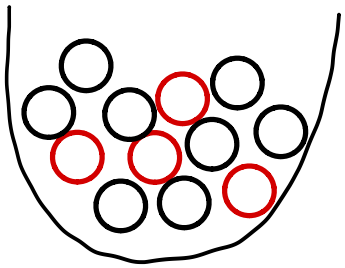
Stat 2911 Lecture Notes

Class 4, 2017

Uri Keich

Class 4: Hypergeometric RV, Fisher's Exact Test, joint PMF, independent discrete RVs, deriving the Poisson distribution

5) A <u>hypergeometric</u> RV X models the number of red balls in a sample of m balls drawn <u>without</u> replacement from an urn with r red balls and $n-r$ black balls.



A 3-parameter family (r, n, m)

$X(\Omega) \subset \{0, 1, \ldots, r\}$

All samples of size m are assumed equally likely: equiprobable space.

For $k \in \{0, 1, \ldots, r\}$,

$$P(X = k) = \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}}$$

Fisher's Exact Test (Rice 13.2)

Statistical Methods for Research Workers (1925)

30 convicted criminals with same sex twin, of which 13 are monozygotic (identical) twins and 17 are dizygothic twins.

10/13 monoz. twin siblings and 2/17 diz. twin siblings were themselves convicted.

Fisher: is there evidence of a genetic link?

twin's fate:

	convicted	not convicted	
mono	10	3	13
di	2	15	17
	12	18	

Assuming that, whether or not the sibling of a convicted criminal is also convicted does <u>not</u> depend on the type of the twin, we have a sample from a hypergeometric dist:

13 red balls (monoz. twin siblings)

17 black balls (diz. " ")

We randomly sample 12 balls (convicted siblings).

X = # of convicted monoz. siblings $\sim$ HyperG(13, 17, 12)

$\binom{r}{} \binom{n-r}{} \binom{m}{}$

What is the prob. we will observe 10 or more red balls in the sample? (p-value)

$$P(X \geq 10) = \sum_{k=10}^{12} P(X=k)$$

$$= \sum_{k=10}^{12} \frac{\binom{13}{k}\binom{17}{12-k}}{\binom{30}{12}} \simeq 0.000465$$

It seems very unlikely that there is no relation between the twin type and their conviction.

But did we establish that criminal mind is inherited?

- probabilistic statement
- conviction vs. truth "that face is up to no-good"
- sampling/ascertainment bias: finger print already in DB
- Identical twins feel tighter connections

Joint distribution

The joint pmf of the RVs X & Y specifies their interaction:

$$p_{XY}(x,y) = P(X=x, Y=y) \quad x,y \in \mathbb{R}$$

Note $\{X=x, Y=y\} = \{\omega : X(\omega)=x, Y(\omega)=y\}$

How do you know this is an event?

If X & Y are discrete RVs with

$$\{x_i\} = X(\Omega) \quad \text{and} \quad \{y_j\} = Y(\Omega)$$

then

$$p_X(x) = P(X=x) \overset{?}{=} \sum_j P(X=x, Y=y_j) = \sum_j p_{XY}(x, y_j).$$

Similarly,

$$p_Y(y) = \sum_i p_{XY}(x_i, y).$$

p_X and p_Y are called the <u>marginal pmfs</u>.

Example (Rice 3.2)

A fair coin is tossed 3 times.

Let $X = 1\{H \text{ on first toss}\} = \#$ of heads in first toss

$Y =$ total # of heads

$\Omega = \{HHH, HHT, \ldots\}$ equiprobable space

$|\Omega| =$

x \ y	0	1	2	3	
0					
1					

$\{X=0, Y=0\} = \{$ $\}$

$\Rightarrow P(X=0, Y=0) =$

x \ y	0	1	2	3	
0	1/8	2/8	1/8	0	1/2
1	0	1/8	2/8	1/8	1/2
	1/8	3/8	3/8	1/8	

$X \sim$

$Y \sim$

Are X and Y "independent"?

Independent RVs

The RV X is ind. of Y if knowing only the value of Y does not give us any information about X, in particular, it does not change the dist. of X:

$$P(X = x_i \mid Y = y_j) = P(X = x_i) \quad \forall i, j$$

Def. $\Rightarrow$

$$\boxed{P(X = x_i, Y = y_j) = P(X = x_i) P(Y = y_j) \quad \forall i, j}$$

$\Leftrightarrow$

$$p_{XY}(x_i, y_j) = p_X(x_i) p_Y(y_j) \quad \forall i, j$$

The joint pmf factors into the product of the marginals.

More generally, the RVs $X_1, \ldots, X_n$ are ind. if

$$P(X_1 = x_1, \ldots, X_n = x_n) = p_{\underline{X}}(\underline{x}) = \prod_{i}^{n} p_{X_i}(x_i) \quad \forall \underline{x} \in \mathbb{R}^n$$

$\underline{X} = (X_1, \ldots, X_n)$ is a random vector and $\underline{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$.

Warning. Pairwise ind. of the RVs does not imply (mutual) ind.

The Poisson Dist.

Recall: $X \sim \text{Poisson}(\lambda)$ if $p_X(k) = e^{-\lambda}\lambda^k/k!$ $k \in \mathbb{N}$

A 1-parameter family: $\lambda > 0$

Where does it come from?

A Poisson dist. models the number of "events" that are registered in a certain time interval. For example, the number of

- particles emitted by a radioactive source in an hour
- incoming calls to a service center between 1-2 pm
- light bulbs burnt in a year
- fatalities from horse kicks in the Prussian cavalry in one corp year (Bortkiewicz 1898, Rice p. 215):

20 corps × 10 years = 200 Corp years

# of deaths	observed count	frequency	Poisson approx.
0	109	0.545	0.543
1	65	0.325	0.331
2	22	0.110	0.101
3	3	0.015	0.021
4	1	0.005	0.003

What are the assumptions that lead to this pmf?

1) The dist. of the number of events at any time interval depends only on the interval's length/duration.
2) The # of events recorded in two disjoint time intervals are ind. of one another.
3) No two events are recorded at exactly the same time point.

Let $\overline{X}_{t,s}$ = # of events in the time interval $(t,s]$ and let $\overline{X}_t = \overline{X}_{0,t}$.

Goal: find the dist. of $\overline{X}_1$.

We start with $P(\overline{X}_1=0)$ by studying $f(t) \stackrel{d}{=} P(\overline{X}_t=0)$.

$$\begin{aligned} f(t+s) &= P(\overline{X}_{t+s}=0) \\ &= P(\overline{X}_t=0, \overline{X}_{t,t+s}=0) \\ &\overset{?}{=} P(\overline{X}_t=0)\, P(\overline{X}_{t,t+s}=0) \\ &\overset{?}{=} P(\overline{X}_t=0)\, P(\overline{X}_{0,s}=0) = f(t)\, f(s) \end{aligned}$$

$$\Rightarrow \quad f(t+s) = f(t)\, f(s) \qquad \forall\, t,s \geq 0$$

What are the solutions of this functional equation?

One type is $f(t) = e^{\alpha t}$ $\quad \alpha \in \mathbb{R}$

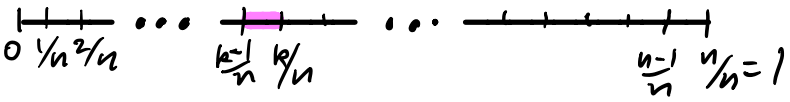
Other solutions are a "complete mess" and, in particular, they are unbounded, but $f(t) \in [0,1]$ $\forall\, t \geq 0$ (why?).

For the same reason $\alpha \le 0$ so $\alpha = -\lambda$ for $\lambda > 0$.

$\Rightarrow P(X_t = 0) = f(t) = e^{-\lambda t}$

In particular, for $t=1$ $P(X_1 = 0) = e^{-\lambda}$ $\left(e^{-\lambda}\lambda^k/k!,\ k=0\right)$

To find $P(X_1 = j)$ for $j > 0$ consider dividing the time interval $(0,1]$ to n intervals of size $1/n$:



Let Y_n = # of intervals $\left(\frac{k-1}{n}, \frac{k}{n}\right]$ in $(0,1]$ in which an event occurred.

$$= \sum_{k=1}^{n} Y_n^k \quad \text{where} \quad Y_n^k = 1_{\{X_{\frac{k-1}{n}, \frac{k}{n}} \ge 1\}}$$

$Y_n^1, \ldots Y_n^n$ are iid Bernoulli (p_n) RVs where

$$p_n = P\left(X_{\frac{k-1}{n}, \frac{k}{n}} \ge 1\right)$$
$$= P\left(X_{1/n} \ge 1\right)$$
$$= 1 - P\left(X_{1/n} = 0\right)$$
$$= 1 - e^{-\lambda/n}$$

$\Rightarrow Y_n \sim \text{Binomial}\left(n, p_n = 1 - e^{-\lambda/n}\right)$

Note: Generally, $Y_n \le X_1$ and $Y_n < X_1$ if two events occur in the same interval.

However, $\lim_{n\to\infty} Y_n(\omega) = X_1(\omega)$ $\forall \omega \in \Omega$ because no two events can occur at the same time.