Uniwersytet Wrocławski
Wydział Matematyki i Informatyki
Instytut Matematyczny
specjalność: analiza danych

Szymon Czop

Cumulants and χ^2-conjecture

Praca licencjaca
napisana pod kierunkiem
dr inż. Wiktor Ejsmont

Wrocław, 2019

Contents

1 Introduction

Let $X_1, X_2, X_3, ..., X_n$ be i.i.d random variables, from standard normal distribution. Then the sample variance $Q_n = nS_n^2 = \sum_{i=1}^n (X_i - \overline{X})^2$ has χ^2 distribution with n-1 degrees of freedom. It is still unknown for classical probability, if there exist any characterisation for all distributions with this property and especially whether it characterise the variables from normal distribution. This problem was announced in work of Kagan, Linnik and Rao [7, page 466] and it's name nowadays is known as $\chi^2 - conjecture$.

Conjecture 1.1. If $X_1, X_2, \dots, X_n$ are independently and identically distributed classical random variables with finite non-zero variance σ^2, then a necessary and sufficient condition for X_1 to be normal is that $\sum_{i=1}^n (X_i - \overline{X})^2/\sigma^2$ be distributed as classical chi-square distribution with $n-1$ degrees of freedom.

The problem above was previously analysed by several authors. Ruben[11] was the first who proved this conjecture under condition that X_i is symmetric or number of variables is equal to two (without assumption about the symmetry). It is not proved yet, if symmetry hypothesizes for the variables, can be dropped for the n $>$ 3.
Later on [12] Ruben, using combinatorial tools, showed that the general symmetry assumption can be dropped, under condition that sum of squares of observations in sample about the sample mean, divided by σ^2 , give us chi-square distribution for two samples with sizes m and n, where $n \neq m$ and $m, n \geq 2$.
Proof of Ruben was firstly made on cumulants, and is is considered as intricate and complicated. Bonderson [1] made more direct proof of this, but he based his computations on moments of the sample variance.
Under additional assumption that $X_1, ..., X_n$ are independent infinitely divisible random variables, Golikova and Kruglow [5] showed that $\chi^2 - conjectute$ is always true.
Recently Ejsmont and Lehner [4] present a solution for this question in free probability.
In this paper we will take a look at the definition of cumulants as mathematical object. Consider it's analytic and combinatoric properties and some useful facts, that will help us in further part of the paper. First proof of Ruben for two variables will lead us true basic usage of cumulants. Second proof for symmetric variables will be based on some properties showed in first proof and mostly combinatoric properties of cumulants. New thing presented in paper is **cumulants based approach** to Golikova and Kruglow proof, for infinitely divisible variables.

2 Preliminaries

In the whole work we assume that all moments exists. This is necessary condition for existence cumulants.

Definition 2.1. *We will denote* ***sample variance*** *of finite random variables X_i as*

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\Big(1 - \frac{1}{n}\Big)\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\sum_{i,j=1,\ i\neq j}^{n} X_iX_j = \frac{1}{n^2}\sum_{1\leq i<j\leq n}(X_i - X_j)^2. \quad (1)$$

In order to simplify this complicated notation in the paper we will think about and call "sample variance " the quadric form $Q_n = nS_n^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2$.
Upon the assumption, that our random variables are from normal distribution it is true that sample variance and sample mean are independent statistics

Definition 2.2. *If X is a random variable then, the* ***expected value*** *is defined as Lebesegue integral:*

$$\mathbb{E}[X] = \int_{\mathbb{R}} x d\mu(x) \quad (2)$$

Definition 2.3. *If X is a random variable from symmetric distribution then:*

$$\forall_{n\geq 1}\ \mathbb{E}[(X - C)^{2n+1}] = 0 \quad (3)$$

Where C is the mean of X.

Fact 2.4. If random variable X comes from standard normal distribution than all it's even moments are equal to:

$$\forall_{n\geq 1}\ \mathbb{E}[X^{2n}] = (2n-1)!! \quad (4)$$

Fact 2.5. [2] If X_1 and X_2 have same distributions and are independent then $X_1 - X_2$ is a symmetric variable.

Definition 2.6. *A probability measure μ on $\mathbb{R}$ is said to be* ***infinitely divisible*** *(ID) if for $n \in \{1, 2 \ldots\}$ there exists a probability measure μ_n such that $\mu = \underbrace{\mu_n * \cdots * \mu_n}_{n\text{-times}}$.*

Definition 2.7. *Cumulants of random variable X can be calculated by* ***cumulant-generating function*** *$C(t)$, which is a natural logarithm of the moment generating function:*

$$C(t) := log(\mathbb{E}[e^{tX}]) \quad (5)$$

3 Cumulants

3.1 Analytic approach

Cumulants were firstly studied by Danish scientist T. N Thiele. At the beginning Thiele called them semi-invariants. The importance of the cumulants comes from observation, that many properties of random variables can be better represented by cumulants rather than moments. We refer to Peccati, Taqqu [10] and Nica,Speicher [9] for further detailed probabilistic aspects of this topic.

With given random variable X the i-th cumulant K_i is defined as:

$$K_i(X) := K_i(\underbrace{X, \dots, X}_{i-times}) = \frac{d^i}{dt^i}\Big|_{t=0} log(M(t))$$

where $M(t)$ is moment generating function for the variable.

This formula lead us to another analytic equation.

$$M(t) = \sum_{i=0}^{\infty} \frac{\mathbb{E}(X^i)}{i!} t^i = exp\Big(\sum_{i=1}^{\infty} \frac{K_i}{i!} t^i \Big)$$

The joint cumulant of several random variables $X_1, \dots, X_n$ of order $(i_1, \dots, i_n)$, where i_j are non-negative integers, is defined by a similar generating function $g(t_1, \dots, t_n) = E\Big(e^{\sum_{i=1}^n t_i X_i}\Big)$

$$K_{i_1 + \dots + i_n}(\underbrace{X_1, \dots, X_1}_{i_1-times}, \dots, \underbrace{X_n, \dots, X_n}_{i_n-times}) = \frac{d^{i_1 + \dots + i_n}}{dt_1^{i_1} \dots dt_n^{i_n}}\Big|_{t=0} log(g(t_1, \dots, t_n)),$$

where $t = (t_1, \dots, t_n)$.

3.2 Combinatoric approach

Let S be a finite subset of $\mathbb{N}$. By partition of S we, mean subsets (or *blocks*)$B_1, B_2, \dots, B_k, \subseteq S$ such that $B_i \cap B_j = \emptyset$, when $i \neq j$ and $B_1 \cup B_2 \cup \dots \cup B_k = S$. To make it more clear we will consider following example.
Set $\{1, 2, 3\}$ has following partitions :
$\pi_1 = \{\{1\}, \{2\}, \{3\}\}$
$\pi_2 = \{\{1, 2\}, \{3\}\}$
$\pi_3 = \{\{1, 3\}, \{2\}\}$
$\pi_4 = \{\{1\}, \{2, 3\}\}$
$\pi_5 = \{\{1, 2, 3\}\}$

Where π_i mean $i-th$ partition of our set.

Any partition defines an equivalence relation on S, denotes $\sim_\pi$ that such the equivalence classes are the blocks of π.That is, $i \sim_\pi j$ if i and j belong to the same block of π. In our example set $\{1,2\} \sim_\pi \{3\}$ for partition number two. The set of all S partitions is denoted as $\mathcal{P}(S)$. In case when $S = \{1,\dots,n\}$ we write $\mathcal{P}(n)$. $\mathcal{P}(n)$ is a poset under refinement order, where we say $\pi \leq \rho$ if every block of π is contained in a block of ρ. The maximal element of $\mathcal{P}(n)$ under this order is denoted by $\hat{1}_n$. Partition marked as this has only one block, in given example this is partition π_5 from above list. The opposite situation occurs when we define minimal element $\hat{0}_n$ .This sign defines us unique partition where every block of singleton is unique, π_1 in our example.

Definition 3.1. *Let $X_1,\dots,X_n$ be random variables . The cumulants are multi linear maps defined implicitly by the relation (connecting them with mixed moments)*

$$\mathbb{E}(X_1X_2\dots X_n) = \sum_{\pi\in\mathcal{P}(n)} K_\pi(X_1,X_2,\dots,X_n), \tag{6}$$

where

$$K_\pi(X_1,X_2,\dots,X_n) := \Pi_{B\in\pi}K_{|B|}(X_i : i \in B) \tag{7}$$

Sometimes we will write $K_r(X) = K_r(X,\dots,X)$. If all X are from the same distribution then:

$$K_1 = \mathbb{E}(X), \qquad K_2 = \mathbb{E}(X^2)-[\mathbb{E}(X)]^2 = \sigma^2, \qquad K_3 = \mathbb{E}(X^3)-3\mathbb{E}(X^2)\mathbb{E}(X)+2[\mathbb{E}(X)]^3$$

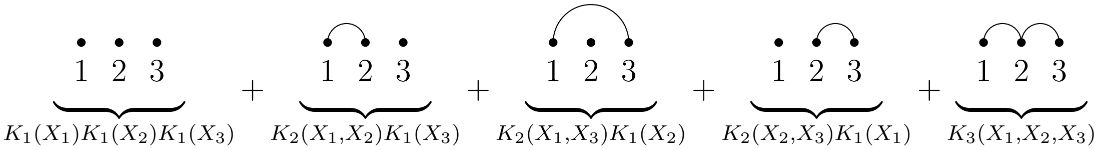
The best way to get the idea is to look at some examples:

$$\mathbb{E}(X_1) = K_1(X_1) \tag{8}$$

$$\mathbb{E}(X_1X_2) = K_1(X_1)K_1(X_2) + K_2(X_1,X_2) \tag{9}$$

$$\begin{aligned}\mathbb{E}(X_1X_2X_3) =& K_3(X_1,X_2,X_3) + K_2(X_1,X_2)K_1(X_3)\\ &+ K_2(X_2,X_3)K_1(X_1) + K_2(X_1,X_3)K_1(X_2)\\ &+ K_1(X_1)K_1(X_2)K_1(X_3)\end{aligned} \tag{10}$$

In further proofs and explanations we will mostly use graphical representation of partitions. Dots with a number will represent a random variable. Connection between two dots will mean that they are in the same partition. To make it more understandable we will show the example where we want to use cumulants to describe $\mathbb{E}(X_1X_2X_3)$.



Theorem 3.2. *[James,Leonov and Shiryaev [6, 8]] Let, $r, n \in \mathbb{N}$ and $i_1 < i_2 < \cdots < i_r = n$ be given and*

$$\rho = \{(1, \ldots, i_1), \ldots, (i_{r-1} + 1, \ldots, i_r)\} \in \mathcal{P}(n)$$

be the induced interval partition. Consider now random variables $X_1, \ldots, X_n$, then the cumulant of the products can be expanded as follows:

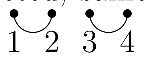
$$K_r(X_1 \ldots X_{i_1}, \ldots, X_{i_{r-1}+1} \ldots X_n) = \sum_{\substack{\pi \in \mathcal{P}(n) \\ \pi \vee \rho = \hat{1}_n}} K_\pi(X_1, \ldots, X_n). \tag{11}$$

This Theorem 3.2 may look somehow complicated so we will discuss example on it.

Lets have cumulant $K_2(X_1X_2, X_3X_4)$.

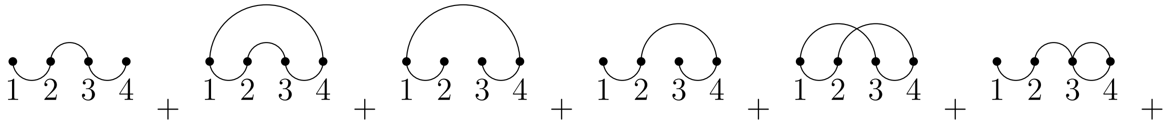
The theorem above can help us to make this hard to compute cumulant more understandable due to the previous definition. According to theorem, we can rewrite this cumulant as :

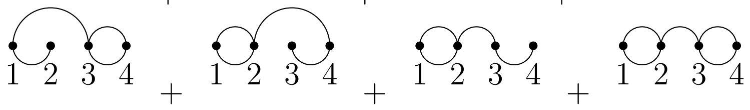
$$\sum_{\substack{\pi \in \mathcal{P}(4) \\ \pi \vee \cup\cup = \hat{1}_4}} K_\pi(X_1, X_2, X_3, X_4)$$

To see how the operator under the sum is working let us imagine that at the beginning random variables X_1 and X_2 are connected, same occurs fore variables X_3, X_4. Graphical representation of this looks like . Do not confuse it with K_2 cumulants, for now they are just connection between certain variables. Now to use theorem, one has to find all combinations of paths that make these two 'islands' connect. The simplest one would be connection between variable X_2 and X_3. Now this new connections will be cumulants. If we connect above mentioned two variables we will have $K_2(X_2, X_3)K_1(X_1)K_1(X_4)$ and this will be the first part of our sum. Let's try to find it all.

Note 3.3. On drawings bellow we always mark the start connection between cumulants (the lines upside-down) so only new upper lines are new made cumulants. If there is no upper connection from the certain variable, it means that is in K_1.

$$K_2(X_1X_2, X_3X_4) = \sum_{\substack{\pi \in \mathcal{P}(4) \\ \pi \vee \cup\cup = \hat{1}_4}} K_\pi(X_1, X_2, X_3, X_4) =$$





in terms of cumulants mean:

$$\begin{aligned}
&K_1(X_1)K_1(X_4)K_2(X_2,X_3) + K_2(X_1,X_4)K_2(X_2,X_3) + K_2(X_1,X_4)K_1(X_2)K_1(X_3) \\
&+ K_2(X_2,X_4)K_1(X_1)K_1(X_3) + K_2(X_1,X_3)K_2(X_2,X_4) + K_3(X_2,X_3,X_4)K_1(X_1) \\
&+ K_3(X_1,X_3,X_4)K_1(X_2) + K_3(X_1,X_2,X_4)K_1(X_3) + K_3(X_1,X_2,X_3)K_1(X4) \\
&+ K_4(X_1,X_2,X_3,X_4)
\end{aligned} \tag{12}$$

3.3 Additional properties

Theorem 3.4. *Random variables $X_1,\dots,X_n$ are independent if and only if, for every $n \geq 1$ and every non-constant choice of $Y_i \in \{X_1,\dots,X_n\}$, where $i \in \{1,\dots,k\}$ (for some positive integer $k \geq 2$) we get $K_k(Y_1,\dots,Y_k) = 0$.*

This is one of the most important property of the cumulants and reason why they are in use by so many in different mathematical solutions.

Cumulants of some important random distributions are listed as follows:

- The Gaussian distribution $N(\mu,\sigma)$ possesses the simplest list of cumulants: $K_1 = \mu$, $K_2 = \sigma$ and $K_n = 0$ for $n \geq 3$

Proof. For the normal distribution $\mathcal{N}(\mu,\sigma^2)$ moment generating function(MGF) is defined as $M_x(t) = e^{\mu t + \frac{t^2\sigma^2}{2}}$. According to earlier definition we can obtain cumulant generating function by logarithm of MGF. This procedure will give us $K_X(t) = \mu t + \frac{t^2\sigma^2}{2}$. Then we can easily compute subsequent values of cumulants.

$$K_1 = \frac{d}{dt}(\mu t + \frac{t^2\sigma^2}{2})\Big|_{t=0} = \mu$$

$$K_2 = \frac{d^2}{dt^2}(\mu t + \frac{t^2\sigma^2}{2})\Big|_{t=0} = \sigma^2$$

$$K_3 = \frac{d^3}{dt^3}(\mu t + \frac{t^2\sigma^2}{2})\Big|_{t=0} = 0$$

There is no need to calculate further cumulants because next derivatives of our cumulant generating function will be always equal to 0 □

- for the Poisson distribution with mean λ we have $K_n = \lambda$

Proof. We will use the same approach as above. Firstly we compute cumulant generating function for Poisson distribution. $log(M_X(t)) = log(e^{\lambda(e^t-1)}) = \lambda(e^t - 1)$.

$$K_1 = \lambda \frac{d}{dt}(e^t - 1)\Big|_{t=0} = \lambda$$

$$K_2 = \lambda \frac{d^2}{dt^2}(e^t - 1)\Big|_{t=0} = \lambda$$

All derivatives will give as exactly the same and in the end we will always obtain λ as a result. □

- for the χ^2 distribution with $n-1$ degrees of freedom we have $K_r = 2^{r-1}(r-1)!(n-1)$

Proof. Now we will calculate the formula for $n-th$ cumulant for χ^2 distribution. Let's have X as a random variable with a $\chi^2_{(1)}$ distribution.Then:

$$K_{\chi^2_{(1)}} = logM_X(t) = -\frac{1}{2}log(1-2t)$$

For the following equation we will use Taylor series:

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

We assume that $f(x)$ is infinitely differentiable, then $f^{(n)}(x_0)$ is n-th derivative of function f over point x_0. In our calculation x_0 will be 0 so actually we are using Maclurin series. To get the unique formula for cumulants from χ^2 with 1 degree of freedom we will focus on part $log(1-2t)$ and try to find unique equation.

$$\begin{aligned}
f(t_0) &= log(1-2t_0) \\
f'(t_0) &= \frac{-2}{1-2t_0} \\
f''(t_0) &= \frac{-4}{(1-2t_0)^2} \\
f'''(t_0) &= \frac{-16}{(1-2t_0)^3} \\
f^{(4)}(t_0) &= \frac{-96}{(1-2t_0)^4}
\end{aligned} \tag{13}$$

We can rewrite equation using Taylor's formula and setting $t_0 = 0$:

$$\begin{aligned} logM_{\chi^2}(t) &= -\frac{1}{2}log(1-2t) = -\frac{1}{2}[-2t - \frac{4t^2}{2} - \frac{16t^3}{3!} - \frac{96t^4}{4!} - \dots] = \\ &= \frac{1}{2}[2t + \frac{(2t)^2}{2} + \frac{(2t)^3}{3} + \frac{(2t)^4}{4} + \dots] \end{aligned} \tag{14}$$

Now to get the r-th cumulant we need to differentiate r-times above equation and set $t = 0$. One will see that simplified form of this can be rewrite as $2^{r-1}(r-1)!$ so we get unique formula for each cumulants from χ^2_1:

$$K_r(Y^2) = 2^{r-1}(r-1)!$$

When we want to calculate any n-th cumulant for $X \sim \chi^2_{n-1}$, instead of $-\frac{1}{2}log(1-2t)$ we will write $-\frac{(n-1)}{2}log(1-2t)$ and following all steps above get at the end:

$$2^{r-1}(r-1)!(n-1)$$

□

These examples shows that by using cumulants we can look at the random variables in a bit different way. To make it even more clear we will describe additional properties for all kind of variables that we will use for further proofs.

- (Additivity) Let $X_1, \dots, X_m$ be any independent random variables.
  Then, $K_r(X_1 + \dots + X_m) = K_r(X_1) + \dots + K_r(X_m)$, $r \geq 1$.
- (Translation Invariance) For any constant c, $K_1(X + c) = c + K_1(X)$ and
  $K_r(X + c) = K_r(X)$, $n \geq 2$.
- If X is a variable from symmetric distribution, then $K_{2r+1}(X) = 0$ when $r > 0$

4 Ruben's proof for two variables

In this section we will focus on Ruben's proof for two independent variables. In proof we assume that $\sigma^2 = 1$. We do not assume any symmetry or any other additional properties of the variables, than existence of all moments. Reasoning in the proof is somehow easier than in original paper of Ruben but bring us the same result. This proposition will lead us through the basic cumulants properties and help in further analyse of $\chi^2 -$ conjecture

Theorem 4.1. *If for Q_n, $n = 2$,random variables X_1 and X_2, are independent and identically distributed. Then if $Q_2 \sim \chi^2_{(1)}$, X_1 and X_2 come from normal distribution*

Proof. We want to remind the definition of Q_n:

$$Q_n = \sum_{i=1}^{n}(X_i - \overline{X})^2$$

Then equation needed in this section comes as follow:

$$\begin{aligned}
Q_2 &= (X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 = \left(X_1 - \frac{X_1+X_2}{2}\right)^2 + \left(X_2 - \frac{X_1+X_2}{2}\right)^2 = \\
&= X_1^2 - 2X_1\frac{(X_1+X_2)}{2} + \left(\frac{X_1+X_2}{2}\right)^2 + X_2^2 - 2X_2\frac{(X_1+X_2)}{2} + \left(\frac{X_1+X_2}{2}\right)^2 = \\
&= X_1^2 - X_1^2 - X_1X_2 + \left(\frac{X_1+X_2}{2}\right)^2 + X_2^2 - X_2^2 - X_1X_2 + \left(\frac{X_1+X_2}{2}\right)^2 = \\
&= -2X_1X_2 + 2\frac{(X_1^2 + 2X_1X_2 + X_2^2)}{4} = -2X_1X_2 + \frac{X_1^2}{2} + X_1X_2 + \frac{X_2^2}{2} = \\
&= \frac{X_1^2}{2} - X_1X_2 + \frac{X_2^2}{2} = \frac{X_1^2 - 2X_1X_2 + X_2^2}{2} = \frac{(X_1-X_2)^2}{2}
\end{aligned}$$

This according to the subject of our work lead us to assumption that

$$\frac{(X_1-X_2)^2}{2} \sim \chi^2_{(1)}$$

We define new random variable Y.

$$Y \sim \frac{X_1 - X_2}{\sqrt{2}} \text{ then } Y^2 \sim \chi^2_{(1)}$$

Now we can easily notice that variable Y has symmetric distribution and all odd cumulants of Y variable are equal to 0. Additionally using property that cumulants with mixed random variables are 0, we can simplify as follow:

$$K_{2r+1}(X_1 - X_2, X_1 - X_2, \ldots, X_1 - X_2) = K_{2r+1}(X_1) - K_{2r+1}(X_2) = 0$$

Now let's have a look at some additional properties of Y variable.

$$\mathbb{E}(Y^2) = K_1(Y^2) = 1$$

On the other hand we know that each odd cumulant of Y variable is equal to 0

$$E(YY) = \underbrace{K_1(Y)K_1(Y)}_{0} + K_2(Y) = K_2(Y) = 1$$

So far we know that

- $Y^2 \sim \chi^2_{(1)}$
- Y has symmetric distribution
- $K_2(Y) = 1$

Now we want to show how behave other even cumulants of Q_n.

$$K_2(Y^2) = Var(Y^2) = 2$$

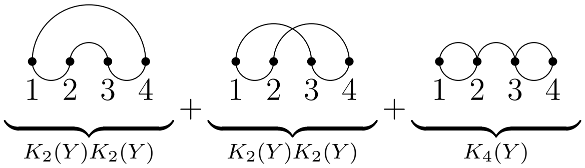
$$K_2(Y^2) = K_2(YY, YY)$$

This as previously explained can be described as a problem, which require to find all combination of connections that :

$$\pi \vee \cup\cup = \hat{1}_4$$

But this time wee need to keep in mind that all odd cumulants are 0

$$K_2(Y^2) =$$



$$\rightarrow 1 + 1 + K_4(Y) = 2$$

$$K_4(Y) = 0$$

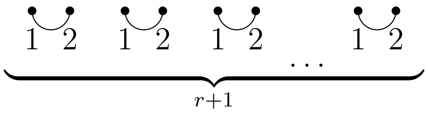
So far we know that every odd cumulant and 4-th cumulant of variable Y is zero. Below we prove that every even cumulant greater than 2 is also equal to zero.

$$K_4(Y) = 0$$

Assumption: $\forall_{r>1} \quad K_{2r}(Y) = 0$
Induction: $K_{(2r+1)}(Y) \overset{?}{=} 0$

$$2^r r! = K_{r+1}(Y^2) = K_{r+1}(\underbrace{YY, \ldots, YY}_{r+1}) = K_{2r+2}(Y) + C_2$$

Where C_2 is defined as all possible connections $\pi \vee \cup \cup \cdots \cup = \hat{1}_{2r+2}$, without usage of full partition. The additional limit that we have is a fact that only double connections are allowed because by assumption and previous fact, any other different ones are equal to zero. The main problem now is how to calculate number of such connections.
We start from this:



Now from the first edge of a first part we can choose $\binom{r}{1}$ other parts where we want to go and multiply it by 2, because when we choose part we want to go we also have to pick edge we want to be connected. Let's say we choose to join peak 1 from part 1 with peak 2 from part 3 now lets take peak that have left from part 3 and join it with another part, different than part 1. We have $\binom{r-1}{1}$ parts to choose and again 2 different edges. Following this instruction and each time starting new connection from the part that we just landed and connecting it to part that still is outside of the overall connection will use number of possibilities equal to :

$$\underbrace{\binom{r}{1}2}_{\text{1-st part}}\underbrace{\binom{r-1}{1}2}_{\text{2-nd part}}\dots\underbrace{\binom{r-(r+1)}{1}2}_{\text{penultimate part}} = 2^r r!$$

We conclude two things

$$C_2 = 2^r r!$$

$$K_{2r+2}(Y) = 0$$

All even cumulants except of K_2 are equal to zero.
By having this information about the cumulants (only K_2 different than zero), one can see that they behave exactly the same way as cumulants of normal distribution i.e

$$\frac{X_1 - X_2}{\sqrt{2}} \sim \mathcal{N}(0,1)$$

Theorem 4.2. *[3] (Levy Cramer) If the sum of two independent variables X_1 and X_2 is normal that each of them is normally distributed*

By expanded theorem:

$$X_1 \sim \mathcal{N}(\mu,1) \quad \text{and} \quad X_2 \sim \mathcal{N}(\mu,1)$$

□

5 Symmetric variables

As tittle say in this proof we will focus on variables that by assumption are symmetric, all their moment's exists and $\sigma^2 = 1$. We are going to show that with these assumptions and information that Q_n has $\chi^2_{(n-1)}$ distribution, we are able unambiguously state that random variables from Q_n have normal distribution.
In this part we will use definition of sample variance S_n^2 as in Definition 2.1.

Theorem 5.1. *If Q_n is made of symmetric i.i.d random variables X_i ,where $i \in \{1,\dots,n\}$ and Q_n has $\chi^2_{(n-1)}$ distribution , then X_i comes from normal distribution.*

Proof.

$$Q_n = \underbrace{\left(1 - \frac{1}{n}\right)}_{a} \underbrace{\sum_{i=1}^{n} X_i^2}_{A} \underbrace{- \sum_{i,j=1,\ i \neq j}^{n} X_i X_j}_{B} = aA + B$$

$$Q_n \sim \chi^2_{(n-1)}$$

$$K_r(Q_n) = 2^{r-1}(r-1)!(n-1)$$

We will also use fact that all odd cumulants for symmetric distributions and cumulants consist of two different independent variable from the same distribution are equal to zero.

We start with writing off the equation for the r-th cumulant:

$$K_r(aA + B) = \sum_{j=0}^{r} \binom{r}{j} K_r(\underbrace{aA \dots aA}_{r-j}, \underbrace{B \dots B}_{j})$$

Then it is good to see that such cumulants behave the same way as equation $(A + B)^r$ with some additional properties of cumulants.To show that we will use Newton binomial formula. To see how it works let's see an example for $K_3(aA + B)$.

$$\begin{aligned} K_3(aA + B) = \binom{3}{0} K_3(aA) + \binom{3}{1} K_3(aA, aA, B) + \binom{3}{2} K_3(aA, B, B) + \binom{3}{3} K_3(B) = \\ = a^3 K_3(A) + 3a^2 K_3(A, A, B) + 3a K_3(A, B, B) + K_3(B) \end{aligned} \tag{15}$$

This will help us in further part of the proof

Now lets perform $K_r(Q_n)$ as follow:

$$K_r(Q_n) = \underbrace{\sum_{\substack{\pi \in \mathcal{P}(2r) \\ \pi = F}} K_\pi(X_i)}_{\text{I}} + \underbrace{\sum_{\substack{\pi \in \mathcal{P}(2r) \\ \pi \vee \cup \dots \cup = F \\ \pi = \text{pairs}}} K_\pi(X_i)}_{II} + \underbrace{\sum_{\substack{\pi \in \mathcal{P}(2r) \\ \pi \vee \cup \dots \cup = F \\ \pi > \text{pairs} \\ \pi < F}} K_\pi(X_i)}_{III}$$

Where F means $\hat{1}_{2r}$ (full partition).

The first part of our sum are all K_π for which exists full partitions. Such cumulants can be made only by usage of A part in K_r because having even one B in cumulant and connecting all variables into one partition will give us two i.i.d variables in same cumulant thus make it zero. Having only A in cumulant is also not solving our problem. A is a sum so it has many i.i.d variables in it, knowing how such sum is behaving according to Newton formula, we can take only cumulants consist of the same variable in each place. There will be of course other shorter cumulants, which are not equal to zero, but for now we are looking for full partitions.

$$K_r(aA) = a^r K_r\Big(\sum_{i=1}^{n} X_i^2, \dots, \sum_{i=1}^{n} X_i^2\Big) = a^r K_r\Big(X_1^2{+}X_2^2{+}\cdots{+}X_n^2, \dots, X_1^2{+}X_2^2{+}\cdots{+}X_n^2\Big)$$

$$\overset{\text{independence}}{=} a^r \sum_{i=1}^{n} K_r(X_i^2) = a^r n K_r(X_1^2) = \underbrace{a^r n K_{2r}(X_1)}_{\pi=F} + a^r n \sum_{\substack{\pi\in\mathcal{P}(2r)\\ \pi<F}} K_\pi(X_i)$$

After the part, where $\pi = F$ there is the sum that will partly goes to sum II and III in definition of $K_r(Q_n)$

Now lets consider the second part of $K_r(Q_n)$ sum. To get to the proper value of this part for the sake of simplicity, we will introduce new variable Y_i an by use of features regarding cumulants, that were proved in previous subsection, get the proper value of the second part of the sum.

$$\begin{aligned} & Y_i \sim \mathcal{N}(0,1), \qquad K_2(Y_i) = 1, \qquad \forall_{i\neq 2} K_n(Y_i) = 0, \qquad Q_n \sim \chi^2, \\ & K_r(Q_n(Y_i)) = (r-1)!2^{r-1}(n-1) \\ & K_r(Q_n(Y_i)) = \underbrace{\sum_{\substack{\pi\in\mathcal{P}(2r)\\ \pi\vee\cup\cdots\cup=F}}}_{(1)} K_\pi(Y_i) = \underbrace{\sum_{\substack{\pi\in\mathcal{P}(2r)\\ \pi\vee\cup\cdots\cup=F\\ \pi=\text{pairs}}}}_{(2)} K_\pi(Y_i) = (r-1)!2^{r-1}(n-1) \end{aligned} \tag{16}$$

Part (1) is obvious because after usage of additivity we will get many cumulants K_r with different products of cumulants e.g $K_r(Y_1^2, Y_5Y_{19}, \dots, Y_4^2, \dots, Y_{77}^2)$. According to Theorem 3.2 this can be rewrite as the sum (1). In sum (2) we are taking under account that we can only have double connection because of Y_i properties as variable from normal distribution and allow only pairs as a connection. This bring us to exactly the same object as second sum in $K_r(Q_n)$. We are not able to count how many such connections is possible because we need to manage mixed cumulants what is extremely hard when their number is increasing. We know by the definition the value of r-th cumulant when, variables come from χ^2 distribution. Knowing this, without complicated calculations we know the exact value of this sum.

Before we will calculate the proper value of the most complicate sum where $\pi <$ F and $\pi >$ pairs, it will be easier to see what information we have from previous sums and compare it to the value of the r-th cumulant of χ^2 distribution.

$$K_r(Q_n) = (r-1)!2^{r-1}(n-1) = \Big(1-\frac{1}{n}\Big)^r nK_{2r}(X_1)+(r-1)!2^{r-1}(n-1)+\sum_{\substack{\pi\in\mathcal{P}(2r)\\ \pi\vee\sqcup\cdots\sqcup=F\\ \pi>\text{pairs}\\ \pi<F}} K_\pi(X_i)$$

.
By definition of $K_r(Q_n)$ value we can simplify:

$$\Big(1-\frac{1}{n}\Big)^r nK_{2r}(X_1)+\sum_{\substack{\pi\in\mathcal{P}(2r)\\ \pi\vee\sqcup\cdots\sqcup=F\\ \pi>\text{pairs}\\ \pi<F}} K_\pi(X_i) = 0$$

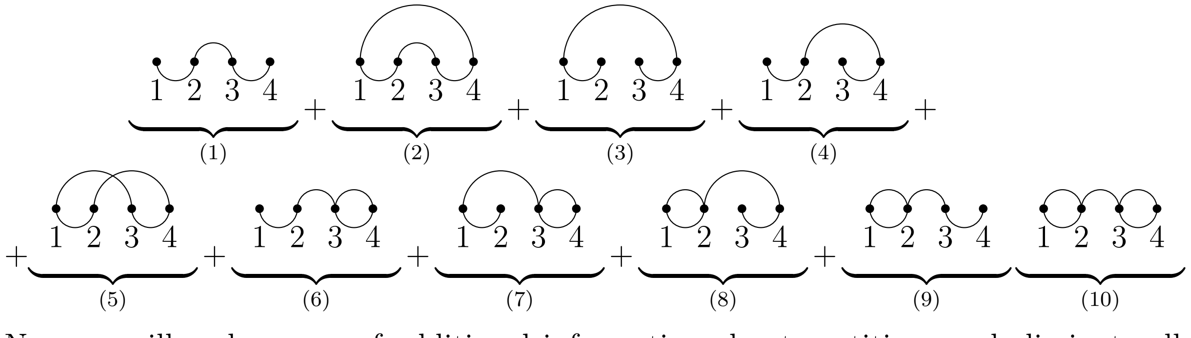
Odd cumulants are equal to zero, thus we care only about even ones:

For r = 2:

$$\Big(1-\frac{1}{n}\Big)^2 nK_4(X_1)+\sum_{\substack{\pi\in\mathcal{P}(4)\\ \pi\vee\sqcup\sqcup=F\\ \pi>\text{pairs}\\ \pi<F}} K_\pi(X_i) = 0$$

We have situation where we have two pairs of cumulants and we should connect then without usage of π as a full partition and pairs. All possible connections without taking under account any additional requirements about the length of partition state as follow:

$$\sum_{\substack{\pi\in\mathcal{P}(2r)\\ \pi\vee\sqcup\sqcup=F}} K_\pi(X_i) \rightarrow$$



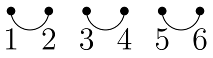
Now we will make usage of additional information about partitions and eliminate all parts where π has pairs and $\pi =$ F.
Cumulants number 1,2,3,4,5,7,8 have connection consist of pairs thus are not allowed

in this sum. Cumulants number 6,9 have odd π and cumulant number 10 is a full one so not allowed due to the sum connection definition. We have no type of cumulants left so the sum is equal to zero. The main conclusion that we want to find out is the fact that, because the sum is zero than $K_4(X_1)$ has to be zero.

For r = 3:

$$\Big(1-\frac{1}{n}\Big)^3 nK_6(X_1)+\sum_{\substack{\pi\in\mathcal{P}(6)\\ \pi\vee\sqcup\sqcup\sqcup=F\\ \pi>\text{pairs}\\ \pi<F}} K_\pi(X_i)=0$$

Now let's look at the situation and analyse possible solutions. We have to connect three pairs without using odd, full partitions and double connections.



Without further long writing one can easily noticed that this is impossible to connect these three pair without the usage of odd full and K_4 cumulant. Again sum and $K_6(X)$ are both equal to zero.

By induction we want to show that above mentioned reasoning is easy to follow for any r.

$$K_6(X_i)=0$$

Assumption: $\forall_{r>1} \quad K_{2r}(X_i)=0$

Induction: $K_{2(r+1)}(X_i)\overset{?}{=}0$

$$\Big(1-\frac{1}{n}\Big)^{2r+2} nK_{2r+2}(X_1)+\sum_{\substack{\pi\in\mathcal{P}(2r+2)\\ \pi\vee\sqcup\cdots\sqcup=F\\ \pi>\text{pairs}\\ \pi<F}} K_\pi(X_i)=0$$

In sum we again have to consider quiet similar situation to these mentioned above. We have to consider :

$$\underbrace{\sqcup\sqcup\cdots\sqcup}_{r+1}\vee\pi=\hat{1}_{2r+2}$$

Again we can't use full partition odd connections and any even connections smaller than $2r+2$. It comes out that we don't have any options thus such connection does not exists. This mean than $K_{2r+2}=0$

We showed that

$$\sum_{\substack{\pi\in\mathcal{P}(2r)\\ \pi\vee\sqcup\cdots\sqcup=F\\ \pi>\text{pairs}\\ \pi<F}} K_\pi(X_i)=0 \quad \text{and} \quad K_{2r+2}(X_i)=0$$

□

By this ending conclusion one will see that cumulants of X_i that comes from Q_n have exactly the same properties as the ones from normal distribution (only $K_2 \neq 0$) and thus $X_i \sim \mathcal{N}(\mu, 1)$

6 ID Variables

In this part we will show **new proof** for ID variables based on cumulants.

Theorem 6.1. *If $Q_n \sim \chi^2_{(n-1)}$, X_i where $i \in \{1, \dots, n\}$ are infinitely divisible variables with the same expected value, then X_i have a normal distribution.*

In terms of cumulants we can say that random variable X_i is infinitely divisible if and only if

$$K_{n+2}(X_i) = \int_{\mathbb{R}} x^n d\rho_i(x)$$

for positive finite measure $d\rho_i$ on $\mathbb{R}$. Our main focus will be to show that measure $\rho_i = \delta_0$. It means that $X_i \sim \mathcal{N}(\mu, 1)$ and to show that that this is true, it is sufficient to calculate that $\int_{\mathbb{R}} x^2 d\rho_i(x) = K_4(X_i) = 0$.

Proof. Under assumption that $\mathbb{E}(Y_1) = \mathbb{E}(Y_2) = \mathbb{E}(Y_3) = \mathbb{E}(Y_4) = 0$ then :

$$K_2(Y_1Y_2, Y_3Y_4) = K_2(Y_1, Y_4)K_2(Y_2, Y_3) + K_4(Y_1, Y_2, Y_3, Y_4)$$

By assumption $K_2(X_1) = 1$, this fact will be used on computation on Q_n. Firstly we will compute value of the formula $K_2(X_i - \overline{X}, X_j - \overline{X})$ for cases when $i = j$ or $i \neq j$.

$i = j$

$$K_2(X_1 - \overline{X}, X_1 - \overline{X}) = K_2\Big(X_1 - \frac{X_1}{n}, X_1 - \frac{X_1}{n}\Big) + (n-1)K_2\Big(\frac{-X_2}{n}, \frac{-X_2}{n}\Big) =$$

$$= \Big(1 - \frac{1}{n}\Big)^2 K_2(X_1) + (n-1)\frac{1}{n^2}K_2(X_2) = \Big(1 - \frac{1}{n}\Big)^2 + \frac{n-1}{n^2} = \frac{(n-1)^2 + n - 1}{n^2} = \frac{n-1}{n}$$

$i \neq j$

$$K_2(X_1 - \overline{X}, X_2 - \overline{X}) = K_2\Big(X_1 - \frac{X_1}{n}, -\frac{X_1}{n}\Big) + K_2\Big(-\frac{X_2}{n}, X_2 - \frac{X_2}{n}\Big) + (n-2)K_2(-\frac{X_3}{n}) =$$

$$= (1 - \frac{1}{n})(-\frac{1}{n})K_2(X_1) - \frac{1}{n}(1 - \frac{1}{n})K_2(X_2) + (n-1)\frac{1}{n^2}K_3(X_3) = -2(\frac{1}{n} - \frac{1}{n^2}) + (\frac{1}{n} - \frac{2}{n^2}) =$$

$$= -\frac{1}{n} \tag{17}$$

By using similar methods lets consider behave of a Q_n sum made only of K_4 cumulants. Firstly:

$$\sum_{i=1}^{n} K_4(X_i-\overline{X}) = \sum_{i=1}^{n}\Big(\Big(1-\frac{1}{n}\Big)^4 K_4(X_i) + \sum_{l\neq i}\Big(-\frac{1}{n}\Big)^4 K_4(X_l)\Big) =$$

$$= \Big(\Big(1-\frac{1}{n}\Big)^4 + \frac{n-1}{n^4}\Big)\sum_{i=1}^{n} K_4(X_i)$$

Then situation for $i \neq j$:

$$\sum_{\substack{i,j=1\\ i\neq j}}^{n} K_4(X_i-\overline{X}, X_i-\overline{X}, X_j-\overline{X}, X_j-\overline{X}) =$$

$$\sum_{\substack{i,j=1\\ i\neq j}}^{n}\Big(\Big(1-\frac{1}{n}\Big)^2\Big(-\frac{1}{n}\Big)^2(K_4(X_i)K_4(X_j)) + \sum_{l\neq i,j}^{n}\Big(-\frac{1}{n}\Big)^4 K_4(X_l)\Big) =$$

$$\Big(2(n-1)\Big(1-\frac{1}{n}\Big)^2\Big(\frac{1}{n}\Big)^2 + \frac{(n-1)(n-2)}{n^4}\Big)\sum_{i=1}^{n} K_4(X_i) = \frac{(n-1)^2}{n^2}\sum_{i=1}^{n} K_4(X_i)$$

This simplification will help us in calculating the following sum :

$$\begin{aligned}
&2(n-1) = K_2(Q_n, Q_n) = \\
&\sum_{i,j=1}^{n} K_2((X_i-\overline{X})^2, (X_j-\overline{X})^2) = 2\sum_{i,j=1} K_2(X_i-\overline{X}, X_j-\overline{X})K_2(X_j-\overline{X}, X_i-\overline{X}) + \\
&+\sum_{i,j=1} K_4(X_i-\overline{X}, X_i-\overline{X}, X_j-\overline{X}, X_j-\overline{X}) = 2\Big(\sum_{i=1}[K_2((X_i-\overline{X}, (X_i-\overline{X})]^2 + \\
&+\sum_{i=1,j=1,i\neq j}^{n}[K_2((X_i-\overline{X}, (X_j-\overline{X})]^2\Big) + \sum_{i=1}^{n} K_4(X_i-\overline{X}) + \\
&+\sum_{i=1,j=1,i\neq j} K_4(X_i-\overline{X}, X_i-\overline{X}, X_j-\overline{X}, X_j-\overline{X}) = \\
&= 2\Big(\frac{(n-1)^2}{n} + \frac{n-1}{n}\Big) + \frac{(n-1)^2}{n^2}\sum_{i=1} K_4(X_i) = 2(n-1) + \frac{(n-1)^2}{n^2}\sum_{i=1} K_4(X_i) \\
&\qquad \frac{(n-1)^2}{n^2}\sum_{i=1} K_4(X_i) = 0 \Leftrightarrow K_4(X_i)
\end{aligned} \tag{18}$$

Now it's easy to see that $\sum_{i=1}^{n} K_4(X_i) = \sum_{i=1}^{n}\int_{\mathbb{R}} x^2 d\rho_i(x) = 0$ and this mean that $\rho_i(x) = \delta_0(x)$. Again we obtain cumulants of variables that behave exactly the same way as ones from normal distribution and thus each X_i from Q_n has $\mathcal{N}(\mu, 1)$ distribution. □

References

[1] Lennart Bondesson, *The sample variance, properly normalized, is χ^2-distributed for the normal law only*, Sankhyā Ser. A **39** (1977), no. 3, 303–304.

[2] George E. P. Box and George C. Tiao, *Bayesian inference in statistical analysis*, Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., 1973, Addison-Wesley Series in Behavioral Science: Quantitative Methods. MR 0418321

[3] Harald Cramér, *Über eine Eigenschaft der normalen Verteilungsfunktion*, Math. Z. **41** (1936), no. 1, 405–414. MR 1545629

[4] Wiktor Ejsmont and Franz Lehner, *Sample variance in free probability*, J. Funct. Anal. **273** (2017), no. 7, 2488–2520. MR 3677831

[5] Nina N. Golikova and Victor M. Kruglov, *A characterisation of the Gaussian distribution through the sample variance*, Sankhya A **77** (2015), no. 2, 330–336.

[6] G. S. James, *On moments and cumulants of systems of statistics*, Sankhyā **20** (1958), 1–30.

[7] A. M. Kagan, Yu. V. Linnik, and C. Radhakrishna Rao, *Characterization problems in mathematical statistics*, John Wiley & Sons, New York-London-Sydney, 1973, Translated from the Russian by B. Ramachandran, Wiley Series in Probability and Mathematical Statistics.

[8] V. P. Leonov and A. N. Shiryaev, *On a method of calculation of semi-invariants*, Theor. Prob. Appl. **4** (1959), 319–328.

[9] Alexandru Nica and Roland Speicher, *Lectures on the combinatorics of free probability*, London Mathematical Society Lecture Note Series, vol. 335, Cambridge University Press, Cambridge, 2006. MR 2266879

[10] Giovanni Peccati and Murad S. Taqqu, *Wiener chaos: moments, cumulants and diagrams*, Bocconi & Springer Series, vol. 1, Springer, Milan; Bocconi University Press, Milan, 2011, A survey with computer implementation, Supplementary material available online. MR 2791919

[11] Harold Ruben, *A new characterization of the normal distribution through the sample variance*, Sankhyā Ser. A **36** (1974), no. 4, 379–388.

[12] ______, *A further characterization of normality through the sample variance*, Sankhyā Ser. A **37** (1975), no. 1, 72–81.