χ^2 Distributation

Basic Information

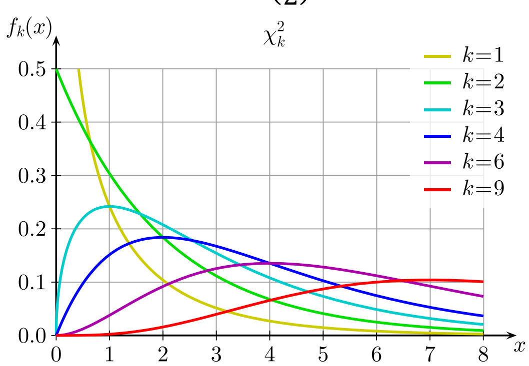
χ^2distributation has only one parameter: degree of freedom $k \in N_*$, written as $\chi^2(k)$ or χ_k^2.
It is very important in the field of statistics and probability, and used extensively for hypothesis testing and/or constructing confidence intervals. Unlike other well-known distributions such as normal distribution or exponential distributions, chi-squared distribution is not often applied in direct sampling; however, it is widely used in hypothesis testing and at times construction of t or F distributions.

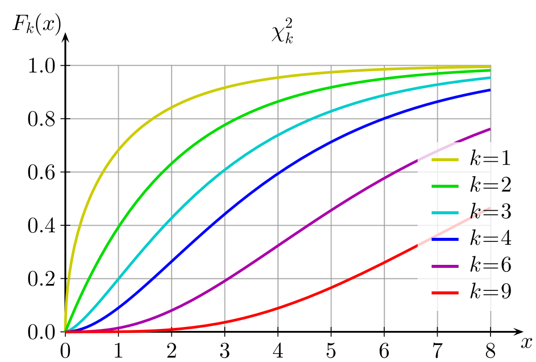
Definition:
$\chi^2(k)$ is defined as the square sum of n stand normal distributation:
If $Z_1, Z_2, \cdots\cdots, Z_k$ are independent, standard normal random variables, then the sum of their squares

$$Q \equiv \sum_{i=0}^{k} (Z_k)^2 \sim \chi^2(k)$$

PDF: $f_k(x) = \left. \dfrac{x^{\frac{k}{2}-1} \cdot e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \cdot \Gamma\left(\frac{k}{2}\right)} \right| x \geq 0$

CDF: $F_n(x) = \displaystyle\int_0^x \frac{x^{\frac{\tau}{2}-1} \cdot e^{-\frac{\tau}{2}}}{2^{\frac{\tau}{2}} \cdot \Gamma\left(\frac{k}{2}\right)} d\tau = \left. \frac{\gamma\left(\frac{k}{2}, \frac{x}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \right| x \geq 0$





Mean: k

Varience: $2k$

Median: $\approx \ k\left(1 - \frac{2}{9k}\right)^3$

Support: $x \in (0, \infty)\ if$ k $= 1; x \in [1, \infty)\ if\ k > 1.$

Entropy: $\frac{k}{2} + \ln\left[2\Gamma\left(\frac{k}{2}\right)\right] + \left(1 - \frac{k}{2}\right)\psi\left[\frac{k}{2}\right]$

Skewness: $\sqrt{8/k}$

Ex. kurtosis: $\frac{12}{k}$

Relation to other distributions

Gamma Distribution: if $X \sim \chi^2(\nu)$ and $c > 0$, then $cX \sim \Gamma(k = \nu/2, \theta = 2c)$

Exponential distribution: if $X \sim \chi^2(2)$, then $X \sim \mathrm{Exp}(1/2)$

Erlang distribution: if $X \sim \chi^2(2k)$, then $X \sim \mathrm{Erlang}(k, 1/2)$

Rayleigh distribution: if $X \sim Rayleigh(1)$, then $X^2 \sim \chi^2(2)$

Maxwell distribution: if $X \sim Maxwell(1)$ then $X^2 \sim \chi^2(3)$

Beta distribution: $if\ X \sim \chi^2(\nu_1)\ and\ Y \sim \chi^2(\nu_2)$ are independent, then $\frac{X}{X+Y} \sim \text{Beta}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)$

Uniform distribution: $if X \sim \text{U}(0,1)$ then $-2\log(X) \sim \chi^2(2)$

Laplace distribution: $if\ X_i \sim \text{Laplace}(\mu, \beta)$ then $\sum_{i=1}^{n} \frac{2|X_i - \mu|}{\beta} \sim \chi^2(2n)$

Application

χ^2distributation is frequently used in testing goodness of fit (Pearson's chi-squared test), as it is reasonable to assume the errors are in independent normal distributions.

In Pearson's chi-squared test, we have

1. A set of data $S \equiv \{\{x_n, \overrightarrow{y_n}\}\}\ has\ a + k$ entries, for which we are trying to fit into $\overrightarrow{x_n}\ based\ on\ \overrightarrow{y_n}$.
2. A function $\vec{f}$ calculated based on S with a degrees of freedom / parameters.

and assume the magnitude of error for x_n, $\epsilon_n \sim N\left(f(\overrightarrow{y_n}), \sqrt{f(\overrightarrow{y_n})}\right)^{*1}$,thus $\frac{f(\overrightarrow{y_n}) - x_n}{\sqrt{f(\overrightarrow{y_n})}} \sim N(0,1)$.

choose h_0: f fits the data set well; h_a: f does not fits the data set well.

Here, define statistic

$$W = \sum_{n=0}^{k+a} \left(\frac{f(\overrightarrow{y_n}) - x_n}{\sqrt{f(\overrightarrow{y_n})}}\right)^2 = \sum_{n=0}^{k} \frac{(f(\overrightarrow{y_n}) - x_n)^2}{f(\overrightarrow{y_n})} \sim \chi^2(k)$$

And it is expected to be in χ^2distributation of degree of freedom k^{*2}, and clearly higher W implies less possible our function fits the data well. Then we can choose some α as significance and perform p-test using this W.

*1- Actually, $\epsilon_n' \sim N\left(f(\overrightarrow{y_n}), h \cdot \sqrt{f(\overrightarrow{y_n})}\right) | h \in \mathrm{R}^+$ is also sufficient, for W' corresponds to ϵ_n' is linear to W.

*2-Just as a intuition, the reason for degree of freedom to be $k\ but\ not\ (k+a) x_n\ is\ that\ f(\overrightarrow{y_n})$ takes a variables dependent on S thus losing a degrees of freedom.