MCS-511
Department of Computer Science
University of Karachi.

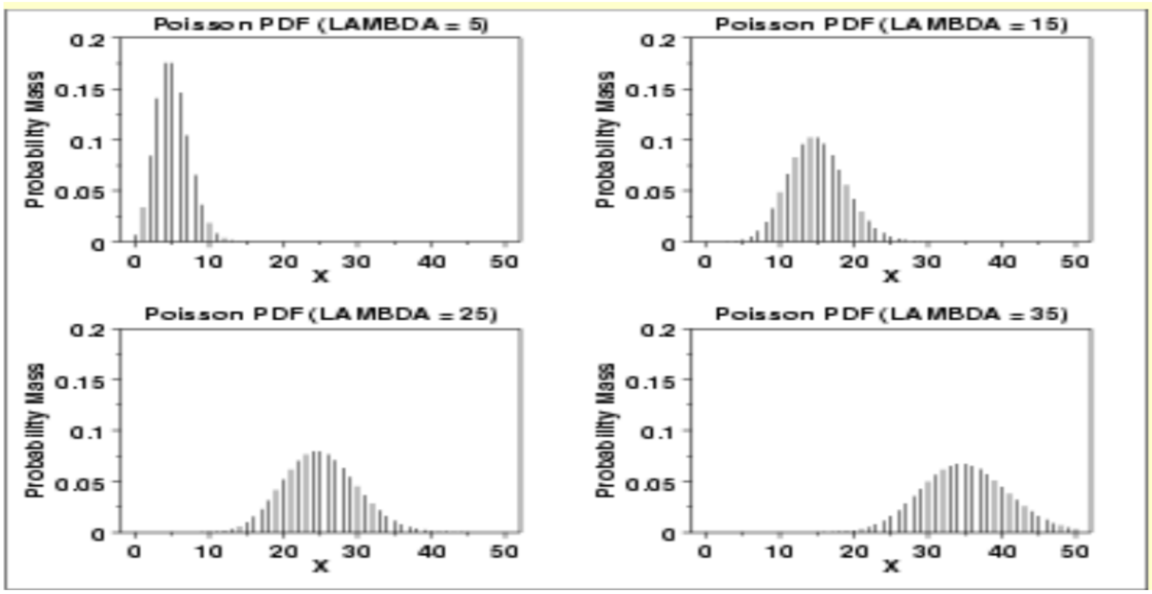
The Poisson distribution is used to model the number of events occurring within a given time interval.

The formula for the Poisson probability mass function is

$$p(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \text{ for } x = 0, 1, 2, \cdots$$

λ is the shape parameter which indicates the average number of events in the given time interval.

The following is the plot of the Poisson probability density function for four values of λ.



The formula for the Poisson cumulative probability function is

$$F(x;\lambda) = \sum_{i=0}^{x} \frac{e^{-\lambda}\lambda^i}{i!}$$

ML for Poisson

Suppose that $X = (X_1, X_2, \ldots, X_n)$ are iid observations from a Poisson distribution with unknown parameter λ. The likelihood function is:

$$\begin{aligned} L(\lambda; x) &= \prod_{i=1}^{n} f(x_i; \lambda) \\ &= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \\ &= \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{x_1! x_2! \cdots x_n!} \end{aligned}$$

By differentiating the log of this function with respect to λ, that is by differentiating the *Poisson loglikelihood function*

$$l(\lambda; x) = \sum_{i=1}^{n} x_i \log \lambda - n\lambda$$

ignoring the constant terms that do not depend on λ, one can show that the maximum is achieved at $\hat{\lambda} = \sum_{i=1}^{n} x_i / n$. Thus, for a Poisson sample, the MLE for λ is just the sample mean.

Chi-square statistic

$$\boxed{\chi^2 = \sum_{k=1}^{N} \frac{\{Obs(k) - Exp(k)\}^2}{Exp(k)}.} \qquad (10.1)$$

Here the sum goes over N categories or groups of data defined depending on our testing problem; $Obs(k)$ is the actually observed number of sampling units in category k, and $Exp(k) = \mathbf{E}\{Obs(k) \mid H_0\}$ is the expected number of sampling units in category k if the null hypothesis H_0 is true.

This is always a *one-sided, right-tail* test. That is because only the low values of χ^2 show that the observed counts are close to what we expect them to be under the null hypotheses, and therefore, the data support H_0. On the contrary, large χ^2 occurs when Obs are far from Exp, which shows inconsistency of the data and the null hypothesis and does not support H_0.

Therefore, a level α rejection region for this chi-square test is

$$R = [\chi^2_\alpha, +\infty),$$

and the P-value is always calculated as

$$P = \boldsymbol{P}\left\{\chi^2 \geq \chi^2_{\text{obs}}\right\}.$$

Pearson showed that the null distribution of χ^2 converges to the Chi-square distribution with $(N-1)$ degrees of freedom, as the sample size increases to infinity. This follows from a suitable version of the Central Limit Theorem. To apply it, we need to make sure the sample size is large enough. The rule of thumb requires an *expected count of at least 5 in each category*,

$$Exp(k) \geq 5 \quad \text{for all } k = 1, \ldots, N.$$

If that is the case, then we can use the Chi-square distribution to construct rejection regions and compute P-values. If a count in some category is less than 5, then we should *merge* this category with another one, recalculate the χ^2 statistic, and then use the Chi-square distribution.

Here are several main applications of the chi-square test.

MCS-511
Department of Computer Science
University of Karachi.

10.1.1 Testing a distribution

The first type of applications focuses on testing whether the data belong to a particular distribution. For example, we may want to test whether a sample comes from the Normal distribution, whether interarrival times are Exponential and counts are Poisson, whether a random number generator returns high quality Standard Uniform values, or whether a die is unbiased.

In general, we observe a sample $(X_1, \ldots, X_n)$ of size n from distribution F and test

$$H_0 : F = F_0 \qquad \text{vs} \qquad H_A : F \neq F_0 \tag{10.2}$$

for some given distribution F_0.

To conduct the test, we take all possible values of X under F_0, the *support* of F_0, and split them into N bins $B_1, \ldots, B_N$. A rule of thumb requires anywhere from 5 to 8 bins, which is quite enough to identify the distribution F_0 and at the same time have sufficiently high expected count in each bin, as it is required by the chi-square test ($Exp \geq 5$).

The observed count for the k-th bin is the number of X_i that fall into B_k,

$$Obs(k) = \# \{i = 1, \ldots, n \; : \; X_i \in B_k\} .$$

If H_0 is true and all X_i have the distribution F_0, then $Obs(k)$, the number of "successes" in n trials, has Binomial distribution with parameters n and $p_k = F_0(B_k) = \boldsymbol{P}\{X_i \in B_k \mid H_0\}$. Then, the corresponding expected count is the expected value of this Binomial distribution,

$$Exp(k) = np_k = nF_0(B_k).$$

After checking that all $Exp(k) \geq 5$, we compute the χ^2 statistic (10.1) and conduct the test.

Chi-square test is basically used for two purposes :

1. to test the distribution of the data /variable

2. it is used for test of independence