Lecture 22: Transformations and Convolutions

Var of Hypergeometry (w, b, n), $p = \frac{w}{w+b}$, $w+b = N$ → population size, not a random variable

$$\mathrm{Var}\left(\sum_{j=1}^{n} X_j\right) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \dots + \mathrm{Var}(X_n) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j)$$

Unconditional variances

By symmetry, they are all the same

$$= n\,\mathrm{Var}(X_1) + 2\binom{n}{2}\mathrm{Cov}(X_1, X_2)$$

X_1 is just a Bernoulli p, so, $\mathrm{Var}(X_1)$ is $p(1-p)$

$$= np(1-p) + 2\binom{n}{2}\left[E(X_1X_2) - E(X_1)E(X_2)\right]$$

(Any of these X_j, the jth ball, before we draw any balls, that's just equally likely to be any of the balls.)

$$= np(1-p) + 2\binom{n}{2}\left(\frac{w}{w+b}\cdot\frac{w-1}{w+b-1} - p^2\right)$$

$$= \frac{N-n}{N-1}\, np(1-p)$$

$\frac{N-n}{N-1}$ → finite population correction

Let's pick $n = 1$,

$$\mathrm{Var}\left(\sum_{j=1}^{n} X_j\right) = p(1-p)$$ → variance of a Bernoulli p

If we are picking one ball, what difference does it make if it's with replacement or without replacement, there is only one ball.

If N is much much larger than n, then $\frac{N-n}{N-1} \approx 1$

$$\mathrm{Var}\left(\sum_{j=1}^{n} X_j\right) = np(1-p)$$ → binomial variance

If the sample is so minuscule compared to the population it's very very unlikely that we would sample the same individual more than once. We are not doing replacement, but what difference does it make, because it's unlikely to get the same person twice anyway, in ~~your~~ our sample.

Transformations

A function of a random variable is a random variable and we use LOTUS to get expected value but LOTUS only gives us the expected value of that transformed random variable. It doesnot give us the whole distribution. A lot of times, we don't just want the mean or the variance, we want the entire distribution.

Theorem: Let X be a continuous r.v. with PDF f_X, $Y = g(X)$, where g is differentiable, strictly increasing.

Then, the PDF of Y is given by

in case of strictly decreasing g, $f_Y(y) = f_X(x) \left|\frac{dx}{dy}\right|$

$$f_Y(y) = f_X(x) \frac{dx}{dy}, \text{ where } y = g(x),\ x = g^{-1}(y)$$

and this is written in terms of y.

Also, $\frac{dx}{dy} = \left(\frac{dy}{dx}\right)^{-1}$. choose the one which is easier to do.

proof: CDF of Y is $P(Y \leq y) = P(g(X) \leq y)$: assuming g has an inverse

$$= P(X \leq g^{-1}(y))$$

$$= F_X(g^{-1}(y))$$

CDF of X evaluated at $g^{-1}(y)$

$$= F_X(x) \quad x = g^{-1}(y)$$

$$f_X(x) = \frac{d F_X(x)}{dx}$$

$$\Rightarrow f_Y(y) = f_X(x)\, dx/dy \text{ (Chain Rule)}$$

Example (Log Normal distribution)

does not mean log of a normal
we can't take log of a negative value
It means that log is normal, not log of the normal.

$Y = e^Z, \; Z \sim N(0,1)$

differentiable, ~~incre~~
strictly increasing

$Y = e^Z$
$\log Y = \log e^Z$
$= Z$
which is normal

$\boxed{Z = \log Y}$

~~$f_Y(y) =$~~

$\frac{dy}{dz} = e^z = y$ (in terms of y)

$f_Z(y) = \underbrace{\frac{1}{\sqrt{2\pi}} e^{-(\ln y)^2/2}}_{f_Z(z) \text{ in terms of } y}$

$\therefore f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-(\ln y)^2/2} \cdot 1/y \quad , y > 0$

Transformations in $\mathbb{R}^n$ (multidimensional version)

$\vec{Y} = g(\vec{X}), \; g: \mathbb{R}^n \to \mathbb{R}^n$

Assume $\vec{X} = (X_1, \ldots, X_n)$ is continuous.

We want Joint PDF of $\vec{Y}$ in terms of Joint PDF of $\vec{X}$.

~~So~~ Joint PDF of $\vec{Y}$ is

$$f_{\vec{Y}}(\vec{y}) = f_{\vec{X}}(\vec{x}) \left| \frac{d\vec{x}}{d\vec{y}} \right|$$

Jacobian (matrix of all possible partial derivatives)

take absolute value of determinant of Jacobian matrix

$$\frac{d\vec{x}}{d\vec{y}} = \begin{pmatrix} \partial x_1/\partial y_1 & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & & & \\ \partial x_n/\partial y_1 & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix}$$

$$\left|\frac{d\vec{y}}{d\vec{x}}\right|^{-1} = \left|\frac{d\vec{x}}{d\vec{y}}\right|$$

Convolution (Sums) ~~We wan~~

We want the distribution of sum of random variables.

Let $T = X + Y$. We want to know the distribution of T, assuming we know the distribution of X and Y.

Discrete Case

X, Y independent

$$P(T=t) = \sum_x P(X=x)\,P(Y=t-x)$$

$$P(X=x, Y=t-x) = P(X=x)\cdot P(Y=t-x)$$

Continuous Case

$$\underset{\text{PDF}}{f_T}(t) = \int_{-\infty}^{\infty} f_X(x)\, f_Y(t-x)\, dx$$

since $\underset{\text{CDF}}{F_T}(t) = P(T \leq t)$

$$= \int_{-\infty}^{\infty} P(X+Y \leq t \mid X=x)\, f_X(x)\, dx$$

(Continuous LOTP)

$$= \int_{-\infty}^{\infty} P(Y \leq t-x)\, f_X(x)\, dx$$

$$= \int_{-\infty}^{\infty} F_Y(t-x)\, f_X(x)\, dx$$

$$f_T(t) = \frac{d}{dt} F_T(t) = \frac{d}{dt}\int_{-\infty}^{\infty} F_Y(t-x)\, f_X(x)\, dx$$

$$= \int_{-\infty}^{\infty} f_X(x)\, f_Y(t-x)\, dx$$

Idea: prove existence of objects with desired property A using probability.
Show $P(A) > 0$ for a random object.

Strategy

→ We don't actually have to compute $P(A)$ exactly, we only need a bound that shows that it's greater than 0.

Suppose each object has a "score". Show there is an object with "good" score.

Theorem: There is an object ~~with score~~ whose score is atleast the average, $E(X)$.

X → score of random object

Shannon's Theorem (Communication / Information Theory)

→ If we are trying to communicate over a noisy channel (so we are trying to send messages from one place to another, but bits get corrupted or there is a lot of noise or interference, etc.), there is something called the capacity of the channel and we can communicate at rates arbitrarily close to the capacity, with arbitrarily small chance of error, i.e., even if we have a very, very noisy channel, we can make the error probability very, very low.

Example

100 people; 15 committees of 20 people; each person is on 3 committees

Existence problem

Show that there exists two committees whose overlap is atleast 3. In other words, there exist 2 committees, where a group of 3 people is on both committees.

Soln:- Idea : find average ~~intersection~~ overlap of 2 random committees.

$$E(\text{overlap}) = 100 \cdot \left[\frac{\binom{3}{2}}{\binom{15}{2}}\right]$$

[fundamental bridge]

We create Indicator random variable for each person. 100 people

choose 2 random committees out of 15

prob. that person no. 1 is on both of those randomly chosen committees

person no 1 is on 3 committees, so chose 2 out of the 3 committees

$$= \frac{3 \cdot \cancel{20}^{10} \cdot 2}{\cancel{15} \cdot \cancel{14}_{7}}$$

$$= 20/7$$

$\Rightarrow$ there exist a pair of committees with overlap of $\geq 20/7 \Rightarrow$ have overlap of ≥ 3 (since overlap is an integer)