Inferential Statistics

Statistical inference is the branch of statistics concerned with drawing conclusions and/or making decisions concerning a population based only on sample data.

Statistical inference: example 1

A clothing store chain regularly buys from a supplier large quantities of a certain piece of clothing. Each item can be classified either as good quality or top quality. The agreements require that the delivered goods comply with standards predetermined quality. In particular, the proportion of good quality items must not exceed 25% of the total.

From a consignment 40 items are extracted and 29 of these are of top quality whereas the remaining 11 are of good quality.

INFERENTIAL PROBLEMS:
1. provide an estimate of \( \pi \) and quantify the uncertainty associated with such estimate;
2. provide an interval of “reasonable” values for \( \pi \);
3. decide whether the delivered goods should be returned to the supplier.

Statistical inference: example 2

A machine in an industrial plant of a bottling company fills one-liter bottles. When the machine is operating normally the quantity of liquid inserted in a bottle has mean \( \mu = 1 \) liter and standard deviation \( \sigma = 0.01 \) liters.

Every working day 10 bottles are checked and, today, the average amount of liquid in the bottles is \( \bar{x} = 1.0065 \) with \( s = 0.0095 \).

INFERENTIAL PROBLEMS:
1. provide an estimate of \( \mu \) and quantify the uncertainty associated with such estimate;
2. provide an interval of “reasonable” values for \( \mu \);
3. decide whether the machine should be stopped and revised.
Statistical inference: example 2

Formalization of the problem:

- **POPULATION:** “all” the bottles filled by the machine;
- **VARIABLE OF INTEREST:** amount of liquid in the bottles (continuous variable);
- **PARAMETERS OF INTEREST:** mean $\mu$ and standard deviation $\sigma$ of the amount of liquid in the bottles;
- **SAMPLE:** 10 bottles.

The values of the parameters $\mu$ and $\sigma$ are unknown, but they affect the sampling values. Sampling evidence provides information on the parameter values.

The sample

- Census survey: attempt to gather information from each and every unit of the population of interest;
- sample survey: gathers information from only a subset of the units of the population of interest.

Why using a sample?

1. Less time consuming than a census;
2. less costly to administer than a census;
3. measuring the variable of interest may involve the destruction of the population unit;
4. a population may be infinite.

Probability sampling

A probability sampling scheme is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined.

**SIMPLE RANDOM SAMPLING:** every unit has an equal probability of being selected and the selection of a unit does not change the probability of selecting any other unit. For instance:

- extraction with replacement;
- extraction without replacement.

For large populations (compared to the sample size) the difference between these two sampling techniques is negligible. In the following we will always assume that samples are extracted with replacement from the population of interest.

Probabilistic description of a population

- Units of the population;
- variable $X$ measured on the population units;
- sometimes the distribution of $X$ is known, for instance

(i) $X \sim N(\mu, \sigma^2)$;
(ii) $X \sim \text{Bernoulli}(\pi)$. 
Probabilistic description of a sample

- The observed sampling values are 
  \[ x_1, x_2, \ldots, x_n; \]

- BEFORE the sample is observed the sampling values are unknown and the sample can be written as a sequence of random variables 
  \[ X_1, X_2, \ldots, X_n \]

- for simple random samples (with replacement):
  1. \( X_1, X_2, \ldots, X_n \) are i.i.d.;
  2. the distribution of \( X_i \) is the same as that of \( X \) for every \( i = 1, \ldots, n \).

Sampling distribution of a statistic (1)

- Suppose that the sample is used to compute a given statistic, for instance 
  (i) the sample mean \( \bar{X} \); 
  (ii) the sample variance \( S^2 \); 
  (iii) the proportion \( P \) of units with a given feature;

- generically, we consider an arbitrary statistic 
  \[ T = g(X_1, \ldots, X_n) \]
  where \( g(\cdot) \) is a given function.

Sampling distribution of a statistic (2)

- Once the sample is observed, the observed value of the statistic is given by 
  \[ t = g(x_1, \ldots, x_n); \]

- suppose that we draw all possible samples of size \( n \) from the given population and that we compute the statistic \( T \) for each sample;

- the sampling distribution of \( T \) is the distribution of the population of the values \( t \) of all possible samples.

Normal population

- Suppose that \( X \sim N(\mu, \sigma^2) \)

- in this case the statistics of interest are:
  (i) the sample mean \( \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \)
  (ii) the sample variance \( S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \)

- the corresponding observed values are 
  \[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \] and 
  \[ s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \]
  respectively.
The sample mean

The sample mean is a linear combination of the variables forming the sample and this property can be exploited in the computation of

• the expected value of $\bar{X}$, that is $E(\bar{X})$;
• the variance of $\bar{X}$, that is $\text{Var}(\bar{X})$;
• the probability distribution of $\bar{X}$.

Expected value of the sample mean

For a simple random sample $X_1, \ldots, X_n$, the expected value of $\bar{X}$ is

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n} E(X_1 + X_2 + \cdots + X_n) = \frac{1}{n} [E(X_1) + E(X_2) + \cdots + E(X_n)] = \frac{1}{n}(n \times \mu) = \mu$$

Variance of the sample mean

For a simple random sample $X_1, \ldots, X_n$, the variance of $\bar{X}$ is

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n^2} \text{Var}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n^2} [\text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n)] = \frac{1}{n^2}(n \times \sigma^2) = \frac{\sigma^2}{n}$$

Sampling distribution of the mean

For a simple random sample $X_1, X_2, \ldots, X_n$, the sample mean $\bar{X}$ has

• expected value $\mu$ and variance $\sigma^2/n$;
• if the distribution of $X$ is normal, then

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

• more generally, the central limit theorem can be applied to state that the distribution of $\bar{X}$ is APPROXIMATIVELY normal.
The sample variance

- The sample variance is defined as
  \[ S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \]
- If \( X_i \sim N(\mu, \sigma^2) \) then
  \[ \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \]
- \( \bar{X} \) and \( S^2 \) are independent.

The chi-squared distribution

- Let \( Z_1, \ldots, Z_r \) be i.i.d. random variables with distribution \( N(0; 1) \);
- The random variable
  \[ X = Z_1^2 + \cdots + Z_r^2 \]
  is said to follow a CHI-SQUARED distribution with \( r \) degrees of freedom (d.f.);
- We write \( X \sim \chi^2_r \);
- \( E(X) = r \) and \( \text{Var}(X) = 2r \).

Counting problems

- The variable \( X \) is binary, i.e. it takes only two possible values; for instance "success" and "failure";
- The random variable \( X \) takes values "1" (success) and "0" (failure);
- The parameter of interest is \( \pi \), the proportion of units in the population with value "1";
- Formally, \( X \sim \text{Bernoulli}(\pi) \) so that
  \[ E(X) = \pi \quad \text{and} \quad \text{Var}(X) = \pi(1 - \pi). \]
The sample proportion (2)

- The sample proportion is such that
  \[ E(P) = \pi \quad \text{and} \quad \text{Var}(P) = \frac{\pi(1-\pi)}{n} \]

- for the central limit theorem, the distribution of \( \bar{X} \) is approximately normal;
- sometimes the following empirical rules are used to decide if the normal approximation is satisfying:
  1. \( n \times \pi > 5 \) and \( n \times (1 - \pi) > 5 \).
  2. \( np(1-p) > 9 \).

Estimation

- Parameters are specific numerical characteristics of a population, for instance:
  - a proportion \( \pi \);
  - a mean \( \mu \);
  - a variance \( \sigma^2 \).

- When the value of a parameter is unknown it can be estimated on the basis of a random sample.

Point estimation

A point estimate is an estimate that consists of a single value or point, for instance one can estimate

- a mean \( \mu \) with the sample mean \( \bar{x} \);
- a proportion \( \pi \) with a sample proportion \( p \);

A point estimate is always provided with its standard error that is a measure of the uncertainty associated with the estimation process.

Estimator vs estimate

- An estimator of a population parameter is
  - a random variable that depends on sample information,
  - whose value provides an approximation to this unknown parameter.

- A specific value of that random variable is called an estimate.
Estimation and uncertainty

- Parameter \( \theta \);
- the sampling statistics \( T = g(X_1, \ldots, X_n) \) on which estimation is based is called the estimator of \( \theta \) and we write \( \hat{\theta} = T \);
- the observed value of the estimator, \( t \), is called an estimate of \( \theta \) and we write \( \hat{\theta} = t \);
- it is fundamental to assess the uncertainty of \( \hat{\theta} \);
- a measure of uncertainty is the standard deviation of the estimator, that is \( SD(T) = SD(\hat{\theta}) \). This quantity is called the STANDARD ERROR of \( \hat{\theta} \) and denoted by \( SE(\hat{\theta}) \).

Point estimation of a mean (\( \sigma^2 \) known)

- Consider the case where \( X_1, \ldots, X_n \) is a simple random sample from \( X \sim N(\mu, \sigma^2) \);
- Parameters:
  - \( \mu \), unknown;
  - assume that the value of \( \sigma^2 \) is known.
- the sample mean can be used as estimator of \( \mu \): \( \hat{\mu} = \bar{X} \);
- the distribution of the estimator is normal with
  \[ E(\hat{\mu}) = \mu \quad \text{and} \quad Var(\hat{\mu}) = \frac{\sigma^2}{n} \]
- STANDARD ERROR \( \hat{\mu} \)
  \[ SE(\hat{\mu}) = \frac{\sigma}{\sqrt{n}} \]

Point estimation of a mean with \( \sigma^2 \) known: example

In the “bottling company” example, assume that the quantity of liquid in the bottles is normally distributed. Then a point estimate of \( \mu \) is

- \( \hat{\mu} = 1.0065 \)
- and the standard error of this estimate is

\[ SE(\hat{\mu}) = \frac{\sigma}{\sqrt{n}} = \frac{0.01}{\sqrt{10}} = 0.0032 \]

Point estimation of a mean (\( \sigma^2 \) unknown)

- Typically the value of \( \sigma^2 \) is not known;
- in this case we estimate it as \( \hat{\sigma}^2 = s^2 \);
- this can be used, for instance, to estimate the standard error of \( \hat{\mu} \)
  \[ SE(\hat{\mu}) = \frac{\hat{\sigma}}{\sqrt{n}} \]
- In the “bottling company” example, if \( \sigma \) is unknown it can be estimated as
  \[ SE(\hat{\mu}) = \frac{0.0095}{\sqrt{10}} = 0.0030 \]
Point estimation for the mean of a non-normal population

- $X_1, \ldots, X_n$ i.i.d. with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$;
- the distribution of $X_i$ is not normal;
- for the central limit theorem the distribution of $\bar{X}$ is approximatively normal.

Point estimation of a proportion

- Parameter: $\pi$;
- the sample proportion $P$ is used as an estimator of $\pi$

\[
\hat{\pi} = P
\]

- this estimator is approximately normally distributed with

\[
E(\hat{\pi}) = \pi \quad \text{and} \quad \text{Var}(\hat{\pi}) = \frac{\pi(1-\pi)}{n}
\]

- the STANDARD ERROR of the estimator is

\[
\text{SE}(\hat{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}}
\]

and in this case the value of standard error is never known.

Estimation of a proportion: example

For the “clothing store chain” example the estimate of the proportion $\pi$ of good quality items is

- $\hat{\pi} = \frac{11}{40} = 0.275$

and an ESTIMATE of the standard error is

\[
\hat{SE}(\hat{\pi}) = \sqrt{\frac{0.275(1-0.275)}{40}} = 0.07
\]

Properties of estimators: unbiasedness

A point estimator $\hat{\theta}$ is said to be an unbiased estimator of the parameter $\theta$ if the expected value, or mean, of the sampling distribution of $\hat{\theta}$ is $\theta$, formally if

\[
E(\hat{\theta}) = \theta
\]

Interpretation of unbiasedness: if the sampling process was repeated, independently, an infinite number of times, obtaining in this way an infinite number of estimates of $\theta$, the arithmetic mean of such estimates would be equal to $\theta$. However, unbiasedness does not guarantees that the estimate based on one single sample coincides with the value of $\theta$. 
Point estimator of the variance

The sample variance $S^2$ is an unbiased estimator of the variance $\sigma^2$ of a normally distributed random variable

$$E(S^2) = \sigma^2.$$ 

On the other hand $\tilde{S}^2$ is a biased estimator of $\sigma^2$

$$E(\tilde{S}^2) = \frac{(n-1)}{n} \sigma^2.$$ 

Bias of an estimator

Let $\tilde{\theta}$ be an estimator of $\theta$. The bias of $\tilde{\theta}$, $Bias(\tilde{\theta})$, is defined as the difference between the expected value of $\tilde{\theta}$ and $\theta$

$$Bias(\tilde{\theta}) = E(\tilde{\theta}) - \theta$$

The bias of an unbiased estimator is 0.

Properties of estimators: Mean Squared Error (MSE)

For an estimator $\tilde{\theta}$ of $\theta$ the (unknown) estimation "error" is given by

$$|\theta - \tilde{\theta}|$$

The Mean Squared Error (MSE) is the expected value of the square of the "error"

$$MSE(\tilde{\theta}) = E[(\theta - \tilde{\theta})^2]$$

$$= Var(\tilde{\theta}) + [\theta - E(\tilde{\theta})]^2$$

$$= Var(\tilde{\theta}) + Bias(\tilde{\theta})^2$$

Hence, for an unbiased estimator, the MSE is equal to the variance.

Most Efficient Estimator

- Let $\tilde{\theta}_1$ and $\tilde{\theta}_2$ be two estimator of $\theta$, then the MSE can be use to compare the two estimators;
- if both $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are unbiased then $\tilde{\theta}_1$ is said to be more efficient than $\tilde{\theta}_2$ if
  $$Var(\tilde{\theta}_1) < Var(\tilde{\theta}_2)$$
- note that if $\tilde{\theta}_1$ is more efficient than $\tilde{\theta}_2$ then also $MSE(\tilde{\theta}_1) < MSE(\tilde{\theta}_2)$ and $SE(\tilde{\theta}_1) < SE(\tilde{\theta}_2)$;
- the most efficient estimator or the minimum variance unbiased estimator of $\theta$ is the unbiased estimator with the smallest variance.
Interval estimation

- A point estimate consists of a single value, so that
  - if \( \bar{X} \) is a point estimator of \( \mu \) then it holds that
    \[ P(\bar{X} = \mu) = 0 \]
  - more generally, \( P(\hat{\theta} = \theta) = 0 \).

- Interval estimation is the use of sample data to calculate an interval of possible (or probable) values of an unknown population parameter.

Confidence interval for the mean of a normal population (\( \sigma \) known)

- \( X_1, \ldots, X_n \) simple random sample with \( X_i \sim N(\mu, \sigma^2) \);
- assume \( \sigma \) known;
- a point estimator of \( \mu \) is \( \hat{\mu} = \bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \);
- the standard error of the estimator is \( SE(\hat{\mu}) = \frac{\sigma}{\sqrt{n}} \).

Before the sample is extracted...

- The sample distribution of the estimator is completely known but for the value of \( \mu \);
- the uncertainty associated with the estimate depends on the size of the standard error. For instance, the probability that \( \hat{\mu} = \bar{X} \) takes a value in the interval \( \mu \pm 1.96 \times SE \) is 0.95 (that is 95%).

Confidence interval for \( \mu \)

- The probability that \( \bar{X} \) belongs to the interval \( (\mu - 1.96 \times SE, \mu + 1.96 \times SE) \) is 95%;
- this can be also stated as: the probability that the interval \( (\bar{X} - 1.96 \times SE, \bar{X} + 1.96 \times SE) \) contains the parameter \( \mu \) is 95%.
Formal derivation of the 95% confidence interval for $\mu$ ($\sigma$ known)

It holds that

$$\frac{\bar{X} - \mu}{SE} \sim N(0, 1) \quad \text{where} \quad SE = \frac{\sigma}{\sqrt{n}}$$

so that

$$0.95 = P \left( -1.96 \leq \frac{\bar{X} - \mu}{SE} \leq 1.96 \right)$$

$$= P \left( -1.96 \leq \bar{X} - \mu \leq 1.96 \cdot SE \right)$$

$$= P \left( -\bar{X} - 1.96 \cdot SE \leq -\mu \leq -\bar{X} + 1.96 \cdot SE \right)$$

$$= P \left( \bar{X} - 1.96 \cdot SE \leq \mu \leq \bar{X} + 1.96 \cdot SE \right)$$

Confidence interval for $\mu$ with $\sigma$ known: example

In the bottling company example, if one assumes $\sigma = 0.01$ known, a 95% confidence interval for $\mu$ is

$$\left( 1.0065 - 1.96 \frac{0.01}{\sqrt{10}}, 1.0065 + 1.96 \frac{0.01}{\sqrt{10}} \right)$$

that is

$$\left( 1.0065 - 0.0062, 1.0065 + 0.0062 \right)$$

so that

$$\left( 1.0003, 1.0126 \right)$$

After the sample is extracted...

On the basis of the sample values the observed value of $\bar{\mu} = \bar{x}$ is computed. $\bar{x}$ may belong to the interval $\mu \pm 1.96 \times SE$ or not. For instance

A different sample...

A different sample may lead to a sample mean $\bar{x}$ that, as in the example below, does not belong to the interval $\mu \pm 1.96 \times SE$ and, as a consequence, also the interval $(\bar{x} - 1.96 \times SE; \bar{x} + 1.96 \times SE)$ will not contain $\mu$.

The interval $(\bar{x} - 1.96 \times SE; \bar{x} + 1.96 \times SE)$ will contain $\mu$ for the 95% of all possible samples.
Interpretation of confidence intervals

Probability is associated with the procedure that leads to the derivation of a confidence interval, not with the interval itself. A specific interval either will contain or will not contain the true parameter, and no probability involved in a specific interval.

Confidence intervals for five different samples of size $n = 25$, extracted from a normal population with $\mu = 368$ and $\sigma = 15$.

A wider confidence interval for $\mu$

- Since it also holds that

$$P(\mu - 2.58 \, SE \leq \bar{X} \leq \mu + 2.58 \, SE) = 0.99$$

- then the probability that $(\bar{X} - 2.58 \, SE, \bar{X} + 2.58 \, SE)$ contains $\mu$ is 99%.

Confidence interval: definition

A confidence interval for a parameter is an interval constructed using a procedure that will contain the parameter a specified proportion of the times, typically 95% of the times.

A confidence interval estimate is made up of two quantities:

- **interval**: set of scores that represent the estimate for the parameter;
- **confidence level**: percentage of the intervals that will include the unknown population parameter.

Confidence level

The confidence level is the percentage associated with the interval. A larger value of the confidence level will typically lead to an increase of the interval width. The most commonly used confidence levels are

- 68% associated with the interval $\bar{X} \pm 1 \, SE$;
- 95% associated with the interval $\bar{X} \pm 1.96 \, SE$;
- 99% associated with the interval $\bar{X} \pm 2.58 \, SE$.

Where the values 1, 1.96 and 2.58 are derived from the standard normal distribution tables.
Notation: standard normal distribution tables

• \( Z \sim N(0, 1) \);
• \( \alpha \) value between zero and one;
• \( z_\alpha \) value such that the area under the \( Z \) pdf between \( z_\alpha \) and \( +\infty \) is equal to \( \alpha \);
• formally

\[
P(Z > z_\alpha) = \alpha \quad \text{and} \quad P(Z < z_\alpha) = 1 - \alpha
\]

• furthermore

\[
P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha
\]

Confidence interval for \( \mu \) with \( \sigma \) known: formal derivation (1)

It holds that

\[
\frac{\bar{X} - \mu}{SE} \sim N(0, 1) \quad \text{where} \quad SE = \frac{\sigma}{\sqrt{n}}
\]

so that

\[
P\left(\mu - z_{\alpha/2} SE \leq \bar{X} \leq \mu + z_{\alpha/2} SE\right) = 1 - \alpha
\]

or, equivalently,

\[
P\left(-z_{\alpha/2} SE \leq \bar{X} - \mu \leq z_{\alpha/2} SE\right) = 1 - \alpha
\]

Confidence interval at the level \( 1 - \alpha \) for \( \mu \) with \( \sigma \) known

A confidence interval at the (confidence) level \( 1 - \alpha \), or \( (1 - \alpha)\% \), for \( \mu \) is given by

\[
(\bar{X} - z_{\alpha/2} SE; \quad \bar{X} + z_{\alpha/2} SE)
\]

Since \( SE = \frac{\sigma}{\sqrt{n}} \) then

\[
(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}; \quad \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}})
\]
Margin of error

- The confidence interval
  \[ \bar{x} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \]

- can also be written as \( \bar{x} \pm ME \) where
  \[ ME = z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \]
  is called the margin of error.

- the interval width is equal to twice the margin of error.

Confidence interval for \( \mu \) with \( \sigma \) unknown

- \( \bar{X} \sim N \left( \mu, \frac{\sigma^2}{\sqrt{n}} \right) \);

- \( \frac{\bar{X} - \mu}{SE} \sim N(0, 1) \);

- in this case the standard error is unknown and needs to be estimated.
  \[ SE(\bar{\mu}) = \frac{\sigma}{\sqrt{n}} \text{ is estimated by } \bar{SE}(\bar{\mu}) = \frac{\bar{\sigma}}{\sqrt{n}} \text{ where } \bar{\sigma} = S \]

- and it holds that
  \[ \frac{\bar{X} - \mu}{SE} \sim t_{n-1} \]

Reducing the margin of error

\[ ME = z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \]

The margin of error can be reduced, without changing the accuracy of the estimate, by increasing the sample size (\( n \uparrow \)).

The Student's \( t \) distribution (1)

- For \( Z \sim N(0; 1) \) and \( X \sim \chi^2_r \), independent;

- the random variable
  \[ T = \frac{Z}{\sqrt{X/r}} \]
  is said to follow a Student's \( t \) distribution with \( r \) degrees of freedom;

- the pdf of the \( t \) distribution differs from that of the standard normal distribution because it has "heavier tails".
The Student’s t distribution (2)

- For \( r \to +\infty \) the Student’s t distribution converges to the standard normal distribution.

For the bottling company example, if the value of \( \sigma \) is not known, then

\[
s = 0.0095 \quad \text{and} \quad t_{9;0.025} = 2.2622
\]

and a 95% confidence interval for \( \mu \) is

\[
\left( \bar{X} - t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}; \bar{X} + t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} \right)
\]

that is

\[
(1.0065 - 0.0068; \quad 1.0065 + 0.0068)
\]

so that

\[
(0.9997; \quad 1.0133)
\]

Confidence interval for \( \mu \) with \( \sigma \) unknown

\[
1 - \alpha = P \left( -t_{n-1,\alpha/2} \leq \frac{\bar{X} - \mu}{SE} \leq t_{n-1,\alpha/2} \right)
\]

\[
= P \left( -t_{n-1,\alpha/2} S\bar{E} \leq \bar{X} - \mu \leq t_{n-1,\alpha/2} S\bar{E} \right)
\]

\[
= P \left( -\bar{X} - t_{n-1,\alpha/2} S\bar{E} \leq -\mu \leq -\bar{X} + t_{n-1,\alpha/2} S\bar{E} \right)
\]

\[
= P \left( \bar{X} - t_{n-1,\alpha/2} S\bar{E} \leq \mu \leq \bar{X} + t_{n-1,\alpha/2} S\bar{E} \right)
\]

where \( t_{n-1,\alpha/2} \) is the value such that the area under the \( t \) pdf, with \( n-1 \) d.f. between \( t_{n-1,\alpha/2} \) and \( +\infty \) is equal to \( \alpha/2 \).

Hence, a confidence interval at the level \((1 - \alpha)\) for \( \mu \) is

\[
\left( \bar{X} - t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}; \bar{X} + t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} \right)
\]

Confidence interval for the mean of a non-normal population

- \( X_1, \ldots, X_n \) i.i.d. with \( E(X_i) = \mu \) and \( \text{Var}(X_i) = \sigma^2 \);

- the distribution of \( X_i \) is not normal;

- for the central limit theorem the distribution of \( \bar{X} \) is approximatively normal;

- if one uses the procedures described above to construct a confidence interval for \( \mu \) the nominal confidence level of the interval is only an approximation of the true confidence level.
Confidence interval for $\pi$

For the central limit theorem

$$\frac{P - \pi}{SE} \approx N(0, 1) \quad \text{where} \quad SE = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

so that

$$1 - \alpha \approx P\left( -z_{\alpha/2} \leq \frac{P - \pi}{SE} \leq z_{\alpha/2} \right)$$

$$= P\left( -z_{\alpha/2} SE \leq P - \pi \leq z_{\alpha/2} SE \right)$$

$$= P\left( -P - z_{\alpha/2} SE \leq -\pi \leq -P + z_{\alpha/2} SE \right)$$

$$= P\left( P - z_{\alpha/2} SE \leq \pi \leq P + z_{\alpha/2} SE \right)$$

Since $\pi$ is always unknown, it is always necessary to estimate the standard error.

Confidence interval for $\pi$: example

For the clothing store chain example, a 95% confidence interval for $\pi$ is

$$\left( \frac{11}{40} - 1.96 \times \hat{SE}(\hat{\pi}) ; \frac{11}{40} + 1.96 \times \hat{SE}(\hat{\pi}) \right)$$

so that

$$\hat{SE}(\hat{\pi}) = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}$$

is estimated by

$$\bar{SE}(\bar{\pi}) = \sqrt{\frac{\bar{\pi}(1 - \bar{\pi})}{n}}$$

where $\bar{\pi} = \bar{x} = \frac{11}{40}$

and one obtains

$$\left( \frac{11}{40} - 1.96 \times 0.07 ; \frac{11}{40} + 1.96 \times 0.07 \right)$$

so that

$$\left( 0.137 ; 0.413 \right)$$

Example of decision problem

Problem: in the example of the bottling company, the quality control department has to decide whether to stop the production in order to revise the machine.

Hypothesis: the expected (mean) quantity of liquid in the bottles is equal to one liter.

The standard deviation is assumed known and equal to $\sigma = 0.01$.

The decision is based on a simple random sample of $n = 10$ bottles.

Statistical hypotheses

A decisional problem in expressed by means of two statistical hypotheses:

- the null hypothesis $H_0$
- the alternative hypothesis $H_1$

the two hypotheses concern the value of an unknown population parameter, for instance $\mu$,

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$
Distribution of $\bar{X}$ under $H_0$

If $H_0$ is true (that is under $H_0$) the distribution of the sample mean $\bar{X}$
- has expected value equal to $\mu_0 = 1$;
- has standard error equal to $SE = \sigma/\sqrt{10} = 0.00316$.
- if $X_1, \ldots, X_{10}$ is a normally distributed i.i.d. sample than also $\bar{X}$ follows a normal distribution, otherwise the distribution of $\bar{X}$ is only approximatively normal (by the central limit theorem).

Observed value of the sample mean

The observed value of the sample mean is $\bar{x}$.

$\bar{x}$ is almost surely different form $\mu_0 = 1$.

under $H_0$, the expected value of $\bar{X}$ is equal to $\mu_0$ and the difference between $\mu_0$ and $\bar{x}$ is uniquely due to the sampling error.

HENCE THE SAMPLING ERROR IS

$$\bar{x} - \mu_0$$

that is

observed value “minus” expected value

Decision rule

The space of all possible sample means is partitioned into a
- rejection region also said critical region;
- nonrejection region.

Outcomes and probabilities

There are two possible “states of the world” and two possible decisions. This leads to four possible outcomes.

<table>
<thead>
<tr>
<th>$H_0$ TRUE</th>
<th>$H_0$ FALSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ IS REJECTED</td>
<td>Type I error $\alpha$</td>
</tr>
<tr>
<td>$H_0$ IS NOT REJECTED</td>
<td>OK</td>
</tr>
</tbody>
</table>

The probability of the type I error is said **significance level** of the test and can be arbitrarily fixed (typically 5%).
**Test statistic**

A test statistic is a function of the sample, that can be used to perform a hypothesis test.

- for the example considered, $\bar{X}$ is a valid test statistics, which is equivalent to the, more common, "$z$" test statistic

$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$

**2-tail test, equal or not equal**

- $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$

**Hypothesis testing: example**

- 5% significance level (arbitrarily fixed);

- $Z = \frac{\bar{X} - 1}{0.00316} \sim N(0, 1)$

- the observed value of $Z$ is $z = \frac{1.0065 - 1}{0.00316} = 2.055$;

- the empirical evidence leads to the rejection of $H_0$.

**p-value approach to testing**

- the $p$-value, also called observed level of significance is the probability of obtaining a value of the test statistic more extreme than the observed sample value, under $H_0$.

- decision rule: compare the $p$-value with $\alpha$:
  - $p$-value $< \alpha \implies$ reject $H_0$
  - $p$-value $\geq \alpha \implies$ nonreject $H_0$

- for the example considered
  
  $p$-value $= P(Z \leq -2.055) + P(Z \geq 2.055) = 0.04$

- $p$-value $< 5\% = $ statistically significant result.

- $p$-value $< 1\% = $ highly significant result.

**$z$ test for $\mu$ with $\sigma$ known**

- $X_1, \ldots, X_n$ i.i.d. with distribution $N(\mu, \sigma^2)$;

- the value of $\sigma$ is known.

- Hypotheses:
  
  $H_0 : \mu = \mu_0$

  $H_1 : \ldots$

- test statistic:

  $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

- under $H_0$ the test statistic $Z$ has distribution $N(0; 1)$
**z test: two-sided hypothesis**

\[ H_1 : \mu \neq \mu_0 \]

In this case

\[ p - \text{value} = P(Z > |z|) \]

**z test: one-sided hypothesis (right)**

\[ H_1 : \mu > \mu_0 \]

In this case

\[ p - \text{value} = P(Z > z) \]

**z test: one-sided hypothesis (left)**

\[ H_1 : \mu < \mu_0 \]

In this case

\[ p - \text{value} = P(Z < z) \]

**t test for \( \mu \) with \( \sigma \) unknown**

- Hypotheses:

\[
\begin{cases}
H_0 : \mu = \mu_0 \\
H_1 : \mu \neq \mu_0
\end{cases}
\]

- Test statistic:

\[ t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \]

- \( p \)-value: \( P(T_{n-1} > |t|) \) where \( T_{n-1} \) follows a Student's \( t \) distribution with \( n - 1 \) degrees of freedom.
Test for a proportion

- Null hypotheses: $H_0 : \pi = \pi_0$.

- Under $H_0$ the sampling distribution of $P$ is approximately normal with expected value $E(P) = \pi_0$ and standard error

\[ SE(P) = \sqrt{\frac{\pi_0(1-\pi_0)}{n}} \]

Note that under $H_0$ there are no unknown parameters.

---

$z$ test for $\pi$

- Hypotheses:

\[ \begin{cases} H_0 : \pi = \pi_0 \\ H_1 : \pi \neq \pi_0 \end{cases} \]

- Test statistic:

\[ Z = \frac{P - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \]

- P-value: $P(Z > |z|)$

---

$z$ test for $\pi$: example

For the clothing store chain example, the hypotheses are

\[ \begin{cases} H_0 : \pi = 0.25 \\ H_1 : \pi > 0.25 \end{cases} \]

Hence, under $H_0$ the standard error is

\[ SE = \sqrt{\frac{0.25(1-0.25)}{40}} = 0.068 \]

so that

\[ z = \frac{0.275 - 0.25}{0.068} = 0.37 \]

and the p-value is $P(Z \geq 0.37) = 0.36$ and the null hypothesis cannot be rejected.