Statistic Analysis in SAS

Course material

Basic and advanced programming and statistics in SAS

dr Karol Przanowski

Structured data analysis

Structured data analysis

- Measure of location
- Measure of statistical dispersion
- Measure of asymmetry
- Measure of shape Miary koncentracji

Descriptive statistics

Miary struktury	Classical	Positional
Scope	Regular distribution	Irregular distribution
	(low variation,	(high variation,
	low asymmetry,	high assymetry,
	low kurtosis)	high kurtosis
Mean	arytmetical mean	mode
		quintile
		(MEDIAN|P50, Q1,Q3,P1,P5,P10,P90,P95,P99)
Variation	variance (VAR)	range=max-min
		RANGE
	standard deviation StdDev	interquartile range
		QRANGE=Q3 - Q1
	Coefficient of variation=StdDev/mean	mid-quartile range Q= QRANGE/2
	CV (coefficient of variation)	
	typical variation	positional coefficient of variation=Q/Mediana
	range (mean–StdDev : mean+StdDev)	
Asymmetry	measure of asymmetry	positional measure of asymmetry
	SKEWNESS|SKEW	
concentration	kurtosis KURTOSIS|KURT	

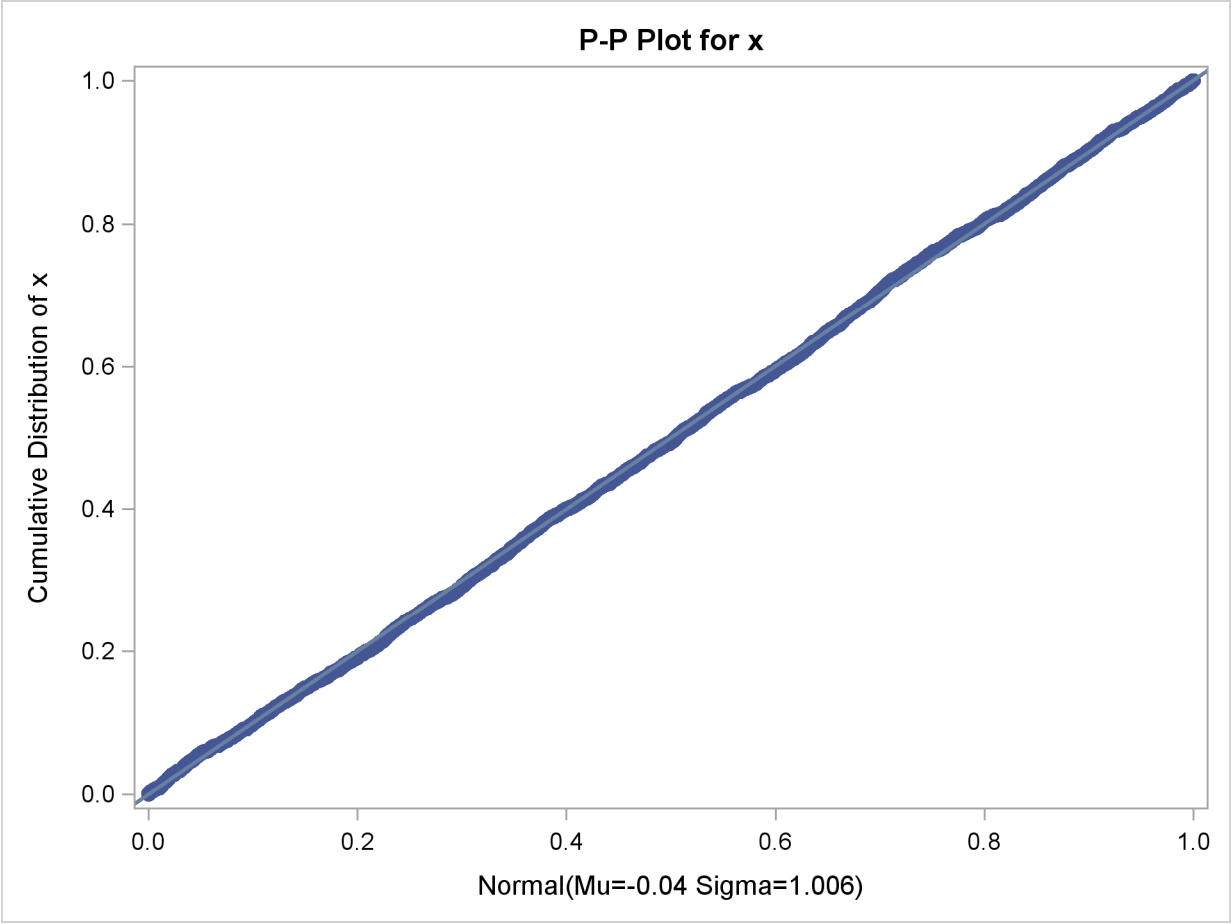
Tests to compare empirical and theoretical distribution

Goodness-of-Fit Tests for Lognormal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.06441431	**Pr > D**	>0.150
Cramer-von Mises	**W-Sq**	0.02823022	**Pr > W-Sq**	>0.500
Anderson-Darling	**A-Sq**	0.24308402	**Pr > A-Sq**	>0.500

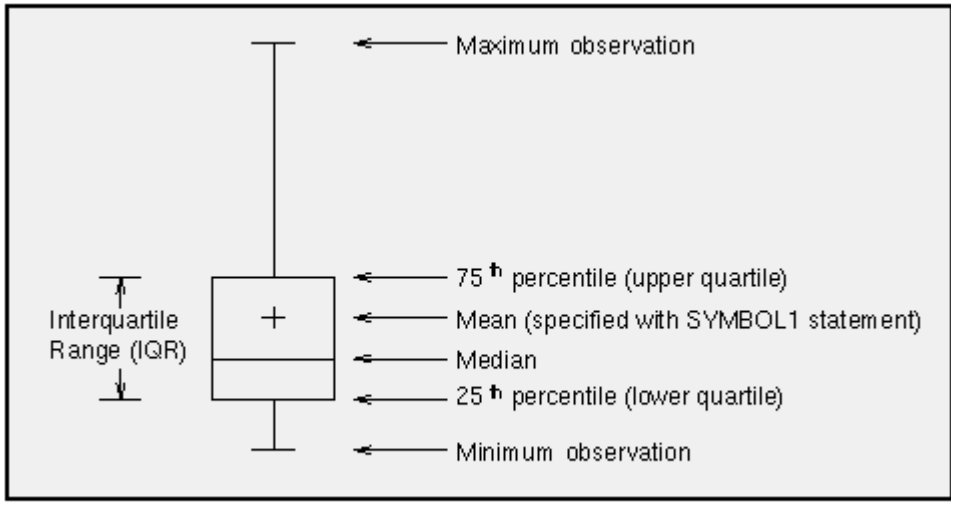
$$D = \sup_x |F_n(x) - F(x)|$$

Probability plot (pp-plot, qq-plot)

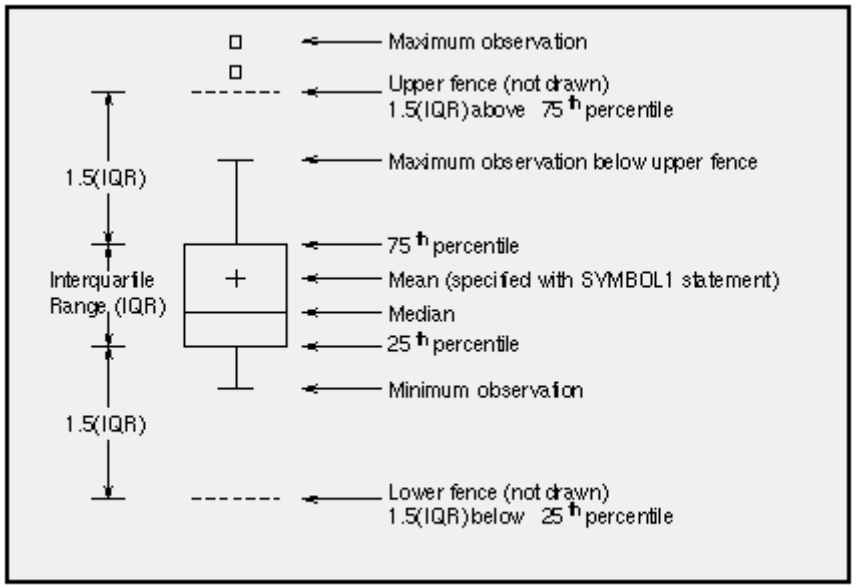
- Plot:
- $(X, F_n^{-1}(F_e(X)))$



Box Plot



Skeletal Box-and-Whisker Plot



BOXSTYLE= SCHEMATIC

Robust estimation of location

- Percentiles:
  - IQR vs. Range
  - Median (P50) vs. Mean

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

- Winsored mean

$$\bar{x}_{wk} = \frac{1}{n}\left((k+1)x_{(k+1)} + \sum_{i=k+2}^{n-k-1} x_{(i)} + (k+1)x_{(n-k)}\right)$$

- Trimmed mean

$$\bar{x}_{tk} = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} x_{(i)}$$

Statistical inference

Statistical hypothesis

Statistical hypothesis is any assumption regarding to population distribution (statistic or parameter). Hypothesis is verified on random sample.

Statistical Test is set of rules and methods used to test statistical hypothesis. Statistical test is a rule to make decision whether hypothesis is true or false based on results calculated on random sample of the population.

Alpha i Beta

- Alpha – significance level
- 1-Beta – test's power.

- Verification :
  - H0 i H1
  - Statistical test and statistic value
  - significance level or p-value
  - Decision

Hypothesis testing

W – all possible results for n-elements sample
W_n – sample from W [$W_n = (x_1, x_2, \ldots, x_n)$]
w - area for obszar przestrzeni próby

Type I Error – incorrect rejection true null hypothesis

$$P(W_n \in w \mid H_0) = \alpha(w)$$

Type II Error – failure to reject false null hypothesis

$$P(W_n \in (W - w) \mid H_1) = \beta(w)$$

The grater α means higher probability to reject H_0
In practice α value is *set between <0,01:0,1>*

Decision

- Reject H0
- There are no enough evidences to reject H0 – by default it doesn’t mean that H1 true.

Test for location of a mean

- We want to verify “real” mean in a population based on statistic calculated on a random sample of the population.

$$\frac{\overline{x} - \mu_0}{s / \sqrt{\sum w_i}}$$

One and two-tailed tests

$$Pr(X \geq x|H)$$
$$Pr(X \leq x|H)$$
$$2\min\{Pr(X \leq x|H), Pr(X \geq x|H)\}$$

Association between variables

PROC CORR

```
proc corr data=…
pearson/spearman/kendall/hoeffding
vardef=df;
var …;
with …;
run;
```

Pearson's correlation coefficient

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$r_{xy} = \frac{\mathrm{COV}_{xy}}{s_x s_y}$$

Spearman's rank correlation coefficient

$$r_d = \frac{\sum \left(R_i - \overline{R}\right)\left(S_i - \overline{S}\right)}{\sqrt{\sum \left(R_i - \overline{R}\right)^2 \sum \left(S_i - \overline{S}\right)^2}}$$

R_i – rank x_i
S_j – rank y_j

$$r_d = 1 - \frac{6\sum_{i=1}^{n} \left(R_i - S_i\right)^2}{n\left(n^2 - 1\right)}$$

Regression analysis

$$Y_i = \beta + \sum_i \alpha_i X_i + \varepsilon_i$$

```
proc reg data=…;
model dependent_variable= independent_variable(s);
run;
```

ANOVA in Regression analysis

Coefficient of determination R^2

Źródło zmienności	SS – Sum of squares	Df – degree of freedom	Mean SS
1. Variability explained by model	SSB	*r* - 1	MSB
2. Variability unexplained by model	SSE	*n* - *r*	MSE
3. Total variability	SST	*n* - 1	_

$$R^2 = \frac{SSB}{SST}$$

Parameter estimation

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\mathrm{COV}_{xy}}{s_x^2}$$

$$\hat{\beta} = \bar{y} - \hat{\alpha}\bar{x}$$

Hypothesis testing

$$H_0 : \alpha = \alpha_0$$

$$H_1 : \alpha \neq \alpha_0$$

$$t = \frac{\hat{\alpha} - \alpha_0}{s_{\hat{\alpha}}} \xrightarrow{\text{standard}} t = \frac{\hat{\alpha}}{s_{\hat{\alpha}}}$$

$$P\left(|t| \geq t_{\alpha,n-2}\right) = \alpha$$

Prediction

Standard estimation error

$$s(\hat{y}_x) = s_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Standard prediction error

$$s(\hat{y}_x^P) = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Least squares estimation assumptions

1. Model is linear against parameters
2. Dependent variables are not random.
3. Every residual εi expected value is equal 0.
4. Every residual εi variance is constant.
5. All residual pairs εi and εj are uncorrelated.
6. For multiple regression all independent variables are uncorrelated

Association between qualitative variables

- Test Chi2, is based on independence event probability $p_{ij}=p_{i.}*p_{.j}$

$$Q_P = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

$$e_{ij} = \frac{n_{i.}\ n_{.j}}{n}$$

$$n_{i.} = \sum_j n_{ij} \quad \text{(row totals)}$$

$$n_{.j} = \sum_i n_{ij} \quad \text{(column totals)}$$

$$n = \sum_i \sum_j n_{ij} \quad \text{(overall total)}$$

$$p_{ij} = n_{ij}/n \quad \text{(cell percentages)}$$

$$p_{i.} = n_{i.}/n \quad \text{(row percentages of total)}$$

$$p_{.j} = n_{.j}/n \quad \text{(column percentages of total)}$$

- Test V-Cramera, D-Somersa

Statistical model automation

Agenda

- Variable selection
- Model fit summary statistics
- Collinearity Diagnostics
- Outliers and influence statistics
- Variable transformation
- Partial correlation
- Variable clustering
- Principal component analysis

Variable selection

- **Forward Selection (FORWARD)**
- The forward-selection technique begins with no variables in the model. For each of the independent variables, the FORWARD method calculates statistics that reflect the variable's contribution to the model if it is included. The p-values for these statistics are compared to the SLENTRY= value that is specified in the MODEL statement (or to 0.50 if the SLENTRY= option is omitted). If no statistic has a significance level greater than the SLENTRY= value, the FORWARD selection stops. Otherwise, the FORWARD method adds the variable that has the largest statistic to the model. The FORWARD method then calculates statistics again for the variables still remaining outside the model, and the evaluation process is repeated. Thus, variables are added one by one to the model until no remaining variable produces a significant statistic. Once a variable is in the model, it stays.

Variable selection

- **Backward Elimination (BACKWARD)**
- The backward elimination technique begins by calculating statistics for a model which includes all of the independent variables. Then the variables are deleted from the model one by one until all the variables remaining in the model produce statistics significant at the SLSTAY= level specified in the MODEL statement (or at the 0.10 level if the SLSTAY= option is omitted). At each step, the variable showing the smallest contribution to the model is deleted.

Variable selection

- **Stepwise (STEPWISE)**
- The stepwise method is a modification of the forward-selection technique and differs in that variables already in the model do not necessarily stay there. As in the forward-selection method, variables are added one by one to the model, and the statistic for a variable to be added must be significant at the SLENTRY= level. After a variable is added, however, the stepwise method looks at all the variables already included in the model and deletes any variable that does not produce an statistic significant at the SLSTAY= level. Only after this check is made and the necessary deletions are accomplished can another variable be added to the model. The stepwise process ends when none of the variables outside the model has an statistic significant at the SLENTRY= level and every variable in the model is significant at the SLSTAY= level, or when the variable to be added to the model is the one just deleted from it.

Model selection

- **Maximum R Improvement (MAXR)**
- The maximum improvement technique does not settle on a single model. Instead, it tries to find the "best" one-variable model, the "best" two-variable model, and so forth, although it is not guaranteed to find the model with the largest for each size.
- The MAXR method begins by finding the one-variable model producing the highest . Then another variable, the one that yields the greatest increase in , is added. Once the two-variable model is obtained, each of the variables in the model is compared to each variable not in the model. For each comparison, the MAXR method determines if removing one variable and replacing it with the other variable increases . After comparing all possible switches, the MAXR method makes the switch that produces the largest increase in . Comparisons begin again, and the process continues until the MAXR method finds that no switch could increase . Thus, the two-variable model achieved is considered the "best" two-variable model the technique can find. Another variable is then added to the model, and the comparing-and-switching process is repeated to find the "best" three-variable model, and so forth.
- The difference between the STEPWISE method and the MAXR method is that all switches are evaluated before any switch is made in the MAXR method. In the STEPWISE method, the "worst" variable might be removed without considering what adding the "best" remaining variable might accomplish. The MAXR method might require much more computer time than the STEPWISE method.

Model selection

- **R Selection (RSQUARE)**
- The RSQUARE method finds subsets of independent variables that best predict a dependent variable by linear regression in the given sample. You can specify the largest and smallest number of independent variables to appear in a subset and the number of subsets of each size to be selected. The RSQUARE method can efficiently perform all possible subset regressions and display the models in decreasing order of magnitude within each subset size. Other statistics are available for comparing subsets of different sizes. These statistics, as well as estimated regression coefficients, can be displayed or output to a SAS data set.
- The subset models selected by the RSQUARE method are optimal in terms of for the given sample, but they are not necessarily optimal for the population from which the sample is drawn or for any other sample for which you might want to make predictions. If a subset model is selected on the basis of a large value or any other criterion commonly used for model selection, then all regression statistics computed for that model under the assumption that the model is given a priori, including all statistics computed by PROC REG, are biased.
- While the RSQUARE method is a useful tool for exploratory model building, no statistical method can be relied on to identify the "true" model. Effective model building requires substantive theory to suggest relevant predictors and plausible functional forms for the model.
- The RSQUARE method differs from the other selection methods in that RSQUARE always identifies the model with the largest for each number of variables considered. The other selection methods are not guaranteed to find the model with the largest . The RSQUARE method requires much more computer time than the other selection methods, so a different selection method such as the STEPWISE method is a good choice when there are many independent variables to consider.

Model selection

- **Minimum R (MINR) Improvement**
- The MINR method closely resembles the MAXR method, but the switch chosen is the one that produces the smallest increase in . For a given number of variables in the model, the MAXR and MINR methods usually produce the same "best" model, but the MINR method considers more models of each size.
- **Adjusted R Selection (ADJRSQ)**
- This method is similar to the RSQUARE method, except that the adjusted statistic is used as the criterion for selecting models, and the method finds the models with the highest adjusted within the range of sizes.
- **Mallows' C Selection (CP)**
- This method is similar to the ADJRSQ method, except that Mallows' statistic is used as the criterion for model selection. Models are listed in ascending order of . If $C_p >$ p then model is underspecified or overspecified. When the right model is chosen, the parameter estimates are unbiased, and this is reflected C_p in near p.

Model fit statistics

MODEL Option or Statistic	Definition or Formula
n	the number of observations
p	the number of parameters including the intercept
i	1 if there is an intercept, 0 otherwise
$\hat{\sigma}^2$	the estimate of pure error variance from the SIGMA= option or from fitting the full model
SST_0	the uncorrected total sum of squares for the dependent variable
SST_1	the total sum of squares corrected for the mean for the dependent variable
SSE	the error sum of squares
MSE	$\dfrac{\mathrm{SSE}}{n-p}$
R^2	$1-\dfrac{\mathrm{SSE}}{\mathrm{SST}_i}$
ADJRSQ	$1-\dfrac{(n-i)(1-R^2)}{n-p}$

Model fit statistics

AIC	$n \ln\left(\frac{\text{SSE}}{n}\right) + 2p$
BIC	$n \ln\left(\frac{\text{SSE}}{n}\right) + 2(p+2)q - 2q^2$ where $q = \frac{n\hat{\sigma}^2}{\text{SSE}}$
CP (C_p)	$\frac{\text{SSE}}{\hat{\sigma}^2} + 2p - n$
GMSEP	$\frac{\text{MSE}(n+1)(n-2)}{n(n-p-1)} = \frac{1}{n} S_p (n+1)(n-2)$
JP (J_p)	$\frac{n+p}{n}\text{MSE}$
PC	$\frac{n+p}{n-p}(1-R^2) = J_p\left(\frac{n}{\text{SST}_i}\right)$
PRESS	the sum of squares of predr_i (see Table 74.9)
RMSE	$\sqrt{\text{MSE}}$
SBC	$n \ln\left(\frac{\text{SSE}}{n}\right) + p \ln(n)$
SP (S_p)	$\frac{\text{MSE}}{n-p-1}$

Model Diagnostic Statistics

PRED ($\widehat{\mathbf{Y}}_i$)	$\mathbf{X}_i\mathbf{b}$
RES (r_i)	$\mathbf{Y}_i - \widehat{\mathbf{Y}}_i$
H (h_i)	$\mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-}\mathbf{x}_i'$
STDP	$\sqrt{h_i\widehat{\sigma}^2}$
STDI	$\sqrt{(1+h_i)\widehat{\sigma}^2}$
STDR	$\sqrt{(1-h_i)\widehat{\sigma}^2}$
LCL	$\widehat{Y}_i - t_{\frac{\alpha}{2}}\text{STDI}$
LCLM	$\widehat{Y}_i - t_{\frac{\alpha}{2}}\text{STDP}$
UCL	$\widehat{Y}_i + t_{\frac{\alpha}{2}}\text{STDI}$
UCLM	$\widehat{Y}_i + t_{\frac{\alpha}{2}}\text{STDP}$
STUDENT	$\dfrac{r_i}{\text{STDR}_i}$
RSTUDENT	$\dfrac{r_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}}$

COOKD	$\dfrac{1}{p}\text{STUDENT}^2\dfrac{\text{STDP}^2}{\text{STDR}^2}$
COVRATIO	$\dfrac{\det(\hat{\sigma}^2_{(i)}(\mathbf{x}'_{(i)}\mathbf{x}_{(i)})^{-1}}{\det(\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1})}$
DFFITS	$\dfrac{(\widehat{\mathbf{Y}}_i - \widehat{\mathbf{Y}}_{(i)})}{(\hat{\sigma}_{(i)}\sqrt{h_i})}$
DFBETAS$_j$	$\dfrac{\mathbf{b}_j - \mathbf{b}_{(i)j}}{\hat{\sigma}_{(i)}\sqrt{(\mathbf{X}'\mathbf{X})_{jj}}}$
PRESS(predr$_i$)	$\dfrac{r_i}{1-h_i}$

Outliers and influence observation

- | Rstudent | >3
- Cookd > 4/n
- H > 2p/n
- | Dffits | > 2sqrt(p/n)
- | Dfbetas | > 2/sqrt(n)

| CovRatio-1 | > 1+3p/n

Box-Cox transformation

$$y_i^{(\lambda)} = \begin{cases} y_i^{\lambda}; & \lambda \neq 0 \\ \log y_i; & \lambda = 0 \end{cases}$$

Partial correlation

- Resitual X10 = X1…X9 X11-X20
- What is correlation with Y?

Variable Clustering

- Proc Varclus

Principal Component

- Proc Princomp
- Proc PLS

Other regression models

- SAS/STAT docummentation
- Introduction to regression procedures