STATISTICS

People sometimes use statistics to describe the results of an experiment or an investigation. This process is referred to as data analysis or **descriptive statistics**.

Statistics is also used in another way: if the entire population of interest is not accessible for some reason, often only a portion of the population (a sample) is observed and statistics is then used to answer questions about the whole population. This process is called **inferential statistics**.

Statistical inference will be the main focus of this lesson.

Example 1

We choose ten people from a population and for each individual measure his or her height. We obtain the following results in cm:

178, 180, 158, 166, 190, 180, 177, 178, 182, 160

It is known that the random variable describing the height of an individual from that particular population has a normal distribution $N(\mu,100)$

---

1. *What is an unbiased estimate of the average height μ of the population?*

2. *Considering that the arithmetic mean of the sample arithmetic mean is 174.9, can we maintain that, with a probability of μ , the hypothesis that the average height of the population is 176 cm holds ?*

The answer to the first question is related to point estimation First we will introduce the notion of a random sample.

We view the raw data $x_1, x_2, \ldots, x_n$ as an implementation of a random vector $X_1, X_2, \ldots, X_n$ with the random variables being independent and each random variable having the same probability distribution $F(x)$. This random vector is called a **random sample** or a **sample**.

In a similar way, we also define a **multidimensional random sample**, for example $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.

A function $Y = S(X_1, X_2, \ldots, X_n)$ that does not explicitly depend on any parameter of the distribution $F(x)$ is called a **statistic** of the sample $X_1, X_2, \ldots, X_n$

Various statistics are then used as **point estimators** of parameters of the distribution $F(x)$. The point estimation is obtained by implementing the point estimator statistic, that is, by substituting the raw data for the random variables in the formula.

The following are examples of the most frequently used statistics:

Sample arithmetic mean

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

Sample variance

$$S'^2 = \frac{\sum_{i}^{n} \left(X_i - \overline{X} \right)^2}{n-1}$$

Sample standard deviation

$$S' = \sqrt{\frac{\sum_{i}^{n} \left(X_i - \overline{X}\right)^2}{n-1}}$$

Sample correlation coefficient

$$R = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right)\left(\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2\right)}}$$

It is desirable for a point estimate to be:

(1) **Consistent**. The larger the sample size, the more accurate the estimate.

(2) **Unbiased**. For each value ϑ of the parameter of the population distribution $F(x)$ the sample expectance $E(Y)$ of the point estimator Y equals ϑ.

(3) Most efficient or **best unbiased**: of all consistent, unbiased estimates, the one possessing the smallest variance.

It can be shown that the following statistics are consistent, unbiased, and best unbiased estimators.

- sample arithmetic mean for $E(X)$
- sample variance for $D(X)$
- sample standard deviation for $\sqrt{D(X)}$
- sample correlation coefficient for $\rho(X,Y)$

Returning to Example 1, we can now say that, based on the data given, 174.9 is a consistent, best unbiased estimation of the average height of the entire population.

Example 2

We choose ten people from a population and for each individual measure his or her height. We obtain the following results in cm:

178, 180, 158, 166, 190, 180, 177, 178, 182, 160

It is known that the random variable describing the height of an individual from that particular population has a normal distribution $N(\mu, 100)$

For a given probability $1-\alpha$, is there an interval I such that the average height μ of the population lies in I with probability $1-\alpha$?

Let us take the sample arithmetic mean

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

If each X_i has a normal distribution $N\left(\mu, \sigma^2\right)$, it can be proved that $\overline{X}$ has distribution $N\left(\mu, \frac{\sigma^2}{n}\right)$

We will show this for $n = 2$:

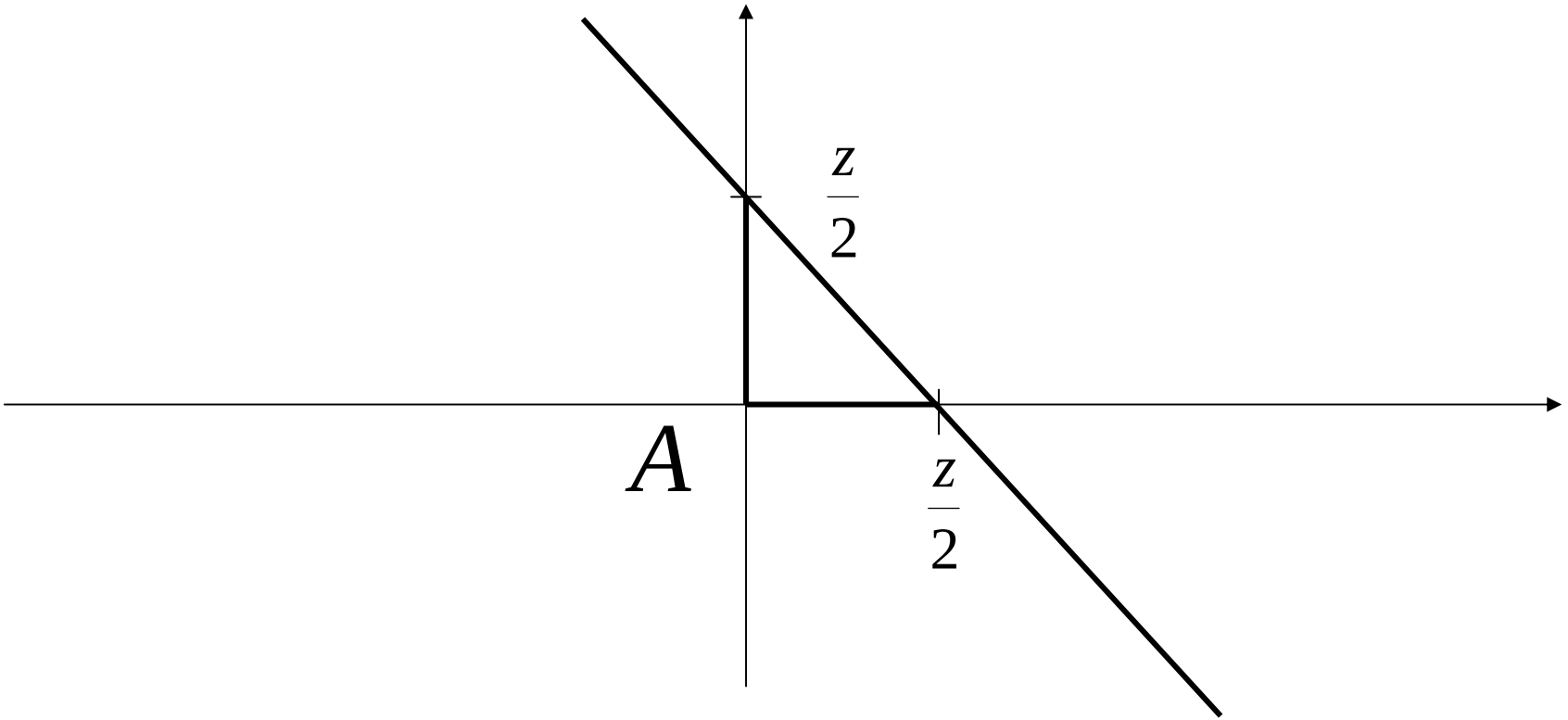
Let X_0 has a distribution $N\left(0,\sigma_x^2\right)$ and $Y_0 \;\; N\left(0,\sigma_y^2\right)$

Put $Z_0 = \dfrac{X_0 + Y_0}{2}$ denote by $F(z)$ the distribution of Z_0 and calculate : $F(z) = P(Z_0 < z) = P(X_0 + Y_0 < z)$

Since X_0 and Y_0 are independent, we can write

$$P\left(\frac{X_0 + Y_0}{2} < z\right) = \iint\limits_{x+y<\frac{z}{2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}} dxdy$$

$$\iint\limits_{x+y<\frac{z}{2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}} dxdy = \iint\limits_{A} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}} dxdy$$



To calculate $$I = \iint_A \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}} dxdy$$

we will use the transformation $x = t,\ y = s - t$ where $J = 1$.

$$I = \frac{1}{2\pi\sigma^2} \int_{-\infty}^{z/2} ds \int_{-\infty}^{\infty} e^{-\frac{t^2+(s-t)^2}{2\sigma^2}} dt = \frac{1}{2\pi\sigma^2} \int_{-\infty}^{z/2} ds \int_{-\infty}^{\infty} e^{-\frac{2t^2-2ts+s^2}{2\sigma^2}} dt =$$

$$= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{z/2} ds \int_{-\infty}^{\infty} e^{-\frac{t^2-ts+s^2/4}{\sigma^2}} e^{-\frac{s^2}{4\sigma^2}} dt = \frac{1}{2\pi\sigma^2} \int_{-\infty}^{z/2} e^{-\frac{s^2}{4\sigma^2}} ds \int_{-\infty}^{\infty} e^{-\frac{\left(t-s/2\right)^2}{\sigma^2}} dt =$$

$$= \frac{1}{\pi\sigma^2} \int_{-\infty}^{u} e^{-\frac{u^2}{\sigma^2}} ds \int_{-\infty}^{\infty} e^{-\frac{t^2}{\sigma^2}} dt = \frac{1}{\sqrt{\pi}\sigma} \int_{-\infty}^{u} e^{-\frac{u^2}{\sigma^2}} ds = \frac{1}{\sqrt{2\pi}\frac{\sigma}{\sqrt{2}}} \int_{-\infty}^{u} e^{-\frac{u^2}{2\left(\frac{\sigma}{\sqrt{2}}\right)^2}} ds$$

The second equation from the left is due to the fact that

$$\int_{-\infty}^{\infty} e^{-\frac{t^2}{\sigma^2}} dt = \sqrt{\pi}\sigma$$

Thus we can conclude that Z_0 has a distribution $N\left(0, \frac{\sigma^2}{2}\right)$

To conclude the proof, we will show that if $X \approx N\left(\mu, \sigma^2\right)$, then

$$Y = X + c \approx N\left(\mu + c, \sigma^2\right)$$

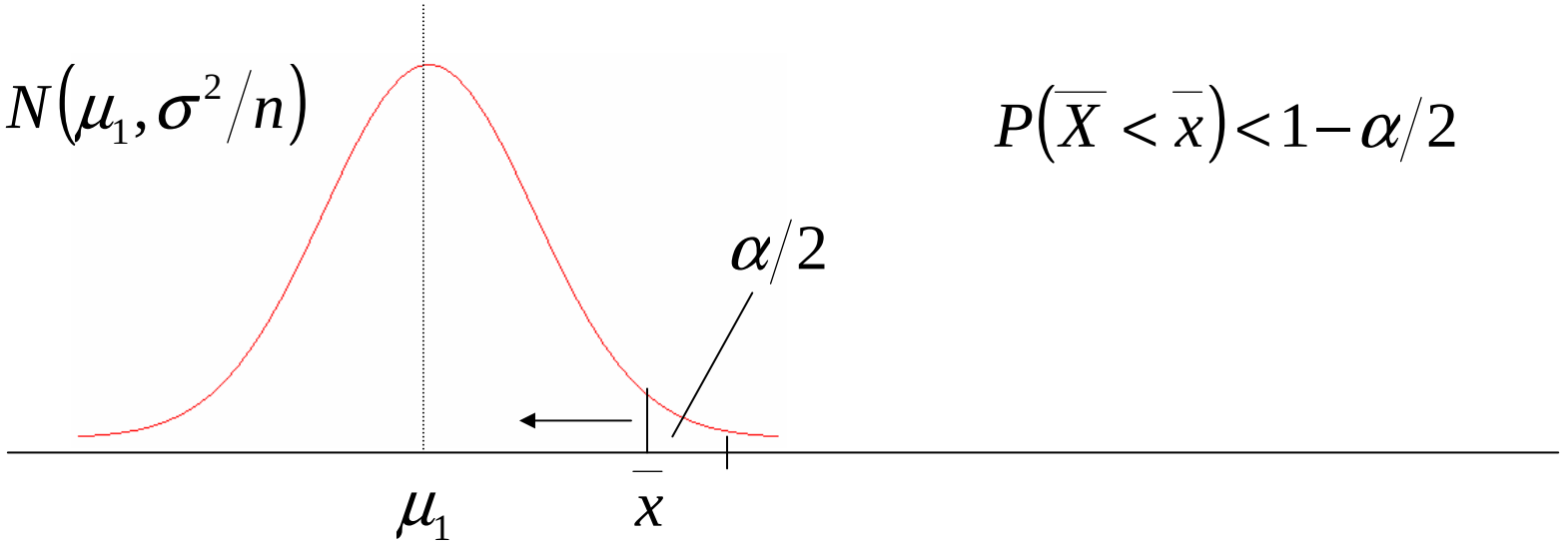
$$F(y) = P(Y < y) = P(X + c < y) = P(X < y - c) =$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{y-c} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \quad x = t - c \quad = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{y} e^{-\frac{(t-(\mu+c))^2}{2\sigma^2}} dt$$

Let us calculate what expectancy μ_1 the population must have for the probability that the value of the sample arithmetic mean is less than $\bar{x}$ is less than $1-\alpha/2$ and, what expectancy μ_2 the population must have for the probability that the value of the sample arithmetic mean is greater than $\bar{x}$ is less than $\alpha/2$.

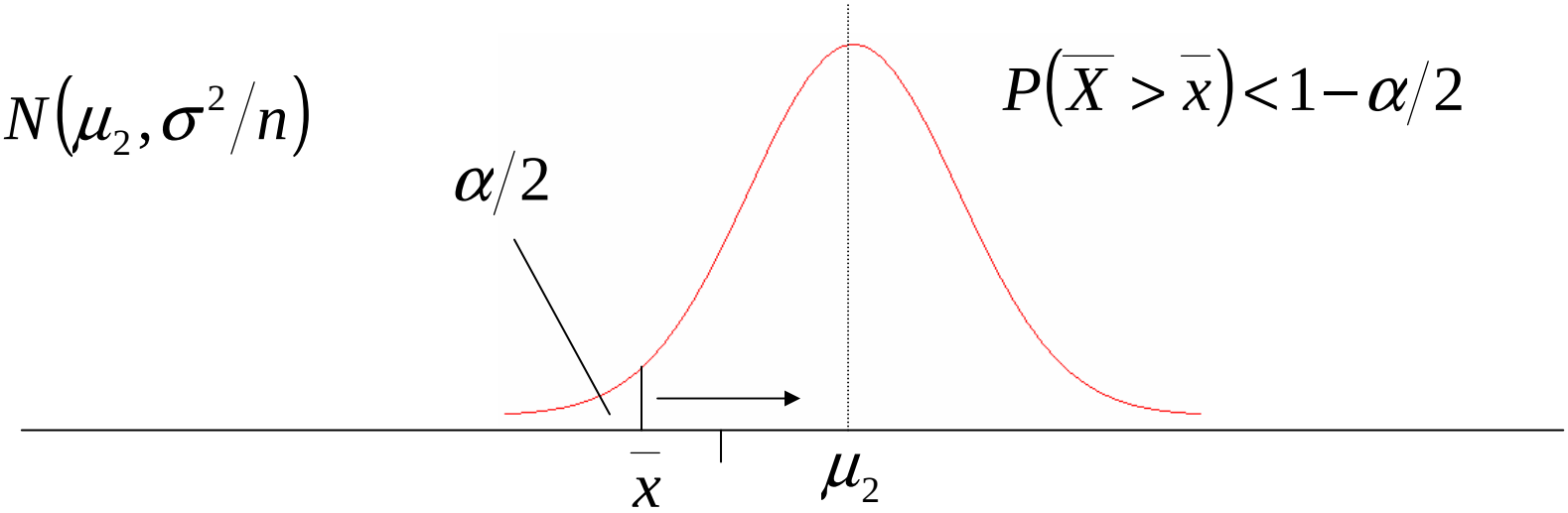
$\overline{X} \approx N\left(\mu_1, \sigma^2/n\right)$

$P\left(\overline{X} < \overline{x}\right) < 1-\alpha/2$



$\overline{X} \approx N\left(\mu_2, \sigma^2/n\right)$

$P\left(\overline{X} > \overline{x}\right) < 1-\alpha/2$



$$P\left(\overline{X} < \bar{x}\right) = P\left(\frac{\overline{X} - \mu_1}{\sigma/\sqrt{n}} < \frac{\bar{x} - \mu_1}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{\sqrt{n}\left(\bar{x} - \mu_1\right)}{\sigma}\right)$$

$$P\left(\overline{X} > \bar{x}\right) = 1 - P\left(\frac{\overline{X} - \mu_2}{\sigma/\sqrt{n}} < \frac{\bar{x} - \mu_2}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{\sqrt{n}\left(\bar{x} - \mu_2\right)}{\sigma}\right)$$

$$\Phi\left(\frac{\sqrt{n}\left(\bar{x}-\mu_1\right)}{\sigma}\right)\le 1-\alpha/2 \qquad 1-\Phi\left(\frac{\sqrt{n}\left(\bar{x}-\mu_2\right)}{\sigma}\right)\le 1-\alpha/2$$

$$\Phi\left(\frac{\sqrt{n}\left(\bar{x}-\mu_1\right)}{\sigma}\right)\le 1-\alpha/2 \qquad \Phi\left(-\frac{\sqrt{n}\left(\bar{x}-\mu_2\right)}{\sigma}\right)\le 1-\alpha/2$$

$$\frac{\sqrt{n}\left(\bar{x}-\mu_1\right)}{\sigma}\le u_{1-\alpha/2} \qquad -\frac{\sqrt{n}\left(\bar{x}-\mu_2\right)}{\sigma}\le u_{1-\alpha/2}$$

$$\mu_1\ge \bar{x}-\frac{\sigma}{\sqrt{n}}u_{1-\alpha/2} \qquad \mu_2\le \bar{x}+\frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}$$

$$\bar{x}-\frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}\le \mu\le \bar{x}+\frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}$$

Now we can solve Example 2:

We have $n = 10, \sigma = 10, \bar{x} = 174.9$

If we choose $1 - \alpha = 0.95$ we get $u_{1-\alpha/2} = u_{0.975} = 1.96$

and so we can write

$$174.9 - \frac{10}{\sqrt{10}} 1.96 \leq \mu \leq 174.9 + \frac{10}{\sqrt{10}} 1.96$$

$$168.7 \leq \mu \leq 181.1$$

Thus, we can conclude that, with probability 0.95, the average height of the population in question lies between 168.7 and 181.1

We can think of the formulas

$$\bar{x} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} \quad \text{and} \quad \mu \le \bar{x} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}$$

as of implementations of the statistics

$$\overline{X} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} \quad \text{and} \quad \overline{X} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}$$

$\left[\overline{X} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} \right]$ is called a confidence interval and its value $\left[\bar{x} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, \bar{s} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} \right]$ for a particular sample is called an interval estimate at confidence level $1-\alpha$

The confidence interval has been derived on the assumption that the population distribution is normal. With such an assumption, other confidence intervals can also be derived:

$$\left[\overline{X} - \frac{S}{\sqrt{n}} t_{1-\alpha/2}, \overline{X} + \frac{S}{\sqrt{n}} t_{1-\alpha/2}\right]$$

is a confidence interval for the expectancy if the population variance and expectancy are not known.

Here $t_{1-\alpha/2}$ denotes the quantile of the Student's t-distribution

with $k = n - 1$ degrees of freedom.

$$\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{\alpha/2}}\right]$$

is a confidence interval for the variance if the population variance and expectancy are not known.

Here $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ are quantiles of Pearson's chi-squared distribution with $k = n$-1 degrees of freedom.