AP Statistics – Lesson Notes - Chapter 6: Standard Deviation, Normal Model

Comparing results from different datasets

Scores for college-bound students on the SAT and ACT tests are unimodal and symmetrical with the following means and standard deviations:

SAT: $\bar{x} = 1500$, $s = 250$ **ACT**: $\bar{x} = 20.8$, $s = 4.8$

Which of the following students has a better score?

Student 1: SAT score of 2030 **Student 2**: ACT score of 32

Both seem substantially above the mean. Standard deviation is a measure of spread, so what if we figured out how many standard deviations each is above its mean?

Determining the number of standard deviations a given data value, x, is above its mean is called calculating the data value's ***z-score:***

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

how different this value is from the mean

average difference from the mean

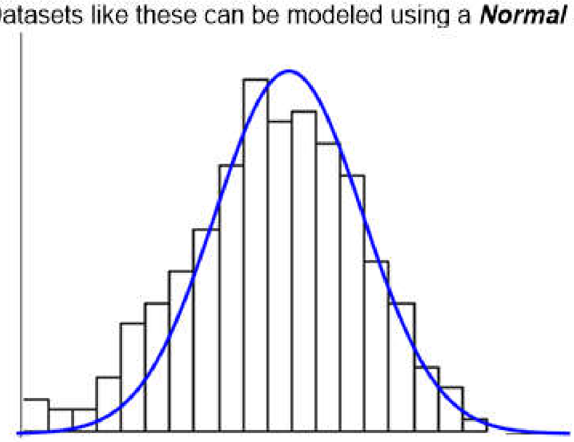
Because z-score has no units, it can be used to compare results from different datasets.

Normal distribution model

How many standard deviations away from the mean does a data value have to be to be considered 'significantly different' than the mean?

In order to answer this, we need a way to model the distribution of the data values. Many datasets are unimodal and symmetrical, with most of the data grouped around the mean and lower frequency as you move away from the mean on both sides.

Datasets like these can be modeled using a ***Normal Distribution Model:***



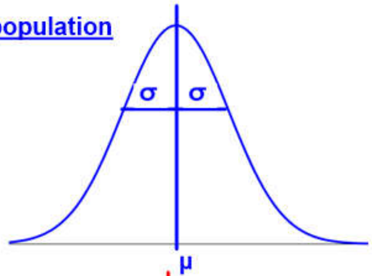
Don't need to know, but in case you're interested here is the probability density function model for a Normal distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

(geogebra demo 'normalparameters.ggb')

Normal distribution model

The population



Greek letters for populations

μ = mean (mu 'mew')

σ = standard (sigma) deviation

These are called the ***parameters.***

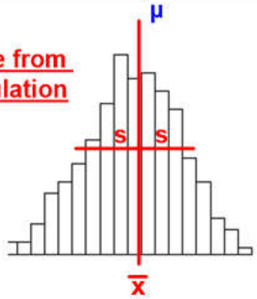
Parameters are (usually unknown) constants

μ = mean σ = standard deviation

estimates

$\bar{x}$ = mean s = standard deviation

A sample from the population



English alphabet letters for samples

$\bar{x}$ = mean

s = standard deviation

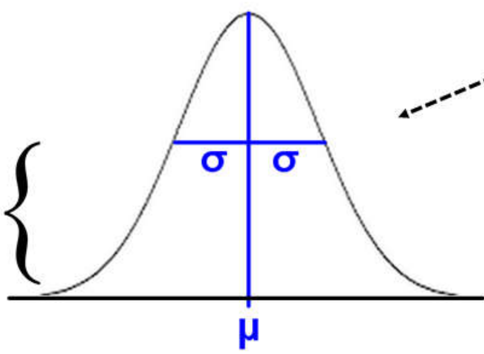
These are called ***statistics.***

Statistics vary for each sample and are estimates of the population parameters

Notation for Normal distribution model: $N(\mu, \sigma)$

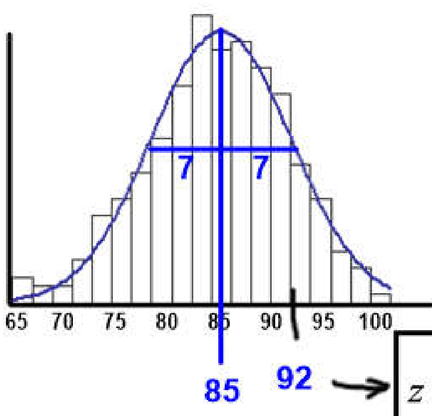
Every time you work a problem involving a Normal distribution, you should sketch the distribution and mark the mean +/- 1 standard deviation, like this:

60.7% of the way up (a little more than halfway up is good for sketching).

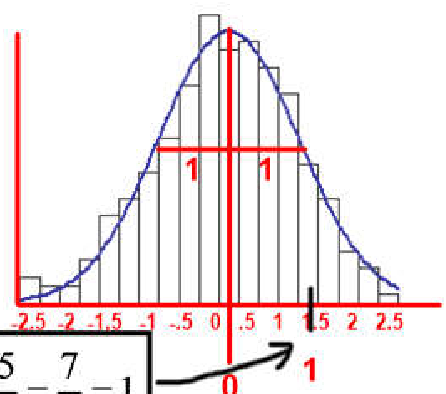


(draw the 'ninja')

We can represent the Normal model data values using the original data values (x) or we can <u>standardize</u> by transforming all the x data values into corresponding z-scores. This produces a **Standard Normal Distribution.**



$$\boxed{z = \frac{92-85}{7} = \frac{7}{7} = 1}$$



Normal Distribution (x values) ***N(85, 7)***

Standard Normal Distribution (z values) ***N(0, 1)***

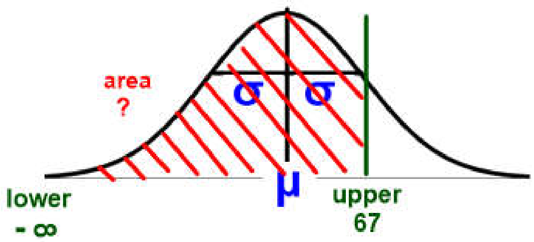
Normal distribution model

The Normal distribution model is a ***probability density function***, which means:

- The area under the whole Normal curve is 1.
  (The sum of all probabilities/percentages = 100%)
- The area under the curve between two x or z values is the percentage of the data in this range (or the probability that one value selected at random will be in this range).

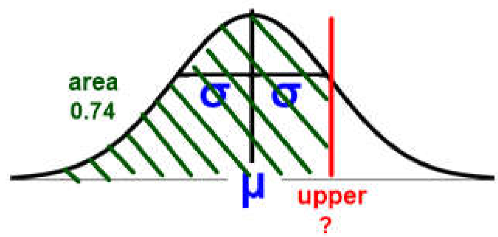
Two calculator functions for normal distributions:

normalcdf(lower, upper, mean, std dev)



use normalcdf when you have boundaries and need area, probability, or percentage

invNorm(area to the left, mean, std dev)



use invNorm when you have an area and need a boundary

Example: The scores on a test are Normally distributed, with a mean of 85 and a standard deviation of 7.

a) What percentage of scores are between 70 and 86?

b) What percentage of scores are between 78 and 92?

c) What percentage of scores are between z=-1 and z=1?

d) What percentage of scores are below z=-1.5?

e) What score is required to be in the top 1%?
(work both ways, using z then x)

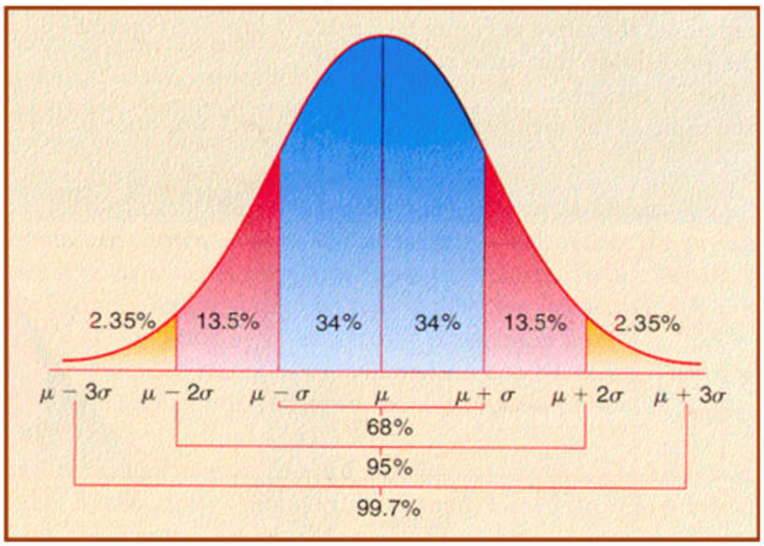
Example: A data set is Normally distributed.

a) What percentage of the data is within 1 standard deviation of the mean?

b) What percentage of the data is within 2 standard deviations of the mean?

c) What percentage of the data is within 3 standard deviations of the mean?

The "Empirical Rule"... ***(memorize the 68-95-99.7 % values)***



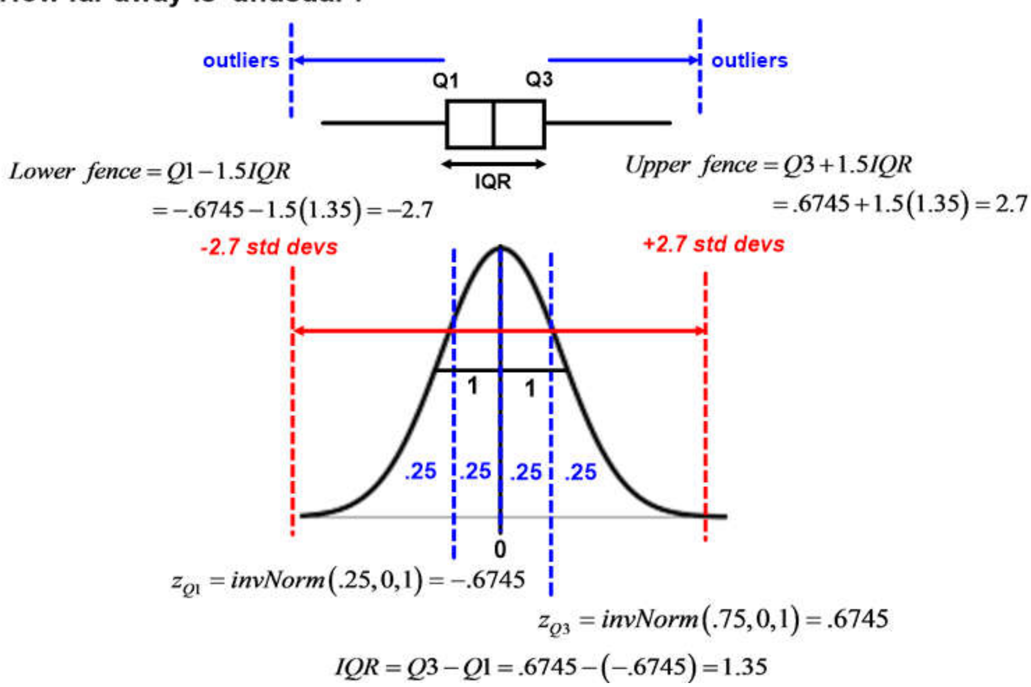
We still haven't answered this question...

"How many standard deviations away from the mean does a data value have to be to be considered 'significantly different' than the mean?"

2σ from the mean
= top/bottom 2.5%

3σ from the mean
= top/bottom 0.15%

How far away is 'unusual'?



How far away is 'unusual'?

For normal distributions...

- +/- 2 standard deviations from the mean is the top/bottom 2.5% (for normal distributions).
- The IQR rule for outliers ('fences') are at +/- 2.7 standard deviations from the mean (for normal distributions).
- +/- 3 standard deviations from the mean is the top/bottom 0.15% (for normal distributions).

Outlier rules-of-thumb (for all distributions)...

1) A data point is an outlier if its value is: $< Q1 - 1.5(IQR) \quad or \quad > Q3 + 1.5(IQR)$

2) A data point is an outlier if its value is more than 2 standard deviations above or below the mean.

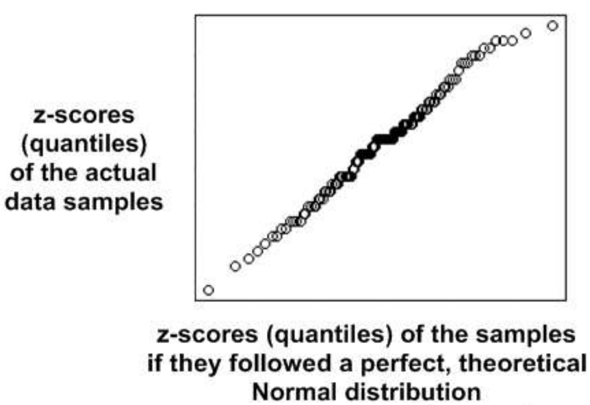
Determining if it is appropriate to use a Normal distribution

In earlier chapters, we've said that we should only use mean and standard deviation if the distribution is **unimodal and symmetrical**. This is referred to as ***Nearly Normal Condition.***

Two ways to determine if a dataset is nearly Normal condition:

- Construct a histogram.
- Construct a Normal Probability Plot (NPP), also known as a Quantile Plot.

Normal Probability (Quantile) Plots

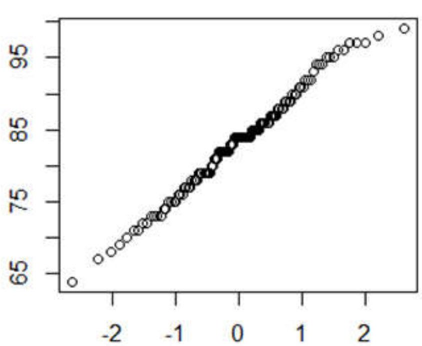


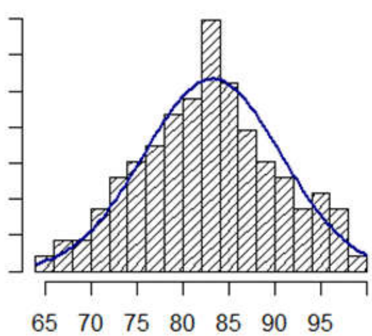
Use calculator:

1) Enter data into list L1.
2) Clear Y= equations.
3) 2nd Y= (Stat Plot):

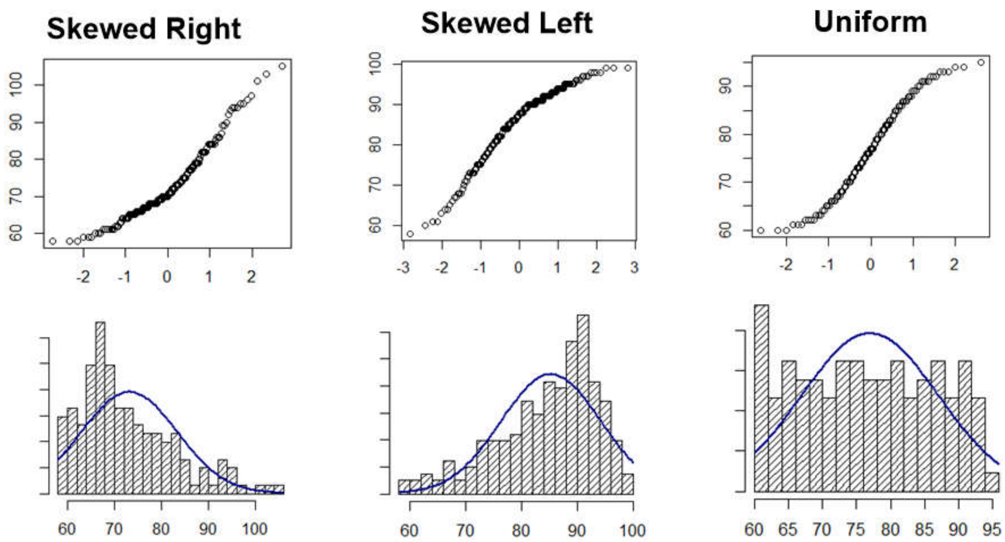
4) Zoom 9: ZoomStat

Interpreting Normal Probability (Quantile) Plots



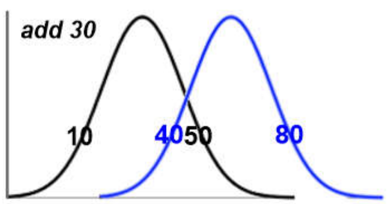


If the NPP is close to a straight line, then the data is close to a Normal distribution shape.



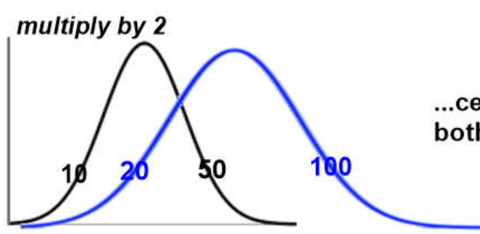
How does a distribution change if we add or multiply all data by a constant?

Add a constant to each data value...



...center changes, but spread does not change.

Multiply each data value by a constant...



...center and spread both change.

$$If\ \ y = ax + b\ \ (multiply\ by\ a, add\ b\ to\ each\ data\ value)$$

$$\mu_y = a\mu_x + b$$

$$med_y = a\,med_x + b$$

$$\sigma^2{}_y = a^2\sigma^2{}_x$$

$$\sigma_y = |a|\,\sigma_x$$

$$IQR_y = |a|\,IQR_x$$

47. First steps. While only 5% of babies have learned to walk by the age of 10 months, 75% are walking by 13 months of age. If the age at which babies develop the ability to walk can be described by a Normal model, find the parameters (mean and standard deviation).