This lemma was used in my old clunky proof of chi squared tests from level 6 so it has be moved to this section.

We need to define some theory to state the theorem.

Say I have a vector of random variables $X_1, X_2, \dots X_k$. These do not have to be independent nor identically distributed. Then the covariance matrix, typically denoted by Σ is such that the entry in the i'th row and j'th column is equal to $E((X_i - E(X_i))(X_j - E(X_j))$ which we showed at the beginning of Level 6 stats equals $E(x_i x_j) - E(x_i)E(x_j)$. First, we know that Σ is symmetric. Also, by definition, the n'th diagonal entry of Σ is given by $Var(X_n)$, and the off diagonal entries, say in the i'th row and j'th volumn, are given by $Cov(X_i, X_j)$. Alternatively, you can think of the covariance matrix of a random vector Z as $E\left((Z - E(Z))(Z - E(Z))^T\right)$ as this is equivalent. We will consider the chi squared table to be a random vector.

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

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

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

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

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

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

Proof:

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

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

2. **Expand Z:**

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

So:

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

3. **Expand the product:**

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

4. **Take expectations:**

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

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

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

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

Theorem: Take n samples of a random vector X. As n goes to infinity, $\sqrt{n}(\bar{X} - p)$ converges in distribution to a normal distribution with mean 0 and finite covariance Σ. Furthermore assume that the distribution of X is at worst a mix of discrete and continuous parts since the supporting parts from this website assume this. Note that here, $\bar{X}$ is the component-wise sample mean of n random vectors, and p is the vector with entries equal to the expected values of each of the components.

Recommended levels for proof: 6 (Normal central limit theorem is used) and also the proof of the Cramer Wold theorem from the misc results section. Once we have these, the result is quite easy.

If we have a random vector X, then we first shift X so that its mean is 0, then we have the following

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

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

Also we have finite covariance by assumption so Var(a.x) is finite.

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

Example of an application: We give an alternative proof from the one in level 6 that chi squared tests work in the simple case with no additional constraints other than the total.

Recommended levels for this: 7 (Just 6 + the projection matrix formula which we cover in the level 7 vectors and matrices course)

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

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

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

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

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

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

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

$$I - \begin{pmatrix} \sqrt{p_1} & \sqrt{p_2} & \dots & \sqrt{p_k} \end{pmatrix} \begin{pmatrix} \sqrt{p_1} \\ \sqrt{p_2} \\ \vdots \\ \sqrt{p_k} \end{pmatrix}$$

because that is the standard formula for a projection matrix. The proof for this is in the technical results document.

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

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

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