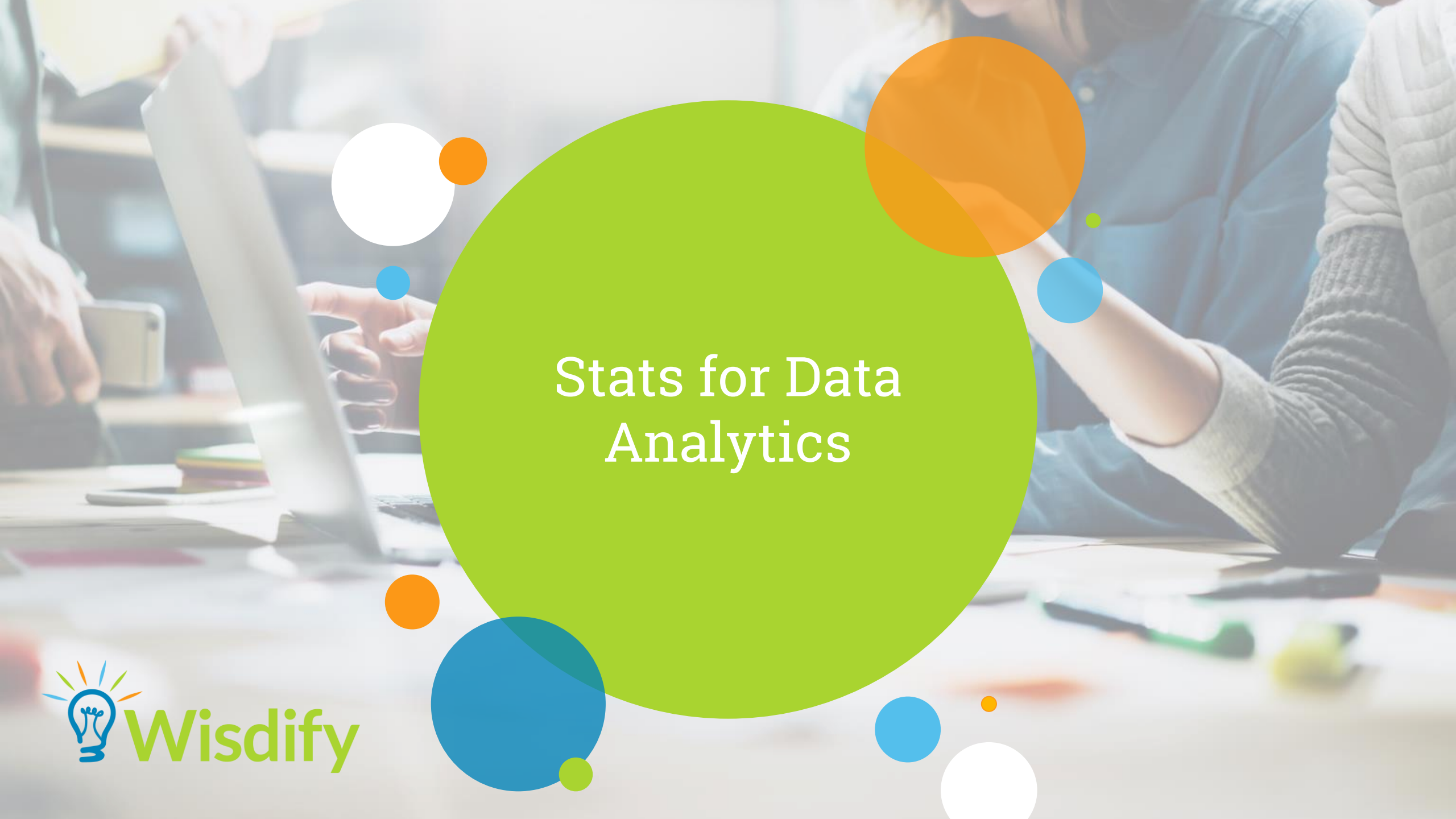


Stats for Data Analytics

Wisdify

The fundamentals of statistics

What is statistics?

Definition: The science that deals with the collection, analysis, and interpretation of data.

Types of Statistics

Descriptive

*Describe what is going on in a data set. Results **cannot** be generalized to any other group.*

Baseball player has a .305 batting average

I have a 3.45 GPA

Inferential

Allow us to infer/predict trends about a larger population based on a study of a sample taken from it.

Pollster samples voters to predict election

Use past data to forecast inventory

Population vs. Sample

Population: Represents *ALL* possible data points or measurements.

Most accurate but it's rare to actually get a true population.

Example

Polling every single voter to see who they will be voting for next election.

Sample: A portion of the population representing the characteristics of the population.

Less accurate, but it's much easier to get a sample.

Example

Randomly polling 3,000 on their voting preference to predict the next election.

Types of data

Categorical: Represents groups or categories

Example

Computer brand

Answer to a yes/no question

All categorical data can be represented as a number (ex. men and women can be represented with 0s and 1s).

Numerical: Represents numbers

Example

Price of a product

Temperature

Ways to visually show our data

Bar chart

Line graph

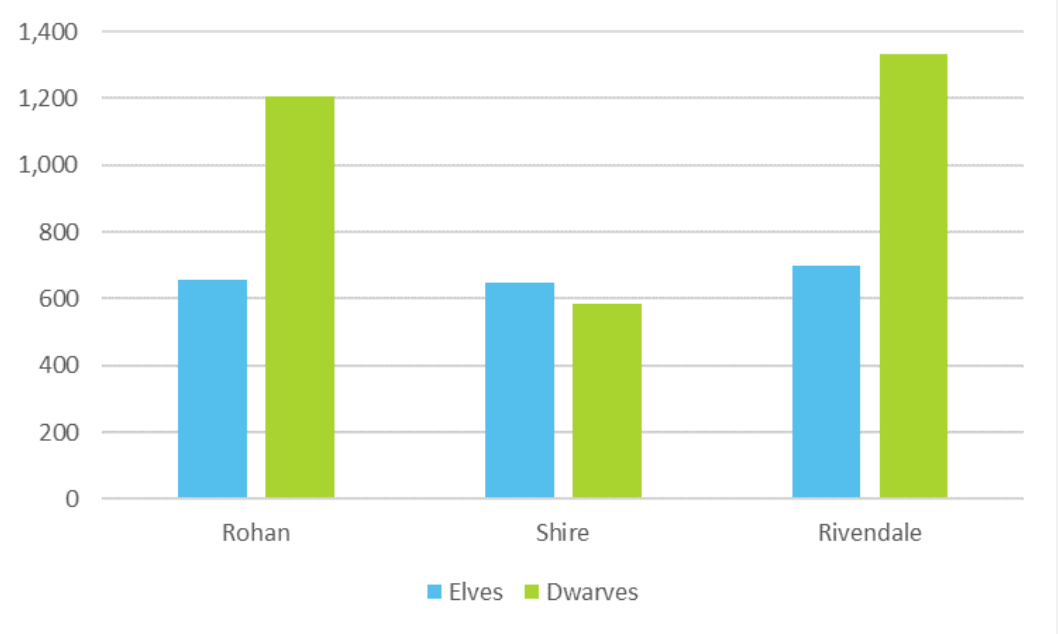
Pie chart

Histogram

Scatter Plot

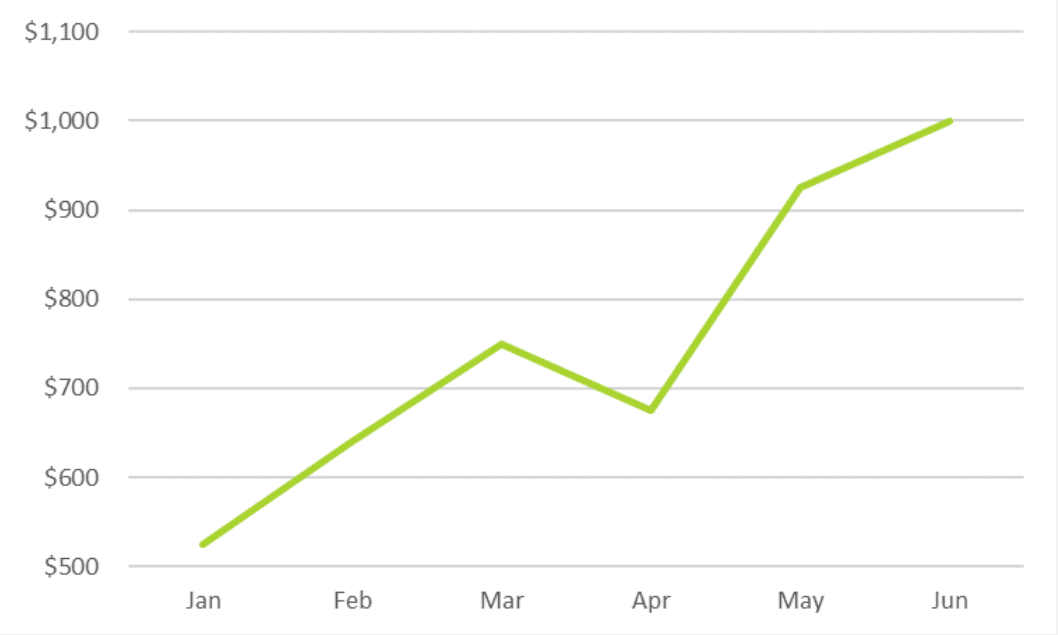
Ways to visually show our data

Bar/Column charts: Used to compare item(s) between different groups. Although not ideal, can also compare items over time.



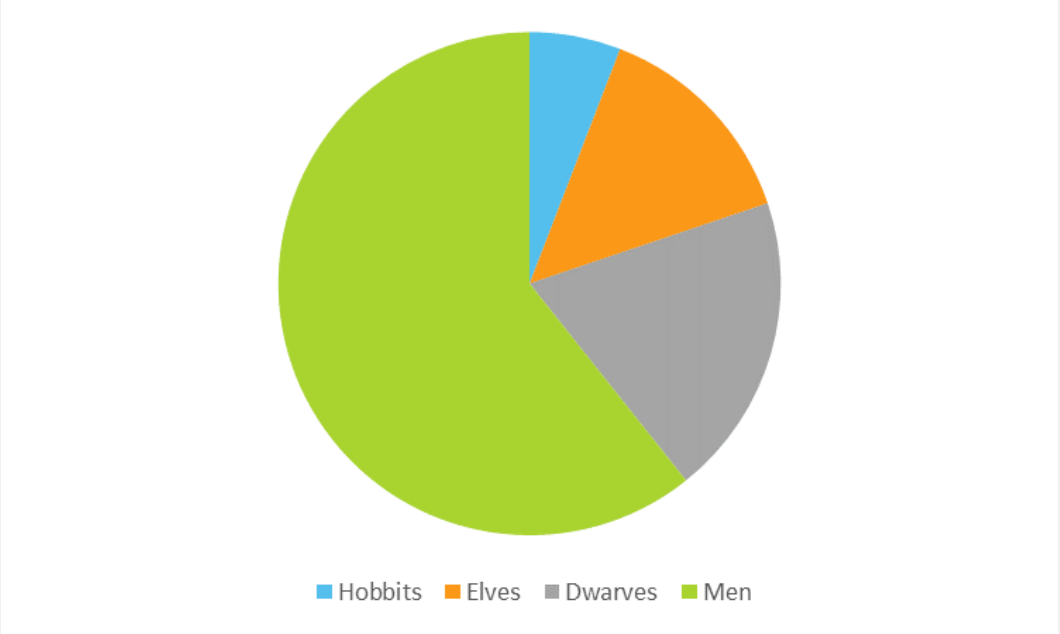
Ways to visually show our data

Line chart: Almost identical to the column chart except the layout is a line instead of a bar which makes spotting trends much easier to read. Line chart is best for a time-series.



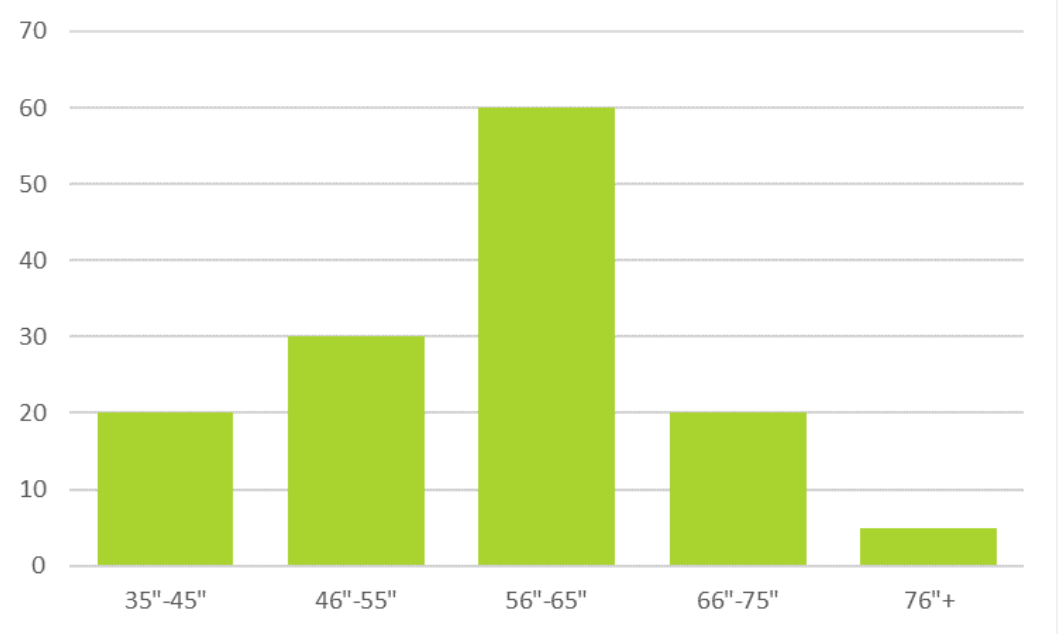
Ways to visually show our data

Pie chart: Used to compare parts of a whole. Do not show changes over time.



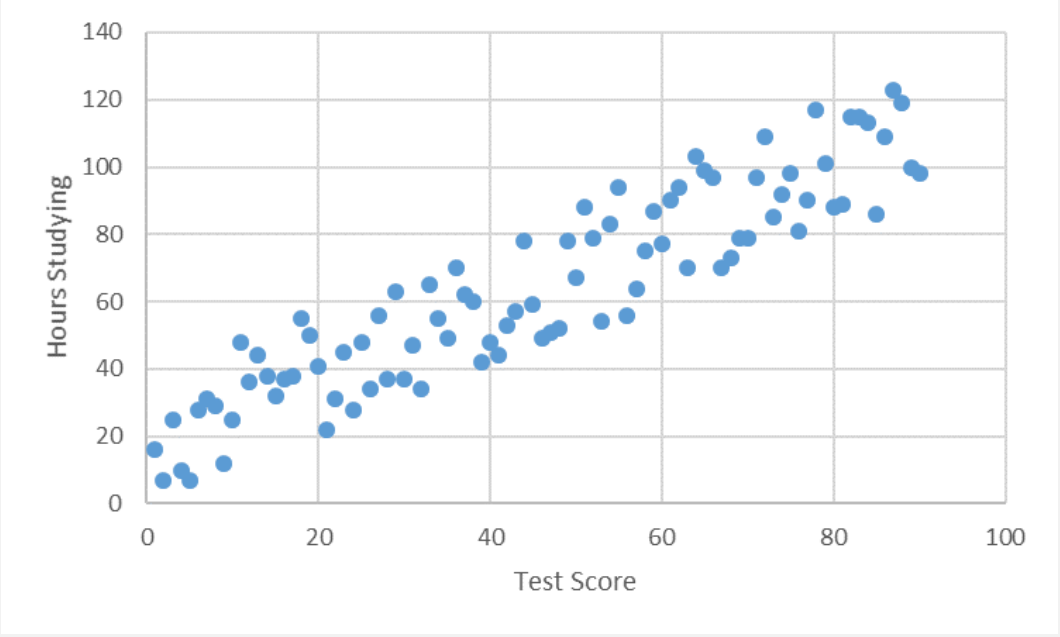
Ways to visually show our data

Histogram: Bar chart that shows the frequency in which groups of data occur. The chart represents the distribution of the data.



Ways to visually show our data

Scatter plot: Similar to line/bar chart, scatter plots have horizontal and vertical axes to plot data points. However, its purpose is to show how much one variable is affected by another.



The toolbox vs. Silver bullet

In statistics, there is not “silver bullet” measure. In fact, we should never rely on one single measure to explain our data. Instead, we use many different tools to dissect out data and reveal insights.

Just like you cannot use only a hammer to build a house, we need many different tools to help us explain and analyze our data.

VS.

The ultimate goal of statistics

The ultimate goal of statistics is to be able to **predict outcomes,** without using astrology. Simply describing our data is a minor goal.

Descriptive statistics

Mean

Mean (simple average): The average of all numbers. To calculate, add all of the numbers in a set and then divide the sum by the total count of numbers. Outliers can have a major effect on the mean.

Example

Employee age
24
25
25
26
27
29
33
35
41
43
63

$$\text{Mean} = \frac{24+25+25+26+27+29+33+35+41+43+63}{11} = \mathbf{33.7}$$

How to calculate it in Excel

=AVERAGE(number1, [number2], …)

You can select an entire range instead of selecting each individual number.

Mean

Outliers can have a major effect on the mean…

The oldest employee is 63	The oldest employee is 43
Employee age	Employee age
24	24
25	24
25	25
26	25
27	26
29	27
33	29
35	33
41	35
43	41
63	43
Mean 33.7	**Mean 30.2**

Average age drops by 3.5 years

Median

Median: After arranging the numbers from smallest to largest value, the median is the middle value. If there's an even number of values, the median is the average of the two middle values. Median is not affected by outliers.

Example

Employee age
24
25
25
26
27
29
33
35
41
43
63

Median = **27**

How to calculate it in Excel

=MEDIAN(number1, [number2], …)

You can select an entire range instead of selecting each individual number.

Mode

Mode: The most frequently occurring number found in a set of numbers.

Example

Employee age
24
25
25
26
27
29
33
35
41
43
63

Mode = **25** (occurs 2x while everything else occurs 1x)

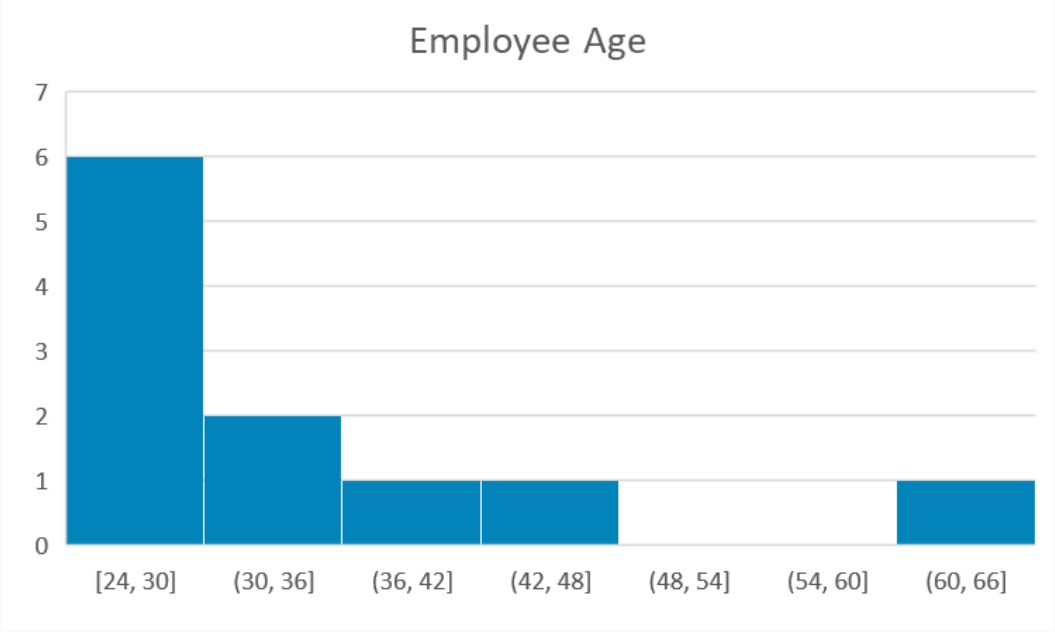
How to calculate it in Excel

=MODE(number1, [number2], …)

You can select an entire range instead of selecting each individual number.

Histogram

Histogram: Helps us visually show how our data is distributed and gives better meaning to mean, median, and mode.



Mean = 33.7

Median = 29

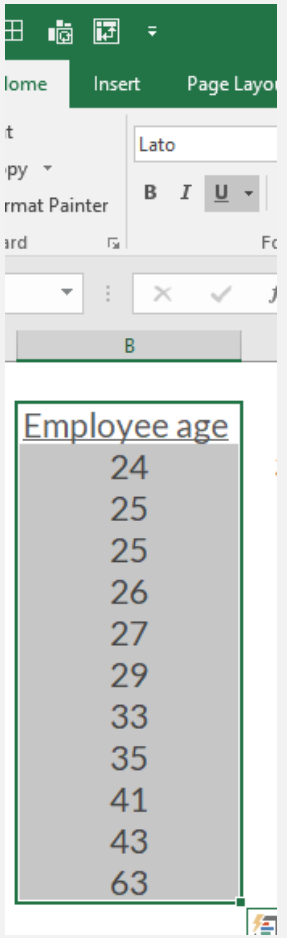
Mode = 25

Age is heavily skewed to the left

Histogram

How to create a histogram in Excel

1. Select all the data you want to include in your histogram. With newer versions of Excel, you don't need to create a frequency table. Score!

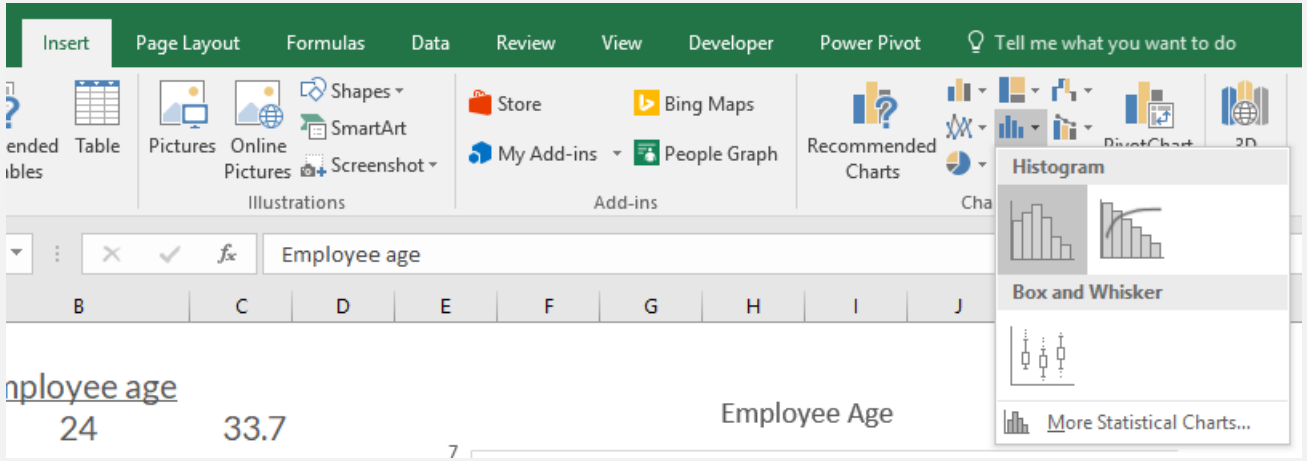


Histogram

How to create a histogram in Excel

2

On the Ribbon, go to Insert and under the Charts area, select Histogram. You may need to view all chart options to see Histogram.

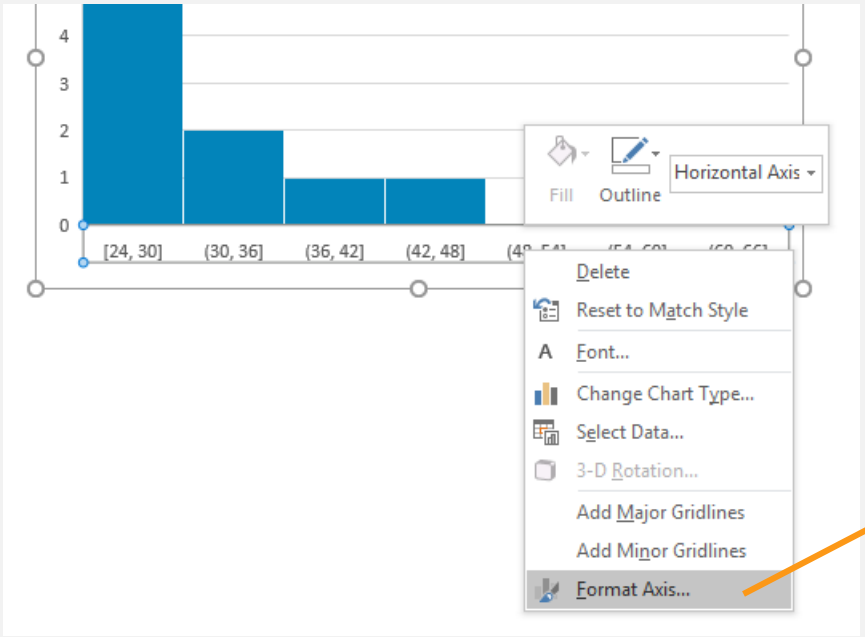


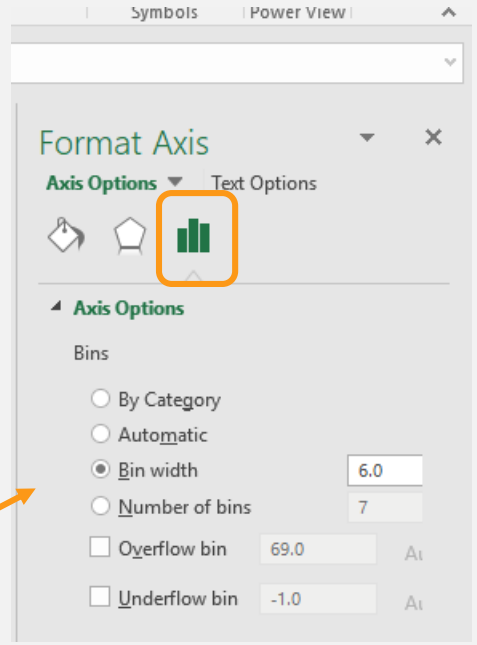
Histogram

How to create a histogram in Excel

3

To change the interval size, right click on the horizontal axis and select Format Axis. Under Axis Option, you can specify the interval (“Bin width”) or # of groups (“Number of bins”)

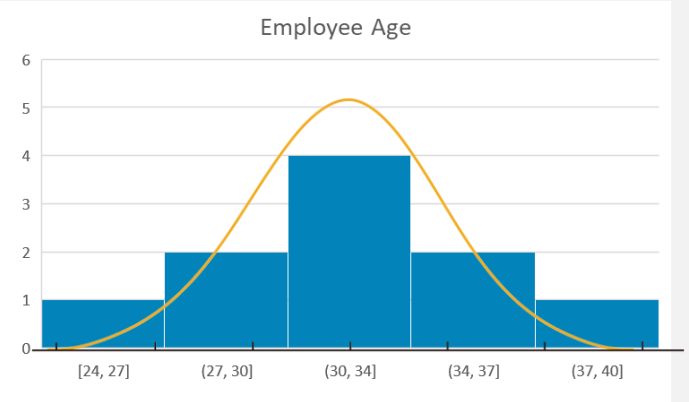




Skewness

Skewness: Measures of the lopsidedness of a histogram

Normal skew: Mean = Median



The histogram is symmetric, or in other words, it has about the same shape on each side. Notice that the shape looks like the famous bell-curve, which we overlaid on this example.

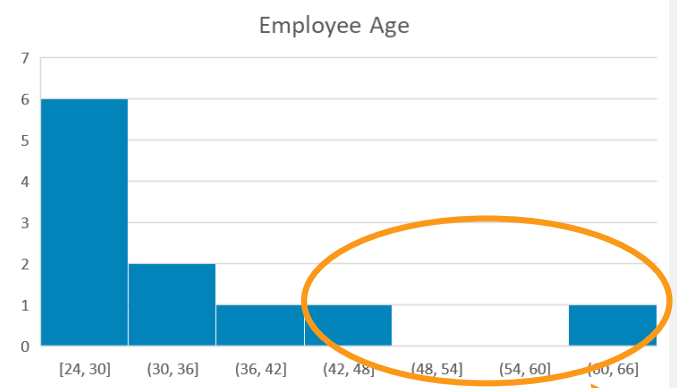
Mean = 33

Median = 33

Skewness

Skewness: Measures of the lopsidedness of a histogram

Right skew: Mean > Median



The name is a bit misleading. Right skew indicates most of the data is to the LEFT with a few large values to the RIGHT.

Example: Pay at a large company. Most employees are paid near the low end with a few executives being paid significantly more.

Mean = 34

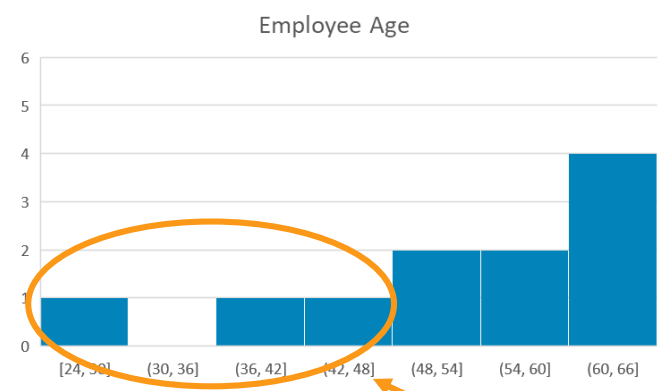
Median = 29

In a histogram or distribution, the ends of the chart are called "tails". In a right skew, the tail is on the RIGHT.

Skewness

Skewness: Measures of the lopsidedness of a histogram

Left skew: Mean < Median



Left skew indicates most of the data is to the RIGHT with a few large values to the LEFT.

Example: Time it takes to complete an exam. Some people finish quickly while most will take the entire time allotted.

Mean = 53

Median = 58

In a left skewed histogram, the tail is on the LEFT of the chart.

Skewness

Calculating skewness is super easy, but the number is fairly useless so we won't worry about it.

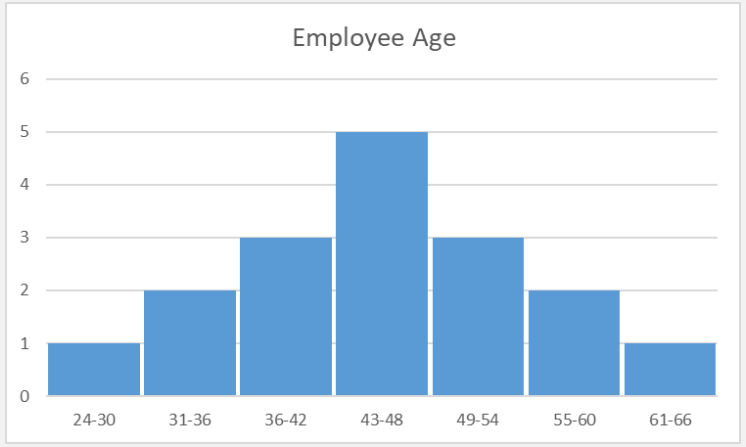
The skewness number doesn't help us with our analysis and doesn't really give any good information.

Are you still concerned about calculating skewness? Use the SKEW() function in Excel to calculate it.

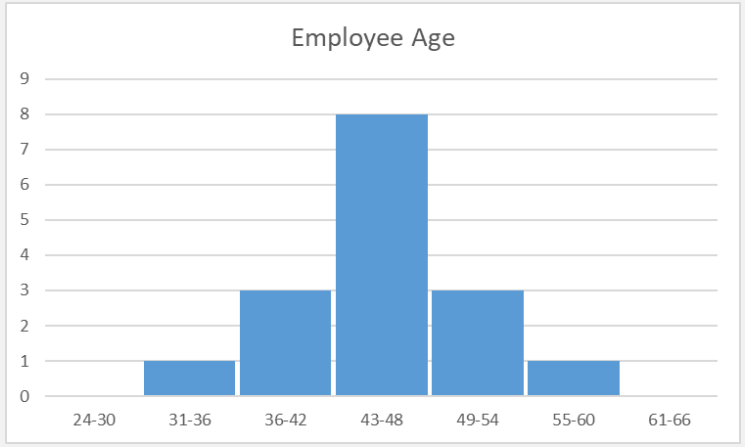
What is variance?

A histogram can *show* us how our data is distributed, but how can we *know* how much it's distributed?

Example: Both charts have almost identical means and medians, but clearly the second chart's data is grouped much closer to the mean. But how much so? That is where variance comes in…



Mean = 45.1
Median = 45.0



Mean = 45.6
Median = 45.0

What is variance?

Variance: How spread out the data is from the mean

Example: Two sets of test scores with the same mean (average)

Test Scores		Test Scores	
0%		45%	
50%		50%	
100%		55%	
Mean	**50%**	**Mean**	**50%**

In the first group, the test scores are extremely spread out from the mean. You have everything from 0% to 100% = **HIGH VARIANCE**

In the second group, all the test scores are close to the mean = **LOW VARIANCE**

Measuring variance: Standard deviation

Standard deviation: How we measure variance. Standard deviation is a number that tells us how much variance our data has (how spread out our numbers are from the mean).

Example: Two sets of test scores with the same mean (average)

	Test Scores
	0%
	50%
	100%
Mean	**50%**
Standard Deviation	**50%**

	Test Scores
	45%
	50%
	55%
Mean	**50%**
Standard Deviation	**5%**

Standard deviation tells us the first set of test scores are distributed much more widely from the mean than the second set

Standard deviation

Is a higher standard deviation good or bad?

High standard deviation	Low standard deviation
=	=
Harder to predict outcomes	**Easier to predict outcomes**

Example: I want to see if my salary is inline with my peers. Which would be better/easier for me to benchmark my salary?

1) Average salary = $50k / Standard deviation = $7k.
2) Average salary = $50k / Standard deviation = $18k.

Standard deviation

Calculating standard deviation

The actual formula for standard deviation can get ugly and knowing it won't add a ton of value...

$$SD = \sqrt{\frac{\sum |x - \mu|^2}{N}}$$

Luckily Excel has a built in function to calculate this number for us...

Standard deviation Excel formulas

=STDEV.S()

Use this if you are working with a sample, which will almost always be the case.

=STDEV.P()

You should only use this if you are working with a true population, which is rare.

Coefficient of variation

Coefficient of variation: Don't let the name scare you. It's simply a way to "standardize" our standard deviation so we can compare the standard deviation across different data sets.

Example: We want to compare the standard deviation of sales and cost of goods sold (COGS). Since sales are likely quite a bit higher than COGS, comparing the pure standard deviation isn't helpful. The coefficient of variation let's us to the comparison.

Formula

$$Coefficient\ of\ variation = \frac{Standard\ deviation}{Mean}\ x\ 100$$

Reminder: Types of statistics

Types of Statistics

Descriptive

*Describe what is going on in a data set. Results **cannot** be generalized to any other group.*

Baseball player has a .305 batting average

I have a 3.45 GPA

Inferential

Allow us to infer/predict trends about a larger population based on a study of a sample taken from it.

Pollster samples voters to predict election

Use past data to forecast inventory

Inferential statistics

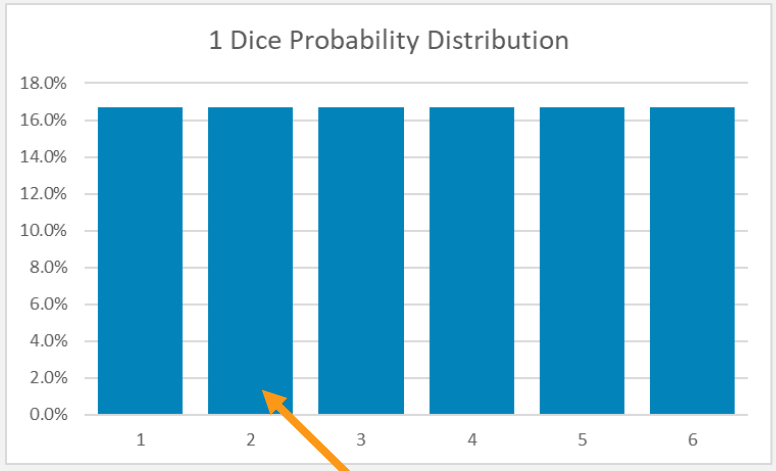
Probability Distribution

Probability distribution: Chart showing all the possible values and the probability they occur. It's similar to a histogram, except instead of showing the number of times a value occurs, it shows the probability it will occur.

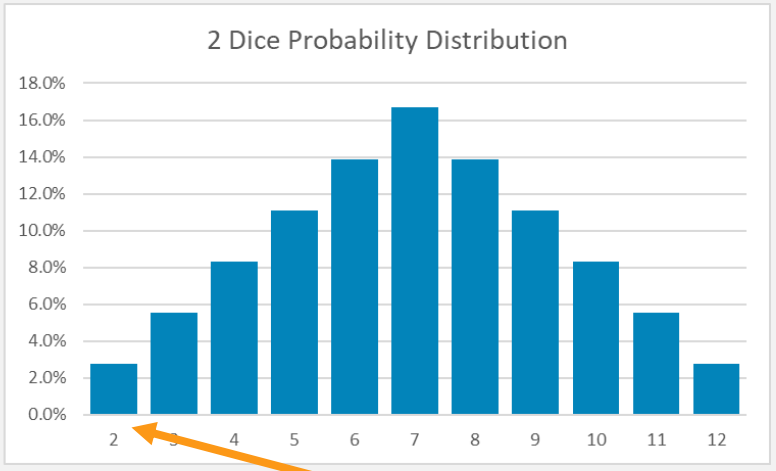
Don't worry, you'll never, ever need to calculate the actual probability of all possible values occurring. That would be sheer madness!

Probability Distribution

Example: Below is the probability distributions for rolling a 1 through 6 on a single dice and rolling a 2 through 12 on two dice.



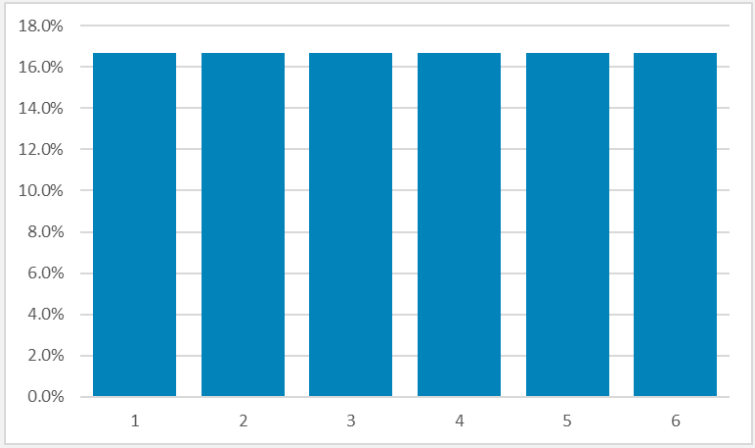
With one dice, the probability of rolling a 2 is 1/6 = 16.7%.



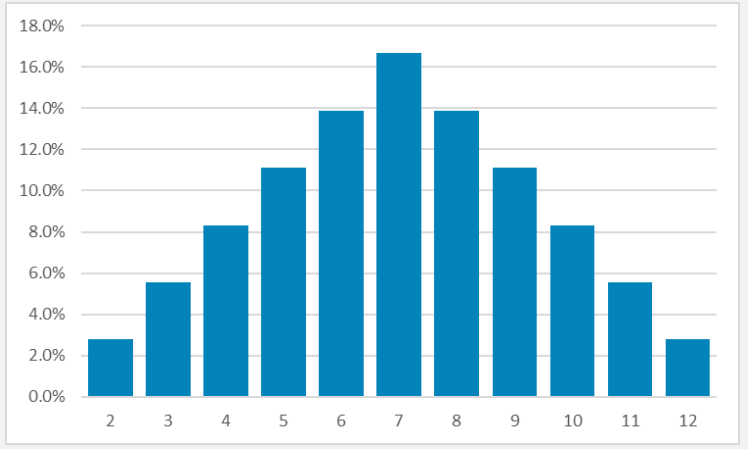
With two dice, the probability of rolling a 2 is 1/36 = 2.7%. The only way to get 2 is to roll two 1s (snake eyes) and there are 36 possible combinations (6x6).

Common types of distributions

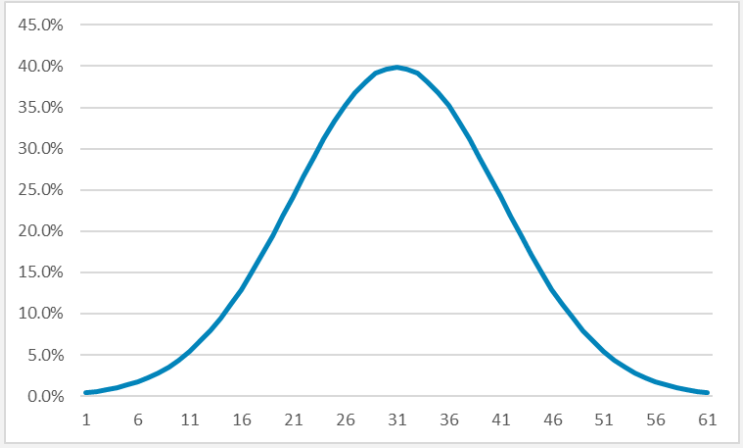
Uniform distribution



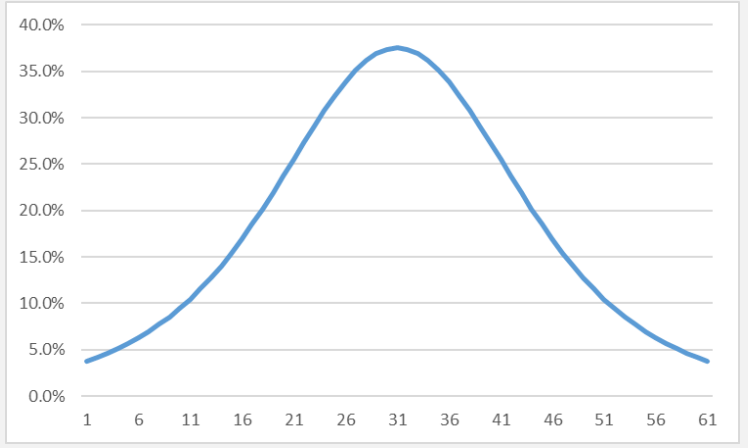
Binomial distribution



Standard distribution



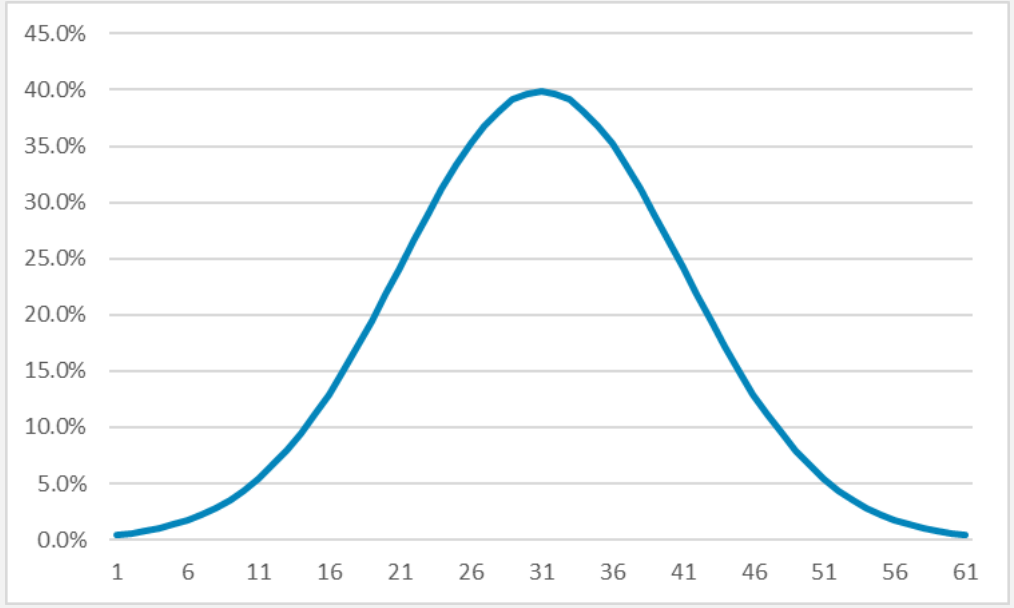
Student's T distribution



Normal distribution

Normal distribution: Most commonly referred to as the “bell curve”. The normal distribution has the following qualities:

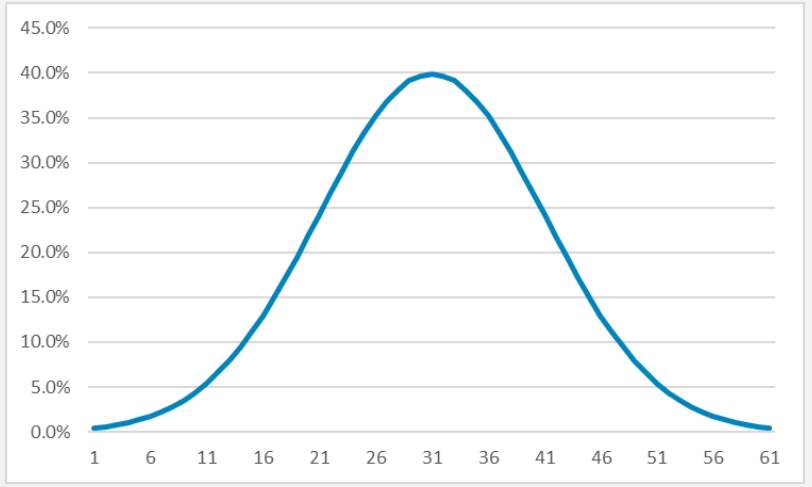
1. The mean, median, and mode are all the same
2. The graph is symmetric on both sides and is bell-shaped



Normal distribution

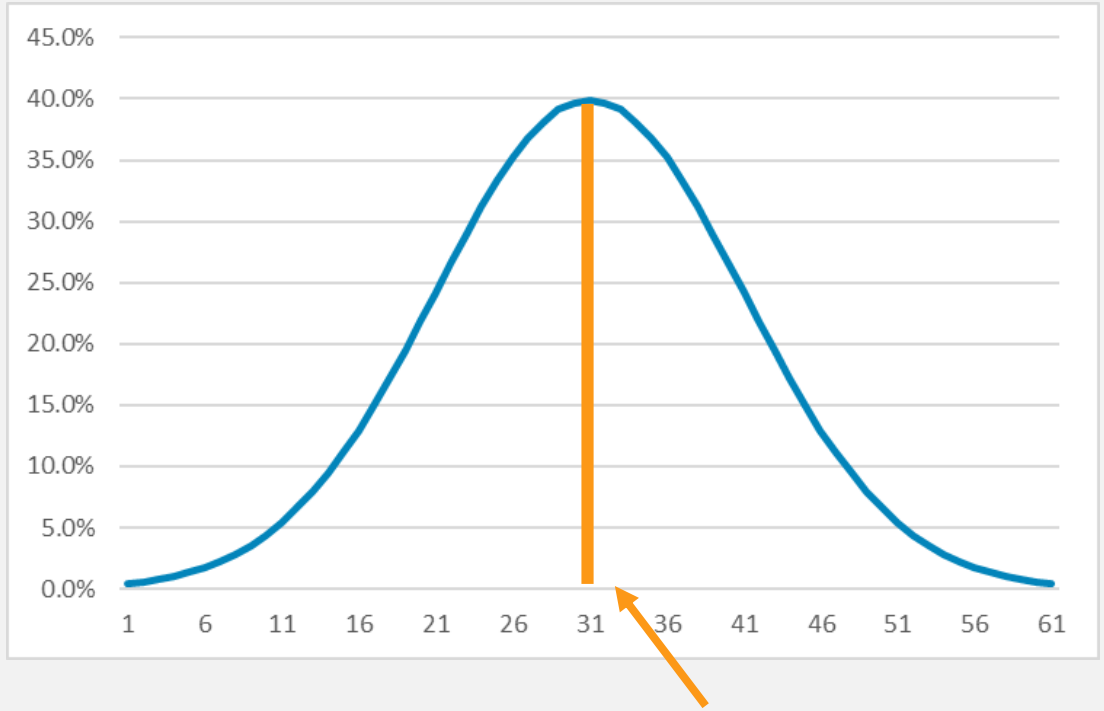
The normal distribution is the most common distribution because:

1. The statistics behind it are “easy” to compute because of its symmetry and rules
2. Basis for a regression analysis (which we’ll explain at the end)
3. Many things naturally follow the bell-curve (height, temperature, etc)
4. Central limit theorem (explained in a bit)



Normal distribution

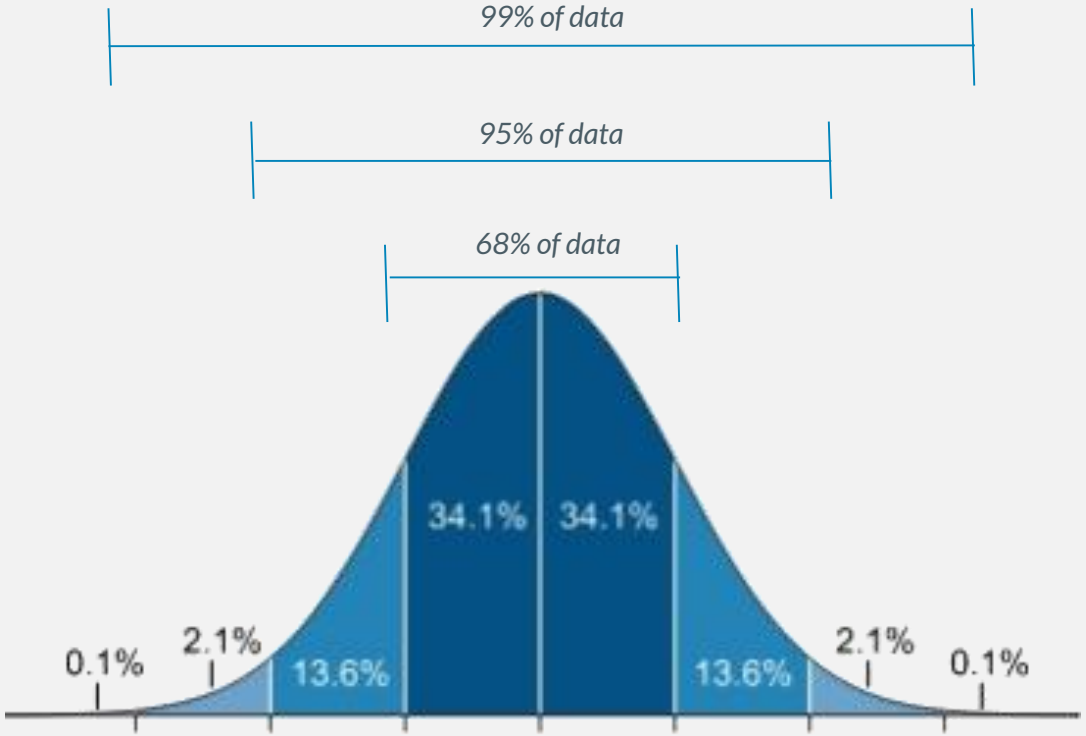
Tying normal distribution with our other stats…mean, median, & mode



The men, median, and mode are approximately the same number and are right smack in the middle of the curve.

Normal distribution

Tying normal distribution with our other stats…standard deviation



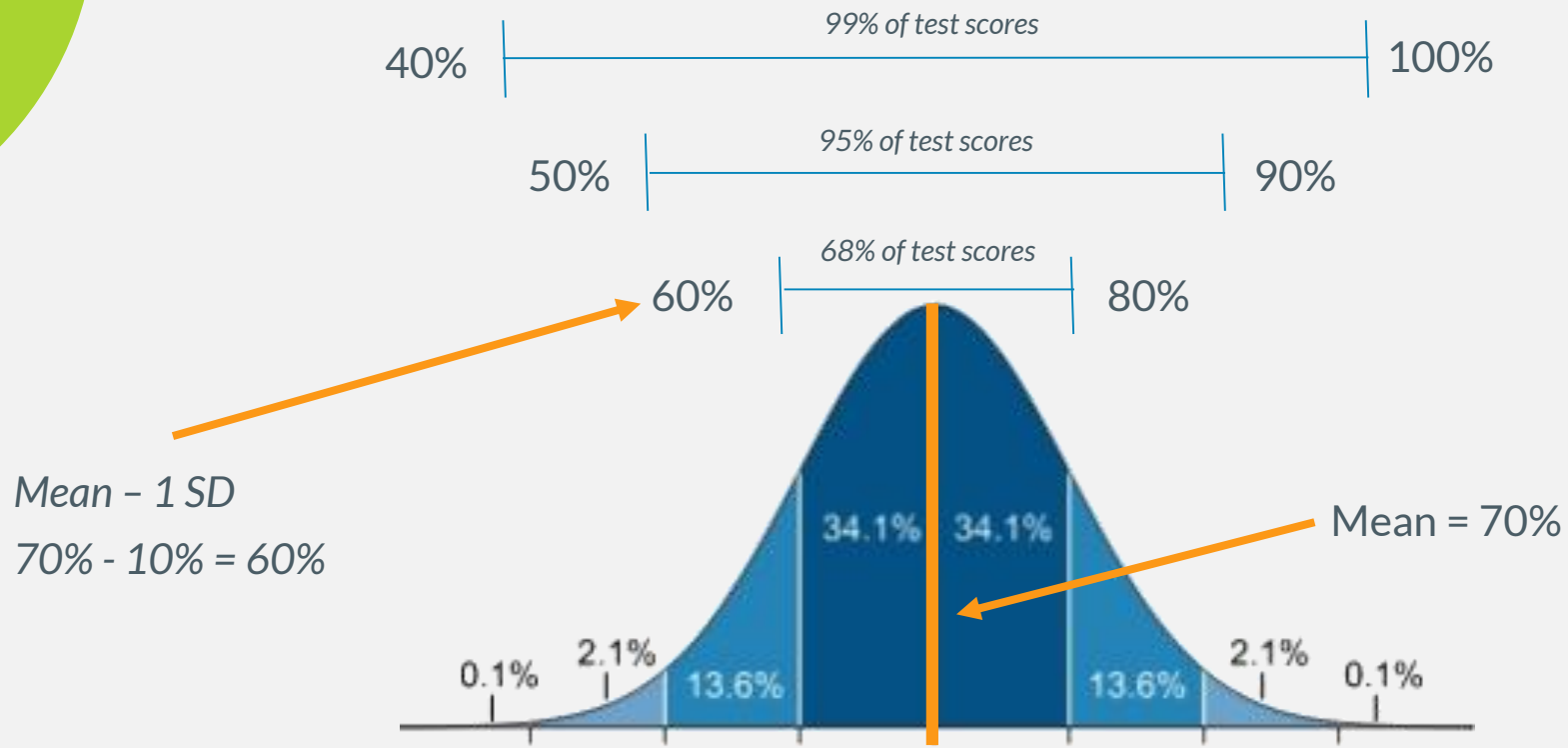
68% of all data falls within ± 1 standard deviation

95% of all data falls within ± 2 standard deviation

99% of all data falls within ± 3 standard deviation

Normal distribution

Example: The average test score was a 70% with a standard deviation of 10%. The test scores would fall in this range.



Central limit theorem

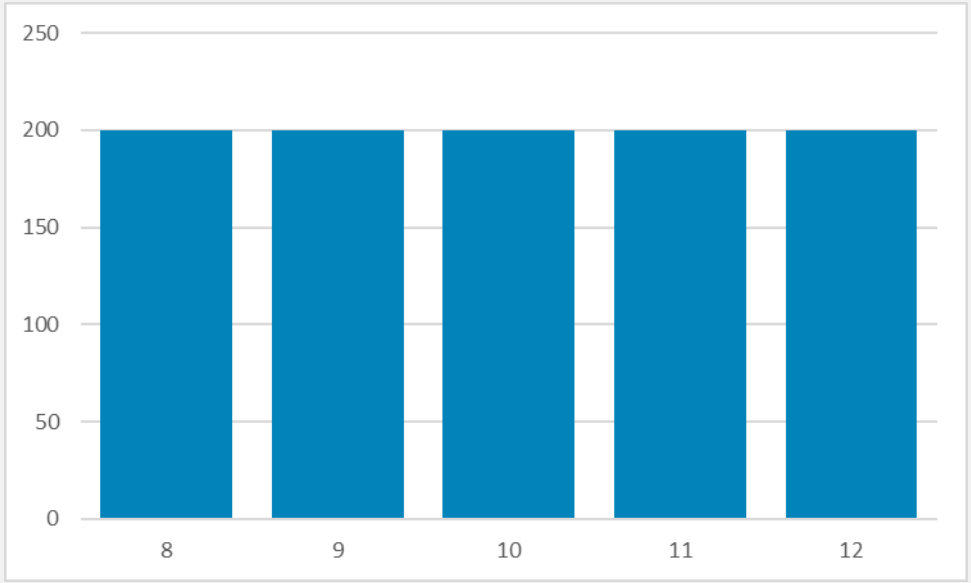
Central limit theorem (CLT): This is a foundational principle of statistics. The CLT explains why we only need a relatively small sample to be able to make assumptions about the entire dataset AND why we can assume a normal distribution with our data (even when it's not!).

Instead of giving you a boring definition, let's explain it to you…

Central limit theorem

Central limit theorem (CLT) explained

Example: Let's say a high school has 1,000 students in grades 8-12 with 200 students in each grade.



The histogram clearly shows this is not a normal distribution.

Central limit theorem

Central limit theorem (CLT) explained

Now let's say, of those 1,000 students, we take a random sample of 20 students and calculate their average grade level…

8
11
9
10
10
12
8
…

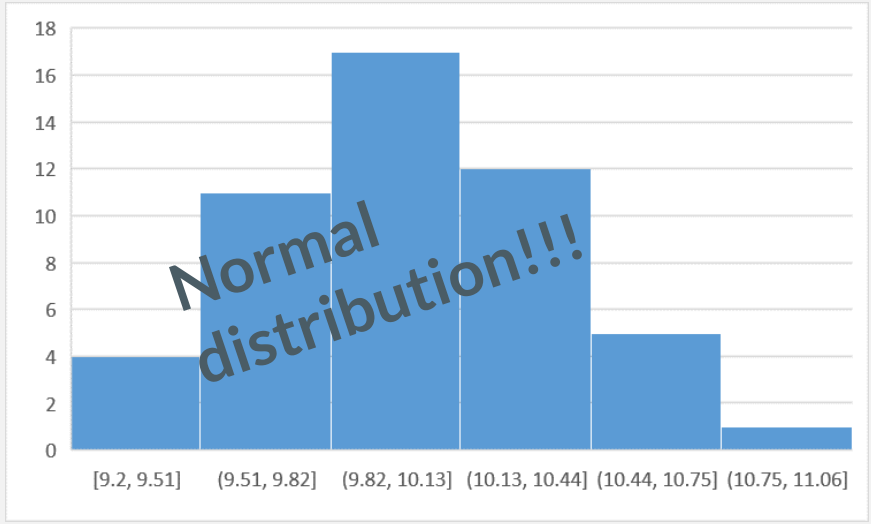
Mean = 9.6

And then we do that 49 more times and collect all the averages of these random 20-person samples...

Central limit theorem

Central limit theorem (CLT) explained

Well, if we take the 50 averages from our samples and throw them in a histogram, guess what their distribution looks like?



So, no matter what your original distribution looks like, the distribution of the sample averages will be normally distributed! Sweet!

Central limit theorem

Wow, that's awesome! So what does that mean?

1. This new distribution we just created will have approximately the same mean as the total population.

2. This new distribution will also have approximately the same standard deviation as the total population. A minor adjustment will need to be made to the sample standard deviation.

3. Given 1 and 2, we can make **estimates** purely with a large sample size and be very confident in our results.

Making estimates

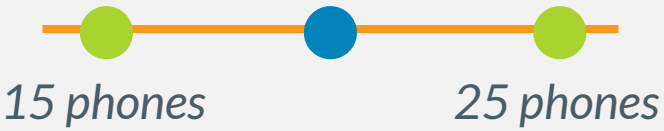
Intro to estimates

Point estimate: An estimate that is a single number

Example: "I estimate we will sell 20 phones next month"

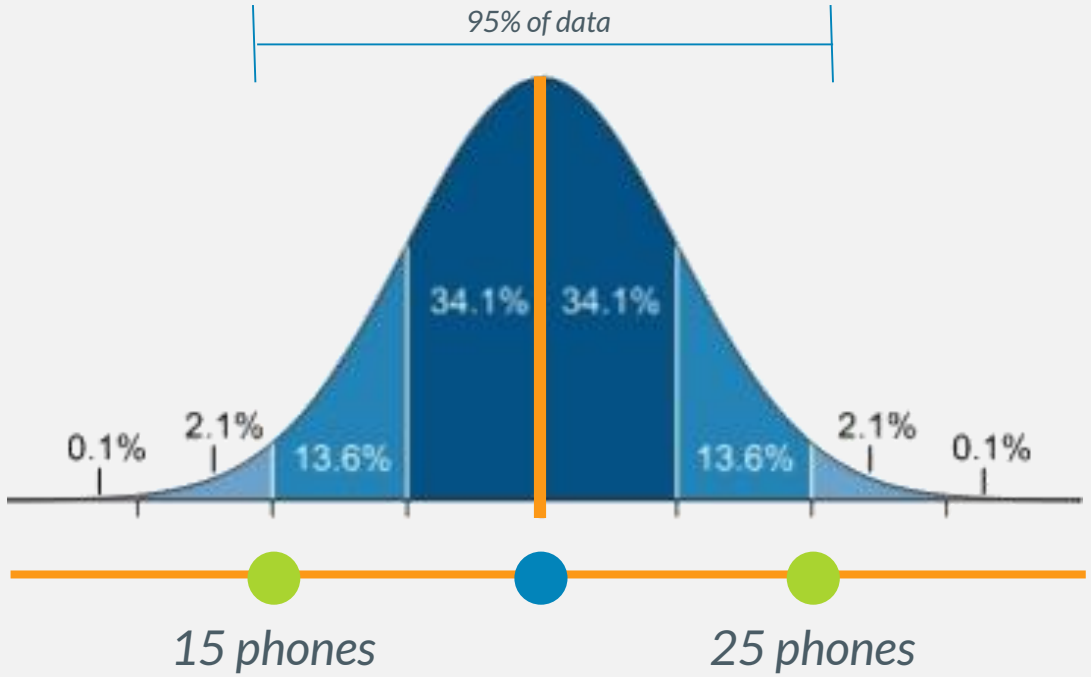
Confidence interval estimate: An estimate that is in a range such that you are 90%, 95%, or 99% confident the actual will be within that range. **You want to make these types of estimates!**

Example: "With 95% confidence, I estimate we will sell 15-25 phones next month"



Intro to estimates

This all ties back to what we've learned about the normal distribution and standard deviation…



If we want to have 95% confidence that the actual number will fall within a range, we need to make sure it captures data ± 2 standard deviations from the mean.

Calculating confidence interval estimate

Formula

*Mean ± reliability factor * standard error*

Basically how many standard deviations from the mean you need to go.

However, because we're using a sample, you need to use a t-table to get this number.

The # of "standard deviations" for a sample, or reliability factor, is called a t-score.

Think of this as the standard deviation of the sample. To calculate, use the following formula:

$$\textbf{Standard error} = \frac{Standard\ deviation\ of\ the\ sample}{\sqrt{The\ number\ of\ samples\ - 1}}$$

Calculating confidence interval estimate

This table shows what reliability factor, or t-score, you should use in your formula. Just find your sample size and the confidence level you want, and pull the number.

Sample size	90%	95%	98%	99%	99.80%	99.90%
5	2.015	2.571	3.365	4.032	5.893	6.869
6	1.943	2.447	3.143	3.707	5.208	5.959
7	1.895	2.365	2.998	3.499	4.785	5.408
8	1.860	2.306	2.896	3.355	4.501	5.041
9	1.833	2.262	2.821	3.250	4.297	4.781
10	1.812	2.228	2.764	3.169	4.144	4.587
11	1.796	2.201	2.718	3.106	4.025	4.437
12	1.782	2.179	2.681	3.055	3.930	4.318
13	1.771	2.160	2.650	3.012	3.852	4.221
14	1.761	2.145	2.624	2.977	3.787	4.140
15	1.753	2.131	2.602	2.947	3.733	4.073
16	1.746	2.120	2.583	2.921	3.686	4.015
17	1.740	2.110	2.567	2.898	3.646	3.965
18	1.734	2.101	2.552	2.878	3.610	3.922
19	1.729	2.093	2.539	2.861	3.579	3.883
20	1.725	2.086	2.528	2.845	3.552	3.850
21	1.721	2.080	2.518	2.831	3.527	3.819
22	1.717	2.074	2.508	2.819	3.505	3.792
23	1.714	2.069	2.500	2.807	3.485	3.768
24	1.711	2.064	2.492	2.797	3.467	3.745
25	1.708	2.060	2.485	2.787	3.450	3.725
26	1.706	2.056	2.479	2.779	3.435	3.707
27	1.703	2.052	2.473	2.771	3.421	3.690
28	1.701	2.048	2.467	2.763	3.408	3.674
29	1.699	2.045	2.462	2.756	3.396	3.659
30	1.697	2.042	2.457	2.750	3.385	3.646
40	1.684	2.021	2.423	2.704	3.307	3.551
60	1.671	2.000	2.390	2.660	3.232	3.460
80	1.664	1.990	2.374	2.639	3.195	3.416
100	1.660	1.984	2.364	2.626	3.174	3.390
1000	1.646	1.962	2.330	2.581	3.098	3.300

Calculating confidence interval estimate

Remember, in a normal distribution (bell-curve), to capture 95% of the data we need to go ± 2 standard deviations from the mean.

Well, look at our table. Notice that, the more samples we get (in other words, the more reliable our data is), the closer the t-score gets to 2 standard deviations.

Sample size	90%	95%	98%	99%	99.80%	99.90%
5	2.015	2.571	3.365	4.032	5.893	6.869
6	1.943	2.447	3.143	3.707	5.208	5.959
7	1.895	2.365	2.998	3.499	4.785	5.408
8	1.860	2.306	2.896	3.355	4.501	5.041
9	1.833	2.262	2.821	3.250	4.297	4.781
10	1.812	2.228	2.764	3.169	4.144	4.587
11	1.796	2.201	2.718	3.106	4.025	4.437
12	1.782	2.179	2.681	3.055	3.930	4.318
13	1.771	2.160	2.650	3.012	3.852	4.221
14	1.761	2.145	2.624	2.977	3.787	4.140
15	1.753	2.131	2.602	2.947	3.733	4.073
16	1.746	2.120	2.583	2.921	3.686	4.015
17	1.740	2.110	2.567	2.898	3.646	3.965
18	1.734	2.101	2.552	2.878	3.610	3.922
19	1.729	2.093	2.539	2.861	3.579	3.883
20	1.725	2.086	2.528	2.845	3.552	3.850
21	1.721	2.080	2.518	2.831	3.527	3.819
22	1.717	2.074	2.508	2.819	3.505	3.792
23	1.714	2.069	2.500	2.807	3.485	3.768
24	1.711	2.064	2.492	2.797	3.467	3.745
25	1.708	2.060	2.485	2.787	3.450	3.725
26	1.706	2.056	2.479	2.779	3.435	3.707
27	1.703	2.052	2.473	2.771	3.421	3.690
28	1.701	2.048	2.467	2.763	3.408	3.674
29	1.699	2.045	2.462	2.756	3.396	3.659
30	1.697	2.042	2.457	2.750	3.385	3.646
40	1.684	2.021	2.423	2.704	3.307	3.551
60	1.671	2.000	2.390	2.660	3.232	3.460
80	1.664	1.990	2.374	2.639	3.195	3.416
100	1.660	1.984	2.364	2.626	3.174	3.390
1000	1.646	1.962	2.330	2.581	3.098	3.300

Calculating confidence interval estimate

Example: We want to estimate inventory for next month. Looking at the last 24 months, the average units sold is 150 with a standard deviation of 10. We want to estimate inventory with 95% confidence.

Mean ± reliability factor * standard error

t-score from table (for sample size use n-1) = 2.069

Standard error = $\frac{10}{\sqrt{24-1}}$ = 2.04

150 + 2.069 * 2.04 = **155** (rounded up)

150 - 2.069 * 2.04 = **145** (rounded down)

We can now tell our boss, that with 95% confidence, we estimate inventory needs to be between 145 and 155 units.

Intro to regression analysis and correlation

What is a regression analysis?

Regression analysis: Used to find **LINEAR** trends in our data. The regression provides us with an equation so that we can make predictions about our data. It also helps us understand how accurate our model is.

Linear

Not linear

Example: You want to predict next month's sales. Obviously there are dozens of factors that impact this number. A regression analysis is a way of sorting out which of these dozens of variables have an impact and add to your predictive model. From your regression, you can figure out which factors matter most, which ones you can ignore, and how certain are you about all of those factors.

What is a regression analysis?

Major components of a regression analysis:

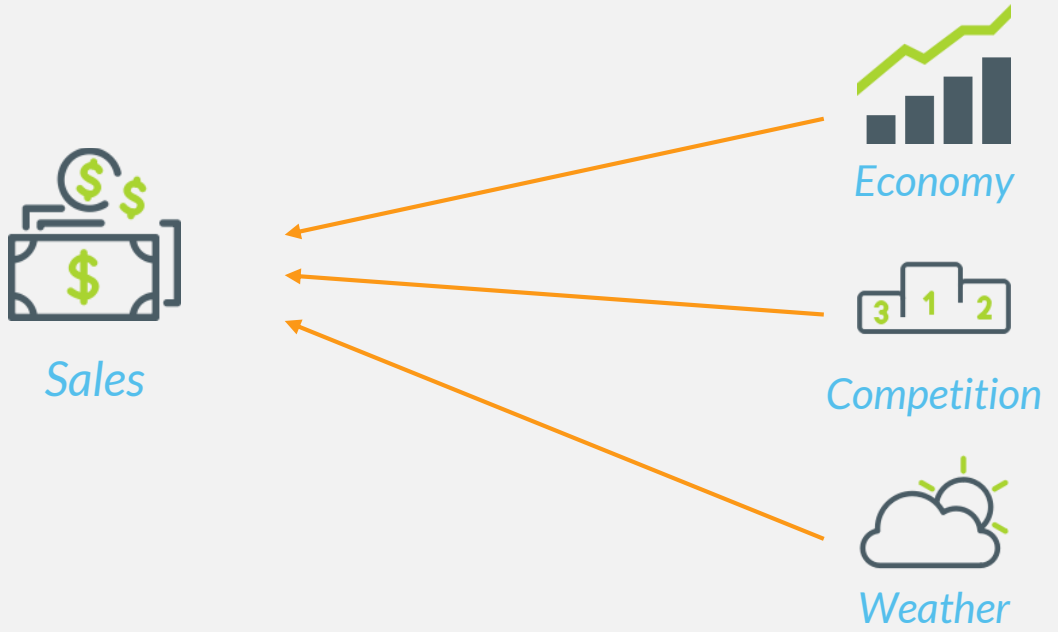
Dependent variable

The main thing you're trying to understand or predict.

Independent variables

The factors you believe have an impact on your dependent variable.

But how can we know if these independent variables (ex. the economy) actually have a meaningful impact on the dependent variable (ex. sales)?



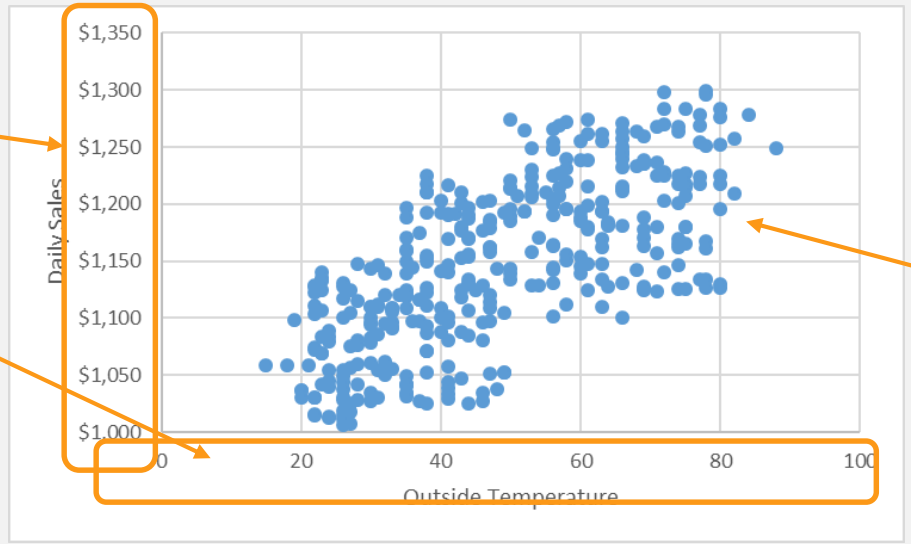
Scatter plot

Scatter plot: A chart that has a bunch of points showing the relationship between two sets of data.

Example: The scatter plot below shows the relationship between the outside temperature (independent variable) and daily sales (dependent variable).

Y (Vertical) axis = *Dependent variable*

X (Horizontal) axis = *Independent variable*



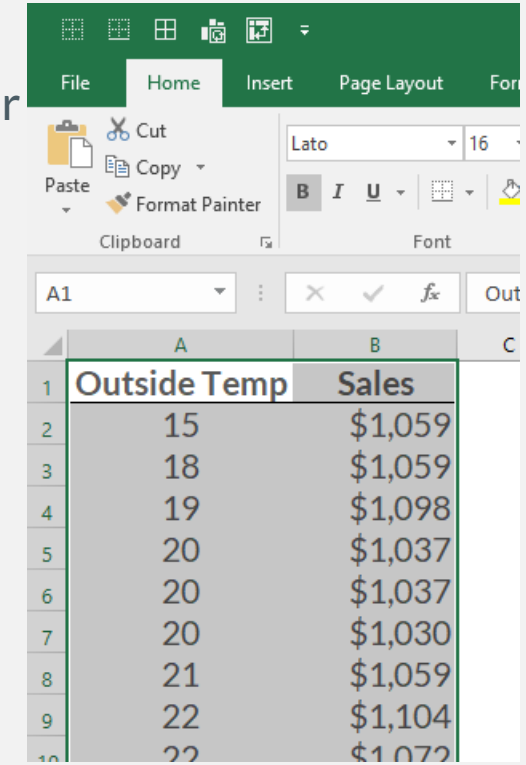
It is pretty clear that there is some type of relationship between the temperature increasing and daily sales increasing. But how do we quantify that? Correlation (which we'll discuss in a minute).

Scatter plot

Creating a scatter plot in Excel

1. Make sure your independent variable is in the first column (ex. outside temperature) and your dependent variable (ex. sales) is in the second column. This will save you time later.

   Select all your data, including the headers.

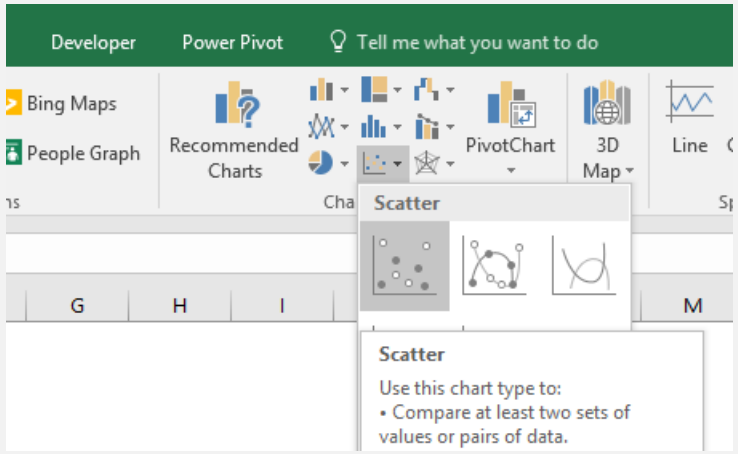


Scatter plot

Creating a scatter plot in Excel

2

On the Ribbon, go to Insert and under the Charts area, select Scatter (the very first option).

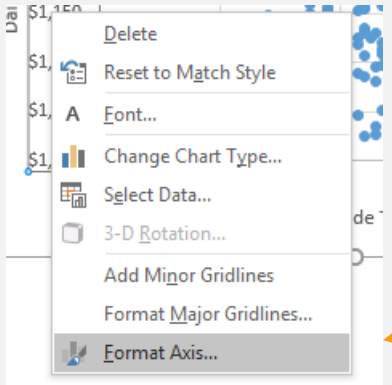


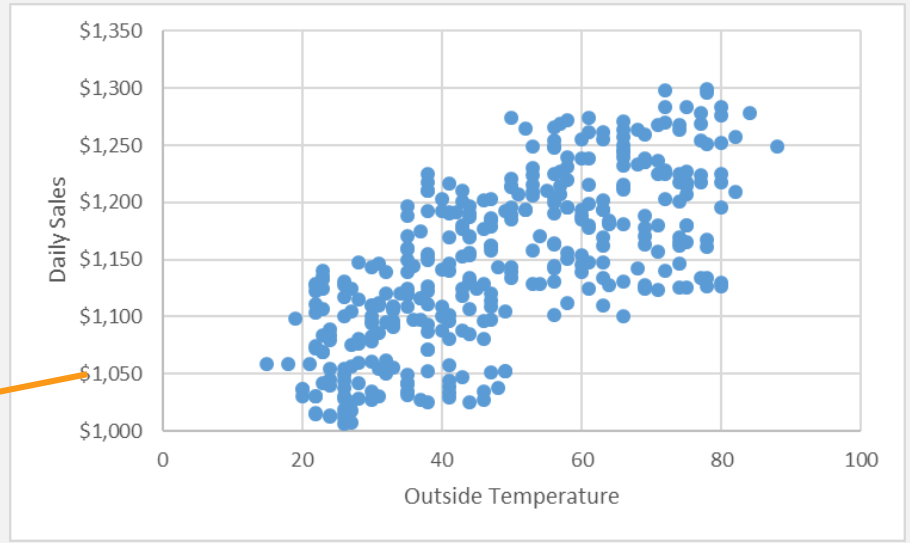
Scatter plot

Creating a scatter plot in Excel

3

Adjust the x and y axis as needed to zoom in on your data. To do this, right click on the axis and select Format Axis. From there, you can change the Bounds.





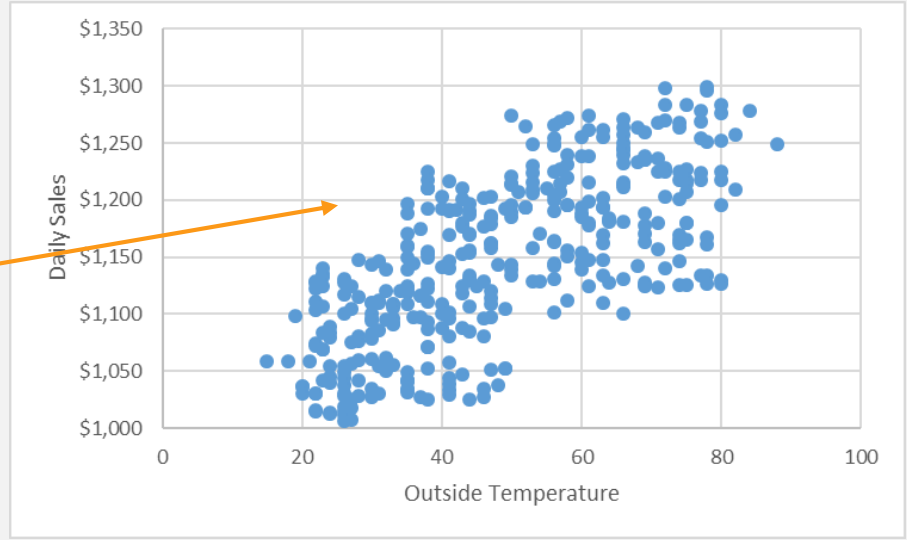
Correlation

Correlation: A quantity (between -1.0 and 1.0) measuring how much two items are dependent upon each other. The closer to 1.0 or -1.0, the higher the correlation.

Example: Using the same example from before…

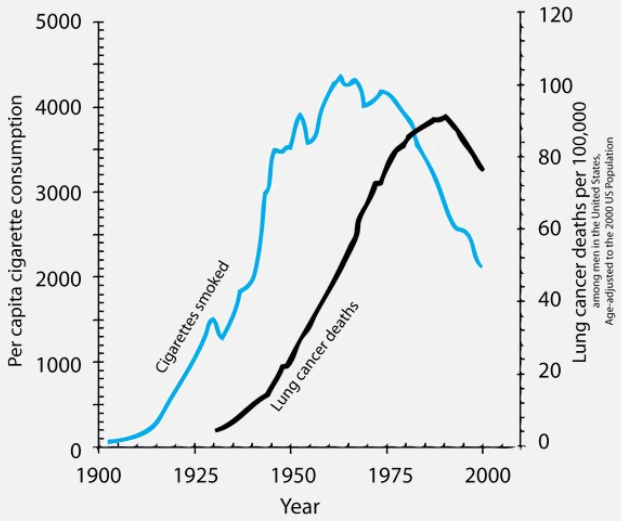
Correlation = .71

In other words, there is a high, positive relationship between the increasing temperature outside and the increase in sales.



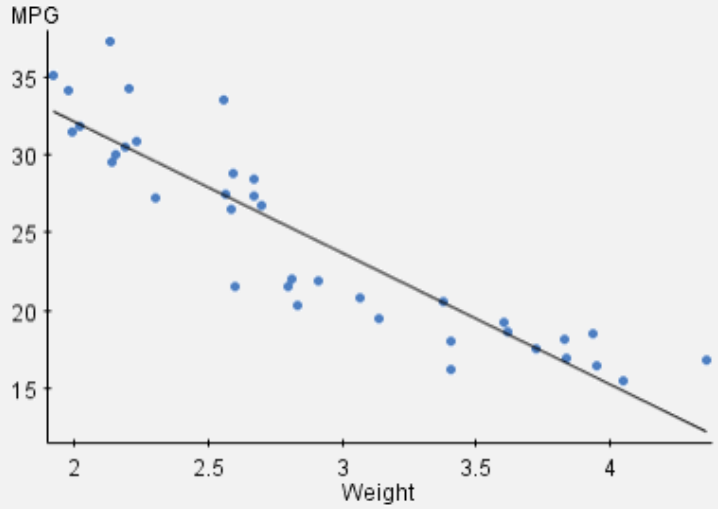
Positive correlation (0-1.0)

An increase in X predicts there will be an increase in Y. (ex. smoking more cigarettes predicts higher lung cancer deaths.)

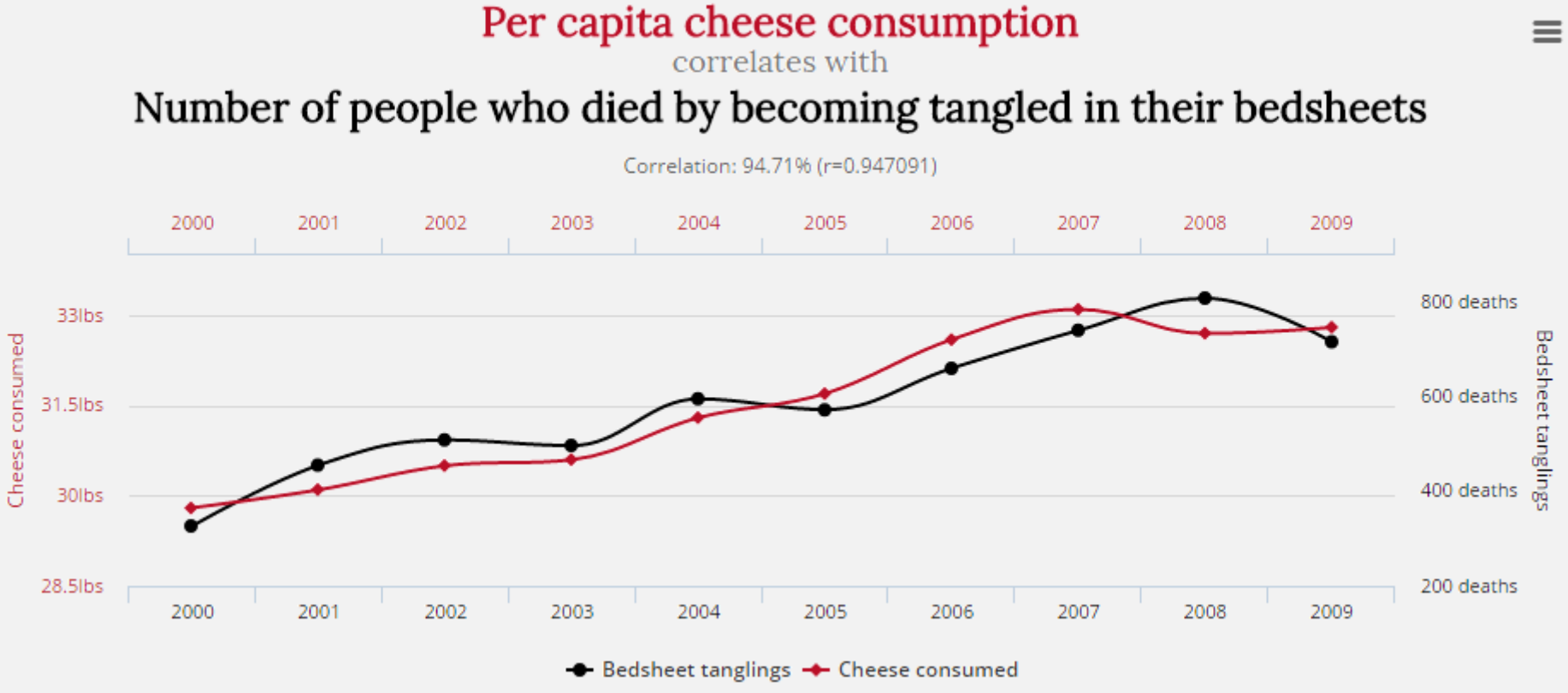


Negative correlation (-1.0-0)

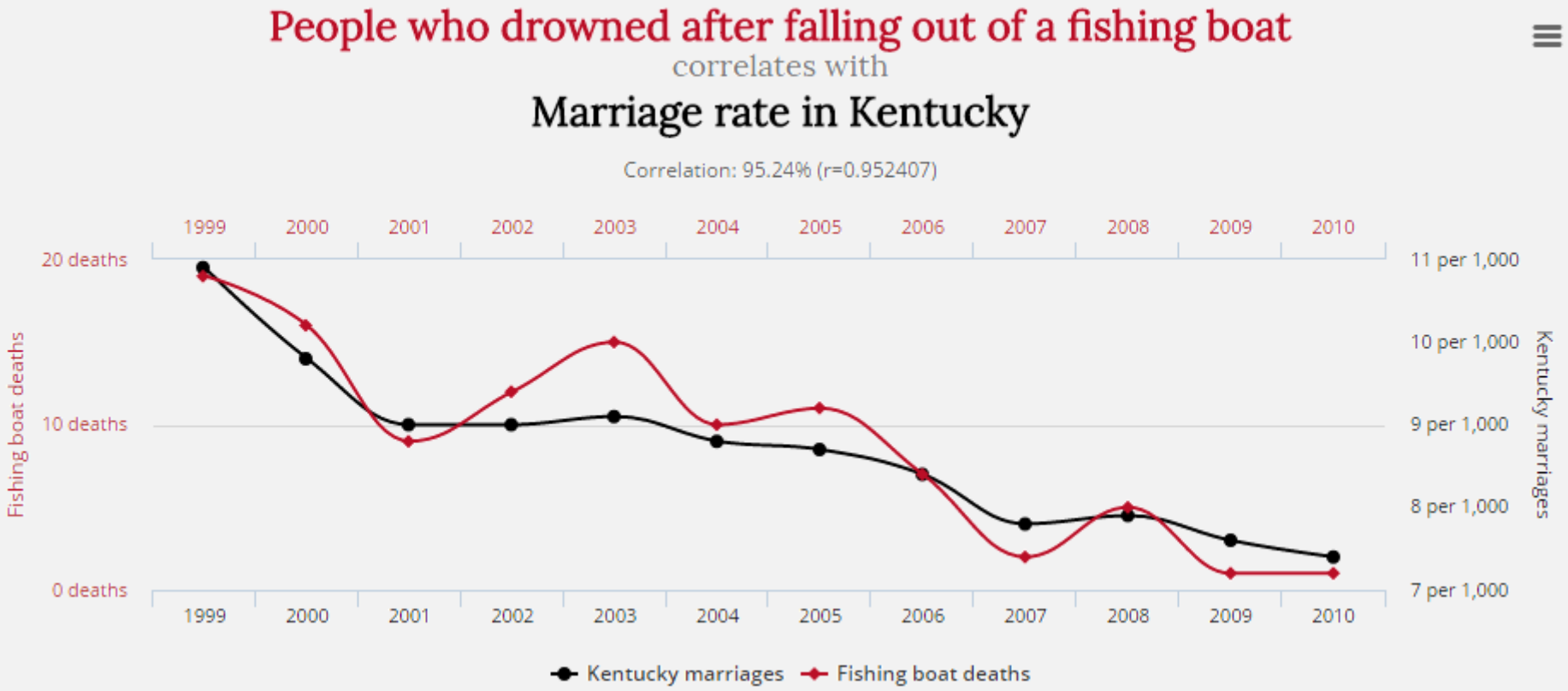
An increase in X predicts there will be a *decrease* in Y. (ex. the heavier the vehicle, the worse the MPG is predicted to be.)



Be careful with correlation!!! Correlation does NOT mean causation!



Be careful with correlation!!! Correlation does NOT mean causation!



Correlation

Calculating correlation in Excel

The actual calculation for correlation is gnarly and knowing the formula won't add much value…

$$Correlation = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Luckily Excel has a built in function to calculate this number for us…

Correlation Excel formula

=CORREL(array1,array2)

It does not matter which order to select your data, you will get the same result.

Correlation

What's considered a "strong" correlation? It depends!

In general:

- 1.0 or -1.0 = Perfect relationship
- 0.7 or -0.7 = Strong relationship
- 0.5 or -0.5 = Moderate relationship
- 0.3 or -0.3 = Weak relationship
- 0.0 = No relationship

But "strong" depends on your field:

Social science: 0.3 - 0.5

Finance: 0.5 - 0.7

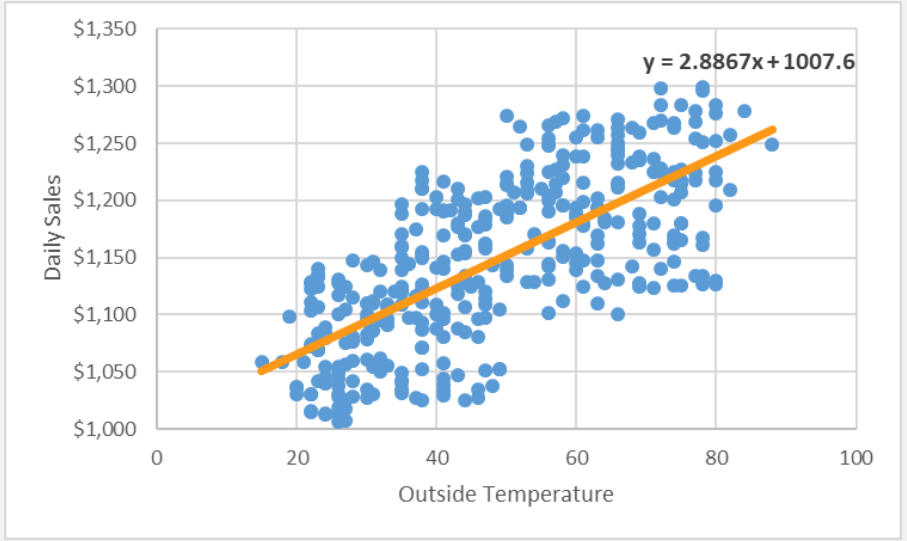
Pharma: 0.7 - 0.9

Trendline or line of best fit

Trendline (line of best fit): A straight line that best represents the data on a scatter plot. This line and its equation help us forecast what future outcomes will be.

Example: Still using our temperature and sales example…

To add a trendline in Excel, select your chart then go to Design -> Add Chart Element -> Trendline -> Linear



Equation of the line

Equation of the line: From the equation of the trendline, we can forecast other results. You'll never have to calculate the equation, Excel or other software will give it to you.

Components of the equation

The infamous equation: ***y = mx + b***

y = The dependent variable we are solving for. This value is on the y axis. In our example, daily sales.

m = The slope of the line. In our example, our slope is 2.89. This means sales increase by $2.89 when it is 1.0 degree warmer outside.

x = The independent variable. This value is on the x axis. In our example, the outside temperature.

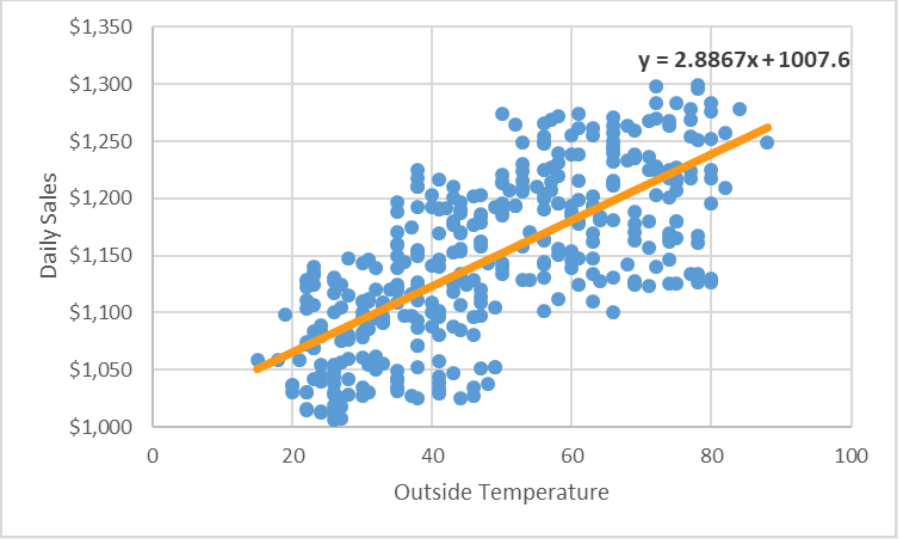
b = The intercept.

Equation of the line

Using the equation of the line to predict sales…

Example: Temperatures are expected to drop significantly next week with highs around 25 degrees. What can we expect daily sales to be?

y ("sales") = 2.89(25 degrees) + 1,007.6 = $1,080



Lingering questions

We are left with a few key lingering questions:

- How accurate is our model?
- How good of a predictor is our independent variable (ex. outside temperatures)?
- What conclusions can we draw?
- Should we keep or throw out our model?

Regression analysis

Running a regression in Excel

Let's first learn how to run a single-variable regression in Excel, then we'll dive in to the individual components of a regression.

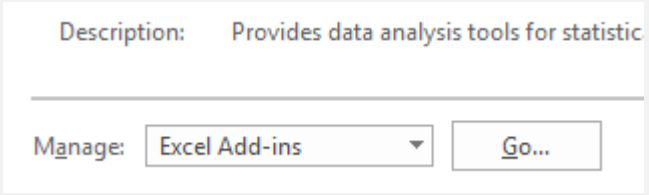
We will only learn the single-variable regression in this course. But, if you can understand this, you can easily run a multivariable regression. We will do this in our Intro to Data Analytics course.

Running a regression in Excel

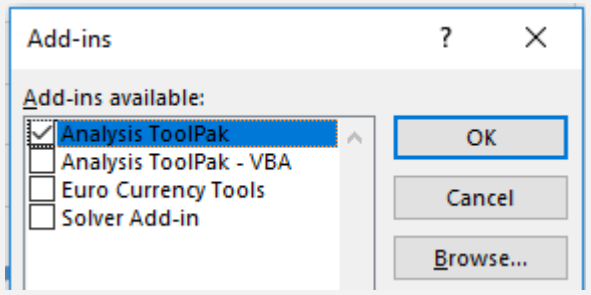
Running a regression analysis in Excel

1. Activate the Analysis Toolpak add-in (no downloading required)

   To get it, go to File -> Options -> Add-ins. At the bottom of the Add-ins screen, select “Excel Add-ins from the dropdown and click “Go”



   Check the “Analysis Toolpak” box and click “OK”

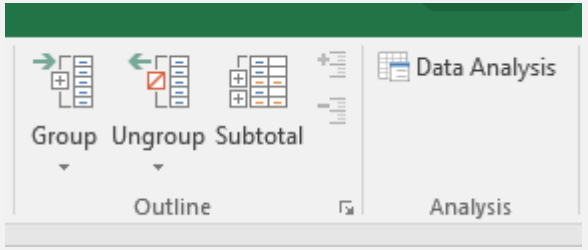


Running a regression in Excel

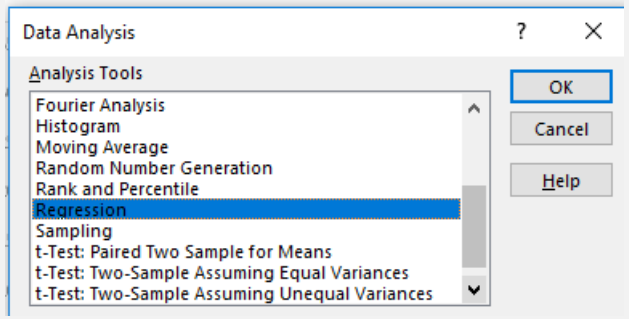
Running a regression analysis in Excel

2

On the Data tab on the ribbon to the far right, click the new Data Analysis icon.



On the popup menu, select “Regression”

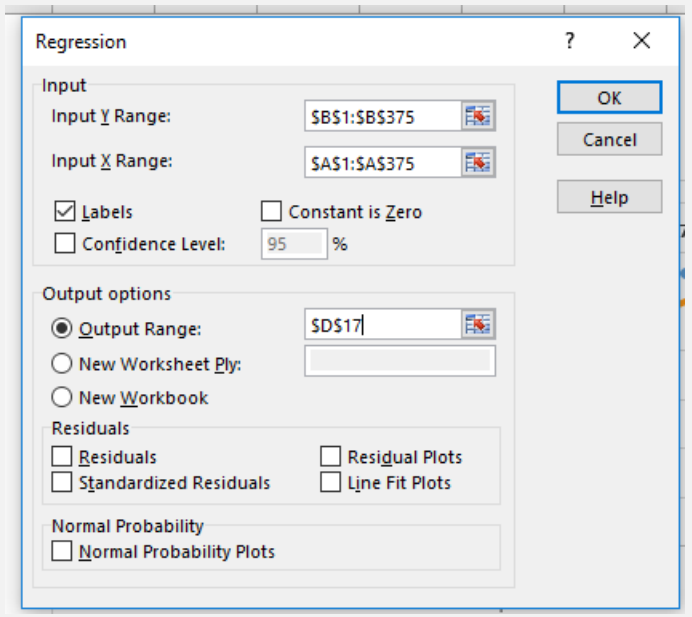


Running a regression in Excel

Running a regression analysis in Excel

3

- Select the appropriate Y variable (the variable you're trying to predict (ex. daily sales). Be sure you include the header.
- Select the appropriate X variable (the variable that predicts the outcome (ex. temperature). Be sure you include the header.
- Check the "Labels" box.
- Select where you want to put your regression output.



The regression output

The output has a bunch of information that looks like gibberish. We only really care about 3 main areas which we'll break down…

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.706
R Square	0.498
Adjusted R Square	0.496
Standard Error	51.421
Observations	374

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	974,890.848	974,890.848	368.696	0.000
Residual	372	983,626.767	2,644.158		
Total	373	1,958,517.615			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1,007.648	7.728	130.397	0.000	992.453	1,022.843	992.453	1,022.843
Outside Temp	2.887	0.150	19.201	0.000	2.591	3.182	2.591	3.182

The regression output

Regression statistics

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.706
R Square	0.498
Adjusted R Square	0.496
Standard Error	51.421
Observations	374

Multiple R: This is just a fancy word for correlation.

R Square: This tells us how well our data fits the model. In our example, our R Square is .498 or 50%. This number means that 50% of the change in sales can be attributed to the change in temperature.

Adjusted R Square: When we have a multivariable regression model, this is the R Square we really care about. This R Square is adjusted based on how many variables we have in the model. We'll explore this more in our Intro to Data Analytics course.

A deeper dive into R-square

R-square: Tells us how close the data is to the trendline (i.e. line of best fit or regression line). In general, the higher the R-squared, the better the model fits your data. The way to calculate R-square is to take the correlation and square it.

As we add in more variables that help explain our data, the higher the R-square will be.

Example: What factors contribute to your income? Many! Just to name a few…

Education

Skills

Marital status

Location

A deeper dive into R-square

Example cont.: As we add more meaningful variables, our R-square increases. This means our model is getting more and more accurate. In a moment we'll explore how to tell if a variable is meaningful or not.

$R^2 = 20\%$

$R^2 = 40\%$

$R^2 = 60\%$

$R^2 = 80\%$

A quick overview of p-values

P-value: At its core, this values helps us understand if a variable in our model is likely to have occurred randomly. If the variable is just some random number, we would not want to include it in our model, so we'd get rid of it.

In general, a value less than 0.05 is good. This means we have 95% confidence that the variable is not random. If the p-value is above 0.05, the probability our variable is random is just too high, so we get rid of it.

Example: This can be confusing so let's break this down into a simple example using a coin.

Let's say we want to check to see if a coin is rigged or not. So we start flipping it.

A quick overview of p-values

Example cont.: Let's say we flip the coin 8 times…

Flip	Result	Probability
1	Heads	50.0%
2	Heads	25.0%
3	Heads	12.5%
4	Heads	6.3%
5	Heads	3.1%
6	Heads	1.6%
7	Heads	0.8%
8	Heads	0.4%

As we keep getting more and more heads, we are more and more certain that the coin is rigged. In other words, we are more and more certain that our results are not random.

In fact, by the 4th head, there is a 6.3% chance that flipping 4 heads in a row is purely random. By the 5th head, there is a 3.1% chance.

This is what the p-value is telling us. We don't want random variables in our model.

The regression output

Back to our regression statistics…

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	974,890.848	974,890.848	368.696	0.000
Residual	372	983,626.767	2,644.158		
Total	373	1,958,517.615			

As a "p-value", should be below 0.05!

There is a lot of stuff here waaaay beyond this course. For us, we only really care about significance f.

Significance F: This number is basically the p-value for the model. This number let's us know if our model is reliable or if it is just a bunch of random numbers. A value less than 0.05 is generally good.

For a simple regression, this number doesn't help us much. This is number is much more important once we have 2+ variables in our model.

The regression output

Regression statistics

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	1,007.648	7.728	130.397	0.000
Outside Temp	2.887	0.150	19.201	0.000

Coefficients: These tell us the equation of our line. So our equation would be y=2.887x + 1,007.648.

Standard Error: The "standard deviation" of our sample.

P-value: The p-value for each variable. **We want a value below 0.05**. For a single variable regression, if this number is above 0.05, we need a whole new variable. For a multivariable regression, if any variable is above 0.05, we should get rid of it.

The p-value for the intercept doesn't really matter, even if it's .999.

The regression output

Tying it all together

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.706
R Square	0.498
Adjusted R Square	0.496
Standard Error	51.421
Observations	374

We want the R-square (or adjusted R-square for a multivariable regression) to be as high as possible. We increase this number by adding variables with low p-values and removing variables with high p-values.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	974,890.848	974,890.848	368.696	0.000
Residual	372	983,626.767	2,644.158		
Total	373	1,958,517.615			

We want the Significance F to be below 0.05.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	1,007.648	7.728	130.397	0.000	992.453
Outside Temp	2.887	0.150	19.201	0.000	2.591

We use the coefficients to create the equation of our line so we can make predictions/forecasts.

The p-value for each independent variable should be below 0.05. We don't really care about the intercept's p-value.

What's next

Where we go from here

Now that we have a foundation in statistics, we can take this knowledge and expand upon it in our **Intro to Data Analytics** course. In that course, we'll apply these principles and expand upon them. For example, we'll actually create a multi-variable regression model.

Courses in this series

SQL for Data Analytics

Learn the code to grab all the relevant data from you database

Stats for Data Analytics

Learn the key statistical tools to analyze our data

Intro to Data Analytics

Apply our statistical tools to real-life business problems

Dashboards and Power BI

Visually represent your analyses and gain more insights from them.