STOR 435.001 Lecture 14

Jointly distributed Random Variables - II

Jan Hannig

UNC Chapel Hill

⚠

Discrete case: Independence is equivalent to

$$p(x, y) = p_X(x)p_Y(y), \quad \text{all } x, y,$$

where $p(x, y)$ is the joint p.m.f. of X and Y, p_X is the p.m.f. of X, p_Y is the p.m.f. of Y.

Jointly continuous case: Independence is equivalent to

$$f(x, y) = f_X(x)f_Y(y), \quad \text{all } x, y,$$

where $f(x, y)$ is the joint p.d.f. of X and Y, f_X is the p.d.f. of X, f_Y is the p.d.f. of Y.

Problem from Final Exam 2011

Suppose that 2 balls are chosen without replacement from an urn containing 5 white and 8 red balls. Let X_1 equal 1 if the 1st ball selected is red, and let it equal 0 otherwise. Let X_2 equal 1 if the 2nd ball selected is red, and let it equal 0 otherwise.

1. Give the joint probability mass function of X_1 and X_2.
2. Are X_1 and X_2 independent? (Provide a mathematical argument.)

Example: Suppose that the number of people who enter Starbucks on a given day is a Poisson random variable with parameter λ. Each person who enters the post office orders a single Latte with probability p and orders something else with probability $1-p$. Let L and NL be the total number of single order Latte and Non single order Lattes in a day. Show that these are independent Poisson random variables with respective parameters λp and $\lambda(1-p)$.

☼

Problem: Two points are selected randomly on a line of length L so as to be on opposite sides of the midpoint of the line. (In other words, the two points X and Y are independent random variables such that X is uniformly distributed over $(0, L/2)$ and Y is uniformly distributed over $(L/2, L)$.) Find the probability that the distance between the two points is greater than $L/3$.

☼

Problem continued

Example: Let $X = U(0,1)$ and $Y = Exp(1)$ be independent. Find cdf and pdf of $Z = X + Y$.

☼

Problem continued

Independence: Random variables $X_1, X_2, \ldots, X_n$ are *independent* if, for any sets $A_1, A_2, \ldots, A_n$,

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n)$$

$$= P(X_1 \in A_1)P(X_2 \in A_2) \ldots P(X_n \in A_n)$$

(that is, the events $\{X_1 \in A_1\}, \{X_2 \in A_2\}, \ldots, \{X_n \in A_n\}$ are independent). For example, in the jointly continuous case, this is equivalent to

$$f(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \ldots f_{X_n}(x_n), \quad \text{all } x_1, x_2, \ldots, x_n,$$

where $f(x_1, x_2, \ldots, x_n)$ is the joint p.d.f. and $f_{X_i}(x_i)$ is a marginal p.d.f. of X_i.

An important operation with combining independent random variables

- Now suppose we had a bunch of independent random variables say $X_1, X_2, \dots X_n$.
- One important way to combine these is *adding* them i.e. making a new random variable $S_n = X_1 + X_2 + \cdots X_n$.
- Why important?
- **Setting one:** We take a random sample of 100 UNC voters. Let X_i be a 1 if i-th individual in sample is voting for Hillary and $X_i = 0$ otherwise. So now we have $X_1, X_2, \dots, X_{100}$. The total number of people in your sample who are voting for Hillary is?
- **Setting two:** Researchers are comparing study habits in two universities (call these U and D) and in particular want to study the average amount of time juniors spend studying per day. They take a sample of 40 students from U and 30 from D and see how much time they spend studying in a week. Let $X_1, X_2, \dots X_{40}$ denote the respective amount of time for the students at U and $Y_1, Y_2, \dots Y_{30}$ denote the same but for students at D. What would they base the conclusions of their study on?
- **Punchline:** sums of random variables super important! I will show you two methods for calculting the distribution of such objects.
- **Method 1: Direct method**
- **Method 2: Moment generating function**

Sums of independent random variables: E.g. jointly continuous case: if X and Y are independent, X has density f_X and Y has density f_Y, then the density of $X + Y$ is:

$f_{X+Y}(a) =$

Name: *convolution* of f_X and f_Y.

Example 3a: If X and Y are independent random variables, both uniformly distributed on $(0, 1)$, calculate the probability density of $X + Y$.

☼

R code to simulate this

```
x<- runif(10^5)
y<- runif(10^5)
z<- x+y
hist(z)
```

Note: For independent discrete random variables we could carry out similar calculations but **not** integrating but "summing":
Example: If $X = Pois(\lambda_1)$ and $Y = Pois(\lambda_2)$ are independent, then

$P(X + Y = n) =$

☼

Method II for finding distributions of sums of independent random variables

- The previous slides we directly calculated the distributions of sums. Turns out there is another powerful tool to do the same. The tool uses two ingredients.
- **Ingredient 1: Moment generating functions** $M_X(t) := E(e^{tX})$.
- **Amazing math fact:** Moment generating functions uniquely characterize the distribution. So suppose we do not know the distribution of a random variable but have somehow managed to compute the MGF and recognize it as the MGF of a known distribution, this means that the random variable has that distribution.

We have computed earlier:

MGF of some standard distributions

1. $X = Bin(n, p)$
$$M_X(t) = (1 - p + pe^t)^n$$
2. $X = Pois(\lambda)$
$$M_X(t) = e^{\lambda(e^t - 1)}.$$
3. $X = N(\mu, \sigma^2)$
$$M_X(t) = e^{t\mu + t^2\sigma^2/2}$$
4. $X =$ Gamma with parameters (α, λ) then $M_X(t)$ is finite only when $t < \lambda$ and then
$$M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^\alpha$$

Ingredient 2

Suppose X and Y are **independent** random variables with marginal pdf f_X and f_Y. Let $g(\cdot), h(\cdot)$ be functions. Show that

$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$$

The same fact holds for discrete independent random variables and not just 2 but any number n independent random variables.

☼

Jointly distributed random variables

Combining the two ingredients

Suppose $X_1, X_2, \dots X_n$ are independent random variables with the mgf of X_i given by M_{X_i}. Let S be the sum of these random variables namely $S = X_1 + X_2 + \cdots X_n$. Then MGF of S

$$\begin{aligned} M_S(t) = E(e^{tS}) = E(e^{t(X_1+X_2+\cdots X_n)}) &= E(e^{tX_1+tX_2+\cdots+tX_n}) \\ &= E(e^{tX_1}e^{tX_2}\cdots e^{tX_n}) \\ &= E(e^{tX_1})E(e^{tX_2})\cdots E(e^{tX_2}) \quad \textbf{by independence} \\ &= M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t). \end{aligned}$$

Punchline: MGF of the **sum** of a bunch of **independent** random variables is the **product** of the individual MGFs.

Method II of computing distribution of sums of independent rvs

1. Compute the individual MGfs $M_{X_i}(t)$.
2. Compute the MGF of the sum S using the above formula.
3. See if M_S matches the MGF of one of the known distributions. If it does then S has that distribution.

Let $0 < p < 1$. If $X_i, i = 1, 2, \ldots, m$ are independent with $X_i = \text{Bin}(n_i, p)$ what is the distribution of $\sum_1^m X_i$? **Note: Same p for all the rvs.**

⚠

Let $\lambda_1, \lambda_2, \ldots \lambda_m > 0$. If $X_i, i = 1, 2, \ldots, m$ are independent Poisson λ_i what is the distribution of $\sum_{i=1}^{n} X_i$?

☼

If $X_i, i = 1, 2, \ldots, n$ are independent and identically distributed exponential random variables with parameter λ, what is the distribution of $\sum_{j=1}^{n} X_i$?

⚠

If X_i are independent normal random variables with parameters μ_i, σ_i^2, what is the distribution of $\sum_{i=1}^{n} X_i$?

☼

Jointly distributed random variables

Problem: Motivated by
`https://www.tkiryl.com/Elementarytatistics/Chapter_9.pdf`.

The following is taken almost verbatim from the above site. One of the larger species of **tarantulas** is the Grammostola mollicoma, whose common name is the Brazilian giant tawny red. A tarantula has two body parts. The anterior part of the body is covered above by a shell, or carapace. From a recent article by F. Costa and F. Perez-Miles titled Reproductive Biology of Uruguayan Theraphosids (The Journal of Arachnology, Vol. 30, No. 3, pp. 571-587):
The carapace length of the adult male G. mollicoma is normally distributed with mean 18.14 mm and standard deviation 1.76 mm.

Suppose you find three such males and measure them. Assume the lengths are independent random variables. Find the probability that the **average** length of the carpace is greater than 19 mm.

☼

Solution continued.

Conclusion

- We thought about operations on independent random variables.
- Focussed on sums of independent random variables.
- Found that in a number of cases the sum of such random variables can be explicitly evaluated.
- Two methods. Method 1: Direct calculation of the pmf or pdf of the sum.
- Method 2: Using moment generating functions.