Standard deviation

We have seen that mean and median both measure the "middle" of a set of data. But the mean is easier to compute, while the median is reliably in the middle.

Standard deviation

We have seen that mean and median both measure the "middle" of a set of data. But the mean is easier to compute, while the median is reliably in the middle.

The situation is similar when measuring the "spread" of data. Last time we defined the inter-quartile ratio $Q_3 - Q_1$, which gives us how much of a spread is needed to account for half of the data.

Standard deviation

We have seen that mean and median both measure the "middle" of a set of data. But the mean is easier to compute, while the median is reliably in the middle.

The situation is similar when measuring the "spread" of data. Last time we defined the inter-quartile ratio $Q_3 - Q_1$, which gives us how much of a spread is needed to account for half of the data. The way to measure spread in a computable way, akin to the average, is with the *standard deviation*.

Standard deviation

We have seen that mean and median both measure the "middle" of a set of data. But the mean is easier to compute, while the median is reliably in the middle.

The situation is similar when measuring the "spread" of data. Last time we defined the inter-quartile ratio $Q_3 - Q_1$, which gives us how much of a spread is needed to account for half of the data. The way to measure spread in a computable way, akin to the average, is with the *standard deviation*.

Definition 1. *Let*

$$x_1, .., x_n$$

be a list of data.

Definition 1. *Let*

$$x_1, .., x_n$$

be a list of data. Let $\overline{x}$ *be the mean.*

Definition 1. *Let*

$$x_1, .., x_n$$

be a list of data. Let $\overline{x}$ *be the mean. The standard deviation is given by*

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Why is this a reasonable measure of spread?

Why is this a reasonable measure of spread? If it is small then one expects the quartiles to be close together,

Why is this a reasonable measure of spread? If it is small then one expects the quartiles to be close together, and if it is large then the quartiles should be spread apart.

Why is this a reasonable measure of spread? If it is small then one expects the quartiles to be close together, and if it is large then the quartiles should be spread apart.

Example 2. [Excel example] *Excel can compute standard deviation with the STDEV command.*

Why is this a reasonable measure of spread? If it is small then one expects the quartiles to be close together, and if it is large then the quartiles should be spread apart.

Example 2. [Excel example] *Excel can compute standard deviation with the STDEV command. We can see how the deviation changes for data sets with larger and smaller "spread."*

Which description of data is better: five-number summary or mean and standard deviation?

Which description of data is better: five-number summary or mean and standard deviation?

The mean and standard deviation are always easier to compute; the five-number summary is always more accurate.

Which description of data is better: five-number summary or mean and standard deviation?

The mean and standard deviation are always easier to compute; the five-number summary is always more accurate. Use $\overline{x}$ and σ when you have a symmetric distribution of data. Use the five-number summary otherwise.

Which description of data is better: five-number summary or mean and standard deviation?

The mean and standard deviation are always easier to compute; the five-number summary is always more accurate. Use $\overline{x}$ and σ when you have a symmetric distribution of data. Use the five-number summary otherwise.

If the distribution is approximately symmetric, the median and the mean will be close, and the quartiles will

be about equally placed around the mean. In that case, the mean, and the standard deviation provide a similar level of information.

First manipulations with normal distributions

First manipulations with normal distributions

Last time we were introduced to the all-important "Bell curve," otherwise known as a normal distribution.

First manipulations with normal distributions

Last time we were introduced to the all-important "Bell curve," otherwise known as a normal distribution. There are only three numbers needed to describe a normal distribution:

First manipulations with normal distributions

Last time we were introduced to the all-important "Bell curve," otherwise known as a normal distribution. There are only three numbers needed to describe a normal distribution:

The total number of data points.

First manipulations with normal distributions

Last time we were introduced to the all-important "Bell curve," otherwise known as a normal distribution. There are only three numbers needed to describe a normal distribution:

The total number of data points. This is usually dealt with by "setting it to one and multiplying at the end."

First manipulations with normal distributions

Last time we were introduced to the all-important "Bell curve," otherwise known as a normal distribution. There are only three numbers needed to describe a normal distribution:

The total number of data points. This is usually dealt with by "setting it to one and multiplying at the end."

μ, its mean, which is the center of the distribution, around which it is symmetric

First manipulations with normal distributions

Last time we were introduced to the all-important "Bell curve," otherwise known as a normal distribution. There are only three numbers needed to describe a normal distribution:

The total number of data points. This is usually dealt with by "setting it to one and multiplying at the end."

μ, its mean, which is the center of the distribution, around which it is symmetric

and σ, its standard deviation (hinted at towards the end of the first lecture).

and σ, its standard deviation (hinted at towards the end of the first lecture).

We will learn better what σ is, but now let's see how it can be used.

Theorem 3. *In a normal distribution, with mean μ and standard deviation σ,*

1. *68% of the observations fall within σ of μ (within one standard deviation of the mean).*

Theorem 3. *In a normal distribution, with mean μ and standard deviation σ,*

1. *68% of the observations fall within σ of μ (within one standard deviation of the mean).*

2. *95% of the observations fall within 2σ of μ (within two standard deviations of the mean).*

Theorem 3. *In a normal distribution, with mean μ and standard deviation σ,*

1. *68% of the observations fall within σ of μ (within one standard deviation of the mean).*

2. *95% of the observations fall within 2σ of μ (within two standard deviations of the mean).*

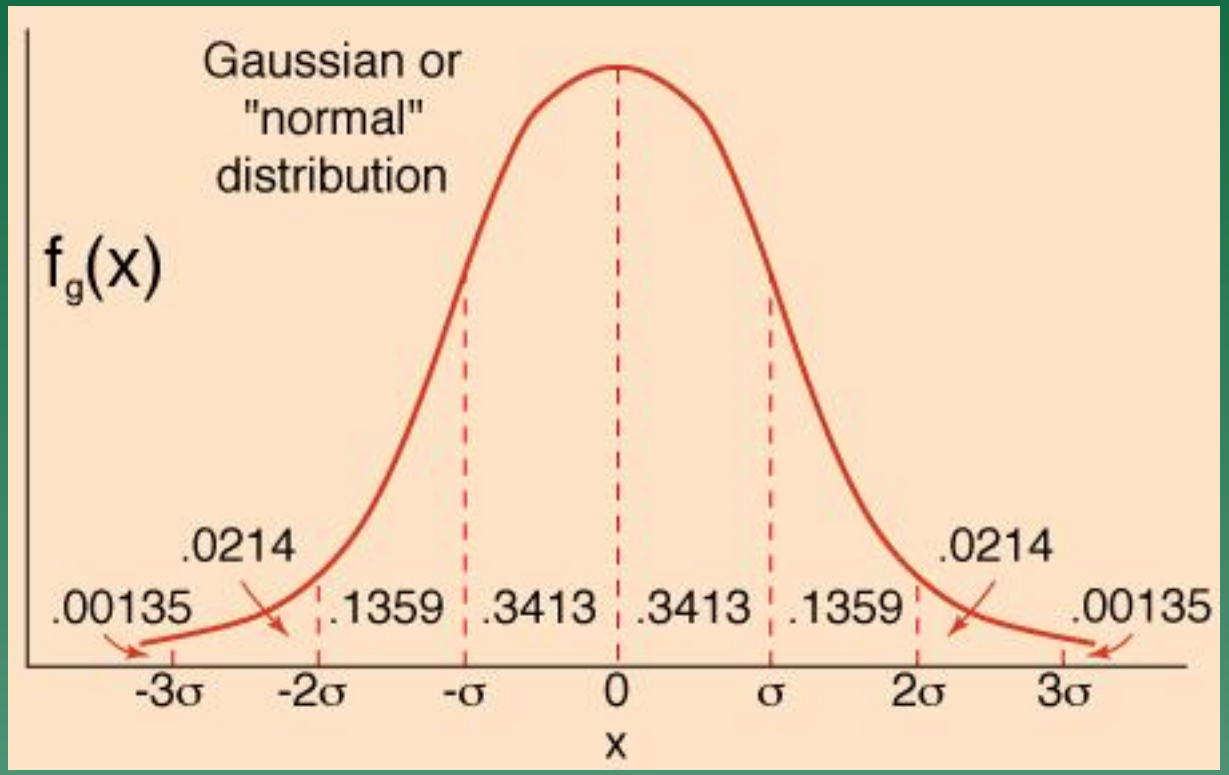
3. *99.7% of the observations fall within 3σ of μ (within 3 standard deviations of the mean).*

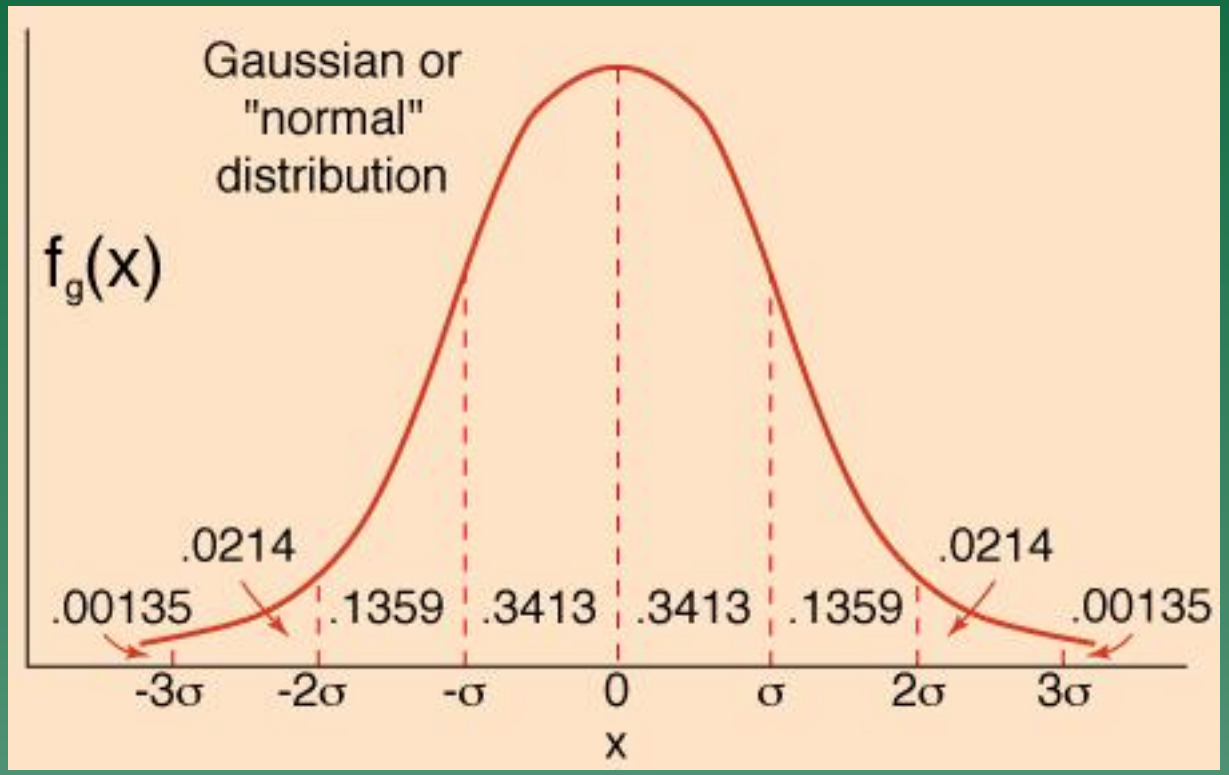
Theorem 3. *In a normal distribution, with mean μ and standard deviation σ,*

1. *68% of the observations fall within σ of μ (within one standard deviation of the mean).*

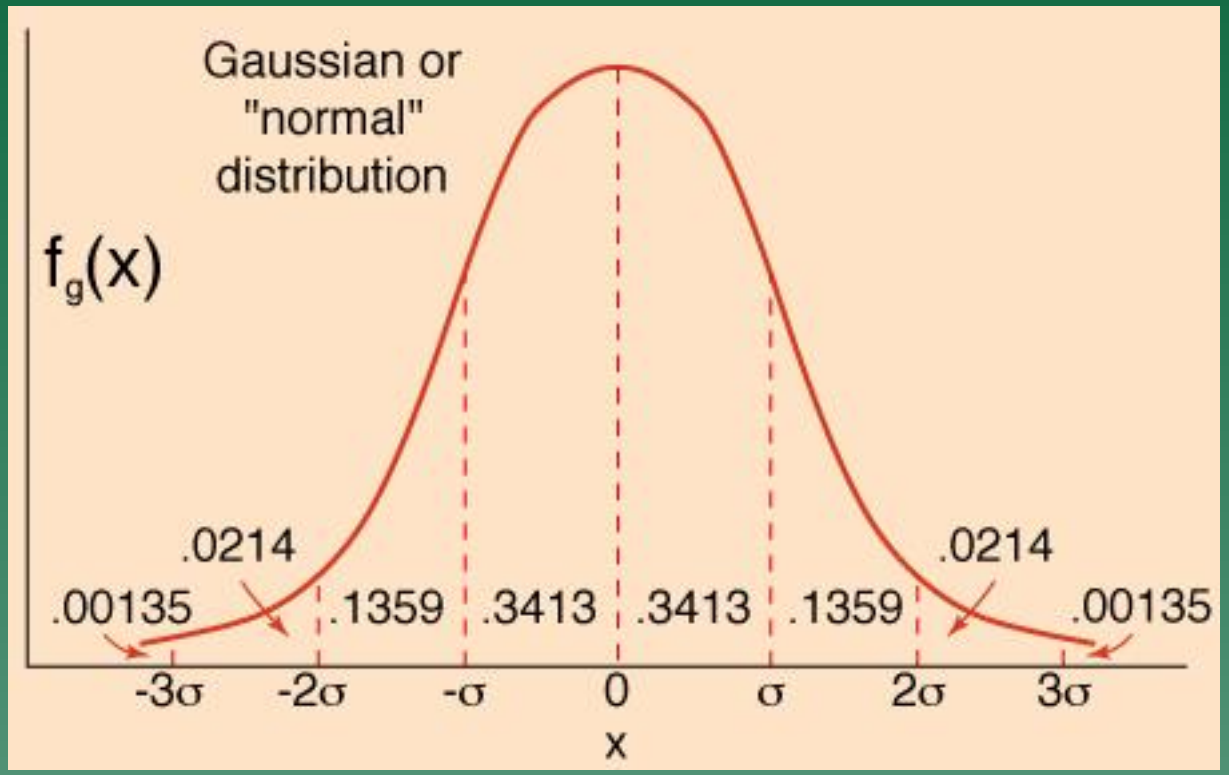
2. *95% of the observations fall within 2σ of μ (within two standard deviations of the mean).*

3. *99.7% of the observations fall within 3σ of μ (within 3 standard deviations of the mean).*


Gaussian or
"normal"
distribution
$f_g(x)$
.0214
.0214
.00135
.1359
.3413
.3413
.1359
.00135
-3σ
-2σ
-σ
0
σ
2σ
3σ
x



Example 4. *The height of adult males in the U.S. is normally distributed with (measured in inches)* $\mu = 69.3$ *and* $\sigma = 2.8$*. Based on this,*



Example 4. *The height of adult males in the U.S. is normally distributed with (measured in inches)* $\mu = 69.3$ *and* $\sigma = 2.8$*. Based on this,*

1. *In our class, how many men should be between 5'6" and 6'? (And how many men are between those heights?)*

1. *In our class, how many men should be between 5'6" and 6'? (And how many men are between those heights?)*

2. *What percentage of men are over six feet tall?*

1. *In our class, how many men should be between 5'6" and 6'? (And how many men are between those heights?)*

2. *What percentage of men are over six feet tall?*

3. *If you were designing a piece of sports equipment with a minimum height needed (golf clubs, hockey sticks), where should you set that height so that over 95% of men could use your equipment?*

Example 5. *Birth weight of babies born in the US is normally distributed with*

$$\overline{x} = 7.31 \text{ pound, and } s = 1.26 \text{ pounds.}$$

Example 5. *Birth weight of babies born in the US is normally distributed with*

$$\overline{x} = 7.31 \text{ pound, and } s = 1.26 \text{ pounds.}$$

Example 5. *Birth weight of babies born in the US is normally distributed with*

$$\overline{x} = 7.31 \text{ pound, and } s = 1.26 \text{ pounds.}$$

Prof. Sinha's daughter Kiri was born at 6.12 pounds (6 pounds, 2 ounces). Roughly, what percentage of babies are born smaller than she?

z-scores and comparing data on different distributions

z-scores and comparing data on different distributions

Let's start with an example question: Who's taller for their gender? A 75 inch tall man, or a 72 inch tall woman?

z-scores and comparing data on different distributions

Let's start with an example question: Who's taller for their gender? A 75 inch tall man, or a 72 inch tall woman? (Who ranks more highly in terms of percentiles?)

z-scores and comparing data on different distributions

Let's start with an example question: Who's taller for their gender? A 75 inch tall man, or a 72 inch tall woman? (Who ranks more highly in terms of percentiles?)

Male height distribution is $N(69.3, 2.8)$. So our man is height 5.7 inches taller than the mean μ.

z-scores and comparing data on different distributions

Let's start with an example question: Who's taller for their gender? A 75 inch tall man, or a 72 inch tall woman? (Who ranks more highly in terms of percentiles?)

Male height distribution is $N(69.3, 2.8)$. So our man is height 5.7 inches taller than the mean μ. But in order to figure out where he is in terms of percentiles, we would need to know how many σ's (standard deviations) he was

away from average.

away from average. This is a matter of arithmetic: $5.7 = 2.04\sigma$. So our man's height is $\mu + 2.04\sigma$ because $75 = 69.3 + 2.04 \times 2.8$.

away from average. This is a matter of arithmetic: $5.7 = 2.04\sigma$. So our man's height is $\mu + 2.04\sigma$ because $75 = 69.3 + 2.04 \times 2.8$.

Female height distribution is $N(64, 2.7)$. So our woman is 8 inches taller than average. We similarly compute that $8 = 2.96\sigma$, and thus our woman's height is $\mu + 2.96\sigma$.

away from average. This is a matter of arithmetic: $5.7 = 2.04\sigma$. So our man's height is $\mu + 2.04\sigma$ because $75 = 69.3 + 2.04 \times 2.8$.

Female height distribution is $N(64, 2.7)$. So our woman is 8 inches taller than average. We similarly compute that $8 = 2.96\sigma$, and thus our woman's height is $\mu + 2.96\sigma$.

We don't need to compute the actual percentages to deduce that our woman is taller for a woman (more standard deviations above the mean) than our man is for a man.

away from average. This is a matter of arithmetic: $5.7 = 2.04\sigma$. So our man's height is $\mu + 2.04\sigma$ because $75 = 69.3 + 2.04 \times 2.8$.

Female height distribution is $N(64, 2.7)$. So our woman is 8 inches taller than average. We similarly compute that $8 = 2.96\sigma$, and thus our woman's height is $\mu + 2.96\sigma$.

We don't need to compute the actual percentages to deduce that our woman is taller for a woman (more standard deviations above the mean) than our man is for a man.

Notice that the numbers we are looking (the multiples of the standard deviation by which our observation is above or below the mean) are given by subtracting the mean and then dividing by the standard deviation.

Notice that the numbers we are looking (the multiples of the standard deviation by which our observation is above or below the mean) are given by subtracting the mean and then dividing by the standard deviation. The resulting numbers are called *z-scores*.

Notice that the numbers we are looking (the multiples of the standard deviation by which our observation is above or below the mean) are given by subtracting the mean and then dividing by the standard deviation. The resulting numbers are called *z-scores*.