1.6 The distribution of the average

As another example from the doping case, we now consider the mean value of each individual. As it turns out, each individual may have his own mean value and these values can vary considerably between individuals. Say that we introduce a 6 months test period to achieve a higher sensitivity of the antidoping control. In this period, each athlete is tested 5 times, and this average is taken as the estimate of the personal mean and shall be used for setting the limits in the future controls. Say that we know that the values of each individual are normally distributed around his personal mean with variance $\sigma^2 = 5^2$ and that the measurements are mutually independent.

- Let $X_i,\ i = 1, \ldots, n$ be independent, equally distributed random variables. What can we say about the expected value and the variance of their mean? Denote $E(X_i) = \mu$ and $\mathrm{var}(X_i) = \sigma^2$ for each i.
  Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$:

$$\begin{aligned} E[\bar{X}] &= E[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu \\ \mathrm{var}[\bar{X}] &= \mathrm{var}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(X_i) = \frac{\sigma^2}{n} \end{aligned}$$

  Note: The independence was needed to calculate the variance, the expected value of a sum is always the sum of expected values.

- Calculate the limits around the estimated mean, in which an undoped athlete should stay with probability 0.99 at his next measurement.
  Hint: use the result that the sum of independent normal variables is again normal.
  The values of each individual are expressed as $X \sim N(\mu, \sigma^2)$ $(\sigma = 5)$. We are interested in the difference $Z = X_6 - \frac{1}{5}\sum_{i=1}^{5} X_i$. This is the difference of two normal variables with equal mean and the variances equal to σ^2 and σ^2/n respectively. The variable Z is therefore distributed as $Z \sim N(0, \sigma^2 + \sigma^2/n) = N(0, 30)$. With the value $z_{0,005} = 2,57$, the limits equal $\frac{1}{5}\sum_{i=1}^{5} X_i \pm 2,57 \cdot \sqrt{30}$.

- Is the average of i.i.d. random variables always distributed with the same distribution?

No, this is not true in general. A counterexample is the sum of Bernoulli variables, which is binomially distributed, so the average is obviously not a Bernoulli variable.

Understanding the ideas in R:

- Assume that the means of the athletes are normally distributed as $N(148, 7.5^2)$ and generate the personal averages for 100 athletes. Further, use the normal distribution $N(0, 5^2)$ and add it to the personal mean of each athlete to generate 6 values for each individual. Estimate the personal means using the first five values and compare their variance to the theoretical result. Check the distribution of the difference between the 6th value and the estimated mean.

1.7 Conditional expected value and variance

Previous research has shown that the hemoglobin of a cyclist outside the competition phase is distributed as $N(150, 7^2)$, whereas the distribution in the competition phase equals $N(140, 11^2)$. Say that the competition phase lasts for 9 months. We are interested in the expected value and the variance of a randomly taken measurement.
Hint: We are interested in the random variable Y, we know that $\{Y|X = 0\} \sim N(150, 7^2)$ *and* $\{Y|X = 1\} \sim N(140, 11^2)$, $P(X = 1) = 0.75$

- Sketch the distribution of Y and try guessing the expected value and the standard deviation
  The expected value should be between the two values, closer to the competition phase since a measurement is more likely to come from that phase.

- Use the example to explain the formula $E(Y) = E[E(Y|X)]$. Is the expected value $E(Y|X)$ a random variable or a constant? Calculate the expected value of the variable Y.
  $Z = E(Y|X)$ is a random variable that can take two values: $P(Z = 140) = 0.75$, $P(Z = 150) = 0.25$. The expected value of this variable is thus

  $$E(Z) = 140 \cdot P(Z = 140) + 150 \cdot P(Z = 150) = 140 \cdot 0.75 + 150 \cdot 0.25 = 142.5.$$

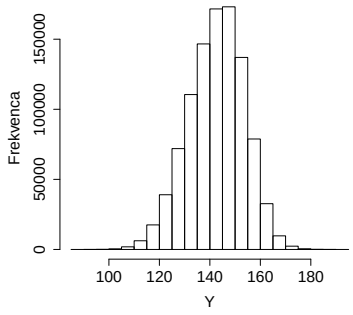

Figure 1: *The distribution of the new variable Y.*

Therefore,

$$E[E(Y|X)] = \sum_x E(Y|X = x) \cdot P(X = x)$$

- Calculate the variance of Y
  We shall use the formula

  $$\text{var}(Y) = \text{var}[E(Y|X)] + E[\text{var}(Y|X)].$$

  The first part of the above formula is the variance of the random variable $Z = E(Y|X)$:

  $$\begin{aligned} \text{var}(Z) &= E([Z - E(Z)]^2) \\ &= 7.5^2 \cdot P[(Z - E(Z)) = 7.5] + 2.5^2 \cdot P[(Z - E(Z)) = 2.5] \\ &= 56.25 \cdot 0.25 + 6.25 \cdot 0.75 = 18.75 \end{aligned}$$

  The standard deviation of the seasonal averages equals 4.33.
  The term $E[\text{var}(Y|X)]$ is the expected value of the variance of Y for a given value of X. We know that $\text{var}(Y|X = 0) = 49$ and $\text{var}(Y|X = 1) = 121$. The expected value equals

  $$E[\text{var}(Y|X)] = 49 \cdot 0.25 + 121 \cdot 0.75 = 103.$$

  Joining the two parts, we get $\text{var}(Y) = 121.75$, $\text{sd}(Y) = 11.03$. variance.

- Find the general expression for the variance ($\{Y|X=0\} \sim N(\mu_0, \sigma_0^2)$, $\{Y|X=1\} \sim N(\mu_1, \sigma_1^2)$, $P(X=1)=p$)
  The value $E(Y|X=0)=\mu_0$ is the expected value of Y when $X=0$, i.e. in the out-of-competition phase. Similarly, $\mu_1 = E(Y|X=1)$ denotes the expected value of Y during the competition phase. We use p to denote the probability of the athlete being in the competition phase. Since X is a Bernoulli distributed variable, we have $E(X) = P(X=1) = p$. The function

$$E(Y|X) = \begin{cases} \mu_0; & X=0 \\ \mu_1; & X=1 \end{cases}$$

  can be written as $E(Y|X) = \mu_0(1-X) + \mu_1 X$. The expected value of the random variable $Z = E(Y|X)$ equals

$$E(Z) = E(E(Y|X)) = \sum_{X=x} E(Y|X=x)P(X=x) = \mu_0(1-p) + \mu_1 p$$

  The variance of the random variable Z equals

$$\begin{aligned} \mathrm{var}(Z) &= \sum_{X=x} \left[E(Y|X=x) - E(Y)\right]^2 P(X=x) \\ &= [\mu_0 - \mu_0(1-p) - \mu_1 p]^2(1-p) + [\mu_1 - \mu_0(1-p) - \mu_1 p]^2 p \\ &= [-p(\mu_0 - \mu_1)]^2(1-p) + [(1-p)(\mu_1 - \mu_0)]^2 p \\ &= [\mu_1 - \mu_0]^2 p^2(1-p) + [\mu_1 - \mu_0]^2(1-p)^2 p \\ &= [\mu_1 - \mu_0]^2 p(1-p)(p+1-p) \\ &= [\mu_1 - \mu_0]^2 p(1-p) \end{aligned}$$

  In the right hand part, the random variable $\mathrm{var}(Y|X)$ equals

$$\mathrm{var}(Y|X) = \begin{cases} \sigma_0^2; & \text{with probability } (1-p) \\ \sigma_1^2; & \text{with probability } p \end{cases}$$

  The variable $\mathrm{var}(Y|X)$ is Bernoulli distributed, its expected value equals $E(\mathrm{var}(Y|X)) = \sigma_0^2(1-p) + \sigma_1^2 p$.
  We join the two parts to get

$$\mathrm{var}(Y) = [\mu_1 - \mu_0]^2 p(1-p) + \sigma_0^2(1-p) + \sigma_1^2 p$$

- Calculate the covariance of X and Y. Find the general expression ($\{Y|X=0\} \sim N(\mu_0, \sigma_0^2)$, $\{Y|X=1\} \sim N(\mu_1, \sigma_1^2)$, $P(X=1)=p$). How does the covariance depend on the parameters? What about the correlation?

$$\begin{aligned}\text{cov}(X,Y) = & \\ &= E[(X-E(X))(Y-E(Y))] \\ &= \int\int (x-E(X))(y-E(Y))f_{X,Y}(x,y)dxdy\end{aligned}$$

We are interested in the expected value with respect to the joint distribution of X and Y (we could denote it as $E_{X,Y}$). We use $f_{X,Y}(x,y) = f_{Y|X}(x,y)f_X(x)$ and start integrating by y

$$\begin{aligned}\text{cov}(X,Y) = & \\ &= \int \left[\int (x-E(X))(y-E(Y))f_{Y|X}(x,y)dy\right] f_X(x)dx \\ &= \int (x-E(X)) \left[\int (y-E(Y))f_{Y|X}(x,y)dy\right] f_X(x)dx\end{aligned}$$

The value $E(Y)$ is a constant in the integral $\int E(Y)f_{Y|X}(x,y)dy$ and can be taken out. The function $f_{Y|X}(x,y)$ represents the conditional density - for each value x, we have a random variable $U = Y|_{X=x}$ with density $f_U(u) = f_{Y|X}(x,y)$. The integral at a given value x thus equals $\int f_{Y|X}(x,y)dy = 1$, and

$$\begin{aligned}\text{cov}(X,Y) = & \\ &= \int (x-E(X)) \left[\int y f_{Y|X}(x,y)dy - E(Y)\right] f_X(x)dx \\ &= \int (x-E(X)) \left[E(Y|x) - E(Y)\right] f_X(x)dx\end{aligned}$$

In our case, X is a discrete covariate, and the integral with respect to X can be replaced by a sum of two terms:

$$\begin{aligned}\text{cov}(X,Y) = & \\ &= (0-E(X))[E(Y|X=0) - E(Y)]P(X=0) + \\ &\quad (1-E(X))[E(Y|X=1) - E(Y)]P(X=1)\end{aligned}$$

The value $E(Y|X=0)=\mu_0$ is the expected value of Y when $X=0$, this is the expected value in the out-of-competition phase, similarly, $\mu_1 = E(Y|X=1)$ denotes the expected value of Y during the competition phase. The probability that the athlete is in the competition phase is denoted by p. X is a Bernoulli random variable, thus $E(X) = P(X=1) = p$. Therefore,

$$\begin{aligned}\mathrm{cov}(X,Y) = & \\ = & -p[\mu_0-\mu](1-p)+(1-p)[\mu_1-\mu]p \\ = & p(1-p)(-\mu_0+\mu_1)\end{aligned}$$

and

$$\mathrm{cor}(X,Y) = \frac{p(1-p)(\mu_1-\mu_0)}{\sqrt{\mathrm{var}Y}\sqrt{p(1-p)}}$$

In our example, $\mathrm{cov}(X,Y) = 0.75{\cdot}0.25{\cdot}(140{-}150) = -1.875$, $\mathrm{cor}(X,Y) = -\frac{1.875}{11.03\sqrt{0.75{\cdot}0.25}} = -0.392$.

Both the covariance and the correlation depend on the difference between the two averages - the larger the difference, the higher the absolute value of covariance and correlation. The two measures have a negative sign if the larger X implies a smaller Y. They also depend on the value of p - the maximum is achieved if $p=0.5$, i.e. when both phases have the same impact on the overall mean. The correlation further also depends on the variability in one and the other phase. If the variability is large compared to the difference between means, the variables do not have a strong correlation.

- What are the values of variance, covariance and correlation if the averages in both phases equal? Are the variables X and Y independent in that case?

  If the average is equal in both phases and hence the difference equal to 0, the variance of Y equals $\mathrm{var}(Y) = \sigma_0^2(1-p)+\sigma_1^2 p$, while both the correlation and the covariance equal 0. Nevertheless, this does not imply that the variables X and Y are independent - the variance of Y depends on the value taken by X. The distribution of Y thus depends on X, even if X does not affect the mean. To conclude - we know that the correlation of two independent variables equals 0, but the reverse is not necessarily true.

Understanding the ideas in R:

- Use R to generate a large number of values and check their distribution

```
> set.seed(1)
> a <- rnorm(1000000,mean=140,sd=11)    #generate values Y|X for comp. phase
> b <- rnorm(1000000,mean=150,sd=7)     #values Y|X for out-of-comp. phase
> x <- sample(0:1,size=1000000,         #generate phase (distribution of X)
+ replace=T,prob=c(0.25,0.75))
> y <- a*x+b*(1-x)                      #random variable Y
> hist(y,prob=T)                        #plot the values
> mean(y)                               #estimate of the mean
> var(y)                                #variance estimate
```

- Try checking each of the above results with R, compare the theoretical values with their estimates.