10
Inferential statistics

10.1 Population and sample

x is a sample of numerosity n: $x = (x_1, x_2, \ldots, x_n)$. *Sample* is defined as a subset of the investigated population, where the population can be finite or infinite. *Population* is the set of the statistical units of interest. It can also be defined as the set of all measurements of interest to the sample collector. It could be the number of freshmen students enrolling at the department of physics, the number of companies in a region, the total residents of a nation, the concentration of nitrogen in a soil over a given region, the results of an experiment and so forth. In some cases the population can be known, for instance the number of students enrolled in the department of physics at the University of Bologna. In this case, if we want to know the average age of the students, it is possible to compute it. The 7 mean is a number, just one number. We do not need to use inferential statistics to know this information, it is sufficient to use descriptive statistics. On the other hand, in many cases the population cannot be measured. The reasons can be many. The number of elements are too large, such as the size of the sand particles in the soils of a given region, the concentration of a pollutant in the atmosphere, the number of leaves in a forest. In some cases, the experiment can be performed only on a limited number of sample for financial, practical and time constrain. For instance during a political pool it is not possible to interview all the possible voters in an election, therefore only samples are used. In all these cases, the statistical analysis becomes inferential.

To infer means to deduce or conclude (something) from evidence and reasoning rather than from explicit statements. Statistical inference is the process of drawing conclusions about populations or scientific evidence from data. It is the process of using data analysis to infer properties of an underlying distribution of probability. Commonly, the inference is performed about a population while having data only on samples.

It is possible to look at the numbers again, $(x_1, x_2, \ldots, x_n)$ as realizations of the discrete random variables(*drv*) $(X_1, X_2, \ldots, X_n)$, in the following way.

A sample of numerosity equal to 10, is extracted from the finite population made by all the final graduation votes (in 110/110) of the graduated students in Statistical Science in the year 2007-2008. For instance the series below:

$$109, 110, 100, 107, 99, 109, 110, 98, 100, 105$$

Another sample, still of numerosity equal to 10, could be:

$$100, 108, 96, 110, 98, 104, 110, 99, 100, 102$$

and so forth. In the example the *drv* X_1 took the value of 109 in the first sample, 110 in the second and so forth. The (*drv*) X_2 took the value of 100 in the first sample, 108 in the second sample, etc.. The same is repeated for the other values that the *drv* can take (X_i), in the range 66 to 110.

Given the numerosity of the sample, n, the vector of the *drv*) is:

$$\mathbf{X} = (X_1, X_2, \ldots, X_n) \tag{10.1}$$

which is a multi−variate random variable. All the possible realizations of the set of *drv* $(X_1, X_2, \ldots, X_n)$ creates the space of the samples (S). S is the set of all the samples that can be extracted from the population. It is possible to think at S as a n-dimensional space where the samples are point in space.

Let us suppose that the *drv* $(X_1, X_2, \ldots, X_n)$ are independent random variables and identically distributed (IID)(it may not be always the case), then the sample is said to be *random*. The definition said that the *drv* $(X_1, X_2, \ldots, X_n)$ have the same distribution, but did not specified which distribution. Indeed, often the distribution is not known, but it is possible to find a solution with the procedures of *statistical inference*.

The hypothesis that the *drv* $(X_1, X_2, \ldots, X_n)$ are (IID) means that the observed data or the results of the measurements are not correlated. Obviously this is not always true and often it is not possible to determine this *a priori*. Therefore the assumption is that the correlation are *weak enough*, such that are not going to significantly affect the results of the statistical analysis. It is assumed, otherwise, that the sample is *representative*, meaning that it will 'represent' the entire population, or somehow a faithful images (hologram).

10.2 Deriving the mean and variance of different random variables

Let's suppose we have two independent random variables X and Y. As described above the expected values for these two random variables are:

$$E[X] = \mu_X \tag{10.2}$$

and

$$E[Y] = \mu_Y \tag{10.3}$$

The expected value of a random variable X, denoted $E(X)$ is a generalization of the weighted average, and is intuitively the arithmetic mean of a large number of independent realizations of X. The variance is:

$$\mathrm{Var}(X) = E[(X - \mu_X)^2] = \sigma_X^2 \tag{10.4}$$

and

$$\mathrm{Var}(Y) = E[(Y - \mu_Y)^2] = \sigma_Y^2 \tag{10.5}$$

We introduce a third random variable Z defined as $Z = X + Y$. The expected value is:

$$E[Z] = E[X + Y] \tag{10.6}$$

which means that the mean of X plus the mean of Y:

$$\mu(Z) = \mu_X + \mu_Y \tag{10.7}$$

and if we have another r.v. A such that:

$$E[A] = E[X - Y] \tag{10.8}$$

which means that the mean of X minus the mean of Y:

$$\mu(A) = \mu_X - \mu_Y \tag{10.9}$$

So what is the variance of random variable Z and A.

$$\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y) \tag{10.10}$$

which is

$$\sigma_Z^2 = \sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 \tag{10.11}$$

The variance of the r.v. A is exactly the same thing:

$$\text{Var}(A) = \text{Var}(X) + \text{Var}(Y) \tag{10.12}$$

$$\sigma_A^2 = \sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 \tag{10.13}$$

why it is not minus as for the mean ?

$$\sigma_A^2 = \sigma_{X-Y}^2 = \sigma_{X+(-Y)}^2 = \sigma_X^2 + \sigma_{-Y}^2 \tag{10.14}$$

now we will see that the variance of negative $-Y$ is:

$$\sigma_{-Y}^2 = \text{Var}(-Y) = E[(-Y - E(-Y))^2] = E[(-1)^2(Y + E(-Y))^2] \tag{10.15}$$

since $E(-Y) = -E(Y)$, therefore

$$\sigma_Y^2 = E[(Y - E(Y))^2] \tag{10.16}$$

Therefore the variance of the difference of two independent random variable is equal to the sum of the variances.

$$\sigma_A^2 = \sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 \tag{10.17}$$

Indeed, it does not matter if we take the negative or the positive of the variable, since we are measuring the absolute distance, so it makes sense. So the quantity σ_Y^2 and

σ^2_{-Y} are the same thing. The important aspect of this derivation is that the mean of differences is the same as the differences of their means:

$$\mu_{X-Y} = \mu_X - \mu_Y \tag{10.18}$$

and the variance of differences is the same as the sum of their variances:

$$\sigma_{X-Y} = \sigma_X + \sigma_Y \tag{10.19}$$

Now, let us consider two random variable X and Y:

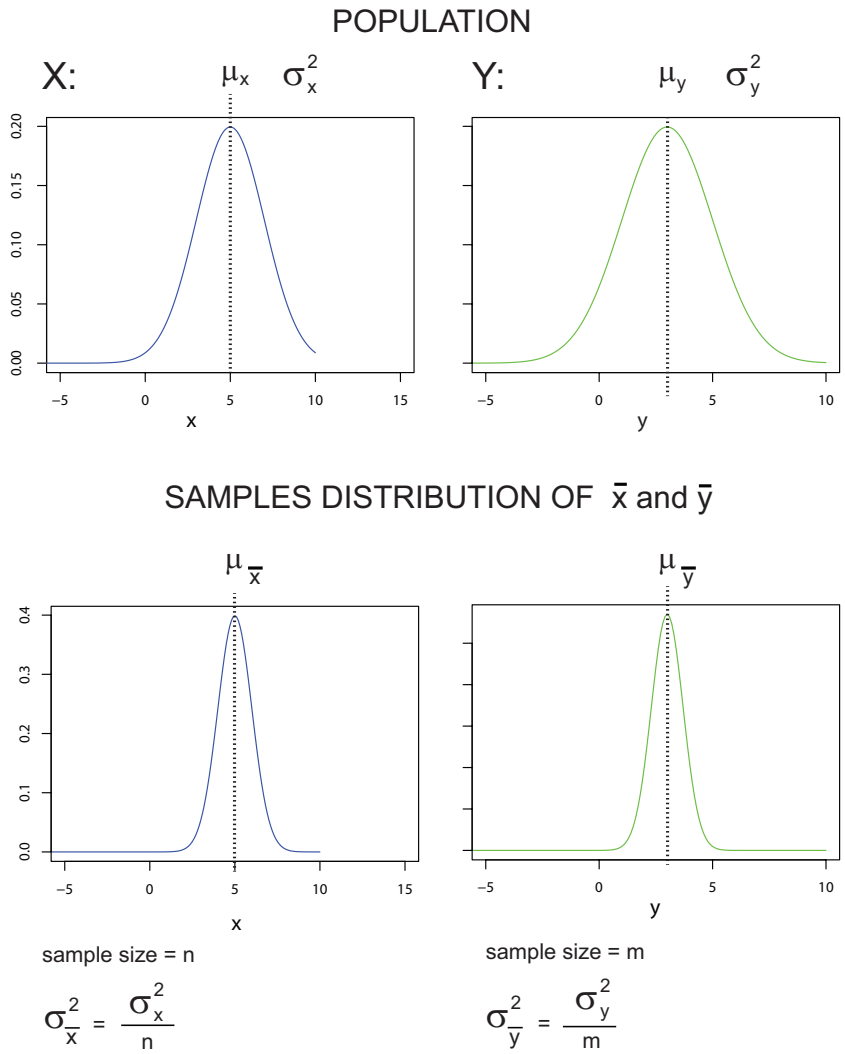

Fig. 10.1 Sampling process from a population

If we consider the difference between there two random variables:

$$Z = \bar{X} - \bar{Y} \tag{10.20}$$

where $\bar{X}$ and $\bar{Y}$ are now the random variables of the means, we will have

$$\sigma^2_{\bar{X}-\bar{Y}} = \sigma^2_{\bar{X}} + \sigma^2_{\bar{Y}} = \frac{\sigma^2_X}{n} + \frac{\sigma^2_Y}{m} \tag{10.21}$$

a new distribution with larger variance will be obtained since the variances of the two distribution are obtained:

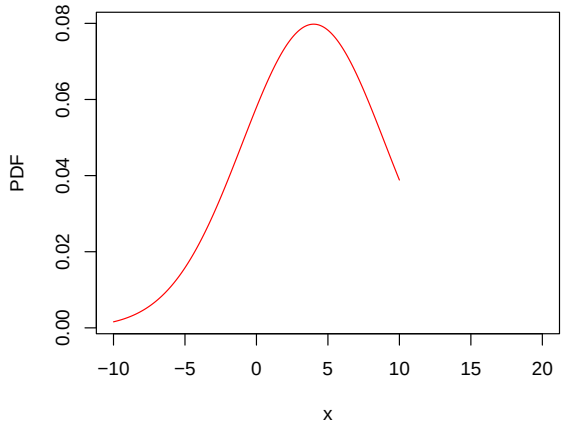

Fig. 10.2 Difference between sampling populations

Therefore the standard deviation is:

$$\sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma^2_X}{n} + \frac{\sigma^2_Y}{m}} \tag{10.22}$$

it is possible to notice that this equation looks like a distance formula, which will help to visualize these concepts in a geometrical way. Moreover, these concepts will be the basis for starting to compare different populations means to identify stastically significant differences among populations.

10.2.1 Example

In this example, the precipitation at Nevada City are used. The population are the precipitation at Nevada City from 1873 to 1978. mean: $\mu = 1337.57$ standard deviation: $\sigma = 391.83$

It is better to write as $\sigma = 390$ e $\mu = 1300$ e $\sigma = 390$ since in the error it is enough one decimal points. These parameters are referred to the population and they

are called *parameters*. The computation of these parameters is shown below, where the data are imported, the values are converted into [mm], and then mean, variance and standard deviation are computed.

```
#CODE Ch10_1.R
setwd("~/Didattica/R_class_4/exercises/Ch9_statistical_inference")
Nevada_prec <- read.table("Nevada_prec.dat"
, sep = "", check.names = FALSE, header = T, na.strings = c("NA", "NAN"))
#Convert precipitation from inches to mm
Nevada_prec$Prec_mm = Nevada_prec$Prec * 25.4
Nevada_prec$Prec_mm < numeric()
mu<- mean(Nevada_prec$Prec_mm) #mean
V<- var(Nevada_prec$Prec_mm) #variance
S<- sd(Nevada_prec$Prec_mm) #standard deviation
```

The output is

```
> mu
[1] 1337.574
> V
[1] 153534.4
> S
[1] 391.8347
>
```

The parameters described above (mean, variance and standard deviation) are known fixed numbers representing a **population**, since the entire data set for that period is known. In many cases the population is unknown and, as described above, the purpose of **inferential statistics** is to derive the unknown parameters from a **sample**. For instance a sample is:

$$n = 20: \; 1422.4, 1752.6, 762.0, 1854.2, 1143.0, 1371.6, 1041.4, 1320.8, 812.8, 812.8,$$
$$914.4, 889.0, 1600.2, 1803.4, 1778.0, 1016.0, 2108.2, 1092.2, 1270.0, 762.0$$

In this case 1422.4 is the realization of the discrete random variable X_1, ..., 762.0 is the realization of the discrete random variable X_{20}, where they are all IID. Now the mean of the population (μ) is estimated through the sample mean, note that as we described before, the symbol for the mean is now $\bar{y}$ where the mean for this sample is $\bar{y} = (1422.4 + 1752.6 + \cdots + 762.0)/20 = 1276.4$. The standard deviation (s) can also be estimated:

$$s = \left[\sum_{i=1}^{20} (x_i - \bar{x})^2 / (n-1) \right]^{1/2} = 418.9 \tag{10.23}$$

Now $\bar{y}$ and s are computed from data from the sample, therefore they are *estimated* values, not true values.

As discussed above, the simplest statistical inference problem is **point estimation**, where a single value (for instance the mean) from the sample data is used to *estimate* a population parameter. In general, define θ as a parameter of the population described by random variables X. X is unknown (the distribution is unknown), and θ is unknown. The quantity θ is called *parameter* and it is a constant of the population. θ is a measure of the distribution of one or more character of the population, for instance the mean μ or the variance σ^2.

The **estimation** of the parameter θ is function of the observations of the sample $\hat{\theta} = t(x_1, \ldots, x_n)$. *When another sample is used, another value is obtained* $\hat{\theta}$. Therefore $\hat{\theta}$ is a realization of the *drv* $\Theta = t(X_1, \ldots, X_n)$.

The discrete random variable Θ is called **estimator**. Since Θ is a random variable, it has a distribution. This is called the principle of *plug-in* (substitution). In other words: *the mean of the population is estimated from the sample mean, the population variance is obtained from the sample variance and the population standard deviation is obtained from the sample standard deviation.* Note that not always the plug−in principle is a good choice.
An example is given where a sample with $n = 20$ is considered:

$$914.4, 1371.6, 762.0, 1117.6, 1092.2, 1905.0, 1600.2, 1320.8, 1016.0, 1244.6,$$
$$1701.8, 1270.0, 1524.0, 889.0, 1625.6, 1117.6, 1066.8, 1143.0, 1752.6, 1473.2$$

The mean is $\bar{y} = 1295.4$ and the standard deviation is $s = 315.7$. The values of $\bar{y}$ are realizations of the estimator *sample mean*:

$$\overline{X} = \sum_{i=1}^{n} X_i / n \tag{10.24}$$

Now $\Theta = \overline{X}$. In the following figures different distributions made by sample averages of 5, 20 and 50 (still with $n = 20$) are depicted. Because of the *central limit theorem*, the realizations of $\overline{X}$, by increasing the number, they tend to assume the well−known bell shape, typical of the normal distribution.

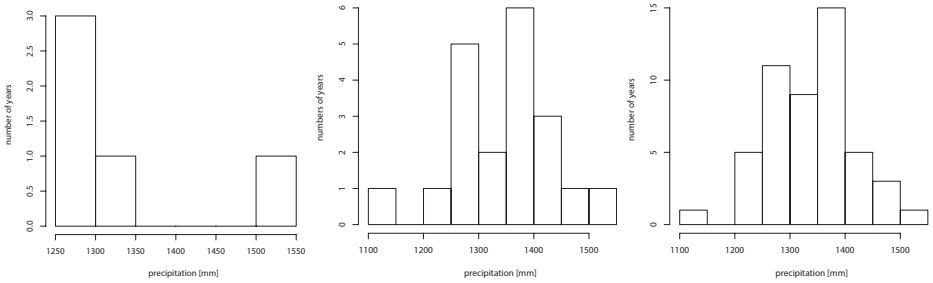

Fig. 10.3

Figure 10.2.1 shows the normal distribution, superimposed over the distribution of the sample means for the precipitation example. The histogram is normalized (total area = 1) for comparison against the normal distribution. The total number of sample means is 100, but the sample size is 25 on the left and 75 on the right.

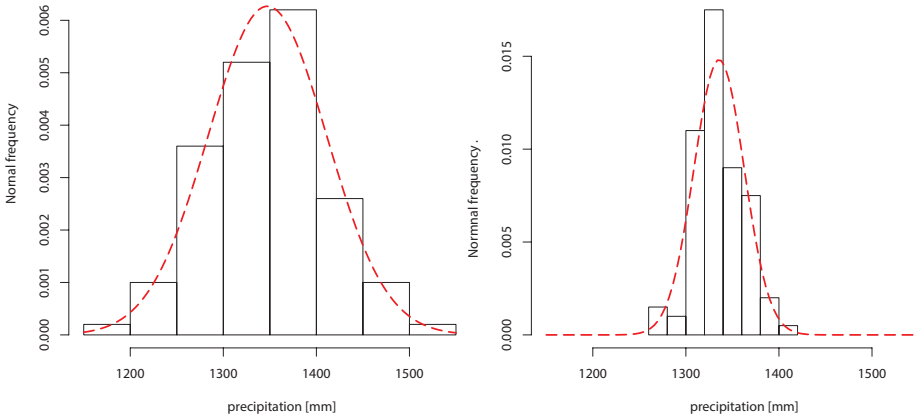

Fig. 10.4

The numerical results are Left: mean = 1347.4, standard deviation = 63.6 ; Right: mean = 1335.7, standard deviation = 26.9.

10.3 Confidence Intervals

In the section above the concept of **point estimator** was provided. However a point estimation is not a reliable assessment of the population without a measure of **uncertainty** about the estimation. How close is the estimated mean close to the population mean. There are several ways to quantify uncertainty. A common method is the concept of **confidence interval**. The confidence interval can be conceptualized as:

$$\bar{x} \pm \text{a Margin of Error} \tag{10.25}$$

where $\bar{x}$ is the sample mean, which is used as **estimator** for the population mean μ. We start from a Normal distribution with mean (μ) and variance (σ^2/n):

$$\mathcal{N}(\mu, \frac{\sigma^2}{n}) \tag{10.26}$$

To properly quantify the error a few assumptions are made: (a) the sample is randomly selected from the population, (b) the population is normally distributed, standardizing $\bar{X}$ by subtracting the mean and divide by the standard deviation:

$$\mathcal{Z} = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1) \tag{10.27}$$

where $\mathcal{Z}$ is a normalized normal distribution with mean = 0 and variance= 1. $\bar{X}$ is a random variable with a normal distribution with mean μ and variance σ. $\bar{X}$ represents the many possible means that are obtained by drawing different samples. The third assumption (c) is that the population standard deviation σ is known. In practice this is commonly rare since in many cases the population is unknown and therefore its standard deviation as well, therefore the population standard deviation will be estimated from samples as well. For now, the assumption is that the population σ is known.

The general equation for the $(1-\alpha)$ confidence interval for μ is:

$$\bar{X} \pm \frac{z_\alpha}{2}\frac{\sigma}{\sqrt{n}} \tag{10.28}$$

where the term on the right end side is the margin of error. $(1-\alpha)$ is the confidence level and it can be 95% or 99%. The 95% is the most common choice. The confidence interval is computed by choosing the correct z_α value. Figure 10.3 shows the normalized normal distribution for a confidence interval of 95%. The graph shows the indicated values of $z_\alpha/2$ ($\pm$1.96 in this case), the area $(1-\alpha)$ and the two tails which are $\alpha/2$. Each tail is obtained by splitting α evenly in the right and left tails. For any confidence level, the appropriate $z_\alpha/2$ value must be computed. In R it is possible to compute the value by typing the `qnorm` which returns the quantile given a fraction value.

```
qnorm(0.975)
[1] 1.959964
```

Since 1-0.025 = 0.975.

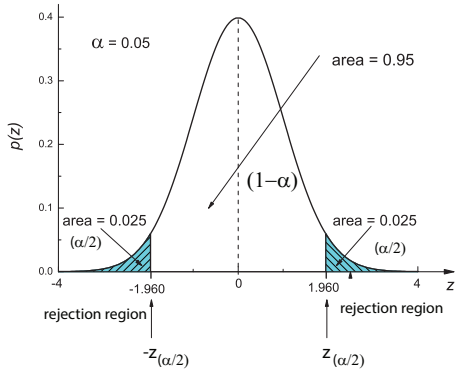

Fig. 10.5 Confidence intervals for α = 0.05.

The standard confidence interval α equal to 0.05 (95% confidence interval, $1-\alpha$ = 0.95) is:

$$\mathrm{P}\left(-1.960 \leq \frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}} \leq 1.960\right) = 0.95 \tag{10.29}$$

multiplying by $(\sigma/\sqrt{n})$ and subtracting $\bar{X}$ from each term:

$$\mathrm{P}\left(\bar{X} - 1.960\,\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.960\,\frac{\sigma}{\sqrt{n}}\right) = 0.95 \tag{10.30}$$

where $\bar{X}$ is a random variable and also $\bar{X_-} = \bar{X} - 1.960\,\sigma/\sqrt{n}$ and $\bar{X_+} = \bar{X} + 1.960\,\sigma/\sqrt{n}$ are random variables.

This equation describes that 95% of the values determined by the realizations $\bar{X_-}$ and $\bar{X_+}$, cover the value of μ. If a single sample is observed with mean $\bar{x}$, which is a realization of $\bar{X}$ it can be said that (with 95% confidence), that:

$$\bar{x} - 1.960\,\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.960\,\frac{\sigma}{\sqrt{n}} \tag{10.31}$$

the interval $[\bar{x} - 1.960\,\sigma/\sqrt{n}, \bar{x} - 1.960\,\sigma/\sqrt{n}]$ is the confidence interval at 95%.

Obviously, the **population mean** μ is a fixed number, although it may be unknown, *it is not a random variable*. Therefore it is wrong to say that the mean of the population has a probability of 95% to be within the interval $[\bar{x} - 1.960\,\sigma/\sqrt{n}, \bar{x} - 1.960\,\sigma/\sqrt{n}]$. The variable that changes is $\bar{x}$, that varies from sample to sample. It is better to say that the *confidence interval* covers the true value of the mean. Indeed, it is better to use the word *coverage interval*, meaning that the interval described above has a probability 95% to *cover* the unknown value of μ. Therefore it exists a probability equal to α that the sample comes from a population where the `mean` is outside of the interval.

Example: Let us assume to have a sample $(9.0, 7.0, 14.0, 13.0, 5.0, 10.4, 6.6, 8.5, 7.5)$ collected from a population with normal distribution and known $\sigma = 1$. The sample mean is $\bar{x} = 81/9 = 9$. Substituting the values of sample mean $(\bar{x})$ and variance (σ) into eq. 10.31, into the confidence interval at 95%, leads to:

$$9 - 1.960\,\frac{1}{\sqrt{9}} \leq \mu \leq 9 + 1.960\,\frac{1}{\sqrt{9}} = [8.35, 9.65] \tag{10.32}$$

There is 95% confidence that this interval will cover the *unknown true mean* of the population. In this case, since the true mean (μ) was 9, the outcome was correct.

In the example below, a sequence of numbers for a sample is created. An estimator plug-in is computed.

```
#Code Ch10_2.R
#Confidence Intervals for a random distribution and
#plug-in estimator for a randomly selected sample
# Define sample elements
y <- c(52, 104, 146, 10, 50, 31, 40, 27, 46)

#Confidence intervals (for a normal distribution)

#Variance
sig2p <- function(x) {var(x)}
# Confidence interval 90%
```

```
conf.level<-0.90

# Standard Interval
int.stand <- function(x,conf.level) {
theta.hat<-mean(x) # theta.hat = estimated mean
SE <- sqrt(sig2p(x)/length(x)) # plug-in sigma/sqrt(n)
alpha<-1-conf.level
# Density, distribution function,
#quantile function and random generation for the normal
#distribution with mean equal to
#mean and standard deviation equal to sd.
#half on one side and one on the othertheta.hat
#+ c(-z.1malpha*SE,+z.1malpha*SE) #average +- standard error
z.1malpha <- qnorm(1-alpha/2)
       }

       #Sample data
       theta.hat<-mean(y)
       theta.hat
       SE <- sqrt(sig2p(y)/length(y))

       int.stand(y,conf.level)

       plot(function(x) dnorm(x,theta.hat,SE),0,110)
       quant<-quantile(y, probs = c(0.025, 0.1587, 0.5, 0.8413, 0.975))
       segments(quant[1],0,quant[1],2,lty=2,col="red",lwd=2)
       segments(quant[4],0,quant[4],2,lty=2,col="red",lwd=2)
```

Here the previous example about precipitation in Nevada city is discussed. The standard deviation σ is unknown and the population from which the sample was collected is not a normal distribution. Nevertheless, it is possible to compute an "approximated level of confidence" based on the convergence toward a normal distribution, described by the central limit theorem. Moreover, it is possible to estimate σ by using the sample standard deviation s. Here the sample that was used before is considered again with $n = 20$: $1422.4, 1752.6, \ldots, 762.0$.

The sample mean is $\bar{x} = 1276.4$. So the population mean ($\mu = 1337.57$) was estimated using the sample mean $\bar{x}$, interpreted as a realization of a random variable $\bar{X}$. Now, it is of interest to determine the accuracy of the estimation. The standard deviation of the population ($\sigma = 391.83$), was estimated from the sample standard deviation $s = 418.9$, again a realization of the random variable sample standard deviation.

$$Var\left(\bar{X}\right) = \sigma^2/n \tag{10.33}$$

where σ^2 was replaced by s^2. The confidence interval at 95% is then:

$$\left[1276.4 - 1.960\,\frac{418.9}{\sqrt{20}}, 1276.4 + 1.960\,\frac{418.9}{\sqrt{20}}\right] = [1092.76, 1459.94] \tag{10.34}$$

Note that the size of the interval $2 \times z_{\alpha/2}\,(\sigma/\sqrt{n})$ is independent from the mean, while it is function of the standard deviation σ (or from the estimator s), therefore from the intrinsic variability of the sample, and from the sample size (number n of elements). In the code below the R code is presented.

The function `sample()` is used to perform an operation of re-sampling. The idea is to simulate a collections of samples, without knowing the entire population. Obviously in this case, we know the total number of cumulative precipitation for each year, so the population is known. In this exercise we pretend to create a certain number of collected samples, generated randomly.

The `for` loop is generating a series of random numbers starting always from the same seed, therefore the series is always the same. It is useful to employ this method, such that differences in various `statistical experiments` are not due to the variation in the generated random numbers. An important feature of the function `sample`

Let us imagine to extract numbers from the bingo, where numbers are in the interval (1-90). If samples of 25 elements are collected each time, but then I put the collected numbers back into the box, those numbers could be picked up again. This method is called with re−introduction. If there are many numbers (maybe 10,000), the computed mean will not be very different if the reintroduction method is selected. However, if the number is small (like the example of the bingo), then the reintroduction method can generate biased means. The selection of this method is possible by using the instruction `replace=FALSE`, there is not reintroduction. The overall procedure is a permutation.

The instruction `set.seed(seed)` set the seed of R‘s random number generator, which is useful for creating simulations or random objects that can be reproduced. For example to create simulated values that are reproducible.

```
> set.seed(4)
> rnorm(4)
[1] 0.2167549 -0.5424926 0.8911446 0.5959806
> set.seed(4)
> rnorm(4)
[1] 0.2167549 -0.5424926 0.8911446 0.5959806
```

The results keep being the same every time, otherwise the rnorm function would return different values each time. The code below employs this function.

```
setwd("~/Didattica/R_class_4/exercises/Ch9_statistical_inference")
Nevada_prec <- read.table("Nevada_prec.dat"
, sep = "", check.names = FALSE, header = T, na.strings = c("NA", "NAN"))
#Convert precipitation from inches to mm
Nevada_prec$Prec_mm = Nevada_prec$Prec * 25.4
Nevada_prec$Prec_mm < numeric()
mu<- mean(Nevada_prec$Prec_mm) #mean
V<- var(Nevada_prec$Prec_mm) #variance
```

```
S<- sd(Nevada_prec$Prec_mm) #standard deviation

xx<-numeric()
mu<-numeric()
sigma<- numeric()

ltot<- 100

for(l in 1:ltot){
       #generate the seed to create a random series,
       #that is always the same (with the same seed)
       set.seed(l+400)
       #number of elements for each sample,
       #with permutation because replace = FALSE
       B<- 25
       xx<- sample(Nevada_prec$Prec_mm,B,replace=F)
       mu[l]<- mean(xx)
       sigma[l]<- sd(xx)
}
mu
sigma

hist(mu, xlab="mm di pioggia",ylab="freq. norm.",main=" ",
prob=T,plot=T,xlim=c(1150,1550))#,ylim=c(0,0.0065))
#,ylim=c(0,0.018))

m.mu<- mean(mu) #means of the sample means
sd.mu<- sd(mu)
m.mu
sd.mu
#density of the Gaussian function
curve(dnorm(x,m.mu,sd.mu),add=T,lty=5,lwd=2,col="red")
#density of the function
```

Figure 10.3 shows the distribution (histogram) of randomly selected samples from the precipitation values at Nevada city, superimposed over a normal distribution.

10.3.1 Confidence intervals for t-Student distribution

The estimation and test procedures about μ presented earlier in this chapter were based on the assumption that the population variance was known or that we had enough observations (samples) to allow the sample standard deviation s to be a reasonable estimate of the population standard deviation σ. In this section a test is presented to be used when σ is unknown, no matter the sample size. For example, to determine the average concentration of a drug into a patient blood stream one hour after the patient

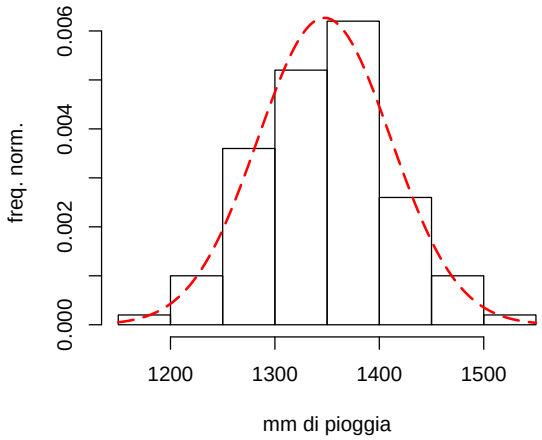

Fig. 10.6 Ramdomly generated samples superimposed over a normal distribution for the Nevada Precipitation values.

suffering from a rare disease was treated with that drug, may not be possible to obtain a random sample of 30 or more observations at a given time.

This test was derived by W.S. Gosset who faced the problem of estimating the mean quality of beer brews, but based on small samples. He thought that using a normal distribution with small σ would lead to falsely reject the null hypothesis at a slightly higher rate than that specified by α. He derived the distribution and percentage points of the test statistic for normal distribution for $n < 30$. He published the results under the pen name Student, because against the company policy to publish his results.

Therefore if σ is unknown, it is not possible to write:

$$\mathcal{Z} = (\bar{X} - \mu)/(\sigma/\sqrt{n}) \sim \mathcal{N}(0,1) \tag{10.35}$$

but it is possible to write:

$$\mathcal{Z} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \tag{10.36}$$

It means that the statistics of the first member of the equation is distributed like a random variable for a t Student with $n-1$ degrees of freedom. In the equation above, the rv are $[\bar{X}]$ and S.

With the t Student (19 degrees of freedom), the quantile corresponds to $1-\alpha = 0.95$, which for the normal distribution is $z_{\alpha/2} = 1.960$, now it is $t_{\alpha/2}^{[n-1]} = 2.093$ $(n-1 = 19)$, a little larger than 1.960.

The confidence interval is then written as:

$$\left[\bar{x} - t_{\alpha/2}^{[n-1]} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}^{[n-1]} \frac{s}{\sqrt{n}}\right] \tag{10.37}$$

In this example is $[1080.3, 1472.4]$, larger than the one before.

If $n \gtrsim 20 - 30$, then $t_{\alpha/2}^{[n-1]} \approx z_{\alpha/2}$. It means that the t Student tends to a normal distribution, with increasing the number of elements in the sample.

It is possible to compute the confidence intervals for the t-student distribution by using

```
qt(0.975,10)
[1] 2.228139...
```

To obtain a value of 1.96 (corresponding to the normal distribution) we must have above 1000 degree of freedom.

```
> qt(0.975,1000)
[1] 1.962339
```

10.3.2 Terminology

The standard deviation of an estimator is often called *standard error*, and it is indicated with $se(\cdot)$. For example, for a *sample mean* of a *rv*, $\bar{X}$, the standard error of the mean $\text{se}(\bar{X})$ is

$$\text{se}(\bar{X}) = \sqrt{\text{Var}\left[\bar{X}\right]} = \sigma/\sqrt{n} \tag{10.38}$$

If σ is unknown, standard error is estimated with $s/\sqrt{n}$. It meansthat the standard error is a way to call the ratio:

$$\frac{\text{sample standard deviation}}{\text{sample size}} \tag{10.39}$$

In the example about precipitation $\bar{x} = 1276.4$, the standard error is written $se(\bar{x}) = 418.9/\sqrt{20} = 93.67$. A common way to refer to the results is to write Un modo frequente per riferire il risultato della stima intervallare è scrivere $\bar{x} \pm \text{se}(\bar{x})$. In the example:

$$\bar{x} = 1276 \pm 94 \tag{10.40}$$

The standard error is written with two significant digits $\text{se}(\bar{x})$. With the previous terminology the confidence interval is $[1182, 1370]$ with an approximation level of $\approx 68\%$.

10.4 Hyphotesis tests

The sample $\mathbf{y}$ of numerosity $n_y = 9$, is analyzed with respect to another sample $\mathbf{z}$ of numerosity $n_z = 7$. The two samples are:

$$\begin{array}{ll} \mathbf{ze} = (94, 197, 16, 38, 99, 141, 23) & n_z = 7 \\ \mathbf{y} = (52, 104, 146, 10, 50, 31, 40, 27, 46) & n_y = 9 \end{array} \tag{10.41}$$

The hypothesis test is formulated when the question is: *are the two sample different?* or better *Are the elements of the sample A larger than the elements of the sample B* ?

Within an experiment framework, performed with different methods, then the questions formulated above are equivalent to ask: *is method x better than method y* ?

The data described above are collected from an experiment by Efron et al., where 16 rats were subjected to a treatment with a new drug. Seven rats were the treatment (they received the drug) while nine rats were the control group (they did not receive the drug). To explore if the group **ze** are larger than the group **y** it is possible to compare the sample means.

$$\overline{ze} = 86.857 \qquad \text{and} \qquad \overline{y} = 56.222$$

from which $\overline{ze} - \overline{y} = 30.635$, where d^{obs}, is the observed difference. The d^{obs} is an estimate of what is usually called $\hat{\theta}$.

Now it was observed that the sample means were different. However is this difference a real difference between the samples, or the difference is due to a random or a systematic error ? In other words: does the drug work or not ? The problem is well known and it brings to the concept of hypothesis test. The **first step** is to formulate a null hypothesis H_0. In our case we assume that the sample ze was extracted from a population described by the random variable Z, and that the sample y was extracted from a population described by the random variable Y. The enunciation of the null hypothesis is:

$$H_0 : Z = Y$$

which means that the two samples are collected from two populations with features that have the same distribution. In other words there are no differences. H_0 states that the probabilistic behavior of the sample ze is the same of the sample y, no matter which sample is collected from the the two populations.

The **second step** is to build a statistical test. For instance the difference between the sample means:

$$D = \overline{Z} - \overline{Y}$$

Note that D is a random variable, which is the difference between the two random variables $\overline{Z}$ e $\overline{Y}$.

In our case, more realizations of D are observed if H_0 is not true, than if H_0 is true. In our case, the more realizations of D occurs, more we assume that H_0 is not true. Obviously to quantify the realizations of D, the probability density of D should be known or at least we can formulate a conjecture about it. If the distribution of D is known, then it is possible to compute the probability that the realizations of D are larger than the observed values of d^{obs}.

Then, what is the probability that a realization of D is larger than d^{obs}?

To define the *probability that the rejection of the null hypothesis is only given by change we use the* "p-value".

Another example is to test if the income of a family from Bologna is significantly different from 2000 euros a month. So the null hypothesis is

$$H_0 : \mu = 2000$$

while

$$H_1 : \mu \neq 2000$$

10.5 Example

Here we are presenting an example where the concept presented in the sections above is applied.

We are trying to test if two low sugar diet will help overweight people to loose weight. One group of 100 people are assigned to a low sugar diet, while another group of 100 people ($n = m$) are kept on the same diet with lower calories but with the same amount of sugar, usually the latter is called the *control* group. After 3 months the first group lost (as average) 4.2 kg, while the second group lost 3.3 kg. At a first look it seems that the first diet was indeed effective in reducing weight. So the mean weight loss $\bar{x}$ and standard deviation $\bar{s}$ were:

$$\text{Low sugar} = \bar{x} = 4.2\,,\ \bar{s_x} = 2.11[\text{kg}]$$

and for the control group:

$$\text{Control} = \bar{y} = 3.3\,,\ \bar{s_y} = 1.83[\text{kg}]$$

at a first superficial look it looks like the low sugar group lost more weight that the control. If the difference is computed:

$$\bar{x} - \bar{y} = 4.2 - 3.3 = 0.9\,[\text{kg}]$$

Now the question is to get a 95 % confidence interval around this number. So, as explained above, we want to look at the distribution (assuming that it is normal) of the difference of the means.
This is going to have a mean and a standard deviation as shown in Figure 10.5. We want to make some inferences about this distribution, based on our samples. We want to define an interval where we know that the true mean will be *covered by this interval*. How many standard deviations we need to go to cover this interval ? This is done with the so called Z table that provides the values for the just one tail. For instance the 0.975 gives a value of z = 1.96, as described above. Or, only 2.5 % of the samples are going to be more that 1.96 standard deviations away from the mean.

It can also be written as there are 95 % changes that the value $\mu_{\bar{X}-\bar{Y}} = 1.91$ will be covered within the distance:

$$\sigma_{\bar{X}-\bar{Y}} \times 1.96 \tag{10.42}$$

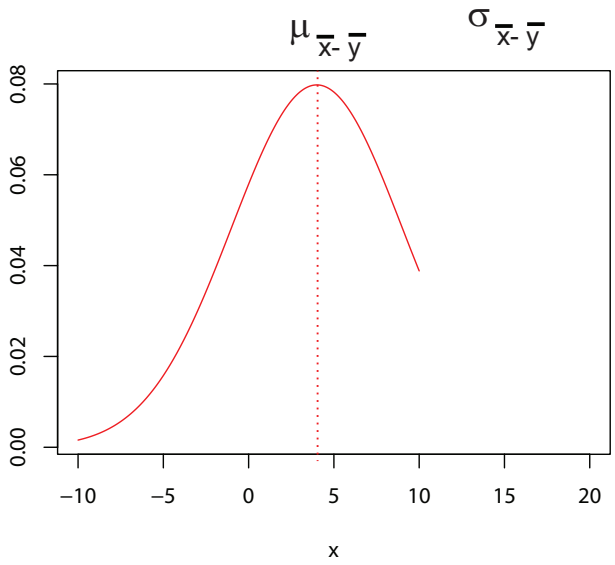

Fig. 10.7 Distribution of the differences of the means.

So the question is how to calculate $\sigma_{\bar{X}-\bar{Y}} \times 1.96$, therefore to compute the standard deviation of the distribution. So the standard deviation will be:

$$\sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \tag{10.43}$$

so by replacing the values the computation will be:

$$\sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_{\bar{X}}^2}{100} + \frac{\sigma_{\bar{Y}}^2}{100}} \tag{10.44}$$

since σ_X and σ_Y are unknown, we can approximate it with the sample standard deviations:

$$s_{\bar{X}-\bar{Y}} = \sqrt{\frac{s_{\bar{X}}^2}{100} + \frac{s_{\bar{Y}}^2}{100}} \tag{10.45}$$

leading to:

$$s_{\bar{X}-\bar{Y}} = \sqrt{\frac{2.11^2}{100} + \frac{1.83^2}{100}} = 0.27 \tag{10.46}$$

now the interval can be computed:

$$0.27 \times 1.96 = 0.52 \tag{10.47}$$

so the confidence interval will be the difference of the means $\pm$ the value computed above:

now the interval can be computed:

$$0.9 \pm 0.52 = 0.52 \tag{10.48}$$

now the interval can be computed:

$$0.38 \leq \text{CI} \leq 1.42 \tag{10.49}$$

the expected value of the sample means and the expected values of the population is the same, therefore this interval gives us an interval that will cover the expected value of the population.