Requirement for proof: Knowledge of multivariate CLT from this section and its shown application, itself depending on knowledge of cramer wold theorem from this section, itself depending on knowledge from all of level 6.

Where we are: We proved chi squared tests for a standard multinomial model by showing that $\sqrt{n}\left(\frac{\bar{x}-p}{\sqrt{p}}\right)$ converged to a standard unit normal under the null hypothesis but need to prove that this holds if $\bar{x}$ is constrained to some linear space, which should work because it is just projecting the standardized vector which is a normal distribution to get a lower dimensional normal distribution. An important idea we will use a lot is that to standardize we multiply by $\sqrt{n}$ and enlarge by constants on the different axes.

The problem: There are several possible ways that the distribution when confined to a slice would not be standard normal even if the unconfined distribution is. An example is if just one slice were wrong as that slice has vanishingly small probability as n gets large.

Terminology: I refer to a plane or clice as the set of observed counts we are constraining ourselves to and I am considering hopping between parallel planes.

Goal. Show that inside the high-probability “central band”, the probability density function cannot vary enough across adjacent lattice slices to allow an alternating (“slice-by-slice”) pathology, so conditioning/projection won’t hide a different limit distribution. We will do this by considering adding a single constraint, as if we had multiple, we could consider ourselves to be adding one more constraint at a time and repeatedly arguing that we have a normal distribution.

Another problem is if the distributions oscillated between the slices but as more slices were added the cumulative distribution still converged to the standard normal (which we know it does), ie the cumulative distribution for varying values of n would look like the image below: it really does converge to the limit but for any fixed value of n we could not interchange when we project and when we take the limiting distribution.

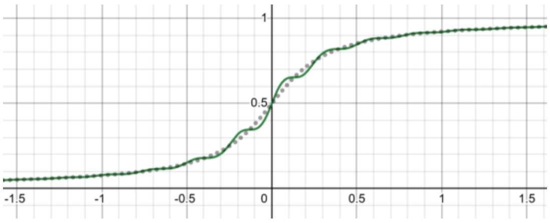

Image to demonstrate the mode of failure with a cdf plot, imagine the oscillations got more frequent as n got large so we genuinely converged pointwise to the dotted line.

Note that if the probabilites are all off from the limiting distribution by a factor (for example the total probability mass is more than 1.1* what the limiting distribution predicts) in some rectangular region, this will be reflected in the limiting distribution if and only if there exists a constant C such that the volume of this region after scaling by $\sqrt{n}$ to get to the standardized normal world is more than a constant C no matter how large n is, and therefore there must exist a constant C such that the total probability mass is (in this example) more than 1.1* what the model predicts in a circle of radius $C\sqrt{n}$ in the raw counts world.

We will define a plane hop by doing the minimal number of observed count swaps to get from one plane to the nearest parallel plane. For example, if our constraint is about estimating the mean, then the planes we hop between correspond to the different estimated means, and moving something from the i'th cell to the j'th cell changes the estimated mean, and some combination of such movements changes the estimated mean to its nearest possible value, and this would be the hop. THE HOP PROCEDURE DOES NOT DEPEND ON N.

We will compute ratio of the probabilities between two chi squared tables differing only by moving one count from one cell to another

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

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

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

Now lets suppose we are always within, say, 10 standard deviations of the mean, as this is almost certain, and we could increase "10" as we want to get that the theorem holds with any tolerance of error we want.

Consider doing trials from the chi squared table repeatedly until we happen to be on our constraint slice (We will call this S), and let A be a rectangular subset of S. Then $P(x \in A) = \frac{P(x\in A \mid x\in S)}{P(x\in S)}$. Note that $P(x \in A \mid x \in S)$ is what we care about – we want it to be that for all rectangular regions A that have non-zero volume in our k-1 dimensional space (Since we suppose that before adding an additional constraint we are in k dimensions and we proved as above that the limiting distribution there is a normal), the limit of the probability mass is what the standard normal would predict.

As an example, suppose are constraint comes from estimating the mean based on data, meaning that S is the condition Mean=m. There are many such planes corresponding to different values of m, and there exists a sequence of cell moves (We will call this a hop) to get from the Mean=m to the Mean=m+ε plane where m+ε is the smallest possible mean greater than m.

Note that there is a one-to-one correspondance between cells in the Mean=m and the Mean=m+ε planes (at least, whenever there isn't that is only because the cells are so close to 0 that for large enough n this has vanishingly small probability). This is because the hop we do from one point in the Mean=m plane increases the mean by ε, and since the mean changes linearly with each cell, it will increase by ε if we start from any point in the Mean=m plane and apply the hop which just changes the cell counts by a fixed amount. To justify this carefully, say we are starting in our slice by only considering things where all observed counts are within a factor of 2 from the expected and then move them by hops the hops give a one to one correspondence between subslices that all contain their part in the 10 standard deviations region – for large enough n we will leave the 10 standard deviations region (On the order of $\sqrt{n}$ counts) way before our subslice will move out of it (This requires on the order of n counts).

Concretely, we take our slice, cut it so it only has the vectors between $\frac{E}{2}$ and E and translate it by hops until we are out of the region. A translation is a one-to-one correspondence and for n large enough, since we do on the order of $\sqrt{n}$ fixed translations, the $\frac{E}{2}$ vectors will not go below 0.

Let the shortest distance between the starting and destination planes after a hop be T. Then there exists a constant T' such that in the standardized world, this distance is $\frac{T'}{\sqrt{n}}$.

Lemma: Inside the "within 10 standard deviations" region, if we want to change the probability of a cell by a factor of D by doing hops there exists a constant C such that it takes $C\sqrt{n}$ hops to do so.

Proof: $\log(O_i) - \log(P_i) - \log(O_j + 1) + \log(P_j)$ is bounded above by

$$\log(E_i + C_1\sqrt{n}) - \log(P_i) - \log(E_j + C_2\sqrt{n}) + \log(P_j)$$

And below by $\log(E_i - C_1\sqrt{n}) - \log(P_i) - \log(E_j - C_2\sqrt{n}) + \log(P_j)$

Because the "10 standard deviations circle" in the standardized world exists as some ellipse in the counts world with dimensions proportional to $\sqrt{n}$, so C_1 and C_2 are the bounds to stay inside this ellipse.

Thus our bounds are $\log(n \pm C_1'\sqrt{n}) - \log(n \pm C_2'\sqrt{n})$ since $\frac{E_i}{P_i} = n$. By standard logarithm rules, this is just $\log\left(1 \pm \frac{C_1'}{\sqrt{n}}\right) - \log\left(1 \pm \frac{C_2'}{\sqrt{n}}\right)$. Since if x is small enough there exists constants R_1 and R_2 with 1 between them $R_1 x < \log(1 + x) < R_2 x$ by the definition of the derivative and the fact that the derivative of $\log(1 + x)$ at 0 is 1. Therefore there exists a constant C' such that the log ratio after one cell move is bounded above in absolute value by $\frac{C'}{\sqrt{n}}$. Since there is a fixed number of cell moves in a hop, the log ratio after one hop is bounded in absolute value above by $\frac{C''}{\sqrt{n}}$ for some C'' still not depending on n.

Now to complete the proof of the lemma, note that to change the probability by a factor of D, we will need to change the log probability by $\log(D)$ and thus will need $\frac{\log(D)}{c''}\sqrt{n}$ hops. Therefore, we can set $C = \frac{\log(D)}{c''}$ and the proof of the lemma is completed.

Now suppose that we have exactly what we want to prove cannot happen: That there exists an ε>0 such that we have a rectangular patch A with k-1-dimensional-volume V in the standardized world such that under the limiting standardized distribution of the constrained to S vector we have $\left|\frac{P(A|S)}{P'(A)} - 1\right| > \varepsilon$ where P'(A) is the probability we would predict A to have under the normal distribution thing we are trying to prove. means that our distribution is actually a standard normal. As discussed, we have convergence to a normal after constraining if this does not happen, since if the total probability mass in all regions were eventually within an arbitrarily small factor of a normal distribution, then the cdf must converge pointwise to the cdf of a normal distribution.

So suppose $\left|\frac{P(A|S)}{P'(A)} - 1\right| > \varepsilon$ happens. Then we aim to derive a contradiction: If we have no bad rectangular patches in the slice we will be done.

To proceed we need another lemma:

Lemma: Inside the "within 10 standard deviations" region, if we want to change the probability mass of a region of cells (in particular S, or any hypothetical "bad region") by a factor of D by doing hops there exists a constant C such that it takes $C\sqrt{n}$ hops to do so.

Proof: By the previous lemma, the log of the probability of each specific table in the region changes by a ratio of at most $\frac{D}{\sqrt{n}}$ for some constant D on each hop. Set $\frac{D}{\sqrt{n}} \coloneqq \delta_n$. Now take the sum of all the probabilities in our rectangular region A: This will be $P(A) = P_{A1} + P_{A2} + \cdots + P_{Af(A,n)}$ where f(A,n) is the number of possible tables in the region A. Now if we take all these from a hop, since the log ratio changes by at most $\frac{D}{\sqrt{n}}$, the actual ratio of each new probability from the old one is between $e^{-\delta_n}$ and e^{δ_n}. Therefore, if the possibility Ak goes to Ak' after a hop, and the region A goes to region A' after a hop, we have the inequality $P_{A1}e^{-\delta_n} + P_{A2}e^{-\delta_n} + \cdots + P_{An(A)}e^{-\delta_n} < P(A') < P_{A1}e^{\delta_n} + P_{A2}e^{\delta_n} + \cdots + P_{An(A)}e^{\delta_n}$. Therefore, $\left|\log\left(\frac{P(A)}{P(A')}\right)\right| < \delta_n = \frac{D}{\sqrt{n}}$. Again I emphasise that C does not depend on n. Then applying the same argument as in the above lemma it takes $C\sqrt{n}$ hops for some constant C.

So if our bad region had $\left|\frac{P(A|S)}{P'(A)} - 1\right| > \varepsilon$, for example $\frac{P(A|S)}{P'(A)} > 1 + \varepsilon$, which we can rewrite as $\frac{P(A)}{P(S)P'(A)} > 1 + \varepsilon$, then what happens is that if hop-adjacent regions to A have $\frac{P(A)}{P(S)P'(A)} < 1 + \frac{\varepsilon}{2}$ then the log of $\frac{P(A)}{P(S)P'(A)}$ must have changed by at least $\left|\log\left(\frac{1+\varepsilon}{1+\frac{\varepsilon}{2}}\right)\right|$ which is constant. Therefore, the log of one of $P(A)$, $P(S)$ and $P'(A)$ must have changed by at least $\frac{1}{3}\left|\log\left(\frac{1+\varepsilon}{1+\frac{\varepsilon}{2}}\right)\right|$ which is constant. The first 2 take at least $C\sqrt{n}$ hops for some fixed C by the lemma, but we need to investigate the rate of change of $\log\big(P'(A)\big)$. If our hop that moved us from A to A' perpendicular to the slice in the standardized world, then by symmetry of the multivariate normal, $P'(A) = P'(A')$ so the log does not change and we are done. However, in reality, it may drift by some fixed number of count units in distance from the perpendicular point on the next plane over. This corresponds to $\frac{B}{\sqrt{n}}$ units of distance in the standardized world for some fixed constant B.

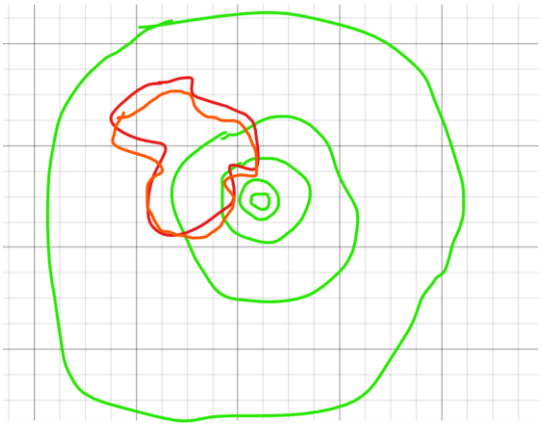

Image: Attempt to visualize moving an arbitrary region by a tiny amount (down from red to orange) with contour plots of a 2D normal drawn. Don't worry if this does not make any sense because of my bad drawing.

Note that the maximum derivative of the 2D – or anyD standard normal – is fixed and depends only on the dimension. Call this M. Then as the area A with volume V moves by $\frac{B}{\sqrt{n}}$ units, its probability mass will change by at most $\frac{BVM}{\sqrt{n}}$, as since the maximum derivative is M, there is no place where moving A by $\frac{B}{\sqrt{n}}$ units changes the probability density of the normal under it by more than $\frac{MB}{\sqrt{n}}$ units and so since the area of A is V the total probability mass of A will change by at most $\frac{BVM}{\sqrt{n}}$. Then again, there exists a constant K (I'm running out of letters) such that if n is large enough, and the probability mass of A is originally X, then in order to change this X by a factor of D, we have to do at least $\frac{X(|D-1|)}{\frac{BVM}{\sqrt{n}}}$ hops, which is a fixed multiply of $\sqrt{n}$. In our case, $D = e^{\frac{1}{3}\left|\log\left(\frac{1+\varepsilon}{1+\frac{\varepsilon}{2}}\right)\right|}$.

Therefore, to change any 3 of the probabilities sufficiently by hopping to make A not be a bad region anymore, there exists a constant C such that we need $C\sqrt{n}$ hops to get the A discrepency from normal ratio to be less than $1 + \frac{\varepsilon}{2}$.

Recall that at the beginning we let the shortest distance between the starting and destination planes after a hop be $T'n^{-\frac{1}{2}}$.

Recall we are working in standardized coordinates $X_n = \sqrt{n}(\bar{x} - p)/\sqrt{p} \in \mathbb{R}^k$ (where this notation means we are working with k dimensional vectors), and we are considering "slices" perpendicular to the hop direction. Let $u$denote the standardized coordinate in the hop direction (perpendicular to the slice), and let $v \in \mathbb{R}^{k-1}$be coordinates within the slice.

Let $A \subset S$be a fixed $(k-1)$-dimensional rectangular region inside a slice (in the standardized world), and suppose $A$is "bad" in the sense that for infinitely many n,

$$\frac{P(X_n \in A \mid X_n \in S)}{P'(V \in A)} > 1 + \varepsilon,$$

where P'denotes the $(k-1)$-dimensional standard normal prediction in the slice coordinates v. (Here $V \sim N_{k-1}(0, I)$.)

By the previous lemmas, within the central band (say $| \ u \ | \leq 10$), it takes at least $C\sqrt{n}$hops to reduce the discrepancy below $1 + \varepsilon/2$. Therefore, for sufficiently large n, we can find $m = \lfloor C\sqrt{n} \rfloor$ (this means the largest integer less than $C\sqrt{n}$) consecutive hop-adjacent slice-pieces $A_1, \dots, A_m$, each congruent to A up to the small drift described earlier, such that each remains "bad" with ratio $> 1 + \varepsilon/2$. Let

$$Q_n \ = \ \bigcup_{j=1}^{m} A_j \, ,$$

which is a "slab" obtained by stacking A along the hop direction. In standardized coordinates, the hop step has length $T'/\sqrt{n}$, hence the slab thickness is $m \cdot (T'/\sqrt{n}) \to CT'$. Thus Q_n converges (in the obvious geometric sense) to a fixed slab

$$Q = \{(u, v) : u \in [u_0, u_0 + CT'], v \in A\}$$

Where u and v are coordinates that mean u is the hop number and v is a vector in A,

for some $u_0$lying in the central band.

Let $Z \sim N_k(0, I_k)$denote the k-dimensional standard normal limit in standardized coordinates. Fix the hop direction as a **unit** vector $a \in \mathbb{R}^k$, and define the scalar "slice coordinate"

$$U \coloneqq a^{\mathsf{T}} Z \in \mathbb{R}.$$

Next choose any linear map $B: \mathbb{R}^k \to \mathbb{R}^{k-1}$(represented by a $(k-1) \times k$ matrix) whose rows span (ie, adding and multiplying the rows together gievs you any vector in) the directions we are using to coordinatize a slice.)

Since (u, v) is a linear image of a multivariate normal vector, it is jointly normal. Compute the covariance blocks:

$$\mathrm{Var}(U) = a^{\top} I a = 1, \mathrm{Cov}(V) = BB^{\top}.$$

$$\mathrm{Cov}\ (V_i, U) = \mathrm{Cov}\ (BZ, a^{\top} Z) = B\mathrm{Cov}\ (Z, Z) a = Ba.$$

Denote $c \coloneqq \mathrm{Cov}\ (V, U) = Ba \in \mathbb{R}^{k-1}$and $\Sigma \coloneqq \mathrm{Cov}\ (V) = BB^{\top}$.

Define the **drift-corrected residual**

$$R \coloneqq V - cU \in \mathbb{R}^{k-1}.$$

Then $R$is uncorrelated with U:

$$\mathrm{Cov}\ (R, U) = \mathrm{Cov}\ (V - cU, U) = \mathrm{Cov}\ (V, U) - c\, \mathrm{Var}\ (U) = c - c \cdot 1 = 0.$$

Because $(U’ R)$is jointly normal, “uncorrelated” implies “independent”, hence

$$U\ \perp\ \boxed{,}\boxed{,}\boxed{,} \perp\ R.$$

Also,

$$\mathrm{Cov}\ (R) = \mathrm{Cov}\ (V - cU) = \Sigma - cc^{\top}.$$

(This matrix is positive semidefinite; in our setting we may restrict to a $(k{-}1)$-dimensional slice coordinate system so that it is positive definite on that subspace.)

Now choose an invertible $(k-1) \times (k-1)$matrix Lsuch that

$$L(\Sigma - cc^{\top})L^{\top} = I_{k-1}(\mathrm{e.g.}\ L = (\Sigma - cc^{\top})^{-1/2}).$$

Define

$$W \coloneqq LR.$$

Then $W \sim N_{k-1}(0, I_{k-1})$and, since $W$is a function of R, we still have

$$U\ \perp\ \boxed{,}\boxed{,}\boxed{,} \perp\ W.$$

Interpretation (this is the drift): the conditional law of Vgiven $U = u$is a translated normal. Indeed for a jointly normal pair, the conditional mean is linear:

$$\mathbb{E}[V \mid U = u] = \mathrm{Cov}\ (V, U)\mathrm{Var}\ (U)^{-1} u = cu,$$

and the conditional covariance is constant:

$$\operatorname{Cov}(V \mid U = u) = \Sigma - cc^{\top}.$$

So the dependence between "slice coordinate" $U$and "within-slice coordinates" $V$is exactly the linear drift cu. Subtracting cUremoves this drift, and $W$is the **drift-free within-slice standard normal coordinate**.

Now fix a within-slice region $A \subset \mathbb{R}^{k-1}$**expressed in the W-coordinates**. For any thin u-interval $I_k = [u_k, u_k + \Delta u]$, define the corresponding Gaussian slice-piece

$$A_k^{(G)} := \{ Z : U \in I_k, W \in A \}.$$

Since $U$and $W$are independent, its Gaussian probability factorizes **without any orthogonality assumption**:

$$P(Z \in A_k^{(G)}) = P(U \in I_k)\, P(W \in A).$$

This is the correct "Gaussian baseline" for a slice-piece: the u-dependence is entirely in the weight $P(U \in I_k)$, while the within-slice factor $P(W \in A)$is constant (because we removed the drift).

Finally, if $Q_n$is a union of disjoint slice-pieces $A_1, \dots, A_m$(disjoint because the $I_k$are disjoint), then

$$P(Z \in Q_n) = \sum_{k=1}^{m} P(Z \in A_k^{(G)}) = P(W \in A) \sum_{k=1}^{m} P(U \in I_k),$$

and this is the weighted sum we will compare against the corresponding probabilities for X_n.

Now we may sum over slice pieces.
Since the slice pieces A_k(hence $\tilde{A}_k$) lie in disjoint U-intervals I_k, the sets $\tilde{A}_k$are disjoint. Therefore

$$P(Y_n \in \tilde{Q}_n) = \sum_{k=1}^{m} P(Y_n \in \tilde{A}_k),$$

where $\tilde{Q}_n := \bigcup_{k=1}^{m} \tilde{A}_k$.
(Continue from here with your ratio/lower-bound step, comparing each $P(Y_n \in \tilde{A}_k)$to its own Gaussian baseline $P(Y \in C_k)$, and then summing.)

On the other hand, by disjointness of slices,

$$P(X_n \in Q_n) = \sum_{k=1}^{m} P(X_n \in A_k).$$

And the statement “A_k is bad by a factor $> 1 + \varepsilon/2$” means **relative to the correct Gaussian baseline for that slice-piece**. Concretely, writing $I_k$for the u-interval corresponding to the k-th slice (of width $T'/\sqrt{n}$), the Gaussian prediction for $A_k \subset \{u \in I_k\}$is

$$P(Z \in A_k) = P(U \in I_k)\, P(V \in A) \approx \phi_1(u_k)\frac{T'}{\sqrt{n}}\, P(V \in A),$$

with a uniform approximation because $\phi_1$is smooth and we are inside the central band. Therefore for all large $n$and all $k = 1, \dots, m$,

$$P(X_n \in A_k) \geq (1 + \varepsilon/2)\, P(Z \in A_k),$$

hence summing gives

$$P(X_n \in Q_n) \geq (1 + \varepsilon/2) \sum_{k=1}^{m} P(Z \in A_k) = (1 + \varepsilon/2)\, P(Z \in Q_n).$$

Passing to the limit $n \to \infty$, $Q_n \to Q$and $P(Z \in Q_n) \to P(Z \in Q)$. But the k-dimensional CLT says $P(X_n \in Q) \to P(Z \in Q)$for such slab/rectangular sets (boundary has Gaussian measure 0), contradicting the inequality above. This contradiction shows the assumed “bad slice” cannot exist, so the conditional distribution on the slice must converge to the $(k-1)$-dimensional standard normal as desired.