Probability, Statistics and Applications

20

Key Topics

Sample Spaces
Random Variables
Mean, Mode and Median
Variance
Normal Distributions
Histograms
Hypothesis Testing
Software Reliability Models
Queueing Theory

20.1 Introduction

Statistics is an empirical science that is concerned with the collection, organization, analysis, interpretation and presentation of data. The data collection needs to be planned and this may include surveys and experiments. Statistics are widely used by government and industrial organizations, and they may be employed for forecasting as well as for presenting trends. They allow the behaviour of a population to be studied and inferences to be made about the population. These inferences may be tested (*hypothesis testing*) to ensure their validity.

The analysis of statistical data allows an organization to understand its performance in key areas, and to identify problematic areas. Organizations will often examine performance trends over time, and will devise appropriate plans and

G. O'Regan, *Guide to Discrete Mathematics*, Texts in Computer Science,
DOI 10.1007/978-3-319-44561-8_20

actions to address problematic areas. The effectiveness of the actions taken will be judged by improvements in performance trends over time.

It is often not possible to study the entire population, and instead a representative subset or sample of the population is chosen. This *random sample* is used to make inferences regarding the entire population, and it is essential that the sample chosen is indeed random and representative of the entire population. Otherwise, the inferences made regarding the entire population will be invalid.

A statistical experiment is a causality study that aims to draw a conclusion regarding values of a *predictor variable*(s) on a *response variable*(s). For example, a statistical experiment in the medical field may be conducted to determine if there is a causal relationship between the use of a particular drug and the treatment of a medical condition such as lowering of cholesterol in the population. A statistical experiment involves the following:

- Planning the research
- Designing the experiment
- Performing the experiment
- Analyzing the results
- Presenting the results

Probability is a way of expressing the likelihood of a particular event occurring. It is normal to distinguish between the frequency interpretation and the subjective interpretation of probability [1]. For example, if a geologist states that 'there is a 70 % chance of finding gas in a certain region' then this statement is usually interpreted in two ways:

- The geologist is of the view that over the long run 70 % of the regions whose environment conditions are very similar to the region under consideration have gas (*Frequency Interpretation*).
- The geologist is of the view that it is likely that the region contains gas, and that 0.7 is a measure of the geologist's belief in this hypothesis (*Personal Interpretation*).

However, the mathematics of probability is the same for both the frequency and personal interpretation.

20.2 Probability Theory

Probability theory provides a mathematical indication of the likelihood of an event occurring, and the probability of an event is a numerical value between 0 and 1. A probability of 0 indicates that the event cannot occur whereas a probability of 1 indicates that the event is guaranteed to occur. If the probability of an event is greater than 0.5 then this indicates that the event is more likely to occur than not to occur.

A *sample space* is the set of all possible outcomes of an experiment, and an *event* E is a subset of the sample space. For example, the sample space for the experiment of tossing a coin is the set of all possible outcomes of this experiment: i.e., head or tails. The event that the toss results a tail is a subset of the sample space.

$$S = \{h, t\} \qquad E = \{t\}$$

Similarly, the sample space for the gender of a newborn baby is the set of outcomes: i.e., the newborn baby is a boy or a girl. The event that the baby is a girl is a subset of the sample space.

$$S = \{b, g\} \qquad E = \{g\}$$

For any two events E and F of a sample space S, we can also consider the union and intersection of these events. That is,

- E ∪ F consists of all outcomes that are in E or F or both.
- E ∩ F (normally written as EF) consists of all outcomes that are in both E and F.
- E^c denotes the complement of E with respect to S and represents the outcomes of S that are not in E.

If EF = ∅ then there are no outcomes in both E and F, and so the two events E and F are mutually exclusive. The union and intersection of two events can be extended to the union and intersection of a family of events E_1, E_2, …, E_n (i.e., $\cup_{i=1}^{n} E_i$ and $\cap_{i=1}^{n} E_i$).

20.2.1 Laws of Probability

The laws of probability essentially state that the probability of an event is between 0 and 1, and that the probability of the union of a mutually disjoint set of events is the sum of their individual probabilities.

i. P(S) = 1
ii. P(∅) = 0
iii. $0 \leq P(E) \leq 1$
iv. For any sequence of mutually exclusive events E_1, E_2, …, E_n. (i.e., $E_i\ E_j$ = ∅ where $i \neq j$) then the probability of the union of these events is the sum of their individual probabilities: i.e.,

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i).$$

The probability of the union of two events (not necessarily disjoint) is given by:

$$P(E \cup F) = P(E) + P(F) - P(EF)$$

The probability of an event E not occurring is denoted by E^c and is given by $1 - P(E)$. The probability of an event E occurring given that an event F has occurred is termed the *conditional probability* (denoted by P(E|F)) and is given by

$$P(E|F) = \frac{P(EF)}{P(F)} \qquad \text{where } P(F) > 0$$

This formula allows us to deduce that

$$P(EF) = P(E|F)P(F)$$

Bayes formula enables the probability of an event E to be determined by a weighted average of the conditional probability of E given that the event F occurred and the conditional probability of E given that F has not occurred:

$$\begin{aligned} E &= E \cap S = E \cap (F \cup F^c) \\ &= EF \cup EF^c \end{aligned}$$

$$\begin{aligned} P(E) &= P(EF) + P(EF^c) \qquad (\text{since } EF \cap EF^c = \varnothing) \\ &= P(E|F)P(F) + P(E|F^c)P(F^c) \\ &= P(E|F)P(F) + P(E|F^c)(1 - P(F)) \end{aligned}$$

Two events E, F are *independent* if the knowledge that F has occurred does not change the probability that E has occurred. That is, P(E|F) = P(E) and since P(E|F) = P(EF)/P(F) we have that two events E, F are independent if:

$$P(EF) = P(E)P(F)$$

Two events E and F that are not independent are said to be *dependent*.

20.2.2 Random Variables

Often, some numerical quantity determined by the result of the experiment is of interest rather than the result of the experiment itself. These numerical quantities are termed *random variables*. A random variable is termed *discrete* if it can take on a finite or countable number of values; otherwise it is termed *continuous*.

The *distribution function* of a random variable is the probability that the random variable X takes on a value less than or equal to x. It is given by

$$F(x) = P\{X \leq x\}$$

All probability questions about X can be answered in terms of its distribution function F. For example, the computation of P $\{a < X < b\}$ is given by

$$\begin{aligned} P\ \{a < X < b\} &= P\{X \leq b\} - P\{X \leq a\} \\ &= F(b) - F(a) \end{aligned}$$

The probability mass function for a discrete random variable X (denoted by $p(a)$) is the probability that it is a certain value. It is given by

$$p(a) = P\{X = a\}$$

Further, F(a) can also be expressed in terms of the probability mass function

$$F(a) = \sum_{\forall x \leq a} p(x)$$

We may also define a probability density function and a probability distribution function X for a continuous random variable X [2], and all probability statements about X can be answered in terms of its density function $f(x)$, and the derivative of the probability distribution function yields the probability density function.

The expected value (i.e., the *mean*) of a discrete random variable X (denoted E[X]) is given by the weighted average of the possible values of X, and the expected value of a function of a random variable is given by E[g(X)]. These are given by

$$\begin{aligned} E[X] &= \sum_{i x_i} P\{X = x_i\} \\ E[g(X)] &= \sum_{i} g(x_i)\ P\{X = x_i\} \end{aligned}$$

The *variance* of a random variable is a measure of the spread of values from the mean, and is defined by

$$\mathrm{Var}(X) = E\left[X^2\right] - (E[X])^2$$

The standard deviation σ is given by the square root of the variance. That is,

$$\sigma = \sqrt{\mathrm{Var}(X)}$$

The *covariance* of two random variables is a measure of the relationship between two random variables X and Y, and indicates the extent to which they both change (in either similar or opposite ways) together. It is defined by

$$\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y].$$

It follows that the covariance of two independent random variables is zero. Variance is a special case of covariance (when the two random variables are identical). This follows since Cov(X, X) = E[X · X] − (E[X])(E[X]) = E[X^2] − (E [X])^2 = Var(X).

A positive covariance (Cov(X, Y) $\geq$ 0) indicates that Y tends to increase as X does, whereas a negative covariance indicates that Y tends to decrease as X increases.

The *correlation* of two random variables is an indication of the relationship between two variables X and Y. If the correlation is negative then Y tends to decrease as X increases, and if it is positive number then Y tends to increase as X increases. The correlation coefficient is a value that is between ±1 and it is defined by

$$\mathrm{Corr(X, Y)} = \frac{\mathrm{Cov(X,Y)}}{\sqrt{\mathrm{Var(X)Var(Y)}}}$$

Once the correlation between two variables has been calculated the probability that the observed correlation was due to chance can be computed. This is to ensure that the observed correlation is a real one and not due to a chance occurrence.

There are a number of special random variables, and these include the Bernoulli trial, where there are just two possible outcomes of an experiment: i.e., success or failure. The probability of success and failure is given by

$$\begin{aligned} \mathrm{P}\{\mathrm{X}\ =\ 0\}\ &= 1 - p \\ \mathrm{P}\{\mathrm{X} = 1\}\ &= p \end{aligned}$$

The mean of the Bernoulli distribution is given by p and the variance by p (1 − p). The *Binomial distribution* involves n Bernoulli trials, each of which results in success or failure. The probability of i successes from n trials is then given by

$$\mathrm{P}\{\mathrm{X} = i\}\ =\ ({}^{n}{}_{i})p^{i}(1-p)^{n-i}$$

with the mean of the Binomial distribution given by np, and the variance is given by np(1 − p).

The *Poisson distribution* may be used as an approximation to the Binomial Distribution when n is large and p is small. The probability of i successes is given by

$$\mathrm{P}\{\mathrm{X} = i\}\ = \mathrm{e}^{-\lambda}\lambda^{i}/i!$$

and the mean and variance of the Poisson distribution is given by λ.

There are many other well-known distributions such as the *hypergeometric distribution* that describes the probability of i successes in n draws from a finite population without replacement; the *uniform distribution*; the *exponential distribution*, the *normal distribution* and the *gamma* distribution. The mean and variance of important probability distributions are summarized in Table 20.1.

The reader is referred to [1] for a more detailed account of probability theory.

Table 20.1 Probability distributions

Distribution name	Density function	Mean/variance
Binomial	$\mathrm{P}\{\mathrm{X}=i\} = \binom{n}{i} p^i (1-p)^{n-i}$	$np,\ np(1-p)$
Poisson	$\mathrm{P}\{\mathrm{X}=i\} = \mathrm{e}^{-\lambda}\lambda^i/i!$	λ, λ
Hypergeometric	$\mathrm{P}\{\mathrm{X}=i\} = \binom{\mathrm{N}}{i}\binom{\mathrm{M}}{n-i}/\binom{\mathrm{N}+\mathrm{M}}{i}$	$n\mathrm{N}/\mathrm{N}+\mathrm{M}, np(1-p)[1-(n-1)/\mathrm{N}+\mathrm{M}-1]$
Uniform	$f(x) = 1/(\beta-\alpha)\alpha \leq x \leq \beta,\ 0$	$(\alpha+\beta)/2,\ (\beta-\alpha)^2/12$
Exponential	$f(x) = \lambda\mathrm{e}^{-\lambda x}$	$1/\lambda,\ 1/\lambda^2$
Normal	$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-(x-\mu)^2/2\sigma^2}$	μ, σ^2
Gamma	$f(x) = \lambda\mathrm{e}^{-\lambda x}(\lambda x)^{\alpha-1}/\Gamma(\alpha)\ (x \geq 0).$	$\alpha/\lambda, \alpha/\lambda^2$

20.3 Statistics

The field of statistics is concerned with summarizing, digesting and extracting information from large quantities of data. Statistics provide a collection of methods for planning an experiment, and analyzing data to draw accurate conclusions from the experiment. We distinguish between descriptive statistics and inferential statistics:

Descriptive Statistics
This is concerned with describing the information in a set of data elements in graphical format, or by describing its distribution.

Inferential Statistics
This is concerned with making inferences with respect to the population by using information gathered in the sample.

20.3.1 Abuse of Statistics

Statistics are extremely useful in drawing conclusions about a population. However, it is essential that the random sample is valid and that the experiment is properly conducted to enable valid conclusions to be inferred. Some examples of the abuse of statistics include

- The sample size may be too small to draw conclusions.
- It may not be a genuine random sample of the population.
- Graphs may be drawn to exaggerate small differences.
- Area may be misused in representing proportions.
- Misleading percentages may be used.

The quantitative data used in statistics may be discrete or continuous. *Discrete data* is numerical data that has a finite number of possible values, and *continuous data* is numerical data that has an infinite number of possible values.

20.3.2 Statistical Sampling

Statistical sampling is concerned with the methodology of choosing a random sample of a population, and the study of the sample with the goal of drawing valid conclusions about the entire population. The assumption is that if a genuine representative sample of the population is chosen, then a detailed study of the sample will provide insight into the whole population. This helps to avoid a lengthy expensive (and potentially infeasible) study of the entire population.

The sample chosen must be random and the sample size must be sufficiently large to enable valid conclusions to be made for the entire population.

Random Sample

A *random sample* is a sample of the population such that each member of the population has an equal chance of being chosen.

There are various ways of generating a random sample from the population including (Table 20.2).

Once the random sample group has been chosen the next step is to obtain the required information from the sample. This may be done by interviewing each member in the sample; calling each member; conducting a mail survey and so on (Table 20.3).

Table 20.2 Sampling techniques

Sampling technique	Description
Systematic sampling	Every *k*th member of the population is sampled
Stratified sampling	The population is divided into two or more strata and each subpopulation (stratum) is then sampled. Each element in the subpopulation shares the same characteristics (e.g., age groups, gender)
Cluster sampling	A population is divided into clusters and a few of these clusters are exhaustively sampled (i.e., every element in the cluster is considered)
Convenience sampling	Sampling is done as convenient and often allows the element to choose whether or not it is sampled

Table 20.3 Types of survey

Survey type	Description
Direct measurement	This may involve a direct measurement of all in the sample (e.g., the height of students in a class)
Mail survey	This involves sending a mail survey to the sample. This may have a lower response rate and may thereby invalidate the findings
Phone survey	This is a reasonably efficient and cost effective way to gather data. However, refusals or hang-ups may affect the outcome
Personal interview	This tends to be expensive and time consuming, but it allows detailed information to be collected
Observational study	An observational study allows individuals to be studied, and the variables of interest to be measured
Experiment	An experiment imposes some treatment on individuals in order to study the response

20.3.3 Averages in a Sample

The term ‘average’ generally refers to the arithmetic *mean* of a sample, but it may also refer to the statistical *mode* or *median* of the sample. These terms are defined below:

Mean
The *arithmetic mean* of a set of n numbers is defined to be the sum of the numbers divided by n. That is, the arithmetic mean for a sample of size n is given by

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

The actual mean of the population is denoted by μ, and it may differ from the sample mean.

Mode
The mode is the data element that occurs most frequently in the sample. It is possible that two elements occur with the same frequency, and if this is the case then we are dealing with a bi-modal or possibly a multi-modal sample.

Median
The median is the middle element when the data set is arranged in increasing order of magnitude.

If there are an odd number of elements in the sample the median is the middle element. Otherwise, the median is the arithmetic mean of the two middle elements.

Mid Range
The midrange is the arithmetic mean of the highest and lowest data elements in the sample. That is, $(x_{\text{max}} + x_{\text{min}})/2$.

The arithmetic mean is the most widely used average in statistics.

20.3.4 Variance and Standard Deviation

An important characteristic of a sample is its distribution, and the spread of each element from some measure of central tendency (e.g., the mean). One elementary measure of dispersion is that of the sample *range*, and it is defined to be the difference between the maximum and minimum value in the sample. That is, the sample range is defined to be

$$\text{range} \quad = x_{\text{max}} - x_{\text{min}}.$$

The sample range is not a reliable measure of dispersion as only two elements in the sample are used, and extreme values in the sample can distort the range to be very large even if most of the elements are quite close to one another.

The standard deviation is the most common way to measure dispersion, and it gives the average distance of each element in the sample from the mean. The sample standard deviation is denoted by s and is defined by

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

The population standard deviation is denoted by σ and is defined by

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

Variance is another measure of dispersion and it is defined as the square of the standard deviation. The sample variance is given by

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

The population variance is given by

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

20.3.5 Bell-Shaped (Normal) Distribution

The German mathematician Gauss (Fig. 20.1) originally studied the normal distribution, and it is also known as the *Gaussian distribution* (Fig. 20.2). It is shaped like a bell and so is popularly known as the *bell-shaped* distribution. The empirical frequencies of many natural populations exhibit a bell-shaped (*norma*l) curve.

The *normal distribution* N has mean μ, and standard deviation σ. Its density function $f(x)$ where (where $-\infty < x < \infty$) is given by

Fig. 20.1 Carl Friedrich Gauss

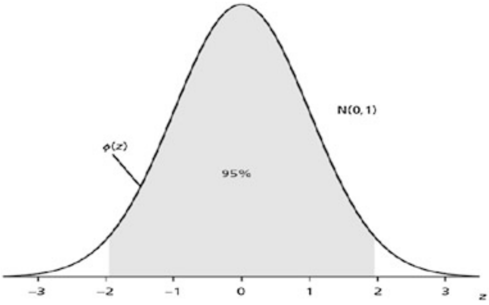


Fig. 20.2 Standard normal bell curve (Gaussian distribution)

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

The *unit* (or *standard*) normal distribution *Z*(0, 1) has mean 0 and standard deviation of 1. Every normal distribution may be converted to the unit normal distribution by $Z = (X - \mu)/\sigma$, and every probability statement about X

$$f(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

has an equivalent probability statement about Z. The unit normal density function is given by

For a normal distribution 68.2 % of the data elements lie within one standard deviation of the mean; 95.4 % of the population lies within two standard deviations of the mean; and 99.7 % of the data lies within three standard deviations of the mean. For example, the shaded area under the curve within two standard deviations of the mean represents 95 % of the population.

A fundamental result in probability theory is the *Central Limit Theorem*, and this theorem essentially states that the sum of a large number of independent and identically distributed random variables has a distribution that is approximately normal. That is, suppose X_1, X_2, …, X_n is a sequence of independent random variables each with mean μ and variance σ^2. Then for large *n* the distribution of

$$\frac{x_1 + x_2 + \cdots + x_n - n\mu}{\sigma\sqrt{n}}$$

is approximately that of a unit normal variable Z. One application of the central limit theorem is in relation to the binomial random variables, where a binomial random variable with parameters (*n*, *p*) represents the number of successes of *n* independent trials, where each trial has a probability of *p* of success. This may be expressed as

$$X = X_1 + X_2 + \cdots + X_n$$

where $X_i = 1$ if the *i*th trial is a success and is 0 otherwise. $E(X_i) = p$ and $Var(X_i) = p(1 - p)$, and then by applying the central limit theorem it follows that for large *n*

$$\frac{\mathrm{X} - np}{\sqrt{np(1-p)}}$$

will be approximately a unit normal variable (which becomes more normal as n becomes larger).

The sum of independent normal random variables is normally distributed, and it can be shown that the sample average of X_1, X_2, …, X_n is normal, with a mean equal to the population mean but with a variance reduced by a factor of $1/n$.

$$\mathrm{E}(\bar{\mathrm{X}}) = \sum_{i=1}^{n} \frac{E(X_i)}{n} = \mu$$

$$\mathrm{Var}(\bar{\mathrm{X}}) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(\mathrm{X}_i) = \frac{\sigma^2}{n}$$

It follows that from this that the following is a unit normal random variable.

$$\sqrt{n}\,\frac{(X-\mu)}{\sigma}$$

The term *six-sigma* (6σ) is a methodology concerned with continuous process improvement and aims for very high quality (close to perfection). A 6σ process is one in which 99.9996 % of the products are expected to be free from defects (3.4 defects per million).

20.3.6 Frequency Tables, Histograms and Pie Charts

A frequency table is used to present or summarize data (Tables 20.4 and 20.5). It lists the data classes (or categories) in one column and the frequency of the category in another column.

A histogram is a way to represent data in bar chart format (Fig. 20.3). The data is divided into intervals where an interval is a certain range of values. The horizontal axis of the histogram contains the intervals (also known as buckets) and the vertical axis shows the frequency (or relative frequency) of each interval. The bars represent the frequency and there is no space between the bars.

Table 20.4 Frequency table—salary

Profession	Salary	Frequency
Project manager	65,000	3
Architect	65,000	1
Programmer	50,000	8
Tester	45,000	2
Director	90,000	1

Table 20.5 Frequency table —test results

Mark	Frequency
0–24	3
25–49	10
50–74	15
75–100	2

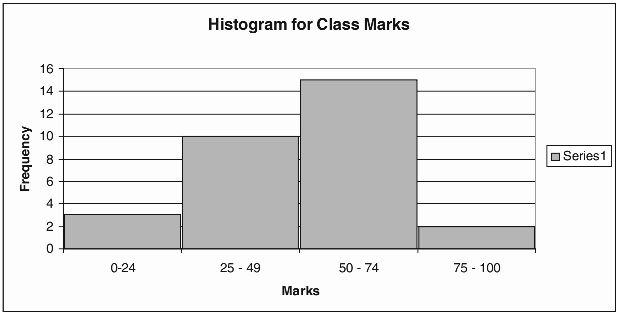
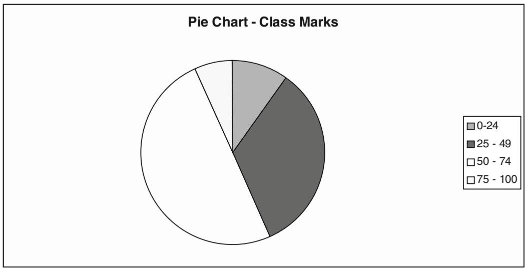


Fig. 20.3 Histogram test results

Fig. 20.4 Pie chart test results



A histogram has an associated shape. For example, it may resemble a normal distribution, a bi-modal or multi-modal distribution. It may be positively or negatively skewed. The construction of a histogram first involves the construction of a frequency table where the data is divided into disjoint classes and the frequency of each class is determined.

A pie chart (Fig. 20.4) offers an alternate way to histograms in the presentation of data. A frequency table is first constructed, and the pie chart presents a visual representation of the percentage in each data class.

20.3.7 Hypothesis Testing

The basic concept of inferential statistics is *hypothesis testing*, where a hypothesis is a statement about a particular population whose truth or falsity is unknown.

Hypothesis testing is concerned with determining whether the values of the random sample from the population are consistent with the hypothesis. There are two mutually exclusive hypotheses: one of these is the null hypothesis H_0 and the other is the alternate research hypothesis H_1. The null hypothesis H_0 is what the researcher is hoping to reject, and the research hypothesis H_1 is what the researcher is hoping to accept.

Statistical testing is then employed to test the hypothesis, and the result of the test is that we either reject the null hypothesis (and therefore accept the alternative hypothesis), or that we fail to reject (i.e., we accept) the null hypothesis. The rejection of the null hypothesis means that the null hypothesis is highly unlikely to be true, and that the research hypothesis should be accepted.

Statistical testing is conducted at a certain level of significance, with the probability of the null hypothesis H_0 being rejected when it is true never greater than α. The value α is called the level of significance of the test, with α usually being 0.1, 0.05, 0.005. A significance level β may also be applied to with respect to accepting the null hypothesis H_0 when H_0 is false, and usually $\alpha = \beta$.

The objective of a statistical test is not to determine whether or not H_0 is actually true, but rather to determine whether its validity is consistent with the observed data. That is, H_0 should only be rejected if the resultant data is very unlikely if H_0 is true.

The errors that can occur with hypothesis testing include type 1 and type 2 errors. Type 1 errors occur when we reject the null hypothesis when the null hypothesis is actually true. Type 2 errors occur when we accept the null hypothesis when the null hypothesis is false (Table 20.6).

For example, an example of a false positive is where the results of a blood test comes back positive to indicate that a person has a particular disease when in fact the person does not have the disease. Similarly, an example of a false negative is where a blood test is negative indicating that a person does not have a particular disease when in fact the person does. Both errors can potentially be very serious.

The terms α and β represent the level of significance that will be accepted, and normally $\alpha = \beta$. In other words, α is the probability that we will reject the null hypothesis when the null hypothesis is true, and β is the probability that we will accept the null hypothesis when the null hypothesis is false.

Testing a hypothesis at the $\alpha = 0.05$ level is equivalent to establishing a 95 % confidence interval. For 99 % confidence α will be 0.01, and for 99.999 % confidence then α will be 0.00001.

Table 20.6 Hypothesis testing

Action	H_0 true, H_1 false	H_0 false, H_1 true
Reject H_1	Correct	False positive—type 2 error $P(\text{accept } H_0 \mid H_0 \text{ false}) = \beta$
Reject H_0	False negative—type 1 error $P(\text{reject } H_0 \mid H_0 \text{ true}) = \alpha$	Correct

The hypothesis may be concerned with testing a specific statement about the value of an unknown parameter θ of the population. This test is to be done at a certain level of significance, and the unknown parameter may, for example, be the mean or variance of the population. An estimator for the unknown parameter is determined, and the hypothesis that this is an accurate estimate is rejected if the random sample is not consistent with it. Otherwise, it is accepted.

The steps involved in hypothesis testing include the following:

1. Establish the null and alternative hypothesis,
2. Establish error levels (significance),
3. Compute the test statistics (often a *t*-test),
4. Decide on whether to accept or reject the null hypothesis.

The difference between the observed and expected test statistic, and whether the difference could be accounted for by normal sampling fluctuations is the key to the acceptance or rejection of the null hypothesis.

20.4 Software Reliability

The design and development of high-quality software has become increasingly important for society. Many software companies desire a sound mechanism to predict the reliability of their software prior to its deployment at the customer site, and this has led to a growing interest in software reliability models.

Definition 12.1 (*Software Reliability*) *Software reliability* is defined as the probability that the program works without failure for a specified length of time, and is a statement of the future behaviour of the software. It is generally expressed in terms of the *mean time to failure* (MTTF) or the *mean time between failure* (MTBF).

Statistical sampling techniques are often employed to predict the reliability of hardware, as it is not feasible to test all items in a production environment. The quality of the sample is then used to make inferences on the quality of the entire population, and this approach is effective in manufacturing environments where variations in the manufacturing process often lead to defects in the physical products.

There are similarities and differences between hardware and software reliability. A hardware failure may arise due to a component wearing out due to its age, and often a replacement is required. Most hardware components are expected to last for a certain period of time, and the variation in the failure rate of a hardware component are often due to the manufacturing process and to the operating environment of the component. Good hardware reliability predictors have been developed, and each hardware component has an expected mean time to failure. The reliability of a product may be determined from the reliability of the individual components of the hardware.

Software is an intellectual undertaking involving a team of designers and programmers. It does not physically wear out and software failures manifest themselves from particular user inputs. Each copy of the software code is identical and the software is either correct or incorrect. That is, software failures are due to design and implementation errors rather than to physically wearing out. The software community has not yet developed a sound software reliability predictor model.

The software population to be sampled consists of all possible execution paths of the software, and since this is potentially infinite it is generally not possible to perform exhaustive testing.

The way in which the software is used (i.e., the inputs entered by the users) will impact upon its perceived reliability. Let I_f represent the fault set of inputs (i.e., $i_f \in I_f$ if and only if the input of i_f by the user leads to failure). The randomness of the time to software failure is due to the unpredictability in the selection of an input $i_f \in I_f$… It may be that the elements in $I_{f.}$ are inputs that are rarely used, and that therefore the software will be perceived as reliable.

Statistical testing may be used to make inferences on the future performance of the software. This requires an understanding of the expected usage profile of the system, as well as the population of all possible usages of the software. The sampling is done in accordance with the expected usage profile.

20.4.1 Software Reliability and Defects

The release of an unreliable software product may result in damage to property or injury (including loss of life) to a third party. Consequently, companies need to be confident that their software products are fit for use prior to their release. The project team needs to conduct extensive inspections and testing of the software prior to its release.

Objective product quality criteria may be set (e.g., 100 % of tests performed and passed) to be satisfied prior to release. This provides a degree of confidence that the software has the desired quality, and is safe and fit for purpose. However, these results are historical in the sense that they are a statement of past and present quality. The question is whether the past behaviour provides a sound indication of future behaviour.

Software reliability models are an attempt to predict the future reliability of the software, and to assist in deciding on whether the software is ready for release.

A defect does not always result in a failure, as it may be benign and may occur on a rarely used execution path. Many observed failures arise from a small proportion of the existing defects. Adam's 1984 case study [3] indicated that over 33 % of the defects led to an observed failure with mean time to failure greater than 5000 years; whereas less than 2 % of defects led to an observed failure with a mean time to failure of less than 50 years. This suggests that a small proportion of defects led to almost all of the observed failures (Table 20.7).

Table 20.7 Adam's 1984 study of software failures of IBM products

	Rare				Frequent			
	1	2	3	4	5	6	7	8
MTTF (years)	5,000	1,580	500	158	50	15.8	5	1.58
Avg. % fixes	33.4	28.2	18.7	10.6	5.2	2.5	1.0	0.4
Prob failure	0.008	0.021	0.044	0.079	0.123	0.187	0.237	0.300

The analysis shows that 61.6 % of all fixes (Group 1. and 2.) were made for failures that will be observed less than once in 1580 years of expected use, and that these constitute only 2.9 % of the failures observed by typical users. On the other hand, groups 7 and 8 constitute 53.7 % of the failures observed by typical users and only 1.4 % of fixes.

This showed that *coverage testing* is not cost effective in increasing MTTF. *Usage testing*, in contrast, would allocate 53.7 % of the test effort to fixes that will occur 53.7 % of the time for a typical user. Harlan Mills has argued [4] that the data in the table shows that usage testing is 21 times more effective than coverage testing.

There is a need to be careful with *reliability growth models*, as there is no tangible growth in reliability unless the corrected defects are likely to manifest themselves as a failure.[1] Many existing software reliability growth models assume that all remaining defects in the software have an equal probability of failure, and that the correction of a defect leads to an increase in software reliability. These assumptions are questionable.

The defect count and defect density may be poor predictors of operational reliability, and an emphasis on removing a large number of defects from the software may not be sufficient in itself to achieve high reliability.

The correction of defects in the software leads to newer versions of the software, and reliability models assume reliability growth: i.e., the new version is more reliable than the older version as several identified defects have been corrected. However, in some sectors such as the safety critical field the view is that the new version of a program is a new entity, and that no inferences may be drawn until further investigation has been done. The relationship between the new version and the previous version of the software needs to be considered (Table 20.8).

The safety critical industry (e.g., the nuclear power industry) takes the conservative viewpoint that any change to a program creates a new program. The new program is therefore required to demonstrate its reliability.

[1] We are assuming that the defect has been corrected perfectly with no new defects introduced by the changes made.

Table 20.8 New and old version of software

Similarities and differences between new/old version
• The new version of the software is identical to the previous version except that the identified defects have been corrected
• The new version of the software is identical to the previous version, except that the identified defects have been corrected, but the developers have introduced some new defects
• No assumptions can be made about the behaviour of the new version of the software until further data is obtained

20.4.2 Cleanroom Methodology

Harlan Mills and others at IBM developed the Cleanroom methodology to assist in the development of high-quality software. The software is released only when the probability of zero-defects is very high.

The way in which the software is used will impact upon its perceived quality and reliability. Failures will manifest themselves on certain input sequences only, and as users will generally employ different input sequences, each user will have a different perception of the reliability of the software. Knowledge of the way that the software will be used allows the software testing to be focused on verifying the correctness of the common everyday tasks carried out by users.

This means that it is important to determine the operational profile of users to allow effective testing of the software to take place. The operational environment may not be stable as users may potentially change their behaviour over time. The collection of operational data involves identifying the operations to be performed and the probability of that operation being performed.

The Cleanroom approach [4] applies statistical techniques to enable a software reliability measure to be calculated, and it is based on the expected usage of the software. It employs *statistical usage testing* rather than coverage testing, and applies statistical quality control to certify the mean time to failure of the software. The statistical usage testing involves executing tests chosen from the population of all possible uses of the software in accordance with the probability of expected use.

Coverage testing involves designing tests that cover every path through the program, and this type of testing is as likely to find a rare execution failure as well as a frequent execution failure. It is highly desirable to find failures that occur on frequently used parts of the system.

The advantage of usage testing (that matches the actual execution profile of the software) is that it has a better chance of finding execution failures on frequently used parts of the system. This helps to maximize the expected mean time to failure.

20.4.3 Software Reliability Models

Models are simplifications of the reality and a good model allows accurate predictions of future behaviour to be made. The adequacy of the model is judged by

Table 20.9 Characteristics of good software reliability model

Characteristics of good software reliability model
Good theoretical foundation
Realistic assumptions
Good empirical support
As simple as possible (Ockham's razor)
Trustworthy and accurate

model exploration, and determining if its predictions are close to the actual manifested behaviour. More accurate models are sought to replace inadequate models.

A model is judged effective if there is good empirical evidence to support it. Models are often modified (or replaced) over time, as further facts and observations lead to aberrations that cannot be explained by the current model. A good software reliability model will have the following characteristics (Table 20.9):

There are several software reliability predictor models employed (with varying degrees of success). Some of them just compute defect counts rather than estimating software reliability in terms of mean time to failure. They include (Table 20.10):

- *Size and Complexity Metrics*

These are used to predict the number of defects that a system will reveal in operation or testing.

- *Operational Usage Profile*

These predict failure rates based on the expected operational usage profile of the system. The number of failures encountered is determined and the software reliability predicted.

- *Quality of the Development Process*

These predict failure rates based on the process maturity of the software development process in the organization.

The extent to which the software reliability model can be trusted depends on the accuracy of its predictions. Empirical data will need to be gathered to determine the accuracy of the predictions. It may be acceptable to have a little inaccuracy during the early stages of prediction, provided the predictions of operational reliability are close to the observations. A model that gives overly optimistic results is termed 'optimistic,' whereas a model that gives overly pessimistic results is termed 'pessimistic.'

Table 20.10 Software reliability models

Model	Description	Comments
Jelinski/moranda model	The failure rate is a Poisson process and is proportional to the current defect content of program. The initial defect count is N; the initial failure rate is Nφ; it decreases to (N − 1)φ after the first fault is detected and eliminated, and so on. The constant φ is termed the proportionality constant	Assumes defects corrected perfectly and no new defects are introduced Assumes each fault contributes the same amount to failure rate
Littlewood/verrall model	Successive execution time between failures independent exponentially distributed random variables. Software failures are the result of the particular inputs and faults introduced from the correction of defects	Does not assume perfect correction of defects
Seeding and Tagging	This is analogous to estimating the fish population of a lake (Mills). A known number of defects is inserted into a software program and the proportion of these identified during testing determined Another approach (Hyman) is to regard the defects found by one tester as tagged and then to determine the proportion of tagged defects found by a second independent tester	Estimate of the total number of defects in the software but not a not s/w reliability predictor Assumes all faults equally likely to be found and introduced faults representative of existing
Generalized Poisson Model	The number of failures observed in *i*th time interval τ_i has a Poisson distribution with mean $\phi(N - M_{i-1}) \tau_i^{\alpha}$ where N is the initial number of faults; M_{i-1} is the total number of faults removed up to the end of the $(i - 1)$th time interval; and ϕ is the proportionality constant	Assumes faults removed perfectly at end of time interval

The assumptions in the reliability model need to be examined to determine whether they are realistic. Several software reliability models have questionable assumptions such as

- All defects are corrected perfectly
- Defects are independent of one another
- Failure rate decreases as defects are corrected.
- Each fault contributes the same amount to the failure rate

20.5 Queuing Theory

The term '*queue*' refers to waiting in line for a service, such as waiting in line at a bakery or a bank, and *queuing theory* is the mathematical study of waiting lines or queues. The origins of queuing theory are in work done by Erlang at the Copenhagen Telephone Exchange in the early twentieth century where he modelled the number of telephone calls arriving as a Poisson process.

Queuing theory has been applied to many fields including telecommunications and traffic management. This section aims to give a flavour and a very short introduction to queuing theory, and it has been adapted from [5]. The interested reader may consult the many other texts available for more detailed information [e.g., 6].

A supermarket may be used to illustrate the ideas of queuing theory, as it has a large population of customers some of whom may enter the supermarket and queuing system (the checkout queues). Customers will generally wait for a period of time in a queue before receiving service at the checkout, and they wait for a further period of time for the actual service to be carried out. Each service facility (the checkouts) contains identical servers, and each server is capable of providing the desired service to the customer (Fig. 20.5).

Clearly, if there are no waiting lines then immediate service is obtained. However, in general, there are significant costs associated with the provision of an immediate service, and so there is a need to balance cost with a certain amount of waiting.

Some queues are *bounded* (i.e., they can hold only a fixed number of customers), whereas others are *unbounded* and can grow as large as is required to hold all waiting customers. The customer source may be finite or infinite, and where the customer source is finite but very large it is often considered to be infinite.

Random variables (described by probability distribution functions) arise in queuing problems, and these include the random variable *q,* which represents the time that a customer spends in the queue waiting for service; the random variable *s*, which represents the amount of time that a customer spends in service; and the

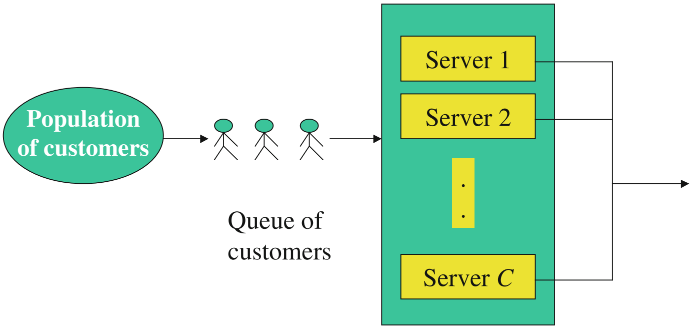


Fig. 20.5 Basic queuing system

random variable w, which represents the total time that a customer spends in the queuing system. Clearly,

$$w = q + s$$

It is assumed that the customers arrive at a queuing system one at a time at random times ($t_0 < t_1 < \cdots < t_n$) with the random variable $\tau_k = t_k - t_{k-1}$ representing the *interarrival times* (i.e., it measures the times between successive arrivals). It is assumed that these random variables are independent and identically distributed, and it is usually assumed that arrivals form a Poisson arrival process (Fig. 20.6).

A Poisson arrival process is characterized by the fact that the interarrival times are distributed exponentially. That is,

$$P(\tau \le t) \;=\; 1 - \mathrm{e}^{-\lambda t}$$

Further, the probability that exactly n customers will arrive in any time interval of length t is given by

$$\frac{\mathrm{e}^{-\lambda t}(\lambda t)^n}{n!} \qquad (\text{where}\, n = 0, 1, 2, \ldots)$$

where λ is a constant average arrival rate of customers per unit time, and the number of arrivals per unit time is Poisson distributed with mean λ.

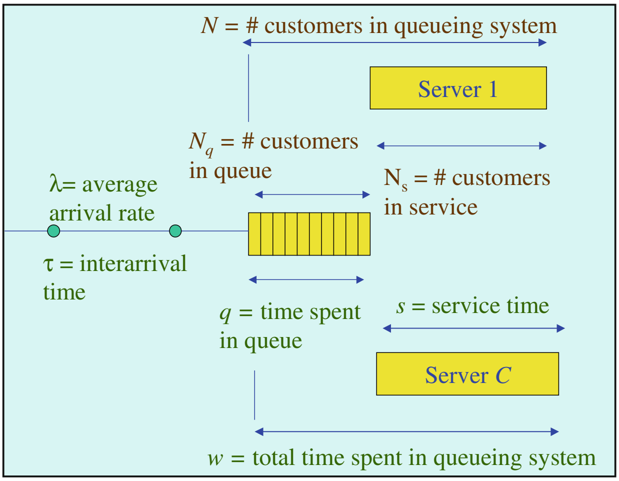

N = # customers in queueing system
Server 1
N_q = # customers in queue
λ= average arrival rate
N_s = # customers in service
τ = interarrival time
q = time spent in queue
s = service time
Server C
w = total time spent in queueing system

Fig. 20.6 Sample random variables in queuing theory

Similarly, it is usual to assume in queuing theory that the service times are random with μ denoting the average service rate, and let s_k denote the service time that the *k*th customer requires from the system. The distribution of service times is given by

$$W_s(t) = P(s \leq t) = 1 - \mathrm{e}^{-\mu t}$$

The capacity of the queues may be *infinite* (where every arriving customer is allowed to enter the queuing system no matter how many waiting customers are present), or finite (where arriving customers may wait only if there is still room in the queue).

Queuing systems may be *single server* (one server serving one customer at a time) systems or *multiple servers* (several identical servers that can service *c* customers at a time). The method by which the next customer is chosen from the queue to be serviced is termed the *queue discipline*, and the most common method is *first-come-first-served* (FCFS). Other methods include the last-in-first-out (LIFO); the shortest job first; or the highest priority job next.

Customers may exhibit various behaviours in a queuing system such as deciding not to join a queue if it is too long; switching between queues to try to obtain faster service; or leaving the queuing system if they have waited too long. There are many texts on queuing theory and for a more detailed account on queuing theory see [6].

20.6 Review Questions

1. What is probability? What is statistics? Explain the difference between them.

2. Explain the laws of probability.
3. What is a sample space? What is an event?
4. Prove Boole's inequality $\mathrm{P}\left(\cup_{i=1}^{n} \mathrm{E}_i\right) \leq \sum_{i=1}^{n} \mathrm{P}(\mathrm{E}_i)$ where the E_i are not necessarily disjoint.
5. A couple has 2 children. What is the probability that both are girls if the eldest is a girl?
6. What is a random variable?
7. Explain the difference between the probability density function and the probability distribution function
8. Explain expectation, variance, covariance and correlation.
9. Describe how statistics may be abused.

10. What is a random sample? Describe methods available to generate a random sample from a population. How may information be gained from a sample?
11. Explain how the average of a sample may be determined, and discuss the mean, mode and median of a sample.
12. Explain sample variance and sample standard deviation.
13. Describe the normal distribution and the central limit theorem.
14. Explain hypothesis testing and acceptance or rejection of the null hypothesis.
15. What is software reliability? Describe various software reliability models.
16. Explain queuing theory and describe its applications to the computing field.

20.7 Summary

Statistics is an empirical science that is concerned with the collection, organization, analysis and interpretation and presentation of data. The data collection needs to be planned and this may include surveys and experiments. Statistics are widely used by government and industrial organizations, and they may be used for forecasting as well as for presenting trends. Statistical sampling allows the behaviour of a random sample to be studied, and inferences to be made about the population.

Probability theory provides a mathematical indication of the likelihood of an event occurring, and the probability is a numerical value between 0 and 1. A probability of 0 indicates that the event cannot occur, whereas a probability of 1 indicates that the event is guaranteed to occur. If the probability of an event is greater than 0.5, then this indicates that the event is more likely to occur than not to occur.

Software has become increasingly important for society and professional software companies aspire to develop high-quality and reliable software. Software reliability is the probability that the program works without failure for a specified length of time, and is a statement on the future behaviour of the software. It is generally expressed in terms of the mean time to failure (MTTF) or the mean time between failure (MTBF), and the software reliability measurements are an attempt to provide an objective judgment of the fitness for use of the software.

There are many reliability models in the literature and the question as to which is the best model or how to evaluate the effectiveness of the model arises. A good model will have good theoretical foundations and will give useful predictions of the reliability of the software.

Queuing theory is the mathematical study of waiting lines or queues, and its origins are in work done Erlang in the early twentieth century. Customers will generally wait for a period of time in a queue before receiving service at, and they wait for a further period of time for the actual service to be carried out. Each service facility (the checkouts) contains identical servers, and each server is capable of providing the desired service to the customer. Queuing theory has been applied to many fields including telecommunications and traffic management.

References

1. Introduction to Probability and Statistics for Engineers and Scientists. Sheldon M. Ross. Wiley Publications. New York. 1987.
2. Mathematics in Computing. Second Edition, Gerard O'Regan. Springer. 2012.
3. Optimizing preventive service of software products. E. Adams. IBM Research Journal, 28(1), pp. 2–14, 1984.
4. Engineering Software under Statistical Quality Control. Richard H. Cobb and Harlan D. Mills. IEEE Software. 1990.
5. Operating Systems. H.M. Deitel. 2nd Edition. Addison Wesley.1990.
6. Fundamentals of Queueing Theory. 4th Edition. Donald Gross and John Shortle. Wiley Interpress. 2008.