1.3 Sum of discrete random variables

Let X and Y represent independent Bernoulli distributed random variables $B(p)$.

- Find the distribution of their sum

  Let $Z = X + Y$. The probability $P(Z = z)$ for a given z can be written as a sum of all the possible combinations $X = x$ in $Y = y$, that result in a given z (summation is not a problem, since different values of z cannot happen simultaneously):

$$\begin{aligned} P(Z=z) &= P(X+Y=z) = \sum_y P(X=z-y, Y=y) \\ &= \sum_y P(X=z-y|Y=y)P(Y=y) \end{aligned}$$

  For independent X and Y we get

$$P(Z=z) = \sum_y P(X=z-y, Y=y) = \sum_y P(X=z-y)P(Y=y)$$

  In our example:

$$\begin{aligned} P(Z=0) &= P(X+Y=0) = P(X=0)P(Y=0) = (1-p)^2 \\ P(Z=1) &= P(X=0)P(Y=1) + P(X=1)P(Y=0) \\ &= (1-p)p + p(1-p) = 2p(1-p) \\ P(Z=2) &= P(X=1)P(Y=1) = p^2 \end{aligned}$$

  Therefore

$$P(Z=z) = \binom{2}{z} p^z (1-p)^{2-z}$$

- What is the distribution of the sum of n independent identically distributed (i.i.d.) Bernoulli random variables?

  Let $Z = \sum_{i=1}^{n} X_i$, $i = 1, \ldots, n$, $X_i \sim B(p)$. We have already shown

that the sum of two Bernoulli variables is a Binomial variable with parameters 2 and p. We now use mathematical induction to show that $Z \sim Bin(n, p)$: we need to show that for independent variables $U \sim Bin(n, p)$ and $X \sim Ber(p)$, the sum $U + X$ is Bernoulli distributed with parameters $n + 1$ and p.
Let $1 \leq z \leq n$, then

$$
\begin{aligned}
P(Z = z) = & \\
= & \sum_x P(U = z - x)P(X = x) \\
= & P(U = z)P(X = 0) + P(U = z - 1)P(X = 1) \\
= & \binom{n}{z} p^z (1-p)^{n-z} \cdot (1-p) + \binom{n}{z-1} p^{z-1}(1-p)^{n-z+1} \cdot p \\
= & p^z(1-p)^{n-z+1}\left[\binom{n}{z} + \binom{n}{z-1}\right] \\
= & p^z(1-p)^{n+1-z}\left[\frac{n!}{z!(n-z)!} + \frac{n!}{(z-1)!(n-z+1)!}\right] \\
= & p^z(1-p)^{n+1-z}\left[\frac{n!(n+1)(n-z+1)}{z!(n-z)!(n-z+1)(n+1)} + \frac{n!(n+1)z}{(z-1)!(n-z+1)!(n+1)z}\right] \\
= & p^z(1-p)^{n+1-z}\frac{(n+1)n!}{z(z-1)!(n-z+1)(n-z)!}\left[\frac{n-z+1}{n+1} + \frac{z}{n+1}\right] \\
= & p^z(1-p)^{n+1-z}\binom{n+1}{z}\frac{n-z+1+z}{n+1} \\
= & \binom{n+1}{z} p^z(1-p)^{n+1-z}
\end{aligned}
$$

The proof for $z = 0$ and $z = n + 1$ is left for the reader.

Understanding the ideas in R:

- Use the function `sample` to generate 100 realizations of two Bernoulli variables and check the distribution of their sum.

1.4 Sum of continuous random variables

While individual values give some indication of blood manipulations, it would be interesting to also check a sequence of values through the whole season. We wish to look at the distribution of the sum of squared standardized departures from the mean value (under the null hypothesis that the athlete is not doped). Let Z denote the standardized difference from the mean value (we assume it is normally distributed), we wish to know the sum $\sum Z^2$ (we square the values, since departures in both directions are of interest). We assume that the measurements were made in intervals long enough to ensure independence between them.

- Find the formula for the distribution of the sum of two independent continuous variables ($Z = X + Y$), compare it with the formula in the discrete case

  We write the cumulative distribution function as the integral:

$$\begin{aligned} P(Z \leq z) &= P(X+Y \leq z) = \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{z-y} f_{X,Y}(x,y)dxdy \\ &= \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{z} f_{X,Y}(v-y,y)dvdy \end{aligned}$$

  where the substitution $x = v - y$ was made. We now interchange the order of integration and find the derivative (we assume that the outer integral is continuous in z):

$$\begin{aligned} P(Z \leq z) &= \int\limits_{-\infty}^{z} \int\limits_{-\infty}^{\infty} f_{X,Y}(v-y,y)dydv \\ f_Z(z) &= \int\limits_{-\infty}^{\infty} f_{X,Y}(z-y,y)dy \\ f_Z(z) &= \int\limits_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy \end{aligned}$$

  (the last line is true if X and Y are independent). The result is analogous to the discrete version.

- Find the distribution of the sum $S = Z_1^2 + Z_2^2$, if Z_1 and Z_2 are standard normal variables?

Hint: Use the equality

$$\int_0^1 \frac{1}{\sqrt{(1-x)x}} dx = \pi$$

We've already shown that $Z^2 \sim \chi_1^2$ ($Z^2 \sim \Gamma(\frac{1}{2}, \frac{1}{2})$), therefore

$$f_{Z^2}(z) = \frac{1}{\sqrt{2\pi z}} \exp\left\{-\frac{z}{2}\right\};\ z > 0$$

We express the density of the sum $Z_1^2 + Z_2^2$. For a given s, the value z_2 must be between 0 (z_1 and z_2 cannot be negative) and s, this defines the limits of integration.

$$\begin{aligned} f_S(s) &= \int_0^s f_{Z_1^2}(s - z_2) f_{Z_2^2}(z_2) dz_2 \\ &= \frac{1}{2\pi} \int_0^s \frac{1}{\sqrt{s - z_2}} \exp\left\{-\frac{s - z_2}{2}\right\} \frac{1}{\sqrt{z_2}} \exp\left\{-\frac{z_2}{2}\right\} dz_2 \\ &= \frac{1}{2\pi} \exp\left\{-\frac{s}{2}\right\} \int_0^s \frac{1}{\sqrt{s - z_2}} \frac{1}{\sqrt{z_2}} dz_2 \end{aligned}$$

We use the substitution $z_2 = sv$, $dz_2 = sdv$, the limits are between 0 and 1:

$$\begin{aligned} f_S(s) &= \frac{1}{2\pi} \exp\left\{-\frac{s}{2}\right\} \int_0^1 \frac{1}{\sqrt{s - sv}} \frac{1}{\sqrt{sv}} sdv \\ &= \frac{1}{2\pi} \exp\left\{-\frac{s}{2}\right\} \int_0^1 \frac{1}{\sqrt{1 - v}} \frac{1}{\sqrt{v}} dv \\ &= \frac{1}{2\pi} \exp\left\{-\frac{s}{2}\right\} \pi \\ &= \frac{1}{2} \exp\left\{-\frac{s}{2}\right\} \end{aligned}$$

The gamma distribution density equals

$$f_X(x) = \frac{\lambda^a x^{a-1} e^{-\lambda x}}{\Gamma(a)};\ x > 0, \lambda > 0, a > 0$$

Our result is the gamma distribution with parameters $\lambda = \frac{1}{2}$ and $a = 1$.

- Say that we get the following five standardized values for a certain athlete (Z values): $1.6, 1.5, -1.6, 1.8, 1.4$. What can we conclude about the athlete?

  We use the result that the sum of n i.i.d. variables with $X_i \sim \Gamma(\frac{1}{2}, \frac{1}{2})$ is distributed as $\sum_{i=1}^{n} X_i \sim \Gamma(\frac{n}{2}, \frac{1}{2})$.

  ```
  > vr <- c(1.6,1.5,-1.6, 1.8, 1.4)
  > rez <- sum(vr^2)
  > rez
  [1] 12.57
  > 1-pgamma(rez, 2.5, 0.5)
  [1] 0.02775943
  ```

  Our null hypothesis is that an athlete is not guilty. Under this hypothesis, the sum of squared standardized departures from the mean is distributed as $\Gamma(\frac{5}{2}, \frac{1}{2})$. The probability of getting the value $12,57$ or more is approximately 0.03.

Understanding the ideas in R:

- Generate 10 values from the $X \sim N(148, 85)$ distribution, these values represent 10 doping test in one athlete. Standardize the values to get a $N(0, 1)$ variable, square them and sum. This is the value of the first athlete, generate values for 1000 athletes in the same way. Plot a histogram of the values. Use the function `pgamma` to find the limit that is exceeded by the $\Gamma(\frac{10}{2}, \frac{1}{2})$ distribution with probability less than 0.01. Calculate the proportion of the athletes in your sample that exceed this limit.

- Say that a doped athlete has the same average, but a larger variance (the values vary more due to blood manipulation). Generate the values for 1000 athletes with a larger variance and check the proportion that exceeds the limits.

1.5 Sum of normal variables

As we have already shown, the linear transformation of a normal variable remains normally distributed. In this exercise, we shall show that the sum of two independent (standard) normal distributions in again normally distributed, but that this may not be true for two dependent normal variables.

- Show that a sum of two independent standardized normal variables is a normally distributed random variable, find its mean and standard deviation.
  Let $X \sim N(0,1)$ and $Y \sim N(0,1)$, using the formula for the density of the sum of two independent random variables, we get

$$\begin{aligned} f_U(u) &= \int_{-\infty}^{\infty} f_X(u-y) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(u-y)^2}{2}\right\} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} dy \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left\{-\frac{u^2}{2}\right\} \exp\left\{-y^2\right\} \exp\left\{uy\right\} dy \end{aligned}$$

We now try to express the above density as a normal density - we write the parts containing y as a square of a sum:

$$\begin{aligned} y^2 - uy &= y^2 - 2y\frac{u}{2} + \left(\frac{u}{2}\right)^2 - \left(\frac{u}{2}\right)^2 \\ &= \left(y - \frac{u}{2}\right)^2 - \frac{u^2}{4} \end{aligned}$$

The above integral can be rewritten as

$$\begin{aligned} f_U(u) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left\{-\frac{u^2}{2}\right\} \exp\left\{-\left(y - \frac{u}{2}\right)^2\right\} \exp\left\{\frac{u^2}{4}\right\} dy \\ &= \frac{1}{2\pi} \exp\left\{-\frac{u^2}{4}\right\} \int_{-\infty}^{\infty} \exp\left\{-\frac{(y-\frac{u}{2})^2}{2\cdot\frac{1}{2}}\right\} dy \\ &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{4}\right\} \frac{1}{\sqrt{2}} \left[\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\frac{1}{2}}} \exp\left\{-\frac{(y-\frac{u}{2})^2}{2\cdot\frac{1}{2}}\right\} dy\right] \end{aligned}$$

The expression under the integral is a density of a normally distributed variable ($N(\frac{u}{2}, (\sqrt{\frac{1}{2}})^2)$), the value of the integral is therefore equal to 1 (regardless of the value of u, which is a constant within this interval). The variable U is normally distributed, $U \sim N(0, (\sqrt{2})^2)$.

- Let the variable Z equal $|Y|$ if $X \geq 0$, and $Z = -|Y|$ if $X < 0$. Find the distribution of Z?
  First plot the simulated values with R:

```
> set.seed(1)
> x <- rnorm(1000,0,1)          #1000 realizacij normalne spr.,
> y <- rnorm(1000,0,1)          #povprecje=0, sd=1
> z <- abs(y)                   #z = |y|
> z[x<0] <- -z[x<0]             #z =-|y|, ce je x<0
> hist(z,main="",ylab="Frekvenca") #histogram z
```

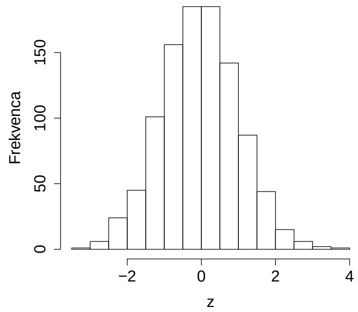

Figure 1: *The Z variable distribution.*

We now express the cumulative distribution function theoretically. Consider the case $z < 0$ (i.e. X negative) and use the fact that X and Y are independent:

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(X < 0, -|Y| \leq z) \\ &= P(X < 0)[P(Y \leq z) + P(Y \geq -z)] \\ &= \frac{1}{2}[2 \cdot P(Y \leq z)] \\ &= P(Y \leq z) = F_Y(z) \end{aligned}$$

Analogously, we get $P(0 \leq Z \leq z) = P(0 \leq Y \leq y)$ for $z > 0$. We have shown that the distribution of Z equals that of Y, i.e. Z is a standardized normal variable.

- Sketch the joint distribution of X and Z. Are the two variables independent?

```
> plot(x,y)
> plot(x,z)
```

It is clear that the variables are not independent, their sign is always equal.

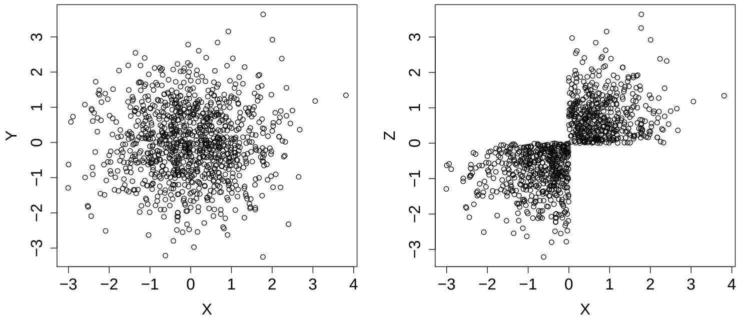

Figure 2: *Scatter plots of the variables.*

- Is the sum $X + Z$ normally distributed?

We see that the sum is not normal - if the variables are dependent, we cannot expect the sum to be normal (we've just found a counterexample).

```
> hist(x+z,main="",ylab="Frekvenca")
```

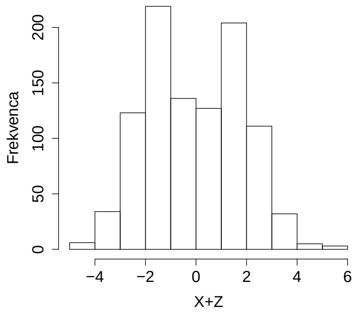

Figure 3: *The distribution of the sum $X + Z$.*