Index

How to use graphs to describe data? - Chapter 1

1.1 Decision making in an uncertain environment

A *population* is the complete set of all items that interest an investigator. Population size, N, can be very large or even infinite. Example of a population is all potential buyers of a new product.

A *sample* is an observed subset of a population. The sample must represent the whole population. Sample size given by n.

Simple random sampling: is a procedure used to select a sample of n objects from a population in such a way that each member of the population is chosen strictly by chance. The selection of one member does not influence the selection of any other member.

Systematic sampling: involves the selection of every jth item in the population. J is the ratio of the population size N to the desired sample size, n (J=N/n). Starting point is randomly selected number between 1 and j.

A *parameter* is a numerical measure that describes a specific characteristic of a population. A *statistic* is a numerical measure that describes a specific characteristic of a sample.

Nonsampling errors:
- The population actually sampled is not the relevant one.
- Survey subjects may give inaccurate or dishonest answers.
- There may be no response to survey questions.

Problem definition:
What information is required? What is the relevant population? How should sample members be selected? How should information be obtained from the sample members?

Descriptive statistics: focus on graphical and numerical procedures that are used to summarize and process data.
Inferential statistics: focus on using the data to make predictions, forecasts, and estimates to make better decisions.

1.2 Classification of variables

A *variable* is a specific characteristic of an individual or object (for example age or weight).

Categorical variables produce responses that belong to groups or categories. For example: yes/no, gender, ranking.

Numerical variables:

Discrete variables have a finite number of values, often comes from a counting process.
Continuous variables may take on any value within a given range of real numbers. Usually arises from a measurement process. For example length, weight, time, distance, temperature.

With *qualitative data* there is no measurable meaning to the "difference" in numbers. For example a football player with number 20 on his shirt, is not twice as good as a player with number 10 on his shirt.

Levels of measurement:
Nominal, data obtained from categorical questions. For example: 1=yes, 2=no or 1=male, 2=female.
Ordinal, data indicate the rank ordering of items. For example: 1=bad, 2=average, 3=good.
With *quantitative data* there is a measurable meaning to the difference in numbers. For example the score on an exam.

Levels of measurement:
Interval, indicates rank and distance from an arbitrary zero measured in unit intervals. For example temperature: 80 degrees Celsius is not four times as warm as 20 degrees Celsius.
Ratio, data indicates both rank and distance from a natural zero. Difference between two measures has meaning (age, weight, length). 200 pounds is twice the weight as 100 pounds.

1.3 Graphs to describe categorical variables

*Frequency distributio*n is a table used to organize data. The left column includes all possible responses on a variable being studied. The right column is a list of the frequencies, or number of observations, for each class. A *relative frequency distributio*n is obtained by dividing each frequency by the number observations and multiplying the resulting proportion by 100%.

- *Bar chart*, the height of a rectangle represents each frequency.
- *Cross tables*, list the number of observations for every combination of values for two categorical or ordinal variables. For example, gender separation.
- *Pie chart*, if we want to draw attention to the proportion of frequencies in each category.
- *Pareto diagram*, is a bar chart that displays the frequency of defect causes. The bar at the left indicates the most frequent cause and the bars to the right the causes with decreasing frequencies.

1.4 Graphs to describe time-series data

A time series is a set of measurements, ordered over time, on a particular quantity of interest. In a time series the sequence of the observations is important. For example; annual university enrollment, annual interest rates, monthly product sales.

A *line chart*, also called a time-series plot, is a series of data plotted at various time intervals. Measuring time along the horizontal axis and the numerical quantity of interest along the vertical axis yields a point on the graph for each observation. Joining points adjacent in time by straight lines produces a time-series plot.

1.5 Graphs to describe numerical variables

A *frequency distribution* for numerical data is a table that summarizes data by listing the classes in the left column and the number of observations in each class in the right column.

The construction of a frequency distribution:
1. Determine k, the number of classes
2. Determine the width of the classes:
   w = class width = (largest observation – smallest observation) / number of classes
3. Classes must be inclusive and non-overlapping

A *cumulative frequency distribution* contains the total number of observations whose values are less than the upper limit for each class.

Relative cumulative frequency distribution, cumulative frequencies are expressed as cumulative proportions or percents of the total.
Histogram, a graph that consists of vertical bars constructed on a horizontal line that is marked off with intervals for the variable being displayed. The intervals correspond to the classes in a frequency distribution table.

Ogive (cumulative line graph) is a line that connects point that are the cumulative percent of observations below the upper limit of each interval in a cumulative frequency distribution.

The shape of a distribution is said to be *symmetric* if the observations are balanced, or approximately evenly distributed, about its center. The center of the data divides a graph of the distribution into two "mirror images".

A distribution is skewed, or *asymmetric*, if the observations are not symmetrically distributed on either side of the center. A skewed-right distribution has a tail that extends farther to the right (positively skewed). A skewed-left distribution has a tail that extends farther to the left (negatively skewed).

Stem-and-leaf display is an EDA graph that is an alternative to the histogram. Data are grouped according to their leading digits (stems), and the final digits (leaves) are listed separately for each member of a class. The leaves are displayed individually in ascending order after each of the stems. The number of digits in each class indicates the class frequency.

A *scatter plot* is used to search for a possible relationship between two numerical variables. Prepared by locating one point for each pair of two variables that represent an observation in the data set. The scatter plot provides a picture of the data.
We can discover: the range of each variable, the pattern of values over the range, a suggestion as to a possible relationship between to variables or an indication of outliers.

1.6 Data presentation errors

Graphs must be as clear and accurate as possible. Just looking at the shape of the line is not enough to get a clear image of the data because of misleading time-series.

How can numerical measures be used to describe data? – Chapter 2

2.1 Measures of central tendency and location

Numerical measures (mean, median, mode) in response to questions concerning the location of the center of a data set. These numerical measures provide information about a "typical" observation in the data and are referred to as measures of central tendency.

Arithmetic mean is the sum of the data values divided by the number of observations.

$$\mu = \frac{\chi_1 + \chi_2 + \ldots\ldots + \chi_n}{N}$$

Median is the middle observation of a set of observations that are arranged in order. If the sample size is an even number, the median is the average of the two middle observations.

Mode (if exists) is the most frequently occurring value. *Unimodal*; there is one mode. *Bimodal*; there are two modes. *Multimodal;* with more than two modes. Most commonly used with categorical data.

Categorical data are best described by the median or the mode. However, the mode may not represent the true center of numerical data.

Numerical data are usually best described by the mean. However, the mean is affected by outliers.

Shape of a distribution, describe the shape of a distribution by computing a measure of skewness. *Skewness* is positive if a distribution is skewed to the right, negative for distributions skewed to the left and 0 for distributions that are symmetric about their mean. In a symmetric distribution, the mean and median are the same.

For continuous numerical unimodal data, the mean is usually less than the median in a skewed-left distribution and the mean is usually greater than the median in a skewed-right distribution.

Geometric mean, $\bar{\chi}_g$, is the nth root of the product of n numbers:

$$\bar{\chi}_g = \sqrt[n]{(\chi_1 \chi_2 \cdots\cdots \chi_{(n)})}$$

Geometric mean rate of return gives the mean percentage return of an investment over time:

$$\bar{r}_g = (\chi_1 \chi_2 \ldots\ldots \chi_n)^{1/n} - 1$$

Percentiles and quartiles are measures that indicate the location of a value relative to the entire set of data. To find percentiles and quartiles, data must first be arranged in order from the smallest to the largest values. The Pth percentile is a value such that approximately P% of the observations are at or below that number. The 50th percentile is the median.

Quartiles separate large data sets into four quarters:
Q1 = the value in the 0.25 (n+1)th ordered position
Q2 = the value in the 0.50 (n+1)th ordered position
Q3 = the value in the 0.75 (n+1)th ordered position

The five number summary refers to the five descriptive measures: minimum, first quartile, median, third quartile, and maximum:

Minimum < Q1 < median < Q3 < maximum

2.2 Measures of variability

While two data sets could have the same mean, the individual observations in one set could vary more from the mean than do the observations in the second set. *Range* is the difference between the largest and smallest observations.
However, the range may be an unsatisfactory measure of variability because outliers influence it. One way to avoid this difficulty is to arrange the data in ascending or descending order and discard a few of the highest and a few of the lowest numbers. If we remove the lowest 25% and the highest 25% of the data, we can measure the spread of the middle 54% of the data. To do this we use the *interquartile range*: Q3 – Q1.

A *box-and-whisker plot* (boxplot) is a graph that describes the shape of a distribution in terms of the five-number summary. The inner box shows the numbers that span the range from the first to the third quartile. A line is drawn through the box at the median. There are two "whiskers" (lines): one from the minimum to the 25th percentile and the other from the 75th percentile to the maximum value.

Variance:

The *population variance*, σ^2 , is the sum of the squared differences between each observation and the population mean divided by the population size, N:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (\chi_i - \mu)^2}{N}$$

The *sample variance*, s^2 , is the sum of the squared differences between each observation and the sample mean divided by the sample size, n, minus 1:

$$s^2 = \frac{\sum_{i=1}^{N} (\chi_i - \bar{\chi})^2}{n-1}$$

Standard deviation:

The population standard deviation, σ , is the square root of the population variance.
The sample standard deviation, s, is the square root of the sample variance.
The coefficient of variation, CV, is the measure of relative dispersion that expresses the standard deviation as a percentage of the mean:

Population coefficient of variation: $CV = \frac{\sigma}{\chi} \times 100\%$ if $\mu > 0$

Sample coefficient of variation: $CV = \frac{s}{\bar{\chi}} \times 100\%$ if $\bar{\chi} > 0$

Chebyshev's theorem: For any population with mean, standard deviation and k > 1, th percent of observations that lie within the interval $[\mu \pm k\sigma]$ is at least $100(1 - \frac{1}{k^2})\%$ if k > 1.

K stands for the number of standard deviations.

For many large populations (bell-shaped) the *Empirical Rule* provides an estimate of the approximate percentage of observations that are contained within one, two or three standard deviations of the mean:

Approximately 68% of the observations are in the interval $\mu \pm 1\sigma$
Approximately 95% of the observations are in the interval $\mu \pm 2\sigma$
Almost all of the observations are in the interval $\mu \pm 3\sigma$

A *z-score* is a standardized value that indicates the number of standard deviations a value is from the mean. A z-score greater than zero indicates that the value is greater than the mean, a z-score less than zero indicates that the value is less than the mean and a z-score of zero indicates that the value is equal to the mean.

$$z = \frac{\chi_i - \mu}{\sigma}$$

The standardized z-score is often used with admission tests for colleges and universities.

2.3 Weighted mean and measures of grouped data

Weighted mean = $\bar{\chi} = \frac{\sum w_i x_i}{n}$

One important situation that requires the use of a weighted mean is the calculation of grade point average (GPA).

2.4 Measures of relationships between variables

Covariance is a measure of the linear relationship between two variables. A positive value indicates a direct or increasing linear relationship, and a negative value indicates a decreasing linear relationship.

Population covariance: $Cov(x, y) = \sigma_{xy} = \frac{\sum_{i=1}^{N} (\chi_i - \mu_x)(y_i - \mu_y)}{N}$

Sample covariance: $Cov(x, y) = s_{xy} = \frac{\sum_{i=1}^{N} (\chi_i - \bar{\chi})(y_i - \bar{y})}{n-1}$

Covariance does not provide a measure of the strength of the relationship between two variables. The *correlation coefficient* also gives the strength of the relationship. We can compute the correlation coefficient by dividing the covariance by the product of the standard deviations of the two variables.

Population correlation coefficient: $\rho = \frac{cov(x, y)}{\sigma_x \sigma_y}$

Sample correlation coefficient: $r = \frac{cov(x, y)}{s_x s_y}$

The correlation coefficient ranges from -1 to +1. When r=0, there is no linear relationship.

It is important to understand that correlation does not imply causation. It is possible for two variables to be highly correlated, but that does not mean that one variable causes the other.

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

Elements of chance: how to use probability methods? – Chapter 3

3.1 Random experiment, outcomes, and events

A random experiment is a process leading to two or more possible outcomes, without knowing exactly which outcome will occur. For example; a coin is tossed and the outcome is either a head or a tail.

The possible outcomes from a random experiment are called the basic outcomes, and the set of all basic outcomes (O) is called the sample space. We use the symbol S to denote the sample space.

An event, E, is any subset of basic outcomes from the sample space. An event occurs if the random experiment results in one of its constituent basic outcomes. The null event represents the absence of a basic outcome and is denoted by Ø.

Intersection of events; is the set of all basic outcomes in S that belong to both event A and B. We use the term joint probability of A and B to denote the probability of the intersection of A and B.
Given K events $E_1, E_2, \ldots, E_k$, their intersection $E_1 \cap E_2 \cap \ldots \cap E_k$, is the set of all basic outcomes that belong to every $E_i (i = 1,2,\ldots, K)$.

Mutually exclusive events: If the events A and B have no common basic outcomes, they are called mutually exclusive, and their intersection, $A \cap B$, is said to be the empty set, indicating that $A \cap B$ has no members.

Union of events: The union of A and B, $A \cup B$, is the set of all basic outcomes in S that belong to at least one of these two events. Hence, the union $A \cup B$ occurs if and only if either A or B or both occur.

The events are *collectively exhaustive* if the union of several events covers the entire sample space, S.
$E_1 \cup E_2 \cup \ldots \cup E_k = S$

Complement: Let A be an event in the sample space, S. The set of basic outcomes of a random experiment belonging to S but not to A is called the complement of A and is denoted by $\bar{A}$. Events A and $\bar{A}$ are mutually exclusive.

Example:
When a batter is up, two events of interest are "the batter reaches base safely" $(Event\ A[O_1, O_2, O_6])$ and "the batter hits the ball" $(Event\ B[O_1, O_4, O_5, O_6])$:

1. Complements of these events: $\bar{A} = [O_3, O_4, O_5]$ and $\bar{B} = [O_2, O_3]$
2. Intersection of events: $A \cap B = [O_1, O_6]$
3. Union of events: $A \cup B = [O_1, O_2, O_4, O_5, O_6]$
4. Mutually exclusive: events A and $\bar{A}$, events B and $\bar{B}$

3.2 Probability and its postulates (three definitions)

- Classical probability: is the proportion of times that an event will occur, assuming that all outcomes in a sample space are equally likely to occur.

The probability of an event A is: $P(A) = \frac{N_A}{N}$

Where N_A is the number of outcomes that satisfy the condition of event A.
Permutations and Combinations:

1. *Number of possible orderings x objects*. $x(x-1)(x-2)\ldots(2)(1)=x!$ Suppose that we have some number x of objects that are to be placed in order. We can view this problem as a requirement to place one of the objects in each of x boxes arranged in a row. The first box, there are x different ways to fill it. Once an object is put in that box, there are (x-1) possible ways to fill the second box. Etc.
2. *Permutations*. We have a number n of objects with which the x ordered boxes could be filled (with n > x). The number of possible orderings is called the number of permutations of x objects chosen from n and is denoted by the symbol P_x^n .

   $$P_x^n=\frac{n!}{(n-x)!}$$
3. *Combinations*: When order is not important. Number of combinations of x objects chosen from n: $C_x^n=\frac{n!}{x!(n-x)!}$

- Relative frequency: indicate how often an event will occur compared to other events. An event with a probability of 0.40 will more often occur than an event with a probability of 0.30.
- Subjective probability: expresses an individual's degree of belief about the chance that an event will occur.

Probability postulates:

1. $0\leqslant P(A)\leqslant 1$ 2. $P(A)=\sum_A P(O_i)$ 3. $P(S)=1$

2. The probability lies between 0 and 1.
3. In terms of relative frequencies; the sum of basic outcomes of an event is equal to the probability of the event.
4. The sum of the probabilities for all basic outcomes in the sample space is 1.

3.3 Probability rules

- Complement rule: $P(\bar{A})=1-P(A)$
- Addition rule: probability of union: $P(A\cup B)=P(A)+P(B)-P(A\cap B)$
- Conditional probability: Probability of event A, given that even B has occurred:
  $$P(A|B)=\frac{P(A\cap B)}{P(B)}\ provided\ that\ P(B)>0$$
- Multiplication rule: Probability of the intersection of two events: $P(A\cap B)=P(A|B)P(B)$
- Statistical independence: $P(A\cap B)=P(A)P(B)$ or $P(A|B)=P(A)(if\ P(B)>0)$

3.4 Bivariate probabilities

Joint probabilities: intersection probability: $P(A_i\cap B_v)$

Marginal probability: the probabilities for individual events: $P(A_i)$ or $P(B_v)$

Conditional probability: $P(A_1|B_1)=\frac{P(A_1\cap B_1)}{P(B_1)}$

The odds in favor of a particular event are given by the ratio of the probability of the event divided by

the probability of its complement. Odds in favor of A: $Odds = \frac{P(A)}{1-P(A)} = \frac{P(A)}{P(\bar{A})}$

Over involvement ratios: the ratio of the probability of an event that occurs under two mutually exclusive and complementary outcome conditions. Example:

$B_1 : Buyers$ $A_1 : have\ seen\ an\ advertisement$

$B_2 : not-buyers$ $A_2 : not\ have\ seen\ an\ advertisement$

Odds of a buyer being conditional on the event "have seen an advertisement" : $\frac{P(B_1|A_1)}{P(B_2|A_1)}$

Advertising considered effective if: $\frac{P(B_1|A_1)}{P(B_2|A_1)} > \frac{P(B_1)}{P(B_2)}$

3.5 Bayes' Theorem

A way of revising conditional probabilities by using available information and a procedure for determining how probability statements should be adjusted, given additional information.

Multiplication rule: $P(A_1 \cap B_1) = P(A_1|B_1)P(B_1)$

Steps for Bayes' Theorem:

1. Define the subset events from the problem
2. Define the probabilities and conditional probabilities for the events defined in Step 1.
3. Compute the complements of the probabilities
4. Formally state and apply Bayes' theorem to compute the solution probability

The conditional probability of E_i , given A:

$$P(E_i|A_1) = \frac{P(A_1|E_i)P(E_i)}{P(A_1|E_1)P(E_1) + P(A_1|E_2)P(E_2) + \ldots + P(A_1|E_K)P(E_K)}$$

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

What are discrete probability distributions and how en when to use them? – Chapter 4

4.1 Random variables

A random variable is a variable that takes on numerical values realized by the outcomes in the sample space generated by a random experiment. A random variable is a discrete random variable if it can take on no more than a countable number of values. For example, the number of defective items in a sample of 20 items from a large shipment or the number of customers arriving at a checkout counter in an hour.

A random variable is a continuous random variable if it can take any value in an interval. For example yearly income, change in price, time, temperature.

4.2 Probability distributions for discrete random variables

The probability distribution function, P(x), of a discrete random variable X represents the probability that X takes the value x, as a function of x. $P(x)=P(X=x),\ for\ all\ values\ of\ x$

Required properties of probability distribution for discrete random variables:

1. $0\leqslant P(x)\leqslant 1\ for\ any\ value\ x$
2. the individual probabilities sum to 1: $\sum_x P(x)=1$

The cumulative probability distribution, $F(x_0)$, of a random variable X, represents the probability that X does not exceed the value x_0 , as a function of x_0 . $F(x_0)=P(X\leqslant x_0)$

Derived properties of cumulative probability distributions for discrete random variables:

1. $0\leqslant F(x_0)\leqslant 1\ for\ every\ number\ x_0$
2. $if\ x_0\wedge x_1\ are\ two\ numbers\ with\ x_0<x_1,\ then\ F(x_0)\leqslant F(x_1)$

4.3 Properties of discrete random variable

The expected value, $E[X]$, of a discrete random variable X is also called its mean μ is defined as $E[X]=\mu=\sum_x xP(x)$.

The variance of a discrete random variable can be denoted as: $\sigma^2=E[(X-\mu)^2]=\sum_x (x-\mu)^2 P(x)$

The standard deviation, σ, is the positive square root of the variance.

Expected value of function of random variables: $E[g(X)]=\sum_x g(x)P(x)$

Summary of properties for linear functions of a random variable: The random variable Y is a+bX.

The mean: $\mu_Y=E[a+bX]=a+b\mu_X$

The variance: $\sigma_Y^2=Var(a+bX)=b^2\sigma_X^2$

Standard deviation: $\sigma_Y=|b|\sigma_X$

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

4.4 Binomial distribution

Bernoulli distribution: a random experiment that can give rise to just two possible mutually exclusive and collectively exhaustive outcomes, which for convenience we label "success" and "failure". The probability of success is P and the probability of failure is (1-P). X is 1 if the outcome of the experiment is success and 0 otherwise. The probability distribution of this random variable is then

$P(0)=(1-P)$ and $P(1)=P$

Mean: $\mu_X=E[X]=\sum_x xP(x)$

Variance: $\sigma^2=E[(X-\mu)^2]=\sum_x (x-\mu)^2 P(x)$

Number of sequences with x successes in n independent trials is:

$C_x^n=\frac{n!}{x!(n-x)!}$ where $n!=nx(n-1)\times(n-2)\times\ldots\times 1$ and $0!=1$

The binomial distribution: P(x successes in n independent trials) =

$P(x)=\frac{n!}{x!(n-x)!}P^x(1-P)^{(n-x)}$ for x = 0,1,2, …., n

Mean: $\mu=E[X]=nP$

Variance: $\sigma_X^2=E[(X-\mu_X)^2]=nP(1-P)$

Binomial distribution:

1. The application involves several trials, each of which has only two outcomes: success or failure.
2. The probability of the outcome is the same for each trial.
3. The probability of the outcome on one trial does not affect the probability on other trials.

Example:

Fleur is a real estate agent, she has 5 contacts, and she believes that for each contact the probability of making a sale is 0.40. What is the probability that she makes at most 1 sale?

$P(X\leqslant 1)=P(X=0)+P(X=1)$

$P(0\,sales)=P(0)=\frac{5!}{0!\,5!}(0.4)^0(0.6)^5=0.078$

$P(1\,sale)=P(1)=\frac{5!}{1!\,4!}(0.4)^1(0.6)^4=0.259$ so $P(X\leqslant 1)=0.078+0.259=0.337$

4.5 Poisson distribution

An interval is divided into a very large number of equal subintervals so that the probability of the occurrence of an event in any subinterval very small is.

- the probability of the occurrence of an event is constant for all subintervals.
- There can be no more than one occurrence in each subinterval.
- Occurrences are independent; that is, an occurrence in one interval does not influence the probability of an occurrence in another interval.

The Poisson distribution function: $P(x)=\frac{e^{-\lambda}\lambda^x}{x!}$ for $x=0,1,2,\ldots$

$P(x)$ = the probability of x successes over a given time or space, given λ.

λ = the expected number of successes per time or space unit, λ>0

λ = 2,71828 (the base for natural logarithms)

Mean: $\mu_x = E[X] = \lambda$

Variance: $\sigma_x^2 = E[(X-\mu_x)^2] = \lambda$

The Poisson distribution has been found to be particularly useful in waiting line, or queuing, problems. Poisson approximation to the binomial distribution: when the number of trials, n, is large and at the same time the probability, P, is small.

If the number of trials, n, is large and nP is of only moderate size (preferably $nP \leqslant 7$), this distribution can be approximated by the Poisson distribution with $\lambda = nP$.

When P is less than 0.05 and n is large, we can approximate the binomial distribution by using the Poisson distribution.

4.6 Hypergeometric distribution

The hypergeometric distribution is used for situations similar to the binomial with the important exception that sample observations are not replaced in the population when sampling from a "small population". Therefore, the probability, P, of a success is not constant from one observation to the next.

A random sample of n objects is chosen from a group of N objects, S of which are successes. The distribution of the number of successes, X, in the sample is called the hypergeometric distribution. Its probability distribution:

$$P(x) = \frac{C_x^S C_{n-x}^{N-s}}{C_n^n} = \frac{\frac{S!}{x!(S-x)!} \times \frac{(N-S)!}{(n-x)!(N-S-n+x)!}}{\frac{N!}{n!(N-n)!}}$$

4.7 Jointly distributed discrete random variables

It is important that the effects of relationships is included in the probability model.

For example: Products at different quality levels have different prices and age groups have different preferences.

In 3.4 the probability of the intersection of bivariate events was presented, $P(A_i \cap B_j)$.

Here we use random variables: $P(x,y) = P(X = x \cap Y = y)$

Marginal probability distribution: $P(x) = \sum_y P(x,y)$ and $P(y) = \sum_x P(x,y)$

Properties of joint probability distributions of discrete random variables:

- $0 \leqslant P(x,y)\ for\ any\ pair\ of\ values\ x \wedge y$
- The sum of the joint probabilities P(x,y) over all possible pairs of values must be 1

Conditional probability distribution: $p(y|x) = \frac{p(x,y)}{p(x)}$ and $p(x|y) = \frac{p(x,y)}{p(y)}$

The jointly distributed random variables X and Y are said to be independent if and only if their joint probability distribution is the product of their marginal probability distributions:
$P(x,y) = P(x)P(y)$

Expected values of functions of jointly distributed random variables:

$$E[g(X,Y)]=\sum_x \sum_y g(x,y)P(x,y)$$

Covariance: is a measure of linear association between two random variables. The expected value of $(X-\mu_x)(Y-\mu_Y)$ is called the covariance between X and Y, denoted as $Cov(X,Y)$:

$$Cov(X,Y)=E[(X-\mu_x)(Y-\mu_Y)]=\sum_x \sum_y (X-\mu_x)(Y-\mu_Y)P(x,y)$$

Correlation coefficient: provides a measure of the strength of the linear relationship between two random variables, with measure being limited to the range from -1 to +1.

$$p=Corr(X,Y)=\frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

If the two random variables are statistically independent, the correlation is equal to 0.

Portfolio analysis: the linear combination of the mean values of the stocks in the portfolio.

Example: a portfolio that consists of a shares of stock A and b shares of stock B. We want to use the mean and variance for the market value, W, of a portfolio, where W is the linear function $W=aX+bY$.

Mean value for W: $\mu_w=E[W]=E[aX+bY]=a\mu_x+b\mu_y$

Variance for W:

$\sigma_W^2=a^2\sigma_X^2+b^2\sigma_Y^2+2abCov(X,Y)$ or $\sigma_W^2=a^2\sigma_X^2+b^2\sigma_Y^2+2abCorr(X,Y)\sigma_X\sigma_Y$

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

What kind of continuous probability distributions exist? – Chapter 5

5.1 Continuous random variables

Cumulative distribution function, (F(x), for a continuous random variable X expresses the probability that X does not exceed the value of x, as a function of x: $P(X \leqslant x) = F(x)$

Graph: vertical f(x), horizontal x.

Probability that a continuous random variable X falls in a specified range:
$P(a < X < b) = F(b) - F(a)$

Example: the cumulative distribution function is F(x) = 0.001x. The probability of sales between 250 and 750 gallon: $P(250 < X < 750) = (0.001)(750) - (0.001)(250) = 0.75 - 0.25 = 0.50$

Properties probability density function:

1. f(x) > 0 for all values of x
2. The area under the probability density function, f(x), over all values of the random variable, X within its range, is equal to 1.0.
3. Suppose that this density function is graphed. Let a and b be two possible values of random variable X, with a < b. Then, the probability that X lies between a and b is the area under the probability density function between these points: $P(a \leqslant X \leqslant b) = \int_a^b f(x)\,dx$
4. The cumulative distribution function, $F(x_0)$, is the area under the probability density function, f(x), up to x_0 : $F(x_0) = \int_{x_m}^{x_0} f(x)\,dx$

5.2 Expectations for continuous random variables

A uniform distribution defined over the range from a to b: $f(x) = \frac{1}{b-a} \qquad a \leqslant X \leqslant b$

Mean: $\mu_X = E[X] = \frac{a+b}{2}$

Variance: $\sigma_X^2 = E[(X - \mu_X)^2] = \frac{(b-a)^2}{12}$ → standard deviation: $\sigma_X = \sqrt{\sigma_X^2}$

Linear functions of random variables: $W = a + bX$

$\mu_W = E[a+bX] = a + b\mu_X \qquad \sigma_W^2 = Var[a+bX] = b^2\sigma_X^2 \qquad \sigma_W = |b|\sigma_X$

5.3 The normal distribution

Properties of the normal distribution:

1. The mean of the random variable: $E[X] = \mu$
2. The variance of the random variable: $Var(X) = \sigma^2$
3. The shape of the probability density function is a symmetric bell-shaped curve centered on the mean.

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

4. Notation normal distribution: $X \sim N(\mu, \sigma^2)$

Cumulative distribution function of the normal distribution: $F(x_0) = P(X \leqslant x_0)$ This is the area under the normal probability density function to the left of X_0 The total area under the curve is 1. The probability that X is between a and b: $P(a < X < b) = F(b) - F(a)$

Standard normal distribution: If Z is a normal random variable with mean 0 and variance 1:: $Z \sim N(0,1)$. Z follows the standard normal distribution.

Relationship between normally distributed random variable and Z: $Z = \frac{X - \mu}{\sigma}$

Example: A client has an investment portfolio whose mean value is equal to £1.000.000 with a standard deviation of £30.000. He asked you to determine the probability that the value of his portfolio is between £970.000 and £1.060.000.

1. Determine the corresponding Z values:

$$Z_{970.000} = \frac{970.000 - 1.000.000}{30.000} = -1.0 \qquad Z_{1.060.000} = \frac{1.060.000 - 1.000.000}{30.000} = +2.0$$

2. The probability that the portfolio value, X, is between £970.000 and £1.060.000 is equal to the probability that Z is between -1 and _2: $P(970.000 \leqslant X \leqslant 1.060.000) = P(-1 \leqslant Z \leqslant +2)$
3. $1 - P(Z \leqslant -1) - P(Z \geqslant +2) = 1 - 0{,}1587 - 0{,}0228 = 0{,}8185$

Other way: $P(a < X < b) = P(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma})$

5.4 Normal distribution approximation for binomial distribution

$X = X_1 + X_2 + \ldots + X_n$ With probabilities "success" P and "failure" 1-P

Mean: $E[X] = \mu = nP$

Variance: $Var(X) = \sigma^2 = nP(1-P)$

If the number of trials n is large, such that $nP(1-P) > 5$, then the distribution of the random variable is approximately a standard normal distribution. $Z = \frac{X - E[X]}{\sqrt{Var(X)}} = \frac{X - nP}{\sqrt{nP(1-P)}}$

Probability that the number of successes will be between a and b:

$$P(a \leqslant X \leqslant b) = P(\frac{a - nP}{\sqrt{nP(1-P)}} \leqslant Z \leqslant \frac{b - nP}{\sqrt{nP(1-P)}})$$

Probabilities for proportion or percentage intervals. A proportion random variable, P, can be computed by dividing the number of successes, X, by the sample size, n:

$\mu = P$ en $\sigma^2 = P(1-P)/n$

5.5 The exponential distribution

Exponential random variable T(t > 0) has a probability density function: $f(t) = \lambda e^{-\lambda t}$

λ is the mean number of independent arrivals per time unit.

Cumulative distribution function: $F(t) = 1 - e^{-\lambda t}$ for t > 0.

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

Mean: $1/\lambda$. Variance: $1/\lambda^2$.

$P(T \leqslant t_a) = (1 - e^{-\lambda . t_a})$ $P(t_b \leqslant T \leqslant t_a) = e^{-\lambda . t_b} - e^{-\lambda . t_a}$

5.6 Jointly distributed continuous random variables

The joint cumulative distribution , $F(x_1, x_2, \ldots, x_K)$ defines the probability that simultaneously X_1 is less than x_1 , X_2 is less than x_2 , and so on..

So: $F(x_1, x_2, \ldots, x_K) = P(X_1 < x_1 \cap X_2 < x_2 \cap \ldots \cap X_K < x_K)$

Covariance: $Cov(X, Y) = E[XY] - \mu_X \mu_Y$

Correlation: $p = Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$

Sums of random variables: $E[(X_1 + X_2 + \ldots + X_K)] = \mu_1 + \mu_2 + \ldots + \mu_K$

Sum of variance when covariant 0: $Var(X_1 + X_2 + \ldots + X_K) = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_K^2$

when covariance no 0: $Var(X_1 + X_2 + \ldots + X_K) = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_K^2 + 2 \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} Cov(X_i, X_j)$

Differences between a pair of random variables:

1. Mean: $E[X - Y] = \mu_X - \mu_Y$
2. Variance when covariance 0: $Var(X - Y) = \sigma_X^2 + \sigma_Y^2$
3. Variance when covariance not 0: $Var(X - Y) = \sigma_X^2 + \sigma_Y^2 - 2\,Cov(X, Y)$

Linear combinations of random variables:

$W = aX + bY$ $\mu_W = E[W] = E[aX + bY] = a\mu_X + b\mu_Y$ $\sigma_W^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2\,abCov(X, Y)$

$W = aX - bY$ $\mu_W = E[W] = E[aX - bY] = a\mu_X - b\mu_Y$ $\sigma_W^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 - 2\,abCov(X, Y)$

If both X and Y are joint normally distributed random variables, then the resulting random variable, W, is also normally distributed.

What are distributions of sample statistics? - Chapter 6

6.1 Sampling from a population

A simple random sample is chosen by a process that 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. The selection of one member is independent of the selection of any other member and every possible sample has the same probability of selection. It is important that a sample represent the population as a whole.

Often very difficult to obtain and measure every item in a population and the cost would be very high.

Sampling distribution example:

Florine is a supervisor with six employees, whose years of experience are: 2,4,6,6,7,8.

The mean: $\mu=(2+4+6+6+7+8)/6=5.5$

Two of these employees are to be chosen randomly for a particular work group. In this example we are sampling without replacement in a small population. There are fifteen possible different random samples that could be selected. Each of the 15 samples has the same probability, 1/15, of being selected. The sample mean 5.0 occurs three times, so the probability of obtaining a sample mean of 5.0 is 3/15.

6.2 Sample distributions of sample means

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:

$$E[\bar{X}]=E[\frac{1}{n}(X_1+X_2+\ldots+X_n)]=\frac{n\mu}{n}=\mu \qquad Var(\bar{X})=\frac{\sigma^2}{n} \qquad \sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}$$

If the sample size, n, is not small compared to the population size, N, then the standard error of $\bar{X}$ is as follows: $\sigma_x=\frac{\sigma}{\sqrt{n}}\times\sqrt{\frac{N-n}{N-1}}$

Standard normal distribution of the sample means: $Z=\frac{\bar{X}-\mu}{\sigma_{\bar{x}}}$

Central Limit Theorem: $X_1, X_2, \ldots, X_n$ is a set of n independent random variables having identical distributions with mean μ, variance σ^2 and $\bar{X}$ as the mean of these random variables. 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 a high probability of occurring. If the sample mean is within that interval, then we can accept the conclusion that the random sample came from the population with the known population mean and variance. Symmetric acceptance interval: $\mu\pm z_{\alpha/2}\sigma_{\bar{x}}$.

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

6.3 Sampling distributions of sample proportions

Sample proportion: $\hat{p}$ is the proportion of the population members that have a characteristic of interest. The sample proportion is: $\hat{p} = X/n$.

Sampling distribution: $\hat{p}$ is the sample proportion of successes in a random sample from a population with proportion of success P.

1. The sampling distribution of $\hat{p}$ has mean P: $E[\hat{p}] = P$
2. The sampling distribution of $\hat{p}$ has standard deviation: $\sigma_{\hat{p}} = \sqrt{\frac{P(1-P)}{n}}$
3. If the sample size is large, the random variable: $Z = \frac{\hat{p} - P}{\sigma_{\hat{p}}}$ is approximately distributed as a standard normal. This approximation is good if: $nP(1-P) > 5$

6.4 Sampling distributions of sample variances

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 :

1. The sampling distribution of s^2 has mean σ^2 : $E[s^2] = \sigma^2$
2. The variance of the sampling distribution of s^2 depends on the underlying population distribution. If that distribution is normal, then: $Var(s^2) = \frac{2\sigma^4}{n-1}$
3. If the population distribution is normal, then: $X^2_{n-1} = \frac{(n-1)s^2}{\sigma^2}$ is distributed as the chi-squared distribution with n-1 degrees of freedom, $(X^2_{(n-1)})$

Confidence interval estimation: one population – Chapter 7

7.1 Properties of point estimators

An estimator of a population parameter is a random variable that depends on the sample information. A specific value of that random variable is called an estimate. For example: the sample mean $\bar{X}$ is a point estimator of the population mean, μ, and the value that $\bar{X}$ assumes for a given set of data is called the point estimate.

A point estimator $\hat{\theta}$ is said to be an unbiased estimator of a population parameter θ if its expected value is equal to that parameter: $E(\hat{\theta})=\theta$

The bias in $\hat{\theta}$: $E(\hat{\theta})-\theta$ 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. $\hat{\theta}_1 \, meer \, efficiënt \, dan \, \hat{\theta}_2 \, wanneer : Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$

The relative efficiency: $\frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}$

7.2 Confidence interval estimation for the mean of a normal distribution: population variance known.

A confidence interval estimator for a population parameter is a rule for determining an interval that is likely to include the parameter.

The interval from a to b is called a 100(1-α)% confidence interval of θ . The quantity % is called the confidence level of the interval. The confidence interval is written as: $a<\theta<b \, , met \, 100(1-\alpha)\%$ confidence.

ME, the margin of error, is the error factor: $\hat{\theta} \pm ME$

Consider a random sample of n observations from a normal distribution with mean μ and variance σ^2 . If the sample mean is $\bar{x}$, then a 100(1-α)% confidence interval for the population mean with known variance is given by:

$$\bar{x} \pm ME \rightarrow ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad width \, , w = 2(ME) \qquad UCL = \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad LCL = \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

ME; margin of error. UCL; Upper confidence limit. LCL; Lower confidence limit.

Reducing the margin of error by reducing the standard deviation, increasing the sample size or decreasing the confidence level.

7.3 Confidence interval estimation for the mean of a normal distribution: population variance unknown.

Student's t distribution with (n-1) degrees of freedom: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

A random variable having the Student's t distribution with v degrees of freedom is denoted t_v .

Then $t_{v,\alpha/2}$ is the reliability factor, defined as the number for which: $P(t_v > t_{v,\alpha/2}) = \alpha/2$.

An random sample of n observations from a normal distribution with mean μ and variance unknown. If the sample mean and standard deviation are, $\bar{x}$ and s, then the degrees of freedom is v = n – 1, and a 100(1-α)% confidence interval for the population mean with unknown variance, is given by:

$\bar{x} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$ of $\bar{x} \pm ME$

7.4 Confidence interval estimation for population proportion (large samples)

Let $\hat{p}$ denote the observed proportion of "successes" in a random sample of n observations from a population with a proportion of successes P. Then, if $nP(1-P) > 5$, a 100(1-α)% confidence interval for the population proportion is given by: $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ or $\hat{p} \pm ME$.

7.5 Confidence interval estimation for the variance of a normal distribution

A random sample of n observations from a normally distributed population with variance σ^2 . If the observed sample variance is s^2 , then the lower and upper confidence limits of a 100(1-α)% confidence interval for the population variance is given by:

$LCL = \frac{(n-1)s^2}{x^2_{n-1,\alpha/2}}$ and $UCL = \frac{(n-1)s^2}{x^2_{n-1,1-\alpha/2}}$

where $x^2_{n-1,\alpha/2}$ is the number for which $P(x^2_{n-1} > x^2_{n-1,\alpha/2}) = \frac{\alpha}{2}$

and $x^2_{n-1,1-\alpha/2}$ is the number for which $P(x^2_{n-1} < x^2_{n-1,1-\alpha/2}) = \frac{\alpha}{2}$

The random variable x^2_{n-1} follows a chi-square distribution with (n – 1) degrees of freedom.

7.6 Confidence interval estimation: finite populations

- The point estimate: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

  Variance of the sample mean: $\hat{\sigma}^2_{\bar{x}} = \frac{s^2}{n}\left(\frac{N-1}{N-1}\right)$

  A 100(1-α)% confidence interval for the population mean is given by: $\bar{x} \pm t_{n-1,\alpha/2}\hat{\sigma}_{\bar{x}}$
- An unbiased estimation procedure for the variance of our estimator of the population total yields the point estimate: $N^2\hat{\sigma}^2_{\bar{x}} = N^2\frac{s^2}{n}\left(\frac{N-n}{N-1}\right)$ so $N\hat{\sigma}_{\bar{x}} = \frac{Ns}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N-1}\right)}$

  A 100(1-α)% confidence interval for the population total is obtained from: $N\bar{x} \pm t_{n-1,\alpha/2}N\hat{\sigma}_{\bar{x}}$

- An unbiased estimation procedure for the variance or our estimator of the population proportion yields the point estimate: $\hat{\sigma}^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n-1}\left(\frac{N-n}{N-1}\right)$

  When the sample size is large 100(1-α)% confidence intervals for the population proportion are given by $\hat{p} \pm z_{\alpha/2}\hat{\sigma}_{\hat{p}}$

7.7 Sample-size determination: large populations

Random sample from a normally distributed population with known variance is selected. A 100(1-α)% confidence interval for the population mean extends a distance ME (sampling error) on each side of the sample mean if the sample size, n, is as follows: $n = \frac{z^2_{\alpha/2}\sigma^2}{ME^2}$ = number of sample observations to achieve a certain interval.

Sample size for population proportion: $n = \frac{0{,}25(z_{\alpha/2})^2}{(ME)^2}$

7.8 Sample-size determination: finite population

$n = \frac{n_0 N}{n_0 + (N-1)}$ with $n_0 = \frac{z^2_{\alpha/2}\sigma^2}{ME^2}$ with $Var(\bar{X}) = \sigma\frac{2}{X} = \frac{\sigma^2}{2}\left(\frac{N-n}{N-1}\right)$

Sample size for population proportion: $n = \frac{NP(1-P)}{(N-1)\sigma^2 p + P(1-P)}$

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

Confidence interval estimation: further topics – Chapter 8

8.1 Confidence interval estimation of the difference between two normal population means: dependent samples

Samples are considered to be dependent if the values in one sample are influenced by the values in the other sample.

Suppose that there is a random sample of n matched pairs of observations from normal distributions with means μ_x and μ_y. That is, let $x_1, x_2, \ldots, x_n$ denote the values of observations from the population with mean μ_x and let $y_1, y_2, \ldots, y_n$ denote the matched sampled values from the population with the mean μ_y.
Let $\bar{d}$ and s_d denote the observed sample mean and standard deviation for the n differences $d_i = x_i - y_i$. If the population distribution of the differences is assumed to be normal, then a 100(1-α)% confidence interval for the difference between two means, dependent samples $(\mu_d = \mu_x - \mu_y)$ is given by $\bar{d} \pm t_{n-1,\,\alpha/2} \frac{s_d}{\sqrt{n}}$ or, equivalently, $\bar{d} \pm ME$
The random variable t_{n-1} has a Student's t distribution with (n-1) degrees of freedom.

$$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$$

Example:

Pair	Drug X	Drug Y	Difference $d_i = x_i - y_i$
1	29	26	3
2	32	27	5
3	31	28	3
4	32	27	5
5	30		
6	32	30	2
7	29	26	3
8	31	33	-2
9	30	36	-6

99% confidence level

$$\bar{d} = 1.625 \qquad s_d = 3.777 \qquad t_{n-1,\alpha/2} = t_{7,0.005} = 3.499 \qquad n = 8$$

Estimate the mean difference in the effectiveness of the two drugs, X and Y:

$$\bar{d} \pm t_{n-1,\,\alpha/2} \frac{s_d}{\sqrt{n}} \quad \rightarrow \quad 1.625 \pm 3.499 \frac{3.777}{\sqrt{8}}$$ → LCL = -3.05 en UCL = 6.30

Since the confidence interval contains the value of zero, it is not possible, based on this data, to determine if either drug is more effective..
Possibilities:

1. $\mu_x - \mu_y$ could be positive, suggesting that drug X is more effective.
2. $\mu_x - \mu_y$ could be negative, suggesting that drug Y is more effective .
3. $\mu_x - \mu_y$ could be zero, suggesting that drug X and drug Y are equally effective.

8.2 Confidence interval estimation of the difference between two normal population means: independent samples.

Three situations:

1. Two means, independent samples, and known population variances.

Suppose that there are two independent random samples of n_x and n_y observations from normally distributed populations with means μ_x and μ_y and variances σ_x^2 and σ_y^2 . If the observed sample means are $\bar{x}$ and $\bar{y}$, then 100(1-α)% a confidence interval for the difference between two means, independent samples, and known population variances is given by:

$(\bar{x}-\bar{y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}$ or, equivalently, $(\bar{x}-\bar{y}) \pm ME$

2. Two means, independent samples, and unknown population variances assumed to be equal.

Suppose that there are two independent random samples with n_x and n_y observations from normally distributed populations with means μ_x and μ_y , and a common, but unknown, population variance. If the observed sample means are $\bar{x}$ and $\bar{y}$, and the observed sample variances are s_x^2 and s_y^2 then a 100(1-α)% confidence interval for the difference between two means, independent samples, and unknown population variances assumed to be equal is given by: $(\bar{x}-\bar{y}) \pm t_{n_x+n_y-2,\,\alpha/2}\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}$ or, equivalently, $(\bar{x}-\bar{y}) \pm ME$

the pooled sample variance s_p^2 is given by: $s_p^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}$

3. Two means, independent samples, and unknown population variances not assumed to be equal.

Suppose that there are two independent random samples of n_x and n_y observations from normally distributed populations with means μ_x and μ_y and it is assumed that the population variances are not equal. If the observed sample means and variances are $\bar{x}$ and $\bar{y}$ and s_x^2 and s_y^2 , then a 100(1-α)% confidence interval for the difference between two means, independent samples, and unknown population variances not assumed to be equal is given by:

$(\bar{x}-\bar{y}) \pm t_{v,\alpha/2}\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}$ or, equivalently $(\bar{x}-\bar{y}) \pm ME$

the degrees of freedom, v, is given by: $v = \frac{\left(\frac{s_x^2}{n_x}\right)+\left(\frac{s_y^2}{n_y}\right)^2}{\left(\frac{s_x^2}{n_x}\right)^2/(n_x-1)+\left(\frac{s_y^2}{s_y^2}\right)^2/(n_y-1)}$

If the sample sizes are equal, then the degrees of freedom reduces to: $v=(1+\frac{2}{\frac{s_x^2}{s_y^2}+\frac{s_y^2}{s_x^2}})\times(n-1)$

8.3 Confidence interval estimation of the difference between two population proportions (large samples)

P_x is the observed proportion of successes in a random sample of n_x observations from a population with proportion P_x of successes. $\hat{p}_y$ Is the observed proportion of successes in an independent random sample of n_y observations from a population with proportion P_y of successes. If the sample sizes are large, a 100(1-α)% confidence interval for the difference between population proportions (large samples) P_x - P_y, is given by:

$(\hat{p}_x-\hat{p}_y)\pm z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x}+\frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$ or, equivalently, $(\hat{p}_x-\hat{p}_y)\pm ME$

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

How to conduct a hypothesis tests of a single population? - Chapter 9

9.1 Concepts of hypothesis testing

Null hypothesis, H0: A maintained hypothesis that is considered to be true unless sufficient evidence to the contrary is obtained.

Alternative hypothesis, H1: A hypothesis against which the null hypothesis is tested and which will be held to be true if the null is declared to be false.

Simple hypothesis: A hypothesis that specifies a single value for a population parameter.

Composite hypothesis: A hypothesis that specifies a range of values for a population parameter.

One-sided alternative: An alternative hypothesis involving all possible values of a population parameter on either one side or the other of the value specified by a simple null hypothesis (either greater than or less than). $H_1: \mu < 27$ or $H_1: \mu > 27$

Two-sided alternative: an alternative hypothesis involving all possible values of a population parameter other than the value specified by a simple null hypothesis (both greater than or less than).
$H_1: \mu \neq 27$

Two possible errors:

Type I error: The rejection of a true null hypothesis, α, significance level.

Type II error: The failure to reject a false null hypothesis, β, power.

9.2 Tests of the mean of a normal distribution: population variance known.

One-sided alternative:

$$H_0: \mu = \mu_0 \quad H_1: \mu > \mu_0 \rightarrow \text{Reject } H_0 \text{ if } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$$

$$H_0: \mu = \mu_0 \quad H_1: \mu < \mu_0 \rightarrow \text{Reject } H_0 \text{ if } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha$$

Two-sided alternative:

$$H_0: \mu = \mu_0 \quad H_1: \mu \neq \mu_0 \rightarrow \text{Reject } H_0 \text{ if } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leftarrow z_{\alpha/2} \text{ or } \text{Reject } H_0 \text{ if } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2}$$

P-value: $\text{Reject } H_0 \text{ if } : p - value < \alpha \qquad p - value = P\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geqslant z_p\right)$

P-value < 0.01, enough evidence to reject H0, high significance level.

P-value between 0.01 and 0.05, strong evidence to reject H0, significant.

P-value >0.05, little evidence to reject H0, not statistical significant.

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

9.3 Tests of the mean of a normal distribution: population variance unknown.

$H_0: \mu = \mu_0 \quad H_1: \mu > \mu_0 \rightarrow \quad Reject\ H_0\ if: t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} > t_{n-1,\alpha}$

$H_0: \mu = \mu_0 \quad H_1: \mu < \mu_0 \rightarrow \quad Reject\ H_0\ if: t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} < -t_{n-1,\alpha}$

9.4 Tests of the population proportion (large samples)

$H_0: P = P_0 \quad H_1: P > P_0 \rightarrow \quad Reject\ H_0\ if: \dfrac{\hat{p} - P_0}{\sqrt{P_0(1-P_0)/n}} > z_\alpha$

$H_0: P = P_0 \quad H_1: P < P_0 \rightarrow \quad Reject\ H_0\ if: \dfrac{\hat{p} - P_0}{\sqrt{P_0(1-P_0)/n}} < -z_\alpha$

$H_0: P = P_0 \quad H_1: P \neq P_0 \rightarrow \quad Reject\ H_0\ if: \dfrac{\hat{p} - P_0}{\sqrt{P_0(1-P_0)/n}} < -z_{\alpha/2}$ or $\dfrac{\hat{p} - P_0}{\sqrt{P_0(1-P_0)/n}} > z_{\alpha/2}$

9.5 Assessing the power of a test

Determining the probability of type II error:

$\beta = P(\bar{x} < \bar{x}_c | \mu = \mu^*) = P\left(z < \dfrac{\bar{x}_c - \mu^*}{\sigma/\sqrt{n}}\right)$ Power = $1 - \beta$

The value of the beta and the power will be different for every μ^*

Steps:

1. Form the test decision rule, find the range of values of the sample proportion leading to failure to reject the null hypothesis.
2. Using the value P_1 for the population proportion to find the probability that the sample proportion will be in the non rejection region.

Example:

$H_0: P = P_0 = 0.50 \quad H_1: P \neq 0.50 \quad n = 600 \quad \alpha = 0.05$

The decision rule, reject H0 if: $\dfrac{\hat{p} - P_0}{\sqrt{P_0(1-P_0)/n}} < -1.96$ or $\dfrac{\hat{p} - P_0}{\sqrt{P_0(1-P_0)/n}} > 1.96$

$\hat{p} > 0.50 + 1.96\sqrt{0.50(1-0.50)/600} = 0.50 + 0.04$ or $\hat{p} < 0.50 - 0.04 = 0.46$

We want to determine the probability of a Type II error when this decision rule is used. Suppose that the true population proportion was $P_1 = 0.55$.

We want to determine the probability that the sample proportion is between 0.46 and 0.54 if the population proportion is 0.55. Thus the probability of Type II error is as follows:

$$P(0.46 \leq \hat{p} \leq 0.54) = P\left[\frac{0.46 - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}} \leq Z \leq \frac{0.54 - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}}\right] \quad = \quad P(-4.43 \leq Z \leq -0.49) = 0.3121$$

The probability of a Type II error is $\beta = 0.3121$ The power of the test: $Power = 1 - \beta = 0.6879$

9.6 Tests of the variance of a normal distribution

$$H_0: \sigma^2 = \sigma_0^2 \quad H_1: \sigma^2 > \sigma_0^2 \quad \rightarrow \quad Reject\ H_0 if: \frac{(n-1)s^2}{\sigma_0^2} > X^2_{n-1,\alpha}$$

$$H_0: \sigma^2 = \sigma_0^2 \quad H_1: \sigma^2 < \sigma_0^2 \quad \rightarrow \quad Reject\ H_0 if: \frac{(n-1)s^2}{\sigma_0^2} < X^2_{n-1,1-\alpha}$$

$$H_0: \sigma^2 = \sigma_0^2 \quad H_1: \sigma^2 \neq \sigma_0^2 \quad \rightarrow \quad Reject\ H_0 if: \frac{(n-1)s^2}{\sigma_0^2} > X^2_{n-1,\alpha/2} \quad \text{or} \quad \frac{(n-1)s^2}{\sigma_0^2} < X^2_{n-1,\alpha/2}$$

Example:

Determine if the variance of impurities in shipments of fertilizer is within the established standard. This standard states that 100-pound bags of fertilizer, the variance in the pounds of impurities cannot exceed 4.

A random sample of 20 bags is obtained, the sample variance is computed to be 6.62

$$H_0: \sigma^2 \leq \sigma_0^2 = 4 \quad H_1: \sigma^2 > 4 \quad \rightarrow \quad Reject\ H_0 if: \frac{(n-1)s^2}{\sigma_0^2} > X^2_{n-1,\alpha}$$

$$\frac{(n-1)s^2}{\sigma_0^2} = \frac{(20-1)(6.62)}{4} = 31.445 \quad > \quad X^2_{n-1,\alpha} = 30.144$$

Therefore, we reject the null hypothesis and conclude that the variability of the impurities exceeds the standard. As a result, we recommend that the production process should be studied and improvements made to reduce the variability of the product components.

Published on *WorldSupporte*r (www.worldsupporter.org)

How to conduct a two population hypothesis tests? - Chapter 10

10.1 Tests of the difference between two normal population means: dependent samples.

A random sample of n matched pairs of observations from distributions with means μ_x and μ_y. Let *d* and s_d denote the observed sample mean and standard deviation for the n differences $(x_i - y_i)$. If the population distribution of the differences is a normal distribution, then the following tests have significance level α :

1. $H_0: \mu_x - \mu_y = 0 \quad H_1: \mu_x - \mu_y > 0$
   Reject H0 if: $\frac{\bar{d}}{s_d/\sqrt{n}} > t_{n-1,\alpha}$
2. $H_0: \mu_x - \mu_y = 0 \quad H_1: \mu_x - \mu_y < 0$
   Reject H0 if: $\frac{\bar{d}}{s_d/\sqrt{n}} < -t_{n-1,\alpha}$
3. $H_0: \mu_x - \mu_y = 0 \quad H_1: \mu_x - \mu_y \neq 0$
   Reject H0 if: $\frac{\bar{d}}{s_d/\sqrt{n}} < -t_{n-1,\alpha/2}$ or $\frac{\bar{d}}{s_d/\sqrt{n}} > t_{n-1,\alpha/2}$

10.2 Tests of the difference between two normal population means: Independent samples.

Three situations::

1. Independent samples, known population variances.

Testing the null hypothesis: $H_0: \mu_x - \mu_y = 0 \quad H_1: \mu_x - \mu_y \neq 0$

Reject H0 if: $\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} < -z_{\alpha/2}$ of $\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} > z_{\alpha/2}$

One-sided test → < $-z_\alpha$ or > z_α

2. Independent samples, unknown population variances, assumed to be equal.

Testing the null hypothesis: $H_0: \mu_x - \mu_y = 0 \quad H_1: \mu_x - \mu_y \neq 0$

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

One-sided test → < $-t_{n_x+n_y-2,\alpha}$ of > $t_{n_x+n_y-2,\alpha}$

3. Independent samples, unknown population variances, not assumed to be equal

Testing the null hypothesis: $H_0: \mu_x - \mu_y = 0 \quad H_1: \mu_x - \mu_y \neq 0$

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 $\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}} > t_{v,\alpha/2}$

One-sided test → $< -t_{v,\alpha}$ of $> t_{v,\alpha}$

10.3 Tests of the difference between two population proportions (large samples)

Independent random samples of size n_x and n_y with proportion of successes $\hat{p}_x$ and $\hat{p}_y$
When we assume that the population proportions are equal, an estimate of the common proportion is as follows: $\hat{p}_0 = \frac{n_x \hat{p}_x + n_y \hat{p}_y}{n_x + n_y}$

For large sample sizes the following tests have significance level α:

Testing the null hypothesis: $H_0: P_x - P_y = 0$ $H_1: P_x - P_y \neq 0$

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 $\frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x}+\frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} > z_{\alpha/2}$

One-sided test → $< -z_\alpha$ of $> z_\alpha$

10.4 Tests of the equality of the variances between two normally distributed populations.

We have two independent samples with n_x and n_y observations from two normal populations with variances σ_x^2 and σ_y^2, sample variances s_x^2 and s_y^2.

The F distribution $F = \frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2}$ with numerator degrees of freedom $(n_x - 1)$ and denominator degrees of freedom $(n_y - 1)$

Tests of equality of variances from two normal populations:
$H_0: \sigma_x^2 = \sigma_y^2$ $H_1: \sigma_x^2 \neq \sigma_y^2$

Reject H0 if: $\frac{s_x^2}{s_y^2} > F_{n_x-1, n_y-1, \alpha/2}$

One-sided test → $F_{n_x-1, n_y-1, \alpha}$

10.5 Some comments on hypothesis testing

Defining the null and alternative hypotheses requires careful consideration of the objectives of the analysis

The tests developed in this chapter are based on the assumption that the underlying distribution is normal or that the central limit theorem applies for the distribution of sample means or proportions.

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

How to conduct a two variable regression analysis? - Chapter 11

11.1 Overview of linear models

Least squares regression line: $\hat{y}: b_0 + b_1 x$ → Slope $b_1 = \frac{Cov(x,y)}{s_x^2} = r\frac{s_y}{s_x}$

y-intercept $b_0 = \bar{y} - b_1 \bar{x}$

11.2 Linear regression model

Assumptions:

1. The Y's are linear functions of X plus a random error term: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
2. The x values are fixed numbers, $x_i (i = 1, \ldots, n)$
3. The error terms are random variables; $E[\varepsilon_i] = 0$ $E[\varepsilon_i^2] = \sigma^2$ for $(i = 1, \ldots, n)$
4. The random error terms, ε_i , are not correlated with one another, so that $E[\varepsilon_i \varepsilon_j] = 0$ for all.

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

11.3 Least squares coefficient estimators

Estimates of the linear equation coefficients b_0 and b_1 → $\hat{y}_i = b_0 + b_1 x_i$

11.4 The explanatory power of a linear regression equation

Analysis of variance; SST = SSR (explained by the regression) + SSE (unexplained error)

Sum of squares total: $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$

Sum of squares error: $SSE = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} e_i^2$

Sum of squares regression: $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$

Coefficient of determination, $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

Correlation → $R^2 = r^2$

Estimation of model error variance : $\hat{\sigma}^2 = s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2}$

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

11.5 Statistical inference: hypothesis tests and confidence intervals

Sampling distribution of the least squares coefficient estimator:

If the standard least squares assumptions hold, then b_1 is an unbiased estimator for β_1 and has a population variance: $\sigma_{b_1}^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{(n-1)s_x^2}$

And an unbiased sample variance estimator: $\sigma_{b_1}^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{(n-1)s_x^2}$

The variance of the slope coefficient depends on two important quantities:

1. The distance of the points from the regression line measured by s_e^2 . Higher values imply greater variance for b_1 .
2. The total deviation of the X values from the mean, which is measured by $(n-1)s_x^2$. Greater deviations in the X values and larger sample sizes result in a smaller variance for the slope coefficient.

Basis for inference about the population regression slope:
Let β_1 be a population regression slope and b_1 be its least squares estimate based on n pairs of sample observations. Then, if the standard regression assumptions hold and it can also be assumed that the errors, ε_i , are normally distributed, the random variable $t = \frac{b_1 - \beta_1}{s_{b_1}}$ I distributed as Student's t with (n – 2) degrees of freedom.

Tests of the population regression slope:

$H_0: \beta_1 = \beta_1^* \quad H_1: \beta_1 > \beta_1^* \quad \rightarrow \quad Reject\ H_0 if\ \frac{b_1 - \beta_1^*}{s_{b_1}} \geq t_{n-2,\alpha}$

$H_0: \beta_1 = \beta_1^* \quad H_1: \beta_1 < \beta_1^* \quad \rightarrow \quad Reject\ H_0 if\ \frac{b_1 - \beta_1^*}{s_{b_1}} \leq -t_{n-2,\alpha}$

$H_0: \beta_1 = \beta_1^* \quad H_1: \beta_1 \neq \beta_1^* \quad \rightarrow \quad Reject\ H_0 if\ \frac{b_1 - \beta_1^*}{s_{b_1}} \geq t_{n-2,\alpha/2}$ or $\frac{b_1 - \beta_1^*}{s_{b_1}} \leq -t_{n-2,\alpha/2}$

Confidence interval for the population regression slope β_1 :
if the regression errors, ε_i , are normally distributed or if the distribution of b_1 is approximately normal and the standard regression assumptions hold, a $100(1-\alpha)\%$ confidence interval for the population regression slope β_1 is given by:
$b_1 - t_{(n-2,\alpha/2)} s_{b_1} < \beta_1 < b_1 + t_{(n-2,\alpha/2)} s_{b_1}$
where $t_{n-2,\alpha/2}$ is the number for which: $P(t_{n-2} > t_{n-2,\alpha/2}) = \alpha/2$ and the random variable t_{n-2} follows a Student's t distribution with (n – 2) degrees of freedom.

Hypothesis test for population slope coefficient using the F distribution:

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

F test for simple regression coefficient:

$H_0: \beta_1 = 0 \quad H_1: \beta_1 \neq 0$

The F statistic: $F = \frac{MSR}{MSE} = \frac{SSR}{s_e^2}$

The decision rule is as follows: $Reject\ H_0\ if\ F \geqslant F_{1, n-2, \alpha}$

11.6 Prediction

Regression models can be used to compute predictions or forecasts for the dependent variable, given an assumed future value for the independent variable.

Forecast prediction intervals and confidence intervals for predictions:
Suppose that the population regression model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (i = 1, …, n), the standard regression assumptions hold, and the ε_i are normally distributed. Let b_0 and b_1 be the least squares estimates of β_0 and β_1, based on $(x_1, y_1), \ldots, (x_n, y_n)$. Then it can be shown that the following are $100(1-\alpha)\%$ intervals.

1. For the forecast of the single outcome value resulting for Y_{n+1}, the prediction interval is as follows: $\hat{y}_{n+1} \pm t_{n-2, \alpha/2} \sqrt{[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}]} s_e$
2. For the forecast of the mean or conditional expectation $E(Y_{n+1} | X_{n+1})$, the confidence interval for predictions is: $\hat{y}_{n+1} \pm t_{n-2, \alpha/2} \sqrt{[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}]} s_e$

   where: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $\hat{y}_{n+1} = B_0 + b_1 x_{n+1}$

The probability is $1-\alpha$ that this interval includes the true prediction of Y.

- 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 implies that we have information for a wide range of values of this variable, which allows more precise estimates of the population regression line and correspondingly narrower confidence intervals and narrower prediction intervals.
- Larger values of the quantity $(x_{n+1} - \bar{x})^2$ result in wider confidence intervals and wider prediction intervals.

11.7 Correlation analysis

Hypothesis test for correlation:

Let r be the sample correlation coefficient, calculated from a random sample of n pairs of observations from a joint normal distribution. The following tests for zero population correlation use the null hypothesis: $H_0: \rho = 0$

$H_0: \rho = 0 \quad H_1: \rho > 0 \quad \rightarrow \quad Reject\ H_0 if\ \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} > t_{n-2,\alpha}$

$H_0: \rho = 0 \quad H_1: \rho < 0 \quad \rightarrow \quad Reject\ H_0 if\ \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} < -t_{n-2,\alpha}$

$H_0: \rho = 0 \quad H_1: \rho \neq 0 \quad \rightarrow \quad Reject\ H_0 if\ \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} < -t_{n-2,\alpha/2}$ or $\frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} > t_{n-2,\alpha}$

11.8 Beta measure of financial risk

Measures and analysis procedures to help investors measure and control financial risk in the development of investment portfolios.

Diversifier risk is that risk associated with specific firms and industries and includes labor conflicts, new competition, consumer market changes, and many other factors. This risk can be controlled by larger portfolio sizes and by including stocks whose returns have negative correlations.

Diversification risk is that risk associated with the entire economy. The overall effect is measured by the average return on stocks. The effect on individual firms is measured by the beta coefficient.

The beta coefficient for a specific firm is the slope coefficient. This slope coefficient indicates how responsive the returns for a particular firm are to the overall market returns.

If the firm's returns follow the market exactly, then the beta coefficient will be 1.
If the firm's returns are more responsive to the market, the beta would be greater than 1.
Less responsive to the market, beta will be less than 1.

The required return on an investment =
$(Risk-free\ rate) + [(beta\ for\ investment) \times ((Market\ return) - (risk-free\ rata))]$

The higher beta value, the higher required return on investment. This higher required return would adjust for the fact that the stock return is influenced more heavily by the diversification market risk.

11.9 Graphical analysis

Graphical analysis is used to show the effect on regression analysis of points that have extreme X values and points that have Y values that deviate considerably from the least squares regression equation.

Extreme points are points that have X values that deviate substantially from the X values for the other points.

The leverage for a point is defined as: $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

Outlier points are those that deviate substantially in the Y direction from the predicted value.

$e_{is} = \frac{e_i}{s_e\sqrt{1-h_i}}$

Published on *WorldSupporte*r (www.worldsupporter.org)

How to conduct a multiple variable regression analysis? - Chapter 12

Simple regression (see chapter 11) can predict a dependent variable as a function of a single independent variable. But often there are multiple variables at play. In order to determine the simultaneous effect of multiple independent variables on a dependent variable, multiple regression is used. The least squares principle fit the model.

12.1 The model

As with simple regression, the first step in the model development is model specification, the selection of the model variables and functional form of the model. This is influenced by the model objectives, namely: (1) predicting the dependent variable, and/or (2) estimating the marginal effect of each independent variable. The second objective is hard to achieve, however, in a model with multiple independent variables, because these variables are not only related to the dependent variable but also to each other. This leaves a web of effects that is not easily untangled.

To make multiple regression models more accurate an error term "ε" is added, as a way to recognize that none of the described relationships in the model will hold exactly and there are likely to be variables that affect the dependent variable, but are not included in the model.

12.2 Estimating Coefficients

Multiple regression coefficients are calculated with the least squares procedure. However, again this is more complicated than with simple regression, as the independent variables not only affect the dependent variable but also each other. It is not possible to identify the unique effect of each independent variable on the dependent variable. This means that the higher the correlations between two or more of the independent variables in a model are, the less reliable the estimated regression coefficients are.

There are 5 assumptions to standard multiple regression. The first 4 are the same as are made for simple regression (see chapter 11). The 5 th states that it is not possible to find a set of nonzero numbers such that the sum of the coefficients equals 0. This assumption excludes the cases in which there is a linear relationship between a pair of independent variables. In most cases this assumption will not be violated if the model is properly specified.

Whereas in simple regression the least squares procedure finds a line that best represents the set of points in space, multiple regression finds a plane that best represents these points (as each variable is represented with its own dimension).

It is important to be aware of the fact that in a multiple regression it is not possible to know which independent variable predicts which change in the dependent variable. After all, the slope coefficient estimated is affected by the correlations between all independent and dependent variables. This also means that any multiple regression coefficient is dependent on all independent variables in the model. These coefficients are thus referred to as conditional coefficients. This is the case is all multiple regression models unless there are two independent variables with a sample correlation of zero (but this is very unlikely). Because of this effect highly correlated independent variables should be avoided if possible, so as to minimize the influence on the estimated coefficients. This is also the reason that proper model specification, based on an adequate understanding of the problem context and theory, is crucial in multiple regression models.

12.3 Inferences with Multiple Regression Equations

Multiple regression isn't exact, and the variability of the dependent variable is only in part explained by the linear function of the independent variables. Therefore often a measure is used to show the proportion of said variability that can be explained by the multiple regression model: Mean square regression (MSR). This measure needs to be adjusted for the number of independent variable. It is

calculated as follows:

$$MSR = \frac{\text{Sum } of\ squares\ regression}{number\ of\ independent\ variables}$$

In multiple regression the sum-of squares decomposition is performed as follows:

Sum of squares total = sum of squares regression + sum of squares error

Which can be interpreted as:

Total sample variability = explained variability + unexplained variability

As with simple regression the SSE can be used to calculate the *estimated variance of population model errors*, which is used for statistical inference.

Another useful measure is R^2, the coefficient of determination. R^2 can be used to describe the strength of the linear relationship between the independent variables and the dependent variable.

The equation is as follows:

$$R^2 = 1 - \frac{SSE}{SST}$$

Be aware that if you want to used R^2 to compare regression models, you can only do so if they have the same set of sample observations of the dependent variable.

Using R^2 as an overall measure of the quality of a fitted equation, has one potential problem. Namely, the SSR (explained sum of squares) will increase as more independent variables are added to the model, even if these added variables are not important predictors. This increase in SSR leads to a misleading increase in R^2.

This problem can be avoided by calculating the adjusted coefficient of determination:

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

In which K stands for the number of independent variables.

Lastly, R, the coefficient of multiple correlation, is the correlation between the observed and predicted value of the dependent variable. This is equal to the root of the multiple coefficient:

$R = r(\hat{y}, y) = \sqrt{R^2}$

12.4 Confidence Intervals & Hypothesis Tests for Coefficients

Both confidence intervals and hypothesis tests for estimated regression coefficients in multiple regression models depend on the variance of the coefficients and the probability distribution of the coefficient.

In a multiple regression model the dependent variable has the same normal distribution and variance as the error term, ε. This means that the regression coefficients have a normal distribution, and their variance can be derived from the linear relationship between the dependent variable and the regression coefficients. Again this involves the same calculation as with simple regression, but more complicated.

The error term is made up of a large number of components with random effects, which is why it can generally be assumed that it is normally distributed. Interestingly, because of the central limit

theorem, the coefficient estimates are generally normally distributed even if ε is not, meaning that the use of ε does not affect the developed hypothesis tests and confidence intervals.

The problematic factor that remains with multiple regression models is that the multitude of relations often leads to interpretation errors. Mainly the correlations between the independent variables influence the confidence intervals as well as the hypothesis tests, and increase the variance of the coefficient estimators. This variance is thus conditional on the entire set of independent variables in the model.

In order to get a good coefficient estimate there should be, if possible: (1) a wide range for the independent variables, (2) independent variables that have low correlations, and (3) a model that is close to all data points. It is not always possible to make such choices, but by being aware of these effects good judgments can be made about the applicability of available models.

Other effects of the multitude of independent variables are:

- An increase in the correlation between the independent variables causes the variance of the coefficient estimators to increase. This has to do with the fact that is becomes more difficult to separate the individual effects of the independent variables on the dependent variable.
- An increase in the number of independent variables makes the algebraic structure of the model more complex. The importance of the influences on the coefficient variance remains the same.

A coefficient variance estimator is shown as s^2_b , and calculated as follows:

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

The square root of a variance estimator is known as the coefficient standard error.

Hypothesis tests for regression coefficients are developed using the coefficient variate estimates. In a multiple regression model the hypothesis test H0: $\beta_j=0$ is used most, as it can determine whether a specific independent variable is conditionally important in the model. Given the other variables in the model a conclusion can be drawn immediately in this case, using the printed Student's t-statistic or the p-value.

It is important to note that hypothesis tests are only valid when only the particular set of variables included in the regression model are used, as these are the variables the tests are based on. Including additional predictor variables makes the tests non-valid.

12.5 Testing Regression Coefficients

It can also occur that the focus of interest lies on the effect of the combination of several variables. This is then calculated as follows:

1. Hypothesis tests are presented to determine if sets of coefficients are simultaneously equal to zero. If this hypothesis is accepted this means that none of the independent variables in the model are statistically significant (aka none provide useful information). This hypothesis is almost always rejected in an applied regression situation. This hypothesis is tested using "the partitioning of variability" (SST=SSR+SSE; see 12.3). Using this a critical value F is calculated, then compared with the F in table 9 (Appendix A) at significance level α. If the calculated value is larger than the value in the table, the null hypothesis can be rejected. This leads to the conclusion that at least one coefficient is not equal to zero.
2. Next a hypothesis test is developed for the subset of regression parameters that bode looking into. This test can be sued to determine whether the combined effect of several independent variable is significant within the regression model. The test is conducted by comparing the SSE from the complete regression model to the SSE(R) from a restricted model that includes only the selected independent variables. If the calculated F is larger than the critical value of F, then the null hypothesis can be rejected and it can be concluded that the variables should be included in the model.

It is also possible to test the hypothesis that a single independent variable, given the other independent variables in the model, does not improve on the prediction of the dependent variable, using the same method of calculation as in step 2 above. This can also be done with a Student's t test, which will yield the same conclusion as a F-test.

12.6 Predicting the Dependent Variable

An important feature of a regression model is to predict the value for the dependent variable, given the values for the independent variables. These forecasts can be calculated using the coefficient estimates. Besides the predicted value itself it is also desirable to have a confidence interval (expected value with probability 1-α) or a prediction interval (expected values + ε). To calculate these intervals the estimates of the standard deviations for the expected values, and the individual points need to be calculated. These calculations are, again, the same as used in simple regression, but more complicated. For this reason the calculations are only done using statistical software.

12.7. Non-linear Models

Regression models always assume a linear relationship, but sometimes a nonlinear relationship needs to be analysed. Luckily there are ways to transform regression models so they can be used for broader applications. This can be done because the assumptions about the independent variables within multiple regression are very loose. Non-linear models that can be used are as follows:

- Quadratic models: This is the more simple non-linear equation. To estimate the coefficients in a quadratic model, however, the variables need to be transformed back into a linear model. This can be done simply as follows:
  Quadratic function: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$
  Transformation: $x_1 = z_1$ & $x_1^2 = z_2$
  Linear function: $y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \varepsilon$
  Transforming the variables means the model can be estimated as a linear multiple regression model, but the results can be used as from a nonlinear model. For adequate interpretation, however, the linear and quadratic coefficients need to be combined.
- Logarithmic models: Exponential functions have constant elasticity, and are most widely used in the analyses of market behaviour. The transformations and calculations associated with this model are more complicated, but are included in any quality statistical software.

12.8 Dummy Variables

Thus far the independent variables have been assumed to exist over a range. But it is also possible for a variable to be categorical. In this case the variable only takes two values: x=0 and x=1 (it is of course possible to take more values, according to the number of categories in the categorical variable). This structure is called a dummy variable, or an indicator variable.

Introducing a dummy variable into a multiple regression model shifts the linear relationship between the dependent variable and the other independent variables by the coefficient β of that dummy variable.

When there is a shift of the linear function by identifiable categorical factors, a dummy variable with values of 1 and 0 can estimate this shift effect. A dummy variable can also be used to model and test for differences in the slope coefficient, by adding an interaction variable.

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

Several Aspects of Regression Analysis – Chapter 13

This chapter focuses on topics that add to the understanding of regression analysis. This includes alternative specifications for these models, and what happens in the situations where basic regression assumptions are violated.

13.1 Developing models

The goal when developing a model is to approximate the complex reality as close as possible with a relatively simple model, which can then be used to provide insight into reality. It is impossible to represent all of the influences in the real situation in a model, instead only the most influential variables are selected.

Building a statistical model has 4 stages:

1. Model Specification: This step involves the selection of the variables (dependent and independent), the algebraic form of the model, and the required data. In order to this correctly it is important to understand the underlying theory and context for the model. This stage may require serious study and analysis. This step is crucial to the integrity of the model.
2. Coefficient Estimation: This step involves using the available data to estimate the coefficients and/or parameters in the model. The desired values are dependent on the objective of the model.
   Roughly there are two goals:
   1. Predicting the mean of the dependent variable: In this case it is desirable to have a small standard error of the estimate, s e . The correlations between independent variables need to be steady, and there needs to be a wide spread for these independent variables (as this means that the prediction variance is small).
   2. Estimating one or more coefficients: In this case a number of problems arise, as there is always a trade-off between estimator bias and variance, within which a proper balance must be found. Including an independent variable that is highly correlated with other independent variables decreases bias but increases variance. Excluding the variable decreases variance, but increases bias. This is the case because both these correlations and the spread of the independent variables influence the standard deviation of the slope coefficients, s_b.
3. Model Verification: This step involves checking whether the model is still accurate in its portrayal of reality. This is important because simplifications and assumptions are often made while constructing the model, this can lead to the model becoming (too) inaccurate). It is important to examine the regression assumptions, the model specification, and the selected data. If something is wrong here, we return to step 1.
4. Interpretation and Inference: This step involves drawing conclusions from the outcomes of the model. Here it is important to remain critical. Inferences drawn from these outcomes can only be accurate if the previous 3 steps have been completed properly. If these outcomes differ from expectations or previous findings you must be critical about whether this is due to the model or whether you really have found something new.

13.2 Further Application of Dummy Variables

Dummy variables were introduced in chapter 12 as a way to include categorical variables in regression analysis. Further uses for these variables will be discussed here.

Dummy variables have values of either 1 or 0, to represent two categories. It is also possible to represent more than two categories by using a combination of multiple dummy variables. The rule is: number of categories -1 = number of dummy variables. So for three categories, two dummy variables are used. For example:

Yes:	$x_1 = 1$	$x_2 = 1$
Maybe:	$x_1 = 1$	$x_2 = 0$
No:	$x_1 = 0$	$x_2 = 0$

Time series data can be portrayed by dummy variables in the same manner. In this case time periods are the categories.

Dummy variables are becoming more popular as a tool in experimental designs. Here again similar specification is used to represent several levels of the treatment. In experimental designs there are also so-called blocking variables, which are part of the environment and cannot be randomized or preselected. By using dummy variables these can be included in the model in such a manner that it's variability can be removed from the independent variables.

13.3 Values in Time-Series Data

When measurements are taken over time the values of the dependent variable are referred to as lagged. Such time-series observations are specified in formulates with the subscript "t".

It is important to be aware of lagged values because the value of the dependent variable in one time period is often related to the value of a previous time period. This previous value of the dependent variable is called a lagged dependent variable.

With lagged dependent variables there is no difference in coefficient estimation, confidence intervals and hypothesis tests in comparison to a regular dependent variable. It is, however, advised to be cautious with the use of confidence intervals and hypothesis tests, as it is possible that the equation errors, ε_i, are no longer independent, which leads to the coefficient estimates no longer being efficient (though unbiased), which means that confidence intervals and hypothesis tests are no longer valid. If the equation errors remain independent, the quality of the approximation will improve as the number of sample observations increases.

13.4 Inclusion of proper independent variables

Models can never be as complete as reality because a model cannot contain al variables that are likely to affect the dependent variable in the real world, so a selection must be made. The joint influence of the variables that are not selected are then absorbed in the error term. If, however, an important variable is omitted this means that the estimated coefficients of the other independent variables will be different, and any conclusions drawn from this model may be faulty.

13.5 Multicollinearity of independent variables

It is possible for two independent variables to be highly correlated. In this case the estimated coefficients of the model can be very misleading. This phenomenon is referred to as multicollinearity. This problem arises out of the data itself, and sadly there is little that can be done about it. It is still important to be watchful of this though. Indications for multicollinearity are:

- The regression coefficients are very different from previous research or expectations.
- Coefficients of variables that are believed to have a strong influence, have a small students t-statistic.
- The student's t-statistics are small for all coefficients, even though there is a large F-statistic (indicating no individual effect but strong effect for the total model).
- There are high correlations between individual independent variables and/or there is a strong linear regression relationship between independent variables.

There are several possibilities to correct multicollinearity:

- Removing one or more of the highly correlated independent variables (this may also have side effects, see 13.4).
- Changing the specification of the model by including a new variable that is a function of the correlated variables.
- Obtaining additional data where the correlation between the independent variables is weaker.

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

13.6 Variance of error terms

When one or more of the regression assumptions are violated the least squared method will lead to inefficient estimated coefficients and misleading conclusions. One of these assumptions is the assumption of homoscedasticity, that the error terms are uniformly distributed and are not correlated.

This assumptions is violated if a model exhibits heteroscedasticity. There are various ways to check for this:

- Relating the error variance to an alternative explanation.
- Making a scatterplot of the residuals versus the independent variables and the predicted values from the regression. A visible relationship (like errors increasing with increasing X-values) is a sign of heteroscedasticity.
- Testing the null hypothesis that the error terms have the same variance, against the alternative hypothesis that their variances depend on the expected values. This procedure can be used when the predicted value of the dependent variable has a linear relationship with the variance of the error term.

It is also possible that there is simply an appearance of heteroscedasticity, for example if logarithmic model is more appropriate but a linear regression model was estimated instead.

13.7 Correlated error terms

The error term represents all variables that influence the dependent variable, outside of the independent variables. In time-series data this term functions differently, as many of these variables may behave similarly over time. This can thus result in a correlation between error terms, also referred to as auto correlated errors. This means that the estimated standard errors for the coefficients are biased, null hypotheses might be falsely rejected, and confidence intervals would be too narrow.

Assuming all errors have the same variance, the structure for autocorrelation, or the first-order autoregressive model of auto correlated behaviour, is:

$\varepsilon_t = \rho\ \varepsilon_{t-1} + u_t$

Where ρ is a correlation coefficient, and u t is a random variable (thus not auto correlated). The coefficient ρ varies from -1 to +1, where a ρ of 0 signifies no autocorrelation, while a -1 or +1 signifies strong autocorrelation.

Autocorrelation can also be found by time plotting the residuals, a jagged plot signifies no autocorrelation.

A more formal test of autocorrelation is the Durbin-Watson test, based on model residuals. It is calculated as follows:

$$d = \frac{\sum_{t-2}^{n} (e_t - e_{t-1})^2}{\sum_{t-1}^{n} e_{t^2}}$$

The Durbin-Watson statistic can be written as: d=2(1-r). Where r is the sample estimate of the population correlation between adjacent errors.

If the errors are not auto correlated, r≈0 and d≈2.

A positive correlation is shown as: 0≤d<2.

A negative correlation is shown as: $2<d\leq4$.

When there are auto correlated errors then the regression procedure needs to be modified to remove its effects. Estimating the coefficients of such a model follows two steps:

- The model is estimated using least squares, which obtains the Durbin-Watson d-statistic. The r, as autocorrelation parameter, can then be calculated.
- Use least squares to estimate a second regression with:
  - Dependent variable: $y_t \square (y_t - ry_{t-1})$
  - Independent variable: $\beta_{1x1t} \square (x_{i1} - rx_{1,t-1})$
- Divide the estimated intercept from this second model by (1-r) to get the correct estimated intercept for the original model.
- Use the output from the second model to carry out hypothesis tests and confidence intervals.

An even more severe problem presents itself when there is a model with lagged dependent variables and auto correlated errors. Here the model also needs to be modified, using a variation on the procedure explained above.

Introduction to nonparametric statistics – Chapter 14

14.1 Goodness-of-fit tests: specified probabilities

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

Chi-square random variable
A random sample of n observations, each of which can be classified into exactly one of K categories, is selected. Suppose the observed numbers in each category are $O_1, O_2, \ldots, O_K$. If a null hypothesis (H0) specifies probabilities $P_1, P_2, \ldots, P_K$ for an observation falling into each of these categories, the expected numbers in the categories, under H0, would be as follows: $E_i = nP_i$ for $i = 1,2,\ldots,K$
If the null hypothesis is true and the sample size is large enough that the expected values are at least 5, then the random variable associated with $X^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$ is known as a chi-square random variable, and has, to a good approximation, a chi-square distribution with (K – 1) degrees of freedom.

A goodness-of-fit test with specified probabilities, of significance level , of H0 against the alternative that the specified probabilities are not correct is based on the decision rule:

$$Reject\ H_0\ if \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} > X^2_{K-1,\alpha}$$

The random variable X^2_{K-1} follows a chi-square distribution with K – 1 degrees of freedom.

14.2 Goodness-of-fit tests: population parameters unknown

Suppose that a null hypothesis specifies category probabilities that depend on the estimation of m unknown population parameters. The appropriate goodness-of-fit test with estimated population parameters is precisely as before, except that the number of degrees of freedom for the chi-square random variable is (K – m – 1) where K is the number of categories and m is the number of unknown population parameters.

Test for the normal distribution:

Jarque-Bera test for normality: Suppose that we have a random sample $x_1, x_2, \ldots, x_n$ of n observations from a population. The test statistic for the Jarque-Bera test for normality is

$$JB = n\left[\frac{(skewness)^2}{6} + \frac{(kurtosis - 3)^2}{24}\right]$$

with $skewness = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{ns^3}$ and $kurtosis = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{ns^4}$

If the number of sample observations becomes very large, this statistic has a chi-square distribution with 2 degrees of freedom.

Unfortunately, the chi-square approximation to the distribution of the Jarque-Bera test statistic is close only for very large sample sizes.

14.3 Contingency tables

Suppose that a sample of n observations is cross-classified according to two characteristics in an r x c contingency table. Denote by O_{ij} the number of observation in the cell that is in the i'th row and the j'th column. If the null hypothesis is H0: No association exists between the two characteristics in the population, then the estimated expected number of observations in each cell under H0 is:

$$E_{ij}=\frac{R_i C_j}{n}$$

where Ri and Cj are the corresponding row and column totals. A test of association at a significance level α is based on the following decision rule:

$$reject\ H_0\ als: \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(O_{ij}-E_{ij})^2}{E_{ij}} > x^2_{(r-1)(c-1),\alpha}$$

14.4 Nonparametric tests for paired or matched samples

The simplest nonparametric test to carry out is the sign test. 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. The sign test is used to test:
$H_0: P=0.5$

The p-value for a sign test is found using the binomial distribution with n = number of nonzero differences, S = number of positive differences and P = 0.5.

a. Upper-tail test: $H_1: P>0.5 \qquad p-value=P(x\geqslant S)$
b. Lower-tail test: $H_1: P<0.5 \qquad p-value=P(x\leqslant S)$
c. Two-tail test: $H_1: P\neq 0.5 \qquad p-value=2P(x\geqslant S)$

A disadvantage of the sign test is that it takes into account only a very limited amount of information; the signs of the differences.

The *Wilcoxon Signed rank test* provides a method for incorporating information about the magnitude of the differences between matched pairs.

- Discard pairs for which the difference is zero
- Rank the remaining n absolute differences in ascending order
- Reject H0 if $T\leqslant T\ Appendix\ tabel\ 10$ met $T=min(T_+, T_-)$
  Where: n= number of nonzero differences, T_+ = sum of the positive ranks T_- = sum of the negative ranks

The *sign test: normal approximation* (large samples)
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.

Mean: $\mu=np=0.5n$

Standard deviation: $\sigma=\sqrt{np(1-p)}=\sqrt{0.25n}=0.5\sqrt{n}$

The test statistic: $Z=\frac{S*-\mu}{\sigma}=\frac{S*-0.5n}{0.5\sqrt{n}}$

a. Two-tail test: $S*=S+0.5 \; if \; S<\mu$ or $S*=S-0.5 \; if \; S>\mu$
b. Upper-tail test: $S*=S-0.5$
c. Lower-tail test: $S*=S+0.5$

The *Wilcoxon Signed Rank test: normal approximation* (large samples)

Mean: $E(T)=\mu_T=\frac{n(n+1)}{4}$

Variance: $Var(T)=\sigma_T^2=\frac{n(n+1)(2n+1)}{24}$

The test statistic: $Z=\frac{T-\mu_T}{\sigma_T}$

If the number, n, of nonzero differences is large and T is the observed value of the Wilcoxon statistic, then the following test have significance level α .

1. If the alternative hypothesis is one-sided, reject H0 if: $\frac{T-\mu_T}{\sigma_T}<-z_\alpha$
2. If the alternative hypothesis is two-sided, reject H0 if: $\frac{T-\mu_T}{\sigma_T}<-z_{\alpha/2}$

The sign test can also be used to test hypotheses about the central location (median) of a population distribution.

14.5 Nonparametric tests for independent random samples

Two tests that compare the central locations of two population distributions when independent random samples are taken from the two populations.

Mann-Whitney U test
Approaches the normal distribution quite rapidly as the number of sample observations increases. The approximation is adequate if each sample contains at least 10 observations.

$$U=n_1n_2+\frac{n_1(N_1+1)}{2}-R_1$$

R_1 denotes the sum of the ranks of the observations from the first population.

Mean: $E(U)=\mu_U=\frac{n_1n_2}{2}$

Variance: $Var(U)=\sigma_U^2=\frac{n_1n_2(n_1+n_2+1)}{12}$

The distribution of the random variable $Z=\frac{U-\mu_U}{\sigma_U}$ is approximated by the normal distribution.

Example:

Number of hours per week spent studying for Finance and Accountings:

Finance	10	6	8	10	12	13	11	9	5	11		
Accounting	13	17	14	12	10	9	15	16	11	8	9	7

Mann-Whitney U test Ranks for hours of study per week:

Finance	(Rank)	Accounting	(Rank)
10	(10.0)	13	(17.5)
6	(2.0)	17	(22.0)
8	(4.5)	14	(19.0)
10	(10.0)	12	(15.5)
12	(15.5)	10	(10.0)
13	(17.5)	9	(7.0)
11	(13.0)	15	(20.0)
9	(7.0)	16	(21.0)
5	(1.0)	11	(13.0)
11	(13.0)	8	(4.5)
		9	(7.0)
		7	(3.0)
	Rank sum = 93.5		Rank sum = 159.5

n1 = 10, n2 = 12, R1= 93.5

$$U=(10)(12)+\frac{(10)(11)}{2}-93.5=81.5$$

$$E(U)=\frac{(10)(12)}{2}=60$$

$$Var(U)=\sigma_U^2=\frac{(10)(12)(23)}{12}=230$$

$Z=\frac{81.5-60}{\sqrt{230}}=1.42$ and p-value = 0.1556

With the usual 0.05 significance level, the test result is not sufficient to conclude that students spend more time studying for one of these subjects than the other.

Wilcoxon rank total statistics T

Mean: $E(T)=\mu_T=\frac{n_1(n_1+n_2+1)}{2}$

Variance: $\sigma_T^2=\frac{n_1 n_2(n_1+n_2+1)}{12}$ Random variable $Z=\frac{T-\mu_T}{\sigma_T}$

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

14.6 Spearman rank correlation

Suppose that a random sample $(x_1, y_1), \ldots, (x_n, y_n)$ of n pairs of observations is taken. 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.

If there are no tied X or Y ranks, an equivalent formula for computing this coefficient is:

$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$ where the d_i are the differences of the ranked pairs.

1. Test against alternative of positive association: $Reject\ H_0 if \quad r_s > r_{s,\alpha}$
2. Test against alternative of negative association: $Reject\ H_0 if \quad r_s < -r_{s,\alpha}$
3. Test against two-sided alternative: $Reject\ H_0 if \quad r_s < -r_{s,\alpha}$ of $r_s > r_{s,\alpha}$

14.7 A nonparametric test for randomness

Runs test: Small sample size.
R is the number of runs in the sequence of n observations with $n \leqslant 20$. The null hypothesis is that the series is a set of random variables.
Appendix table14 gives the smallest significance level at which this null hypothesis can be rejected against the alternative of positive association between adjacent observations, as a function of R and n.

If the alternative is the two-sided hypothesis on non randomness, the significance level must be doubled if it is less than 0.5. If the significance level read from the table is greater than 0.5, the appropriate significance level for the test against the two-sided alternative is $2(1-\alpha)$.

Runs test: Large sample size
Given that we have a time series with n observations and n > 20, define the number of runs, R, as the number of sequences above or below the median.

$H_0: the\ series\ is\ random$
The distribution of the number of runs under the null hypothesis can be approximated by a normal distribution. Under the null hypothesis, $Z = \frac{R - \frac{n}{2} - 1}{\sqrt{\frac{n^2-2n}{2(n-1)}}}$ has a standard normal distribution.

1. If the alternative hypothesis is positive association between adjacent observations:
$$Reject\ H_0 if\ Z = \frac{R - \frac{n}{2} - 1}{\sqrt{\frac{n^2-2n}{4(n-1)}}} < -z_\alpha$$
2. Two-sided hypothesis of nonrandomness:
$$Reject\ H_0 if\ Z = \frac{R - \frac{n}{2} - 1}{\sqrt{\frac{n^2-2n}{4(n-1)}}} < -z_{\alpha/2} \quad \text{or} \quad Z = \frac{R - \frac{n}{2} - 1}{\sqrt{\frac{n^2-2n}{4(n-1)}}} > z_{\alpha/2}$$

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

How to analyse variance? - Chapter 15

There are situations and experiments that require processes to be compared at more than two levels. Data from such experiments can be analysed using analysis of variance or ANOVA.

15.1 Comparing Population Means

There are other ways to compare population means than ANOVA, but these are based on the assumption of either paired observations or independent random samples, and can only be used to compare two population means. ANOVA can be used to compare more than two populations, and also uses assessments of variation, which forms a large problem in other methods.

15.2 One-Way ANOVA

The procedure for testing the equality of population means is called a one-way ANOVA. This procedure is based on the assumption that all included populations have a common variance.

The total sum of squares (SST) in this procedure is made up of a within-group sum of squares (SSW) and a between groups sum of squares (SSG): SST = SSW + SSG
This division of the SST forms the basis of the one-way ANOVA, as it expresses the total variability around the mean for the sample observations.

If the null hypothesis is true (all population means are the same) then both SSW and SSG can be used to estimate the common population variance. This is done by dividing by the appropriate number of degrees of freedom.

Because SSW and SSG both provide an unbiased estimate of the common population variance if the null hypothesis is true, a difference between the two values indicates that the null hypothesis is false. The test of the null hypothesis is thus based on the ratio of mean squares:

$$F = \frac{MSG}{MSW}$$

Where $MSW = \frac{SSW}{n-K}$ and $MSG = \frac{SSG}{K-1}$

With the assumptions that the population variances are equal and the population distributions are normal.
The closer the ratio is to 1, the less indication there is that the null hypothesis is false.

These results are also summarized in a one-way ANOVA table, which has the following format:

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Squares**	**F-ratio**
Between groups	SSG	K – 1	MSG	MSG/MSW
Within groups	SSW	n – K	MSW	
Total	SST	N – 1		

It is also possible to calculate a minimum significant difference (MSD) between two sample means, as evidence to conclude whether the population means are different. This is done:

$$MSD(K) = Q\frac{S_p}{\sqrt{n}}$$

With s_p being the estimate of variance $s_p = \sqrt{MSW}$, n the number of observations, K the number p of populations, and Q being a factor.

15.3 Kruskal-Wallis test

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. Like the majority of nonparametric tests this test is based on the ranks of the sample observations. In this test the null hypothesis is based on the calculation:

$$W = \frac{12}{n(n+1)} \frac{\sum_{i=1}^{k} R_{i^2}}{n_i} - 3(n+1)$$

Where R are the ranks for the sample observations. The hypothesis is rejected if W is larger than $X^2_{k-1,\alpha}$ (a number with probabilityα, by a random χ^2 variable, with (K-1) degrees of freedom.

15.4 Two-way ANOVA

If there is a situation where a second factor also influences the outcome, it is best to design the experiment in such a manner that the influence of this factor can also be taken into account. This additional variable is then called a blocking variable and this design is called a randomized block design, the outcomes of which can be analysed using a two-way ANOVA.

In a randomized block design because the several categories from the two independent variables are randomly combined.
Using the observation for the i^{th} group and the j^{th} block, the population model can be portrayed as following: $X_{ij} = \mu + G_i + B_j + \varepsilon_{ij}$.

Here X_{ij} is the random variable, μ is the overall mean, the parameter G i measures the discrepancy between the mean of group i and μ, the parameter B j measures the discrepancy between the mean of block j and μ, and ε_{ij} represents the experimental error.

In a two-way ANOVA the SST is split up in the between-blocks sum of squares (SSB) and the between-groups sum of squares (SSG), and also contains the error sum of squares (SSE). It is thus split up as: SST = SSB + SSG + SSE.

The null hypothesis of the population group means being equal is then tested through the ratio of the mean square for groups to the mean square error: $F = \frac{MSG}{MSE}$

The results of a two-way ANOVA are also best summarized in a two-way ANOVA table. This has the same set-up as a one-way ANOVA table, except for the sources of variation (between groups, between blocks, error, and total).

15.5 Two-Way ANOVA with multiple observations per cell

It is also possible to have more than one observation per cell. This has two advantages:

1. More sample data leads to more precise estimates meaning that the differences among the population means can be distinguished better.
2. The interaction between groups and blocks, as a source of variability, can be isolated.

This model thus has three null hypothesis: no difference between group means, no difference between block means, and no group-block interaction.

In this model the SST consists of one more factor: the interaction sum of squares (SSI), corresponding with the extra source of variation: Interaction.

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

How to calculate predictions with the use of Time-Series Data? - Chapter 16

Time series data involves measurements that are ordered over time, in which the sequence of observations is important. Most procedures for data analysis cannot be used for this data, as these procedures are based on the assumption that the errors are independent. Thus, different forms of analysis are needed.

The main goal of analysing time-series data is to make predictions. An important assumption here is that the relations between variables remain constant.

16.1 Time-Series Components

Most time-series have the following four components:

1. Trend component: Values grow or decrease steadily over long periods of time.
2. Seasonality component: An oscillatory patterns that is specific per season (quarter year) repeats itself.
3. Cyclical component: And oscillatory or cyclical pattern that is not related to seasonal behaviour.
4. Irregular component: No pattern is regular enough to only exist through these predictable trends; each series of data will also have irregular components (similar to the random error term).

Analysis of time-series data involves constructing a formal model in which most of these components are explicitly or implicitly present, in order to describe the behaviour of the data series. In building this model the series components can either be regarded as being fixed over time, or as steadily evolving over time.

16.2 Moving Averages

Moving averages are the basis for many practical adjustment procedures. It can be used to remove the irregular component or smooth seasonal component:

- Removing the irregular component: This is done by replacing each observation with the average of itself and its neighbours. The theory is that this will decrease the effect of the irregular component on each data point.
- Smoothing the seasonal component: This is done by producing four-period moving averages in such a manner that the seasonal values become one single seasonal moving average. This does mean that the values have shifted in time (in comparison to the original series), but this can be corrected by centring the averages. The specific procedure always depends on the amount of stability the pattern is assumed to have, and whether seasonality is thought to be additive or multiplicative (in the latter case: use logarithms).
  If there is an assumption of a stable seasonal pattern a further seasonal-adjustment approach can be used: the seasonal index method. Here the original series is expressed as a percentage of the centred 4-point moving average series.

Additionally moving averages are very suitable for detecting cyclical components and/or trends.

16.3 Predictions using smoothing

There are a various prediction methods, and the choice you make should always depend on the resources, the objectives, and the available data.

Simple exponential smoothing is a more basic prediction method that is appropriate when the series is non-seasonal and has no consistent trends. It predicts future values on the basis of an estimate of the current level of the time series. This estimate is comprised of a weighted average of current and

past values, where most weight is given to the most recent observations (with decreasing weight the older the observation is).

The smoothed series is then ^x, with ^x t = (1 – α)^x_{t-1} + αx_t . Where t signifies t the moment in the time series, and α is the smoothing constant. The smoothing constant is a value between 0 and 1 and is different per situation. It is possible to rely on experience or judgment to choose this value, or to try several different values and see which is more successful.

The Holt-Winters exponential smoothing procedure is a more advanced prediction method that allows for trend. It functions just like the simple exponential smoothing procedure, but with the added variable for the trend estimate T t-1 .

An extension of this method also allows for seasonality. This is done by using a set of recursive estimates from the time-series. For this a level factor (α), a trend factor (β) and a multiplicative seasonal factor (γ) are used.

16.4 Predictions using Auto-Regression

The procedure of autoregressive models uses the available time-series data to estimate the parameters of a model of the process that could have generated the time series. This is based on autocorrelation, correlation patterns between adjacent periods. The model that is formed by this is: $x_t = \gamma + \varphi_1 x_{t-1} + \varepsilon_t$. Where γ and φ_1 are fixed parameters. The parameter γ allows for rhe mean of the series x_t to be other than 0. The random variables ε_t have a mean of 0, fixed parameters and are not correlated with each other.

This is called a first-order autoregressive model. It is possible to extend this model by making the current value of the series dependent on the two most recent observations, this is then called a second-order autoregressive model.

16.5 The Box-Jenkins approach

It is good to briefly mention the Box-Jenkins approach to predictions in time-series data. In this procedure one (1) defines a broad class of models for predictions, and then (2) develop a methodology for picking a suitable model on the basis of the characteristics of the available data. This has three general stages:

1. Selecting a specific model that might be appropriate, based on summary statistics.
2. Estimated the unknown coefficients in this model.
3. Applying checks to determine whether the model adequately represents the available data.

This approach is useful due to its flexibility.

A general model class that can be used here is that of autoregressive integrated moving average models (ARIMA models).

Published on *WorldSupporte*r (www.worldsupporter.org)

How to sample a population? - Chapter 17

There are various ways of sampling a population, according to research and analysis goals.

17.1 Stratified Sampling

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. A question that arises here Is how to allocate the sampling effort among the strata. There are various possibilities:

- Proportional allocation: The proportion of the sample from a stratum is the same as the proportion of that stratum to the population. This is used if there is little to nothing known about the population and there are no strong requirements for the production of information.
- Optimal allocation: More sample effort is allocated to strata with a higher population variance. This is used if the objective is to estimate an overall population parameter (such as mean, total, or proportion) as precisely as possible. This method is only optimal with this goal in mind.

Analysing the results of stratified random samples is relatively straightforward, and any stratum sample mean (m_j) can be used as an unbiased estimator of the population mean (μ_j). It can also be sued to estimate the population total, as this is the product of the population mean and the number of population members.

17.2. Other Ways to Sample

Various other sampling methods are:

- Cluster Sampling: This method can be used when a population can be subdivided into small geographical units, or clusters. A simple random sample of clusters is then selected, and each member of these clusters is contacted for data. Using this method very little prior information of the population is needed.
- Two-Phase Sampling: In this method the regular data-collection is preceded by a smaller pilot study, in which a smaller sample is used. This cost more time but allows for methods and procedures to be improved, and can provide some estimations for the true study.
- Non-random sampling: There are two main methods:
  - Non-probabilistic sampling: Sample members are selected by convenience. This often means that the sample is not representative of the population and lacks proper statistical validity.
  - Quota sampling: There are specified numbers of people of certain characteristics (race, age, gender etc.) that are contacted. This usually produces quite accurate estimates of population parameters, but it is not possible to determine the reliability of these estimates, because the sample was not randomly chosen.