Modeling Risk and Realities: Week 3

Senthil Veeraraghavan
Operations, Information and Decisions Department

Week 3: Choosing Distributions that fit your Data

- Data and visualization: Graphical representation

- Choosing among the family of distributions: Discrete and continuous distributions.

- How good does a certain distribution fit? Hypothesis testing and goodness of fit.

Session 3

Fitting distributions to data

- We made a case that it is important to fit a “right” distribution by visualizing the data.
- We generated histograms for our two datasets.
  - Dataset1_histogram.xlsx
  - Dataset2_histogram.xlsx
- Now, we can use those files to test goodness of fit.
- Before we do that, let us understand the concept behind testing the goodness of fit of a distribution.

Goodness of Fit tests

- After evaluating the histograms and summary statistics (mean, standard deviation, etc), we can explore distributions that can provide a good fit.
- Goodness of fit tests provide statistical evidence to test hypotheses about the nature of distribution that can fit the data.
- Two popular statistical goodness-of-fit tests are
  - Chi-Square test (χ^2 test)
  - Kolmogorov-Smirnov test.
- Anderson-Darling test is another test that is used less frequently.
- We will focus on the Chi-Square test.

Chi-Square test

- The Chi-Square tests the following null hypothesis against the alternate hypothesis.
  - Null Hypothesis: the studied data comes from a random variable following a specified distribution (e.g. uniform or normal).
  - Alternate Hypothesis: The sample data does not come from the specified distribution.
- Note: this is a one-sided test.
- In other words, you can disprove that data came from a specific distribution, but you cannot prove it came from that distribution.

Running a Chi-Square Test on Your Data

- We will run the Chi-Square tests on our datasets. However, we will first look at some thumb rules to run the test.
- Ideally, you should have at least 50 data points.
- Divide your data into n "buckets" with at least 5 observations in each bucket.
- Every Chi-Square test has "degrees of freedom" = number of buckets – parameters of specified distribution – 1.
- For example, if you have n=10 buckets and try to fit a normal distribution with 2 parameters (mean and standard deviation),
  - Degrees of freedom = 10 – 2 – 1 = 7.

Chi-Square Test

- For each Chi-Square test with some degree of freedom, you can reject the null hypothesis with some confidence.
  - This could be set at 99%, 95%, etc.

- Chi-Square confidence tables are available at lots of sources.
  - For example, see the table at the following online link.

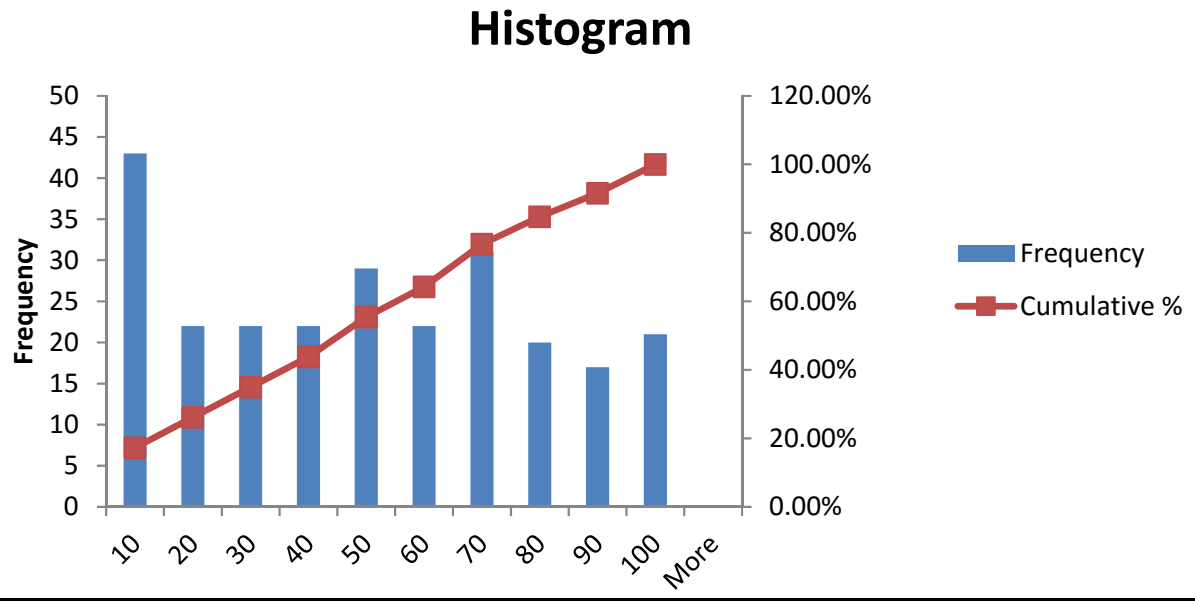
  https://www.medcalc.org/manual/chi-square-table.php

- We will explore Chi-Square test on our two-data sets
  - Dataset1_histogram.xlsx
  - Dataset2_histogram.xlsx

Data Set 1

- The figure below gives the histogram
  - pdf in blue bars
  - CDF in red curve.

- Given the visualization of the pdf suggests a uniform distribution
  - We run a chi-square test for uniform distribution based on calculated min and max values from the data



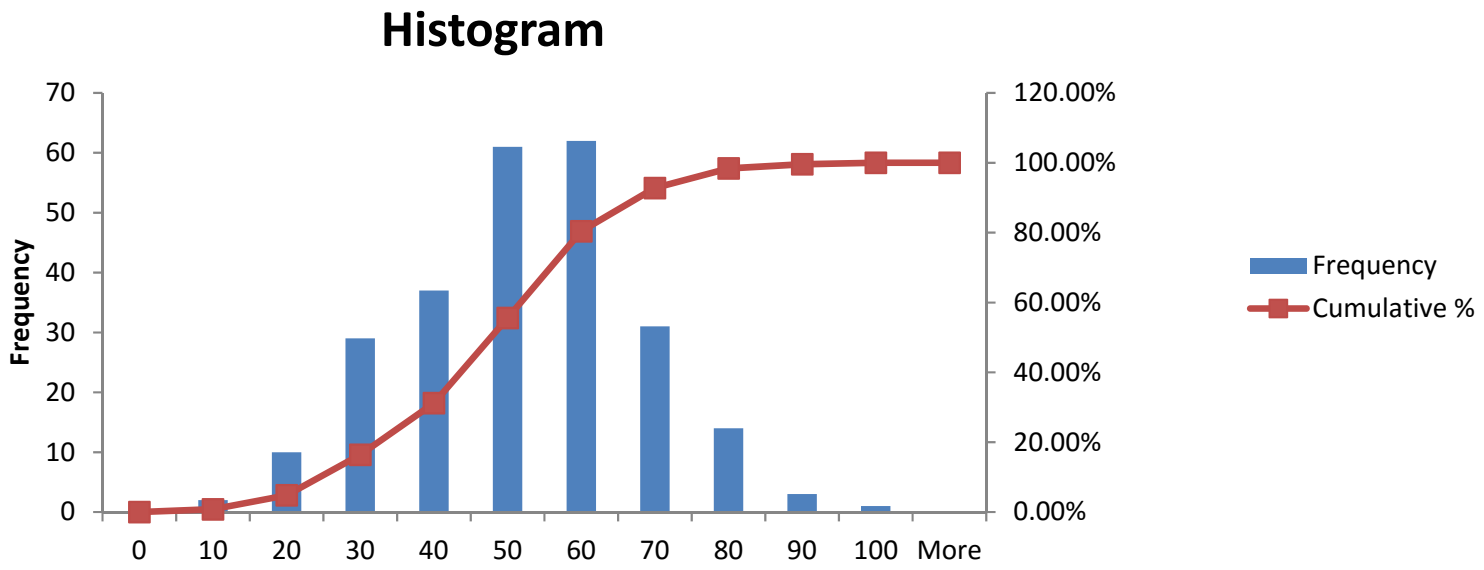
Data Set 1

- Descriptors for Uniform distribution (2 parameters).
  - MIN value = 0.09
  - MAX value = 99.87
- Recall that our Null Hypothesis is that data comes from a uniform distribution.
- Degrees of Freedom = Number of bins – Number of parameters -1
  = 10 – 2 -1 =7.
- Chi-squared test gives a value of 0.013
- Looking at the tables (e.g. link), for degree of freedom 7.
  - We fail to reject the null hypothesis that data came from the uniform distribution (with confidence of 99.5%).

Data Set 2

- The figure below gives the histogram
  - pdf in blue bars
  - CDF in red curve.

- Given the visualization of the pdf suggests a normal distribution
  - We run a chi-square test for normal distribution based on calculated average and standard deviation from the data



Data Set 2

- Descriptors for Normal distribution (2 parameters).
  - Sample average (sample mean) = 47.20
  - Standard deviation = 15.78
- Recall that our Null Hypothesis is that data comes from a normal distribution.
- Degrees of Freedom = Number of bins – Number of parameters -1
  = 10 – 2 – 1 = 7.
- Chi-squared test gives a value of 0.8851
- Looking at the tables (e.g. link), for degree of freedom 7.
  - We fail to reject the null hypothesis that data came from normal distribution (with confidence of 99.5%).

Goodness of Fit Files

- The tabulated excel files are now reported in
  - Dataset1_FIT.xlsx
  - Dataset2_FIT.xslx

Kolmogorov-Smirnov test (K-S test)

- For small samples, K-S test is more suitable.

- Basic Idea of K-S test:
  - Arrange the data values in ascending order
  - Arrange theoretical values similarly (from cumulative distribution function).
  - Find the maximal difference between the data value and its corresponding theoretical value.
  - If this maximal difference value is low, the fit is good.

- Typically, a value of 0.03-0.04 or lower is considered good.

Modeling Using Continuous Distributions

- Depending on size and nature of data, modeling reality using continuous distributions and choosing the correct distribution can be a challenging task.
- It is mathematically elegant to use a continuous distribution, but the approach creates complexities.
- Hence, often simulation is used.
- This will be our focus in Week 4. Congrats on ending Week 3.
- Best wishes for Week 4!

Conclusion

Senthil Veeraraghavan
Operations, Information and Decisions Department
@senthil_veer