2.2 Business Example

Sunday, 02 October 2022 8:22

Summary	• Interpreting Descriptive Statistics • Goodness-of-fit

Business example

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

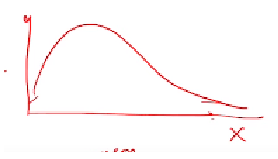
- What is the relation between mean, mode and median for symmetric distributions?
- Try to interpret the given statistics.
- How do you interpret skewness?
  - If skewness is positive, the data is skewed to the _____?
    - How will its shape look like?

Summary statistics

VAR00001

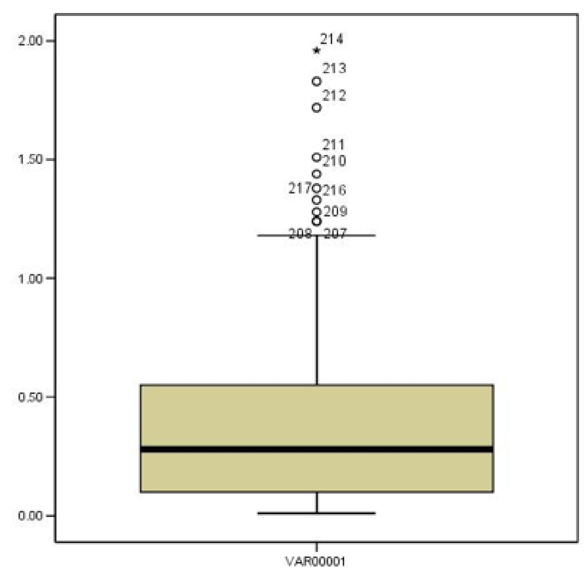
N	Valid	217
	Missing	1
Mean		.4012
Median		.2800
Mode		.05[a]
Std. Deviation		.38093
Variance		.145
Skewness		1.466
Std. Error of Skewness		.165
Range		1.95
Minimum		.01
Maximum		1.96
Percentiles	25	.1000
	50	.2800
	75	.5500

a. Multiple modes exist. The smallest value is shown

- For symmetric distributions: Mean = Mode = Median (or very close to each other) (eg, normal distribution)
- But here, mean, mode and median are different. It means it's not a symmetric distribution.
  So it rules all the symmetric theoretical distributions for us.
- Looking at the min-max of the data points, it can be observed that none of the data point takes on negative value.
  Rules out all the distributions that go to the negative side of the line.
- Skewness tells us about the symmetry of the distribution.
  The value of skewness is positive, so it's a positive skew, i.e., the data is skewed to the right.
  Right Tail > Left Tail
- 

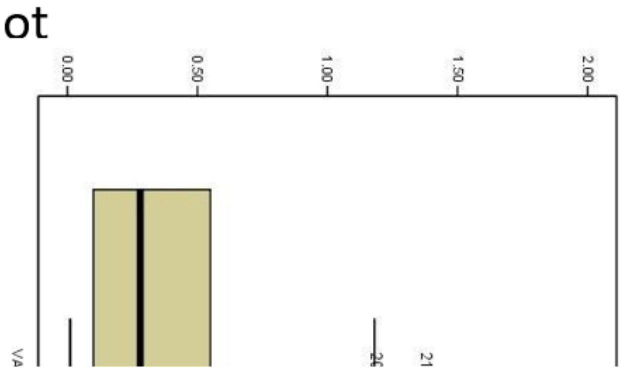
- So consider all the positive skewed distributions as the potential distributions here.

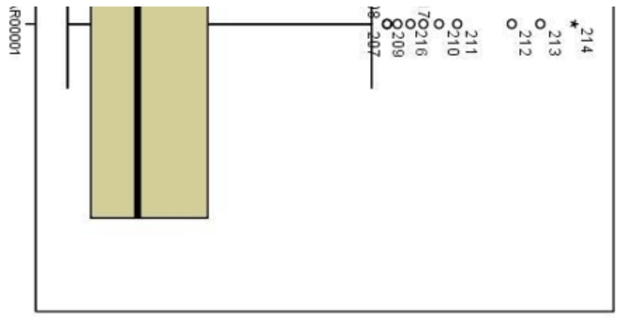
Box plot



- Below is given the same plot tilted 90 degrees, to get a better picture of the points on x-axis.

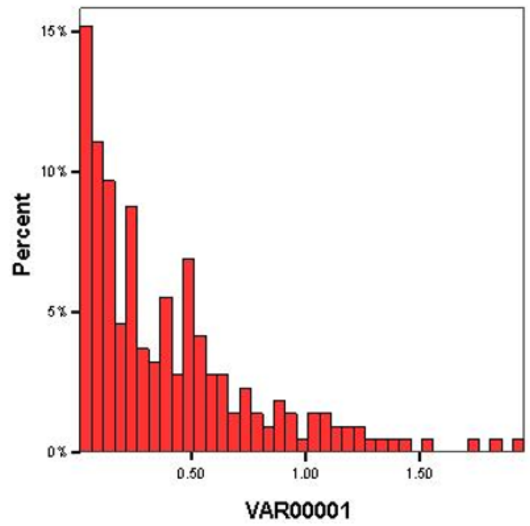
Box plot





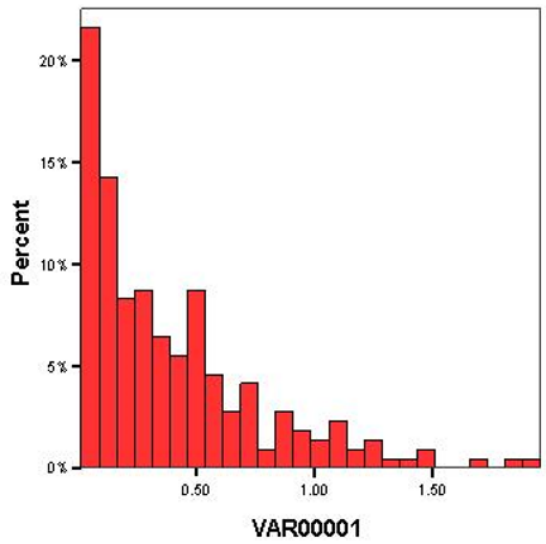
- It's clear from the box plot that this distribution has a positive skew.

Histograms



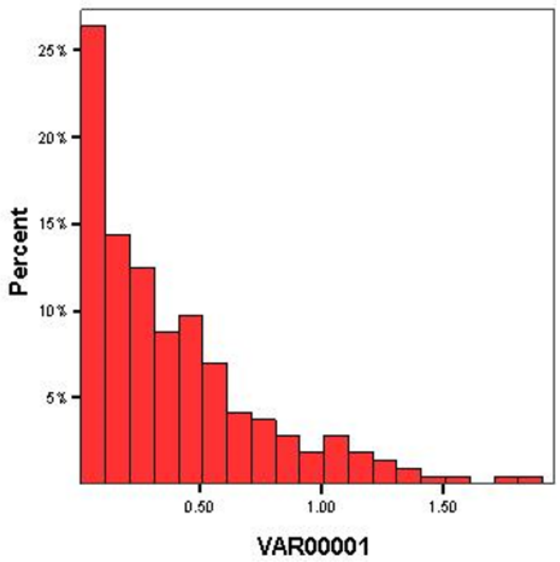
- Large number of values close to 0.
- Very few values more than 1.5.

Histograms



- We increased the width of the bars here.
- Frequency seems to be dropping as you go to the right.

Histograms



- Even thicker bars here.

- What is Coefficient of variation, cv?
  - This statistic is for continuous or discrete distributions?
  - cv = 1 for?
  - cv > 1 for?

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?

- $Coefficient\ of\ variation,\ (cv) = \frac{Standard\ Dev}{Mean} = \frac{\sigma}{\mu}$
- $cv = 1$ :: Exponential Distribution
  for slightly right-skewed curve with $cv > 1$ :: Lognormal Distribution
- In some cases, cv may not be defined.
  Take an example of standard normal distribution where $\mu = 0$.
- Here, $cv = \frac{0.38093}{0.4012} = 0.9495 \approx 1$

- Lexis ratio
- Skewness (v)
  - $v = 0$ for?
  - $v > 0$?
    - $v = 2$ for?
  - $v < 0$?

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.

- If Skewness,
  - v = 0 :: Normal distribution
  - v > 0 :: Distribution is skewed towards right
    - v = 2 :: Exponential Distribution
  - v < 0 :: Distribution is skewed towards left
- We'll try to fit exponential distribution here.

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?)

- How can we check the goodness-of-fit of the fitted distribution?
  - Name the methods
  - Explain

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).

- We're trying to fit exponential distribution here.

1. In frequency comparison, we can compare the frequency that comes from exponential distribution with the frequency you have observed from the dataset. If the frequencies match, we say that exponential distribution is good fit.
2. Probability plots are visual tools that tell you if the observed frequency/quartiles/percentiles matches with the frequency/quartiles/percentiles of the distribution. If it fits, you'll get a line , if it doesn't fit, you'll be far away from that line.
3. Goodness of fit tests: Many of them uses Chi-Square tests.