UNIVERSITY OF OSLO

FACULTY OF MATHEMATICS AND NATURAL SCIENCES

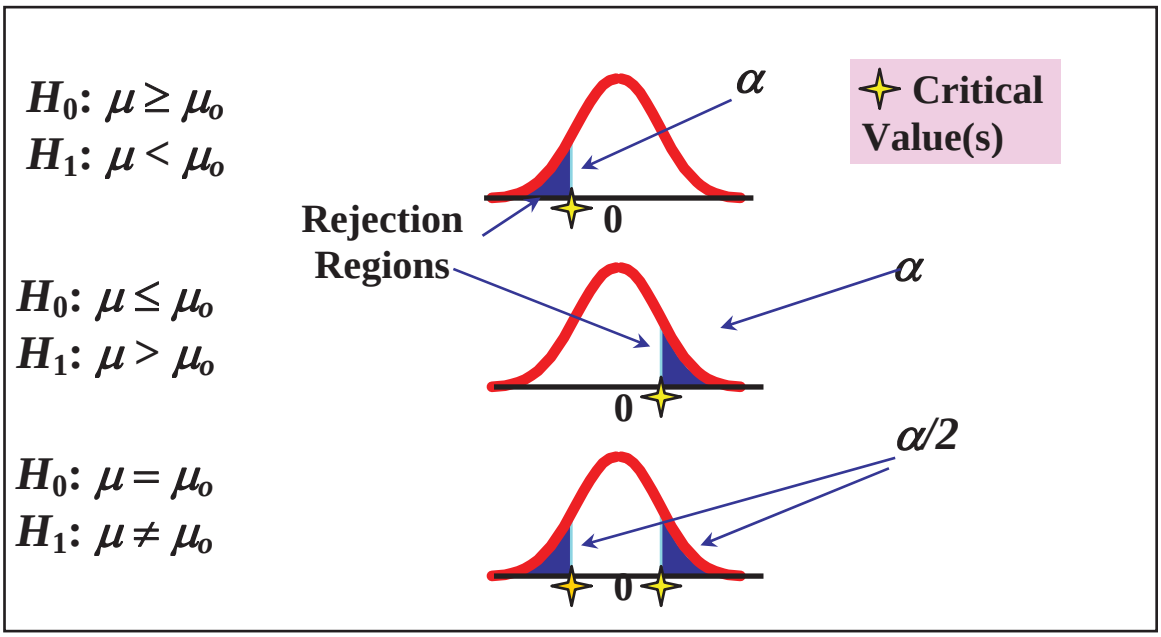
Professor Chongyu Xu
Department of Geosciences
P.O.Box 1047, Blindern
N-0316 Oslo, Norway

Telephone: +47-22 85 58 25
Telefax: +47-22 85 42 15
E-mail: chongyu.xu@geo.uio.no
WWW: http://folk.uio.no/chongyux

STATISTICAL AND STOCHASTIC METHODS IN HYDROLOGY

(Lecture notes)

Chongyu Xu



(2011)

With modifications from Nino Amvrosiadi, 2013
Department of Earth Sciences, Uppsala University
Nino.Amvrosiadi@geo.uu.se

Note:

This lecture notes is used for class discussion. Most materials are abstracted from the course book (Haan, C.T., 2002, Statistics Methods in Hydrology, Iowa State Univ Press, Ames, Iowa), from my experiences and other sources (with no detailed reference). More details can be found from above references.

This course differs from other mathematical statistics/stochastic courses in such a way that we do not pay very much attention in proving the definitions, theories and mathematic laws. Instead, we focus on:

- Where/when to use statistics/stochastic methods in hydrology?
- Which probability distributions and statistical theory/methods are useful/used in solving hydrological problems?
- How to use these modern statistical/stochastic methods/models in problem solving.

Contents pages

Chapter 1

Probability and probability distributions – basic concepts

1. Basic concepts

- **Outcome** - is the result of an experiment.
- **Sample Space** - a list of all the possible outcomes of an experiment.
- **Element** – any particular point in the sample space.
- **Event** - is any collection of outcomes of an experiment. Formally, any subset of the sample space is an event.
- **Frequency**: how many times something happens
- **Relative frequency**: number of times something happens relative to the number of times it could have happened, i.e. the total number of times an experiment is carried out (proportion of times)

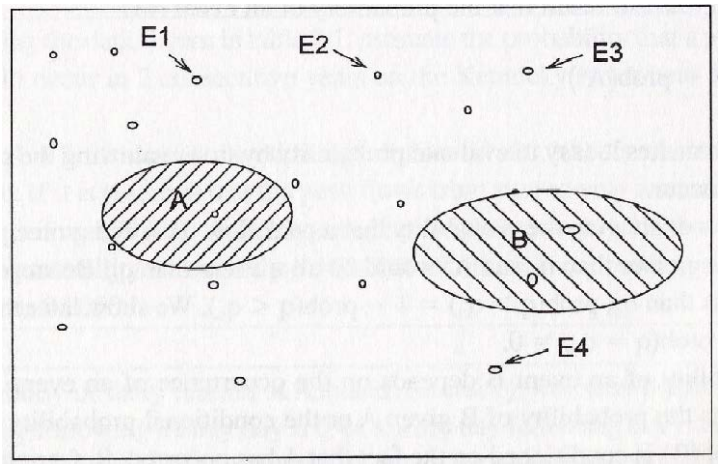


Figure 1.1: Venn diagram illustrating a sample space (whole box), elements (E_i), and events (shaded areas)

2. Probability

- **Definition 1 (Classical definition)**: If a random event can occur in ***n*** **equally likely** and **mutually exclusive** ways, and if $\mathbf{n_a}$ of the ways have an attribute A, then the probability of the occurrence of the event having attribute A is written as

$$P(A) = \frac{\text{number of outcomes corresponding to event A}}{\text{total number of outcomes}} = n_a/n \qquad (1)$$

- **Definition 2: the expected relative frequency of a particular outcome** - The probability of an event has also been defined as its long-run relative frequency. If an experiment is repeated many times without changing the experimental conditions, the relative frequency of any particular event will settle down to some value. The probability of the event can be defined as the limiting value of the relative frequency as $n \to \infty$, Prob(E) => relative frequency (E).

$$P(A) = \lim_{n \to \infty} n_a / n \qquad (2)$$

Example 1.1

A classic example is the probability of getting a head in flipping a coin.

If we know the coin is balanced and not biased toward "heads" or "tails", we can apply the first definition and a single flip is enough. There are 2 possible outcomes – heads or tails – so n is 2. There is one outcome with a head so n_a is 1. The probability is then ½.

If we do not know if the coin is balanced, we cannot use the first definition. We have to test the coin if it is balanced. This is not the case of definition 2 either, since we cannot flip the coin an infinite number of times. We have to resort to a finite sample of flips. Figure 1.2 (from Haan's book) shows how the estimates of the probability of a head changes as the number of trials changes. A trend toward ½ is noted, this is called stochastic convergence towards ½. To answer 'is the coin unbiased'? It seems more trials are needed. This is the saturation of the hydrologist. He/she many times needs more observations but does not have them and cannot get them. The data at hand cannot clearly indicate a single answer. **This is where probability and statistics come to play**.

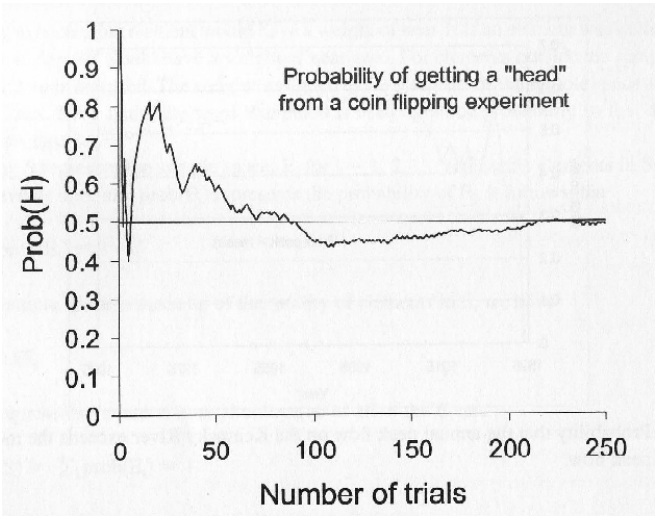


Figure 1.2: Coin flipping experiment

An advantage is that Equation (2) allows to estimate probabilities based on observations and does not require that outcomes be equally likely. This advantage is somewhat offset in that estimates of probability based on observations are empirical and will only stochastically converge to the true probability as the number of observations gets large (in many situations are not very large)

Example 1.2

Table 2.1 in Haan lists the annual peak discharges of Kentucky river. Fig.1.3 shows the probability of an annual peak flow exceeding the mean annual flow as a function of time starting in 1895. Note that each year additional data becomes available to determine both the mean annual flow and the probability of exceeding that value. Here a convergence toward 0.56 is noted yet not assured.

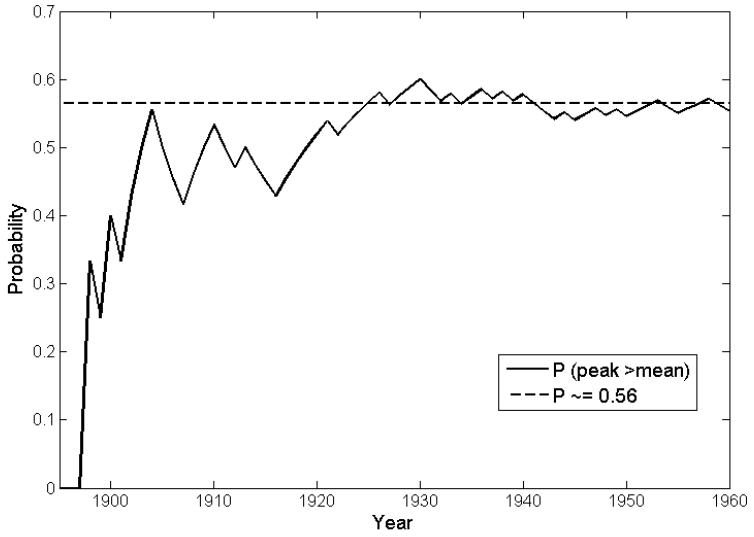


Figure 1.3: Probability that the annual peak flow on the Kentucky River exceeds the mean annual peak flow.

How to make this plot:

- Calculate the average peak discharge Q_{avg}
- For each time step calculate P as:
- P = (number of years up to this time step with $Q>Q_{avg}$) / (total number of years up to this time step)
- Plot P vs. Years

Conditional probability: Let *A* and *B* be events in a random experiment with $P(B) > 0$. For example: A = 'it rains', B= 'it is spring'. The *conditional probability* of *A* to occur, given that *B* occurs is defined to be:

$$P(A \mid B) = P(A \cap B) / P(B)$$

where:

P(A|B) = the (conditional) probability that event A will occur given that event B has occurred already

P(A∩ B) = the (unconditional) probability that event A **and** event B occur

P(B) = the (unconditional) probability that event B occurs

- The **multiplication rule** is a result used to determine the probability that two events, A and B, both occur, follows from the definition of conditional probability.

  The result is often written as follows, using set notation

  $$P(A \cap B) = P(A|B) \cdot P(B) \qquad \text{or} \qquad P(A \cap B) = P(B|A) \cdot P(A)$$

  And is read as: The probability of **both** A and B to occur is the probability of A given that B occurs, multiplied with the probability of B.

- **Independent probability**
  Two events are independent if the occurrence of one of the events gives us no information about whether or not the other event will occur; that is, the events have no influence on each other, thus

  If $P(A \mid B) = P(A)$ then we say A is independent of B, thus we have:

  $$P(A \cap B) = P(A) \cdot P(B)$$

Example 1.3

It has been calculated that the probability of peak flow exceeding a certain limit is 0.05. What is the probability that this will occur in 2 successive years assuming the peak flows from year to year are independent.

Solution:

Let A be event: peak flow exceeds the certain limit in one year

Let B be event: peak flow exceeds the certain limit in the next year

Since A and B are independent events:

$P(A \cap B) = P(A) \cdot P(B) = 0.05 \cdot 0.05 = 0.0025$

- **Mutually exclusive events**
  Two events are mutually exclusive (or disjoint) if it is impossible for them to occur together. Formally, two events A and B are mutually exclusive if and only if

  $$P(A \cap B) = 0$$

- **The addition rule** is a result used to determine the probability that event A **or** event B occurs, **or** both occur.
  The result is often written as follows, using set notation
  The probability of event A **or** event B occurs:

  $$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

We have:

$P(A \cup B) = P(A)+P(B)- P(A \mid B)\cdot P(B)$ if A and B are dependent

$P(A \cup B) = P(A)+P(B)- P(A)\cdot P(B)$ if A and B are independent

$P(A \cup B) = P(A)+P(B)$ if A and B are Mutually exclusive

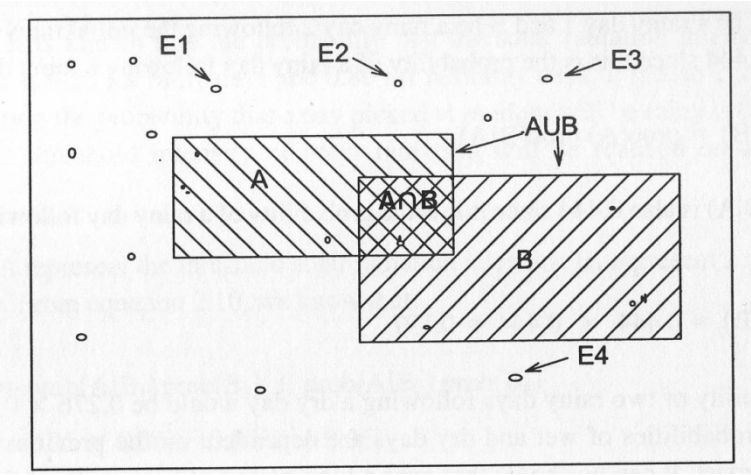


Figure 1.4: Venn diagram showing A∩B and A∪B

- **Complement event**

  P(A') or P(A complement) =probability of 'event A does not occur'

  $P(A') = 1-P(A)$ probability of A not to occur

  $P(A \cap A') = 0$ probability of A occur ***and*** A not occur =0

  $P(A \cup A') = P(A)+P(A')=1$ probability of A occur ***or*** A not occur =1

Example 1.4

River ***a*** and river ***b*** are two adjacent tributaries of a bigger river. River ***a*** has a yearly discharge smaller than its mean u_a in 50% of the years, while river ***b*** has a yearly discharge smaller than its mean u_b in 60% of the times. At any year the probability of river **b** having discharge smaller than u_b, given that the rive ***a*** has discharge smaller than u_a, is 70%.

=> A event: $Q_a < u_a$; P(A)=0.5

B event: $Q_b < u_b$; P(B)=0.6

P(B|A) = 0.7

Calculate the probability:

- that both rivers have a discharge smaller than their mean value.
  P(A∩B)=…
- that at least one river has a discharge smaller than its mean.
  P(A∪B)=…
- that river A has a discharge smaller than its mean given that river B has a discharge smaller than its mean.
  P(A/B)=P(∩B)/P(B)=…

- that at least one river has a discharge higher than its mean.
  $P(A' \cup B') = 1 - P(A \cap B) = \ldots$
- That both rivers have a discharge higher than their mean.
  $P(A' \cap B') 1 - P(A \cup B) = \ldots$

Example 1.5

(*from Haan, Example 2.2 on page 23*)

A study of daily rainfall at Station AK in July based on many years of observation has shown that:

Probability of a rainy-rainy day (i.e. a rainy day following a rainy day) is 0.444, a dry-dry is 0.724, a dry-rainy is 0.276, a rainy-dry is 0.556.

If it is observed that a certain July day (day_1) is rainy, what is the probability that the next **two** days (day_2 and day_3) will also be rainy?

Solution:
Let A be a rainy day_2 and B a rainy day_3 following the initial rainy day_1. The probability of A is 0.444 since this is the probability of rainy-rainy day.

The probability of two rainy days following a rainy day:
$P(A \cap B) = P(A) \cdot (B|A)$
Now the $P(B|A)$ is also 0.444 since this is the probability of a rainy day following a rainy day, therefore :
$P(A \cap B) = 0.444 * 0.444 = 0.197$

The probability of two rainy days following a dry day would be $0.276 \cdot 0.444 = 0.122$.

- **Total probability theorem (*optional*)**

Events that belong to a set which includes all the possible outcomes are called collectively exhaustive events. When rolling a die for example, the outcomes 1,2,3,4,5,6 are both mutually exclusive and collectively exhaustive because **only one** of these numbers can come up per roll, and **only a number belonging to this set** can come up.

If B_1, B_2, …, B_n represent a set of *mutually exclusive* and *collectively exhaustive* events, one can determine the probability of another event A from:

$$P(A) = \sum_{i=1}^{n} p(A / B_i) P(B_i)$$

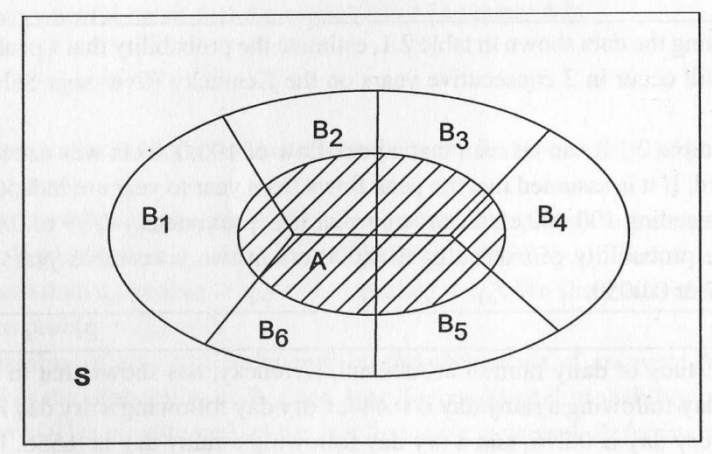


Figure 1.5: Venn diagram for theorem of total probability

Example 1.6

(*from Haan, Example 2.3 on page 25*)

It is known that the probability that solar radiation intensity will reach a threshold value is 0.25 for rainy days and 0.80 for non-rainy days. It is also known that for this particular location, the probability that a day picked at random will be rainy is 0.36. What is the probability that the threshold intensity of solar radiation will be reached on a day picked at random?

Solution:
Let A be the threshold solar radiation intensity, B1 be a rainy day and B2 a non-rainy day. From equation above we know

$P(A) = \Sigma P(A|B_i)\cdot P(B_i) = P(A|B1)\cdot P(B1)+P(A|B2)\cdot P(B2)$
$=0.25\cdot 0.36+0.8\cdot (1-0.36) = 0.602$

Bayes' Theorem (*optional*)

Bayes theorem provides a means of estimating probabilities of one event by observing a second event. Assume that you want to calculate the probability of event A using the probabilities of the collectively exhaustive events B_j.

Rewriting the conditional probability equation:

$$P(A \mid B_j) = P(A \cap B_j) / P(B_j) \text{ and } P(B \mid A) = P(A \cap B) / P(A)$$

We get:

$$P(A)\cdot P(B_j|A)=P(B_j)\cdot P(A|B_j)$$

Substituting from the total probability equation for P(A) = $\sum_{i=1}^{n} P(A/B_i)P(B_i)$,

$$P(B_j|A) = \frac{P(B_j)\cdot P(A|B_j)}{\sum_{j=1}^{n} P(B_j)\cdot P(A|B_j)}$$

where, $P(B_j)$ is called ***prior probability*** and $P(B_j|A)$ ***posterior probabilities***.

For application, see example 2.4 on page 25.

3. Sampling and Counting

Factorial Rule: For *n* different items, there are *n*! **arrangements**. The factorial rule is used when you want to find the number of arrangements for **ALL** objects.
For example, if you have three numbers – 1,2,3 – you can arrange them in 3!=1·2·3=6 different ways: 123, 132, 231, 213, 312, 321

Sampling: In calculating probability, we often need to deal with the problem of sampling or selecting a sample of *r* items from *n* items (e.g. select randomly *r*=3 pencils from a box that has *n*=20 pencils) Sampling can be done in four different ways:

a. Ordered with replacement: select randomly a pencil, write down its color, put it back in the box. In this case the order of colors that you draw is important.
b. Ordered without replacement: select a pencil, write down its color, don't put it back in the box.
c. Unordered with replacement: select a pencil, write down its color, put it back in the box. In this case the order of colors is not important (it doesn't matter if I got red, blue, red, but that I got two red and one blue).
d. Unordered without replacement: select a pencil, write down its color, don't put it back in the box.

- In the Ordered with replacement case, r items from *n* items can be selected in
  $\mathbf{n^r}$
  ways.
- In Ordered without replacement case, the first item has *n* ways of selection, and second has *n-1* ways, …. Thus *r* ordered items can be selected from *n* without replacement in:
  $(n)_r = n(n-1)(n-2)\ldots(n-r+1) = n!/(n-r)!$

  (The is called the number of **Permutations** of n items taken r at a time)

- In Unordered without replacement case, the number of ways in selecting r items from n items is
  $\binom{n}{r} = \frac{(n)_r}{r!} = \frac{n!}{(n-r)!r!}$
  (Combinations are arrangements of elements without regard to their order or position).

- In unordered with replacement case, the number of ways in selecting r items from n items is

$$\binom{n+r-1}{r} = \frac{(n+r-1)!}{(n-1)!r!}$$

Example 1.7

(*from Haan, example 2.5 page 28*)

For a particular watershed, 10 rain gages are available. Records from 3 are known to be bad. If 4 records are selected at random from 10 records,
(a) what is the prob that 1 bad is selected?
(b) Prob that 3 bad will be selected?
(c) Prob at least 1 bad is selected?

Solution:

The total number of ways of selecting 4 records (unsorted, without replacement) from 10 is $\binom{n}{r} = \frac{10!}{6!4!} = 210$

(a) the number of ways of selecting 1 bad from 3 bad and 3 good from 7 good is

$$\binom{3}{1}\binom{7}{3} = \frac{3!}{2!1!}\frac{7!}{4!3!} = 105$$

the prob of (a) is 105/210 = 0.5

(b) the number of ways of selecting 3 bad and 1 good is $\binom{3}{3}\binom{7}{1} = 1 \times 7 = 7$

the prob of (b) is 7/210 = 0.033

(c) prob of at least 1 bad record = 1- prob of no bad record

$$= 1 - \frac{\binom{3}{0}\binom{7}{4}}{210} = 1 - \frac{35}{210} = 0.833$$

(Note: Calculating the factorial of a number n in Matlab: >> factorial(n))

4. Frequency histogram and cumulative frequency distribution

Hydrologists are often faced with large quantities of data. Since it is difficult to grasp the total data picture from tabulation, a useful first step in data analysis is to use graphical procedures. Frequency histogram is a graphical presentation of the data. This is done by grouping the data into classes and then plotting a bar graph with the number or the relative frequency (proportion) of observations in a class versus the midpoint (called class mark) of the class interval.

Cumulative frequency distribution shows the frequency of events less than (greater than) some given value. They are formed by ranking the data from the smallest (largest) to the largest (smallest), dividing the rank by the number of data points and plotting this ratio against the corresponding data value.
Procedure:

- Suppose you have a data series of size n (e.g. table 2.1 on page 17)
- Assigning data to classes (depends on range of data, number of data and data behavior, normally the number of classes Nc = 5~20)
- Count the number of observations in each class
- Calculate the relative frequency of each class = number in class/total number
- Calculate the cumulative relative frequency $\sum$
- Plot the observed relative frequency on the graph.

Number of classes, Nc, could be determined from

$Nc = 1+3.3\log n$

Too few classes will eliminate detail and obscure the basic pattern of the data. Too many classes result in erratic patterns of alternating high and low frequencies (e.g. fig.1.6 b)

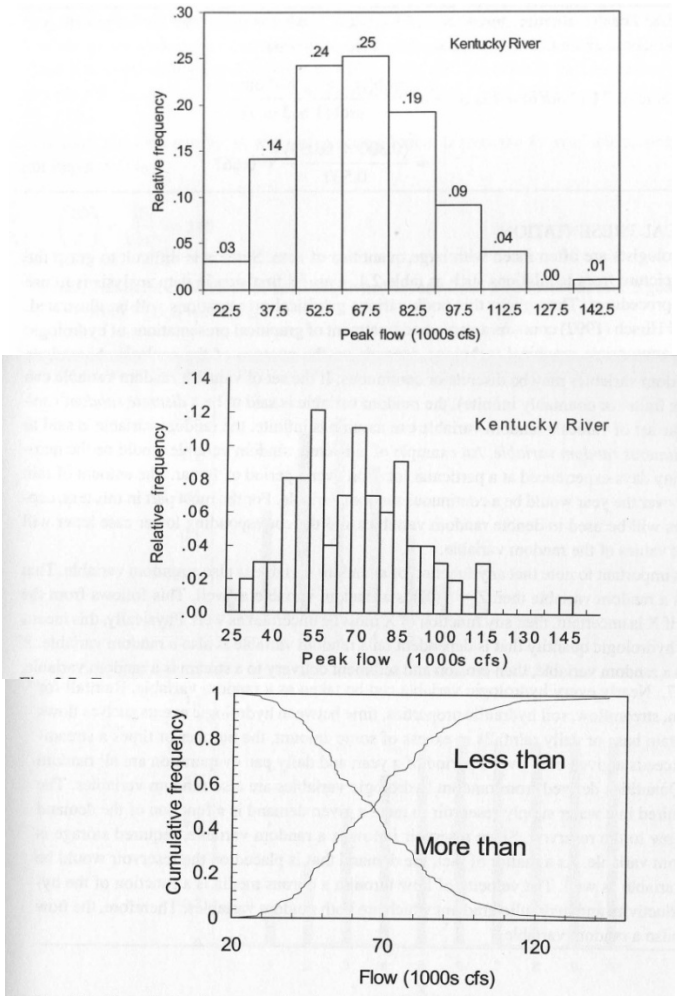


Figure 1.6: Kentucky River pick flow data: frequency histogram (top), frequency histogram with too many classes (middle), cumulative frequency distribution (bottom).

5. Probability distributions - Discrete random variable

A random variable is a real-valued function defined on a sample space. There are two types of random variable; *discrete* and *continuous*.

The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values. It is also sometimes called the probability function or the probability mass function.

More formally, the probability distribution of a discrete random variable X is a function which gives the probability $f(x_i)$ that the random variable equals x_i, for each value x_i:

$f(x_i) = P(X = x_i)$.

It satisfies the following conditions:
a) $0 \leq f(x_i) \leq 1$ for all x_i;
b) $\sum f(x_i) = 1$ for the whole possible range of x

A typical plot of the distribution of probability (*pdf=probability distribution function*) and the cumulative probability distribution (*cdf=cumulative distribution function*) associated with the values that a discrete random variable can assume:

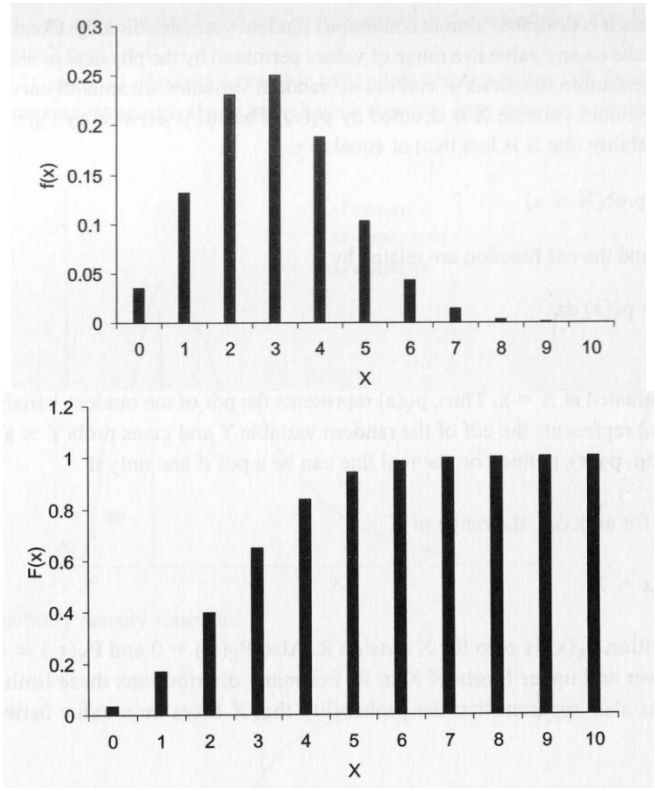


Figure 1.7: Top: A discrete probability distribution function – the probability that a random variable X equals x_i is $f(x_i)$. Bottom: A discrete cumulative distribution function – the probability that a random variable X is smaller than x_i is $F(x_i)$.

6. Cumulative probability distribution - Discrete random variable

It is a function giving the probability that the random variable X is less than or equal to x, for every value x.

Formally, the cumulative distribution function F(x) is defined to be

$F(x) = P(X \leq x)$, for $-\infty < x < \infty$

For a discrete random variable, the cumulative distribution function is found by summing up the probability distribution function:

$$F(x_n) = \sum_{i=1}^{n} f(x_i) \ , n=1,2,3\ldots$$

Example 1.8

Discrete case: A die is tossed 6 times, the expected probability for each x_i is $P(x_i)$:

x_i:	1	2	3	4	5	6
$P(x_i)=f(x_i)$:	1/6	1/6	1/6	1/6	1/6	1/6

The cumulative distribution function F(x) is then

x_i:	1	2	3	4	5	6
$F(x_i)$:	$f(x_1)=1/6$	$f(x_1)+ f(x_2)=2/6$ …	3/6	4/4	5/6	6/6

F(x) does not change at intermediate values. For example, F(1.3) = F(1.86) = F(1) = 1/6

Example1.9

Discrete case: Suppose a random variable X has the following probability distribution $f(x_i)$:

x_i	0	1	2	3	4	5
$f(x_i)$	1/32	5/32	10/32	10/32	5/32	1/32

The cumulative distribution function F(x) is then

x_i	0	1	2	3	4	5
$F(x_i)$	1/32	6/32	16/32	26/32	31/32	32/32

F(x) does not change at intermediate values. For example, F(1.3) = F(1.86) = F(1) = 6/32

7. Probability density function – Continuous random variables

For a continuous random variable X, if the function f(x) satisfies:

(1) f(x)>=0 for all x_i
(2) $\int f(x)dx = 1$ for the whole range of x

then f(x) is a probability density function.

(Discrete functions are said to have probabilities; continuous functions have probability densities.)

An important property is that if the *pdf* of a variable x is known, then the *pdf* of any g(x), where g(x) any function of x, can be derived:

$$f\big(g(x)\big) = f(x) \cdot \left|\frac{dx}{dg}\right|$$

8. Cumulative distribution function

It is a function giving the probability that the random variable X is less than or equal to x, for every value x.
Formally, the cumulative distribution function F(x) is defined to be:

$F(x) = P(X \leq x)$, for $-\infty < x < \infty$

For a continuous random variable, the cumulative distribution function is the integral of its probability density function.

Relationship between f(x) and F(x) is

$F(x) = \int f(x)dx$
And $f(x) = dF(x)/dx$

Since $F(x) = P(X \leq x)$ it follows that

$\int_a^b f(x)dx = P(a < X < b) = F(b) - F(a)$

Note that $P(X=d) = \int_d^d f(x)dx = F(d)-F(d) = 0$.

i.e. for continuous variables, probability of x getting any fixed value is zero!

Figure 1.8 illustrates a possible pdf and its corresponding cdf.

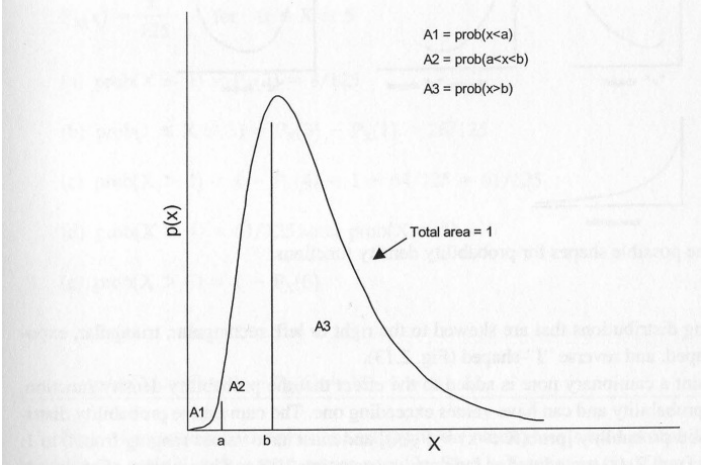


Figure 1.8: a) Probability density function

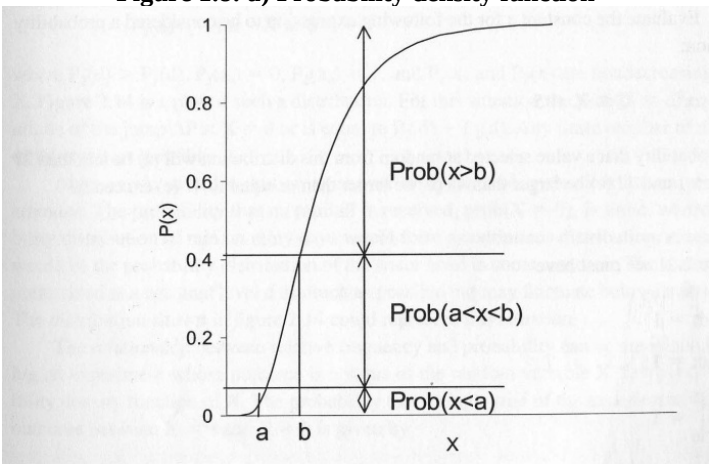


Figure 1.8: b) Cumulative probability distribution function

Example 1.10

(*from Haan, Example 2.8, pages 36-37*)

Evaluate the constant c for the following expression to be considered a probability density function

$f(x) = cx^2$ $\quad 0 \le x \le 5$

a) P (x < 2) =?
b) P(x ≤ 2) =?
c) P(1<x<3) =?
d) P(x>6) =?

Solution:

For f(x) to be a density function, it has to fulfill two conditions, i.e. f(x) >0, and $\int f(x)dx = 1$, then:

$$\int_0^5 f(t)dt = 1 \Rightarrow \int_0^5 ct^2 dt = \frac{ct^3}{3}\Big|_0^5 = 1 \Rightarrow c = 3/125$$

so f(x) = $\frac{3x^2}{125}$

and

$F(x) = \frac{x^3}{125}$ for 0≤x≤5

From above F(x) function, we get:

(a) P(x<2) = F(2) = 8/125

(b) P(x≤2) = F(2) =8/125

(c) P(1<x<3)=F(3)-F(1) = 26/125

(d) P(x>6) = 1-prob(x≤6)= 1-1=0 (This is obvious also because 0≤x≤5)

9. Expected relative frequency

If x_i represents the midpoint of an interval of $X = [x_i - \frac{\Delta x_i}{2}, x_i + \frac{\Delta x_i}{2}]$, then the expected relative frequency of outcomes in this interval of repeated, independent trials of the experiment is given by:

$$f_{x_i} = F(x_i + \Delta x_i / 2) - F(x_i - \Delta x_i / 2)$$

Because the right-hand side of this equation represents the area under f(x) between $x_i - \Delta x_i / 2$ and $x_i + \Delta x_i / 2$, it can be approximated by

$$f_{x_i} = \Delta x_i f(x_i)$$

This equation can be used to determine the expected relative frequency. See example below:

Example 1.11

Plot the expected frequency histogram using the probability density function: $(x) = \frac{3 \cdot x^2}{125}$, and a class interval of $\Delta x = ½$.

Solution:

$f_{x_i} = \Delta x_i \cdot f(x_i) = \frac{3 \cdot x^2}{250}$

x_i (*midpoint of class intervals*)	$f(x_i)$ (*pdf*)	f_{xi} (*expected relative frequency*)
0.25	0.0015	0.00075
0.75	0.0135	0.00675
1.25	0.0375	0.01875
1.75	0.0735	0.03675
2.25	0.1215	0.06075
2.75	0.1815	0.09075
3.25	0.2535	0.12675
3.75	0.3375	0.16875
4.25	0.4335	0.21675
4.75	0.5415	0.27075
		Sum: 0.9975

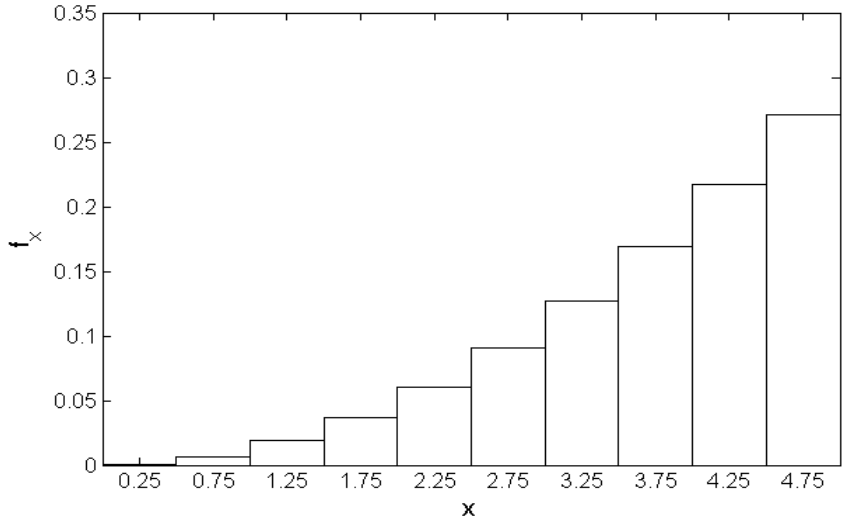


Figure 1.9: expected relative frequency

10. Bivariate distributions

For continuous random variables X and Y their joint probability distribution function is f(x,y) and corresponding cumulative probability distribution is F(x,y), and they are related by:

$$f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y)$$

and:

$$F(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v) du dv$$

The corresponding relationship for X and Y being discrete variables is as follows

$$f(x_i, y_j) = P(X = x_i \text{ and } Y = y_j)$$

$$F(x, y) = P(X \le x \text{ and } Y \le y) = \sum_{x_i \le x} \sum_{y_j \le y} f(x_i y_j)$$

Properties of cumulative distribution function:

- $F(x, \infty)=P(X \le x, Y \le \infty) = P(X \le x)=F(x)$ cumulative marginal distribution of X
- $F(\infty, y) = F(y)$
- $F(x, y) \ge 0$
- $F(\infty,\infty)=1$ (100% probability that X and Y less than ∞)
- $F(-\infty,y)=F(x, -\infty)=0$ (0% probability that X or Y less than $-\infty$)

Chapter 2

Properties of random variables

1. Basic concept

- **Statistical Inference** - use information from a sample to draw conclusions (inferences) about the population from which the sample was taken.
- **Population** - a complete assemblage of all values representative of a random process
- **Sample** – a group of values selected for studying the property of population
- **Parameter** - a parameter is a value, usually unknown (and which therefore has to be estimated), used to represent a certain population characteristic.
- **Statistic** - A statistic is a quantity that is calculated from a sample of data use to estimate the parameters of the population
- **Estimator** - An estimator is any quantity calculated from the sample data which is used to give information about an unknown quantity in the population. For example, the sample mean is an estimator of the population mean.
- **Estimate** - an estimate is the particular value of an estimator that is obtained from a particular sample of data and used to indicate the value of a parameter. If the value of the estimator in a particular sample is found to be 5, then 5 is the estimate of the population mean μ.
- **Estimation** - Estimation is the process by which sample data are used to indicate the value of an unknown quantity in a population

2. Moments

There are three commonly used characteristics to describe the shape of a probability density function:

a) Central tendency, the measures of which are: expected value, mean, median and mode.
b) Dispersion, the measures of which are: range, variance, standard deviation and coefficient of variation.
c) Symmetry, the measures of which are: skewness and coefficient of skewness.

The calculation of the above measures is based on a concept from classical mechanics: the moments about an axis. The first moment of an object about an axis tells us how easy is to rotate the object around this axis (fig.2.1), and is calculated as $\mu_1' = x \cdot W$, where x the distance from the rotation axis and W the weight of the object.

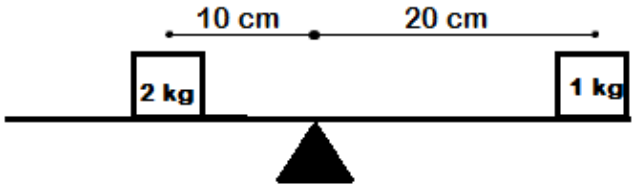


Figure 2.1: The system above is in equilibrium and is not rotating around the tip of the triangle, because the moments on the left and the right side of the bar are equal.

The same concept can be applied using area instead of weight (imagine an object like a thin plate), like in figure 2.2.

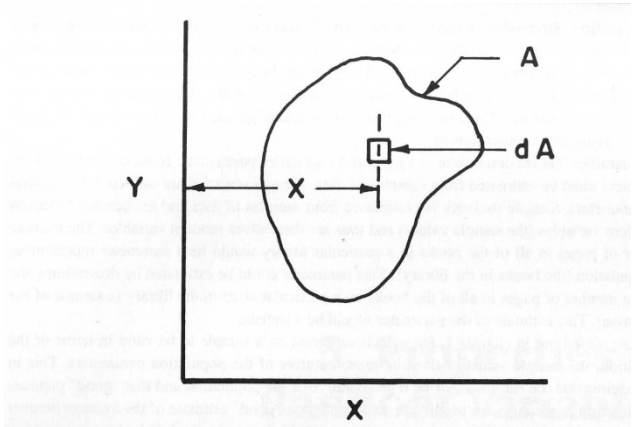


Figure 2.2: First moment of an infinitesimal area (dA) about the y-axis: $\mu_1' = x \cdot dA$. First moment of the whole area (A) about the y-axis: $\mu_1' = \int xdA$

Instead of an arbitrary shape we could have a probability distribution function. In this case the infinitesimal area is dA=dx·f(x) (fig.2.3), and the first moment about the y-axis is:

$$\mu_1' = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

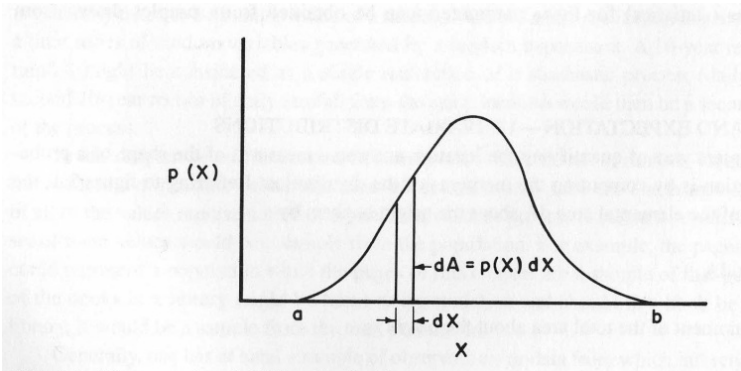


Figure 2.3: First moment of probability distribution function about the y-axis.

Definition of moments about y-axis: In general, if X is a random variable, the i^{th} moment of X about the origin is defined as:

$$\mu_i^{'} = \int_{-\infty}^{\infty} x^i f(x)dx \qquad \text{for continuous variables}$$

$$\mu_i^{'} = \sum_{j=1}^{n} x_j^i f(x_j) \qquad \text{for discrete variables}$$

Definition of central moments (about a vertical axis passing through the mean): The i^{th} central moment (moment about the mean) is given by

$$\mu_i = \int_{-\infty}^{\infty} (x-\mu)^i f(x)dx \qquad \text{for continuous variables}$$

$$\mu_i = \sum_{j=1}^{n} (x_j-\mu)^i f(x_j) \qquad \text{for discrete variables}$$

3. Expected value, E(x), expectation

The expected value (or population mean) of a random variable indicates its average or central value. It is a useful summary value (a number) of the variable's distribution. (It is usually unknown).

The expected value of a random variable X is symbolized by E(X) or μ.

If X is a **continuous** random variable with probability density function *f(x)*, then the expected value of X is defined by

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\ dx$$

If X is a **discrete** random variable with possible values x_1,.....,x_n, and corresponding probabilities $f(x_j)$ denote $P(X = x_j)$, then the expected value of X is defined by

$$E(X) = \Sigma x_j\ f(x_j)$$

where the elements are summed over all values of the random variable X.

Thus, the expected value (expectation) of X is the first moment about the y-axis ($\mu_1^{'}$).

It is apparent that:

- The expected value of $(x-\mu)^i$ is equal to the i^{th} central moment $E[(x-\mu)^i] = \mu_i$; later on we will show that when i = 2, μ_i is the variance
- The expected value of x^i is equal to the i^{th} moment about origin $E[x^i] = \mu_i'$

Properties:

If g(x) is a function of x

$E(g(x)) = \int g(x)f(x)dx$
$E(c) = c$
$E(c\, g(x)) = cE(g(x))$
$E(g1(x)+g2(x)) = E(g1(x))+E(g2(x))$
$E(x-\mu)^i = \mu_i$

For the central moment we have that

$$\mu_0 = \int (x-\mu)^0 f(x)dx = \int f(x)dx = 1$$

$$\mu_1 = \int (x-\mu)^1 f(x)dx = E(x-\mu) = E(x) - E(\mu) = \mu - \mu = 0$$

4. Measures of central tendency

- **Arithmetic mean:** A sample estimate of the population mean is the arithmetic average.

$$\overline{X} = \frac{1}{n}\sum x_i$$

- **Geometric mean:** Used in lognormal distribution

$$X_G = (\prod_{i=1}^{n} x_i)^{1/n}, \quad \text{where } \prod x_i = x_1 x_2 ... x_n$$

- **Weighted mean**

$$\overline{X}_w = \frac{\sum_{i=1}^{n} W_i X_i}{\sum_{i=1}^{n} W_i}$$

Where W_i is the weight associated with i^{th} observation, X_i, and n is the number of observation. In the special case $W_i = \frac{1}{n}$ the above equation becomes the Arithmetic mean.

- **Median**

The median is that value of x for which P(X≤x) = P(X≥x) = 0.5. In the case of a continuous distribution the median corresponds to an ordinate which separates density curve into two parts having equal areas of ½ each.

$$\int_{-\infty}^{\mu_{md}} f(x)dx = \int_{\mu_{md}}^{\infty} f(x)dx = 0.5$$

For discrete random variable, the median is the value halfway through the **ordered** data set, below and above which there is an equal number of data values.

Or $\mu_{md} = x_p$ where p is determined from

$$\sum_{i=1}^{p} f(x_i) = 0.5$$

It is generally a good descriptive measure of the location which works well for skewed data, or data with outliers. Median may not exist.

Example 2.1

With an odd number of data values, for example 11, we have:

Data:
7 4 7 3 9 7 5 8 3 6 5

Ordered data:
3 3 4 5 5 6 7 7 7 8 9

Median:
6: leaving 5 values below and 5 above.

With an even number of data values, for example 10, we have:

Data
7 4 7 3 9 7 5 8 3 6

Ordered data:
3 3 4 5 6 7 7 7 8 9

Median:
The value that is halfway between the two 'middle' data points, that is, halfway between 6 and 7 = 6.5 (If x is discrete variable that can only take integer values, median does not exist).

Thus, the median is the middle value or the mean of the middle two values, when the data is arranged in numerical order

- **Mode**

  The mode is the most frequently occurring value in a set of discrete data.
  For continuous variables, the probability distribution function f(x) has a maximum for x=mode:

$$\frac{df(x)}{dx} = 0 \quad \text{and} \quad \frac{\mathrm{d}^2 f(x)}{dx^2} < 0$$

  For discrete variable mode is the x value associated with $Max_{i=1}^{n} f(x_i)$

  A sample or population may have none, one or more than one mode. (Can you give an example of each case?)

Example 2.2

Suppose the results of the Statistics exam were distributed as follows:

Student	Score
1	94
2	81
3	56
4	90
5	70
6	65
7	90
8	90
9	30

Then the mode (most common score) is 90, the median (middle score) is 70, and the mean is 74. If the mean and median are not the same, the distribution of the sample is skewed (see next page).

Example 2.3

A discrete random variable has probability function

$f(x) = \frac{1}{2^x}$, where x = 1, 2, 3, …

Calculate the median and the mode

Solution:

In this case the **mode** is x = 1, for which the probability is ½ which is the maximum.

Median is a value x for which $P(X \le x) = 1/2$ and $P(X \ge x) = 1/2$

This means that 50% of the area under the f(x) curve lies on each side of median (fig.2.4).

If we assume that x can get continuous values:

$$\int f(x)dx = \int \frac{1}{2^x} dx = \frac{1}{log2 \cdot 2^x} = 0.5$$

=> x≈2.732
(*Given only integer values for x, one can assume that any number between 2 and 3 could represent the median. For convenience one can chose the midpoint of the interval, i.e. 5/2*)

If we assume that x can get only discrete integer values, then the median does not exist.

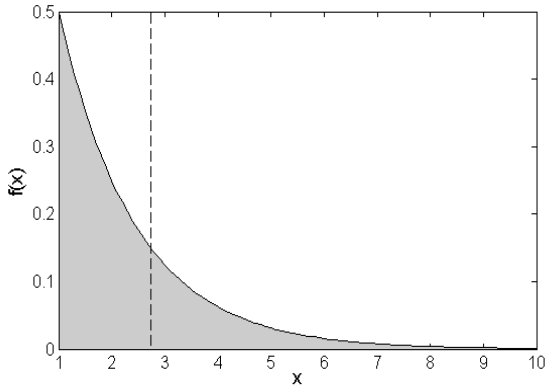


Figure 2.4: probability distribution function $f(x)=1/2^x$

5. Measures of dispersion

- **Range** = maximum value – minimum value

- **Variance**

  The (population) variance of a random variable is a non-negative number which gives an idea of how widely spread the values of the random variable are likely to be; the larger the variance, the more scattered the observations on average.
  Stating the variance gives an impression of how closely concentrated round the expected value the distribution is; it is a measure of the 'spread' of a distribution about its average value.

  Variance is symbolized by V(X) or Var(X) or σ^2.

  The variance of the random variable X is defined to be:

  $V(X) = \sigma^2 = E[X - E(X)]^2 = E(x-\mu)^2 = \int (x-\mu)^2 f(x)dx = u_2$ = **2nd central moment**

  For discrete variable: $$\sigma_x^2 = \frac{\sum (x_i - \mu)^2}{n}$$
  (i.e. average squared deviation from the mean).

  Sample estimation of population variance, σ_x^2: $$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Comparing equation for s^2 with that for σ_x^2, there are two differences; (1) μ (population mean, unknown) is replaced by $\bar{x}$ (sample mean, an estimate of μ), (2) *n* is replaced by *n-1*, in order to have unbiased estimation of population variance. This can be proven by derive $E(s^2) = \ldots \sigma^2$, (see appendix A)

Properties:

V(c)=0
V(c+x) = V(x)
$V(cx)=c^2V(x)$
$V(a+bx)=b^2V(x)$

- Standard deviation

  Taking the square root of the variance gives the standard deviation. Standard deviation has the same unit as the original variable x, which is an advantage over variance.

  That is: $\sqrt{V(X)} = \sqrt{\sigma^2} = \sigma$

  Sample estimation:

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

Notes:

a) The larger the variance, the further that individual values of the random variable (observations) tend to be from the mean, or average;
b) The smaller the variance, the closer that individual values of the random variable (observations) tend to be to the mean, or average;
c) The variance and standard deviation of a random variable are always non-negative.

- Coefficient of variation

  The coefficient of variation measures the spread of a set of data as a proportion of its mean. It is often expressed as a percentage. It is dimensionless, so it can be used for cross comparison, i.e. for comparison of the variability of the same variable at different places, etc.
  It is the ratio of the sample standard deviation to the sample mean:

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

  There is an equivalent definition for the coefficient of variation of a population, which is based on the expected value and the standard deviation of a random variable:

$$\frac{\sqrt{V(x)}}{E(x)}$$

6. Measures of symmetry

- Skewness
  Skewness is a parameter that describes asymmetry in a random variable's probability distribution.
  - $skewness = \int (x-\mu)^3 f(x)dx = \mu_3$ = **3rd moment about mean**

- Coefficient of skewness (dimensionless, more commonly used). Several types of skewness are defined, the terminology and notation of which are unfortunately rather confusing. "The" coefficient of skew of a distribution is defined to be
  - Population $\gamma = \frac{\mu_3}{(\mu_2)^{3/2}} = \frac{\mu_3}{\sigma^3}$
    Where μ_3 is the 3rd moment about mean, and σ is the standard deviation.
  - Sample unbiased estimate $Cs = \frac{n\sum(x_i-\bar{x})^3}{(n-1)(n-2)s^3}$

  Cs > 0 Positive skew (long tail in the right, most hydrological variables, e.g. discharge hydrograph, rainfall, etc), Cs<0 negative skew.

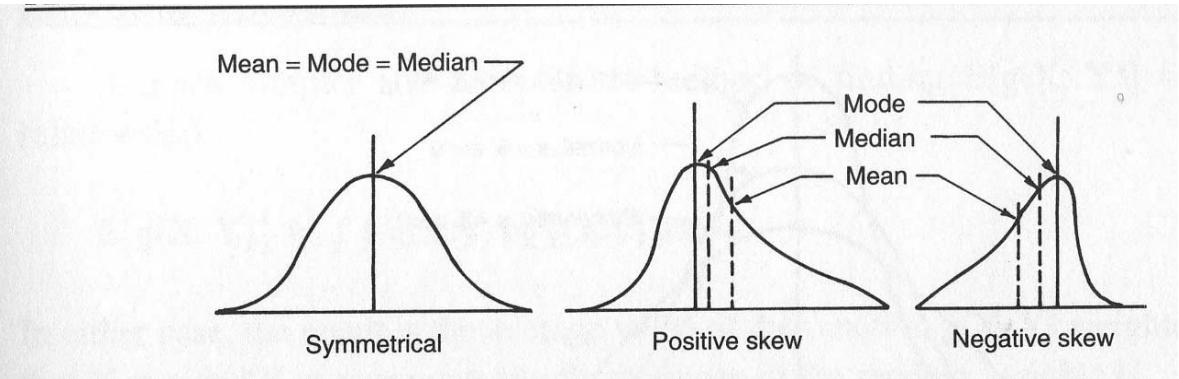


Figure 2.5: Location of mean, median and mode of symmetrical, positively and negatively skewed distributions.

7. Moments and expectation - Jointly distributed random variables

- Covariance (*measure of linear dependence only*)
  In probability theory and statistics, **covariance** is a measure of how much two variables change together. (Variance is a special case of the covariance when the two variables are identical.)

  If two variables tend to vary together (that is, when one of them is above its expected value, then the other variable tends to be *above* its expected value too), then the covariance between the two variables will be positive. On the other hand, if one of them tends to be above its expected value when the other variable is *below* its expected value, then the covariance between the two variables will be negative.

The covariance between two real-valued random variables *X* and *Y*, with expected values $E(X) = \mu_x$ and $E(Y) = \mu_y$ is defined as

$$
\begin{aligned}
&Cov(X,Y) = \sigma_{x,y} = \mu_{1,1} \\
&= \int\int (x - \mu_x)(y - \mu_y)\, f_{x,y,}(x,y)dxdy \\
&= E\left[(X - \mu_x)(Y - \mu_y\right] \\
&= E(X \cdot Y) - E(X) \cdot E(Y) \\
&= E(X \cdot Y) - \mu_x \mu_y
\end{aligned}
$$

where E is the expected value operator.

- Random variables whose covariance is zero are called uncorrelated

  If *X* and *Y* are independent, then their covariance is zero. This follows because under independence,

  $$E(X \cdot Y) = E(X) \cdot E(Y) = \mu_x \mu_y$$

  Recalling the final form of the covariance derivation given above, and substituting, we get

  $$Cov(X,Y) = \mu_x \mu_y - \mu_x \mu_y = 0$$

- The converse, however, is generally not true: Some pairs of random variables have covariance zero although they are not independent. Under some additional assumptions, covariance zero sometimes does entail independence, as for example in the case of multivariate normal distributions

i.e. if x and y are independent $\Rightarrow \sigma_{x,y} = 0$

But $\sigma_{x,y}$ =0 does not imply that x and y are independent (since Cov is a measure of linear dependence only).

Sample estimate $S_{x,y} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$

The units of measurement of the covariance Cov(*X*, *Y*) are those of *X* times those of *Y*. By contrast, correlation, which depends on the covariance, is a dimensionless measure of linear dependence.

Properties of Covariance

If *X*, *Y*, *W*, and *V* are real-valued random variables and *a*, *b*, *c*, *d* are constant ("constant" in this context means non-random), then the following facts are a consequence of the definition of covariance

$$Cov(X, a) = 0$$
$$Cov(X, X) = Var(X)$$
$$Cov(X, Y) = Cov(Y, X)$$
$$Cov(aX, bY) = abCov(X, Y)$$
$$Cov(X + a, Y + b) = Cov(X, Y)$$
$$Cov(aX + bY, cW + dV) = acCov(X, W) + adCov(X, V), + bcCov(Y, W) + bdCov(Y, V)$$

- Correlation coefficient – a normalized covariance (dimensionless)

$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y} \qquad \text{population}$$

Where σ_x and σ_y are standard deviation of population x and y

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y} \qquad \text{sample estimate}$$

Where s_x and s_y are standard deviation of sample x and y

Note: correlation coefficient so defined is a measure of linear dependence only!
If r =0, we cannot say x and y are independent. That means they do not have linear dependence (fig.2.6).

Note also that a high correlation does not necessarily mean a cause-and-effect relationship existence between the correlated variables, e.g. discharges of two adjacent catchments.

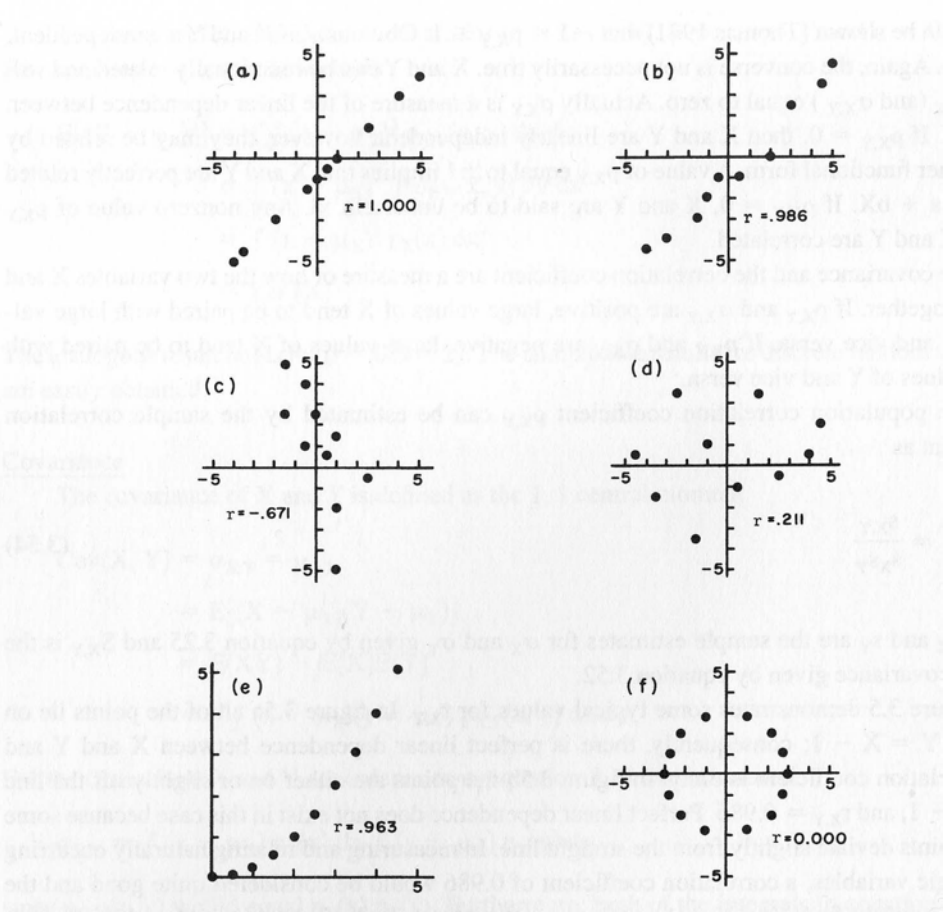


Figure 2.6: Examples of correlation coefficient

8. Further properties

If Z is a linear function of two random variables X and Y, then

- Z =aX+bY
- E(Z) = E(aX+bY) = aE(X)+bE(Y)
- Var(Z) = Var(aX+bY) = E(aX+bY)-E^2(aX+bY)

- A special case if $X_1, X_2, \ldots X_n$ is a random sample, then: $Y = \sum_{i=1}^{n} X_i / n$ is also a random variable, and

$$E(Y) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

$$Var(Y) = Var(\overline{X}) = \sum_{i=1}^{1} \frac{1}{n^2} Var(X) = \frac{n}{n^2} Var(X) \quad \text{or } S_{\bar{x}}^2 = \frac{1}{n} S_x^2$$

=> The variance of the mean of a random sample is equal to the variance of the sample divided by the number of observations in the sample.

Chapter 3

Some discrete distributions and applications in earth sciences

1. Hypergeometric distribution

Assumptions for hypergeometric distribution:

1. There is a population of N items divided into two groups (success or failure), k of which are belonging to one group ("success") and N-k of which are belonging to another group ("failure").

2. A random sample (without replacement) of n items is taken from the N items.

3. Probability of getting x successes (elements with property k) in n trials?

In the above definition, there are 4 parameters:

- N – the total number of items, size of population
- n – total number of trials, size of sample
- k – number of items in the population belonging to one group (success)
- x – number of items in the sample having the property of k

Solution:

- The total number of ways of selecting a sample of size n from N (*unordered, without replacement*) is:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

- The number of ways of selecting x successes and n-x failures from the population N contains k successes and N-k failures is

$$\binom{k}{x}\binom{N-k}{n-x} = \frac{k!}{x!(k-x)!}\frac{(N-k)!}{(n-x)!(N-k-n+x)!}$$

- The probability of getting x successes in a sample of size n drawing from population of size N contains k successes is

$$f_x(x;N,n,k) = \binom{k}{x}\binom{N-k}{n-x} \Big/ \binom{N}{n}$$

- Cumulative distribution: probability of getting less than or equal x successes is

$$F_x(x;N,n,k) = \sum_{i=0}^{x} \binom{k}{i}\binom{N-k}{n-i} \Big/ \binom{N}{n}$$

The expected number of events during n trials is:

$$E(x) = \frac{n \cdot k}{N}$$

And the variance is:

$$Var(x) = \frac{n \cdot k(N-k)(N-n)}{N^2(N-1)}$$

There are many such examples:

Example 3.1

Historical data show that there are 10 rainy days in June in Uppsala. Assume the rainy days are independent (this is usually not a good assumption). You want to go to field for one week to measure rainfall intensity.
(a) What is the probability of 4 rainy days in 7 randomly selected days in June?
(b) What is the probability that less than 4 of these randomly selected days are rainy?

Solution:

N=30, n=7, x=4, k=10

a)

$$f_x(x;N,n,k) = \frac{\binom{10}{4}\cdot\binom{20}{3}}{\binom{30}{7}} = \frac{\frac{10!}{4!\,(10-4)!}\cdot\frac{20!}{3!\,(20-3)!}}{\frac{30!}{7!\,(30-7)!}} = 0.12$$

b)

$$F_x(x;N,n,k) = \frac{\binom{10}{0}\cdot\binom{20}{7} + \binom{10}{1}\cdot\binom{20}{6} + \binom{10}{2}\cdot\binom{20}{5} + \binom{10}{3}\cdot\binom{20}{4}}{\binom{30}{7}} = 0.86$$

Example 3.2

Uppsala University has 3500 students, 200 of them have a name Anna, randomly select 100 students, what is the probability of getting 10 Anna?
N=3500, n=100, x=10, k=200

$f_x(x;N,n,k) = f_x(x;3500,100,200) = \ldots$

Example 3.3

A box contains 1000 balls, of which 200 are red balls, let's perform an experiment in such a way that each time a ball is taken from the box in random, its color is observed and do not put the ball back. What is the probability of getting 10 red balls in the 100 trials?

Solution:

N = 1000, k=200, x =10, n=100

$f_x(x;N,n,k)= f_x(x;1000,100,200) = \ldots$

Example 3.4

Assume we are playing a card game with a regular deck of 52 cards, where 16 of these are "face cards" and each "hand" consists of 10 randomly selected cards. Using combinations, find the probability of getting 4 face cards in a hand of 10 cards.

Solution:

We know there are $\binom{52}{10}$ possible "hands" and $\binom{16}{4}\binom{36}{6}$ possible hands with 4 face cards, therefore

P(4 face cards) = $\binom{16}{4}\binom{36}{6}/\binom{52}{10}$=0.224

2. Bernoulli Trials

(1) Binomial distribution

The *Bernoulli trials process*, named after James Bernoulli, is one of the simplest yet most important random processes in probability.

Assumptions for binomial distribution:
1. There are n trials, each classifiable as "success" or "failure".
2. The trials are independent (draw a sample with replacement, which is different from hypergeometric. The probability of success, p, is constant from trial to trial.)
3. Probability of getting x successes (elements with property k) in n trials?

The points 1 and 3 are the same as hypergeometric distribution, the condition 2 is different.

General form of Binomial distribution:

$$f_x(x;n,p) = \binom{n}{x} p^x q^{n-x}$$

where x- number of success, n – total number of trials, p – probability of one success, q – probability of one failure

Example 3.5

The probability of runoff exceeds a certain limit for any given year is p. Suppose the exceedance of runoff in any year is independent, what is probability that the runoff will exceed this limit once in three years?

Solution:

Let q=1-p be the probability runoff not exceeding its limit in a year.

Then we have three possible arrangements:

pqq (exceed in year 1) qpq (in year 2) qqp (in year 3)

The probability of exceeding one time in 3 years = pqq+qpq+qqp = $3pq^2$

Or more general $= \binom{3}{1} p^1 q^{3-1} = \frac{3!}{1!(3-1)!} p^1 q^2$

Example 3.6

A box contains 1000 balls, of which 200 are red balls. Let's perform an experiment in such a way that each time a ball is taken from the box in random, its color is observed and then the ball is put back. What is the probability of getting 10 red balls in the 100 trials?

Compare the result with the Example 3.3, which is the same experiment but **without** replacement.

Solution:

n = 100 total number of trials, x = 10 number of success, p = 200/1000 = 0.2, and q=1-0.2 = 0.8

$$f_x(x;n,p) = \binom{100}{10} 0.2^{10} 0.8^{90} =$$

The expected number of events during n trials and variance of x for Binomial distribution are:

$$E(x) = np$$

$$Var(x) = npq$$

The coefficient of skewness: $\gamma_s = \frac{p-q}{\sqrt{npq}}$

When p ≈ q or n is large, binomial approaches symmetrical ($\gamma_s \approx 0$)

Cumulative distribution: The probability of getting less than and equal x successes:

$$F_x(x;n,p)=\sum_{i=0}^{x}\binom{n}{i}p^i q^{n-i}$$

Example 3.7

(a) What is the probability of a 10-year flood will occur 4 times in 40 years?
(b) What is the probability of a 10-year flood will occur less than 4 times in 40 years
(c) On the average how many times a 10-year flood will occur 4 times in 40 years?

Solution:

10-year flood => return period T=10 years => probability to occur during a year= 1/T =0.1

(a) $f_x(x;n,p)=f_x(4;40,0.1)=\binom{40}{4}0.1^4 0.9^{36}=0.205$

(b) $$F_x(x;n,p)=F_x(4;40,0.1)=\sum_{i=0}^{3}\binom{40}{i}p^i q^{40-i}=\binom{40}{0}p^0 q^{40}+\binom{40}{1}p^1 q^{39}+\binom{40}{2}p^2 q^{38}+\binom{40}{3}p^3 q^{37}=0.0147+0.0656+0.142+0.200=0.422$$

(c) E(x) = n·p = 40·0.1 = 4 times

Note: The binomial distribution can be used to approximate the hypergeometric distribution if n (sample size) is small compared with N (population). In such a case, drawing a sample with or without replacement is nearly equivalent.

Example 3.8

Compare the hypergeometric and binomial for N=40, n=5, k=10, x=[0,5]

Solution:

x	**Hypergeometric** $f_x(x;N,n,k)=f_x(x;40,5,10)$	**Binomial** $f_x(x;n,p)=f_x(x;5,10,40)$
0	0.2166	0.2373
1	0.4165	0.3955
2	0.2777	0.2637
3	0.0793	0.0879
4	0.0093	0.0146
5	0.0004	0.0010

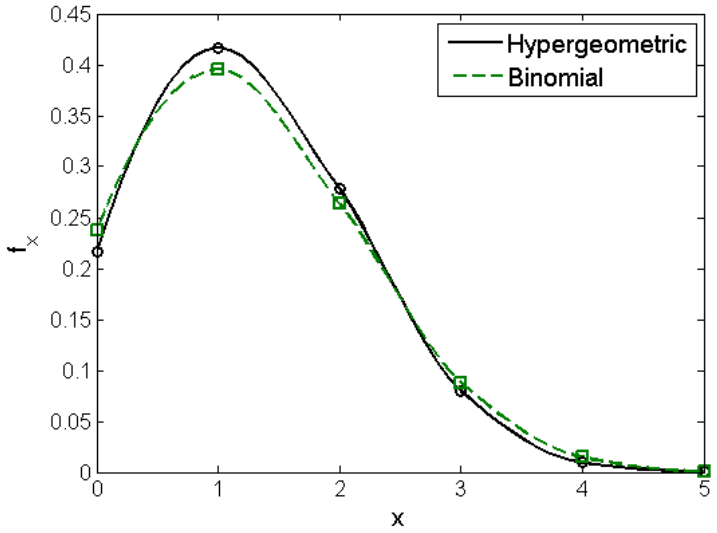


Figure 3.1: Hypergeometric and Binomial probability distribution functions

Comment: This merely indicates that drawing a small sample without replacement from a large population and drawing the same sample with replacement (so probabilities in each trial are constant) are nearly equivalent.

(2) Geometric distribution

Similar to the Binomial distribution, but now we are interested in the probability of first success.

Assumptions for geometric distribution:

1. There are n trials, each classifiable as "success" or "failure".
2. The trials are independent. (draw a sample with replacement, which is different from **hypergeometric.** The probability of success, p, is constant from trial to trial.)
3. What is the probability the 1st success occurs on the xth trial.

Note: conditions 1 and 2 are identical with **Binomial distribution,** but the condition 3, i.e. the question is different.

Example 3.9

Change the Example 3.4 slightly: the probability of runoff exceeds a certain limit for any given year is p; suppose the exceedance of runoff in any year is independent, what is the probability that the runoff will exceed this limit first time in the third year?

Solution:

First time in the third year means that no exceed in the first and second years, therefore the probability is qqp

General equation:

$$f_x(x; p) = pq^{x-1}$$

The expected number of events and variance of x for Geometric distribution:

E(x)=1/p

Var(x)=q/p2

Example 3.10

Compare:

(a) What is the probability of a 10 year flood will occur once in 10 years?

Solution: (**binomial**): $f_x(x; n, p) = f_x(1; 10, 0.1) = \binom{10}{1} \cdot 0.1^1 * 0.9^9 = 0.387$

(b) What is the probability that a 10 year flood will occur first time in the 10th year?

Solution: (**geometric**): $f_x(x; p) = p \cdot q^{x-1} = 0.0387$

(3) Negative Binomial Distribution

Assumptions for negative Binomial distribution:

1. There are n trials, each classifiable as "success" or "failure".

2. The trials are independent (draw a sample with replacement, which is different from **hypergeometric).** The probability of success, p, is constant from trial to trial.)

3. The probability that the k^{th} success occurs on the x^{th} trial (x > k) is given by the negative binomial distribution.

Note: conditions 1 & 2 are identical with **Binomial distribution and Geometric distribution,** but condition 3, i.e. the question, is different. For geometric distribution we asked 'what is the probability the 1^{st} success occurs on the x^{th} year', but here we ask 'what is the probability the k^{th} success occurs on the x^{th} year'.

The probability that the k^{th} exceedance (success) occurs on the x^{th} trial (x > k) of a Bernoulli process can be found by noting that there must be k-1 exceedances in x-1 trials preceding the k^{th} exceedance on the x^{th} trial.

$$f_x(x;k,p) = \binom{x-1}{k-1} p^k q^{x-k} \qquad x = k, k+1, \ldots$$

Mean and variance of Negative Binomial distribution:

$$E(X) = k/p$$

$$Var(X) = kq/p^2$$

Example 3.11

What is the probability that the fourth occurrence of a 10-year flood will be on the fortieth year?

Solution: $f_x(40;4,0.1) = \binom{39}{3} \cdot 0.1^4 \cdot 0.9^{36} = 0.0206$

3. Poisson Process and distribution

In hydrology, the Poisson distribution is often used to approximate the Binomial distribution with parameters *n* and *p*. Consider a Bernoulli process defined over an interval of time (or space) so that *p* is the probability that an event may occur during the time interval. If the time interval is allowed to become shorter and shorter so that the probability, *p*, of an event occurring in the interval gets smaller and smaller and the number of trials, *n*, increases in such a fashion that *np* remains constant, then the expected number of occurrence in any total time interval remains the same.

Let $\lambda = np$. The Bernoulli process can be approximate by Poisson distribution with parameter λ. This is useful since the computations involved in calculating Binomial probabilities are greatly reduced.
The probability of *x* events in a particular unit is,

$$P(X = x) = f_x(x;\lambda) = \frac{\lambda^x \cdot e^{-\lambda}}{x!} \quad , \quad x=1,2,\ldots; \lambda>0$$

Expected number of events: $E(x) = \lambda$

Variance: $Var(x) = \lambda$

Coefficient of skew: $\gamma_s = \frac{1}{\sqrt{\lambda}}$

Cumulative distribution: $F_x(x;\lambda) = \sum_{i=0}^{x} \frac{\lambda^i \cdot e^{-\lambda}}{i!}$

When λ is large, the distribution goes from a positively skewed distribution to a nearly symmetrical distribution.

Example 3.12

Calculate the probability that a 100 year flood will occur once in 20 years.

Solution:

Binomial: $f_x(x; n, p) = f_x(1; 20, 0.01) = \binom{20}{1}(0.01)^1(0.99)^{19} = 0.16$

Poisson: $f_x(x;\lambda) = f_x(1;0.2) = 0.2^1 e^{-0.2}/1! \approx 0.16$

Example 3.13

Calculate the probability that a 2 year flood will occur 5 times in 10 years.

Solution:

Binomial: $= f_x(x; n, p) = f_x(5; 10, 0.5) = \binom{10}{5} \cdot 0.5^5 * 0.5^5 = 0.246$

Poisson = $f_x(x;\lambda) = f_x(5;0.2) = \frac{5^5 * e^{-5}}{5!} = 0.176$

The difference is big in the second example, because, p is not small and n is not larger, in this case Binomial gets exact answer.

Chapter 4

Normal distribution and other continuous distributions

1. Normal distribution - Gaussian distribution

It is the most widely used and most important continuous probability distribution, because it has many applications and many other distributions (both discrete and continuous) can be converted to normal distribution under certain conditions.
Application in hydrology:

- model the error (measurement error, model simulation error) series
- annual maximum discharge
- a good approximation to many other distributions, both discrete and continuous

- **Density function** $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ (1)
- **Character**:
  - Two parameters μ (population mean), σ^2 (variance)
  - Bell shaped
  - Continuous and $-\infty < x < \infty$
  - symmetric about μ

A common notation is $N(\mu,\sigma^2)$

The normal density can be actually specified by means of equation (1). The height of the density at any value *x* is given by

$$f(\mathrm{x})=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Graph of normal distribution (fig. 4.1):

- The graph has a single peak at the center, this peak occurs at μ (the mean).
- The graph is symmetrical about μ (the mean).
- The graph never touches the horizontal axis.
- The area under the graph is equal to 1.

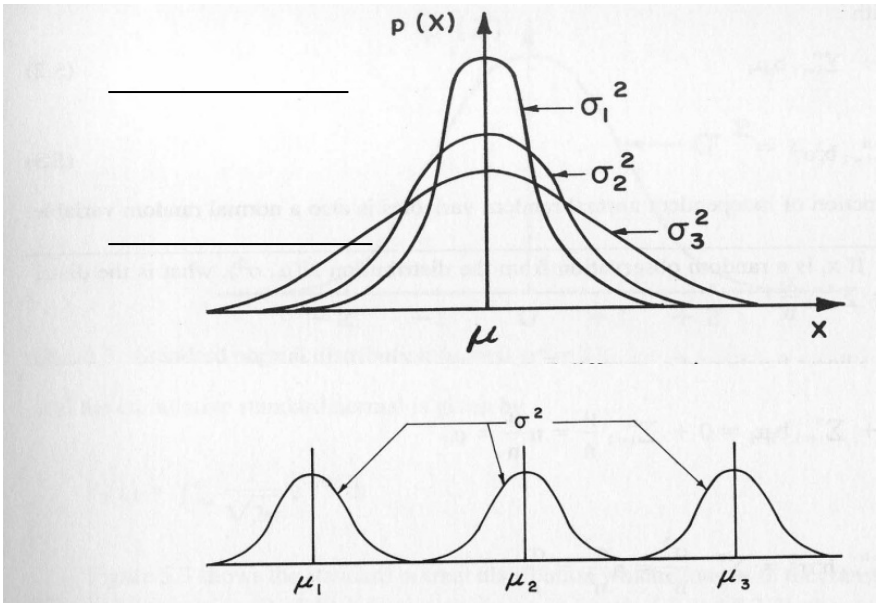


Figure 4.1: Top: Normal distributions with same mean and different variances; Bottom: Normal distributions with same variances and different means.

Although there are many normal curves, they all share an important property that allows us to treat them in a uniform fashion:

- **The 68-95-99.7% Rule**

  All normal density curves satisfy the following property which is often referred to as the *Empirical Rule*.

  - **68.27%** of the observations fall within **1 standard deviation** of the **mean**, that is, between $\mu - \sigma$ and $\mu + \sigma$.
  - **95.45%** of the observations fall within **2 standard deviations** of the **mean**, that is, between $\mu - 2\sigma$ and $\mu + 2\sigma$.
  - **99.73%** of the observations fall within **3 standard deviations** of the **mean**, that is, between $\mu - 3\sigma$ and $\mu + 3\sigma$.

  Thus, for a normal distribution, almost all values lie within **3 standard deviations** of the mean.

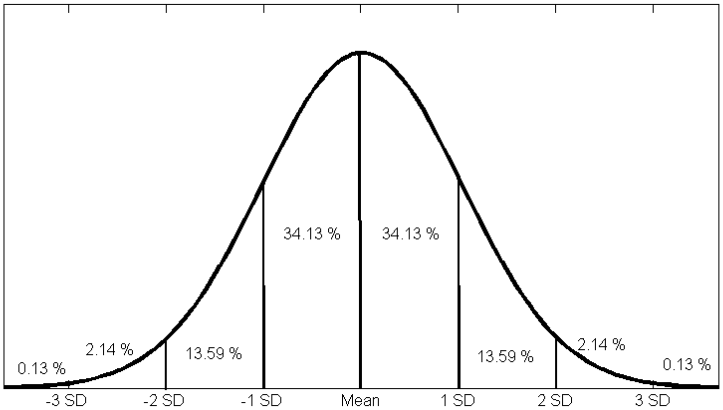


Figure 4.2: Normal distribution

Example 4.1

By definition, if a random variable x is normally distributed it has a range of $(-\infty, +\infty)$. This is because the probability density curve approaches but never reaches zero. However, most hydrological variables are bounded by X = 0. Explain/justify the use of normal distribution in some instances even though the hydrological variable under consideration may be bounded by X = 0.

Answer:

By studying 68-99% rule it can be seen that 68.26% of the normal distribution is within 1 standard deviation of the mean, 95.44% within 2 standard deviations of the mean and 99.74% within 3 standard deviations of the mean. These are called the 1, 2, and 3 sigma bounds of the normal distribution. The fact that only 0.26% of the area of the normal distribution lies outside the 3 sigma bounds demonstrates that the probability of a value less than μ-3σ is only 0.0013 and is the justification for using the normal distribution in some instances even through the random variable under consideration may be bounded by X=0. If μ is greater than 3σ, the chances of an X less than zero are negligible.

Consider for example that x is the annual precipitation, and that data from 100 years are available. Assume that x is normally distributed around 600mm, and has standard deviation 100mm (fig.4.3a). By plotting the cumulative probability function (fig.4.3b) we can derive the probability of x being less than three standard deviations from the mean:

$P_1(x < \bar{x} - 3 \cdot \sigma) = 0.0013$

And the probability of x being negative:

$P_2(x < 0) = 10^{-10}$

P_2 is small enough to be considered as zero.

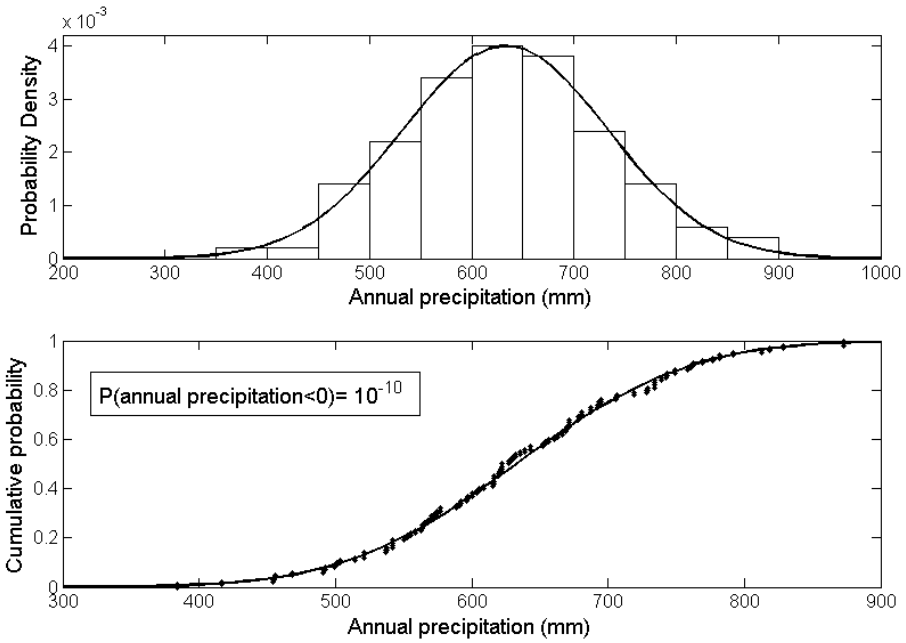


Figure 4.3: Theoretical PDF and CDF fitted to normally distributed annual precipitation data.

2. Reproductive properties

Let $X_1, X_2, ...X_n$ be independent random variables where X_i is $N(\mu_i, \sigma_i^2)$-distributed and let Y = a+bX; then Y is normally distributed with mean: a+bμ and variance: $(b\sigma)^2 = b^2\sigma^2$, i.e. $Y \sim N(a + b\mu, b^2\sigma^2)$.

In general, if X_i for i=1,2,…n are independently and normally distributed with mean μ_i and variance σ_i^2 then $Y=a+b_1X_1+b_2X_2 + \ldots+b_nX_n$ is normally distributed with:

Mean: $\mu_Y = a + \sum_{i=1}^{n} b_i\mu_i$

and variance: $\sigma_Y^2 = \sum_{i=1}^{n} b_i^2\sigma_i^2$

Example 4.2

Applying the properties above prove that if X~N(μ, σ^2) then $\bar{X}$ is $N(\mu, \frac{\sigma^2}{n})$

Solution:

$\bar{X} = (x_1 + x_2 + \cdots + x_n)/n$

=> $a = 0,\ b = 1/n$

$b_i=1/n$, $\mu_i = \mu$, $\sigma_i=\sigma\ \forall$ i

$$\mu_{\bar{X}} = 0 + \sum_{i=1}^{n} b_i \cdot \mu_i = \frac{1}{n} \cdot (\mu_1 + \mu_2 + \cdots + \mu_n) = \frac{1}{n} \cdot \mu \cdot n = \mu$$

$$\sigma_{\bar{X}}^2 = \sum_{i=1}^{n} \sigma_i^2 \cdot b_i^2 = \frac{1}{n^2} \cdot (\sigma_3^2 + \sigma_2^2 + \cdots \sigma_n^2) = \frac{1}{n^2} \cdot \sigma^2 \cdot n = \frac{\sigma^2}{n}$$

Example 4.3

Assume annual maximum discharge in a river station follows normal distribution. From 20 years observation data we get, the mean =10m^3/s and standard deviation = 3 m^3/s. what is the probability for any given year the annual maximum discharge is less than 8 m3/s?

Solution:

$$P(x<8) = F(8) = \int_{-\infty}^{8} f(x)dx = \int_{-\infty}^{8} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

This is not analytically integrable, it can be solved by standard normal distribution in next section.

3. Standard normal distribution

By using linear transformation:

$$Z = \frac{X - \mu}{\sigma}$$

the random variable Z will be normally distributed: N(μ=0,σ^2= 1).

The probability density function will then be transformed to: $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

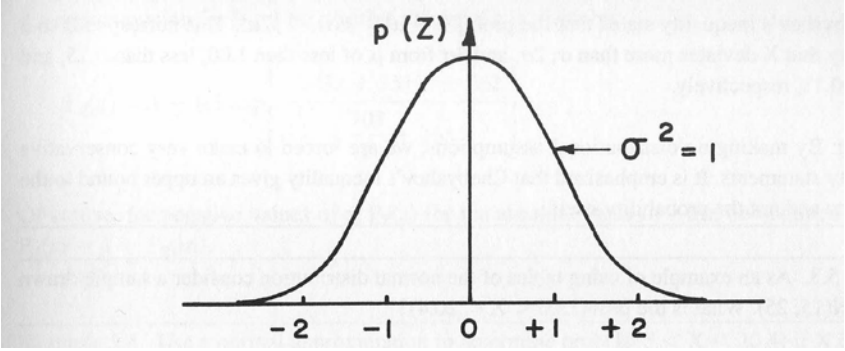


Figure 4.4: Standard normal distribution (μ=0,σ^2=1)

The 68-95-99.7% Rule applies for standard normal distribution.

Compare with general normal distribution.
Table A11 and A12 on page 461 – 462.

To solve the Example 4.3 using standard normal distribution table:

$$P(x < 8) = P\left(Z < \frac{x - 10}{3}\right) = P\left(Z < \frac{8 - 10}{3}\right) = P(Z < -0.67) = 1 - P(z < 0.67)$$

⇨ $P(x < 8) = 1 - 0.7486 = 0.251$

Example 4.4

30 years observation showed that in a station the mean annual temperature is normally distributed with mean 15 and variance 25. What is the probability for any given year the mean annual temperature is between 10.5 and 20.4?

Solution:

We know X is N(15, 5^2)

$$P(10.5 < X < 20.4) = P\left(\frac{10.5 - 15}{5} < Z < \frac{20.4 - 15}{5}\right) = P(-0.9 < Z < 1.08)$$
$$= P(Z < 1.08) - P(Z < -0.9) = 0.8599 - \big(1 - P(Z < 0.9)\big)$$
$$= 0.8599 - (1 - 0.8159) = 0.6758$$

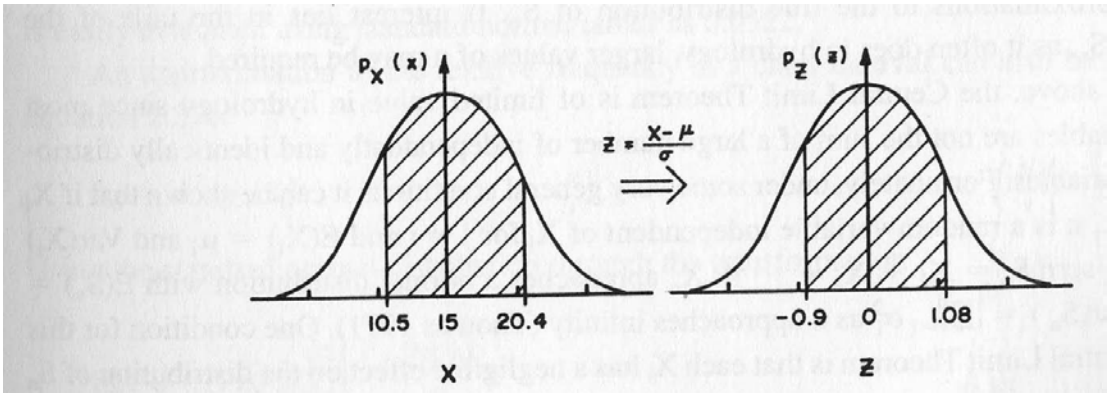


Figure 4.5: The shaded area is the probability: *P(10.5<x<20.4)=P(-0.9<z<1.08)*

4. Constructing Probability distribution function curves for data (very important!!!)

Procedure:

- Suppose you have a data series of size n (see Example 4.5 below).
- Assigning data to classes (depends on range of data, number of data and data behavior, Nc = 5~20)
- Count the number of observations in each class
- Calculate the relative frequency of each class = number in class/total number
- Calculate the cumulative relative frequency $\sum$
- Calculate for each class the expected (theoretical) relative frequency according to the density function of the assigned probability distribution. The relative frequency in class i is calculated: $f_{x_i} = \int_{x_{i-1}}^{x_i} f(x_i)dx$ or approximated by $f_{x_i} = \Delta x_i \cdot f(x_i)$. For normal distribution it can easily be calculated: $f_{x_i} = \Delta x_i \cdot \frac{f(z_i)}{\sigma}$ An example for normal distribution is shown on page 107.
- Calculate the empirical cumulative distribution function.
- Plot the observed relative frequency and expected relative frequency on the same graph.

Example 4.5

- Suppose you have a data series of size n=100. You can create for example a synthetic normally distributed data series $N(10,3^2)$ in Matlab as: randn(100,1)·3+10

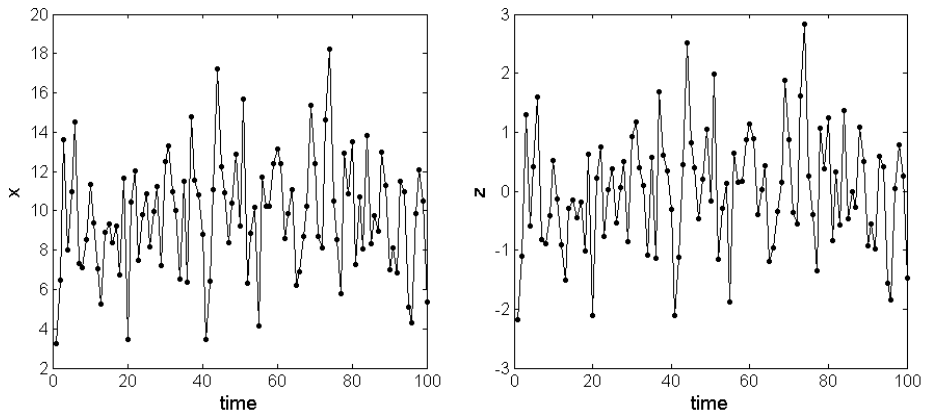


Figure 4.6: Data series $N(10,3^2)$

- Class mark: $Cm = \frac{class_{max} - class_{min}}{2}$
- Number of observations in each class: *No*

- Observed relative frequency: $f(x_i) = \frac{number\ of\ observations}{total\ number}$

- Cumulative relative frequency: $F(x_i) = \sum f(x_i)$
- Theoretical relative frequency: $f_{x_i} = \Delta x_i \cdot \frac{f(z_i)}{\sigma}, \quad f(z_i) = p_z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$

  Note that $p_z(z)$ is the definition of probability density function for standard normal distribution.

- Cumulative expected relative frequency: $F_{x_i} = \sum f_{x_i}$

Table 4.1: See above for the description of each column

***Cm* (for x)**	***Cm* for (z)**	***No***	***f(x$_i$)***	***F(x$_i$)***	f_{x_i}	F_{x_i}
4	-1.93	5	0.05	0.05	0.04	0.04
6	-1.26	14	0.14	0.19	0.12	0.16
8	-0.59	23	0.23	0.42	0.22	0.39
10	0.08	26	0.26	0.68	0.27	0.65
12	0.75	20	0.20	0.88	0.20	0.86
14	1.42	8	0.08	0.96	0.10	0.95
16	2.09	2	0.02	0.98	0.03	0.98
18	2.76	2	0.02	1	0.01	0.99

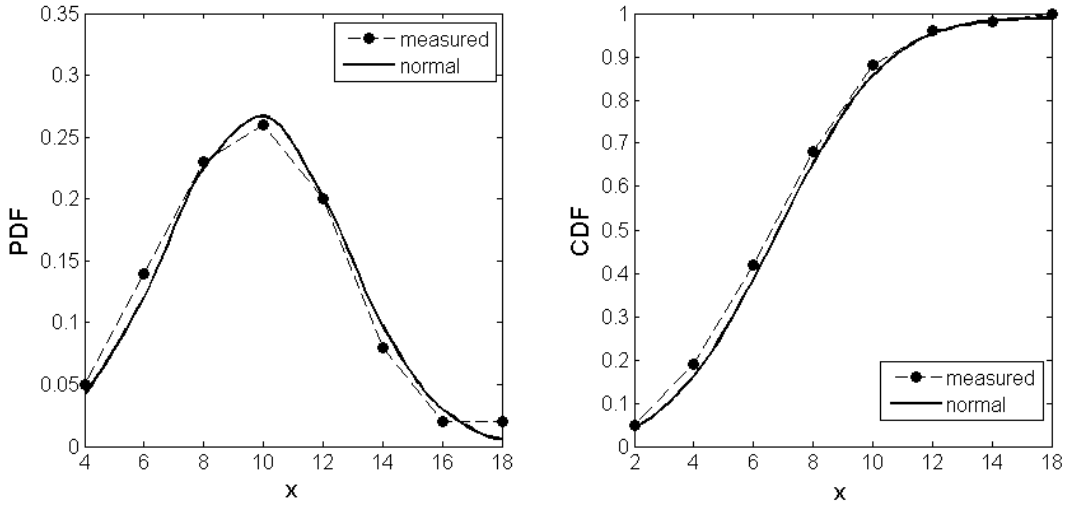


Figure 4.7: Measured and normal PDF and CDF

5. Central limit theorem

Let $X_1, X_2, ...X_n$ be independent, identically distributed (need not to be normal) random variables having mean μ and finite nonzero variance σ^2.
Let $S_n = X_1 + X_2 + ...X_n$. Then S_n is normally distributed with mean $n \cdot \mu$ and variance $n \cdot \sigma^2$.

The central limit theorem explains why many distributions tend to be close to the normal distribution.

6. Normal approximation for other distributions

- Binomial distribution

The Binomial distribution has an additive property (Gibra1973). If X is a binomial random variable with parameters n1 and p (n=number of trials, p=probability of a success) and Y is a binomial random variable with parameters n2 and p, then Z = X+Y is a binomial random variable with parameters n = n1+n2 and p.
Extending this to the sum of several binomial random variables, the Central Limit Theorem would indicate that the normal distribution approximates the binomial distribution if n is large.
If X is a variable with binomial distribution, with mean $\mu = n \cdot p$ and variance $\sigma^2 = n \cdot p \cdot q = n \cdot p(1-p)$, then as n gets larger the distribution of:

$$Z = \frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{np(1-p)}}$$

approaches a standard normal distribution N(0, 1).

Example 4.6

X is a binomial discrete random variable with n=25 and p=0.3. Compare the binomial and normal approximation to the binomial for evaluating the $P(5<X\leq 8)$

Solution:

Using binomial distribution:

$$P(5 < X \leq 8) = F_x(x; n, p) = \sum_{x_i=6}^{8} f(x_i; n, p)$$

$$P(5 < X \leq 8) = \sum_{x_i=6}^{8} f(x_i; 25, 0.3) = \sum_{x_i=6}^{8} \binom{25}{x_i} \cdot 0.3^{x_i} \cdot (1-0.3)^{25-x_i}$$

$$P(5 < X \leq 8) = 0.483$$

When using normal approximation, a correction should be applied for approximating a discrete random variable by a continuous random variable (table4.2):

Table 4.2: Corrections for approximating a discrete random variable by a continuous random variable:

Discrete	Continuous
X = x	x-½ < X < x+½
x ≤ X ≤ y	x-½ < X < x+½
X ≤ x	X < x+½
X ≥ x	X ≥ x-½
X< x	X ≤ x-½
X> x	X ≥ x+½

According to table4.2 and $Z = \frac{X - n \cdot p}{\sqrt{np(1-p)}}$, we have:

$$P(5.5 \leq X < 8.5) = P\left(\frac{5.5 - n \cdot p}{\sqrt{np(1-p)}} \leq Z < \frac{8.5 - n \cdot p}{\sqrt{np(1-p)}}\right)$$

$$= P\left(\frac{5.5 - 25 \cdot 0.3}{\sqrt{25 \cdot 0.3(1-0.3)}} \leq Z < \frac{8.5 - 25 \cdot 0.3}{\sqrt{25 \cdot 0.3(1-0.3)}}\right)$$

$$= P(-0.873 \leq Z < 0.435)$$

$$= P(Z < 0.435) - \big(1 - P(Z \leq 0.873)\big)$$

$$= 0.6664 - (1 - 0.8078) = 0.474$$

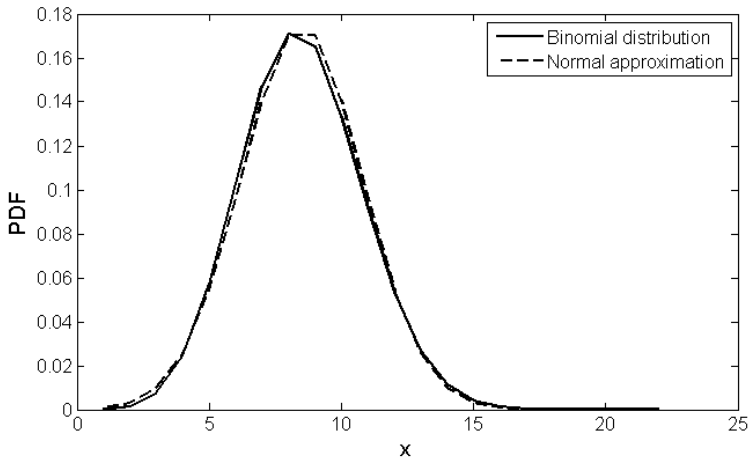


Figure 4.8: Binomial: f(x;25,0.3) and Normal: N(7.5,5.25) distributions.

- Poisson distribution with λ:

$\mu=\lambda$, $\sigma^2=\lambda$

$$\Rightarrow Z=\frac{X-\mu}{\sigma}=\frac{X-\lambda}{\sqrt{\lambda}} \qquad \text{and} \quad Z\sim N(0,1)$$

Example 4.7

Calculate and plot individual terms of the Poisson distribution for λ=2 and λ=9. Approximate the Poisson by the normal and plot the normal approximations on the same graph.

Solution:
The probability density function of Poisson distribution is:

$$f(x_i;\lambda)=\frac{\lambda^{x_i}\cdot e^{-\lambda}}{x!}$$

The probability density function of normal distribution is:

$$f(z)=\frac{1}{\sigma\cdot\sqrt{2\pi}}\cdot e^{-z^2/2}, \qquad z=\frac{x_i-\lambda}{\sqrt{\lambda}}$$

Table 4.3 Probability density functions for Poisson and Normal distributions

	λ=2			**λ=9**		
x_i	z_i	$f(x_i;\lambda)$	$f(z)$	z_i	$f(x_i;\lambda)$	$f(z)$
0	-1.41	0.1400	0.1038	-3.00	$13\cdot10^{-4}$	$15\cdot10^{-4}$
1	-0.71	0.2707	0.2197	2.67	$11\cdot10^{-4}$	$38\cdot10^{-4}$
2	0	0.2707	0.2821	-2.33	$50\cdot10^{-4}$	$87\cdot10^{-4}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
20	12.02	0.0014	0	3.33	$14\cdot10^{-4}$	$15\cdot10^{-4}$

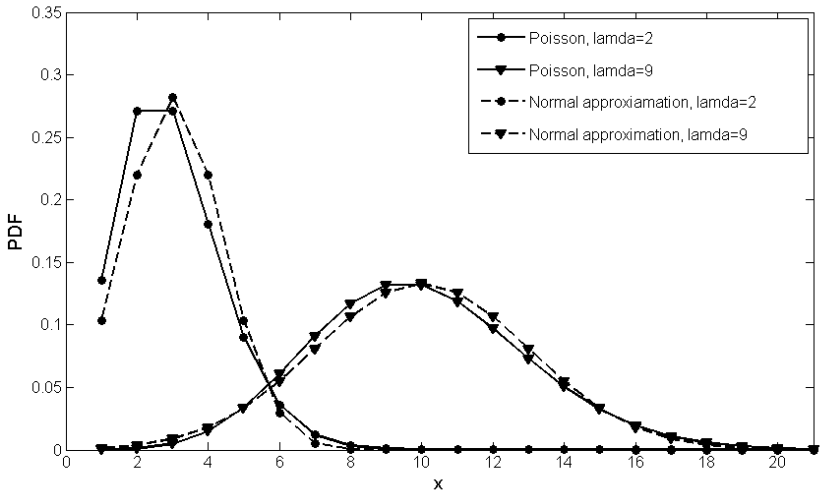


Figure 4.9: Poisson distribution (solid lines) and its normal approximation (dashed lines), for λ=2 (circles) and λ=9 (triangles)

The Poisson distribution is positively skewed, with coefficient of skew: $\gamma_s = \frac{1}{\sqrt{\lambda}}$. It becomes more symmetric and approaches better the normal distribution for decreasing γ_s, thus for increasing λ (fig.4.9).

7. Lognormal distribution

Applying the central limit theorem, X is a random variable and let $X = X_1 X_2 ... X_n$, then the logarithm of X, lnX, can be expected to be normally distributed. This can be seen by letting
$Y = \ln X = \ln(X_1 X_2 ... X_n) = \ln X_1 + \ln X_2 + ... + \ln X_n = Y_1 + Y_2 + ... + Y_n$.
From central limit theorem, Y can be expected to be normally distributed with mean μ_Y and variance σ_Y^2

$$f(y) = \frac{1}{\sigma_y \sqrt{2\pi}} e^{\frac{(y-\mu_y)^2}{2\sigma_y^2}} \qquad -\infty < y < \infty$$

Now we say, Y is normally distributed with parameters μ_Y and variance σ_Y^2
and X is lognormally distributed with parameters μ_Y and variance σ_Y^2.

We can also get the density function of X from:

$$f(y) = f(x) \cdot \left|\frac{dx}{dy}\right| \quad \Rightarrow \quad f(x) = f(y) \cdot \left|\frac{dy}{dx}\right|$$

$$Y = lnX \;\Rightarrow\; \left|\frac{dy}{dx}\right| = \frac{1}{x}$$

$$\Rightarrow \quad f(x) = \frac{1}{x} \frac{1}{\sigma_y \sqrt{2\pi}} e^{-\frac{(lnx-\mu_y)^2}{2\sigma_y^2}}, \quad x > 0$$

- Parameters μ_Y and σ_Y^2 can be estimated by $\bar{Y}$ and S_Y^2 in the usual manner by first transforming all of the X_i's to Y_i's by: $y_i = lnx_i$

  Then:

  $$\bar{y} = \frac{\sum y_i}{n} \qquad \rightarrow \mu_y$$

  And:

  $$s_Y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} \qquad \rightarrow \sigma_y^2$$

- Parameters μ_Y and variance σ_Y^2 can also be determined without taking the logarithm of all the data from

$$\overline{Y} = \frac{1}{2}\ln\left[\frac{\overline{X}^2}{C_v^2+1}\right] \qquad \rightarrow \mu_y$$
$$s_Y^2 = \ln(C_v^2+1) \qquad \rightarrow \sigma_y^2$$

where C_v is the coefficient of variation of the original data $= S_x / \overline{X}$

Note: *f(y)* is symmetrical, but *f(x)* is not! The parameters for *f(x)* are:

$$E(X) = \exp\left[\mu_Y + \sigma_Y^2/2\right]$$
$$Var(X) = \mu_X^2\left[\exp(\sigma_Y^2)-1\right]$$
$$C_v = \left[\exp(\sigma_Y^2)-1\right]^{1/2}$$

and the coefficient of skew is:

$$C_s = 3C_v + C_v^2$$

Table of standard normal distribution can be used to evaluate the lognormal distribution.

8. Chi-square distribution

If Z is N(0, 1), i.e. $Z = (X - \mu)/\sigma$, then:

$$Y = \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{x})^2}{\sigma^2}$$

has a Chi-square distribution with density function

$$f_{\chi^2}(x) = \frac{x^{-(1-\nu/2)}e^{-x/2}}{2^{\nu/2}\Gamma(\upsilon/2)}$$

where ν is the degree of freedom, and Γ is the Gamma function defined as:

$$\Gamma(n) = (n-1)!\ , \qquad if\ n = positive\ and\ even$$
$$\Gamma(n) = \sqrt{\pi}\frac{(n-2)!!}{2^{(n-1)/2}}\ , \qquad if\ n\ poditive\ and\ odd$$

Chi-square distribution is widely used in calculating confidence interval of variance and in testing the goodness-of-fit of observed data to a specified theoretical probability distribution, i.e. Chi-square test. This will be discussed more extensively in Lecture 7.

9. Student's t-Distribution

A Student's t-Distribution is a statistical distribution published by William Gosset in 1908.

Suppose X_1, ..., X_n are independent random variables that are normally distributed with expected value μ and variance σ^2. Let the **sample** mean be:

$$\overline{X}_n = (X_1 + ... + X_n)/n$$

And the **sample** variance be:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

It is readily shown that the quantity:

$$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

is normally distributed with mean 0 and variance 1. Gosset studied a related quantity:

$$T = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}$$

and showed that T has the probability density function

$$f(t) = \frac{\Gamma[(\nu+1)/2]\left(1+t^2/\nu\right)^{-(\nu+1)/2}}{\sqrt{\pi\nu}\,\Gamma(\nu/2)} \qquad -\infty < t < \infty;\ \nu > 0$$

The distribution of T is now called the ***t*-distribution**. The parameter v is conventionally called the number of **degrees of freedom**. The distribution depends on v, but not μ or σ; the lack of dependence on μ and σ is what makes the *t*-distribution important in both theory and practice

The mean, variance, skewness, of Student's *t*-distribution are

$\mu = 0$; $\sigma^2 = \nu/(\nu-2)$; $\gamma=0$;

When $\nu \to \infty$:
$\sigma^2 \to 1$
and t-distribution → Standard normal distribution

One application of the t distribution is to test the sampling distribution of the mean from a normal distribution with unknown variance. Details will be discussed in hypothesis testing.

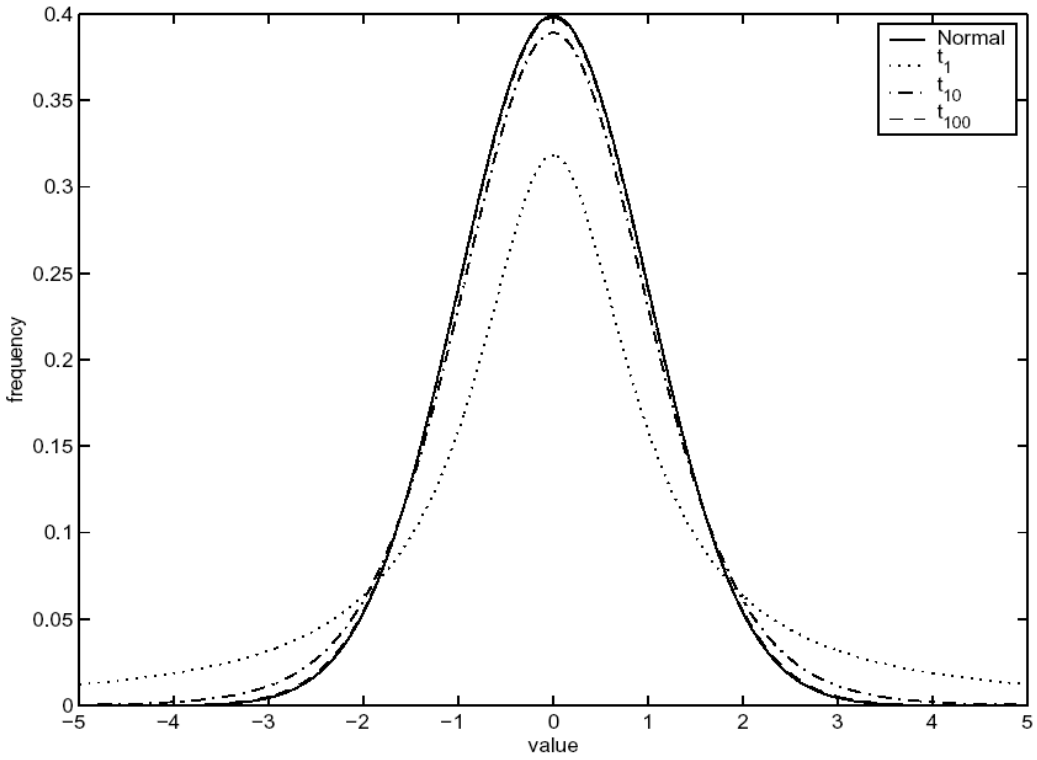


Figure 4.10: Comparison of the standard normal probability density function values (solid line) with probability density functions of t-distribution with 1, 10 and 100 degree of freedom.

10. F distribution

If U is a chi-square variate with m degrees of freedom and V is a chi-square variate with n degrees of freedom, and U and V are independent, then

$$X = \frac{U/m}{V/n}$$

Has an F distribution with γ_1= m and γ_2= n degree of freedom. The probability function of F distribution is given by:

$$f(x) = \Gamma(\frac{\gamma_1 + \gamma_2}{2}) \cdot \frac{\gamma_1^{\frac{\gamma_1}{2}} \cdot \gamma^{\frac{\gamma_2}{2}} x^{\frac{(\gamma_1 - 2)}{2}} \cdot (\gamma_1 + \gamma_2 \cdot x)^{-\frac{(\gamma_1+\gamma_2)}{2}}}{\Gamma\left(\frac{\gamma_1}{2}\right) \cdot \Gamma(\frac{\gamma_2}{2})}$$

The F distribution is often used to test the difference of variances of two normally distributed samples. Details will be discussed in hypothesis testing.

11. Extreme value distributions

In hydrology the extreme values are of specific interest. For example dams' construction requires the knowledge of extreme flow and precipitation; if a community's water supply is based on an unregulated river or on rainwater collection, it is necessary to know the probability of extremely low water supply (e.g. minimum monthly discharge and minimum monthly precipitation).

- Extreme value Type I – Gumbel distribution:

  Has been used for rainfall depth – duration – frequency studies (Hershfield 1961) and as the distribution of the yearly maximum of daily river flows. Its probability density function is:

$$f(x) = \frac{1}{\sigma} e^{-z-e^{-z}}, \quad z = \frac{x-\mu}{\sigma}$$

- Extreme value Type II – Frechét distribution:

  Is bound on the lower side (X>0) and has a heavy upper tail. Its probability density function is:

$$f(x) = \frac{\alpha}{\beta} \cdot \left(\frac{\beta}{x}\right)^{\alpha+1} \cdot e^{-\left(\frac{\beta}{x}\right)^{\alpha}}, \quad \alpha, \beta > 0$$

  α= shape parameter, β=scale parameter

- Extreme value Type III – Weibull distribution:

  Frequently used as the distribution of low stream flows, as it relates to extremely small values. Its probability density function is:

$$f(x) = \frac{\alpha}{\beta} \cdot \left(\frac{x}{\beta}\right)^{\alpha-1} \cdot e^{-\left(\frac{x}{\beta}\right)^{\alpha}}, \quad \alpha, \beta > 0$$

  α= shape parameter, β=scale parameter

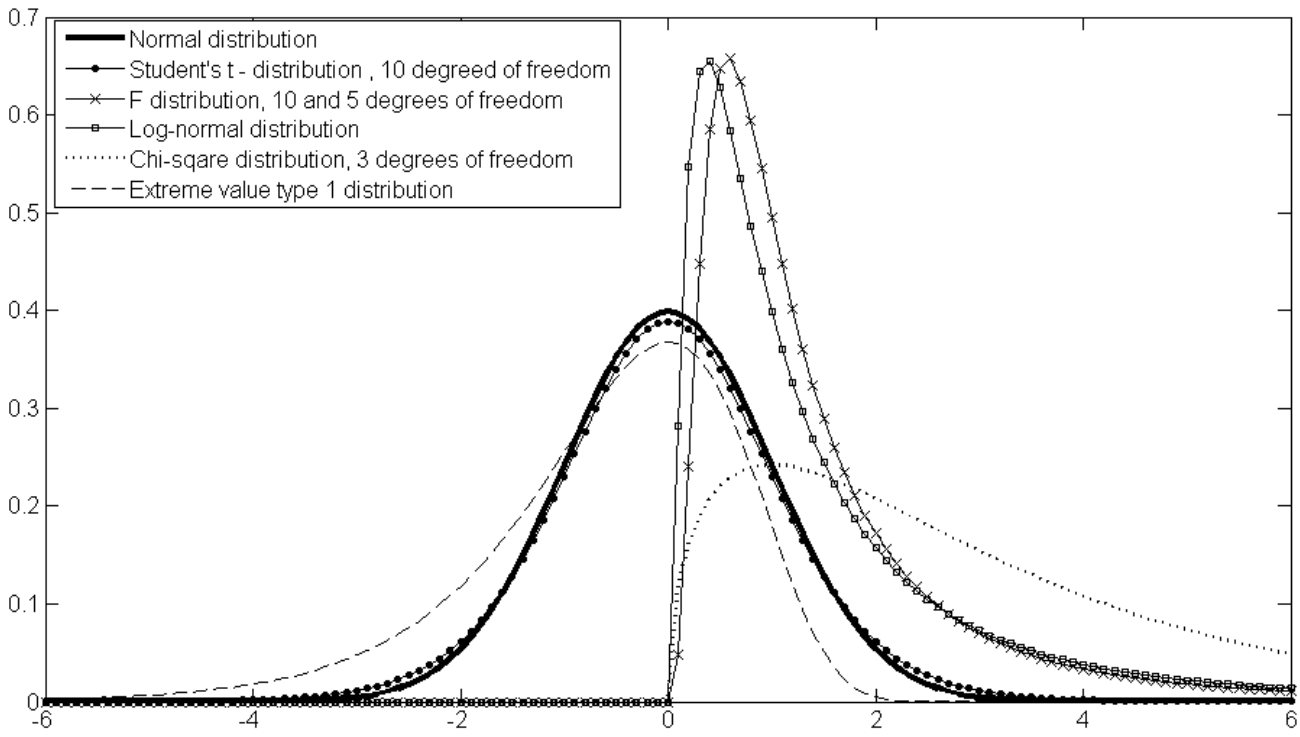


Figure 4.11: Summary of the probability distributions discussed above.

Chapter 5

Frequency analysis

1. Introduction

- Important tool in Engineering hydrology especially for design and operation of hydraulic structures.
- Design flood, Q_T – is the magnitude of the peak flow for a selected/predetermined design level as measured by return period, T (= magnitude of an event that has return period T)
- Return period, T – the average length of time between two events (i.e. floods) of a given magnitude or greater. Roughly, mean value of a series has a T = 2, smallest value has a T=1, and largest value has a T >= n.
- Probability of exceedance
  The probability that a given event, Q_T, will be exceeded, $P(Q>Q_T)$

  We have

$$T = \frac{1}{P(Q > Q_T)}$$

2. Purpose of Frequency analysis

- To calculate the design flow Q_T for a given return period, T, or to calculate the probability of exceedance.
- To estimate the return period or probability of exceedance for a given Q

3. Assumptions

- Data relevant – Q_{max} for flood, Q_{min} for water supply,
- Data sufficient – n>25 (requirement of statistical analysis, i.e. estimation of second and higher moments)
- Data independent – requirement for statistical model

The procedures for calculating design flood with different distributions are shown as an example, but the method of frequency analysis is not limited to design flood. It is also used in calculating

1. The design low flow (i.e. minimum one day, minimum 7 days flow, etc.),
2. The design storm (i.e. maximum 1 hour, maximum 24hours, maximum 3 days rainfall, etc.)

4. Methods

- Graphic method – probability plotting
- Analytical method

5. Graphic methods – procedure

- Select a Q_{max} value from each year,
- Range the data in a decreasing order, i.e. $Q_1 \geq Q_2 \geq Q_3 \ldots$
- Assign a frequency/probability of exceedance to each Qi. The most common method is the Weibull formula:

$$P(Q > Q_m) = \frac{m}{n+1}$$

where n is the total number of data and m is the order of the Q. This means that $m(Q_{max}) = 1$, and $m(Q_{min}) = n$.

- Plot Q ~$P(Q>Q_m)$ or plot Q~$T = 1/P(Q>Q_m)$
    - On millimeter paper (without distribution assumption) – fit a curve
    - On probability paper (with distribution assumption) – fit a straight line if the data fits the probability distribution as the probability paper presents

- Knowing the probability of $P(Q>Q_T)$ or T , Q_T can be read from plot or knowing the magnitude of Q_T, we can read $P(Q>Q_T)$ or T from the plot.

Example 5.1

100 measurements of daily peak flow are available (*here used a synthetic dataset, N(10,9)*). Calculate the flow with 10% probability of exceedance.

Solution:

Table 5.1: Synthetic dataset of daily specific discharge

Q (mm)	Sorted Q	Rank = m	Exceedance probability =m/(100+1)
15.7	15.7	1	0.04
8.8	13.9	2	0.08
11.2	13.8	3	0.12
6.5	13.6	4	0.19
3.4	11.3	5	0.23
⋮	⋮	⋮	⋮
7.7	3.4	100	0.96

From figure 5.1 Q=13.28.

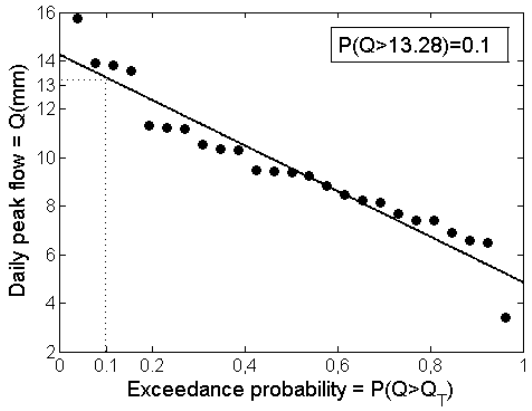


Figure 5.1: Exceedance probability

6. Analytical method – frequency factor method

The working equation:

Any random variable X can be written as:

$$X = \bar{x} + \Delta X$$

where ΔX is the deviation from the mean, $\bar{x}$,which is calculated as: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

A new quantity, the frequency factor *K*, can be defined as: $K = \frac{\Delta X}{s}$

where s is the standard deviation, calculated as: $s=\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

$$\Rightarrow \quad X = \bar{x} + sK$$

For a design value X_T, we can write

$$\boxed{\begin{array}{l} X_T = \bar{x} + K_T \cdot s \\ or \\ X_T = \bar{x}(1 + C_v K_T) \end{array}}$$

where C_v is the coefficient of variation = $s / \bar{x}$

The frequency factor K_T depends on the *probability distribution* being used and on the *return period*, T.

Application 1: given T to get X_T.
Procedure:
- Calculate $\bar{x}$ and s from the data series
- Get K_T from the tables or equations for a given probability distribution and T
- Calculate X_T from the equation

Application 2: given X_T to get T
Procedure:
- Calculate $\bar{x}$ and s from the data series
- Calculate K_T by rewriting the equation with K_T as unknown
- T can be read from the K_T~T table or calculated from the equations for a given probability distribution

6.1 Example for normal distribution

If X is normally distributed, i.e. $N(\bar{x}, s^2)$:

$$X_T = \bar{x} + K_T S$$

$$\Rightarrow \quad K_T = \frac{X_T - \bar{x}}{s}$$

That means that K_T is the standardized normal variate Z ($Z = \frac{X - \mu}{\sigma}$):

$$K_T = Z \sim N(0,1)$$

K_T can then be read from the standard normal distribution table for a given T or probability. Note that in the table, the probability $P(Z<z_T)$ is shown, while in frequency analysis we used probability of exceedance, $P(Z>z_T)$:

$$p(Z > z_T) = 1 - p(Z < z_T)$$

The standard normal distribution table can be used to:

- Get the value of Z_T (i.e. K_T) knowing $P(Z < z_T)$
- Get the value of $P(Z < z_T)$ knowing Z_T (i.e. K_T)

The value of Z_T (i.e. K_T) can also be obtained using the equation given in the book:

$$P(Z < z_T) = F(z_T) = \int_{-\infty}^{z_T} f(Z)dZ = \int_{-\infty}^{z_T} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

(The values in the standard normal distribution table were also calculated from this equation).

Example 5.1:

For a normally distributed annual maximum discharge, we have calculated that:
$\overline{Q} = 475$ m^3/s, s = 167 m^3/s
What is the design flood with a return period T = 100 years? i.e. $Q_{T=100}$ = ?

Solution:

T = 100
$\Rightarrow P(Q>Q_T) = 1/T = 0.01$
$\Rightarrow P(Q<Q_T) = 1 - P(Q>Q_T) = 0.99$
$\Rightarrow Z = K_T = 2.33$ (normal distribution table)

Then:
$Q_T = \bar{Q} + K_T.s = 475 + 2.33 \cdot 167 = 864$ m^3/s

6.2 Example for lognormal distribution

We say X is lognormally distributed if Y=ln(X) is normally distributed with mean μ_Y and variance σ_Y^2

If the return period T is known, in order to calculate X_T:

- Let $y_i = \ln x_i$ for all x_i

- Calculate $\bar{y} = \frac{1}{n}\sum y_i$ and $s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$
- Read K_T from normal distribution table for a given T
- Calculate $Y_T = \bar{y} + K_T S_y$
- Calculate $X_T = e^{Y_T}$

!Note: If only the mean and standard deviation $(\bar{x}, s_x)$ of a lognormally distributed variable x are available, then the mean and standard deviation $(\bar{y}, s_y)$ of the associated normally distributed variable y=ln(x) are calculated as:

$$\bar{y} = \ln\left(\frac{\bar{x}^2}{\sqrt{s_x^2 + \bar{x}^2}}\right), \qquad s_y = \sqrt{\ln\left(\frac{s_x^2}{\bar{x}^2} + 1\right)}$$

Example 5.2

For a lognormally distributed annual maximum discharge Q we have calculated that:
$\bar{Q} = 475$ m^3/s, s=167 m^3/s
What is the design flood with a return period T = 100 years? i.e. $Q_{T=100}$ =?

Solution:

Let $y_i = \ln(Q_i)$ for all Q_i

$$\bar{y} = \ln\left(\frac{\bar{x}^2}{\sqrt{s_x^2 + \bar{x}^2}}\right) = \ln\left(\frac{475^2}{\sqrt{167^2 + 475^2}}\right) = 6.1$$

$$s_y = \sqrt{\ln\left(\frac{s_x^2}{\bar{x}^2} + 1\right)} = \sqrt{\ln\left(\frac{167^2}{475^2} + 1\right)} = 0.34$$

T=100
$\Rightarrow P(Q>Q_T) = 1/T = 0.01$
$\Rightarrow P(Q<Q_T) = 1 - P(Q>Q_T) = 0.99$
$\Rightarrow Z = K_T = 2.33$ (normal distribution table)

$Y_T = \bar{y} + K_T S_y = 6.1 + 0.34 \cdot 2.33 = 6.89$
$X_T = e^{Y_T} = 984.6$ m^3/s

6.3 Other probability distributions

- Extreme value type I distribution:
  The K_T value can either be calculated by using equation below or read from Extreme value type I distribution given in the appendix (table7.5, appendix).

$$K_T = -\frac{\sqrt{6}}{\pi}\left\{0.5772 + \ln\left[\ln\left(\frac{T}{T-1}\right)\right]\right\}, \quad \Rightarrow \quad T = \frac{1}{1-\exp\left\{-\exp\left[-\left(0.5772 + \frac{\pi K_T}{\sqrt{6}}\right)\right]\right\}}$$

Example 5.3

The annual maximum discharge Q has an Extreme value Type I distribution; $\overline{Q} = 475$ m^3/s, s=167 m^3/s.
What is the design flood with a return period T = 100 years? i.e. $Q_{T=100}$ =?

Solution:
From the equation above:
$K_T = 3.14$
$\Rightarrow Q_T = 475+167\cdot 3.14 = 999$ m3/s

- Pearson Type III distribution:
  Pearson Type III distribution has three parameters, λ, β and ε, which can be estimated through calculation of mean, standard deviation and coefficient of skewness.

  Procedure:

  - Compute the mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \rightarrow \lambda$
  - Compute the standard deviation, $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \rightarrow \beta$
  - Compute the coefficient of skewness, $Cs = \frac{n\sum(x_i - \overline{x})^3}{(n-1)(n-2)s^3} \rightarrow \varepsilon$
  - Compute K_T by equations below or read from Table 7.3 given in the appendix.

    $$K_T = z + (z^2-1)k + \frac{1}{3}(z^3 - 6z)k^2 - (z^2-1)k^3 + zk^4 + \frac{1}{3}k^5$$

    Where $k = C_s/6$

    $$z = w - \frac{2.516 + 0.8029w + 0.01033w^2}{1 + 1.4328w + 0.1893w^2 + 0.00131w^3}$$

    $$w = \left[\ln\left(\frac{1}{p^2}\right)\right]^{1/2}$$

    $$p = \frac{1}{T}$$

  - Compute $x_T = \overline{x} + K_T s$

Example 5.4

Same as example 5.3, but assume that Q has Pearson Type III distribution:

K_T is read from the appendix – table 7.3a, using recurrence interval 100 years and skewness coefficient Cs=0.093:

$K_T = 2.4$
$Q_T = 475+167 \cdot 2.4 = 875$ m3/s

- Log-Pearson Type III distribution:

  Procedure:

  - Transformation x to Y=log(x)
  - Compute the mean, $\bar{y}$
  - Compute the standard deviation, s_y
  - Compute the coefficient of skewness, Cs
  - Compute K_T by equations above or read from Table 7.3a given in the appendix.
  - Compute $y_T = \bar{y} + K_T s$
  - Compute $x_T = 10^{y_T}$ or $x_T = e^{y_T}$

Example 5.5

The same as in example 5.3 but for log-Pearson Type III distribution:

y=ln(Q)
$\bar{y} = 6.1$
$s_y = 0.378$
Cs = -0.274
$K_T = 2.13$ from appendix – table 7.3
$Y_T = \bar{y} + K_T S_y$ =6.1+0.378*2.13 = 6.9
$Q_T = e^{Y_T} = 1002$ m3/s

7. Relationship between, design level (return period), design life and allowable risk

Design life = expected life of the structure

Example 5.6

From the above example (5.5) we know that the design return period for the Q = 1000 m^3/s at this station is 100 years. That means on the average the probability of the value will be exceeded (failure of the structure) is P(Q>1000) = 1/T = 0.01

What is the probability that the flow Q=1000 will be exceeded at least once in 100 years?

Solution:

P(Q>1000 in any year) = p = 0.01 *probability of failure in 1 year*
P(Q≤1000 in any year) = 1-p *probability of non-failure in 1 year*
P(Q≤1000 every year in 100 years) = $(1-p)^{100}$

P(Q>1000 at least one year in 100 year) = 1- Prob(Q≤1000 every year in 100 years) = $1-(1-p)^{100} = 1-(1-0.01)^{100} = 0.633$

Thus if the design life and its design return period are the same, the chances are very great that the capacity of the structure will be exceeded during it design life. The risk associated with a return period over n years is

$$\text{Risk} = 1-(1-p)^N = 1-\left(1-\frac{1}{T}\right)^N$$

Example 5.7

If the allowable risk for the structure in 20 years is 0.20, what return period shall be used to do the design?

Solution:

$$\text{Risk} = 1-(1-p)^N = 1-\left(1-\frac{1}{T}\right)^N$$

$\Rightarrow \quad 0.2 = 1-(1-1/T)^{20}$
$\Rightarrow \quad 1-1/T = 0.9889$
$\Rightarrow \quad T = 90$ years

If change 20 years to 100 years, then the return period becomes: T = 448 years

More Discussion and comparison will be made during the exercise time.

8. Other important notes

The accuracy of the calculation depends on:

1. how well the data fits the assigned probability distribution
2. the length of the data series

How do I know which distribution fits my data best?
Three methods:

2. probability paper – straight line
3. compare observed and expected (cumulative) relative frequency curves
4. goodness-of-fit testing: Chi-square methods and Kolmogorov-Smirnov tests

In order to get better fit of the probability distribution, the frequency analysis method may be applied to seasonal data. For example if you can show that the rainfall in different season may come from different origin and have different characters, it is then difficult to fit one distribution; different season may have different distributions

Chapter 6

Confidence interval and hypothesis testing

The confidence intervals and hypothesis testing are used for parameter estimation, forecasting and hypothesis testing, etc.

1. Definition

A confidence interval is a range of values that has a specified probability of containing the parameter being estimated. This statement may be written as

$$P(L<\theta<U) = 1-\alpha \qquad (1)$$

Where: -- L and U are lower and upper confidence limits; L,U are random variables.
-- [L, U] is the confidence interval (C.I).
-- α is the significance level; this is a probability and the values usually used are: 1% (α=0.01), 5% (α=0.05), or 10% (α=0.1).
-- 1- α is the confidence level, a probability, this is a probability and the values usually used are: 99% (1- α=0.99), 95% (1 - α=0.95), or 90% (1 - α=0.90).
-- θ is the parameter; this is a constant.

Equation (1) reads as: the probability that the interval L to U contains θ is 1-α.
It is NOT correct to read: the probability that θ is between L and U is 1-α.

2. Calculation of confidence intervals

Let the probability that the interval (v1,v2) includes the variable V is 1-*a:*

$$P(v1<V<v2)=1-\alpha \qquad (2)$$

- **C.I. for the mean of a Normal distribution with unknown variance (n small)**

Assume that we select randomly a small sample (n=small) from a population, and calculate the mean of the sample ($\bar{x}$). We want to know if the calculated mean is representative of the population mean (μ). Therefore we set the probability of the interval ($\bar{x}$-Δ, $\bar{x}$+Δ) including μ, and then calculate this interval. Δ is a number (name randomly chosen here) depending on the assumed distribution, significance level and sample's standard deviation.

If $x_i \sim N(\mu, s^2)$ then $\bar{x} \sim N(\mu, s_{\bar{x}}^2) = N(\mu, \frac{s^2}{n})$

From (2), let $V = \frac{(\bar{x}-\mu)}{s_{\bar{x}}}$. Then V has a **t distribution** with n-1 degree of freedom.

$$P(v_1 < \frac{(\bar{x}-\mu)}{s_{\bar{x}}} < v_2) = 1-\alpha \qquad (3)$$

$$\Rightarrow P(t_{\alpha/2,n-1} < \frac{(\bar{x}-\mu)}{s_{\bar{x}}} < t_{1-\alpha/2,n-1}) = 1-\alpha$$
$$\Rightarrow P(-t_{1-\alpha/2,n-1} < \frac{(\bar{x}-\mu)}{s_{\bar{x}}} < t_{1-\alpha/2,n-1}) = 1-\alpha$$
$$\Rightarrow P(\bar{x} - t_{1-\alpha/2,n-1}s_{\bar{x}} < \mu < \bar{x} + t_{1-\alpha/2,n-1}s_{\bar{x}}) = 1-\alpha$$

This latter equation is in the form of (1), so the confidence limits are:

$$l = \bar{x} - t_{1-\alpha/2,n-1}s_{\bar{x}}$$
$$u = \bar{x} + t_{1-\alpha/2,n-1}s_{\bar{x}} \qquad (5)$$

where $t_{1-\alpha/2,n-1}$ can be read from t distribution table (*note that this is the Δ mentioned above*).

- **C.I. for the mean of a Normal distribution with known variance (n large)**

  If the population variance is known or the sample size is large, i.e. $x_i \sim N(\mu,\sigma^2)$**,** then $\bar{x} \sim N(\mu,\sigma_{\bar{x}}^{\ 2}) = N(\mu,\frac{\sigma^2}{n})$

  The pivotal quantity V in (3) becomes $\frac{(\bar{x}-\mu)}{\sigma_{\bar{x}}}$, which has a **standard normal** distribution

The confidence limits (5) become:

$$l = \bar{x} - z_{1-\alpha/2}\sigma_{\bar{x}}$$
$$u = \bar{x} + z_{1-\alpha/2}\sigma_{\bar{x}} \qquad (6)$$

where $z_{1-\alpha/2}$ can be read from standard normal distribution table

Example 6.1

From 99 years annual discharge data it was calculated that the mean is 66.54 and the standard deviation is 22.32. (i.e. $\bar{x} = 66.54$, $s_x = 22.32$)
Calculated the 95% confidence interval for the mean assuming that:
(a) the true variance is unknown
(b) the true variance is known as $(22.32)^2$

Solution:

(a) $s_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \frac{22.32}{\sqrt{99}} = 2.243$

From the t table we get $t_{1-\alpha/2,n-1} = t_{0.975,98} = 1.99$

From equation 5:

$$l = \bar{x} - t_{1-\alpha/2,n-1}s_{\bar{x}} = 66.54 - 1.99(2.243) = 62.076$$
$$u = \bar{x} + t_{1-\alpha/2,n-1}s_{\bar{x}} = 66.54 + 1.99(2.243) = 71.004$$

Thus, we say we are 95% confident that interval 62.076 to 71.004 contains the true population mean.

(b) $\sigma_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \frac{22.32}{\sqrt{99}} = 2.243$

From the Z table we get $z_{1-\alpha/2} = z_{0.975} = 1.96$

From equation 6:

$$l = \bar{x} - z_{1-\alpha/2}\sigma_{\bar{x}} = 66.54 - 1.96(2.243) = 62.144$$
$$u = \bar{x} + z_{1-\alpha/2}\sigma_{\bar{x}} = 66.54 + 1.96(2.243) = 70.911$$

Thus, we say we are 95% confident that interval 62.144 to 70.911 contains the true population mean.

Question: why using Z distribution is confidence interval is smaller than using t distribution?

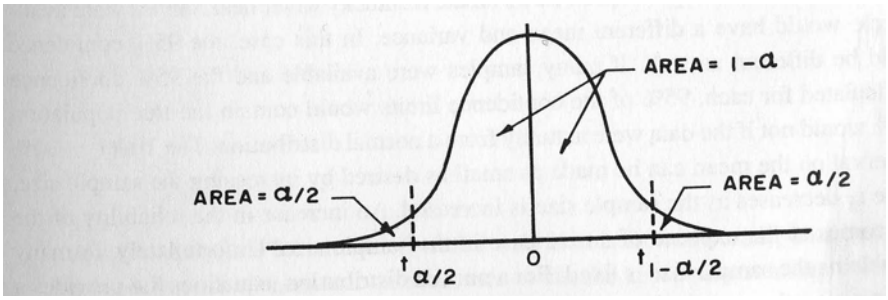


Figure 6.1: Illustration of confidence intervals using t distribution.

- **C.I. for the variance of a Normal distribution**

  We know that if $x_i \sim N(\mu, \sigma^2)$ then the quantity $\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2}$ has a Chi-square distribution (Lecture 4.9). Then:

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

Letting this quantity equal V in (3) results in:

$$P(v_1 < \frac{(n-1)s_x^2}{\sigma_x^2} < v_2) = 1-\alpha$$

Since we have Chi-square distribution, choose v1 equal $\chi^2_{\alpha/2,n-1}$ and v2 equal $\chi^2_{1-\alpha/2,n-1}$. Then:

$$P(\chi^2_{\alpha/2,n-1} < \frac{(n-1)s_x^2}{\sigma_x^2} < \chi^2_{1-\alpha/2,n-1}) = 1-\alpha$$

$$\Rightarrow\ P(\frac{(n-1)s_x^2}{\chi^2_{1-\alpha/2,n-1}} < \sigma_x^2 < \frac{(n-1)s_x^2}{\chi^2_{\alpha/2,n-1}}) = 1-\alpha$$

which is in form of (1) and the confidence limits on σ^2 are:

$$l = \frac{(n-1)s_x^2}{\chi^2_{1-\alpha/2,n-1}}$$

$$u = \frac{(n-1)s_x^2}{\chi^2_{\alpha/2,n-1}}$$

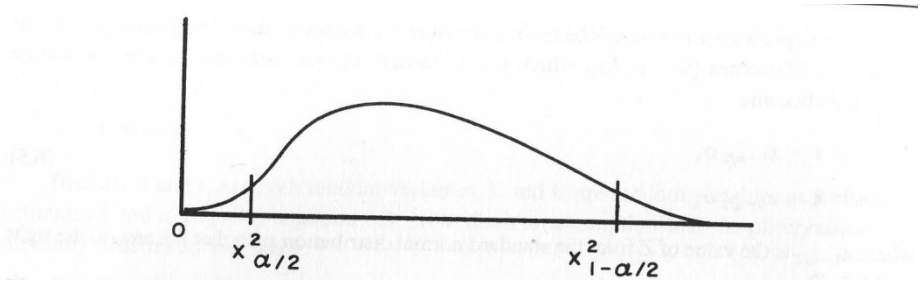


Figure 6.2: Confidence limits on a chi-square distribution

Example 6.2

Determine the 90% confidence interval on the variance for the situation described in example 6.1.

Solution:

$\chi^2_{\alpha/2,n-1} = \chi^2_{0.05,98} = 76.1$ (Chi-square distribution table, appendix)

$\chi^2_{1-\alpha/2,n-1} = \chi^2_{0.95,98} = 122.1$ (Chi-square distribution table, appendix)

$$l = \frac{(n-1)s_x^2}{\chi^2_{1-\alpha/2,n-1}} = \frac{98(22.32)^2}{122.1} = 400$$

$$u = \frac{(n-1)s_x^2}{\chi^2_{\alpha/2,n-1}} = \frac{98(22.32)^2}{76.1} = 642$$

The 90% confidence intervals for the variance are found to be 400 to 642.

Or: The 90% confidence intervals for the standard deviation are found (by taking the square roots of the above limits) to be 20 to 25.

- **One-sided confidence intervals**

  Sometimes one is only interested in an interval estimate on one side of a parameter. Then equation (1) becomes:

  Prob(L<θ) = 1-α,
  Following the same procedure we gets,

  $l = \bar{x} - t_{1-\alpha,n-1}s_{\bar{x}}$

  The analogous results would hold for any one-sided, lower or upper confidence limit.

- **Confidence intervals for proportions/probabilities**

Sometimes we are interested to know what fraction of the population has a certain characteristic. Since we cannot examine the whole population, we choose a random sample (sample size= n), and examine what proportion ($\hat{p}$) of this sample bares this characteristic. The confidence interval **population proportion** is calculated as:

$$P = \hat{p} \pm z_{1-\frac{a}{2}}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$$

Note that when the sample size increases P approaches $\hat{p}$, which means that the sample estimate becomes more reliable.

Example 6.3

500 volunteers participated in a drug test, 8 of which reported a particular side effect.
a) Find the 95% confidence interval for the population of all people that are going to use this drug to have the same side effect.
b) Find the number of volunteers needed, so that we are 95% confident that the **estimate for the population proportion** will not exceed the **estimate for the sample proportion** more than 0.5%.

Solution:

$\hat{p}$ =8/500=0.016
$z_{1-\frac{a}{2}} = z_{1-\frac{1-0.95}{2}} = z_{0.975} = 1.96$ (from normal distribution table)

a) n=500

$$P = \hat{p} \pm z_{1-\frac{a}{2}}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} = 0.016 \pm 1.96\sqrt{\frac{0.016(1-0.016)}{500}}$$

P=(0.005, 0.027)

We are 95% confident that the drug will have side effects on 0.5 to 2.7 percent of the whole population that will use it.

b) $z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} \leq 0.005$

$n \geq \left(\frac{z_{0.975}}{0.005}\right)^2 \cdot 0.016 \cdot (1-0.016) = 2420$

P=0.016±0.005= (0.011, 0.021)

After testing the drug on at least 2420 people, we will be 95% confident that the drug will have side effects on not more than 2.1% of the whole population that will use it.

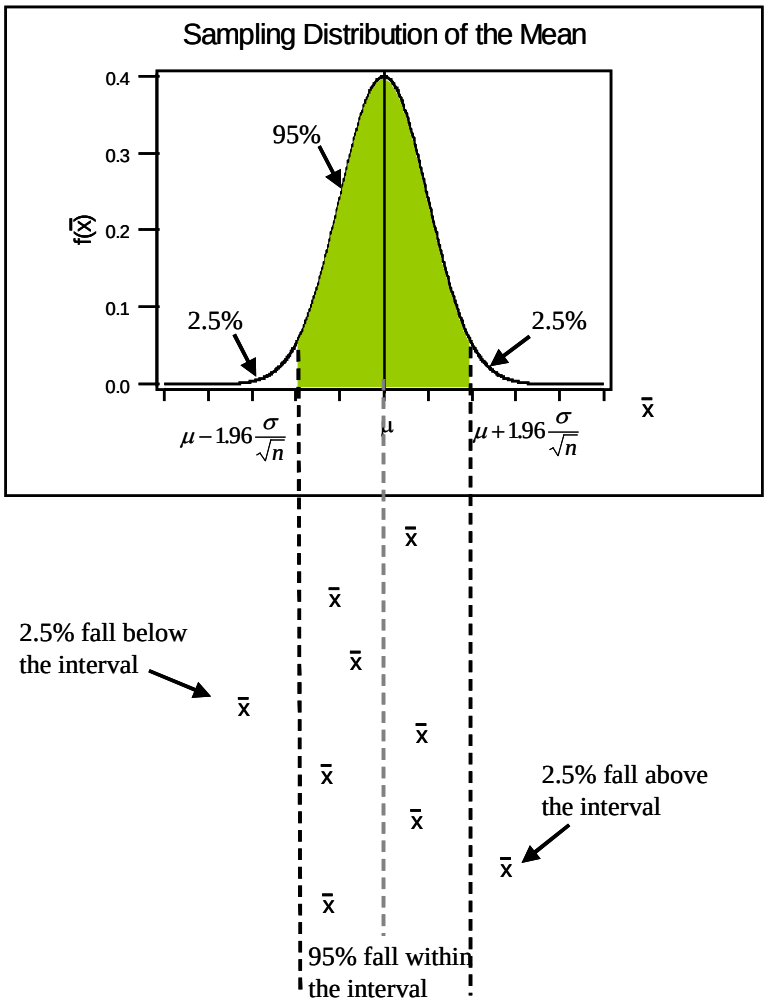
3. Summary of confidence interval

- **HOW CONFIDENCE INTERVALS BEHAVE**
  - Ideal situation – high confidence and small margin of error
  - Margin of error (E) = $z_{\alpha/2}\,\sigma/\sqrt{n}$
  - The smaller the margin of error, the more precise our estimation of μ
  - The margin error equation can also be used to estimate sample size in order to achieve a given margin of error:

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$$

- **Properties of error**
  - **Error increases with smaller sample size**
    For any confidence level, large samples reduce the margin of error (sample size)
  - **Error increases with larger standard Deviation**
    As variation among the individuals in the population increases, so does the error of our estimate (data behavior)
  - **Error increases with larger z values**
    Tradeoff between confidence level and margin of error (significance level)

- **A 95% Interval around the Population Mean**

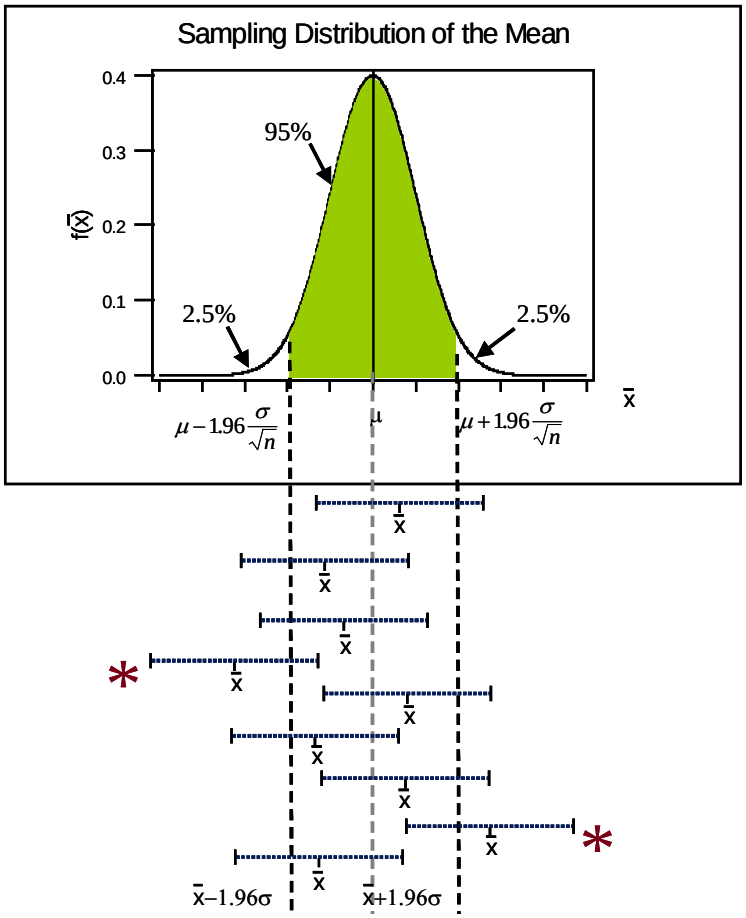


Approximately 95% of sample means can be expected to fall within the interval $\left[\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}\right]$.

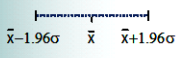
Conversely, about 2.5% can be expected to be above $\mu + 1.96\frac{\sigma}{\sqrt{n}}$ and 2.5% can be expected to be below $\mu - 1.96\frac{\sigma}{\sqrt{n}}$.

So 5% can be expected to fall outside the interval $\left[\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}\right]$.

- **95% Intervals around the Sample Mean**



Approximately 95% of the intervals $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$ around the sample mean can be expected to include the actual value of the population mean, μ. (When the sample mean falls within the 95% interval around the population mean.)



*5% of such intervals around the sample mean can be expected ***not*** to include the actual value of the population mean. (When the sample mean falls outside the 95% interval around the population mean.)

- **The 95% Confidence Interval for μ**

A 95% confidence interval for μ when σ is known and sampling is done from a normal population, or a large sample is used:

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

The quantity $1.96\frac{\sigma}{\sqrt{n}}$ is often called the ***margin of error*** or the ***sampling error***.

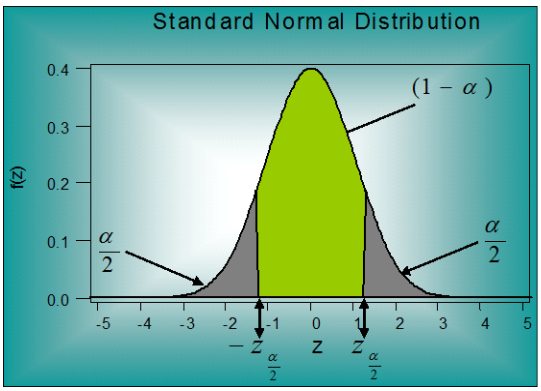
For example, if: n = 25
σ = 20
$\bar{x}$ = 122

A 95% confidence interval:

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}} = 122 \pm 1.96\frac{20}{\sqrt{25}}$$
$$= 122 \pm (1.96)(4)$$
$$= 122 \pm 7.84$$
$$= [114.16, 129.84]$$

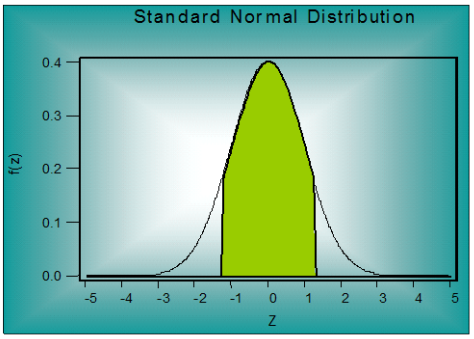
- **Critical Values of z and Levels of Confidence**

$(1-\alpha)$	$\frac{\alpha}{2}$	$z_{\frac{\alpha}{2}}$
0.99	0.005	2.576
0.98	0.010	2.326
0.95	0.025	1.960
0.90	0.050	1.645
0.80	0.100	1.282



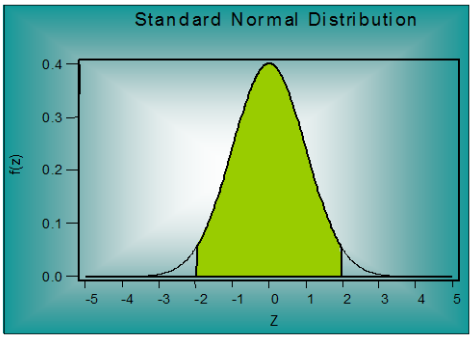
- **The Level of Confidence and the Width of the Confidence Interval**

When sampling from the same population, using a fixed sample size, the ***higher the confidence level, the wider the confidence interval.***



80% Confidence Interval:

$$\bar{x} \pm 1.28\frac{\sigma}{\sqrt{n}}$$

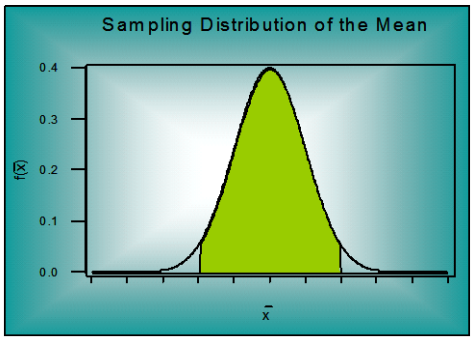


95% Confidence Interval:

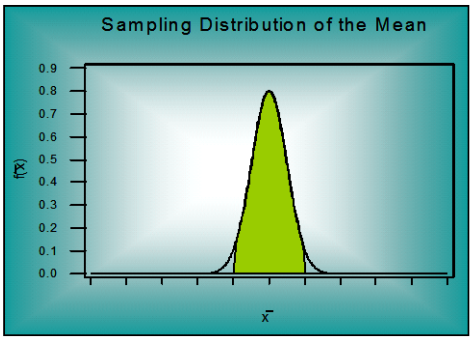
$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

- **The Sample Size and the Width of the Confidence Interval**

When sampling from the same population, using a fixed confidence level, the ***larger the sample size, n, the narrower the confidence interval.***



95% Confidence Interval: n = 20



95% Confidence Interval: n = 40

Hypothesis testing

(1) Definition: Hypothesis tests are procedures for making rational decisions about the reality of effects.

- *Statistical decision* – decision made about the population on the basis of sample information
- The *null hypothesis*, H_0 represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved.
- The *alternative hypothesis*, H_a, is a statement of what a statistical hypothesis test is set up to establish.
- The *significance level* (fixed level measured by α) of a statistical hypothesis test is a fixed probability of wrongly rejecting the null hypothesis H_0, if it is in fact true.
- A *test statistic* is a quantity calculated from our sample of data. Its value is used to decide whether or not the null hypothesis should be rejected in our hypothesis test. The choice of a test statistic will depend on the assumed probability model and the hypotheses under question.
- The *critical region* CR, or *rejection region* RR, is a set of values of the test statistic for which the null hypothesis is rejected in a hypothesis test.
- The *probability value (p-value*) is the probability of wrongly rejecting the null hypothesis if it is in fact true. It is similar as α but smaller, $\alpha = 0.05$, this would be reported as '$p < 0.05$'. Reject the null hypothesis if the p-value is less than the level of significance. You will fail to reject the null hypothesis if the p-value is greater than or equal to the level of significance. The P value is the **observed significance level** (or **P value**)--**the smallest fixed level at which the null hypothesis can be rejected**. If your personal fixed level is greater than or equal to the P value, you would reject the null hypothesis. If your personal fixed level is less than to the P value, you would fail to reject the null hypothesis. For example, if a P value is 0.027, the results are significant for all fixed levels greater than 0.027 (such as 0.05) and not significant for all fixed levels less than 0.027 (such as 0.01). A person who uses the 0.05 level would reject the null hypothesis while a person who uses the 0.01 level would fail to reject it.

(2) Procedure – steps

1. formulate the null hypothesis, H_0, to be tested
2. formulate the alternative hypothesis, H_a
3. determine an test statistic
4. determine the distribution of the test statistic
5. define the rejection region or critical region of the test statistic
6. calculate the test statistic
7. determine if the calculated value of the test statistic falls in the rejection region of the distribution of the test statistic

(3) Examples in test

- **one sample two tailed**

(a) Test if the mean of a normal distribution is μ_o, knowing that the variance is σ^2 (Z-test).

Solution:
Hypothesis: $H_0: \mu = \mu_0$ against $H_a: \mu \neq \mu_0$

Get a sample and calculate the mean and standard deviation ($\bar{x}, \sigma_x^2$)

Test statistic: $Z = \dfrac{(\bar{x} - \mu_o)}{\sigma_{\bar{x}}}$

H_0 is rejected if $|Z| = \left| \dfrac{(\bar{x} - \mu_o)}{\sigma_{\bar{x}}} \right| > z_{1-\alpha/2}$

Equivalent to say if $\bar{x}$ is outside of the confidence interval:

$\bar{x} < \mu_o - z_{1-\alpha/2}\sigma_{\bar{x}}$

or

$\bar{x} > \mu_o + z_{1-\alpha/2}\sigma_{\bar{x}}$

(b) Test if the mean of a normal distribution is μ_o, with unknown variance

(This is the famous T-test in the lecture)

$H_o : \mu = \mu_o,\ H_a : \mu \neq \mu_o$

$$t = \frac{(\bar{x} - \mu_o)}{s_{\bar{x}}}$$

H_0 is rejected if :

$$|t| = \left| \frac{(\bar{x} - \mu_o)}{s_{\bar{x}}} \right| > t_{1-\alpha/2}$$

Example 6.4

The annual runoff for a station for the period 1953-1970 has been calculated as 14.65cm and the standard deviation 4.75cm. Test the hypothesis that the true mean annual discharge is 16.5cm.

Solution:

Assume normal distribution and unknown variance since n is small.

$H_0: \mu = 16.5$; $H_a: \mu \neq 16.5$

$$t = \frac{(\bar{x} - \mu)}{s_{\bar{x}}} = \frac{(14.65 - 16.5)}{4.75/\sqrt{18}} = -1.65$$

Using α = 5% and $t_{1-\alpha/2,n-1} = t_{0.975,17} = 2.11$ (from appendix, table A13)

Since $|t| < t_{1-\alpha/2,n-1}$ we do not reject the hypothesis that the true mean might be 16.5

(4) Errors and power in hypothesis testing

Decision	Hypothesis true	Hypothesis false
Reject	Type I error	No error
Don't reject (accept)	No error	Type II error

- Type I error = significance level, α
- Type II error is denoted by β. The exact probability of a type II error is generally unknown. A type II error is frequently due to sample sizes being too small.
- Power of hypothesis test - measures the test's ability to reject the null hypothesis when it is actually false - that is, to make a correct decision. In other words, the power of a hypothesis test is the probability of not committing a type II error. It is calculated by subtracting the probability of a type II error from 1, usually expressed as:

Power = 1 - P(type II error) =(1-β).

The maximum power a test can have is 1, the minimum is 0.

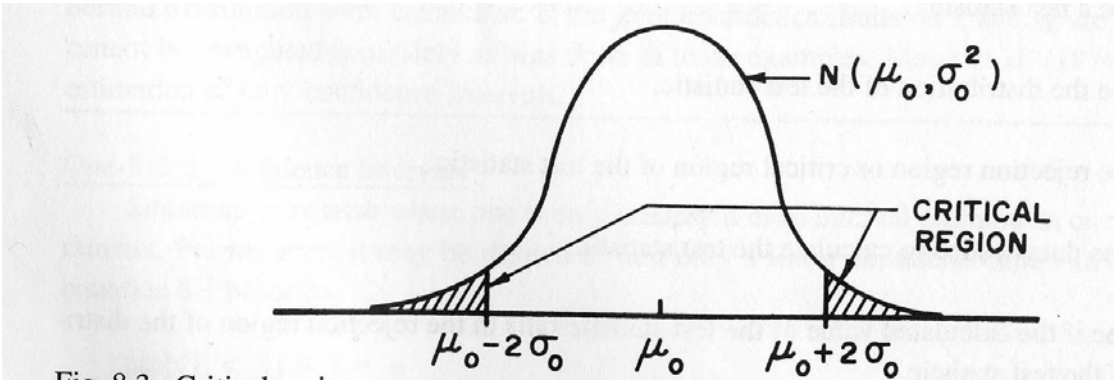


Figure 6.3: The critical region (shaded area) is the probability of rejecting a hypothesis that is actually true.

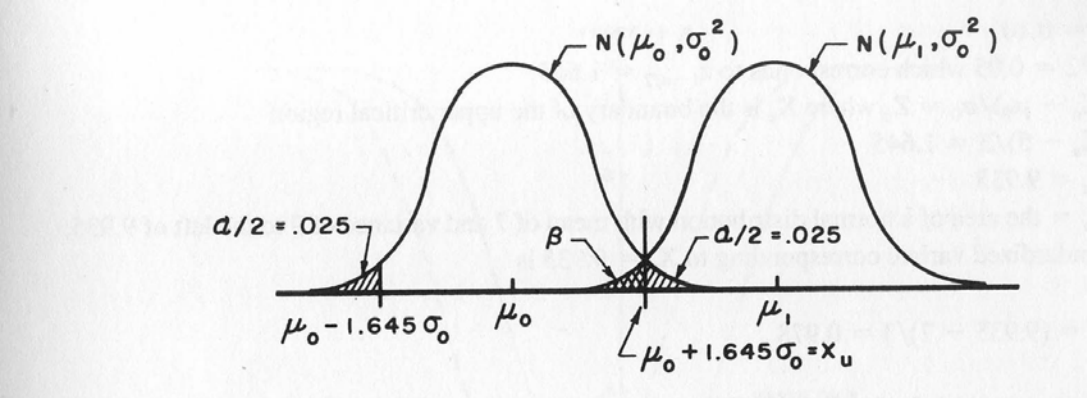
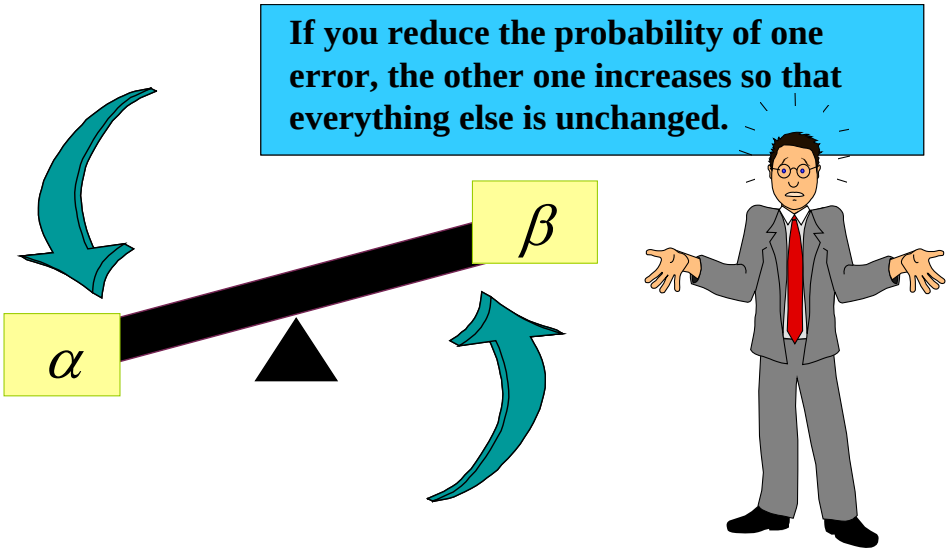


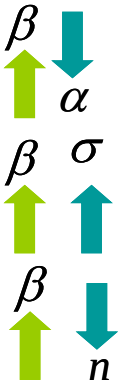
Figure 6.4: Illustration of α and β. β is (double-shaded area) the probability to fail to reject a hypothesis that is actually false.

Type I & II Errors Have an Inverse Relationship (FOR FIXED SAMPLE SIZE)



Factors Affecting Type II Error

- True value of population parameter
  - β Increases when the difference between hypothesized parameter and its true value decrease
- Significance level
  - β Increases when α decreases
- Population standard deviation
  - β Increases when σ increases
- Sample size
  - β Increases when n decreases



Example 6.5

Assume a single observation is selected from a normal distribution with mean μ=7 and variance σ_0^2=9. It is hypothesized that μ= μ_0 =5. If the test is conducted at the 10% significance level, what is β?

Solution:

α/2 =0.05, which corresponds to $z_{1-\frac{\alpha}{2}} = 1.645$

$(X_u-\mu_0)/s = z_{1-\alpha/2}$, where X_u is the boundary of the upper critical region

$(X_u - 5)/3 = 1.645$

X_u=9.935

A_u = the area of a normal distribution with mean of 7 and variance 9 to the left of 9.935.

The standardized variate corresponding to X_u = 9.935 is:

$z_u = (9.035 - 7)/3 = 0.978$

The area to the left of z_u = 0.978 from a standard normal distribution is 0.8365.

Similarly, if X_l is the boundary of the lower critical region, we have:

$(x_l -5)/3 = -1.645=0.0645$.

A_1 is the area of a normal distribution with mean 7 and variance 9 to the left of 0.0645.

$z_l=(0.0645 -7)/3= -2.31$

A_l=0.0104

Now, $\beta = A_u - A_l = 0.8365 - 0.0104 = 0.8261$.

Thus, yje probability of accepting the hypothesis that $\mu=5$ when in fact $\mu=7$ is 0.8261 when α is 0.1.
The probability of Type II error is 0.8261.

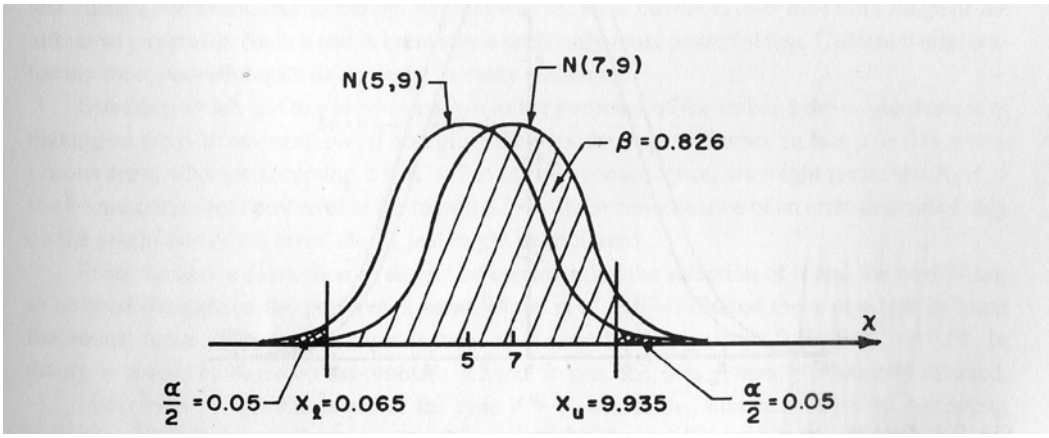


Figure 6.5: Illustration of β

(5) More examples of hypothesis testing

- **One sample, one tail. Test on mean, normal distribution,**

  (a) Known variance:

  $\boxed{H_o : \mu = \mu_1,\ H_a : \mu = \mu_2}$ or $\boxed{H_o : \mu = \mu_o,\ H_a : \mu > \mu_o \text{ or } \mu < \mu_o}$

$$Z = \frac{(\bar{x} - \mu_1)}{\sigma_{\bar{x}}}$$

If $\mu_1 > \mu_2$, then H_0 is rejected if:

$\bar{x} \le \mu_1 - z_{1-\alpha}\sigma_{\bar{x}}$

if $\mu_1 < \mu_2$, then H_0 is rejected if:

$\bar{x} \ge \mu_1 + z_{1-\alpha}\sigma_{\bar{x}}$

  (b) Unknown variance:

  $\boxed{H_o : \mu = \mu_1,\ H_a : \mu = \mu_2}$ or $\boxed{H_o : \mu = \mu_o,\ H_a : \mu > \mu_o \text{ or } \mu < \mu_o}$

Test statistic: $t = \dfrac{(\bar{x} - \mu_1)}{s_{\bar{x}}}$

If $\mu_1 > \mu_2$, then H_0 is rejected if:

$\bar{x} \le \mu_1 - t_{1-\alpha,n-1} s_{\bar{x}}$

If $\mu_1 < \mu_2$, then H_0 is rejected if :

$\bar{x} \ge \mu_1 + t_{1-\alpha,n-1} s_{\bar{x}}$

Example 6.6

The annual runoff for a station for the period 1953-1970 has been calculated as 14.65cm and the standard deviation 4.75cm. Can the mean annual discharge be 16.5 in reality? (i.e. can the population mean be 16.5 instead of 14.65 as sample shows).

Solution:

$\alpha = 5\%$, $n=18$
Assume normal distribution and unknown variance since n is small.
H_0: $\mu=\mu_1= 16.5$; H_α: $\mu=\mu_2=14.65$; $\mu_1>\mu_2$

$$t = \frac{(\bar{x} - \mu)}{s_{\bar{x}}} = \frac{(14.65 - 16.5)}{4.75/\sqrt{18}} = -1.65$$

$t_{1-\alpha,n-1} = t_{0.95,17} = -1.74$ (from appendix, table A13):

$$\mu_1 - t_{1-a,n-1} \cdot s_{\bar{x}} = 16.5 - 1.74 \cdot \frac{4.75}{\sqrt{18}} = 14.55$$

$\Rightarrow \quad \bar{x} > \mu_1 - t_{1-a,n-1} \cdot s_{\bar{x}}$

Thus, we do not reject the hypothesis that the population mean may be 16.5

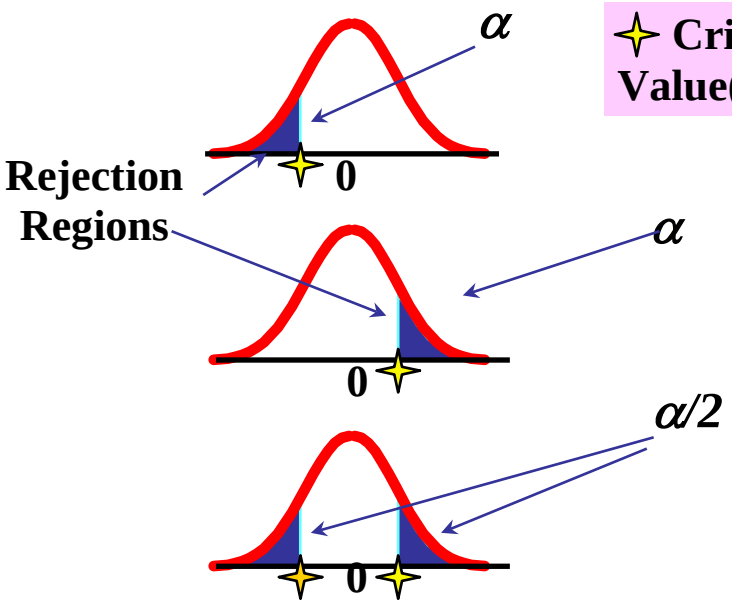
Compare with the previous example, here we test if the mean 16.5 instead of 14.65, in the previous example, we test if the mean is 16.5 nothing else..

Example: level of significance and the rejection region

$H_0: \mu \geq \mu_o$
$H_1: \mu < \mu_o$

$H_0: \mu \leq \mu_o$
$H_1: \mu > \mu_o$

$H_0: \mu = \mu_o$
$H_1: \mu \neq \mu_o$



- Two-sample, two-tail: if the means of two normal distributions are significantly different.

(a) Known variances:

$x_1 \sim N(\mu_1, \sigma_1^2)$ $\quad x_2 \sim N(\mu_2, \sigma_2^2)$

$$H_o: \mu_1 - \mu_2 = \delta,\ H_a: \mu_1 - \mu_2 \neq \delta$$ $\quad \delta = 0$

Test statistic: $$z = \frac{(\bar{x}_1 - \bar{x}_2 - \delta)}{\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}}$$

H_0 rejected if: $|z| > z_{1-\alpha/2}$

(b) Unknown variances:

Test statistic: $$t = \frac{(\bar{x}_1 - \bar{x}_2 - \delta)}{\sqrt{(s_1^2/n_1 + s_2^2/n_2)}}$$

H_0 rejected if: $|t| > \frac{w_1 t_1 + w_2 t_2}{w_1 + w_2}$

where: $w_1 = s_1^2 / n$

$w_2 = s_2^2 / n$

$t_1 = t_{1-\alpha/2, n_1 - 1}$

$t_2 = t_{1-\alpha/2, n_2 - 1}$

- One sample, two tail test for the variance of normal distribution

$H_o : \sigma^2 = \sigma_o^2$; $H_a : \sigma^2 \neq \sigma_o^2$

Test statistic: $\chi_c^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma_o^2} = \frac{(n-1)S^2}{\sigma_o^2}$

H_0 rejected if: $\chi_c^2 > \chi_{1-\alpha/2, n-1}^2$ or $\chi_c^2 < \chi_{\alpha/2, n-1}^2$

Example 6.7

For the example 6.6 test the hypothesis that the variance is 36.

Solution:

$H_o : \sigma^2 = \sigma_o^2 = 36$; $H_a : \sigma^2 \neq \sigma_o^2$

$$\chi_c^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma_o^2} = \frac{(n-1)S^2}{\sigma_o^2} = \frac{17(4.75)^2}{36} = 10.65$$

From a Chi-square table:

$\chi_{\alpha/2, n-1}^2 = \chi_{0.025,17}^2 = 7.6$

$\chi_{1-\alpha/2, n-1}^2 = \chi_{0.975,17}^2 = 30.2$

Since 10.65 is neither smaller than 7.6 nor bigger than 30.2, H_0 is not rejected.

- **The F-test –** testing whether two samples come from the populations with the same variance

Used to test a null hypothesis that the variance of one population is equal to that of a second population $\sigma_{pop1}^2 = \sigma_{pop2}^2$

$H_o : \sigma_1^2 = \sigma_2^2$; $H_a : \sigma_1^2 \neq \sigma_2^2$

The F statistic is a simple ratio of the two unbiased estimator of population variance:

$$F_c = s_1^2 / s_2^2$$

where F is distributed as F distribution with n_1-1 and n_2-1 degrees of freedom, and $s_1^2 > s_2^2$.

The F value will be close to one if the variances of two populations are equal or close.
H_0 is rejected if :

$F_c > F_{1-\alpha,n1-1,n2-1}$ (table A15)

- Test for equality of variances from several Normal distributions

  To test the $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ for k independent samples each form a normal population with mean μ_i and variance σ_i^2, it is first necessary to calculate the k sample variances S_i^2. The quantity Q/h is approximately distributed as a chi-square distribution with k-1 degrees of freedom, where:

$$Q = \sum_{i=1}^{k}(n_i - 1) \cdot \ln[\sum_{i=1}^{k}(n_i - 1) \cdot \frac{S_i^2}{N-k}] - \sum_{i=1}^{k}(n_i - 1) \cdot \ln(S_i^2)$$

$$h = 1 + \frac{\sum_{i=1}^{k}\left[\frac{1}{n_i - 1}\right] - \frac{1}{N-k}}{3(k-1)}$$

$$N = \sum_{i=1}^{k} n_i$$

  H_0 is rejected if:
  $\frac{Q}{h} > \chi^2_{1-\alpha,k-1}$

  In this test H_a is that at least one σ_i^2 is different from the other σ_i^2. The test is known as Bartlett's test for homogeneity of variance.

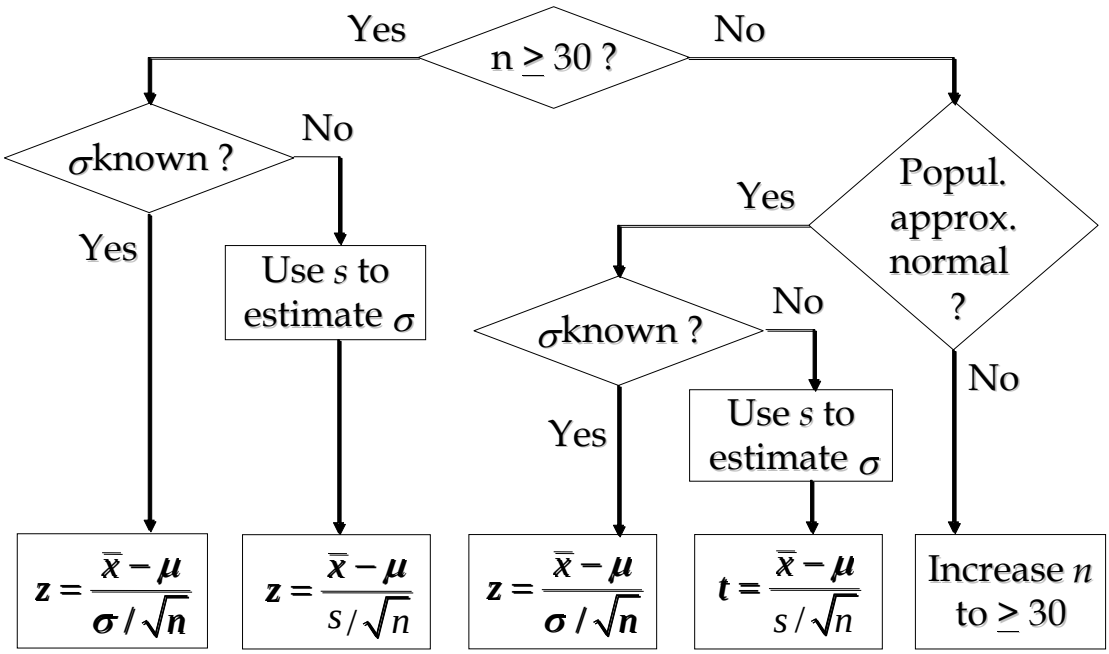
Summary and Review: Hypothesis Testing Facts

- **Hypotheses**:
  - *Null Hypothesis* H_0: The accepted explanation, status quo. This is what we're trying to disprove.
  - *Alternate Hypothesis* H_a: What the researcher or scientist thinks might really be going on, a (possibly) better explanation than the null hypothesis.
- **Test**:
  - The goal of the test is to reject H_0 in favor of H_a. We do this by calculating a *test statistic* and comparing its value with a value from a table in the book, the *critical value*.

- If our test statistic is more extreme than our critical value, then it falls within the *rejection region* of our test and we reject H_0. We can set up the rejection region before computing our test statistic.

- **Decisions**:
  - Reject H_0.
  - Fail to reject H_0.
- **Errors**:
  - *Type I:* Reject H_0 when H_0 is really true.
  - *Type II:* Fail to reject H_0 when H_0 is really false.
- **Z-test or T-test:**
  When we test the mean, both one sample test and two sample test, we used either Z-test (when we assume the variance is known) or T-test (when the variance is unknown). These are called parametric test.
- **The basic assumptions of T test are**:
  - Normality – the population from which the samples are drawn was assumed to be normally distributed.
  - Homogeneity of variance – the variance of the population from which the samples are drawn was assumed to be constant.
  - Independence – each individual score is assumed not to be influenced by any other score.

Summary of Test Statistics to be Used in a Hypothesis Test about a Population



Procedure of one sample parametric test

Hypothesis test for population mean μ_x		Hypothesis test for population variance σ_x^2
z-test • population is normally distributed and sample size is large $n \geq 30$ • σ_x^2 is known	t-test • population is normally distributed and sample size is small $n < 30$ • σ_x^2 is <u>not</u> known	χ^2 -test • population is normally distributed • σ_x^2 is known
Step 1: state the null and alternative hypothesis Two tail test: $H_0: \mu = \mu_0$ against $H_\alpha: \mu \neq \mu_0$ One tail test: $H_0: \mu = \mu_0$ against $H_\alpha: \mu < \mu_0$ or $\mu > \mu_0$		**Step 1**: state the null and alternative hypothesis Two tail test: Ho: $\sigma^2 = \sigma_o^2$ against $H_\alpha: \sigma^2 \neq \sigma_o^2$ One tail test: $H_0: \sigma^2 = \sigma_o^2$ against $H_\alpha: \sigma^2 < \sigma_o^2$ or $\sigma^2 > \sigma_o^2$
Step 2: choose the level of significance: α, the probability of making a type I error if H_0 is true		
Step 3: determine the <u>critical values</u> for the level of significance α, Two tail test: find $z_{\alpha/2}$ **or** $z_{1-\alpha/2}$ in the z-table One tail test: find z_α **or** $z_{1-\alpha}$ in the z-table Illustrate the non-rejection and rejection regions	**Step 3**: determine the <u>critical values</u> for the level of significance α, and degree of freedom = (n-1) Two tail test: find $t_{\alpha/2,n-1}$ **or** $t_{1-\alpha/2,n-1}$ in the t-table One tail test: $t_{\alpha,n-1}$ **or** $t_{1-\alpha,n-1}$ in the t-table Illustrate the non-rejection and rejection regions	**Step 3**: determine the <u>critical values</u> for the level of significance α, Two tail test: find $\chi^2_{\alpha/2,n-1}$ **and** $\chi^2_{1-\alpha a/2,n-1}$ in the χ^2 -table One tail test: find $\chi^2_{\alpha,n-1}$ **or** $\chi^2_{1-\alpha,n-1}$ in the χ^2 -table Illustrate the non-rejection and rejection region
Step 4: calculate test statistic $Z_c = \frac{(\bar{x}-\mu_o)}{\sigma_{\bar{x}}} = \frac{(\bar{x}-\mu_o)}{\sigma_x/\sqrt{n}}$ Where $\bar{x}, \sigma_x$ are calculated from sample data	**Step 4**: calculate test statistic $t_c = \frac{(\bar{x}-\mu_o)}{s_{\bar{x}}} = \frac{(\bar{x}-\mu_o)}{s_x/\sqrt{n}}$ Where $\bar{x}, s_x$ are calculated from sample data	**Step 4**: calculate test statistic $\chi_c^2 = \frac{(n-1)s_x^2}{\sigma_o^2}$ Where s_x is calculated from sample data
Step 5: compare the test statistic value with the critical values: where does your test statistic fall? In a rejection or non-rejection region?	**Step 5**: compare the test statistic value with the critical values: where does your test statistic fall? In a rejection or non-rejection region?	**Step 5**: compare the test statistic value with the critical values: where does your test statistic fall? In a rejection or non-rejection region?
Step 6: decision For two tail test: if $\lvert Z_c \rvert > z_{1-\alpha/2}$ reject H_0 For one tail test: if $Z_c > z_{1-\alpha}$ or $Z_c < z_\alpha$ reject H_0	**Step 6**: decision For two tail test: if $\lvert t_c \rvert > t_{1-\alpha/2,n-1}$ reject H_0 For one tail test: if $t_c > t_{1-\alpha,n-1}$ or $t_c < t_{\alpha,n-1}$ reject H_0	**Step 6**: decision For two tail test: if $\chi_c^2 > \chi^2_{1-\alpha/2,n-1}$ or $\chi_c^2 < \chi^2_{\alpha/2,n-1}$ reject H_0 For one tail test: if $\chi_c^2 > \chi^2_{1-\alpha,n-1}$ or $\chi_c^2 < \chi^2_{\alpha,n-1}$ reject H_0

Chapter 7

Testing the goodness of fit of data to probability distributions

1. The chi-square test

- Chi-square distribution

If we draw all possible samples of size *n* from a normal population and plot $\sum z^2$, they would form the Chi-square χ^2 distribution as mentioned earlier.

The χ^2 distribution in hydrology is used to compare whether the distribution of the data is same with or different from the predetermined distribution and to test the significance of the similarity/dissimilarity.

$$\chi_c^2 = \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^2$$

The χ^2 test is most often used on data that have been grouped into classes. Assuming that our observations have been grouped into *k* classes, the test statistic is found as:

$$\chi_c^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

where O_j and E_j are the number of observed and expected values in the *j*'th class.

- **Procedure of using χ^2 test**
  - Group data into class (5~10)
  - Count the number of observations in each class, denote it as Oj
  - Determine the expected number of observations in each class which equals to the expected relative frequency multiplying by the total number of observations (for details see Lecture 4 and Table 5.1 and Figure 5.5 in page 108, Haan), denote it as Ej
  - Calculate the test statistic:

$$\chi_c^2 = \sum_{j=1}^{n} \frac{(O_j - E_j)^2}{E_j}$$

  - The null hypothesis that the data are from the specified distribution is rejected if:

$$\chi_c^2 > \chi_{1-\alpha,k-p-1}^2$$

where: α is significance level, k-p-1 degrees of freedom, k the number of classes, p the number of parameters of the specified distribution.

Example 7.1

The annual peak flow of Kentucky River is measured for 66 years (Haan, *table 2.1*). At significance level α=0.05 test the null hypothesis that the annual peak flow is normally distributed.

Solution:

H_0: data normally distributed; H_a: data not normally distributed

Test statistic: $\chi_c^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i$

Reject H_0 if: $\chi_c^2 = \sum \frac{(O-E)^2}{E} > \chi_{1-\alpha,k-p-1}^2$

N= 66
Q_{mean}= 67.5 (calculated from data)
σ=21.0 (calculated from data)

The expected relative frequency, e.g. of the first class, is calculated as:

$$f_{x_i} = P(20 < x < 30) = P\left(z < \frac{30 - 67.5}{21}\right) - P\left(z < \frac{20 - 67.5}{21}\right)$$

$$\Rightarrow f_{x_i} = P(z < -1.78) - P(z < -2.26) = [1 - P(z < 1.78)] - [1 - P(z < 2.26)]$$

$$\Rightarrow f_{x_i} = 0.026$$

O = observed number of data in each class
E = expected number of data in each class= $f_{x_i} \cdot N$

Table 7.1:Calculation of χ_c^2 test statistic

Class limits	Class mark	O	f_{x_i}	E	$\frac{(O-E)^2}{E}$
<30	25	2	0.026	1.7	0.00
30 - 40	35	3	0.057	3.8	0.25
40 - 50	45	10	0.107	7.1	1.29
50 - 60	55	9	0.159	10.5	0.10
60 - 70	65	11	0.189	12.5	0.08
70 - 80	75	10	0.178	11.7	0.33
80 - 90	85	12	0.135	8.9	1.00
90 - 100	95	6	0.08	5.3	0.20
100 - 110	105	0	0.038	2.5	3.00
>110	115	3	0.015	1.0	4.00
			Sum:	64.9	**10.25**

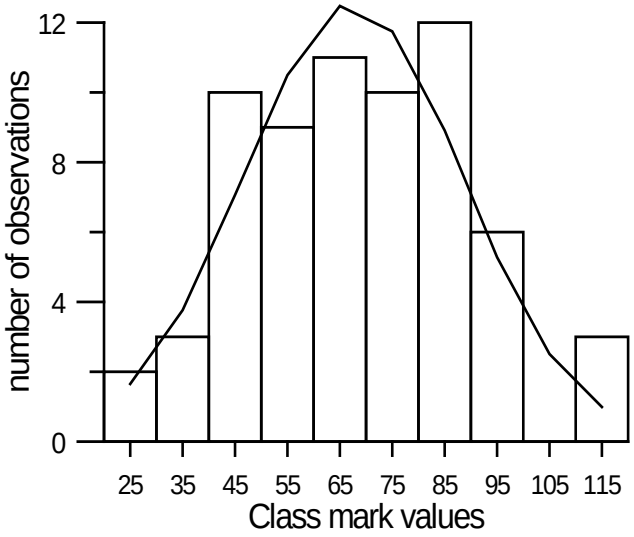


Figure 7.1: observed (histogram) and expected (line) number of observations in each class

$\Rightarrow \quad \chi_c^2 = 10.25$

k= number of classes = 10
p= number of parameters in the assumed distribution (normal) = 2 (mean and variance)

!**Note** that the chi-square table in the appendix gives the values for $\chi^2_{\alpha,k-p-1}$, therefore we should look up:

$$\chi^2_{\alpha,k-p-1} = \chi^2_{0.05,10-2-1} = \chi^2_{0.05,7} = 14.1$$

$\Rightarrow \chi_c^2 < \chi^2_{0.05,7}$,

and therefore we cannot reject the hypothesis that the data are normally distributed.

Example 7.2

Test if the die is fair.
If the die is fair, the chance for each side up shall be the same if it is tossed many times (the expected number is uniformly distributed). An example is done by tossing 60 time and the outcomes are:

Table 7.2: Outcomes of tossing the die 60 times

Side	Number of times
1	4
2	6
3	17
4	16
5	8
6	9

Solution:

H_0: the die is fair; H_a: the die is not fair

Decision rule: Reject H_0 if $\chi_c^2 > \chi^2_{1-\alpha,k-p-1}$
where α=0.05, k=6, p=0

Table 7.3: Calculation of χ_c^2 test statistic

Side	Observed frequency (O)	Expected frequency (E)	$\frac{(O-E)^2}{E}$
1	4	10	3.6
2	6	10	1.6
3	17	10	4.9
4	16	10	3.6
5	8	10	0.4
6	9	10	0.1
		Sum:	**14.2**

$\Rightarrow \chi_c^2 = 14.2$
$\chi^2_{0.95,5} = 11.1$

Since $\chi_c^2 > \chi^2_{0.95,5}$, the hypothesis that the die is fair is rejected.

Example 7.3

Test if the sex of the new born babies in a city is biased.
In a city there are 442 babies born in two months, of which 224 are boys and 218 girls, test if the sex of babies is biased at 5% significance level. (i.e. test if the sample is uniformly distributed).

Solution:

H_0: sample is not biased; H_a: sample is biased
Test statistic: $\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

Decision rule: Reject H_0 if $\chi_c^2 > \chi^2_{1-\alpha,k-p-1}$
where α=0.05, k=2, p=0

Table 7.4:Calculation of χ_c^2 test statistic

Class	Expected number (E)	Observed number (O)	$\frac{(O-E)^2}{E}$
Female	221	224	0.041
Male	221	218	0.041
		Sum:	**0.082**

$\Rightarrow \chi_c^2 = 0.082$
$\chi^2_{1-0.05,2-0-1} = \chi^2_{0.95,1} = 3.8$

Since $\chi_c^2 < \chi^2_{0.95,1}$, we cannot reject the hypothesis that the sample is not biased.

2. Kolmogorov-Smirnov test

The **Kolmogorov-Smirnov test** is used to compare whether the distribution of the data is same with or different from the predetermined distribution and to test the significance of the similarity/dissimilarity.

The test is based on the comparison between the cumulative observed and expected theoretical frequency curve of the distribution to be tested.

Procedure:

- Calculate the expected cumulative frequency (values of cumulative distribution function), $F^e(x)$
- Calculate the sample cumulative frequency, $F^o(x)$=k/n, where k is the number of observation less than or equal to x, and n is the total number of observation
- Determine the maximum deviation, D
  $D = \max\left|F^e(x) - F^o(x)\right|$
- If, for the chosen significance level, the value D is greater than or equal to critical tabulated value of the Kolmogorov-Smirnov statistic, the hypothesis that the data fit the tested distributed is rejected.

$$\mathbf{H_0: F^o(x) = F^e(x)\ ;\ \ H_a: F^o(x) \neq F^e(x)}$$

In the following example, we test if the data is normally distributed. The method can be used to test if the data fits other distributions, like, lognormal, exponential, Pearson, etc. In this case, when calculating $F^e(x)$, the distribution function that is to be tested is used.

Example 7.4

The annual peak flows Q(l/s) for 11 years are given in table 7.5. Test the null hypothesis that Q is normally distributed, using Kolmogorov – Smirnov test.

Solution:

Q_{mean}=53.83
s=24.63

In the table 7.5 are shown:
$F^o(x)$ = observed cumulative frequency = $m_i/11$
z = standardized Q = $\frac{Q_i - Q_{mean}}{s}$
$F^e(x)$ = expected cumulative frequency = $P(Z<z_i)$ (from normal distribution table)
D=deviation of observed from the expected cumulative frequency = $|F^e(Q) - F^o(Q)|$

Table 7.5: Calculating the deviation

year	Q	Q ordered	m	$F^o(Q)$	z	$F^e(Q)$	D
1981	32.1	9.2	1	0.091	-1.81	0.036	0.055
1982	41.8	22.5	2	0.182	-1.27	0.102	0.079
1983	77.6	32.1	3	0.273	-0.88	0.189	0.0833
1984	71.8	41.8	4	0.364	-0.49	0.312	0.051
1985	82.5	50.3	5	0.455	-0.14	0.444	0.010
1986	50.3	58.5	6	0.545	0.19	0.575	0.029
1987	22.5	71.8	7	0.636	0.73	0.767	**0.131**
1988	71.8	71.8	8	0.727	0.73	0.767	0.040
1989	9.2	74	9	0.818	0.82	0.791	0.027
1990	74	77.6	10	0.909	0.97	0.834	0.075
1991	58.5	82.5	11	1	1.16	0.877	0.123

$\Rightarrow \max|D| = 0.131$
For $\alpha=0.05$ and $n=11$: $D_{critical} = 0.391$ (from appendix, table A16)

$\Rightarrow \max|D| < D_{critical}$

And therefore the hypothesis that the data are normally distributed cannot be rejected.
Note though that the hypothesis is rejected when performing the test in Matlab using the internal function kstest(Q).

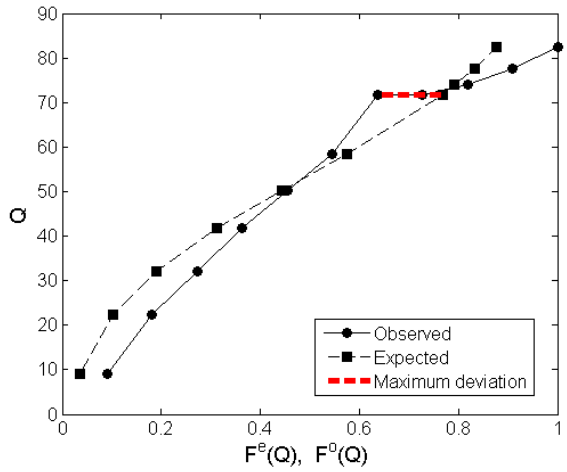


Figure 7.2: Observed and expected cumulative frequencies.

Chapter 8

Correlation and simple regression

1. Simple linear regression

Regression analysis is a statistical tool that utilizes the relation between two or more quantitative variables so that one variable (dependent variable) can be predicted from the others (independent variables).

The simplest model of such is:

$$Y_i = a + bX_i + e_i$$

and:

$$\hat{Y}_i = a + bX_i$$

Where: e_i is the error term or residual of the regression line; X_i and Y_i are the observed independent and dependent variables respectively; $\hat{Y}_i$ are the values estimated from the regression line; *a* and *b* are regression coefficients where *b* is called the *slope* of the line and *a* is the *y-intercept*. The slope measures the amount *Y* increases/decreases when *X* increases/decreases by one unit. The *y-intercept* is the value of *Y* when *X*=0.

Our objective is to fit a straight line to points on a scatterplot (see fig.8.1). So we want to find *a* and *b* such that the line $\hat{Y}_i = a + bX_i$ fits the data as well as possible.

We need to define what we mean by a "best" fit. We want a line that is in some sense closest to all of the data points simultaneously. In statistics, we define a *residual*, e_i, as the vertical distance between a measured point and the fitted line:

$$e_i = Y_i - \hat{Y}_i = Y_i - (a + bX_i)$$

Since residuals can be positive or negative, we will square them to remove the sign. By adding up all of the squared residuals, we get a measure of how far away from the data our line is. Thus, the "best" line will be the one which has the minimum sum of squared residuals, i.e., *min*($\sum e_i^2$). This method of fitting a line is called *least squares*.

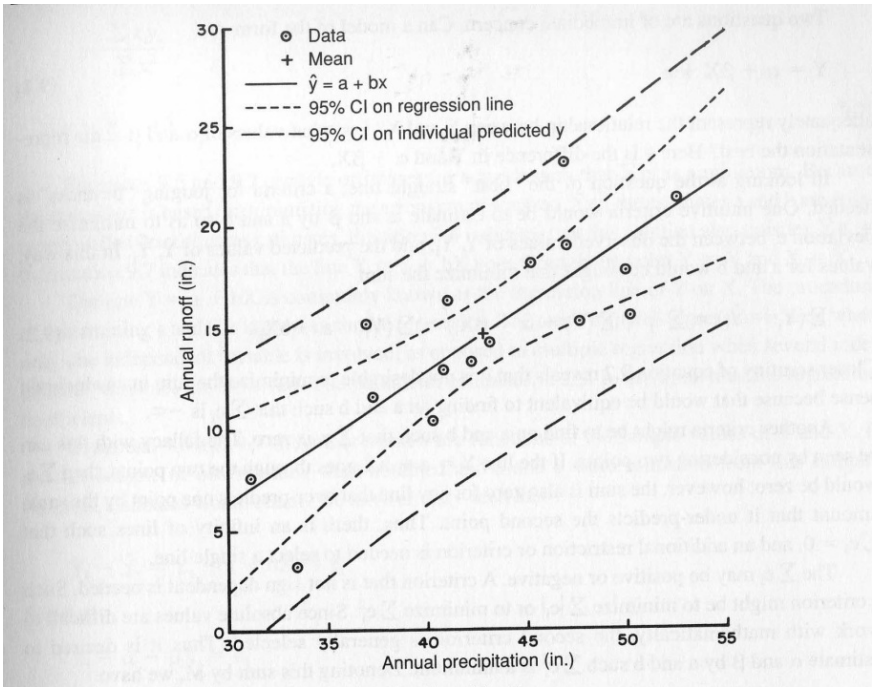


Figure 8.1: A typical plot of simple regression: Annual rainfall –runoff relation for Cave Creek

2. Tasks in regression

a) Calculation of regression coefficients: estimation of the coefficients *a* and *b.*
b) Model evaluation
c) Calculation of confidence interval – testing of regression line, coefficients, etc
d) Significance of coefficients

a) Parameter estimation:

If we denote the sum of squared residuals as M, we have:

$$\mathrm{M} = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - a - bX_i)^2$$

We know that the best fit is achieved when this sum is minimized. Also, M has a minimum for:

$$\frac{\partial M}{\partial a} = 0, \quad \frac{\partial^2 M}{\partial a^2} > 0$$

and:

$$\frac{\partial M}{\partial b} = 0, \quad \frac{\partial^2 M}{\partial b^2} > 0$$

For the first derivative condition we have:

$$\frac{\partial M}{\partial a} = -2\sum(Y_i - a - bX_i) = 0$$

$$\frac{\partial M}{\partial b} = -2\sum X_i(Y_i - a - bX_i) = 0$$

For the second derivative condition we have:

$$\frac{\partial^2 M}{\partial a^2} = 2 > 0$$

$$\frac{\partial^2 M}{\partial b^2} = 2\sum X_i^2 > 0$$

From the first derivative condition we get:

$$\sum(Y_i - a - bX_i) = 0$$

$$\sum X_i(Y_i - a - bX_i) = 0$$

Solving the linear system of these two equations we have:

$$b = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{cov}(X,Y)}{\text{var}(X)}$$

$$a = \overline{Y} - b\overline{X}$$

$$\boxed{\begin{array}{c} b = \dfrac{\text{cov}(X,Y)}{\text{var}(X)} \\ \\ a = \overline{Y} - b\overline{X} \end{array}}$$

b) Model evaluation

- **Coefficient of determination**

After fitting a line to the data-points, we want to know how much of the variability in the dependent variable (Y) is explained by the regression. For this, the coefficient of determination (R^2) is often used, and is expressed as:

$$R^2 = \frac{Explained\ variance}{Total\ variance}$$

The variability in the dependent variable Y is quantified as a sum of squares:

$\sum (Y_i - \bar{Y})^2$ = total sum of squares corrected for the mean = total variance

$\sum (\hat{Y}_i - \bar{Y})^2$ = The squared deviations of the predicted values from the mean value, explained variance by the regression line

$\sum (Y_i - \hat{Y}_i)^2$ = the sum of squares of deviation from the regression = unexplained variance

A relationship between the above variances can be derived by using the definitions of *a* and *b* coefficients and a small mathematical trick: $Y_i = Y_i + \hat{Y}_i - \hat{Y}_i + \bar{Y} - \bar{Y}$

$$\Rightarrow \sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

Finally, R^2 can be expressed as:

$$R^2 = \frac{\text{explained variance}}{\text{total variance}} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

Another way to express R^2 is:

$$R^2 = b \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2} = \frac{\left(\sum (X_i - \bar{X}) \sum (Y_i - \bar{Y})\right)^2}{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}$$

$$= b^2 \frac{\sum (X_i - \bar{X})^2}{\sum (Y_i - \bar{Y})^2} = b^2 \frac{s_x^2}{s_y^2}$$

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2} = b^2 \frac{s_x^2}{s_y^2}$$

- R^2 ranges from 0 to 1 ($r = b\frac{s_x}{s_y}$ ranges from -1 to +1)
- Expressed as a percentage, it represents the proportion that can be predicted by the regression line.
- The value 1 - R^2 is therefore the proportion of variance contributed by other factors.

- **Standard Error of Estimate (SEE)**

  - A measure of the variability of the regression line, i.e. the dispersion around the regression line.
  - It tells how much variation there is in the dependent variable between the raw value and the expected value in the regression:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

  - This SEE allows us to generate the confidence interval on the regression line as we did in the estimation of means

- **Standard error (deviation) for *a* (σ_a)**

$$\sigma_a = \sqrt{\text{var}(a)}, \qquad \text{var}(a) = S^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{t=1}^{n}(x_t - \overline{x})^2}\right]$$

- **Standard error (deviation) for *b* (σ_b)**

$$\sigma_b = \sqrt{\text{var}(b)}, \qquad \text{var}(b) = \frac{S^2}{\sum_{t=1}^{n}(x_t - \overline{x})^2}$$

c) Confidence intervals

- **Confidence interval for *a*:**

$$L = a - t_{1-\alpha/2,n-2}s_a$$
$$U = a + t_{1-\alpha/2,n-2}s_a$$

- **Confidence interval for b:**

$$L = b - t_{1-\alpha/2,n-2}s_b$$
$$U = b + t_{1-\alpha/2,n-2}s_b$$

- **Confidence interval on regression line**

$$l = \hat{y}_k - s_{\hat{y}_k} t_{1-\alpha/2,n-2}$$
$$u = \hat{y}_k + s_{\hat{y}_k} t_{1-\alpha/2,n-2}$$

where: $\hat{y}_k = a + bx_k$

$$s_{\hat{y}_k} = s\sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2}}$$

By substituting various values of x_k into these equations, the desired confidence limits are obtained. The confidence intervals are the narrowest at $x_k = \bar{x}$

- **Confidence interval on individual points**

$$l = \hat{y}_k - s_{\hat{y}_k} t_{1-\alpha/2,n-2}$$
$$u = \hat{y}_k + s_{\hat{y}_k} t_{1-\alpha/2,n-2}$$

where: $\hat{y}_k = a + bx_k$

$$s_{\hat{y}_k} = s\sqrt{1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2}}$$

- **Confidence interval for standard error and variance**

$$\frac{(n-2)S^2}{\chi^2_{1-\alpha/2,n-2}} < \sigma^2 < \frac{(n-2)S^2}{\chi^2_{\alpha/2,n-2}}$$

where: $S^2 = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$

d) Significance of coefficients

- **Test for significance of *a***

H_o: $a = a_o$; H_a: $a \neq a_o$

For example $a_o = 0$; test statistic: $t = \dfrac{a-0}{s_a}$

H_o rejected if $|t| \geq t_{1-\alpha/2,n-2}$

- **Test for significance of *b***

H_o: $b = b_o$; H_a: $b \neq b_o$ (for example $b_o = 0$)

For example $b_o = 0$; test statistic: $t = \dfrac{b-0}{s_b}$

H_o rejected if $|t| \geq t_{1-\alpha/2,n-2}$

Example 8.1

Here is an example of simple regression.

1) **Calculation of regression coefficients**
2) **Calculation of R^2 value**
3) **Calculation of confidence intervals**
4) **Testing the significance of coefficients**

1) Calculation of regression coefficients:

Day	y_i	x_i	$y_i - \bar{y}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	8.2	6.5	5.83	6.02	36.29	35.14
2	6.4	7.8	4.03	4.82	23.27	19.46
3	5.9	5.7	3.53	2.72	7.42	9.62
4	6.3	6.6	3.93	3.62	13.13	14.25
5	2.8	4.5	0.43	1.52	2.32	0.66
6	4.3	4.0	1.93	1.02	1.05	1.98
7	6.0	6.8	3.63	3.82	14.62	13.89
...	...	...	...	...	...	...
mean	2.4	3.0				
∑	0.6	3.3	0	0	488	454

Regression formula:

$$y = a + b \cdot x$$

$$b = \frac{\sum (y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \Rightarrow b = \frac{453.75}{488.32} = 0.929$$

$$a = \bar{y} - b \cdot \bar{x} \Rightarrow a = 2.4 - 0.929 \cdot 3.0 = -0.399$$

2). Calculation of R2 value

Day	y_i	x_i	$y_i - \bar{y}$	$\hat{y}_i - \bar{y}$	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$
1	8.2	6.5	5.83	5.56	34.03	31.33
2	6.4	7.8	4.03	4.45	16.27	20.09
3	5.9	5.7	3.53	2.5	12.48	6.41

⋮

⋮

Day	y_i	x_i	$y_i - \bar{y}$	$\hat{y}_i - \bar{y}$	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$
4	6.3	6.6	3.93	3.33	15.47	11.34
5	2.8	4.5	0.43	1.38	0.19	2.00
6	4.3	4.0	1.93	0.92	3.74	0.91
7	6.0	6.8	3.63	3.52	13.20	12.62
...	...	...	...	...	...	...
mean	2.4	3.0				
∑	0.6	3.3	0	0	435.83	421.64

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \Rightarrow \frac{421.64}{435.83} = 0.9674$$

3) Calculation of confidence intervals

- Confidence limits on regression line:

$$L_{lower}^{upper} = a + b \cdot x_k \pm s \cdot \left(\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right)^{\frac{1}{2}} \cdot t_{1-\alpha/2;n-2}$$

- Confidence limits for individual predicted values:

$$L_{lower}^{upper} = a + b \cdot x_k \pm s \cdot \left(1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right)^{\frac{1}{2}} \cdot t_{1-\alpha/2;n-2}$$

where:

a and *b* – regression coefficients

x_k – any given value (in computations were used measured values of temperature)

$s = \sqrt{\frac{1}{n-2} \cdot \sum(y_i - \hat{y}_i)^2}$ = standard error of regression equation

$t_{1-\alpha/2;n-2} = 2.09$ for $\alpha = 0.05$ and $n = 21$ (from t-distribution table)

Day	y_i	x_i	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$	L_{lower} reg.l.	L_{upper} reg.l.	L_{lower} ind.val.	L_{upper} ind.val.
1	8.2	9.0	7.96	0.06	7.33	8.59	6.05	9.88
2	6.4	7.8	6.85	0.2	6.29	7.41	4.96	8.74
3	5.9	5.7	4.9	1	4.44	5.35	3.04	6.76
4	6.3	6.6	5.73	0.32	5.24	6.23	3.86	7.61
5	2.8	4.5	3.78	0.97	3.37	4.20	1.93	5.64
6	4.3	4.0	3.32	0.96	2.92	3.72	1.47	5.17
7	6.0	6.8	5.92	0.06	5.42	6.42	4.04	7.79
8	7.4	8.4	7.41	0.2	6.81	8.00	5.51	9.31
...	...	...	...	...	...	...	...	...
Σ	49.7	62.5		14.19				

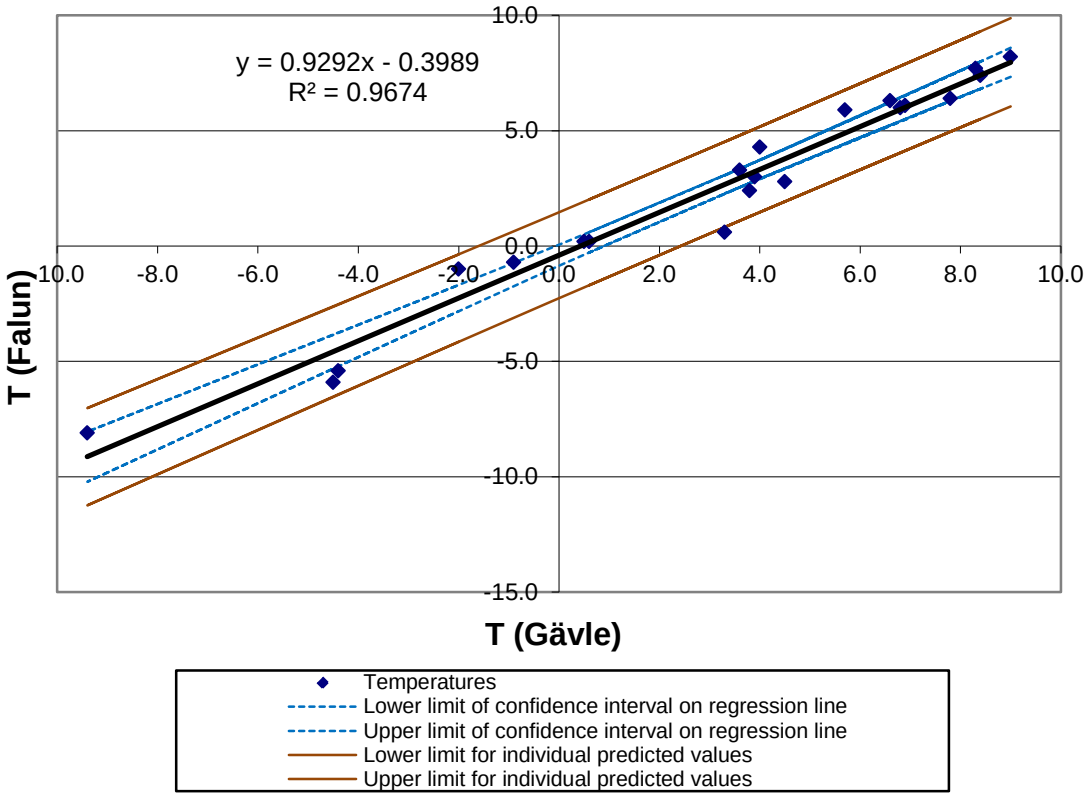


Figure 8.2: linear regression between the temperature measurements in Gävle and Falun, and the corresponding confidence intervals

4) Testing the significance of regression coefficients

- Test for significance of *a:*

  H_o: $a = 0$; H_a: $a \neq 0$

  $$t = \frac{a - 0}{s_a} = ...$$

  If $|t| \geq t_{1-\alpha/2,n-2}$ it mean that H_0 can be rejected, therefore coefficient *a* is significant and shall be retained.

- Test for significance of *b:*

  H_o: $b = 0$; H_a: $b \neq 0$

  $$t = \frac{b - 0}{s_b} = ...$$

  If $|t| \geq t_{1-\alpha/2,n-2}$ it mean that H_0 can be rejected, therefore coefficient *b* is significant and shall be retained.

Chapter 9

Multiple regression analysis

1. General Purpose

The general purpose of multiple regression (the term was first used by Pearson, 1908) is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable.

In general then, multiple regression procedures will estimate a linear equation of the form:

$$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + ... + b_p \cdot X_p + e$$

where X_i are the independent variables and Y is the dependent variable.

2. Assumptions of multiple regression

- The **independent variables** can **predict the dependent variable**, but the dependent variable **cannot** be used to predict the independent variables. It is a **one-way analysis**.
- Independent variables should be **justified theoretically**.
- Independent variables selected should have **strong correlations** with the **dependent variable** but **only weak correlations** with other **independent variables**.
- **Each independent variable** has **the same relationship** with the **dependent variable** at any value of other independent variables ($Y - X_i$ relationships are consistent).
- Dependent variables' **scores** come from a **normal distribution**

3. Common objectives in multiple regression analysis

1) Determine parameters of the regression equation
2) Determine the percent of the variability in ***y*** that is accounted for by multiple regression, R^2
3) Test statistical significance of the multiple regression
4) Determine the relative importance of different independent variables **x** in explaining **y**.

4. Parameter estimation

An example is the two independent variables x and y and one dependent variable z in the linear relationship case:

$$z = a + bx + cy$$

The linear regression in the above equation produces a plane in the three dimensional space; the sum of squared distances between the plane and the measured points is smallest for the best fit

(fig.9.1). This is analogous with the simple regression, which produces a straight line in the two dimensional space with the smallest sum of squared distances from the measured points.

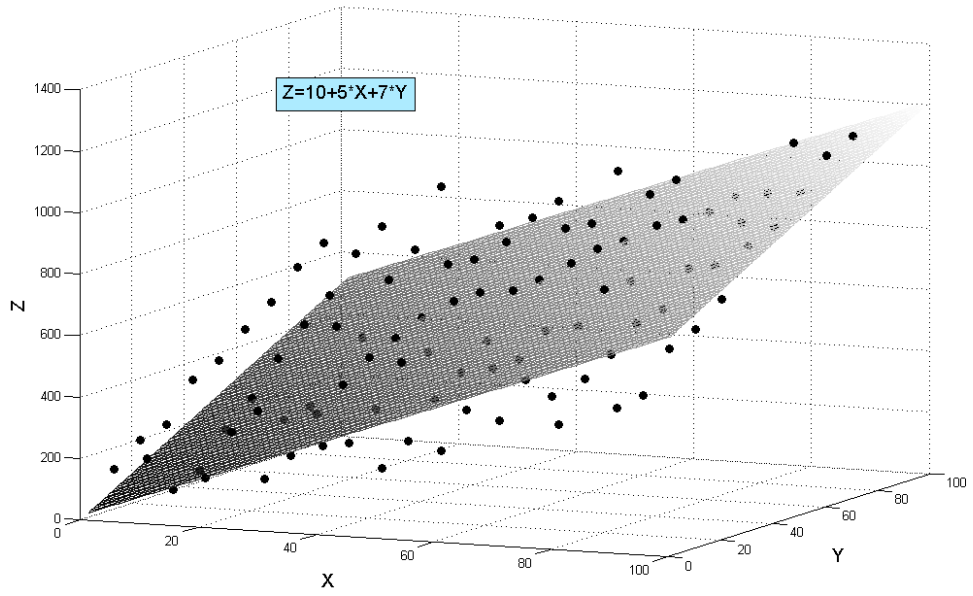


Figure 9.1: Multiple linear regression between one dependent variable (Z), and two independent variables (X, Y).

For a given data set: (x_1, y_1, z_1), (x_2, y_2, z_2), ..., (x_n, y_n, z_n), where $n \geq 3$, the best fitting plane f(X,Y) has the least square error:

$$\prod = \sum_{i=1}^{n} [z_i - f(x_i, y_i)]^2 = \sum_{i=1}^{n} [z_i - (a + bx_i + cy_i)]^2 = \min$$

Please note that *a*, *b*, and *c* are unknown coefficients while all x_i, y_i, and z_i are given. To obtain the least square error, the unknown coefficients *a*, *b*, and *c* must yield zero first derivatives and positive second derivatives:

$$\begin{cases} \dfrac{\partial \prod}{\partial a} = -2\sum_{i=1}^{n} [z_i - (a + bx_i + cy_i)] = 0 \\ \dfrac{\partial \prod}{\partial b} = -2\sum_{i=1}^{n} x_i [z_i - (a + bx_i + cy_i)] = 0 \\ \dfrac{\partial \prod}{\partial c} = -2\sum_{i=1}^{n} y_i [z_i - (a + bx_i + cy_i)] = 0 \end{cases}$$

$$\begin{cases} \dfrac{\partial^2 \prod}{\partial a^2} = 2 > 0 \\ \dfrac{\partial^2 \prod}{\partial b^2} = 2\sum_{i=1}^{n} x_i^{\,2} > 0 \\ \dfrac{\partial^2 \prod}{\partial c^2} = 2\sum_{i=1}^{n} y_i^{\,2} > 0 \end{cases}$$

Expanding the first derivatives equations system, we have:

$$\begin{cases} \sum_{i=1}^{n} z_i = a\sum_{i=1}^{n} 1 + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i z_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\,2} + c\sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i z_i = a\sum_{i=1}^{n} y_i + b\sum_{i=1}^{n} x_i y_i + c\sum_{i=1}^{n} y_i^{\,2} \end{cases}$$

The unknown coefficients *a, b*, and *c* can hence be obtained by solving the above linear equations.

For a general case where more than two independent variables are used in the regression, see page 243-245

5. Evaluation of the multiple regression model

- **R^2**, also called *multiple correlation* or the *coefficient of multiple determination*, is the percent of the variance in the dependent explained uniquely or jointly by the independent variables.
- **R^2 is still a measure of fit: same interpretation as in simple regression**

$$R^2 = \frac{explained\ variance}{total\ variance} = \frac{\sum(\hat{Z}_i - \bar{Z})^2}{\sum(Z_i - \bar{Z})^2} = 1 - \frac{\sum(Z_i - \hat{Z}_i)^2}{\sum(Z_i - \bar{Z})^2} = 1 - \frac{\sum e_i^2}{\sum(Z_i - \bar{Z})^2}$$

Where the relation between $\hat{Z}_i$, $\bar{Z}$ *and* Z_i is:

$$\sum(Z_i - \bar{Z})^2 = \sum(\hat{Z}_i - \bar{Z})^2 + \sum(Z_i - \hat{Z}_i)^2$$

And:

$\sum(Z_i - \bar{Z})^2$= total sum of squares correct for the mean = total variance

$\sum(\hat{Z}_i - \bar{Z})^2$= The squared deviations of the predicted values from the mean value, explained variance by the regression line

$\sum(Z_i - \hat{Z}_i)^2$ = the sum of squares of deviation from the regression
= unexplained variance

6. Adjusted R-square

The adjusted value for R^2 will be equal or smaller than the regular R^2. The adjusted R^2 adjusts for a bias in R^2. R^2 tends to overestimate the variance accounted for compared to an estimate that would be obtained from the population.

There are two reasons for the overestimate: a large number of predictors and a small sample size. So, with a large sample and with few predictors, adjusted R^2 should be very similar to the R^2 value.

Researchers and statisticians differ on whether to use the adjusted R^2. It is probably a good idea to look at it to see how much your R^2 might be inflated, especially with a small sample and many predictors.

$$R^2_{adjusted} = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$$

where n is the number of data, and k the number of independent variable used in the regression.

7. Partial-correlation coefficient

The partial-correlation coefficient measures the net correlation between the dependent variable and one independent variable after excluding the common influence of (i.e., holding constant) the other independent variables in the model. For example, $r_{YX_1X_2}$ is the partial correlation between Y and X_1, after removing the influence of X_2 from both Y and X_1

$$r_{YX_1X_2} = \frac{r_{YX_1} - r_{YX_2}r_{X_1X_2}}{\sqrt{1 - r^2_{X_1X_2}}\sqrt{1 - r^2_{YX_2}}}$$

$$r_{YX_2X_1} = \frac{r_{YX_2} - r_{YX_1}r_{X_1X_2}}{\sqrt{1 - r^2_{X_1X_2}}\sqrt{1 - r^2_{YX_1}}}$$

where r_{YX_1} = simple-correlation coefficient between Y and X_1 and r_{YX_2} and $r_{X_1X_2}$ are accordingly defined. Partial-correlation coefficient range in value from -1 to +1 (as do simple-correlation coefficients), have the sign of the corresponding estimated parameter and are used to determine the relative importance of the different explanatory variables in a multiple regression.

8. How many and which independent variables shall be used?

- **Principle:**
  - All variables that give an important increase of R^2
  - All variables that for theoretical or other reasons should be included
  - As few as possible (principle of parsimony) – the more variables, the greater uncertainty, larger type II error and the fewer degrees of freedom

- **Methods:**

1) Standard regression

All relevant variables are used in the regression equation

2) Stepwise multiple regression, also called *statistical regression*

a. In stage one, the independent best correlated with the dependent is included in the equation.
b. In the second stage, the remaining independent with the highest **partial correlation** with the dependent, after the first independent is removed, is entered.
c. Continue until no variables "significantly" explain residual variation. At each step the increment of R^2 is tested by performing the F-test

$$F_c = \frac{\left(1-R_{n-1}^2\right)\cdot\left(N-n-1\right)}{\left(1-R_n^2\right)\cdot\left(N-n-2\right)}$$

where N is number of data and n is used number of independent variables.

If $F_c > F_{1-\alpha,N-n-1,N-n-2}$, we know that addition of X_n is significant.

3) Backward Stepwise multiple regression

The process of stepwise multiple regression can also work backward, starting with all variables and eliminating independents one at a time until the elimination of one makes a significant difference in R-squared.

4) Hierarchical multiple regression

This is similar to stepwise regression, but the researcher, not the computer, determines the order of entry of the variables. F-tests are used to compute the significance of each added variable (or set of variables) to the explanation reflected in R^2.

Example 9.1

Assume that we want to relate the growth rate (G) of a bamboo plant during the growing season to air temperature (T) and precipitation (P). Estimate the regression coefficients using the data in table 9.1, and estimate the growth rate for a day with P=20mm and T=20 ^{o}C

Table 9.1: Measured growth rate, temperature and precipitation

G (mm/day)	T (^{o}C)	P (mm)
39.7	15.6	8.9
27.8	19.7	7.8
3.5	18.68	8.8
35.9	18.2	12.4
42.8	21.7	12.8.
37.9	21.1	13.8
35.1	24.1	16.8
69.4	25.8	16.2
76.7	25.7	15.6
58.9	25.9	17.6
80.5	25.4	20.3
75.2	27.1	21.9
60.7	29.2	22.2
95.5	28.0	20.7
116.3	30.1	23.7

By making a simple scatter plot of G versus T and P (fig.9.2), we decide that linear fitting can be applied.

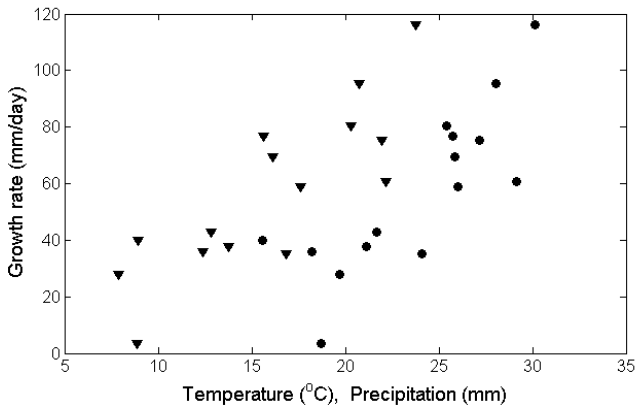


Figure 9.2: Growth rate vs temperature and precipitation

The model applied will be of the form:

$$G(T,P) = a + b \cdot T + c \cdot P$$

The best fit will be the one that has the least square error:

$$\Pi = \sum \left(G_i - \hat{G}_i\right)^2 = minimum$$

As discussed above, this is achieved by equating the first derivatives of Π with respect to a, b and c to zero:

$$\frac{\partial \Pi}{\partial a} = \frac{\partial \sum\left(G_i - \hat{G}_i\right)^2}{\partial a} = \frac{\partial \sum[G_i - (a + b \cdot T_i + c \cdot P_i)]^2}{\partial a} = 0$$

$$\frac{\partial \Pi}{\partial b} = \frac{\partial \sum\left(G_i - \hat{G}_i\right)^2}{\partial b} = \frac{\partial \sum[G_i - (a + b \cdot T_i + c \cdot P_i)]^2}{\partial b} = 0$$

$$\frac{\partial \Pi}{\partial c} = \frac{\partial \sum\left(G_i - \hat{G}_i\right)^2}{\partial c} = \frac{\partial \sum[G_i - (a + b \cdot T_i + c \cdot P_i)]^2}{\partial c} = 0$$

$$\Rightarrow \begin{cases} \sum_{i=1}^{n} G_i = a\sum_{i=1}^{n} 1 + b\sum_{i=1}^{n} T_i + c\sum_{i=1}^{n} P_i \\ \sum_{i=1}^{n} T_i G_i = a\sum_{i=1}^{n} T_i + b\sum_{i=1}^{n} T_i^2 + c\sum_{i=1}^{n} T_i P_i \\ \sum_{i=1}^{n} P_i G_i = a\sum_{i=1}^{n} P_i + b\sum_{i=1}^{n} T_i P_i + c\sum_{i=1}^{n} P_i^2 \end{cases}$$

Table 9.2: All the sums from the above system of equations

$\sum G_i$	$\sum T_i$	$\sum P_i$	$\sum T_i G_i$	$\sum P_i G_i$	$\sum T_i P_i$	$\sum T_i^2$	$\sum P_i^2$
856.2	356.6	239.8	$2.18 \cdot 10^4$	$1.54 \cdot 10^4$	$5.99 \cdot 10^3$	$8.75 \cdot 10^3$	$4.21 \cdot 10^3$

Consequently, the above equations' system will be rewritten to:

$$856.2 = 15 \cdot a + 356.6 \cdot b + 239.8 \cdot c$$
$$2.18 \cdot 10^4 = 356.6 \cdot a + 8.75 \cdot 10^3 \cdot b + 5.99 \cdot 10^3 \cdot c$$
$$1.54 \cdot 10^4 = 239.8 \cdot a + 5.99 \cdot 10^3 \cdot b + 4.21 \cdot 10^3 \cdot c$$

If we intend to solve the system using matrix algebra, we can rewrite it as:

$$\begin{pmatrix} 856.2 \\ 2.18 \cdot 10^4 \\ 1.54 \cdot 10^4 \end{pmatrix} = \begin{pmatrix} 15 & 356.6 & 239.8 \\ 356.6 & 8.75 \cdot 10^3 & 5.99 \cdot 10^3 \\ 239.8 & 5.99 \cdot 10^3 & 4.21 \cdot 10^3 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 15 & 356.6 & 239.8 \\ 356.6 & 8.75 \cdot 10^3 & 5.99 \cdot 10^3 \\ 239.8 & 5.99 \cdot 10^3 & 4.21 \cdot 10^3 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 856.2 \\ 2.18 \cdot 10^4 \\ 1.54 \cdot 10^4 \end{pmatrix}$$

The calculations can be done manually, or using matrix division in Matlab:
If $A = B \cdot x$, then: $x = B^{-1} \cdot A = B \backslash A$, where x, A and B are matrices.

$\Rightarrow a = -40.09$, $b = 1.95$, $c = 3.18$

$$\Rightarrow \boldsymbol{G = -40.09 + 1.95 \cdot T + 3.18 \cdot P}$$

The estimated growth rate during a day with temperature 20°C and precipitation 20mm is 62.46 mm.
The result is illustrated in figure 9.3.

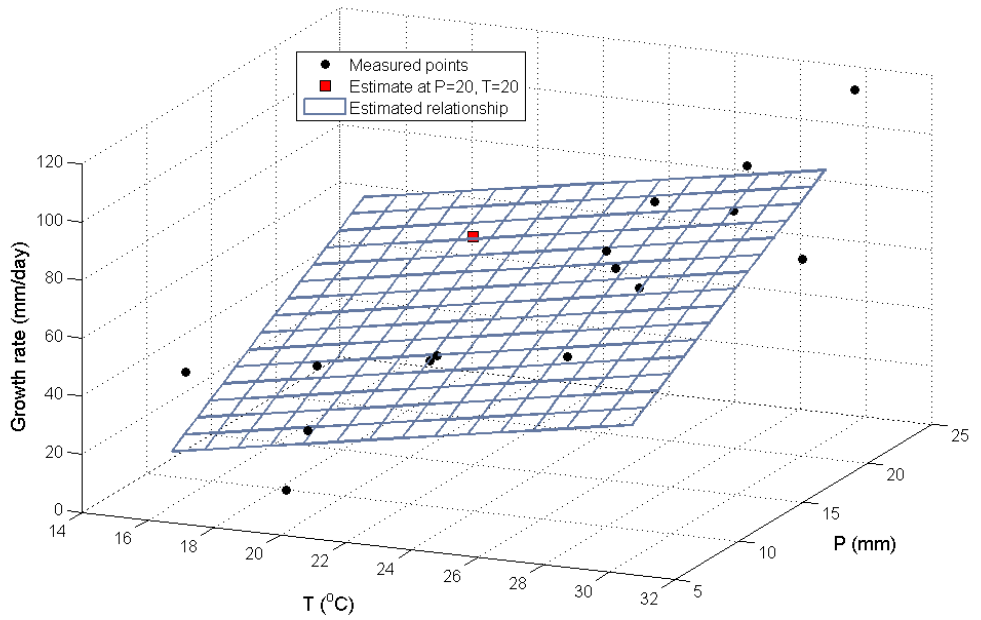


Figure 9.3: Measured points, fitted plane and estimated growth rate at P=20mm and T=20 °C

9. Multiple regression in Excel

Multiple regression can be performed in Excel through:

Data → Data analysis → Regression.

Choose as *Input Y Range* the dependent variable measurements.
Choose as *Input X Range* **both** independent variables measurements.

Using the Data from the Example 9.1 we have the following outputs:

Table 9.3: Regression statistics

Regression Statistics		*Explanation*
Multiple R	0.840	Correlation between the measured (G_i)and predicted ($\hat{G}_i$)dependent variable.
R Square	0.705	Fraction of the dependent variable's (G_i) variation that is explained by the independent variables (T,P)
Adjusted R Square	0.656	$R^2_{adjusted} = 1-(1-R^2)\left(\frac{n-1}{n-k-1}\right)$, n=sample size, k= number of independent variables =number of regressors
Standard Error	17.126	The typical deviation between the measured and predicted values
Observations	15	n=sample size

Table 9.4: ANOVA

		df	*SS*	*MS*	*F*	*Significance F*
	Regression	2	8417.99	4209.00	14.35	0.00065
	Residual	12	3519.54	293.29		
	Total	14	11937.53			
Explanation		*df*= degrees of freedom	*SS*= sum of squares	*MS=SS/df*	F statistic	At confidence level 1-0.00065, at least one of the coefficients if the model is significant.
	Regression	df= k	$SS_R = \sum(\hat{Z}_i - \bar{Z})^2$	MS_R=SSR/k	MS_R / MS_T	
	Residual	df= n-k-1	$SS_E = \sum(Z_i - \hat{Z}_i)^2$	$MS_E = SS_E/(n-k-1)$		
	Total	df=n-1	$SS_T = \sum(Z_i - \bar{Z})^2$	$MS_T = SS_T/(n-1)$		

Table 9.5: Coefficients

		Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
	Intercept	-40.1	35.1	-1.1	0.3	-116.5	36.4
	X Variable 1	2.0	2.8	0.7	0.5	-4.2	8.1
	X Variable 2	3.2	2.4	1.3	0.2	-2.0	8.4
Explanation	**Intercept**	*a*	S_a	$t = \frac{a-0}{S_a}$	Significance level that the tested parameter is significant	$a \pm t_{1-\frac{\alpha}{2},n-1} \cdot S_a$	
	Independent variable 1	b_1 =Slope of ind.var.1	S_{b1}	$t = \frac{b-0}{S_b}$		$b_1 \pm t_{1-\frac{\alpha}{2},n-1} \cdot S_{b1}$	
	Independent variable 2	b_2 =Slope of ind.var.2	S_{b2}	$t = \frac{b-0}{S_b}$		$b_2 \pm t_{1-\frac{\alpha}{2},n-1} \cdot S_{b2}$	

Chapter 10

Parameter estimation theory and methods

1. Parameter estimation

A probability distribution *f(x)* or any model function shall be written as: $f(x; \theta_1, \theta_2,...,\theta_m)$, where θ_i are the parameters. This indicates that in general the distribution is a function of the random variables as well as the set of parameters. To use the probability function to estimate the probabilities, the values of parameters need to be estimated. The usual procedure for estimating a parameter is to obtain a random sample $x_1, x_2,\ldots, x_n$ from the population X. Thus $\hat{\theta}_i$, an estimate of θ_i is a function of observations. Since $\hat{\theta}_i$ is a function of random variables, $\hat{\theta}_i$ is itself a random variable.

Figure 10.1 shows an example of the probability distribution of a variable x, and the probability distribution function that we want to fit to it. In this case the parameter 'a' needs to be estimated.

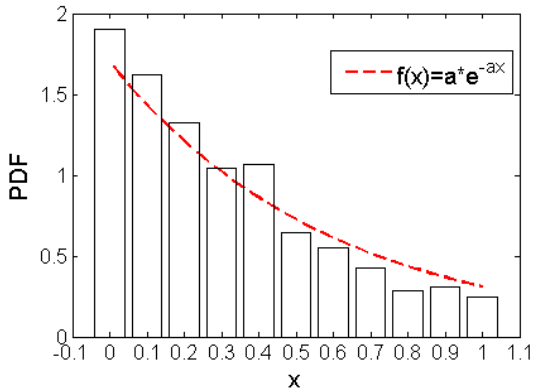


Figure 10.1: Probability distribution and fitted function

2. Estimation criteria

- Unbiaseness: an estimator $\hat{\theta}$ of a parameter θ is said to be unbiased if $E(\hat{\theta}) = \theta$. The bias, if any, is given by $E(\hat{\theta}) - \theta$. This does not mean that $\hat{\theta}$ is equal to θ, or even close to θ. It simply means that the average of many independent estimates of θ will equal θ.
- Consistency: If the sample is large enough, the probability that a parameter θ and its estimator $\hat{\theta}$ differ more than an arbitrary constant ε is zero: $P(|\hat{\theta} - \theta| > \varepsilon) \approx 0$
- Efficiency: If $\hat{\theta}$ is an unbiased estimator of θ then Var($\hat{\theta}$) is the smallest compared with any other estimator.

- Sufficiency: An estimator $\hat{\theta}$ of a parameter θ is said to be sufficient if $\hat{\theta}$ uses all the information relevant to θ that is contained in the sample.

3. Estimation methods

- Method of moments: Use the moments to estimate parameters. For a distribution with m parameters, the procedure is to equal the first m moments of the distribution to first m sample moments.

Example 10.1

Estimate the parameter λ of the distribution $f(x) = \lambda \cdot e^{-\lambda \cdot x}\ ,\ x > 0$

Solution:

The first moment about the origin is the population mean, μ. If we assume that the sample is large enough, then μ will equal the sample mean, $\bar{x}$:

$$\mu = \bar{x} = \int x \cdot f(x)dx = \lambda \cdot \int_0^{\infty} x \cdot e^{-\lambda \cdot x}dx = \lambda \frac{1}{\lambda^2} = \frac{1}{\lambda}$$

So, the estimate for λ is: $\hat{\lambda} = 1/\bar{x}$

- Least square method

  Estimation of parameters by using least squares method for linear regression was discussed in Chapter 8.

- Maximum likelihood method

  Maximum likelihood method begins with writing a mathematical expression known as the ***Likelihood function*** of the sample data. Loosely speaking, the likelihood of a set of data is the probability of obtaining that particular set of data, given the chosen probability distribution model. This expression contains the unknown model parameters. The values of these parameters that maximize the sample likelihood are known as ***Maximum Likelihood Estimates*** or **MLE's**.

  If we have *n* random observations *x*, the joint probability density of *x* as function of parameters θ will be expressed as:

$$f(x_1, x_2, \ldots, x_n \,|\theta_1, \theta_2, \ldots, \theta_m)$$

Since the measurements are independent, the joint probability density can be rewritten to:

$$f(x_1|\theta_1, \theta_2,\ldots, \theta_m)\cdot f(x_2|\theta_1, \theta_2,\ldots, \theta_m)\cdot\ldots\cdot f(x_n|\theta_1, \theta_2,\ldots, \theta_m)$$

This latter expression is proportional to the probability that the particular random sample would be obtained from the population and is known as the likelihood function:

$$L(\theta_1, \theta_2, \ldots, \theta_m) = \prod_{i=1,n} f(x_i|\theta_1, \theta_2, \ldots, \theta_m)$$

The objective is to find for which set of parameters $\theta_1, \theta_2, \ldots, \theta_m$ the likelihood function is maximized.
The maximum likelihood can be obtained by taking the derivative of L with respect to θ and setting it equal to zero. For this purpose it is convenient to first take the logarithms and then take the derivatives. From this we can obtain θ_i in terms of x_k.

Advantages:

+ The method has a good intuitive foundation. The underlying concept is that the best estimate of a parameter is the one that gives the highest probability that the observed set of measurements will be obtained.

+ The least-squares method and various approaches to combining errors or calculating weighted averages, etc., can be derived or justified in terms of maximum likelihood approach.

+ The method is of sufficient generality that most problems are amenable to a straightforward application of this method, even in cases where other techniques become difficult. Inelegant but conceptually simple approaches often provide useful results where there is no easy alternative.

Disadvantages:

- Maximum likelihood method is very CPU intensive and thus extremely slow.

Example 10.2

Find teh maximum likelihood estimator for the parameter λ of the distribution:
$f(x) = \lambda \cdot e^{-\lambda \cdot x}\ ,\ x > 0$

Solution:
$L(\lambda) = \prod_{i=1}^{n} \lambda \cdot \mathrm{e}^{-\lambda \cdot \mathrm{x_i}} = \lambda^{\mathrm{n}} e^{-\lambda \Sigma x_i} \qquad \Rightarrow\ \ln(L) = n \cdot \ln(\lambda) - \lambda \cdot \Sigma x_i$

$\Rightarrow \frac{\partial[\ln(L)]}{\partial\lambda} = \frac{n}{\lambda} - \Sigma x_i = 0$

$\Rightarrow \hat{\lambda} = \frac{n}{\Sigma x_i} = \frac{1}{\bar{x}}$

Chapter 11

Introduction to Geostatistics

1. Introduction

1.1 Problem:

- **Sample data** for many physical parameters is often **sparse** on **irregular grids.**
- Frequently, the values of the variable are desired on a regular and significantly finer grid than is practical to sample.

1.2 Objective:

- Point estimates at locations where measurements are not available.
- Estimates for the areal averages over the entire area or some part of the area.
- Estimates of the uncertainty associated with point or areal average estimates.
- The distribution of values over areas of various sizes and locations.

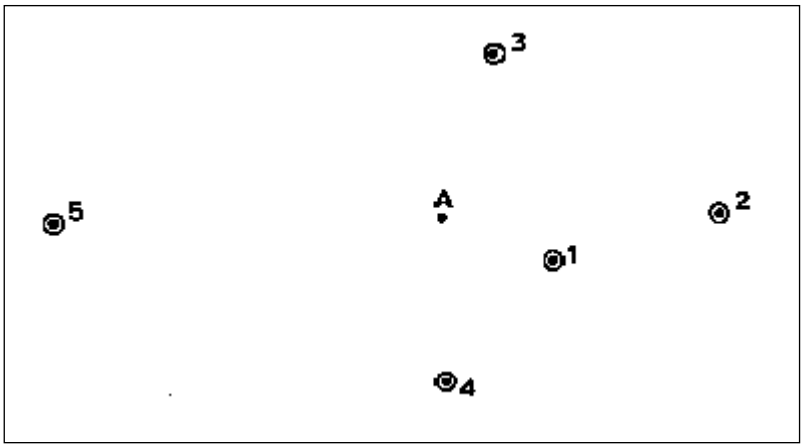


Figure11.1: Hypothetical sampling and estimation situation

1.3 Possible ways:

- Use un-weighted mean $\hat{X}(A) = \frac{1}{n}\sum_{i=1}^{n} x_i$ (no spatial continuity of variables is considered, this is not a good way)
- A better estimation technique use **weighted linear combinations** of nearby sample values to estimate the value at a point:

$$\hat{X}(A) = W_1 X_1 + W_2 X_2 + \ldots + W_n X_n \qquad \textbf{(1)}$$

Here, W_i are weights of the sampled data X_i, $\hat{X}(A)$ is an estimate of the value of the variable at **X(A),** the location that is not sampled.

The estimation problem is to choose the weights $\mathbf{W_i}$**, i=1,...,n** to minimize the error in estimation, i.e. $\min(\hat{X}(A) - X(A))$

1.4 Difficult:

- However, this goal is impossible to achieve in practice, since it would require prior knowledge of the true values at the unsampled locations, i.e., observed values at the unsampled locations.

1.5 How to solve the problem?

- As a result, most estimation techniques make substantial **simplifying assumptions** regarding the behavior of the variable in order to predict unsampled values.

2. Random function theory

2.1 Random function

- Geostatistics is based on **Random Function** [RF] theory.
- The samples $\mathbf{V_j}$ are viewed as outcomes of a random variable that is a function of the spatial coordinates $\mathbf{V(x_j)}$**.**
- The linear estimator $\hat{\mathbf{v}}_0$ is then viewed as an outcome of the RF $\hat{\mathbf{V}}(\mathbf{x_0})$ which approximates the true value at $\mathbf{x_0}$**.**
- The true value is also assumed to be a RF $\mathbf{V(x_0)}$**.**
- The actual error incurred $\hat{\mathbf{v}}_0\mathbf{-v_0}$ can then be viewed as an outcome of the RF: $\mathbf{R(x_0)= \hat{V}(x_0)-V(x_0)}$ which is called the **estimation error.**

2.2 How it works?

- You sample some quantity at a number of locations.
- Applying the linear estimator [with some fixed choice of weights] you produce and estimate at some unsampled location.
- Suppose now that you return to the **same** sample locations and resample.
- Due to measurement error, the sample values are slightly different.
- Applying the previous linear estimator you produce a second estimate for the same unsampled location.
- And so on. Each time, you are in fact producing a realization of the estimation error RF.

2.3 What would be the desirable properties of this estimation error RF?

- Ideally, it should be identically zero! Failing this [and you will!] one should aim for two things.
    - Firstly, the estimation error should be **unbiased.** That is, its expected value is zero: **E[R(x0)]=0.** This loosely means that on average the error incurred is zero.

- The second property is that the estimation error should have **minimum spread.** That is, its variance **Var[R(x0)]** is minimal. Linear estimators which produce estimation error RF's having the properties of unbiasedness and minimum variance are called **BLUE** [Best Linear Unbiased Estimators].
- The geostatistical estimation problem is to choose the weights **wi, i=1,...,n** to obtain a BLUE estimator.

3. Inverse Distance estimation methods

3.1 Equation

The weights are inversely proportional to the distances, h

$$w_i = \frac{h(x_0, x_i)^{\{-p\}}}{W} \qquad \textbf{(2)}$$

The non-negative parameter **p** is chosen to reflect the assumed measure of spatial continuity of the variable. For example, with **p=5** much more weight is given to the nearest sample, than if using **p=1.**
The factor **W** is a normalization factor chosen so that the weights sum to unity. This gives the estimate the desirable property of unbiasedness [see later].

3.2 Four variants of Inverse Distance estimation:

- Local Average: p=0
  [LA] technique assumes **no spatial continuity** of the variable, and simply uses the arithmetic mean of the sample values.

- Inverse Distance: p=1

- Inverse Distance Squared: p=2
  [ID] and [IDS] techniques assume a loss in spatial continuity over greater distances, the loss assumed greater for the latter.

- Polygonal: p-->infinity
  [P] technique is perhaps the simplest and uses the value of only the **nearest sample** as the estimate

Which of these techniques is best? Tend to rely heavily on the practitioners experience and judgment. The art of estimation becomes just that.

However, Inverse Distance estimators are not BLUE

4. Kriging – a BLUE estimator

Kriging, after D. G. Krige from South Africa, is an interpolation technique that uses statistical nature of the variability of geo-spatial data. The spatial variability in geological data consists of three sources – (1) a structural part consisting of a trend, (2) a random, spatially correlated component, and (3) a random error term, i.e.

$$P(x) = m(x) + g(x) + \varepsilon \qquad (3)$$

where m(x) is the spatially dependent mean, g(x) is the random, spatially correlated part, and □is the Gaussian error term (figure 1). The nature of m(x) can be determined by techniques such as trend surface analysis and will not be discussed here. Once the structural component, m(x), is removed, we are left with two random terms one of which is spatially correlated [g(x)] and the other is a pure noise term (ε).

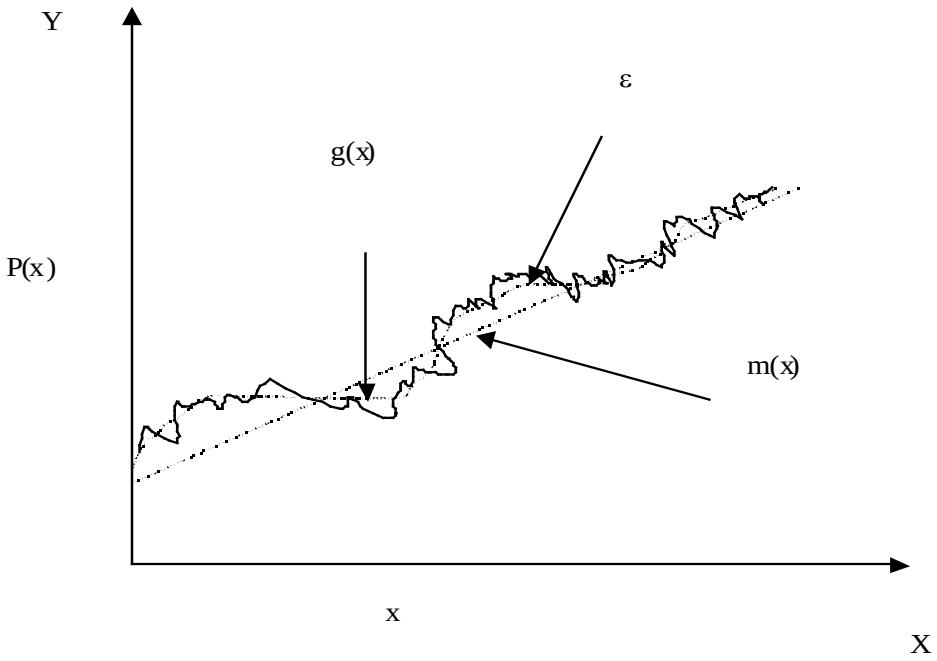


Figure 11.2: Three major components in geo-spatial data.

In the following discussion we make the following assumptions[1]:

1) Once the structural component is removed, the mean or expectation of the parameter is independent of the location, i.e. $E(P_i)=E(P_{i+h})$,

[1] Review of basic statistics:
If Z is a random variable then its mean or expectation, $\mu = E(Z)$ and its variance, $\sigma^2=E(Z-\mu)^2 =E(Z^2)-\mu^2$.
Covariance between any two random variables, U and V is $C(U,V)=E(UV-\mu_U\mu_V) =E(UV) -\mu_U\mu_V$. If Uand V are drawn from the same distribution, then $\mu_U=\mu_V=\mu$ and $C(UV)=E(UV)-\mu^2$

2) Variation of parameter values separated by a distance h depends only on the separation distance h but not on their location, i.e. $C[P_i, P_{i+h}]$ is independent of location x,

where E stands for expectation or mean and C stands for covariance. These assumptions are known as second order stationarity assumptions.
If the separation distance, h is zero we expect all data to fall on a 45 degree line, OO'. An observed data P_{i+h} which is separated by a nearby data point P_i by a distance h may be a distance $AC=AB\cos(45^o)= (P_{i+h} - P_{i)})/\sqrt{2}$ away from OO'. The moments of all such data points which are separated by a distance h is given by:

$$M = \sum_{i=1}^{n} (P_{i+h} - P_i)^2 / 2 = \frac{1}{2} \sum_{i=1}^{n} (P_{i+h} - P_i)^2 \qquad (4)$$

The average value of M given in equation (4) is a measure of dissimilarity between nearby points separated by a distance h and is known as semi-variance, $\gamma(h)$,

$$\gamma(h) = \frac{1}{2n(h)} \sum_{i=1}^{n} (P_{i+h} - P_i)^2 \qquad (5)$$

The separation distance, h, is often referred to as lag. As separation distance approaches zero, one would expect the semi-variance to approach zero. However, the semi-variance usually approaches a non-zero value, known as nugget as h->0. This is due to the Gaussian error component of the geo-spatial data. As lag increases, $\gamma(h)$ also increases and becomes relatively constant for lag values greater than certain distance, a, known as range. This constant value of semi-variance for $h \geq a$ is known as sill. Figure (3) shows a typical semivariogram along with nugget, sill, and range for this case.

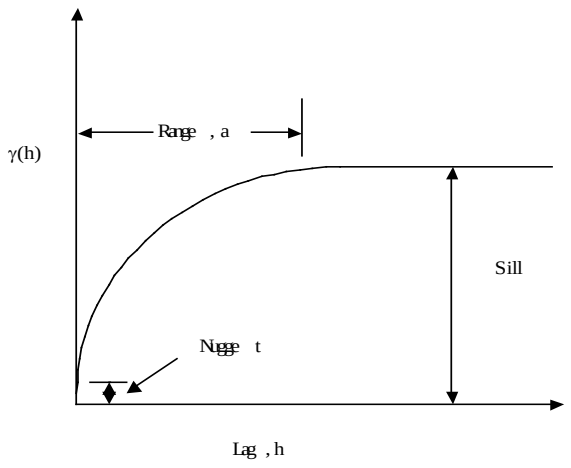


Figure 11.3: A typical semivariogram which shows a clear nugget, sill, and range.

Instead of the measure of dissimilarity of nearby data points, we could have looked at the similarity or correlation between points separated by a distance h. Covariance functions provide a measure of similarity between points separated by lag, h. It is defined as:

$$C(h) = \frac{1}{n(h)}\sum_{i=1}^{n}(P_i - \mu_{P_i})(P_{i+h} - \mu_{P_{i+1}}) \cong \frac{1}{n(h)}\sum_{i=1}^{n}(P_iP_{i+h} - \mu^2) \qquad (6)$$

where $\mu = \frac{1}{N}\sum_{k=1}^{N} P_k$ which is the mean of all data points N.

Sometimes an alternative to covariance function known as correlation function is used (note that μ_{Pi} refers to the mean of P_is). Correlation function is essentially a normalized version of covariance function and ranges between 0 and 1. The covariance function tends to be high when h=0 (i.e. correlation function is 1) and goes to zero for points which are separated by distances greater or equal to the range (i.e. uncorrelated). Figure 11.4 is schematic of the covariance function, C(h) and correlogram γ(h). These are related by following relationships:

$$C(h) = C(0) - \gamma(h) \qquad (7)$$

and

$$\text{Correlogram, } \rho(h) = C(h)/C(0) = 1 - \gamma(h)/C(0) \qquad (8)$$

Because:

$$\begin{aligned}
\gamma(h) &= \frac{1}{2}E(P_i - P_{i+1})^2 \\
&= \frac{1}{2}E(P_i^2 - 2P_iP_{i+1} - P_{i+1}^2) \\
&= \frac{1}{2}\left\{E(P_i^2) - 2E(P_iP_{i+1}) + E(P_{i+1}^2)\right\} \\
&= E(P_i^2) - E(PP_{i+1}) \qquad \left\{\because E(P_i^2) = E(P_{i+1}^2)\right\} \\
&= \left\{E(P_i^2) - \mu^2\right\} + \left\{E(PP_{i+1}) - \mu^2\right\} \\
&= C(0) - C(h)
\end{aligned}$$

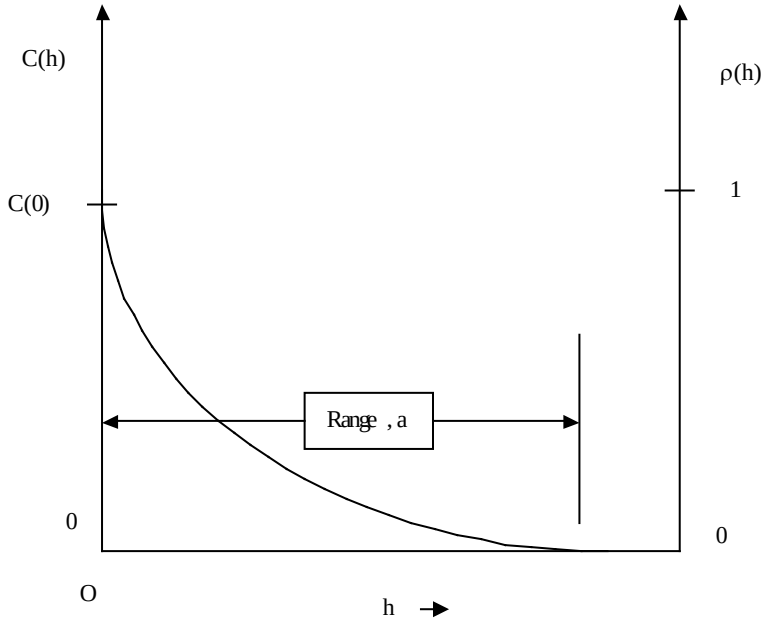


Figure 11.4. Covariance and correlogram representation of geo-spatial variability.

As distance increases C(h) decreases and approaches zero and γ(h) increases and becomes asymptotic to the sill value which is equal to the variance,σ^2, of the random function. From equation (7), we get:

$$
\begin{aligned}
&C(\infty) = 0 = C(0) - \gamma(\infty) \\
&But \;\; \gamma(h) = \sigma^2 \\
&\therefore \;\; C(0) = \sigma^2 \\
&and \;\; C(h) = \sigma^2 - \gamma(h)
\end{aligned}
\qquad (9)
$$

Therefore semivariogram, covariance, and correlogram are similar (figure 11.5) and for historical reasons semivariogram is widely used in geostatistics.

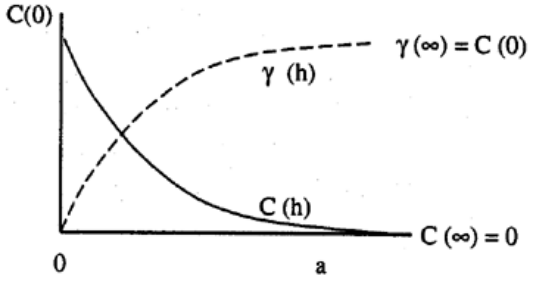


Figure 11.5 Comparison of autocovariance and semivariance

Models of Semivariogram:

Spherical model:

$$\gamma(h) = c_0 + c_1\left\{\frac{3h}{2a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right\} \qquad for \;\; 0 < h < a$$
$$= c_0 + c_1 \qquad for \;\; h \geq a \qquad (10)$$
$$\gamma(0) = 0$$

Exponential model:

$$\gamma(h) = c_0 + c_1\left\{1 - \exp\left(\frac{-h}{a}\right)\right\} \qquad (11)$$

Linear model:

$$\gamma(h) = c_0 + c_1 h \qquad (12)$$

The linear model (equation 12) is used when the semivariogram does not appear to have a clear sill. Note c_0 and c_1 are constants obtained from curve fitting the experimental data. Moreover, c_0 is the nugget and $(c_0 + c_1)$ is the sill or variance, σ^2.

Kriging: Interpolation technique in which weighting functions are derived from the geostatistical spatial analysis of the sample variogram is known as kriging. Let $\hat{P}_0$ be the estimated value of the geological data at point O. It can be written as:

$$\hat{P}_0 = \sum_{i=1}^{n} w_i P_i \qquad (13)$$

where $\sum_{i=1}^{n} w_i = 1$. The error or residual, R, in interpolation is given by:

$$R = P_0 - \hat{P}_0 = P_0 - \sum_{i=1}^{n} w_i P_i \qquad (14)$$

Variance of this residual can be expressed as:

$$Var(R) = s_R^2 = E(P_0 - \sum_{i=1}^{n} w_i P_i)^2 \qquad (15)$$

Since mean of the residual is zero, i.e.,

$$E(R) = E\left\{P_0 - \sum_{i=1}^{n} w_i P_i\right\} = E(P_0) - \sum_{i=1}^{n} w_i E(P_i)$$

$$= \mu - \mu \sum_{i=1}^{n} w_i = \mu - \mu = 0 \qquad \left\{\square \sum_{i=1}^{n} w_i = 1\right\}$$

The weights w_i's are obtained by minimizing the variance of residual subject to the constraint $\sum_{i=1}^{n} w_i = 1$ with respect to w_is and the Lagrange multiplier, λ, i.e.

$$\Phi = E(P_0 - \sum_{i=1}^{n} w_i P_i)^2 + 2\lambda(\sum_{i=1}^{n} w_i - 1) \qquad (16)$$

and

$$\frac{\partial \Phi}{\partial w_i} = \frac{\partial \Phi}{\partial \lambda} = 0 \qquad for \ \ all \ \ i \qquad (17)$$

where Φ is the objective function.

This minimization procedure results in,

$$\sum_{j=1}^{n} w_j C_{ij} + \lambda = C_{io}$$

$$\text{for i = 1, 2, …,n} \qquad (18)$$

$$\sum_{i=1}^{n} w_i = 1$$

or in the following matrix equation for the weights:

$$\begin{bmatrix} C_{11} & C_{12} & C_{13} & . & . & . & C_{1n} & 1 \\ C_{21} & C_{22} & C_{23} & . & . & . & C_{2n} & 1 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ C_{n1} & C_{n2} & C_{n3} & . & . & . & C_{nn} & 1 \\ 1 & 1 & 1 & . & . & . & 1 & 0 \end{bmatrix} \begin{Bmatrix} w_1 \\ w_2 \\ w_3 \\ . \\ . \\ w_n \\ \lambda \end{Bmatrix} = \begin{Bmatrix} C_{10} \\ C_{20} \\ C_{30} \\ . \\ . \\ C_{n0} \\ 1 \end{Bmatrix} \tag{19}$$

where C_{ij} is the covariance between parameter values P_i and P_j. Similarly C_{io} is the covariance between the parameter value at the unvisited point O (P_o) and the parameter value at point i (P_i).

For example: consider a simple two parameter case. The variance of estimation is

$$\begin{aligned} Var(R) &= E(P_0 - w_1P_1 - w_2P_2)^2 \\ &= E(P_0^2 + w_1^2P_1^2 + w_2^2P_2^2 - 2w_1P_0P_1 - 2w_2P_0P_2 + 2w_1w_2P_1P_2) \\ &= E(P_0^2) + w_1^2E(P_1^2) + w_2^2E(P_2^2) - 2w_1E(P_0P_1) - 2w_2E(P_0P)_2 + 2w_1w_2E(P_1P_2) \\ &= C_{00} + \mu^2 + w_1^2(C_{11} + \mu^2) + w_2^2(C_{22} + \mu^2) - 2w_1(C_{10} + \mu^2) - 2w_2(C_{20} + \mu^2) + 2w_1w_2(C_{12} + \mu^2) \\ &= C_{00} + w_1^2C_{11} + w_2^2C_{22} - 2w_1C_{10} - 2w_2C_{20} + 2w_1w_2C_{12} \\ &\quad + \mu^2 + \mu^2(w_1^2 + w_2^1 + 2w_1w_2) - 2\mu^2(w_1 + w_2) \\ &= C_{00} + w_1^2C_{11} + w_2^2C_{22} - 2w_1C_{10} - 2w_2C_{20} + 2w_1w_2C_{12} + \mu^2 + \mu^2(w_1 + w_2)^2 - 2\mu^2(w_1 + w_2) \\ &= C_{00} + w_1^2C_{11} + w_2^2C_{22} - 2w_1C_{10} - 2w_2C_{20} + 2w_1w_2C_{12} + (\mu^2 + \mu^2 - 2\mu^2) \\ &= C_{00} + w_1^2C_{11} + w_2^2C_{22} - 2w_1C_{10} - 2w_2C_{20} + 2w_1w_2C_{12} \end{aligned}$$

$\therefore$ The objective function Φ is given by:

$$\Phi = C_{00} + w_1^2C_{11} + w_2^2C_{22} - 2w_1C_{10} - 2w_2C_{20} + 2w_1w_2C_{12} + 2\lambda(w_1 + w_2 - 1)$$

$$\frac{\partial\Phi}{\partial w_1} = 2w_1C_{11} - 2C_{10} + 2w_2C_{12} + 2\lambda = 0 \text{ or } w_1C_{11} + w_2C_{12} + \lambda = C_{10}$$

$$\frac{\partial\Phi}{\partial w_2} = 2w_2C_{22} - 2C_{20} + 2w_1C_{12} + 2\lambda = 0 \text{ or } w_1C_{12} + w_2C_{22} + \lambda = C_{20}$$

$$\frac{\partial\Phi}{\partial \lambda} = 2(w_1 + w_2 - 1) = 0 \text{ or } w_1 + w_2 = 1$$

Or in the matrix form:

$$\begin{bmatrix} C_{11} & C_{12} & 1 \\ C_{12} & C_{22} & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{Bmatrix} w_1 \\ w_2 \\ \lambda \end{Bmatrix} = \begin{Bmatrix} C_{10} \\ C_{20} \\ w \end{Bmatrix}$$

The procedure to use the ordinary kriging method is as follows:

1) From the experimental data determine the experimental variogram,

   Suppose that the value to be interpolated is referred to as x. The experimental variogram is found by calculating the variance (γ) of each point in the set with respect to each of the other points and plotting the variances versus distance (h) between the points.

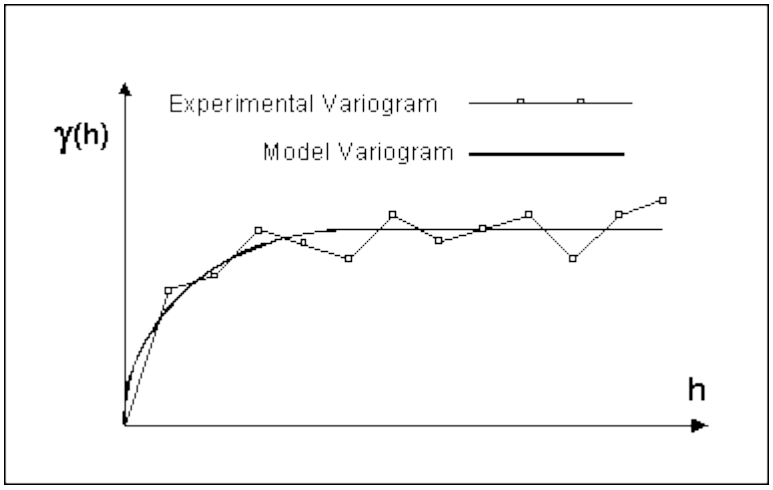


Figure 11.6: Experimental and Model Variogram Used in Kriging

   As can be seen in the above figure, the shape of the variogram indicates that at small separation distances, the variance in x is small. In other words, points that are close together have similar x values. After a certain level of separation, the variance in the x values becomes somewhat random and the model variogram flattens out to a value corresponding to the average variance

2) Determine the variogram model. The variogram models are discussed in the previous section. Such as linear model, Spherical model, exponential model, etc.

   Use the variogram model to determine $C_{ij}(h)$ and $C_{i0}(h)$ through $\gamma_{ij}(h)$ and $\gamma_{i0}(h)$ for use in equation (19),

   Equation (19) can also be written directly in terms of semivariogram,

$$\sum_{j=1}^{n} w_j \gamma_{ij} - \lambda = \gamma_{io}$$

$$\text{for } i = 1, 2, \ldots \qquad (20)$$

$$\sum_{i=1}^{n} w_i = 1$$

   and variance

$$s_R^2 = \sum_{i=1}^{n} w_i \gamma_{io} - \lambda \qquad (21)$$

3) Solve equation (19) or (20) for weights,

4) Use equation (13) to predict desired value, $P_{o.}$ The variance of the residuals is calculated by equation (21).

5) The standard error of the estimate is the square root of the estimation variance and equals:

$$S_e = \sqrt{\hat{\sigma}_R^2}$$

Therefore, the probability that the true elevation is within one standard error above or below the estimated value is 68%. Two standard errors away would give a confidence of 95%. For this example, the water table elevation at location p is

$$\hat{P}_o \pm 2S_e$$

The prediction obtained by kriging has the property of minimum variance. Moreover, range,a, provides an effective radius of influence. If geological data shows anisotropy, which is often the case, the technique discussed here can be extended to two dimensions (Issaaks and Srivastava, 1989).

http://www.cee.vt.edu/program_areas/environmental/teach/smprimer/kriging/kriging.html#TheoryEqns
http://www.ems-.com/gmshelp/Interpolation/Interpolation_Schemes/Kriging/Kriging.htm
http://www.pssc.ttu.edu/pss5231/Class%20Days/week_8.htm

Chapter 12

Time series analysis

1. Time series and time series analysis

- Time series – a sequence of values collected over time on a particular variable
    - Observed at discrete times
    - Recorded continuously over time
    - Averaged over a time interval
    - Regular time step or irregular time step

  We discuss discrete times series with regular time step, daily, monthly, yearly, etc.

- Time series analysis
    - Tasks
        - Identify different components
        - Model different component
    - Purposes
        - Understand the physical mechanism
        - Describe the process
        - Simulate (extend) data series
        - Forecast/predict future values

2. **First step in time series analysis:** Plot the data and look for some important characteristics.

   What to look for?

   - Is there a **trend** over time? That is, does the series increase or decrease with time?
   - Is there any regular **seasonality** (or cyclic components)?
   - Is there **constant variability** over time?
   - Are there any other **systematic features** of the data?

   In figure 12.1 ate least three features can be detected:
   - an increasing trend,
   - an increasing variance
   - a periodic component (seasonal pattern)

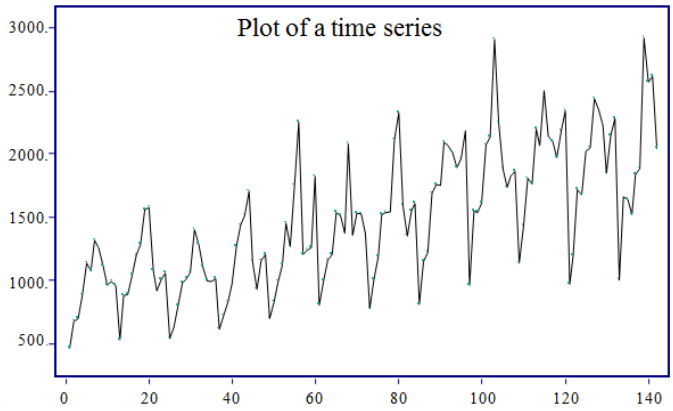
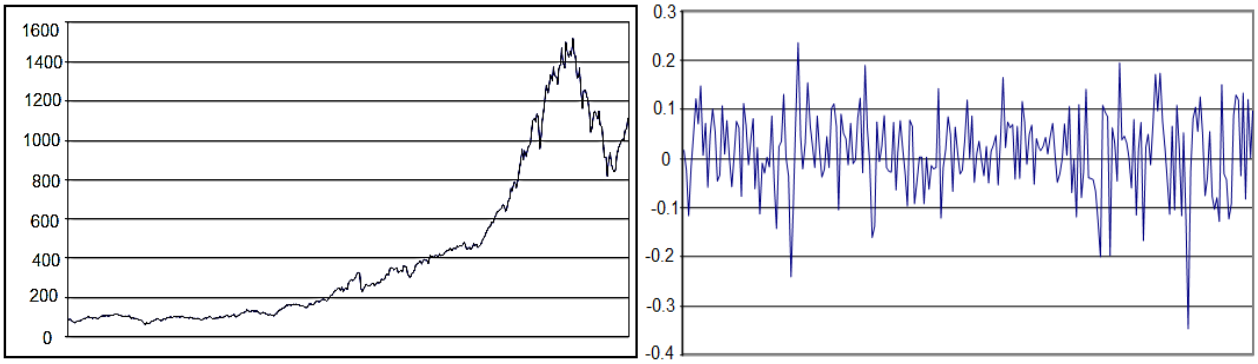


Figure 12.1: An example of time series

But what can you say about these?



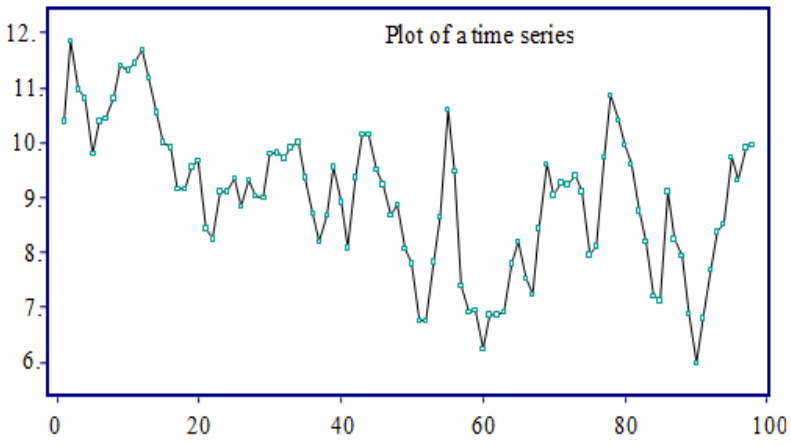


Figure 12.2 Examples of time series

To summarize the statistical properties of the above time series, we need to apply some time series analysis methods.

3. Statistics used to describe a time series

Name	*Sample estimation*	*Notation for population*
Mean	$E(X_t) = \overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$	μ
Variance	$S^2 = \frac{1}{n-1}\sum_{t=1}^{n}(X_t - \overline{X})^2$	σ^2
Covariance	$\mathrm{cov}(X_t, X_{t+L}) = \frac{1}{n-L}\sum_{t=1}^{n-L}(X_t - \overline{X})(X_{t+L} - \overline{X})$	λ_L

where L is the time lag.

Autocorrelation – is the correlation between two copies of the same time-series but with a time shift of L:

$$\rho_L = cor(x_t, x_{t+L}) = \frac{\mathrm{cov}(x_t, x_{t+L})}{\sqrt{Var(x_t)Var(x_{t+L})}}$$

and sample estimation is ρ_L

$$r_L = \frac{\frac{1}{n-L}\sum_{t=1}^{n-L}(X_t - \overline{X})(X_{t+L} - \overline{X})}{\sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(X_t - \overline{X})^2 \frac{1}{n-L}\sum_{t=1}^{n-L}(X_{t+L} - \overline{X}_L)^2}}$$

When n is much larger than L (n>>L):

$$\frac{1}{n-1}\sum_{t=1}^{n}(X_t - \overline{X})^2 \approx \frac{1}{n-L}\sum_{t=1}^{n-L}(X_{t+L} - \overline{X}_L)^2$$

$$\Rightarrow r_L = \frac{\frac{1}{n-L}\sum_{t=1}^{n-L}(X_t - \overline{X})(X_{t+L} - \overline{X})}{\frac{1}{n-1}\sum_{t=1}^{n}(X_t - \overline{X})^2}$$

4. Types of time series

Stationary time series

If the statistics of the sample (mean, variance, covariance, autocorrelation, etc.) as calculated by above equations do not change with timing or the *length* of the sample, then the time series is said to be stationary to the second order moment (recall from lecture 2 that the variance of a variable X is defined to be the 2nd central moment), weakly stationary, or stationary in the broad sense. Mathematically one can write as:

$$E(X_t) = \mu$$
$$Var(X_t) = \sigma^2$$
$$Cov(X_t, X_{t+L}) = \lambda_L$$
$$Autocor(X_t, X_{t+L}) = \rho_L$$

In hydrology, moments of the third and higher orders are rarely considered because of the unreliability of their estimates. Second order stationarity, also called covariance stationarity, is usually sufficient in hydrology.

A process is strictly stationary when the distribution of X_t does not depend on time and when all simultaneous distributions of the random variables of the process are only dependent on their mutual time-lag. In another words, a process is said to be strictly stationary if its n-th (n for any integers) order moments do not depend on time and are dependent only on their time lag.

Nonstationary time series

If the values of the statistics of the sample (mean, variance, covariance, etc.) as calculated by above equations are dependent on the *timing* or the *length* of the sample, i.e. if a definite trend is discernible in the series, then it is a non-stationary series. Similarly, periodicity in a series means that it is non-stationary. Mathematically one can write as:

$$E(X_t) = \mu_t$$
$$Var(X_t) = \sigma_t^2$$
$$Cov(X_t, X_{t+L}) = \lambda_{L,t}$$
$$Autocor(X_t, X_{t+L}) = \rho_{L,t}$$

That means, all the calculated statistics have t as a subscript.

White noise time series

For a stationary time series, if the process is purely random and stochastically independent, the time series is called a white noise series. Mathematically one can write as:

$$E(X_t) = \mu$$
$$Var(X_t) = \sigma^2$$
$$Cov(X_t, X_{t+L}) = 0 \qquad \text{for all } L \neq 0$$
$$Autocor(X_t, X_{t+L}) = 0 \qquad \text{for all } L \neq 0$$

Gaussian time series

A Gaussian random process is a process (not necessarily stationary) of which all random variables are normally distributed, and of which all simultaneous distributions of random variables of the process are normal. When a Gaussian random process is weakly stationary, it is also strictly stationary, since the normal distribution is completely characterised by its first and second order moments.

5. Components of a time series

The time series X(t) can be analyzed into two components, the deterministic and the stochastic:

$$X(t) = D(t) + E(t)$$

The stochastic component E represents the random fluctuations of the. The deterministic component D represents all the variations that can be explained and expressed as a function of time. For example:

D(t) = Trend component + Periodic component

Trend is the tendency of the values of a time series to increase or decrease monotonically with time (e.g. the atmospheric CO_2 has an increasing trend the last 100 years). The trend can be linear or non-linear. Periodicity describes a repetitive tendency of the data, and is usually described by combination of sine functions.

Figure 12.3 shows all the components of a given time series. Note that here we have one more component, the catastrophic event. This could be for example an unusually high flow event.

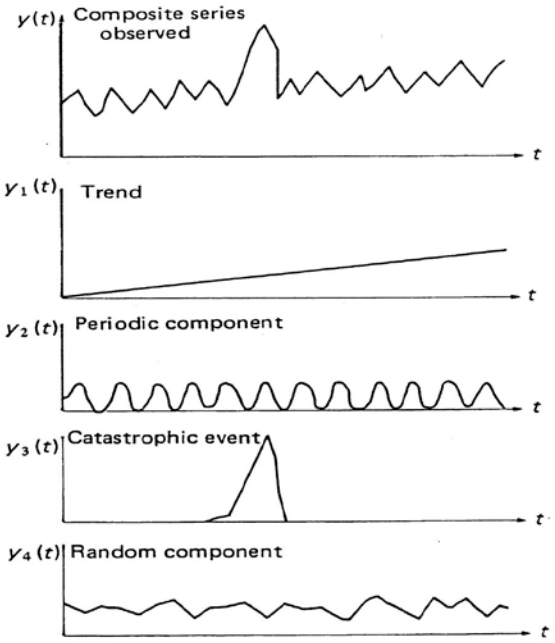


Figure 12.3: The time series components

6. Trend component

6.1 Identify trend

- **Run test method**

  It has been used in testing the stationarity of time series, randomness of series, and trend of a series

  Procedure:

  - For a time series, X, calculate the mean, $\overline{X}$ or median
  - Compare Xi with $\overline{X}$
    - If Xi > $\overline{X}$ then give a sign "+"
    - If Xi < $\overline{X}$ then give a sign "-"

  - Calculate the number of values with "+" and "-", and denote them as n_1 and n_2, respectively
  - Calculate the number of runs, and denote it as R
    - A 'run' is the sequence of + or – signs, for example
    - ++ - +++ -- + --- the number of run is 6
    - If the time series is characterized by noise only (no trend), the lengths of the runs will be randomly distributed
  - For large n_1, n_2 > 10 the distribution of R is approximated by a normal distribution with a mean of:

    $$\mu = \frac{2 \cdot n_1 \cdot n_2}{n_1 + n_2} + 1$$

    and variance of:

    $$\sigma^2 = \frac{2 \cdot n_1 \cdot n_2 \cdot (2 \cdot n_1 \cdot n_2 - n_1 - n_2)}{(n_1 + n_2)^2 \cdot (n_1 + n_2 - 1)}$$

  - We may then simply use the *z*-statistic

    $Z = \frac{(R - \mu)}{\sigma}$ then Z~N(0, 1)
  - If $|Z| > Z_{1-\alpha/2}$ then there is a trend

    where α is the significance level.

Example 12.1

Monthly average air temperature measurements are available form a station (fig.12.4). Check if there is a trend using run test.

The data were created on Matlab using: x=3*sin((0:(2*pi/12):10*2*pi))+0.005*(1:121)+10+2*rand(1,121); According to this dataset there is annual periodicity and the pre-assigned trend equals to +0.005 ^{o}C/month.

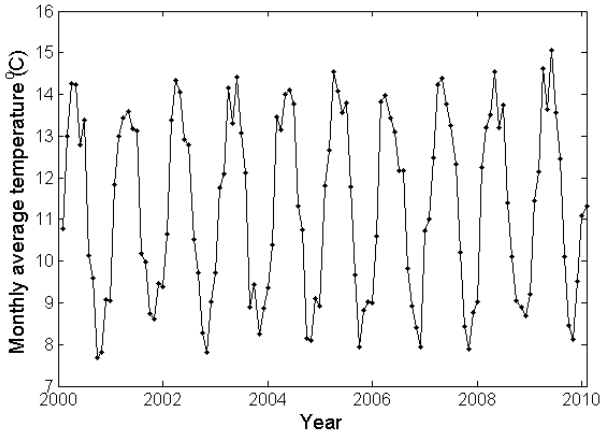


Figure 12.4: Monthly average air temperature

Solution:

$\bar{x} = 11.297$
Assign '+' and '-' to each x that is greater than or smaller than $\bar{x}$.

Count the number of positive and negative runs: n_1=11, n_2=11.

Total number of runs: R= n_1+n_2=22

Since n_1, n_2 >10, calculate μ and σ^2 as:

$$\mu = \frac{2 \cdot n_1 \cdot n_2}{n_1 + n_2} + 1 = 12$$

$$\sigma^2 = \frac{2 \cdot n_1 \cdot n_2 \cdot (2 \cdot n_1 \cdot n_2 - n_1 - n_2)}{(n_1 + n_2)^2 \cdot (n_1 + n_2 - 1)} = 5.2381$$

Calculate the z-statistic:

$$Z = \frac{R - \mu}{\sigma} = \frac{22 - 12}{\sqrt{5.2381}} = 4.37$$

Critical value of Z for significance level α=0.05 (from normal distribution table):

$$Z_{1-\frac{\alpha}{2}} = Z_{0.975} = 1.96$$

$\Rightarrow Z > Z_{1-\frac{\alpha}{2}}$ **According to the run test, the trend is significant.**

- **Mann-Kendall test method**

The test uses the raw (un-smoothed) hydrologic data to detect possible trends. The Kendall statistic was originally devised by Mann (1945) as a non-parametric test for trend (used without specifying if the trend is linear or non-linear). Later the exact distribution of the test statistic was derived by Kendall (1975).
There have been many different variants for Mann-Kendall method in the literature, one of which is described as follows.

The Mann-Kendall test is based on the test statistic S defined as follows:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sgn}(x_j - x_i) \qquad \text{(A)}$$

where x_j are the sequential data values, *n* is the length of the data set, and

$$\operatorname{sgn}(\theta) = \begin{cases} 1 & \text{if } \theta > 0 \\ 0 & \text{if } \theta = 0 \\ -1 & \text{if } \theta < 0 \end{cases}$$

Mann (1945) and Kendall (1975) have documented that when $n \geq 8$, the statistic S is approximately normally distributed with the mean and the variance as follows:

$$E(S) = 0$$

$$Var(S) = \frac{n(n-1)(2n+5) - \sum_{p=1}^{q} t_p (t_p - 1)(2t_p + 5)}{18}$$

where:
n = number of data

q = the number of tied groups. This is the number of groups with equal values/ties; for example in: 1 2 3 3 5 1 1, q=2 because there are two values (1 and 3) that repeat several times in the dataset.

t_p = the number of ties for the p^{th} value. This is the number of data in the p^{th} group, where p goes from 1 to q. In the example above: t_1=3 (because 1 appears three times), and t_2=2 (because 3 appears two times). Note that the sum is independent of the order, i.e. the result would be the same if we said t_2=3 (1 appears 3 times) and t_1=2 (3 appears two times).

The standardized Mann-Kendall test statistic Z_{MK} is computed by:

$$Z_{MK} = \begin{cases} \dfrac{S-1}{\sqrt{Var(s)}} & S > 0 \\ 0 & S = 0 \\ \dfrac{S+1}{\sqrt{Var(s)}} & S < 0 \end{cases}$$

The standardized MK statistic Z follows the standard normal distribution with mean of zero and variance of one.

The hypothesis that there is not trend will be rejected if

$$|Z_{MK}| > Z_{1-\alpha/2}$$

where $Z_{1-\alpha/2}$ is the value read from a standard normal distribution table with α being the significance level of the test.

Table 12.1: An example showing how the Mann-Kendall statistic S (equation A above) is calculated:

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Number of	
6.5	6.2	6.8	7.0	7.1	6.9	7.5	7.1	7.6	7.8	Positive (+) signs	Negative (-) Signs
	6.2-6.5 (-)	6.8-6.5 (+)	7.0-6.5 (+)	7.1-6.5 (+)	6.9-6.5 (+)	7.5-6.5 (+)	7.1-6.5 (+)	7.0-6.5 (+)	7.8-6.5 (+)	8	1
		6.8-6.2 (+)	7.0-6.2 (+)	7.1-6.2 (+)	6.9-6.2 (+)	7.5-6.2 (+)	7.1-6.2 (+)	7.6-6.2 (+)	7.8-6.2 (+)	8	0
			7.0-6.8 (+)	7.1-6.8 (+)	6.9-6.8 (+)	7.5-6.8 (+)	7.1-6.8 (+)	7.6-6.8 (+)	7.8-6.8 (+)	7	0
				7.1-7.0 (+)	6.9-7.0 (-)	7.5-7.0 (+)	7.1-7.0 (+)	7.6-7.0 (+)	7.8-7.0 (+)	5	1
					6.9-7.1 (-)	7.5-7.1 (+)	7.1-7.1 (0)	7.6-7.1 (+)	7.8-7.1 (+)	3	1
						7.5-6.9 (+)	7.1-6.9 (+)	7.6-6.9 (+)	7.8-6.9 (+)	4	0
							7.1-7.5 (-)	7.6-7.5 (+)	7.8-7.5 (+)	2	1
								7.6-7.1 (+)	7.8-7.1 (+)	2	0
									7.8-7.6 (+)	1	-
										S = 40 – 4 = 36	

Example 12.2

Apply Mann-Kendall test to the dataset from example 12.1 to check if the trend is significant.

Solution:

The length of the dataset is n=121, therefore it would be very time consuming to manually calculate table 12.1.
Using Matlab, and naming our variable as y, we have:

- To produce the differences in table 12.1:

```
for i=1:120
for j=i+1:121
    MK(i,j)=y(j)-y(i);
end
end
```

- To get the signs (+) and (-) in table 12.1:

```
MK=sign(MK)
```

- To get S in table 12.1:

```
S=sum(sum(MK))
```

$\Rightarrow$ S = 84 > 0

Since random function was used to produce the data, there are no ties. Otherwise we would count all the values that repeat in the dataset.
$\Rightarrow t_p = 0$

$$\Rightarrow V(S) = \frac{n(n-1)(2n-5)-\sum_{p=1}^{q} t_p(t_p-1)}{18} = \frac{n(n-1)(2n-5)}{18} = \frac{121(121-1)(2\cdot 121-5)}{18}$$

$\Rightarrow Var(S) = 1.9037 \cdot 10^5$

The standardized Mann-Kendall test statistic Z_{MK} is:

$$Z_{MK} = \frac{S-1}{\sqrt{Var(S)}} = \frac{84-1}{\sqrt{1.9037 \cdot 10^5}} = 0.19$$

From standard normal distribution table and significance level α=0.05 we have:

$$Z_{1-\frac{a}{2}} = Z_{0.975} = 1.96$$

$$\Rightarrow Z_{MK} < Z_{1-\frac{a}{2}}$$

Therefore the hypothesis that there is no trend cannot be rejected.

According to Mann-Kendall test the trend is not significant.

- **Linear-regression method**

The simple linear regression method is a parametric t-test method that can be used for trend detecting. It consists of two steps:
a. Fitting a linear simple regression equation with the time T as independent variable and the hydrological variable to be tested as dependent variable, Y, (i.e. $Y_i = \alpha + \beta T_i$)
b. Testing whether the slope, β of the regression equation is statistically different from zero (trend exists) or not statistically different from zero (no trend exists).
The parametric t-test requires the data to be tested to be normally distributed. The normality of the data series can be tested by applying the Kolmogorov–Smirnov test.

Procedure:

(1) Fit a linear regression equation with the time T as independent variable and the hydrologic data, Y as dependent variable, i.e.

$$Y = \alpha + \beta \cdot T$$

(2) Test if regression coefficient β is statistically significantly different from zero or not. If yes, there is a trend; if β >0 there is an increasing trend, if β<0 there is a decreasing trend.

Test of hypothesis concerning β can be made by noting that $(\beta - 0)/S_\beta$ has t distribution with n-2 degrees of freedom.

Thus the hypothesis $H_o : \beta = 0$ versus $H_a : \beta \neq 0$ is tested by computing the test statistic:

$$t = \frac{\beta - 0}{S_\beta}$$

where S_β is the standard deviation of the coefficient β, with:

$$S_\beta = S \Big/ \sqrt{\sum_{i=1}^{n} (T_i - \overline{T})^2}$$

and

$$S = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}$$

where S is the standard error of the regression, Y_i and $\hat{Y}_i$ are observed and estimated hydrologic variable from the regression equation, respectively.

(3) The hypothesis H_o, i.e. no trend, is rejected if $|t| > t_{1-\alpha/2,\, n-2}$

6.2 Modeling and removing the trend

As mentioned above, the trend can be linear or non-linear.
In the first case the trend line will be of the form: $Y(t)=a + b\cdot t$
In the second case the trend will be of the form: $Y(t) = a + bt + ct^2 + \ldots$

Removing the trend simply means to subtract it from the initial dataset. The remaining part E_t is the unexplained component:

$$E_t = X_t - T_t$$

7. Periodic component

7.1 Identify periodic component

In most annual series of data, there is no cyclical variation in the annual observations, but in the sequences of monthly data distinct periodic seasonal effects are at once apparent. The existence of periodic components may be investigated quantitatively by (1) Fourier analysis, (2) spectral analysis, and (3) autocorrelation analysis. Of which, the autocorrelation analysis method is widely used by hydrologists and will be discussed briefly in this section.

Autocorrelation method:
The procedure consists of two steps, calculating the autocorrelation coefficients and testing their statistical significance. For a series of data, X_t, the autocorrelation coefficient r_L between X_t and X_{t+L} are calculated and plotted against values of L (known as the lag), for all pairs of data L time units apart in the series:

$$r_L = \frac{1}{n-L}\sum_{t=1}^{n-L}(X_t - \overline{X})(X_{t+L} - \overline{X}) / \frac{1}{n}\sum_{t=1}^{n}(X_t - \overline{X})^2$$

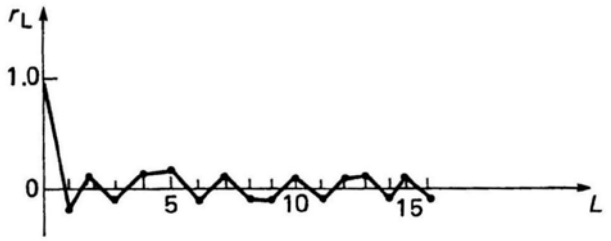
where $\overline{X}$ is the mean of the sample of n values of X_t and L is usually taken for values from zero up to n/4. A plot of r_L versus L forms the **correlogram**. The characteristics of a time series can be seen from the correlogram. Examples of correlograms are given in figure 11.5. Calculation of equation for different L gives the following cases:

- $-1 < r_L < 1$ for all L≠0
- $r_L = 1$ if L = 0. That is, the correlation of an observation with itself is one. This is also a direct result of the above equation.
- If $r_L \approx 0$ for all L ≠ 0, then the process is said to be a ***purely random process***. This indicates that the observations are linearly independent of each other. The correlogram for such a complete random time series is shown in fig.12.5(a).
- If $r_L \neq 0$ for some L ≠ 0, but after L > τ, then $r_\tau \approx 0$; the time series is ***stationary but not independent***, i.e. it still referred to as simply a random one (not purely random) since it has

a 'memory' up to L = τ. When $r_\tau \approx 0$, the process is said to have no memory for what occurred prior to time t-τ. The correlogram for such a non-independent stochastic process is shown in fig.12.5(b).

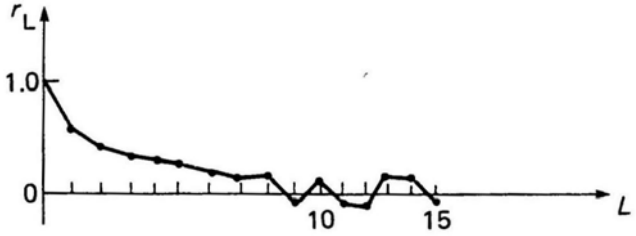
- If $r_L \neq 0$ for all L ≠ 0, it is ***a non-stationary time series*** since in the case the data containing a cyclic (deterministic) component. The correlogram would appear as in fig.12.5(c). Where T is the period of the cycle.

$r_L \approx 0$ for all $L \neq 0$



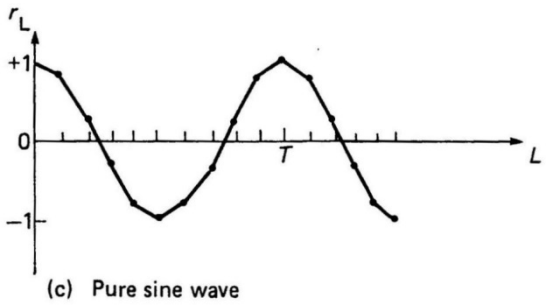
(a) Random, independent noise

$r_L \approx 0$ when L gets large



(b) Autoregressive, Markov process

r_L never approaches zero even for large k values ($r_L \neq 0$ for all $L \neq 0$)



(c) Pure sine wave

Figure 12.5: Examples of correlograms. (a) Purely random (stochastic) series – no memory; (b) Stationary series with memory; (c) Periodic series (non-stationary)

Testing the statistical significance of r_L:

If n gets larger, $\boldsymbol{r_L}$ will be approximately normally distributed with mean $\frac{-1}{n-1}$ and variance $\frac{(n-2)}{(n-1)^2}$ (see Haan, 2002). The confidence limits of r_L are estimated by:

$$l = \left(-1 - z_{1-\alpha/2}\sqrt{n-2}\right)/(n-1)$$
$$u = \left(-1 + z_{1-\alpha/2}\sqrt{n-2}\right)/(n-1)$$

H_0: $\mathbf{r_L} = 0$; H_a: $\mathbf{r_L} \neq 0$
If the calculated $\mathbf{r_L}$ falls outside these confidence limits, then the hypothesis that $\mathbf{r_L}$ is zero is rejected for significant level α and time lag L.

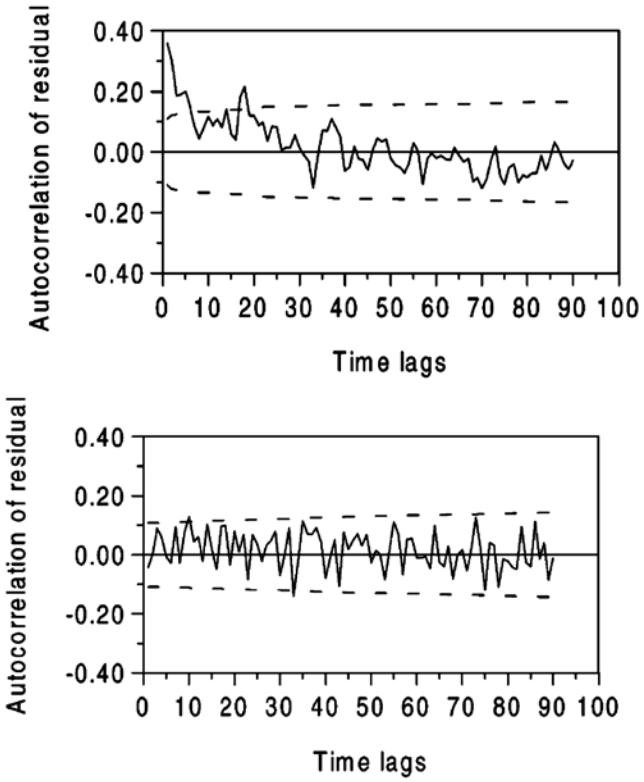


Figure 12.6: Examples of autocorrelation and its confidence interval. Correlated case (top); Independent case (bottom).

7.2 Modeling and removing the periodic component

Example 12.3

The periodic component of the monthly temperature can be modeled with a model as:

$$\hat{T}(t) = a + b\sin\left[\left(\frac{2\pi}{12}\right)(t-c)\right]$$

where ***t*** is the time in month, parameter ***a*** equals the mean monthly temperature, ***b*** is an amplification parameter, the parameter ***c*** is a phase parameter. For monthly temperature data in a station near Uppsala, I have briefly manually calibrated that a = 5.32, b = 10 and c = 4.1.

$$\Rightarrow \hat{T}(t) = a + b(\sin\left[\left(\frac{2\pi}{12}\right)(t-c)\right] = 5.3 + 10\sin[\left(\frac{2\pi}{12}\right)(t-4.1)]$$

The result of the *sine* curve model is as follows (fig.11.7), where the blue curve is the observed temperature, *T(t)*, and the red curve the modeled temperature, $\hat{T}(t)$.

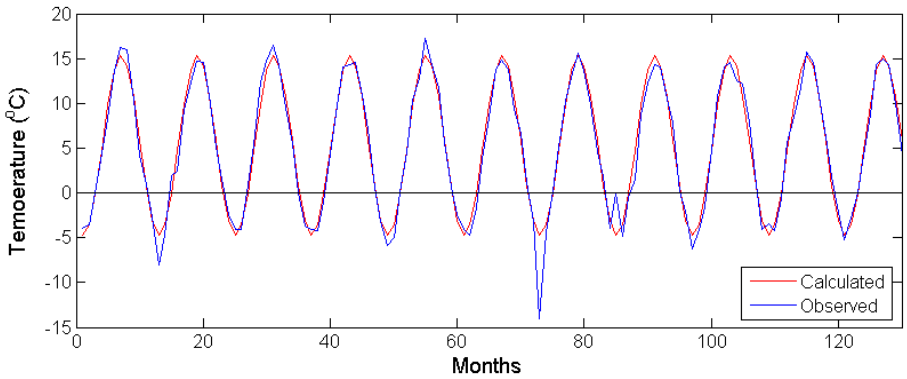


Figure 12.7: Observed and calculated temperatures.

A stationary residual time series can be obtained using $T(t) - \hat{T}(t)$:

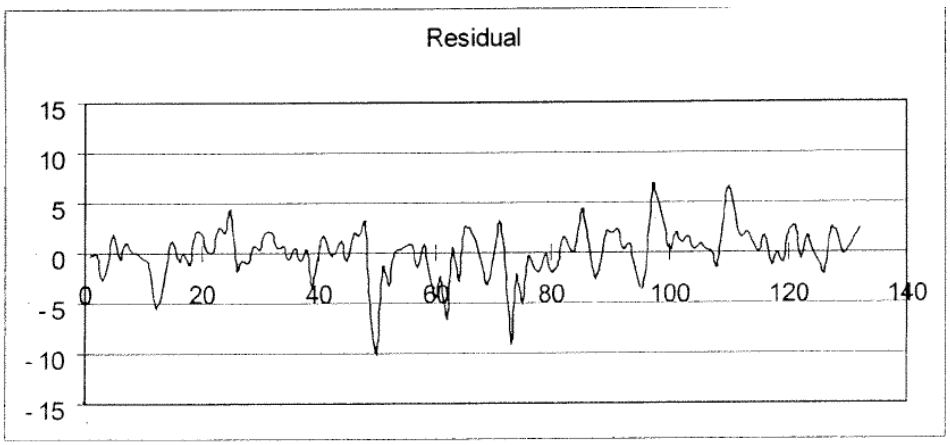


Figure 12.8: Residuals= $T(t) - \hat{T}(t)$

Example 12.4

The periodic component in the upper curve of figure 12.9 is modeled by harmonic regression, using the following equation:

$$s_t = \alpha_0 + \sum_{k=1}^{h} \left(\alpha_k \cos(\lambda_k t) + \beta_k \sin(\lambda_k t) \right)$$

The lower curve in figure 12.9 is the plot of the equation above, after fitting the parameters. The periodic component can be removed by subtracting the lower curve from the upper curve; then a stationary residual time series is obtained.

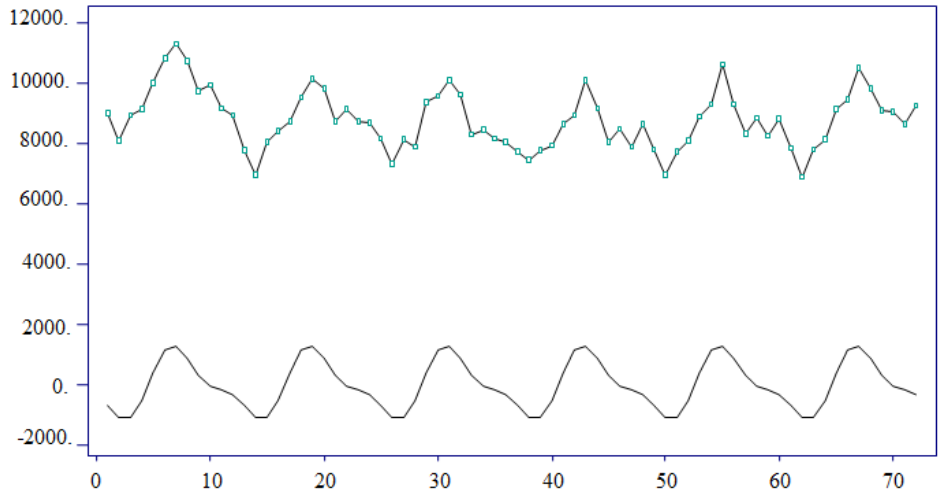


Figure 12.9: An example of the time series and the fitted periodic component

Theoretically speaking, any time series no matter how complex can be modeled by harmonic regression involving many cosine and sine terms. A very complicate time series can be modeled by a combination of several deterministic and stochastic models (fig.12.10).

In order to remove the periodic component, subtract the modeled periodic term form the original data:

$$E_t = X_t - P_t$$

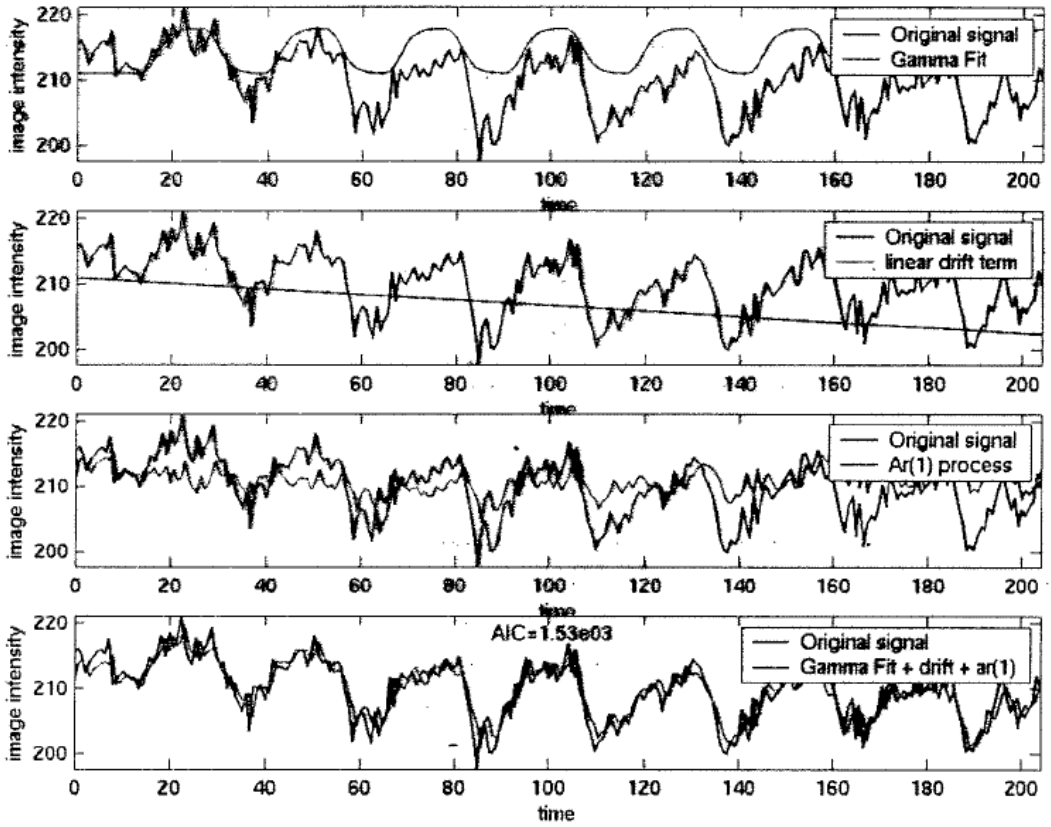


Figure 12.10: An example of complex time series, for which three modeled components are used. As a result the modeled curve is very close to the original (lowest subplot).

8. Homoscedasticity test – the Kruskal-Wallis Test

The Kruskal-Wallis test, or H test, enables us to test the null hypothesis that k independent random samples come from identical populations. It is a nonparametric test. The method assumes that the variable has a continuous distribution, but nothing is said about the form of the population distribution or distributions from which the samples were drawn. The test is based on the statistic:

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

In the test, all observations are ranked jointly, and R_i is the sum of the ranks occupied by the n_i observations of the ith sample, and $n_1 + n_2 + ... + n_k = n$. When $n_i > 5$ for all i and the null hypothesis is true, the sampling distribution of the H statistic is well approximated by the chi-square distribution with k-1 degrees of freedom. The null hypothesis of homoscedasticity will be rejected for a given significance level, α if computed H is bigger than $\chi^2_{1-\alpha,k-1}$. An application example of this method can be seen from (**Xu, C-Y**, 2001. Statistical analysis of a conceptual water balance model, methodology and case study, *Water Resources Management* 15: 75-92. (pdf file))

9. Steps for constructing a time series model

Step 1: Model any *deterministic* trend and/or seasonal components that may be present, and remove these from the data.
Step 2: Choose a model (from among a family of probability models) that *best* represents the distribution of residuals from step 1.
Step 3: Estimate the parameters of the chosen model.
Step 4: Check model for goodness of fit.

Example 12.5

Assume that we have temperature measurements (n=401), plotted on figure 12.11. Identify and remove the different components of the dataseries, so that the remaining part is a stationary series.
(Note that here a synthetic dataset is used, and t is an arbitrary time step.)

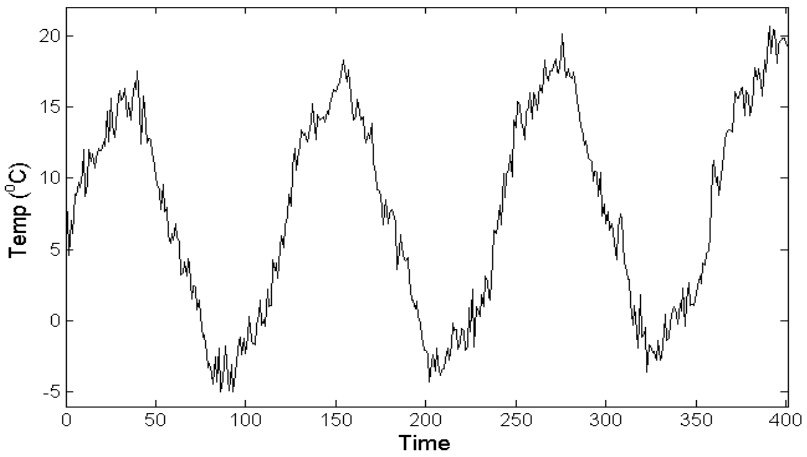


Figure 12.11: Temperature measurements

Solution:

Three components can be identified above:

- T(t) : trend component
  (deterministic, changes slowly with t);
- S(t) : seasonal component
  (deterministic, period d);
- E(t) : noise component
  (random, stationary).

The model will be: **Xt = Tt + St + Et**

First perform simple linear regression and fit a straight line of the form: $T(t) = c_1 + c_2 \cdot t$.

$\Rightarrow T(t)=x+0.1\cdot t$
Subtract *T(t)* form the dataset (fig.12.12,top): *Temp(t)-T(t)*

Next, try to fit a periodic component of the form: $S(t)= c_3 \cdot sin(c_4 \cdot t+c_5)$
$\Rightarrow S(t) = 10 \cdot \sin(\frac{2\pi}{12} \cdot t)$
Subtract *S(t)* from previous step (fig.11.12, bottom): *Temp(t)-T(t)-S(t).*

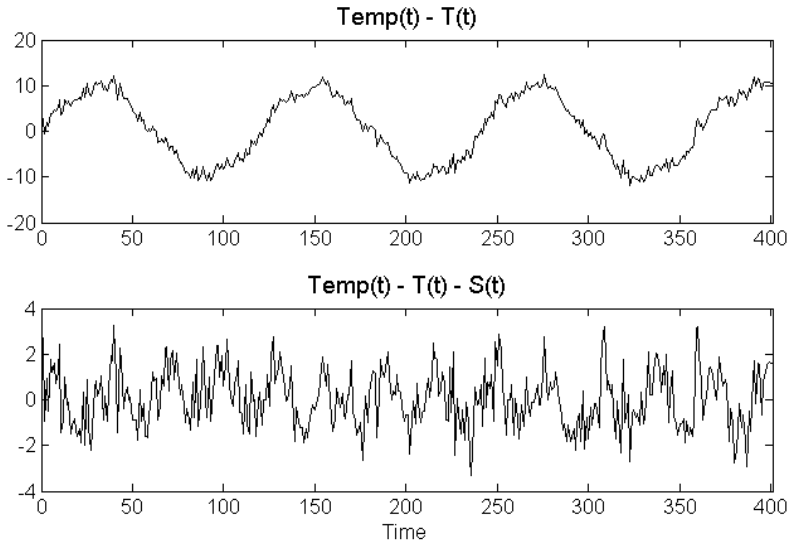


Figure 12.12: Initial time series with removed trend (top); Initial time series with removed trend and periodic component (bottom).

In figure 12.12(bottom) it seems that there is still some periodicity, but it's difficult to identify due to the noise. At this point we perform autocorrelation test to see if the remaining component E(t) is stationary (fig.12.13). The upper and lower confidence limits are estimated at α=0.05 as:

$$C.I. = \frac{-1 \pm z_{1-a/2} \cdot \sqrt{n-2}}{n-1}$$

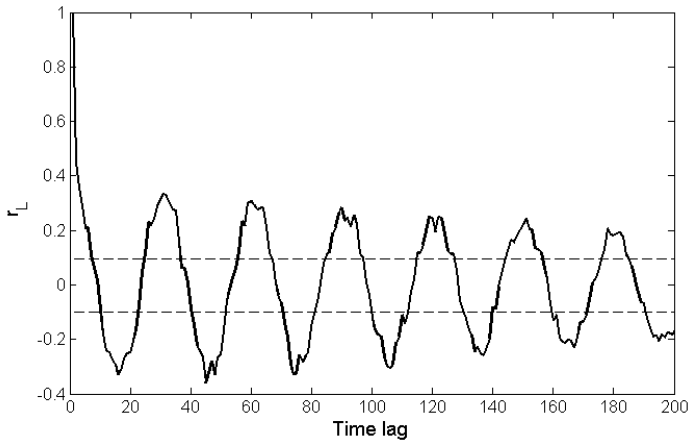


Figure 12.13: Autocorrelation of Temp(t)-T(t)-S(t)

From the plot above it is obvious that there is some periodic component left.

Find the remaining periodic component: $S'(t) = \sin(\frac{2\pi}{3} \cdot t)$

Subtract S´(t) from the previous dataset (fig.12.14,top) and perform autocorrelation analysis again (fig.12.14bottom).

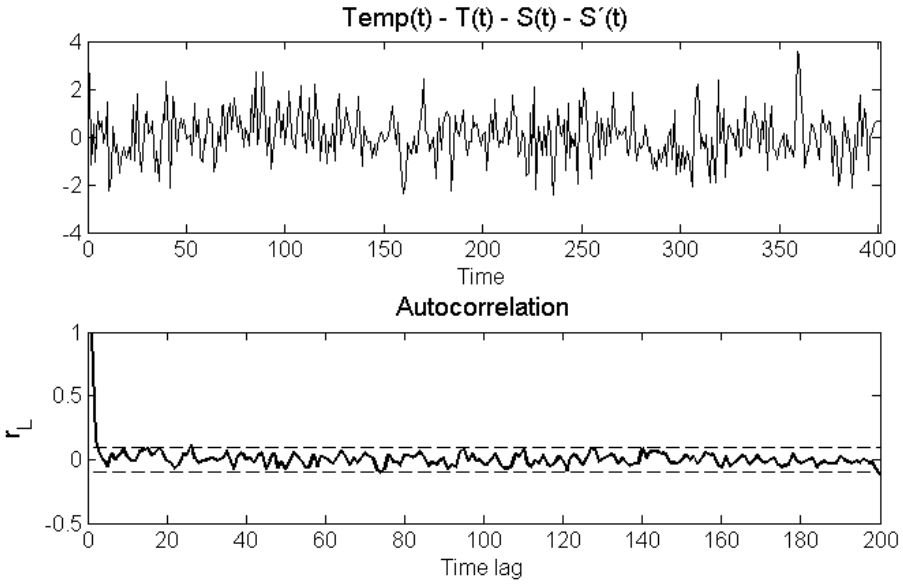


Figure 12.14: Initial time series with removed trend and two periodic components (top); Autocorrelation of *E´(t) = Temp(t) – T(t) – S(t) – S´(t)* (bottom)

r_L remains inside the 95% confidence interval around 0 for all L≠0, which means that it is not significant for any of the time lags. Therefore E´(t) is white noise.

We can write the equation of the initial time series as:

$$Temp(t) = 5 + 0.1 \cdot t + 10 \cdot \sin\left(\frac{2\pi}{12} \cdot t\right) + \sin\left(\frac{2\pi}{3} \cdot t\right) + E'(t)$$

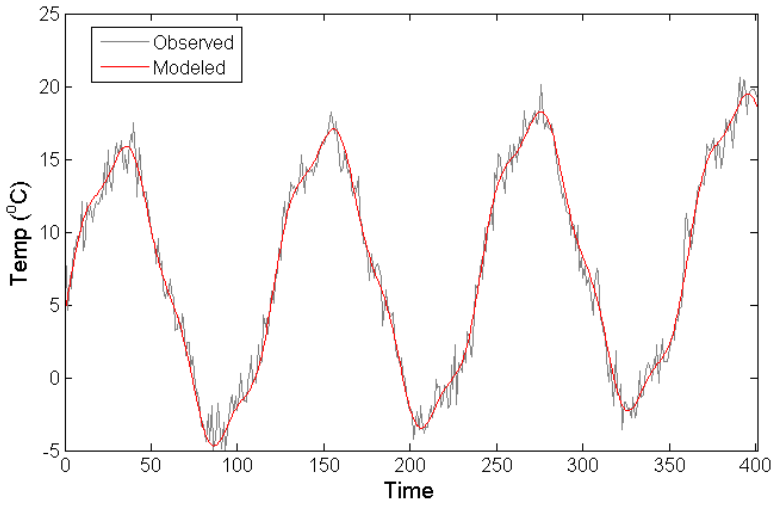


Figure 12.15: Observed and modeled time series

10. Stochastic Models

- In time series analysis, the goal is to develop a model that provides a reasonably close approximation of the underlying process that generates the time series data.

- These models can then be used to predict future values of the time series variable.

- Models:
  - Statistical model for purely random data with knowing probability distribution
  - Models for stationary time series
  - Models for non-stationary time series

To be discussed in next lecture.

Chapter 13

Stochastic models

From the last lecture we have classified a time series into a purely random series (white noise), a stationary series and a non-stationary series. This lecture we will discuss the difference of stochastic model for these time series.

1. Objectives

a. To extend the short record for design and operational purposes
b. To forecast the future values

2. Stochastic models for purely random series with known probability distribution

a. White noise series, i.e. stationary + independent (you shall remember what is a stationary times series and an independent series.)
b. Probability distribution is known

Example: if X_t is a white noise series and normally distributed

i.e. $X_t \sim N(\mu_x, \sigma_x^2)$

then the model can be

$$X_t = \mu_x + \sigma_x Z \qquad \text{(A)}$$

where μ_x and σ_x are the mean and standard deviation of the X_t series, which can be estimated by:

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

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

and $Z \sim N(0,1)$ is a random series having standard normal distribution which can be generated by Monte Carlo simulation.

Note, in the Frequency Analysis chapter, we have derived that if X is normally distributed, i.e. $N(\bar{x}, s^2)$, then:

$$X_T = \bar{x} + K_T S \qquad \text{(B)}$$

$$\Rightarrow K_T = \frac{X_T - \bar{x}}{s}$$

That means that K_T is the standardized normal variate Z ($Z = \frac{X - \mu}{\sigma}$), i.e.

$K_T = Z \sim N(0,1)$

So equations (A) and (B) are the same model for modeling normally distributed white noise time series.

3. Stochastic models for stationary time series

When persistence (memory) is present, synthetic sequences cannot be constructed by taking a succession of a sample values from a probability distribution, since this will not take account of the relation between each number of the sequence and those preceding it. Different models are available for such times series.

General Model

Model Output	=	Linear combination of past output	+	Linear combination of present and past inputs
		AR part		**MA part**

3.1 Autoregressive models – AR models

The AR model is used for autocorrelated stationary time series.
The general form of a p-th order autoregressive model, AR(p), is

$$y_t = \mu + \beta_1(y_{t-1} - \mu) + \beta_2(y_{t-2} - \mu) + ... + \beta_p(y_{t-p} - \mu) + \varepsilon_t$$
$$= \mu + \sum_{i=1}^{p} \beta_i (y_{t-i} - \mu) + \varepsilon_t \qquad (1)$$

where μ is mean value of the series,
p is the order of AR model, written as AR(P)
β_i are the regression coefficients,
ε_t noise or prediction error, normally assumed as $N(0, \sigma_\varepsilon^2)$

Often, equation (1) is written as

$$y_t = c + \beta_1 y_{t-1} + \beta_2 y_{t-2} + ... + \beta_p y_{t-p} + \varepsilon_t$$
$$= c + \sum_{i=1}^{p} \beta_1 y_{t-1} + \varepsilon_t \qquad (2)$$

Where constant $c = \left(1 - \sum_{i=1}^{p} \beta_i\right)\mu$

There are p+2 parameters to be estimated: β_1, β_2, …β_p, μ (=mean) and σ_ε^2 (=variance of residuals).

Sometimes, a centralized form of equation (1) is used:
Let $x_t = y_t - \mu$ to get centralised AR(p) model

Then (1) can be written as

$$x_t = \underbrace{\beta_1 x_{t-1} + \beta_2 x_{t-2} + ... + \beta_p x_{t-p}}_{\text{deterministic}} + \underbrace{\varepsilon_t}_{\text{stochastic}}$$

or

$$x_t = \sum_{i=1}^{p} x_{t-i}\beta_i + \varepsilon_t \qquad (3)$$

So equations (1) – (3) are three different forms of the same model.

The output of the model, x_t, is a function of the past output values, x_{t-1}, …, x_{t-p}, and a random term.

- **The first order autoregression model – Markov model**

  $y_t = \mu + \beta_1(y_{t-1} - \mu) + \varepsilon_t$
  or:
  $y_t = c + \beta_1 y_{t-1} + \varepsilon_t$ where $c = (1-\beta_1)\mu$
  or:
  $x_t = \beta_1 x_{t-1} + \varepsilon_t$ (for centralised)

  has found particular application in hydrology for modelling annual mean discharge.

It has three parameters to be estimated: μ, β_1 and σ_ε^2 for the ε_t term.

Parameter estimation:

The Yule-Walker equations shows the relation between regression coefficient β and autocorrelation coefficient ρ

$$\rho_k = \sum_{j=1}^{P} \beta_j \rho_{k-j}, \qquad k=0, 1, \ldots$$

where ρ_k is the autocorrelation coefficient

note that: $\rho_o = 1$

$$\rho_1 = \beta_1 \cdot \rho_o = \beta_1$$

$$\sigma_y^2 = \sigma_\varepsilon^2 + \beta_1^2 \sigma_y^2$$

From above equations we get:

$$\hat{\beta}_1 = \rho_1$$

$$\hat{\sigma}_\varepsilon^2 = \sigma_y^2(1-\beta_1^2)$$

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The procedure for generating a series of values for y_t using AR(1) model is:

(1) estimate μ_y, σ_y, and β_1 by $\bar{y} = \mu_y$, $s_y = \sigma_y$, and $r_1 = \beta_1$ respectively, and $\sigma_\varepsilon^2 = \sigma_y^2(1-\beta_1^2)$
(2) select a z_t at random from a N(0, 1) distribution, and
(3) select an initial value for y_{t-1},
(4) calculate y_t based on $\bar{y}$, s_y, and β_1, and y_{t-1} by

$$y_t = \mu_y + \beta_1(y_{t-1} - \mu_y) + z_t\sigma_y\sqrt{(1-\beta_1^2)}$$

(5) delete the first 50 values to get rid of the influence from initial value

- **The second order autoregression model**

$$y_t = \mu + \beta_1(y_{t-1} - \mu) + \beta_2(y_{t-2} - \mu) + \varepsilon_t$$

Parameter estimation: $\mu_y, \beta_1, \beta_2, \sigma_\varepsilon$

By using the Yule-Walker equation we get:

$\rho_1 = \beta_1 \cdot \rho_o + \beta_2\rho_{-1}$
$\rho_2 = \beta_1 \cdot \rho_1 + \beta_2\rho_0$
or, because of symmetry of the autocorrelation function ($\rho_o = 1$, $\rho_{-1} = \rho_1$)

$\rho_1 = \beta_1 + \beta_2\rho_1$ i.e. $\rho_1 = \dfrac{\beta_1}{1-\beta_2}$

$\rho_2 = \beta_1 \cdot \rho_1 + \beta_2$ i.e. $\rho_2 = \beta_2 + \dfrac{\beta_1^2}{1-\beta_2}$

from which parameters β_1 and β_2 can be estimated as:

$$\hat{\beta}_1 = \frac{\rho_1(1-\rho_2)}{1-\rho_2^2} \qquad \hat{\beta}_2 = \frac{\rho_2 - \rho_1^2}{1-\rho_1^2}$$

for AR(2) model,

$$\sigma_\varepsilon^2 = \sigma_x^2(1-\beta_1\rho_1 - \beta_2\rho_2)$$

substitute above equations for ρ_1 and ρ_2 we get the estimation of σ_ε.

$$\hat{\sigma}_\varepsilon^2 = \sigma_x^2 \frac{1+\beta_2}{1-\beta_2}\left[(1-\beta_2)^2 - \beta_1^2\right]$$

and $\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

The procedure for using AR(2) models is the same as for AR(1) model.

3.2 The Moving average model - MV(q)

The sequence is MA type if

$$y_t = \mu_y + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + ... + \theta_q\varepsilon_{t-q}$$

or

$$x_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + ... + \theta_q\varepsilon_{t-q}$$

where ε_t is a white noise with $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$

θ_i are parameters of order q, i.e. parameter $\theta_k = 0$ for k > q

Above equations give the definition of MA model: For a white noise or purely random series, ε_t with mean zero and standard deviation σ_ε.

That is, a moving average model is conceptually a linear regression of the current value of the series against previous (unobserved) white noise error terms or random shocks. The random shocks at each point are assumed to come from the same distribution, typically a normal distribution, with location at zero and constant scale. The distinction in this model is that these random shocks are propagated to future values of the time series. Fitting the MA estimates is more complicated than with autoregressive models because the error terms are not observable. This means that iterative non-linear fitting procedures need to be used in place of linear least squares.

The output of the model x_t, is a function of the current and past of the inputs, ε_t, ε_{t-1}, …, ε_{t-q}.

The MA(1) is then

$$x_t = \varepsilon_t + \theta_1\varepsilon_{t-1}$$

There are two parameters, σ_ε^2 and θ_1

In stochastic hydrology, the moving average process describes the deviations of a sequence of events from their mean.

Parameter estimation for MV(1)

It can be shown that (for details see page 356-357 of Haan, 2002)

$$\sigma_x^2 = (1+\theta_1^2)\sigma_\varepsilon^2 \qquad \rho_1 = -\frac{\theta_1}{1+\theta_1^2}$$

where σ_x^2 and ρ_1 are variance and autocorrelation of lag 1 of the sample data series.

solve the above equations simultaneously we get

$$\hat{\theta}_1 = \frac{-1 \pm \sqrt{1-4\rho_1^2}}{2\rho_1} \qquad \hat{\sigma}_\varepsilon^2 = \sigma_x^2 \Big/ \left(1+\theta_1^2\right)$$

Autocorrelation coefficient ρ_1 and variance σ_x^2 can be calculated from the observed data.

3.3 The ARMA model

In statistics and time series analysis, autoregressive moving average (ARMA) models, sometimes called Box-Jenkins models after the iterative Box-Jenkins methodology usually used to estimate them, are typically applied to time series data. Given a time series of data X_t, the ARMA model is a tool for understanding and, perhaps, predicting future values in this series. The model consists of two parts, an autoregressive (AR) part and a moving average (MA) part. The model is usually then referred to as the ARMA(p,q) model where p is the order of the autoregressive part and q is the order of the moving average part (as defined below).

In the case, x(t) is a mixed process where the output is a function of past outputs and current/past inputs

$$x_t = c + \sum_{i=1}^{p} x_{t-i}\beta_i + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

All notations have same meaning an before. The error terms ε_t are generally assumed to be independent identically-distributed random variables (i.i.d.) sampled from a normal distribution with zero mean: $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ where σ_ε^2 is the variance of the error.
There are p+q+2 parameters.

The ARMA (1,1) model is:

$$x_t = \beta_1 x_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

Parameters $\beta_1, \theta_1, and \sigma_\varepsilon^2$ can be estimated by solving the following equations:

$$\rho_1 = \frac{(\beta_1 - \theta_1)(1-\theta_1\beta_1)}{1+\theta_1^2 - 2\beta_1\theta_1}$$

$$\rho_2 = \beta_1 \rho_1$$

$$\sigma_\varepsilon^2 = \frac{(1-\beta_1^2)\sigma_x^2}{1-2\beta_1\theta_1+\theta_1^2}$$

3.4 Stochastic model for nonstationary time series
First order Markov process with periodicity: Thomas - Fiering model

The first order Markov model of the previous section assumes that the process is stationary in its first three moments. It is possible to generalise the model so that the periodicity in hydrologic data is accounted for to some extent. The main application of this generalisation has been in generating monthly streamflow where pronounced seasonality in the monthly flows exists. In its simplest form, the method consists of the use of twelve linear regression equations. If, say, twelve years of record are available, the twelve January flows and the twelve December flows are abstracted and January flow is regressed upon December flow; similarly, February flow is regressed upon January flow, and so on for each month of the year.

$$q_{jan} = \overline{q}_{jan} + b_{jan}(q_{dec} - \overline{q}_{dec}) + \varepsilon_{jan}$$
$$q_{feb} = \overline{q}_{feb} + b_{feb}(q_{jan} - \overline{q}_{jan}) + \varepsilon_{feb}$$
$$\ldots\ldots$$

Fig.8 shows a regression analysis of q_{j+1} on q_j, pairs of successive monthly flows for the months (j+1) and j over the years of record where j = 1, 2, 3, ..., 12 (Jan, Feb, ... Dec) and when j = 12, j+1 = 1 = Jan (there would be 12 such regressions). If the regression coefficient of month j+1 on j is b_j, then the regression line values of a monthly flow, $\hat{q}_{j+1}$, can be determined from the previous months flow q_j, by the equation:

$$\hat{q}_{j+1} = \overline{q}_{j+1} + b_j(q_j - \overline{q}_j)$$

To account for the variability in the plotted points about the regression line reflecting the variance of the measured data about the regression line, a further component is added:

$$Z \cdot s_{j+1}\sqrt{(1-r_j^2)}$$

where s_{j+1} is the standard deviation of the flows in month j+1, r_j is the correlation coefficient between flows in months j+1 and j throughout the record, and Z = N(0, 1), a normally distributed random deviate with zero mean and unit standard deviation. The general form may written as

$$\hat{q}_{j+1,i} = \overline{q}_{j+1} + b_j(q_{j,i-1} - \overline{q}_j) + Z_{j+1,i} \cdot s_{j+1}\sqrt{(1-r_j^2)}$$

Where $b_j = r_j \times s_{j+1} / s_j$, there are 36 parameters for the monthly model ($\overline{q}$, r and s for each month). The subscript *j* refers to month. For monthly synthesis *j* varies from 1 to 12 throughout the year. The subscript *i* is a serial designation from year 1 to year *n*. Other symbols are the same as mentioned earlier.

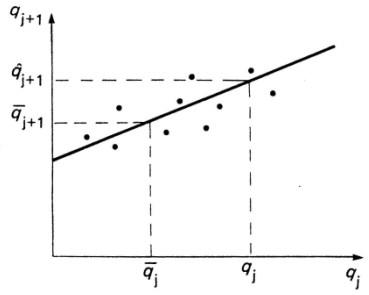


Figure 13.1 Regression analysis of q_{j+1} on q_j

The procedure for using the model is as follows:

(1) For each month, j = 1, 2, … 12, calculate

(a) the mean flow $\bar{q}_j = \frac{1}{n}\sum_i q_{j,i}$; $\quad (i = j, 12+j, 24+j, \cdots)$

(b) the standard deviation $S_j = \sqrt{\frac{\sum_i (q_{j,i} - \bar{q}_j)^2}{n-1}}$

(c) the correlation coefficient with flow in the preceding month,

$$r_j = \frac{\sum_{i=1} (q_{j,i} - \bar{q}_j)(q_{j+1,i} - \bar{q}_{j+1})}{\sqrt{\sum_i (q_{j,i} - \bar{q}_j)^2 \sum_i (q_{j+1,i} - \bar{q}_{j+1})^2}}$$

(d) the slope of the regression equation relating the month's flow to flow in the preceding month:

$$b_j = r_j \frac{S_{j+1}}{S_j}$$

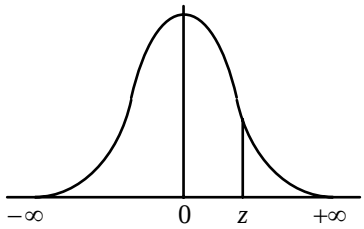
(2) The model is then set of twelve regression equations

$$\hat{q}_{j+1,i} = \bar{q}_{j+1} + b_j (q_{j,i} - \bar{q}_j) + Z_{j+1,i} \cdot s_{j+1} \sqrt{(1 - r_j^2)}$$

where Z is a random Normal deviate N(0, 1).

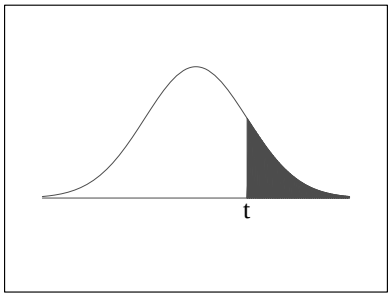
(3) To generate a synthetic flow sequence, calculate (generate) a random number sequence {Z_1, Z_2, … }, and substitute in the model.

NORMAL DISTRIBUTION TABLE



	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

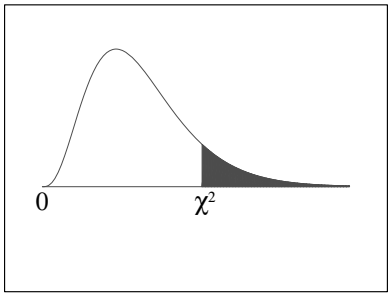
t-Distribution Table



The shaded area is equal to α for $t = t_\alpha$.

df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
32	1.309	1.694	2.037	2.449	2.738
34	1.307	1.691	2.032	2.441	2.728
36	1.306	1.688	2.028	2.434	2.719
38	1.304	1.686	2.024	2.429	2.712
∞	1.282	1.645	1.960	2.326	2.576

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

F distribution critical value landmarks

Table entries are critical values for *F** with probably *p* in right tail of the distribution.

Figure of *F* distribution (like in Moore, 2004, p. 656) here.

Degrees of freedom in denominator (df2)		Degrees of freedom in numerator (df1)											
	p	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**12**	**24**	**1000**
1	0.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	60.71	62.00	63.30
	0.050	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	243.9	249.1	254.2
	0.025	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.6	976.7	997.3	1017.8
	0.010	4052	4999	5404	5624	5764	5859	5928	5981	6107	6234	6363
	0.001	405312	499725	540257	562668	576496	586033	593185	597954	610352	623703	636101
2	0.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.41	9.45	9.49
	0.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.41	19.45	19.49
	0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.41	39.46	39.50
	0.010	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.42	99.46	99.50
	0.001	998.38	998.84	999.31	999.31	999.31	999.31	999.31	999.31	999.31	999.31	999.31
3	0.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.22	5.18	5.13
	0.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.74	8.64	8.53
	0.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.34	14.12	13.91
	0.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.05	26.60	26.14
	0.001	167.06	148.49	141.10	137.08	134.58	132.83	131.61	130.62	128.32	125.93	123.52
4	0.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.90	3.83	3.76
	0.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.91	5.77	5.63
	0.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.75	8.51	8.26
	0.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.37	13.93	13.47
	0.001	74.13	61.25	56.17	53.43	51.72	50.52	49.65	49.00	47.41	45.77	44.09
5	0.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.27	3.19	3.11
	0.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.68	4.53	4.37
	0.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.52	6.28	6.02
	0.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	9.89	9.47	9.03
	0.001	47.18	37.12	33.20	31.08	29.75	28.83	28.17	27.65	26.42	25.13	23.82
6	0.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.90	2.82	2.72
	0.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.00	3.84	3.67
	0.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.37	5.12	4.86
	0.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.72	7.31	6.89
	0.001	35.51	27.00	23.71	21.92	20.80	20.03	19.46	19.03	17.99	16.90	15.77
7	0.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.67	2.58	2.47
	0.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.57	3.41	3.23
	0.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.67	4.41	4.15
	0.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.47	6.07	5.66
	0.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	13.71	12.73	11.72
8	0.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.50	2.40	2.30
	0.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.28	3.12	2.93
	0.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.20	3.95	3.68
	0.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.67	5.28	4.87
	0.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.19	10.30	9.36
9	0.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.38	2.28	2.16
	0.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.07	2.90	2.71
	0.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	3.87	3.61	3.34
	0.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.11	4.73	4.32
	0.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	9.57	8.72	7.84

Critical values computed with Excel 9.0

Degrees of freedom in denominator (df2)		Degrees of freedom in numerator (df1)										
	p	1	2	3	4	5	6	7	8	12	24	1000
10	0.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.28	2.18	2.06
	0.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.91	2.74	2.54
	0.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.62	3.37	3.09
	0.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.71	4.33	3.92
	0.001	21.04	14.90	12.55	11.28	10.48	9.93	9.52	9.20	8.45	7.64	6.78
12	0.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.15	2.04	1.91
	0.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.69	2.51	2.30
	0.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.28	3.02	2.73
	0.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.16	3.78	3.37
	0.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.00	6.25	5.44
14	0.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.05	1.94	1.80
	0.050	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.53	2.35	2.14
	0.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.05	2.79	2.50
	0.010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	3.80	3.43	3.02
	0.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.13	5.41	4.62
16	0.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	1.99	1.87	1.72
	0.050	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.42	2.24	2.02
	0.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	2.89	2.63	2.32
	0.010	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.55	3.18	2.76
	0.001	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.20	5.55	4.85	4.08
18	0.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	1.93	1.81	1.66
	0.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.34	2.15	1.92
	0.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.77	2.50	2.20
	0.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.37	3.00	2.58
	0.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.13	4.45	3.69
20	0.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.89	1.77	1.61
	0.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.28	2.08	1.85
	0.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.68	2.41	2.09
	0.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.23	2.86	2.43
	0.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	4.82	4.15	3.40
30	0.100	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.77	1.64	1.46
	0.050	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.09	1.89	1.63
	0.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.41	2.14	1.80
	0.010	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	2.84	2.47	2.02
	0.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.00	3.36	2.61
50	0.100	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.68	1.54	1.33
	0.050	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	1.95	1.74	1.45
	0.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.22	1.93	1.56
	0.010	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.56	2.18	1.70
	0.001	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.44	2.82	2.05
100	0.100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.61	1.46	1.22
	0.050	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.85	1.63	1.30
	0.025	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.08	1.78	1.36
	0.010	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.37	1.98	1.45
	0.001	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.07	2.46	1.64
1000	0.100	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.55	1.39	1.08
	0.050	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.76	1.53	1.11
	0.025	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20	1.96	1.65	1.13
	0.010	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.20	1.81	1.16
	0.001	10.89	6.96	5.46	4.65	4.14	3.78	3.51	3.30	2.77	2.16	1.22

Use StaTable, WinPepi > WhatIs, or other reliable software to determine specific *p* values

A.16. Critical values for the Kolmogorov–Smirnov test statistic

Sample size (n)	Significance level				
	.20	.15	.10	.05	.01
1	.900	.925	.950	.975	.995
2	.684	.726	.776	.842	.929
3	.565	.597	.642	.708	.829
4	.446	.474	.510	.624	.734
5	.446	.474	.510	.563	.669
6	.410	.436	.438	.486	.577
7	.381	.405	.438	.486	.577
8	.358	.381	.411	.457	.543
9	.339	.360	.388	.432	.514
10	.322	.342	.368	.409	.486
11	.307	.326	.352	.391	.468
12	.295	.313	.338	.375	.450
13	.284	.302	.325	.361	.433
14	.274	.292	.314	.349	.418
15	.266	.283	.304	.338	.404
16	.258	.274	.295	.328	.391
17	.250	.266	.286	.318	.380
18	.244	.259	.278	.309	.370
19	.237	.252	.272	.301	.361
20	.231	.246	.26-4	.294	.352
25	.21	.22	.24	.264	.32
30	.19	.20	.22	.242	.29
35	.18	.19	.21	.23	.27
40				.21	.25
50				.19	.23
60				.17	.21
70				.16	.19
90				.14	
100				.14	
Asymptotic	$1.07/\sqrt{n}$	$1.14/\sqrt{n}$	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.63/\sqrt{n}$

Adapted from Journal American Statistical Association 47:425–441, 1952, Z. W. Birnbaum.

Table 7.5. Frequency factors for extreme value type I distribution

Sample size n	Return period								
	5	10	15	20	25	50	75	100	1000
15	0.967	1.703	2.117	2.410	2.632	3.321	3.721	4.005	6.265
20	0.919	1.625	2.023	2.302	2.517	3.179	3.563	3.836	6.006
25	0.888	1.575	1.963	2.235	2.444	3.088	3.463	3.729	5.842
30	0.866	1.541	1.922	2.188	2.393	3.026	3.393	3.653	5.727
35	0.851	1.516	1.891	2.152	2.354	2.979	3.341	3.598	
40	0.838	1.495	1.866	2.126	2.326	2.943	3.301	3.554	5.576
45	0.829	1.478	1.847	2.104	2.303	2.913	3.268	3.520	
50	0.820	1.466	1.831	2.086	2.283	2.889	3.241	3.491	5.478
55	0.813	1.455	1.818	2.071	2.267	2.869	3.219	3.467	
60	0.807	1.446	1.806	2.059	2.253	2.852	3.200	3.446	
65	0.801	1.437	1.796	2.048	2.241	2.837	3.183	3.429	
70	0.797	1.430	1.788	2.038	2.230	2.824	3.169	3.413	5.359
75	0.792	1.423	1.780	2.029	2.220	2.812	3.155	3.400	
80	0.788	1.417	1.773	2.020	2.212	2.802	3.145	3.387	
85	0.785	1.413	1.767	2.013	2.205	2.793	3.135	3.376	
90	0.782	1.409	1.762	2.007	2.198	2.785	3.125	3.367	
95	0.780	1.405	1.757	2.002	2.193	2.777	3.116	3.357	
100	0.779	1.401	1.752	1.998	2.187	2.770	3.109	3.349	5.261
∞	0.719	1.305	1.635	1.866	2.044	2.592	2.911	3.137	4.936

Table 7.3a. K_T values for positive skew coefficients Pearson type III distribution[1]

Skew coef.	Recurrence interval (years) 1.0101	2	5	10	25	50	100	200
	Percent chance (≥)							
γ	99	50	20	10	4	2	1	0.5
3.0	−0.667	−0.396	0.420	1.180	2.278	3.152	4.051	4.970
2.9	−0.690	−0.390	0.440	1.195	2.277	3.134	4.013	4.904
2.8	−0.714	−0.384	0.460	1.210	2.275	3.114	3.973	4.847
2.7	−0.740	−0.376	0.479	1.224	2.272	3.093	3.932	4.783
2.6	−0.769	−0.368	0.499	1.238	2.267	3.071	3.889	4.718
2.5	−0.799	−0.360	0.518	1.250	2.262	3.048	3.845	4.652
2.4	−0.832	−0.351	0.537	1.262	2.256	3.023	3.800	4.584
2.3	−0.867	−0.341	0.555	1.274	2.248	2.997	3.753	4.515
2.2	−0.905	−0.330	0.574	1.284	2.240	2.970	3.705	4.444
2.1	−0.946	−0.319	0.592	1.294	2.230	2.942	3.656	4.372
2.0	−0.990	−0.307	0.609	1.302	2.219	2.912	3.605	4.298
1.9	−1.037	−0.294	0.627	1.310	2.207	2.881	3.553	4.223
1.8	−1.087	−0.282	0.643	1.318	2.193	2.848	3.499	4.147
1.7	−1.140	−0.268	0.660	1.324	2.179	2.815	3.444	4.069
1.6	−1.197	−0.254	0.675	1.329	2.163	2.780	3.388	3.990
1.5	−1.256	−0.240	0.690	1.333	2.146	2.743	3.330	3.910
1.4	−1.318	−0.225	0.705	1.337	2.128	2.706	3.271	3.828
1.3	−1.383	−0.210	0.719	1.339	2.108	2.666	3.211	3.745
1.2	−1.449	−0.195	0.732	1.340	2.087	2.626	3.149	3.661
1.1	−1.518	−0.180	0.745	1.341	2.066	2.585	3.087	3.575
1.0	−1.588	−0.164	0.758	1.340	2.043	2.542	3.022	3.489
.9	−1.660	−0.148	0.769	1.339	2.018	2.498	2.957	3.401
.8	−1.733	−0.132	0.780	1.336	1.993	2.453	2.891	3.312
.7	−1.806	−0.116	0.790	1.333	1.967	2.407	2.824	3.223
.6	−1.880	−0.099	0.800	1.328	1.939	2.359	2.755	3.132
.5	−1.955	−0.083	0.808	1.323	1.910	2.311	2.686	3.041
.4	−2.029	−0.066	0.816	1.317	1.880	2.261	2.615	2.949
.3	−2.104	−0.050	0.824	1.309	1.849	2.211	2.544	2.856
.2	−2.178	−0.033	0.830	1.301	1.818	2.159	2.472	2.763
.1	−2.252	−0.017	0.836	1.292	1.785	2.107	2.400	2.670
0	−2.326	0	0.842	1.282	1.751	2.054	2.326	2.576

[1] Interagency Advisory Committee on Water Data (1982).

Table 7.3b. K_T values for negative skew coefficients Pearson type III distribution[1]

Skew coef γ	Recurrence interval (years) 1.0101	2	5	10	25	50	100	200
	Percent chance (≥) 99	50	20	10	4	2	1	0.5
.0	−2.326	0	0.842	1.282	1.751	2.054	2.326	2.576
−.1	−2.400	0.017	0.846	1.270	1.716	2.000	2.252	2.482
−.2	−2.472	0.033	0.850	1.258	1.680	1.945	2.178	2.388
−.3	−2.544	0.050	0.853	1.245	1.643	1.890	2.104	2.294
−.4	−2.615	0.066	0.855	1.231	1.606	1.834	2.029	2.201
−.5	−2.686	0.083	0.856	1.216	1.567	1.777	1.955	2.108
−.6	−2.755	0.099	0.857	1.200	1.528	1.720	1.880	2.016
−.7	−2.824	0.116	0.857	1.183	1.488	1.663	1.806	1.926
−.8	−2.891	0.132	0.856	1.166	1.448	1.606	1.733	1.837
.9	−2.957	0.148	0.854	1.147	1.407	1.549	1.660	1.749
−1.0	−3.022	0.164	0.852	1.128	1.366	1.492	1.588	1.664
−1.1	−3.087	0.180	0.848	1.107	1.324	1.435	1.518	1.581
−1.2	−3.149	0.195	0.844	1.086	1.282	1.379	1.449	1.501
−1.3	−3.211	0.210	0.838	1.064	1.240	1.324	1.383	1.424
−1.4	−3.271	0.225	0.832	1.041	1.198	1.270	1.318	1.351
−1.5	−3.330	0.240	0.825	1.018	1.157	1.217	1.256	1.282
−1.6	−3.388	0.254	0.817	0.994	1.116	1.166	1.197	1.216
−1.7	−3.444	0.268	0.808	0.970	1.075	1.116	1.140	1.155
−1.8	−3.499	0.282	0.799	0.945	1.035	1.069	1.087	1.097
−1.9	−3.553	0.294	0.788	0.920	0.996	1.023	1.037	1.044
−2.0	−3.605	0.307	0.777	0.895	0.959	0.980	0.990	0.995
−2.1	−3.656	0.319	0.765	0.869	0.923	0.939	0.946	0.949
−2.2	−3.705	0.330	0.752	0.844	0.888	0.900	0.905	0.907
−2.3	−3.753	0.341	0.739	0.819	0.855	0.864	0.867	0.869
−2.4	−3.800	0.351	0.725	0.795	0.823	0.830	0.832	0.833
−2.5	−3.845	0.360	0.711	0.771	0.793	0.798	0.799	0.800
−2.6	−3.889	0.368	0.696	0.747	0.764	0.768	0.769	0.769
−2.7	−3.932	0.376	0.681	0.724	0.738	0.740	0.740	0.741
−2.8	−3.973	0.384	0.666	0.702	0.712	0.714	0.714	0.714
−2.9	4.013	0.390	0.651	0.681	0.683	0.689	0.690	0.690
−3.0	4.051	0.396	0.636	0.660	0.666	0.666	0.667	0.667

[1] Interagency Advisory Committee on Water Data (1982).

TABLE 11.3.1
Population parameters and sample statistics

Population parameter	Sample statistic
1. *Midpoint*	
Arithmetic mean	
$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Median	
x such that $F(x) = 0.5$	50th-percentile value of data
Geometric mean	
antilog $[E(\log x)]$	$\left(\prod_{i=1}^{n} x_i\right)^{1/n}$
2. *Variability*	
Variance	
$\sigma^2 = E[(x-\mu)^2]$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Standard deviation	
$\sigma = \{E[(x-\mu)^2]\}^{1/2}$	$s = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}$
Coefficient of variation	
$CV = \frac{\sigma}{\mu}$	$CV = \frac{s}{\bar{x}}$
3. *Symmetry*	
Coefficient of skewness	
$\gamma = \frac{E[(x-\mu)^3]}{\sigma^3}$	$C_s = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)(n-2)s^3}$

The *variability* of data is measured by the *variance* σ^2, which is the second moment about the mean:

$$E[(x-\mu)^2] = \sigma^2 = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)\,dx \tag{11.3.3}$$

TABLE 11.5.1
Probability distributions for fitting hydrologic data

Distribution	Probability density function	Range	Equations for parameters in terms of the sample moments
Normal	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$-\infty \leq x \leq \infty$	$\mu = \bar{x},\ \sigma = s_x$
Lognormal	$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right)$ where $y = \log x$	$x > 0$	$\mu_y = \bar{y},\ \sigma_y = s_y$
Exponential	$f(x) = \lambda e^{-\lambda x}$	$x \geq 0$	$\lambda = \frac{1}{\bar{x}}$
Gamma	$f(x) = \frac{\lambda^\beta x^{\beta-1} e^{-\lambda x}}{\Gamma(\beta)}$ where Γ = gamma function	$x \geq 0$	$\lambda = \frac{\bar{x}}{s_x^2}$ $\beta = \frac{\bar{x}^2}{s_x^2} = \frac{1}{CV^2}$

TABLE 11.5.1 (*cont.*)

Probability distributions for fitting hydrologic data

Distribution	Probability density function	Range	Equations for parameters in terms of the sample moments
Pearson Type III (three parameter gamma)	$f(x) = \dfrac{\lambda^{\beta}(x-\epsilon)^{\beta-1}e^{-\lambda(x-\epsilon)}}{\Gamma(\beta)}$	$x \geq \epsilon$	$\lambda = \dfrac{s_x}{\sqrt{\beta}}$, $\beta = \left(\dfrac{2}{C_s}\right)^2$ $\epsilon = \bar{x} - s_x\sqrt{\beta}$
Log Pearson Type III	$f(x) = \dfrac{\lambda^{\beta}(y-\epsilon)^{\beta-1}e^{-\lambda(y-\epsilon)}}{x\,\Gamma(\beta)}$ where $y = \log x$	$\log x \geq \epsilon$	$\lambda = \dfrac{s_y}{\sqrt{\beta}}$, $\beta = \left[\dfrac{2}{C_s(y)}\right]^2$ $\epsilon = \bar{y} - s_y\sqrt{\beta}$ (assuming $C_s(y)$ is positive)
Extreme Value Type I	$f(x) = \dfrac{1}{\alpha}\exp\left[-\dfrac{x-u}{\alpha} - \exp\left(-\dfrac{x-u}{\alpha}\right)\right]$	$-\infty < x < \infty$	$\alpha = \dfrac{\sqrt{6}s_x}{\pi}$ $u = \bar{x} - 0.5772\alpha$

Appendixes

A.1. COMMON DISTRIBUTIONS

Hypergeometric Distribution

$$f_X(x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \qquad X = 0, 1, 2, \ldots, n$$

$$\mu = \frac{nk}{N} \qquad \sigma^2 = \frac{nk(N-k)(N-n)}{N^2(N-1)} \qquad \gamma = \frac{(N-2k)(N-2n)(n-1)^{1/2}}{(N-2)[nk(N-k)(N-n)]^{1/2}}$$

Binomial Distribution

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad X = 0, 1, 2, \ldots, n$$

$$\mu = np \qquad \sigma^2 = np(1-p) \qquad \gamma = \frac{(1-2p)}{\sqrt{np(1-p)}}$$

Geometric Distribution

$$f_X(x) = p(1-p)^{x-1} \qquad X = 1, 2, 3, \ldots$$

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2} \qquad \gamma = \frac{1-2p}{\sqrt{(1-p)}}$$

Negative Binomial Distribution

$$f_X(x) = \binom{x-1}{k-1} p^k (1-p)^{x-k} \qquad X = k, k+1, \ldots$$

$$\mu = \frac{k}{p} \qquad \sigma^2 = \frac{k(1-p)}{p^2}$$

Poisson Distribution

$$f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad X = 0, 1, 2, \ldots$$

$$\mu = \lambda \qquad \sigma^2 = \lambda \qquad \gamma = \frac{1}{\sqrt{\lambda}}$$

Uniform Distribution

$$p_X(x) = \frac{1}{\beta - \alpha} \qquad \alpha \le X \le \beta$$

$$\mu = \frac{\beta + \alpha}{2} \qquad \sigma^2 = \frac{(\beta - \alpha)^2}{12} \qquad \gamma = 0$$

Triangular Distribution

$$p_X(x) = \frac{2}{\beta - \alpha}\left(\frac{x - \alpha}{\delta - \alpha}\right) \qquad \alpha < X < \delta$$

$$= \frac{2}{\beta - \alpha}\left(\frac{\beta - x}{\beta - \delta}\right) \qquad \delta < X < \beta$$

$$\mu = \frac{\alpha + \beta + \delta}{3} \qquad \sigma^2 = \frac{\alpha^2 + \beta^2 + \delta^2 - \alpha\beta - \alpha\delta - \beta\delta}{18}$$

$$\gamma = \frac{18\alpha\beta\delta - 3[\alpha\beta(\alpha + \beta) + \beta\delta(\beta + \delta) + \delta\alpha(\delta + \alpha)] + 108C_V^2\mu^3}{10C_V^3\mu^3}$$

Exponential Distribution

$$p_X(x) = \lambda e^{-\lambda(x-\epsilon)} \qquad X > \epsilon; \quad \beta > 0$$

$$\mu = \epsilon + \frac{1}{\lambda} \qquad \sigma = \frac{1}{\lambda^2} \qquad \gamma = 2$$

Normal Distribution

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2\right) \qquad -\infty \le X \le \infty$$

$\mu, \qquad \sigma^2, \qquad \gamma = 0$

Lognormal Distribution

$$p_X(x) = \frac{1}{x\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{1}{2}\left[\frac{\ln(x)-\mu_y}{\sigma_y}\right]^2\right) \qquad X \ge 0 \quad Y = \ln(X)$$

$$\mu = \exp(\mu_y + \sigma_y^2/2) \qquad \sigma^2 = \mu^2(\exp(\sigma_y^2) - 1) \qquad \gamma = 3c_v + c_v^3$$

Lognormal Distribution (3-parameter)

$$p_X(x) = \frac{1}{(x-\epsilon)\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{1}{2}\left[\frac{\ln(x-\epsilon)-\mu_y}{\sigma_y}\right]^2\right) \qquad X \ge \epsilon \qquad Y = \ln(X-\epsilon)$$

$$\mu = \epsilon + \exp(\mu_y + \sigma_y^2/2) \qquad \sigma^2 = (\mu-\epsilon)^2(\exp(\sigma_y^2) - 1)$$

$$\gamma = 3\kappa + \kappa^3 \qquad \text{where } k = \sqrt{\exp(\sigma_y^2) - 1}$$

Gamma Distribution

$$p_X(x) = \frac{\lambda^\eta x^{\eta-1} e^{-\lambda x}}{\Gamma(\eta)} \qquad X > 0$$

$$\mu = \eta/\lambda \qquad \sigma^2 = \eta/\lambda^2 \qquad \gamma = 2/\sqrt{(\eta)}$$

Extreme Value Type I Distribution for Maximums

$$p_X(x) = \frac{1}{\alpha}\exp(-y - e^{-y}) \qquad \text{for} \quad y = (x-\beta)/\alpha \qquad -\infty < X < \infty \qquad \alpha > 0$$

$$\mu = \beta + 0.577\alpha \qquad \sigma^2 = 1.645\alpha^2 \qquad \gamma = 1.1396$$

Extreme Value Type I Distribution for Minimums

$$p_X(x) = \frac{1}{\alpha}\exp(y - e^{y}) \qquad \text{for} \quad y = (x-\beta)/\alpha \qquad -\infty < X < \infty \qquad \alpha > 0$$

$$\mu = \beta - 0.577\alpha \qquad \sigma^2 = 1.645\alpha^2 \qquad \gamma = -1.1396$$

Extreme Value Type III for Minimums

$$p_X(x) = \frac{\alpha \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right]}{x^{1-\alpha}\beta^{\alpha}} \qquad X \geq 0; \quad \alpha, \beta > 0$$

$$\mu = \beta\Gamma(1 + 1/\alpha) \qquad \sigma^2 = \beta^2[\Gamma(1 + 2/\alpha) - \Gamma^2(1 + 1/\alpha)]$$

$$\gamma = \frac{\Gamma(1 + 3/\alpha) - 3\Gamma(1 + 2/\alpha)\Gamma(1 + 1/\alpha) + 2\Gamma^3(1 + 1/\alpha)}{[\Gamma(1 + 2/\alpha) - \Gamma^2(1 + 1/\alpha)]^{3/2}}$$

Extreme Value Type III for Minimums (3-parameter)

$$p_X(x) = \frac{\alpha \exp\left[-\left(\frac{x - \epsilon}{\beta - \epsilon}\right)^{\alpha}\right]}{(x - \epsilon)^{1-\alpha}(\beta - \epsilon)^{\alpha}} \qquad X \geq \epsilon; \qquad \alpha, \beta > 0$$

$$\mu = \epsilon + (\beta - \epsilon)\Gamma(1 + 1/\alpha) \qquad \sigma^2 = (\beta - \epsilon)^2 [\Gamma(1 + 2/\alpha) - \Gamma^2(1 + 1/\alpha)]$$

$$\gamma = \frac{\Gamma(1 + 3/\alpha) - 3\Gamma(1 + 2/\alpha)\Gamma(1 + 1/\alpha) + 2\Gamma^3(1 + 1/\alpha)}{[\Gamma(1 + 2/\alpha) - \Gamma^2(1 + 1/\alpha)]^{3/2}}$$

HYDROLOGIC DATA

A.2. Monthly runoff (in.), Cave Creek near Fort Spring, Kentucky

Year	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Total
1953	0.02	0.05	0.19	2.40	0.86	4.16	1.47	3.54	0.31	0.18	0.07	0.01	13.26
1954	0.00	0.02	0.04	0.54	0.22	0.04	1.39	0.35	0.08	0.07	0.06	0.14	2.95
1955	0.02	0.04	0.30	0.73	4.63	5.79	0.59	1.97	0.55	0.24	0.28	0.03	15.17
1956	0.04	0.06	0.13	0.59	6.37	4.69	1.92	0.28	0.32	0.64	0.38	0.08	15.50
1957	0.07	0.10	1.72	3.08	3.25	1.03	3.92	0.68	0.24	0.06	0.05	0.02	14.22
1958	0.03	1.06	4.32	2.00	2.21	1.17	2.35	2.36	0.19	3.69	1.70	0.12	21.20
1959	0.06	0.09	0.17	2.70	1.95	1.12	1.02	0.24	0.24	0.05	0.04	0.02	7.70
1960	0.03	0.36	2.69	2.19	3.13	2.91	0.68	0.19	3.64	1.38	0.14	0.30	17.64
1961	0.12	0.52	0.79	2.04	2.95	5.32	4.76	4.14	1.59	0.48	0.18	0.04	22.93
1962	0.02	0.06	0.76	3.46	4.01	5.08	3.30	0.79	0.96	0.30	0.08	0.07	18.89
1963	0.39	1.41	1.24	1.50	1.46	5.48	0.52	0.25	0.14	0.29	0.11	0.03	12.82
1964	0.01	0.04	0.03	0.87	1.73	7.88	0.45	0.21	0.11	0.08	0.02	0.16	11.59
1965	0.15	0.07	3.47	2.76	2.30	4.49	1.46	0.31	0.08	0.05	0.01	0.02	15.17
1966	0.04	0.02	0.02	0.48	2.81	0.79	2.02	3.32	0.25	0.14	0.41	0.11	10.41
1967	0.07	1.19	3.57	0.97	1.61	4.66	0.50	4.76	0.33	0.14	0.15	0.07	18.02
1968	0.09	0.38	2.71	1.35	0.98	4.25	2.38	1.99	0.91	0.29	0.75	0.16	16.24
1969	0.14	0.22	1.12	2.78	2.16	0.73	2.37	0.74	0.40	0.27	0.66	0.17	11.76
1970	0.07	0.25	0.91	1.30	3.89	2.91	5.68	2.06	0.38	0.14	0.06	0.27	17.92
Mean	0.08	0.33	1.34	1.76	2.58	3.47	2.04	1.57	0.60	0.47	0.29	0.10	14.63
St Dev	0.09	0.44	1.40	0.96	1.50	2.22	1.52	1.52	0.85	0.86	0.41	0.09	4.79

Source: U. S. Geological Survey.