Basics of Multivariate Modelling and Data Analysis

Kurt-Erik Häggblom

4. Statistical concepts
4.1 Measurement, population and sample
4.2 Stochastic variables
4.3 Probability distributions
4.4 Sample statistics
4.5 Statistical tests
4. Statistical concepts

4.1 Measurement, population and sample

4.1.1 Measurement errors

If the measurement of a continuous (process) variable is repeated a number of times at constant conditions, the measured values will not be identical. If we assume that the variable has a “true” value, most measurements will be erroneous to some degree.

Measurement errors can be classified as follows:

- **Systematic error**: An overall deviation from the true value in some systematic way (too high or too low), also called bias. This type of error *can not be detected* by repeated measurements with the same instrument when the true value is unknown.

- **Gross error**: An occasional, very large, deviation from the true value, also called outlier. This type of error *can be detected* by repeated measurements. If detected, the measurement is usually not used.

- **Random error**: A “normal” fluctuation about the true (or biased) value. The error is a stochastic variable. When measurements are repeated, *stochastic analysis* can be used to infer properties of the true variable.
4. Statistical concepts

In practice, process conditions may not remain constant when measurements are repeated. This introduces further variations in the measured variable. Often this is the major source for random variations.

4.1.2 Precision and accuracy

A set of measurements made at constant conditions is

- **accurate**, if the measurements (or their mean value) are close to the true value of the variable — this indicates a small bias
- **precise**, if all individual measurements are close to the mean value of the measurements — indicates small random errors

As the figure shows

- good accuracy does not imply good precision
- good precision does not imply good accuracy

[ From Miller (2005) ]

![Diagram showing accuracy and precision](image-url)
4.1.3 Population and sample

A population is the
- entire collection of objects of some kind
- all possible values the measurement of a variable can give

A sample is a subset of the population, i.e. a number of
- objects (chosen randomly) from the population
- measurements of a variable

Thus, a sample (usually) contains several objects or measurements.

Although this definition of sample is the “official” one in statistics, a sample is sometimes taken to mean *one* object, on which measurements are made.

To avoid this confusion, and also to make a distinction between objects and measurements, we can use the terms
- set of objects
- set of measurements

instead of sample.
4. Statistical concepts

4.2 Stochastic variables

Random errors are stochastic variables and because measurements contain random errors, they are stochastic variables, too.

A stochastic variable is

- **discrete**, if it can have only a **countable** number of distinct values, e.g. integer values (1, 2, 3, …)
- **continuous**, if it can have any **value** (with infinite precision) in a certain interval of values

Measurements of continuous variables are usually taken to be continuous stochastic variables even though the measurement has a finite precision.

A stochastic variable has the property that

- a future value **cannot be predicted** with certainty
- a future value, or an interval of values in the case of a continuous variable, occurs with some probability

The properties of a stochastic variable are **completely defined by the probability** by which any possible value, or interval of values for a cont. var., occurs.
4.2.1 Probability distribution

**Discrete distribution**

The *Probability Mass Function* (pmf) of a discrete random variable $X$ is the function $f(k)$ defined by

$$f(k) = P(X = k) \geq 0, \quad \sum_{k=-\infty}^{\infty} P(X = k) = 1,$$

where $P(X = k)$ is the probability that $X$ assumes the value $k$. In practice, $k$ is often restricted to some discrete set $K, k \in K$. This means that $P(i \notin K) = 0$.

The *Cumulative Distribution Function* (cfd) of $X$ is defined by

$$F(k) = P(X \leq k) = \sum_{i=-\infty}^{k} P(X = i).$$

A pmf $f(k)$ and its cdf $F(k)$ of a discrete random variable $X$ [Ogunnaike, 2009].
4.2 Stochastic variables

4.2.1 Probability distribution

**Continuous distribution**

The *Probability Density Function* (pdf) of a continuous random variable $X$ is a function $f(x)$ defined by

$$f(x) \geq 0, \quad \int_{-\infty}^{\infty} f(x) \, dx = 1.$$  

Sometimes, $f(x)$ is called *probability function*, *probability distribution*, or *density function*.

The *Cumulative Distribution Function* (cdf) of $X$ is given by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t) \, dt.$$  

A pdf $f(x)$ and its cdf $F(x)$ of a continuous random variable $X$ [Spiegel et al., 2009].
4.2.2 Characteristic parameters

**Expected value**

The expected value, also called *(mathematical) expectation* or the *mean*, of a random variable $X$ is defined

$$
\mu = E(X) = \sum_{k \in K} kf(k) \quad \text{for discrete } X
$$

$$
= \int_{-\infty}^{\infty} xf(x)dx \quad \text{for continuous } X
$$

- $E(|X - \mu|)$ is called *mean deviation*

**Variance**

The variance of a random variable $X$ is defined

$$
\mu_2 = \sigma^2 = \text{Var}(X) = E[(X - \mu)^2] = \sum_{k \in K} (k - \mu)^2 f(k), \text{ discrete } X
$$

$$
= \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx, \text{ continuous } X
$$

- $\sigma$ is called *standard deviation*
- $\sigma / \mu$ is called *coefficient of variation*

The variance is a measure of the *dispersion* or the *scatter* of the values about the mean.
4.2 Stochastic variables

4.2.2 Characteristic parameters

**Skewness**

The skewness of a random variable $X$ is defined

$$
\mu_3 = E[(X - \mu)^3] = \sum_{k \in K} (k - \mu)^3 f(k), \text{ discrete } X
$$

$$
= \int_{-\infty}^{\infty} (x - \mu)^3 f(x) dx, \text{ continuous } X
$$

- $\gamma_3 = \mu_3 / \sigma^3$ is called the *coefficient of skewness*

The skewness is a *measure of asymmetry*. It provides information about the relative difference between negative and positive deviations from the mean.

**Negatively skewed** (skewed left).  **Positively skewed** (skewed right).

[Ogunnaike, 2009]
4.2 Stochastic variables

4.2.2 Characteristic parameters

**Kurtosis**

The skewness of a random variable $X$ is defined

$$
\mu_4 = E[(X - \mu)^4] = \sum_{k \in K} (k - \mu)^4 f(k), \text{ discrete } X
$$
$$
= \int_{-\infty}^{\infty} (x - \mu)^4 f(x) \, dx, \text{ continuous } X
$$

- $\gamma_4 = \mu_4 / \sigma^4$ is called the coefficient of kurtosis

The kurtosis is a *measure of how flat or peaked* a probability distribution is; $\gamma_4 = 3$ is *normal* kurtosis, $\gamma_4 < 3$ is *mild* kurtosis, $\gamma_4 > 3$ is *high* kurtosis.

---

*Mild kurtosis* (dashed line).

[Ogunnaike, 2009]

*High kurtosis* (dashed line).
4. Statistical concepts

4.3 Important probability distributions

4.3.1 The binomial distribution

Suppose that we have an experiment such as tossing a coin or die repeatedly. Each toss is called a trial. In any single trial there will be a probability associated with a particular event such as head on the coin or 4 on the die. This probability will not change from one trial to the next. Such trials are then said to be independent trials.

Let $p$ be the probability that an event will happen in any single trial (called the probability of success). Then $q = 1 - p$ is the probability that the event will fail to happen in any single trial (called the probability of failure). The probability that the event will happen exactly $x$ times in $n$ trials is given by the probability function

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

where $x = 0, 1, \ldots, n$ is the number of successes in $n$ trials. This probability distribution of a discrete variable is called the binomial distribution.
Some of the properties of the binomial distribution are listed in the table.

<table>
<thead>
<tr>
<th>Property</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\mu = np$</td>
</tr>
<tr>
<td>Variance</td>
<td>$\sigma^2 = npq$</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>$\sigma = \sqrt{npq}$</td>
</tr>
<tr>
<td>Coefficient of skewness</td>
<td>$\alpha_3 = \frac{q - p}{\sqrt{npq}}$</td>
</tr>
<tr>
<td>Coefficient of kurtosis</td>
<td>$\alpha_4 = 3 + \frac{1 - 6pq}{npq}$</td>
</tr>
</tbody>
</table>
### 4.3.2 The normal distribution

One of the most important examples of a continuous probability distribution is the normal distribution, sometimes called the Gaussian distribution. The density function for this distribution is given by

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

where \( \mu \) and \( \sigma \) are the mean and standard deviation, respectively.

If we introduce the standardized stochastic variable \( Z \) defined by

\[ Z = \frac{X - \mu}{\sigma} \]

the standard normal density function takes the form

\[ f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \]

Note that \( E(Z) = 0 \) and \( \text{Var}(Z) = 0 \).
Some of the properties of the normal distribution are summarized in the table.

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\mu$</td>
</tr>
<tr>
<td>Variance</td>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>Coefficient of skewness</td>
<td>$\alpha_3 = 0$</td>
</tr>
<tr>
<td>Coefficient of kurtosis</td>
<td>$\alpha_4 = 3$</td>
</tr>
</tbody>
</table>

**Relation between binomial and normal distributions**

If $n$ is “large” ($n \geq 30$) and if neither $p$ nor $q$ is too close to zero, the binomial distribution can be closely approximated by a normal distribution with standardized random variable given by

$$Z = \frac{X - np}{\sqrt{npq}}$$
4.3.3 Hypergeometric distribution
4.3.4 Poisson distribution
4.3.5 Geometric distribution
4.3.6 Negative binomial distribution
4.3.7 Chi-square distribution
4.3.8 $F$ distribution
4.3.9 Student’s $t$ distribution
4.3.10 Logistic distribution
4.2.11 Lognormal distribution
4.2.12 Weibull distribution

…
4. Statistical concepts

4.4 Sample statistics

As we have seen, a population or a stochastic variable can be characterized by “bulk” parameters such as

- mean
- variance
- standard deviation

A sample of a population or a stochastic variable can be characterized by similar parameters, called

- sample mean
- sample variance
- sample standard deviation

Since it is important to make a distinction between sample parameters and the theoretically true parameters of a population or a stochastic variable, we will use different mathematical symbols for them.

<table>
<thead>
<tr>
<th></th>
<th>Population</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\mu$</td>
<td>$\overline{x}$</td>
</tr>
<tr>
<td>Variance</td>
<td>$\sigma^2$</td>
<td>$s^2$</td>
</tr>
<tr>
<td>St.dev.</td>
<td>$\sigma$</td>
<td>$s$</td>
</tr>
</tbody>
</table>
4. Statistical concepts

4.4 Sample statistics

Statistical sample parameters

Sample mean:
\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

Sample variance:
\[ s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

Sample covariance:
\[ s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \]

Sample correlation:
\[ r_{xy} = \frac{s_{xy}}{s_x s_y} \], where \( s_x = \sqrt{s_x^2} \), \( s_y = \sqrt{s_y^2} \)
4. Statistical concepts

4.5 Statistical tests

There are a large number of statistical test that can be performed, e.g.

- Test for (normal) distribution
- Test for mean values ($t$-test)
- Test for comparison of variances ($F$-test)

We are interested in the properties of the whole population or the true stochastic variable.

Based on tests like these, one can say if the population or the stochastic variable has a certain property with some given probability, e.g. 95%