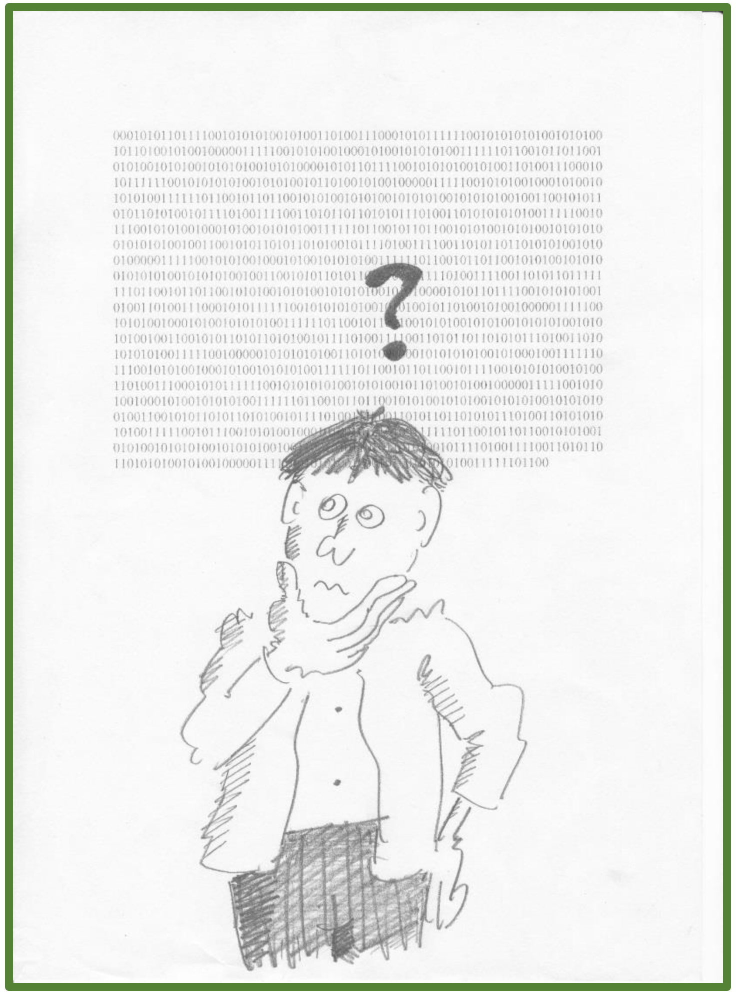
Statistics and data analysis

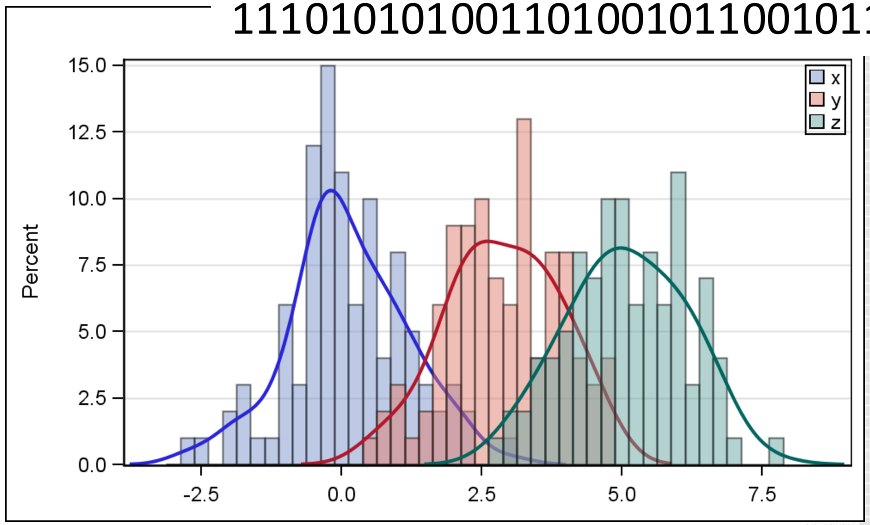
Zohar Yakhini, Leon Anavy

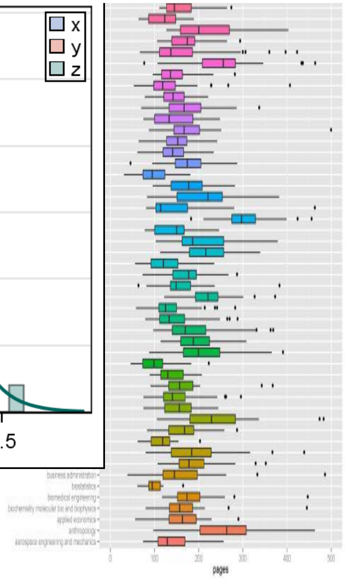
IDC, Herzeliya

Independence and variations, convolution, computer age statistics



0010011101010100101010100100100010
1010100010101111101011010011001001
1110101010011010010110010110110010





Third central moment of Poisson (\lambda)

Sample/Coupon collection

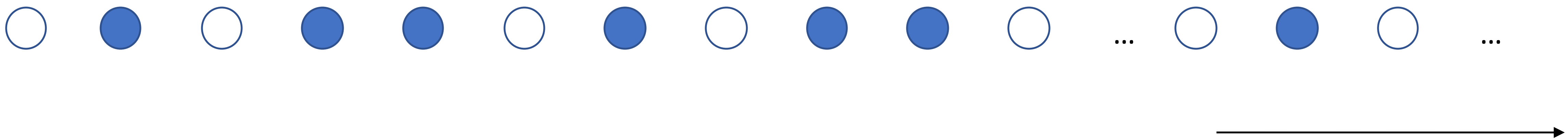
The RV T counts the number of observations required to see at least m=1 users from each country of the n=100. How many visits will it take if every visit comes from each of the countries with equal probabilities and independent of all previous visits?

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

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.

The Geometric distribution

$\omega \epsilon \Omega$:



Continue to infinity …

$$X(\omega) = \text{time of first success}$$

$$X \sim Geom(p)$$

$$P(X = k) = ?$$

Sample/Coupon collection

We saw

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

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:

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

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

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

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

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

$$\geq P(T \geq E(T)+\lambda)$$

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

Sample/Coupon collection

What about the Variance of T?

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

$X_i \sim Geom\left(p_i = \frac{100-i+1}{100}\right)$ 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
```

Computer age statistics

Let T_N denote the waiting time for full single coupon collection with N different equiprobable coupon types

5.A

Write code to compute the exact value of $E(T_N)$

```
def single_coupon_probabilities(n):
    return [(n - i) / float(n) for i in range(n)]

def single_coupon_mean(n):
    """
    Returns the mean of the single coupon problem, i.e. E(T_N).
    """
    return sum([1.0 / p for p in single_coupon_probabilities(n)])
```

5.B

Write code to compute the exact value of $V(T_N)$

```
def single_coupon_variance(n):
    """
    Returns the variance of the single coupon problem, i.e. V(T_N).
    """
    return sum([(1.0 - p) / (p ** 2) for p in single_coupon_probabilities(n)])
```

Taking $n = 100$:

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

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

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

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.

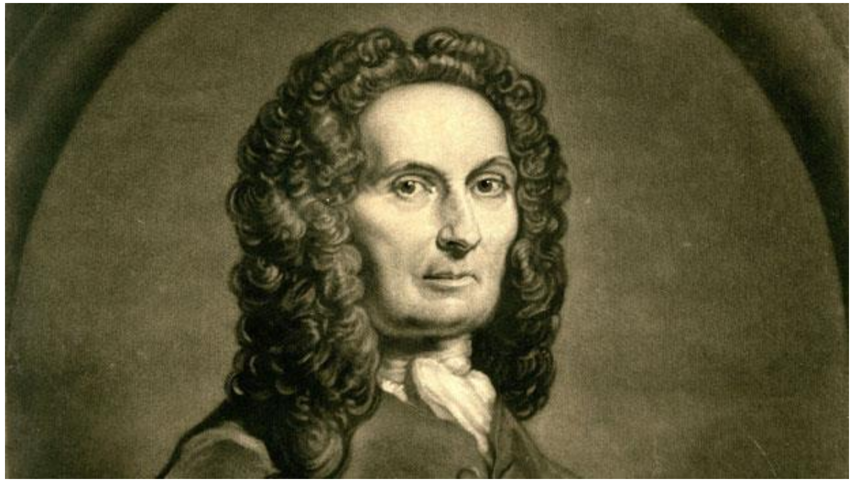
Exact coupon collector waiting time

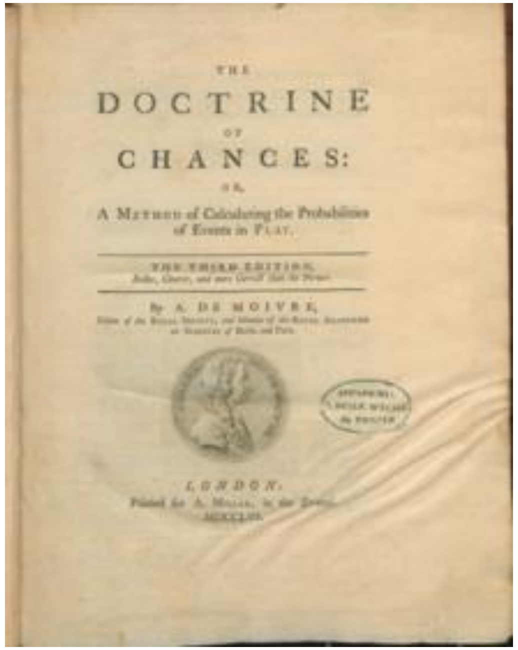
```
exact P(T_100>718) = 0.12
exact P(T_100>818) = 0.05
exact P(T_100>918) = 0.02
```

Comparison slide

	Chebichef w bound	**Strong Chebichef**	**Convolution (exact)**
P(T_100>718)	0.41	0.3958	0.12
P(T_100 > 818)	0.18	0.1759	0.05
P(T_100 > 918)	0.10	0.0989	0.02

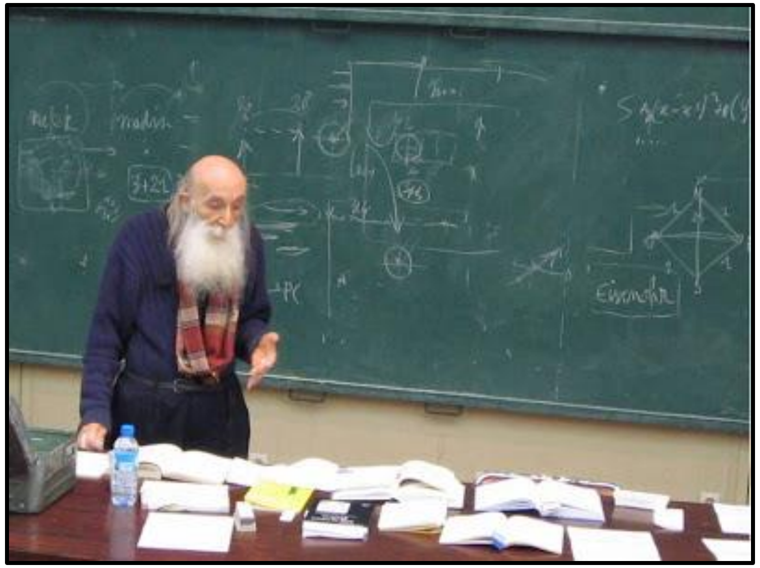
Sample/coupon collector and computers: Historical notes





- 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.



Collect 80% of the coupons

Independence – a broader point of view

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

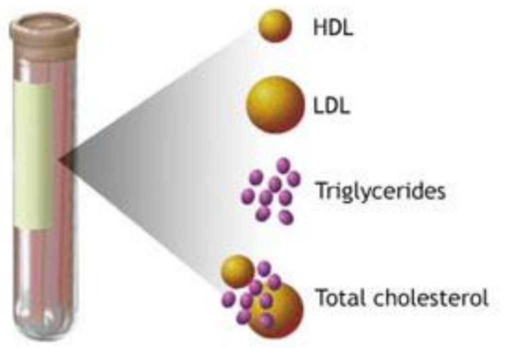
$X + Y + Z$ vs $Binom(0.5,3)$

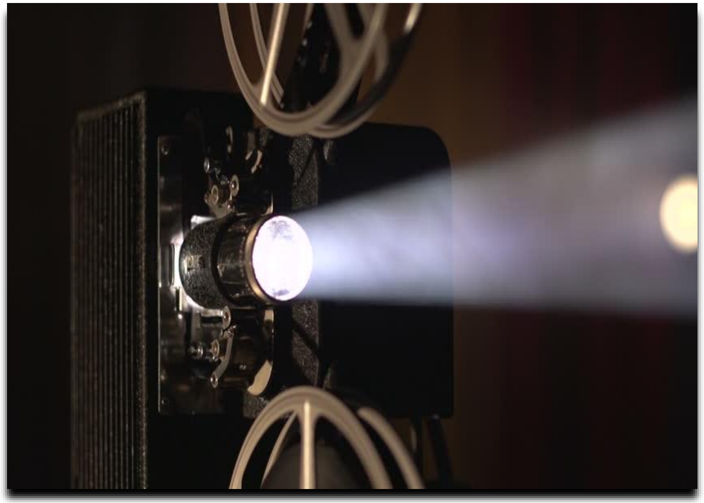
$$E(X + Y + Z) = ?$$
$$V(X + Y + Z) = ?$$

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

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:

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

Independent but not conditionally independent?

Conditionally independent but not independent?

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

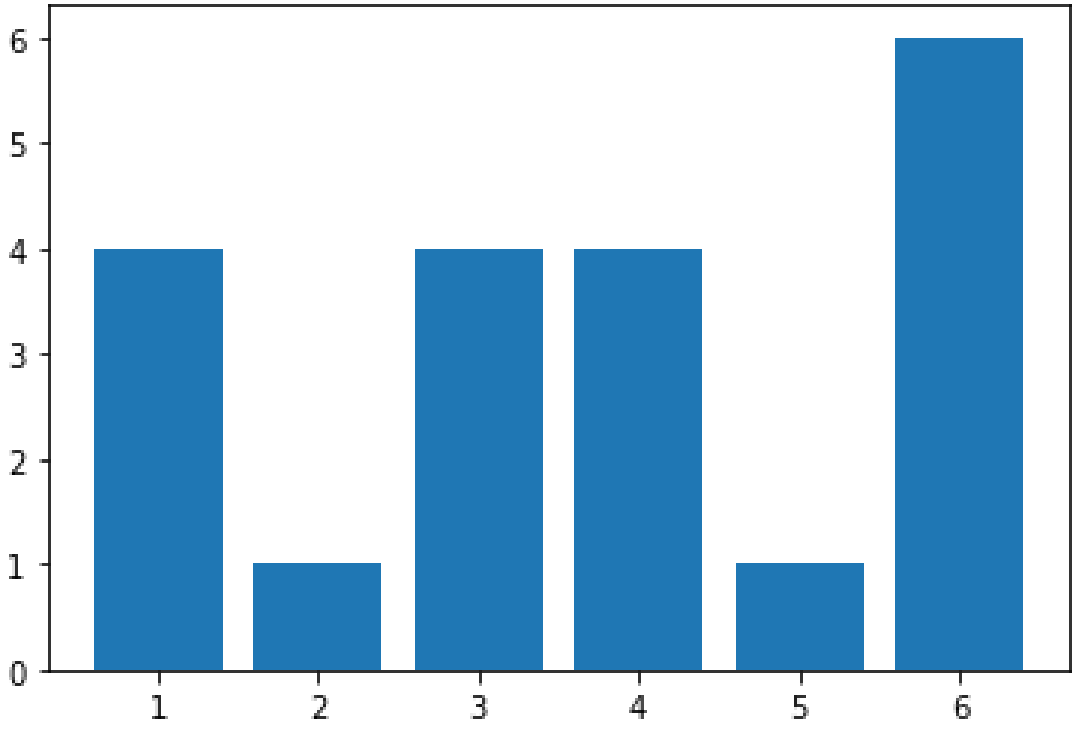
Now roll a die n times and define $X = (X_1, X_2, \ldots, X_6)$ where X_i counts the number of i's

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

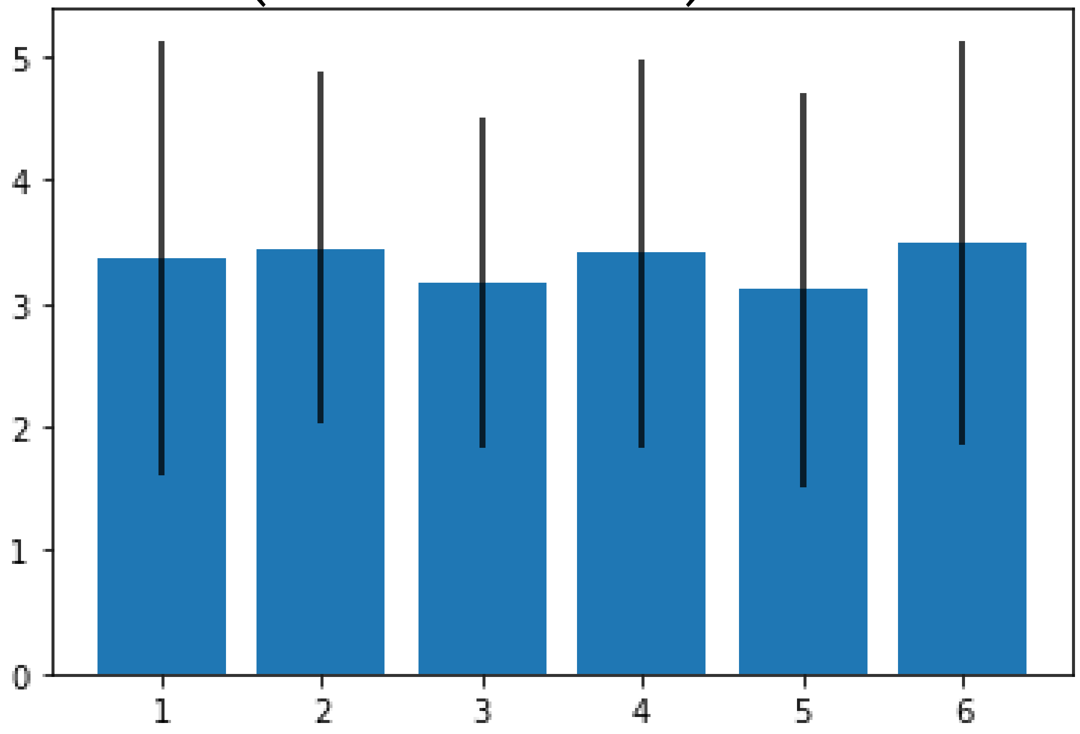
- What is d?
- What is p?
- What is $\mathrm{E}(X_i)$?
- What is $\mathrm{V}(X_i)$?

Multinomial Distribution

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



One experiment



100 repeated experiments, with averaging

Multinomial Distribution

We just defined: 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 $X \sim \mathrm{MNom}(N, P)$, $X = (X_1, X_2, \ldots, X_d)$.

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

Now observe that $X_i + X_j \sim \mathrm{Binom}(N, p_i + p_j)$ and therefore, from the above identity we get:

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

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

$$= -2Np_ip_j$$

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

Summary

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