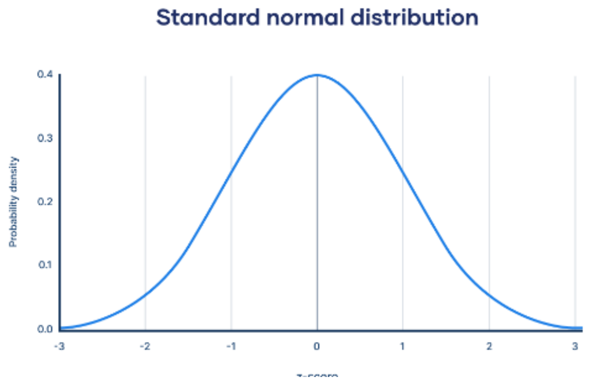
3/11/2023

Normal Distribution
Central Limit Theorem
Confidence Intervals?

Next Saturday we are on spring break. I forgot to put it in the calendar!

Normal distribution: bell curve



Standard score: transforms information from a general normal distribution (which a different mean and different standard deviation) into a value on the standard normal distribution.

Standard score = z score

Use x when in a general normal distribution (out in the world), use z when talking specifically about the standard normal distribution (mean =0, st.dev=1).

Using the standard score:

$$z = \frac{x - \mu}{\sigma}$$

If you want to describe a general normal distribution, you need to specify the mean and the standard deviation.

We can use the standard score to compare values in different distributions.

The SAT is designed to have a mean of 500 (on each part) and a standard deviation of around 100. The ACT has a mean of 21, and a standard deviation of around 5.

Suppose you take both tests (the math portion), the SAT you get 580 and on the ACT you get 28. Which is the better score?

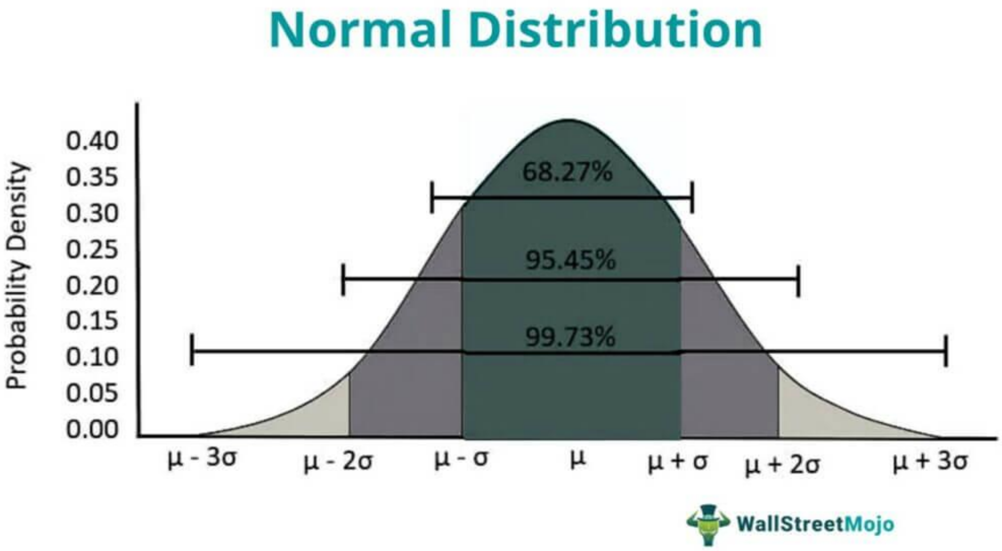
We can compare them with the standard score.

$$z_{SAT} = \frac{580 - 500}{100} = \frac{80}{100} = 0.8$$

$$z_{ACT} = \frac{28 - 21}{5} = \frac{7}{5} = 1.4$$

In this example, the ACT score is better (the z-score is higher), and so that’s the test score to submit.

Empirical Rule



Empirical Rule = 68-95-99.7 Rule
Depends on whole number multiples of the standard deviation.
The rule says:
Within 1 standard deviation of the mean (between z=-1 and z=1), 68% of the population is between these values.

Women have a mean height of 5’4’’ and a standard deviation of 3’. So 68% of the women have a height between 5’1’’ and 5’7’’.

Within 2 standard deviations of the mean (between z=-2 and z=2), 95% of the population is between these values.

95% of women are between 4’10’’ (6’’ shorter than the mean) and 5’10’’ (6’’ taller than the mean).

Within 3 standard deviations of the mean (between z=-3, and z=3), 99.7% of the population is between these values.

99.7% of women are between 4’7’’ (9’’ less than the mean) and 6’1’’ (9’’ taller than the mean).

Suppose the mean test score on an exam is 78% with a standard deviation of 5%. What percentage of the test-takers scored between 73% and 83%? -- 68%

What percentage of the test-takers were between 68% and 88%? –95%

What percentage of the test-takers are between 73% and 88%? (68+95)/2 = 81.5%

What percentage of the test-takers are between 73% and 88%? – we know that 68% are between 73% and 83%... what percent are between 83% and 88%?

95-68 = 27… there is an additional 27% between 1 standard deviation and 2 standard deviations according to the empirical rule. I only want half of this. 27/2 = 13.5%

68%+13.5% = 81.5%

To calculate probabilities that are not whole multiples of the standard deviation, we need technology.

See Excel. (you can do this in the calculator—the calculator function works a little differently than Excel).

Basic rules for NORM.DIST:
Always set cumulative to TRUE
Calculate the probability of being less than the specified value.
To get more than, use the complement rule
To do between: calculate the probability at the higher value, then subtract the probability at the lower value.

If you are calculating probabilities for the standard normal distribution, you can use NORM.S.DIST, specify x (z), and TRUE. If you use NORM.DIST, specify mean 0 and standard deviation 1.

Inverse calculations:
If you have a probability and you want the value that that corresponds to.
For example, What score on the SAT represents the top 10% of the test-takers?

Top 10% means that 90% of the population did less well. The inverse function wants the percent below, never the percent above.

The inverse function takes the percentage and gives you the score.

In Excel the function NORM.INV. This is always just cumulative, so we don’t need to specify that.

Central Limit Theorem
At this point, we begin doing inferential statistics…

As we take samples, the mean of the sample is going to tend toward the mean of the population, the distance from the true population mean will tend to get smaller as the sample size grows. And the distribution of means from repeated sampling will have the shape of the normal distribution (regardless of whether the original population is normal or not).

We are not actually going to take repeated samples, but because of this property, we can take one sample, and use that to estimate how far we are likely to be from the true population value.

How good that estimate is is going to depend on the sample size.

Distributions of sample means that we are thinking about are called sampling distributions.

And the standard error is the standard deviation of the sampling distribution: is a measure of the variability in the mean of our sample.

Standard error formula (for means):

$$SE = \frac{\sigma}{\sqrt{n}}$$

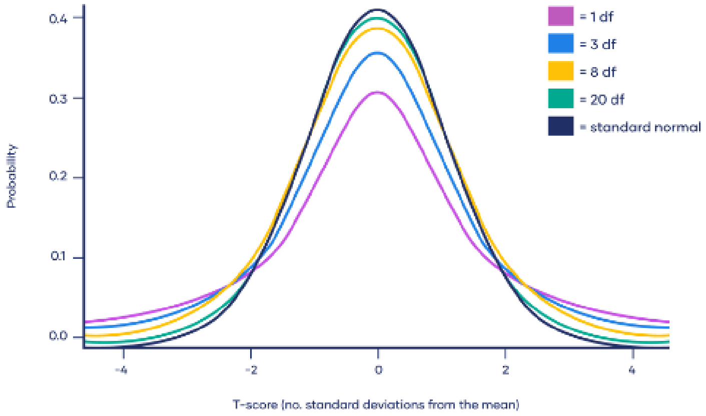
Sigma is the standard deviation of a single observation in the population.
n is the sample size.

For a proportion problem: the mean is the sample proportion.
The standard error:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

For means problems, generally we want a sample of 30 or 40 or bigger. Proportion problems require that $np(1-p) > 10$

Student T distribution.
Is very similar to the standard normal distribution, but there is more probability in the tails of the T-distribution. And there is a different T-distribution for each sample size. Around 30 or 40 sample size, the two distributions are almost equal.



The T-Distribution: T.DIST, T.INV