2.3 Guessing the Distribution

Sunday, 09 October 2022 12:01

<table>
<tr>
<td>Summary</td>
<td>

- Q-Q plot
- P-P plot
- Chi-Square test

</td>
</tr>
<tr>
<td>

- What are the two probability plots we use to check goodness-of-fit?
- How do you interpret Q-Q plot?
- What points to we plot in Q-Q plot?

</td>
<td>

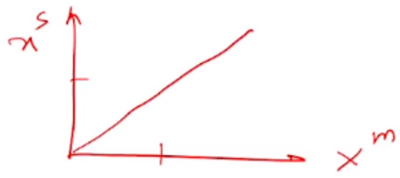
Probability plots

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.

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

- 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.

- $F^\wedge$ :: indicates the distribution that we're trying to fit
  $F^\sim{}_n$ :: indicates the distribution that come from the sample
- If $x^M{}_{qi}$ that comes from the fitted/model distribution matches with the $x^S{}_{qi}$ that comes from the sample distribution, then you're going to get a line.
- 

- You'll plot all the x that comes from X^M, that is model distribution on x-axis, and plot that against the x that comes from sample distribution, X^S.
- If x_i from model distribution matches with the sample distribution, you'll get a nice 45° line in Q-Q plot.
- This line will have an intercept of 0 and a slope of 1.
- Practically we'll get a line that is around this 45° line. And how far away we are from this 45° line determines how good is the model distribution.
- If very far from this line, then the model distribution doesn't match the sample distribution.

</td>
</tr>
<tr>
<td>

- How do you interpret P-P plot?
- What points do we plot in P-P plot?

</td>
<td>

Probability plots

- P-P plot: Probability-Probability plot.

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

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

- Here, we compare probability distributions: $F^\wedge$ and $F^\sim$ (CDF).

</td>
</tr>
<tr>
<td>

- Q-Q plot amplifies the differences between the _____.
- P-P plot amplifies the differences between the _____.

</td>
<td>

Probability plots

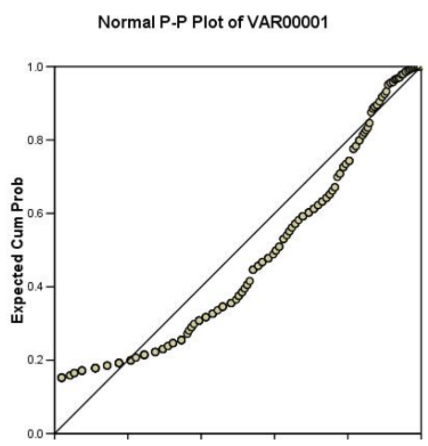
- The *Q-Q* plot will amplify the differences between the tails of the model distribution and the sample distribution.
- Whereas, the *P-P* plot will amplify the differences at the middle portion of the model and sample distribution.

</td>
</tr>
<tr>
<td>

- In general, for both probability plots, what do we plot on the:
  1. x-axis?
  2. y-axis?

</td>
<td>

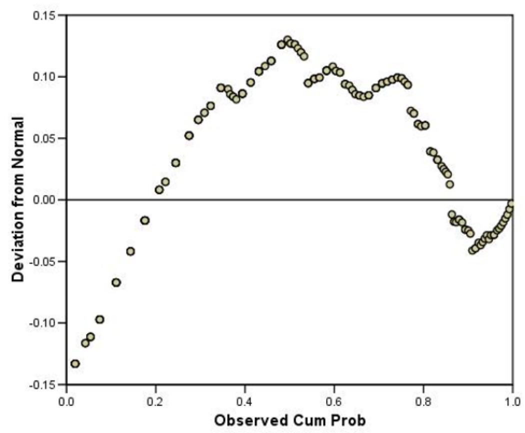
Probability plots: Dataset



</td>
</tr>
</table>

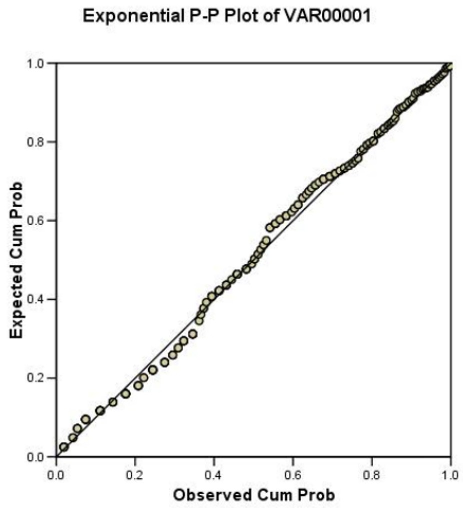
- On the observed data Var1, we've tried to fit Normal distribution using P-P Plot.
- X-axis :: Observed cumulative frequency
  Y-axis :: Expected(model) cumulative frequency
- So, both x- and y-axis will go from 0 to 1.
- Here, observed points don't seem to be close to expected points.

Probability plots: Dataset



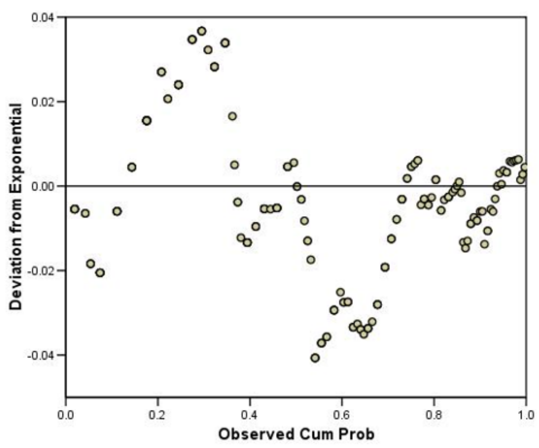
- Here is shown the deviation of observed points from Normal.
- The deviation seems to be high in P-P plot, of the order of 0.15.

Probability plots: Dataset



- Here, we're trying to fit exponential distribution in the dataset using P-P plot.
- The observed points seem to be very close to the 45^{o} line.

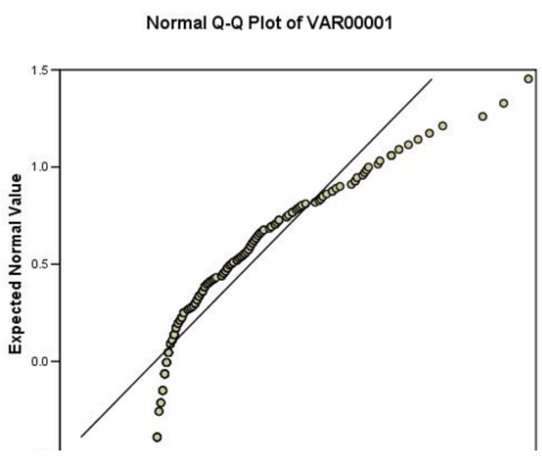
Probability plots: Dataset



- Here, the deviation of the dataset points from the exponential distribution is shown.
- Notice the scale of Y-axis, it's of the order of 0.04, which is small.

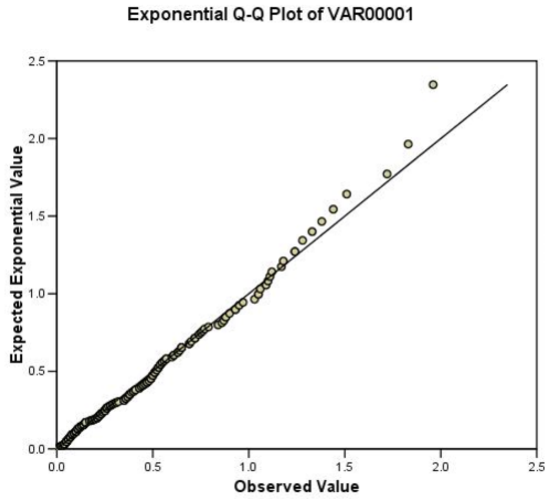
- Conclusion: In P-P plots, the exponential distribution seem to be a better fit than the normal distribution.

Probability plots: Dataset



- In this Q-Q plot, we're trying to fit normal distribution to the dataset Var1.
- X-axis :: X-values from the sample
  Y-axis :: Y-values from the model
- The observed points seem to be deviating from the 45° line.

Probability plots: Dataset



- Q-Q plot with the exponential distribution.
- Seems very close to the 45° line.
  There are some deviations in the upper portion, these deviations are for higher observed values.

- Conclusion: In Q-Q plots as well, the exponential distribution seem to be a better fit than the normal distribution.

Q. Do we now conclude that the exponential distribution is a good fit for the data?
A. No. We also need to look at statistical goodness-of-fit tests.

- Name the two famous statistical goodness-of-fit tests.

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 $F^\wedge$.

H_0: The X_i's are IID random variables with distribution function $F^\wedge$.

- Two famous tests:

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

- How do you calculate chi-square test?
  - Look here.

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:

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

	• Better formula is given here.
• Based on chi-square test statistic when do we: • Accept H_0? • Reject H_0?	**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. $If \quad \chi^2 > \chi^2_{k-1,1-\alpha}$ Reject H_0 $If \quad \chi^2 \leq \chi^2_{k-1,1-\alpha}$ Do not Reject H_0
	• The data given to us was a time data. Time to get a service in a bank. • The exponential distribution has a strong association with the queuing theory.