Variances add when independent

$$Var(X+Y) = (E(X+Y)^2) - E((X+Y)^2)$$

$$= \left(E(X)+E(Y)\right)^2 - E(X^2+2XY+Y^2)$$

$$E(X)^2 + E(Y)^2 + 2E(X)E(Y) - E(X^2) - 2E(XY) - E(Y^2)$$

$$= Var(X) + Var(Y) + 2(E(X)E(Y) - E(XY))$$

So it remains to show that the quantity E(X)E(Y)-E(XY)=0 when X and Y are independent. In fact, this quantity, which is easily shown to be equal to E((X-E(X))(Y-E(Y))) by expanding, is called Covariance (Cov(X,Y)).

When X and Y are independent, then for each possible value of X, the expected value of Y-E(Y)=0. Since the expectation when there are a bunch of possible outcomes is intuitively equal to the sum of the expected value from those outcomes times the probability of those outcomes, we can apply this to the possible values of X, so the quantity E((X-E(X))(Y-E(Y))) becomes a weighted sum of terms like E((x-E(X))P(X=x)(Y-E(Y))), which is a bunch of terms like E(constant(Y-E(Y))) which is 0 because Y-E(Y) is always 0 regardless of the value of X due to the fact that X and Y are independent.

An important idea is that independence allows you to take E(X)E(Y)=E(XY).

Normal distribution formula and variance

Formula:

So we have a probability distribution based on e^{-x^2}, this is an assumption. We change this to $e^{-\frac{x^2}{2}}$ since it turns out (we will prove this after) that this will make the variance equal 1. Then we have to rescale by a factor, which turns out to be $\sqrt{2\pi}$ (This is what we will prove, but really we just need to know that the area exists which is easy to prove, however the square root of pi is so fun that I can't just not show the proof) so that the area under the curve equals 1. We will prove that the total area under e^{-x^2} is $\sqrt{\pi}$ then it is clear that by shifting the mean the area will not change and that if we divide x by a constant σ then we are stretching the curve and therefore its area by a factor of σ so we need to divide the correction factor by σ to compensate and keep the area under the curve at 1.

Now, we will consider what happens if we take the e^{-x^2} curve and rotate it about the y axis, like this:

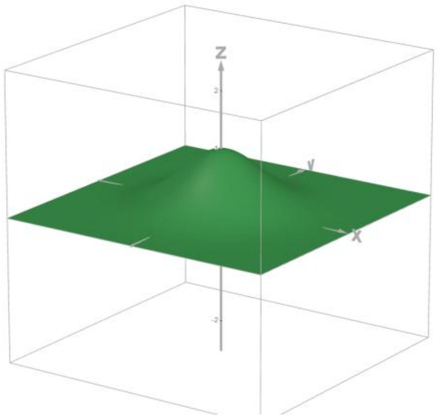


Image: 2 dimensional normal plotted on 3d graph

As you can see, it now essentially becomes a function of the distance from the origin d, ie $z = e^{-d^2}$. Now by pythagoras this becomes $z = e^{-(x^2+y^2)}$. We want to show that the volume of this surface is π, because then we know that if the value $\int_{-\infty}^{\infty} e^{-x^2}dx$ is k then the value of, the strip of this surface for a certain y value with a small width dy (As illustrated in the figure below) is approximately $(\int_{x=-\infty}^{\infty} e^{-(y^2+x^2)}dx)dy = (e^{-y^2}\int_{-\infty}^{\infty} e^{-x^2}dx)dy = (e^{-y^2}k)dy,$

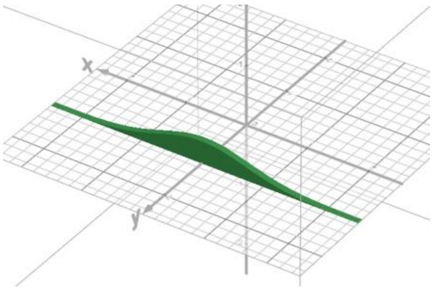


Image: A thin slice of the 2d normal in the above image

Then the volume under the surface is the sum of the volumes of these tiny strips (as dy gets smaller), ie, as we have seen in the interpretation of the integral as a sum, $\int_{-\infty}^{\infty}(e^{-y^2}k)dy = k^2$. Therefore, if we can show that the volume under the surface which is k^2 is π, then k=$\sqrt{\pi}$ as required.

The normal distribution is very special in the sense that if we take x and y independently normally distributed (which makes the probability density become the product of the probability density for x and y), the resulting distribution has rotational symmetry and depends only on the instance, as it becomes $e^{-(x^2+y^2)}$, which is a function of the distance by pythagoras.

Now, to prove that the volume is indeed π, we will consider what happens if we split the volume into concentric rings, like this:

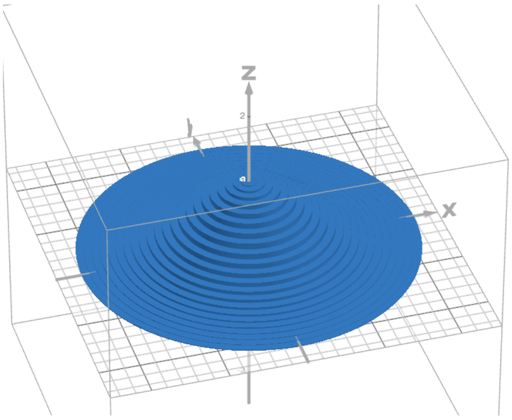


Image: 2d normal on 3d graph approximated by concentric rings of the right height

Suppose the width of the rings is dr and their distance from the origin is r, then the height of the rings is e^{-r^2} and the circumference of them is 2πr so the volume of each ring is 2πre^{-r^2}dr. So, the total volume under the figure is $\int_0^{\infty} 2\pi\mathrm{r}e^{-r^2}\mathrm{dr}$, again by the interpretation of integration as a sum. Note that you can check by differentiating that an antiderivative of $2\pi\mathrm{r}e^{-r^2}$ is $-\pi e^{-r^2}$. So, we have to evaluate $\left[-\pi e^{-r^2}\right]_0^{\infty}$ which is 0-[- π] which is π as required.

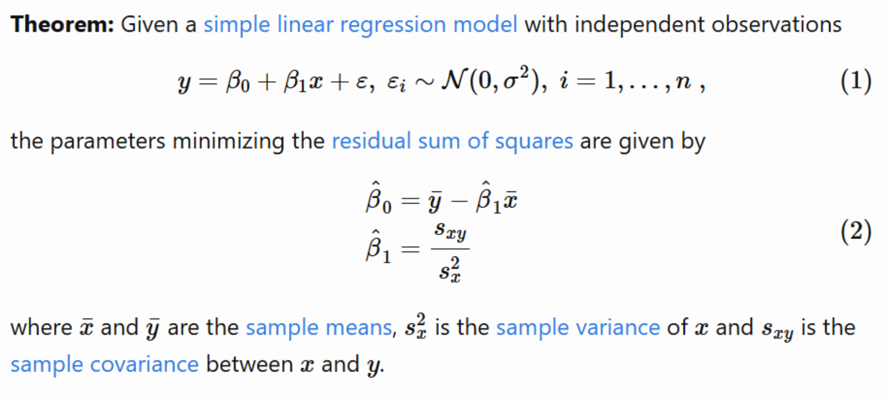
Comment: This is often the case in mathematics, the idea seeing things in 2 different ways is extremely common (like here with strips vs rings).

Variance:

If we have a standard normal $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx$ then the variance is given by $E(x^2)$ since the mean is 0 by symmetry so we need to work out $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}}\,dx$ and show that it is indeed 1, then it becomes clear that the variance is always σ^2 due to scaling properties of variances (ie, if we were to the stretch the normal distribution by 2x, we get this effect by multiplying σ by 2 in the formula and the variance multiplies by 4 by variance scaling properties and the variance is unchanged by shift in the mean since it is related to distance from mean)

Now, we try integrating $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx$ (which, keep in mind, is 1) by parts with $e^{-\frac{x^2}{2}}$ being the part to differentiate and 1 being the part to integrate. This will give us $\frac{1}{\sqrt{2\pi}}\left[\left[xe^{-\frac{x^2}{2}}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}}\,dx\right]$ (the – in the integration by parts formula cancels with the – from differentiating the exponential). Now, since $xe^{-\frac{x^2}{2}}$ goes to 0 as x goes to positive and negative infinity (This is because clearly the derivative of a normal distribution must go to 0 on either side and this is just minus that, or alternatively because the exponential term decays much faster than x grows), this formula, which is equal to 1, reduces to the integral $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}}\,dx$ that we needed to find.

Regression line formula minimizes squares



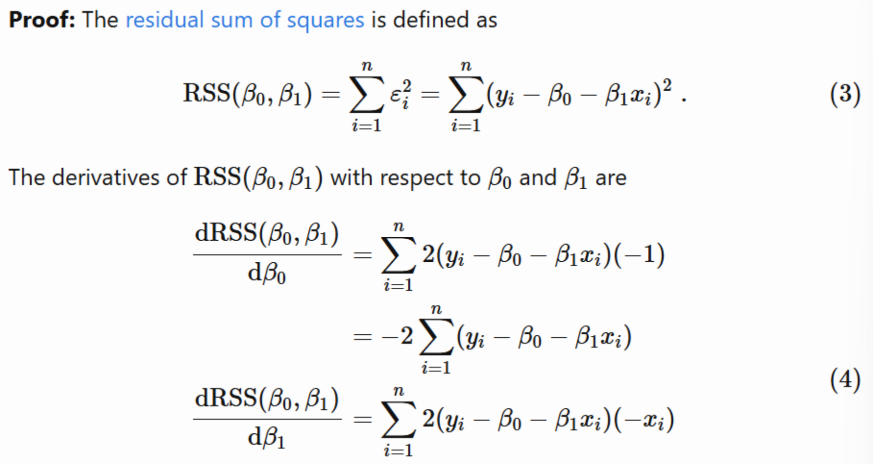
Theorem: Given a simple linear regression model with independent observations

$$y = \beta_0 + \beta_1 x + \varepsilon,\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2),\ i = 1, \ldots, n\ , \qquad (1)$$

the parameters minimizing the residual sum of squares are given by

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \end{aligned} \qquad (2)$$

where $\bar{x}$ and $\bar{y}$ are the sample means, s_x^2 is the sample variance of x and s_{xy} is the sample covariance between x and y.



Proof: The residual sum of squares is defined as

$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2\ . \qquad (3)$$

The derivatives of $\mathrm{RSS}(\beta_0, \beta_1)$ with respect to β_0 and β_1 are

$$\begin{aligned} \frac{\mathrm{dRSS}(\beta_0, \beta_1)}{\mathrm{d}\beta_0} &= \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1) \\ &= -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \\ \frac{\mathrm{dRSS}(\beta_0, \beta_1)}{\mathrm{d}\beta_1} &= \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) \end{aligned} \qquad (4)$$

$$= -2 \sum_{i=1}^{n} (x_i y_i - \beta_0 x_i - \beta_1 x_i^2)$$

and setting these derivatives to zero

$$\begin{aligned}
0 &= -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\
0 &= -2 \sum_{i=1}^{n} (x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2)
\end{aligned} \tag{5}$$

yields the following equations:

$$\begin{aligned}
\hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_0 \cdot n &= \sum_{i=1}^{n} y_i \\
\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + \hat{\beta}_0 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} x_i y_i \; .
\end{aligned} \tag{6}$$

From the first equation, we can derive the estimate for the intercept:

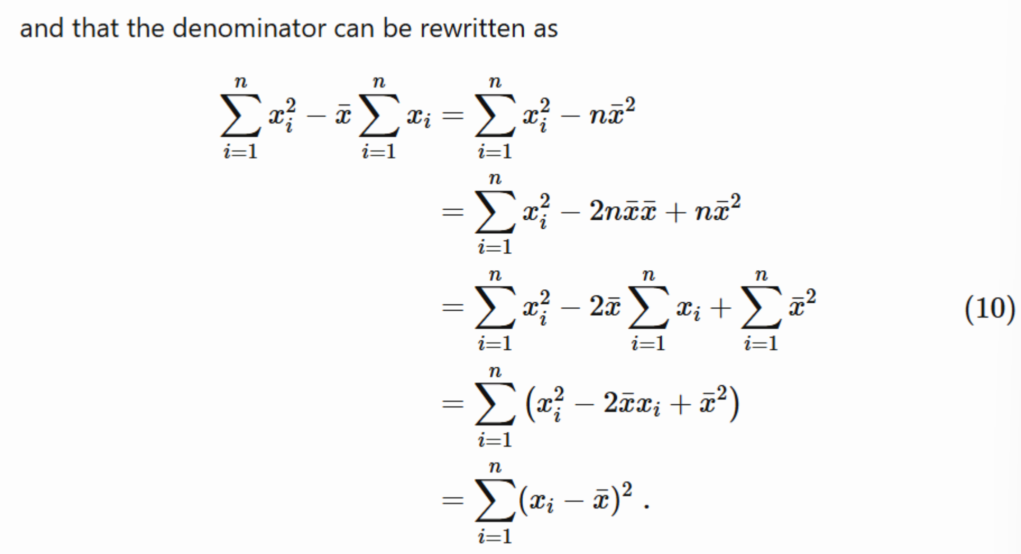
$$\begin{aligned}
\hat{\beta}_0 &= \frac{1}{n} \sum_{i=1}^{n} y_i - \hat{\beta}_1 \cdot \frac{1}{n} \sum_{i=1}^{n} x_i \\
&= \bar{y} - \hat{\beta}_1 \bar{x} \; .
\end{aligned} \tag{7}$$

From the second equation, we can derive the estimate for the slope:

$$\begin{aligned}
\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + \hat{\beta}_0 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} x_i y_i \\
\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + \left(\bar{y} - \hat{\beta}_1 \bar{x} \right) \sum_{i=1}^{n} x_i &\overset{(7)}{=} \sum_{i=1}^{n} x_i y_i \\
\hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i \right) &= \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i} \; .
\end{aligned} \tag{8}$$

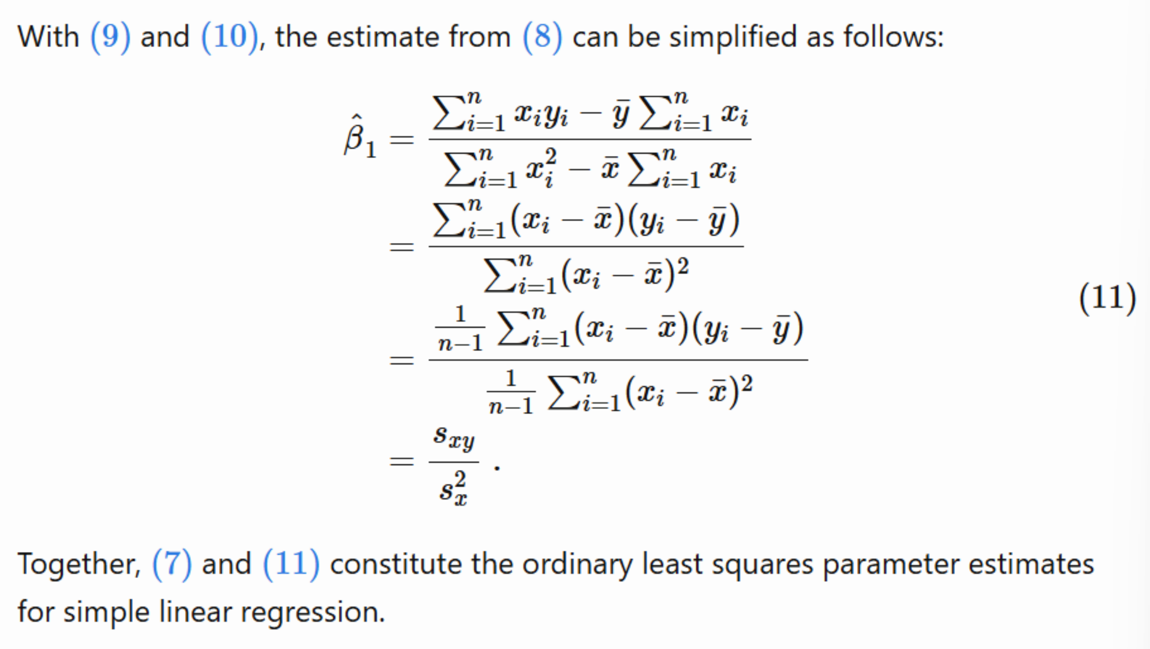
Note that the numerator can be rewritten as

$$\begin{aligned}
\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y} - n \bar{x} \bar{y} + n \bar{x} \bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i - \bar{x} \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \bar{x} \bar{y} \\
&= \sum_{i=1}^{n} (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y}) \\
&= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})
\end{aligned} \tag{9}$$



and that the denominator can be rewritten as

$$\begin{aligned} \sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}\bar{x} + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2 \\ &= \sum_{i=1}^{n} \left(x_i^2 - 2\bar{x}x_i + \bar{x}^2\right) \\ &= \sum_{i=1}^{n} (x_i - \bar{x})^2 \; . \end{aligned} \tag{10}$$

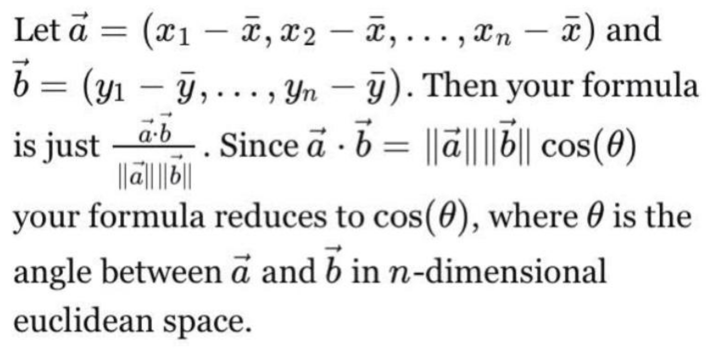


With (9) and (10), the estimate from (8) can be simplified as follows:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{s_{xy}}{s_x^2} \; . \end{aligned} \tag{11}$$

Together, (7) and (11) constitute the ordinary least squares parameter estimates for simple linear regression.

PMCC between -1 and 1



Let $\vec{a} = (x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x})$ and $\vec{b} = (y_1 - \bar{y}, \ldots, y_n - \bar{y})$. Then your formula is just $\frac{\vec{a}\cdot\vec{b}}{||\vec{a}||\,||\vec{b}||}$. Since $\vec{a}\cdot\vec{b} = ||\vec{a}||\,||\vec{b}||\cos(\theta)$ your formula reduces to $\cos(\theta)$, where θ is the angle between $\vec{a}$ and $\vec{b}$ in n-dimensional euclidean space.

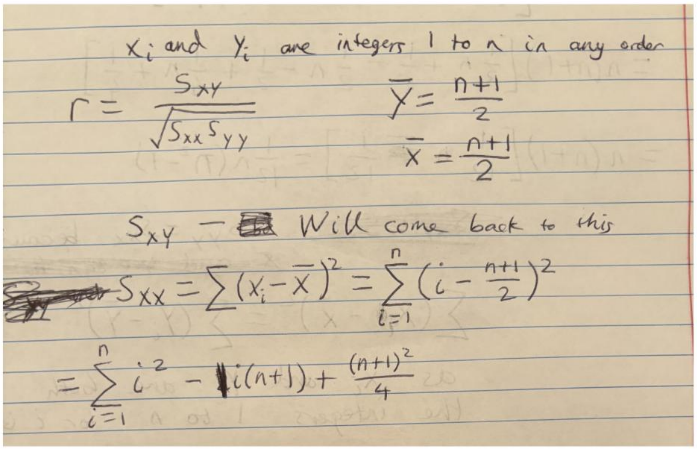
Noting that cos(θ) is between -1 and 1 for real θ

PMCC measures correlation

I mean, if you look at the proof above, you see that it's literally the high dimensional cosine of the angle between your vector of x deviations from the mean and y deviations from the mean! It should then be obvious that if the deviations are close together, the cosine of the angle between them should be high, and vice versa, and furthermore the property should hold that scaling both of them shouldn't change the angle between them, which takes care of that part. It also follows that PMCC is only -1 or 1 when x against y makes a straight line, as the deviation from the sample means of x and y in that case will always be the same, just scaled by a constant, thus the cosine of the angle between them will be -1 or 1.

Rank correlation coefficient formula

We compute the PMCC between the rank numbers.



x_i and y_i are integers 1 to n in any order

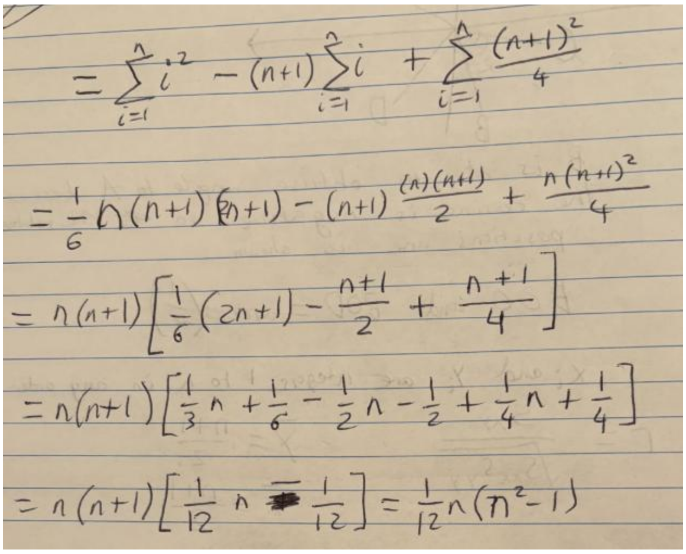
$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \qquad \bar{y} = \frac{n+1}{2} \qquad \bar{x} = \frac{n+1}{2}$$

S_{xy} — Will come back to this

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum_{i=1}^{n} \left(i - \frac{n+1}{2}\right)^2$$

$$= \sum_{i=1}^{n} i^2 - i(n+1) + \frac{(n+1)^2}{4}$$

Images: Shows my derivation on paper



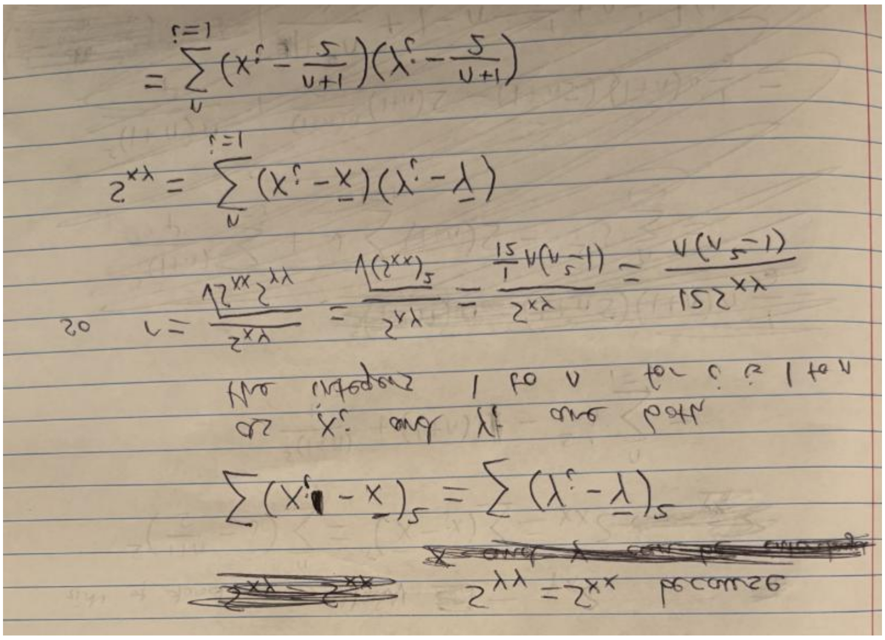
$$= \sum_{i=1}^{n} i^2 - (n+1)\sum_{i=1}^{n} i + \sum_{i=1}^{n} \frac{(n+1)^2}{4}$$

$$= \frac{1}{6}n(n+1)(2n+1) - (n+1)\frac{(n)(n+1)}{2} + \frac{n(n+1)^2}{4}$$

$$= n(n+1)\left[\frac{1}{6}(2n+1) - \frac{n+1}{2} + \frac{n+1}{4}\right]$$

$$= n(n+1)\left[\frac{1}{3}n + \frac{1}{6} - \frac{1}{2}n - \frac{1}{2} + \frac{1}{4}n + \frac{1}{4}\right]$$

$$= n(n+1)\left[\frac{1}{12}n - \frac{1}{12}\right] = \frac{1}{12}n(n^2-1)$$



$S_{yy} = S_{xx}$ because

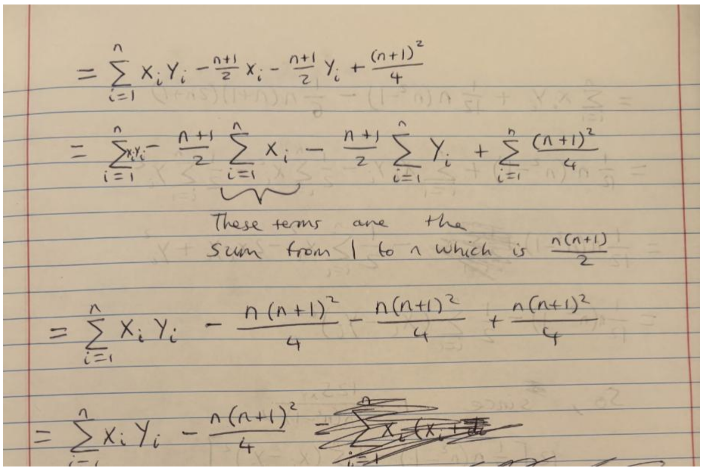
$$\sum (x_i - \bar{x})^2 = \sum (y_i - \bar{y})^2$$

as x_i and y_i are both the integers 1 to n for i = 1 to n

$$\text{so} \quad r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{S_{xy}}{\sqrt{(S_{xx})^2}} = \frac{S_{xy}}{\frac{1}{12}n(n^2-1)} = \frac{12S_{xy}}{n(n^2-1)}$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$= \sum_{i=1}^{n} \left(x_i - \frac{n+1}{2}\right)\left(y_i - \frac{n+1}{2}\right)$$



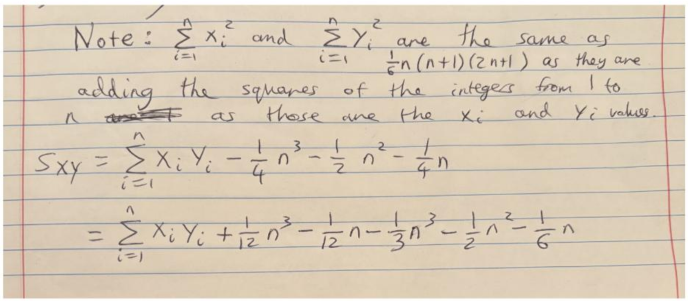
$$= \sum_{i=1}^{n} X_i Y_i - \frac{n+1}{2} X_i - \frac{n+1}{2} Y_i + \frac{(n+1)^2}{4}$$

$$= \sum_{i=1}^{n} X_i Y_i - \frac{n+1}{2} \underbrace{\sum_{i=1}^{n} X_i}_{} - \frac{n+1}{2} \sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} \frac{(n+1)^2}{4}$$

These terms are the sum from 1 to n which is $\frac{n(n+1)}{2}$

$$= \sum_{i=1}^{n} X_i Y_i - \frac{n(n+1)^2}{4} - \frac{n(n+1)^2}{4} + \frac{n(n+1)^2}{4}$$

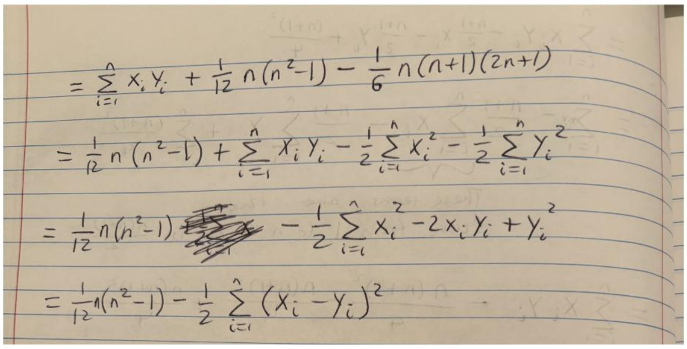
$$= \sum_{i=1}^{n} X_i Y_i - \frac{n(n+1)^2}{4}$$



Note: $\sum_{i=1}^{n} X_i^2$ and $\sum_{i=1}^{n} Y_i^2$ are the same as $\frac{1}{6}n(n+1)(2n+1)$ as they are adding the squares of the integers from 1 to n as those are the X_i and Y_i values.

$$S_{xy} = \sum_{i=1}^{n} X_i Y_i - \frac{1}{4}n^3 - \frac{1}{2}n^2 - \frac{1}{4}n$$

$$= \sum_{i=1}^{n} X_i Y_i + \frac{1}{12}n^3 - \frac{1}{12}n - \frac{1}{3}n^3 - \frac{1}{2}n^2 - \frac{1}{6}n$$

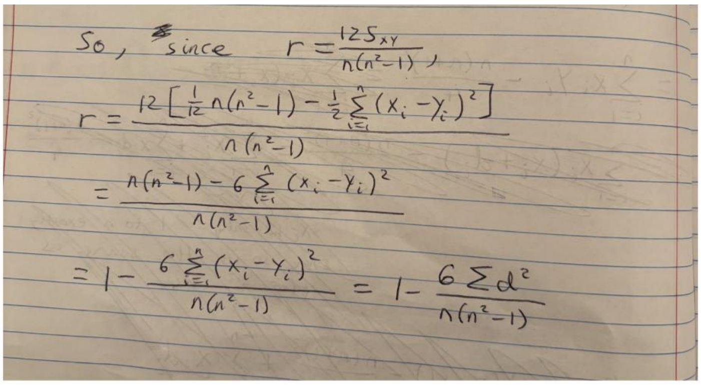


$$= \sum_{i=1}^{n} X_i Y_i + \frac{1}{12}n(n^2-1) - \frac{1}{6}n(n+1)(2n+1)$$

$$= \frac{1}{12}n(n^2-1) + \sum_{i=1}^{n} X_i Y_i - \frac{1}{2}\sum_{i=1}^{n} X_i^2 - \frac{1}{2}\sum_{i=1}^{n} Y_i^2$$

$$= \frac{1}{12}n(n^2-1) - \frac{1}{2}\sum_{i=1}^{n} X_i^2 - 2X_iY_i + Y_i^2$$

$$= \frac{1}{12}n(n^2-1) - \frac{1}{2}\sum_{i=1}^{n} (X_i - Y_i)^2$$



So, since $r = \frac{12 S_{XY}}{n(n^2-1)}$,

$$r = \frac{12\left[\frac{1}{12}n(n^2-1) - \frac{1}{2}\sum_{i=1}^{n}(X_i - Y_i)^2\right]}{n(n^2-1)}$$

$$= \frac{n(n^2-1) - 6\sum_{i=1}^{n}(X_i - Y_i)^2}{n(n^2-1)}$$

$$= 1 - \frac{6\sum_{i=1}^{n}(X_i - Y_i)^2}{n(n^2-1)} = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Binomial distribution formula

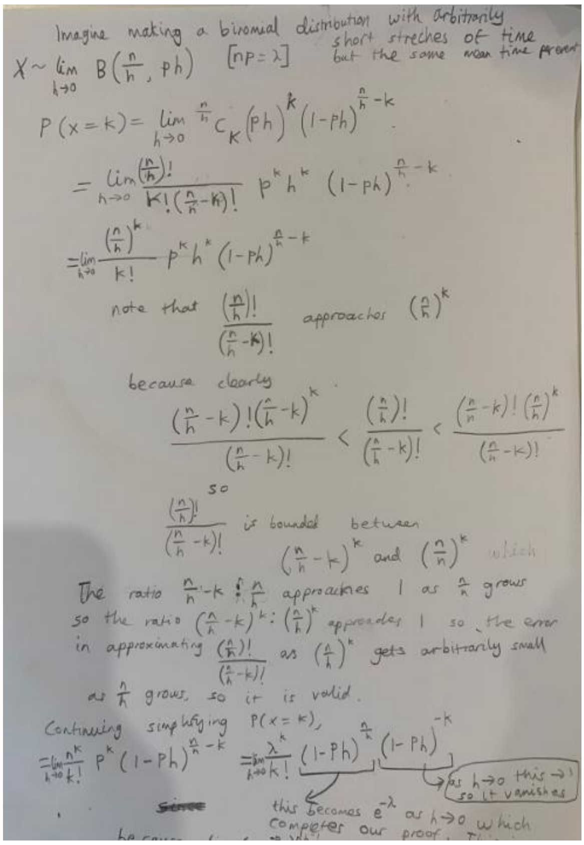
This formula comes from the fact that there are $\binom{n}{r}$ ways to have r out of n trials be “successful”, and the probability of each way is clearly $p^r(1-p)^{n-r}$ since clearly we need “success” r times and “failure” n-r times. So the final probability becomes $\binom{n}{r}p^r(1-p)^{n-r}$.

Poisson distribution formula

Lemma: $\lim_{x \to \infty}(1 + \frac{a}{x})^{bx} = e^{ab}$

Proof of lemma: $\ln(\lim_{x\to\infty}(1+\frac{a}{x})^{bx}) = \lim_{x\to\infty}(\ln((1+\frac{a}{x})^{bx}))$ (Since ln is continuous so intuitively it should therefore commute with limits. This is a standard result in analysis but since it's intuitively true I won't bother proving it)

This limit is equal to $\lim_{x\to\infty}(bx\ln(1+\frac{a}{x}))$ = $b\lim_{\frac{1}{x}\to 0}(\frac{\ln(1+\frac{a}{x})}{\frac{1}{x}}) = ab\lim_{\frac{1}{x}\to 0}(\frac{\ln(1+\frac{a}{x})}{\frac{a}{x}}) = ab\lim_{\frac{1}{x}\to 0}(\frac{\ln\left(1+\frac{a}{x}\right)-\ln(1)}{\frac{a}{x}})$ since ln(1) is 0. But this is just ab times the derivative of ln(x) when x is 1, which is just ab. Since ab is the natural logarithm of the limit, the limit is e^{ab} as required. Now we can derive the poisson formula, by considering a scenario where the expected number of times for an event to occur in a time interval stays constant but the amount of chances for the event to happen gets larger and the probability of it happening in each time interval gets smaller, and taking a limit. This is what we get. We assume that h is n divided by an integer so that the binomial distribution in question is actually defined.



Imagine making a binomial distribution with arbitrarily short streches of time but the same mean time present

$X \sim \lim_{h\to 0} B(\frac{n}{h}, ph)$ [$np = \lambda$]

$P(x=k) = \lim_{h\to 0} {}^{\frac{n}{h}}C_K (ph)^k (1-ph)^{\frac{n}{h}-k}$

$= \lim_{h\to 0} \frac{(\frac{n}{h})!}{K!(\frac{n}{h}-k)!} p^k h^k (1-ph)^{\frac{n}{h}-k}$

$= \lim_{h\to 0} \frac{(\frac{n}{h})^k}{k!} p^k h^k (1-ph)^{\frac{n}{h}-k}$

note that $\frac{(\frac{n}{h})!}{(\frac{n}{h}-k)!}$ approaches $(\frac{n}{h})^k$

because clearly

$$\frac{(\frac{n}{h}-k)!(\frac{n}{h}-k)^k}{(\frac{n}{h}-k)!} < \frac{(\frac{n}{h})!}{(\frac{n}{h}-k)!} < \frac{(\frac{n}{h}-k)!(\frac{n}{h})^k}{(\frac{n}{h}-k)!}$$

so

$\frac{(\frac{n}{h})!}{(\frac{n}{h}-k)!}$ is bounded between $(\frac{n}{h}-k)^k$ and $(\frac{n}{h})^k$ which

The ratio $\frac{n}{h}-k : \frac{n}{h}$ approaches 1 as $\frac{n}{h}$ grows so the ratio $(\frac{n}{h}-k)^k : (\frac{n}{h})^k$ approaches 1 so the error in approximating $\frac{(\frac{n}{h})!}{(\frac{n}{h}-k)!}$ as $(\frac{n}{h})^k$ gets arbitrarily small as $\frac{n}{h}$ grows, so it is valid.

Continuing simplifying $P(x=k)$,

$$= \lim_{h\to 0}\frac{n^k}{k!} p^k (1-ph)^{\frac{n}{h}-k} = \lim_{h\to 0}\frac{\lambda^k}{k!} \underbrace{(1-ph)^{\frac{n}{h}}}_{\text{this becomes } e^{-\lambda} \text{ as } h\to 0 \text{ which completes our proof.}} \underbrace{(1-ph)^{-k}}_{\boxed{\text{as } h\to 0 \text{ this} \to 1 \text{ so it vanishes}}}$$

~~Since~~

Image: Shows my derivation on paper

So our final formula is $\frac{e^{-\lambda}\lambda^k}{k!}$ As required.

In fact, since the poisson probabilities must sum to 1, multiplying by e^{λ} gives another proof for the exponential function taylor series.

Continuous distribution expected value intuition

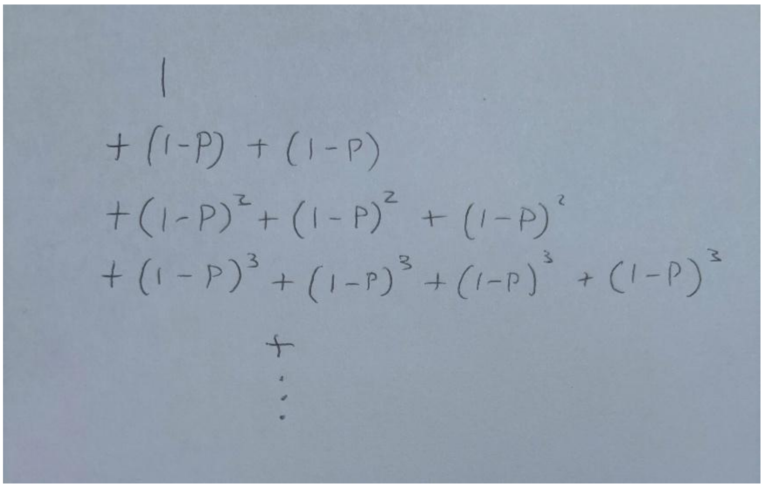
The reason you can think of the expected value of a continuous random variable X with probability density function f(x) as $\int_{-\infty}^{\infty} xf(x)dx$ is because indeed the intuition of expectation as "the sum over all possible values of the value times how likely you are to get that value" is essentially being taken to the limit as the values get closer together and the probabilities get smaller for each one (ie, dx → 0). It is for the same reason that $\int_{-\infty}^{\infty} g(x)f(x)dx$. There isn't really a mathematical statement we need to prove here, it's just a definition that we are providing an intuition for.

Geometric/Negative binomial distribution mean and variance

Mean of geometric should be intuitive. Although I will give a formal proof, think: If something has a 1/3 chance of happening, it should take on average 3 tries for it to happen!

Mean of geometric variable x with probability p (0<p<1) = $\sum_{n=0}^{\infty} nP(x=n)$ = $\sum_{n=0}^{\infty} np(1-p)^{n-1}$ (This is intuitive, it has to happen n times and happen 1 time) = $p\sum_{n=0}^{\infty} n(1-p)^{n-1}$.

This sum is equivalent to adding the rows of the following:



$$\begin{array}{llll} 1 & & & \\ +(1-P) & +(1-P) & & \\ +(1-P)^2 & +(1-P)^2 & +(1-P)^2 & \\ +(1-P)^3 & +(1-P)^3 & +(1-P)^3 & +(1-P)^3 \\ & + & & \\ & \vdots & & \end{array}$$

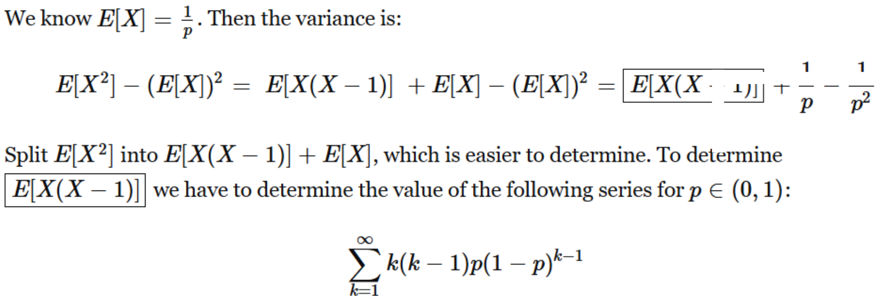
Image: Shows a sum rewritten so that all the terms with a particular exponent are on the same row

We now consider instead adding the columns rather than the rows.

Technical note: Note that this does converge absolutely, since all terms are non-negative and the sum $\sum_{n=0}^{\infty} n(1-p)^{n-1}$ converges by the ratio test. For details on what this means, see the appendix with technical justification of the generalized binomial theorem in the A level maths supplementary material document. The reason this matters is that this means we are allowed to change the order in which we sum terms.

Now we add the columns of the image above rather than the rows, giving us that the mean of a geometric random variable with probability p is now equal to $p\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}(1-p)^n(1-p)^m =$ $p\sum_{n=0}^{\infty}\frac{(1-p)^n}{1-(1-p)}$ (geometric series) = $p\sum_{n=0}^{\infty}\frac{(1-p)^n}{p}$= $\sum_{n=0}^{\infty}(1-p)^n$=$\frac{1}{1-(1-p)}$(geometric series)=$\frac{1}{p}$ as required.

Mean of negative binomial follows from additive property of means, since a negative binomial distribution is the sum of geometric distributions. Variance of negative binomial will follow from additive property of variances when we prove the variance of a geometric distribution, which we will do below:



We know $E[X] = \frac{1}{p}$. Then the variance is:

$$E[X^2] - (E[X])^2 = E[X(X-1)] + E[X] - (E[X])^2 = \boxed{E[X(X-1)]} + \frac{1}{p} - \frac{1}{p^2}$$

Split $E[X^2]$ into $E[X(X-1)] + E[X]$, which is easier to determine. To determine $\boxed{E[X(X-1)]}$ we have to determine the value of the following series for $p \in (0,1)$:

$$\sum_{k=1}^{\infty} k(k-1)p(1-p)^{k-1}$$

Now we substitute q=1-p

$$\sum_{k=1}^{\infty} k(k-1)p(1-p)^{k-1} = p\sum_{k=1}^{\infty} k(k-1)(1-p)^{k-1}$$

$$= p\sum_{k=1}^{\infty}(k-1)kq^{k-1}$$

$$= p\frac{d}{dq}\left(\sum_{k=1}^{\infty}(k-1)q^{k}\right)$$

$$= p\frac{d}{dq}\left(q^2\sum_{k=1}^{\infty}(k-1)q^{k-2}\right)$$

$$= p\frac{d}{dq}\left(q^2\sum_{k=2}^{\infty}(k-1)q^{k-2}\right)$$

$$= p\frac{d}{dq}\left(q^2\frac{d}{dq}\left(\sum_{k=2}^{\infty}q^{k-1}\right)\right)$$

$$= p\frac{d}{dq}\left(q^2\frac{d}{dq}\left(\sum_{k=1}^{\infty}q^{k}\right)\right)$$

Justification for differentiating infinite power series inside their radius of convergence (Which this is, the ratio between consecutive terms approaches 1-p which is between 0 and 1) is given in the A level maths supplementary material document under the generalized binomial theorem formal justification appendix.

$$= p\frac{d}{dq}\left(q^2\frac{d}{dq}\left(\frac{1}{1-q}-1\right)\right)$$

$$= p\frac{d}{dq}\left(\frac{q^2}{(1-q)^2}\right)$$

$$= p\left(\frac{-2q}{(q-1)^3}\right) \qquad \text{Backsub. } q=(1-p)$$

$$= p\left(\frac{-2(1-p)}{((1-p)-1)^3}\right) = p\left(\frac{-2+2p}{-p^3}\right)$$

$$= \frac{-2+2p}{-p^2} = \frac{2(p-1)}{-p^2} = \frac{2(1-p)}{p^2}.$$

Now putting the result back into the equation for $Var[X]$ gives us:

$$Var[X] = E[X(X-1)] + E[X] - (E[X])^2 = \frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{2-2p+p-1}{p^2}$$
$$= \frac{1-p}{p^2}.$$

Generating functions multiply when variables add

I will demonstrate why this works with an example. Suppose we have the following random variables x and y:

K	0	1	2	3
P(x=K)	0.1	0.2	0.3	0.4

K	0	1	2
P(y=K)	0.3	0.5	0.2

Then the generating function of x is $0.1 + 0.2t + 0.3t^2 + 0.4t^3$ and the generating function of y is $0.3 + 0.5t + 0.2t^2$. Now consider how we would find the following:

1. The t^4 coefficient of the product of the generating functions
2. P(x+y=4)

For the first one, when we expand it, we will have a t^4 term from the $0.3t^2$ from the first term times the $0.2t^2$ from the second term, and also the $0.4t^3$ term from the first term times the $0.5t$ from the second term. The products that will give a t^4 term are exactly those whose exponents add up to 4.

Now, to find P(x+y=4), consider all the possible cases: Either x=2 and y=2 or x=3 and y=1. So, since x and y are assumed to be independent, we have that P(x+y=4)=P(x=2)P(y=2)+P(x=3)P(y=1). Crucially, multiplying the probabilities that the variable equals k is the same as multiplying the coefficients of t^k by the definition of generating functions, and we do it on exactly those where x+y, which corresponds to the sum of the exponents, equals 4.

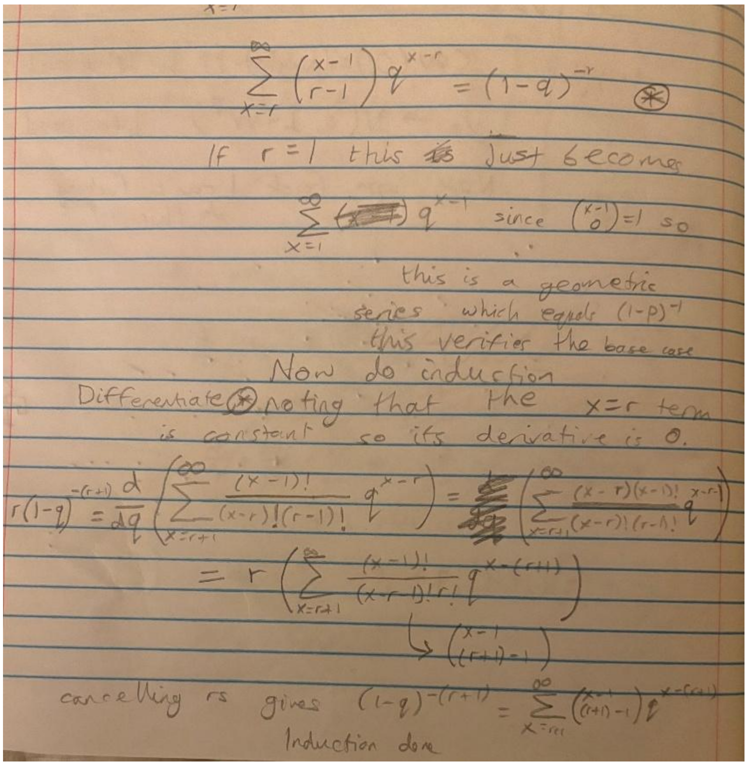
Hopefully this is convincing enough that the product identity holds.

Linear coding property of generating functions

We want to prove that if a random variable X has probability generating function g(t) then aX+b has probability generating function $t^b g(t^a)$. The reason this is true is because when we have aX all the possible values for the variables are multiplied by a and these correspond to the exponents in the generating function which therefore must be multiplied by a. Then by adding b we add b to all the possible values for the variables which corresponds to adding b to the exponents in the generating function, thus it is exactly like multiplying each term by t^b.

Negative binomial generating function

We will first prove that $\sum_{x=r}^{\infty} \binom{x-1}{r-1} q^{x-r} = (1-q)^{-r}$ by induction on r, noting that q=1-p in this context. Note that the differentiation of the power series we will do is allowed since in this context |q|<1 so we are inside the radius of convergence by the ratio test and the fact that the result of a ratio test is the same under differentiation (See the appendix on the binomial theorem in the A level maths document to see what I'm talking about).



$$\sum_{x=r}^{\infty}\binom{x-1}{r-1}q^{x-r} = (1-q)^{-r} \quad \boxed{*}$$

If $r=1$ this ~~is~~ just becomes

$$\sum_{x=1}^{\infty} q^{x-1}$$ since $\binom{x-1}{0}=1$ so

this is a geometric series which equals $(1-p)^{-1}$
this verifies the base case

Now do induction

Differentiate $\boxed{*}$ noting that the $x=r$ term is constant so its derivative is 0.

$$r(1-q)^{-(r+1)} = \frac{d}{dq}\left(\sum_{x=r+1}^{\infty}\frac{(x-1)!}{(x-r)!(r-1)!}q^{x-r}\right) = \left(\sum_{x=r+1}^{\infty}\frac{(x-r)(x-1)!}{(x-r)!(r-1)!}q^{x-r-1}\right)$$

$$= r\left(\sum_{x=r+1}^{\infty}\frac{(x-1)!}{(x-r-1)!r!}q^{x-(r+1)}\right)$$

↳ $\binom{x-1}{(r+1)-1}$

cancelling rs gives $(1-q)^{-(r+1)} = \sum_{x=r+1}^{\infty}\binom{x-1}{(r+1)-1}q^{x-(r+1)}$

Induction done

Image: Shows my proof by induction.

Now we have the lemma, so we will find the generating function of the negative binomial, using q=1-p. Note that here the sums are over x.

$$P(X=x)=\binom{x-1}{r-1}p^r q^{x-r},\ x=r,r+1,r+2,\ldots$$

$$G_X(t)=\sum\binom{x-1}{r-1}p^r q^{x-r}t^x$$

$$=\sum\binom{x-1}{r-1}p^r q^{x-r}t^x$$

$$=\sum\binom{x-1}{r-1}(pt)^r(qt)^{x-r}$$

$$=(pt)^r\sum\binom{x-1}{r-1}(qt)^{x-r}$$

$$=(pt)^r(1-qt)^{-r}$$

$$=\left(\frac{pt}{1-qt}\right)^r=\left(\frac{pt}{1-(1-p)t}\right)^r \quad \text{as required.}$$

Normal + Normal = Normal

In my proof of the area under a normal distribution curve earlier, I briefly discussed the idea that the normal distribution is special since it is completely rotationally symmetrical if you plot in a higher dimensional space the probability density function of different standard normals. Suppose X and Y are N(0,1) random variables that are independent, then it suffices to show that aX+bY=N(0, a^2+b^2) since if X and Y were shifted by constants the sum would just shift accordingly and it would all be fine. Here is the joint probability density function of X and Y visualised:

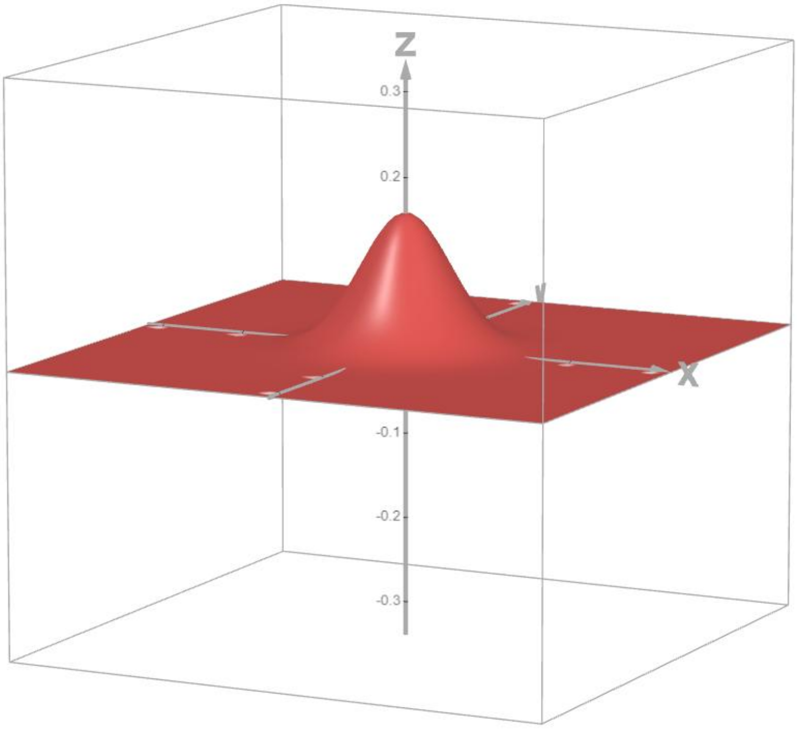


Image: another 2d normal plotted on a 3d graph

As you can see, there is rotational symmetry. We actually just need to show that aX+bY is normal then we will know from mean and variance additive properties the desired result. Suppose, for example, we want to find the probability density function of P(3x+2y), then the probability density that P(3x+2y=k), intuitively, corresponds to the area under the slice of the above diagram corresponding to 3x+2y=k. In the end, we will have to rescale the probability density function of 3x+2y so that the area under that curve is 1, but it's quite obvious we have proportionality, which is what we actually need. Below is the slices I mean for some values of k.

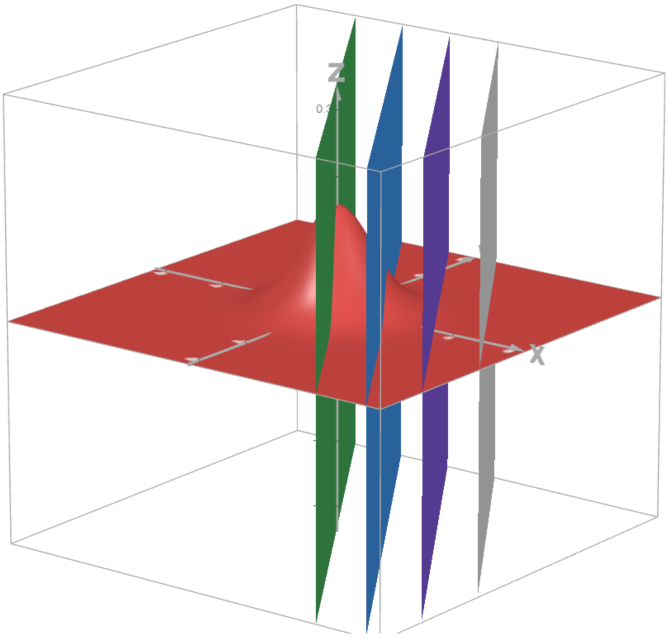
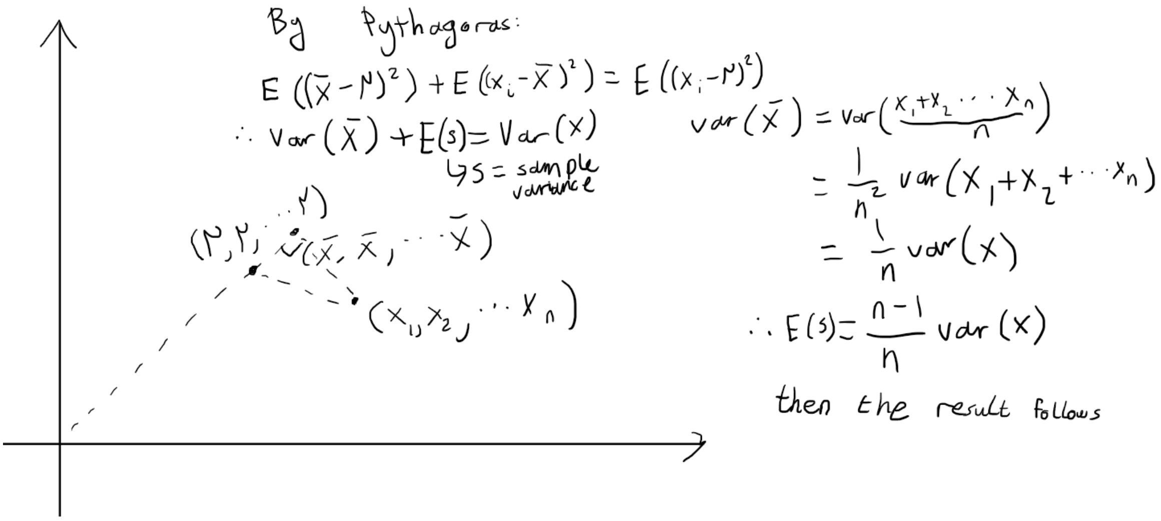


Image: The same 2d normal with parallel vertical planes on the same graph

Visually you can see from the rotational symmetry of the image that the function of the area under the red in each of these planes, which as discussed is what we need, is a normal distribution, completing what is not really a proof but rather a visual argument.

Unbiased variance estimator, vector approach



By Pythagoras:

$$E((\bar{X}-\mu)^2) + E((x_i-\bar{X})^2) = E((x_i-\mu)^2)$$

$$\therefore Var(\bar{X}) + E(s) = Var(X)$$

↳ s = sample variance

$$var(\bar{X}) = var\left(\frac{X_1+X_2 \cdots X_n}{n}\right)$$

$$= \frac{1}{n^2} var(X_1+X_2+\cdots X_n)$$

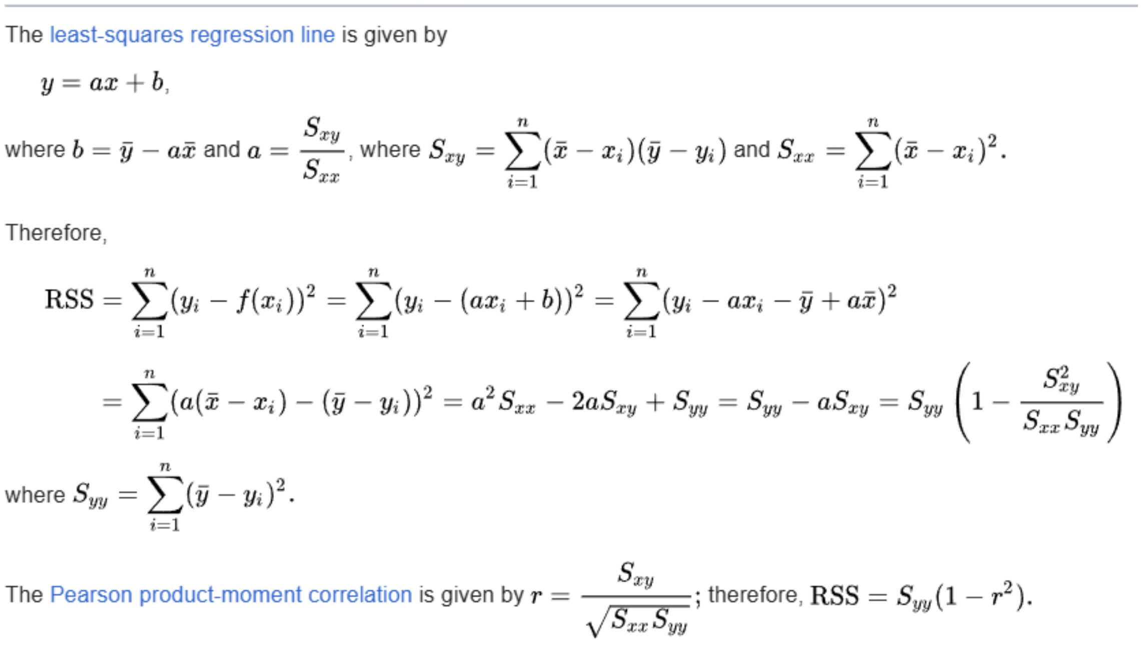
$$= \frac{1}{n} var(X)$$

$$\therefore E(s) = \frac{n-1}{n} var(X)$$

then the result follows

Image: Diagram to show pythagoras vector approach for variance estimator

Residual sum of squares formula



The least-squares regression line is given by

$$y = ax + b,$$

where $b = \bar{y} - a\bar{x}$ and $a = \frac{S_{xy}}{S_{xx}}$, where $S_{xy} = \sum_{i=1}^{n}(\bar{x} - x_i)(\bar{y} - y_i)$ and $S_{xx} = \sum_{i=1}^{n}(\bar{x} - x_i)^2$.

Therefore,

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - f(x_i))^2 = \sum_{i=1}^{n}(y_i - (ax_i + b))^2 = \sum_{i=1}^{n}(y_i - ax_i - \bar{y} + a\bar{x})^2$$

$$= \sum_{i=1}^{n}(a(\bar{x} - x_i) - (\bar{y} - y_i))^2 = a^2 S_{xx} - 2aS_{xy} + S_{yy} = S_{yy} - aS_{xy} = S_{yy}\left(1 - \frac{S_{xy}^2}{S_{xx}S_{yy}}\right)$$

where $S_{yy} = \sum_{i=1}^{n}(\bar{y} - y_i)^2$.

The Pearson product-moment correlation is given by $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$; therefore, $\mathrm{RSS} = S_{yy}(1 - r^2)$.

Note: the fact that $a^2S_{xx} = aS_{xy}$ is because of the definition of a.

Characteristic functions

Here we will define characteristic functions and prove some basic properties of them (although these basic properties have not such basic proofs). These will play a “central” role in our proof. If I have a probability distribution x, then I define the characteristic function of x as $\phi_x(t) = E(e^{itx})$. This is related to the fourier transform of the ditstribution, that is the cf divided by 2π. What we will prove here is that if you know the characteristic function of a distribution you can reverse engineer what the distribution was. To fully understand the proof, you will need to know the results in the “Some technical results” document, which is a separate document because the results are there are not only relevant to statistics, so it would be weird to include them in here. The idea will be to show that from the characteristic function you can reverse engineer the integral from a to b of our probability distribution for any a and b, which determines the entire distribution.

If f is our probability distribution, then

$$\phi_f(t) = \int_{\mathbb{R}} f(y) e^{ity}\, dy$$

And define, for arbitrary a and b with a<b,

$$I_\varepsilon \;:=\; \frac{1}{2\pi} \int_{\mathbb{R}} \phi_f(t)\, e^{-\varepsilon t^2}\, \frac{e^{-iat} - e^{-ibt}}{it}\, dt$$

This is equal to

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}[\int_{-\infty}^{\infty} f(y)e^{ity}dy]e^{-\varepsilon t^2}\frac{e^{-iat}-e^{-ibt}}{it}dt$$

Now we will show that

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left|f(y)e^{ity}e^{-\varepsilon t^2}\frac{e^{-iat}-e^{-ibt}}{it}\right|dydt$$

Is finite so that we can swap the integrals around. Note that for each fixed t, $\int_{-\infty}^{\infty}\left|f(y)e^{ity}e^{-\varepsilon t^2}\frac{e^{-iat}-e^{-ibt}}{it}\right|dy = \int_{-\infty}^{\infty}\left|f(y)e^{ity}\right|\left|e^{-\varepsilon t^2}\right|\left|\frac{e^{-iat}-e^{-ibt}}{it}\right|dy =$ $\left|e^{-\varepsilon t^2}\right|\left|\frac{e^{-iat}-e^{-ibt}}{it}\right|\int_{-\infty}^{\infty}\left|f(y)e^{ity}\right|dy \le \left|e^{-\varepsilon t^2}\right|\left|\frac{e^{-iat}-e^{-ibt}}{it}\right| |\int_{-\infty}^{\infty} f(y)e^{ity}dy|$

Since this third term equals $E(e^{ity})$ which is the expectation of things not outside the complex unit circle, its absolute value must be no greater than 1. Therefore, $\int_{-\infty}^{\infty}\left|f(y)e^{ity}e^{-\varepsilon t^2}\frac{e^{-iat}-e^{-ibt}}{it}\right|dy \le$ $\left|e^{-\varepsilon t^2}\right|\left|\frac{e^{-iat}-e^{-ibt}}{it}\right|$, so it remains to show that $\int_{-\infty}^{\infty}\left|e^{-\varepsilon t^2}\right|\left|\frac{e^{-iat}-e^{-ibt}}{it}\right|dt$ is finite (as the factor of 2π does not change finiteness), which eimplifies our problem massively.

To do this, I will put some bounds on the term$\left|\frac{e^{-iat}-e^{-ibt}}{it}\right|$. This term is equal to $\frac{|e^{-iat}-e^{-ibt}|}{|t|}$.

Now I will use the inequality $\left|e^{ix}-e^{iy}\right| < |x-y|$ if x and y are real numbers not equal to eachother. The reason this inequality is true can be shown geometrically:

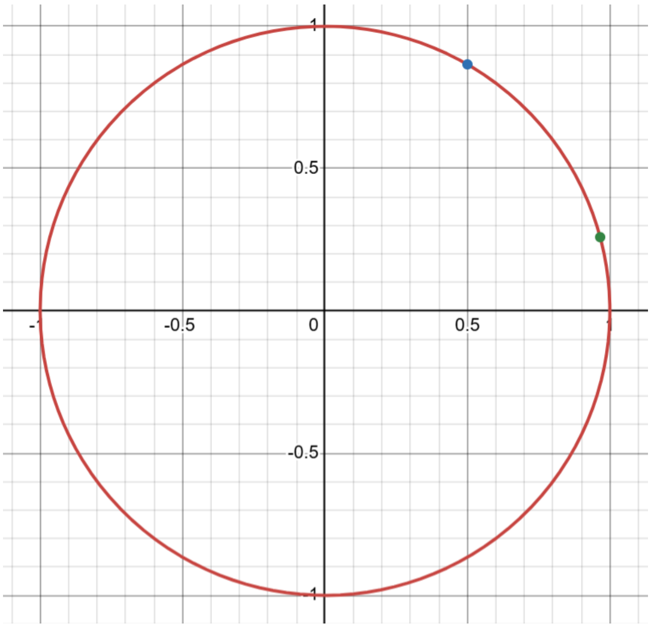


Image: A unit circle on the complex plane with 2 points marked to demonstrate the “shortest distance” principle

Here, the green and blue points represent values of e^{ix} for certain values of x. The distance that they are apart along the unit circle is exactly the difference between these values of x, but the distance between these points is less than this difference, because you could get further by walking in a straight line, as that is more efficient. That is why this inequality holds.

So, if |t|<1, our expression is less than $\frac{|at-bt|}{|t|}$ which is just |a-b|. If |t|≥1, then the numerator is the absolute value of the difference between two points on the unit circle, which cannot be more than the diameter of the unit circle which is 2, so it is at most $\frac{2}{|t|}$. So, the integral $\int_{-\infty}^{\infty}\left|e^{-\varepsilon t^2}\right|\left|\frac{e^{-iat}-e^{-ibt}}{it}\right| dt$ is at most this:

$$\int_{|t|\leq 1} |b-a|\, e^{-\varepsilon t^2}\, dt + \int_{|t|>1} \frac{2\, e^{-\varepsilon t^2}}{|t|}\, dt$$

The first term is at most 2|b-a|. The second term will just get larger if you remove the denominator, and then it turns into an integral which we have already proven has a finite value, which we could give in terms of the square root of pi.

Now we get this after swapping the integrals and combining some exponents:

$$I_\varepsilon = \int_{\mathbb{R}} f(y)\, K_\varepsilon(y)\, dy, \quad K_\varepsilon(y) := \frac{1}{2\pi}\int_{\mathbb{R}} e^{-\varepsilon t^2}\, \frac{e^{it(y-a)} - e^{it(y-b)}}{it}\, dt.$$

Definition: erf(x):=$\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2} dt$. This function goes to 1 as x goes to infinity and -1 as x goes to negative infinity. It is essentially a cdf of the normal rescaled to go from -1 to 1.

I will now define another new function as follows:

$$H_\varepsilon(u) := \frac{1}{2\pi}\int_{\mathbb{R}} e^{-\varepsilon t^2}\frac{e^{itu}}{it}\,dt$$

Note that $H_\varepsilon(u) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\varepsilon t^2}\int_0^u e^{its}dsdt.$

We want to show that $\int_{-\infty}^{\infty}\int_0^u |e^{-\varepsilon t^2}e^{its}|dsdt$ is finite so that we can swap the integrals around.

$\int_{-\infty}^{\infty}\int_0^u |e^{-\varepsilon t^2}e^{its}|dsdt = \int_{-\infty}^{\infty}\int_0^u |e^{-\varepsilon t^2}||e^{its}|dsdt = \int_{-\infty}^{\infty}|e^{-\varepsilon t^2}|\int_0^u |e^{its}|dsdt = \int_{-\infty}^{\infty} e^{-\varepsilon t^2}\int_0^u 1dsdt = \int_{-\infty}^{\infty} ue^{-\varepsilon t^2}dt.$

If we use the substitution $v = t\sqrt{\varepsilon}$ it will follow that this integral is $u\sqrt{\frac{\pi}{\varepsilon}}$ which is finite. Therefore we get that $H_\varepsilon(u)$ is equal to this:

$$\int_0^S \left(\int_{\mathbb{R}} e^{-\varepsilon t^2}e^{ist}\,dt\right)ds$$

We now complete the square as follows:

$$-\varepsilon t^2 + ist = -\varepsilon\left(t - \frac{is}{2\varepsilon}\right)^2 - \frac{s^2}{4\varepsilon}$$

To get

$$\int_{\mathbb{R}} e^{-\varepsilon t^2}e^{ist}\,dt = e^{-\,s^2/(4\varepsilon)}\int_{\mathbb{R}} e^{-\varepsilon\left(t-\frac{is}{2\varepsilon}\right)^2}dt = \sqrt{\frac{\pi}{\varepsilon}}\;e^{-\,s^2/(4\varepsilon)}$$

The reason we can shift by an imaginary constant step and still get the same result is because our function has a global single valued antiderivative defined by the taylor series which converges everywhere, so you can integrate along contours like these and get the same result

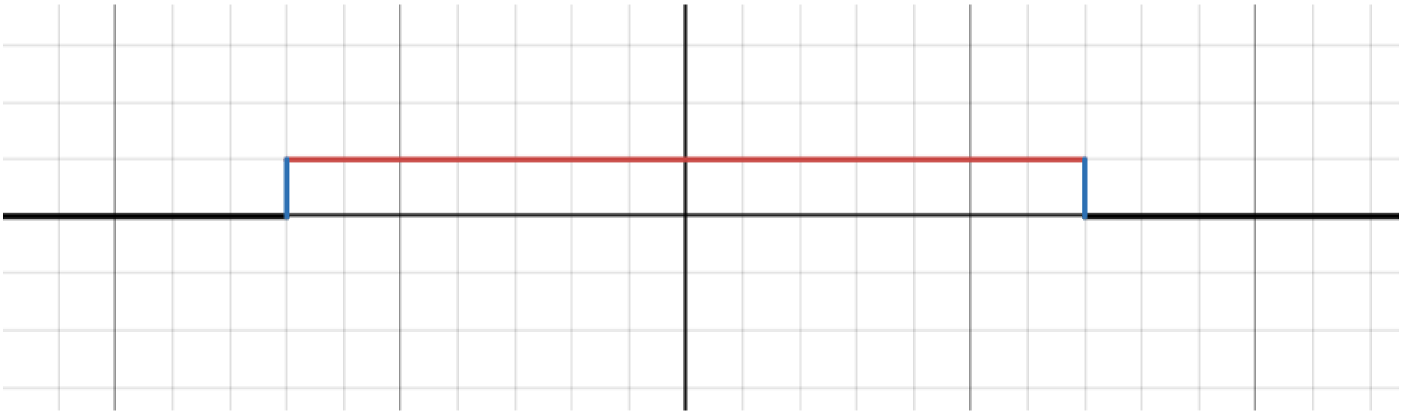


Image: A contour that is the reals shifted by an imaginary constant for a large interval

As we make the red part longer, we know from the shape of the bell curve that the area of the tails will go to 0, so the limit of these integrals which are all the same as the integral along the real line will be the same as the integral along the real line and be equal to the integral with the imaginary shift.

Therefore, $H_\varepsilon(u)$ is equal to $\frac{1}{2\pi}\sqrt{\frac{\pi}{\varepsilon}}\int_0^u e^{-\frac{s^2}{4\varepsilon}}ds$, which can be shown by the substitution $v = \frac{s}{2\sqrt{\varepsilon}}$ to be equal to $\frac{1}{2}\,\text{erf}\,(\frac{u}{2\sqrt{\varepsilon}})$.

Now we have the following identity from the definitions:

$$K_\varepsilon(y) = H_\varepsilon(y-a) - H_\varepsilon(y-b) = \frac{1}{2}\left[\operatorname{erf}\Big(\frac{y-a}{2\sqrt{\varepsilon}}\Big) - \operatorname{erf}\Big(\frac{y-b}{2\sqrt{\varepsilon}}\Big)\right]$$

Now consider what the K function approaches as ε approaches 0: If y<a, then y also is <b since we defined a<b. Therefore, both terms will go to -1 as the inputs to the erf functions will go to -infinity, so the whole thing will go to 0. Same if y>b. But, if a<y<b, then the first term's input will go to infinity so the first term will go to 1, and the second terms input will go to -infinity so the whole thing will go to ½(1-(-1)) which is 1. Therefore, as ε goes to 0, K approaches the function that is 1 when y is between a and b and 0 otherwise, and it approaches ½ at exactly a and b, but this doesn't really matter. What does matter is that K is always between -1 and 1.

Recall, we had this:

$$I_\varepsilon = \int_{\mathbb{R}} f(y)\,K_\varepsilon(y)\,dy, \quad K_\varepsilon(y) := \frac{1}{2\pi}\int_{\mathbb{R}} e^{-\varepsilon t^2}\,\frac{e^{it(y-a)} - e^{it(y-b)}}{it}\,dt.$$

We know that since f is always positive and the integral of f over the reals is 1 since f is a probability distribution, and K is between 0 and 1, define $f_n(y) := f(y)K_{\frac{1}{n}}(y)$, then $f_n(y)$ is bounded above in absolute value by $f(y)$ which has a finite integral, so dominated convergence applies. Since $f_n(y)$ converges pointwise to f when a<y<b and 0 otherwise (and what happens at exactly a and b does not affect the value of the integrals), we know by dominated convergence on f_n that as ε (or 1/n, same thing) goes to 0, I_ε goes to $\int_a^b f(x)dx$. But I_ε was defined only in terms of the characteristic function. So done.

We will now prove that if the characteristic function of a sequence of probability distributions converges, then the distributions converge to the probability distribution with the limit characteristic function. To do this, we will define yet another function. Let $L_\varepsilon(t) := \frac{e^{-\varepsilon t^2}}{2\pi}\frac{e^{-iat}-e^{-ibt}}{it}$, then $I_\varepsilon = \int_{-\infty}^{\infty}\phi(t)L_\varepsilon(t)dt$. L_ε is integrable with the same proof as earlier and bounded by $\frac{b-a}{2\pi}$, so if our sequence of characteristic functions $\phi_n(t)$ converges pointwise to $\phi(t)$ then $\phi_n(t)L_\varepsilon(t)$ converges to $\phi(t)L_\varepsilon(t)$ as n goes to infinity. Also, $|\phi_n(t)|\leq 1$ from earlier, so $|\phi_n(t)L_\varepsilon(t)|<|L_\varepsilon(t)|$. Therefore all the hypotheses for the dominated convergence theorem apply, so the limit of $\int_{-\infty}^{\infty}\phi_n(t)L_\varepsilon(t)dt$ is indeed I_ε. Therefore, letting ε approach 0, we have that the probability our distribution lands between a and b approaches that probability for a pdf with cf ϕ if the cf's converge to ϕ. So done.

Note: This means the distribution converges if the cf converges, in the sense that the weight of the distribution on any interval converges, but the probability density function need not converge pointwise. Convergence in distribution actually means the cumulative distribution function converges pointwise at all points where it is continuous, which this satisfies since we showed that all integrals of the pdf converge so the cdf converges.

We will also prove the converse, ie that convergence in distribution implies convergence in characteristic function. This is needed for the chi squared table derivation, and everything before this was needed for the central limit theorem. This also requires knowledge of some of the results in the technical results document.

Suppose we have a sequence of probability distributions X_n converging in distribution to a probability distribution X. For any ε>0, we can pick a point on the cdf of X with a value strictly less than $\frac{\varepsilon}{2}$, and another point with a value strictly greater than $1-\frac{\varepsilon}{2}$. This is always possible since in the extremes, the cdf approaches 0 and 1, so we can find points as close as we want to 0 and 1. Let M be at least the maximum absolute value of these points we've picked and ensure we pick M such that at M and -M, the cdf of X is continuous and does not jump. Then $P(|X|>M)<\varepsilon$. But since X_n converges in distribution to X, the cdfs converge pointwise, so at -M and M where the cdf of X is within $\frac{\varepsilon}{2}$ of 0 and 1 respectively, by definition of convergence we can find an N such that for all X_n with n>N, the cdf of X_n at M and -M get as close as we want to the cdf of X at M and -M. Specifically, make it so close that it is still within $\frac{\varepsilon}{2}$ of 0 and 1 respectively. Then we have shown that for any ε>0 there exists M and N such that if n>N then $P(|X_n|>M)<\varepsilon$.

Now for any ε>0 pick M and N_1 such that if $n>N_1$ we have that $P(|X_n|>M)<\frac{\varepsilon}{4}$. Now let 1_S be the indicator function of the set S, ie the function that returns 1 for points inside the set S and 0 everywhere else. Then $E\left(e^{itX_n}\right)-E\left(e^{itX}\right)$

$$=E\left(e^{itX_n}1_{|X_n|\leq M}+e^{itX_n}1_{|X_n|>M}\right)-E\left(e^{itX}1_{|X|\leq M}+e^{itX}1_{|X|>M}\right)$$

We simply have that the expectation overall is the expectation when inside a set plus the expectation when outside that set, this is obvious. We split this even further to get the following:

$$=E\left(e^{itX_n}1_{|X_n|\leq M}\right)+E\left(e^{itX_n}1_{|X_n|>M}\right)-E\left(e^{itX}1_{|X|\leq M}\right)-E\left(e^{itX}1_{|X|>M}\right)$$

Now, recall that we have the triangle inequality for both real numbers and integrals. Since expectations are actually defined in terms of integrals, we do have the inequality |E(x)|≤E(|x|) directly from the integral inequality. So, we have:

$$\left|E\left(e^{itX_n}\right)-E\left(e^{itX}\right)\right|$$

$$=\left|E\left(e^{itX_n}1_{|X_n|\leq M}\right)+E\left(e^{itX_n}1_{|X_n|>M}\right)-E\left(e^{itX}1_{|X|\leq M}\right)-E\left(e^{itX}1_{|X|>M}\right)\right|$$

By what we just did

$$\leq\left|E\left(e^{itX_n}1_{|X_n|\leq M}\right)-E\left(e^{itX}1_{|X|\leq M}\right)\right|+\left|E\left(e^{itX_n}1_{|X_n|>M}\right)\right|+\left|E\left(e^{itX}1_{|X|>M}\right)\right|$$

By the normal triangle inequality

$$\leq\left|E\left(e^{itX_n}1_{|X_n|\leq M}\right)-E\left(e^{itX}1_{|X|\leq M}\right)\right|+E\left(\left|e^{itX_n}\right|1_{|X_n|>M}\right)+E\left(\left|e^{itX}\right|1_{|X|>M}\right)$$

By the triangle inequality for expectations. However, $\left|e^{itX_n}\right| = |e^{itX}| = 1$ so since this is only when $|X_n| > M$, we have that

$$\left|E\left(e^{itX_n}\right) - E\left(e^{itX}\right)\right| \leq \left|E\left(e^{itX_n}1_{|X_n|\leq M}\right) - E\left(e^{itX}1_{|X|\leq M}\right)\right| + \frac{\varepsilon}{4}$$

Which helps a lot, since the goal is to make this term less than a full epsilon so we have that the characteristic functions get as close together as we want so we have the desired result.

Now, e^{itx} on [-M,M] is a continuous function on a closed bounded interval so it is uniformly continuous. Therefore we can find a δ small enough that on any interval of length at most δ, e^{itx} is within a ball in the complex plane of diameter $\frac{\varepsilon}{4}$, which we can do by considering real and imaginary parts separately and making sure they are in an interval of length $\frac{\varepsilon}{4\sqrt{2}}$, ensuring the whole thing is in a square within the circle that we need. Since this δ works everywhere by uniform continuity, we can find a finite partition $-M = a_0 < a_1 < \cdots < a_m = M$ where the distance between each x is at most δ. We take care to make sure that each a is a point where the cdf is continuous, and this will have the property that if $g(x) \coloneqq e^{itx}$, then for any x and y between a_{j-1} and a_j we have that $|g(x) - g(y)| < \frac{\varepsilon}{4}$. In particular, for any ξ_j with $a_{j-1} < \xi_j \leq a$, which I will now pick arbitrarily for each interval j, we have that the difference between $c_j \coloneqq g(\xi_j)$ and g(anything else in that interval) is bounded above in absolute value by $\frac{\varepsilon}{4}$. I will now define the simple function $s(x) \coloneqq \sum_{j=1}^{m} c_j 1_{(a_{j-1},a_j]}$, then since $|c_j|$ is 1 since it is equal to $e^{i*stuff}$, we have that it is always the case that $|g(x) - s(x)| < \frac{\varepsilon}{4}$. Now lets work on the $\left|E\left(e^{itX_n}1_{|X_n|\leq M}\right) - E\left(e^{itX}1_{|X|\leq M}\right)\right|$ term which we hope to bound by $\frac{3\varepsilon}{4}$ in order to be done: This term is equal to $\left|\int_{-M}^{M} g(x)f_n(x)dx - \int_{-M}^{M} g(x)f(x)dx\right|$ because of the definition of expected value in terms of integrals and the definition of g(x) and the fact that the indicator function makes this be on the interval [-M,M]. Here f_n is the pdf of X_n and f is the pdf of x. I can write the term as

$$\left|\int_{-M}^{M}(g-s)(x)f_n(x)dx - \int_{-M}^{M}(g-s)(x)f(x)dx + \int_{-M}^{M}s(x)f_n(x)dx - \int_{-M}^{M}s(x)f(x)dx\right|$$

Then both versions of the triangle inequality mean that this is

$$\leq \int_{-M}^{M}|(g-s)(x)|f_n(x)dx + \int_{-M}^{M}|(g-s)(x)|f(x)dx + \left|\int_{-M}^{M}s(x)f_n(x)dx - \int_{-M}^{M}s(x)f(x)dx\right|$$

Since f is positive and real so we can pull it out of the absolute value (this also justifies the inequality for expectations in general). Now using the bounds we got earlier we can simplify this even further:

$$\leq \int_{-M}^{M}\frac{\varepsilon}{4}f_n(x)dx + \int_{-M}^{M}\frac{\varepsilon}{4}f(x)dx + \left|\int_{-M}^{M}s(x)f_n(x)dx - \int_{-M}^{M}s(x)f(x)dx\right|$$

$$\leq \frac{\varepsilon}{4}\int_{-M}^{M}f_n(x)dx + \frac{\varepsilon}{4}\int_{-M}^{M}f(x)dx + \left|\int_{-M}^{M}s(x)f_n(x)dx - \int_{-M}^{M}s(x)f(x)dx\right|$$

$$\leq \frac{\varepsilon}{2} + \left|\int_{-M}^{M}s(x)f_n(x)dx - \int_{-M}^{M}s(x)f(x)dx\right|$$

Since those integrals cannot be greater than 1. So we just have to make sure that

$\left|\int_{-M}^{M}s(x)f_n(x)dx - \int_{-M}^{M}s(x)f(x)dx\right| \leq \frac{\varepsilon}{4}$

Then we're done, since we will have that the characteristic function converges for every fixed t and therefore converges pointwise.

To do this, we note that since s is equal to c_j on each interval (a_{j-1}, a_j), we have that the term in question can be written as

$$\left|\sum_{j=1}^{m}\int_{a_{j-1}}^{a_j} c_j f_n(x)dx - \sum_{j=1}^{m}\int_{a_{j-1}}^{a_j} c_j f(x)dx\right|$$

But the integral of f and f_n is the cdf, which I'll call F and F_n respecively, so the term becomes

$$\left|\sum_{j=1}^{m} c_j[F_n(a_j) - F_n(a_{j-1})] - \sum_{j=1}^{m} c_j[F(a_j) - F(a_{j-1})]\right|$$

$$= \left|\sum_{j=1}^{m} c_j\left[\left(F_n(a_j) - F_n(a_{j-1})\right) - \left(F(a_j) - F(a_{j-1})\right)\right]\right|$$

$$\leq \sum_{j=1}^{m} \left|c_j\left[\left(F_n(a_j) - F_n(a_{j-1})\right) - \left(F(a_j) - F(a_{j-1})\right)\right]\right|$$

By the triangle inequality

$$= \sum_{j=1}^{m} \left|\left[\left(F_n(a_j) - F_n(a_{j-1})\right) - \left(F(a_j) - F(a_{j-1})\right)\right]\right|$$

Since $|c_j| = 1$

$$\leq \sum_{j=1}^{m} \left[|F_n(a_j) - F(a_j)| + |F_n(a_{j-1}) - F(a_{j-1})|\right]$$

By the triangle inequality.

Now, since each a_j is a point where F is continuous, it means F_n converges to F there. So if n is large enough $F_n(a_j)$ will be within $\frac{\varepsilon}{8m}$ of $F(a_j)$ if we pick $n > N_j$. Since there are finitely many (m, the number of intervals in our partition) N_j's, simply pick the largest one, then we have that n is large enough so that our sum is

$$\leq \sum_{j=1}^{m}\left[\frac{\varepsilon}{8m} + \frac{\varepsilon}{8m}\right] = \sum_{j=1}^{m}\left[\frac{\varepsilon}{4m}\right] = \frac{\varepsilon}{4}$$

So done.

We are going to have to prove this stuff for random vectors as well. This is in preparation for the infamous chi squared tables. A random vector is defined by a probability density function in R^k. We define the cf of a random vector Y to be a function that takes in a vector t and outputs $E\left(e^{i(t.Y)}\right)$. If the space is d dimensional, we define

$$I_\varepsilon := \frac{1}{(2\pi)^d}\int_{R^d}\phi_y(t)e^{-\varepsilon|t|^2}\prod_{j=1}^{d}\frac{e^{-ia_jt_j}-e^{-ib_jt_j}}{it_j}dt$$

Where that big scary looking thing just means product.

$$= \frac{1}{(2\pi)^d}\int_{R^d}\int_{R^d} f(y)e^{i(t.y)}\prod_{j=1}^{d}\frac{e^{-ia_jt_j}-e^{-ib_jt_j}}{it_j}dye^{-\varepsilon|t|^2}dt$$

Note: Each $\frac{e^{-ia_jt_j}-e^{-ib_jt_j}}{it_j}$ is bounded for the same reasons as bsfore, so we have the product of bounded things times a thing which integrates to 1 in the inner dy integral, so that's bounded. Now in total we have a bounded thing times the integral of $e^{-\varepsilon|t|^2}$, which is finite, so we have the conditions to swap the integrals around.

Since $e^{-\varepsilon|t|^2}$is just the product of $e^{-\varepsilon(t_j^2)}$, we can simplify I_ε as follows:

$$= \frac{1}{(2\pi)^d}\int_{R^d} f(y)\int_{R^d} e^{i(t.y)}\prod_{j=1}^{d}\frac{e^{-ia_jt_j}-e^{-ib_jt_j}}{it_j}e^{-\varepsilon|t|^2}dtdy$$

$$= \int_{R^d} f(y)\int_{R^d}\prod_{j=1}^{d}\frac{1}{2\pi}e^{i(t_jy_j)}\frac{e^{-ia_jt_j}-e^{-ib_jt_j}}{it_j}e^{-\varepsilon(t_j)^2}dtdy$$

We know what each term looks like from earlier, so we end up with

$$= \int_{R^d} f(y)\int_{R^d}\prod_{j=1}^{d}\frac{1}{2}\left(\operatorname{erf}\left(\frac{y_j-a_j}{2\sqrt{\varepsilon}}\right)-\operatorname{erf}\left(\frac{y_j-b_j}{2\sqrt{\varepsilon}}\right)\right)dtdy$$

Which in the limit vanishes exactly when we are outside the (a,b) high dimensional rectangle and goes to 1 when we are inside it, for the same reasons – Each term goes to 1 or 0 and the product is 1 only when all terms go to 1.

Now at last we have, because of dominated convergence again, the same result for random vectors.

We also need the result for random vectors that convergence in cf implies convergence in distribution.

To do this, we will define yet another function. Let $L_\varepsilon(t) := \frac{e^{-\varepsilon|t|^2}}{2\pi}\prod_{j=1}^{d}\frac{e^{-ia_jt_j}-e^{-ib_jt_j}}{it_j}$, then $I_\varepsilon = \int_{R^d}\phi(t)L_\varepsilon(t)dt$. With exactly the same proof as the univariate case, all the hypotheses for the dominated convergence theorem apply, so the limit of $\int_{-\infty}^{\infty}\phi_n(t)L_\varepsilon(t)dt$ is indeed I_ε. Therefore, letting ε approach 0, we have that the probability our distribution lands between a and b approaches that probability for a pdf with cf ϕ if the cf's converge to ϕ. So done.

Note on this: A valid probability distribution need not have a well defined pdf everywhere, but we have been assuming this. It may have atoms, which are values a such that P(x=a)>0. Say P(x=3)=0.2 but otherwise x has a normal probability density function g. Then we say that $f(x) = g(x) + 0.2\delta(x-3)$ where $\delta(x)$ is not a function that is defined in the traditional sense, but rather a function that when you integrate it returns 0 if your integration interval does not contain 0 and 1 otherwise. Ie, the following hold:

$$\int_{\mathbb{R}} \delta(x)\, h(x)\, dx = h(0)$$

$$\int_a^b \delta(x)\, dx = \begin{cases} 1, & 0 \in [a, b], \\ 0, & \text{otherwise.} \end{cases}$$

This is called the dirac delta function. You can think of it as a spike, or as the limit of normals with mean 0 and tinier variances approaching 0. The integration theory we have used to get to this proof is still logically sound even with dirac deltas allowed. (See technical detail 2) This also allows us to work with continuous random variables, discrete random variables, or combinations of those in one unified framework. We can avoid issues by making sure we pick our a's and b's so that at the ends of an interval, or at the corners of our rectangle, the cdf is continuous, so we can integrate inside the rectangle and recover the cdf.

Technical detail: This kind of "ensuring continuity" thing which we do a lot here is possible because the cdf is increasing so it can't be discontinuous on an entire interval or it wouldn't be bounded on that interval. This is because of the thing in the technical results document that says we cannot add up uncountably many finite numbers and get a finite number. Since there is no interval where our thing is discontinuous, if we end up at a discontinuity we can just move a really really tiny as tiny as we want amount then be fine.

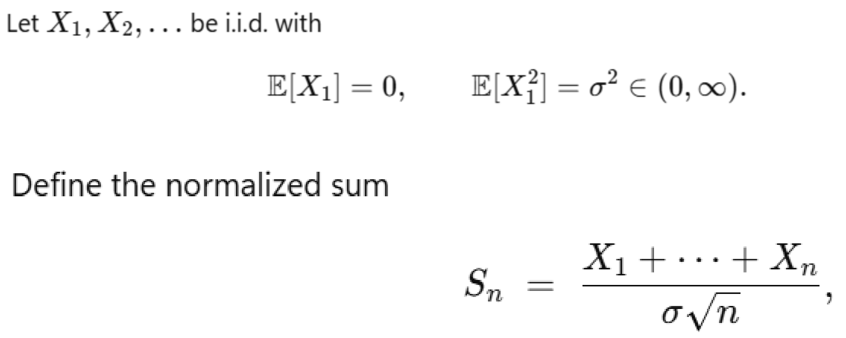
Technical detail 2: The monotone convergence theorem still applies to these "mixed" integrals because the Dirac-delta and continuous parts converge separately, so the usual proof carries over unchanged. Everything built on this theorem therefore remains valid.

- Dirac-delta part: In each step of our increasing sequence of functions, the total mass of the delta spikes is non-decreasing and bounded by 1, hence convergent.
- Continuous part: The continuous (normal-type) component converges in the ordinary way, giving the correct limit.
- Putting them together: It remains to justify that the total contribution from the delta spikes in the limit equals the limit of their partial sums. If there are only finitely many spikes this is immediate. If there are infinitely many, we choose any finite subset whose combined mass is within any ε/2 of the full limit mass. The contributions from these finitely many spikes in the sequence eventually come within ε/2 of their limiting total, and the remaining spikes contribute at most ε/2, so we can eventually get arbitrarily close to the limit we want, ie within ε for any ε we choose.

Therefore the sum of the Dirac-delta parts in the limit agrees with the limit of the sums, so the monotone convergence theorem holds exactly as in the purely continuous case.

Central limit theorem (Normal approximation is a corollary of this)

Setup:



Let $X_1, X_2, \ldots$ be i.i.d. with

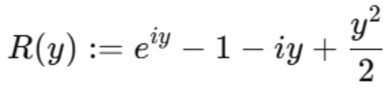
$$\mathbb{E}[X_1] = 0, \qquad \mathbb{E}[X_1^2] = \sigma^2 \in (0, \infty).$$

Define the normalized sum

$$S_n \;=\; \frac{X_1 + \cdots + X_n}{\sigma\sqrt{n}},$$

Note that if our variables do not have mean 0 we can shift them so that they do and still apply this argument. S_n has mean 0 and variance 1 so it is standardized, and the goal is to prove that as n gets large, S_n is well approximated by a standard normal, as this is what the central limit theorem says. The characteristic function of S_n is $E\left(e^{it\left(\frac{X_1}{\sigma\sqrt{n}}\right)+it\left(\frac{X_2}{\sigma\sqrt{n}}\right)+\cdots+it\left(\frac{X_n}{\sigma\sqrt{n}}\right)}\right) = E\left(e^{it\left(\frac{X_1}{\sigma\sqrt{n}}\right)}\right)E\left(e^{it\left(\frac{X_2}{\sigma\sqrt{n}}\right)}\right)\ldots E\left(e^{it\left(\frac{X_n}{\sigma\sqrt{n}}\right)}\right)$ by properties of exponents and the fact that with independent things we can split expectation and product. Therefore, $\phi_{S_n}(t) = \left(\phi_x\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n$. Therefore, we will prove that $\phi_x(u) = 1 - \frac{\sigma^2 u^2}{2} + o(u^2)$ where $o(u^2)$means that as u goes to 0 this gets much smaller than u^2, ie if u is small enough this becomes smaller than an arbitrary constant times u^2. We will then be able to put $u_n = \frac{t}{\sigma\sqrt{n}}$ to get $\phi_x(u_n) = 1 - \frac{t^2}{n} + o(\frac{1}{n})$, where this time $o(\frac{1}{n})$ means that it if n is going to infinity this gets much smaller than $\frac{1}{n}$ by any arbitrarily large constant.

Define for real y



$$R(y) := e^{iy} - 1 - iy + \frac{y^2}{2}$$

$$\int_0^1 (1-s)e^{isy}ds = \left[(1-s)\frac{e^{isy}}{iy}\right]_0^1 + \int_0^1 \frac{e^{isy}}{iy}ds = -\frac{1}{iy} + \left[\frac{e^{isy}}{(iy)^2}\right]_0^1 = -\frac{1}{iy} + \frac{e^{iy}}{(iy)^2} - \frac{1}{(iy)^2}$$

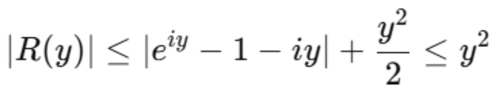
Where I have used integration by parts with u=1-s and v=the exponential thingy

Therefore $e^{iy} - 1 - iy = (iy)^2 \int_0^1 (1-s)e^{isy}ds$

So $|e^{iy} - 1 - iy| = |(iy)^2|| \int_0^1 (1-s)e^{isy}ds\,| \le y^2 \int_0^1 \left|(1-s)e^{isy}\right|ds$

$= y^2 \int_0^1 |1-s|\left|e^{isy}\right|ds = y^2 \int_0^1 |1-s|ds = \frac{y^2}{2}$.

Therefore,



$$|R(y)| \le |e^{iy} - 1 - iy| + \frac{y^2}{2} \le y^2$$

Now notice that $(iy)^2 \int_0^1 (1-s)\left(e^{isy} - 1\right)ds = e^{iy} - 1 - iy - (iy)^2 \int_0^1 (1-s)ds$

$$= e^{iy} - 1 - iy + y^2 \int_0^1 (1-s)ds = e^{iy} - 1 - iy + \frac{y^2}{2} = R(y)$$

So,

$$|R(y)| = \left|(iy)^2\int_0^1 (1-s)\big(e^{isy}-1\big)\,ds\right| \le y^2\int_0^1 (1-s)\,s|y|\,ds = \frac{|y|^3}{6}$$

Now define h for real numbers as follows:

$$h(y) = \begin{cases} R(y)/y^2, & y \neq 0, \\ 0, & y = 0. \end{cases}$$

By our two bounds on R(y), h is always bounded by 1, and since it is bounded by y/6 it goes to 0 as y goes to 0.

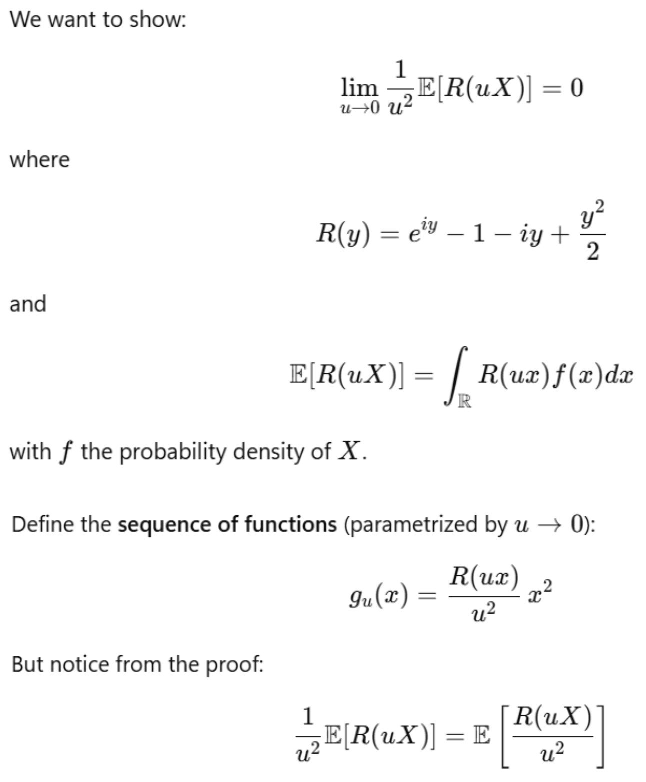
Now, using E(X)=0 and therefore that $E(x^2) = E(x^2) - E(x)^2 = \sigma^2$, we have that

$$\phi(u) = \mathbb{E}[e^{iuX}] = 1 + iu\,\mathbb{E}[X] - \frac{u^2}{2}\,\mathbb{E}[X^2] + \mathbb{E}[R(uX)] = 1 - \frac{\sigma^2u^2}{2} + \mathbb{E}[R(uX)]$$

Now, by definition of h,

$$\frac{\phi(u) - 1 + \frac{\sigma^2u^2}{2}}{u^2} = \mathbb{E}\left[\frac{R(uX)}{u^2}\right] = \mathbb{E}\big[h(uX)\,X^2\big]$$

Therefore if this goes to 0 as u goes to 0, we have $o(u^2)$ so we will be done with finding the cf of x.



We want to show:

$$\lim_{u\to 0}\frac{1}{u^2}\mathbb{E}[R(uX)] = 0$$

where

$$R(y) = e^{iy} - 1 - iy + \frac{y^2}{2}$$

and

$$\mathbb{E}[R(uX)] = \int_{\mathbb{R}} R(ux)f(x)dx$$

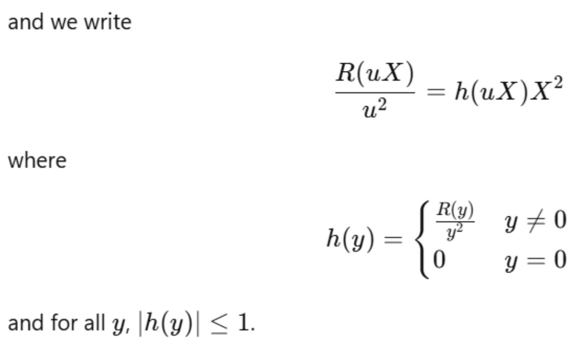
with f the probability density of X.

Define the **sequence of functions** (parametrized by $u \to 0$):

$$g_u(x) = \frac{R(ux)}{u^2}\,x^2$$

But notice from the proof:

$$\frac{1}{u^2}\mathbb{E}[R(uX)] = \mathbb{E}\left[\frac{R(uX)}{u^2}\right]$$



and we write

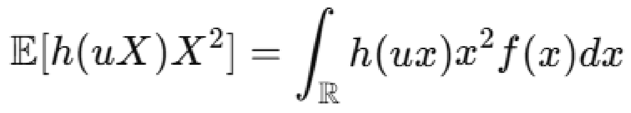
$$\frac{R(uX)}{u^2} = h(uX)X^2$$

where

$$h(y) = \begin{cases} \frac{R(y)}{y^2} & y \neq 0 \\ 0 & y = 0 \end{cases}$$

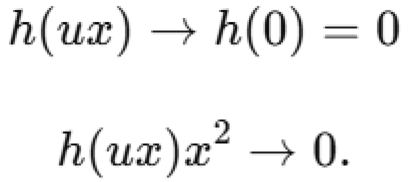
and for all y, $|h(y)| \leq 1$.

Parametrized by u --> 0 essentially means we can define a sequence of functions g_n by setting u=1/n, similar to what we did before. Now, since |h(y)|≤1, we have this



$$\mathbb{E}[h(uX)X^2] = \int_{\mathbb{R}} h(ux)x^2 f(x)dx$$

Where the integrand is bounded above by the function $x^2 f(x)$. This is integrable because Var(x) is finite. This is an assumption of the central limit theorem. If Var(x) is not finite, the central limit theorem is not always true! The cauchy distribution is an example of this. Also, since h(y) goes to 0 as y goes to 0, we can do as follows:



$$h(ux) \to h(0) = 0$$

$$h(ux)x^2 \to 0.$$

So our integral converges pointwise to 0, so all the hypotheses of the dominated convergence theorem are met, so this term that we needed to go to 0 goes to 0, so done.

Ok, now to find the characteristic function of S_n, remember that it is $\left(\phi_x\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n$ so therefore since the characteristic function of x is $1-\frac{t^2}{n}+o(\frac{1}{n})$,the cf of S_n is $\left(1-\frac{t^2}{n}+o(\frac{1}{n})\right)^n$, which by the same limit we used in the poisson distribution proof, goes to $e^{-\frac{t^2}{2}}$ as n goes to infinity. Now it just remains to show that this is indeed the cf of a standard normal, and we will be done at last. The cf of a standard normal is

$$\int_{-\infty}^{\infty} f(y)e^{ity}dy$$

Where f(y) is the formula for a standard normal, so we have

$$\int_{-\infty}^{\infty} \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}} e^{ity}dy = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}e^{ity}dy = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(y^2-2ity)}dy = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}((y-it)^2-t^2)}dy$$

$$= \frac{1}{\sqrt{2\pi}}e^{-t^2}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(y-it)^2}dy = e^{-t^2}$$

Where we have used the same contour integration idea as in the characteristic function proof, so done.

Note: The reason we did not just use the taylor series expansion is because we have not justified swapping expectation with an infinite sum.

What the truck is actually a "degree of freedom"

In the context of chi squared tests, it's the dimension of the space of allowed values. If we estimate the mean, we know that the value we estimated was the sample mean, which can be rewritten as a hyperplane equation that means we have been constrained to a lower dimension.

Note: For estimating mean and variance, it is not as obvious that it is linear, but in fact you can infer exactly the values of $E(X)$ and $E(X^2)$ from your estimates, which are linear constraints/weighted sums of the cell values. If only estimating variance, luckily I've never ever ever ever seen a question on this on an A level textbook or exam, so I have, and you can, get away with not knowing how this would work, I don't think there's a simple explanation. If there is a question like this, it's badly designed.

Chi squared tests

Note: A χ_n^2 distribution is defined as the distribution of the sum of the squares of the values of n standard normals.

We are going to have to define some theory here. Say I have a vector of random variables $X_1, X_2, \dots X_k$. These do not have to be independent nor identically distributed. Then the covariance matrix, typically denoted by Σ is such that the entry in the i'th row and j'th column is equal to $E((X_i - E(X_i))(X_j - E(X_j))$ which we showed at the beginning equals $E(x_i x_j) - E(x_i)E(x_j)$. First, we know that Σ is symmetric. Also, by definition, the n'th diagonal entry of Σ is given by $Var(X_n)$, and the off diagonal entries, say in the i'th row and j'th volumn, are given by $Cov(X_i, X_j)$. Alternatively, you can think of the covariance matrix of a random vector Z as $E\big((Z - E(Z))(Z - E(Z))^T\big)$ as this is equivalent. We will consider the chi squared table to be a random vector.

The distribution with this property (ie that all dot products make normal distributions) with covariance matrix Σ is unique because any such distribution has the same characteristic function. To prove this, note that for a random vector x, $\phi_{a.x}(u) = E\big(e^{iu(a.x)}\big) = \phi_x(ua)$. The characteristic function of a standard normal is $e^{\frac{-t^2}{2}}$ as we showed in the CLT proof, and now I will derive the characteristic function of a normal with mean μ and variance σ^2. We need to evaluate the following integral:

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{\left(\frac{y-\mu}{\sigma}\right)^2}{2}} e^{ity}dy = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{-((y-\sigma^2 it-\mu)^2+\sigma^4t^2+2\mu\sigma^2 it)}{2\sigma^2}}dy$$

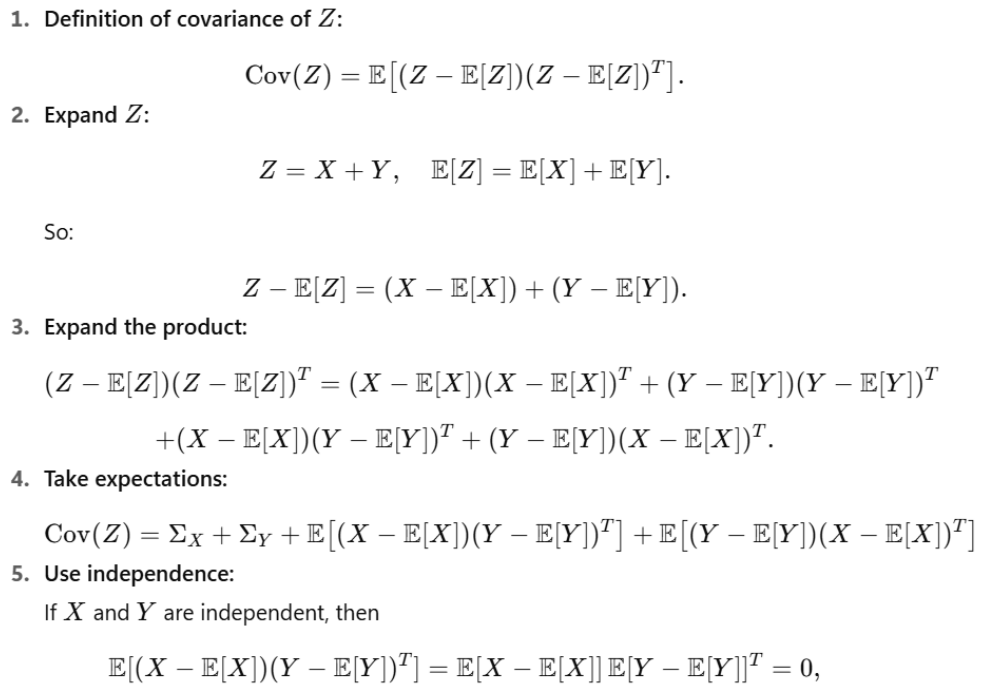
$$= \left(e^{\frac{\sigma^2t^2}{2}+\mu it}\right)\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{-(y-\sigma^2 it-\mu)^2}{2\sigma^2}}dy = \left(e^{\frac{\sigma^2t^2}{2}+\mu it}\right)\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\left(\frac{y}{\sqrt{2}\sigma}-stuff\ independent\ of\ y\right)^2}dy$$

$$= e^{\frac{\sigma^2t^2}{2}+\mu it}$$

But this is $\phi_{a.x}(u)$ if μ and σ^2 are the mean and variance of the variable a.x, so since $\phi_{a.x}(u) = \phi_x(ua)$, this defines the characteristic function of X if X is distributed normally, as required.

Lemma: Covariance matricies of independent vectors ad d like regular variances.

Proof:



1. **Definition of covariance of Z:**

$$\mathrm{Cov}(Z) = \mathbb{E}\left[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^T\right].$$

2. **Expand Z:**

$$Z = X + Y, \quad \mathbb{E}[Z] = \mathbb{E}[X] + \mathbb{E}[Y].$$

So:

$$Z - \mathbb{E}[Z] = (X - \mathbb{E}[X]) + (Y - \mathbb{E}[Y]).$$

3. **Expand the product:**

$$(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^T = (X - \mathbb{E}[X])(X - \mathbb{E}[X])^T + (Y - \mathbb{E}[Y])(Y - \mathbb{E}[Y])^T$$
$$+ (X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T + (Y - \mathbb{E}[Y])(X - \mathbb{E}[X])^T.$$

4. **Take expectations:**

$$\mathrm{Cov}(Z) = \Sigma_X + \Sigma_Y + \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T\right] + \mathbb{E}\left[(Y - \mathbb{E}[Y])(X - \mathbb{E}[X])^T\right]$$

5. **Use independence:**
If X and Y are independent, then

$$\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T] = \mathbb{E}[X - \mathbb{E}[X]]\,\mathbb{E}[Y - \mathbb{E}[Y]]^T = 0,$$

Where this is for the same reason as in the univariate case.

We will now prove a very important lemma (called the multivariate central limit theorem), which says that as n goes to infinity, $\sqrt{n}(\bar{X} - p)$ converges in distribution to a normal distribution with mean 0 and covariance Σ. Note that here, $\bar{X}$ is the component-wise sample mean of n random vectors, and p is the vector with entries equal to the expected values of each of the components.

To prove the multivariate CLT, we will prove a lemma within this lemma which says that for a fixed vector a, if a sequence of random vectors Y_n converges in distribution to a random vector Y, then $a.Y_n$ converges in distribution to $a.Y$, and that if $a.Y_n$ converges in distribution to $a.Y$ for all vectors a, then Y_n converges in distribution to Y. The first part is because a linear combination of the components of the Y_n converges in distribution to the linear combination of the components of Y since Y_n converges to Y. The other direction is the hard part. Suppose we have a vector t, and $Z_n \coloneqq t.Y_n$ and $Z \coloneqq t.Y$. Suppose also that Z_n converges in distribution to Z. Then we will show that Y_n converges in distribution to Y. $\phi_{Y_n}(t) = E\left(e^{i(t.Y_n)}\right) = E\left(e^{i*1*(Z_n)}\right) = \phi_{Z_n}(1)$. Since Z converges in distribution, we have that as n goes to infinity $\phi_{Y_n}(t) = \phi_{Z_n}(1) \to \phi_Z(1) = E\left(e^{iZ}\right) = E\left(e^{i(t.Y)}\right) = \phi_Y(t)$. This is true for every fixed vector t, so since we know that convergence in characteristic functions implies convergence in distribution even for vectors, the result follows.

Now it is time to prove the multivariate CLT. If we have a random vector X, then we first shift X so that its mean is 0, then we have the following

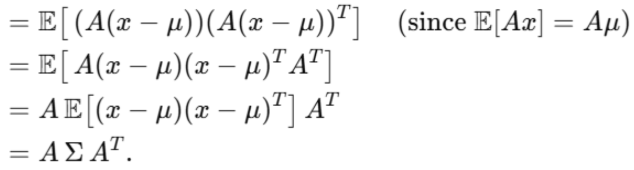
$$Var(a.x) = E((a.x)^2) - \left(E(a.x)\right)^2 = E\left(\left(\sum_{i=1}^k a_i x_i\right)^2\right) - \left(E\left(\sum_{i=1}^k a_i x_i\right)\right)^2$$

$$= E\left(\sum_{i=1}^k \sum_{j=1}^k a_i\, a_j x_i x_j\right) = \sum_{i=1}^k \sum_{j=1}^k a_i\, a_j E(x_i x_j) = \sum_{i=1}^k \sum_{j=1}^k a_i\, a_j \Sigma_{ij}$$

Let's unpack what I mean by Σ_{ij}. Σ is the covariance matrix of X, and this is the i'th row j'th column entry of the covariance matrix, which equals $E(x_i x_j)$ since the $E(x_i)E(x_j)$ part of the covariance matrix terms is 0 by assumption. It is also the case that $\sum_{i=1}^{k}\sum_{j=1}^{k} a_i\, a_j \Sigma_{ij}$ is the exact sum you would get if you compute the product $A^T\Sigma A$, which means $Var(A.x) = A^T\Sigma A$. Therefore, the distribution of $\sqrt{n}(\bar{X} - p)$ is the distribution of $\sqrt{n}(\bar{Y})$ where Y is just X but scaled to have a mean of 0. $\sqrt{n}(a.\bar{Y})$ has a variance of $A^T\Sigma A$ always since the taking the mean and the multiplying by $\sqrt{n}$ cancel the effect of eachother, and by the normal central limit theorem as n increases this approaches a normal distribution. Since this in fact holds always for all vectors a, $\sqrt{n}(\bar{Y})$ converges to a multivariate normal, as a multivariate normal is defined as a distribution that is normally distributed if you take a dot product with that random vector and a fixed vector, and this is unique. So done (with multivariate CLT).

We will also show what I showed above more generally: Suppose A is not just a kx1 vector but a kxr matrix and we want to find the covariance matrix of Ax:

$$Cov(Ax) = E\left(\big(Ax - E(Ax)\big)\big(Ax - E(Ax)\big)^T\right) =$$



$$\begin{aligned}
&= \mathbb{E}\big[\,(A(x-\mu))(A(x-\mu))^T\big] \quad (\text{since } \mathbb{E}[Ax] = A\mu) \\
&= \mathbb{E}\big[\,A(x-\mu)(x-\mu)^T A^T\big] \\
&= A\,\mathbb{E}\big[(x-\mu)(x-\mu)^T\big]\,A^T \\
&= A\,\Sigma\,A^T.
\end{aligned}$$

Notation: k is the number of cells in the chi squared table, and therefore the number of dimensions of the random vector in question. p_k is the probability the k'th cell of this vector is activated on a single trial.

The covariance matrix of this vector X will be denoted by Σ, which in it's i'th row j'th column entry has $E(X_iX_j) - E(X_i)E(X_j)$. If i and j are distinct, then this first term becomes 0 as at least one of X_i and X_j is always 0, so we just have $-p_ip_j$. If i=j, then X_iX_j is 1 with probability p_i, and $E(X_i)E(X_j)$ is p_i^2, so we get $p_i(1-p_i)$ on the diagonal entries. Therefore,

$$\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_2p_1 & \cdots & -p_kp_1 \\ -p_1p_2 & p_2(1-p_2) & & -p_kp_2 \\ & \vdots & \ddots & \vdots \\ -p_1p_k & -p_2p_k & \cdots & p_k(1-p_k) \end{pmatrix}$$

This is also the covariance matrix of $\sqrt{n}(\bar{x} - p)$ because we get there by 1. Adding up everything (multiplies the covariance matrix by n), divide by n to get the mean (divides the covariance matrix by n^2), multiply by $\sqrt{n}$ on the front (multiplies by n again), then subtract a constant vector p (does not do anything), so we're back where we started. Now to prove the chi squared statistic I will prove that geometrically, the vector $\sqrt{n}\frac{\bar{X}-p}{\sqrt{p}}$ (which is the one where squaring the components or equivalently by pythagoras squaring its length gives the chi squared statistic: $n\frac{(\bar{X}-p)^2}{p} = \frac{(n\bar{X}-np)^2}{np} = \frac{(O-E)^2}{E}$) acts approximately as a standard normal in the k-1 dimensional subspace it is constrained to given the constraint that $\bar{X} - p$ has components which sum to 0 because there were n trials and so $\bar{X}$ has components that add up to 1 and p has components adding up to 1 trivially. Now notice that $\sqrt{n}\frac{\bar{X}-p}{\sqrt{p}}$ actually means we divide each component by the square root of that corresponding p, so if we multiply each component by that square root again, then sum the components, we will get 0. Putting it another way, $\sqrt{n}\frac{\bar{X}-p}{\sqrt{p}}$ dot product with the vector with components $\sqrt{p_1}, \sqrt{p_2}, \dots, \sqrt{p_k}$ is 0, so the constraint is that we are perpendicular to $\sqrt{p_1}, \sqrt{p_2}, \dots, \sqrt{p_k}$. Therefore the projection matrix onto the space that $\sqrt{n}\frac{\bar{X}-p}{\sqrt{p}}$ is confined to is given by

$I - \begin{pmatrix}\sqrt{p_1} & \sqrt{p_2} & \dots & \sqrt{p_k}\end{pmatrix}\begin{pmatrix}\sqrt{p_1} \\ \sqrt{p_2} \\ \vdots \\ \sqrt{p_k}\end{pmatrix}$ because that is the standard formula for a projection matrix. The proof for this is in the technical results document.

Now let D be the diagonal matrix with$\sqrt{p_1}, \sqrt{p_2}, \dots, \sqrt{p_k}$ as its entries, then $\sqrt{n}\frac{\bar{X}-p}{\sqrt{p}} = D^{-\frac{1}{2}}\sqrt{n}(\bar{x}-p)$. So, since we know how pre-multiplying by a matrix affects the covariance, and because the transpose of a diagonal matrix equals itself, we have $Cov\left(\sqrt{n}\frac{\bar{X}-p}{\sqrt{p}}\right) = D^{-\frac{1}{2}}(D-pp^T)D^{-\frac{1}{2}} = I - SS^T$, where S is $\begin{pmatrix}\sqrt{p_1} \\ \sqrt{p_2} \\ \vdots \\ \sqrt{p_k}\end{pmatrix}$. This is indeed the projection matrix, and we will show that if the covariance matrix is a projection matrix, then the random vector behaves like a standard normal in the relevant subspace. This vector does indeed become normally distributed as n gets large because of the multivariate central limit theorem. Here is the proof of this: We know that if B is any k*m matrix with orthogonal columns spanning the plane in question and Z is a standard multivariate normal random vector, then $B^T Z \sim N(0, BB^T)$. But notice, BB^T is the projection matrix onto the space spanned by B's columns, and it is the covariance matrix of Z after being projected. So the covariance matrix being a projection matrix makes the vector be the projected vector.

As promised, I will now explain why it would have been ok if we had taken the negative square root. Essentially, everything we took the square root of we eventually took the square of again.

So now we know that the vector $\sqrt{n}\frac{\bar{X}-p}{\sqrt{p}}$, which is the chi squared statistic, is a k-1-dimensional normal, so the square of its length, which is equal to the chi squared is χ^2_{k-1}.

Now this is the part I got famously stuck on (https://sites.google.com/view/giraffelalalandnews-20250712/home) before I had the epiphany to realize what I'm about to write on July 2nd 2025. If we add another linear constraint, such as estimating the mean, we can do a pretty geometric argument. What happens is we are essentially taking a k-1-dimensional standard normal distribution and constraining it to k-2 dimensions, so the square of the distance of a roll of this normal from the mean, which by pythagoras is the sum of the squares, becomes the square of the distance of a roll in a lower dimensional normal, so it is now a chi squared distribution with one less degree of freedom. If we constrain the row and column totals in an r*c contingency table, then we get r-1 constraints for the first r-1 rows, since once we have constrained those the last one is already constrained. Similarly for columns: we get c-1 constraints. So our degrees of freedom is rc-1-(r-1)-(c-1) which is equal to (r-1)(c-1).

Technical justification on the argument above given in an appendix.

As for why we combine cells with expected values under 5, it's just a rule of thumb, not really a rigorous statement with anything special about 5, at least as far as I know. Essentially the reason is that for the central limit theorem to work, it has to be possible for us to get tails in both directions. If the expected value was, say 1, then the distribution would not possibly be able to be symmetrical without going into the negative values. Also, if p is small, the cell values become well approximated by a poisson distribution, and by the central limit theorem, the poisson distribution is better approximated by a normal the larger the mean is. As for why 5? You can see from this picture that is about when the poisson distribution starts to look like a normal curve.

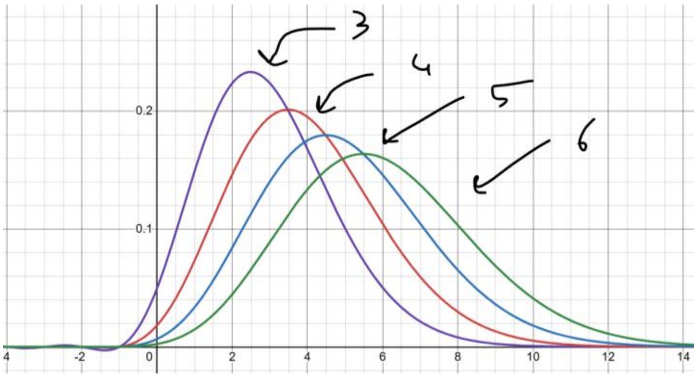
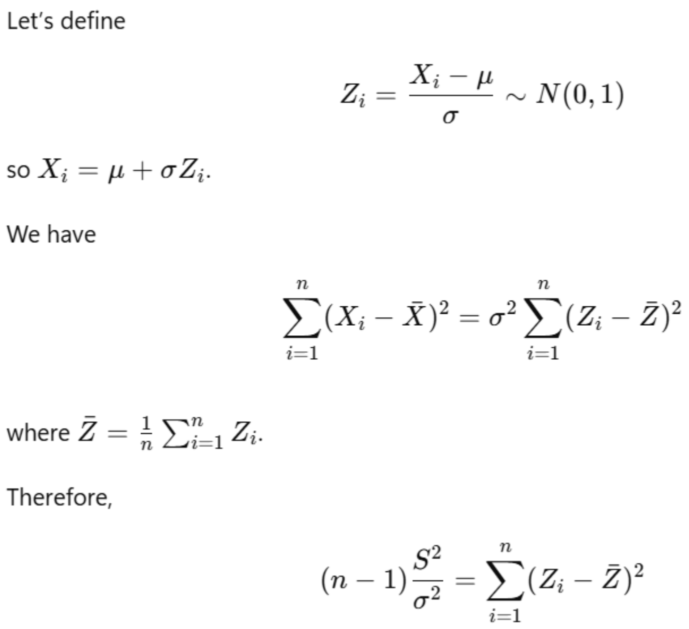


Image: Po(3, 4, 5, 6) on the same graph, demonstrating that around 5 is when it starts to look like a normal distribution.

The formula $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$

This follows from the definition of S^2, σ^2 and χ^2_{n-1}. The result is clear if you were doing this with a N(0,1) variable, then observe that $\frac{S^2}{\sigma^2}$ does not change when we scale or move the variable since s scales with σ and neither change by shifting. Here is why it works for an N(0,1) variable: If I have a bunch of independently distributed random variables, then we can use the same trick of using rotational symmetry of the normal distribution.



Let's define

$$Z_i = \frac{X_i - \mu}{\sigma} \sim N(0,1)$$

so $X_i = \mu + \sigma Z_i$.

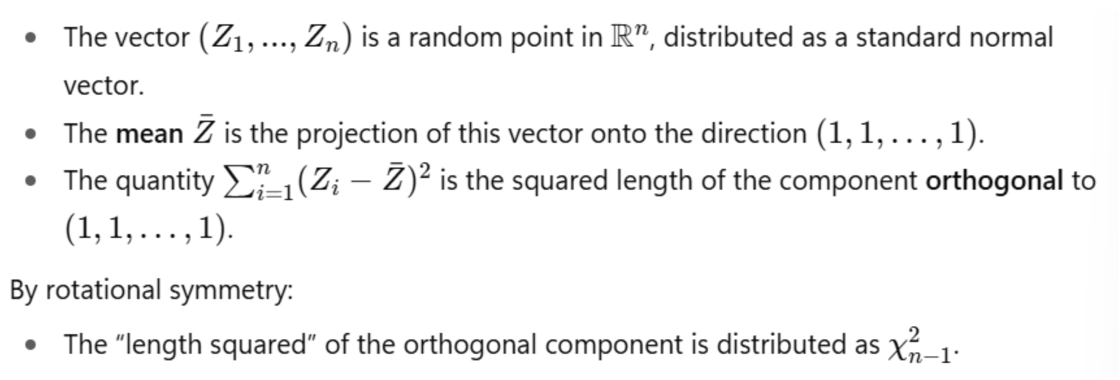
We have

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sigma^2 \sum_{i=1}^{n}(Z_i - \bar{Z})^2$$

where $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$.

Therefore,

$$(n-1)\frac{S^2}{\sigma^2} = \sum_{i=1}^{n}(Z_i - \bar{Z})^2$$



- The vector $(Z_1, \ldots, Z_n)$ is a random point in $\mathbb{R}^n$, distributed as a standard normal vector.
- The **mean** $\bar{Z}$ is the projection of this vector onto the direction $(1, 1, \ldots, 1)$.
- The quantity $\sum_{i=1}^{n}(Z_i - \bar{Z})^2$ is the squared length of the component **orthogonal** to $(1, 1, \ldots, 1)$.

By rotational symmetry:

- The "length squared" of the orthogonal component is distributed as χ^2_{n-1}.

This is because a standard normal confined to the plane such that the sample mean is what it is is just a rotation of a normal with one fewer variable in a normal plane, and in that case chi squared is clearly the distance squared from the origin of that multivariate normal.

t test for difference between means

Note: I'm not really sure how the degrees of freedom relate to the dimension test, so I will merely prove the mathematical properties required.

Note that a t_{n-1} distribution is given by $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ in a sample of size n normally distributed variables, where scaling them will affect the numerator and the denominator in the same way and shifting them will not affect the numerator. In the two-sample t-test, we define the following, assuming X and Y have equal variance:

$$T = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}\, S_p}, \qquad S_p^2 = \frac{\sum_{i=1}^{n_X}(X_i - \bar{X})^2 + \sum_{j=1}^{n_Y}(Y_j - \bar{Y})^2}{n_X + n_Y - 2},$$

Note that S_p^2 is an unbiased estimate for the variance since it is a linear combination of unbiased estimates for Var(X) and Var(Y). We ultimately need to show that T has the same distribution as a $t_{n_x+n_y-2}$ variable. Note that we have the following:

$$Z := \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \quad \text{is } N(0,1)$$

The reason is because the variance of $n_x\bar{X}$, which is the sum of the elements of X, by additivity of variances is the sum of all the variances of the elements of x which is $n_x\sigma^2$. Therefore, since shifting does not affect the variance this is also the variance of $n_x(\bar{X} - \mu)$. Since multiplying a variable by the square of a constant multiplies the variance by that constant, we get that that the variance of $\bar{X} - \mu$ must be $\frac{\sigma^2}{nx}$. Similarly, we have that the variance of $\bar{Y} - \mu$ is $\frac{\sigma^2}{ny}$. Therefore, we have that the expression above is standardized correctly since the variance of the numerator is $\frac{\sigma^2}{nx} + \frac{\sigma^2}{ny}$. We also have the following:

$$U := \frac{(n_X + n_Y - 2)\, S_p^2}{\sigma^2} \quad \text{is } \chi^2_{n_X+n_Y-2}$$

For the same reason as the similar formula we proved above: We simply have to constrain the means of X and Y which reduces the dimension by 2.

Now we have that

$T = \frac{Z}{\sqrt{\frac{U}{n_x+n_y-2}}}$ by definition of T. However, this is equal to $\frac{N(0,1)}{\frac{\sqrt{\chi^2_{n_x+n_y-2}}}{\sqrt{n_x+n_y-2}}}$ which is equivalent to a $t_{n_x+n_y-2}$ variable in the traditional sense using the relation between S^2 and χ^2.

Other stuff about t and f distributions do not require mathematics since the tables are based on numerical data with no other implicit assumptions other than invariance under shifting and scaling.

Appendix: Chi squared projection argument technical justification

Note on the argument above: Even if there were some pathological affine subspace on which the standardized residuals failed to converge to the same limit distribution as elsewhere (while the overall CLT still held), such a subspace would have probability 0 under the limiting Gaussian, so the theory holds almost surely. Crucially, we also would not have weird behavior like alternating distributions on hyperplanes that somehow average out.

Justification for the above claim that we would not have such behavior: This is because the rate of change of the probability of a configuration as I move trials between cells is a tinier fraction of the probability as n gets larger. The reason is because lets suppose I am moving an item from cell i to cell j. The values in the rest of the cells are fixed in this situation, so we will compute the ratio between the probabilities by computing the ratio between the probabilities given that the rest of the cells are in whatever configuration they are in, then multiplying by the probability of that configuration at the end does not matter, since ultimately we are computing a ratio. So we are essentially finding the ratio in a binomial context.

$P(before) = \frac{(O_i+O_j)!}{(O_i)!(O_j)!}\frac{{p_i}^{O_i}}{(p_i+p_j)^{O_i}}\frac{{p_j}^{O_j}}{(p_i+p_j)^{O_j}}$ where the two fractions on the right is because we are working with probabilities like (i'th cell given i'th cell or j'th cell).

$$P(after) = \frac{(O_i+O_j)!}{(O_i-1)!\,(O_j+1)!}\frac{{p_i}^{O_i-1}}{(p_i+p_j)^{O_i-1}}\frac{{p_j}^{O_j+1}}{(p_i+p_j)^{O_j+1}}$$

The ratio between these is $\frac{O_i}{O_j+1}\frac{p_j}{p_i}$

For any function that grows faster than $\sqrt{n}$ but slower than n, say $n^{3/4}$, the probability that O_i is out from E_i by more than $n^{3/4}$ goes to 0 as n gets large as that is on the order of $n^{1/4}$standard deviations (Since Var(O_i)=n*c where c is a constant equal to the variance of O_i in a 1 trial case, so the standard deviation is on the order of $n^{1/2}$ by definition). Therefore, almost surely, the ratio is $\frac{np_i+o(n^{3/4})}{np_j+o(n^{3/4})}\frac{p_j}{p_i} = \frac{n+o(n^{3/4})}{n+o(n^{3/4})} = 1 + o(n^{-1/4}) \to 1$ as n gets large. O is shorthand for "on the order of". However, we still need what feels like the 98273497th technical justification in this proof's chain of dependencies: What if for some reason we had on the order of $n^{1/4}$ oscillations within the band that our plane is likely to fall in? This would still satisfy the limiting distribution and the bound we have, which is a problem. Hence, suppose that we have a chain of nearby planes obtained by moving one cell over. If we had a chain formed by moving multiple cells over, the ratio in question still satisfies being within $1 + o(n^{-1/4})$. This is because the product of many such ratios is still $o(n^{-1/4})$. So now, consider the fact that the probability ratio is given by the product of stuff of the form $\frac{o_i}{o_j+1}\frac{p_j}{p_i}$, where each of these when we factor the p's become of the form $\frac{n+o(n^{3/4})}{n+o(n^{3/4})}$, and the point is that here, since we have only multiplied by constants, we ultimately change the numerator and the denominator by constants as we move between more parallel planes, so the rate of change of this grows like O(1/n) (since the denominator and numerator change each makes it change like O(1/n)). So, as we move between $O(n^{3/4})$ planes (which is how many are in our "almost certain" band), the ratio changes by $O(n^{-1/4})$. The ratio would need to change by a non-vanishing amount for us to have this weird oscillation behavior we're worried about. There can't be oscillations, so if the probability density were off somewhere in our plane, they would be off there in all parallel planes, and this would carry over to the limiting distributin, contradicting the CLT. So done.