Rescaling, Z-scores and normal distributions

Objective

(1) Learn how rescaling a distribution affects its summary statistics.

(2) Understand the concept of normal model.

Concept briefs:

* Rescaling effect = +/− operations shift the median & mean, but not IQR or SD.

  x/÷ operations scale all summary statistics.

* Z-score = Rescaled data value

  = Distance of data from the mean, measured in SD's.

* Standardizing data = Rescaling all data values into Z-scores.

* Normal model = A perfect "bell-shaped" histogram, often used to approximate symmetric, unimodal distributions.

* The 68-95-99.7 rule = Indicates proportions of spread in normal models.

* Standard normal Vs. other normal models.

Effect of rescaling data distributions

Firstly, what does "rescaling" mean?

* It is somewhat like changing units (e.g, change all heights from feet to inches).

* It involves performing the same sequence of arithmetic operations on every data value.

Two key operations exist: (1) addition (OR, subtraction) of a constant.

(2) multiplication (OR, division) by a constant.

Summary of effect of rescaling

Addition / subtraction affects the "center" but not the "spread" of the distribution.

It shifts the position of every data value by the same amount, so the mean, median, and positions of Q1, Q3, etc. change. But this doesn't change the range, IQR or standard deviation.

Multiplication / division affect the "center" and the "spread" of the distribution.

If we multiply by "m," the mean, median, Q1, Q3, get multiplied by "m" as well.

Furthermore, the IQR and standard deviation get multiplied by "m" as well.

Standardizing and z-scores

Z-score is a very, very important concept!

* Each data value in a distribution can be converted (rescaled) into a "z-score."
* Here is the shortest, easiest, insightful way to think of a z-score

z-score = How many SD's away from the mean is this data value?

E.g., Suppose the mean height of 10 students in a group is 68" and the SD is 4"

(i) Beth is one of these 10 students. Her height is 70". Thus, she is one half SD above the mean. Her z-score is 0.5.

(ii) Mary is also part of this group, and she is 64" tall. So, she is 1 SD below the mean. Her z-score is −1.0.

* Important point: Z-scores have no units, even if the original data has units.

Standardizing

This means converting all values in a distribution to z-scores.

Now guess what is the mean & SD of a standardized distribution?

The Normal Model

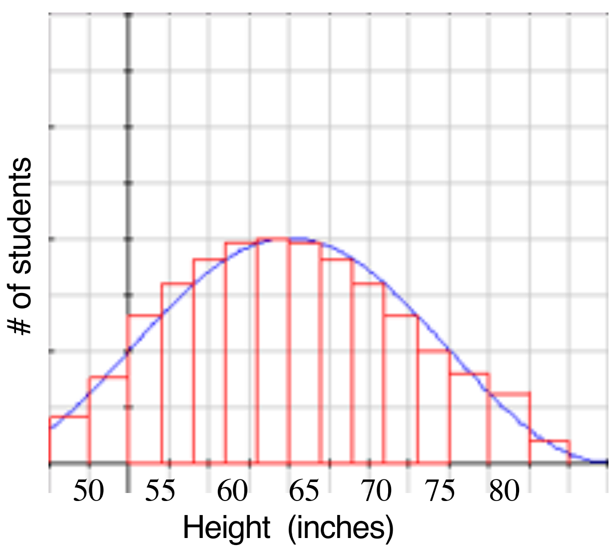
What is the meaning of the term "model" in mathematics?

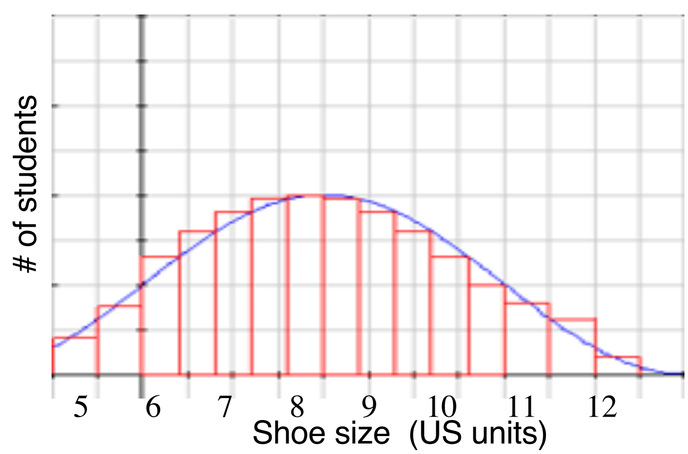
It means an "idealization" of a real-life situation in order to make it feasible to analyze, understand & make useful predictions.

The Normal model refers to a particular graphical shape applicable to analyzing unimodal, symmetric distributions. It looks like the bell curve.

No real-life data distribution is ever perfectly "Normal," but many can be analyzed quite nicely by treating them as normal distributions, if they are roughly symmetric and unimodal.

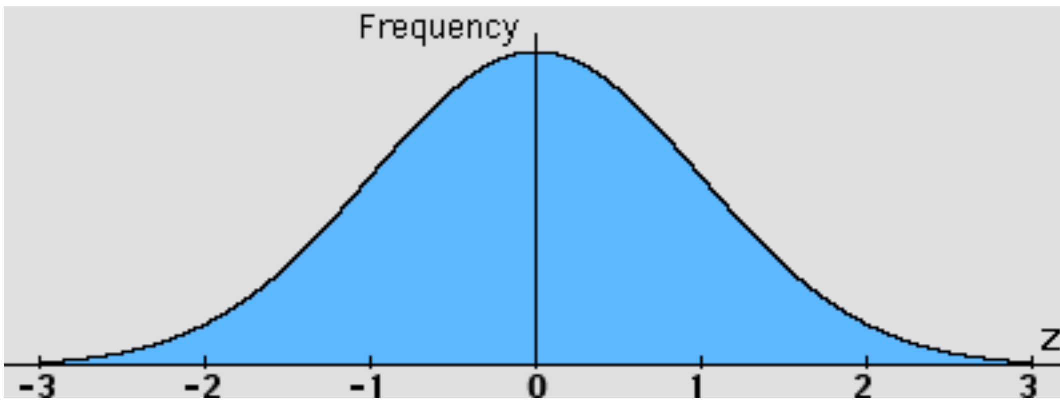
Q: Consider the following 2 distributions: (1) Height of students in this class; and (2) Shoe-size of students in this class. Is it possible to treat both as normal distributions?   How?





Answer: Yes, we can analyze both as normal distributions! Simply standardize & write both distributions in terms of z-scores. Then, if they're both normal, the two graphs will look identical (including same #s on the x-y axes).

IN FACT, any data distribution, if it is "normal," should look like the following graph after standardizing:



Points to note:

* This is a smooth, continuous curve, not a histogram. So, how does it apply to histograms? Answer: We look at the overall profile of histograms.

* The above graph is called the **standard normal model** (because it has mean=0, and standard deviation=1).

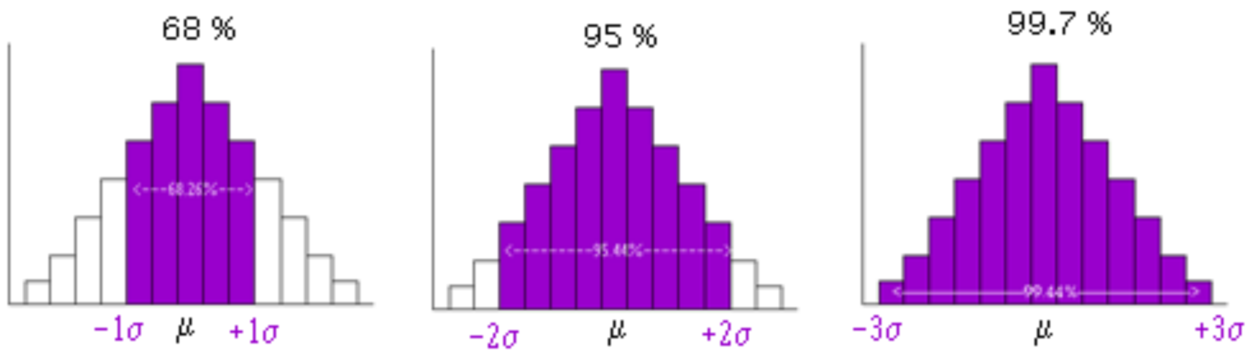
* The normal model can be rescaled to have any other combination of mean and SD. It is not the value of mean/SD that is important, but the shape of the distribution itself.

Key attributes of all normal distributions

* If a distribution is normal, it has some very specific statistical properties.

* This includes the 68-95-99.7 percent rule, which says:

  (1) 68% of all data lies within 1 standard deviation of the mean.
  (2) 95% of all data lies within 2 SD's.
  (3) 99.7% of all data lies within 3 SD's.



* Standard notation: **μ** = mean; **σ** = standard deviation.

The Normal Model

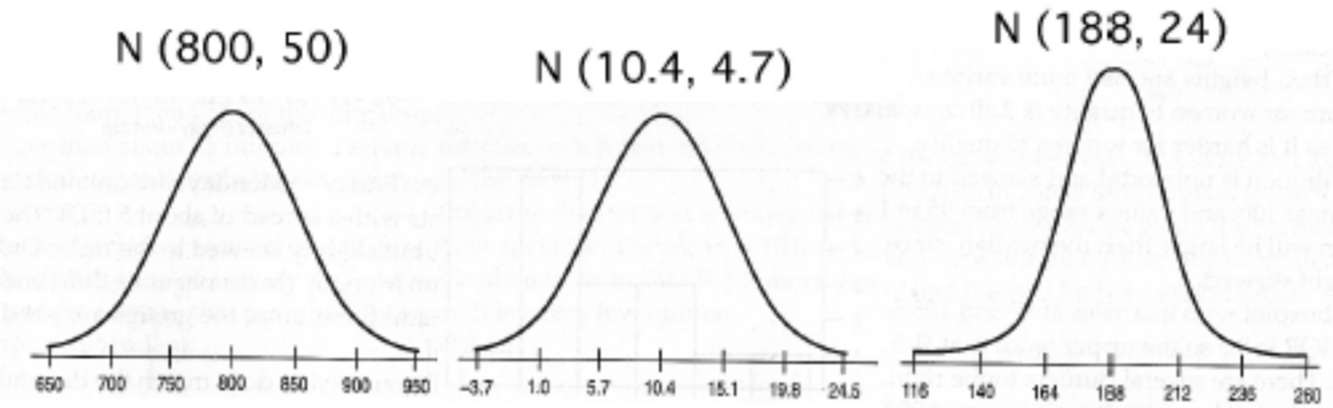
<u>Recall</u>: You can have a normal model with any combination of **μ** and **σ**

* Therefore, there is also special notation for referring to normal models in general

N (**μ**,**σ**) denotes <u>the</u> normal model with mean=**μ** and SD=**σ**

* Note that once you specify **μ**, **σ** there is exactly one normal model with those particular parameters.

Examples of normal models with different choices of **μ** and **σ** :



So how do we use all this stuff in practice?

* The key point here is: We have a mathematical model that is a good approximation to certain kinds of data.

* When a model "fits" some specific data, it becomes very easy to analyze & predict key attributes related to that variable.
E.g., Suppose you had a model for the fluctuations in price of a company's stock. You could make a fortune, quit school & retire in a few months!

* Many real-life distributions are very close to normal:

  * Average local temperature in January for the past 100 years.
  * Weight of peaches harvested in a crop.
  * Student shoe-sizes on campus
  * Volume of soda in 12 oz cans.
  * Blood-pressure or pulse-rates of any (large) group of people.

If we know a distribution is approximately normal, we can estimate many useful details about it.

Illustration: Suppose the distribution of shoe-sizes of students in this class is (roughly) symmetric and unimodal, with mean 8.0 and standard deviation 1.0. Estimate each of the following statistics:

(A) What % of students have shoe-size larger than 9.0?

(B) What % have shoe-size smaller than 6.0?

(C) What % have shoe-sizes between 6.0 and 7.0?

Answers: (a) 32 / 2 = 16%

(b) 5 / 2 = 2.5 %

(c) 27 / 2 = 13.5 %