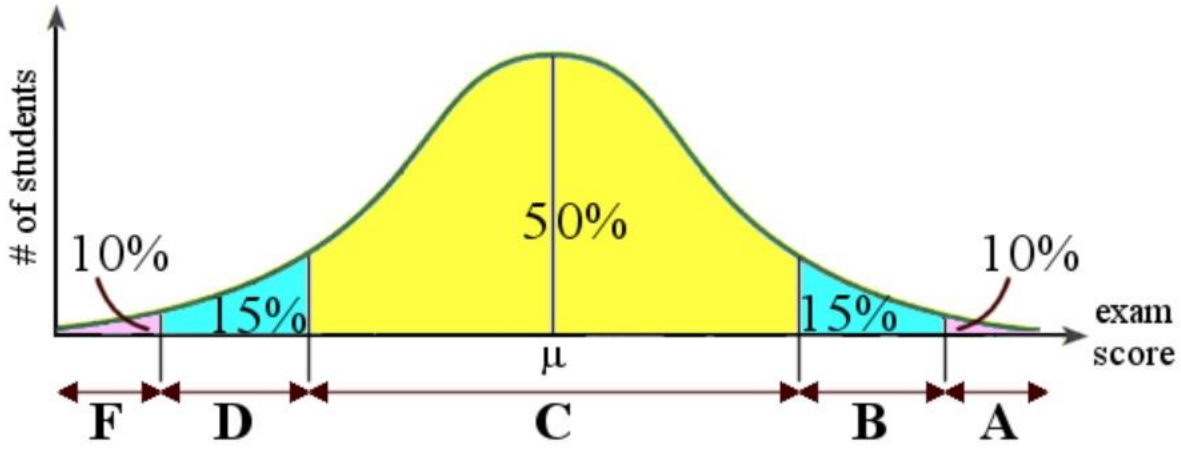
Probability distributions

What are distributions?

How to identify correct distribution of the data?

Probability distributions

- A statistical model that shows possible outcomes of a particular event or course of action as well as the statistical likelihood of each event.
- What do we mean by “Grades in a course follows a normal distribution”?
- What do we mean by “Sales for the next month may be uniformly distributed”?



How to go about this?

How do we use the collected business data (sales volume, loan defaulters, Salary hikes in an organization, etc.)?

1. The data values themselves are used directly in the simulation. This is called trace-driven simulation.
2. "Fit" a theoretical distribution to the data (and check whether that "fit" is good!).
3. The data values could be used to define an empirical distribution function in some way.

What are these empirical distributions?

- Using the data, we build our own distributions.
- How does one build a distribution?

- Essential building blocks:

Define the density/distribution functions.

Estimate the parameters (mean, standard deviation, etc.)

Empirical distributions

For ungrouped data:

Let $X_{(i)}$ denote the *i*th smallest of the X_j's so that: $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$.

$$F(x) = \begin{cases} 0 & if \quad x < X_{(1)} \\ \dfrac{i-1}{n-1} + \dfrac{x - X_{(i)}}{(n-1)\left(X_{(i+1)} - X_{(i)}\right)} & if \quad X_{(i)} \le x < X_{(i+1)} \quad for \quad i = 1,2,\cdots,n-1 \\ 1 & if \quad X_{(n)} \le x \end{cases}$$

Empirical distributions

For grouped data:

- Suppose that n X_j's are grouped in k adjacent intervals $[a_0,a_1), [a_1,a_2),\ldots[a_{k-1},a_k)$ so that jth interval contains n_j observations. $n_1+ n_2+\ldots n_k = n$.
- Let a piecewise linear function G be such that $G(a_0) = 0$, $G(a_j) = (n_1+ n_2+\ldots n_j)/n$, then:

$$G(x)=\begin{cases}0 & if \quad x<a_0\\ G\left(a_{j-1}\right)+\dfrac{x-a_{j-1}}{a_j-a_{j-1}}\left[G\left(a_j\right)-G\left(a_{j-1}\right)\right] & if \quad a_{j-1}\le x<a_j,\ j=1,2,\cdots k\\ 1 & if \quad a_k\le x.\end{cases}$$

The three approaches…

- Approach 1 is used to validate simulation model when comparing model output for an existing system with the corresponding output for the system itself.
- Two drawbacks of approach 1: simulation can only reproduce only what happened historically; and there is seldom enough data to make all simulation runs.
- Approaches 2 and 3 avoid these shortcomings so that any value between minimum and maximum can be generated. So approaches 2 and 3 are preferred over approach 1.
- If theoretical distributions can be found that fits the observed data (approach 2), then it is preferred over approach 3.

Approach 3 v/s Approach 2

- Empirical distribution may have some irregularities if small number of data points are available. Approach 2 smoothens out the data and may provide information on the overall underlying distribution.
- In approach 3, it is usually not possible to generate values outside the range of observed data in the simulation.
- If one wants to test the performance of the simulated system under extreme conditions, that can not be done using approach 3.
- There may be compelling (physical) reasons in some situations for using a particular theoretical distribution. In that case too, it is better to get empirical support for that distribution from the observed data.

Business example

- Data points: 217.
- For these data points, we need to fit a probability distribution.

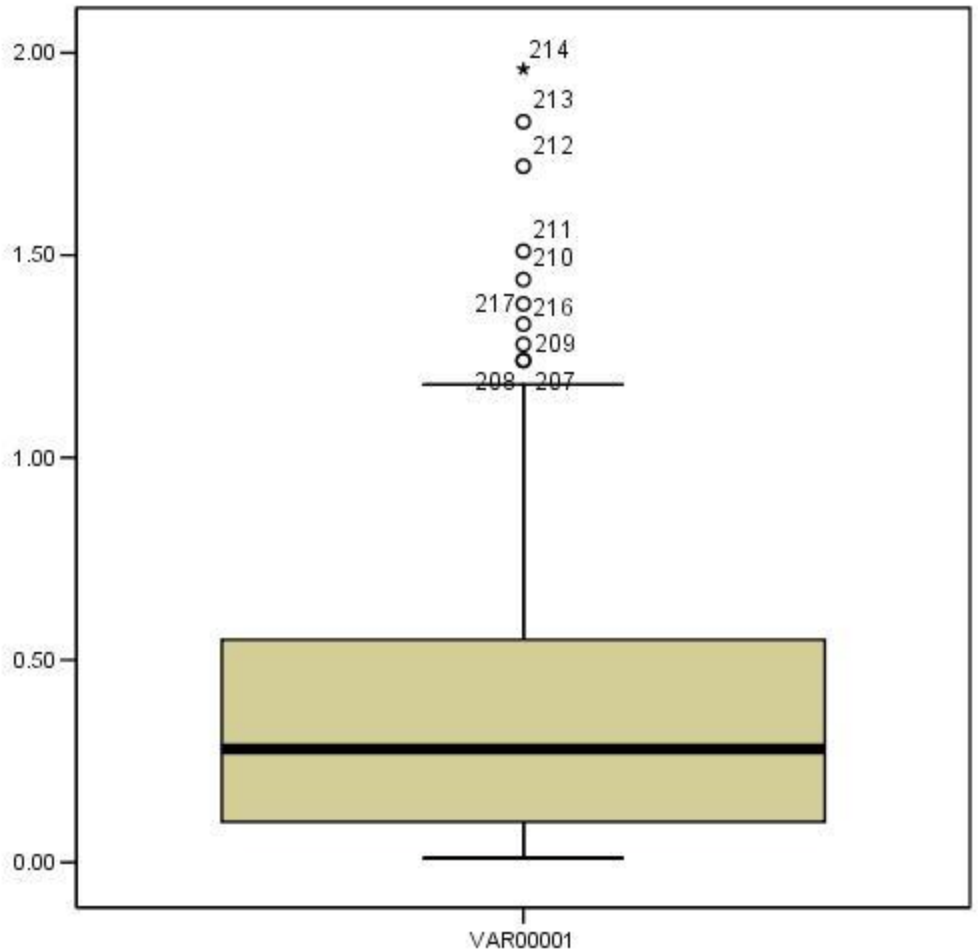
Summary statistics

VAR00001

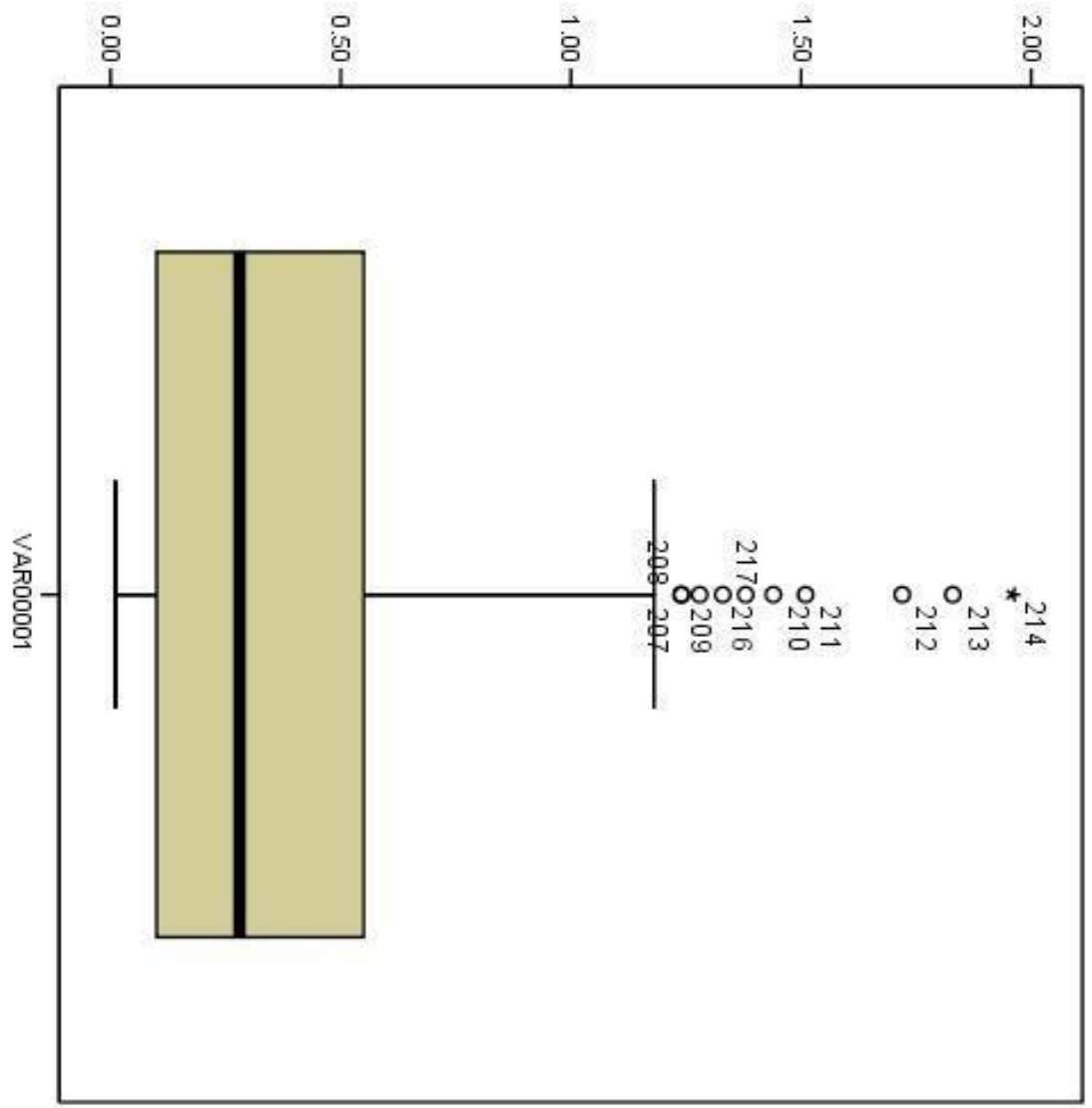
N	Valid	217
	Missing	1
Mean		.4012
Median		.2800
Mode		.05[a]
Std. Deviation		.38093
Variance		.145
Skewness		1.466
Std. Error of Skewness		.165
Range		1.95
Minimum		.01
Maximum		1.96
Percentiles	25	.1000
	50	.2800
	75	.5500

a. Multiple modes exist. The smallest value is shown

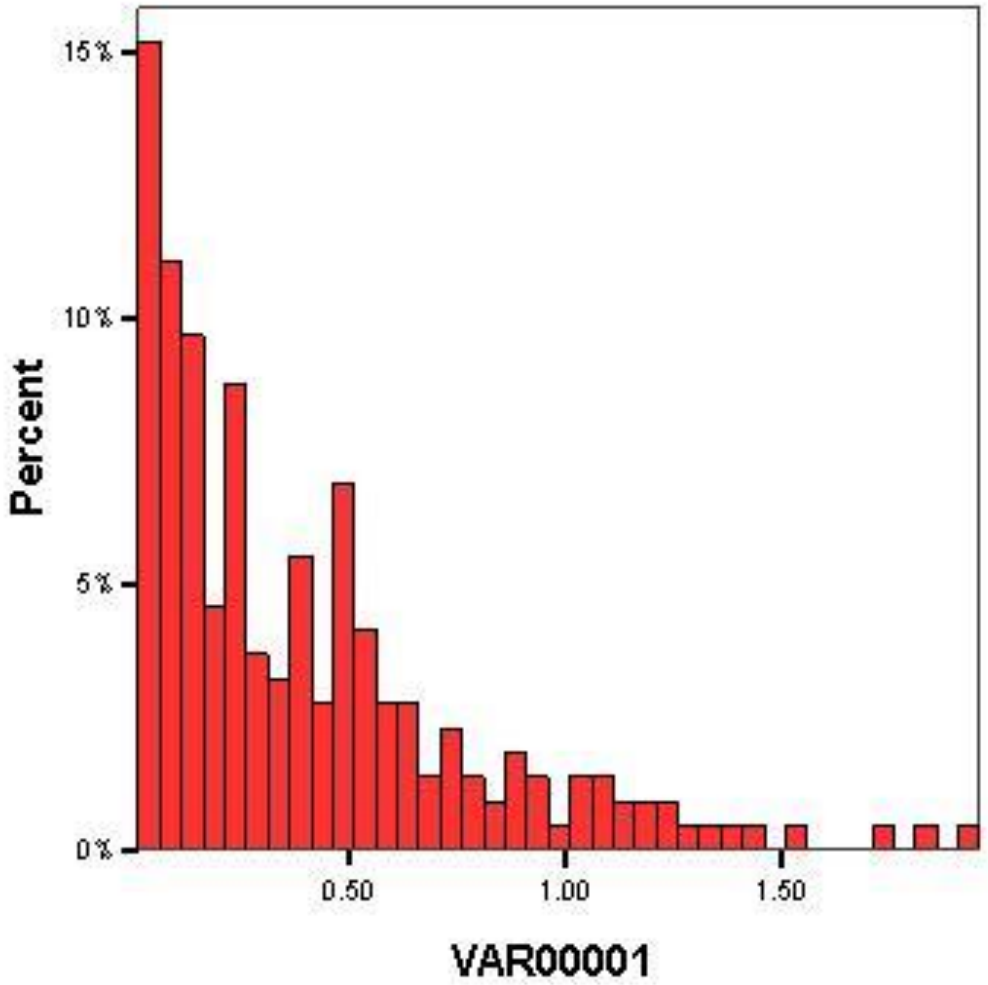
Box plot


2.00
1.50
1.00
0.50
0.00
214
213
212
211
210
217
216
209
208
207
VAR00001

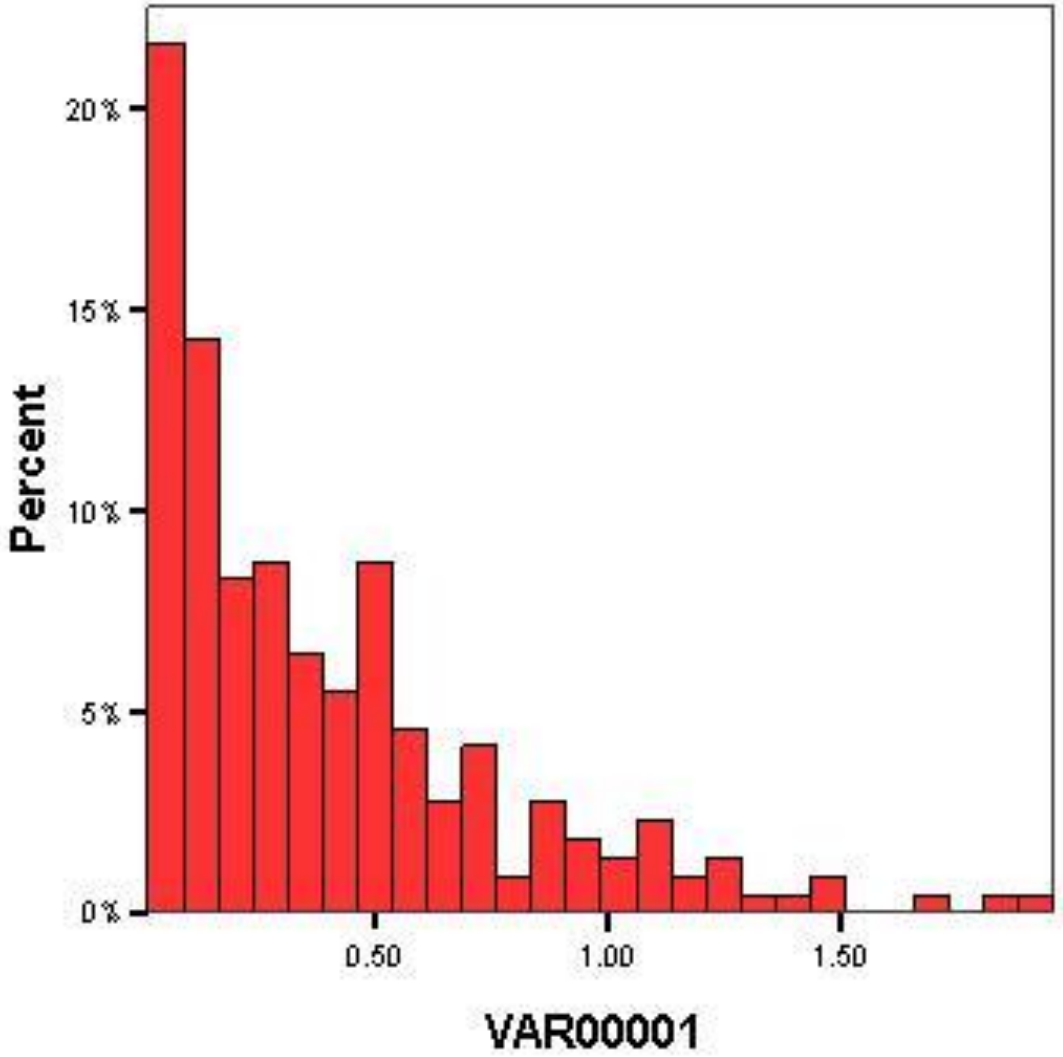
Box plot


0.00
0.50
1.00
1.50
2.00
VAR00001
214
213
212
211
210
216
209
208
207

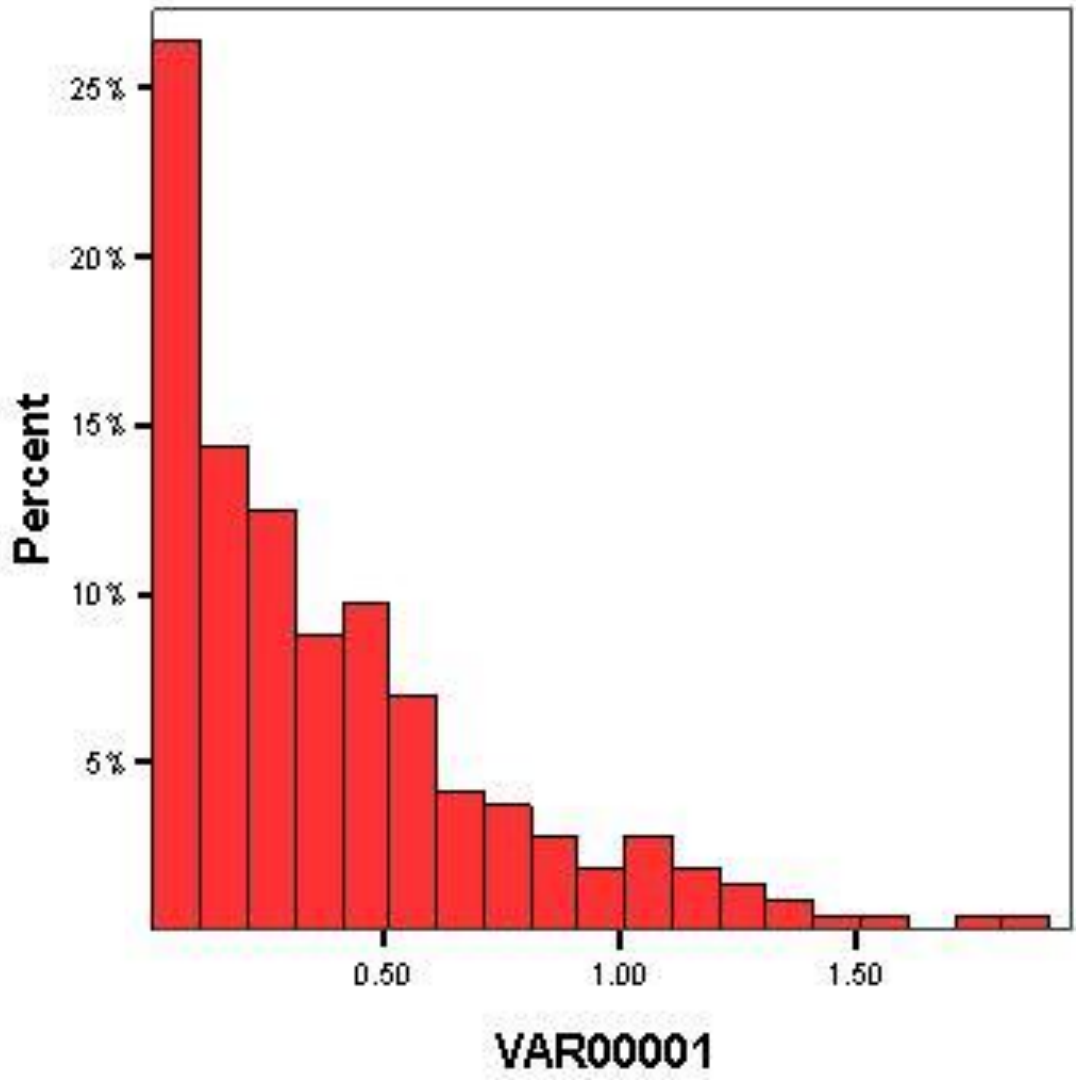
Histograms



Histograms



Histograms



Clues from summary statistics

- For the symmetric distributions mean and median should match. In the sample data, if these values are sufficiently close to each other, we can think of a symmetric distribution (e.g. normal).
- Coefficient of variation (*cv*): (ratio of std dev and the mean) for continuous distributions. The *cv = 1* for exponential dist. If the histogram looks like a slightly right-skewed curve with *cv >1*, then lognormal could be better approximation of the distribution.

Note: For many distributions *cv* may not even be properly defined. When? Examples?

Clues from summary statistics

- Lexis ratio: same as *cv* for discrete distributions.
- Skewness (*v*): measure of symmetry of a distribution. For normal dist. $v = 0$. For $v>0$, the distribution is skewed towards right (exponential dist, $v = 2$). And for $v<0$, the distribution is skewed towards left.

Parameter estimation

- Once distribution is guessed, the next step is estimating the parameters of the distribution.

- Each distribution has a set of parameters.
- ✓ Normal distribution has mean and standard deviation
- ✓ Exponential distribution has a “λ”.

- Most common method of parameter estimation: MLE (What is this?)

Goodness-of-fit

- For the input data we have, we have assumed a probability distribution.
- We also have estimated the parameters for the same.
- How do we know this fitted distribution is “good enough?”

- It can be checked by several methods:
  1. Frequency comparison (a bit technical)
  2. Probability plots (visual tool)
  3. Goodness-of-fit tests (statistical test of goodness. Very widely used).

Probability plots

Q-Q plot: Quantile-quantile plot

- Graph of the q_i-quantile of a fitted (model) distribution versus the q_i-quantile of the sample distribution.

$$x_{q_i}^{M} = \hat{F}^{-1}(q_i)$$

$$x_{q_i}^{S} = \tilde{F}_n^{-1}(q_i) = X_{(i)}, i = 1,2,\ldots n.$$

- If *F^(x)* is the correct distribution that is fitted, for a large sample size, then *F^(x)* and $F_n(x)$ will be close together and the Q-Q plot will be approximately linear with intercept 0 and slope 1.
- For small sample, even if *F^(x)* is the correct distribution, there will some departure from the straight line.

Probability plots

- P-P plot: Probability-Probability plot.

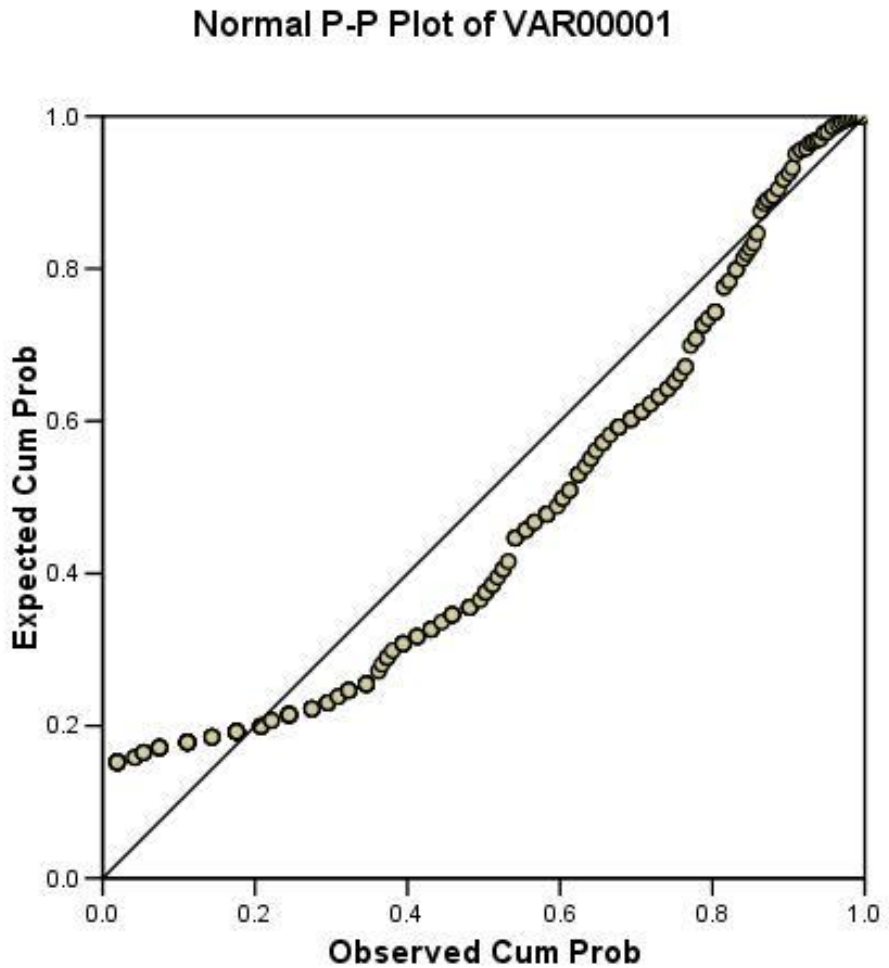
  A graph of the model probability $\hat{F}\left(X_{(i)}\right)$ against the sample probability $\tilde{F}_n\left(X_{(i)}\right) = q_i, i = 1,2,\ldots n.$

- It is valid for both continuous as well as discrete data sets.
- If *F^(x)* is the correct distribution that is fitted, for a large sample size, then *F^(x)* and $F_n(x)$ will be close together and the P-P plot will be approximately linear with intercept 0 and slope 1.

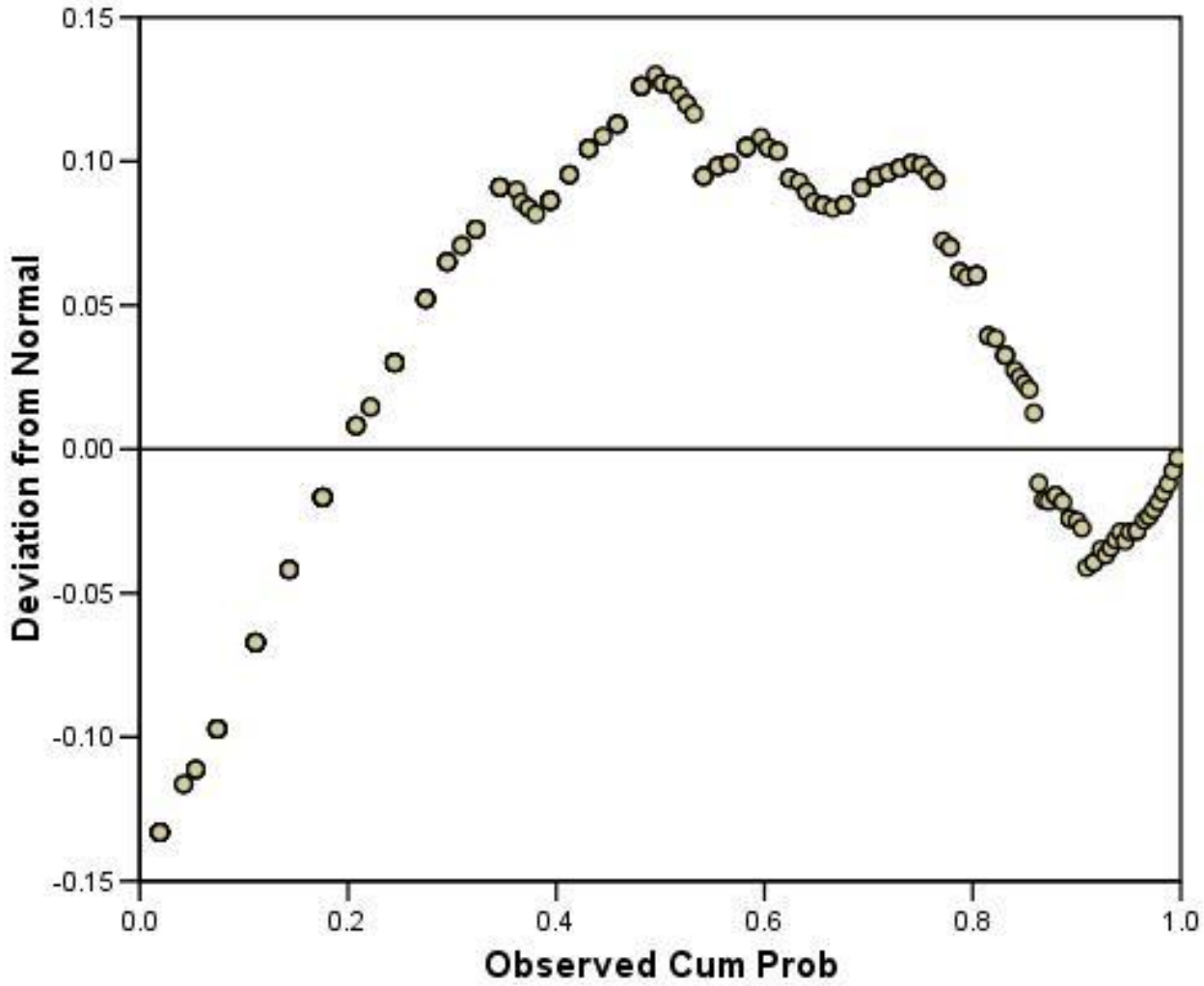
Probability plots

- The *Q-Q* plot will amplify the differences between the tails of the model distribution and the sample distribution.
- Whereas, the *P-P* plot will amplify the differences at the middle portion of the model and sample distribution.

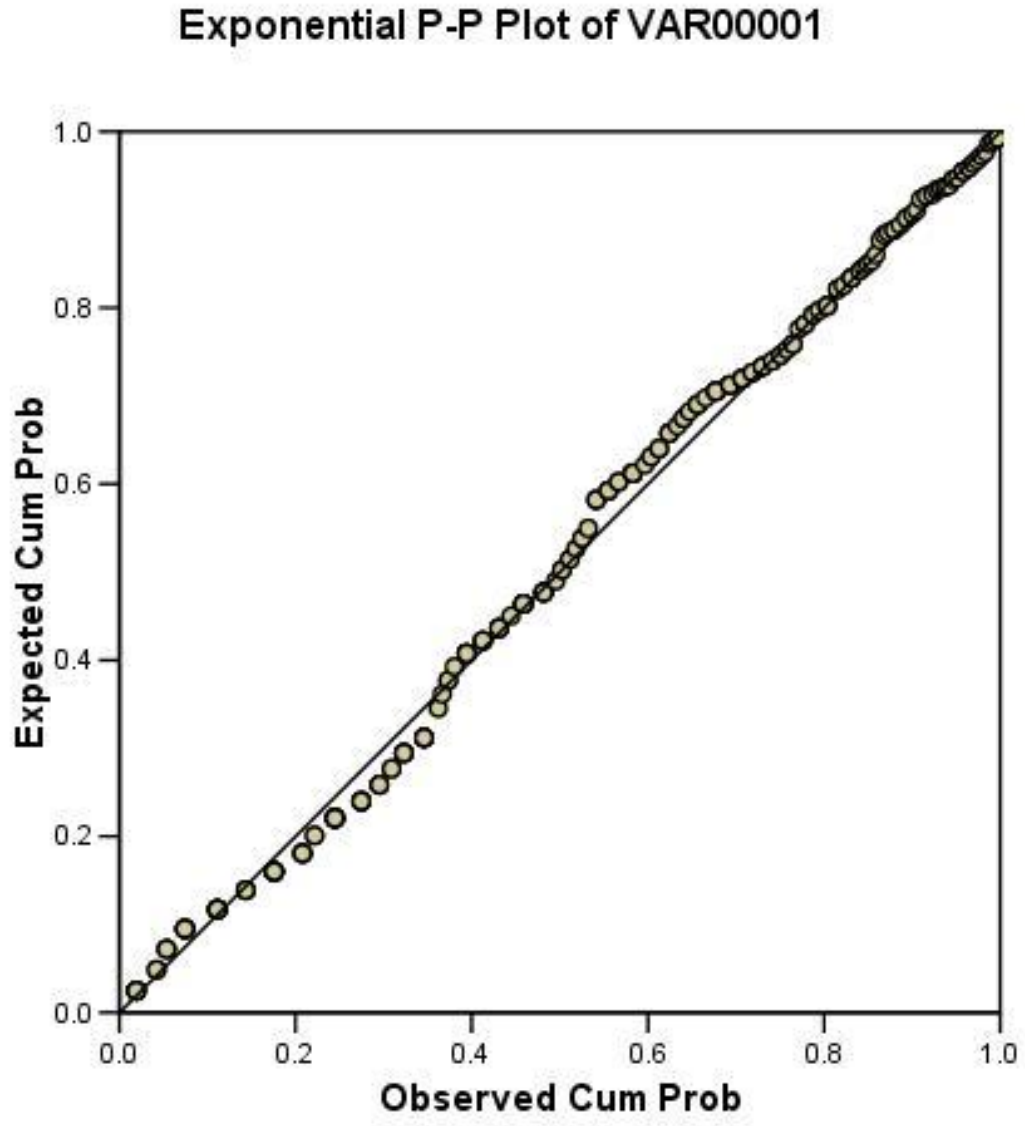
Probability plots: Dataset



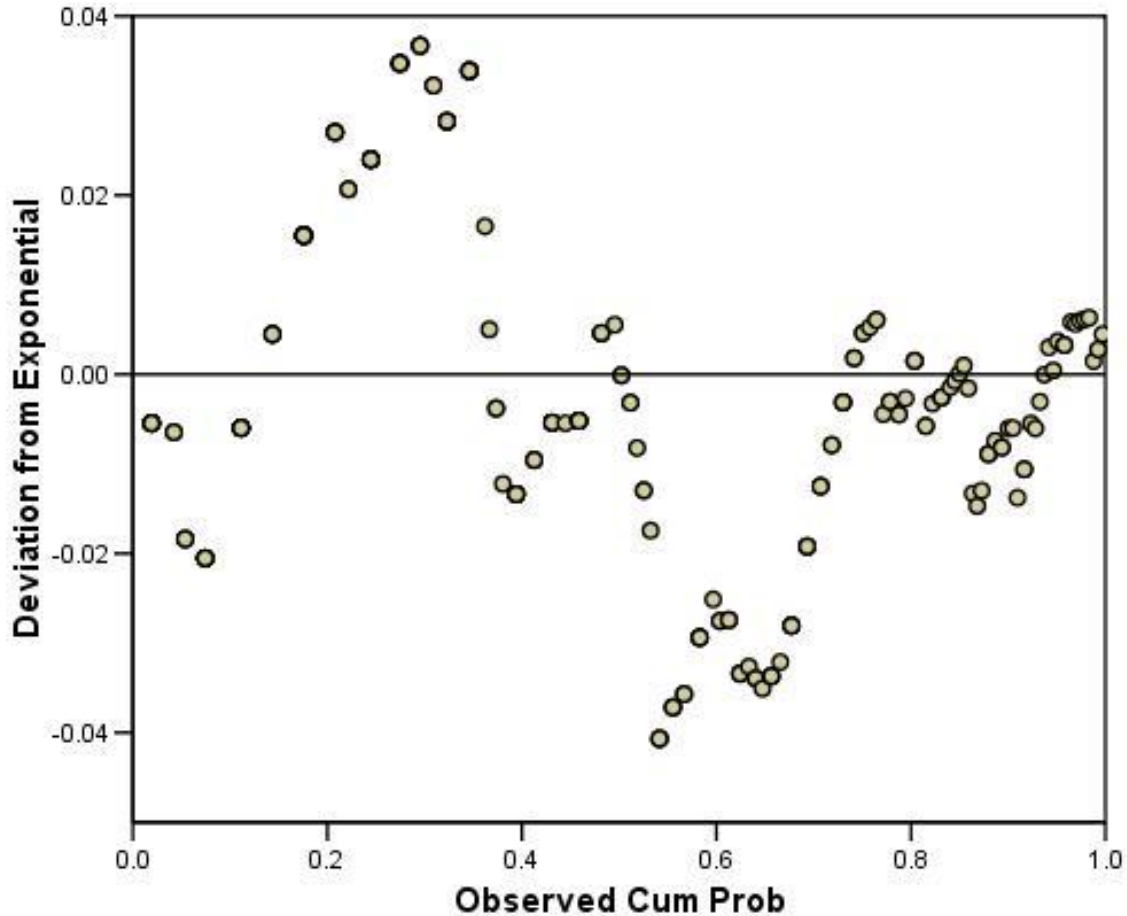
Probability plots: Dataset



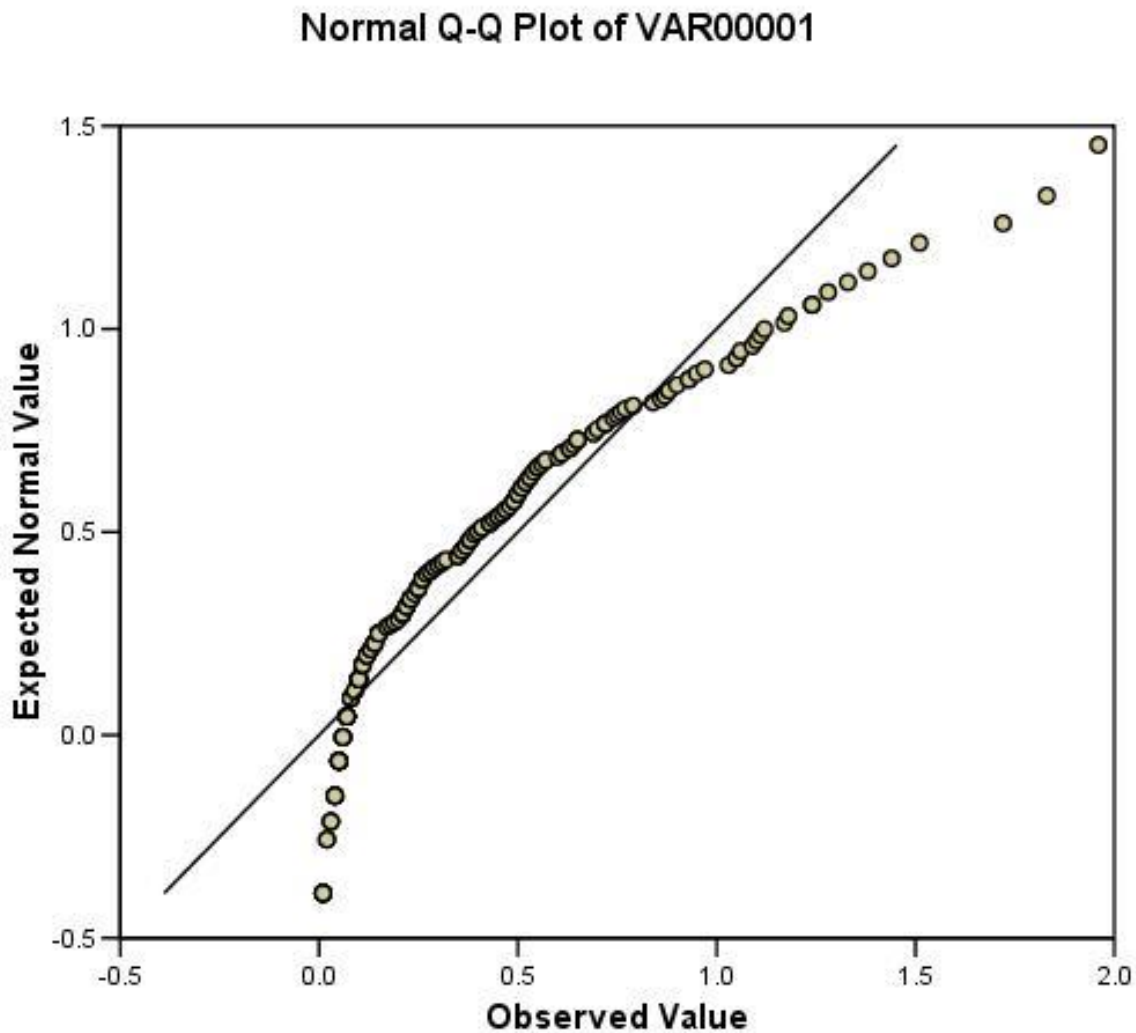
Probability plots: Dataset



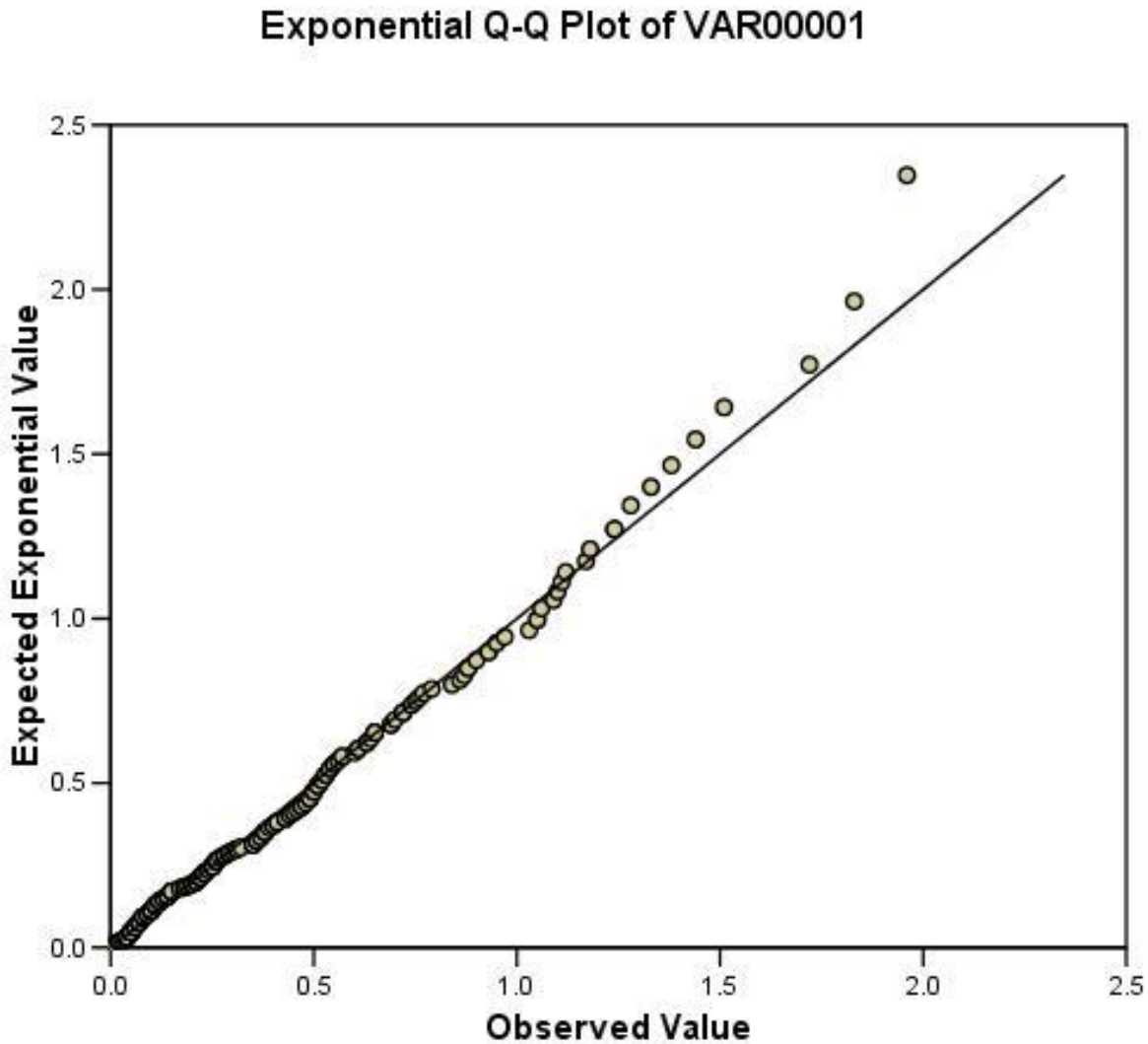
Probability plots: Dataset



Probability plots: Dataset



Probability plots: Dataset



Goodness-of-fit tests

- A goodness-of-fit test is a statistical hypothesis test that is used to assess formally whether the observations $X_1, X_2, X_3 \ldots X_n$ are an independent sample from a particular distribution with function $F\hat{}$.

H_0: The X_i's are IID random variables with distribution function $F\hat{}$.

- Two famous tests:

1. Chi-square test
2. Kolmogorov - Smirnov test

Chi-square test

- Applicable for both, continuous as well as discrete, distributions.
- Method of calculating chi-square test statistic:

1. Divide the entire range of fitted distribution into k adjacent intervals -- $[a_0, a_1)$, $[a_1, a_2), \ldots [a_{k-1}, a_k)$, where it could that $a_0 = -\infty$ in which case the first interval is $(-\infty, a_1)$ and/or $a_k = \infty$.

N_j = # of X_i's in the jth interval $[a_{j-1}, a_j)$, $j = 1, 2 \ldots n$.

2. Next, we compute the expected proportion of X_i's that would fall in the jth interval if we were sampling from fitted distribution

Chi-square test

$$\text{For continuous distributions}: p_j = \int_{a_{j-1}}^{a_j} \hat{f}(x)dx$$

$$\text{For discrete distributions}: p_j = \sum_{a_{j-1} \le x_j < a_j} \hat{p}(x_j).$$

- Finally the test statistic is calculated as:

$$\chi^2 = \sum_{j=1}^{k} \frac{\left(N_j - np_j\right)^2}{np_j}.$$

Chi-square test

- This calculated value of the test statistic is compared with the tabulated value of chi-square distribution with *k-1* df at *1-*α level of significance.

$$If \quad \chi^2 > \chi^2_{k-1,1-\alpha} \quad \text{Reject } H_0$$

$$If \quad \chi^2 \leq \chi^2_{k-1,1-\alpha} \quad \text{Do not Reject } H_0$$