

THE NORMAL SHAPE

One Quantitative Variable
Unit 1: Exploratory Data Analysis

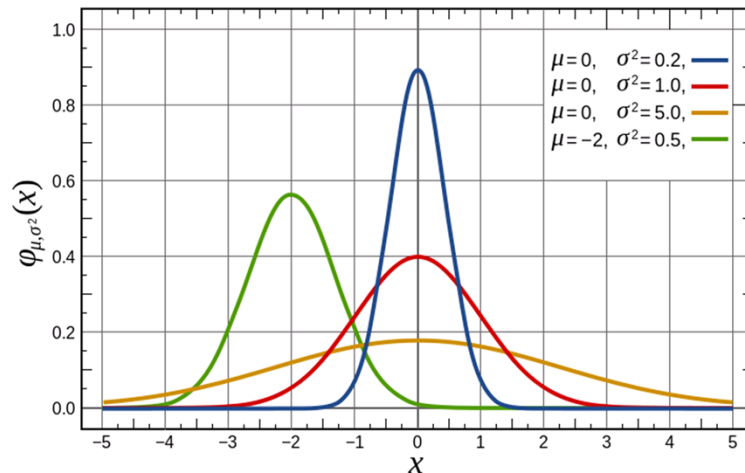


UF UNIVERSITY of
FLORIDA

We will discuss the normal distribution in greater detail in our unit on probability. However, as it is often of use to use exploratory data analysis to determine if the sample seems reasonably normally distributed or not, we will take this opportunity to talk briefly on this topic.

We will also use this topic to help you see how the standard deviation might be useful for distributions which are normally distributed.

The Normal Distribution



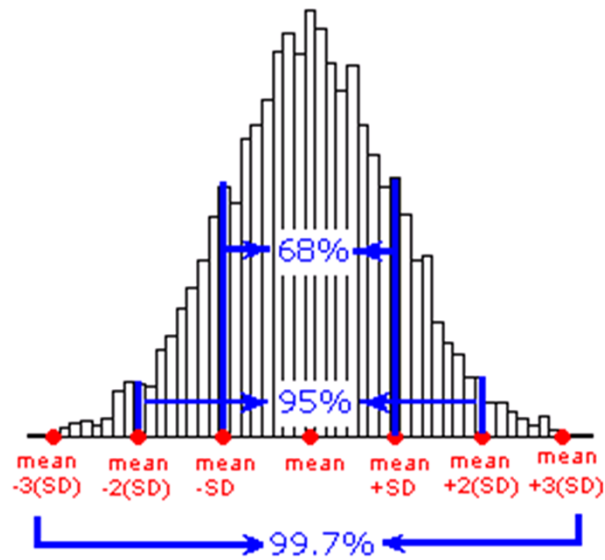
The normal distribution is a distribution which is uniquely identified by its mathematical equation once you know the mean and standard deviation .

It is symmetric and mound shaped and has reasonably small tails (it is rare to find a value very far out in the extremes).

Unfortunately, simply being symmetric and mound shaped isn't all there is to being normally distributed.

For now, we will look at a few simple ways to visually assess the normality of a given quantitative variable.

Standard Deviation Rule



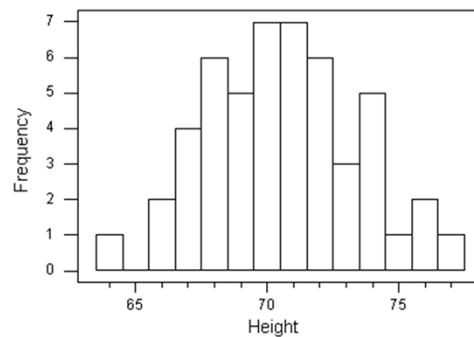
For normally distributed, or approximately normally distributed variables, we can use the standard deviation rule to break up the distribution and determine some percentages within certain ranges.

The standard deviation rule states that:

- Approximately 68% of the observations fall within 1 standard deviation of the mean.
- Approximately 95% of the observations fall within 2 standard deviations of the mean.
- Approximately 99.7% (or virtually all) of the observations fall within 3 standard deviations of the mean.

Example with Data

Interval	Mean-SD, Mean+SD (67.7 , 73.4)	Mean-2(SD), Mean+2(SD) (64.9 , 76.3)	Mean-3(SD), Mean+3(SD) (62 , 79.2)
Percentage of Observations in interval	34 observations 34/50 = 68%	48 observations 48/50 = 96%	All 50 observations 50/50 = 100%
SD Rule says...	68%	95%	99.7%

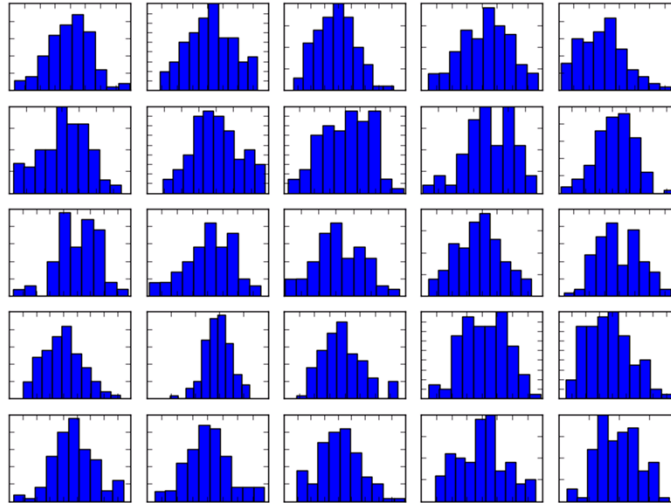


We can use this rule in two directions.

- 1) If we know a distribution is roughly normal and we know the mean and standard deviation, then we can determine approximate percentages for certain ranges
- 2) If we have a dataset and we want to know if it is approximately normally distributed, we can determine what percentage fall within 1, 2, and 3 standard deviations, and compare those to the target values of 68, 95, 99.7.

In the example of heights above, we can see that the data follow the rule fairly closely! We have exactly 68% within 1 standard deviation, we have 96% within 2 standard deviations, and 100% within 3 standard deviations. The histogram does show a distribution with a reasonably normal shape.

Sampling Variability



<http://work.thaslwanter.at/Stats/html/statsDistributions.html>

UF UNIVERSITY of FLORIDA

Even if the population is exactly normally distributed, samples from the population can appear non-normal especially for small sample sizes

Each of these graphs was generate from a normal distribution using samples of size 100. Notice the variation in these graphs. Some of them you might even say are skewed, some look too fat and some too thin.

This is the kind of sampling variation that statistics must conquer!!

Wiki

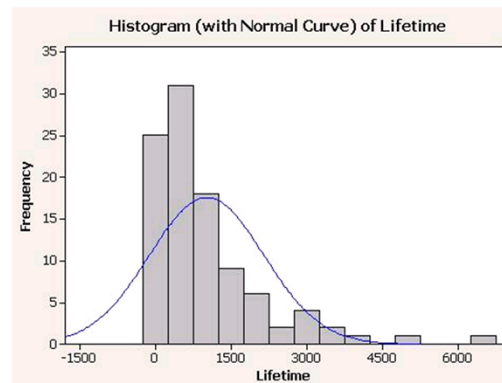
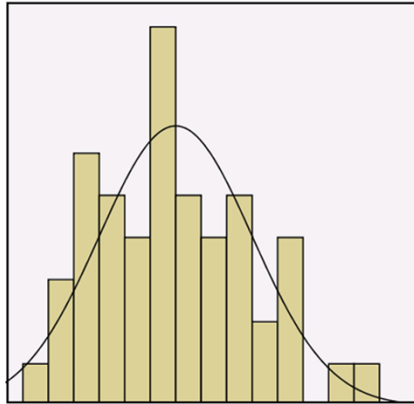
- In [probability theory](#), **Chebyshev's inequality** (also spelled as **Tchebysheff's inequality**) guarantees that in any [data sample](#) or [probability distribution](#), "nearly all" values are close to the [mean](#)
- the precise statement being that no more than $1/k^2$ of the distribution's values can be more than k [standard deviations](#) away from the mean
- $K=2$: no more than $1/4$ are 2+ standard deviations away
- As is often the case, if you cannot apply a "standard" method, there is an alternative!!

Note that the standard deviation rule only works for distributions which are approximately normal.

There is another rule, called Chebyshev's inequality, which handles data of absolutely ANY shape.

We won't cover the rule in this class but here we provide some information and links if you are interested in learning more.

Visually Assessing “Normality”



Getting back to assessing normality.

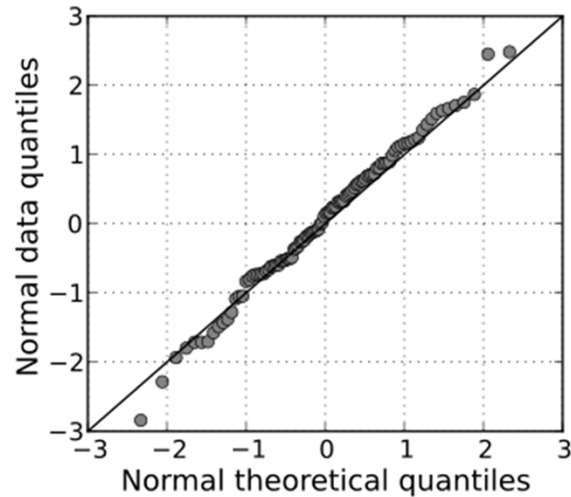
One method is to overlay a normal curve with the same mean and standard deviation as the sample. The closer the data in the histogram fits this curve, the closer the data are to representing a normal distribution. Again remember the issue of sampling variability!!

The graph on the left certainly could have come from a distribution with population outlined in the curve.

However, the distribution on the right does not seem to match the target curve at all.

You might argue that you don't need the curve to make these decisions, and if so, likely you are correct!

QQ-Plot or Probability Plot



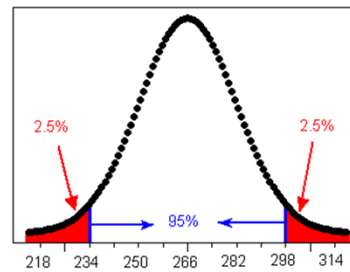
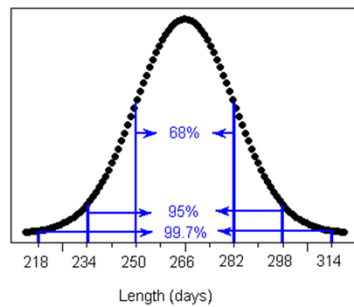
QQ-Plots or Normal probability plots are another plot which is commonly used to visually assess normality.

We won't discuss how they are created.

The idea is that if the data are exactly from a normal distribution, the points would fall exactly on the line.

Specific types of patterns can be recognized by looking at normal probability plots, however, we will generally decide the shape of distributions based upon histograms and boxplots and use QQ-plots only to determine how closely the data resembles a normal distribution.

Training for Later



Using the standard deviation rule to work back and forth between ranges and percentages is excellent practice for a skill we will need later in the semester regarding finding probabilities for normal distributions.

There are numerous examples and activities designed to help you with this skill now and later.

Z-Scores

- How many standard deviations does the raw score deviate from the mean?
- Standardized scores can be used to help identify potential outliers

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

Our final topic is a quick reminder about standardized scores or z-scores.

Recall that z-scores measure how many standard deviations a value is above or below the mean.


Based upon the standard deviation rule, for symmetric distributions, we could use this calculation to determine in which “area” you fall.

- For approximately normal distributions, z-scores greater than 2 or less than -2 are rare (will happen approximately 5% of the time).

Based upon Chebyshev’s rule


- For any distribution, z-scores greater than 4 or less than -4 are rare (will happen less than 6.25% of the time).

In this way, z-scores can also be used to identify potential outliers.



THE NORMAL SHAPE

One Quantitative Variable
Unit 1: Exploratory Data Analysis



The normal distribution is used often in statistics and we will see it again. The more you understand these idea now, the easier it will be to pick up the new ideas we introduce later in the course.

This concludes our discussion of exploratory data analysis for one quantitative variable.

We have learned about

Histograms, stemplots, boxplots, shapes of distributions, measures of center, spread, position, and outliers as well as some specific skills related to the normal distribution.