Lecture 41: Last time we defined, for r.v.'s $X$ and $Y$,

the correlation: \( \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{SD}(X) \text{SD}(Y)} \)

and proved that: \(-1 \leq \text{Cov}(X, Y) \leq 1\).

**Def:** • $X$ and $Y$ are "uncorrelated" if $\text{Corr}(X, Y) = 0$.

**Note:** \( \text{Corr}(X,Y) = 0 \quad \text{Cov}(X,Y) = 0 \quad \text{E}(XY) = \text{E}(X)\text{E}(Y) \)

• $X$ and $Y$ are "completely correlated" if either $\text{Corr}(X, Y) = 1$ or $\text{Corr}(X, Y) = -1$

• $X$ and $Y$ are "positively correlated" if $\text{Corr}(X, Y) > 0$  
  **E.g.:** $X =$ height of a person  
  $Y =$ weight of a person

• $X$ and $Y$ are "negatively correlated" if $\text{Corr}(X, Y) < 0$  
  **E.g.:** $X =$ median income in a neighborhood  
  $Y =$ crime rate in a neighborhood

**Theorem:** let $X$ and $Y$ be random variables (with: \( \text{SD}(X) > 0, \text{SD}(Y) > 0 \))

\[ X, Y \text{ are completely correlated } \iff Y = aX + b \text{ , for some } a \neq 0, b \in \mathbb{R}. \]

**Remark:** the above theorem says that $X$ and $Y$ are completely correlated if and only if one is completely determined by the other, i.e. if and only if they are "completely dependent".  
On the other hand, if $X$ & $Y$ are independent then they're uncorrelated.
Proof: \( \Rightarrow \): from the proof of last time's theorem (formula (\( \cdot \cdot \cdot \))),
if \( E(X Y) = 1 \) then \( E[(X - Y)^2] = 0 \), therefore see note
\( P(X - Y = 0) = 1 \). So, with probability 1,
\[
\frac{X - \mu_X}{\sigma_X} = \frac{Y - \mu_Y}{\sigma_Y}
\]
i.e., \( X = \frac{\sigma_X}{\sigma_Y} Y - \frac{\sigma_X}{\sigma_Y} \mu_Y + \mu_X = aY + b \). Similarly, if
\( E(X Y) = -1 \) then by \( (\cdot \cdot \cdot ) \) of the same proof we
have \( E[(X + Y)^2] = 0 \), therefore \( P(X + Y = 0) = 1 \).
So, with probability 1,
\[
\frac{X - \mu_X}{\sigma_X} = -\frac{Y - \mu_Y}{\sigma_Y}
\]
i.e., \( X = -\frac{\sigma_X}{\sigma_Y} Y + \frac{\sigma_X}{\sigma_Y} \mu_Y + \mu_X = aY + b \).

\( \Leftarrow \): if \( Y = aX + b \) then:
\[
\text{Cov}(X, Y) = E(XY) - E(X)E(Y)
\]
\[
= E[X(aX + b)] - E(X)E(aX + b) = \frac{\sigma^2}{\sigma_Y} \mu_Y + \mu_X = a \text{Var}(X)
\]
Also: \( \text{SD}(Y) = |a| \text{SD}(X) \), therefore:
\[
\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{SD}(X) \text{SD}(Y)} = \frac{a \text{Var}(X)}{|a| \text{SD}(X)^2} = \frac{a}{|a|}.
\]
\[
= \text{sign}(a) = \begin{cases} 1 & \text{if } a > 0 \\ -1 & \text{if } a < 0 \end{cases}
\]

(x) This follows from Chebyshev's inequality: for any \( \alpha \), arbitrary
\[
P(|Z - \mu_Z| \geq \alpha \text{SD}(Z)) \leq \frac{1}{\alpha^2}; \quad \text{choose } \alpha = \frac{\varepsilon}{\text{SD}(Z)}, \text{ and}
\]
\[
P(|Z - \mu_Z| \geq \varepsilon) \leq \frac{\text{SD}(Z)}{\varepsilon^2}. \quad \text{If } \mu_Z = 0 \text{ and } E(Z^2) = 0,
\]
then \( \text{SD}(Z) = \sqrt{\text{Var}(Z)} = \sqrt{E(Z^2) - \mu_Z^2} = 0 \), so, for all \( \varepsilon \),
\[
P(|Z| \geq \varepsilon) = 0, \quad \text{i.e. } Z = 0 \text{ with probability 1.}
Example: play Roulette (typically 38 numbers; 18 red, 18 black, 2 green (0 & 00)).

- \( r \) = proportion of red numbers (typically \( \frac{18}{38} \))
- \( b \) = "black" (typically \( \frac{18}{38} \))
- \( n \) = # of spins (fixed)
- \( N_R \) = # of times I get a red number
- \( N_B \) = # of times I get a black number

\[ \text{Corr}(N_R, N_B) = ? \]

- \( N_R \sim \text{Binomial} \left( n, r \right) \Rightarrow E(N_R) = nr, \ SD(N_R) = \sqrt{nr(1-r)} \)
- \( N_B \sim \text{Binomial} \left( n, b \right) \Rightarrow E(N_B) = nb, \ SD(N_B) = \sqrt{nb(1-b)} \)

- To compute \( \text{Cov}(N_R, N_B) \), note that

\[
\text{Var}(N_R + N_B) = \frac{\text{Var}(N_R)}{nr(1-r)} + \frac{\text{Var}(N_B)}{nb(1-b)} + 2 \text{Cov}(N_R, N_B)
\]

\[ N_R + N_B = \text{number of spins which are either red or black} \]

\[ \sim \text{Binomial} \left( n, r+b \right) \]

\[ \Rightarrow \text{Var}(N_R + N_B) = n(r+b)(1-r-b). \text{ So:} \]

\[
\text{Cov}(N_R, N_B) = \frac{1}{2} \left\{ n(r+b)(1-r-b) - nr(1-r) - nb(1-b) \right\}
\]

\[ = \frac{1}{2} \left\{ nr + nb - nr^2 - nr b - nb^2 - nr b - nr + nr^2 - nb + nb^2 \right\} = -nr b
\]

\[ \Rightarrow \text{Corr}(N_R, N_B) = -\frac{nr b}{\sqrt{nr(1-r)} \sqrt{nb(1-b)}} = -\sqrt{\frac{r}{1-r} \frac{b}{1-b}} < 0
\]

- **Negative!** If \( N_R \) is large, \( N_B \) is small. (independent of \( n \)).

- When \( r = 1-b \) (no green numbers): \( \text{Corr}(N_R, N_B) = -1 \)
  - In this case \( N_R = n - N_B \), i.e. \( N_R = \alpha N_B + \beta \) with \( \alpha = -1 \), \( \beta = n \).
- When \( r = b = \frac{18}{38} \), we have \( \text{Cov}(N_R, N_B) = -0.9 \).
Remark: positive correlation does not imply causation!
I.e. one variable does not necessarily influence the
other directly. More likely, there is an unknown
factor that influences both variables similarly.

For example, the r.v.'s:
X = height of a child
Y = proficiency in math of a child
have Cov(X,Y) > 0. That is not because height
has a direct influence on math proficiency, but
because if a child is older then typically
he's both taller and has had more years of math.
(unknown factor = age).

Another example:
X = CO₂ levels in a city (pollution levels)
Y = average weight in a city

They are positively correlated! Not because
breathing CO₂ makes you fat, but because
the richer a city is, the more likely it is
to be polluted and people are more likely to
eat excessively.

A funny example:
X = Math PhDs awarded a certain year in the US
Y = Suicides by hanging, strangulation,
or suffocation in a year in the US.

Can you think of a possible unknown factor
that influences both? 😄
(Did not cover the following 2 pages in class).

A few words on the last section: **Bivariate normal density.**

Suppose first that $X \sim \text{Normal}(0, 1)$ and $Y \sim \text{Normal}(0, 1)$ are independent. Therefore $\text{Corr}(X, Y) = 0$.

We have $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$, $f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$, so by independence the joint density is given by

\[ (*) \quad z = f(x, y) = f_X(x) f_Y(y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}. \]

We already saw that the level curves of $f(x, y)$ are circles. I.e. if we fix $z = k$, with $k < \frac{1}{2\pi}$, we get $e^{-\frac{1}{2}(x^2+y^2)} = k < 1$.

\[ \Rightarrow -\frac{1}{2}(x^2+y^2) = \ln(2\pi k) \Rightarrow x^2 + y^2 = -2 \ln(2\pi k) > 0 \]

i.e. $x^2 + y^2 = R^2$, circles with radius $R = \sqrt{-2 \ln(2\pi k)}$.

What if $\text{Corr}(X, Y) \neq 0$?

**Theorem:** if $X \sim \text{Normal}(0, 1)$, $Y \sim \text{Normal}(0, 1)$ and $\text{Corr}(X, Y) = \rho$,

\[ (***) \quad f(x, y) = \frac{1}{2\pi \sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2} \frac{x^2-2\rho xy+y^2}{1-\rho^2} \right\} \]

called the **bivariate normal density**.

**Remark:** if $\rho = 0$, we get back $(**)$.
What are the level curves of this? If you fix $0 < k \leq \frac{1}{2\pi \sqrt{1 - q^2}}$ and set $f(x,y) = k$, we get

$$\exp \left\{ -\frac{1}{2} \cdot \frac{x^2 - 2qxy + y^2}{1 - q^2} \right\} = 2\pi k \sqrt{1 - q^2}$$

$$\Rightarrow x^2 - 2qxy + y^2 = -2(1-q^2) \ln(2\pi k \sqrt{1 - q^2})$$

These are equations of ellipses! To see this, rotate the axes by $\frac{\pi}{4}$:

$$\begin{cases}
    x = \frac{\sqrt{2}}{2} u - \frac{\sqrt{2}}{2} v \\
    y = \frac{\sqrt{2}}{2} u + \frac{\sqrt{2}}{2} v
\end{cases}$$

Insert these in $x^2 - 2qxy + y^2 = R^2$ and get:

$$\frac{1}{2} u^2 - uv + \frac{1}{2} v^2 - 2q \left( \frac{1}{2} u^2 - \frac{1}{2} v^2 \right) + \frac{1}{2} u^2 + uv + \frac{1}{2} v^2 = R^2$$

$$u^2 + v^2 - 2qu^2 + 2qv^2 = R^2$$

$$(1-q)u^2 + (1+q)v^2 = R^2$$

$$\frac{u^2}{\frac{R^2}{1-q}} + \frac{v^2}{\frac{R^2}{1+q}} = 1$$

Which is an ellipse with semi-axes:

- along $u$: $a = \frac{R}{\sqrt{1-q}}$ \quad so if $q > 0 \Rightarrow a > b$
- along $v$: $b = \frac{R}{\sqrt{1+q}}$ \quad if $q < 0 \Rightarrow a < b$

The End