Dr Isabella Gollini

Probability and Statistics

7 Random Samples

7.1 Independent random variables

Let X be a r.v. with $E(X) = \mu$ and $\text{var}(X) = \sigma^2$. It follows from the linearity of the expectation operator that, for any constant a,

$$E(aX) = aE(X) = a\mu.$$

Furthermore,

$$\text{var}(aX) = E[(aX - a\mu)^2] = E[a^2(X - \mu)^2] = a^2E[(X - \mu)^2] = a^2\text{var}(X) = a^2\sigma^2.$$

Thus

$$\text{var}(aX) = a^2\text{var}(X) = a^2\sigma^2.$$

If X and Y are two r.v.s and a and b are two constants then by the linearity of the expectation operator, as stated in Section 5,

$$E(aX + bY) = aE(X) + bE(Y).$$

For any pair of r.v.s, X and Y, and any pair of real numbers, x and y, write $\Pr(\{X \leq x\} \cap \{Y \leq y\})$ as $\Pr(X \leq x, Y \leq y)$.

Definition

Two r.v.s, X and Y, are said to be *(statistically) independent* if, for all pairs of real numbers x and y,

$$\Pr(X \leq x, Y \leq y) = \Pr(X \leq x)\Pr(Y \leq y).$$

This mathematical definition of statistical independence corresponds to the intuitive notion of independence that the value taken by X does not influence nor is influenced by the value taken by Y. So, assuming that the conditioning events have probability greater than zero, $\Pr(X \leq x | Y \leq y) = \Pr(X \leq x)$ and $\Pr(Y \leq y | X \leq x) = \Pr(Y \leq y)$.

It turns out that, for statistically independent r.v.s, variances are additive: if the r.v.s X and Y are statistically independent then

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y).$$

More generally, if the r.v.s X and Y are statistically independent then, for any two constants a and b,

$$\text{var}(aX + bY) = a^2\text{var}(X) + b^2\text{var}(Y).$$

Extending the definition of independence to more than two r.v.s, the r.v.s $X_1, X_2, \ldots, X_n$ are said to be *(mutually statistically) independent* if for all sets of real numbers $x_1, x_2, \ldots, x_n$

$$\Pr(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} \Pr(X_i \leq x_i).$$

The linearity property of the expectation operator may be generalized in the following way. For any set of r.v.s $X_1, X_2, \ldots, X_n$ and any set of constants $a_1, a_2, \ldots, a_n$,

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i). \tag{1}$$

Further, if the r.v.s $X_1, X_2, \ldots, X_n$ are independent then, for any set of constants $a_1, a_2, \ldots, a_n$,

$$\operatorname{var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \operatorname{var}(X_i). \tag{2}$$

7.2 Populations and random samples

Recall from Section 4 that a r.v. X is a number attached to the outcome of an experiment. Suppose that the experiment consists of drawing an individual at random from a population and that the r.v. X is a measurement carried out on the individual. We may consider repeating this experiment and carrying out the associated measurement n times, denoting the measurements $X_1, X_2, \ldots, X_n$. In any particular realization of such a sequence of experiments, the values observed are denoted, as in Section 1, by the lower case letters, $x_1, x_2, \ldots, x_n$, a sample of size n of observed values of X.

- Recall also from Section 4 that, in mathematical terms, a r.v. X is a function defined on the sample space. The distinction in notation here is between $X_1, X_2, \ldots, X_n$ thought of as r.v.s, functions on the sample space, and denoted by upper case letters, and $x_1, x_2, \ldots, x_n$, particular values taken by $X_1, X_2, \ldots, X_n$ on a particular occasion, and denoted by lower case letters.

  However, this distinction will sometimes become blurred.

We now suppose that the we are sampling from a large population and that the individual observations are taken independently of each other, in which case we have what is known as a *random sample.* The r.v.s $X_1, X_2, \ldots, X_n$ are independently and identically distributed, each having the same distribution as X, with mean μ and variance σ^2.

We may also be sampling from what is known as an *infinite population*, for example, when we are making repeated measurements of some quantity, or throwing a die an arbitrary number of times. We have a potentially unlimited supply of observations, so it is as if we were sampling from an infinite population of measurements or tosses. Again, if we assume that n measurements or observations are taken independently of each other, we have a random sample of size n.

7.3 The sampling distribution of the mean

Suppose that $X_1, X_2, \ldots, X_n$ are a random sample of size n, i.e., independently and identically distributed r.v.s from some distribution with mean μ and variance σ^2. Define the sample mean $\bar{X}$ by

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

Applying the results of Equations (1) and (2), with $a_i = 1/n, E(X_i) = \mu, \mathrm{var}(X_i) = \sigma^2$ $(1 \leq i \leq n)$, we obtain

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

and

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

Hence

$$E(\bar{X}) = \mu, \tag{3}$$

identical with the mean of the original distribution, and

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n} \ , \tag{4}$$

equal to the variance of the original distribution divided by n.

If we consider taking many samples of size n from the original distribution, the sample mean $\bar{x}$ will vary from sample to sample. Equation (3) tells us that, on average, $\bar{x}$ will be equal to the mean μ of the individual observations x_i. Equation (4) tells us that the variance from sample to sample of $\bar{x}$ will be smaller by a factor of $1/n$ than the variance σ^2 of the individual observations x_i — taking means has the effect of smoothing out the fluctuations.

It turns out to be a property of the normal distributions that, for any set of normally distributed r.v.s $X_1, X_2, \ldots, X_n$ and any set of constants $a_1, a_2, \ldots, a_n$, the r.v. $\sum_{i=1}^{n} a_i X_i$ is also normally distributed.

In particular, bearing in mind the results of Equations (3) and (4), if $X_1, X_2, \ldots, X_n$ is a random sample of size n from the $N(\mu, \sigma^2)$ distribution then

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \ . \tag{5}$$

It follows from the standardization of Equation (8) of Section 6 that

$$\frac{(\bar{X} - \mu)}{\left(\dfrac{\sigma}{\sqrt{n}}\right)} \sim N(0, 1) \ . \tag{6}$$

There is a famous result in probability theory, the *Central Limit Theorem*, which may be used to conclude that, at least for large n, the results of Equations (5) and

(6) are approximately valid for a random sample $X_1, X_2, \ldots, X_n$, of size n from some distribution with mean μ and variance σ^2, even if $X_1, X_2, \ldots, X_n$ are not themselves normally distributed.

Theorem 1 (The Central Limit Theorem) *Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of independently and identically distributed r.v.s, each having a distribution with mean μ and variance σ^2. Define*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (n = 1, 2, 3, \ldots).$$

As $n \to \infty$,

$$\frac{(\bar{X} - \mu)}{\left(\dfrac{\sigma}{\sqrt{n}}\right)} \overset{D}{\to} Z,$$

where $Z \sim N(0, 1)$.

- The notation $\overset{D}{\to}$ is used for "convergence in distribution," which we are not defining formally in this course.
- To put it less precisely, as $n \to \infty$, the distribution of $\bar{X}$ converges to the $N(\mu, \sigma^2/n)$ distribution.

If we consider taking many samples of size n from some distribution, the *sampling distribution of the mean* describes how the sample mean $\bar{x}$ will vary from sample to sample. If we are sampling from a $N(\mu, \sigma^2)$ distribution then Equation (5) or (6) specifies exactly the sampling distribution of the mean. If we are sampling from some other distribution with mean μ and variance σ^2 then Equation (5) or (6) gives us an approximation to the sampling distribution of the mean, especially so if n is large.

7.4 Percentage points of the standard normal distribution

The $P\%$ *percentage point* $x(P)$ of the standard normal distribution is such that if $Z \sim N(0, 1)$ then $\Pr(Z > x(P)) = P\% = P/100$, as illustrated in Figure 1. It follows that

$$\Phi(x(P)) = \Pr(Z \leq x(P)) = 1 - \frac{P}{100}$$

and hence

$$x(P) = \Phi^{-1}\left(1 - \frac{P}{100}\right), \tag{7}$$

where Φ denotes the c.d.f. of the standard normal distribution and Φ^{-1} is its inverse.

Percentage points of the standard normal distribution can be found in Table 5 of the *New Cambridge Statistical Tables.*

We find, for example, that $x(2.5) = 1.9600$, so that, making use of the symmetry of the p.d.f. of the $N(0, 1)$ distribution, if $Z \sim N(0, 1)$ then

$$\begin{aligned} \Pr(Z > 1.96) &= 0.025 = 2.5\% \\ \Pr(Z < -1.96) &= 0.025 = 2.5\% \\ \Pr(-1.96 < Z < 1.96) &= 0.95 = 95\% \,. \end{aligned}$$

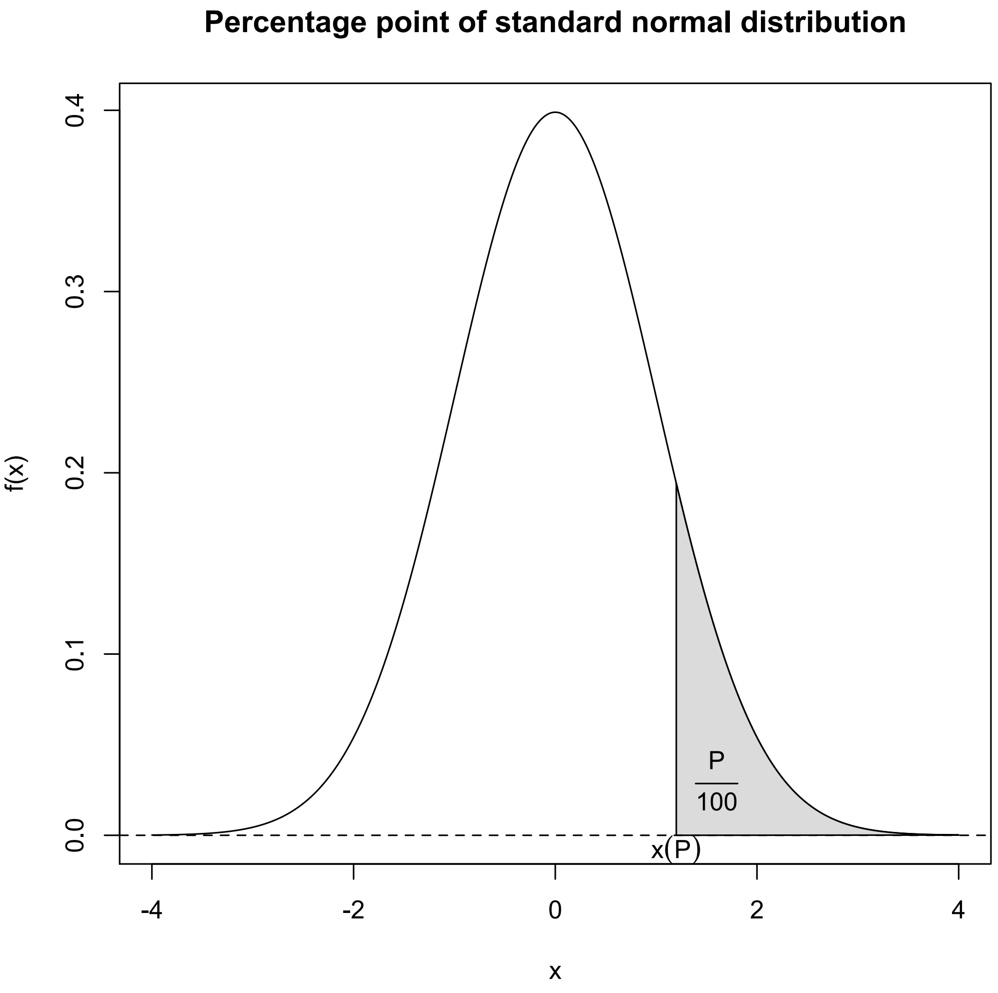

Figure 1: $x(P)$ is the $P\%$ percentage point

More generally, if $X \sim N(\mu, \sigma^2)$ then, using the standardized version of the r.v. X,

$$\begin{aligned} \Pr((X-\mu)/\sigma > 1.96) &= 0.025 \\ \Pr((X-\mu)/\sigma < -1.96) &= 0.025 \\ \Pr(-1.96 < (X-\mu)/\sigma < 1.96) &= 0.95 \end{aligned}$$

or, equivalently, if $X \sim N(\mu, \sigma^2)$ then

$$\begin{aligned} \Pr(X > \mu + 1.96\sigma) &= 0.025 \\ \Pr(X < \mu - 1.96\sigma) &= 0.025 \\ \Pr(\mu - 1.96\sigma < X < \mu + 1.96\sigma) &= 0.95\ . \end{aligned}$$

`R` may also be used to calculate percentage points of the standard normal distribution, using the relationship of Equation (7). In the following `R` session, a list of percentages P has been entered into the vector `P` and, the corresponding values of $1 - P/100$ have been calculated and put into the vector `Pr`. The `qnorm` function has been used to calculate the corresponding values of $\Phi^{-1}(1 - P/100)$ and put them into vector `x(P)`.

- If there are no subcommands for the `qnorm` command then `R` assumes that the standard normal distribution is being referred to.
- In Excel we may use the function `NORMINV` to calculate values of the inverse of a normal c.d.f. or `NORMSINV` for the special case of the standard normal distribution.

```
P <- c(5, 2.5, 1, .5, .1, .05)
Pr <- 1 - P / 100
xP <- qnorm(Pr)
cbind(P, Pr, xP)

##         P     Pr       xP
## [1,] 5.00 0.9500 1.644854
## [2,] 2.50 0.9750 1.959964
## [3,] 1.00 0.9900 2.326348
## [4,] 0.50 0.9950 2.575829
## [5,] 0.10 0.9990 3.090232
## [6,] 0.05 0.9995 3.290527
```

7.5 Control charts

Consider a stream of items coming off a production line, where each item has a continuous variable such as diameter or weight that is being monitored for quality. When the production process is *in control*, the values of the variable measured on each item are assumed to be independently and identically distributed r.v.s, each having an $N(\mu, \sigma^2)$ distribution. Here σ is the standard deviation of the *inherent process variation*, which may have been determined over a long history of observation of the process and which cannot, without a radical overhaul of production methods, be altered.

But from time to time, there may occur some change in the process that alters the distribution of the variable, in particular, its mean or variance. The process is then said to be *out of control*. Such a change may arise from what are sometimes referred to as *assignable causes* — causes that can be identified and remedied. The purpose of quality control is to detect when the process has gone out of control, so that remedial action may be taken.

At regular intervals of time, a random sample of size n is taken from the process, each such sample yielding values $x_1, x_2, \ldots, x_n$, say. To monitor the mean of the process, as successive samples are taken, we record the values of the sample mean $\bar{x}$. If the process is in control then the successive values of $\bar{x}$ are independently and identically distributed r.v.s, where, recalling Equation (5),

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) .$$

So, in the long run, as long as the process is in control, and recalling that $x(2.5) = 1.96$, 95% of the values of $\bar{x}$ should lie in the interval

$$\left(\mu - 1.96\frac{\sigma}{\sqrt{n}}\ ,\ \mu + 1.96\frac{\sigma}{\sqrt{n}}\right)\ ,$$

1 in 40 above the upper limit and 1 in 40 below the lower limit.

Similarly, since $x(0.1) = 3.09$, 99.8% of the values of $\bar{x}$ should lie in the interval

$$\left(\mu - 3.09\frac{\sigma}{\sqrt{n}}\ ,\ \mu + 3.09\frac{\sigma}{\sqrt{n}}\right)\ ,$$

1 in 1000 above the upper limit and 1 in 1000 below the lower limit.

A *Shewhart control chart for means*, or *$\bar{x}$-chart*, is a plot of successive values of $\bar{x}$ on a chart on which are drawn *warning limits* at $\mu \pm 1.96\sigma/\sqrt{n}$ and *action limits* at $\mu \pm 3.09\sigma/\sqrt{n}$.

There is a variety of ways in which an $\bar{x}$-chart may be used for monitoring the process, but the following is a commonly adopted procedure. If a value of $\bar{x}$ falls outside the action limits (an event of probability 1/500 if the process is in control) or if two successive values fall outside the warning limits (an event of probability $(1/20)^2 = 1/400$ if the process is in control) then this is taken as strong evidence that the process is out of control, and action may be initiated to trace and eliminate assignable causes of variation.

A minor modification of the $\bar{x}$-chart, often used, is to have warning limits at $\mu \pm 2\sigma/\sqrt{n}$ and action limits at $\mu \pm 3\sigma/\sqrt{n}$.

For illustration, we have taken a data set from *Clarke and Cooke*, p386, on the diameters (in millimetres) of holes drilled in steel plates. There are 12 samples of size 3. While the process is in control, the process mean is supposed to be 22.5 and the process standard deviation 0.3.

In the control chart the warning limits are at

$$22.5 \pm 2\ \frac{0.3}{\sqrt{3}} \quad \text{i.e.} \quad 22.5 \pm 0.346\ .$$

The action limits are at

$$22.5 \pm 3\ \frac{0.3}{\sqrt{3}} \quad \text{i.e.} \quad 22.5 \pm 0.520\ .$$

The data on the diameters have been entered into a column in a .csv file worksheet, which is of length 12 × 3 = 36. The first column identifies the sample. These data may be plotted on an $\bar{x}$-chart by the `R` commands shown below, followed by the chart that has been produced.

Load and check the data:

```
holes <- read.csv("Holes.csv", header = FALSE)
colnames(holes) <- c("Sample", "Diameter")
head(holes, 7)

##   Sample Diameter
## 1      1     22.7
## 2      1     22.8
## 3      1     23.1
## 4      2     23.4
## 5      2     23.6
## 6      2     23.0
## 7      3     23.1

tail(holes, 7)

##    Sample Diameter
## 30     10     22.3
## 31     11     22.3
## 32     11     22.4
## 33     11     21.9
## 34     12     21.6
## 35     12     21.8
## 36     12     22.2
```

Calculate the means by sample:

```
Xbar <- aggregate(holes[,2], by = list(holes[,1]), mean)
Xbar

##    Group.1        x
## 1        1 22.86667
## 2        2 23.33333
## 3        3 22.80000
## 4        4 22.60000
## 5        5 22.73333
## 6        6 22.30000
## 7        7 22.50000
## 8        8 22.23333
## 9        9 22.10000
## 10      10 22.06667
## 11      11 22.20000
## 12      12 21.86667
```

The mean is 22.5, the standard deviation is $sd = 0.3$, The sample size is $n = 3$. We also store the title for the plot.

```
mean <- 22.5
sd <- 0.3
n <- 3

plot_title <- "Shewhart control chart of diameter"
```

Here we calculate the 2 and 3 “sigma limits” from the target process mean. The “sigma limit” is $\sigma/\sqrt{n}$. We also identify the sample means that are outside the action limits.

```
sample <- Xbar[,1]
xbar <- Xbar[,2]

SL3 <- mean + c(-3, 3) * sd / sqrt(n)
SL2 <- mean + c(-2, 2) * sd / sqrt(n)

outliers <- (xbar > max(SL3)) | (xbar < min(SL3))
sample[outliers]

## [1]  2 12
```

We draw the Shewhart control chart.

```
par(mar = c(4, 4, 4, 7))

plot(sample, xbar, type = "l",
     xlab = "Sample",
     ylab = "Sample Mean",
     main = plot_title,
     ylim = range(xbar, SL3),
     las = 1)

abline(h = mean, col = "green")
abline(h = SL2, col = "red")
abline(h = SL3, col = "red")

points(sample, xbar, pch = 16 + outliers, col = 1 + outliers)

axis(4,
 at = c(SL3, mean, SL2),
 label = paste(c("-3SL =", "+3SL =", "mean =", "-2SL =", "+2SL ="),
              round(c(SL3, mean, SL2), 2), sep = " "),
 las = 1)
```

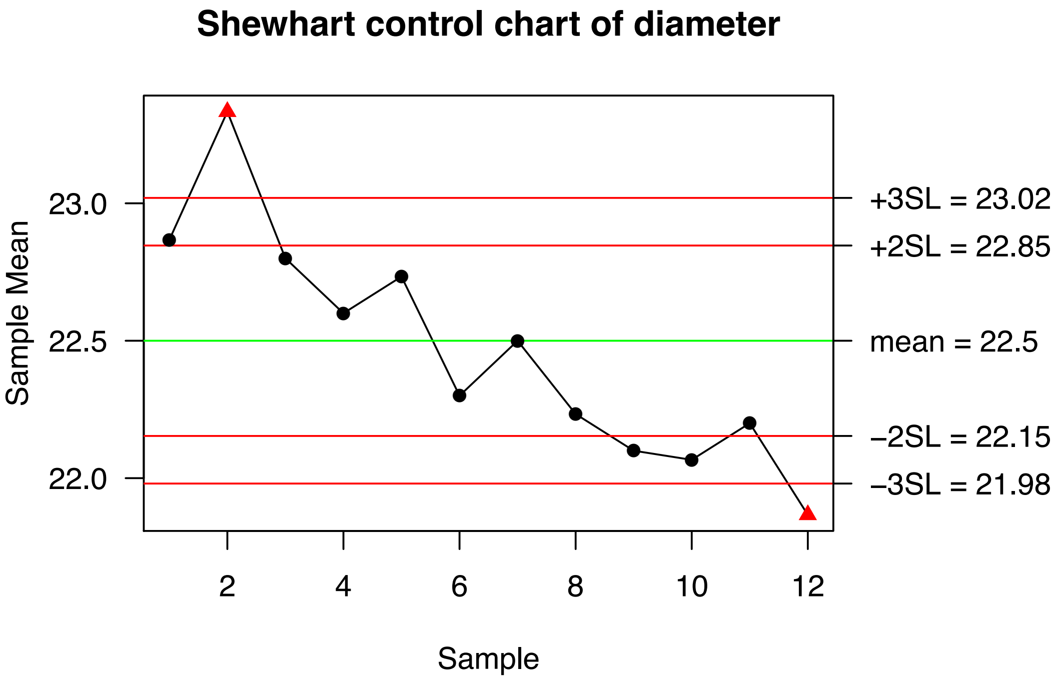

Figure 2: Shewhart control chart for means. Note that the ± 2SL and ± 3SL limits are at the values of the warning and action limits as calculated above.

Looking at Figure 2 we can notice that the first sample mean is outside the warning limits, the second sample mean is outside the action limits, the ninth and tenth sample means are outside the warning limits, and the twelfth sample mean is outside the action limits. So we may suppose that, according to the procedure suggested above, action would have been taken after observation of the second, tenth and twelfth samples.

However, it is also worth remarking that there appears to be a drift over time from higher to lower diameters.