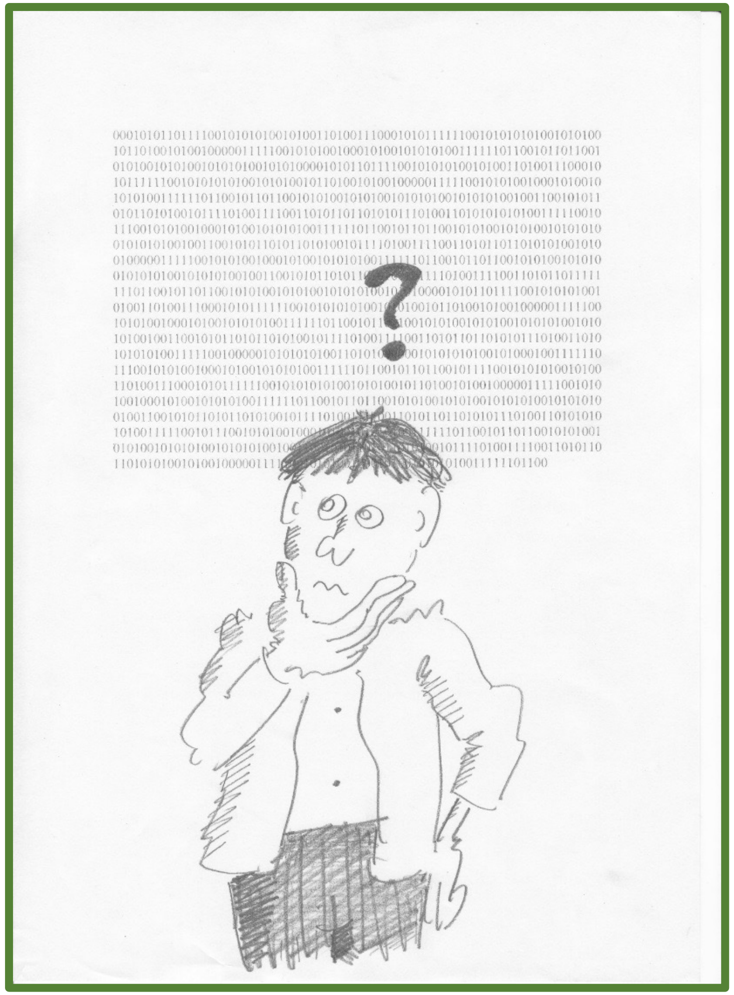
Statistics and data analysis

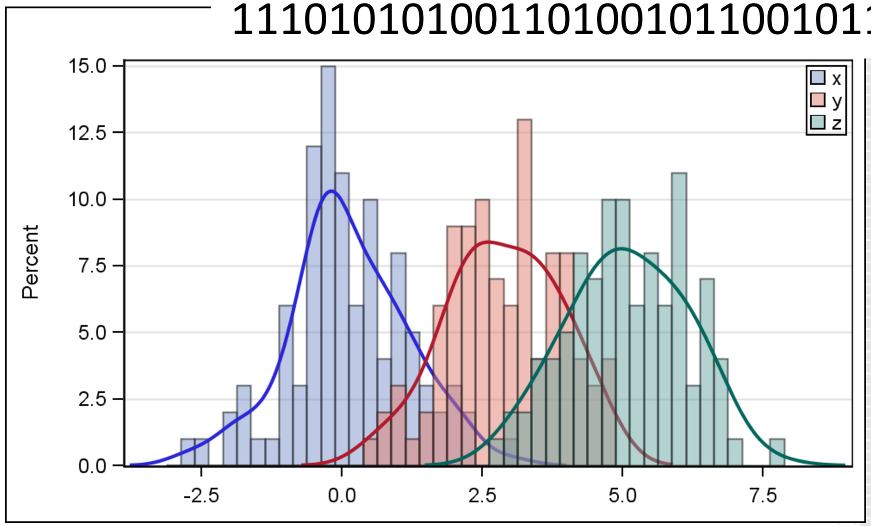
Zohar Yakhini

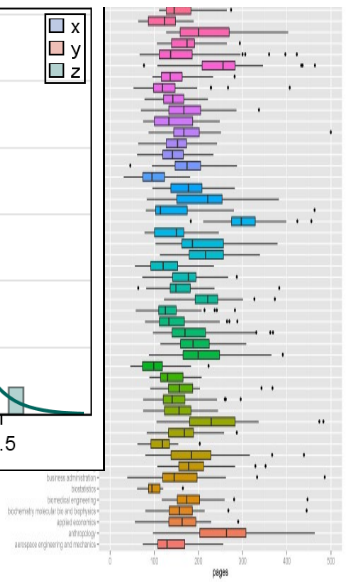
IDC, Herzeliya



More distributions, independence

0010011101010100101010100100100010
1010100010101111101011010011001001
1110101010011010010110010110110010





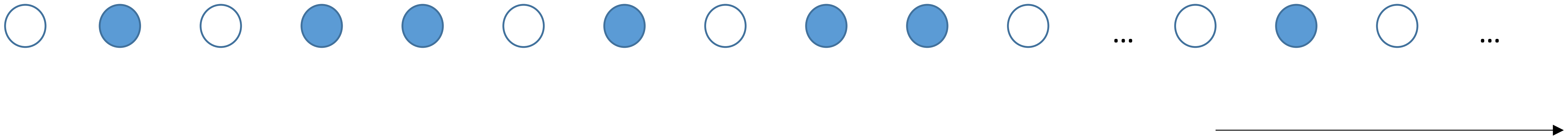
Binomial Distribution

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$\omega \epsilon \Omega$:

The Geometric distribution

$\omega \epsilon \Omega$:



Continue to infinity …

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

$$X \sim Geom(p)$$

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

Geometric Distribution – Expectation and variance

$$E(Y)=\sum_{y=1}^{\infty} y\left[q^{y-1}p\right]=p\sum_{y=1}^{\infty}\frac{dq^{y}}{dq}=p\frac{d}{dq}\sum_{y=1}^{\infty}q^{y}=p\frac{d}{dq}\left[q\sum_{y=1}^{\infty}q^{y-1}\right]=$$

$$=p\frac{d}{dq}\left[\frac{q}{1-q}\right]=p\left[\frac{(1-q)(1)-q(-1)}{(1-q)^2}\right]=\frac{p\left((1-q)+q\right)}{(1-q)^2}=\frac{p}{p^2}=\frac{1}{p}$$

$$E\left(Y(Y-1)\right)=\sum_{y=1}^{\infty} y(y-1)\left[q^{y-1}p\right]=pq\sum_{y=1}^{\infty}\frac{d^2q^{y}}{dq^2}=pq\frac{d^2}{dq^2}\sum_{y=1}^{\infty}q^{y}=pq\frac{d^2}{dq^2}\left[q\sum_{y=1}^{\infty}q^{y-1}\right]=$$

$$=pq\frac{d^2}{dq^2}\left[\frac{q}{1-q}\right]=pq\frac{d}{dq}\frac{1}{(1-q)^2}=pq\left(-2(1-q)^{-3}(-1)\right)=\frac{2pq}{\left(1-q\right)^3}=\frac{2pq}{p^3}=\frac{2q}{p^2}$$

$$\Rightarrow E\left(Y^2\right)=E\left(Y(Y-1)\right)+E(Y)=\frac{2q}{p^2}+\frac{1}{p}=\frac{2(1-p)+p}{p^2}=\frac{2-p}{p^2}$$

$$\Rightarrow V(Y)=E\left(Y^2\right)-[E(Y)]^2=\frac{2-p}{p^2}-\left[\frac{1}{p}\right]^2=\frac{2-p-1}{p^2}=\frac{1-p}{p^2}=\frac{q}{p^2}$$

$$\Rightarrow \sigma=\sqrt{\frac{q}{p^2}}$$

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 k we should have exactly $r - 1$ successes in first $k - 1$ trials, followed by a success

  - $P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$
  - $E(X) = \frac{r}{p}$
  - $V(X) = \frac{r(1-p)}{p^2}$

IDC HERZLIYA

...

Binomial rate vs n

Consider

$X_1 \sim \mathrm{Binom}(1, \lambda)$ and $X_2 \sim \mathrm{Binom}(2, \lambda/2)$

Which is larger:

$P(X_1 \geq 1)$ or $P(X_2 \geq 1)$?

$E(X_1)$ or $E(X_2)$?

Poisson – a limit of binomials with an increasing n and a fixed mean

Consider repeated coin tossing with increasingly smaller success rates

$X_1 \sim Binom(1, \lambda)$

$X_2 \sim Binom(2, \lambda/2)$

$X_3 \sim Binom(3, \lambda/3)$

$$P(X_1 \geq 1) = P(X_1 = 1) = \lambda$$

$$P(X_2 \geq 1) = 1 - P(X_2 = 0)$$

$$= 1 - (1 - \lambda/2)^2 = \lambda - (\lambda/2)^2 < \lambda$$

IDC HERZLIYA

Poisson – a limit of binomials with an increasing n and a fixed mean

$X_n \sim Binom(n, \lambda/n)$

$$P(X_n = k) = \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{n(n-1)\cdots(n-k+1)}{n^k}\cdot\frac{\lambda^k}{k!}\cdot\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}$$

as $n \to \infty$

$$\frac{n(n-1)\cdots(n-k+1)}{n^k} \to 1 \qquad \left(1-\frac{\lambda}{n}\right)^n \to \boxed{} \qquad \left(1-\frac{\lambda}{n}\right)^{-k} \to 1$$

IDC HERZLIYA

Poisson – a limit of binomials with an increasing n and a fixed mean

So,

$\forall k = 0, 1 \dots$

we have

$$P(X_n = k) \xrightarrow[n \to \infty]{} e^{-\lambda} \frac{\lambda^k}{k!}$$

IDC HERZLIYA

Poisson Distribution

- X ~ Poisson(λ) if

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

- Example – a website receives visits distributed as Poisson(0.5) per second.
- What is the probability of no visits at a certain second?
- Answer: exp(-0.5)
- What is the probability of no visits in a stretch of 10 seconds?
- Compute in two ways:
  - Poisson(5) yields exp(-5)
  - 10 independent as above yields (exp(-0.5))^10 = exp(-5)

IDC HERZLIYA

Poisson Distribution

Distribution often used to model the number of incidences in some characteristic unit of time or space:

- Arrivals of customers to a store within one hour
- Numbers of flaws in a roll of fabric of a given length
- Number of visitors to a website in one minute
- Number of calls to a service center in 10 mins

IDC HERZLIYA

Poisson Distribution – Expectation and Variance

$$f(y)=\frac{e^{-\lambda}\lambda^{y}}{y!}\qquad y=0,1,2,\ldots$$

$$E(Y)=\sum_{y=0}^{\infty}y\left[\frac{e^{-\lambda}\lambda^{y}}{y!}\right]=\sum_{y=1}^{\infty}y\left[\frac{e^{-\lambda}\lambda^{y}}{y!}\right]=\sum_{y=1}^{\infty}\frac{e^{-\lambda}\lambda^{y}}{(y-1)!}=\lambda e^{-\lambda}\sum_{y=1}^{\infty}\frac{\lambda^{y-1}}{(y-1)!}=\lambda e^{-\lambda}e^{\lambda}=\lambda$$

$$E\left(Y(Y-1)\right)=\sum_{y=0}^{\infty}y(y-1)\left[\frac{e^{-\lambda}\lambda^{y}}{y!}\right]=\sum_{y=2}^{\infty}y(y-1)\left[\frac{e^{-\lambda}\lambda^{y}}{y!}\right]=\sum_{y=2}^{\infty}\frac{e^{-\lambda}\lambda^{y}}{(y-2)!}=$$

$$=\lambda^{2}e^{-\lambda}\sum_{y=2}^{\infty}\frac{\lambda^{y-2}}{(y-2)!}=\lambda^{2}e^{-\lambda}e^{\lambda}=\lambda^{2}$$

$$\Rightarrow E\left(Y^{2}\right)=E\left(Y(Y-1)\right)+E(Y)=\lambda^{2}+\lambda$$

$$\Rightarrow V(Y)=E\left(Y^{2}\right)-\left[E(Y)\right]^{2}=\lambda^{2}+\lambda-[\lambda]^{2}=\lambda$$

$$\Rightarrow \sigma=\sqrt{\lambda}$$

Statistical independence

Example – Rolling 2 Dice (Red/Green)

Ω = All possible outcomes, that is:

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

- Assuming that all outcomes have P = 1/36 is based on assuming that the result of one dice DOES NOT AFFECT the rolling of the other in any way.
- What is the probability of G = 3 or 6 and R = 5?
- P(G = 3 or 6) = 1/3
- P(R = 5) = 1/6
- The probability of the JOINT event is, assuming 1/36 in each entry, 1/36+1/36 = 1/18.
- This is just the product of the two probabilities: P(G = 3 or 6 and R = 5) = 1/3 * 1/6 = 1/18
- This is called STATISTICAL INDEPENDENCE.
- When we defined 1/36 in every entry we imply that the two rolls are independent random variables

IDC HERZLIYA

Definitions and factoids …

- Two events (subsets of the sample space Ω), A and B, are said to be statistically independent if the occurrence of one doesn't affect the occurrence of the other:

  $P(A \mid B) = P(A)$, where $P(A \mid B) = P(A \cap B)/P(B)$ is the conditional probability of A given B.

- Form here we get

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A \cap B)P(B)}{P(A)P(B)}$$

$$= P(A|B)\frac{P(B)}{P(A)} = P(B)$$

- Show that from here it follows that $P(A \mid B) = P(A \mid \neg B)$
- It also clearly follows that $P(A \cap B) = P(A)P(B)$

Independent random variables

- Two random variables X and Y, defined over the same space Ω have a joint distribution $p(x, y)$.
- They also have marginal distributions
- The same marginal can often be joined (or coupled) in very different ways. The independent copula is only one of them.
- They are called independent if for all numbers x and y we have
$$P(X = x \text{ and } Y = y) = P(X = x) \cdot P(Y = y)$$
- Or – for all x and y as above,
the events $P(X = x)$ and $P(Y = y)$ are independent.
- If X and Y are independent then $E(XY) = E(X) \cdot E(Y)$
(prove this …)
- Is the opposite true?

Linearity of expected values

- $E(X+Y) = E(X) + E(Y)$
- This is true for ANY random variables. They don't have to be independent.
- This generalizes to any sums.

Sample/Coupon collection

- A website is seeking information about users from 100 different cities.
- It needs to observe the action of m users from each city to perform the analysis.
- How many visits will it take if every visit comes from each of the cities with equal probabilities and independent of all previous visits?
- On average?

Poll: 100 countries and m=1

IDC HERZLIYA

Sample/Coupon collection

At this point we will compute the expected value for the case $m = 1$.

We define random variables $X_i,\ i = 1 \dots 100,$ as follows.
Let X_1= the number of visits until the first country is in ($X_1 == 1$)
Let X_2 = the number of visits, after the first country is in, until the second country is also in

…

Let X_i = the number of visits, after the first $i-1$ countries are in, until the i-th country is also in.

Now let

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

Sample/Coupon collection

We are, of course, interested in

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

Now note that $X_i \sim Geom(p = (100 - i + 1)/100)$ and we therefore have $E(X_i) = \frac{1}{p} = \frac{100}{100-i+1}$

So:

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

<u>In general, for n types of coupons, we have $E(T) = nH(n) \sim n \ln n.$</u>

General m and unequal probabilities require a more complex treatment.

IDC HERZLIYA

$$\boxed{Var(X+Y)}$$

IDC HERZLIYA

Covariance

- Consider X and Y defined on the same sample space Ω
- $Cov(X,Y) = E(\,(X - \mu(X))(Y - \mu(Y)))$
- $Cov(X,Y) = \ E(XY) - \ E(X)E(Y)$
- When X and Y are independent, what is $Cov(X,Y)$?
- Is the opposite true?

Binomial Distribution – Variance and S.D.

$$f(y)=\frac{n!}{y!(n-y)!}p^y q^{n-y} \quad y=0,1,\ldots,n \quad q=1-p$$

Note: $E(Y^2)$ is difficult (impossibl e?) to get, but $E(Y(Y-1))=E(Y^2)-E(Y)$ is not :

$$E(Y(Y-1))=\sum_{y=0}^{n} y(y-1)\left[\frac{n!}{y!(n-y)!}p^y q^{n-y}\right]=\sum_{y=2}^{n} y(y-1)\left[\frac{n!}{y!(n-y)!}p^y q^{n-y}\right]$$

(Summand $=0$ when $y=0,1$)

$$\Rightarrow E(Y(Y-1))=\sum_{y=2}^{n}\frac{n!}{(y-2)!(n-y)!}p^y q^{n-y}$$

Let $y^{**}=y-2 \Rightarrow y=y^{**}+2$ Note: $y=2,\ldots,n \Rightarrow y^{**}=0,\ldots,n-2$

$$\Rightarrow E(Y(Y-1))=\sum_{y^{**}=0}^{n-2}\frac{n(n-1)(n-2)!}{y^{**}!\left(n-(y^{**}+2)\right)!}p^{y^{**}+2}q^{n-(y^{**}+2)}=n(n-1)p^2\sum_{y^{**}=0}^{n-2}\frac{(n-2)!}{y^{**}!\left((n-2)-y^{*}\right)!}p^{y^{**}}q^{(n-2)-y^{**}}=$$

$$=n(n-1)p^2(p+q)^{n-2}=n(n-1)p^2\left(p+(1-p)\right)^{n-2}=n(n-1)p^2$$

$$\Rightarrow E(Y^2)=E(Y(Y-1))+E(Y)=n(n-1)p^2+np=np[(n-1)p+1]=n^2p^2-np^2+np=n^2p^2+np(1-p)$$

$$\Rightarrow V(Y)=E(Y^2)-[E(Y)]^2=n^2p^2+np(1-p)-(np)^2=np(1-p)$$

$$\Rightarrow \sigma=\sqrt{np(1-p)}$$

IDC HERZLIYA

Or: linearity of variance for independent variables

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:

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

IDC HERZLIYA

Sum of two independent Poissons is Poisson

Sum of 2 indpt Poissons

$$P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

$$P(Y=k) = e^{-\mu}\frac{\mu^k}{k!}$$

X and Y indpt. Let $Z = X+Y$

$$P(Z=k) = \sum_{i=-\infty}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!}\cdot e^{-\mu}\frac{\mu^{k-i}}{(k-i)!}$$

Here summands are 0 when either of the denominator factorials are negative

$$= e^{-(\lambda+\mu)}\cdot\frac{1}{k!}\sum_{i=0}^{k}\binom{k}{i}\lambda^i\mu^{k-i}$$

$$= e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^k}{k!}$$

Higher moments

The raw kth moment of a random variable X is $E(X^k)$
The central kth moment of a random variable X is $E((X-\mu(X))^k)$

Let $X \sim \mathrm{Binom}(n,p)$. What is the 3^{rd} central moment of X ?

$X = \sum_{i=1}^{n} X_i$, where $X_i \sim \mathrm{Ber}(p)$, independent.

$$\gamma_3 = E\left[\left(\sum_{i=1}^{n}(X_i - p)\right)^3\right] = E\left[\sum_{i,j,k=1 \dots n}(X_i-p)(X_j-p)(X_k-p)\right]$$

$$= \sum_{i,j,k=1 \dots n} E((X_i-p)\big(X_j-p\big)(X_k-p))$$

The terms of the last summation are all 0 except when $i = j = k$. Therefore:

$$\gamma_3 = nE\big((X_1-p)^3\big) = n(p(1-p)^3 + (1-p)(-p)^3) .$$

And, after further simplification: $\gamma_3 = np(1-p)(1-2p)$

Mutual independence vs k-wise independence

Summary

- Geometric distribution
- Negative binomials (next week: how to compare them)
- Poisson distribution
- Coupon collector
- Independence and the covariance of two random variables
- Convolution of pdfs (to be continued)
- Higher moments and an example
- Mutual indpce vs lower order indpce