Chapter 3
Exploratory Data Analysis

INTRODUCTION

Exploratory data analysis is a means for gaining first insights and getting familiar with your data. Data are usually structured in a matrix form where the columns are *variables* (or attributes, characteristics, covariates) and the rows are *objects* or observations (e.g., persons, loans, or years) for the variables. Many databases are very complex. Exploring all data at hand typically requires dealing with masses of data, or big data. Looking at observations of variables object by object is therefore usually too time consuming, too costly, or simply not possible. If we were to examine every one, out of the more than 600,000 entries in our mortgage database for every variable, it would be impossible to draw any meaningful conclusions from this. Therefore we aggregate the information behind each variable and compute some *summary or descriptive statistics* and provide summarizing charts. We can do this in a *one-dimensional (univariate)* or a *multidimensional (multivariate)* way. One-dimensional data analysis treats every variable one by one and explores key measures for each variable separately whereas multidimensional data analysis treats variables jointly and explores dependencies and relations between variables. We start this chapter with univariate analysis and continue then with multivariate (particularly bivariate) analysis.

ONE-DIMENSIONAL ANALYSIS

We begin with exploratory data analysis, looking at variables separately in a one-dimensional way. This means we are only interested in empirical univariate distributions or parameters thereof, variable by variable, and do not yet analyze variables jointly or multivariately. The latter is done in the subsection where we are interested in correlations and dependencies between variables.

Observed Frequencies and Empirical Distributions

First we compute how often a specific value of a variable is observed. This is meaningful only when a variable has a finite set of possible values, that is, intechnical terms, when the variable is measured on a *discrete scale*. A simple example is the default of a mortgage loan, which usually has only two possible values (one for default and another for nondefault). Otherwise, if virtually any value of a variable is possible and each entry has a different value, the variable is measured on a *continuous scale*. An example is gross domestic product (GDP) growth, which can theoretically (if measured fine-grained) take any possible value between $-\infty$ and $+\infty$.

Now, consider we have as a starting point a sample with n observations for a variable X (e.g., thousands of observations for the variable FICO score in the mortgage database). For

each observation we measure a specific value of the FICO score, denoted by $x_1, \ldots, x_n$, which is called the *raw data*. Let the variable be either discrete or continuous, but grouped into classes (e.g., FICO scores from 350 to 370, 370 to 390). Then we denote the values or class numbers by $a_1, \ldots, a_k$ and count the absolute numbers of occurrence of each value or class number by:

$$h_j = h(a_j)$$

and the relative frequencies by:

$$f_j = \frac{h_j}{n}$$

Obviously, it holds that $\sum_{j=1}^{k} h_j = n$ and $\sum_{j=1}^{k} f_j = 1$. Moreover, we define the absolute and relative cumulative frequency $H(x)$ and $F(x)$ for each value x as the number or relative frequency of values being at most equal to x (i.e., being equal to x or lower).

$$H(x) = \sum_{a_j \leq x} h(a_j)$$

$$F(x) = \frac{H(x)}{n}$$

Graphically, this is a (nondecreasing) "stairway" function. In SAS, frequencies can easily be computed using PROC FREQ and graphically plotted using PROC UNIVARIATE. First, we compute the observed frequencies for the defaults in the data set. A default is coded as 1, and a nondefault is coded as 0.

```
PROC FREQ DATA=data.mortgage;
TABLES default_time;
RUN;
```

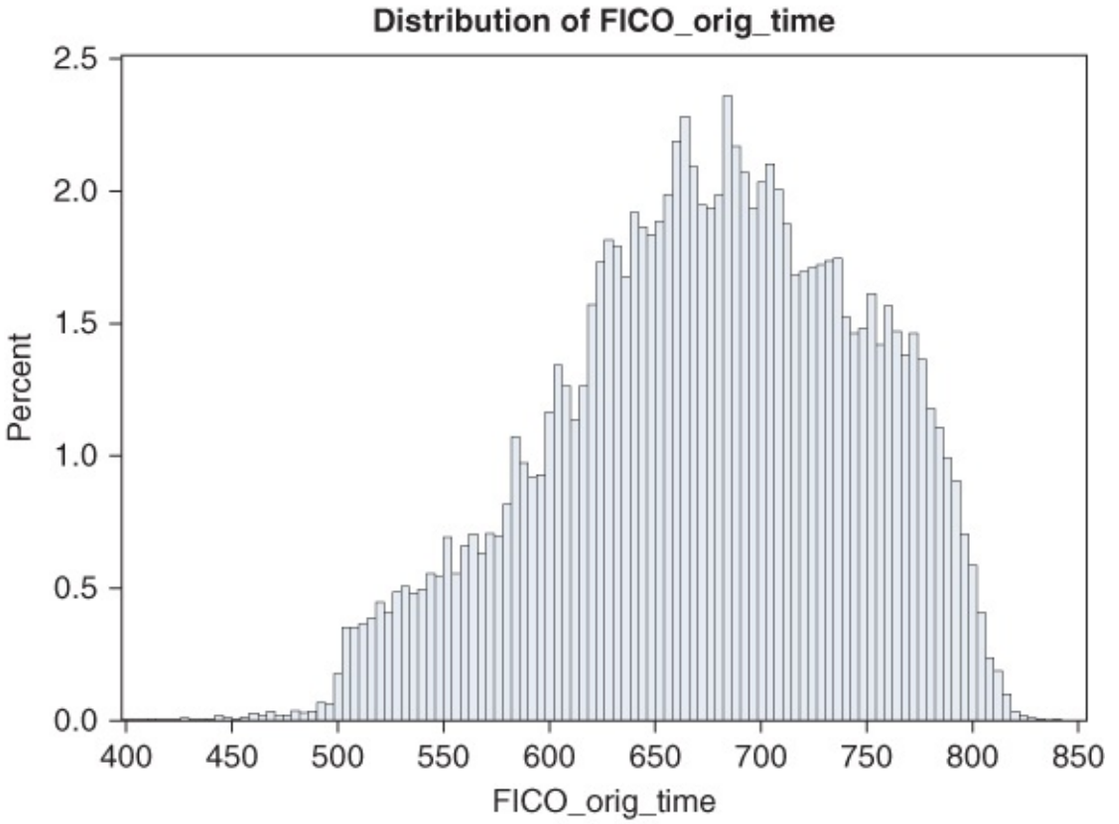
The output table (Exhibit 3.1) shows that there are 607,331 nondefaults and 15,158 defaults, which give relative frequencies of 97.56% and 2.44%, respectively, where the total number of observations is 622,489. Cumulative absolute (relative) frequencies are 607,331 (97.526%) for all values lower than or equal to 0 and 622,489 (100%) for all values lower than or equal to 1.

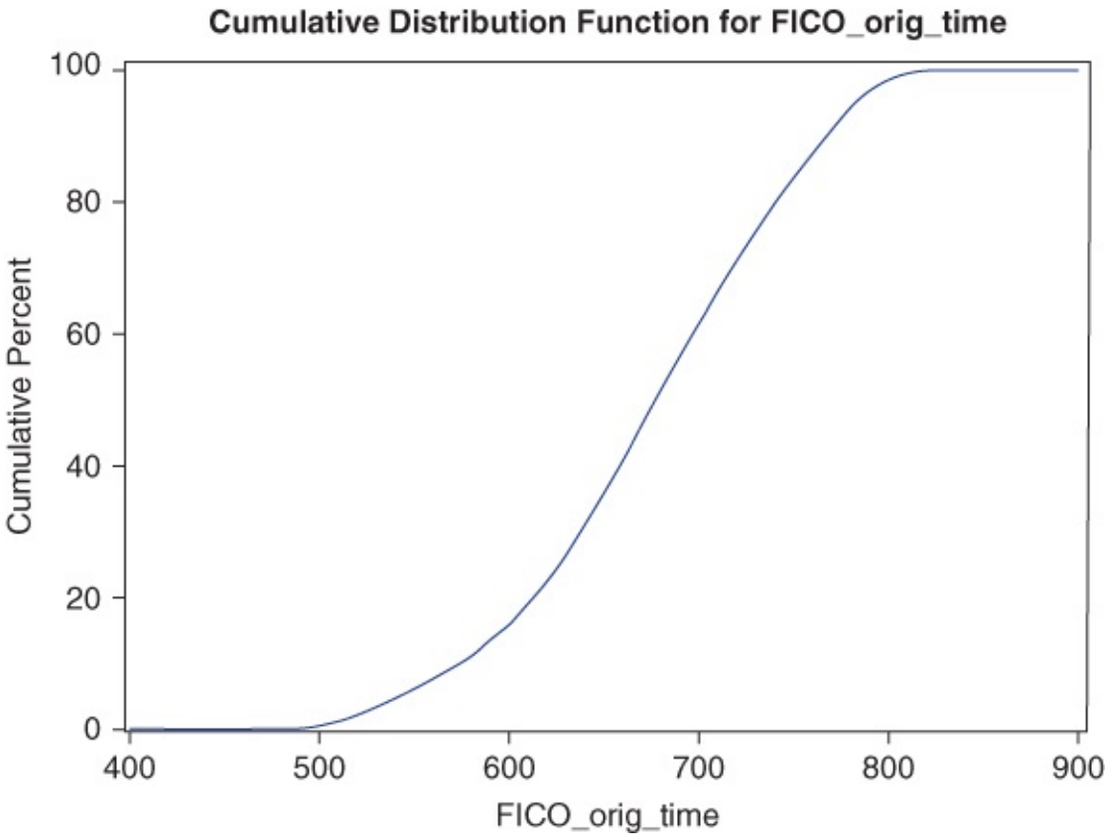
Exhibit 3.1 Absolute and Relative Frequencies

The FREQ Procedure				
default_time	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	607331	97.56	607331	97.56
1	15158	2.44	622489	100.00

Next, we compute histograms, which plot the absolute (or relative) frequencies for values or classes of variables and the empirical cumulative distribution function (CDF) exemplarily for

the variable's FICO score (FICO_orig_time) and LTV at origination (LTV_orig_time). The distribution of relative frequencies (called by the command HISTOGRAM in PROC UNIVARIATE) in Exhibit 3.2 shows increasing percentages for FICO up to a value of almost 700 and then decreasing percentages. The distribution is not symmetric as the tail on the left is longer (fatter) than on the right of the distribution. It is skewed to the left and steep to the right. For LTV we see that similar values have a rather high relative frequency compared to others, particular the value of 80. This is because LTV is the ratio of the outstanding loan amount to the collateral value, and banks traditionally extend loans in the region of 80 percent. The percentages in both histograms add up to 100 percent.





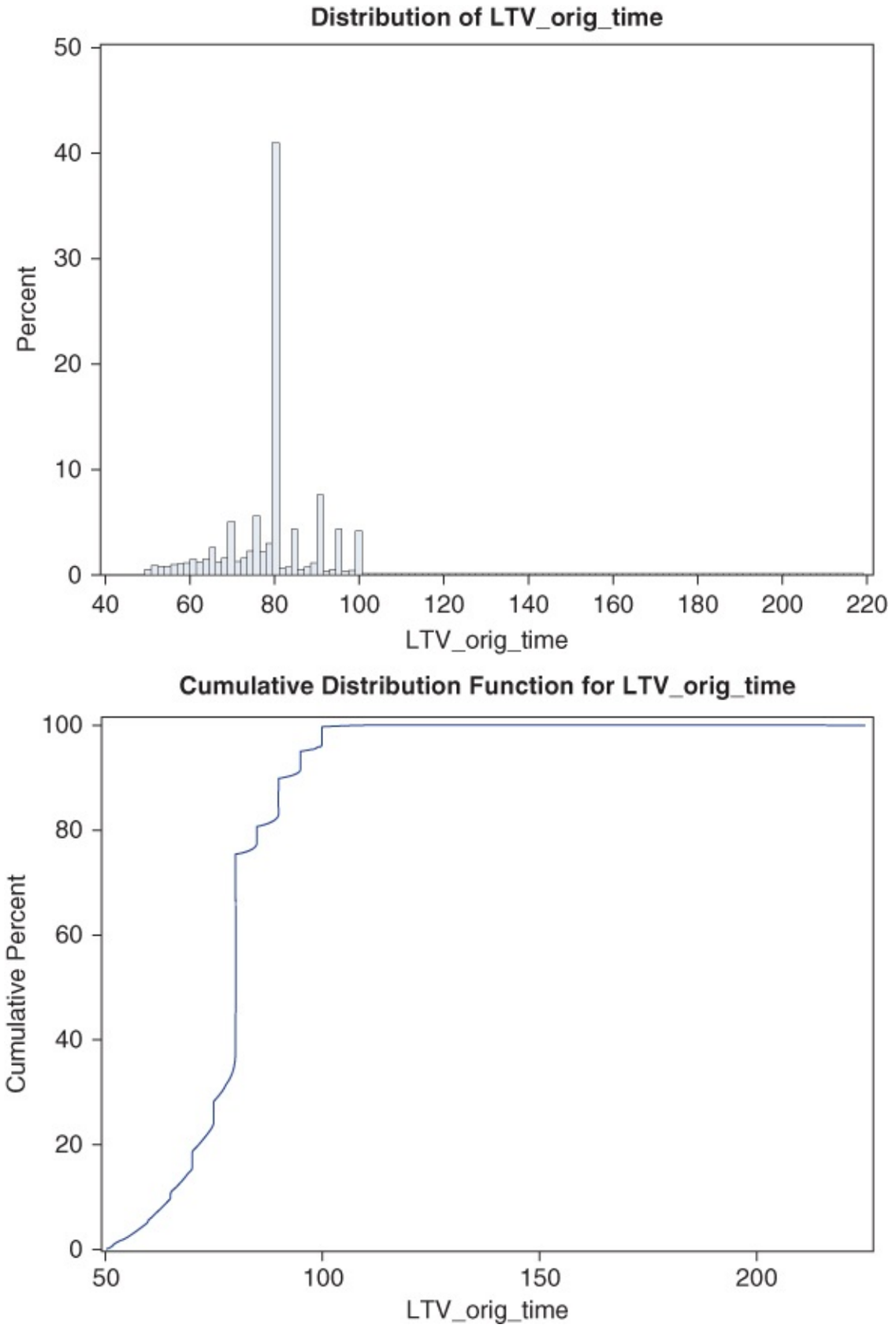


Exhibit 3.2 Histograms and CDF Plots

The CDFs for both variables start at 0 percent to the left and end at 100 percent to the right. It is nondecreasing and for FICO it looks almost continuous, because FICO seems to be an almost continuous variable. For LTV we see some "jump points" (or discontinuity points) at

the values where we identified high relative frequencies.

```
ODS GRAPHICS ON;
PROC UNIVARIATE DATA=data.mortgage;
VAR FICO_orig_time LTV_orig_time;
CDFPLOT FICO_orig_time LTV_orig_time;
HISTOGRAM FICO_orig_time LTV_orig_time;
RUN;
ODS GRAPHICS OFF;
```

Location Measures

In addition, or as an alternative to the description of the entire distribution, we often report summarizing measures. These measures give numerical characterizations about the location of the distribution, its dispersion, and its shape, and are generally called "moments." Three measures for location are commonly used: the mean, the median, and the mode. The mean of a distribution, or arithmetic average, is an equally weighted sum of each value of a variable summed over all observations. Assume we have n values $x_1, \ldots, x_n$; then the mean is:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The mean requires a variable to be measured on a metric, continuous scale. Another measure for location is the median, which requires at least ordinally scaled (i.e., ranked values). Let the raw observations be ordered from lowest to highest; that is, create the ordered raw data as $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ where the values in the parentheses denote the rank number of the observation. Then, if the number n of observations is uneven, the median is defined as:

$$x_{Med} = x_{\left(\frac{n+1}{2}\right)}$$

That is, the variable value of the observation which is exactly in the middle of the ordered list. If n is even, the median is defined as the average of both observations in the middle of the ordered list;

$$x_{Med} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2},$$

Finally, the mode is defined as the observation that is encountered most frequently in the data set; that is,

$$x_{Mod}: \text{Value with highest frequency}$$

Obviously, the mode is useful only when the variable is ordinal or categorical. Mean and median have important properties. The formula for the mean uses all data for its computation whereas the median processes only one (or two) data points. Thus, the mean is affected by extreme observations (outliers), whereas the median is robust with respect to outliers. If a distribution has one mode only and is symmetric, it holds that $\overline{x} = x_{Med} = x_{Mod}$. If it is skewed to

the left (right) it holds that $\bar{x} \leq x_{Med} \leq x_{Mod}$ ($\bar{x} \geq x_{Med} \geq x_{Mod}$).

A more general expression for the median is a quantile. A p-quantile x_p with $0 < p < 1$, is defined as the value for which

- At least a proportion p of sample values is lower than or equal to x_p.
- At least a proportion $1 - p$ of sample values is higher than or equal to x_p.

That is,

$$\frac{number\,(x - values \leq x_p)}{n} \geq p \quad \text{and} \quad \frac{number\,(x - values \geq x_p)}{n} \geq 1 - p$$

Special quantiles are:

$x_{0.5}$:	median
$x_{0.25}, x_{0.75}$:	lower and upper quartiles
$x_{0.1}, x_{0.2}, \ldots, x_{0.9}$:	deciles

Measures of location can be computed via SAS PROC MEANS where the requested measures are specified after the PROC MEANS command; see SAS Institute Inc. (2015). In Exhibit 3.3 we show exemplarily the output for the default indicator, FICO and LTV. The binary default variable has a mean of 0.0244, which corresponds to the default rate of 2.44 percent. Mean and mode are zero, and as there are 2.44 percent defaults (“ones”) in the data set, the 99 percent quantile is one. FICO and LTV have averages of 673.6169 and 78.9755, respectively, and the higher value for the median shows that both distributions are skewed to the left. The modes are 660 and 80, and 1 percent of all values are lower than or equal to 506 and 52.2, or higher than or equal to 801 and 100.

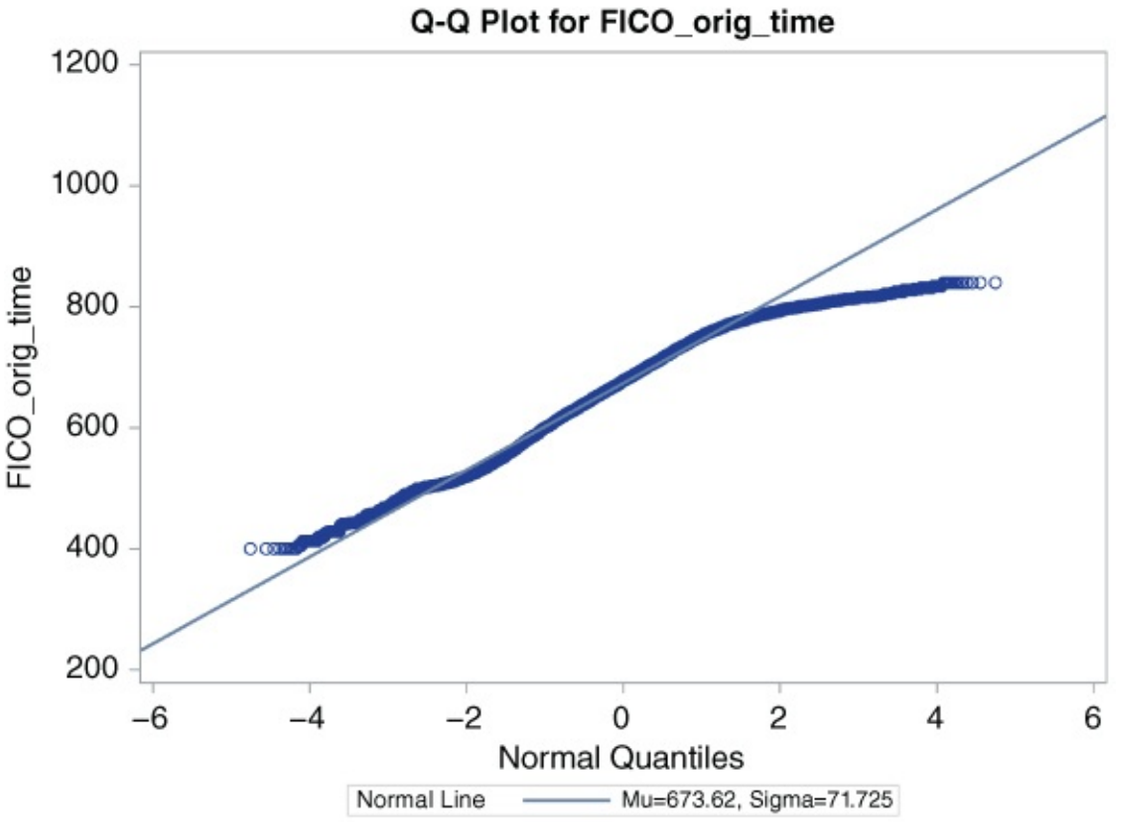
Exhibit 3.3 Location Measures

The MEANS Procedure						
Variable	***N***	**Mean**	**Median**	**Mode**	**1st Pctl**	**99th Pctl**
default_time	622489	0.0244	0.0000	0.0000	0.0000	1.0000
FICO_orig_time	622489	673.6169	678.0000	660.0000	506.0000	801.0000
LTV_orig_time	622489	78.9755	80.0000	80.0000	52.2000	100.0000

```
PROC MEANS DATA=data.mortgage
N MEAN MEDIAN MODE P1 P99 MAXDEC=4;
VAR default_time FICO_orig_time LTV_orig_time;
RUN;
```

Quantiles can be used for a graphical comparison with standard distributions, such as a normal distribution. The normal distribution is widely used in applications and is a symmetric distribution with a single mode. Using PROC UNIVARIATE, a quantile-quantile (Q-Q) plot can be created using the command QQPLOT, which compares for each value its quantile value

with the theoretical value under a specific distribution. Here we use the normal distribution with the same mean and standard deviation as the empirical data. If the data were from a normal distribution, both the empirical and the theoretical quantiles should be roughly equal and lie on the diagonal line. In Exhibit 3.4, we see divergences for FICO and LTV, particularly for extreme observations, which signals that the empirical data have different tails than the theoretical normal distribution.



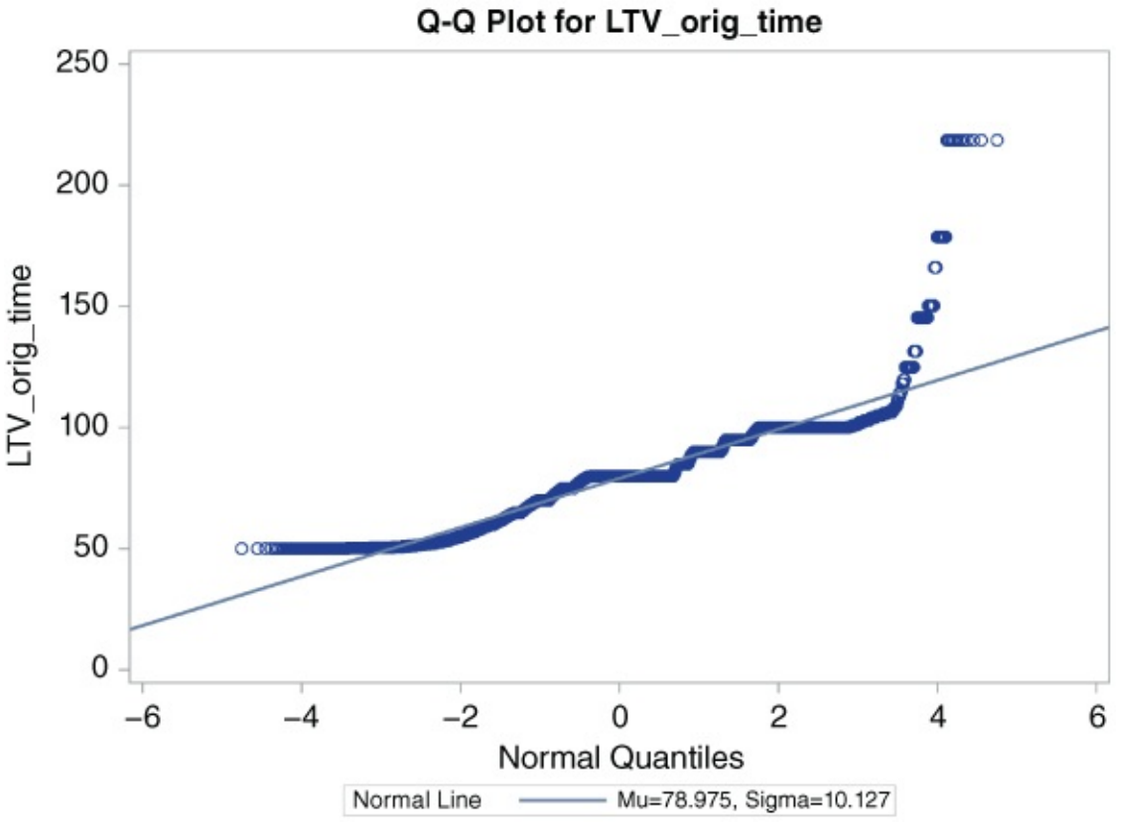


Exhibit 3.4 Q-Q Plot versus Normal Distribution

```
ODS GRAPHICS ON;
PROC UNIVARIATE DATA=data.mortgage NOPRINT;
QQPLOT FICO_orig_time LTV_orig_time
```

```
/NORMAL(MU=EST SIGMA=EST COLOR=LTGREY) ;
RUN;
ODS GRAPHICS OFF;
```

Dispersion Measures

Next, we discuss the most commonly used dispersion measures. The first is the span or range, which is simply the difference between the minimum and the maximum values. If we consider the ordered data set, then:

$$sp = x_{(n)} - x_{(1)}$$

The next two are mean squared error (MSE), sample variance, and standard deviation, which are defined as:

$$MSE = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$s = +\sqrt{s^2}$$

All measures compute an average quadratic distance from the mean (or a square root of it). MSE and sample variance differ only in the factor $1/n$ or $1/(n-1)$, respectively, which has to do with some theoretical properties. For large n, the differences become negligible and both numbers almost coincide. Note that all three figures are computed using all variable values and thus are, like the mean, not robust with respect to outliers. Moreover, as the standard deviation is the square root of the variance (which is a squared distance), it is measured in the same unit as the original variable (and not in squared units). All three measures are scale dependent. In other words, a variable with higher values in general should exhibit a higher dispersion, all else being equal. To control for this, one can use a standardization and compute the coefficient of variation (CV), which is defined as:

$$v = \frac{s}{\bar{x}}$$

These measures can be computed using PROC MEANS with the relevant statistic options. Additionally, we compute the distance between the quartiles. (See Exhibit 3.5)

Exhibit 3.5 Dispersion Measures

The MEANS Procedure								
Variable	***N***	**Minimum**	**Maximum**	**Range**	**Quartile**	**Variance**	**Std Dev**	**Coef**
					Range			**Varia**
default_time	622489	0.0000	1.0000	1.0000	0.0000	0.0238	0.1541	632.9
FICO_orig_time	622489	400.0000	840.0000	440.0000	103.0000	5144.4122	71.7246	10.64
LTV_orig_time	622489	50.1000	218.5000	168.4000	5.0000	102.5572	10.1271	12.82

```
PROC MEANS DATA=data.mortgage
N MIN MAX RANGE QRANGE VAR STD CV MAXDEC=4;
VAR default_time FICO_orig_time LTV_orig_time;
RUN;
```

Skewness and Kurtosis Measures

Next we compute measures for the shape of the distribution: skewness and kurtosis. First, we define the standardized value for each variable as:

$$z_i = \frac{x_i - \bar{x}}{s}$$

Then the skewness is calculated in SAS as:

$$skew = \frac{1}{n}\sum_{i=1}^{n} z_i^3$$

which is the average of the deviations from the mean to the power of three. Similarly, the kurtosis is computed as:

$$kurt = \frac{1}{n}\sum_{i=1}^{n} z_i^4 - 3$$

Note that mean and variance of the standardized values are 0 and 1, respectively. A negative value for the skewness shows that the distribution is skewed to the left; a positive value shows there is a skew to the right. Kurtosis measures the peakedness of the distribution. When you subtract the value of 3, as is the case in SAS, it is sometimes called excess kurtosis since the value is contrasted with the value of 3 for the normal distribution. Thus, a negative value signals a lower kurtosis and a positive value signals a higher kurtosis than the normal distribution. Both statistics can be computed in PROC MEANS using the respective commands. While the distribution for default is obviously strongly skewed to the right, FICO and LTV are left skewed, where FICO has a lower kurtosis than the normal distribution, and default and LTV a (strongly) higher kurtosis. (See Exhibit 3.6.)

Exhibit 3.6 Skewness and Kurtosis Measures

The MEANS Procedure			
Variable	***N***	**Skewness**	**Kurtosis**
default_time	622489	6.1719	36.0920
FICO_orig_time	622489	−0.3213	−0.4684
LTV_orig_time	622489	−0.1964	1.4364

```
PROC MEANS DATA=data.mortgage
N SKEW KURT MAXDEC=4;
VAR default_time FICO_orig_time LTV_orig_time;
RUN;
```

TWO-DIMENSIONAL ANALYSIS

Having explored the empirical data on a one-dimensional basis for each variable, we are usually also interested in interrelations between variables; for example, if and how variables have a tendency to comove together. Thus, variables can be analyzed jointly and the joint empirical distribution can be examined. Moreover, summarizing measures for dependencies and comovements can be computed.

Joint Empirical Distributions

The joint empirical distribution simultaneously computes the frequency distribution of two or more variables. Assume we have a sample size n and we have values $x_1, \ldots, x_n$ of attribute X (e.g., default) and values $y_1, \ldots, y_n$ of attribute Y (FICO) for each loan. Let

$(x_1, y_1), \ldots, (x_n, y_n)$	raw data
n	sample size
$a_1, \ldots, a_k$	attribute values / categories of variable X
$b_1, \ldots, b_l$	attribute values / categories of variable Y
$h_{ij} = h(a_i, b_j)$	joint absolute frequencies
$f_{ij} = f(a_i, b_j) = \frac{hij}{n}$	joint relative frequencies
$h_{i\cdot} = \sum_{j=1}^{l} h_{ij}$	absolute marginal frequencies of X
$f_{i\cdot} = \frac{h_{i\cdot}}{n} = f_1(a_i)$	relative marginal frequencies of X
$f_{\cdot j} = \frac{h_{\cdot j}}{n} = f_2(b_j)$	relative marginal frequencies of Y

Then the joint absolute (analogously relative) frequencies can be summarized in a two-way frequency table such that:

	b_1	...	b_l	
a_1	h_{11}	⋯	h_{1l}	$h_{1\cdot}$
⋮	⋮	⋱	⋮	⋮
a_k	h_{k1}	⋯	h_{kl}	$h_{k\cdot}$
	$h_{\cdot 1}$	⋯	$h_{\cdot l}$	n

where the marginal frequencies of X in each row are the sums across the respective column, and the marginal frequencies of Y in each column are the sums across the respective row, and these are equal to the one-way (stand-alone) frequencies from the earlier subsection. As earlier, this table makes sense only if the variables are ordinal, categorical, or divided into groups if they are metric. In the following example, we compute the two-way frequency table for default versus FICO where FICO is divided into five groups, which is performed by PROC RANK. Then PROC FREQ is used to compute the cross-tabulation of default versus FICO. The output table (Exhibit 3.7) shows the absolute and relative frequencies.

The FREQ Procedure

Table of default_time by FICO_orig_time						
default_time	FICO_orig_time(Values of FICO_orig_time Were Replaced by Ranks)					
	0	1	2	3	4	Total
0	121213 19.47 19.96 96.51	119701 19.23 19.71 96.91	121069 19.45 19.93 97.44	121876 19.58 20.07 98.08	123472 19.84 20.33 98.89	607331 97.56
1	4381 0.70 28.90 3.49	3820 0.61 25.20 3.09	3186 0.51 21.02 2.56	2385 0.38 15.73 1.92	1386 0.22 9.14 1.11	15158 2.44
Total	125594 20.18	123521 19.84	124255 19.96	124261 19.96	124858 20.06	622489 100.00

Exhibit 3.7 Two-Dimensional Contingency Table

For example, 121,213 loans (or 19.47 percent of a total of $n = 622{,}489$ observations) had default status 0 and were in the lowest group (group 0) of the FICO scores. Altogether 607,331 loans were not in default, which is 97.56 percent of all loans (as stated earlier), and of those 607,331 nondefault loans a percentage of 19.96 percent were in FICO group 0. Altogether, 125,594 loans were in FICO group 0, and 121,213 of these (or 96.51 percent) did not default whereas 3.49 percent of all loans in FICO group 0 defaulted. An important result we can infer from the table is that the proportion of loans that defaulted decreases the higher the FICO group becomes (from 3.49 percent in group 0 to 1.11 percent in group 4). This leads to the conclusion that there should be some interrelation between FICO and default (i.e., the higher the FICO score, the lower the relative frequency of default).

```
DATA mortgage1;
```

```
SET data.mortgage;
RUN;
PROC SORT DATA = mortgage1;
BY id time;
RUN;
PROC RANK DATA = mortgage1
GROUPS=5
OUT=quint(KEEP=id time FICO_orig_time);
VAR FICO_orig_time;
RUN;
DATA new;
MERGE mortgage1 quint;
BY id time;
RUN;
PROC FREQ DATA=new  ;
TABLES default_time*FICO_ORIG_TIME;
RUN;
```

Another way of inferring the relation between both variables (without grouping FICO first) is to look at box plots, which can be requested in SAS using PROC BOXPLOT; see SAS Institute Inc. (2015). The following program code plots FICO versus default where a separate box plot for each default category (0 and 1) is computed. A box plot consists of a box and whiskers. (See Exhibit 3.8.) In SAS the standard statistics represented by the box-and-whiskers plot are given as follows.

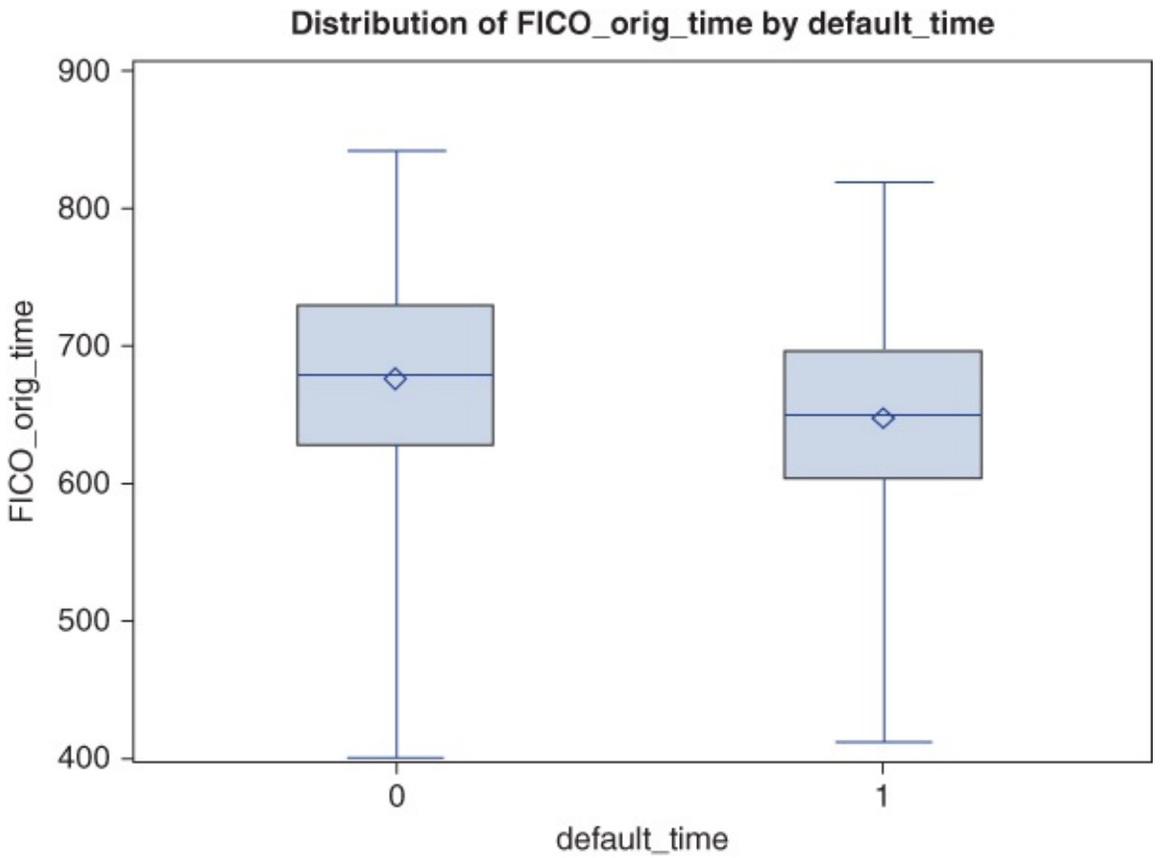


Exhibit 3.8 Box Plot of FICO Grouped by Default

Group summary statistic	*Feature of box-and-whiskers plot*
Maximum	End point of upper whisker
Third quartile (75th percentile)	Upper edge of box
Median (50th percentile)	Line inside box
Mean	Symbol marker
First quartile (25th percentile)	Lower edge of box
Minimum	End point of lower whisker

The following SAS codes then plot the two boxes and whiskers for FICO and LTV, for each category of default separately. As can be seen in Exhibit 3.9, the location of the box is higher for FICO and lower for LTV in the nondefault category compared to the default category, which again shows some interrelation between the variables in the sense that higher FICO scores correspond to a lower default frequency and higher LTVs correspond to a higher default frequency.

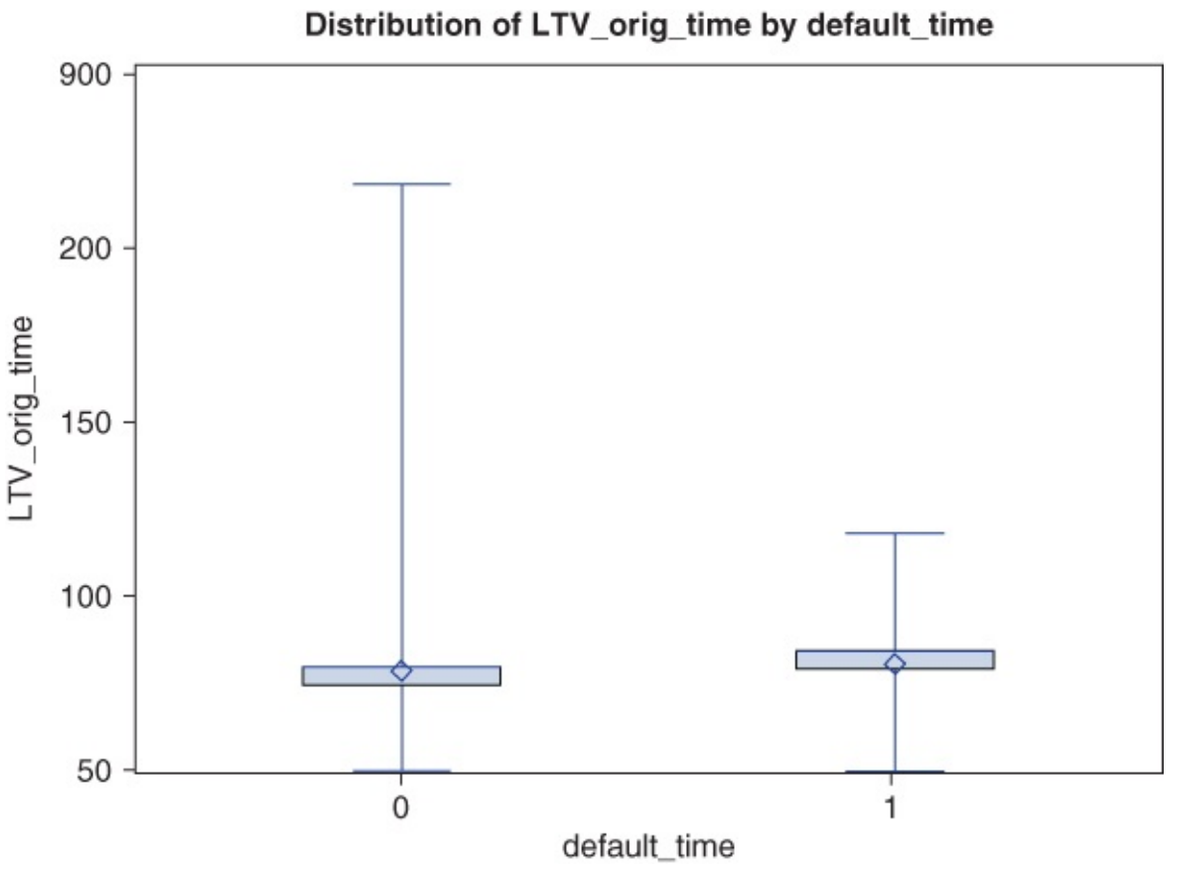


Exhibit 3.9 Box Plot of LTV Grouped by Default

```
PROC SORT DATA= mortgage1;
BY default_time;
RUN;
ODS GRAPHICS ON;
PROC BOXPLOT DATA = mortgage1;
PLOT FICO_orig_time*default_time /IDSYMBOL=CIRCLE
```

```
IDHEIGHT=2 CBOXES=BLACK BOXWIDTH=10 ;
RUN;
ODS GRAPHICS OFF;
ODS GRAPHICS ON;
PROC BOXPLOT DATA = mortgage1;
PLOT LTV_orig_time*default_time / IDSYMBOL=CIRCLE
IDHEIGHT=2 CBOXES=BLACK BOXWIDTH=10 ;
RUN;
ODS GRAPHICS OFF;
```

The joint frequency tables and box plots allow us to draw conclusions about dependencies between variables from a visual perspective. To make stronger statements about dependence, one needs to compute numeric measures. To start, one first needs a reference point, which is obviously the case of independence. Statistically, two variables are empirically independent if the joint frequencies in the two-way table are given by the product of the respective marginal frequencies across the entire table; that is,

$$f_{ij} = \frac{h_{i\cdot}h_{\cdot j}}{n^2}$$
$$h_{ij} = \frac{h_{i\cdot}h_{\cdot j}}{n}$$

Having defined a reference point, we can now start to make assessments about the direction and the strength of the dependence by measuring the deviation from independence. This is done via correlation and dependence measures.

Correlation Measures

The first measure is the χ^2-coefficient, which can be applied to two-way tables and is defined as:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{\left(h_{ij} - \frac{h_{i\cdot}h_{\cdot j}}{n}\right)^2}{\frac{h_{i\cdot}h_{\cdot j}}{n}} = n\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(f_{ij} - f_{i\cdot}f_{\cdot j})^2}{f_{i\cdot}f_{\cdot j}}$$

It measures the sum across all cells in the table of the squared deviations of the observed frequencies from those that would result if the two variables were independent. Thus, in the case of perfect independence a value of $\chi^2 = 0$ would result. It has a lower reference point of zero but is unbounded from above. We can now compute related measures, such as the ϕ-*coefficient*

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

the *contingency coefficient*

$$cc = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

and *Cramer's* V

$$V = \sqrt{\frac{\chi^2/n}{m}}$$

where $m = \min(k-1, l-1)$ and k and l are the numbers of rows and columns of the table. It holds that $V \in [0, 1]$. The measures (plus two other measures that are not discussed here) can be computed using PROC FREQ with the command CHISQ as shown in the following code. All measures show dependence, and a statistical test conducted for χ^2 signals that the value is statistically different from zero (independence). In other words, this gives evidence of dependence between FICO and default. Note that due to the large number of observations n and the minimum of row and column numbers of two, ϕ, cc, and V coincide. (See Exhibit 3.10.)

The FREQ Procedure

Statistics for Table of default_time by FICO_orig_time

Statistic	DF	Value	Prob
Chi-Square	4	1881.6127	<.0001
Likelihood Ratio Chi-Square	4	2037.4607	<.0001
Mantel-Haenszel Chi-Square	1	1848.1334	<.0001
Phi Coefficient		0.0550	
Contingency Coefficient		0.0549	
Cramer's V		0.0550	

Exhibit 3.10 Chi-Square-Related Measures of Association

```
PROC FREQ DATA=new;
TABLES default_time*FICO_ORIG_TIME / CHISQ;
RUN;
```

The χ^2 and related measures can be used for categorical ordered, or grouped metric variables. If we group metric variables, however, this means a partial loss of information in the data. If we want to compute dependency measures for metric variables, we can take account of the full information in the data by using every raw variable value. This can be done by the first metric measure of association, the sample covariance defined as:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

The *sample covariance* is (as is the variance) scale dependent. A standardized measure is the sample correlation defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum x_i^2 - n\bar{x}^2\right)\left(\sum y_i^2 - n\bar{y}^2\right)}} = \frac{s_{xy}}{s_x \cdot s_y}$$

r_{xy} is sometimes called the *Bravais-Pearson correlation* (or *product-moment correlation*) and takes on values $-1 \leq r_{xy} \leq +1$. It is a measure for *linear* dependence, that is, $r_{xy} = \pm 1 \iff y_i = a + b \cdot x_i$. A coefficient of $r_{xy} \approx 0$ does not necessarily mean that the variables are not dependent, only that there is no linear relation.

Another measure for ordered (ranked) data is the *Spearman rank correlation*. Let $rk(x_i)$ be the rank of x_i in the ordered list of X, and $rk(y_i)$ be the rank of y_i in the ordered list of Y. Then we can compute the correlation by applying the Bravais-Pearson correlation to the ranked values $(rg(x_i), rg(y_i))$:

$$r_{sp} = \frac{\sum_{i=1}^{n}(rk(x_i) - \overline{rk}(x))(rk(y_i) - \overline{rk}(y))}{\sqrt{\sum_{i=1}^{n}(rk(x_i) - \overline{rk}(x))^2 \cdot \sum_{i=1}^{n}(rk(y_i) - \overline{rk}(y))^2}}$$

which yields:

$$r_{sp} = 1 - \frac{6 \cdot \sum_{i=1}^{n}(rk(x_i) - rk(y_i))^2}{n \cdot (n^2 - 1)}$$

Another frequently used measure of nonlinear dependence is *Kendall's* τ_b. It is based on the number of concordances and discordances in paired observations. Concordance occurs when paired observations vary together, and discordance occurs when paired observations vary differently. It is defined as:

$$\tau_b = \frac{\sum_{i<j}(sign(x_i - x_j)(sign(y_i - y_j))}{\sqrt{(T_0 - T_1)(T_0 - T_2)}}$$

where $T_0 = n(n-1)/2$, $T_1 = \sum_k t_k(t_k - 1)/2$, and $T_2 = \sum_l u_l(u_l - 1)/2$. t_k is the number of tied x-values in the k th group of tied x-values and u_l is the number of tied y-values in the l th group of tied y-values, and $sign(z)$ is defined as:

$$sign(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z = 0 \\ -1 & \text{if } z < 0 \end{cases}$$

The dependence can also be represented graphically using a scatter plot where each symbol in the plot gives the two-dimensional location of an observation. As the total data set consists of more than 600,000 observations, a scatter plot would be very time consuming. We therefore draw a random sample of 1 percent of observations, and compute the correlation measures

(Exhibit 3.11) and show the scatter plot (Exhibit 3.12) using PROC CORR for FICO and LTV. An observation is included if a uniform variable (equal distribution between zero and one) generated by RANUNI with the seed 123456 is less than 1 percent. Alternatively, PROC SURVEYSELECT may be used for random sampling.

The CORR Procedure

Pearson Correlation Coefficients, *N* = 6073 Prob > |r| under H0: Rho=0				
	FICO_orig_time	LTV_orig_time	PFICO_orig_time	PLTV_orig_time
FICO_orig_time	1.00000	−0.14063		<.0001
LTV_orig_time	−0.14063	1.00000	<.0001	
Spearman Correlation Coefficients, *N* = 6073 Prob > |r| under H0: Rho=0				
	FICO_orig_time	LTV_orig_time	PFICO_orig_time	PLTV_orig_time
FICO_orig_time	1.00000	−0.17170		<.0001
LTV_orig_time	−0.17170	1.00000	<.0001	
Kendall Tau b Correlation Coefficients, *N* = 6073 Prob > |tau| under H0: Tau=0				
	FICO_orig_time	LTV_orig_time	PFICO_orig_time	PLTV_orig_time
FICO_orig_time	1.00000	−0.12363		<.0001
LTV_orig_time	−0.12363	1.00000	<.0001	

Exhibit 3.11 Correlation Measures

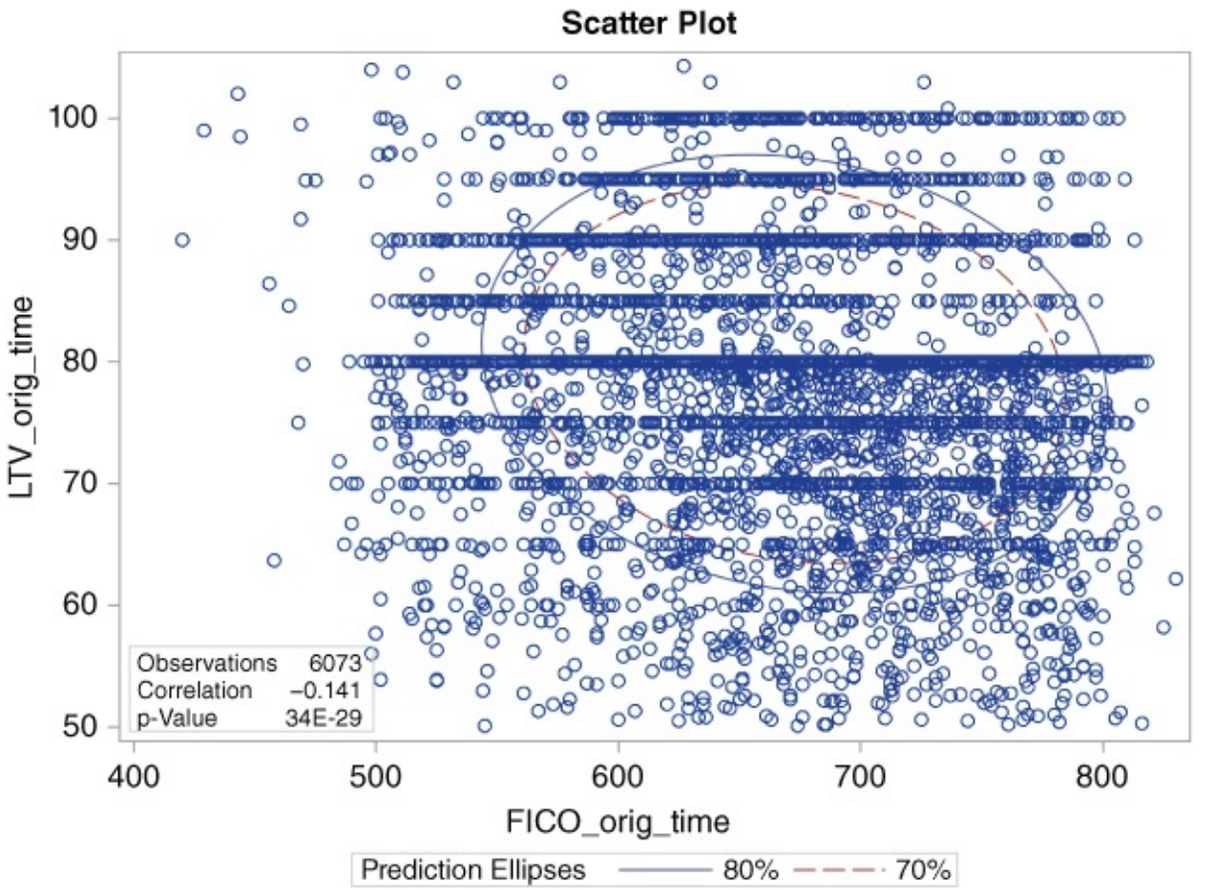


Exhibit 3.12 Scatter Plot of FICO versus LTV (Sample)

The options Kendall and Spearman request these measures to be computed. As can be seen, all three measures return a negative correlation, which means that a higher FICO score corresponds to a lower LTV. The size of the measures (around -0.2) and the scatter plot, however, show that the strength of this relation is not very strong, revealing that both variables provide self-contained information.

```
DATA sample;
SET data.mortgage;
IF RANUNI(123456) < 0.01;
RUN;
ODS GRAPHICS ON;
PROC CORR DATA=sample
PLOTS(MAXPOINTS=NONE)=SCATTER(NVAR=2 ALPHA=.20 .30)
KENDALL SPEARMAN;
VAR FICO_orig_time LTV_orig_time;
RUN;
ODS GRAPHICS OFF;
```

HIGHLIGHTS OF INDUCTIVE STATISTICS

Sampling

The former two sections provided graphical and mathematical measures for exploring data. These data can be from an entire population or from a sample of the population only. Actually, most data at hand are only sample data that are drawn from a larger population. Even the huge data set of mortgage loans can be interpreted as a sample from the entire population of all mortgages in the United States or the whole world. There are various types of sampling designs, which can be roughly divided into:

- *Probability sampling*: Here each unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined.
- *Nonprobability sampling*: Some units in the population have no chance of being selected or the probability cannot be accurately determined. This gives an exclusion bias. Examples are voluntary sampling or convenience sampling where people are selected either by their own or by the surveyor's will to participate.

In most applications one assumes probability sampling. This can be further divided into:

- *Simple random sampling*: It is assumed that the population consists of N objects. The sample consists of n objects, and all possible samples of n objects are equally likely to occur. An example is the lottery method. If the probability of each object is not equal, it is simply named “random sampling.”
- *Stratified sampling*: The population is divided into groups (also called strata), based on some characteristic (e.g., sex or age). Then, within each group, a probability sample (often a simple random sample) is selected.
- *Cluster sampling*: Every member of the population is assigned to one, and only one, group

(called a cluster), for example a municipality. Then a sample of clusters is chosen, using a probability method (often simple random sampling). Only individuals within sampled clusters are surveyed.

- Others, which we do not explain explicitly (e.g., multistage sampling, systematic random sampling).

Let us have a closer look at random sampling. We denote by Ω the population of individuals or objects $(\omega_1,...,\omega_N)$ with attribute X (e.g., age). Next, consider n random draws from that population, so that:

$$\begin{aligned}
&\text{1st draw delivers:} && \omega_1 \xrightarrow[\text{measurement}]{} X_1(\omega_1) = x_1 \\
&\text{2nd draw delivers:} && \omega_2 \xrightarrow[\text{measurement}]{} X_2(\omega_2) = x_2 \\
&\quad\vdots && \quad\vdots \\
&n\text{th draw delivers:} && \omega_n \xrightarrow[\text{measurement}]{} X_n(\omega_n) = x_n
\end{aligned}$$

That is, the objects are randomly selected, and their attributes are measured (e.g., age of the drawn person). The data $x_1, \ldots, x_n$ are realizations of the random variables $X_1, \ldots, X_n$ and represent the same property X. Thus, $X_1, \ldots, X_n$ are *identically* distributed as X.

Hence, if objects and their variables (age, height etc.) are obtained via random sampling, their values can be interpreted as outcomes and realizations of the same random variables. If the random draws are also independent, then $X_1, \ldots, X_n$ are *independent and identically distributed (i.i.d.)*.

Point Estimation

Now, if one has sample data at hand, the goal is often to not only describe and explore the data, but also make inferences about distributions, parameters thereof, or relations between variables, such as correlations, in the underlying population using the sample data. For this, the sampling mechanism has to be taken into account.

The first goal is to infer one or more parameters of the distribution of a variable in the population using the sample. This is called *parameter* or *point estimation*. An example is computing the mean of the sample and taking it as an estimate for the mean (the expectation) in the population. There are various statistical techniques for parameter estimation, the most important of which are:

- *Ordinary least squares (OLS) method*: The parameter(s) are computed such that the sum of the squared differences of the observations from the predictions is minimized.
- *Method of moments (MM)*: The parameters of the population are substituted by their counterparts in the sample.
- *Maximum-likelihood (ML) method*: The parameters are computed such that the likelihood (probability) of observing the sample at hand is maximal among the potential set of parameters.

Most of the methods in this book use the ML method. The method is described in more detail for models in the respective chapters (e.g., for PD models in the chapters on PDs).

Confidence Intervals

Once parameters are estimated, the sampling design comes into play. If we have a random sample, then the number of observations is typically much smaller than the size of the population. Therefore, we will not match the true (and unknown) parameter in the population exactly but instead have a random deviation from that parameter.

A popular way of constructing probability bounds around the parameter estimate is by computing confidence intervals. If one chooses an approach t delivers with probability, say, $1-\alpha$, an interval that contains the true parameter, then this interval is called a *confidence interval*. Formally, let θ be the (unknown) parameter in the population of interest and let $\hat{\theta}$ be its estimate from a random sample. Then we construct lower and upper bounds B_l and B_u such that:

$$P(B_l \leq \theta \leq B_u) = 1 - \alpha$$

where α is called the error probability and $1-\alpha$ is the confidence level.

Many models considered in this book result in approximately normally distributed parameter estimators. In particular, it can be shown that the ML estimation methods approximately yield normally distributed estimators that are furthermore unbiased (i.e., their expectation is the true parameter). Then, confidence intervals can be easily constructed. Let the estimator $\hat{\theta}$ be normally distributed with expectation θ and variance σ_θ^2; that is, $\hat{\theta} \sim N(\theta, \sigma_\theta^2)$. Then $\frac{\hat{\theta}-\theta}{\sigma_\theta}$ is standard normally distributed, and due to the properties of the standard normal distribution we obtain $P(-z_{1-\alpha/2} \leq \frac{\hat{\theta}-\theta}{\sigma_\theta} \leq z_{1-\alpha/2}) = 1-\alpha$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of the standard normal distribution. Rearranging yields:

$$P(\hat{\theta} - z_{1-\alpha/2} \cdot \sigma_\theta \leq \theta \leq \hat{\theta} z_{1-\alpha/2} \cdot \sigma_\theta) = 1 - \alpha$$

This gives the confidence interval

$$[\hat{\theta} - z_{1-\alpha/2} \cdot \sigma_\theta; \quad \hat{\theta} + z_{1-\alpha/2} \cdot \sigma_\theta]$$

Usually, however, the standard deviation σ_θ of the estimator is also unknown and has to be estimated from the sample. Let $\hat{\sigma}_\theta$ be the estimate (e.g., the sample standard error). Then the confidence interval becomes:

$$[\hat{\theta} - t_{n-1,1-\alpha/2} \cdot \hat{\sigma}_\theta; \quad \hat{\theta} + t_{n-1,1-\alpha/2} \cdot \hat{\sigma}_\theta]$$

where $t_{n-1,1-\alpha/2}$ is the $1-\alpha/2$ percentile of Student's t-distribution with $n-1$ degrees of freedom and n is the sample size. For large sample sizes, this converges toward the standard normal distribution.

In SAS, confidence intervals are typically automatically reported in the standard output when model parameters are estimated (e.g., for the PD models from the chapters on PDs). For sample means they can also be computed using PROC UNIVARIATE with option BASICINTERVALS, as shown in the following code for LTV; see SAS Institute Inc. (2015). The option cibasic(alpha=.01) produces a 99 percent confidence interval. (See Exhibit 3.13.)

The UNIVARIATE Procedure
Variable: LTV_orig_time

Basic Confidence Limits Assuming Normality			
Parameter	Estimate	99% Confidence Limits	
Mean	78.97546	78.94240	79.00852
Std Deviation	10.12705	10.10372	.
Variance	102.55718	102.08523	.

Exhibit 3.13 Basic Confidence Intervals

```
ODS SELECT BASICINTERVALS;
PROC UNIVARIATE DATA=data.mortgage CIBASIC(ALPHA=.01);
VAR LTV_orig_time;
RUN;
```

Due to the large sample size, the standard error of the mean (not the standard error of the total observations) is very low. Therefore, the confidence interval for the mean is very narrow. Moreover, note that SAS also produces intervals for the standard deviation and the variance (where it failed for numerical reasons to provide the upper bound). These intervals are not constructed using normal distributions. Because these are not often used throughout the book, we will not go into further detail here.

Hypothesis Testing

Another way of inferring from the sample to the population is hypothesis testing on one or more parameters in the population.

Step 1: Hypothesis Formulation

In a first step, one has to formulate a null hypothesis for the population, which will be tested using the sample data. For example, a hypothesis could be about a specific value for a parameter in the population and could be formulated as H_0: "The population parameter θ is exactly equal to a value θ_0", or in short $H_0 : \theta = \theta_0$. The null hypothesis also has an alternative, say H_1, which is valid if H_0 is not true. Here the alternative could be H_1: "The population parameter θ is not equal to a value θ_0," or in short $H_1 : \theta \neq \theta_0$. This is called a two-sided hypothesis.

There are also one-sided hypotheses, namely $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, and $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$.

The hypotheses are usually derived from economic theories or intuition. For example, a null hypothesis could be that the mean LTV in the population is 60 whereas the alternative could be that it is different from 60.

Step 2: Choice of Test Statistic

In a second step, we determine an appropriate measure for testing the hypothesis, the so-called test statistic. If we want to test a hypothesis for the population mean, an obvious test statistic is the sample mean.

Step 3: Distribution of the Test Statistic

As a third step, we determine the probability distribution of the test statistic under the assumption that H_0 is true. As many model parameters are estimated via maximum likelihood, the estimators that serve as test statistics are approximately normally distributed. For example, if the null hypothesis were about a parameter $H_0 : \theta = \theta_0$ and the estimator $\hat{\theta}$ is normally distributed with variance σ_θ^2, then under the null hypothesis the distribution of $\hat{\theta}$ is $N(\theta_0, \sigma_\theta^2)$.

Step 4: Computing the p-Value

Next, given the distribution under the null hypothesis, and given the sample estimate $\hat{\theta}$, one computes the probability of observing exactly the value $\hat{\theta}$ or greater in the sample data when the null is true. In other words, under the normal distribution assumption one computes for our null hypothesis,

$$p - value = 2 \cdot \left[1 - \Phi \left(\frac{|\hat{\theta} - \theta_0|}{\sigma_\theta} \right) \right]$$

which is the so-called p-value. The term in brackets on the right-hand side of the equation is multiplied by 2 and the numerator is computed as the absolute deviation because we have a two-sided test here. For a one-sided test, the multiplicator and the absolute operator would be dropped. The p-value gives the probability of sampling the observed sample value (or a greater value) assuming the null hypothesis is true. The lower this probability, the more p-value evidence against the null hypothesis (and in favor of the alternative). For small values ($0.05 \leq p$-value < 0.1) we say that the result is weakly significant. For even smaller values ($0.01 \leq p$-value < 0.05) the result is significant, and for very small values (p-value < 0.01) the result is said to be strongly significant against the value under the null hypothesis. The borderline values are sometimes also called significance levels. This two-sided test can also be conducted as one-sided tests, similar to confidence intervals. Moreover, if the standard deviation has to be estimated in addition, the CDF of the standard normal distribution is replaced by the percentile of the Student's t-distribution as for the confidence intervals.

Step 5: Decision

Finally, we have to decide whether to support or reject the null hypothesis as a consequence of the test result (the p-value). This decision has consequences for the error that often occurs. Remember that the test decision is made from sample data only, which are random. That is, even if the data may provide evidence against the null hypothesis, another sample may yield the opposite result. Generally, given a null hypothesis H_0 and an alternative hypothesis H_1, we can differentiate between the true but unknown states of the world (either H_0 or H_1 is true), and the decision based on the statistical test (either rejection of H_0 or no rejection of H_0). Thus, four scenarios can arise:

1. H_0 is true but the test decision is *not* to reject H_0 → This is a correct decision.
2. H_0 is true but the test decision is to (erroneously) reject H_0 → This is a wrong decision.
3. H_0 is not true (and H_1 is true instead) and the test decision is to reject H_0 → This is a correct decision.
4. H_0 is not true (and H_1 is true instead) but the test decision is *not* to reject H_0 → This is a wrong decision.

Situations 1 and 3 are not problematic, but situations 2 and 4 might be, as a wrong decision occurs. We won't go into detail here but will follow up on this in the chapter on model validation.

Similar to confidence intervals, most standard procedures automatically compute p-values when a model is estimated. In PROC UNIVARIATE the p-values can be computed via the following code *p*-Value using the option TESTFORLOCATION. The option MU0=60 specifies the value under the null hypothesis.

```
ODS GRAPHICS ON;
ODS SELECT TESTSFORLOCATION ;
PROC UNIVARIATE DATA=data.mortgage MU0=60;
VAR LTV_orig_time;
RUN;
ODS GRAPHICS OFF;
```

PROC UNIVARIATE provides three different tests where only the first is of interest here. As can be seen in Exhibit 3.14, the p-value is lower than 0.0001 and therefore the mean LTV is significantly different from 60 and we should reject the null hypothesis that the LTV is 60 in the population.

The UNIVARIATE Procedure
Variable: LTV_orig_time

Tests for Location: Mu0=60				
Test	Statistic		p Value	
Student's *t*	*t*	1478.343	Pr > |*t*|	<.0001
Sign	M	277123	Pr >= |M|	<.0001
Signed Rank	S	9.452E10	Pr >= |S|	<.0001

Exhibit 3.14 Test for Location

There are huge numbers of different tests for various hypotheses (e.g., for medians, for standard deviations, for entire distributions) and we will not try to cover this in the introductory section. Most of the models estimated in this book and most of the SAS output will work with the standard tests as shown in this chapter.

REFERENCE

SAS Institute Inc. 2015. *SAS/STAT 14.1 User's Guide: Technical Report*. Cary, NC: SAS Institute.