z-scores

z-scores

It is common sense that for example on a physical fitness test, a 14-year-old doing 50 sit-ups in a minute is not doing as well as a 12-year-old doing 45 sit-ups, if the first performance is below the 14-y.o. mean while the second is above the 12-y.o. mean.

z-scores

It is common sense that for example on a physical fitness test, a 14-year-old doing 50 sit-ups in a minute is not doing as well as a 12-year-old doing 45 sit-ups, if the first performance is below the 14-y.o. mean while the second is above the 12-y.o. mean. In general to compare data from different populations, we needed to understand how many deviations from the mean that data lies. Formally we have the following.

Definition 1. *Given an observation x among normally distributed data with distribution $N(\mu, \sigma)$, the z-score for x is*

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

z-scores are also sometimes called standardized variables.

The z-score measures how many standard deviations x is above the mean μ (or below, if the z-score is negative).

The z-score measures how many standard deviations x is above the mean μ (or below, if the z-score is negative). It is a standard statistical measure.

The z-score measures how many standard deviations x is above the mean μ (or below, if the z-score is negative). It is a standard statistical measure. For example, the class evaluations you fill out at the end of each quarter get compiled within each department, and instructors see their z-scores.

The z-score measures how many standard deviations x is above the mean μ (or below, if the z-score is negative). It is a standard statistical measure. For example, the class evaluations you fill out at the end of each quarter get compiled within each department, and instructors see their z-scores. If an instructor has z-scores above 1 or below -1 he or she is considered particularly good or, respectively, bad - why would this be?

Example 2. *If a professor gets an average score of* 9.1 *on an item where the department average is* 8.6 *and the standard deviation is* 0.45*, what is her* z*-score on this item?*

Example 2. *If a professor gets an average score of* 9.1 *on an item where the department average is* 8.6 *and the standard deviation is* 0.45*, what is her* z*-score on this item? If instead her average was* 8.2*, what would her* z*-score be?*

Example 2. *If a professor gets an average score of* 9.1 *on an item where the department average is* 8.6 *and the standard deviation is* 0.45*, what is her* z*-score on this item? If instead her average was* 8.2*, what would her* z*-score be?*

But the use of z-scores to evaluate teaching might have its problems.

Example 2. *If a professor gets an average score of* 9.1 *on an item where the department average is* 8.6 *and the standard deviation is* 0.45*, what is her* z*-score on this item? If instead her average was* 8.2*, what would her* z*-score be?*

But the use of z-scores to evaluate teaching might have its problems. First of all, will these scores be normally distributed? (Think about whether the scores are "cut off" anywhere, and what that might do to the distribution).

Example 2. *If a professor gets an average score of* 9.1 *on an item where the department average is* 8.6 *and the standard deviation is* 0.45*, what is her* z*-score on this item? If instead her average was* 8.2*, what would her* z*-score be?*

But the use of z-scores to evaluate teaching might have its problems. First of all, will these scores be normally distributed? (Think about whether the scores are "cut off" anywhere, and what that might do to the distribution). Also, the z-scores tend to correlate with the grades students are expecting - why would this be?

(Correlations will be a big topic later in the term.)

(Correlations will be a big topic later in the term.)

Using z-scores to calculate proportions

Using z-scores to calculate proportions

We did some of these kinds of questions informally before, but with z-scores we can answer them precisely.

Using z-scores to calculate proportions

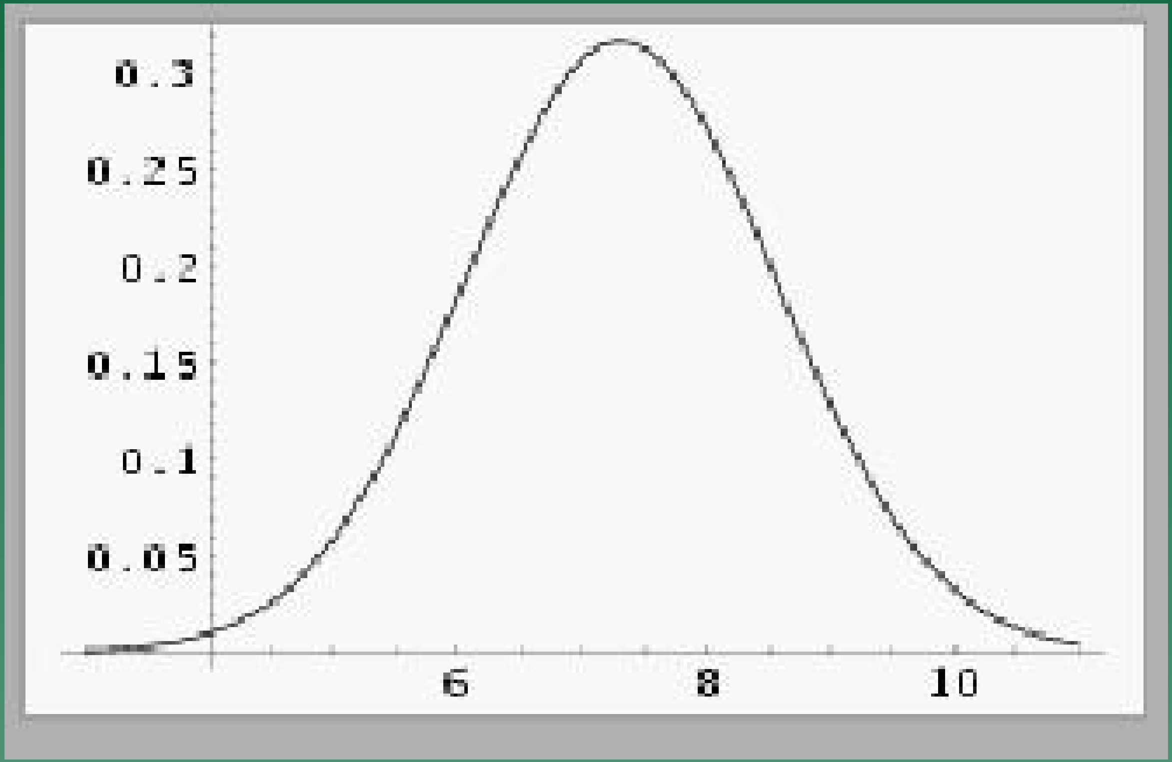
We did some of these kinds of questions informally before, but with z-scores we can answer them precisely.

Example 3. *Birthweight in the United States is normally distributed according to* $N(7.31, 1.26)$. *Precisely what proportion of babies are born over 8 pounds?*

Using z-scores to calculate proportions

We did some of these kinds of questions informally before, but with z-scores we can answer them precisely.

Example 3. *Birthweight in the United States is normally distributed according to* $N(7.31, 1.26)$. *Precisely what proportion of babies are born over 8 pounds?*


0.3
0.25
0.2
0.15
0.1
0.05
6
8
10

The steps we take to solve this problem are:

1. Compute the standardized variable (or z-score) associated to this variable.

The steps we take to solve this problem are:

1. Compute the standardized variable (or z-score) associated to this variable. In this case $z = .548$.

The steps we take to solve this problem are:

1. Compute the standardized variable (or z-score) associated to this variable. In this case $z = .548$.

2. Now we want to know how many babies are born with standardized variable $z \geq .548$.

The steps we take to solve this problem are:

1. Compute the standardized variable (or z-score) associated to this variable. In this case $z = .548$.

2. Now we want to know how many babies are born with standardized variable $z \geq .548$. We look $.548$ up in Table A in the appendix. First we round to $.55$. This number is in the 6th row and 6th column, and is $.7088$.

The steps we take to solve this problem are:

1. Compute the standardized variable (or z-score) associated to this variable. In this case $z = .548$.

2. Now we want to know how many babies are born with standardized variable $z \geq .548$. We look $.548$ up in Table A in the appendix. First we round to $.55$. This number is in the 6th row and 6th column, and is $.7088$. A way to think of this number is that the area under the standardized normal curve with $z \leq .55$ is .7088, while

the area under the whole curve is 1.

the area under the whole curve is 1.

3. Finally, we translate this into a percentage. This means 70.88% of the observations of x satisfy $x \leq \mu + .55\sigma$. So about 71% of babies are born less than 8 pounds, and about 29% of babies are born above 8 pounds.

the area under the whole curve is 1.

3. Finally, we translate this into a percentage. This means 70.88% of the observations of x satisfy $x \leq \mu + .55\sigma$. So about 71% of babies are born less than 8 pounds, and about 29% of babies are born above 8 pounds.

These three steps are worth repeating.

If x is observed in a normal distribution, then to find the percentile associated to x we:

1. Compute the associated z-score, or standardized variable.

2. Look up the z-score in Table A, to get an associated fraction.

3. Multiply by 100 to get a percentage score.

Example 4. *What proportion of babies are born under 5 pounds?*

Example 4. *What proportion of babies are born under 5 pounds?*

Example 5. *What proportion of babies are born between 6 pounds and 8 pounds?*

Example 4. *What proportion of babies are born under 5 pounds?*

Example 5. *What proportion of babies are born between 6 pounds and 8 pounds? (For problems like this, it is especially helpful to draw a picture.)*

Calculating variable ranges from proportions

Calculating variable ranges from proportions

The process we followed above can be reversed in order to find specifications associated to percentages.

Calculating variable ranges from proportions

The process we followed above can be reversed in order to find specifications associated to percentages. We must use Table A in reverse.

Example 6. *You manufacture batteries whose duration times are normally distributed with a mean of 80 hours and a standard deviation of 10 hours. You wish to guarantee to replace batteries that fail before a certain time.*

Calculating variable ranges from proportions

The process we followed above can be reversed in order to find specifications associated to percentages. We must use Table A in reverse.

Example 6. *You manufacture batteries whose duration times are normally distributed with a mean of 80 hours and a standard deviation of 10 hours. You wish to guarantee to replace batteries that fail before a certain time. What time should you choose if you wish to ensure that you replace at most 2.5% of the batteries?*

Calculating variable ranges from proportions

The process we followed above can be reversed in order to find specifications associated to percentages. We must use Table A in reverse.

Example 6. *You manufacture batteries whose duration times are normally distributed with a mean of 80 hours and a standard deviation of 10 hours. You wish to guarantee to replace batteries that fail before a certain time. What time should you choose if you wish to ensure that you replace at most 2.5% of the batteries?*

We need to find a time T so that $x \leq T$ only 2.5% of the time where x is a variable with distribution $N(80, 10)$.

1. Convert 2.5% into a decimal number.

We need to find a time T so that $x \leq T$ only 2.5% of the time where x is a variable with distribution $N(80, 10)$.

1. Convert 2.5% into a decimal number. It is 2.5 hundredths, so it is $.025$.

We need to find a time T so that $x \leq T$ only 2.5% of the time where x is a variable with distribution $N(80, 10)$.

1. Convert 2.5% into a decimal number. It is 2.5 hundredths, so it is $.025$.

2. Find $.025$ in the *inside* of Table A.

We need to find a time T so that $x \leq T$ only 2.5% of the time where x is a variable with distribution $N(80, 10)$.

1. Convert 2.5% into a decimal number. It is 2.5 hundredths, so it is $.025$.

2. Find $.025$ in the *inside* of Table A. This corresponds to -2.81.

We need to find a time T so that $x \leq T$ only 2.5% of the time where x is a variable with distribution $N(80, 10)$.

1. Convert 2.5% into a decimal number. It is 2.5 hundredths, so it is $.025$.

2. Find $.025$ in the *inside* of Table A. This corresponds to -2.81. So the area under the part of the standardized normal curve $z \leq -2.81$ is $.025$.

We need to find a time T so that $x \leq T$ only 2.5% of the time where x is a variable with distribution $N(80, 10)$.

1. Convert 2.5% into a decimal number. It is 2.5 hundredths, so it is $.025$.

2. Find $.025$ in the *inside* of Table A. This corresponds to -2.81. So the area under the part of the standardized normal curve $z \leq -2.81$ is $.025$.

3. Unstandardize:

3. Unstandardize: $z \leq -2.81$ is the same as

$$\frac{x-80}{10} \leq -2.81 \text{ or } x - 80 \leq -28.1 \text{ or } x \leq 51.9.$$

3. Unstandardize: $z \le -2.81$ is the same as

$$\frac{x-80}{10} \le -2.81 \text{ or } x-80 \le -28.1 \text{ or } x \le 51.9.$$

So if we take $T = 51.9$ ($T = 50$ might make better ad copy), you can guarantee to replace batteries that die in less that 51.9 hours and be confident that will be no more that 2.5% of your batteries.

In general (for those who like formulae to follow), suppose x is a normally distributed variable with distribution $N(\mu, \sigma)$ and you wish to find a value C so that K% of the observations of x satisfy $x \leq C$. Take the following steps:

In general (for those who like formulae to follow), suppose x is a normally distributed variable with distribution $N(\mu, \sigma)$ and you wish to find a value C so that K% of the observations of x satisfy $x \leq C$. Take the following steps:

1. Find $K/100$ inside of Table A. Take the correspondong value, call it U.

In general (for those who like formulae to follow), suppose x is a normally distributed variable with distribution $N(\mu, \sigma)$ and you wish to find a value C so that K% of the observations of x satisfy $x \leq C$. Take the following steps:

1. Find $K/100$ inside of Table A. Take the correspondong value, call it U.

2. Then K% of observations of the standard normal variable z will satisfy $z \leq U$.

In general (for those who like formulae to follow), suppose x is a normally distributed variable with distribution $N(\mu, \sigma)$ and you wish to find a value C so that K% of the observations of x satisfy $x \leq C$. Take the following steps:

1. Find $K/100$ inside of Table A. Take the correspondong value, call it U.

2. Then K% of observations of the standard normal variable z will satisfy $z \leq U$.

3. Unstandardize: $z \le U$ is the same as

$$\frac{x-\mu}{\sigma} \le U \text{ or } x-\mu \le \sigma U \text{ or } x \le \sigma U + \mu.$$

3. Unstandardize: $z \leq U$ is the same as

$$\frac{x-\mu}{\sigma} \leq U \text{ or } x-\mu \leq \sigma U \text{ or } x \leq \sigma U + \mu.$$

So $C = \sigma U + \mu$.

Example 7. *Suppose a professor has pre-ordained that 20% of his class should get A's, 40% get B's, 30% get C's, 5% D's and 5 % get F's.*

Example 7. *Suppose a professor has pre-ordained that 20% of his class should get A's, 40% get B's, 30% get C's, 5% D's and 5 % get F's. (Such a system of pre-ordained percentages is called "grading on the (bell) curve", a standard practice at places like MIT.)*

Example 7. *Suppose a professor has pre-ordained that 20% of his class should get A's, 40% get B's, 30% get C's, 5% D's and 5 % get F's. (Such a system of pre-ordained percentages is called "grading on the (bell) curve", a standard practice at places like MIT.) Suppose the final grade averages are normally distributed with an mean of* 72 *and a standard deviation of* 8.

Example 7. *Suppose a professor has pre-ordained that 20% of his class should get A's, 40% get B's, 30% get C's, 5% D's and 5 % get F's. (Such a system of pre-ordained percentages is called "grading on the (bell) curve", a standard practice at places like MIT.) Suppose the final grade averages are normally distributed with an mean of* 72 *and a standard deviation of* 8*. Where should the grade lines be set to achieve the percentages of this system?*