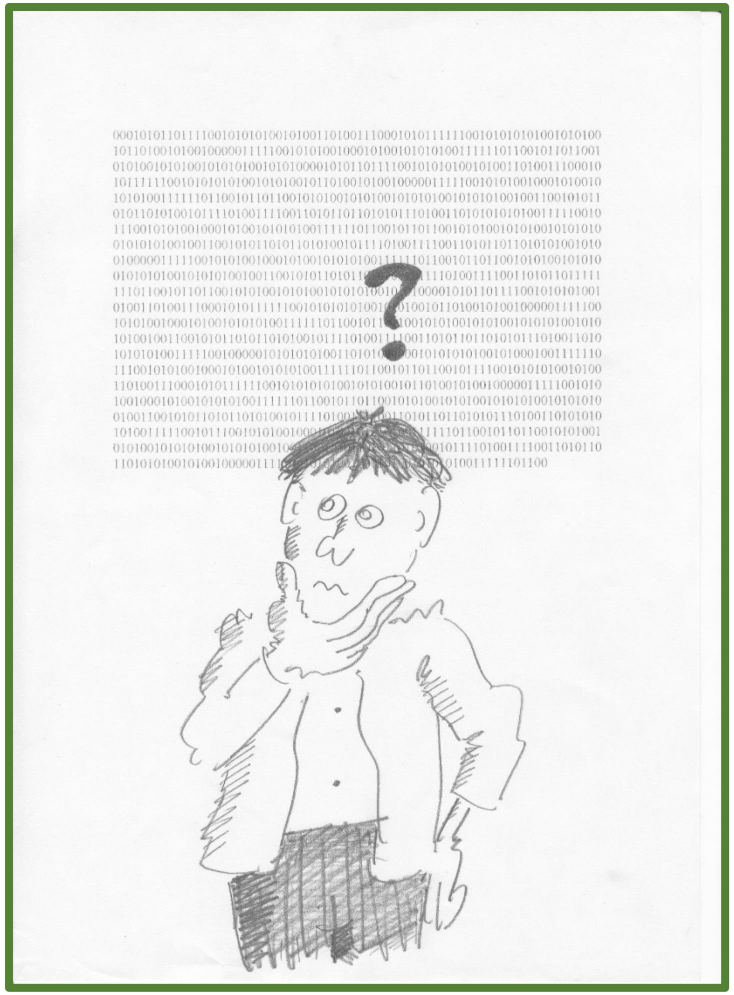
Statistics and data analysis

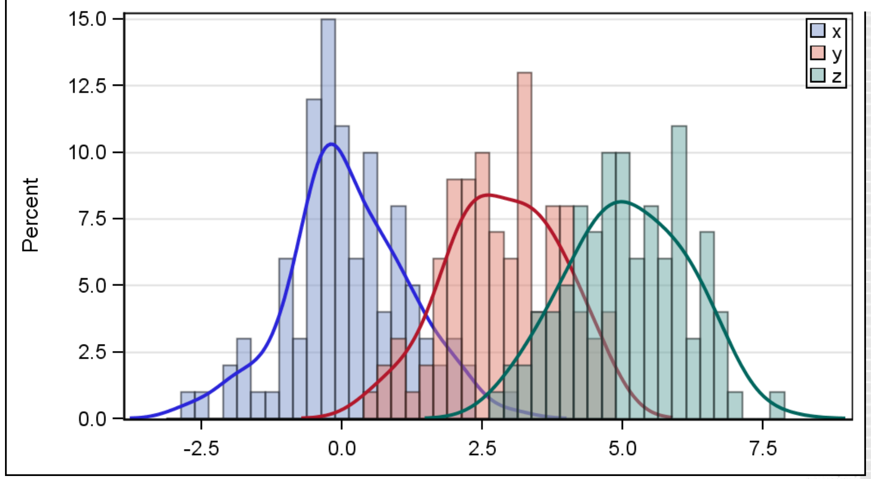
Zohar Yakhini, Leon Anavy

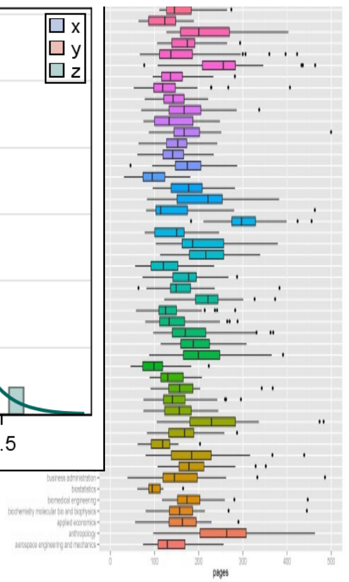
IDC, Herzeliya

Independence and variations, convolution, computer age statistics



001001110101010010101010010010001010101000101011111010110100110010011110101010011010010110010110110010





Negative Binomial Distribution

- In successive Bernoulli(p) instances, what is the distribution of the number of trials (in some versions – failures) needed until the r^{th} success.
  (the Geometric Distribution is equivalent to $r = 1$)
- For this number to equal y we should have exactly $r - 1$ successes in first $y - 1$ trials, followed by a success

$$p(y) = \binom{y-1}{r-1} p^r (1-p)^{y-r} \qquad y = r, r+1, \ldots$$

$$E(Y) = \frac{r}{p}$$

$$V(Y) = \frac{r(1-p)}{p^2}$$

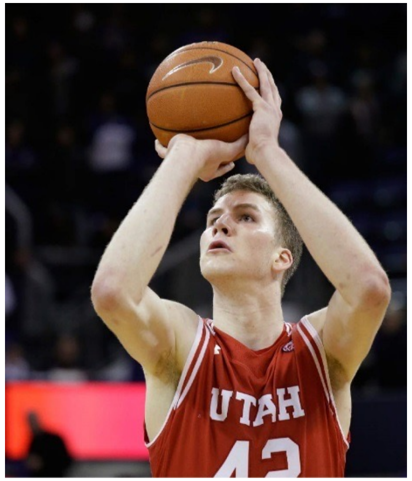
Randomistan basketball, again

Players shoot synchronously.

Player 1:
Probability of scoring = $p < 1/2$
Shoots until he has r successes.
X_1 is the attempt when that happened.

Player 2:
Probability of scoring = mp for some integer $1 < m$ so that $mp < 1$
Shoots until she has mr successes.
X_2 is the attempt when that happened.

From: DeGroot, Probability and Statistics



- Which is higher $\mathrm{E}(X_1)$ or $\mathrm{E}(X_2)$?
- Which is higher $\mathrm{V}(X_1)$ or $\mathrm{V}(X_2)$?
- Placing a bet on $X_1 > X_2$? (Player 2 is better)

$$\mathrm{E}(X_1) = \frac{r}{p} = \frac{mr}{mp} = E(X_2)$$

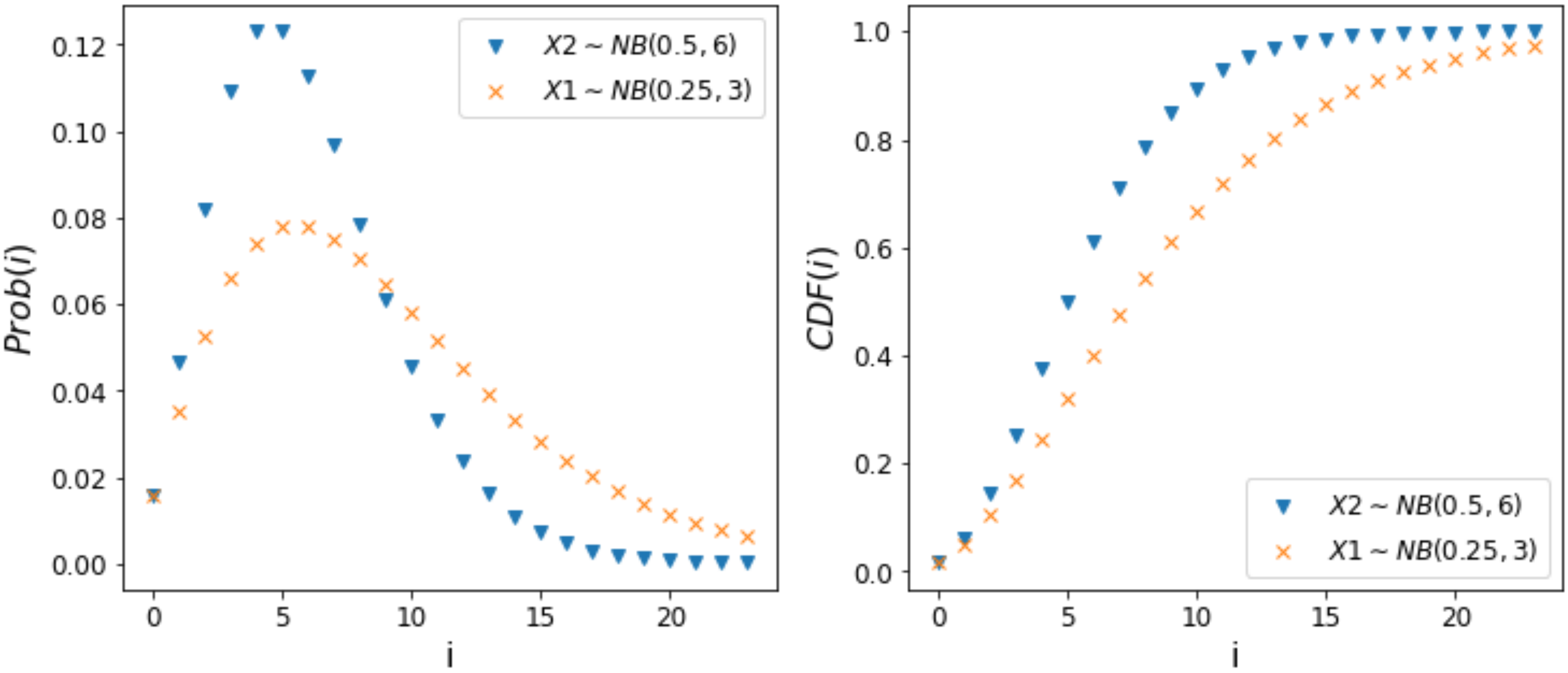
$$\mathrm{V}(X_1) = \frac{r(1-p)}{p^2} \; ? \; \frac{mr(1-mp)}{(mp)^2} = V(X_2)$$

scipy.stats.nbinom

```
from scipy.stats import nbinom
import numpy as np
from matplotlib import pyplot as plt
```

```
r = 3
p = 0.25
m = 2
```

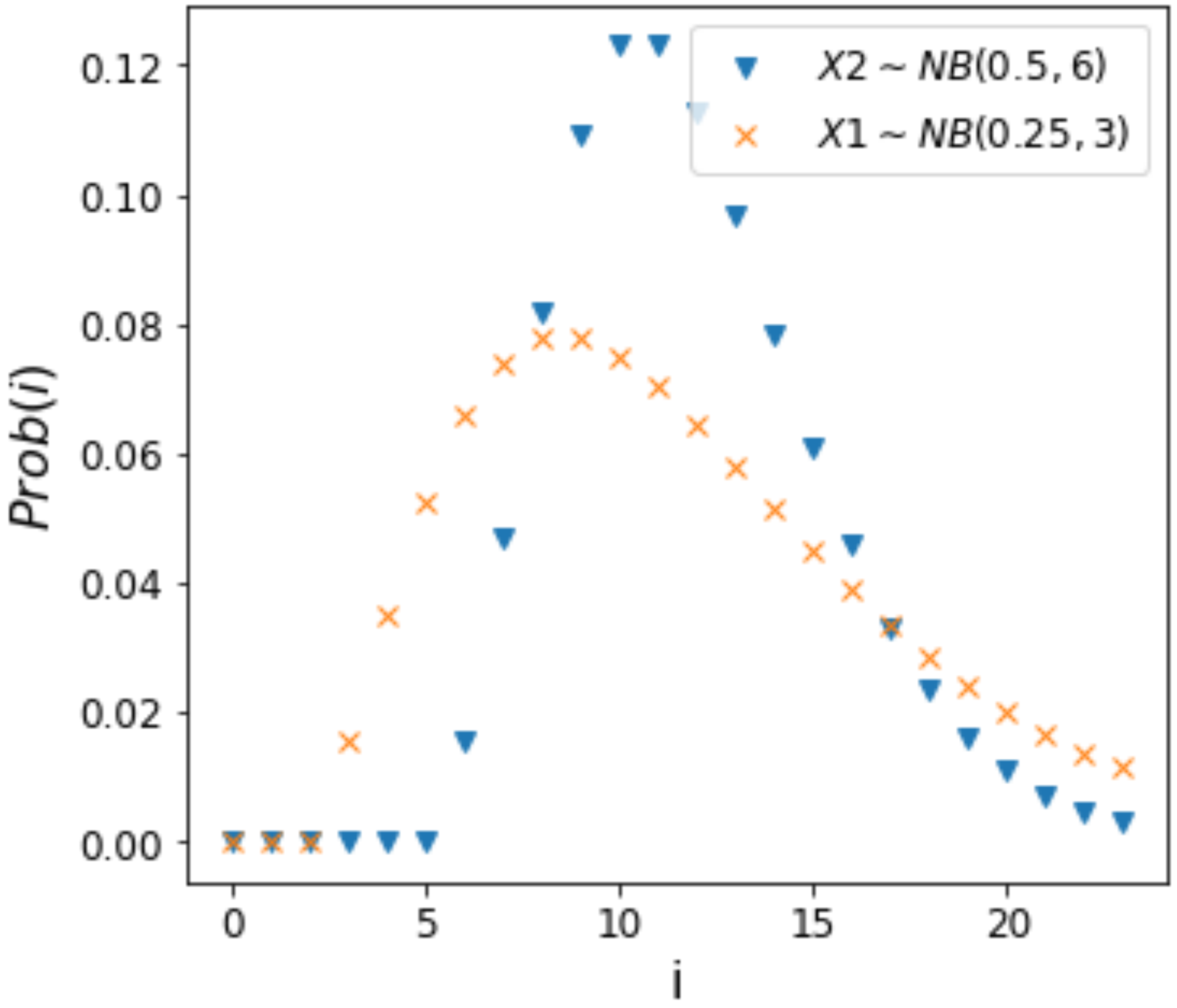
```
X1 = nbinom(r,p)
X2 = nbinom(r*m,p*m)

i = range(0,int(np.round(2*r/p,0)))

p_X1_i = X1.pmf([xx for xx in i])
p_X2_i = X2.pmf([xx for xx in i])

plt.figure(figsize=(12,5))
plt.subplot(1,2,1)
plt.plot(i,p_X2_i,'v',label="$X2\sim NB({{{0}}},{{{1}}})$".format(p*m,r*m))
plt.plot(i,p_X1_i,'x',label="$X1\sim NB({{{0}}},{{{1}}})$".format(p,r))
plt.xlabel("i",fontsize=16)
plt.ylabel('$Prob(i)$',fontsize=16)
plt.legend()
```

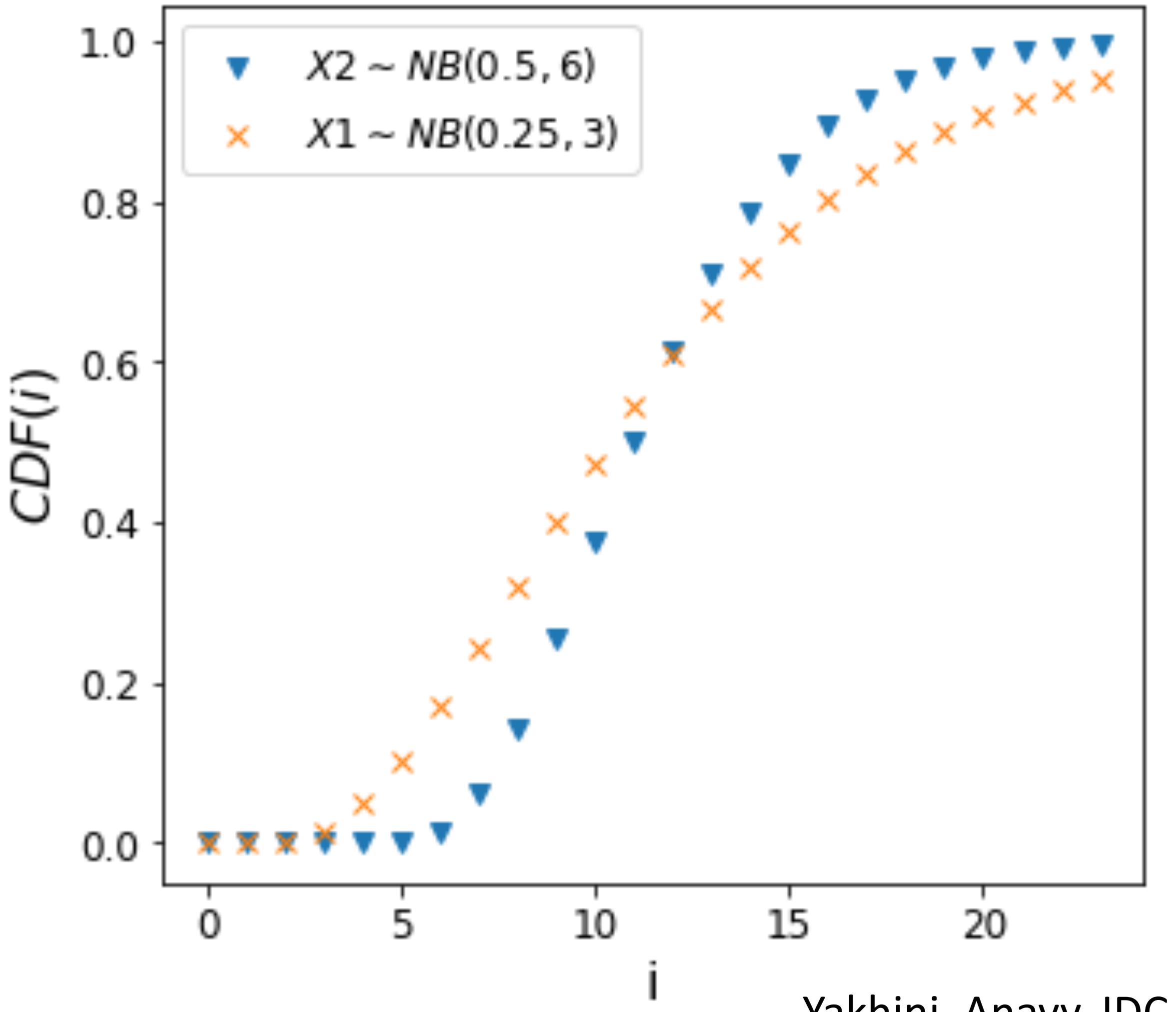


Prob(i)
CDF(i)
i
$X2 \sim NB(0.5, 6)$
$X1 \sim NB(0.25, 3)$

```
# Wrong behaviour
X1 = nbinom(r,p)
X2 = nbinom(r*m,p*m)
```

```
# Correct behaviour
X1 = nbinom(r,p,loc=r)
X2 = nbinom(r*m,p*m,loc=m*r)
```





Note: E vs mod of a distribution

Randomistan basketball story

- Which is higher $\mathrm{E}(X_1)$ or $\mathrm{E}(X_2)$?
- Which is higher $\mathrm{V}(X_1)$ or $\mathrm{V}(X_2)$?
- Placing a bet on $X_1 > X_2$? (Player 2 is better)

```
r = 3
p = 0.25
m = 2
mean_X1, var_X1 = nbinom.stats(r,p,loc=r)
mean_X2, var_X2 = nbinom.stats(r*m,p*m,loc=m*r)
print(f'E(X1_1) = {mean_X1}, Var(X1_1) = {var_X1}')
print(f'E(X1_2) = {mean_X2}, Var(X1_2) = {var_X2}')
```

```
E(X1_1) = 12.0, Var(X1_1) = 36.0
E(X1_2) = 12.0, Var(X1_2) = 12.0
```

How to assess betting on the players?

- Placing a bet on $X_1 > X_2$? (Player 2 is better)

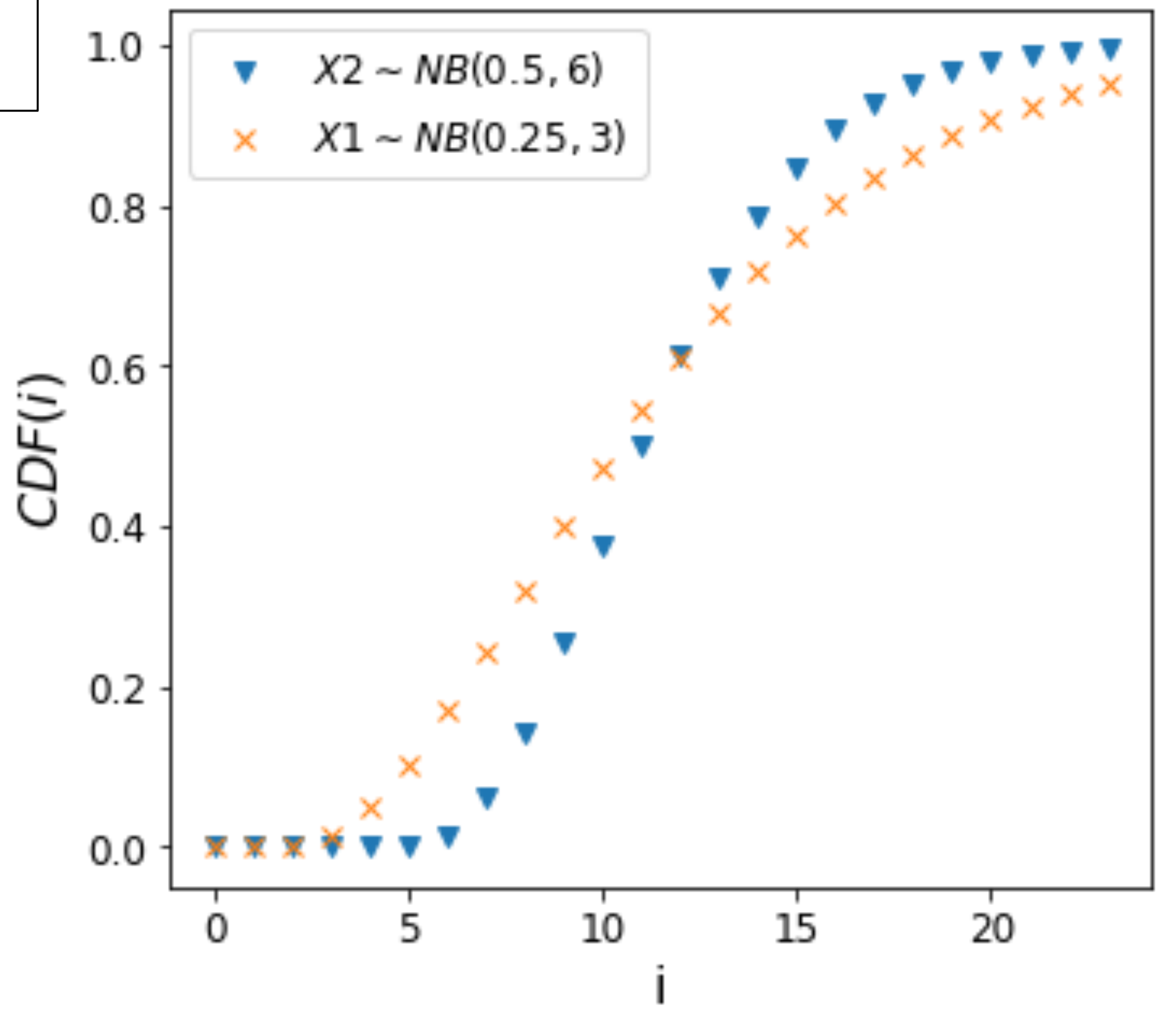
We can choose a player and bet on whether they succeed before 8, before 20. Which player should we prefer?

Calculate $P(X_1 \leq 8)$ and $P(X_2 \leq 8)$

Calculate $P(X_1 \leq 20)$ and $P(X_2 \leq 20)$

```
v1 = 8
v2 = 20
f_X1_v1 = X1.cdf(v1)
f_X2_v1 = X2.cdf(v1)
f_X1_v2 = X1.cdf(v2)
f_X2_v2 = X2.cdf(v2)
```

```
P(X1 <= 8) = 0.32
P(X2 <= 8) = 0.14
X1 wins on 8 trial
P(X1 <= 20) = 0.91
P(X2 <= 20) = 0.98
X2 wins on 20 trial
```



Who completes the task earlier? Computer age statistics

Calculate $P(X_1 > X_2)$

Lower Bound:

$$P(X_1 > X_2) = \sum_{y=m\cdot k}^{\inf} P(X_2 = y)P(X_1 > y) \geq$$

$$\sum_{y=m\cdot k}^{R} P(X_2 = y)P(X_1 > y) = \sum_{y=m\cdot k}^{R} P(X_2 = y)(1 - CDF_{X_1}(y))$$

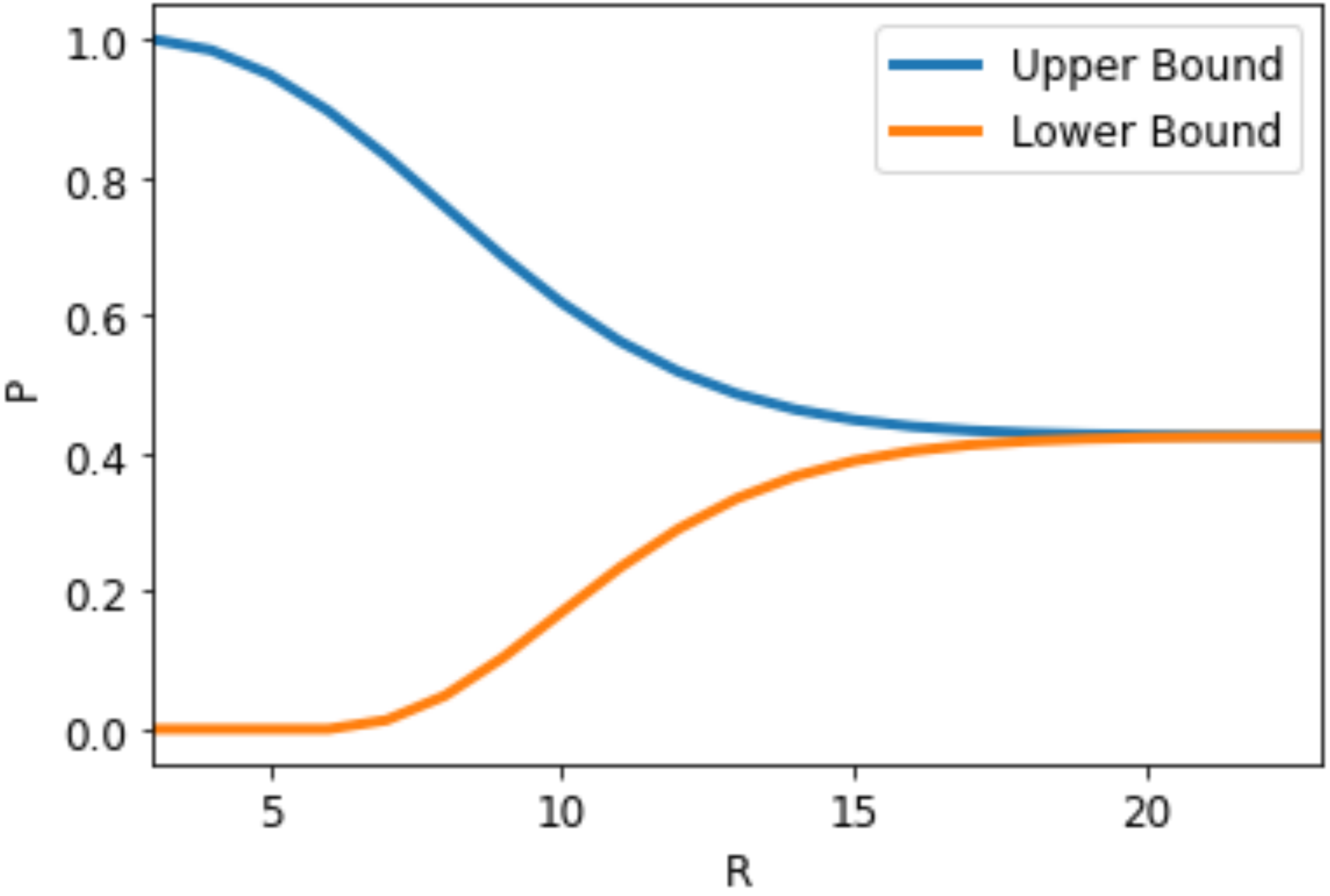
Upper Bound:

$$P(X_1 > X_2) = 1 - P(X_2 \geq X_1) = 1 - \sum_{x=k}^{\inf} P(X_1 = x)P(X_2 \geq x) \leq$$

$$1 - \sum_{x=k}^{R} P(X_1 = x)P(X_2 \geq x) = 1 - \sum_{x=k}^{R} P(X_1 = x)(1 - P(X_2 < x)) = 1 - \sum_{x=k}^{R} P(X_1 = x)(1 - CDF_{X_2}(x - 1))$$

$$P(X_1 > X_2)$$

Calculate $P(X > Y)$



$P(X_1 > X_2) \in [0.4246, 0.4251]$

Sample/Coupon collection

The RV T counts the number of observations required to see at least m=1 users from each country of the n=100. How many visits will it take if every visit comes from each of the countries with equal probabilities and independent of all previous visits?

$$T = X_1 + X_2 + X_3 + \ldots + X_i + \ \ldots + X_{99} + X_{100}$$

Where the random variable X_i counts the number of visits, after the first $i - 1$ countries are in, until the i-th country is also in.

Sample/Coupon collection

We saw

$$E(T) = E(X_1 + X_2 + X_3 + \ldots + X_i + \ \ldots + X_{99} + X_{100}) = \sum_{i=1}^{100} E(X_i)$$

Note that $X_i \sim Geom\left(p_i = \frac{100-i+1}{100}\right)$ and we therefore have $E(X_i) = \frac{1}{p_i} = \frac{100}{100-i+1}$

So:

$$E(T) = 100\left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ \ldots + \frac{1}{100}\right)$$

$$\underline{E(T) = nH(n) \sim n \ln n.}$$

Sample/Coupon collection

How can we use Chebyshev's inequality to get a bound on:

$$P(T > nH(n) + cn) \text{ ?}$$

$$P(|X - \mu| \geq \lambda) \leq \frac{\mathrm{V(X)}}{\lambda^2}$$

Sample/Coupon collection

$$P(|X-\mu| \geq \lambda) \leq \frac{\mathrm{V(X)}}{\lambda^2}$$

$$P(T \geq E(T) + \lambda) \leq P(|T - E(T)| \geq \lambda)$$

$$P(T \geq nH(n) + cn) \leq \frac{\mathrm{V(T)}}{c^2 n^2}$$

Sample/Coupon collection

What about the Variance of T?

$$Var(T) = Var(X_1 + X_2 + X_3 + \dots + X_i + \ \dots + X_{99} + X_{100}) = \sum_{i=1}^{100} Var(X_i)$$

$X_i \sim Geom\left(p_i = \frac{100-i+1}{100}\right)$ and we therefore have $Var(X_i) = \frac{1-p_i}{p_i^2}$

So:

$$Var(T) = \sum_{i=1}^{100} \frac{1-p_i}{p_i^2} = \sum_{i=1}^{100} \frac{1}{p_i^2} - \sum_{i=1}^{100} \frac{1}{p_i} = 100^2 \sum_{i=1}^{100} \frac{1}{i^2} - 100 \sum_{i=1}^{100} \frac{1}{i}$$

$$Var(T) = n^2 \sum_{i=1}^{n} \frac{1}{i^2} - nH(n)$$

Sample/Coupon collection

$$Var(T) = n^2 \sum_{i=1}^{n} \frac{1}{i^2} - nH(n) < n^2 \sum_{i=1}^{n} \frac{1}{i^2} < n^2 \frac{\pi^2}{6}$$

$$P(T \geq nH(n) + cn) \leq \frac{\mathrm{V(T)}}{c^2 n^2} < \frac{\pi^2}{6c^2}$$

Taking $n = 100$:

$c = 2$: $P(T \geq 518 + 200) < \frac{\pi^2}{24} = 0.4112$

$c = 3$: $P(T \geq 518 + 300) < \frac{\pi^2}{54} = 0.1828$

$c = 4$: $P(T \geq 518 + 400) < \frac{\pi^2}{96} = 0.1028$

Computer age statistics

We can calculate the true value of the variance and use this for the Chebyshev bound:

$$Var(T) = n^2 \sum_{i=1}^{n} \frac{1}{i^2} - nH(n)$$

$$P(T \geq nH(n) + cn) \leq \frac{\mathrm{V(T)}}{c^2 n^2}$$

Taking $n = 100$: $Var(T) = 16449$

$c = 2$: $P(T \geq 518 + 200) \leq 0.3958$

$c = 3$: $P(T \geq 518 + 300) \leq 0.1759$

$c = 4$: $P(T \geq 518 + 400) \leq 0.0989$

```
v = single_coupon_variance(100)
print(f'Using exact Variance')
print(f'P(T_100>718) <= {v/4/100**2 :.4f}')
print(f'P(T_100>818) <= {v/9/100**2 :.4f}')
print(f'P(T_100>918) <= {v/16/100**2 :.4f}')

print(f'Using upper bound on the variance')
print(f'P(T_100>718) <= {math.pi ** 2 / 6 / 4 :.4f}')
print(f'P(T_100>818) <= {math.pi ** 2 / 6 / 9 :.4f}')
print(f'P(T_100>918) <= {math.pi ** 2 / 6 / 16 :.4f}')
```

```
Using exact Variance
P(T_100>718) <= 0.3958
P(T_100>818) <= 0.1759
P(T_100>918) <= 0.0989
Using upper bound on the variance
P(T_100>718) <= 0.4112
P(T_100>818) <= 0.1828
P(T_100>918) <= 0.1028
```

Sample/Coupon collection

How can we calculate the probability directly:

$$P(T > nH(n) + cn)\ ?$$

$$T = X_1 + X_2 + X_3 + \ldots + X_i + \ \ldots + X_{99} + X_{100}$$

Sums of independent random variables

Let X and Y be two independent random variables. Let $Z = X + Y$. Then

$$P(Z = z) = \sum_{i=-\infty}^{\infty} P(X = i)P(Y = z - i)$$

For continuous random variables, the density function of Z is:

$$h(z) = \int_{-\infty}^{\infty} f(t)g(z - t)dt$$

Computer age statistics

We can use convolutions to compute the actual FULL (or rather – the interesting part) distribution of T_N :

$$P(T_N = k) = \sum_{i=-\infty}^{\infty} P(G_N = i) P(T_{N-1} = k - i)$$
$$= \sum_{i=1}^{k-1} P(G_N = i) P(T_{N-1} = k - i)$$

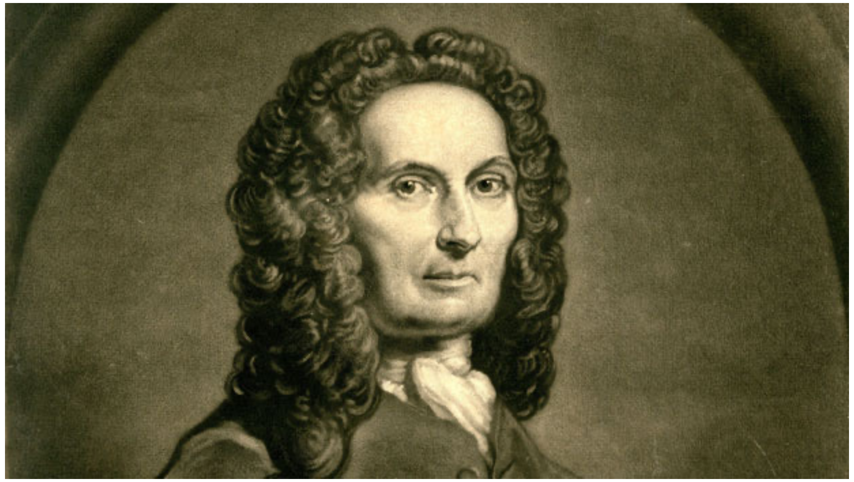
where $G_s \sim Geo(\, p = \frac{N-s+1}{N})$.

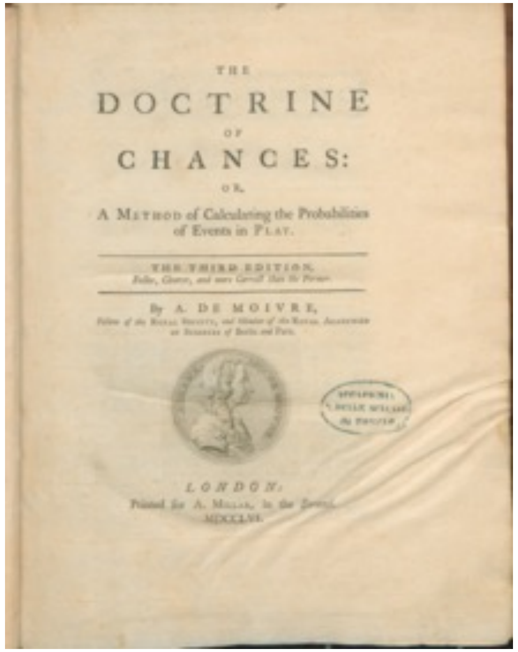
We need to initialize this with $P(T_1 = 1) = 1$ and 0 for all other values.

Exact coupon collector waiting time

```
exact P(T_100>718) = 0.12
exact P(T_100>818) = 0.05
exact P(T_100>918) = 0.02
```

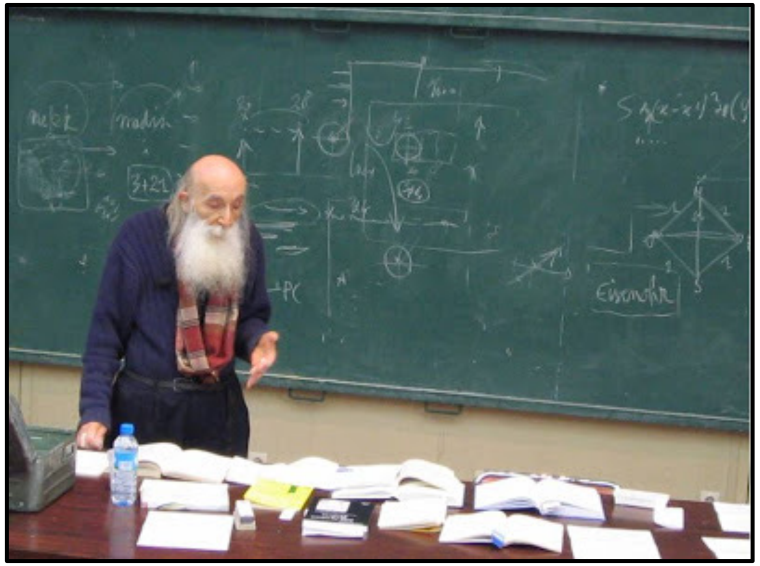
Sample/coupon collector and computers: Historical notes





- The T_N discussion dates back to Abraham de Moivre 1667 (France) – 1754 (England)
- Rigorously treated by William Feller in the 1940s
- Variants are still being studied as an active field of research

Jean Paul Benzecri, French statistician (1932-2019):
"It is unthinkable to use methods conceived before the invention of the computer. Statistics will have to be completely rewritten!"
Stated in 1965.



Pairwise independence

A set of random variables $(X_1, X_2, \dots, X_n)$ is said to be pairwise independent if any two random variables X_i and X_j are independent.
Recall – a set of random variables as above is called (collectively or mutually) independent if

$$\forall (x_1, x_2, \dots, x_n)$$

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i)$$

Equivalence?

- Does collective independence imply pairwise independence?
- Does pairwise independence imply collective independence?

Var of a sum?

Pairwise independence is sufficient for the linearity of variances.

Let X and Y two independent Bernoulli w $p = \frac{1}{2}$.
Let $Z = XOR(X, Y)$.
We work in $\Omega = \{0,1\}^3$.

We have the following joint probability mass function:

X	Y	Z	P
0	0	0	0.25
0	1	1	0.25
1	0	1	0.25
1	1	0	0.25

$$X + Y + Z \text{ vs } Binom(0.5,3)$$

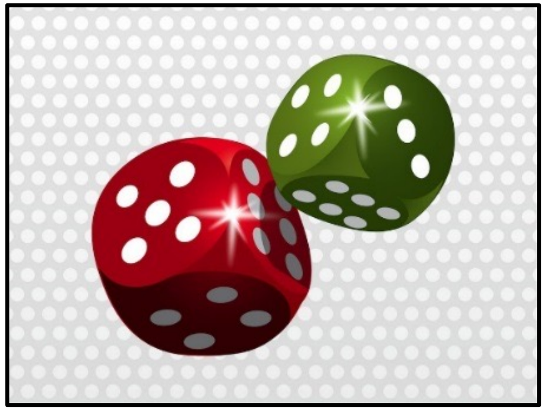
$$E(X+Y+Z) = ?$$
$$V(X+Y+Z) = ?$$

X	Y	Z	P
0	0	0	0.25
0	1	1	0.25
1	0	1	0.25
1	1	0	0.25

Covariance and independence

$$P(X = x) = 1/3 \text{ for } x = -1, 0, 1 \qquad , \qquad Y = X^2$$

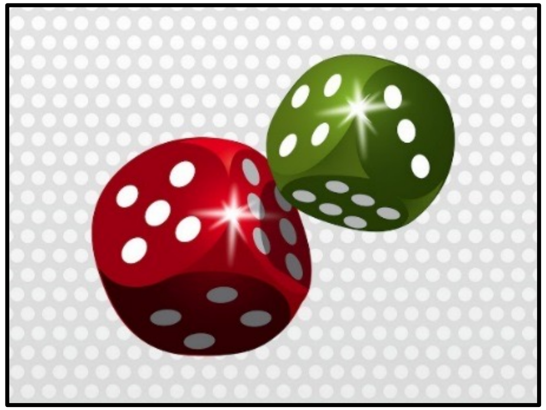
Multinomial Distribution



Roll a die n times, Y counts the number of 5's

- What is the distribution of Y?
- Can you treat this as a coin? What is p?
- What is $\mathrm{E}(Y)$?
- What is $\mathrm{V}(Y)$?

Multinomial Distribution



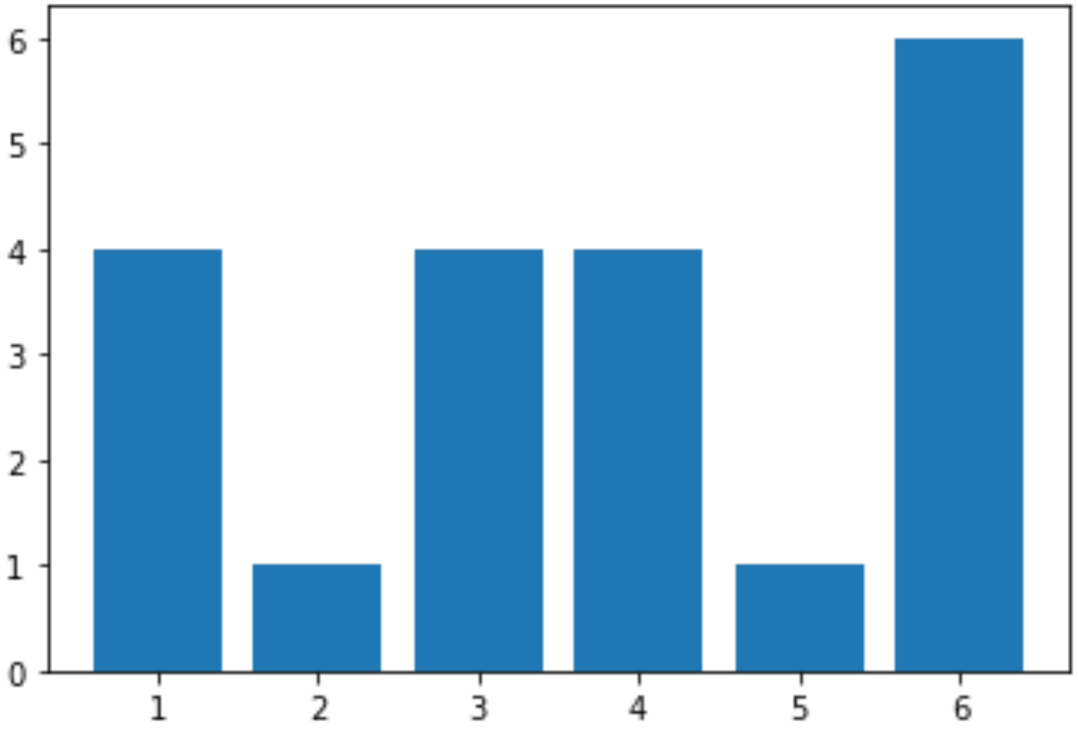
Now roll a die n times and define $X = (X_1, X_2, \ldots, X_6)$ where X_i counts the number of i's

$$X = (X_1, X_2, \ldots, X_d) \sim \text{Multinomial}(n, p)$$

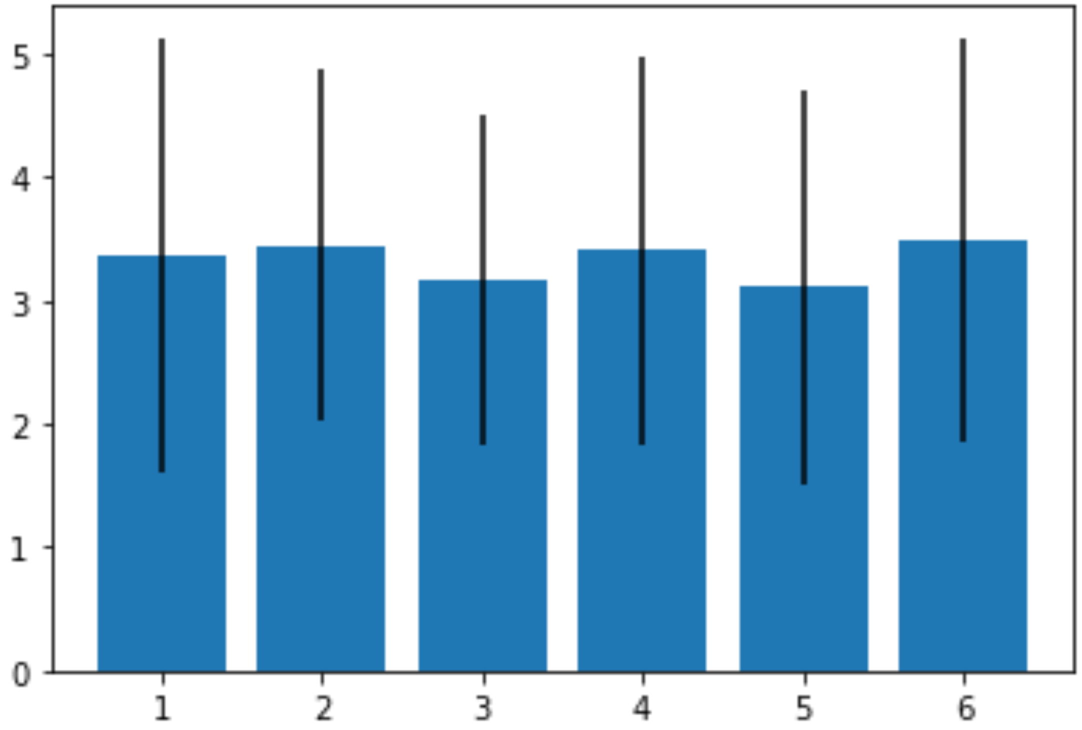
- What is d?
- What is p?
- What is $\mathrm{E}(X_i)$?
- What is $\mathrm{V}(X_i)$?

Multinomial Distribution

$$X = (X_1, X_2, \dots, X_6) \sim \text{Multinomial}\left(20, \left(\frac{1}{6}, \dots, \frac{1}{6}\right)\right)$$

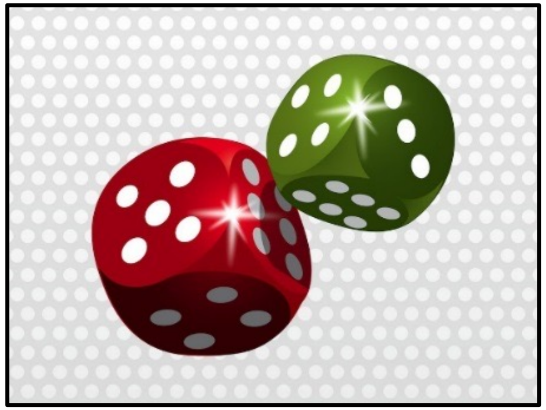


One experiment



100 repeated experiments, with averaging

Multinomial Distribution



Now roll a die n times and define $X = (X_1, X_2, \dots, X_6)$ where X_i counts the number of i's

$$X = (X_1, X_2, \dots, X_d) \sim \text{Multinomial}(n, p)$$

- Are these random variables collectively independent?
- Pairwise independent?

Multinomial Distribution - covariances

Let $X \sim \mathrm{MNom}(N, P)$, $X = (X_1, X_2, \ldots, X_d)$. What is $Cov(X_i, X_j)$?

$$Var(X_i + X_j) = V(X_i) + 2Cov(X_i, X_j) + V(X_j)$$

Now observe that $X_i + X_j \sim \mathrm{Binom}(N, p_i + p_j)$ and therefore we can write:

$$2Cov(X_i, X_j) = Var(X_i + X_j) - V(X_i) - V(X_j) =$$

$$= N\left[(p_i + p_j)(1 - p_i - p_j) - p_i(1 - p_i) - p_j(1 - p_j)\right]$$

$$= -2Np_ip_j$$

Multinomials, example

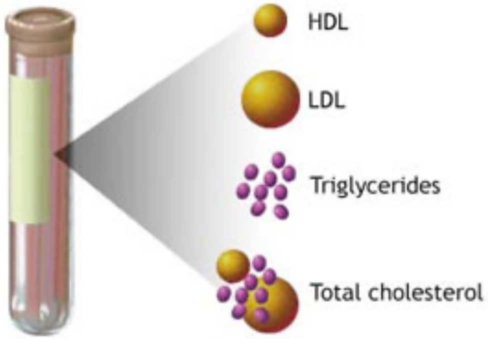
Let $X \sim \mathrm{MNom}(N, P)$, $X = (X_1, X_2 , \ldots , X_{10})$

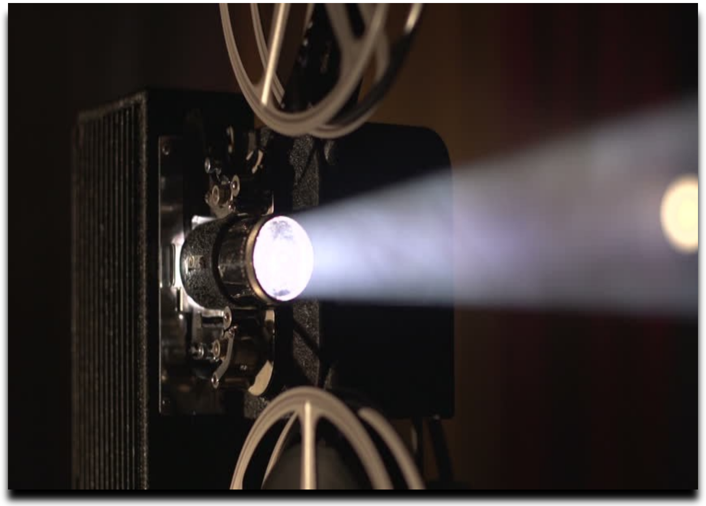
Also assume that P is uniform $\frac{1}{10}$

What is $Cov(X_1 + X_2, X_7 + X_8)$?

Conditional independence

- Is the blood cholesterol level of a person independent of the number of movies watched by that person so far?
- No – they are both related to the age of the person.
- But – they are conditionally independent given the age.
- Presumably …, socioeconomic and behavioral factors ignored …
- Notation: $X \perp Y \mid C$





Conditional Independence - Definition

Two random variables X and Y are conditionally independent given a third rv C if
for all pairs (x, y) AND
for all possible values c of C,
we have:

$$\boxed{P\big((X = x \ \wedge \ Y = y)|C = c\big) = P\big((X = x)|C = c\big) \cdot P\big((Y = y)|C = c\big)}$$

Conditionally independent but not independent?

Independent but not conditionally independent?

Summary

- Computer age statistics:
  - + Comparing negative binomials
  - + Coupon collector – exact calculations
  - + Independence and convolutions
- Mutual indpce vs lower order indpce
- Multinomials
- Conditional independence