Statistics:

Basic intro: visualize, summarize, test, find trends,

Challenges: optional

Programming: optional

Data is the basic, from which people make decisions.

Statistics is universal, useful, and fun

In statistics, the way we find out is data. In the previous database,

Linear line

Do you trust these numbers?

Bar chart (2D)

Relationship between bar chart and scatterplot

Histogram (special case of barchart, 2D, it groups numbers into ranges, and you decide what ranges to use) frequency vs data bin

Pie chart: relative data, it is for comparing

Relationship between histogram and pie chart

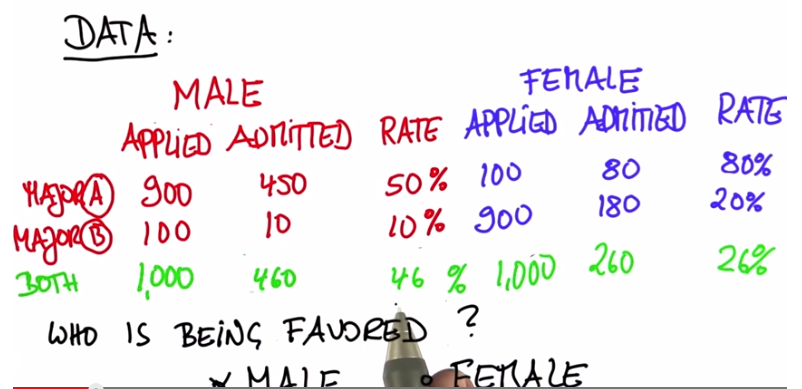
Visualisation: picture data in graphs, including scatter plot, bar chart, histogram, pie chart.

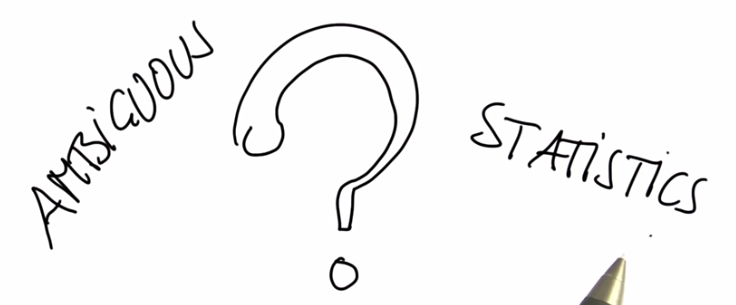
Make you own graphs by programming

By just looking at the data, you can understand a lot.

Simpson’s paradox

Admissions case study



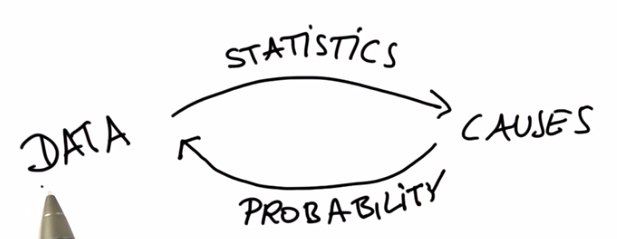


A famous saying goes “I never believe in statistics. I did not doctor myself”

Statistics is deep and often manipulated.

Raw data ---> Decisions/Conclusions

Probability



Probability of event: p

Probability of opposite event: 1-p

Probability of composit event: PPPP (independence)

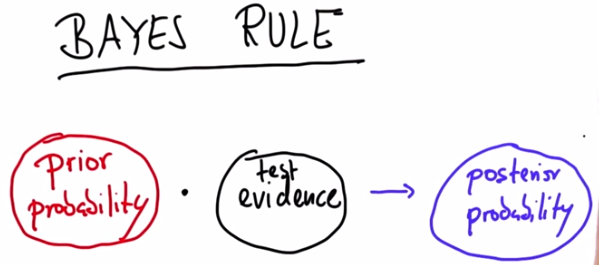
Conditional probability

Dependent things

P(A|B) means the probability of A given B



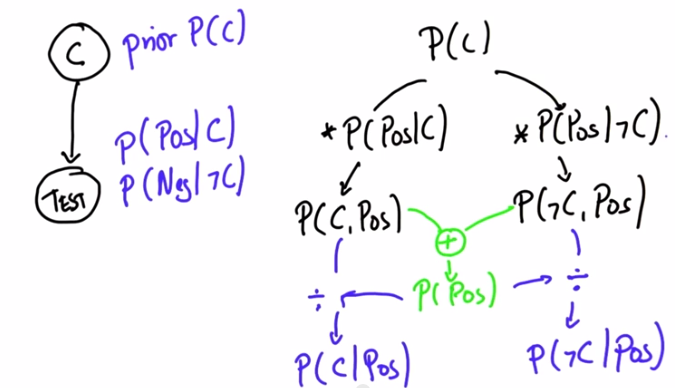
Bayes rule

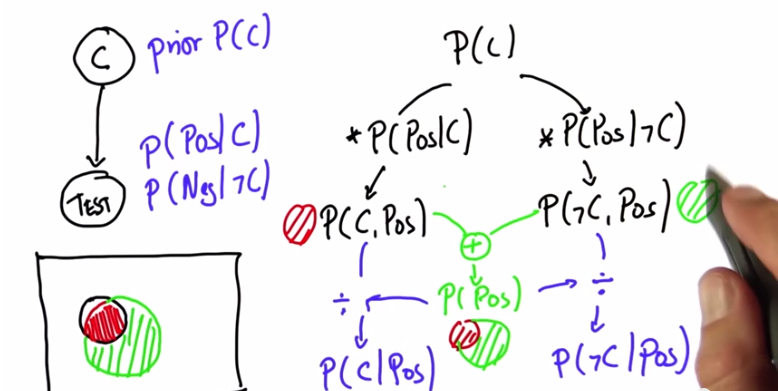


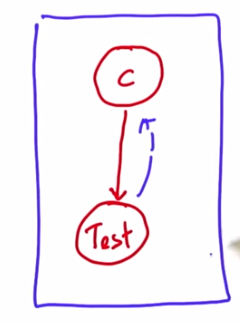


P(A|B) = =

P(B) is the nomaliser







Continuous probability distributions

In continuous distribution, every outcome has probability 0

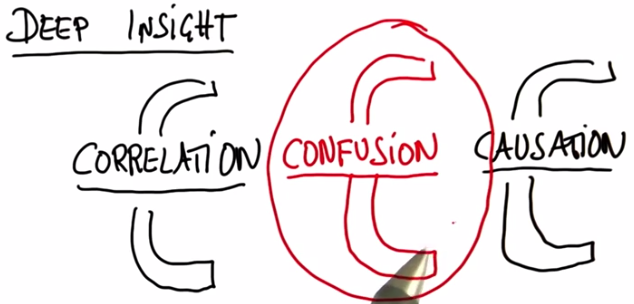
Density: f(x)

Density function is continuous? No.

Density function is positive? No. it is non negative

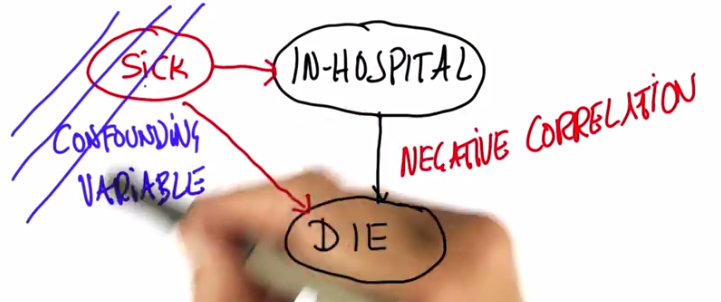
Density can be larger than 1, but probability is always smaller than 1

Gaussian?





In statistics, a confounding variable is an extraneous variable in a statistical model that correlates with both the dependent variable and independent variable.



Is in-hospital correlated with die?

Does in-hospital cause fire? Maybe reverse causation

Correlation ---- causation

Every range of proportions when flipping coins?

Estimators

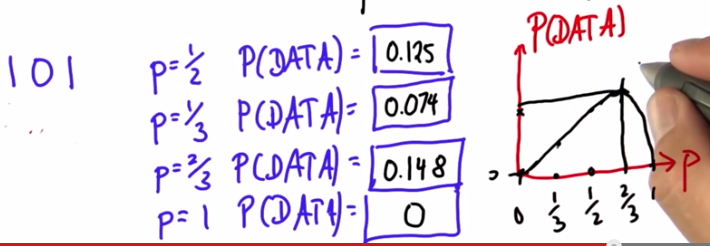
Maximum likelihood estimator

Laplacian estimator

To fake or not to fake?

1. MLE





2. The Laplace Estimator or Laplace Smoothing, also called [Additive Smoothing](http://en.wikipedia.org/wiki/Additive_smoothing) is defined as follows:

*P*=(*C*+*k)/(N*+*k*⋅*D)*

Where:

*P* is the probability of a particular outcome  
*C* is the count for the particular outcome  
*N* is the number experiments (flips of a coin, throws of a dice, etc)  
*D* is the size of domain of outcomes (2 for a coin, 6 for a 6-sided die, etc)  
*k* is the smoothing factor or weight.

The instructor used the particular value *k*=1, which results in:

*P*=(*C*+1)/(*N*+*D)*

The objective of this algorithm is to "average" the distribution obtained empirically (with experiments) with an uniform distribution (i.e., a distribution where all outcomes are equally probable). Note that this was exactly what he did. He faked that he started with exactly one count for each possible outcome (an uniform distribution) and then added the counts from the empirical data to calculate the frequencies.

You are "averaging" your empirical data with an uniform distribution, since you have no prior knowledge about what is the distribution of your data. Therefore, it is a fair assumption that all the outcomes are equally probable. But note that the effect of the "fake data" gets less and less important as you gather more and more experimental data.

You do not necessarily need to equal the importance of the "fake data" to the importance of the empirical data, as the instructor did. You can make the "fake data" more important or less important by changing the value of *k*. Note that greater the value of *k*, greater the weight of the "fake data" compared to the empirical data. For instance, making *k*=2 would be equivalent to add "fake data" twice before adding the empirical data to calculate the frequencies. Finding an ideal value of *k* is something that you do usually by trial and error.

When there is not much data, fake it.

To fake or not to fake?

To fake when you are not sure about the results, it will give you better answer.

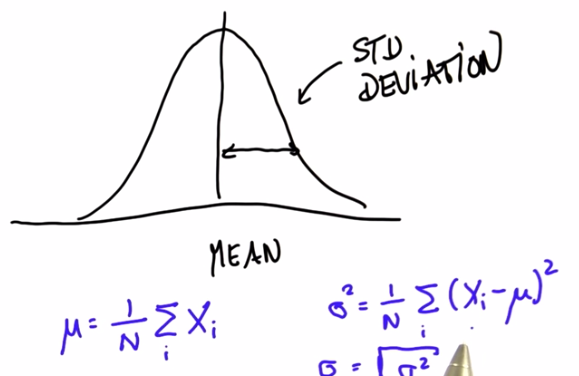


Median: sort, then pick the one in the middle

Variance, standard deviation

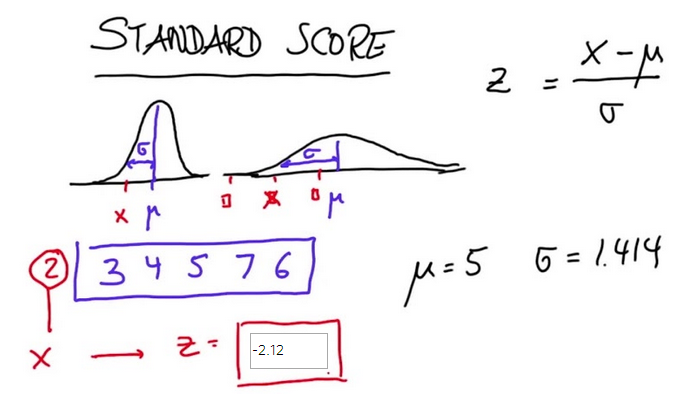
Variance measures how close these data are around mean

Std = square root of Variance



Note that the variance of the distribution describing the sum of two draws from a normal distribution is different from the variance of the distribution of taking a single draw from the distribution, then doubling it.

When you add two Gaussian valuables, mean adds up, and variance adds up. (prove it needs multipages)



Outlier

Quartiles

Interquartile range (IQR) is between Lower quartile and upper quartile. The data outside are generally ignored.

The whole idea of the quartiles is to break your data into four chunks of equal size. These "chunks" are demarcated by three data points: the lower quartile, the median and the upper quartile:

|--------------| L |--------------|M|--------------|U|--------------|

Q D Q

Let's compute the size of our data. The data is composed of the following data points:

* The lower quartile (LQ), the median (MD) and the upper quartile (UQ). That's three data points.
* The chunks of data demarcated by LQ, MD and UQ. All you know about them is that they need to be the same size. Let's say that the size of each chunk is *N*. That's 4⋅*N* (4 "chunks", each one of size *N*)
* <------ N -----><- 1 -><------ N ----><- 1 -><------ N ----><- 1 -><------ N ---->
* |---------------| L |--------------| M |---------------| U |-------------|
* Q D Q

Therefore the size of our data is (4⋅*N*)+3 (We are just adding the data points).

The smallest possible value of each chunk is 1. If we increase the size of our chunks in increments of one we will get the following table:

"CHUNK" SIZE DATA SIZE

1 (4 x 1) + 3 = 7

2 (4 x 2) + 3 = 11

3 (4 x 3) + 3 = 15

4 (4 x 4) + 3 = 19

5 ... ...

And so on...

k-th percentile is k% through the data

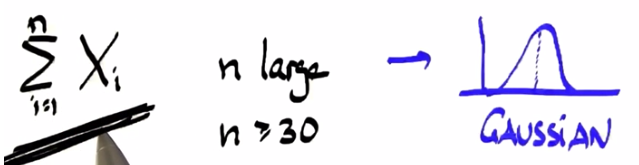
Binomial distribution:

Central Limit Theorem

Many coin flips

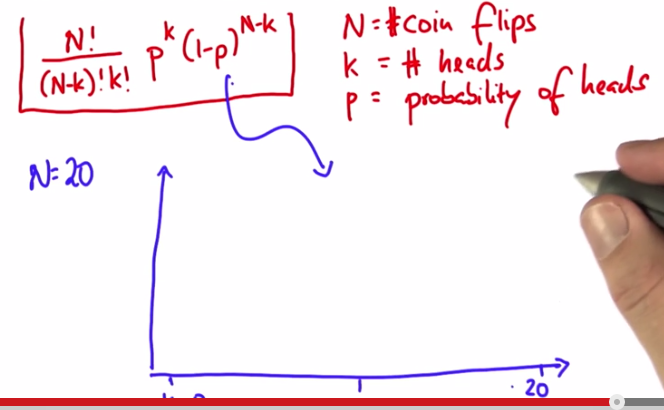
Distribution of the sum of many things

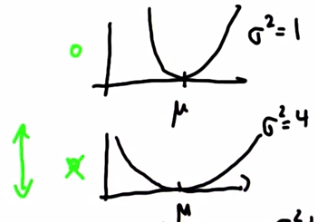
Pascal triangle: The value is sum of the value above and left

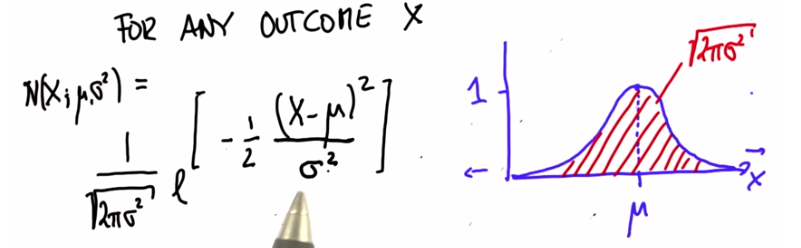


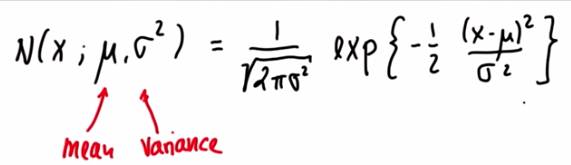
The sum of the outcomes

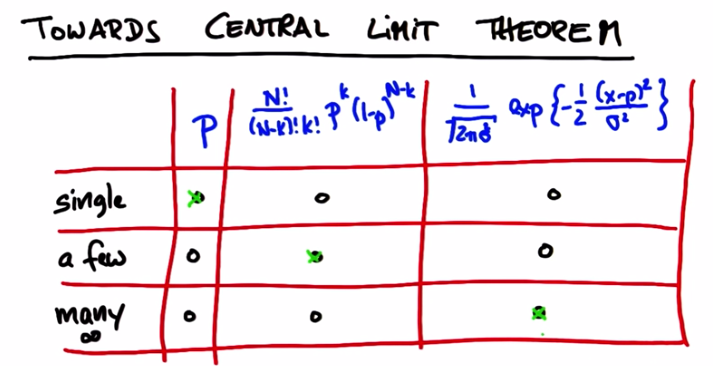
Coin flip 🡪 Binominal distribution 🡪 central limit theorem 🡪 The normal distribution

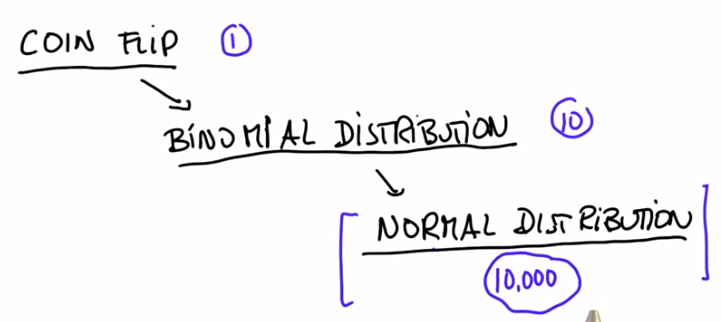




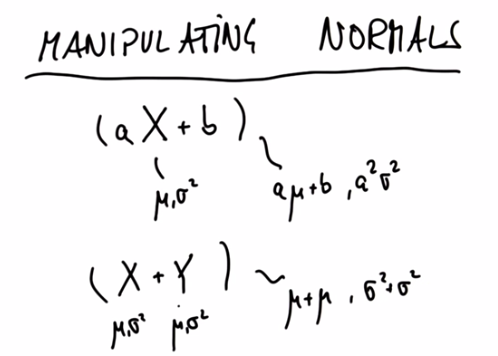




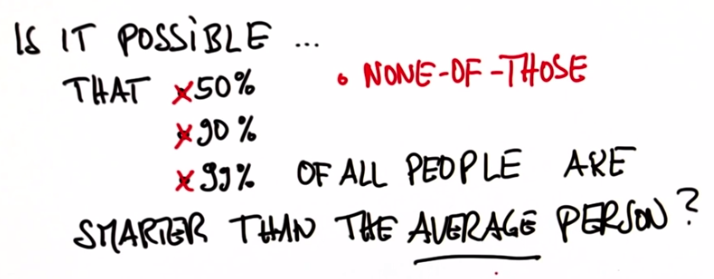


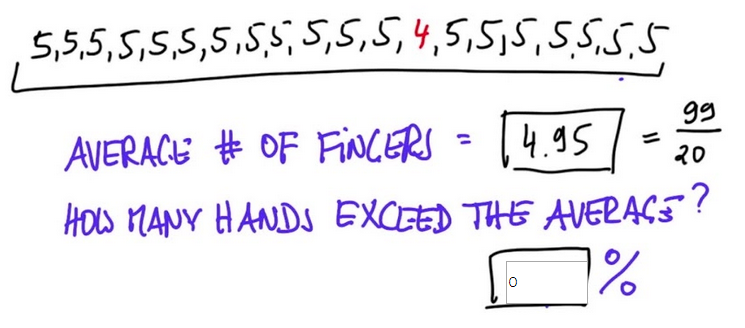


Manipulate normal distribution



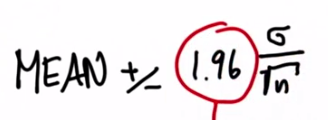
Most better than average





This data doesn't have 4n+3 points as in the examples in the lecture so the quartile does not lie within the data. Try to find the points that are approximately 25%,50%, and 75% of the way through the data.

Guess sebastian’s weight

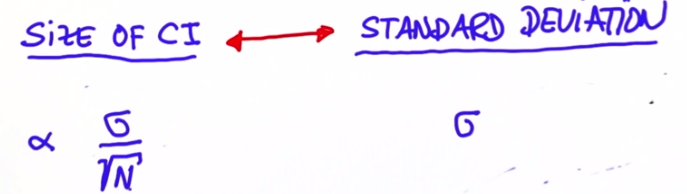
Confidence intervals : 

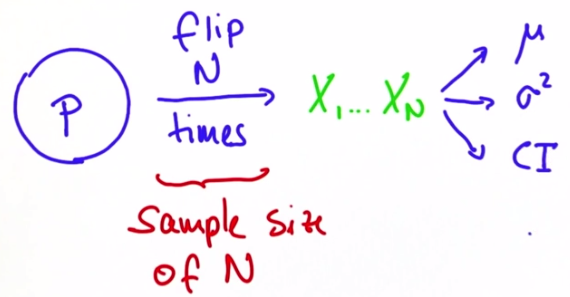
Margin of error is an amount of random sampling error in a survey’s results. The larger the margin of error, the less confidence one should have that the poll’s reported results are close to the true figures.

Margin of error 🡪 confidence interval

More trust 🡪 smaller confidence interval (CI)

Suppose we increase the sample size N: size of CI shrinks, standard deviation stays the same.





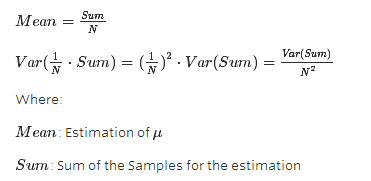
Note that this isn't the true mean, since you're applying the definition of mean over a sample of size *N*, not over the whole space of events. This is, therefore, an estimation of *μ*

The whole idea of confidence interval is to determine how good is this estimation. If I do this estimation (sampling) over and over and average the results, greater the number of samplings, better will be the estimation. If I do this an infinite number of times (or until the space of events is exhausted), this estimation would converge to the true mean. So the mean of the new random variable *M* is the true mean *μ*

As any random variable, *M* has a mean and a variance, and these are the "Mean" (mean of estimation of *μ* or mean of *M*) and "Variance of the Mean" (variance of the estimation of *μ*) that Prof. Thrun is referring to. He's analyzing the distribution of *M*, which represents an estimation or sampling (i.e., you get *N* samples and average these samples).

He wants to know how close this estimation is to the true *μ*. What a 95% confidence interval means? If I do samplings over and over, 95% of my samplings should result in an estimation of *μ* inside this confidence interval.

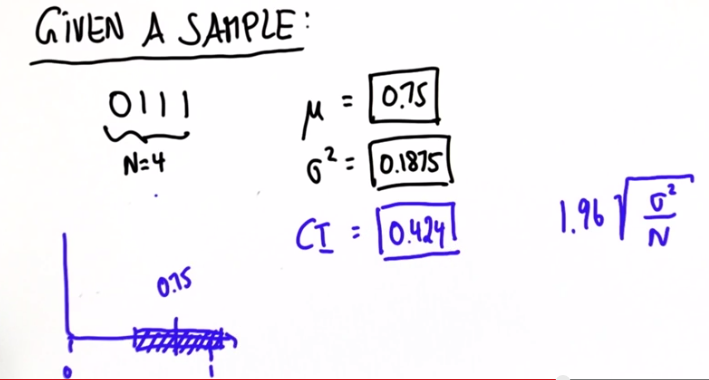
So, "Variance of the Mean" is what he was really going after. "Variance of the Sum" is actually an intermediate step when calculating the "Variance of the Mean". That's because of the properties of mean and variance from unit 21 ("Manipulating Normals"):



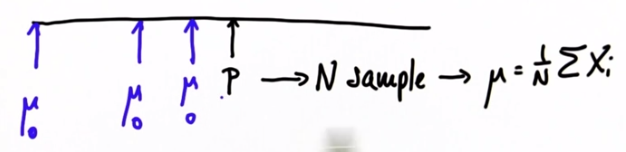
Mean = p

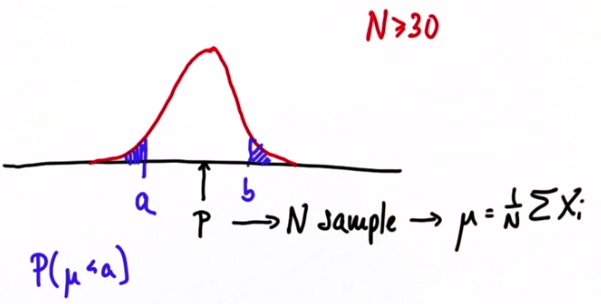
Var (p) = p(1-p)

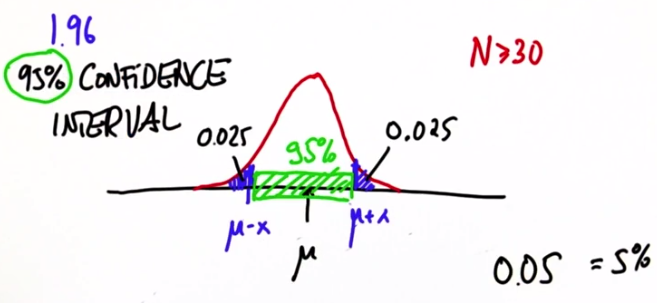
CI = 1.96\*

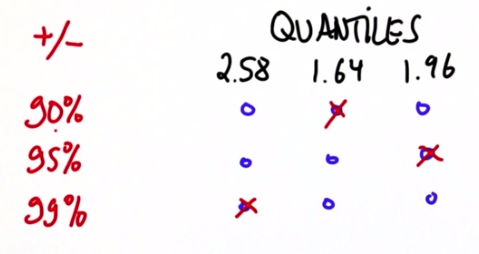


Magic number 1.96



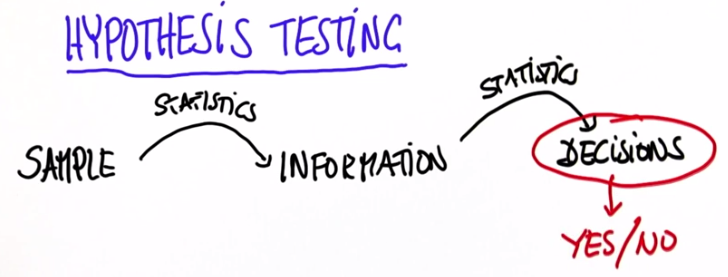


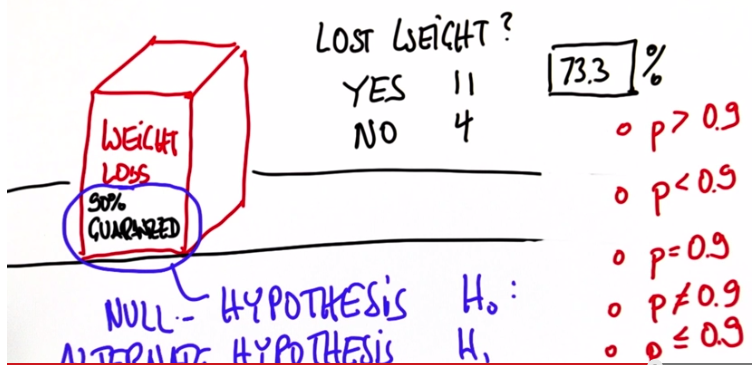




(for large samples N >= 30)

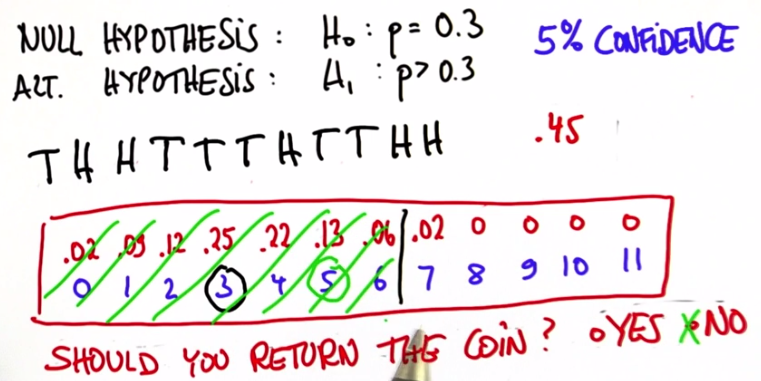
Hypothesis testing





Ho = 0.9;

H1 < 0.9;



the critical region has to agree with the alternate hypothesis (disagrees with the null hypothesis)

In the hypothesis test:

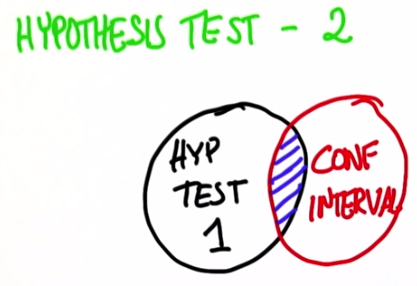
Define H0 and H1, (one-sided test or two-sided test?)

Calculate the binomial probabilities,

Calculated the accumulated sum of probabilities until it reaches 5%

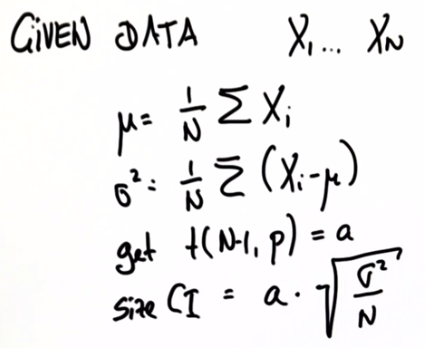
Find the critical region

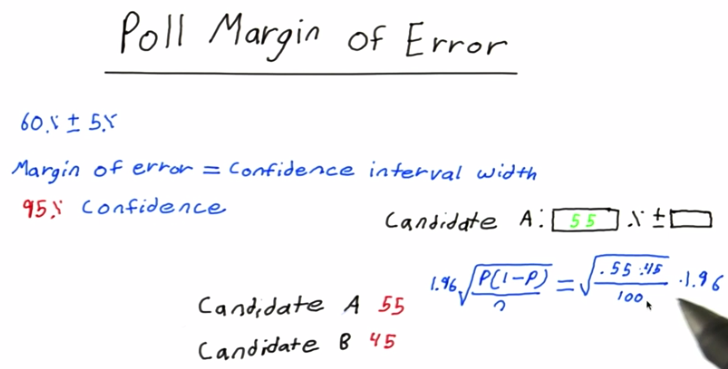
At last, compared it with the tested results



Idea: use the confidence interval, check the number is within or outside of the confidence interval.







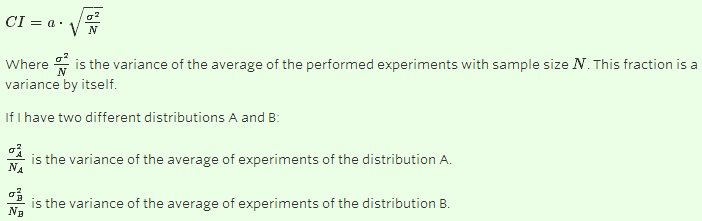
When to use p to calculate variance????

From Unit 21 ("Manipulating Normals"), you know that you can combine the variables wB and wA and create a new variable (wB-wA).

**Additional Comments: Combining two Distributions with Different Sample Sizes**

When combining distributions, the size of the samples for the experiments A and B not necessarily need to be of the same size.

That's the formula for the confidence interval:

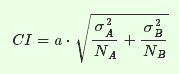


To obtain the variance over the combined distribution, I still have to add the variances (although these are variances for the averages of the experiments, not over the original random variable), so the variance of the

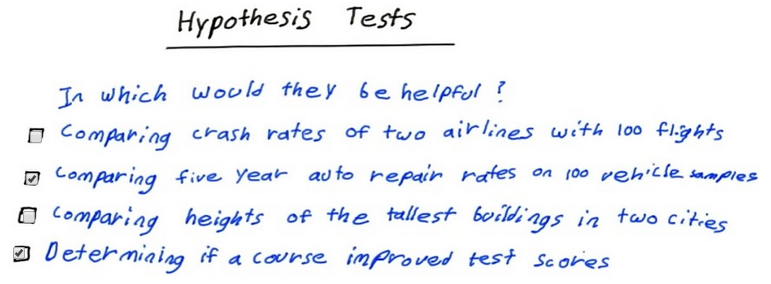


And now I understand why Adam broke the fraction in two, even though they had the same denominator. So the experiments do not need to have the same sample size after all. Although if both distributions have the same size, you can derive the variances the way I did on my solution.

So a general formula for computing the confidence interval for two distributions A and B would be:



One way is to calculate the Confidence Interval, the other way is to calculate the magic number (a) using the “must” CI, then compare the magic number with the table. If "a" is greater than xx, you know that you're outside the confidence interval (i.e., in the blue areas), and therefore your null hypothesis does not hold.



Do not know the conditions where hypothesis tests are used?

You have a *μB*−*A*=0.01% and you would like to know if it could be equal to zero ( *μA*=*μB*). To include zero inside your confidence interval, you would need a confidence interval of 0.01%, therefore:

*σ*'s are not given for the distributions, but you know that the expected variance can be calculated as follows:

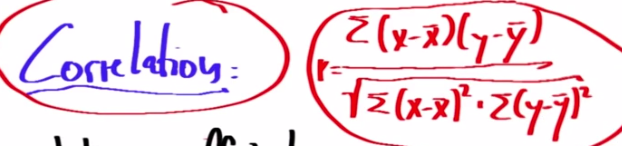


Correlation coefficient (r):

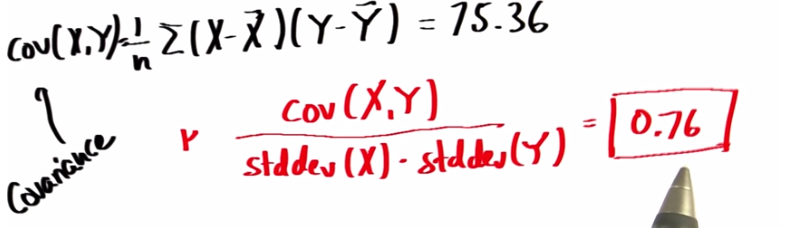
Value between -1 and 1;

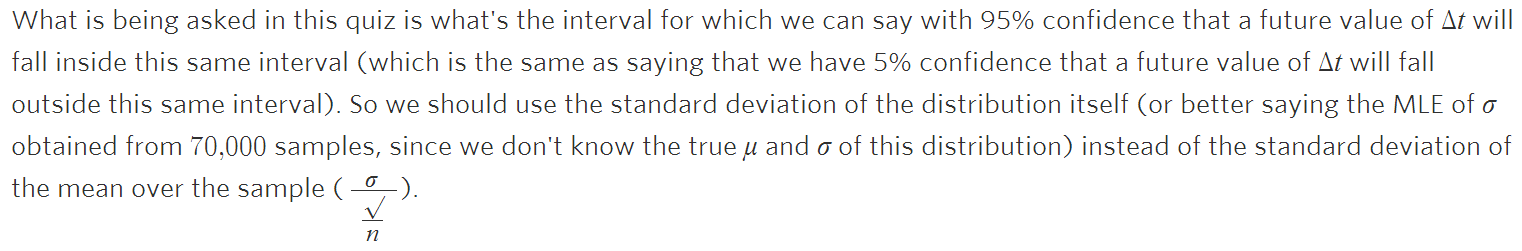
Tell us how related two variables are

Both 1 and -1 stand for perfectly linear data

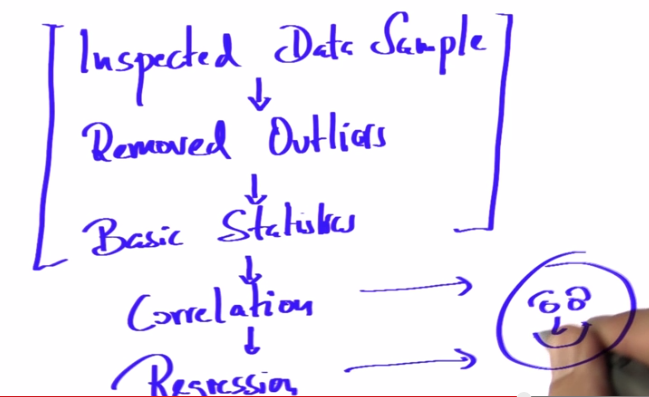


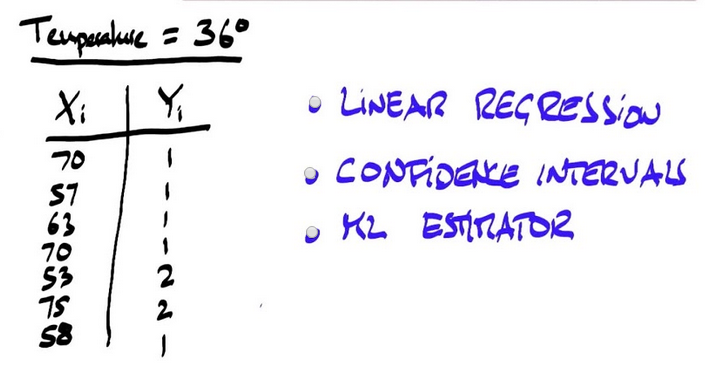
Outliers can be removed by quartiles. However, if data sets are paired, then the quartile method is not adopted.



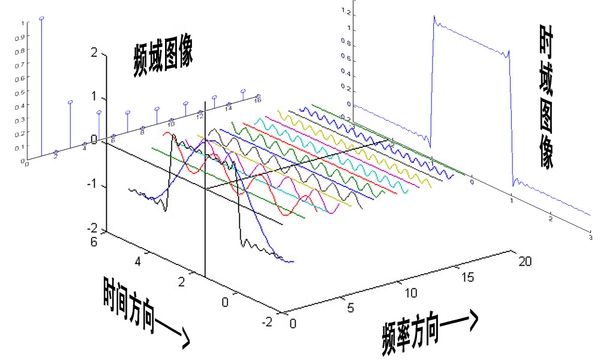


The difference between the standard deviation of the distribution itself and the standard deviation of the mean over the sample () ?????





Fourie Transform



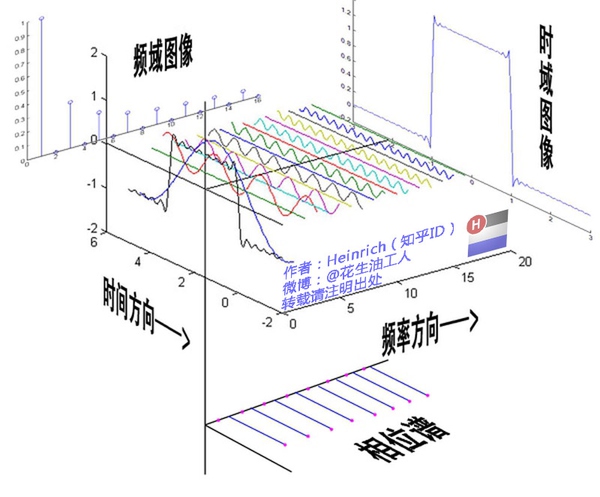
频道频道，就是频率的通道，不同的频道就是将不同的频率作为一个通道来进行信息传输

我把sin（3x）+sin（5x）的曲线给你，但是前提是你不知道这个曲线的方程式，现在需要你把sin（5x）给我从图里拿出去，看看剩下的是什么。这基本是不可能做到的。但是在频域呢？则简单的很，无非就是几条竖线而已。所以很多在时域看似不可能做到的数学操作，在频域相反很容易。这就是需要傅里叶变换的地方。尤其是从某条曲线中去除一些特定的频率成分，这在工程上称为滤波，是信号处理最重要的概念之一，只有在频域才能轻松的做到。

傅里叶变换则可以让微分和积分在频域中变为乘法和除法，大学数学瞬间变小学算术有没有。

通过时域到频域的变换，我们得到了一个从侧面看的频谱，但是这个频谱并没有包含时域中全部的信息。因为频谱只代表每一个对应的正弦波的振幅是多少，而没有提到相位。基础的正弦波A.sin(wt+θ)中，振幅，频率，相位缺一不可，不同相位决定了波的位置，所以对于频域分析，仅仅有频谱（振幅谱）是不够的，我们还需要一个相位谱。那么这个相位谱在哪呢？

这里需要纠正一个概念：时间差并不是相位差。如果将全部周期看作2Pi或者360度的话，相位差则是时间差在一个周期中所占的比例。我们将时间差除周期再乘2Pi，就得到了相位差。



以上我们讲解的只是傅里叶级数的三角函数形式。接下去才是最究极的傅里叶级数——指数形式傅里叶级数。但是为了能更好的理解指数形式的傅里叶级数，我们还需要一个工具来帮忙——欧拉公式。

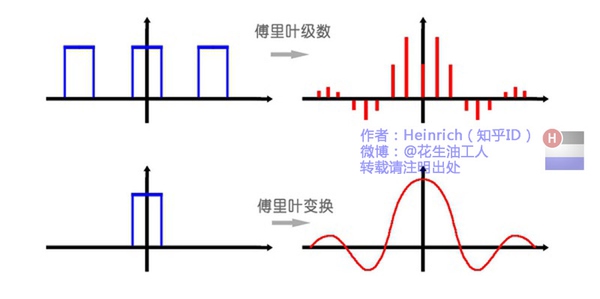
傅里叶变换（Fourier Transform，FT）-> 离散傅里叶变换（Discrete Fourier Transform， DFT）-> 快速傅里叶变换（Fast Fourier Transform）  
FT是理论基础，以FT为理论基础，可以完成从频率估计到求解微分方程各式各样的问题；  
DFT是指信号被采样之后你会得到**离散**（如你需要处理的音频信号被采样）而非**连续**的信号，这个时候就需要DFT来告诉你怎样处理并告知你一些离散情况下的特殊问题；  
FFT是一种计算DFT的算法，计算复杂度很低也就是执行起来很快的意思。

举个例子吧：有人通过在小黑屋按钢琴的一个键不松会产生一个单音信号给你传递情报，  
y(t)=\sin(2\pi ft+\theta)  
信号的频率f取决于他所按的键。你看不见他，却希望获知信号的频率。怎么办？  
1.FT的理论就会告诉你可以通过傅里叶变化获知这个频率。  
但是这个信号飘荡在空中，你需要先通过采样得到一个离散信号  
y[i]=\sin(2\pi \frac{f}{f_{s}}i+\theta) \  \ \ \ \ (i=1,2,...N)  
(f_{s}是采样频率，香农和奈奎斯特告诉我们，需要f_{s}>2f)。  
2.得到离散信号后如何计算f，DFT就会告诉你怎么办;  
3.你嫌DFT太慢了怎么办，FFT就粉墨登场了。

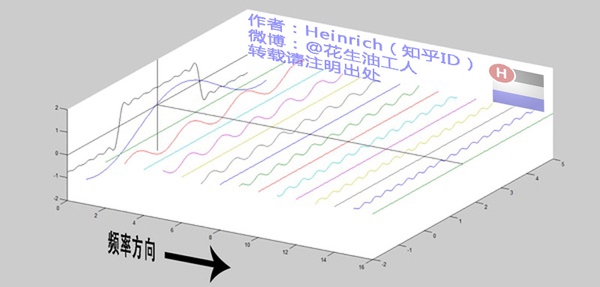
傅里叶级数的本质是将一个周期的信号分解成无限多分开的（离散的）正弦波，但是宇宙似乎并不是周期的。

是否有一种数学工具将连续非周期信号变换为周期离散信号呢？抱歉，真没有。比如傅里叶级数，在时域是一个周期且连续的函数，而在频域是一个非周期离散的函数。

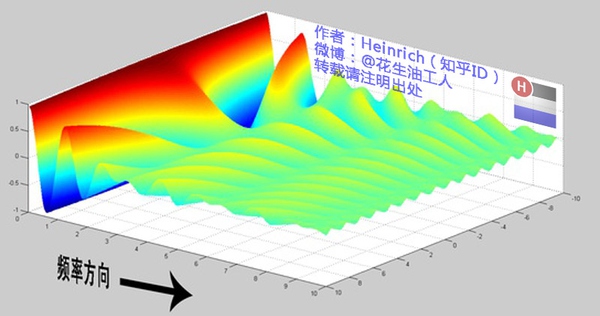
傅里叶变换，是将一个时域非周期的连续信号，转换为一个在频域非周期的连续信号。傅里叶变换实际上是对一个周期无限大的函数进行傅里叶变换。



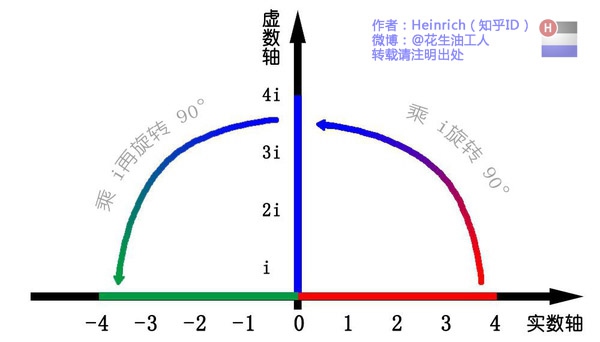
离散谱: 原来离散谱的叠加，变成了连续谱的累积。所以在计算上也从求和符号变成了积分符号。



连续谱:



乘虚数i的一个功能——旋转



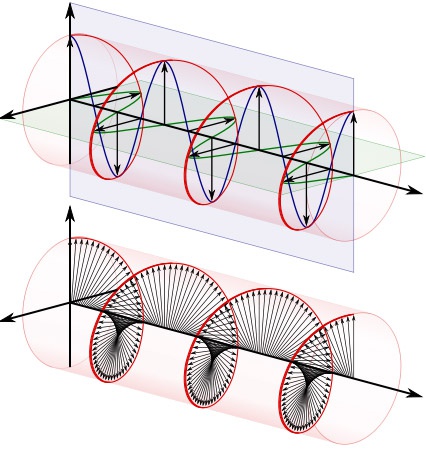
欧拉公式, 这个公式关键的作用，是将正弦波统一成了简单的指数形式





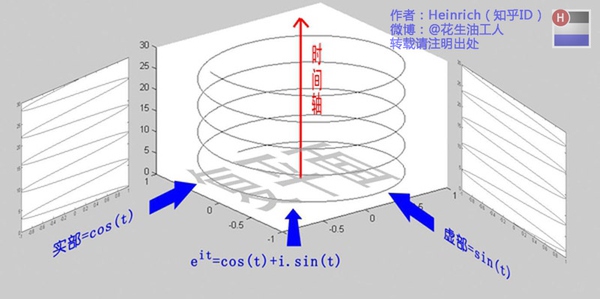
用复数来描述电场与磁场简直完美到爆棚！我们即可以让电场强度与复数磁场强度相加而不损失各自的信息，又满足了电场与磁场90度垂直的要求。另外，一旦我们需要让任何一个场旋转90度，只要乘一个“i”就可以了.

利用欧拉公式，我们可以将任何一个正弦波看作其在实轴上的投影。假如两个不同的正弦波，可以用数学表达为：  
http://p2.zhimg.com/e1/25/e12595ebb6cb8f04e8b066488843ea64_m.jpg  
好了，现在如果我想用第一个正弦波利用线性变换为第二个，我们就只需要将A乘对应的系数使其放大至B（本例为乘2），然后将θ1加上一定的角度使其变为θ2（本例为加30度），然后将得到的第二个虚数重新投影回实轴，就完成了在实数中完全无法做到的变换。  
  
这种利用复指数来计算正弦波的方法也对电磁波极其适用，因为电磁波都是正弦波，当我们需要一个电磁波在时间上延迟/提前，或是在空间上前移/后移，只需要乘一个复指数就可以完成对相位的调整了。



我们可以用两种方法来理解正弦波：

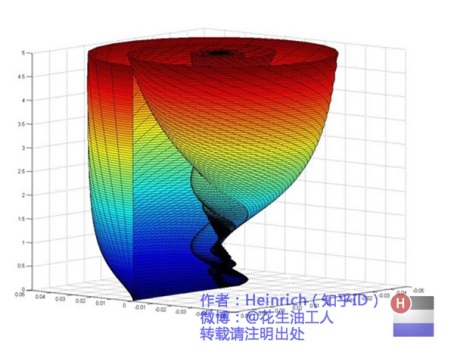
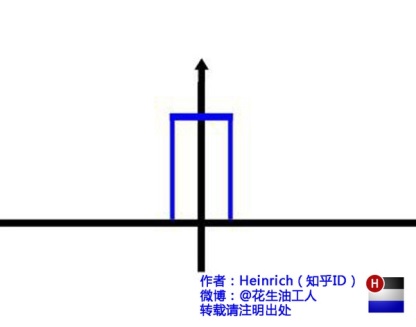
第一种就是螺旋线在实轴的投影:

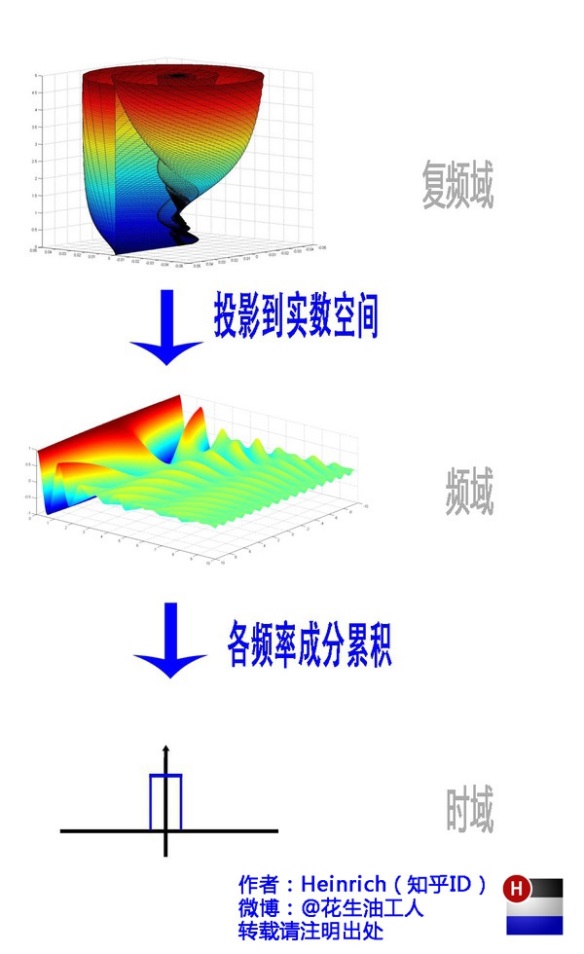


另一种需要借助欧拉公式:



e^(it)可以理解为一条逆时针旋转的螺旋线，那么e^(-it)则可以理解为一条顺时针旋转的螺旋线。而cos(t)则是这两条旋转方向不同的螺旋线叠加的一半，因为这两条螺旋线的虚数部分相互抵消掉了！逆时针旋转的我们称为正频率，而顺时针旋转的我们称为负频率. 举个例子的话，就是极化方向不同的两束光波，磁场抵消，电场加倍。

* 

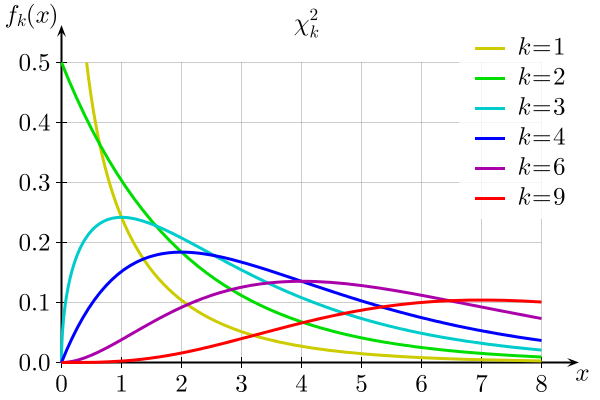


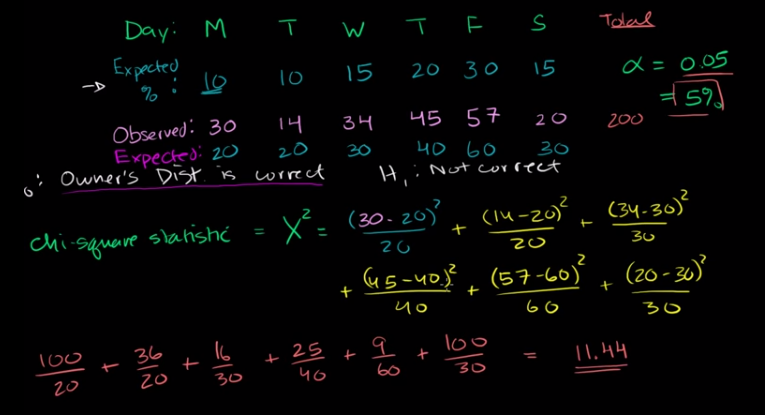
Chi – square distribution

Normal distribution x ~ N (μ,σ)

Q1 = x2 , Q1 ~ χ12 , where degree of freedom is 1

Q2 = x12 + x22 , Q2 ~ χ22 , where degree of freedom is 2





Compare the x2 with the value having the same degree of greedom in the table

χ2 is used to examine the goodness of fit