APPLIED STATISTICAL METHODS

Mathematics Institute

Faculty of Mathematics and Computer Science

Wroclaw University – Fall 2015

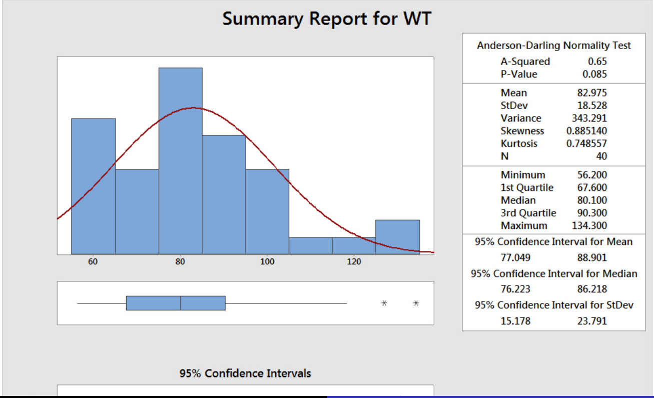
LECTURE 3

Towards data analysis - modeling intro

Wroclaw, 5 - 9 October, 2015

Exploratory Data analysis - recap:

Given a data set $X_1, X_2, \ldots, X_n$, we use numerical and graphical sumamries to describe data sets. These include graphs such as histogram and box plot, and numerical y statistics such as mean, median, standard deviation of a data set. There is one "standard" plot in MINITAB that includes both histogram, boxplot, and other information about a data set. It is called "Graphical Summary" and looks like this:



Summary Report for WT

```
Anderson-Darling Normality Test
   A-Squared          0.65
   P-Value           0.085

   Mean             82.975
   StDev            18.528
   Variance        343.291
   Skewness       0.885140
   Kurtosis       0.748557
   N                    40

   Minimum          56.200
   1st Quartile     67.600
   Median           80.100
   3rd Quartile     90.300
   Maximum         134.300
95% Confidence Interval for Mean
   77.049           88.901
95% Confidence Interval for Median
   76.223           86.218
95% Confidence Interval for StDev
   15.178           23.791
```

Empirical cdf from a sample.

Let $X_1, X_2, \ldots, X_n$ be iid from a distribution with cdf F. The empirical cdf (ecdf) of that sample is defined as

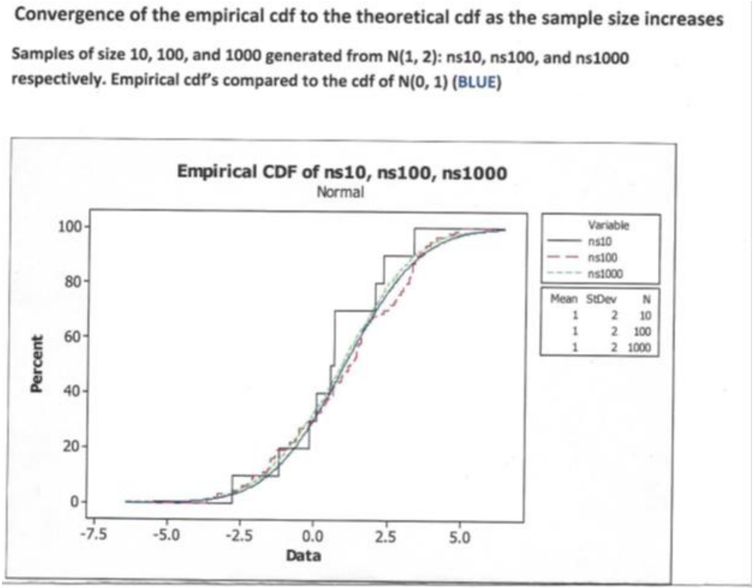
$$\hat{F}_n(x) = \text{ or } = F_{e,n}(x) = \frac{\# X_i' s \leq x}{n}.$$

Fundamental Theorem of Mathematical Statistics. As n increases, the empirical cdf converges uniformly to the theoretical cdf F.

$$\lim_{n \to \infty} max_x |\hat{F}_n(x) - F(x)| = 0.$$

This convergence is illustrated in Figure 1.

Convergence of the ECDF to the theoretical cdf



Convergence of the empirical cdf to the theoretical cdf as the sample size increases

Samples of size 10, 100, and 1000 generated from N(1, 2): ns10, ns100, and ns1000 respectively. Empirical cdf's compared to the cdf of N(0, 1) (BLUE)

How do we use ECDF?

Using ECDF:

- Modeling and checking if a model fits the data, or
- Checking if two (or more) samples are coming from the same distribution, etc.

The idea: If model and empirical cdfs are “close”, then the model “is close” to the distribution that generated the data. Similarly, if ecdf’s from two samples are close, then the distributions the samples came from may also be ”close”. Examples of “good fit” and “poor fit” are below

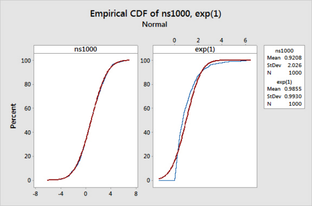


Figure: ECDFs of samples from N(1, 4) (left panel) and exp(1) (right panel) overlayed on the theoretical (model) cdf of N(1, 4).

Fitting models: Probability plots

A probability plot: theoretical quantiles of a model versus empirical quantiles of the data. The data are "empirical quantiles". We find the corresponding quantiles of the model we wish to fit.

Def: Empirical/sample quantiles. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with cdf F and the quantile function Q. To find sample quantiles, we sort the data in the increasing order to get "order statistics": $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. Each $X_{(i)}$ is the i/nth empirical quantile because

$$\text{Relative frequency of observations less than or equal to } X_{(i)} = \frac{i}{n}.$$

Relative frequency approximates probability, so if rv $X \sim F$, then $P(X \leq x_{(i)}) \approx \frac{i}{n}$.

Fitting models: Probability plots, contd.

If a model fits the distribution of the data:

- The theoretical (model) quantiles should be close to the empirical (data) quantiles,
- The plot of empirical quantiles versus theoretical (model) quantiles should be approximately a straight line with slope equal to 1 though the origin (line y=x).

Probability plots, contd.

The figure below illustrates models that fit well and models that fit poorly using probability plots for.

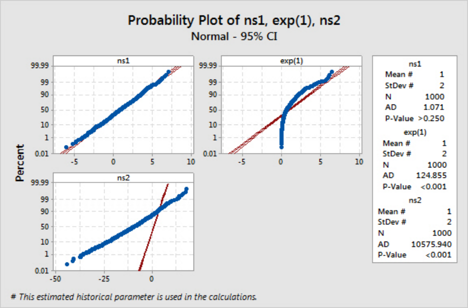


Figure: Probability plots of samples from N(1, 4) (left top panel), and exp(1) (right top panel), and N(-10, 100) overlayed on the theoretical (model) cdf of N(1, 4).

The plot in the bottom row of Figure 3 shows a straight line probability plot, however the line is not y=x. This reflect an important property of the probability plots.

Probability plots, contd.

Let X be a rv, and let $Y = aX + b$, where a and b are real numbers. We call a scale and b (horizontal) shift of X. Note, that $VarY = a^2 VarX$, so standard deviation of Y is the standard deviation of X multiplied (scaled) by $|a|$. Let $Q_X(\cdot)$ and $Q_Y(\cdot)$ be quantile functions of X and Y, respectively. Then

$$Q_X(p) = x_p \text{ iff } P(X \leq x_p) = p \text{ iff } P(\frac{Y-b}{a} \leq x_p) = p$$
$$\text{iff } P(Y \leq ax_p + b) = p \text{ iff } Q_Y(p) = aQ_X(p) + b.$$

Thus, if a rv Y is a linear function of rv X, the quantile function of Y is (the same) linear function of the quantile function of X. Hence the probability plot of a model X for a sample from Y will be close to a straight line, but the line will not be $y = x$. In fact, the line will be $y = ax + b$.

Practical use of probability plots.

- A "good" property of probability plots is that they show a straight line even if the data comes from a distribution that is shifted and scaled distribution of the model.
- That means that **probability plots fit "families" of models**, where the distributions in the family differ only by scale and/or center.

Quantile to Quantile (QQ)-plots.

- QQ-plots are constructed in the same way as probability plots but for two samples.
- We plot empirical quantiles of one sample versus the corresponding empirical quantiles of another sample to check if the samples came from the same distribution.

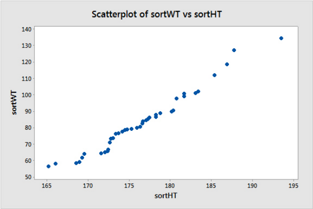


Figure: QQplot of Weight and Height for the MBODY data observations.

To plot this QQplot, I sorted the two columns and then used ”Scatter plot” of the sorted columns.