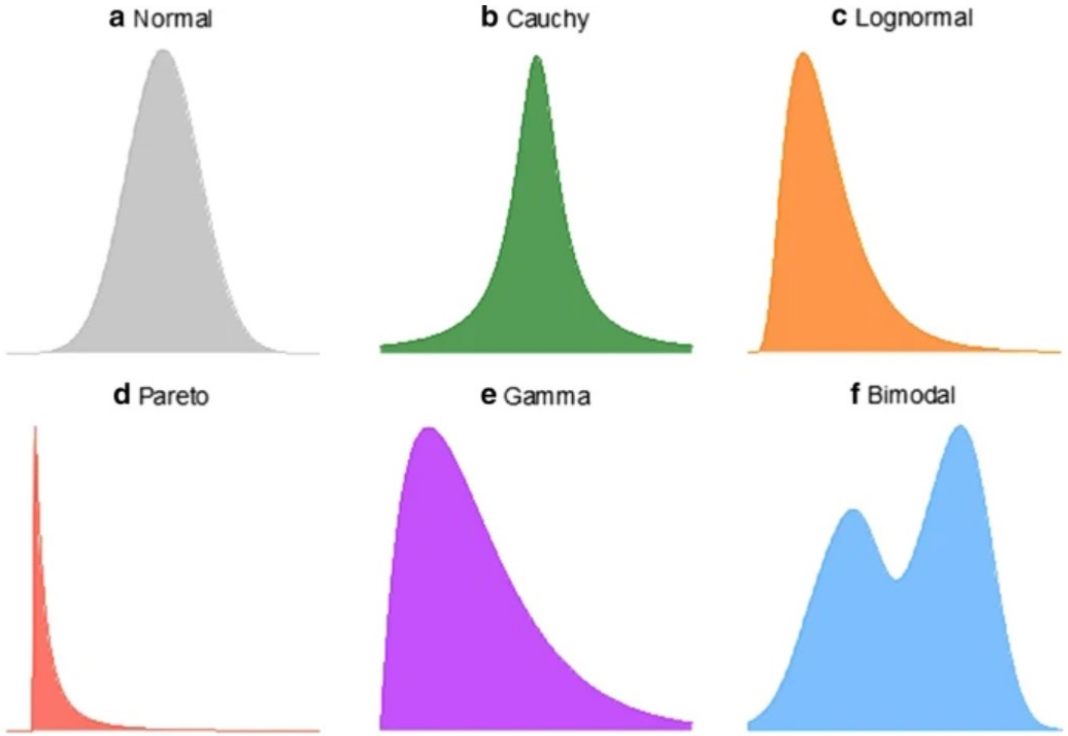
STATISTICAL DISTRIBUTIONS OF DATA



The notion of random variables and statistical distributions is:

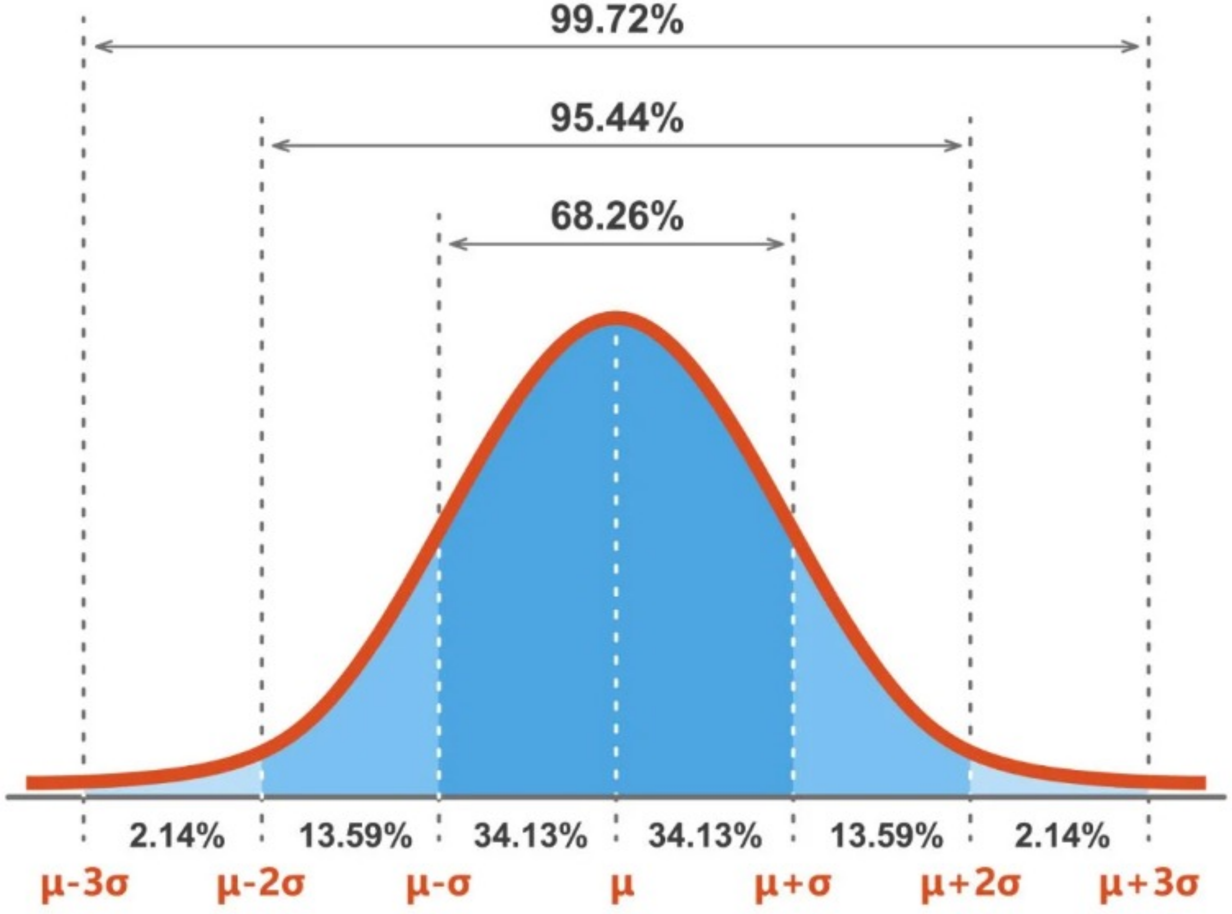
1. The value of a variable (X) is random. (ex. X is a person's weight. Its value could be 112, 175, or 215).
2. But the weight is not totally random. It has a smallest and largest possible value (ex. 2 & 600).
3. The X-values can be displayed on a horizontal number line.
4. The Y–value associated with the X-value is the probability of the X-value occurring.
5. The sum of all the probabilities (the area under the curve) is 1.0.

The shapes to the left are examples.

For the Gamma distribution, there is a higher probability the random variable X will have a low value.

ANALYST BOOT CAMP

NORMAL DISTRIBUTION – A SPECIAL CASE



Many things in the real world are **normally distributed** (ex. IQ, hand size).

The equation of the normal distribution is

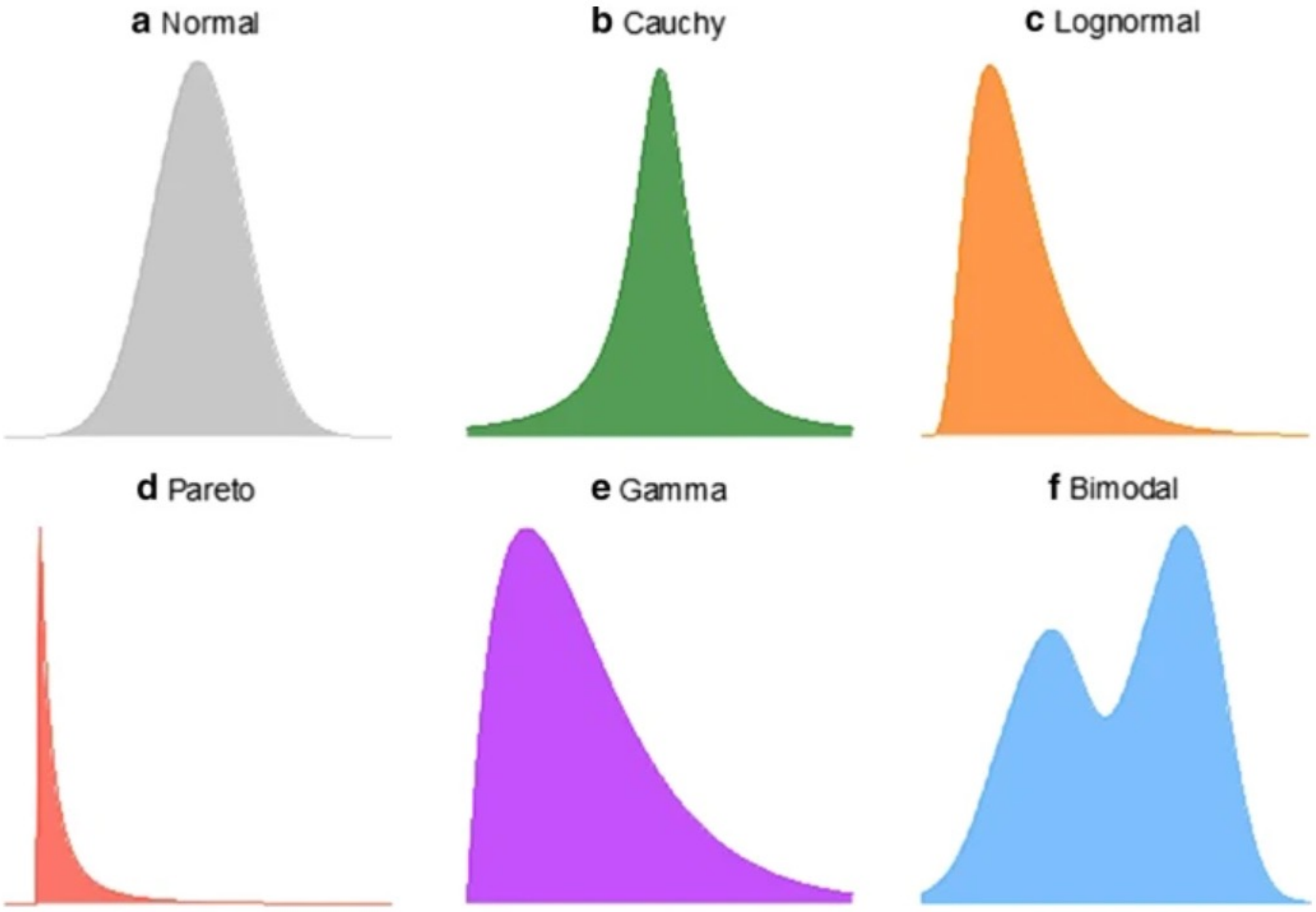
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The total area under the curve is 1.00. Furthermore:

- 68% of the area is within 1 SD of the mean,
- 95% within 2 SD,
- 99.7% within 3 SD.

So, if we have a normal distribution and we know the mean & SD, then we have an excellent idea of the possible values of the random variable.

STATISTICAL DISTRIBUTIONS OF DATA



We pay a lot of attention to Normal distributions, giving the impression it is the only one that ever occurs.

But many real-world data sets are not Normally distributed.

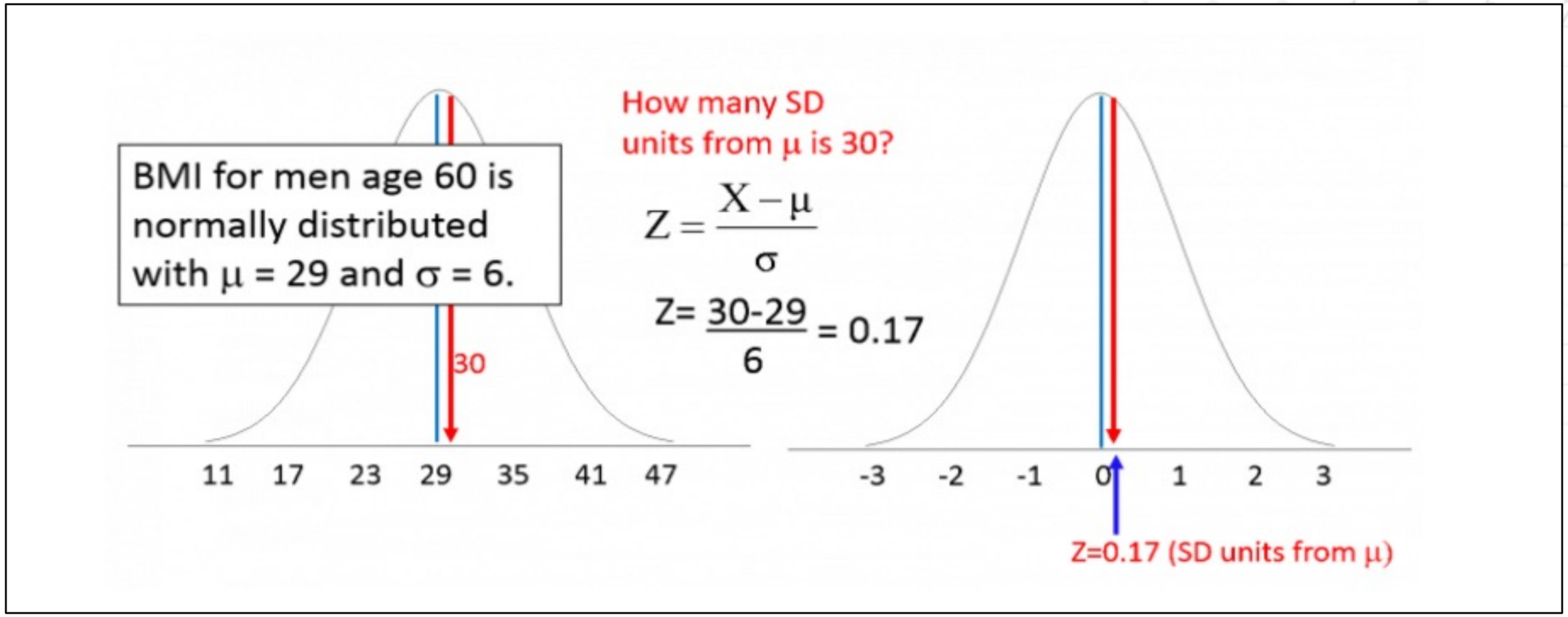
Can you give examples of data that has one of the distributions to the left?

NORMAL DISTRIBUTIONS

Khan Academy

Normal Distributions

Z-SCORES CONVERT ANY NORMAL DISTRIBUTION TO A STANDARD NORMAL DISTRIBUTION (MEAN=0 SD=1)



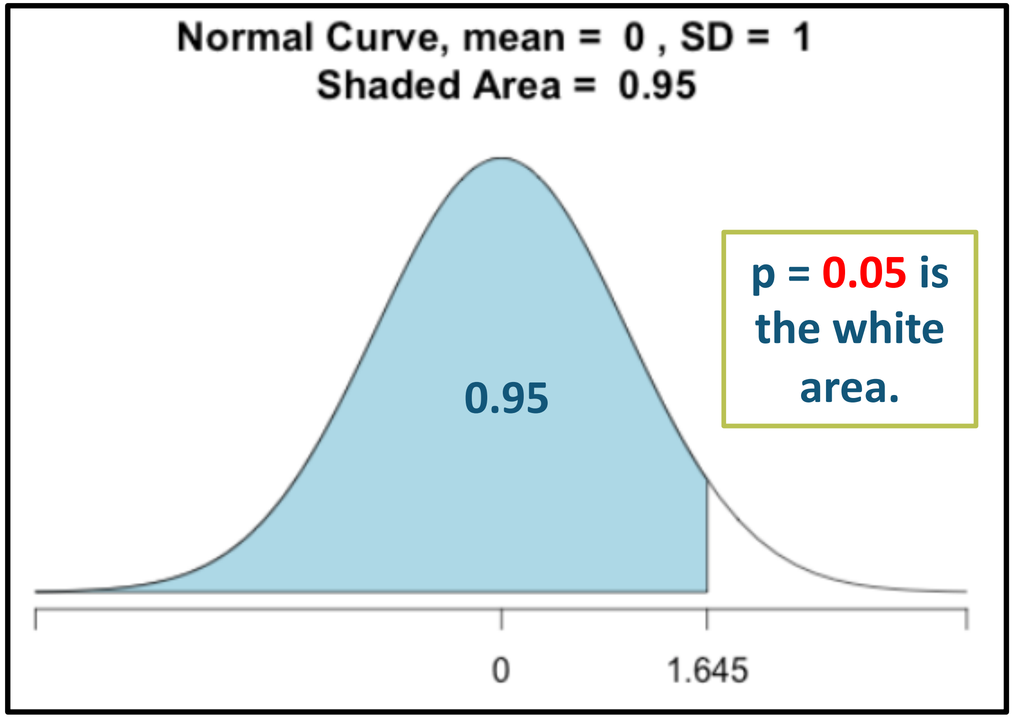
We re-frame (convert) the problem because there are published tables for the standard normal distribution.

The figure on the left is a normal distribution, mean=29, standard deviation=6. What is the probability a BMI is less than 30? This is the area to the left of 30 since the total area = 1.00

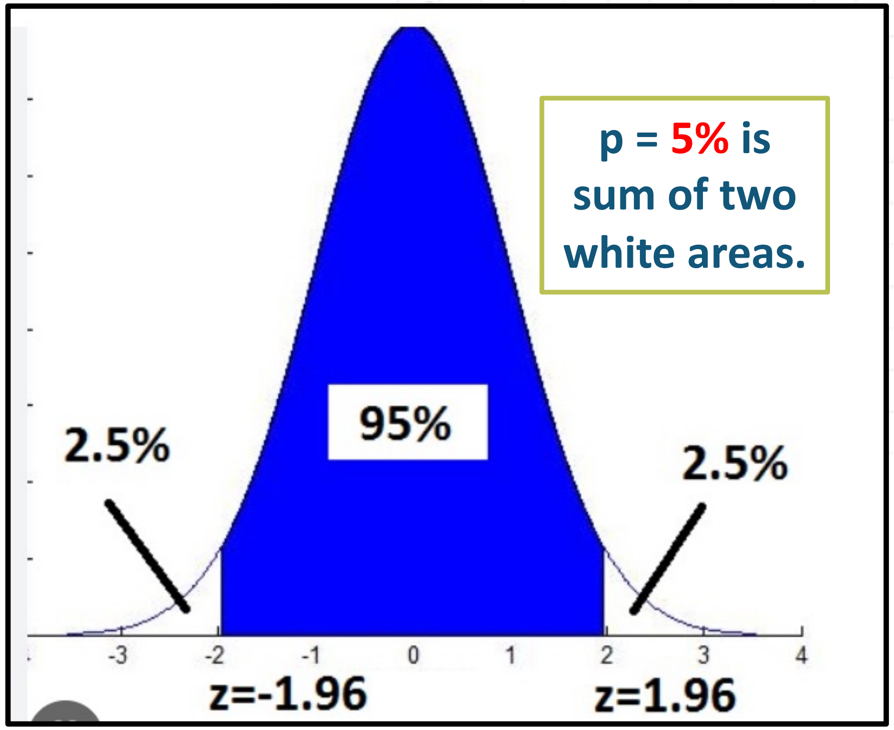
The figure on the right re-frames the problem by using a Z-score to transform to a "standard normal distribution" which has mean=0, standard deviation = 1. We want the area to the left of 0.17.

ANALYST BOOT CAMP

Z-SCORES FOR THE STANDARD NORMAL DISTRIBUTION WHEN P = 0.05



1.645 is the 1-tail Z-score for p=0.05



+1.96, -1.96 are the 2-tail Z-scores for p=5%

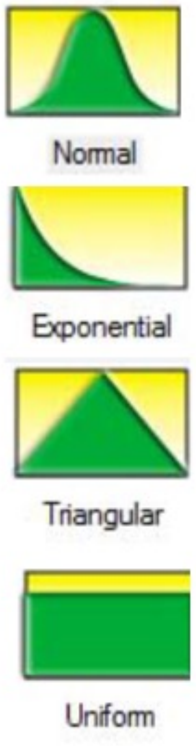
Note that 1.96 ~ 2.00

Z SCORES – KHAN

Khan Academy

Z Score Introductions

THE CENTRAL LIMIT THEOREM

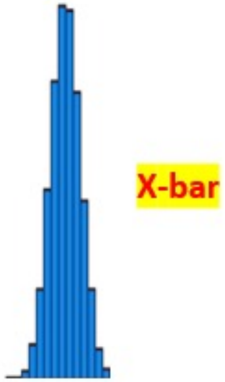


The **Central Limit Theorem (CLT)** allows us to make a powerful statement about **the distribution of the mean (x-bar) of any shaped probability distribution** (top graphs).

Take many samples of size n >30 from the original distribution. Plot the means of those samples on a bar chart. That bar chart will be approximately normally distributed. And it will be skinnier than the original distribution because

the SD(mean) = SD(original distribution) divided by the square root of n.

We will do a class exercise using many samples of size 30 from the uniform distribution and plotting the means of those samples.



ANALYST
BOOT CAMP

CENTRAL LIMIT THEOREM

Khan Academy

[Central Limit Theorem]()

ANALYST BOOT CAMP

A/B TESTING & CENTRAL LIMIT THEOREM

In A/B testing, we usually have a very large sample size. For example, in the earlier A/B presentation, the sample size was 64,000. The SQRT(64,000) is 253.

- The SD(mean) will be extremely small if the SD(sample) is divided by 253. The CI will then also be extremely small. So small that for all practical purposes you can treat the results as being “exact”.
- So, the values you got (below) can be treated as exact numbers when making the A vs. B comparison and you don’t need to compute confidence intervals around the averages.

  - 51.2% for the average of A
  - 51.7% for the average of B

EXCEL EXERCISES

Go to the Excel exercises in your workbook. The answers are to the far right on each tab.

Tabs

- NormDist
- Z-score
- CLT1
- CLT2