

Random variables and probability distributions from khan academic

Definition of random variable: random process, then quantify the outcomes to numbers, then do math about the outcomes

Random variables:

Discrete – distinct and separate value, as long as you can count them

e.g. x = 0 for tails and 1 for heads

Continuous – any value in interval, and it is infinite, you can not list them

A probability distribution for random variable

Y: Probability

X: Outcomes or Value for X



Law of large numbers

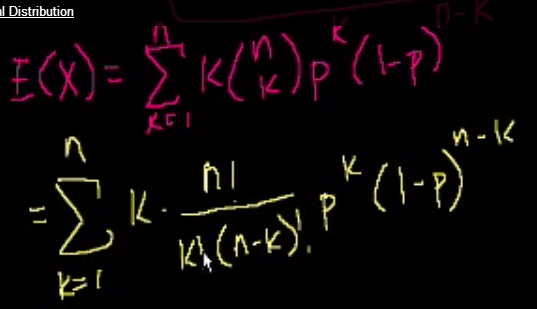
The mean of the sample is approaching the expected value (the true mean) as long as the number of sample is large.

Expected value of binomial distribution (only)

E(X) = np

Where X is the outcome, n is the # of trial, and p is the probability.

Proven process:



Where P(x=k) =

Poisson process

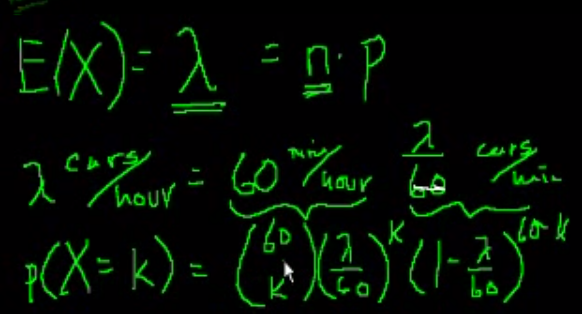
How many cars passed in an hour? In no way this # will affect the # of cars passed in next hour.

X = # of cars passed in an hour.

E(x) = λ

For binomial distribution: E(x) = np

In an approximation, E(x) = λ = np = 60\*(λ/60)



Where is the success in a sum, n is the success in a small interval, p is the probability of success in a small interval

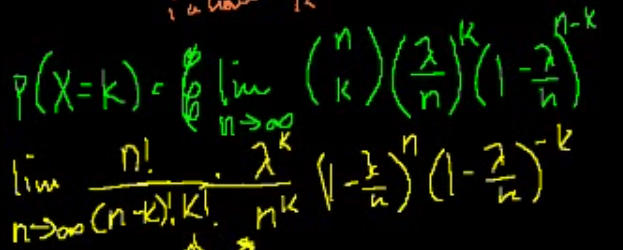
The issue is: in a minute, a car passed is counted as 1 successful trial; if 5 cars passed in a minute, it is still counted as 1 successful trial.

To solve this issue,we divide the interval in a second (60 min = 3600 s)

Then more division

To prove this,

Because λ = np, p = λ/n,



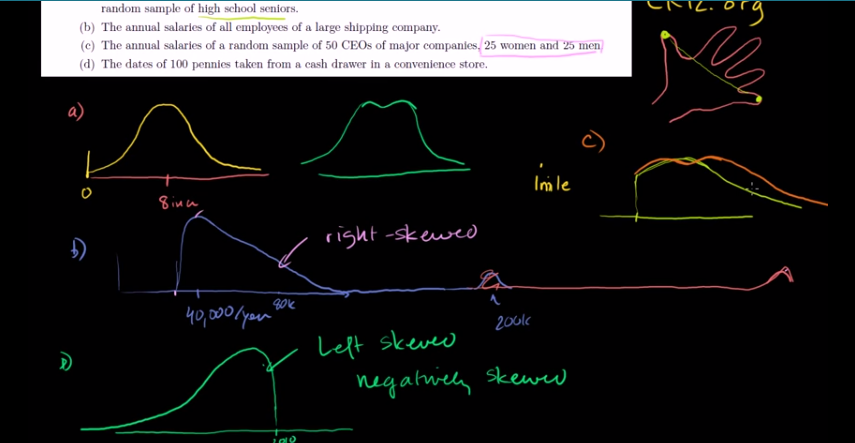
Correlation vs. Causality

Normal distribution

The probability is given by the area under the curve

Cumulative distribution curve

Q1: Qualitive sense of normal distribution: which of the following data sets is most likely to be normally distributed? (ideas: symmetry)



Q2: the probability for the case x < value, using the normal distribution.

(be careful that it requires the consideration of the two tails)

(be careful about the difference between the confidence interval and z-table)

Empirical rule:

68-95-99.7

z-score can be applied to any standard distribution

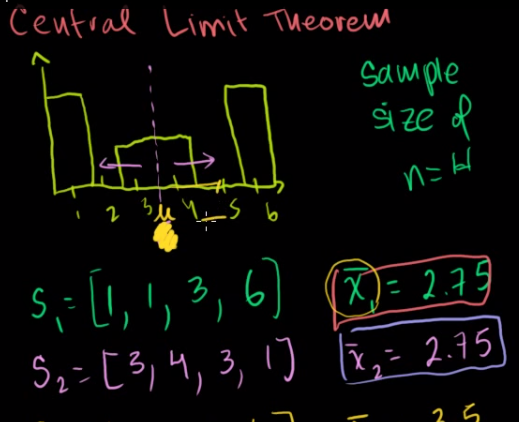
Central limit theorem: sampling distribution

We can start with any crazy distribution with mean and std deviation

As the # of sampling increased, the frequency of the mean of each sample trends to the normal distribution.

As the sample size increases, the frequency of the mean of each sample trends to the normal distribution, and the std deviation becomes smaller.

Actually, not only the mean of each sample, the sum or else also works.

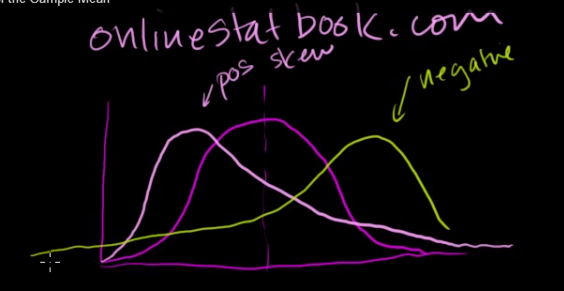


Skew and kutosis help to tell how normal the distribution is

Skew:

Positive skew: long tail going to the right direction

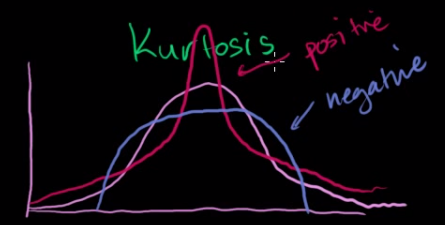
Negative skew: long tail going to the left direction



Kutosis:

Positive kutosis:

Negative kutosis:

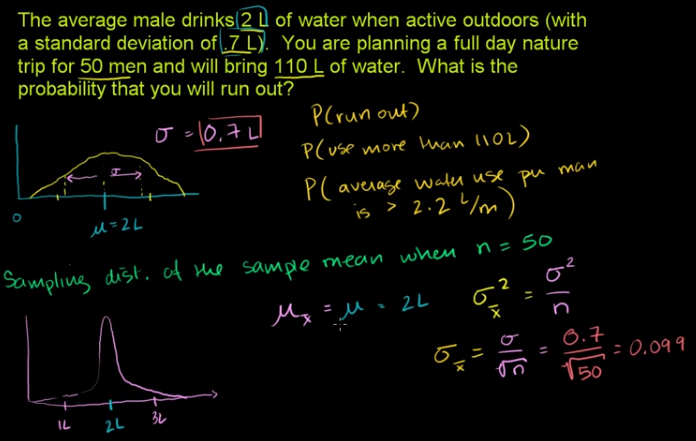


μ, σ^2 are the original mean and variance of the distribution of the probability, respectively;

x bar is the sample mean; μ\_x\_bar (= μ) is the mean of the distribution of the mean (x bar), the variance of the distribution of the mean (x bar) is σ^2.

When people are talking about the sample size, they mean the number of outcomes for each trial in samplings.

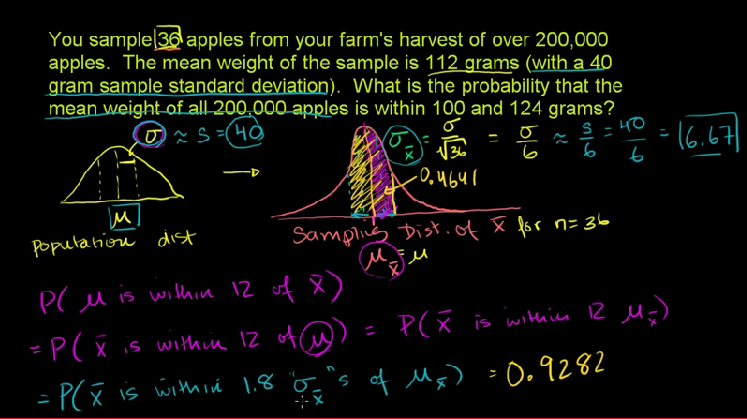
Problem 1:



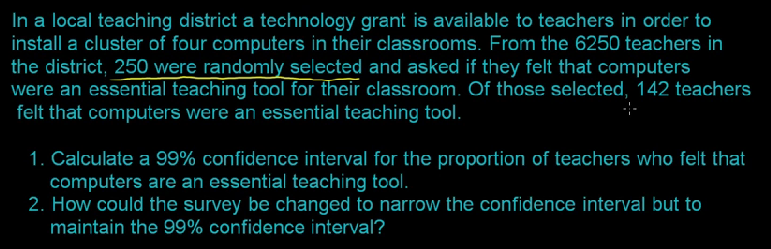
Problem 2:

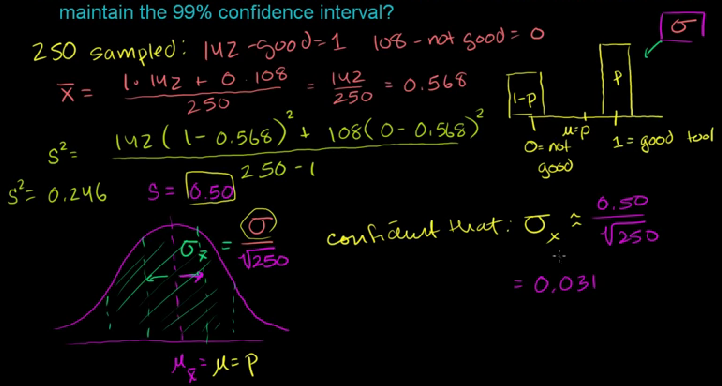
My question is why the 40 gram sample standard deviation is for the original distribution instead of the 36 sampling distribution?

Answer use the sample standard deviation to estimate the standard deviation of original distribution. But why original distribution?



Problem #3:



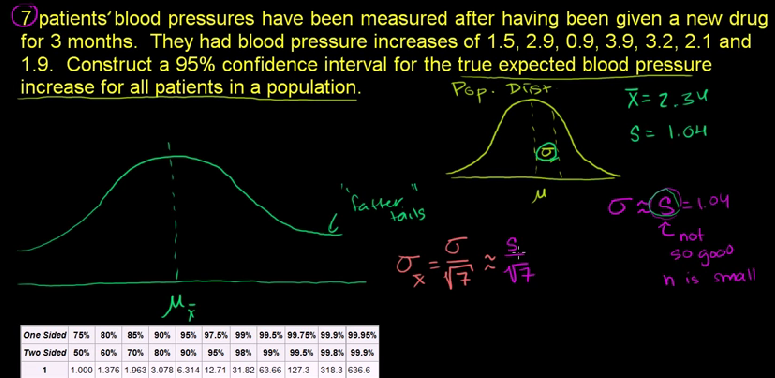


(Again, be careful about the difference between the confidence interval and z-table)

(hint: #1 use the sample mean and std deviation to estimate the mean and std deviation of the original distribution. If sample size > 30, estimate is good, we use the normal distribution. If sample size < 30, it is a bad estimate, we use the t distribution, which has a fatter tail.

#2 the questions is about using the distribution of sample mean.)

Problem #4:



Though we do not know the mean and std deviation of the sampling distribution, but we can estimate it.

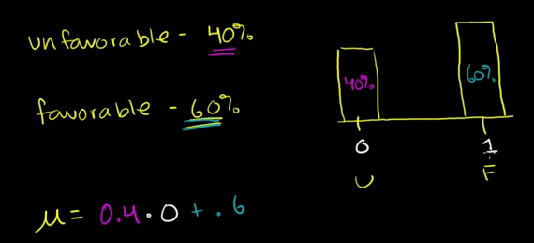
Find an interval such that reasonably confident that there is a 95% chance that the true μ is in the interval.

(note: when using the t table, one sided or two sided, n-1 in the 1st column)

Mean and variance of Bernoulli distribution: B(1,p)

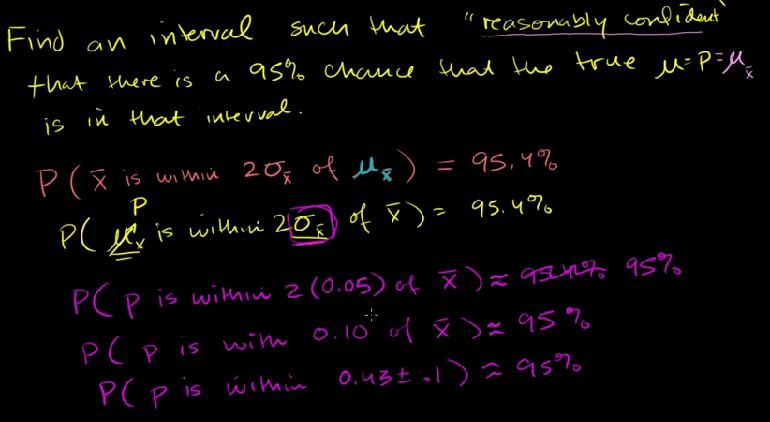
Q: What is the difference between Bernoulli distribution and binominal distribution?

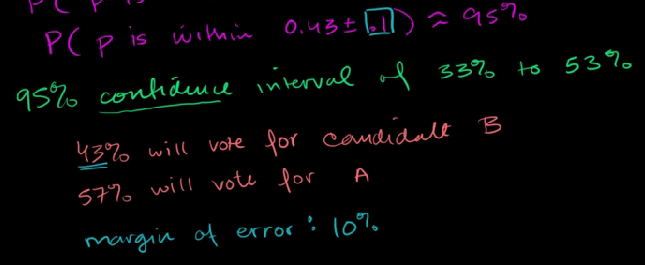
A: it is the simplest situation of the binominal distribution.



Mean:μ = (1-p)\*0 + p\*1 = p

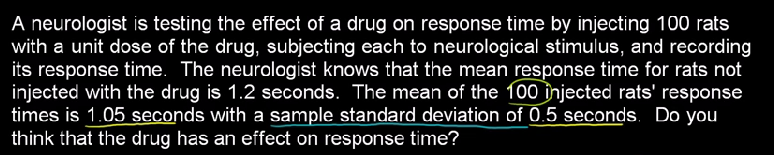
Variance: σ2 = p(1-p)



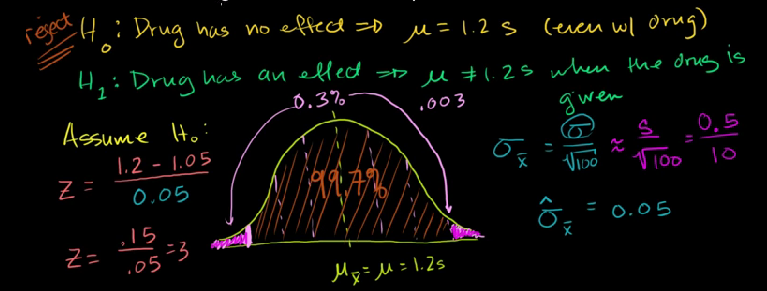


Margin of errors: std deviation

Hypothesis testing and p-values

Problem #5:

The key is to find out if the sample mean (1.05 s) is within the confidence interval of the sample mean distribution. According to the central limit theory, the sample mean distribution is ~ Normal (μ,σ). These two parameters can be obtained from the questions. μ is equal to the mean of the original distribution. σ is related to the std deviation of the original distribution, which can be estimated by the std deviation in the sampling.



The null hypothesis: it refers to a general statement or default position that there is no relationship between two measured phenomena, or that a potential medical treatment has no effect. (The hypothesis in the question is not right) Rejecting H0, thus concluding that there are grounds for believing that there is a relationship between two phenomena.

The alternative hypothesis:

The probability of getting the value given the non hypothesis is called p-value.

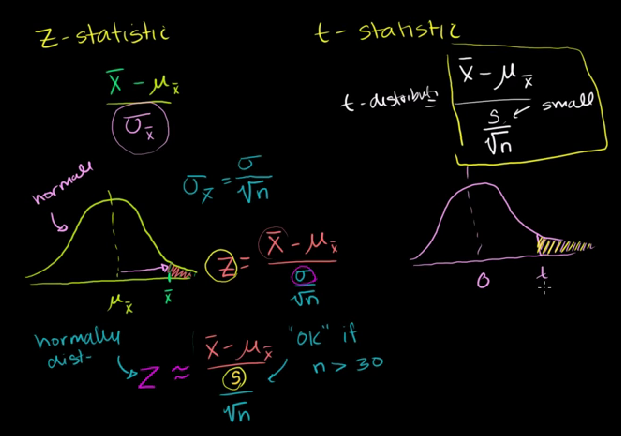
In the question above, p-value is equal to 0.003, which is very small, so we reject the hypothesis Ho. Generally, if p-value is < 5%, I will reject the chance.

Type 1 error: rejecting Ho even though it is true

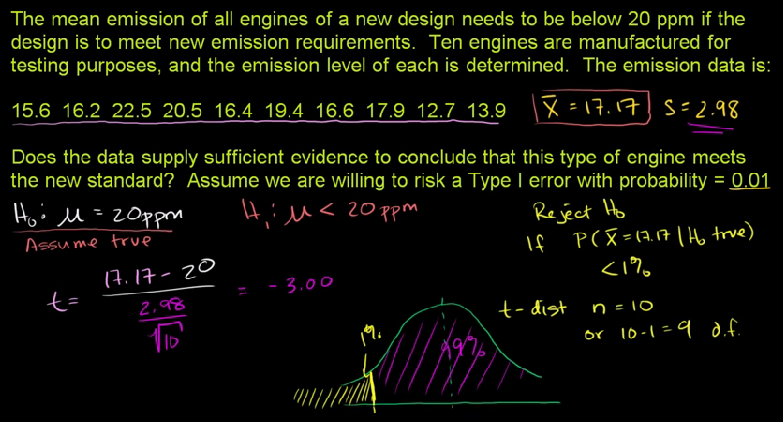
We always assume that the non-hypothesis is true

Z-statistics vs. T-statistics:

If the sample size is > 30, use the z-table; if the sample size is < 30, use the t-table



Problem #6:



Another followed-up question:

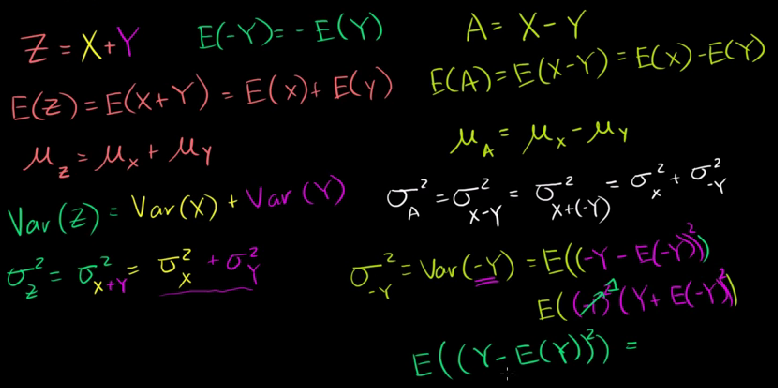
What is the 95% confidence interval?

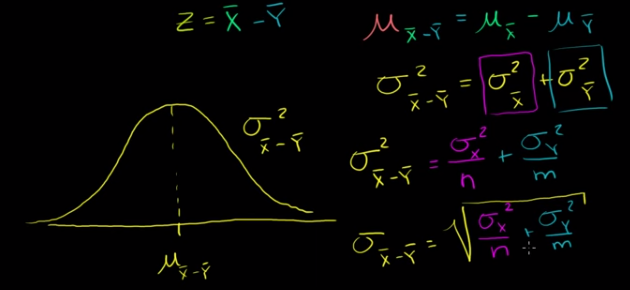
So from these two questions, it is clear that the xx% confidence interval and hypothesis (H0, H1) are quite similar. They are using the same equations, and the same process, but asking for different variables by providing different information (xx% and μ, respectively)

Significance level

Hypothesis testing with two samples

Variance of differences of random variables





Chi-square distribution