Published on *WorldSupporter* (www.worldsupporter.org)

Chapter 1

- A population is the complete set of all items that interest an investigator. A parameter describes a specific characteristic of a population. A statistic describes a specific characteristic of a sample. A variable is a specific characteristic of an individual or object.
- With simple random sampling and systematic sampling each member of the population is chosen strictly by chance.
  Nonsampling errors: not the relevant population, inaccurate or dishonest answers, no response.
- Descriptive statistics to process data. Inferential statistics to use data.
- Two sorts of variables:
  Categorical variables produce responses that belong to groups or categories.
  Numerical variables: - Discrete variables have a finite number of values. - Continuous variables may take on any value within a given range.
- Qualitative data & Quantitative data:
  With qualitative data there is no measurable meaning to the difference in numbers. Levels of measurement: Nominal, ordinal.
  With quantitative data there is a measurable meaning to the difference in numbers. Levels of measurement: Interval, ratio.
- Frequency distribution is a table used to organize data. A relative frequency distribution measures the frequency compared to the total.
  Cumulative frequency distribution: histogram or ogive.
- Nominal measurement: bar chart, cross tables, pie chart, pareto diagram.
- Shape of a distribution: Symmetric if distributed about the center. Asymmetric of not symmetrically distributed on either side of the centre.
- Stem-and-leaf display: data are grouped according to their leading digits.
  Scatter plot: to search for a possible relationship between two numerical variables.

Chapter 2

- Numerical measures in response to questions concerning the location of the center of a data set: the mean, median and mode.
  Categoric data: median or mode.
  Numeric data: mean or median.
- Skewness positive if distribution skewed to the right, negative if skewed to the left.
- Percentiles and quartiles indicate the location of a value relative to the entire set of data.
  Quartiles separate large data sets into four quarters.
- The five number summary: minimum, first quartile, median, third quartile, and maximum. A box-and-whisker plot describes the shape of a distribution in terms of the five-number summary.
- Range is the difference between the largest and smallest observations.
- The population variance is the sum of the squared differences between each observation and the population mean divided by the population size. The sample variance is the sum of the squared differences between each observation and the sample mean divided by the sample size minus 1. The standard deviation is the square root of the variance. The coefficient of variation expresses the standard deviation as a percentage of the mean.
- Z-score indicates the number of standard deviations a value is from the mean:
  Chebyshev's theorem; number of observations within K standard deviations of the mean.
  According to the emperical rule; 68% of the observation are within one standard deviation, 95% within 2 standard deviations and almost all of the observations within 3.

- Covariance is a measure of the linear relationship between two variables. Correlation coefficient also gives the strength of the relationship.

Chapter 3

- A random experiment is a process leading to two or more possible outcomes, without knowing exactly which outcome will occur. The possible outcomes from the random experiment are called the basic outcomes, the set of all basic outcomes is called the sample space.
- An event is any subset of basic outcomes:
  - Intersection of events; the set of all basic outcomes in S that belong to both event A and B.
  - Mutually exclusive events have no common basic outcomes.
  - Union of events is the set of all basic outcomes in S that belong to at least one of the events A or B.
  - Events are collectively exhaustive if the union of several events covers the entire sample.
  - Complements of A is the set of basic outcomes that does not belong to A.
- Classical probability: All outcomes in a sample space are equally likely to occur.
  Number of possible orderings x objects: Permutation or combination.
- Probability postulates: probability between 0 and 1, the sum of basic outcomes of an event is equal to the probability of the event, the sum of basic outcomes in the sample space is 1.
  Probability rules: complement rule, addition rule, conditional probability, multiplication rule, statistical independence.
- Joint probability: probability of intersection. Marginal probability: probabilities for individual events. Conditional probability: probability of event A, given that event B happened.
- Bayes' theorem: how probability statements should be adjusted, given additional information.

Chapter 4

- Random variable: A discrete random variable can take on no more than a countable number of values. A continuous random variable can take any value in an interval. Linear function of a random variable: Y = a + bX
- The probability distribution function, P(x), of a random variable X represents the probability that X takes the value x, as a function of x.
  Cumulative probability distribution, $F(x_0)$, is the sum of probabilities.
  The expected value of a random variable X is also called its mean.
- Bernoulli distribution: a random experiment that can give rise to just two possible mutually exclusive and collectively exhaustive outcomes.
- Combination: number of sequences with x successes in n independent trials.
- Binomial distribution: P(x successes in n independent trials).
- Poisson distribution function: $P(x)=\frac{e^{-\lambda}\lambda^{x}}{x!}$ for $x=0,1,2,....$
  The poisson distribution has been found to be particularly useful in waiting line problems.
- The hypergeometric distribution is used for situation similar to the binomial with the important exception that sample observations are not replaced in the population when sampling from a "small population".
- Covariance: the linear association between two random variables. Correlation coefficient: measure of strength of the linear relationship between two random variables.
- Portfolio analysis: the linear combination of the mean values of the stocks in the portfolio.

Published on *WorldSupporter* (www.worldsupporter.org)

Chapter 5

- Probability that a continuous random variable X falls in a specified range:
  $P(a< X< b)=F(b)-F(a)$
  Properties probability density function: f(x) > 0 for all values of x, area under f(x) over all values of the random variable is equal to 1.0, the probability that X lies between a and b is the area under the probability density function between these points.
- A uniform distribution defined over the range from a to b: $f(x)=\frac{1}{b-a} \quad a\leqslant X\leqslant b$
- The shape of the probability density function is a symmetric bell-shaped curve centered on the mean.
- Notation normal distribution: $X\sim N(\mu,\sigma^2)$
- Probability that the number of successes will be between a and b:
  $$P(a\leqslant X\leqslant b)=P(\frac{a-nP}{\sqrt{nP(1-P)}}\leqslant Z\leqslant\frac{b-nP}{\sqrt{nP(1-P)}})$$
- The exponential distribution: $P(T\leqslant t_a)=(1-e^{-\lambda t_a})$
  Joint cumulative distribution: $F(x_1,x_2,\ldots,x_K)=P(X_1< x_1\cap X_2< x_2\cap\ldots\cap X_K< x_K)$
- Covariance: $Cov(X,Y)=E[XY]-\mu_X\mu_Y$
  Correlation: $p=Corr(X,Y)=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}$
- Linear combinations of random variables: $W=aX+bY$ or $W=aX-bY$
  $\mu_W=E[W]=E[aX+bY]=a\mu_X+b\mu_Y \quad \sigma_W^2=a^2\sigma_X^2+b^2\sigma_Y^2+2\text{abCov}(X,Y)$
  $\mu_W=E[W]=E[aX-bY]=a\mu_X-b\mu_Y \quad \sigma_W^2=a^2\sigma_X^2+b^2\sigma_Y^2-2\text{abCov}(X,Y)$

Chapter 6

- Simple random sampling: selects a sample of n objects from a population in such a way that each member of the population has the same probability of being selected.
- Sample mean of random variables $X_1,X2,\ldots,X_n$ is: $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$
- The sampling distribution of the sample means is the population mean: :
  $$E[\bar{X}]=E[\frac{1}{n}(X_1+X_2+\ldots+X_n)]=\frac{n\mu}{n}=\mu \quad Var(\bar{X})=\frac{\sigma^2}{n} \quad \sigma_x=\frac{\sigma}{\sqrt{n}}$$
- Central limit theorem: as n becomes large, the central limit theorem states that the distribution of $Z=\frac{\bar{X}-\mu_X}{\sigma_X}$ approaches the standard normal distribution.
- Acceptance intervals: an interval within which a sample mean has high probability of occurring.
- Sample proportion is the proportion of the population members that have a characteristic of interest.
  Sample distribution: the sample proportion of successes in a random sample from a population with proportion of success P.
  Sample variance: $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ Sample standard deviation: $s=\sqrt{s^2}$

Sample distribution: s^2 is the sample variance for a random sample of n observations from a population with variance σ^2

- If the population distribution is normal, then: $X^2_{n-1}=\frac{(n-1)s^2}{\sigma^2}$ is distributed as a chi-squared distribution with n – 1 degrees of freedom.

Chapter 7

- A specific value of a random variable is called an estimate. A point estimator is an unbiased estimator of a population parameter if its expected value is equal to that parameter. The bias of an unbiased estimator is 0. If there are several unbiased estimators of a parameter, then the unbiased estimator with the smallest variance is called the most efficient estimator.
- Relative efficiency: $Var(\hat{\theta}_2)/Var(\hat{\theta}_1)$
- A confidence interval estimator for a population parameter is a rule for determining an interval that is likely to include the parameter.
  The quantity % is called the confidence level of the interval.
  ME, the margin of error, is the error factor. Reducing the margin of error by reducing the standard deviation, increasing the sample size or decreasing the confidence level.
- Confidence interval for the population mean: $\bar{x}\pm ME$
  UCL: upper confidence limit. LCL: lower confidence limit
  - Confidence interval estimation for the mean population variance unknown: student's t distribution with n – 1 degrees of freedom. Reliability factor $t_{v,\alpha/2}$
  - Confidence interval estimation for population proportion: $\hat{p}\pm ME$
  - Confidence interval estimation for the variance of a normal distribution: The random variable x^2_{n-1} follows a chi-square distribution with (n – 1) degrees of freedom.
  Confidence interval for the population total: $N\bar{x}\pm t_{n-1,\alpha/2}N\hat{\sigma}_{\bar{x}}$
- Point estimate: $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$
- Number of sample observation to achieve a certain interval: $n=\frac{z^2_{\alpha/2}\sigma^2}{ME^2}$
- Sample size for population proportion: $n=\frac{NP(1-P)}{(N-1)\sigma^2 p+ P(1-P)}$

Chapter 8

- Samples are considered to be dependent if the values in one sample are influenced by the values in the other sample. Difference between two observations: $d_i=x_i-y_i$
- If the the population distribution of the differences is assumed to be normal: $\bar{d}\pm ME$ with ME = $t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}}$
- To determine if either drug is more effective possibilities:
  $\mu_x-\mu_y$ could be positive, suggesting that drug X is more effective.
  $\mu_x-\mu_y$ could be negative, suggesting that drug Y is more effective.

$\mu_x - \mu_y$ could be zero, suggesting that drug X and drug Y are equally effective.

- Three possible situations confidence interval estimation of the difference between to normal population means, independent samples:

  - The population variances are known: $(\bar{x}-\bar{y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}$

  - Population variances unknown, assumed to be equal: $(\bar{x}-\bar{y}) \pm t_{n_x+n_y-2,\alpha/2}\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}$

  - Population variances unknown, not assumed to be equal: $(\bar{x}-\bar{y}) \pm t_{v,\alpha/2}\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}$

- Large sample-size → $(\hat{p}_x-\hat{p}_y) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x}+\frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$

Chapter 9

- Null hypothesis is considered to be true unless sufficient evidence to the contrary is obtained. Alternative hypothesis against which the null hypothesis is tested.
- One-sided alternative: $H_1: \mu < 27$ or $H_1: \mu > 27$
  Two-sided alternative: $H_1: \mu \neq 27$
- Simple hypothesis specifies a single value for a population parameter. Composite hypothesis specifies a range of values for a population parameter.
- Type I error is the rejection of a true null hypothesis.
  Type II error is the failure to reject a false null hypothesis.
  The probability of type II error: $\beta = P(\bar{x} < \bar{x}_c | \mu = \mu^*) = P(z < \frac{\bar{x}_c - \mu^*}{\sigma/\sqrt{n}})$
- The p-value is the significance level.
- Tests of the mean of a normal distribution: population variance unknown:
  $H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$ or $H_1: \mu < \mu_0$
- Tests of the variance of a normal distribution:
  $H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$ or $H_1: \sigma^2 < \sigma_0^2$ or $H_1: \sigma^2 \neq \sigma_0^2$

Chapter 10

- Tests of the difference between two normal population means:
  - Dependent samples: $H_0: \mu_x - \mu_y = 0$ $H_1: \mu_x - \mu_y \neq 0$
  Reject H0 if: $\frac{\bar{d}}{s_d/\sqrt{n}} < -t_{n-1,\alpha/2}$ or $\frac{\bar{d}}{s_d/\sqrt{n}} > t_{n-1,\alpha/2}$
  - Independent samples, three situations:
  Known population variances: $H_0: \mu_x - \mu_y = 0$ $H_1: \mu_x - \mu_y \neq 0$
  Reject H0 if: $\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} < -z_{\alpha/2}$ or $\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} > z_{\alpha/2}$

Unknown population variances, assumed to be equal: $H_0: \mu_x - \mu_y = 0 \quad H_1: \mu_x - \mu_y \neq 0$

Reject H0 if: $\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}} < -t_{n_x+n_y-2,\alpha/2}$ or $\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}} > t_{n_x+n_y-2,\alpha/2}$

Unknown population variances, not assumed equal: $H_0: \mu_x - \mu_y = 0 \quad H_1: \mu_x - \mu_y \neq 0$

Reject H0 if: $\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}} < -t_{v,\alpha/2}$ or $\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}} > t_{v,\alpha/2}$

- Large samples tests: $H_0: P_x - P_y = 0 \quad H_1: P_x - P_y \neq 0$

  Reject H0 if: $\frac{(\hat{p}_x-\hat{p}_y)}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x}+\frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} < -z_{\alpha/2}$ or $\frac{(\hat{p}_x-\hat{p}_y)}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x}+\frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} > z_{\alpha/2}$

- Tests of the equality of the variances between two normally distributed populations:

  $H_0: \sigma_x^2 = \sigma_y^2 \quad H_1: \sigma_x^2 \neq \sigma_y^2$ Reject H0 if: $\frac{s_x^2}{s_y^2} > F_{n_x-1,n_y-1,\alpha/2}$

- The tests developed are based on the assumption that the underlying distribution is normal or that the central limit theorem applies for the distribution.

Chapter 11

- Least squares regression line: $\hat{y}: b_0 + b_1 x$ with slope b1 and y-intersection b0.

  Linear squares regression population model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

  Estimates of the linear equation coefficients b_0 and b_1 → $\hat{y}_i = b_0 + b_1 x_i$
- Analysis of variance: Sum of squares total = Sum of squares regression + sum of squares error. SST = SSR(explained by the regression) + SSE (unexplained error).

  Coefficient of determination $R^2 = SSR/SST = 1 - (SSE/SST)$

  Correlation: $R^2 = r^2$
- The variance of the slope coefficient depends on two important quantities: The distance of the points from the regression line and the total deviation of the X values from the mean.
- Tests of the population regression slope with the student's t distribution:

  $H_0: \beta_1 = \beta_1^* \quad H_1: \beta_1 \neq \beta_1^*$
- Hypothesis test for population slope coefficient using the F distribution:

  $H_0: \beta_1 = 0 \quad H_1: \beta_1 \neq 0$
- Regression models can be used to compute predictions or forecasts for the dependent variable, given an assumed future value for the independent variable.

  The probability is $1-\alpha$ that the interval includes the true prediction of Y.
- Prediction and confidence interval:

  The larger the sample size n, the narrower are both the prediction interval and the confidence interval.

  The larger s_e^2 , the wider are both the prediction interval and the confidence interval.

A large dispersion allows more precise estimates of the population regression line and correspondingly narrower confidence intervals and narrower prediction intervals.

- Hypothesis test for correlation: $H_0: \rho = 0 \quad H_1: \rho \neq 0$
- Diversifiable risk is that risk associated with specific firms and industries.
  Nondiversifiable risk is that risk associated with the entire economy.
- The beta coefficient for a specific firm is the slope coefficient, indicates how responsive the returns for a particular firm are to the overall market returns. The higher beta value, the higher required return on investment.
  The required return on an investment =
  $(Risk-free\,rate)+[(beta\ for\ investment)\times((Market\,return)-(risk-free\,rata))]$
- Extreme points are points that have X values that deviate substantially from the X values for the other points. The leverage for a point. Outlier points are those that deviate substantially in the Y direction from the predicted value.

Chapter 12

- Regression objectives are either to predict the value of the dependent variable, or to estimate the marginal effect of each independent variable.
- A population multiple regression model is a model that includes multiple independent variables.
- Standard multiple regression assumptions include the four standard simple regression assumptions, plus a fifth one: It is not possible to find a set off nonzero numbers such that the sum of the coefficients equals zero.
- Multiple regression models include an error term, ε, that represents variability caused by variables not included in the model.
- In multiple regression coefficients are estimated using least squares, but these estimates become less reliable the higher the correlations between independent variables are.
- Any regression coefficient in a multiple regression model is dependent on all independent variables, and are thus referred to as conditional coefficients.
- Mean square regression (MSR) shows the proportion of the variability by the dependent variable that can be explained by the regression model.
- In a multiple regression model the sum-of-squares (SST; or sample variability) can be split into the sum of squares regression (SSR; or explained variability) and the sum of squares error (SSE; or unexplained variability). This is referred to as sum-of-squares decomposition.
- The coefficient of determination, R^2, describes the strength of the linear relationship between the independent variables and the dependent variables, and is calculated by 1 – SSE/SST.
- Adding more independent variables leads to a misleading increase in R^2, which can be avoided by calculating the adjusted coefficient of determination.

$$-R^2 = 1 - \frac{SSE/(n-K-1)}{SST/(n-1)}$$

- The coefficient variance estimator, s 2b , is calculated as:

$$S_{b1}^2 = \frac{S_{e^2}}{(n-1)S_{x1}^2(1-r_{x1,x2}^2)}$$

- The square root of s^2_b is the coefficient standard error.
- Multiple regression models can be transformed into non-linear models, namely quadratic models and logarithmic models.

- Dummy variables can be used to represent categorical data in a regression model, and have a value of either 0 or 1.

Chapter 13

- Models are developed through four steps: model specification (selecting the variables, the algebraic form, and the data), coefficient estimation, model verification (checking whether the model is still accurate), and interpretation and inference.
- Dummy variables can be used to represent more than two categories by using multiple dummy variables. The rule is: *number of categories -1 = number of dummy variables.*
- In time series data the values of the dependent variable are related, this is then referred to as a *lagged dependent variable.*
- Not including important independent variables in a model can make any conclusions drawn from this model faulty.
- Multicollinearity is the phenomenon of two highly correlated independent variables. This leads to misleading estimated coefficients.
- Correlations between error terms are called auto-correlated errors. This leads to the estimated standard errors for the coefficients being biased, the null hypotheses falsely being rejected, and confidence intervals being too narrow. Autocorrelation can be formally tested with the Durbin-Watson test.

Chapter 14

- Goodness-of-fit test: an assessment of the closeness of the fit to the assumed population distribution of probabilities.

  A goodness-of-fit test: $Reject\ H_0 if \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} > X^2_{K-1,\alpha}$
- The random variable X^2_{K-1} follows a chi-square distribution with K – 1 degrees of freedom. Degrees of freedom: (K – m – 1)
- Jarque-Bera test for normality: $JB = n[\frac{(skewness)^2}{6} + \frac{(kurtosis-3)^2}{24}]$

  with $skewness = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3}$ and $kurtosis = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4}$
- Contingency tables: $Reject\ H_0 if: \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(O_{ij} - E_{ij})^2}{E_{ij}} > x^2_{(r-1)(c-1),\alpha}$ with $E_{ij} = \frac{R_i C_j}{n}$
- The sign test is used in market research studies to determine if consumer preference exists for one of two products. Calculate the difference for each pair of observations and record the sign of this difference: $H_0 : P = 0.5$

  The Wilcoxon signed rank test provides a method for incorporating information about the magnitude of the difference between matched pairs.

  Reject H0 if $T \leqslant T\ Appendix\ table\ 10$ with $T = min(T_+, T_-)$

- As a consequence of the central limit theorem, the normal distribution can be used to approximate the binomial distribution if the sample size is large.
  $S*=S+0.5 \; if \; S<\mu$ or $S*=S-0.5 \; if \; S>\mu$
- Wilcoxon signed rank test large sample: Reject H0 if $\frac{T-\mu_T}{\sigma_T} < -z_{\alpha/2}$
- Mann-whitney U test: $Z=\frac{U-\mu_U}{\sigma_U}$

  $U=n_1 n_2+\frac{n_1(N_1+1)}{2}-R_1 \qquad E(U)=\mu_U=\frac{n_1 n_2}{2} \qquad Var(U)=\sigma_U^2=\frac{n_1 n_2(n_1+n_2+1)}{12}$
- Wilcoxon rank total statistic T: $Z=\frac{T-\mu_T}{\sigma_T}$

  $E(T)=\mu_T=\frac{n_1(n_1+n_2+1)}{2} \qquad \sigma_T^2=\frac{n_1 n_2(n_1+n_2+1)}{12}$
- Spearman rank correlation coefficient: If x_i and y_i are each ranked in ascending order and the sample correlation of these ranks is calculated, the resulting coefficient is called the Spearman rank correlation coefficient.
  Test: $Reject \; H_0 \; if \qquad r_s < -r_{s,\alpha}$ or $r_s > r_{s,\alpha}$
- Runs test: $H_0 : the \; series \; is \; random$
- Large sample size: n > 20

Chapter 15

- An Analysis of Variance (ANOVA) can be used to analyze data at more than two levels. It can compare more than two populations and uses assessments of variation.
- A one-way ANOVA tests the equality of population means.
- The total sum of squares (SST) in a one-way ANOVA is made up of a within- group sum of squares (SSW) and a between-groups sum of squares (SSG).
- The Kruskal-Wallis test is a nonparametric alternative to the one-way ANOVA and is used when there is a strong indication that the parent population distributions are markedly different from the normal.
- A two-way ANOVA test the equality of the population means when there is asecond independent variable.
- The SST in a two-way ANOVA is made up of a between-blocks sum of squares (SSB), a between-groups sum of squares (SSG), and an error sum of squares (SSE).
- It is possible to have a two-way ANOVA with multiple observations per cell.
- This adds another factor that makes up the SST, the interaction sum of squares (SSI).

Chapter 16

- Time series data involves measurements that are ordered over time, in which the sequence of observations is important.
- Most time-series have four components: a trend component (steady increases or decreases), a seasonality component (changes specific to seasons), a cyclical component (a repeating pattern), and an irregular component (representing unpredictability, similar to ε).

- Moving averages can be used to adjust time-series data by removing the irregular component and/or seasonal component. This is done by replacing a value by the average of itself and its two neighbouring values (for removing the irregular component) or producing four-period moving averages (removing the seasonal component).
- The effect of the seasonal component can be calculated through the seasonal index method which compares the smoothed data with the original data.
- When the data series is non-seasonal and has no consistent trends simple exponential smoothing can be used to predict future data in the time series. This is done through estimation of a weighted average of current and past values, where more weight is given to the most recent observations (with decreasing weight the older the observation is).
- The Holt-Winters exponential smoothing procedure allows for trend, by using the added variable of the trend estimateT t-1. This procedure can also be extended to allow for seasonality by using a set of recursive estimates from the time- series.
- Based on autocorrelation patters between adjacent periods, the procedure of autoregressive models can use the available time-series data to estimate the parameters of a model of the process that could have generated this data.
- The Box-Jenkins approach is a flexible approach to prediction time-series data, based on how to choose the appropriate model.
- ARIMA models are autoregressive integrated moving average models.

Chapter 17

- Stratified sampling involves breaking the population into strata (a.k.a. subgroups) according to a specific identifiable characteristic in such a way that each member of the population belongs to only one strata.
- Stratified random sampling is the process of selecting independent simple random samples from each strata.
- Sampling effort can be allocated among the strata by either using proportional allocation (sample-strata proportion is equal to strata-population proportion) or optimal allocation (strata with higher variance receive more sample effort).
- The method of optimal allocation is only optimal when trying to estimate an overall population parameter (mean/total/proportion) as precisely as possible.
- Cluster sampling involves breaking the population into clusters, making a random collection of clusters, and contact each member of these clusters for data.
- Two-Phase sampling involves an additional pilot study with a smaller sample previous to the study itself.
- Non-random sampling can either take the form of non-probabilistic sampling, where a convenient sample is chosen, or of quota sampling, where it is predetermined how many members of each group will be included.