Chi-squared Goodness of Fit

Prasangsha Ganguly

April 2020

Problem Statement:

Say, you are given a sample data of one variable from an unknown population. You want to identify the probability distribution of the population. If you can identify the underlying parametric distribution from where your data comes, any further analysis may become easier.

For example, say you are given 50 values of the waiting times in a restaurant over a month. The owner of the restaurant claims that the distribution of the waiting time follow exponential distribution with an unknown rate λ (or mean $1/\lambda$). You have to verify that and find the unknown λ. Note, that we can find the non-parametric distribution by finding the $ECDF$ of the data and then differentiate it to get the PDF. Also, there are several other methods like Kolmogorov-Smirnov test for this problem.

Solution Method:

In this solution, the Method of Moments have been used to estimate the parameters. The method of moments equates each of the r theoretical moments with the corresponding r^{th} sample moment to obtain a system of r equations in r unknowns:

$$\int_{-\infty}^{\infty} x^i f(x;\theta_1,\theta_2,...\theta_r) = \frac{1}{n}\sum_{j=1}^{n}(x_j)^i \quad i \in \{1,2...r\}$$

The solutions of this system of equations are the estimators of $\theta_1,\theta_2,...\theta_r$ The sample mean ($\bar{x}$) is an estimate of the population mean. Hence, we calculate the sample mean and use that as an estimator of the population mean. For some distributions, if we estimate the parameter according to the mean then the variance of the distribution become implicit. For example, if $X \sim Poisson(\lambda)$ then, $E[X] = \lambda, Var(X) = \lambda$. Hence if the sample mean is used as an estimate of the population mean (λ), the variance become fixed. For other cases such as normal distribution, $X \sim \mathcal{N}(\mu,\sigma^2)$, the sample variance is used as an estimate of the population variance. The corrected sample variance $s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$ is an unbiased estimator of the population variance.

Say, we are given some values, first we sort the data. For example say we have, $1.86, 2.68, ...22.82, 23.85$. Now, say our null hypothesis H_0 is "The data comes from a exponential distribution with $\lambda = 0.1$". Now we create $0.2, 0.4, 0.6$ and 0.8 quantiles of the exponential distribution with $\lambda = 0.1$. These are the values that mark the probability of each area $= 0.2$. For example, if we have $X \sim Exp(0.1)$ and 0.2^{th} quantile is say x, then $P[X \leq x] = 0.2$.

So, we have obtained the $0.2, 0.4, 0.6$ and 0.8 quantiles of the hypothesized distribution. Say the values are $x_{0.2}, x_{0.4}, x_{0.6}, x_{0.8}$. Consider that we are given n data points. So, expect $n * 20/100$ or $n/5$ points to lie in each of the intervals $(-\infty, x_{0.2}]$, $[x_{0.2}, x_{0.4}]$, $[x_{0.4}, x_{0.6}]$, $[x_{0.6}, x_{0.8}]$ and $[x_{0.8}, \infty)$.

However, in the given data sets we calculate the number of observations fall in each of the ranges $(-\infty, x_{0.2}]$, $[x_{0.2}, x_{0.4}]$, $[x_{0.4}, x_{0.6}]$, $[x_{0.6}, x_{0.8}]$ and $[x_{0.8}, \infty)$, and they may not be equal to $n/5$. Say the observed counts are $O_{0.2}, O_{0.4}, O_{0.6}, O_{0.8}, O_{1.0}$ We now create the contingency table as follows:

Table 1: The Contingency Table

Intervals:	$(-\infty, x_{0.2}]$	$[x_{0.2}, x_{0.4}]$	$[x_{0.4}, x_{0.6}]$	$[x_{0.6}, x_{0.8}]$	$[x_{0.8}, \infty)$
Expected counts:	$n/5$	$n/5$	$n/5$	$n/5$	$n/5$
Observed counts:	$O_{0.2}$	$O_{0.4}$	$O_{0.6}$	$O_{0.8}$	$O_{1.0}$

We create the test statistic as $C = \sum_{intervals} \frac{[Expected\_counts - Observed\_counts]^2}{Expected\_counts}$.

For our case, the test statistic is, $c_{obs} = \frac{(n/5-O_{0.2})^2}{n/5} + \frac{(n/5-O_{0.4})^2}{n/5} + \frac{(n/5-O_{0.6})^2}{n/5} + \frac{(n/5-O_{0.8})^2}{n/5} + \frac{(n/5-O_{1.0})^2}{n/5}$

Under null hypothesis, this statistic C follows Chi-square distribution with degree of freedom $= d - l - 1$, where d is the number of intervals and l is the number of parameters estimated.

We calculate the $P - value$ as $P[C \geq c_{obs}]$. If the $P - value$ is more than our significance level α, then we accept our null hypothesis, that is, the data comes from the hypothesized distribution. Otherwise, we conclude that the data doesn't come from the hypothesized distribution.

The Distributions Used:

Here we considered the following distributions. However, the method can be extended for other distributions.

- **Uniform**: $X \sim Unif[a, b]$,

$$f_X(x) = \begin{cases} 0 & x < a \\ \frac{1}{b-a} & a \leq x \leq b \\ 0 & b < x \end{cases}$$

  $E[X] = \frac{b+a}{2}, Var(X) = \frac{(b-a)^2}{12}$

- **Poisson**: $X \sim Poisson(\lambda)$ $f_X(k) = \frac{\lambda^k e^{-\lambda}}{k!}$ $k \in \mathbb{N}_0$ $E[X] = \lambda, Var(X) = \lambda$
- **Exponential**: $X \sim Exponential(\lambda)$ $f_X(x) = 1 - e^{-\lambda x}$ $x \in \mathbb{R}_0^+$ $E[X] = \frac{1}{\lambda}, Var(X) = \frac{1}{\lambda^2}$
- **Normal**: $X \sim \mathcal{N}(\mu, \sigma^2)$ $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ $x \in \mathbb{R}$ $E[X] = \mu, Var(X) = \sigma^2$
- **Gamma**: $X \sim \Gamma(k, \theta)$ k is shape and θ is scale $f_X(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$ $x \in \mathbb{R}_0^+$ $E[X] = k\theta, Var(X) = k\theta^2$
- **Chi-squared**: $X \sim \chi^2(k)$ k is degrees of freedom $f_X(x) = \frac{1}{\Gamma(k/2)2^{k/2}} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$ $x \in \mathbb{R}_0^+$ $E[X] = k, Var(X) = 2k$
- **Beta**: $X \sim Beta(\alpha, \beta)$ α and β are the shape $f_X(x) = \frac{x^{(\alpha-1)}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ $x \in [0, 1]$ $E[X] = \frac{\alpha}{\alpha+\beta}, Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- **Erlang**: $X \sim Erlang(k, \lambda)$ $f_X(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!}$ $x \in R_0^+$ $E[X] = \frac{k}{\lambda}, Var(X) = \frac{k}{\lambda^2}$
- **Weibull**: $X \sim Weibull(k, \lambda)$ $f_X(x) = \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1} e^{(-x/\lambda)^k}$ $x \in R_0^+$ $E[X] = \lambda\Gamma(1 + \frac{1}{k}), Var(X) = \lambda^2[\Gamma(1 + \frac{2}{k}) - (\Gamma(1 + \frac{1}{k}))^2]$

Here, we considered only these distributions as these distributions cover a lot of the parametric distributions in use. Several other distributions can be created from the above basic distributions. Like, $F - Distribution$ is a ratio of two scaled Chi-squared random variables.

Installation:

$install.packages('devtools')$
$devtools :: install\_github("PrasangshaG/FindDistribution")$
$library('FindDistribution')$
$distributionPG(data\_vector)$