Normal Curves

Today's Goals

- Normal curves!
- Before this we need a basic review of statistical terms. I mean basic as in underlying, not easy.
- We will learn how to retrieve statistical data from normal curves.
- As an application, we'll see how to determine the margin of error of a poll.

Statistics basics

Here's some terminology you should be familiar with:

- **Mean/Average**: For a set of N numbers, $d_1, d_2, \ldots, d_N$, the mean is given by $\mu = (d_1 + d_2 + \cdots + d_N)/N$.
- **Median**: Sort the data set from smallest to largest: $d_1, d_2, \ldots, d_N$. The median is the *middle number*. If N is odd, the median is $d_{(N+1)/2}$. If N is even, the median is the average of $d_{N/2}$ and $d_{(N/2)+1}$.
- **Mode**: The *most common number(s)*. A data set can have more than one mode. (We won't really study mode. It was just feeling left out so I put it on the slide.)
- **Range**: The difference between the highest and lowest values of the data ($R = Max - Min$).

Percentiles

The pth **percentiles** of a data set is a number X_p such that $p\%$ is smaller or equal to X_p and $(100 - p)\%$ of the data is bigger or equal to X_p.

To find the pth percentile of a *sorted* data set $d_1, d_2, \ldots, d_N$, first find the *locator* $L = (p/100)\, N$.

If L is a whole number, then $X_p = \frac{d_L + d_{L+1}}{2}$.

If L is not a whole number, then $X_p = d_{L^+}$ where L^+ is L rounded up.

Percentiles

This is Evelyn.

Evelyn is in the 40th percentile for height (40% of babies Evelyn's age weigh as much or less than she does while 60% weigh as much or more).

Quartiles

- The **first quartile** Q_1 is the 25th percentile of a data set.
- The **median** is the 50th percentile of a data set (also technically the *second quartile*).
- The **third quartile** Q_3 is the 75th percentile of a data set.
- The **fourth quartile** is d_N (the last number in the data set).

The **interquartile range (IQR)** is the difference between the third quartile and the first quartile ($IQR = Q_3 - Q_1$).

IQR tells us how spread out the middle 50% of the data values are.

Hold up!

Why aren't we doing any examples?

Because I'm not going to ask you to compute any of these things directly from a set of data. Instead, we will study visual representations of the data called *bell curves*.

But, I want you to be familiar with the terminology and how it's computed. So bear with me.

Standard deviation and variance

Standard deviation tells us how spread out a data set is *from the mean*.

Let A be the mean of a data set. For each value x in the data set, $x - A$ is the *deviation from the mean*. We want to average these values but for technical reasons we actually need to average their *squares*.

This average is called the **variance** V. The **standard deviation** is the square root of the variance, $\sigma = \sqrt{V}$.

Example

Scores (x)	Deviation (x-A)	(x-A)^2
40.00	-37.29	1390.22
41.00	-36.29	1316.65
48.00	-29.29	857.65
48.00	-29.29	857.65
70.00	-7.29	53.08
73.00	-4.29	18.37
73.00	-4.29	18.37
74.00	-3.29	10.80
77.00	-0.29	0.08
77.00	-0.29	0.08
82.00	4.71	22.22
85.00	7.71	59.51
85.00	7.71	59.51
88.00	10.71	114.80
90.00	12.71	161.65
90.00	12.71	161.65
94.00	16.71	279.37
95.00	17.71	313.80
96.00	18.71	350.22
98.00	20.71	429.08
99.00	21.71	471.51
77.29	0.00	330.78

The number 330.78 is the variance V, the average of squared devations.

The standard deviation is then

$$\sigma = \sqrt{V} \approx 18.19$$

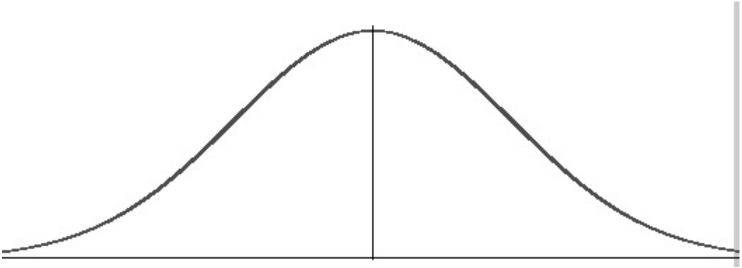
Normal curves

Say we flipped a coin 100 times? We *expect* to get heads 50 times and tails 50 times, but it's also very likely that we would not get this. (For a challenge, compute the probability of this event.)

When John Kerrich was a POW during World War II he wanted to test the probabilistic theory on coin flipping with a real life experiment. He flipped a coin 10,000 times and recorded the number of heads for each 100 trials.

What took Kerrich weeks (months?) we can do in a matter of seconds via computer software like Maple.

Bell curves



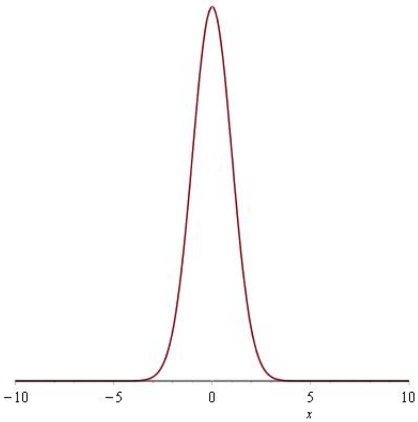
A set of data with **normal distribution** has a bar graph that is perfectly bell shaped.

Properties of normal curves

- **Symmetry**: Every normal curve has a vertical axis of symmetry.
- **Median and mean**: If a data set, then the median and mean are the same and they correspond to the point where the axis of symmetry intersects the horizontal axis.
- **Standard deviation**: The standard deviation is the horizontal distance between the mean and the **point of inflection**, where the graph changes the direction it is bending.

Normal distributions and curve

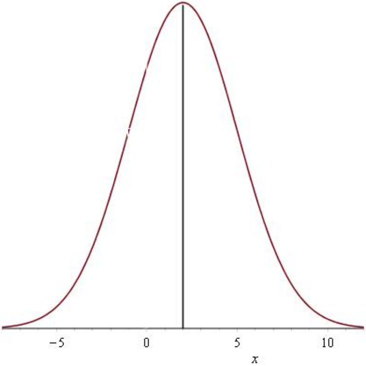
We say a distribution of data is **normal** if its bar graph is perfectly bell shaped.



This type of curve is called **normal**

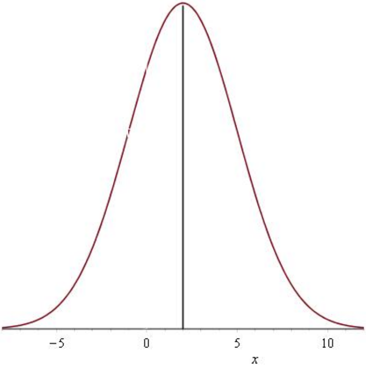
Properties of normal curves

Symmetry: Every normal curve has a vertical axis of symmetry.



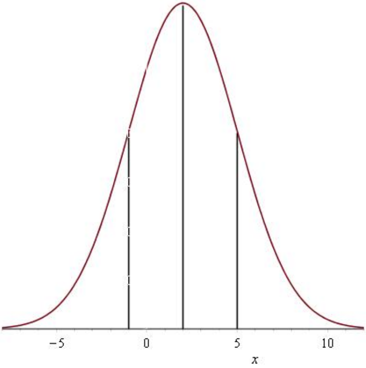
Properties of normal curves

Median and mean: If a data set is normal, then the median and mean are the same and they correspond to the point where the axis of symmetry intersects the horizontal axis.



Properties of normal curves

Standard deviation: The standard deviation is the horizontal distance between the mean and the **point of inflection**, where the graph changes the direction it is bending.



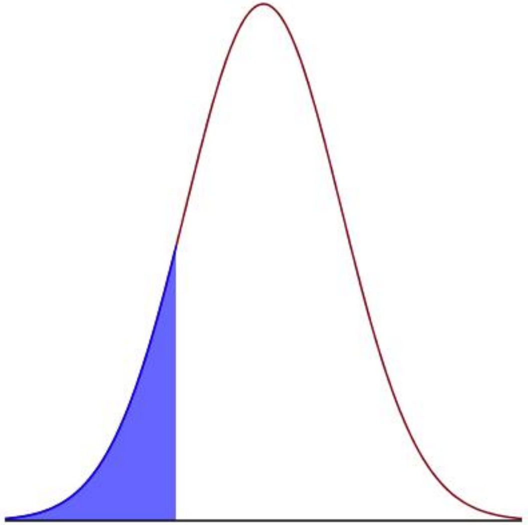
Properties of normal curves

Quartiles: The first and third quartiles can be found using the mean μ and the standard deviation σ.

$$Q_1 = \mu - (.675)\sigma \text{ and } Q_3 = \mu + (.675)\sigma.$$

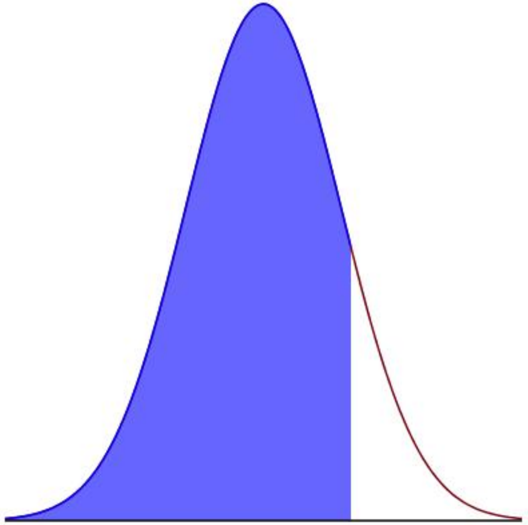
Properties of normal curves

$Q_1 = \mu - (.675)\sigma$



Properties of normal curves

$Q_3 = \mu + (.675)\sigma$



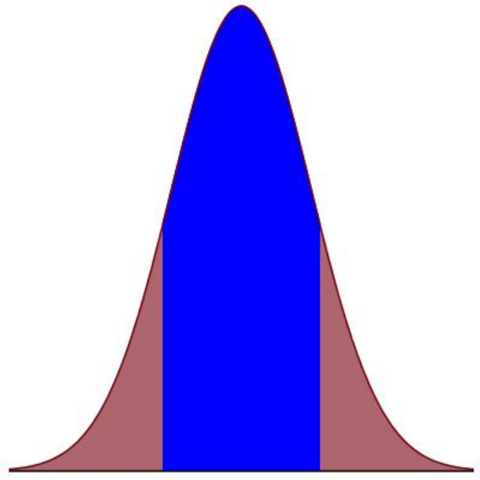
Properties of normal curves

The 68-95-99.7 Rule: In a normal data set,

- Approximately 68% of the data falls between one standard deviation of the mean ($\mu \pm \sigma$). This is the data between P_{16} and P_{84}.
- Approximately 95% of the data falls within two standard deviations of the mean ($\mu \pm 2\sigma$). This is the data between $P_{2.5}$ and $P_{97.5}$.
- Approximately 99.7% of the data falls within three standard deviations of the mean ($\mu \pm 3\sigma$). This is the data between $P_{0.15}$ and $P_{99.85}$.

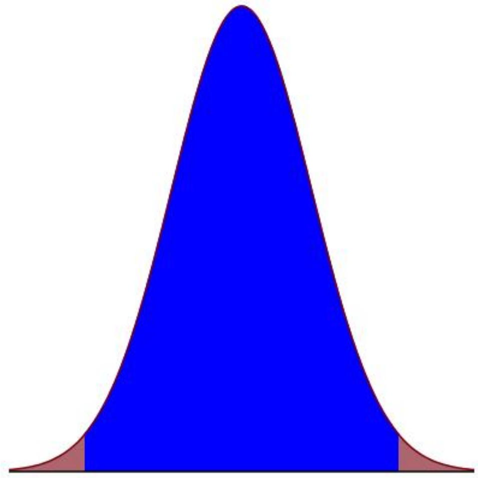
Properties of normal curves

68%



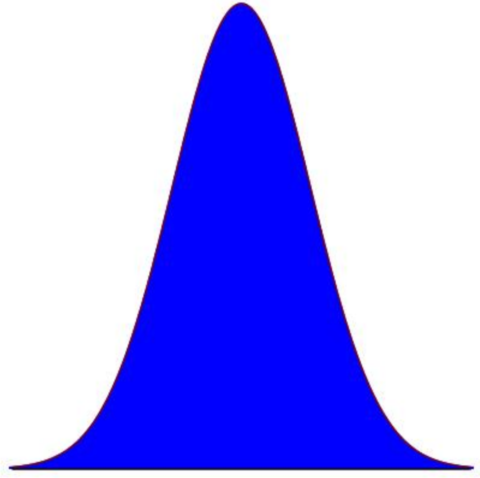
Properties of normal curves

95%



Properties of normal curves

99.7%



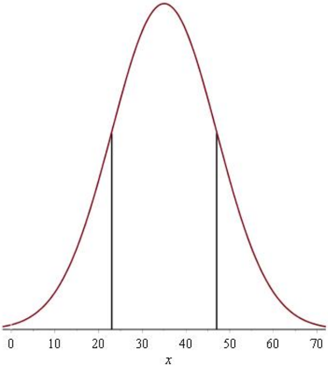
Example

Suppose we have a normal data set with mean $\mu = 500$ and standard deviation $\sigma = 150$. We have the following:

- $Q_1 = 500 - .675 \times 150 \approx 399$
- $Q_3 = 500 + .675 \times 150 \approx 601$
- Middle 68%: $P_{16} = 500 - 150 = 350$, $P_{84} = 500 + 150 = 650$.
- Middle 95%: $P_{2.5} = 500 - 2(150) = 200$, $P_{97.5} = 500 + 2(150) = 800$.
- Middle 99.7%: $P_{0.15} = 500 - 3(150) = 50$, $P_{99.85} = 500 + 3(150) = 850$.

Example

Consider a normal distribution represented by the normal curve with points of inflection at $x = 23$ and $x = 45$. Find the mean and standard deviation. Use them to compute Q_1, Q_3 and the middle 68%, 95%, and 99.7%.



Standardizing normal data

In essence, all normalized data sets are the same. They all have a mean μ and standard deviation σ. The same percentage of data is located in the same increments of σ from the mean. Thus, there is value in *standardizing normal data*.

Psychometry

This is Laura.

Laura is a *psychometrist*. She conducts psychological assessments.

Her patients are adults but their ages range from 18 and up. She uses z-values to standardize her patients' assessment scores.

Standardizing Rule

In a normal distribution with mean μ and standard deviation σ, the standardized value of a data point x is

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

The result of this is the **z-value** of the data point x.

Conversions

Suppose we have a normal data set with mean $\mu = 120$ and standard deviation $\sigma = 30$. If $x = 100$, then the z-value of x is

$$z = \frac{x - 120}{30} = -\frac{2}{3} \approx -.67.$$

If a z-value of some x is .5, what is x (for the data above)?

$$\begin{aligned} .5 &= \frac{x - 120}{30} \\ 15 &= x - 120 \\ 135 &= x \end{aligned}$$

Or, we could recognize that a z-value of .5 means that x is $\frac{1}{2}$ a standard deviation to the right of the mean (so $120 + 15 = 135$).

Variables

In algebra, a variable typically is a placeholder for some type of solution or set of solutions.

Given the equation $x + 3 = 10$, then the variable x represents the number 7.

Given the equation $x^2 + 5 = 21$, then x represents a member of the set of solutions $\{-4, 4\}$.

Random variables

A variable representing a random (probabilistic) event is called a **random variable**.

For example, if we toss a coin 100 times and let X represent the number of times heads comes up, then X is a random variable.

Like an algebraic variable, X represents a number between 0 and 100, but the possible values for X are not equally likely.

The probability of $X = 0$ or $X = 100$ is $(1/2)^{100}$, which is a very small number.

The probability of $X = 50$ is about 8%.

Random variable

Continuing with the example, we know that X has an approximately normal distribution with mean $\mu = 50$ and standard deviation $\sigma = 5$ (for a sufficiently large number of repetitions).

What is the (approximate) probability that X will fall between 45 and 55? This is 1 standard deviation from the mean, so the probability is approximately 68%.

The Honest-Coin Principle

We can now generalize the previous example to a trial with n tosses.

Let X be a random variable representing the number of heads in n tosses of an honest (fair) coin (assume $n \geq 30$).

Then X has an approximately normal distribution with mean $\mu = n/2$ and standard deviation $\sigma = \sqrt{n}/2$.

The Dishonest-Coin Principle

Let X be a random variable representing the number of heads in n tosses of a coin (assume $n \geq 30$), and let p denote the probability of heads on each toss of the coin.

Then X has an approximately normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{np(1-p)}$.

Note that when $p = \frac{1}{2}$ we recover the Honest-Coin Principle.

Margin of Error

In a poll conducted by Public Policy Polling before the recent Democratic primary in Missouri interviewed 839 likely voters.

Their poll found almost a tie between Hillary Clinton and Bernie Sanders.

Therefore, we can use the Honest-Coin Principle to compute the margin of error for the poll.

Margin of Error

According the Honest-Coin Principle, we have

$$\mu = \frac{839}{2} = 419.5 \text{ and } \sigma = \frac{\sqrt{839}}{2} = 14.48.$$

The standard deviation σ is approximately 1.72% of the sample.

This means that the pollsters could assume with 95% confidence that either candidate would get between $(50 \pm 2(1.72))\%$ of the vote. That is, between 46.55% and 53.45%.

The value 2σ is called the **margin of error**.

Margin of Error

On the other hand, in a poll conducted by Public Policy Polling before the recent Democratic primary in North Carolina interviewed 747 likely voters.

Their poll found Hillary Clinton with 60% support and Bernie Sanders with 40%.

Therefore, we can use the Dishonest-Coin Principle to compute the margin of error for the poll.

Margin of Error

According the Dishonest-Coin Principle, we have

$$\mu = 747 * .6 = 448.2 \text{ and } \sigma = \sqrt{747 * .6 * .4} = 13.39.$$

The standard deviation σ is approximately 1.79% of the sample.

This means that the pollsters could assume with 95% confidence that Clinton candidate would get between $(60 \pm 2(1.79))\%$ of the vote. That is, between 56.42% and 63.58%.

The margin of error in this example is $2\sigma = 3.58\%$.