Skewness and Kurtosis, Normal Curve, Standardization

COM 221-ML | CCDATSCL

Central Tendency

In addition to describing the shape of the distribution of a sample or population of measurements, we also describe the data set's central tendency.

A measure of central tendency represents the center or middle of the data. Sometimes we think of a measure of central tendency as a typical value.

1. Mean
2. Mode
3. Median

Population Variance and STDev

The more spread out the population measurements, the larger is the population variance, and the larger is the population standard deviation. And vice versa.

When a population is too large to measure all the population units, we estimate the population variance and the population standard deviation by the **sample variance** and the **sample standard deviation**.

Variance vs Standard Deviation

Standard deviation looks at how spread out a group of numbers is from the mean, by looking at the square root of the variance.

The variance measures the average degree to which each point differs from the mean—the average of all data points.

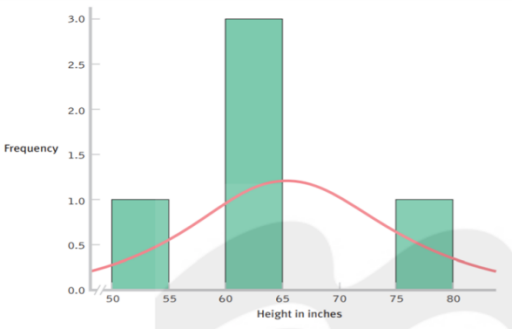
Since, standard deviation is the square root of variance, the value is in the **same units as the mean**.

A normal distribution with mean = 10 and sd = 12 is exactly the same as a normal distribution with mean = 10 and variance = 144.

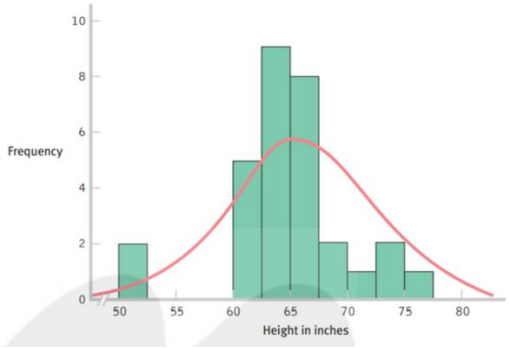
The Normal Distribution / Curve

A normal distribution has a **bell-shaped density** curve described by its mean and standard deviation. The density curve is symmetrical, unimodal, centered about its mean, with its spread determined by its standard deviation.

Identifying the normal curve allows us to **determine probabilities** about data and then **draw conclusions** that we can apply beyond the data.



Here is a histogram of the heights, in inches, of **5 students**. With so few students, the data are unlikely to closely resemble the normal curve that we would see for an entire population of heights.

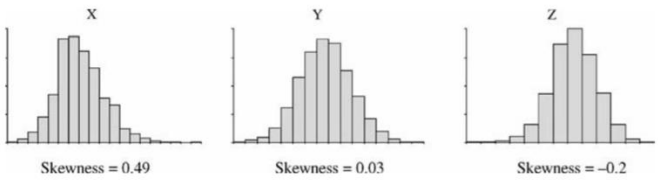


Here is a histogram of the heights, in inches, of **30 students**. With a larger sample, the data begin to resemble the normal curve of an entire population of heights.

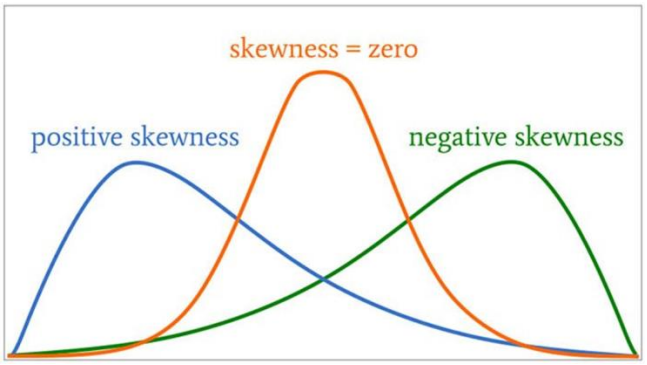
Skewness

There are methods for quantifying the **lack of symmetry** or **skewness** in the distribution of a variable. The formula to calculate skewness, for a variable x, with individual values xi, with n data points, and a standard deviation of s is:

$$\text{skewness} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{(n-1)s^3}$$



- A skewness value of nearer **zero** indicates that the distribution is **symmetrical**.
- A skewness value of **positive** indicates that the right tail is longer than the left tail.
- A skewness value of **negative** indicates that the left tail is longer than the right tail.



- If skewness is less than -1 or greater than 1, the distribution is highly skewed.
- If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed.
- If skewness is between -0.5 and 0.5, the distribution is approximately symmetric

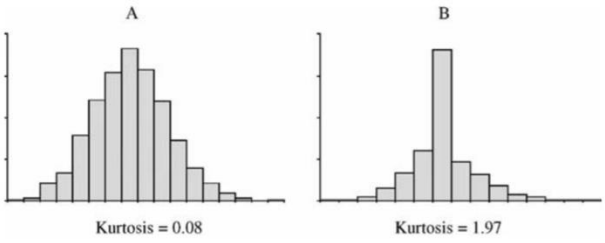
Kurtosis

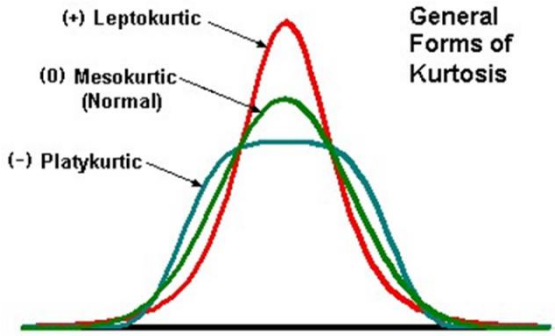
The measurement of how heavily the tails of a distribution differ from the tails of a normal distribution is also important. This measurement is defined as **kurtosis**.

Kurtosis identifies whether the tails of a given distribution contain **extreme values or outliers. The kurtosis of a normal distribution equals / approx. = 1.**

The following formula can be used for calculating kurtosis for a variable x, with xi representing the individual values, with n data points and a standard deviation of s:

$$\text{kurtosis} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{(n-1)s^4}$$

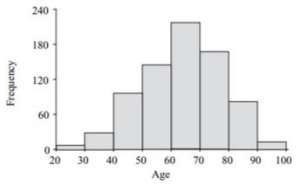




Variables with a **pronounced peak** (leptokurtic) toward the mean have a high kurtosis score (+ excess kurtosis) and variables **with a flat peak** (platykurtic) have a low kurtosis score (- excess kurtosis).

A leptokurtic distribution may be prone to extreme values on either side / outliers while a platykurtic distribution often has flat tails indicating small outliers in a distribution

Describing a variable



Number of observations: 759

Central tendency
Mode: 69, 76, 64, 63
Median: 64
Mean: 63.55

Variation
Minimum: 22
Maximum: 97
Range: 75
Quartiles: 54 (Q1), 64 (Q2), 75 (Q3)
Variance: 194.65
Standard deviation: 13.95

Shape
Skewness: – 0.22
Kurtosis: – 0.47

Figure 5.6. Descriptive statistics for variable **Age**

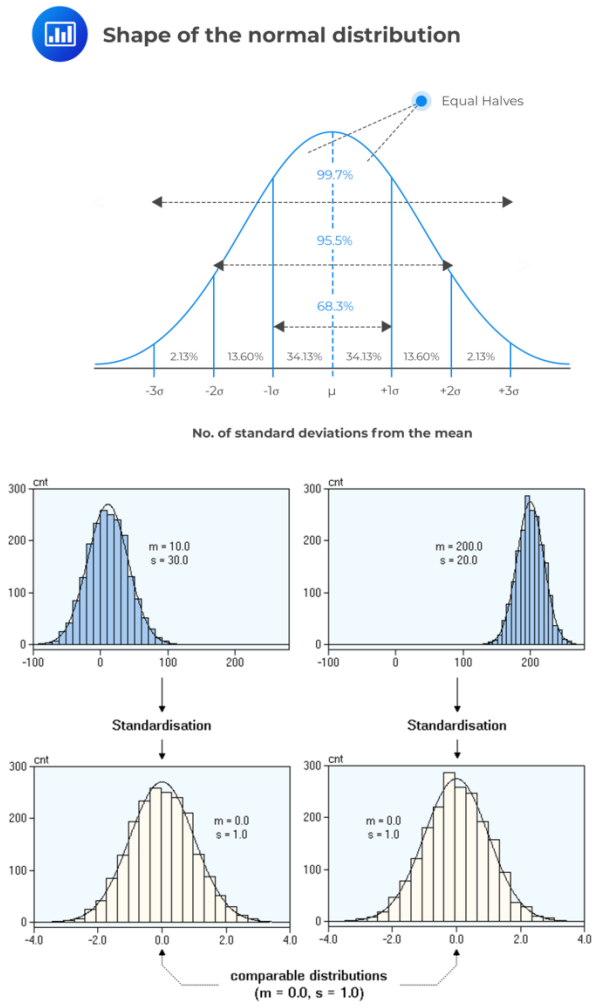
Standardization

Standardization is a way to convert individual scores from different normal distributions to a **shared normal distribution** with a known mean and standard deviation.

We can standardize different variables by using their means and standard deviations to convert any raw score into a **z score**. A **z score** is the **number of standard deviations a particular score is from the mean**.

$$z = \frac{x_i - \bar{x}}{s}$$

Using standardization on your sample data will result to a **z distribution** or a **standard normal distribution (total area under curve = 1).**



Importance of Standardization

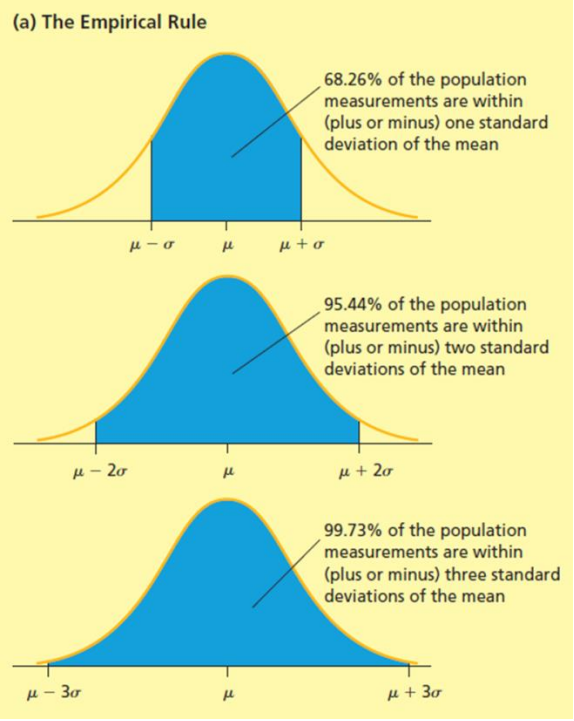
- One of the first problems with making meaningful comparisons is that **variables are measured on different scales**. For example, we might measure height in inches but measure weight in pounds. In order to compare heights and weights, **we need a way to put different variables on the same standardized scale**.
- Let's say you know that after taking the midterm examination, you are 1 standard deviation above the mean in your statistics class. Is this good news? What if you are 0.5 standard deviation below the mean?
- For a test, we know that being above the mean is good; for anxiety levels, we know that being above the mean is usually bad. **z scores create an opportunity to make meaningful comparisons**.

Empirical Rule

One type of relative frequency curve describing a population is the normal curve.

The **normal curve** is a symmetrical, bell-shaped curve and is illustrated on the right.

If a population is described by a normal curve, we say that the population is normally distributed.



The Empirical Rule for a Normally Distributed Population

If a population has **mean** μ and **standard deviation** σ and is **described by a normal curve**, then, as illustrated in Figure 3.14(a),

1. 68.26 percent of the population measurements are within (plus or minus) one standard deviation of the mean and thus lie in the interval $[\mu - \sigma, \mu + \sigma] = [\mu \pm \sigma]$
2. 95.44 percent of the population measurements are within (plus or minus) two standard deviations of the mean and thus lie in the interval $[\mu - 2\sigma, \mu + 2\sigma] = [\mu \pm 2\sigma]$
3. 99.73 percent of the population measurements are within (plus or minus) three standard deviations of the mean and thus lie in the interval $[\mu - 3\sigma, \mu + 3\sigma] = [\mu \pm 3\sigma]$

Example:

Say that you and your friend both took separate quizzes. You earned 92 out of 100; the distribution of your class had a mean of 78.1 and a standard deviation of 12.2. Your friend earned 8.1 out of 10; the distribution of his class had a mean of 6.8 with a standard deviation of 0.74.

Again, we're only interested in the classes that took the test, so these are populations. Who did better?

Again, we're only interested in the classes that took the test, so these are populations. Who did better?

We standardize the scores in terms of their respective distributions.

$$\text{Your score: } z = \frac{(X - \mu)}{\sigma} = \frac{(92 - 78.1)}{12.2} = 1.14$$

$$\text{Your friend's score: } z = \frac{(X - \mu)}{\sigma} = \frac{(8.1 - 6.8)}{0.74} = 1.76$$

Both you and your friend scored above the mean and have positive z scores. Second, we compare the z scores. Although you both scored well above the mean in terms of standard deviations, <u>your friend did better with respect to his class than you did with respect to your class</u>.

z scores are useful because:

- z scores give us a sense of **where a score falls in relation to the mean of its population** (in terms of the standard deviation of its population).
- z scores **allow us to compare scores from different distributions**. Yet we can be even more specific about where a score falls. An additional and particularly helpful use of z scores is that they also have this property.
- z scores **can be transformed into percentiles**.

Z-Table

A z-table, also called the **standard normal table**, is a mathematical table that allows us to know the **percentage of values** below (to the left) a z-score in a standard normal distribution.

Transforming z scores to percentiles

1. Get the z score of a measurement
2. Find the area corresponding to the z-score using a z table
3. Draw a valid conclusion

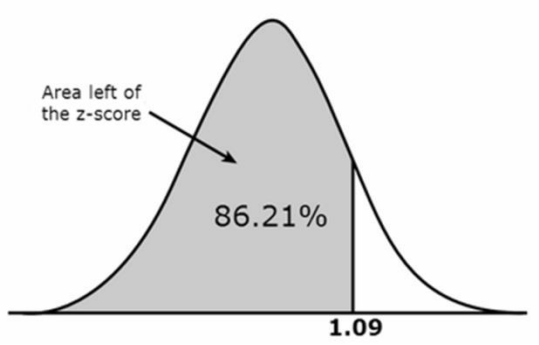
Example #1

50 randomly selected volunteers took an IQ test. Helen, one of the volunteers, scored 74 (x) from maximum possible 120 points. The average score was 62 (μ) and the standard deviation was 11 (σ). How well did she do on the test compared to other volunteers?

1. <u>Get the z score.</u>
   (74 – 62) / 11 = 1.09090909. We can round this number to 1.09 which is the standardized score (same as z-score) that we are going to use.
2. <u>Find the area corresponding to the z-score using a z table.</u>

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830

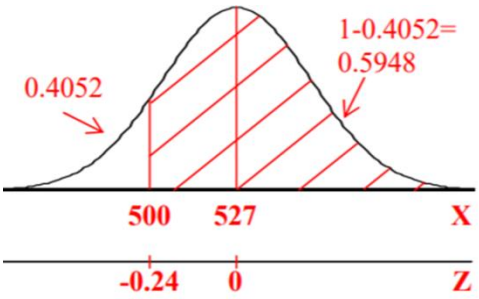
3. <u>Draw a valid conclusion</u>.
   The area that we looked up in the z-table suggests that **Helen received a better score than 86%** of the volunteers who took the IQ test. If you would like to know an exact number of people who Helen outperformed at the test, then just multiply 50 (remember that's how many people took the test) by 0.8621 which is 43.1. As there are no partial human beings, we just round the number to 43. Helen did better than other 43 test takers.



Example # 2

Most graduate schools of business require applicants for admission to take the Graduate Management Admission Council's GMAT examination. Scores on the GMAT are roughly normally distributed with a **mean of 527** and a **standard deviation of 112**. What is the probability of an individual scoring above 500 on the GMAT?

1. Get the z score.
   (500 – 527) / 112 = -0.24197. We can round this number to **-0.24** which is the standardized score (same as z-score) that we are going to use.
2. Find the area corresponding to the z-score using a z table.
   A z score of **-0.24** is equivalent to **0.40517** or **0.4052**.



   We need to get the area on the right side instead of the left side. Thus, **1 - 0.4052 = 0.5948**
3. Draw a valid conclusion.
   Therefore, the probability of scoring above **500** in the GMAT **is 59.48%.**

Example # 3

How high must an individual score on the GMAT in order to score in the highest 5%?

What we know:

mean = 527

standard dev = 112

How do we convert 5% to a z score?

$$z = \frac{(X - \mu)}{\sigma} \qquad X = z(\sigma) + \mu$$

a) **Transform the percentile** to z score. 1 – 0.05 = 0.95 (locate this value in the z table)

b) Use the z score formula to get the value of X.

X = 527 + 1.645 (112)

X = 527 + 184.24

X = 711.24

c) **Draw a valid conclusion**. Thus, you need to have at least 711.24 as a GMAT score to be included in the upper 5%.