Identifying and fitting distributions

Sunday, 09 October 2022 22:49

• What's the approach we take for identifying and fitting distributions?	Task: Identifying Distributions and fitting distributions Approach: 1. Look at visualizations 2. Use descriptive statistics to further identify distributions 3. Go for tests to confirm judgement
	<u>Detailed Approach</u>
What are the steps in the analysis?	1. Plot Histogram • Whenever you want to find a distribution, the first thing you do is plot Histogram. • When you plot a histogram, you can look at it and tell what family of distributions a dataset may belong to. 2. Do some descriptive Analysis 3. Plot Probability Plots to check goodness of fit i. P-P Plot ii. Q-Q Plot 4. Apply statistical tests to check goodness of fit i. Chi square test ○ For this test, we usually need two kind of frequencies: • Observed frequency: sample frequency • Expected frequency: population frequency, the model distribution where we assume the data is coming from. ○ Output: the chi-square statistic and the $p-value$ ○ Compare the $p-value$ and α (significance level) and, conclude about the null hypothesis. ii. Calculated and tabular chi-square statistic comparison ○ Input: confidence level, degrees of freedom (df) ○ Output: Tabular chi-square statistic ○ Compare tabular and calculated chi-square values and, conclude about the null hypothesis.
Defining Hypotheses	Null Hypothesis: Sample distribution follows model distribution. Alternate Hypothesis: Sample distribution does not follow model distribution.
Chi Square Test • How to calculate expected frequency? `stats.chisquare(obs_freq, expec_freq)` • What does this command return? • Chi-square formula? • What is p-value? • What is α? • Comparing both these values, when do we accept the null hypothesis?	`stats.chisquare(obs_freq, expec_freq)` This command returns two values: 1. Chi-Square Statistic 2. P-value • P-value is the probability of observing this particular sample when the null hypothesis is assumed to be true. • α :: Significance level, that can take a value of 0.05, or 0.10, or 0.15. • If P-value > α :: We accept the null hypothesis.
• Tabulated Chi-Square Statistic `scipy.stats.chi2.ppf(0.95, df=9)` • What arguments does this command take? • How to calculate df? • How do we calculate df in a contingency table?	`scipy.stats.chi2.ppf(0.95, df=9)` Above command gives us the tabulated chi-square statistic, that takes two arguments: 1. First argument is confidence level = 95% we have taken 2. Second argument is degrees of freedom, $df = k - p - 1$ where, a. k = number of classes/intervals/buckets b. p = number of parameters we estimate from the sample
• Comparing calculated and tabulated chi-square statistic, when do we accept the null hypothesis?	If Tabulated value ≥ Calculated value » We accept the null hypothesis
Some other observations:	• In Python, whenever we run a code for any of the plots (Q-Q or P-P), by default, it's always comparing it with the **standard normal distribution.** • It converts the data that is fed to it to a standard normal, basically by doing $\frac{x - \mu}{\sigma}$. • It will scale it and then it will compare it with the standard normal(default), and see how the quantiles (or distributions) are fitting. • So, it normalizes the data, and then compare it with the standard normal.
	• Poisson distribution has mean = variance (= λ)
	• Generally for normal distributions the Kurtosis would be around 3.
	• Degree of freedom = k - p - 1 • p = 0 for uniform distribution as there is no parameter. ☐ • This point is doubtful. Needs verification
	☐ Question: What range decides if the skewness is very small or very large to consider?
