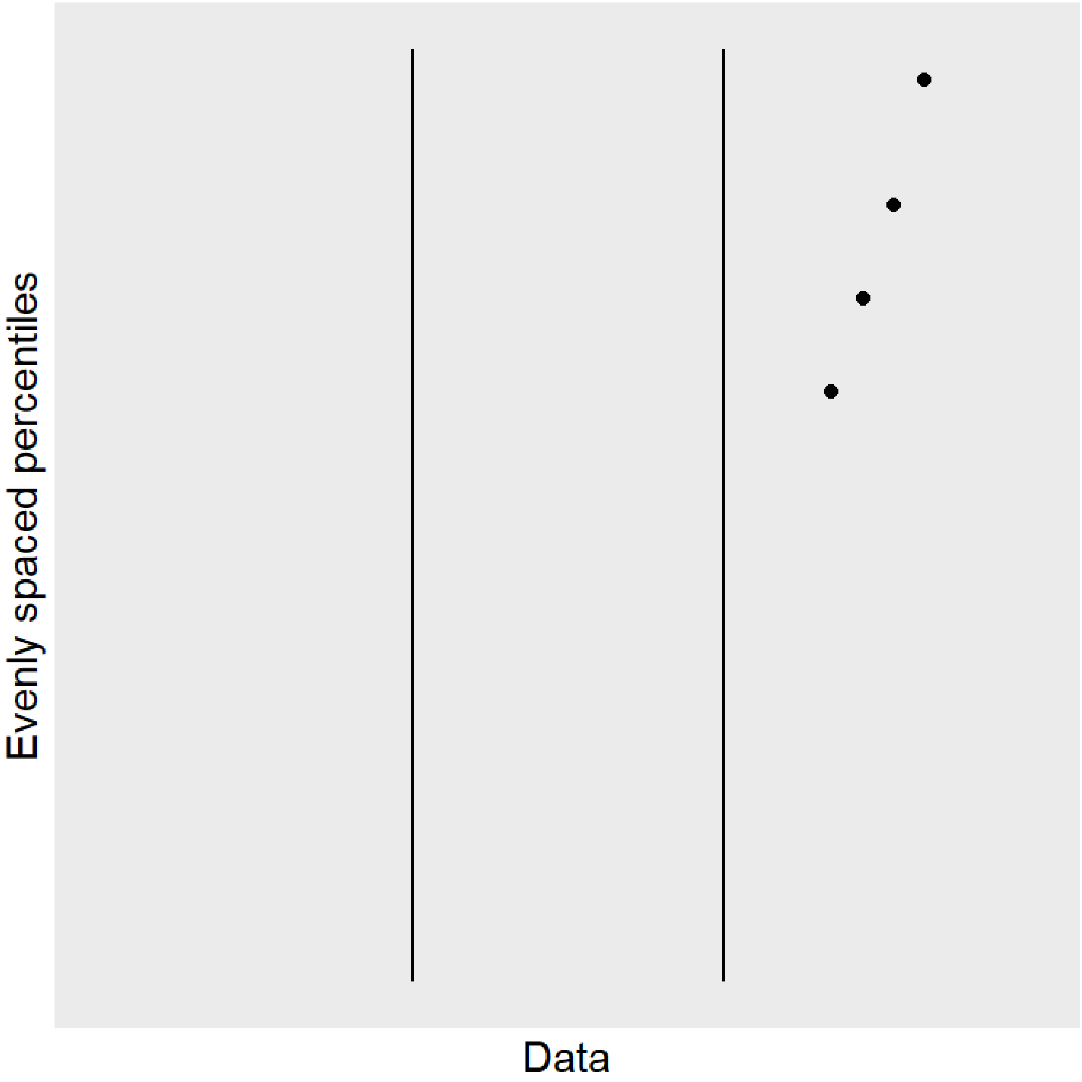
Interpreting, heavy right tail



Speaker notes

Look for the same behavior on the right side. If the right third of the data is relatively flat (less than 45 degrees), this is an indication that the extreme values on the high end are moving faster than the normal percentiles, indicating a tendency towards extreme values on the right. This is evidence of a heavy right tail.

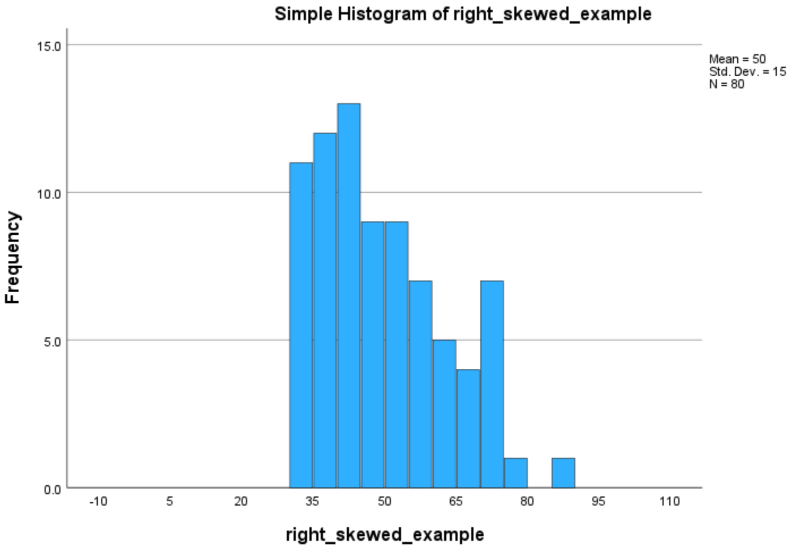
Interpreting, light right tail

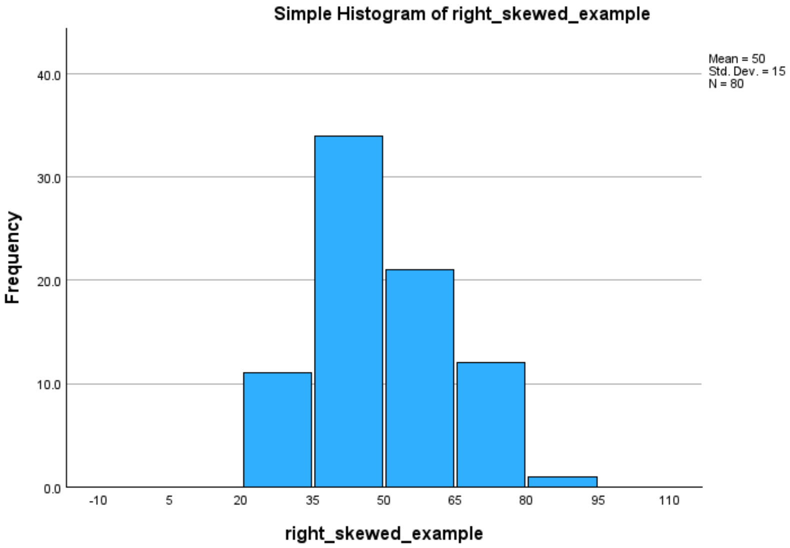


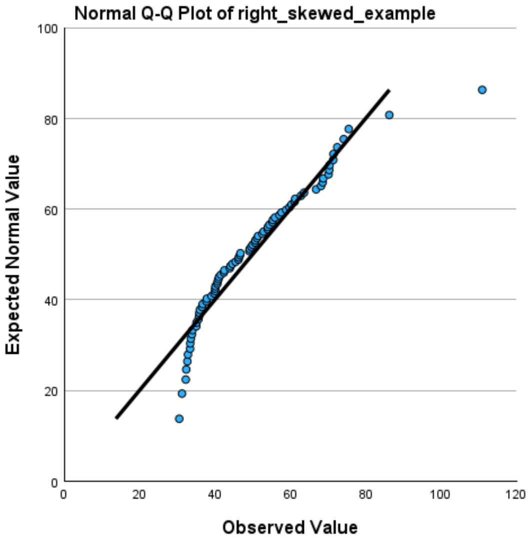
Speaker notes

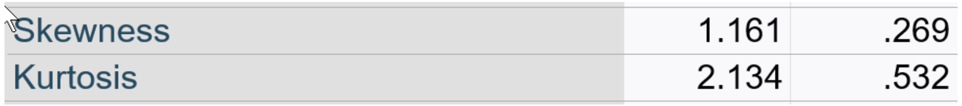
Finally, if the right third of the data is relatively steep (greater than 45 degrees), this is an indication that the extreme values on the high end are moving slower than the normal percentiles, indicating a tendency towards little or no extreme values on the right. This is evidence of a light right tail.

Right skewed data







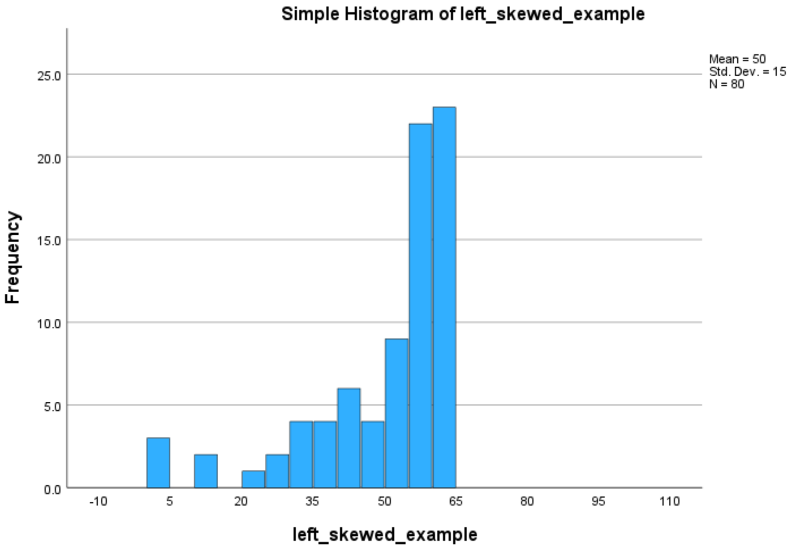


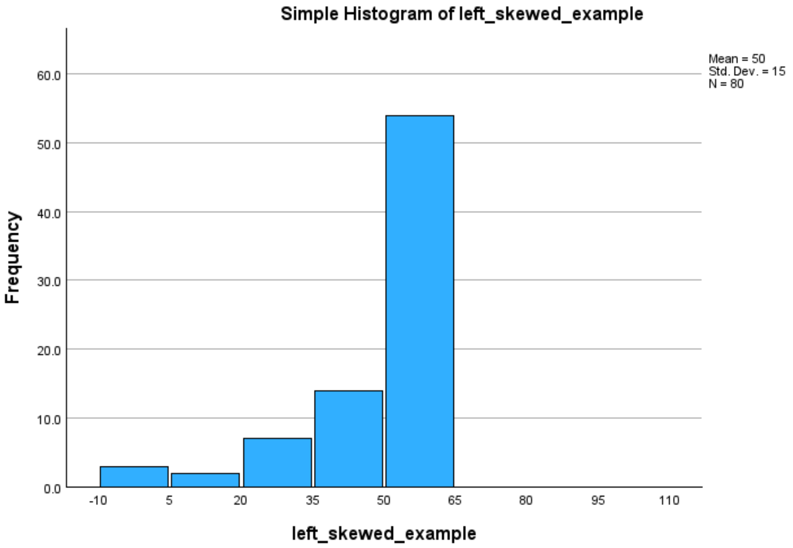
Skewness	1.161	.269
Kurtosis	2.134	.532

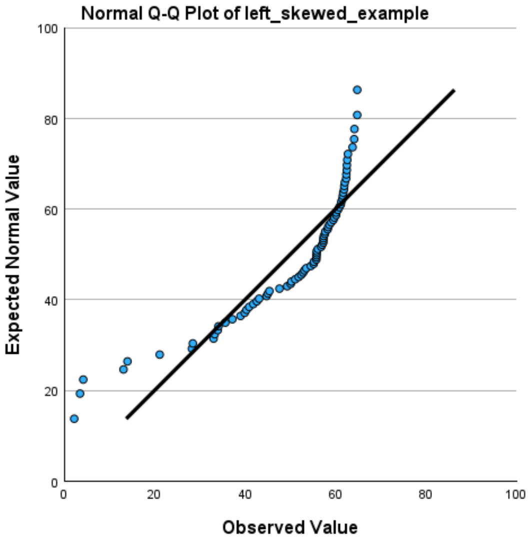
Speaker notes

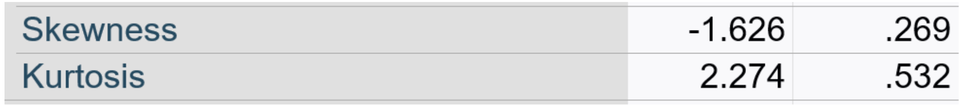
Here is some artificial data the has a right skew.

Left skewed data







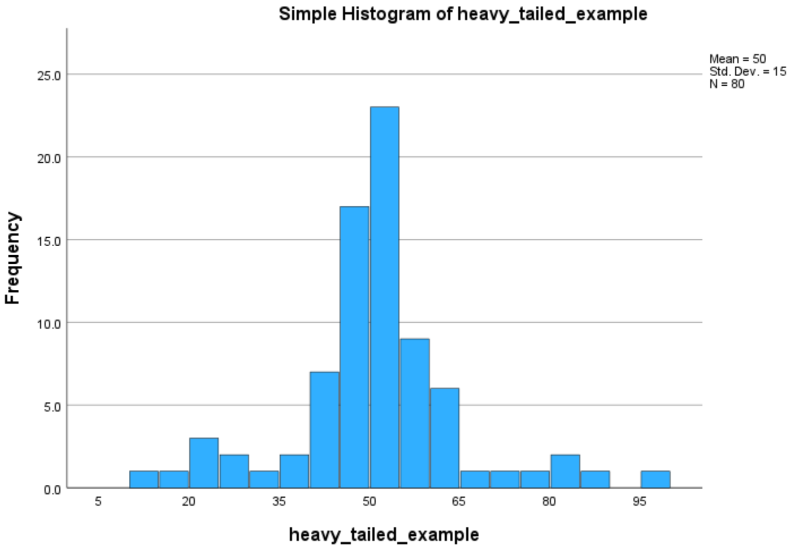


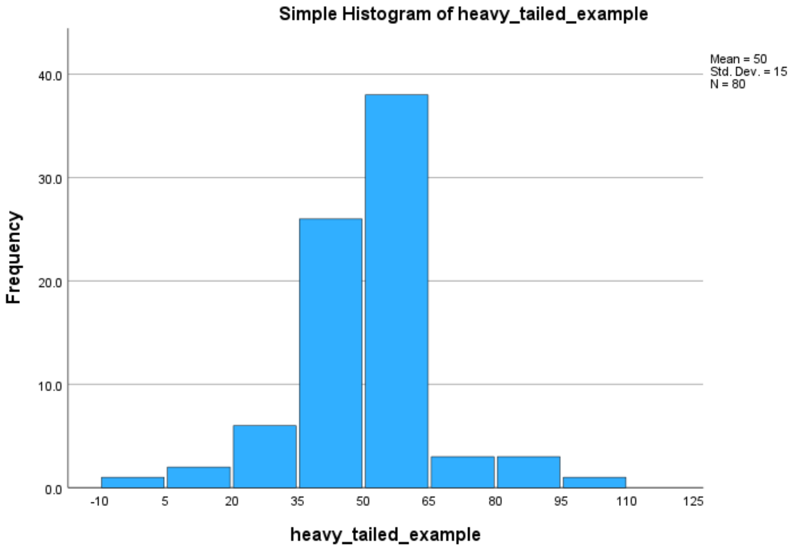
Skewness	-1.626	.269
Kurtosis	2.274	.532

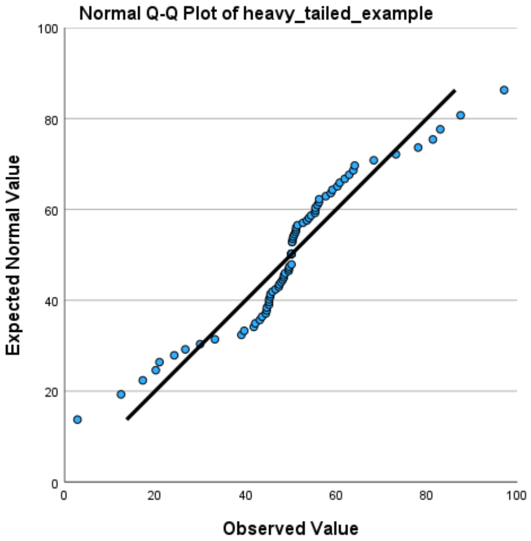
Speaker notes

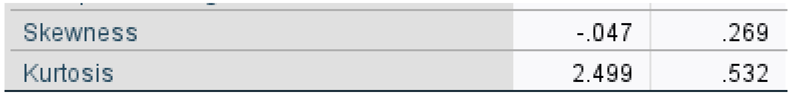
It is less common, but sometimes you may see left skewed data. This is heavy tailed on the left and light tail on the right. This means that if you see outliers, they tend to appear on the left (the low end).

Heavy tailed data







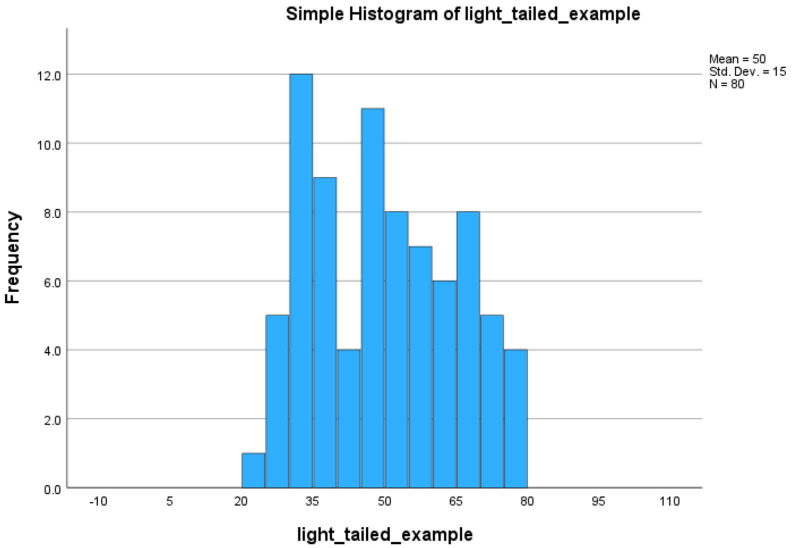


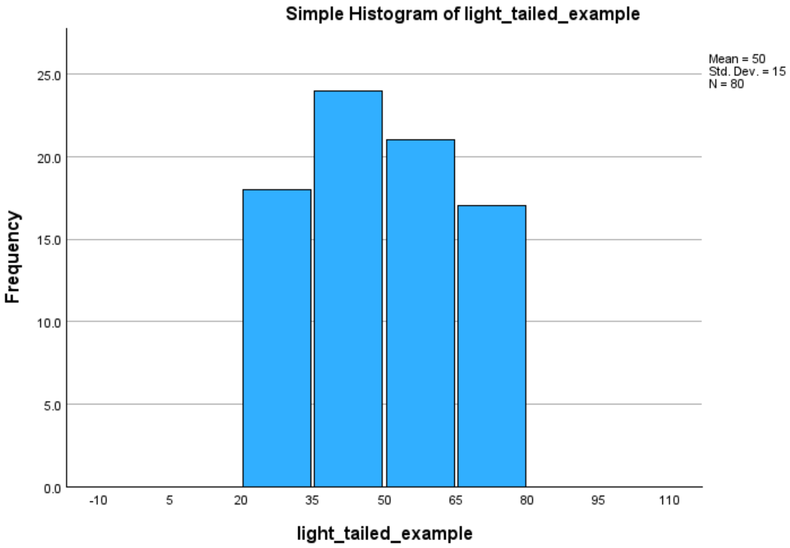
Skewness	-.047	.269
Kurtosis	2.499	.532

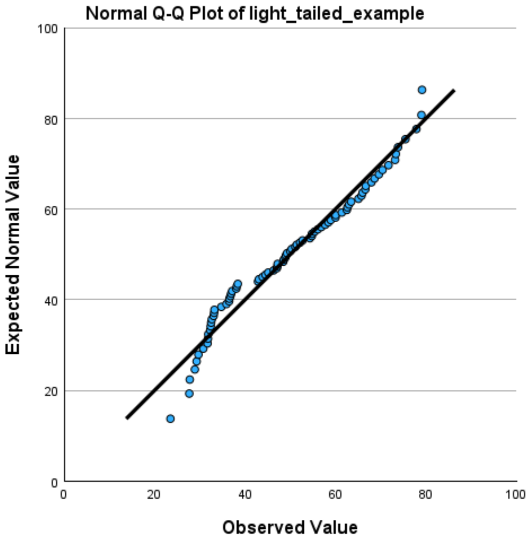
Speaker notes

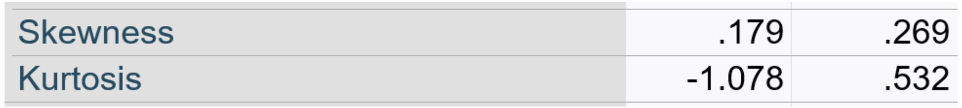
Sometimes you see an excess of extreme values on both ends. This is a heavy tail on the left and a heavy tail on the right.

Light tailed data







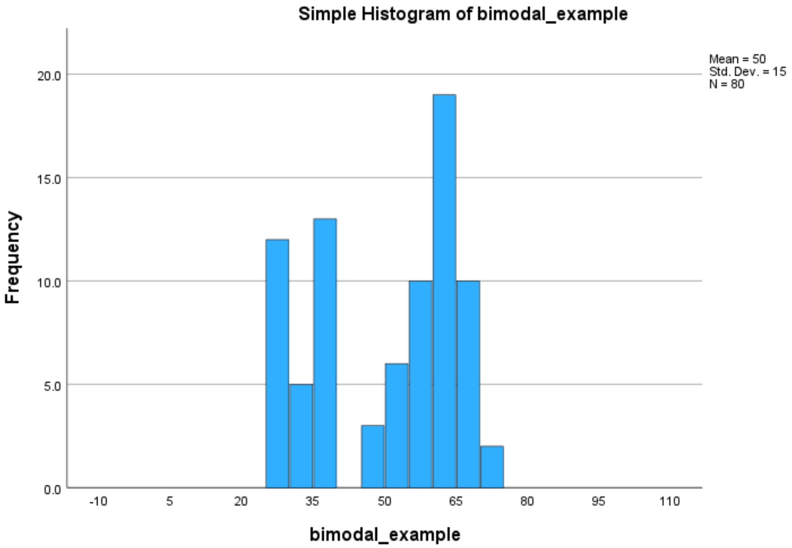


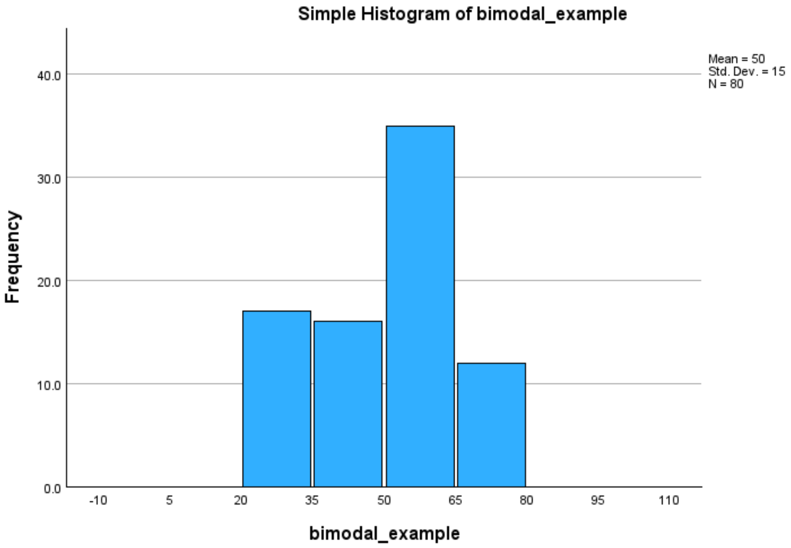
Skewness	.179	.269
Kurtosis	-1.078	.532

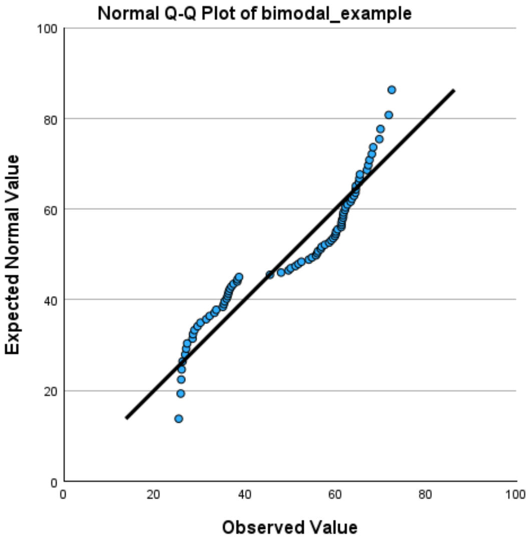
Speaker notes

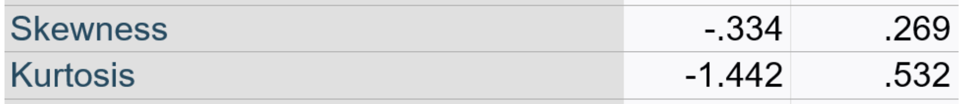
You might see a light tailed distribution. This means there are fewer extremes at either end. The data just drops off suddenly in both directions.

Bimodal data







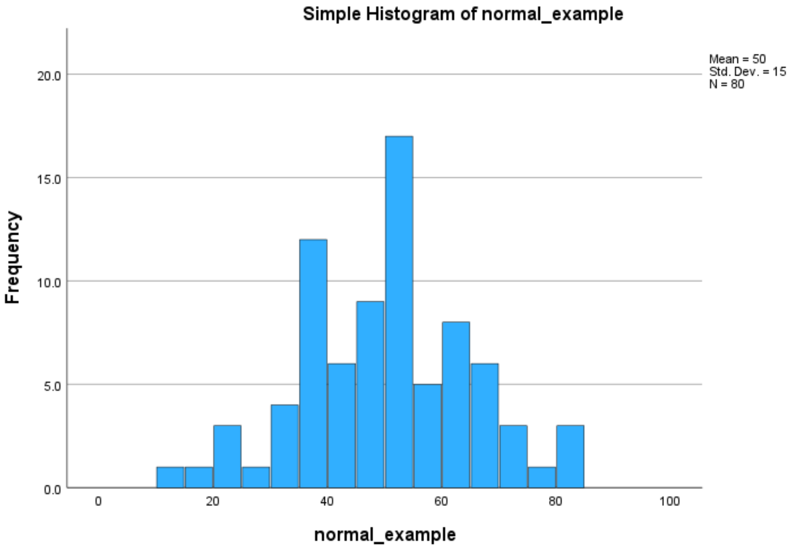


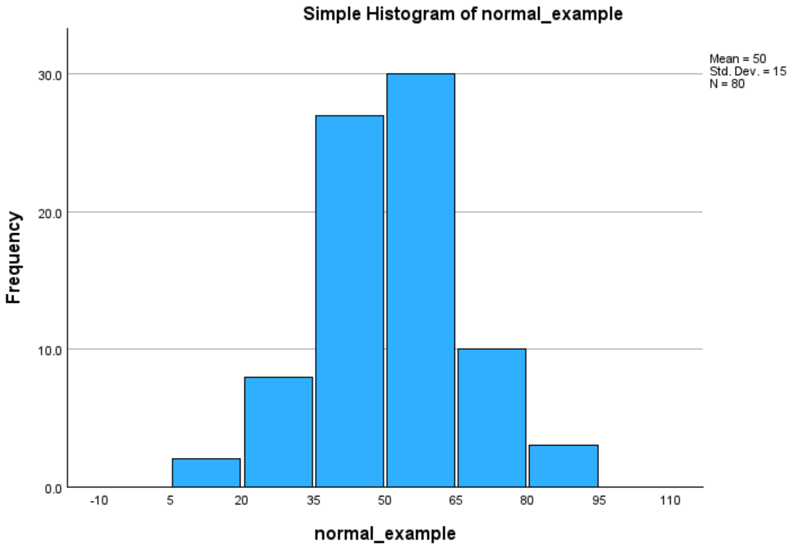
Skewness	-.334	.269
Kurtosis	-1.442	.532

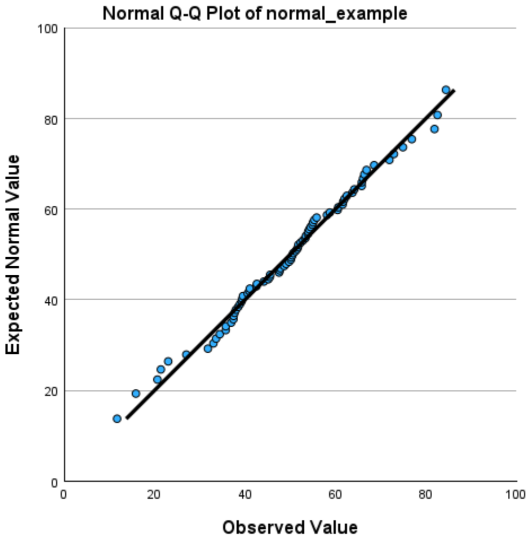
Speaker notes

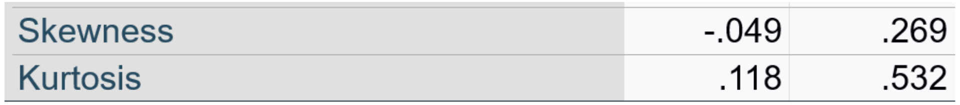
Bimodal data is tricky to diagnose with a histogram because what looks bimodal with a large number of bars can look quite different with fewer bars.

Normal data









Skewness	-.049	.269
Kurtosis	.118	.532

Speaker notes

Not everything is non-normal. Here is what you would see with the various diagnostics for normal data.

Log transformation

- If $a^b = c$, then $log_a(c) = b$
- $log(a \times b) = log(a) + log(b)$
- Three commonly used bases
  - log base 10: (log_{10})
  - log base 2: (log_2)
  - natural log, log base e: (ln)
- Important! SPSS uses lg10, not log10.

Speaker notes

I don’t use a lot of mathematics in this class, but I do want to talk about the logarithm function.

Where the log transformation is routine

- Log units are common in science
    - Richter scale (1 unit equals 10 fold change)
    - Decibel (20 units equals 10 fold change)

Speaker notes

You might be familiar with several measures that are represented on a log scale.

Why use the log function

- Stretches small values
- Squeezes large values
- Possible benefits
  - Removing skew
  - Eliminating outliers
  - Stabilizing variation
  - Model simplification

Speaker notes

The logarithm function tends to squeeze together the larger values in your data set and stretches out the smaller values. This squeezing and stretching can correct one or more of the following problems with your data.

Not all data sets will suffer from these problems. Even if they do, the log transformation is not guaranteed to solve these problems. Nevertheless, the log transformation works surprisingly well in many situations.

Furthermore, a log transformation can sometimes simplify your statistical models. Some statistical models are multiplicative: factors influence your outcome measure through multiplication rather than addition. These multiplicative models are easier to work with after a log transformation.

When to use a log transformation

- Data bounded below by zero
- Data defined as a ratio
- Max / Min > 3

Speaker notes

There are several things that you should look for when deciding whether to use a log transformation.

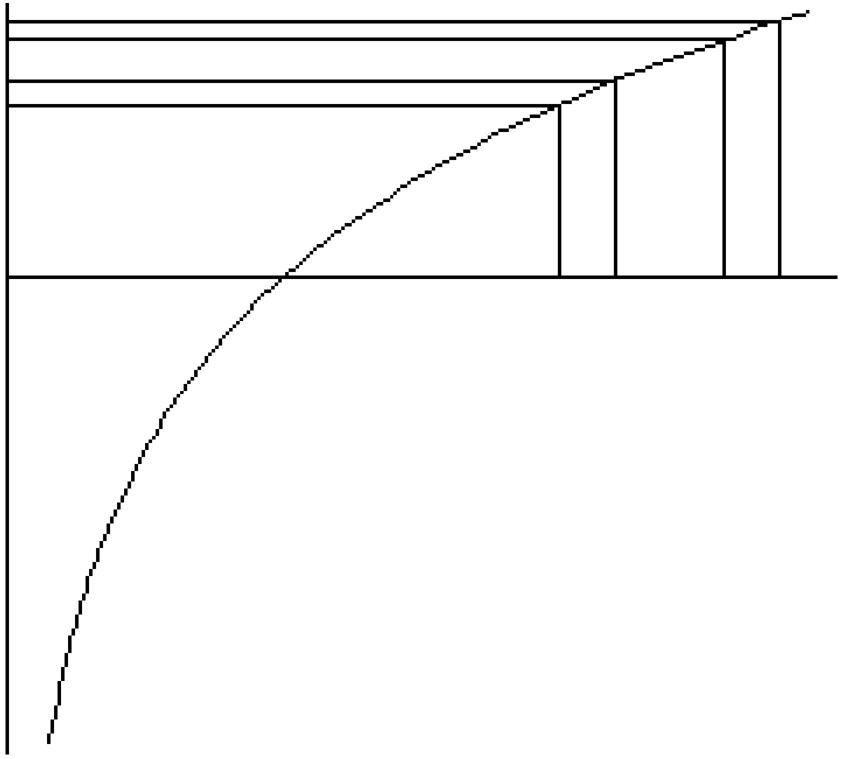
Data that is bounded below by zero is the first thing to look for, especially when your data comes close to, but never equals zero. The lower bound will often cause skewness because outliers can still appear on the high end, but not on the low end.

You should especially consider using the log transformation for data that represents ratios. Ratios are usually bounded below by zero. There is also an asymmetry in ratios. Ratios where then numerator is less than the denominator are confined to be between 0 and 1. When the numerator is larger than the denominator, the ratio can roam freely between 1 and infinity.

The log transformation works for measures other than ratios, so you should still consider it for other types of outcomes.

There's an informal rule about the log transformation, though, that you should consider. Use it only when you have a wide spread of data, where the largest value in your data is more than three times as large as the smallest values in your data. If your data ranges from 200 to 300, the ratio of 1.5 indicates that you are unlikely to see much change with the log transformation.

Squeezing

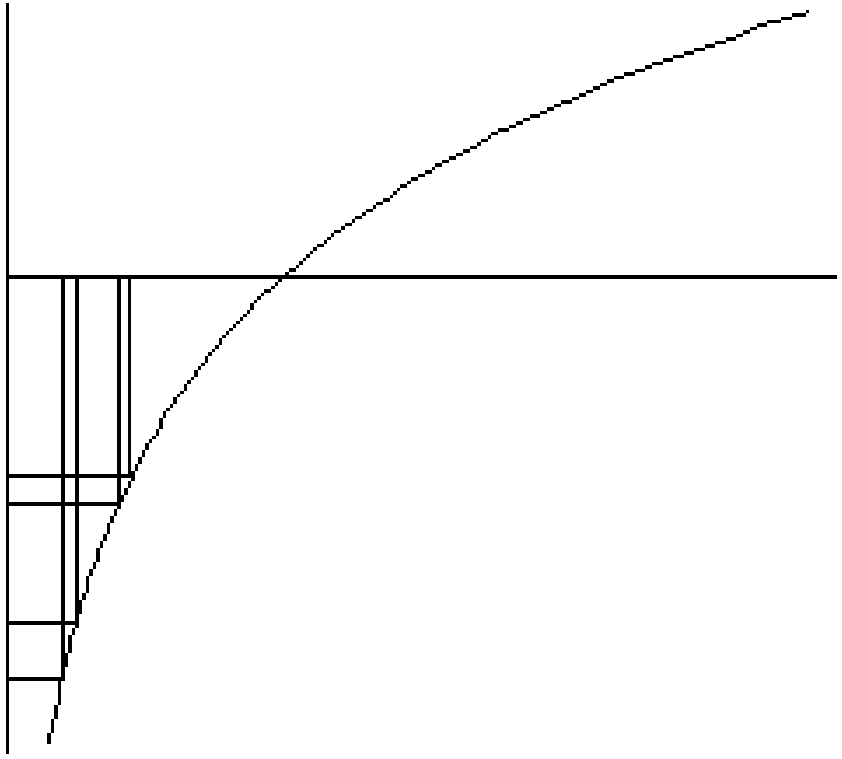


Speaker notes

The logarithm function squeezes together big data values (anything larger than 1). The bigger the data value, the more the squeezing. The graph below shows this effect.

The first two values are 2.0 and 2.2. Their logarithms, 0.69 and 0.79 are much closer. The second two values, 2.6 and 2.8, are squeezed even more. Their logarithms are 0.96 and 1.03.

Stretching

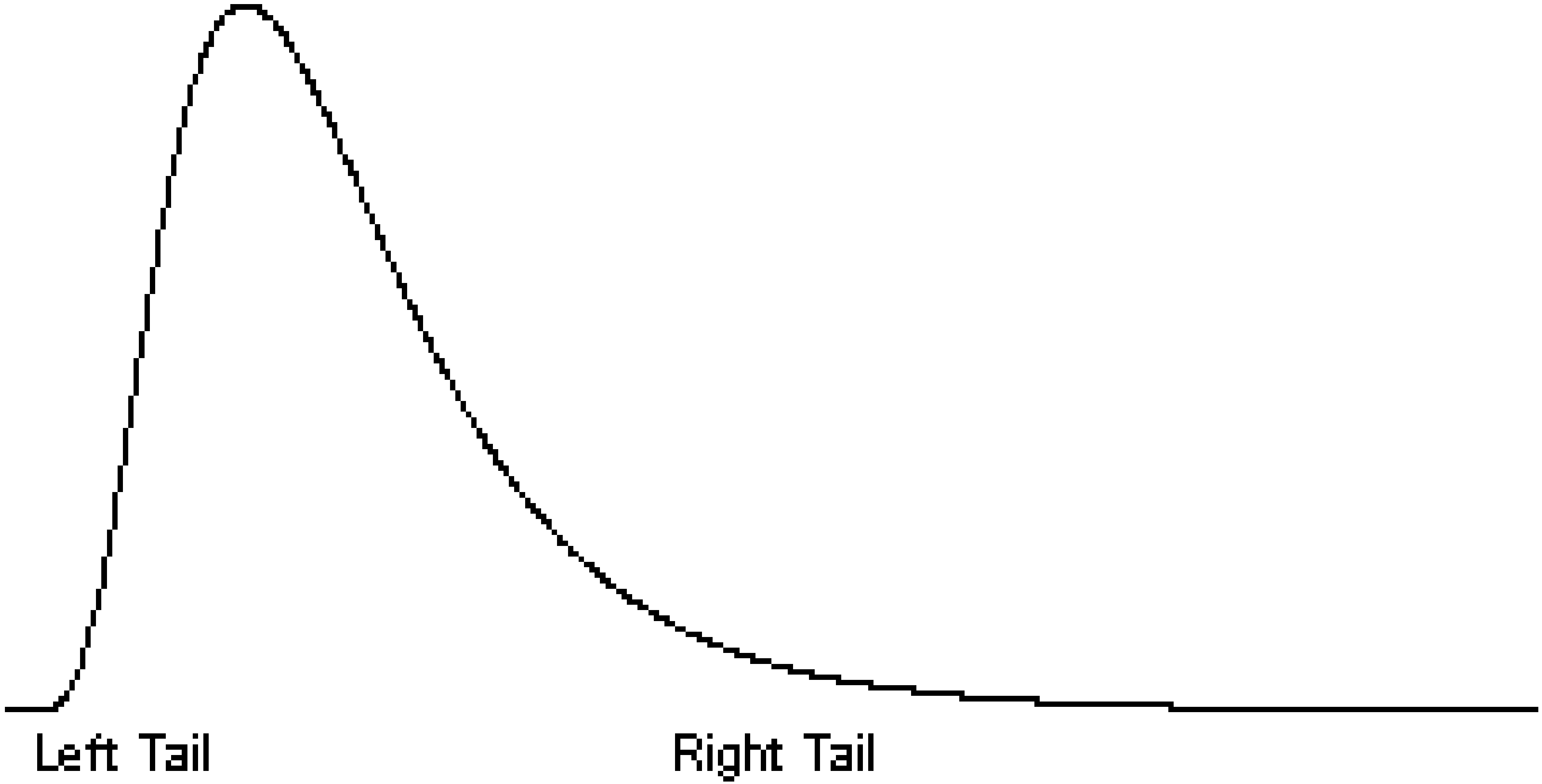


Speaker notes

The logarithm also stretches small values apart (values less than 1). The smaller the values the more the stretching. This is illustrated below.

The values of 0.4 and 0.45 have logarithms (-0.92 and -0.80) that are further apart. The values of 0.20 and 0.25 are stretched even further. Their logarithms are -1.61 and -1.39, respectively.

Skewness



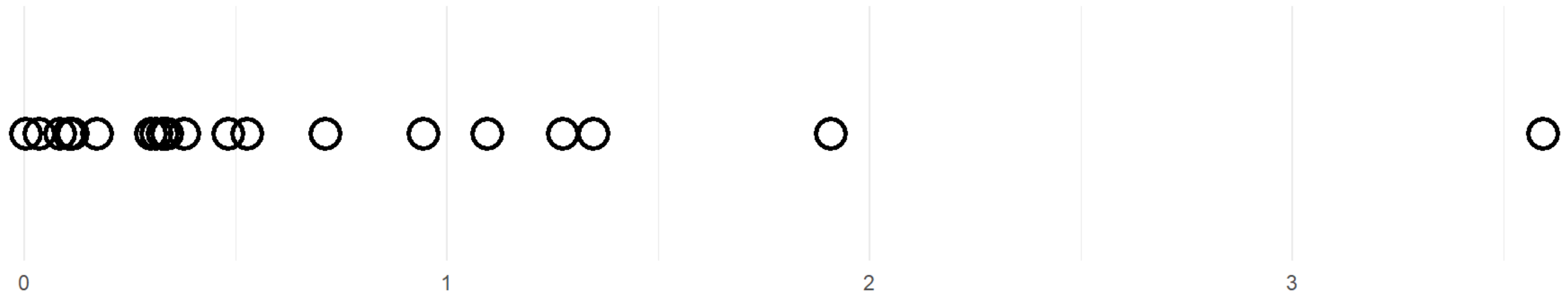
Speaker notes

If your data are skewed to the right, a log transformation can sometimes produce a data set that is closer to symmetric. Recall that in a skewed right distribution, the left tail (the smaller values) is tightly packed together and the right tail (the larger values) is widely spread apart.

The logarithm will squeeze the right tail of the distribution and stretch the left tail, which produces a greater degree of symmetry.

If the data are symmetric or skewed to the left, a log transformation could actually make things worse. Also, a log transformation is unlikely to be effective if the data has a narrow range (if the largest value is not more than three times bigger than the smallest value).

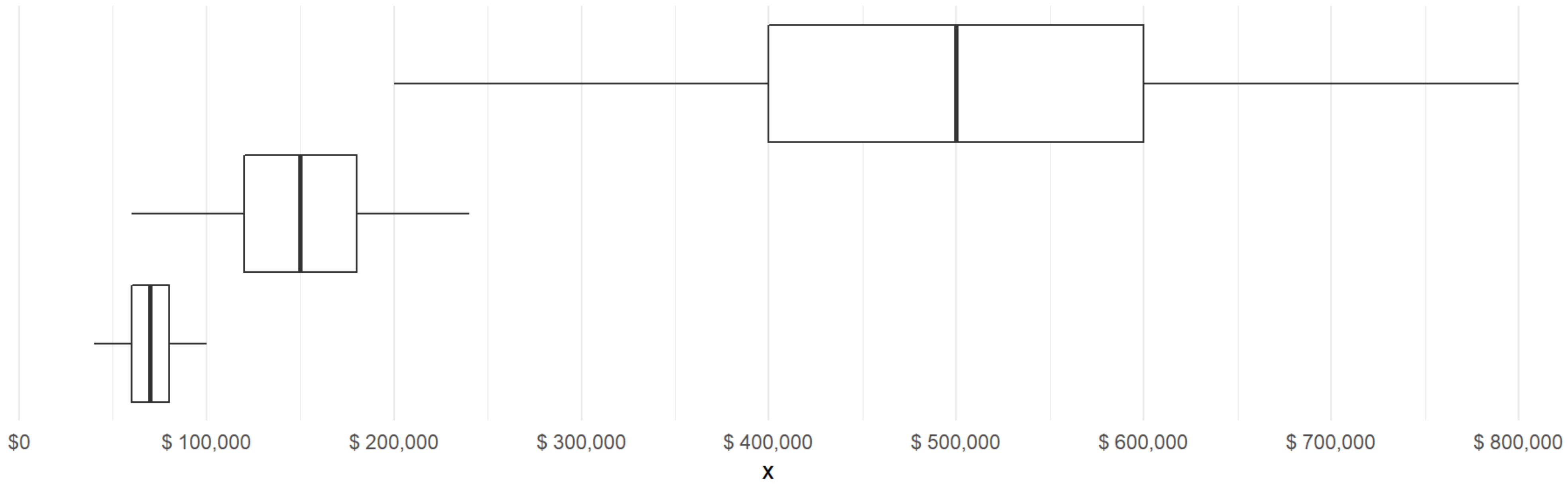
Outliers



Speaker notes

If your data has outliers on the high end, a log transformation can sometimes help. The squeezing of large values might pull that outlier back in closer to the rest of the data. If your data has outliers on the low end, the log transformation might actually make the outlier worse, since it stretches small values.

Unequal variation



Speaker notes

Many statistical procedures require that all of your subject groups have comparable variation. If you data has unequal variation, then the some of your tests and confidence intervals may be invalid. A log transformation can help with certain types of unequal variation.

A common pattern of unequal variation is when the groups with the large means also tend to have large standard deviations. Consider housing prices in several different neighborhoods. In one part of town, houses might be cheap, and sell for 60 to 80 thousand dollars. In a different neighborhood, houses might sell for 120 to 180 thousand dollars. And in the snooty part of town, houses might sell for 400 to 600 thousand dollars. Notice that as the neighborhoods got more expensive, the range of prices got wider. This is an example of data where groups with large means tend to have large standard deviations.

With this pattern of variation, the log transformation can equalize the variation. The log transformation will squeeze the groups with the larger standard deviations more than it will squeeze the groups with the smaller standard deviations. The log transformation is especially effective when the size of a group’s standard deviation is directly proportional to the size of its mean.

Multiplicative models

- Additive model
  - Catalogs +1,000 causes sales +$5,000
- Multiplicative model
  - Rain + 1 inch causes pollen * 0.5
- Multiplicative models are tricky
- Log converts it to an additive model

Speaker notes

There are two common statistical models, additive and multiplicative.

An additive model assumes that factors that change your outcome measure, change it by addition or subtraction. An example of an additive model would when we increase the number of mail order catalogs sent out by 1,000, and that adds an extra $5,000 in sales.

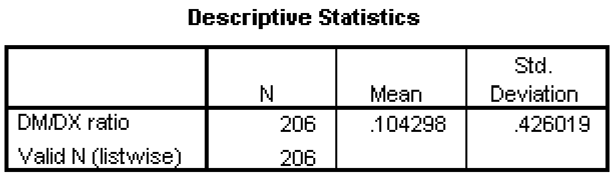
A multiplicative model assumes that factors that change your outcome measure, change it by multiplication or division. An example of a multiplicative model would be when an inch of rain takes half of the pollen out of the air.

In an additive model, the changes that we see are the same size, regardless of whether we are on the high end or the low end of the scale. Extra catalogs add the same amount to our sales regardless of whether our sales are big or small. In a multiplicative model, the changes we see are bigger at the high end of the scale than at the low end. An inch of rain takes a lot of pollen out on a high pollen day but proportionately less pollen out on a low pollen day.

If you remember your high school algebra, you’ll recall that the logarithm of a product is equal to the sum of the logarithms.

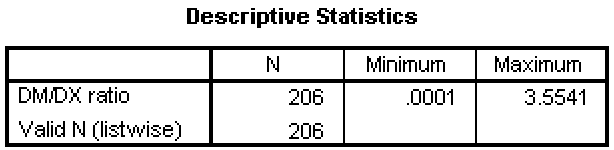
Therefore, a logarithm converts multiplication/division into addition/subtraction. Another way to think about this in a multiplicative model, large values imply large changes and small values imply small changes. The stretching and squeezing of the logarithm levels out the changes.

Example: Metabolic ratio



Descriptive Statistics

	N	Mean	Std. Deviation
DM/DX ratio	206	.104298	.426019
Valid N (listwise)	206		



Descriptive Statistics

	N	Minimum	Maximum
DM/DX ratio	206	.0001	3.5541
Valid N (listwise)	206		

Speaker notes

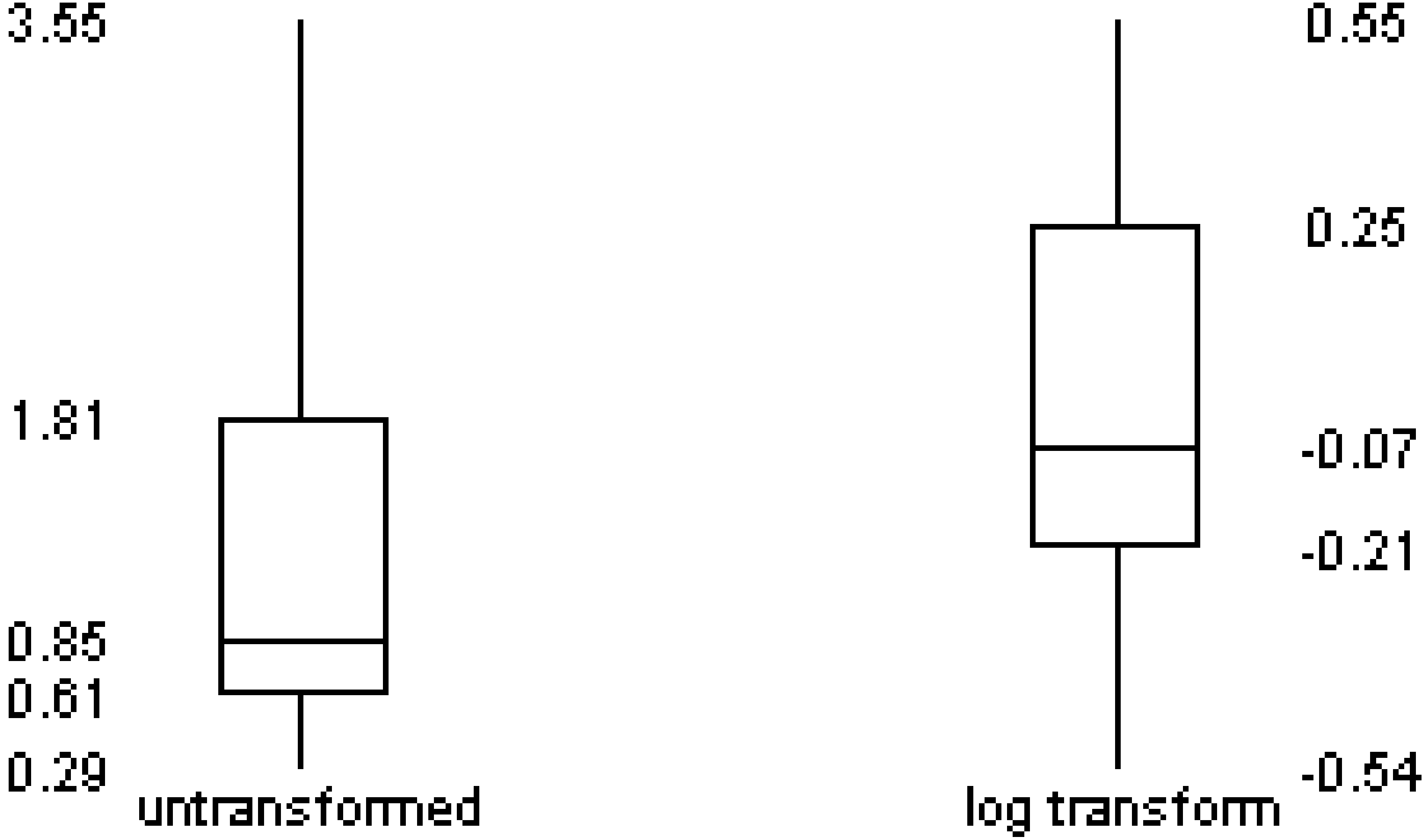
The DM/DX ratio is a measure of how rapidly the body metabolizes certain types of medication. A patient is given a dose of dextrometorphan (DM), a common cough medication. The patients urine is collected for four hours, and the concentrations of DM and DX (a metabolite of dextrometorphan) are measured. The ratio of DM concentration to DX is a measure of how well the CYD 2D6 metabolic pathway functions. A ratio less than 0.3 indicates normal metabolism; larger ratios indicate slow metabolism.

Genetics can influence CYP 2D6 metabolism. In this set of 206 patients, we have 15 with no functional alleles and 191 with one or more functional alleles.

The DM/DX ratio is a good candidate for a log transformation since it is bounded below by zero. It is also obviously a ratio. The standard deviation for this data (0.4) is much larger than the mean (0.1).

Finally, the largest value is several orders of magnitude bigger than the smallest value.

Skewness

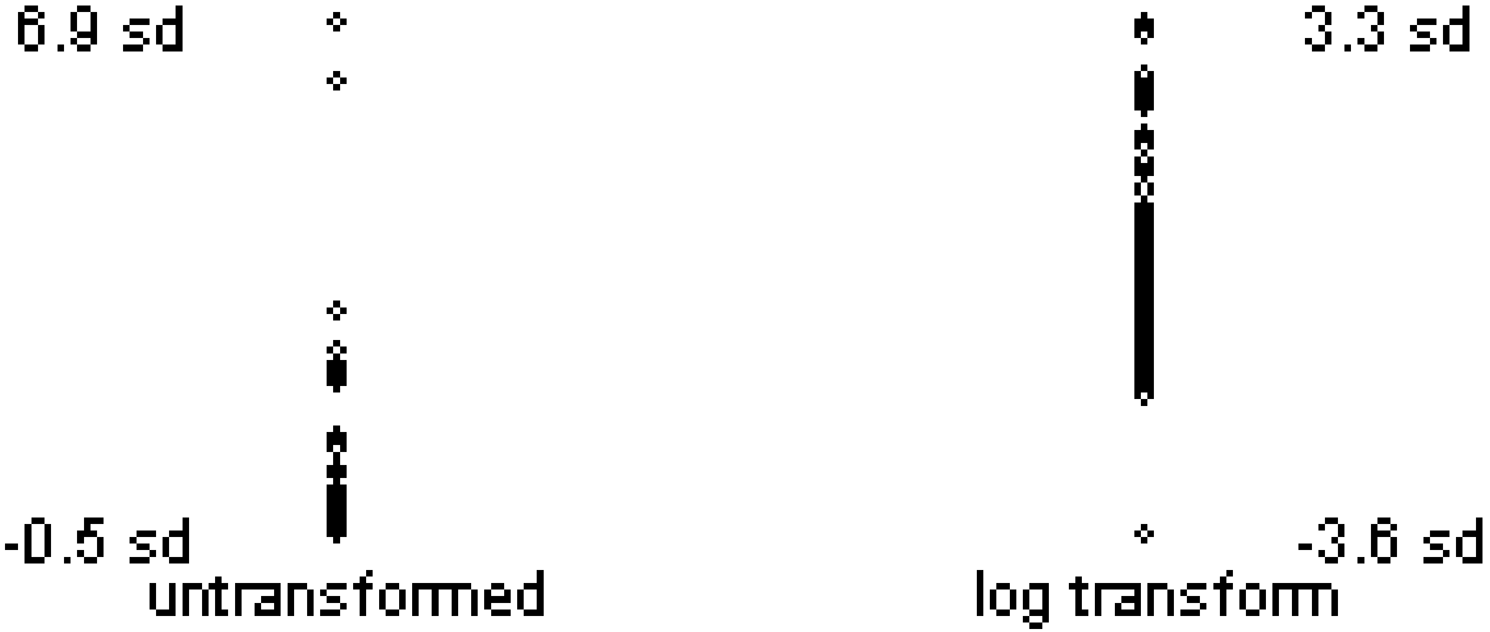

3.55
1.81
0.85
0.61
0.29
untransformed
0.55
0.25
-0.07
-0.21
-0.54
log transform

Speaker notes

The boxplots below show the original (untransformed) data for the 15 patients with no functional alleles. The graph also shows the log transformed data. Notice that the untransformed data shows quite a bit of skewness. The lower whisker and the lower half of the box are much packed tightly, while the upper whisker and the upper half of the box are spread widely.

The log transformed data, while not perfectly symmetric, does tend to have a better balance between the lower half and the upper half of the distribution.

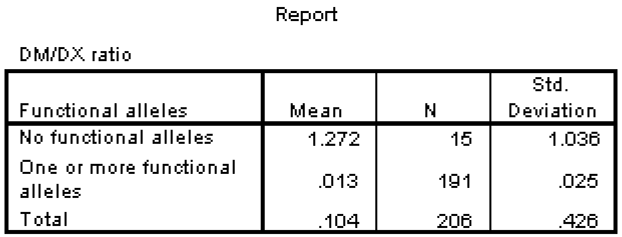
Outliers



Speaker notes

The graph below shows the untransformed and log transformed data for the subset of patients with exactly two functional alleles (n=119). The original data has two outliers which are almost 7 standard deviations above the mean. The log transformed data are not perfect, and perhaps there is now an outlier on the low end. Nevertheless, the worst outlier is still within 4 standard deviations of the mean. The influence of outliers is much less extreme with the log transformed data.

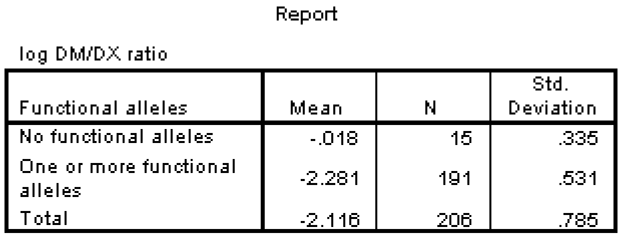
Unequal variation



Report

DM/DX ratio

Functional alleles	Mean	N	Std. Deviation
No functional alleles	1.272	15	1.036
One or more functional alleles	.013	191	.025
Total	.104	206	.426



Report

log DM/DX ratio

Functional alleles	Mean	N	Std. Deviation
No functional alleles	-.018	15	.335
One or more functional alleles	-2.281	191	.531
Total	-2.116	206	.785

Speaker notes

When we compute standard deviations for the patients with no functional alleles and the patients with one or more functional alleles, we see that the former group has a much larger standard deviation. This is not too surprising. The patients with no functional alleles are further from the lower bound and thus have much more room to vary.

After a log transformation, the standard deviations are much closer.

Further reading

- Oliver N. Keene. The log transformation is special. Keene ON. Stat Med 1995: 14(8); 811-9. Article is behind a paywall
- Wikipedia. The Log-normal distribution. Available in html format

Algebra formula for a straight line

- $Y = mx + b$
- $m = \Delta y / \Delta x$
- m = slope
- b = y-intercept

Speaker notes

One formula in algebra that most people can recall is the formula for a straight line. Actually, there are several different formulas, but the one that most people cite is

Y = m X + b

where m represents the slope, and b represents the y-intercept (we’ll call it just the intercept here). They can also sometimes remember the formula for the slope:

$m = \incrementy/\incrementx$

In English, we would say that this is the change in y divided by the change in x.

Linear regression interpretation of a straight line

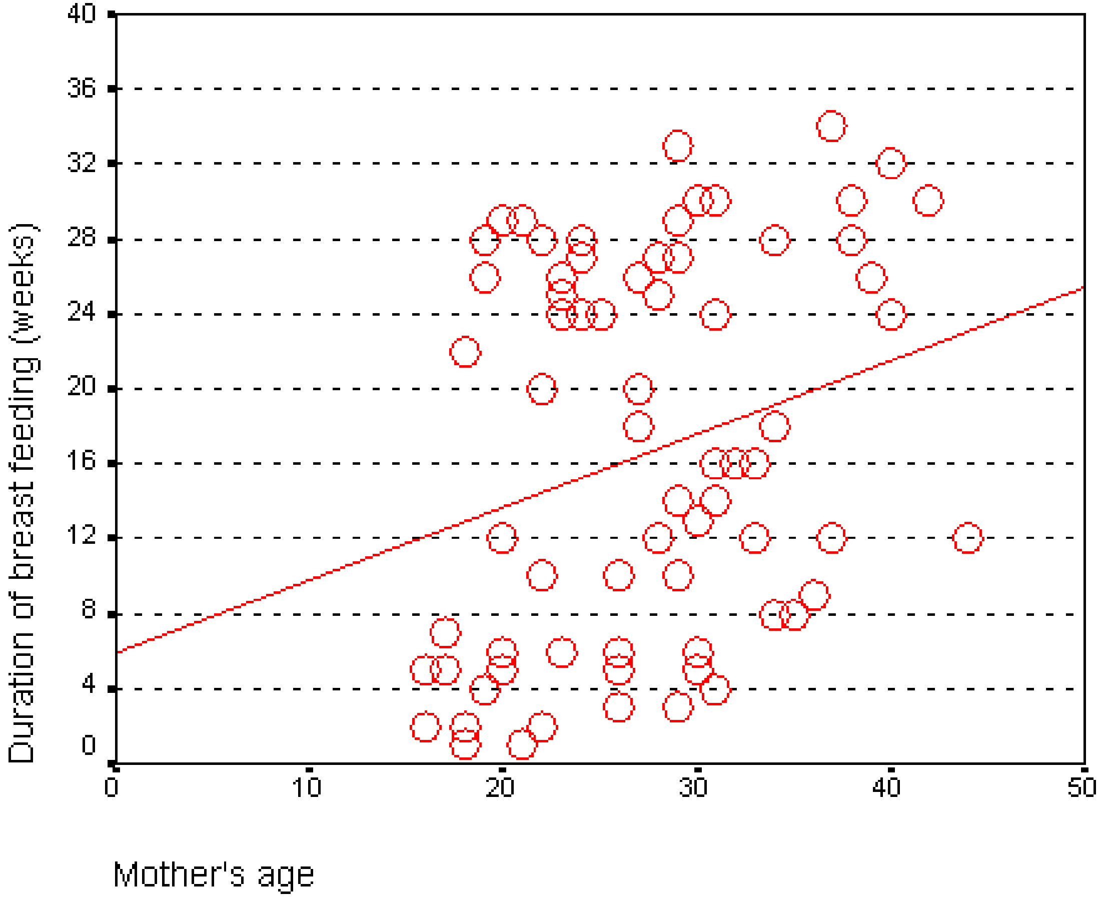
- The slope represents the estimated average change in Y when X increases by one unit.
- The intercept represents the estimated average value of Y when X equals zero.

Speaker notes

In linear regression, we use a straight linear to estimate a trend in data. We can’t always draw a straight line that passes through every data point, but we can find a line that “comes close” to most of the data. This line is an estimate, and we interpret the slope and the intercept of this line as follows:

Be cautious with your interpretation of the intercept. Sometimes the value X=0 is impossible, implausible, or represents a dangerous extrapolation outside the range of the data.

First regression example with interpretation

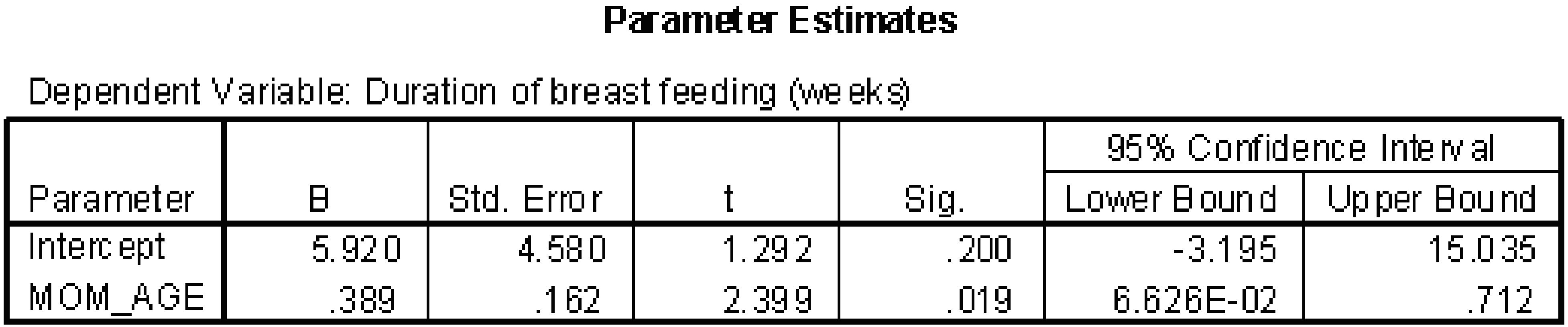


Speaker notes

The graph shown below represents the relationship between mother’s age and the duration of breast feeding in a research study on breast feeding in pre-term infants.

The regression coefficients are shown below. The intercept, 6, is represented the estimated average duration of breast feeding for a mother that is zero years old. This is an impossible value, so the interpretation is not useful. What is useful, is the interpretation of the slope, approximately 0.4. The estimated average duration of breast feeding increases by 0.4 weeks for every extra year in the mother’s age.

Output from SPSS



Parameter Estimates

Dependent Variable: Duration of breast feeding (weeks)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	5.920	4.580	1.292	.200	-3.195	15.035
MOM_AGE	.389	.162	2.399	.019	6.626E-02	.712

Speaker notes

Here is what the output from SPSS looks like.

Interpretation when X is categorical

- Code X as 0-1
- Intercept: Estimated average value of Y for the “0” category.
- Slope: Estimated average change in Y when category changes from “0” to “1”.

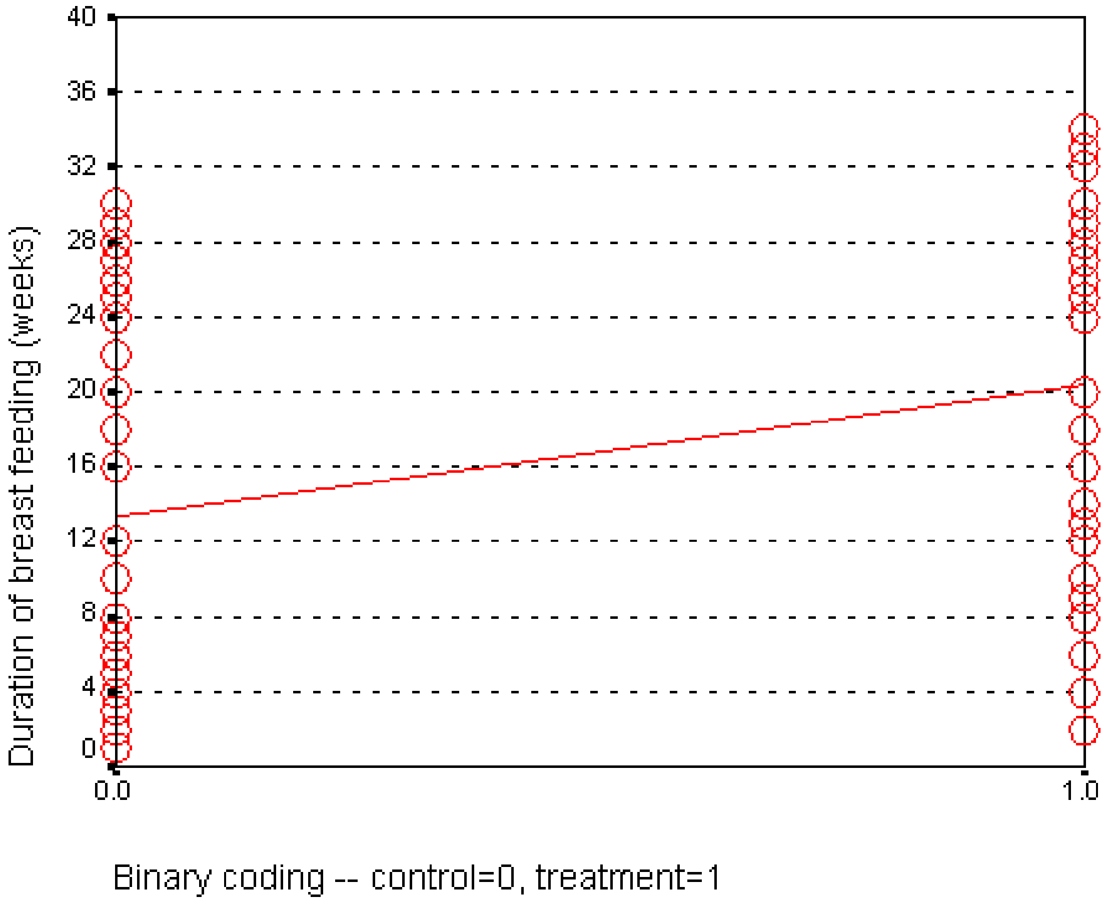
Speaker notes

When X is categorical, the interpretation changes somewhat. Let’s look at the simplest situation, a binary variable. A binary variable can have only two possible categories. Some examples are live/dead, treatment/control, diseased/healthy, male/female. We need to assign number codes to the categories. Most people assign the codes 1 and 2, but it is actually better to assign the codes 0 and 1.

When we represent a binary variable using 0-1 coding, the slope represents the estimated average change in Y when you switch from one group to the other. The intercept represents the estimated average value of Y for the group coded as zero.

The interpretation of the regression coefficient for a categorical variable with more than two values is a bit trickier.

Second regression example with interpretation

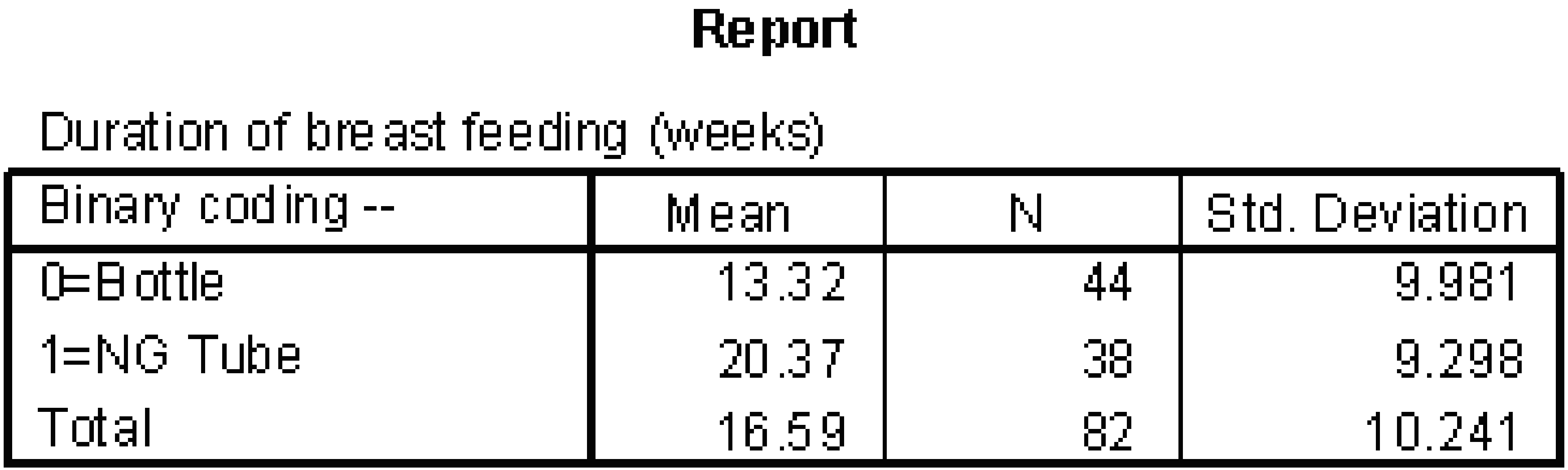


Speaker notes

In a study of breast feeding, we have a treatment group and a control group. Let us label the treatment group as 1 and the control group as 0. The outcome variable is the age when breast feeding stopped.

The control group had a mean duration of breast feeding just a bit larger than 13. The mean for the treatment group is just a bit larger than 20. Notice that the regression line shown above connects the two means.

SPSS output



Report

Duration of breast feeding (weeks)

Binary coding --	Mean	N	Std. Deviation
0=Bottle	13.32	44	9.981
1=NG Tube	20.37	38	9.298
Total	16.59	82	10.241

Speaker notes

In this situation, the intercept, 13, represents the average duration for the control group. The slope is 7, which is the change in the average duration when we move from the control group to the treatment group.

Alternate coding

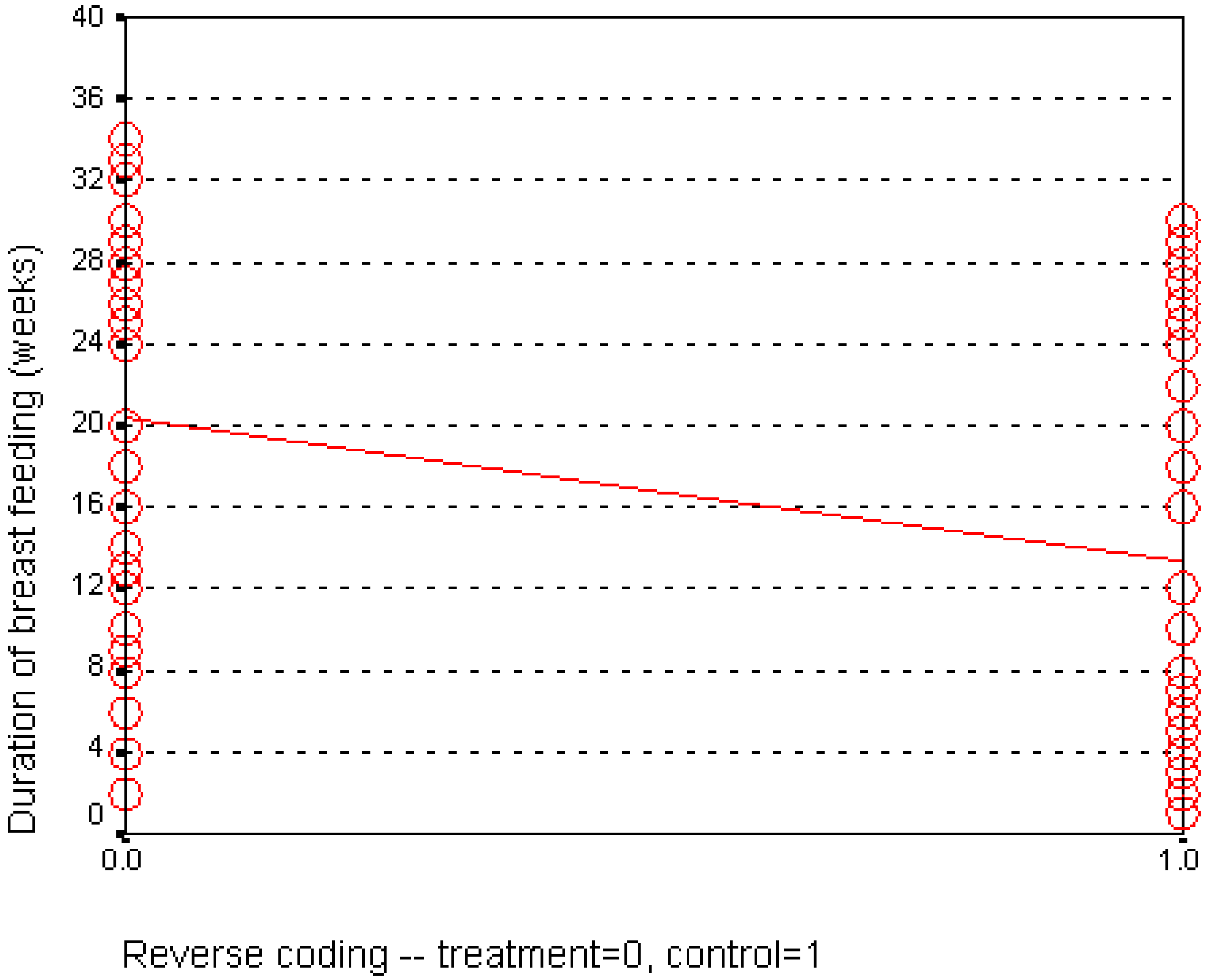


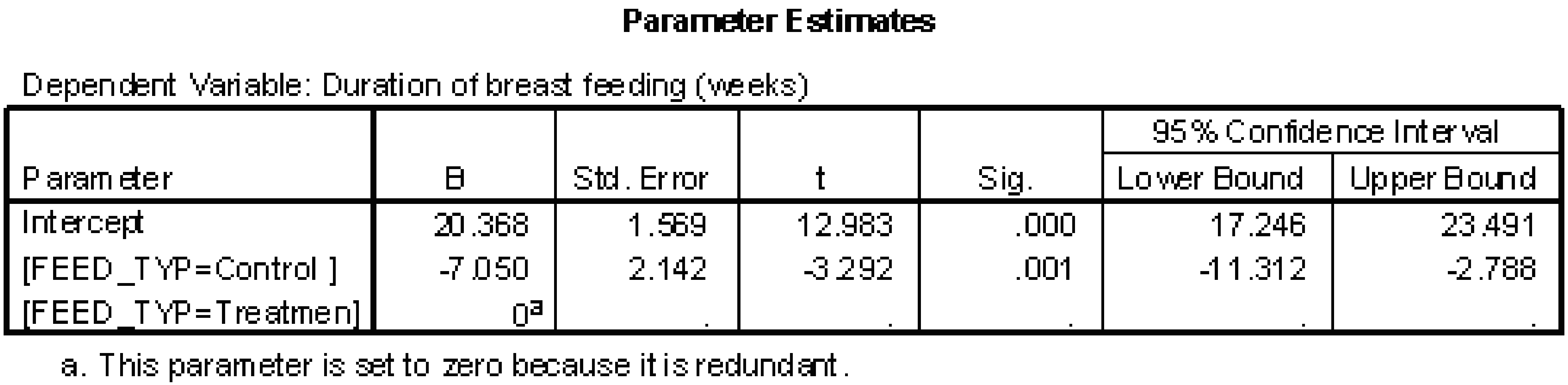
Figure 5. Scatterplot with alternate ordering of treatment

Speaker notes

We could have just as easily labeled the treatment group as 0 and the control group as 1. If we did that, we would get a graph that looks like the following:

Here, the intercept, 20, represents the mean of the treatment group. The slope, -7, represents the change in average duration as we move from the treatment group to the control group. It is actually this reverse coding that SPSS chooses as a default.

SPSS output



Parameter Estimates

Dependent Variable: Duration of breast feeding (weeks)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	20.368	1.569	12.983	.000	17.246	23.491
[FEED_TYP=Control]	-7.050	2.142	-3.292	.001	-11.312	-2.788
[FEED_TYP=Treatmen]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Speaker notes

Neither coding is correct or incorrect. Just make sure that you understand the difference. If you get a slope that is in the opposite direction of what you expected, perhaps it is because your software is using a different coding than what you expected.