WROCLAW UNIVERSITY

APPLIED STATISTICAL METHODS – Fall 2015

LECTURE NOTES - 3nd SET - week 1

TOWARDS DATA ANALYSIS: Modeling and goodness of fit

Towards modeling

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:

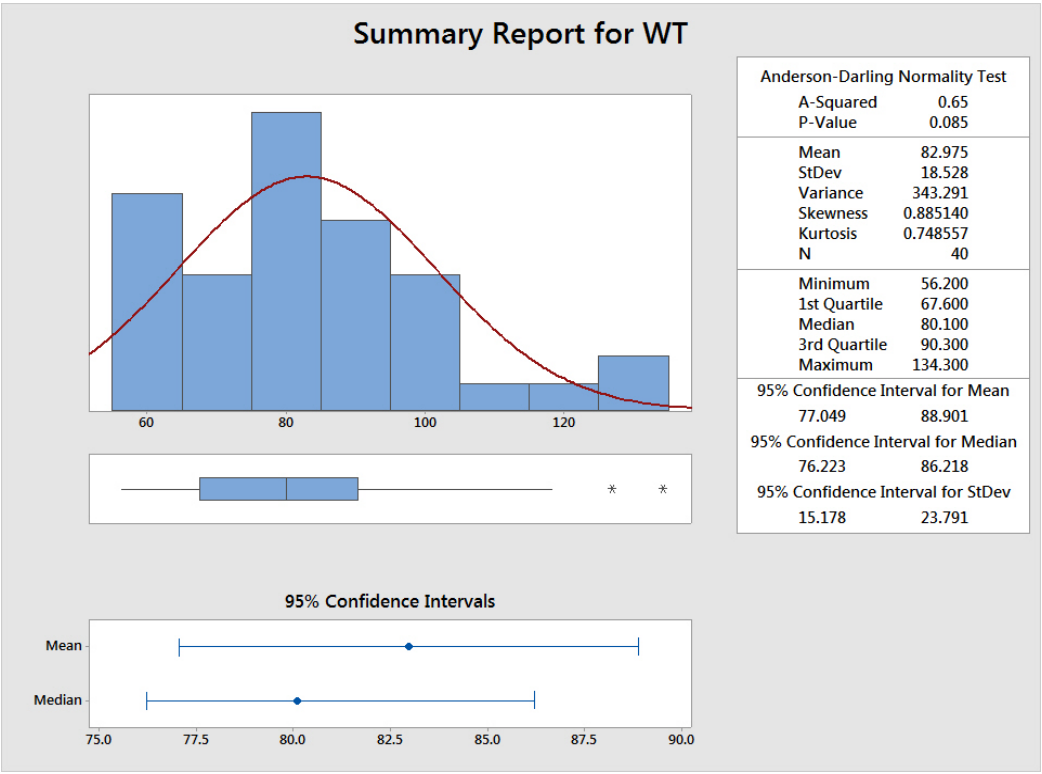

FIGURE 1. Graphical summary of the weight data.

The summary plot contains a histogram and a boxplot of the data. There is a curve on the histogram that approximates the shape of the histogram. This curve is the pdf of the "best fitting" normal distribution. The boxplot is plotted horizontally under the histogram so that the horizontal scale of the histogram and the vertical scale of the boxplot, both showing values of the observation, are aligned. Below the boxplot there are two confidence intervals: one for the mean, the other for the median of the distribution that generated

the data. On the right side of the graph, there is a box with several numerical summaries and test results. We will talk about the test later today. the summaries are the same as those produced by the summary stats command.

TOWARDS MODELING

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.

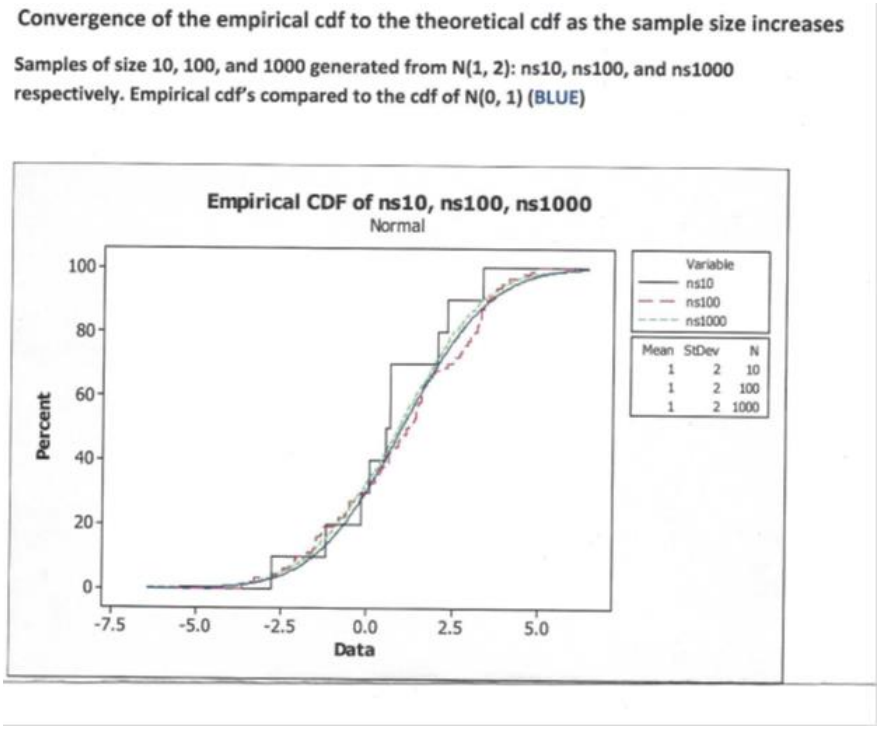

FIGURE 2. Convergence of empirical cdf's to the theoretical cdf of a normal rv with mean 1 and variance 2, as the sample size increases.

Graphical and numerical summaries are commonly produced at the beginning of analysis of a data set, because they provide an overview of the important characteristics of the data: center, spread, shape, and possibly outliers.

Using ecdf. One of the common uses of ecdf is in modeling and checking if a model fits the data, or if two samples are coming from the same distribution, etc. The idea here is that 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". See Figure 2 for examples of "good fit" and "poor fit".

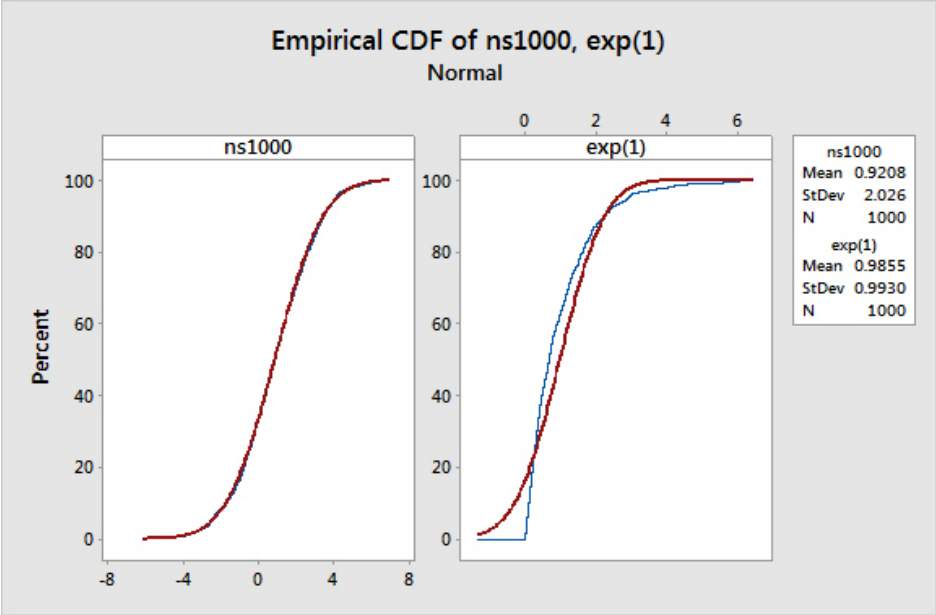

FIGURE 3. ECDFs of samples from N(1, 4) (left panel) and exp(1) (right panel) overlayed on the theoretical (model) cdf of N(1, 4).

Probability plots. Another technique for fitting models to data. A probability plot is plotting theoretical quantiles of a model versus empirical quantiles of the data. We use the data itself as "empirical quantiles", and find the corresponding quantiles of the model we wish to fit.

Empirical 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": $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. The sorted sample satisfies:

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}.$$

Note that each $X_{(i)}$ is the i/nth empirical quantile because

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

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

If a model fits the distribution of the data, then the theoretical (model) quantiles should be close to the empirical (data) quantiles, thus 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). Figure 3 illustrates probability plots for models that fit well and models that fit poorly.

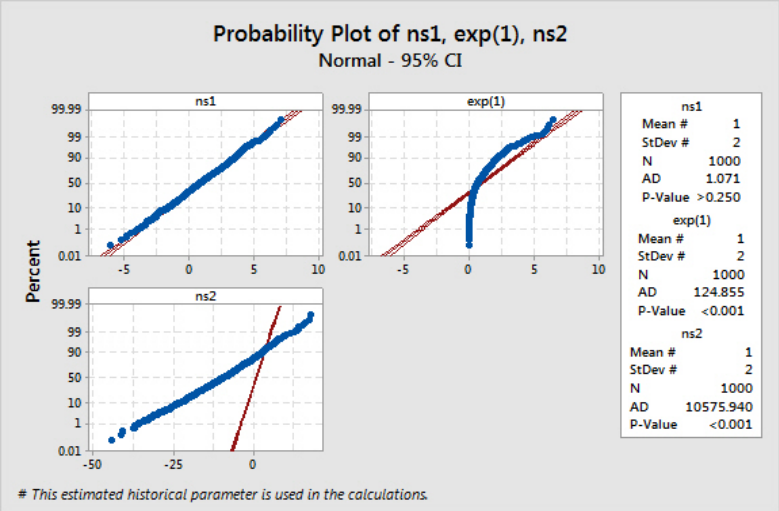

FIGURE 4. Probability plots of samples from N(1, 4) (left top panel), and exp(1) (right top panel), and N(-10, 100) overlayed on the theoretical (model) cdf of N(1, 4).

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.

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. If the model (distribution) we suspect fits the data indeed does fit the data, then the probability plot of the data with that model distribution should approximately follow a straight line. 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.

QQ-plots. QQ-plots or quantile-to-quantile plots are constructed in the same way as probability plots 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.