Contents

1 Variances **1**

2 The normal distribution **2**
2.1 Density and variance 2
2.2 Sums of normals 3

3 Correlation and regression **5**
3.1 Residual sums of squares 5
3.2 Correlation coefficients 6

4 Derivations of standard results **8**

5 Probability generating function properties **10**

6 Unbiased variance estimator, vector approach **11**

7 The central limit theorem **11**
7.1 Characteristic functions 11
7.2 Proof of CLT 15

8 Chi squared tests **18**

9 Other properties of the chi squared and t distributions **30**

1 Variances

Theorem. Let X and Y be independent random variables with well defined variance, then $Var[X]+Var[Y] = Var[X+Y]$

Proof.

$$\begin{aligned} Var\left(X+Y\right) &= \left(E(X+Y)^2\right) - E\left((X+Y)^2\right) \\ &= \left(E\left(X\right)+E\left(Y\right)\right)^2 - E\left(X^2+2XY+Y^2\right) \\ E(X)^2 + E(Y)^2 + 2E\left(X\right)E\left(Y\right) &- E\left(X^2\right) - 2E\left(XY\right) - E(Y^2) \\ &= Var\left(X\right)+Var\left(Y\right)+2(E\left(X\right)E\left(Y\right) - \ E\left(XY\right)) \end{aligned}$$

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

$$E(\text{constant}(Y - E(Y)))$$

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 (for independent variables)

$$E(X)E(Y) = E(XY)$$

□

2 The normal distribution

2.1 Density and variance

Theorem. The function

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

is a valid probability density function with mean μ and variance σ

Proof. 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:

Figure 1 shows a 2 dimensional normal plotted on 3D graph

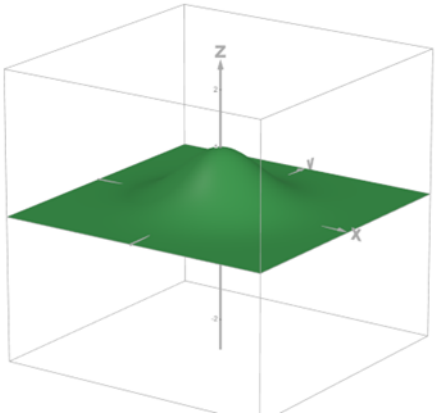

Figure 1

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 $(\int_{x=-\infty}^{\infty} e^{-(y^2+x^2)}dx)dy = (e^{-y^2}\int_{-\infty}^{\infty} e^{-x^2}dx)dy = (e^{-y^2}k)dy$,

Figure 2 shows a thin slice of the 2d normal in the above image

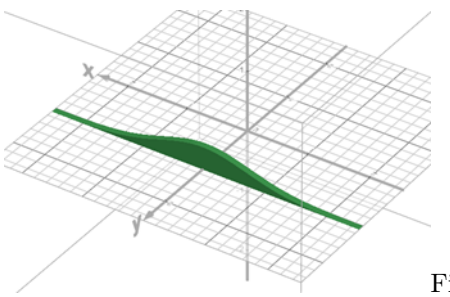

Figure 2

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:

Figure 3 shows a 2d normal on 3d graph approximated by concentric rings of the right height

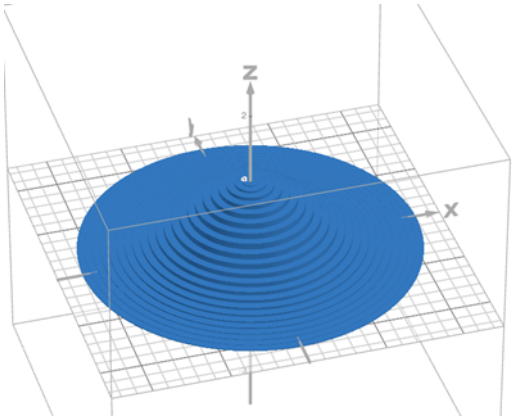

Figure 3

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 $2\pi r$ so the volume of each ring is $2\pi r e^{-r^2} dr$. So, the total volume under the figure is $\int_0^\infty 2\pi r e^{-r^2} \mathrm{dr}$, again by the interpretation of integration as a sum. Note that you can check by differentiating that an antiderivative of $2\pi r e^{-r^2}$ is $-\pi e^{-r^2}$. So, we have to evaluate $\left[-\pi e^{-r^2}\right]_0^\infty$ which is $0 - [-\pi]$ 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 $\frac{1}{\sqrt{2\pi}} \left[\left[x e^{-\frac{x^2}{2}} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}} dx \right]$ (the – in the integration by parts formula cancels with the – from differentiating the exponential). Now, since $x e^{-\frac{x^2}{2}}$ 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.

□

2.2 Sums of normals

Theorem. The sum of two normal distributions is a normal distribution. This is "evidence" for the central limit theorem in the sense that if there is a universal limiting distribution it should have this property.

Proof. 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:

Figure 4 shows another 2d normal plotted on a 3d graph

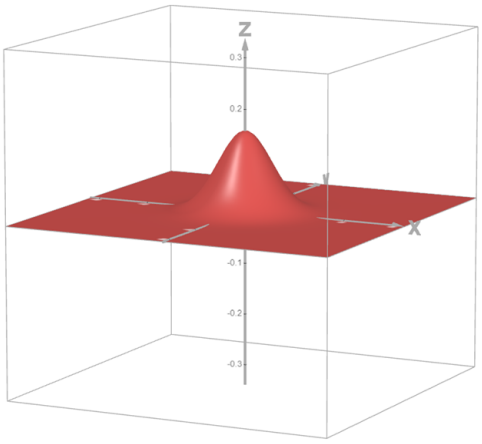

Figure 4

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.

Figure 5 shows the same 2d normal with parallel vertical planes on the same graph

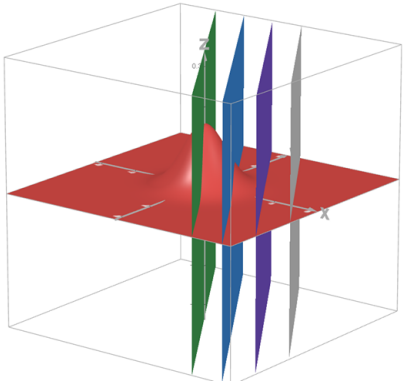

Figure 5

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.

□

3 Correlation and regression

3.1 Residual sums of squares

Theorem. Given a finite data set, and a linear regression line $y = a + bx$, the parameters minimizing the sum of squared distances are given by

$$b = \frac{s_{xy}}{s_{xx}} = \frac{\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum \left[\left(X_i - \bar{X}\right)^2\right]}$$

and

$$a = \bar{y} - b\bar{x}$$

provided $s_{xx} \neq 0$, ie the points do not lie all on a vertical line

Proof. The sum of squares as a function of a and b is defined as $RSS(a,b) := \sum_{i=1}^{n} (y_i - a - bx_i)^2$. Now differentiate with respect to a and b, since we want to find the minimum.

It is clear that the minimum exists and will be a stationary point where the derivatives with respect to a and b are both 0. Both derivatives are clearly positive when you differentiate twice so a stationary point will correspond to a minimum. Here are the derivatives (differentiating wrt one variable implies the other is held constant):

$$\frac{dRSS(a,b)}{da} = \sum_{i=1}^{n} 2(y_i - a - bx_i)(-1) = -2\sum_{i=1}^{n}(y_i - a - bx_i)$$

$$\frac{dRSS(a,b)}{db} = \sum_{i=1}^{n} 2(y_i - a - bx_i)(-x_i) = -2\sum_{i=1}^{n}(x_i y_i - ax_i - bx_i^2)$$

If these are both 0 then

$$0 = \sum_{i=1}^{n}(y_i - a - bx_i) = \sum_{i=1}^{n}(x_i y_i - ax_i - bx_i^2)$$

$$a * n + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

since the a derivative is 0, and

$$b\sum_{i=1}^{n} x_i^2 + a\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \tag{1}$$

since the b derivative is 0. From the first equation, it follows that

$$a = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{b}{n}\sum_{i=1}^{n} x_i = \bar{y} - b\bar{x}$$

Therefore we just need to prove the value of b. We'll do this using equation (1) and the fact that $a = \bar{y} - b\bar{x}$.

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

$$b\left(\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i$$

Therefore,

$$b = \frac{\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - 2n\bar{x}\bar{x} + n\bar{x}^2}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i^2}$$

$$= \frac{\sum_{i=1}^{n} (x_i y_i - \bar{y}x_i - \bar{x}y_i + \bar{x}\bar{y})}{\sum_{i=1}^{n} (x_i^2 - 2\bar{x}x_i + x_i^2)}$$

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

As required.

□

Proposition. The residual sum of squares from the regression line above is given by $S_{yy}(1 - r^2)$ where r is the PMCC

Proof. Using the formula we proved above, we get

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

$$= \sum_{i=1}^{n} (b(\bar{x} - x_i) - (\bar{y} - y_i))^2 = b^2 S_{xx} - 2bS_{xy} + S_{yy} = S_{yy} - bS_{xy}$$

The last equality is by the form we proved of b.

$$= S_{yy}\left(1 - \frac{S_{xy}^2}{S_{xx}S_{yy}}\right) = S_{yy}(1 - r^2)$$

Again since $b = \frac{S_{xy}}{S_{xx}}$.

□

3.2 Correlation coefficients

Theorem. The product moment correlation coefficient is always between -1 and 1.

Proof. I love this one because it was one of the first time I saw different areas of maths come together.

Let $\overrightarrow{a} = (x_1 - \bar{x}, x_2 - \bar{x}, ..., x_n - \bar{x})$ and $\overrightarrow{b} = (y_1 - \bar{y}, y_2 - \bar{y}, ..., y_n - \bar{y})$. Then the formula for the PMCC is just $\frac{a.b}{|a||b|}$. But $a.b = |a||b|cos(\theta)$ so we get $cos(\theta)$ which is clearly between -1 and 1.

□

Corollary. PMCC actually measures correlation.

Proof. 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.

□

Proposition. If we have ranked data x and y then the PMCC between the ranks of x and y is given by $1-\frac{6\sum d^2}{n(n^2-1)}$ where d is the differences between the ranks.

Proof. x_i and y_i are integers 1 to n in any order. Note that

- $r=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ (This is what we want to find).
- $\bar{y}=\frac{n+1}{2}$
- $\bar{x}=\frac{n+1}{2}$

Now

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

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

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

$$=n\left(n+1\right)\left[\frac{1}{6}\left(2n+1\right)-\frac{n+1}{2}+\frac{n+1}{4}\right]=n\left(n+1\right)\left[\frac{1}{12}n-\frac{1}{12}\right]=\frac{1}{12}n\left(n^2-1\right)$$

Note that since x_i and y_i are the same but just reordered, it follows that

$$S_{xx}=\sum\left(X_i-\bar{X}\right)^2=\sum\left(Y_i-\bar{Y}\right)^2=S_{yy}$$

Therefore $\sqrt{S_{xx}S_{yy}}=\frac{1}{12}n\left(n^2-1\right)$. Therefore $r=\frac{12S_{xy}}{n(n^2-1)}$. Now,

$$S_{xy}=\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)=\sum_{i=1}^{n}\left(X_i-\frac{n+1}{2}\right)\left(Y_i-\frac{n+1}{2}\right)$$

$$=\sum_{i=1}^{n}\left[X_iY_i-\frac{n+1}{2}X_i-\frac{n+1}{2}Y_i+\frac{\left(n+1\right)^2}{4}\right]$$

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

Note that $\sum_{i=1}^{n}X_i$ and $\sum_{i=1}^{n}Y_i$ are both $\frac{n(n+1)}{2}$ and so

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

$$=\sum_{i=1}^{n}X_iY_i-\frac{1}{4}n^3-\frac{1}{2}n^2-\frac{1}{4}n$$

Note that $\sum_{i=1}^{n}X_i^2=\sum_{i=1}^{n}Y_i^2=\frac{1}{6}n\left(n+1\right)\left(2n+1\right)=\frac{1}{3}n^3+\frac{1}{2}n^2+\frac{1}{6}n$ so the S_{xy} expression can be rewritten as

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

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

But now

$$r=\frac{12S_{xy}}{n\left(n^2-1\right)}=\frac{12\left(\frac{1}{12}n\left(n^2-1\right)-\frac{1}{2}\sum_{i=1}^{n}\left(X_i-Y_i\right)^2\right)}{n\left(n^2-1\right)}=1-\frac{6\sum_{i=1}^{n}d^2}{n\left(n^2-1\right)}$$

□

4 Derivations of standard results

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

Proof. $\ln(\lim_{x\to\infty}\left(1+\frac{a}{x}\right)^{bx})=\lim_{x\to\infty}\left(\ln((1+\frac{a}{x})^{bx})\right)$ (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

$$\lim_{x\to\infty}\left(bxln\left(1+\frac{a}{x}\right)\right)=b\lim_{\frac{1}{x}\to 0}\left(\frac{ln(1+\frac{a}{x})}{\frac{1}{x}}\right)=ab\lim_{\frac{1}{x}\to 0}\left(\frac{ln(1+\frac{a}{x})}{\frac{a}{x}}\right)=ab\lim_{\frac{1}{x}\to 0}\left(\frac{ln\left(1+\frac{a}{x}\right)-ln(1)}{\frac{a}{x}}\right)$$

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.

□

Proposition. The poisson distribution (as we understand it intuitively) satisfies $P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}$ where λ is the mean.

Proof. Consider 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 stretches of time but the same mean time per event, in the sense of

$$X\sim B\left(\frac{n}{h},ph\right)$$

where $h\to 0$ and $np=\lambda$. We'll consider h to be the reciprocal of an integer just to make this make sense or whatever.

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

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

Because as h gets small, $\frac{\left(\frac{n}{h}\right)!}{\left(\frac{n}{h}-k\right)!}=\left(\frac{n}{h}\right)\left(\frac{n}{h}-1\right)...\left(\frac{n}{h}-k+1\right)\approx\left(\frac{n}{h}\right)^k$ where the ratio between $\frac{\left(\frac{n}{h}\right)!}{\left(\frac{n}{h}-k\right)!}$ and $\left(\frac{n}{h}\right)^k$ approaches 1. Now write the probability as

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

Note that $\left(1-ph\right)^{-k}$ goes to 1 as h goes to 0 since p and k are constant, so the probability is

$$\frac{\lambda^k}{k!}\lim_{h\to 0}\left(1-ph\right)^{\frac{n}{h}}$$

So by the lemma, 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.

Corollary. The variance of a poisson distribution is λ

Proof. In the binomial setting above the variance is $np(1-ph)\to np=\lambda$ (By the known variance of the binomial distribution).

□

Proposition. A geometric distribution has mean $\frac{1}{p}$ and variance $\frac{1-p}{p^2}$. For negative binomial mean and variance we just multiply by n since the negative binomial is the sum of n geometric distributions.

Proof. 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:

1

$+(1-p)+(1-p)$

$+(1-p)^2+(1-p)^2+(1-p)^2$

$+(1-p)^3+(1-p)^3+(1-p)^3+(1-p)^3$

$+$

$\vdots$

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 level 4. 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 (by the geometric series formula)

$$p\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}(1-p)^n(1-p)^m = p\sum_{n=0}^{\infty}\frac{(1-p)^n}{1-(1-p)} = p\sum_{n=0}^{\infty}\frac{(1-p)^n}{p} = \sum_{n=0}^{\infty}(1-p)^n = \frac{1}{1-(1-p)} = \frac{1}{p}$$

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:

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

Split $E[X^2]$ into $E\left[X(X-1)\right]$+$E[X]$. To determine $E\left[X(X-1)\right]$ we will calculate the value of the series $\sum_{k=1}^{\infty} k\left(k-1\right)p\left(1-p\right)^{k-1}$.

Now we substitute $q = 1-p$.

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

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

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

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

We could differentiate because we were inside the radius of convergence because the ratio between consecutive terms approaches 1-p which is between 0 and 1. Continuing the derivation,

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

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

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

□

5 Probability generating function properties

Proposition. Generating functions multiply when variables add

Proof. 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 $x = 2, y = 2$ or $x = 3, y = 1$. 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.

□

Corollary. Since the negative binomial is the sum of geometric distributions we get the generating function for that for free - it's just the pgf of the geometric distribution to the power of n.

Proposition. If a random variable X has probability generating function $g(t)$ then $aX + b$ has probability generating function $t^b g(t^a)$.

Proof. 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.

□

Proposition. $\lim_{z\to 1^-} p'(z) = E[x]$ when E[x] is finite.

Proof. $p'(z) = \sum r p_r z^{r-1}$

This is increasing and bounded by E[x].

Let $\varepsilon > 0$. Then there is N large enuogh such that $\sum_{r=1}^{n} r p_r z^{r-1} \geq E[x] - \varepsilon$ since the infinite sums are the limit of the partial/finite sums and this is how limits are defined.

Now $\lim_{z\to 1^-} p'(z) \geq \lim_{z\to 1^-} \sum_{r=1}^{n} r p_r z^{r-1} \geq E[x] - \varepsilon$

The result follows since ε can be as small as we like.

If E[x] is infinite, then it is the same proof, we say that for each M large enough, $\sum_{r=1}^{n} r p_r z^{r-1} \geq M$ and we do the same limiting argument.

With the same proof, $\lim_{z\to 1^-} p''(z) = E[x(x-1)]$.

□

6 Unbiased variance estimator, vector approach

Figure 6 shows the geometric intuition for the unbiased variance estimator (ie, $\frac{n}{n-1}$ times the sample variance) using the pythagorean theorem.

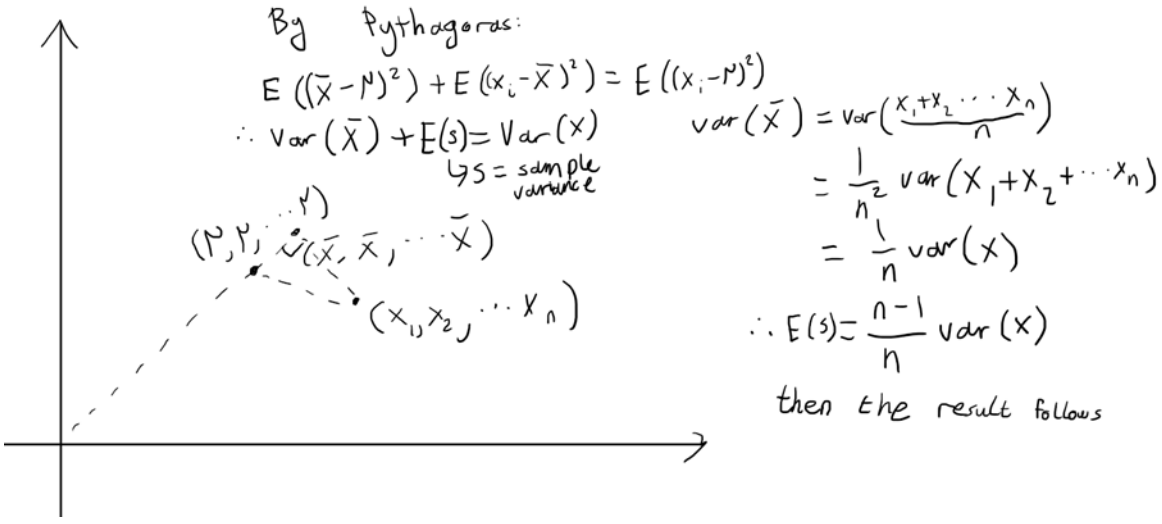

Figure 6

7 The central limit theorem

7.1 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 of the CLT. 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, the fourier transform is just the characteristic function 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 be familiar with level 6.2. The idea will be to show that from the characteristic function there is a formula that lets you 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 density function (We will generalize this shortly), then

$\phi_f(t) = E\left(e^{itx}\right) = \int_{-\infty}^{\infty} f(y)\, e^{ity} dy$ is the characteristic function.

And define, for arbitrary a and b with a<b and the cdf being continuous at a and b (the last constraint is ok by technical results since the set of discontinuities is countable and thus we can dodge them by moving as small an amount away as we like),

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

The first problem we run into is that not all distributions have a probability density function f, however there is a clean way to solve that problem. I will first mention a possible solution that does not work but I will mention it anyway because it introduces important concepts that will be used in future levels, then I will mention the real solution.

The problem is a distribution 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, or the distribution is discrete. Then we say (in the first case) 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:

$\int_{\mathbb{R}} \delta(x) h(x) dx = h(0)$ where this notation means integrating over all real numbers, ie minus infinity to infinity. Also, $\int_a^b \delta(x) dx = \begin{cases} 1 & 0 \in [a, b] \\ 0 & \text{otherwise} \end{cases}$

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.

However, this is not enough, as not all distributions are part discrete part continuous. The solution is as follows: We basically noticed that in the technical results document when we were building integration theory we talked about "length", however we can define length in a non-traditional way, ie we can say that the distance between 2 points on the number line is the probability that one instance of our distribution lies between those 2 points. Then we call this length system (or measure) μ.

Why this is necessary: We could say we just want to prove the CLT for part discrete part continuous distributions and then stick to f(x) notation. For 4 months, that is what this document did. However, this is wrong because it implicitly assumes the limiting distribution is of this type. However, this is no longer an issue as finite variance (our only condition currently) is preserved under limiting distribution.

Example:

$$\int_a^b d\mu = P(a < x < b)$$

$$\int_{-\infty}^{\infty} x d\mu = E(x)$$

We can think of $d\mu$ as $f(x)dx$ where $f(x)$ is a density, it's just that a density may not always exist, but this new length system can be thought of as rescaling the number line by the density so that it ends up having length 1 and our distribution is uniform on it.

$$\int_{-\infty}^{\infty} x^2 d\mu - E(x)^2 = Var(x)$$

$$\phi_f(t) = E\left(e^{ity}\right) = \int_{-\infty}^{\infty} e^{ity} d\mu$$

We will now use this notation.

Now, we come to the hardest theorem yet on this website, a title which will be dethroned in the next section of this very same document.

Theorem. The characteristic function determines the distribution.

Proof. The idea is to try to find a way to determine $P(a < X < b)$, and the dominated convergence theorem is a tool that lets us do this using a limit.

Recall the definition of I_ε from earlier. Now we can say that

$$I_\varepsilon = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} e^{ity}d\mu\right]e^{-\varepsilon t^2}\frac{e^{-iat}-e^{-ibt}}{it}dt$$

Now we will show that

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left|e^{ity}e^{-\varepsilon t^2}\frac{e^{-iat}-e^{-ibt}}{it}\right|d\mu dt$$

is finite so that we can swap the integrals around. Note that for each fixed t,

$$\int_{-\infty}^{\infty}\left|e^{ity}e^{-\varepsilon t^2}\frac{e^{-iat}-e^{-ibt}}{it}\right|d\mu = \int_{-\infty}^{\infty}\left|e^{ity}\right|\left|e^{-\varepsilon t^2}\right|\left|\frac{e^{-iat}-e^{-ibt}}{it}\right|d\mu$$

$$= \left|e^{-\varepsilon t^2}\right|\left|\frac{e^{-iat}-e^{-ibt}}{it}\right|\int_{-\infty}^{\infty}\left|e^{ity}\right|d\mu \leq \left|e^{-\varepsilon t^2}\right|\left|\frac{e^{-iat}-e^{-ibt}}{it}\right|\left|\int_{-\infty}^{\infty}e^{ity}d\mu\right|$$

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,

$$\int_{-\infty}^{\infty}\left|e^{ity}e^{-\varepsilon t^2}\frac{e^{-iat}-e^{-ibt}}{it}\right|d\mu \leq \left|e^{-\varepsilon t^2}\right|\left|\frac{e^{-iat}-e^{-ibt}}{it}\right|$$

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 simplifies 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 (Figure 7):

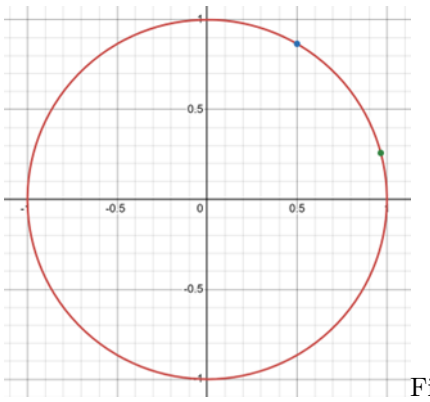

Figure 7

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:

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

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:

$I_\varepsilon = \int_{-\infty}^{\infty} K_\varepsilon(y)\, d\mu$ where $K_\varepsilon(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\varepsilon t^2} \frac{e^{it(y-a)}-e^{it(y-b)}}{it} dt$

Definition: $\operatorname{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$. 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:

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

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

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

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

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:

$$\int_0^u \left(\int_{-\infty}^{\infty} e^{-\varepsilon t^2} e^{ist} dt \right) ds$$

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

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

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

Figure 8 shows a contour that is the reals shifted by an imaginary constant for a large interval

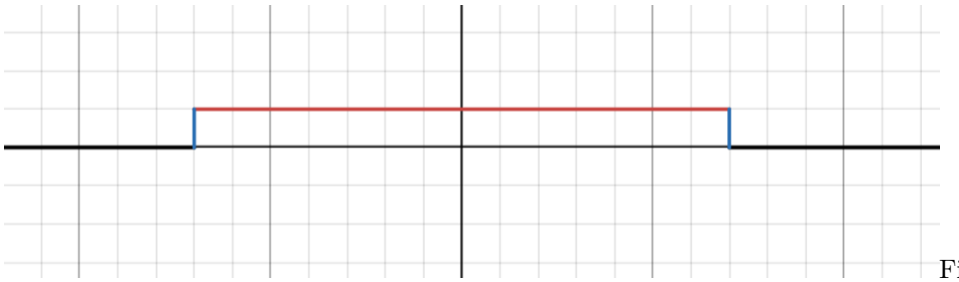

Figure 8

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:

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

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 $\frac{1}{2}$(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 $\frac{1}{2}$ 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:

$I_\varepsilon=\int_{-\infty}^{\infty}K_\varepsilon(y)\,d\mu$ where $K_\varepsilon(y)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-\varepsilon t^2}\frac{e^{it(y-a)}-e^{it(y-b)}}{it}dt$

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 $f_n(y):=K_{\frac{1}{n}}(y)$, then $f_n(y)$ is bounded above in absolute value by 1 which has a finite integral, so dominated convergence applies. Since $f_n(y)$ converges pointwise to 1 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 $\int_a^b d\mu$. But I_ε was defined only in terms of the characteristic function and we just determined $\int_a^b d\mu$ where a and b were chosen arbitrarily.

□

Theorem. If the characteristic function of a sequence of probability distributions converges to a function which is the characteristic function of some distribution, then the distributions converge to the probability distribution with the limit characteristic function.

Proof. 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.

□

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.

7.2 Proof of CLT

We will prove this for distributions that have a probability density function that have finite variance. The cauchy distribution is an example of something that does not work due to not having finite variance – In fact for the cauchy distribution the CLT is not even true.

Setup: Let $X_1, X_2, \ldots$ be identically distributed and independent where each has mean 0 and variance σ which is finite.

Define the normalized sum $S_n = \frac{X_1+\cdots+X_n}{\sigma\sqrt{n}}$.

Theorem. (The central limit theorem) $S_n - \mu$ converges in distribution (in the sense of convergence of the probability of it falling within any interval) to $N(0,1)$.

Proof. 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 $\phi_x(u) = 1 - \frac{\sigma^2u^2}{2} + o(u^2)$ 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 $\phi_x(u_n) = 1 - \frac{t^2}{n} + o\left(\frac{1}{n}\right)$, where this time $o\left(\frac{1}{n}\right)$ means that it if n is going to infinity this gets much smaller than $\frac{1}{n}$ by any arbitrarily large constant.

Note that here, most proofs expand e^{itx} as a taylor series, but this is not justified since after this a swap of expectation and infinite summation that is not justified normally takes place. Below is the rigorous way to do this.

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

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

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

Therefore,

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

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,

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

Now define h for real numbers as follows:

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

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 $E\left(x^2\right) = E\left(x^2\right) - E\left(x\right)^2 = \sigma^2$, we have that

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

Now, by definition of h,

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

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 prove $\lim_{u\to 0} \frac{1}{u^2} E\left[R\left(ux\right)\right] = 0$ where $R\left(y\right) = e^{iy} - 1 - iy + \frac{y^2}{2}$ and $E\left[R\left(ux\right)\right] = \int_0^1 R\left(ux\right) d\mu$

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:

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

and we write

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

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

Parametrized by $u \to 0$ 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 $E\left[h\left(uX\right)X^2\right] = \int_0^1 h\left(ux\right)x^2 d\mu$

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 $1 - \frac{t^2}{n} + o\left(\frac{1}{n}\right)$,the cf of S_n is $\left(1 - \frac{t^2}{n} + o\left(\frac{1}{n}\right)\right)^n$, 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 (since it has a pdf in this case)

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

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

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

□

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

8 Chi squared tests

This is where uniqueness of characteristic functions gets dethroned as the hardest thing we prove in this website.

The problem is defining what a degree of freedom is AND showing that the test statistic $\sum_{r=1}^{k} \frac{(O_r-E_r)^2}{E_r}$ is well approximated for fairly large n by a distribution that depends ONLY on this number. And in the proof we will better understand what this distribution (That you are taught is a chi squared distribution) really is. We will prove 3 claims that will on their own explain quite well why chi squared tests work and then use them to prove a master theorem which says that the cumulative distribution function converges to this distribution as we add more trials to any fixed model.

Definition. A random vector is a vector of random variables that are possibly not independent. It may have its own probability density function of multiple variables and it always has its own cumulative distribution function.

Now the setup of a chi squared test is that we have a list of probabilities of several events determined by the null hypothesis. For example, for testing if a dice if fair, this list is $\left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)$, and in this proof we will call these $(p_1, p_2, p_3, p_4, p_5, p_6)$

Now if we roll the dice 120 times, our probability distribution becomes a lattice in 5D space with the "mean/expected" vector being $(np_1, np_2, \ldots, np_k) = (20,\ 20,\ 20,\ 20,\ 20,\ 20)$ in this case. In general the lattice is in (k-1)-D space where k is the number of cells in the table, since we have a constraint that the sum of all entries in the table equals a fixed total, which causes the dimension to decrease by 1.

If we are testing something like a binomial distribution and use the data to form our null hypothesis, then we are assuming that the true mean is the same as the mean estimated by the data we got, which is effectively finding the probability distribution **given** that we have that linear constraint, and this reduces the number of dimensions that the distribution lives in by 1, provided they are independent from other existing constraints. This dimension count is exactly what we call the degrees of freedom.

Now similarly if we have a contingency table then fixing the first r-1 row totals adds r-1 constraints and fixing the last row does nothing as fixing the rest fixes the last one automatically, and similarly we have c-1 constraints for columns and each one decreases the dimension of 1. Since the unconstrained model has rc-1 degrees of freedom as there are rc cells, the constrained model has rc-r-c+1 degrees of freedom which is (r-1)(c-1) if you expand it.

Therefore the first step of understanding what a degree of freedom is is done.

From now on I will use x_r to denote the r'th entry of a vector x.

Now consider our contingency table or list of observed data to be a random vector X and total n and number of cells k. will consider the random vector with general r'th entry $\sqrt{n}\left(\frac{\overline{x_r}-p_r}{\sqrt{p_r}}\right)$, where we note that the expected value of $\overline{x_r}$ is p_r (In the dice example, the number of ones divided by the number of rolls would be $\overline{x_1}$ and the expected value of this is $p_1 = \frac{1}{6}$). This begs the question of why we are specifically considering the vector given by $\sqrt{n}\left(\frac{\overline{x_r}-p_r}{\sqrt{p_r}}\right)$, which is what I call the standardized version of m, or m moved to the standardized world, where m is what I am calling the random vector of the table itself. The answer is because of what happens when you compute the square of its magnitude.

The square of the magnitude is the sum of the squares of its components. This is $\sum_{r=1}^{k} n\frac{(\overline{x_r}-p_r)^2}{p_r}$. Lets now multiply the numerator and denominator by n to get $\sum_{r=1}^{k} \frac{n^2(\overline{x_r}-p_r)^2}{np_r} = \sum_{r=1}^{k} \frac{(n\overline{x_r}-np_r)^2}{np_r} = \sum_{r=1}^{k} \frac{(O_r-E_r)^2}{E_r}$: Exactly the test statistic.

Now a chi squared distribution is defined as follows:

Take a normal distribution in k dimensions where each component of this k-dimensional random vector is N(0,1) and they are all independent, then note that the probability density of this function is rotationally symmetric (This is a property of the normal distribution we discussed earlier and it comes from pythagoras – this property is why it comes up so much). A χ_k^2 distribution is defined as the distribution of the square of the absolute value of this random vector, or the sum of the squares of its components, or the sum of the squares of k N(0,1) variables.

Therefore, we want to investigate $\sqrt{n}\left(\frac{\overline{x_r}-p_r}{\sqrt{p_r}}\right)$ when constrained to lattices in the plane and hope that we can show that it is standard normal so that we get the result we want. To do this, we will show that the probability of all versions of the contingency table approach what a normal distribution would predict, in a precise sense described below. It is not

enough to merely show that the total-only constraint version converges to a k-1-dimensional normal and then assert that constraining a k-1 dimensional normal to s-dimensional slices gives an s-dimensional normal, as although this is true, if the actual probability distribution of tables had a small number of bad slices, the whole k-1 dimensional thing could still converge to a standard normal. This is essentially what I did in the old version of the proof, although with some bounding arguments to try to rule out this behavior which may have been enough to derive a contradiction if we had a bad slice for infinitely many n but this approach is cleaner and more robust. Therefore we need to control the probability of each table configuration, as if we do that instead of controlling just the limiting distribution which allows for probabilities of nearby tables to oscillate and average out and generally behave badly, it will be enough. Note that we only consider linear constraints (ie, constraints to flat regions, not curved regions) in this proof.

Philosophical remark: The statement "Chi squared tests work" is not very precise so we need to pick a statement that will justify that chi squared tests work. So here is that statement:

Take the vector of expected values and the nearest lattice point in the unconstrained version of the distribution (there is one at most 1 count away in each cell) and call this table m*, and let y be $\left|\sqrt{n}\left(\frac{\overline{x}-p}{\sqrt{p}}\right)\right|$ (the "standardized" vector). Also fix a $0 < \varepsilon < \frac{1}{6}$ (it will be later that we see why the upper bound was $\frac{1}{6}$) and a region R defined by $|x - E| < n^{\frac{1}{2}+\varepsilon}$, and let m be an arbitrary table. Here x is the counts vector.

Then we claim that

1. $p\left(m \in R\right) \to 1$ as $n \to \infty$
2. $\sup_{m \in R}\left(\left|\frac{p(m)}{p(m^*)e^{-\frac{1}{2}|y|^2}} - 1\right|\right) \to 0$ as $n \to \infty$

Note that we are considering the ratio $\frac{p(m)}{p(m^*)e^{-\frac{1}{2}|y|^2}}$ here as the standard normal predicts that the ratio between the probability of an arbitrary table and the probability of the peak table is about $e^{-\frac{1}{2}|y|^2}$

In english, we will show that for large enough n, the probability of any table in a region with probability approaching 1 as n gets large can be made close to the prediction of the standard normal with as narrow of a margin as we want, and also that this is true when conditioning to a lattice.

Note a subtlety here: If we merely showed that $\left|\frac{p(m)}{p(m^*)e^{-\frac{1}{2}|y|^2}} - 1\right| \to 0$ for each m this would not be enough. I know this isn't a good example as m depends on n, but the idea is we want uniform convergence instead of pointwise convergence (which we defined and discussed in Level 6.2). Note that we need uniform convergence for this proof because it allows us to pass the limit through the integral at the end, which is why we add in the supremum operator, because we need to show that there is a fixed n where the normal approximation is fairly good everywhere.

At this point, Claims 1 and 2 give us uniform control over the probability of every table in the central region R. Because this control holds table-by-table, we are allowed to impose a linear constraint and reason exactly as in the continuous normal case: inside R, conditioning simply restricts us to a lower-dimensional normal shape (by symmetry, as explained earlier). What is not yet guaranteed is that conditioning cannot concentrate a significant amount of probability into the tails, since Claim 1 only controls the total probability of the tails before conditioning. A third claim is therefore needed to show that the tail bound from Claim 1 still holds after conditioning, so that almost all of the conditional probability remains inside R.

3. Define R_2 by $|x - E| < An^{\frac{1}{2}+\varepsilon}$ (where A is a constant that depends on the specific probabilities but not on n which we will explain how to find in the proof of the lemma) which still has probability going to 1 by precisely the same proof as part 1 of the claim. Then the claim is that for any S intersecting R_2, $p\left(m \in R \mid m \in S\right) \to 1$ as $n \to \infty$ where S is any lattice we constrain to: We need this because we only have control inside R by parts 1 and 2 so in principle there could be "bad lattices" where this does not hold without contradicting part 1 provided this happens with probability going to 0 as $n \to \infty$.

Subtlety: We need it to be the case that the set of possible points in the lattice we constrain to does not have lower dimension than the lattice we are constraining to. What I mean is, if the table was 2 cells that both have 1 and 1, then a constraint like "$\sqrt{2}$cell 1 + cell 2 = $1 + \sqrt{2}$" seems like we are constraining to 1 dimension, but really the set of tables

actually in that constraint is 0 dimensional. Luckily, in all practical cases, it is easy to show that this dimension thing.

Example. If we have an $r * c$ contingency table then we can change it in $(r-1)(c-1)$ independent ways and still get a valid contingency table, unless one of the cells is 0, but as we will see later in the proof this is not an issue. We can do this by picking a row number A between 1 and r-1 inclusive and a column number b between 1 and c-1 inclusive and doing the following:

- Add 1 to the (A,B) and (r,c) cells

- Subtract 1 from the (A,c) and (r,B) cells

This leaves the constraints satisfied and there are (r-1)(c-1) directions we can move in this way, so that is the dimension of the set of possible tables.

Example. If we have the specific case where a cell counts the number of 1's, the number of 2's, etc, and we are estimating the mean, and there are k total cells, then my k-2 possible moves (for each i from 2 to k-1 inclusive) are to remove 2 counts from the i'th cell and add 1 to both the i-1 and i+1'th cell, leaving the mean and total unchanged.

Since we control the probability of every single possible table as well as the behavior in every slice in the "almost certain" zone, we can say confidently once we have proven this that chi squared tests work better as n gets large as if n is large enough we will eventually have to go far into the "rejection zone" to run into issues anyways.

Note that in the unconstrained model, $|m^* - E| \leq \sqrt{k}$. This is because there exists a possible table within a distance $\sqrt{k}$ of E and the reason is because we can round everything in E down and then add 1 to each component until the total is correct, and then each component will be within 1 of E and it will be a possible table with distance at most $\sqrt{k}$ from E.

Lemma 1. (Stirling's approximation) For all positive integers n, $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n < n! \leq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}$

Proof. Define $d_n \coloneqq \log\left(\frac{n!e^n}{n^{n+\frac{1}{2}}}\right) = \log(n!) - \left(n+\frac{1}{2}\right)\log(n) + n$

Then $d_n - d_{n+1} = \log(n!) - \left(n+\frac{1}{2}\right)\log(n) + n - \log((n+1)!) + \left(n+\frac{3}{2}\right)\log(n+1) - n - 1$

$$d_n - d_{n+1} = -\log(n+1) - \left(n+\frac{1}{2}\right)\log(n) + \left(n+\frac{3}{2}\right)\log(n+1) - 1$$

$$d_n - d_{n+1} = -\left(n+\frac{1}{2}\right)\log(n) + \left(n+\frac{1}{2}\right)\log(n+1) - 1 = \left(n+\frac{1}{2}\right)\left(\log\left(\frac{n+1}{n}\right)\right) - 1$$

Now do the substitution $t \coloneqq \frac{1}{2n+1}$. $n = \frac{1-t}{2t}$ Then $d_n - d_{n+1} = \left(\frac{1}{2} + \frac{1-t}{2t}\right)\log\left(\frac{\frac{1-t}{2t}+1}{\frac{1-t}{2t}}\right) - 1$

$$= \frac{1}{2t}\log\left(\frac{1+t}{1-t}\right) - 1$$

Note that by the log taylor series, $\log(1+t) = t - \frac{t^2}{2} + \frac{t^3}{3} - \frac{t^4}{4} + \dots$ (|t|<1 automatically by how t is defined here so it is valid), and $\log(1-t) = -\left(t + \frac{t^2}{2} + \frac{t^3}{3} + \frac{t^4}{4} + \dots\right)$.

Therefore $\frac{1}{2t}\log\left(\frac{1+t}{1-t}\right) - 1 = \frac{1}{2t}\log(1+t) - \frac{1}{2t}\log(1-t) - 1$

$$= \frac{1}{2t}\left(t - \frac{t^2}{2} + \frac{t^3}{3} + \frac{t^4}{4} + \dots\right) + \frac{1}{2t}\left(t + \frac{t^2}{2} + \frac{t^3}{3} + \frac{t^4}{4} + \dots\right) - 1$$

$$= \frac{1}{2t}\left(2t + \frac{2t^3}{3} + \frac{2t^5}{5} + \frac{2t^7}{7} + \dots\right) - 1 = 1 + \frac{t^2}{3} + \frac{t^4}{5} + \frac{t^6}{7} + \dots - 1 = \frac{t^2}{3} + \frac{t^4}{5} + \frac{t^6}{7} + \dots$$

t is positive so $d_n - d_{n+1} < \frac{t^2}{3} + \frac{t^4}{3} + \frac{t^6}{3} + \dots = \frac{1}{3}\frac{t^2}{1-t^2}$ (geometric series)$= \frac{1}{3((2n+1)^2-1)}$ by definition of t. By simple algebra this is $\frac{1}{12}\left(\frac{1}{n} - \frac{1}{n+1}\right)$.

Now apply the method of differences/telescoping. $d_1 - d_n = d_1 - d_2 + d_2 - d_3 + \cdots + d_{n-1} - d_n$

$$< \frac{1}{12}\left(\frac{1}{1} - \frac{1}{2}\right) + \frac{1}{12}\left(\frac{1}{2} - \frac{1}{3}\right) + \cdots + \frac{1}{12}\left(\frac{1}{n-1} - \frac{1}{n}\right) = \frac{1}{12}\left(1 - \frac{1}{n}\right)$$

Note that $d_n - d_{n+1}$ is positive so d is a decreasing sequence, and for all n, d_n is bounded below by $d_1 - \frac{1}{12}$ for all n by the inequality above. A decreasing sequence that is bounded below converges and therefore our sequence d_n converges to a limit which we will call A which is also a strict lower bound for all terms d_n in our decreasing sequence.

Now let $m > n$, then $d_n - d_m < \frac{1}{12}\left(\frac{1}{n} - \frac{1}{n+1}\right) + \frac{1}{12}\left(\frac{1}{n+1} - \frac{1}{n+2}\right) + \cdots + \frac{1}{12}\left(\frac{1}{m-1} - \frac{1}{m}\right) = \frac{1}{12}\left(\frac{1}{n} - \frac{1}{m}\right)$. Taking the limit as $m \to \infty$ we get that $d_n - A \leq \frac{1}{12n}$ (not strict as a sequence always strictly above a value can converge to that value). Therefore $A < d_n \leq A + \frac{1}{12n}$.

Now we will do some magic by working on a seemingly irrelevant integral where A magically pops out.

$$I_n \coloneqq \int_0^{\frac{\pi}{2}} \sin^n(\theta) d\theta = \left[-\cos(\theta)\sin^{n-1}(\theta)\right]_0^{\frac{\pi}{2}} + (n-1)\int_0^{\frac{\pi}{2}} \sin^{n-2}(\theta)\cos^2(\theta) d\theta$$

Where we did integration by parts integrating sin and differentiating $\sin^{n-1}(\theta)$ and were careful to not flip any minus signs the wrong way even when we had double negatives to deal with.

$I_n = (n-1)\int_0^{\frac{\pi}{2}} \sin^{n-2}(\theta) - \sin^n(\theta) d\theta$ because the breacketed part vanishes to 0 and we have the pythagorean identity $\cos^2(\theta) = 1 - \sin^2(\theta)$. Thus $I_n = (n-1)(I_{n-2} - I_n)$ so $nI_n = (n-1)I_{n-2}$ so we have our recurrence relation $I_n = \frac{n-1}{n} I_{n-2}$ for n at least 2. Since $I_0 = \frac{\pi}{2}$ and $I_1 = 1$ by basic calculus we have the following:

$I_{2n} = \frac{1}{2}\frac{3}{4}\cdots\frac{2n-1}{2n}\frac{\pi}{2}$ and $I_{2n+1} = \frac{2}{3}\frac{4}{5}\cdots\frac{2n}{2n+1}$ by the recurrence relation. It is also easy to see that I_n is a decrerasing sequence since the functions $\sin^n(\theta)$ are decreasing as n increases as $0 < \sin^n(\theta) < 1$ everywhere in the domain except at $\frac{\pi}{2}$. Therefore $1 \leq \frac{I_{2n}}{I_{2n+1}} \leq \frac{I_{2n-1}}{I_{2n+1}} = 1 + \frac{1}{2n}$ by the recurrence relation and therefore $\frac{I_{2n}}{I_{2n+1}} \to 1$ and similarly so does $\frac{I_{2n-1}}{I_{2n}}$.

$$I_{2n} = \frac{1}{2}\frac{3}{4}\cdots\frac{2n-1}{2n}\frac{\pi}{2} = \frac{1*2*3*\cdots*2n}{2*2*4*4*\cdots*2n*2n}\frac{\pi}{2} = \frac{(2n)!}{(n!2^n)^2}\frac{\pi}{2}$$

$$I_{2n+1} = \frac{2}{3}\frac{4}{5}\cdots\frac{2n}{2n+1} = \frac{2*2*4*4*\cdots*2n*2n}{1*2*3*\cdots*2n*(2n+1)} = \frac{(n!2^n)^2}{(2n+1)!}$$

Note that when n is large the ratio between $n!$ and $n^{n+\frac{1}{2}}e^{-n+A}$ converges to 1 as the log of the difference converges to 0 by the definition of A and the definition of d_n.

Now $\frac{I_{2n}}{I_{2n+1}} = \frac{\frac{(2n)!}{(n!2^n)^2}\frac{\pi}{2}}{\frac{(n!2^n)^2}{(2n+1)!}} = \frac{(2n)!(2n+1)!\pi}{(n!2^n)^4 2} = \pi(2n+1)\left[\frac{((2n)!)^2}{2^{4n+1}(n!)^4}\right]$

$$\frac{((2n)!)^2}{2^{4n+1}(n!)^4} = \frac{\left(\frac{(2n)!e^{2n}}{(2n)^{2n+\frac{1}{2}}}\right)^2}{2^{4n+1}(n!)^4}\frac{e^{-4n}}{(2n)^{-1-4n}} = \frac{e^{2d_{2n}}}{2^{4n+1}\left(\frac{n!e^n}{n^{n+\frac{1}{2}}}\right)^4}\frac{e^{4n}}{n^{4n+2}}\frac{e^{-4n}}{(2n)^{-1-4n}} = \frac{e^{2d_{2n}}}{2^{4n+1}e^{4d_n}}\frac{1}{n^{4n+2}}\frac{1}{(2n)^{-1-4n}}$$

$$= \frac{e^{2d_{2n}}}{2^{4n+1}e^{4d_n}}\frac{1}{n^{4n+2}}\frac{1}{2^{-1-4n}n^{-1-4n}} = \frac{e^{2d_{2n}}}{e^{4d_n}}\frac{1}{n}$$

As n gets large the ratio between this and $\frac{1}{ne^{2A}} \to 1$ and thus the limit as n goes to infinity of the ratio $\frac{I_{2n}}{I_{2n+1}} = \pi(2n+1)\left[\frac{((2n)!)^2}{2^{4n+1}(n!)^4}\right]$ (which we know is 1) is exactly the limit of $\pi(2n+1)\frac{1}{ne^{2A}} = \frac{2\pi}{e^{2A}}$, and therefore A must be $\log\left(\sqrt{2\pi}\right)$. I told you this would me magic!

Thus, we are about to get the lemma.

$$A < d_n \leq A + \frac{1}{12n}$$

$$\log\left(\sqrt{2\pi}\right) < \log\left(\frac{n!e^n}{n^{n+\frac{1}{2}}}\right) \leq \log\left(\sqrt{2\pi}\right) + \frac{1}{12n}$$

$$\sqrt{2\pi} < \frac{n!e^n}{n^{n+\frac{1}{2}}} \leq \sqrt{2\pi}e^{\frac{1}{12n}}$$

$$\sqrt{2\pi n} < \frac{n!e^n}{n^n} \leq \sqrt{2\pi n}e^{\frac{1}{12n}}$$

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n < n! \leq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}$$

□

Remark. $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}$ is far closer to n! in practice than the lower bound, in fact it is possible to show that $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n+1}}$ is a lower bound and strict inequalities are true for both bounds but we will not prove this as although it is not too difficult it is quite long and technical and the bound given above is sufficient for the proof we will do.

Proof of part 1 of the claim:

We will work on the unconstrained model for now since the bounds we will get on the probabilities will allow us to condition onto sub-lattices later.

Suppose $|x - E| \geq n^{\frac{1}{2}+\varepsilon}$ but for all r from 0 to k, $|O_r - E_r| < \frac{n^{\frac{1}{2}+\varepsilon}}{\sqrt{k}}$. Then we would have $|x - E| < n^{\frac{1}{2}+\varepsilon}$ by pythagoras ($|x - E| = \sqrt{\sum_{r=1}^{k} (O_r - E_r)^2} < \sum_{r=1}^{k} \frac{n^{1+2\varepsilon}}{k} = \sqrt{n^{1+2\varepsilon}} = n^{\frac{1}{2}+\varepsilon}$). Hence $|x - E| \geq n^{\frac{1}{2}+\varepsilon}$ implies that there is at least one r with $|O_r - E_r| \geq \frac{n^{\frac{1}{2}+\varepsilon}}{\sqrt{k}}$. But the r'th entry of the vector is on its own is distributed like a binomial distribution (minus the mean) so is standard deviation is proportional to $\sqrt{n}$, so the number of standard deviations away the r'th entry of the vector would have to be from its value is proportional to n^{ε} which goes to infinity (slowly) as n goes to infinity.

We want to show that the probability of the r'th component being out by that much goes to 0, as it will then follow that the probability of any of the k components being out by that much is 0 since it is a fixed number (k) of separate events each with probability going to 0. Well, it *seems* obvious that being out by an unbounded number of standard deviations has probability approaching 0 but we want to show this rigorously.

Therefore we will prove an easy inequality (due to markov) that says that $P(y \geq a) \leq \frac{E(y)}{a}$ where y is any distribution.

The proof is that $E(y) \geq E(y \mid y \geq a) \geq E(a \mid y \geq a) = aP(y \geq a)$ so rearranging gives the desired result.

We prove another easy inequality (due to chebyshev) that says that for any distribution x we have that

$$P(|x - E(x)| \geq t) \leq \frac{Var(x)}{t^2}$$

The proof is simple: Just note $P\left(|x - E(x)|^2 \geq t^2\right) \leq \frac{Var(x)}{t^2}$ is equivalent to what we are claiming but it is just the above inequality applied to $|x - E(x)|^2$.

Therefore, since $Var(O_r) = np_r(1 - p_r) < n$, taking $t = \frac{n^{\frac{1}{2}+\varepsilon}}{\sqrt{k}}$. Then the inequality above gives that

$$P(|O_r - E_r| \geq t) \leq \frac{Var(x)}{t^2} < \frac{kn}{n^{1+2\varepsilon}} = kn^{-2\varepsilon} \to 0$$

Note that from now on I use m_r, O_r interchangably to mean the same thing.

Proof of part 2 of the claim:

We need to better understand the probability of a particular table in the unconstrained model. The idea is we can get it in terms of factorials then apply lemma 1.

Suppose the total is n and the counts are $m_1, m_2, \ldots, m_k$. Then we will argue as follows:

We do n trials and the number of ways to interchange the order of these n trials is n!, however for each possible outcome, the order of those is irrelevant (For example, if you roll 120 dice, then we can change the order that we got the 1's in 20! Different ways without changing the overall order). Therefore the number of possible ways to order the trials without distinguishing the same outcome is exactly $\frac{n!}{m_1!m_2!\ldots m_k!}$, and the probability of one particular ordering is $p_1^{m_1}p_2^{m_2}\ldots p_k^{m_k}$. Therefore the exact probability of a particular table is $\frac{n!}{m_1!m_2!\ldots m_k!}p_1^{m_1}p_2^{m_2}\ldots p_k^{m_k}$.

Taking logs of lemma 1 gives $\log(m!) = \left(m+\frac{1}{2}\right)\log(m) - m + \frac{1}{2}\log(2\pi) + r_m$ with $0 < r_m < \frac{1}{12}$.

Now set $h_i \coloneqq m_i - m_i^* = O_i - E_i + O(1)$. Therefore in the standardized world, $y_i = \frac{h_i}{\sqrt{np_i}} + O\left(n^{-\frac{1}{2}}\right)$.

Here when I use say O(f(x)) I mean there exists a constant C not depending on n or m such that for all m in R the thing we are saying is |O(f(x))| is less than Cf(x)

$$\log(P(m)) = \log(n!) - \sum_{i=1}^{k}\log(m_i!) + \sum_{i=1}^{k} m_i\log(p_i)$$

By lemma 1,

$$\log(P(m)) = \left(n+\frac{1}{2}\right)\log(n) - n + \frac{1}{2}\log(2\pi) + r_n - \sum_{i=1}^{k}\left(m_i+\frac{1}{2}\right)\log(m_i) - m_i + \frac{1}{2}\log(2\pi) + r_{m_i} + \sum_{i=1}^{k} m_i\log(p_i)$$

$$\log(P(m^*)) = \left(n+\frac{1}{2}\right)\log(n) - n + \frac{1}{2}\log(2\pi) + r_n - \sum_{i=1}^{k}\left(m_i^*+\frac{1}{2}\right)\log(m_i^*) - m_i^* + \frac{1}{2}\log(2\pi) + r_{m_i^*} + \sum_{i=1}^{k} m_i^*\log(p_i)$$

Then

$$\log(P(m)) - \log(P(m^*)) = -\sum_{i=1}^{k}\left(m_i+\frac{1}{2}\right)\log(m_i) - h_i + r_{m_i} + \sum_{i=1}^{k} h_i\log(p_i) + \sum_{i=1}^{k}\left(m_i^*+\frac{1}{2}\right)\log(m_i^*) + r_{m_i^*}$$

Then the remainder terms contribute $\left|\sum_{i=1}^{k} r_{m_i} - r_{m_i^*}\right| \leq \sum_{i=1}^{k}\frac{1}{12}\left|\frac{1}{m_i} - \frac{1}{m_i^*}\right|$ by the triangle inequality.

If $n > \frac{8}{p_i^3}$ for all i and we are in R, $m_i > m_i^* - n^{\frac{2}{3}} > \frac{1}{2}np_i$ which is a constant times n. Therefore, if n is large enough, $\left|\sum_{i=1}^{k} r_{m_i} - r_{m_i^*}\right| \leq \sum_{i=1}^{k}\frac{1}{12}\left|\frac{h_i}{m_i m_i^*}\right| \leq C\sum_{i=1}^{k}\left|\frac{h_i}{n^2}\right|$ for some fixed C. This is therefore $O\left(n^{\varepsilon-\frac{3}{2}}\right)$.

$m_i = m_i^*(1+u_i)$ with $u_i = \frac{h_i}{m_i^*} = \frac{O\left(n^{\frac{1}{2}+\varepsilon}\right)}{np_i+O(1)} = O\left(n^{\varepsilon-\frac{1}{2}}\right)$

Write $\Delta(m) \coloneqq \log(P(m)) - \log(P(m^*)) = \sum_{i=1}^{k} -\left(m_i+\frac{1}{2}\right)\log(m_i) + h_i + h_i\log(p_i) + \left(m_i^*+\frac{1}{2}\right)\log(m_i^*) + O(n^{\varepsilon-\frac{3}{2}})$

Write $m_i = m_i^*(1+u_i)$ so $u_i \coloneqq \frac{h_i}{m_i^*} < \frac{n^{\frac{1}{2}+\varepsilon}}{np_i+O(1)} = O\left(n^{\varepsilon-\frac{1}{2}}\right)$.

We note that $\sum h_i = 0$ so

$$\Delta(m) = \sum_{i=1}^{k} -\left(m_i+\frac{1}{2}\right)\log(m_i) + h_i\log(p_i) + \left(m_i^*+\frac{1}{2}\right)\log(m_i^*) + O(n^{\varepsilon-\frac{3}{2}})$$

Now $\left(m_i+\frac{1}{2}\right)\log(m_i) = \left(m_i^*+h_i+\frac{1}{2}\right)\left(\log(m_i^*) + \log(1+u_i)\right)$

Therefore

$$\Delta(m) = \sum_{i=1}^{k} -\left(m_i^*+h_i+\frac{1}{2}\right)\left(\log(m_i^*) + \log(1+u_i)\right) + h_i\log(p_i) + \left(m_i^*+\frac{1}{2}\right)\log(m_i^*) + O(n^{\varepsilon-\frac{3}{2}})$$

$$\Delta(m) = \sum_{i=1}^{k} -m_i^*\log\left(m_i^*\right) - h_i\log\left(m_i^*\right) - \frac{1}{2}\log\left(m_i^*\right) - m_i^*\log(1+u_i) - h_i\log(1+u_i) - \frac{1}{2}\log(1+u_i) + h_i\log(p_i) + \left(m_i^*+\frac{1}{2}\right)\log\left(m_i^*\right) + O(n^{\varepsilon-\frac{3}{2}})$$

$$\Delta(m) = \sum_{i=1}^{k} -h_i \log(m_i^*) - m_i^* \log(1+u_i) - h_i \log(1+u_i) - \frac{1}{2}\log(1+u_i) + h_i \log(p_i) + O(n^{\varepsilon - \frac{3}{2}})$$

$$\Delta(m) = \sum_{i=1}^{k} -h_i \log(m_i^*) - m_i \log(1+u_i) - \frac{1}{2}\log(1+u_i) + h_i \log(p_i) + O(n^{\varepsilon - \frac{3}{2}})$$

$$h_i(\log(m_i^*) - \log(p_i)) = h_i\left(\log\left(\frac{m_i^*}{p_i}\right)\right) = h_i(\log(n + O(1))) = h_i(\log(n) + O\left(\frac{1}{n}\right))$$

$$\sum_{i=1}^{k} h_i(\log(m_i^*) - \log(p_i)) = \sum_{i=1}^{k} h_i\left(O\left(\frac{1}{n}\right)\right) = O\left(n^{\varepsilon - \frac{1}{2}}\right)$$

This is because $\sum h_i = 0$

Then

$$\Delta(m) = \sum_{i=1}^{k} -m_i^* \log(1+u_i) - h_i \log(1+u_i) - \frac{1}{2}\log(1+u_i) + O(n^{\varepsilon - \frac{1}{2}})$$

$$\Delta(m) = \sum_{i=1}^{k} -m_i \log(1+u_i) - \frac{1}{2}\log(1+u_i) + O(n^{\varepsilon - \frac{1}{2}})$$

$$R_3(x) \coloneqq \log(1+x) - x + \frac{x^2}{2}$$

$$\Delta(m) = \sum_{i=1}^{k} -m_i\left(u_i - \frac{1}{2}u_i^2 + R_3(u_i)\right) - \frac{1}{2}u_i + \frac{1}{4}u_i^2 - \frac{1}{2}R_3(u_i) + O(n^{\varepsilon - \frac{1}{2}})$$

$$\Delta(m) = \sum_{i=1}^{k} -m_i\left(u_i - \frac{1}{2}u_i^2 + R_3(u_i)\right) + \sum_{i=1}^{k} -\frac{1}{2}u_i + \frac{1}{4}u_i^2 - \frac{1}{2}R_3(u_i) + O(n^{\varepsilon - \frac{1}{2}})$$

Note that by definition $h_i = m_i^* u_i$ and $m_i = m_i^*(1+u_i)$ so $m_i u_i = m_i^*(1+u_i)u_i = h_i + m_i^* u_i^2$

Therefore since $\sum h_i = 0$ we get that

$$\Delta(m) = \sum_{i=1}^{k} -m_i^* u_i^2 + \frac{1}{2}m_i u_i^2 - m_i R_3(u_i) + \sum_{i=1}^{k} -\frac{1}{2}u_i + \frac{1}{4}u_i^2 - \frac{1}{2}R_3(u_i) + O(n^{\varepsilon - \frac{1}{2}})$$

$$\Delta(m) = \sum_{i=1}^{k} -\frac{1}{2}m_i^* u_i^2 + \frac{1}{2}h_i u_i^2 - m_i R_3(u_i) + \sum_{i=1}^{k} -\frac{1}{2}u_i + \frac{1}{4}u_i^2 - \frac{1}{2}R_3(u_i) + O(n^{\varepsilon - \frac{1}{2}})$$

Now use the fact that $u_i = \frac{h_i}{m_i^*}$ to get

$$\Delta(m) = \sum_{i=1}^{k} -\frac{1}{2}\frac{h_i^2}{m_i^*} + \frac{1}{2}\frac{h_i^3}{(m_i^*)^2} - m_i R_3(u_i) + \sum_{i=1}^{k} -\frac{1}{2}u_i + \frac{1}{4}u_i^2 - \frac{1}{2}R_3(u_i) + O(n^{\varepsilon - \frac{1}{2}})$$

$$\Delta(m) = \sum_{i=1}^{k} -\frac{1}{2}\frac{h_i^2}{m_i^*} + \frac{1}{2}\frac{h_i^3}{(m_i^*)^2} - m_i R_3(u_i) - \sum_{i=1}^{k} \frac{1}{2}R_3(u_i) + O(n^{\varepsilon - \frac{1}{2}})$$

Because $\sum u_i = \sum O\left(n^{\varepsilon - \frac{1}{2}}\right) = O\left(n^{\varepsilon - \frac{1}{2}}\right)$ since the sum is finite and $\sum u_i^2$ gets even smaller, in fact it is $O\left(n^{2\varepsilon - 1}\right)$ as n gets large.

Now $\frac{h_i^3}{(m_i^*)^2} = \frac{O\left(n^{\frac{3}{2}+3\varepsilon}\right)}{O(n^2)}$ where crucially for large enough n $(m_i^*)^2 \geq Cn^2$ for some positive C which allows us to say that its reciprocal is $O(n^{-2})$. It is here we use the fact that $\varepsilon < \frac{1}{6}$, as we now have

$$\Delta(m) = \sum_{i=1}^{k} -\frac{1}{2}\frac{h_i^2}{m_i^*} - m_i R_3(u_i) - \sum_{i=1}^{k} \frac{1}{2}R_3(u_i) + O(n^{3\varepsilon - \frac{1}{2}})$$

Now by taylor's theorem, $R_3(u_i)$ from how it was defined is $\frac{1}{3}\xi_i^3$ for some $0 < \xi_i < u_i$, and since by definition $u_i = O\left(n^{\varepsilon - \frac{1}{2}}\right)$ and $m_i = O(n)$ so $R_3(u_i) = O\left(n^{3\varepsilon - \frac{3}{2}}\right)$ and thus so is $\sum_{i=1}^{k} \frac{1}{2} R_3(u_i)$, and also $m_i R_3(u_i) = O\left(n^{3\varepsilon - \frac{1}{2}}\right)$ so at last we have

$$\Delta(m) = \sum_{i=1}^{k} -\frac{1}{2}\frac{h_i^2}{m_i^*} + O(n^{3\varepsilon - \frac{1}{2}})$$

The last step is

$$\Delta(m) = \sum_{i=1}^{k} -\frac{1}{2}\frac{h_i^2}{E_i + O(1)} + O(n^{3\varepsilon - \frac{1}{2}})$$

$$\Delta(m) = \sum_{i=1}^{k} -\frac{1}{2}\frac{h_i^2}{E_i}\left(1 + O\left(\frac{1}{n}\right)\right) + O(n^{3\varepsilon - \frac{1}{2}})$$

$$\Delta(m) = \sum_{i=1}^{k} -\frac{1}{2}\frac{h_i^2}{E_i} + \sum_{i=1}^{k} -\frac{1}{2}\frac{h_i^2}{E_i} O\left(\frac{1}{n}\right) + O(n^{3\varepsilon - \frac{1}{2}})$$

$$\Delta(m) = \sum_{i=1}^{k} -\frac{1}{2}\frac{h_i^2}{E_i} + O(n^{3\varepsilon - \frac{1}{2}})$$

Since that term was $O(n^{2\varepsilon - 1})$.

Now we are basically there because:

$$\Delta(m) = \log(P(m)) - \log(P(m^*))$$

Therefore

$$\log(P(m)) - \log(P(m^*)) = \frac{1}{2}|y|^2 + O(n^{3\varepsilon - \frac{1}{2}})$$

Undoing logs gives

$$\frac{P(m)}{P(m^*)} = e^{-\frac{1}{2}|y|^2} T$$

Where T goes to 1 as n goes to infinity for each fixed set of p values and regardless of m provided m is in R, so done with part 2 of the claim.

Lemma 2. There exists constants $c, C > 0$ such that $\sup_{m \notin R} p(m) \leq C p(m^*) e^{-cn^{2\varepsilon}}$ for sufficiently large n. Assume the probabilities are fixed and non-zero and the number of cells k is fixed.

Proof. We first show the simple inequality tyat says $(1+t)\log(1+t) \geq t + \frac{t^2}{3}$ for all $t \in \left[-\frac{1}{2}, \frac{1}{2}\right]$.

Consider $(1+t)\log(1+t) - t - \frac{t^2}{3}$. The derivative of this is $\log(1+t) - \frac{2t}{3}$ and the second derivative is $\frac{1}{1+t} - \frac{2}{3}$ which in the range we care about is always non-negative. Notice that the derivative at 0 is 0 so we have a graph with positive curvature that flattens at t=0 where the value is 0, and thus the graph is non-negative in the range we care about so we finished the inequality.

Define a probability vector $q \coloneqq (q_1, q_2, \ldots, q_k)$, $\sum q_r = 1$. Define $t_r \coloneqq \frac{q_r}{p_r} - 1$.

Define $D(q||p) \coloneqq \sum_{r=1}^{k} q_r \log\left(\frac{q_r}{p_r}\right) = \sum_{r=1}^{k} p_r(1+t_r)\log(1+t_r)$. Using the inequality above we have the bound $D(q||p) \geq \sum_{r=1}^{k} p_r\left(t_r + \frac{t_r^2}{3}\right)$ whenever all t_r's are between $-\frac{1}{2}$ and $+\frac{1}{2}$ – we will handle the case where this is not true separately.

$$D(q||p) \geq \sum_{r=1}^{k} p_r t_r + \frac{1}{3}\sum_{r=1}^{k} p_r t_r^2 = \sum_{r=1}^{k} q_r - p_r + \frac{1}{3}\sum_{r=1}^{k} p_r t_r^2 = \frac{1}{3}\sum_{r=1}^{k} p_r t_r^2 = \frac{1}{3}\sum_{r=1}^{k} \frac{(q_r - p_r)^2}{p_r}$$

Given a table m set $q_r = \frac{m_r}{n}$, ie the observed probability. The standardized vector in components is given by

$$y_r = \sqrt{n}\frac{q_r - p_r}{\sqrt{p_r}} = \frac{O_r - E_r}{\sqrt{np_r}}$$

Recall that from this we derived

$$|y|^2 = \sum_{r=1}^{k} \frac{(O_r - E_r)^2}{E_r}$$

and also it is the case that

$$|m - E|^2 = \sum_{r=1}^{k} (O_r - E_r)^2$$

Therefore

$$|y|^2 = \sum_{r=1}^{k} \frac{(O_r - E_r)^2}{E_r} \geq \frac{1}{\max_r E_r} \sum_{r=1}^{k} (O_r - E_r)^2 = \frac{|m - E|^2}{\max_r E_r}$$

By definition of R, if $m \notin R$, $|y|^2 \geq \frac{n^{1+2\varepsilon}}{\max_r E_r} = \frac{n^{2\varepsilon}}{p_{max}}$. Here p_{max} is the largest probability.

In the case where all t_r's are between $-\frac{1}{2}$ and $+\frac{1}{2}$ we had $D(q\,|\,|\,p) \geq \frac{1}{3}\sum_{r=1}^{k} \frac{(q_r - p_r)^2}{p_r} = \frac{1}{3n}\sum_{r=1}^{k} \frac{(O_r - E_r)^2}{E_r} = \frac{|y|^2}{3n}$ so $nD\,(q\,|\,|\,p) \geq \frac{y^2}{3}$ so by the previous deduction $nD\,(q\,|\,|\,p) \geq \frac{n^{2\varepsilon}}{3p_{max}}$.

If for some r, $|t_r| > \frac{1}{2}$ then the set of possible q vectors (including cases where some are 0) sastisfying $|t_r| \geq \frac{1}{2}$ is bounded and since it contains all its boundary points it is closed, and trivially it does not contain p. The function $D(q\,|\,|\,p)$ for fixed p is continuous on K (we say 0log(0)=0 for the case when one of the q's is 0 as that is the limit of xlog(x) as x goes to 0), it must be the case (by the technical results document, extreme value theorem) that $D(q\,|\,|\,p)$ attains a minimum value on K d_0 somewhere in K. Note that $d_0 > 0$: The proof of this is as follows.

For all positive u, $\log(u) \leq u - 1$ because those curves are tangent at u=1 and log has negative curvature everywhere (differentiate it twice). Therefore $\log\left(\frac{p_r}{q_r}\right) \leq \frac{p_r}{q_r} - 1$ so $q_r\log\left(\frac{p_r}{q_r}\right) \leq p_r - q_r$, therefore $q_r\log\left(\frac{q_r}{p_r}\right) \geq q_r - p_r$, so $\sum_{r=1}^{k} q_r\log\left(\frac{q_r}{p_r}\right) \geq \sum_{r=1}^{k} q_r - p_r = 0$.

Note that $\log(u) = u - 1$ only happens when $u = 1$ so if p and q differ for any r then that term in the sum becomes strictly positive and the rest are non-negative so the inequality becomes strict, hence the inequality is strict whenever q and p are not the same, so certainly everywhere in K.

Therefore everywhere in K, $nD\,(q\,|\,|\,p) \geq nd_0$.

Note that as explained earlier the number of ways to get a certain table by doing a bunch of trials is exactly $\frac{n!}{\prod m_r!}$, and $\prod p_r^{m_r} = \exp\left(n\sum q_r\log(p_r)\right)$ where exp(x) means e^x. Now define another function $H(q) \coloneqq -\sum q_r\log(q_r)$ with 0log(0) defined as 0.

Now $\frac{n!}{\prod m_r!}$ is the $\prod_{r=1}^{k} m_r^{m_r}$ coefficient in the expansion of $n^n = (m_1 + m_2 + \cdots + m_k)^n$ so we have the inequality $\frac{n!}{\prod m_r!} \leq \frac{n^n}{\prod_{r=1}^{k} m_r^{m_r}}$ so $p(m) = \frac{n!}{\prod m_r!}\prod p_r^{m_r} \leq \prod p_r^{m_r}\frac{n^n}{\prod_{r=1}^{k} m_r^{m_r}}$.

Taking logs gives $\log(p(m)) \leq n\log(n) - \sum m_r\log(m_r) + \sum m_r\log(p_r)$.

$$\log(p(m)) \leq n\log(n) - \sum nq_r\log(q_r) - \sum nq_r\log(n) + \sum m_r\log(p_r)$$

$$\log(p(m)) \leq -\sum nq_r\log(q_r) + \sum nq_r\log(p_r)$$

$$\log(p(m)) \leq n\sum (q_r)\log\left(\frac{p_r}{q_r}\right) = -nD(q\,|\,|\,p)$$

Thus $p(m) \leq e^{-nD(q\,|\,|P)} \leq e^{-nd_0}$

Now we apply lemma 1 and take logs of both sides of the formula for probabilities of particular unconstrained tables.

$$\log(p(m)) = \log(n!) - \sum_{r=1}^{k} \log(m_r!) + \sum_{r=1}^{k} m_r\log(p_r)$$

$$\log(p(m)) = n\log(n) - n + \frac{1}{2}\log(2n\pi) + R_n - \sum_{r=1}^{k} m_r\log(m_r) - m_r + \frac{1}{2}\log(2m_r\pi) + R_{m_r} + \sum_{r=1}^{k} m_r\log(p_r)$$

Where $0 < R_n < \frac{1}{12n}$ and $0 < R_{m_r} < \frac{1}{12m_r} \leq \frac{1}{12np_{min}}$, therefore

$$\log\left(p\left(m\right)\right) = n\log\left(n\right) - n + \frac{1}{2}\log\left(2n\pi\right) - \sum_{r=1}^{k}\left[m_r\log\left(m_r\right) - m_r + \frac{1}{2}\log\left(2m_r\pi\right)\right] + \sum_{r=1}^{k} m_r\log\left(p_r\right) + R$$

Where R goes to 0 as n goes to infinity.

Simplifying,

$$\log\left(p\left(m\right)\right) = n\log\left(n\right) + \frac{1}{2}\log\left(2n\pi\right) - \sum_{r=1}^{k}\left[m_r\log\left(m_r\right) + \frac{1}{2}\log\left(2m_r\pi\right)\right] + \sum_{r=1}^{k} m_r\log\left(p_r\right) + R$$

Because n is the sum of the m_r's

$$\log\left(p\left(m\right)\right) = n\log\left(n\right) + \frac{1}{2}\log\left(2n\pi\right) - \sum_{r=1}^{k}\left[m_r\log\left(m_r\right)\right] - \frac{1}{2}\sum_{r=1}^{k}\left[\log\left(2m_r\pi\right)\right] + \sum_{r=1}^{k} m_r\log\left(p_r\right) + R$$

$$\log\left(p\left(m\right)\right) = \frac{1}{2}\log\left(2n\pi\right) - \sum_{r=1}^{k}\left[m_r[\log\left(q_r\right))]\right] - \frac{1}{2}\sum_{r=1}^{k}\left[\log\left(2m_r\pi\right)\right] + \sum_{r=1}^{k} m_r\log\left(p_r\right) + R$$

$$\log\left(p\left(m\right)\right) = \frac{1}{2}\log\left(2n\pi\right) - \sum_{r=1}^{k}\left[m_r[\log\left(\frac{q_r}{p_r}\right))]\right] - \frac{1}{2}\sum_{r=1}^{k}\left[\log\left(2m_r\pi\right)\right] + R$$

$$\log\left(p\left(m\right)\right) = \frac{1}{2}\log\left(2n\pi\right) - n\sum_{r=1}^{k}\left[q_r\left[\log\left(\frac{q_r}{p_r}\right)\right]\right] - \frac{1}{2}\sum_{r=1}^{k}\left[\log\left(2m_r\pi\right)\right] + R$$

$$\log\left(p\left(m\right)\right) = \frac{1}{2}\log\left(2n\pi\right) - nD\left(q\,|\,|\,p\right) - \frac{1}{2}\sum_{r=1}^{k}\left[\log\left(2m_r\pi\right)\right] + R$$

$$\log\left(p\left(m\right)\right) = \frac{1-k}{2}\log\left(n\right) - nD\left(q\,|\,|\,p\right) - \frac{1}{2}\sum_{r=1}^{k}\left[\log\left(q_r\right)\right] - \frac{k+1}{2}\log(2\pi) + R$$

Note that $\frac{k+1}{2}\log(2\pi)$ is a fixed constant here.

Note that $|m^* - E| \leq \sqrt{k}$ from earlier, therefore $|q^* - p| \leq \frac{\sqrt{k}}{n}$ and therefore $D\left(q\,|\,|\,p\right)$.

By the technical results document (Taylor's theorem),

$$\log\left(\frac{q_r}{p_r}\right) = \log\left(1 + \frac{q_r - p_r}{p_r}\right) = \frac{q_r - p_r}{p_r} - \frac{1}{2}\left(\frac{q_r - p_r}{p_r}\right)^2 + O\left(\left(\frac{q_r - p_r}{p_r}\right)^3\right)$$

But then

$$q_r\log\left(\frac{q_r}{p_r}\right) = \left(q_r\right)\left[\frac{q_r - p_r}{p_r} - 2\left(\frac{q_r - p_r}{p_r}\right)^2 + O\left(\left(\frac{q_r - p_r}{p_r}\right)^3\right)\right]$$

Now set $\delta_r \coloneqq q_r - p_r$, and we know from what we did a few lines ago that $\delta_r = O\left(\frac{1}{n}\right)$.

$$q_r\log\left(\frac{q_r}{p_r}\right) = \left(p_r + \delta_r\right)\left[\frac{\delta_r}{p_r} - \frac{1}{2}\left(\frac{\delta_r}{p_r}\right)^2 + O\left(\left(\frac{\delta_r}{p_r}\right)^3\right)\right]$$

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

Note however that when we sum over r to get $D(q^*||p)$, the δ_r term vanishes since those add to 0, and since $\delta_r = O\left(\frac{1}{n}\right)$, $D(q^*|\,|p) = O\left(\frac{1}{n^2}\right)$ and therefore $nD(q^*|\,|p) = O\left(\frac{1}{n}\right)$. Since $q_r^* - p_r$, the log terms stay bounded.

Therefore we have

$$\log\left(p\left(m\right)\right)=\frac{1-k}{2}\log\left(n\right)-nD\left(q\mid\mid p\right)-\frac{1}{2}\sum_{r=1}^{k}\left[\log\left(q_r\right)\right]+O(1)$$

From earlier, and

$$\log\left(p\left(m^*\right)\right)=\frac{1-k}{2}\log\left(n\right)-\frac{1}{2}\sum_{r=1}^{k}\left[\log\left(q_r^*\right)\right]+O(1)$$

Since the terms $+\frac{1}{2}\sum_{r=1}^{k}\left[\log\left(q_r^*\right)\right]+O(1)$ stay bounded, we reduce to

$$\log\left(p\left(m^*\right)\right)=\frac{1-k}{2}\log\left(n\right)+O(1)$$

Therefore

$$\log\left(\frac{p\left(m\right)}{p(m^*)}\right)=-nD\left(q\mid\mid p\right)-\frac{1}{2}\sum_{r=1}^{k}\left[\log\left(q_r\right)\right]+O(1)$$

Therefore we just need to bound $\sum_{r=1}^{k}\left[\log\left(q_r\right)\right]$. Outside of K, this is fine, as each q is bounded. Therefore we have that

$$\log\left(\frac{p\left(m\right)}{p(m^*)}\right)=-nD\left(q\mid\mid p\right)+O(1)$$

Therefore, if n is large enough, $\frac{p(m)}{p(m^*)}\leq Ce^{-nD(q\mid\mid p)}$ outside K for some C provided n is large enough.

Since $\log\left(p\left(m^*\right)\right)=\frac{1-k}{2}\log\left(n\right)+O(1)$

It means that for n large enough, there exists positive constants A and B such that

$$\mathrm{An}^{\frac{1-\mathrm{k}}{2}}<p\left(m^*\right)<\mathrm{Bn}^{\frac{1-\mathrm{k}}{2}}$$

Therefore, we have the bound (for some positive constants A and B).

Therefore, because outside K, $nD\left(q\mid\mid p\right)\geq\frac{n^{2\varepsilon}}{3p_{max}}$, $\frac{p(m)}{p(m^*)}\leq Ce^{-cn^{2\varepsilon}}$ so the lemma is proved outside K.

Inside K, $p\left(m\right)\leq e^{-d_0n}$ so $\frac{p(m)}{p(m^*)}\leq Cn^{\frac{k-1}{2}}e^{-d_0n}\leq e^{-\frac{d_0}{2}n}\leq e^{-\frac{d_0}{3}n^{2\varepsilon}}$ for some C (because $\varepsilon<\frac{1}{2}$) and any large enough n. The second inequality is because $Cn^{\frac{k-1}{2}}$ eventually is beaten by $e^{\frac{d_0}{2}n}$ (take the ratio and apply lhopital at least $\frac{k-1}{2}$ times as n goes to infinity to see this) This proves the lemma in all cases.

□

Proof of part 3 of the claim:

Pick A from the statement of the lemma to be $\sqrt{cp_{min}}$ where c is as in the lemma 2 statement and p_{min} is the smallest probability in the table.

Define R^C as the outside of R. Pick $m_0\in S\cap R_2$, then $P\left(m\notin R|m\in S\right)=\frac{\sum_{m\in S\cap R^C}p(m)}{\sum_{m\in S}p(m)}\leq\frac{\sum_{m\in S\cap R^C}p(m)}{p(m_0)}$. Let |S| be the number of lattice points in S. Then $\sum_{m\in S\cap R^C}p(m)\leq|S|\sup_{\mathrm{m}\notin\mathrm{R}}p(m)$. By lemma 2, if n is large enough, $\sup_{\mathrm{m}\notin\mathrm{R}}p(m)\leq|S|Cp\left(m^*\right)e^{-cn^{2\varepsilon}}$. By part 2, if n is large enough, since $m_0\in R$ we can force $p\left(m_0\right)\geq\frac{1}{2}p\left(m^*\right)e^{-\frac{1}{2}|y_0|^2}$ where y_0 is the vector m_0 moved to the standardized world.

Now since the standardization is $y_i=\frac{m_i-E_i}{\sqrt{np_i}}$ we have

$$|y_0|^2=\sum_{i=1}^{k}\frac{\left(\left(m_0\right)_i-E_i\right)^2}{E_i}\leq\frac{1}{np_{min}}\sum_{i=1}^{k}\left(\left(m_0\right)_i-E_i\right)^2=\frac{|m_0-E|^2}{np_{min}}$$

Since $m_0\in R_2$, $|y_0|^2\leq\frac{1}{np_{min}}cp_{min}\ n^{1+2\varepsilon}$ by how R_2 is defined, and this is just $|y_0|^2\leq cp_{min}\frac{n^{2\varepsilon}}{p_{min}}$. Therefore we have that $p\left(m_0\right)\geq\frac{1}{2}p\left(m^*\right)e^{-\frac{1}{2}cn^{2\varepsilon}}$. Going back to the original inequality and putting in all our results, we get the following:

$$P\left(m\notin R|m\in S\right)\leq\frac{\sum_{m\in S\cap R^C}p\left(m\right)}{p\left(m_0\right)}\leq|S|\frac{Cp\left(m^*\right)e^{-cn^{2\varepsilon}}}{\frac{1}{2}p\left(m^*\right)e^{-\frac{1}{2}cn^{2\varepsilon}}}=2C\left|S\right|e^{-\frac{1}{2}cn^{2\varepsilon}}$$

Now there are n+1 possibilities for each cell and k cells so we get the trivial upper bound which says, therefore $|S| \leq (n+1)^k$

$$P(m \notin R | m \in S) \leq 2C(n+1)^k e^{-\frac{1}{2}cn^{2\varepsilon}}$$

Again exponentials beat polynomials so the above goes to 0, completing the proof of the proposition. So done.

Proof that this implies chi squared tests work:

Proof. Setup: S is the space we are constrained to which may or may not contain constraints other than only the total. Let $S = a + V$ where V is a space parallel to S and a is a line perpendicular to S. Then every z in S satisfies $z = a + v$ and $z^2 = a^2 + v^2$. Write $L = |a|$. Also if y is in the standardized world, then for any space A contained in S and the central region (in fact we care about the squared distance so we only need A to be a circle, so there is no issue with A being some weird fractal region or anything like that), if Λ_n is the set of lattice points in the standardized world where there are no additional constraints, and R is the central region in the space we are constraining to, then we have the following identity by claims 1, 2 and 3:

$$P(y \in A | y \in S) = \frac{\sum_{z \in \Lambda_n \cap A} \phi(z) + o(1) \sum_{z \in \Lambda_n \cap R} \phi(z)}{\sum_{z \in \Lambda_n \cap S} \phi(z) + o(1) \sum_{z \in \Lambda_n \cap R} \phi(z)}$$

Since the probability of being in the central region R tends to 1 we have the identity in the image above where $\phi(z)$ is proportional to $e^{-\frac{1}{2}|z|^2}$ where I mean the length of z in the standardized world.

Now let A be a circle/sphere in S with center a and A and S both having dimension equal to the degrees of freedom count. Then $P(y \in A \mid y \in S) = \frac{\sum_{z \in A} Ce^{-\frac{1}{2}\left(a^2+L^2\right)}}{\sum_{z \in S} Ce^{-\frac{1}{2}(a^2+L^2)}}(1+o(1)) = \frac{\sum_{z \in A} e^{-\frac{1}{2}\left(L^2\right)}}{\sum_{z \in S} e^{-\frac{1}{2}(L^2)}}(1+o(1))$ which we hope to show is $\frac{\int_{\text{In radius of A}} e^{-\frac{1}{2}L^2}}{\int_S e^{-\frac{1}{2}L^2}} \to P(\text{A genuine normal is in radius of A} \mid \text{In S})$, as then we will have a universal cumulative distribution for this distance A which gives the chi squared distribution when squared. We will simply show that the sums approach the integrals as n goes to infinity then we will be done.

Important point: Subject to a constraint the expected cell is always in the middle of the lattice we are constraining to, ie exactly E+a. Therefore the square of the length away from this point in the standardized world is actually $\frac{(O-E)^2}{E}$ and not off because E is away from the center of the normal distribution we are looking at – hopefully this makes sense.

Start with the situation where all constraints are imposed, then add even more until only a single point is allowed. Now imagine removing the constraints one by one until we are back to where we started.

Each time we remove one constraint, we gain a new freedom of movement: from any allowed point, we can now move in one new straight direction and still remain within the allowed set. Crucially, this direction is genuinely new — it is not just a combination of the directions we already had — because if it were, removing the constraint would not have created any new reachable points.

After all constraints have been removed, we are left with a flat space (the slice we are working in), and there is a fixed collection of directions with the following property:

1. Starting from any allowed point, we can reach any other allowed point by moving some whole number of steps forward or backward along each of these directions.

2. None of these directions is redundant: each one adds a new way of moving that cannot be recreated using the others.

Now fix one allowed point P. From P, draw a line segment to the point obtained by moving one step in the first direction. Do the same for each of the other directions. Using these line segments as edges, we obtain a solid shape (a parallelepiped): it consists of all points you can reach from $P$by moving partway along each direction, but by no more than one step in any of them.

We call this shape a **repeating cell**: it has the property that if we shift it by any whole-number combination of the allowed directions, the shifted copies fill the entire constrained space without gaps or overlaps. Every allowed point lies in exactly one such shifted copy. This has no points in its interior – This is easy to see if you imagine starting from the point and then when you remove the first constraint connect it to the nearest point, when you remove the second one connect

this line to the nearest identical line, and so on. So we can actually tile S with repeating cells such that the points in the lattice are exactly the vertices.

Thus, the constrained space can be tiled by identical parallelepipeds built from these directions, and such a repeating cell always exists.

Now recall that we define the riemann integral in higher dimensions as the limit of the sum of the value of the point at one corner of the tile times the volume of the tile. In the standardized world, distances are divided by $\sqrt{n}$ so the volume converges to 0, and for fixed $\delta > 0$ the parallelopipeds are eventually all contained in a ball of radius δ. Since the normal distribution pdf is uniformly continuous inside R by the technical results document, it means that for any ε we can make the amount the function changes within each repeating cell as small as we want, and therefore we can make the difference of the integrals on our finite region as close as we want to the weighted sum of the points. This uniform continuity is actually why these riemann integrals on continuous functions on a closed set work similarly to what we showed in level 4.

□

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 (which is usually when we have small expected values), 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.

Figure 9 shows Po(3, 4, 5, 6) on the same graph, demonstrating that around 5 is when it starts to look like a normal distribution.

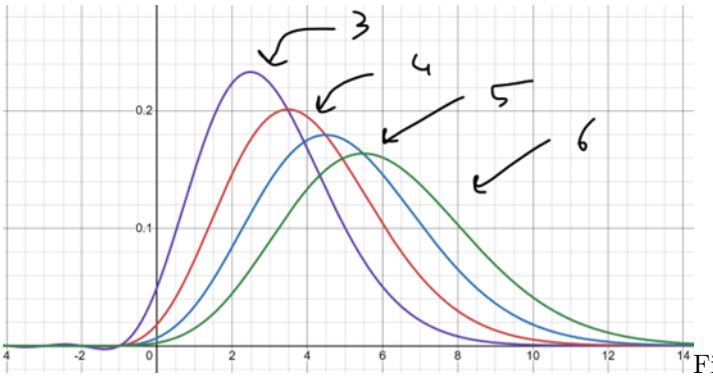

Figure 9

9 Other properties of the chi squared and t distributions

Proposition. The formula $(n-1)\,\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$ is true for normally distributed variables

Proof. 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

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

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

Therefore,

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

- The vector $(Z_1, ..., Z_n)$ 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 $(1, 1, ..., 1)$
- The quantity $\sum_{i=1}^{n} \left(Z_i - \bar{Z}\right)^2$ is the squared length of the component orthogonal to $(1, 1, ..., 1)$

By rotational symmetry, the "length squared" of this orthogonal bit 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. Conditioning is allowed as it is a continuous distribution and issues came from using a discrete distribution.

□

Proposition. Suppose we have observations from 2 normal distributions X and Y with the same variance and we want to test the difference between their means, and we have n_X and n_Y observations from each variable respectively. Then if we define

$$s_p^2 = \frac{(n_x - 1)\, s_{xx} + (n_y - 1)\, s_{yy}}{n_x + n_y - 2}$$

Then the random variable (conditional on μ_x and μ_y being the means of X and Y respectively)

$$T = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{s_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

has a t-distribution with parameter $n_X + n_Y - 2$

Proof. Note: People call the parameter of the t distribution the number of degrees of freedom but I'm not really sure how the degrees of freedom relate to the dimension here, so I will merely consider it a parameter.

Note that a t_{n-1} distribution is given by $\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}$ 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:

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:

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

The reason is because the variance of $n_x\overline{X}$, 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 $n_x(\overline{X} - \mu)$. Since multiplying a variable by the square of a constant multiplies the variance by that constant, we get that that the variance of $\overline{X} - \mu$ must be $\frac{\sigma^2}{nx}$. Similarly, we have that the variance of $\overline{Y} - \mu$ 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:

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. For example, as mentioned in level 5.2 it is easy to see that the t distribution does not depend on the variance - it is clear if we fix the variance to 1, and it stays the same when we rescale it.

□