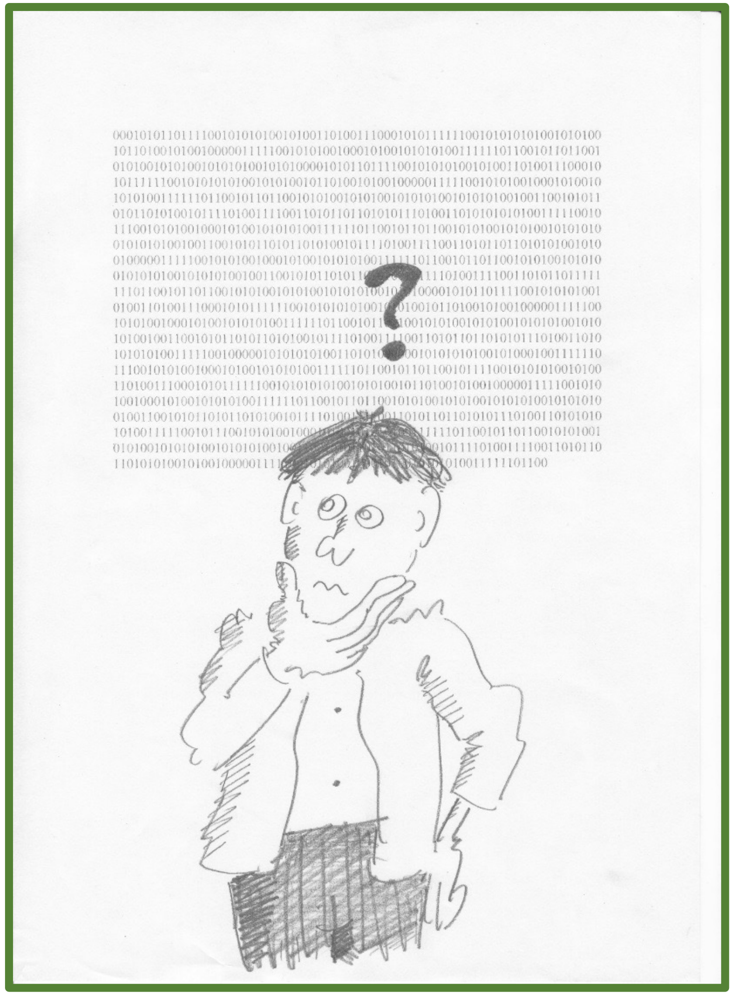
Statistics and data analysis

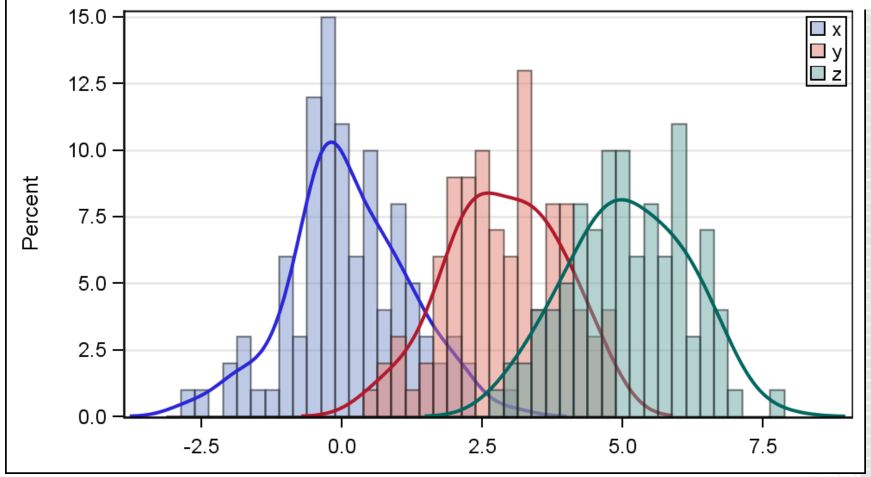
Zohar Yakhini, Leon Anavy, Ben Galili

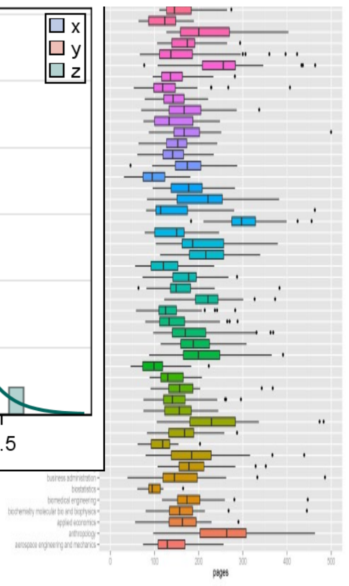
IDC, Herzeliya

Distributions and a brief review of basic concepts



0010011101010100101010100100100010
1010100010101111101011010011001001
1110101010011010010110010110110010





Basics: (Ω, P)

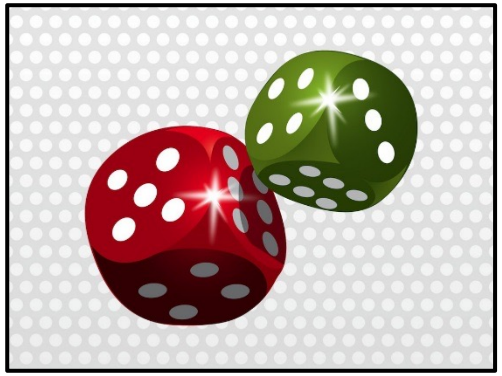
$\Omega =$ the probability sample space

A collection (an algebra) of measurable events – sets of samples

A probability measure, P – assigns a probability to every measurable set of samples

The measure P is additive for disjoint sets, it's non-negative and $P(\Omega) = 1$

Example – Rolling 2 Dice (Red/Green)



Ω = All possible outcomes

Measurable sets = all subsets of this finite space

Red\Green	1	2	3	4	5	6
1	1,1	1,2	1,3	1,4	1,5	1,6
2	2,1	2,2	2,3	2,4	2,5	2,6
3	3,1	3,2	3,3	3,4	3,5	3,6
4	4,1	4,2	4,3	4,4	4,5	4,6
5	5,1	5,2	5,3	5,4	5,5	5,6
6	6,1	6,2	6,3	6,4	6,5	6,6

Random Variables

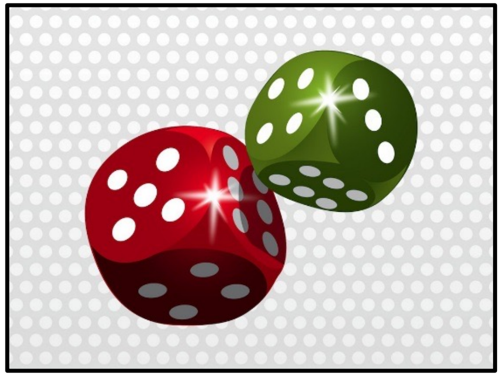
- A Random Variable (RV) is a <u>numerical function</u> defined on the probability sample space.
- For each element of a sample space, the random variable takes on exactly one value
- Random Variables are usually denoted using upper case letters (X, Y)
- Individual outcomes for an RV are usually denoted using lower case letters (x, y)

Example:

- Toss a coin 5 times.
- The sample space –
  Ω = all possible outcomes:
  00000, 00001, 00010, 00011, … , 01111, …, 11101, 11110, 11111
- One possible RV defined on this space – the outcome of the third toss
- Another : Y $=$ total number of 1s (what is $P(Y = 1)$ in this case?)
- Is the number of 1s prime? (a binary RV)
- Another : count how many 1s on even numbered tosses
- Count how many 11 runs

Example – Rolling 2 Dice (Red/Green)

Ω = All possible outcomes



Red\Green	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Probability Distributions

- Probability Distribution: Table, Graph, or Formula that describes values a random variable can take on, and their corresponding occurrence probabilities (discrete RV) or density (continuous RV)
- Discrete Probability Distribution: Assigns probabilities (masses) to the individual outcomes
- Continuous Probability Distribution: Assigns density at individual points, probability of ranges can be obtained by integrating density function
- Discrete probability distribution: $p(y) = P(Y = y)$
- Continuous densities are denoted by $f(y)$.
  We then have
$$P(Y \in I) = \int_I f(y)dy$$
- Cummulative Distribution Function: $F(y) = P(Y \leq y)$
- Probability distributions sum or integrate to 1.
- What values can the CDF, F, of a random variable take?

IDC HERZLIYA

Sum of 2 Dice – Probability Mass Function & CDF

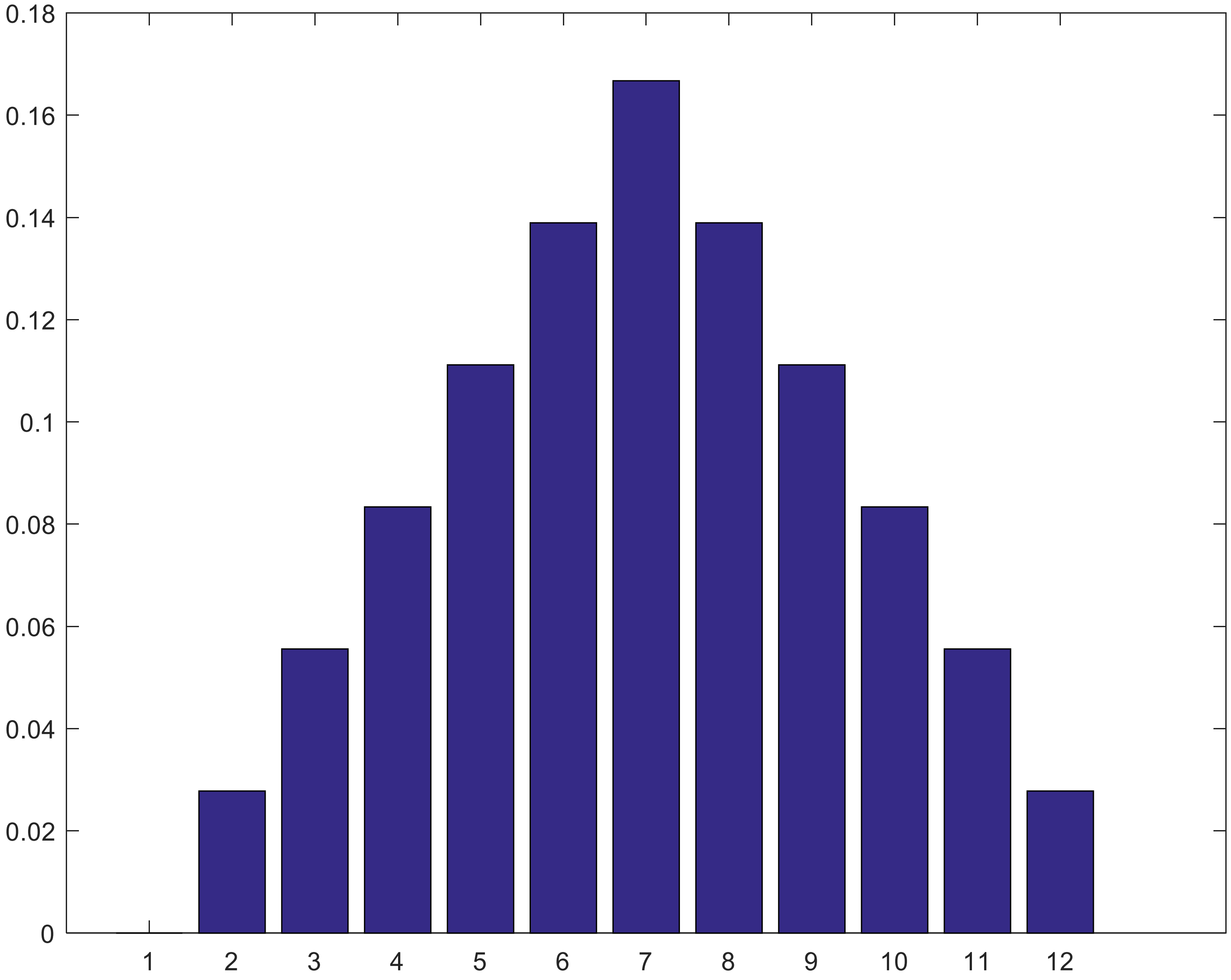
y	p(y)	F(y)
2	1/36	1/36
3	2/36	3/36
4	3/36	6/36
5	4/36	10/36
6	5/36	15/36
7	6/36	21/36
8	5/36	26/36
9	4/36	30/36
10	3/36	33/36
11	2/36	35/36
12	1/36	36/36

$$p(y) = \frac{\#\text{ of ways 2 dice can sum to } y}{\#\text{ of ways 2 dice configurations}}$$

$$F(y) = \sum_{t=2}^{y} p(t)$$

IDC HERZLIYA

Rolling 2 Dice – Probability Mass Function

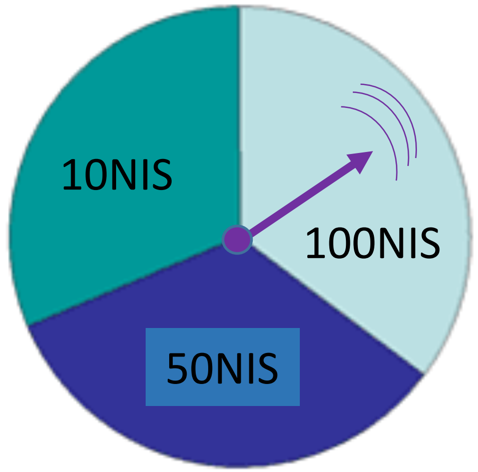


IDC HERZLIYA

Expected Values of Discrete RV's

- Mean (aka Expected Value) – the weighted average value an RV (or function of RV). Weighting is according to the underlying probability space.
- Variance – Average squared deviation between a realization of an RV (or function of RV) and its mean
- Standard Deviation – Positive Square Root of Variance (in same units as the data)
- Notation:
  - Mean: E(Y) = μ
  - Variance: Var(Y) = σ^2
  - Standard Deviation: σ

$$E(X) = \sum_{all\ relevant\ x} x\,p(x)$$



How much will we pay (or not) to play this game?

Expected Value and Variance of Discrete RV's

Mean : $E(Y)=\mu=\sum_{\text{all } y} y p(y)$

Mean of a function $g(Y)$: $E[g(Y)]=\sum_{\text{all } y} g(y)p(y)$

Variance : $V(Y)=\sigma^2=E\left[(Y-E(Y))^2\right]=E\left[(Y-\mu)^2\right]=$

$$=\sum_{\text{all } y}(y-\mu)^2 p(y)=\sum_{\text{all } y}\left(y^2-2y\mu+\mu^2\right)p(y)=$$

$$=\sum_{\text{all } y} y^2 p(y)-2\mu\sum_{\text{all } y} y p(y)+\mu^2\sum_{\text{all } y} p(y)=$$

$$=E\left[Y^2\right]-2\mu(\mu)+\mu^2(1)=E\left[Y^2\right]-\mu^2$$

Standard Deviation : $\sigma=+\sqrt{\sigma^2}$

IDC HERZLIYA

Expected Values of Linear Functions of Discrete RV's

Linear Functions : $g(Y) = aY + b \quad (a, b \equiv \text{constants})$

$$E[aY+b] = \sum_{\text{all } y} (ay+b)p(y) =$$

$$= a\sum_{\text{all } y} yp(y) + b\sum_{\text{all } y} p(y) = a\mu + b$$

$$V[aY+b] = \sum_{\text{all } y} \left((ay+b) - (a\mu+b)\right)^2 p(y) =$$

$$\sum_{\text{all } y} (ay - a\mu)^2 p(y) = \sum_{\text{all } y} \left[a^2 (y-\mu)^2\right] p(y) =$$

$$= a^2 \sum_{\text{all } y} (y-\mu)^2 p(y) = a^2\sigma^2$$

$$\sigma_{aY+b} = |a|\sigma$$

Example – Rolling 2 Dice

y	p(y)	yp(y)	y^2p(y)
2	1/36	2/36	4/36
3	2/36	6/36	18/36
4	3/36	12/36	48/36
5	4/36	20/36	100/36
6	5/36	30/36	180/36
7	6/36	42/36	294/36
8	5/36	40/36	320/36
9	4/36	36/36	324/36
10	3/36	30/36	300/36
11	2/36	22/36	242/36
12	1/36	12/36	144/36
Sum	36/36 =1.00	252/36 =7.00	1974/36= 54.833

$$\mu = E(Y) = \sum_{y=2}^{12} yp(y) = 7.0$$

$$\sigma^2 = E\left[Y^2\right] - \mu^2 = \sum_{y=2}^{12} y^2 p(y) - \mu^2$$

$$= 54.8333 - (7.0)^2 = 5.8333$$

$$\sigma = \sqrt{5.8333} = 2.4152$$

Expectation - another angle

Consider a probability space (Ω, P) and a rv $X\colon\ \Omega \to \mathbb{R}$

An equivalent definition of the expected value is:

$$E(X) = \sum_{\omega\epsilon\Omega} X(\omega)P(\omega)$$

A very important conclusion is:

$$E(X+Y) = \sum_{\omega\epsilon\Omega}\big(X(\omega)+Y(\omega)\big)P(\omega)$$

$$= \sum_{\omega\epsilon\Omega} X(\omega)P(\omega) + \sum_{\omega\epsilon\Omega} Y(\omega)P(\omega)$$

$$= E(X) + E(Y)$$

Linearity of expectations

Red\Green	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

IDC HERZLIYA

$$\forall \lambda > 0 \;\; P(|X - \mu| > \lambda) \leq \frac{V(X)}{\lambda^2}$$

Deviation from the mean

- Tchebysheff's theorem: Suppose Y is any random variable with mean μ and standard deviation σ. Then:
  $P(\mu - b\sigma \le Y \le \mu + b\sigma) \ge 1-(1/b^2)$ for $b > 0$
  - b=1: $P(\mu-1\sigma \le Y \le \mu+1\sigma) \ge 1-(1/1^2) = 0$ (trivial result)
  - b=2: $P(\mu-2\sigma \le Y \le \mu+2\sigma) \ge 1-(1/2^2) = 3/4$
  - b=3: $P(\mu-3\sigma \le Y \le \mu+3\sigma) \ge 1-(1/3^2) = 8/9$
- Note that this is a very conservative bound, but that it works for any distribution

- For Mound Shaped Distributions, aka Gaussians:
  - k=1: $P(\mu-1\sigma \le Y \le \mu+1\sigma) \approx 0.68$
  - k=2: $P(\mu-2\sigma \le Y \le \mu+2\sigma) \approx 0.95$
  - k=3: $P(\mu-3\sigma \le Y \le \mu+3\sigma) \approx 0.995$

IDC HERZLIYA

Proof of Tchebysheff's Theorem

Breaking real line into 3 parts :

$i)\ \ (\text{-}\infty,(\mu\text{-}k\sigma)^{-}]\qquad ii)\ \ [(\mu\text{-}k\sigma),(\mu+k\sigma)]\quad iii)\,[(\mu+k\sigma)^{+},\infty)$

Making use of the definition of Variance :

$$V(Y)=\sigma^2=\sum_{-\infty}^{\infty}(y-\mu)^2p(y)=$$

$$\sum_{-\infty}^{(\mu\text{-}k\sigma)^{-}}(y-\mu)^2p(y)+\sum_{(\mu\text{-}k\sigma)}^{(\mu+k\sigma)}(y-\mu)^2p(y)+\sum_{(\mu+k\sigma)^{+}}^{\infty}(y-\mu)^2p(y)$$

In Region $i)$: $y-\mu\le -k\sigma \Rightarrow (y-\mu)^2\ge k^2\sigma^2$

In Region $iii)$: $y-\mu\ge k\sigma \Rightarrow (y-\mu)^2\ge k^2\sigma^2$

$$\Rightarrow \sigma^2\ge k^2\sigma^2P(Y<\mu-k\sigma)+\sum_{(\mu\text{-}k\sigma)}^{(\mu+k\sigma)}(y-\mu)^2p(y)+k^2\sigma^2P(Y>\mu+k\sigma)$$

$$\Rightarrow \sigma^2\ge k^2\sigma^2P(Y<\mu-k\sigma)+k^2\sigma^2P(Y>\mu+k\sigma)=$$

$$=k^2\sigma^2\left[1-P(\mu-k\sigma\le Y\le\mu+k\sigma)\right]$$

$$\Rightarrow \frac{\sigma^2}{k^2\sigma^2}=\frac{1}{k^2}\ge\left[1-P(\mu-k\sigma\le Y\le\mu+k\sigma)\right]\Rightarrow P(\mu-k\sigma\le Y\le\mu+k\sigma)\ge 1-\frac{1}{k^2}$$

Discrete Uniform Distribution

- Suppose *Y* can take on any integer value between *a* and *b* inclusive, each equally likely (e.g. rolling a dice, where *a*=1 and *b*=6). Then *Y* follows the discrete uniform distribution.

$$f(y) = \frac{1}{b-(a-1)} \qquad a \le y \le b$$

$$F(y) = \begin{cases} 0 & y < a \\ \dfrac{\operatorname{int}(y)-(a-1)}{b-(a-1)} & a \le y < b \qquad \operatorname{int}(x) \equiv \text{integer portion of } x \\ 1 & y \ge b \end{cases}$$

$$E(Y) = \sum_{y=a}^{b} y\left(\frac{1}{b-(a-1)}\right) = \frac{1}{b-(a-1)}\left[\sum_{y=1}^{b} y - \sum_{y=1}^{a-1} y\right] = \frac{1}{b-(a-1)}\left[\frac{b(b+1)}{2} - \frac{(a-1)a}{2}\right] = \frac{b(b+1)-a(a-1)}{2(b-(a-1))}$$

$$E\left(Y^2\right) = \sum_{y=a}^{b} y^2\left(\frac{1}{b-(a-1)}\right) = \frac{1}{b-(a-1)}\left[\sum_{y=1}^{b} y^2 - \sum_{y=1}^{a-1} y^2\right] = \frac{1}{b-(a-1)}\left[\frac{b(b+1)(2b+1)}{6} - \frac{(a-1)a(2a-1)}{6}\right] =$$

$$= \frac{b(b+1)(2b+1)-a(a-1)(2a-1)}{6(b-(a-1))}$$

$$\Rightarrow V(Y) = E\left(Y^2\right) - [E(Y)]^2 = \frac{b(b+1)(2b+1)-a(a-1)(2a-1)}{6(b-(a-1))} - \left[\frac{b(b+1)-a(a-1)}{2(b-(a-1))}\right]^2$$

Note: When $a=1$ and $b=n$:

$$E(Y) = \frac{n+1}{2} \qquad V(Y) = \frac{(n+1)(n-1)}{12} \qquad \sigma = \sqrt{\frac{(n+1)(n-1)}{12}}$$

Bernoulli Distribution

- An experiment consists of one trial. It can result in one of 2 outcomes: Success or Failure (or a property being Present or Absent).
- Probability of Success ($Y = 1$) is p $(0 < p < 1)$
- Example: coin tossing

$$p(y) = \begin{cases} p & y = 1 \\ 1 - p & y = 0 \end{cases}$$

$$E(Y) = \sum_{y=0}^{1} y p(y) = 0(1-p) + 1p = p$$

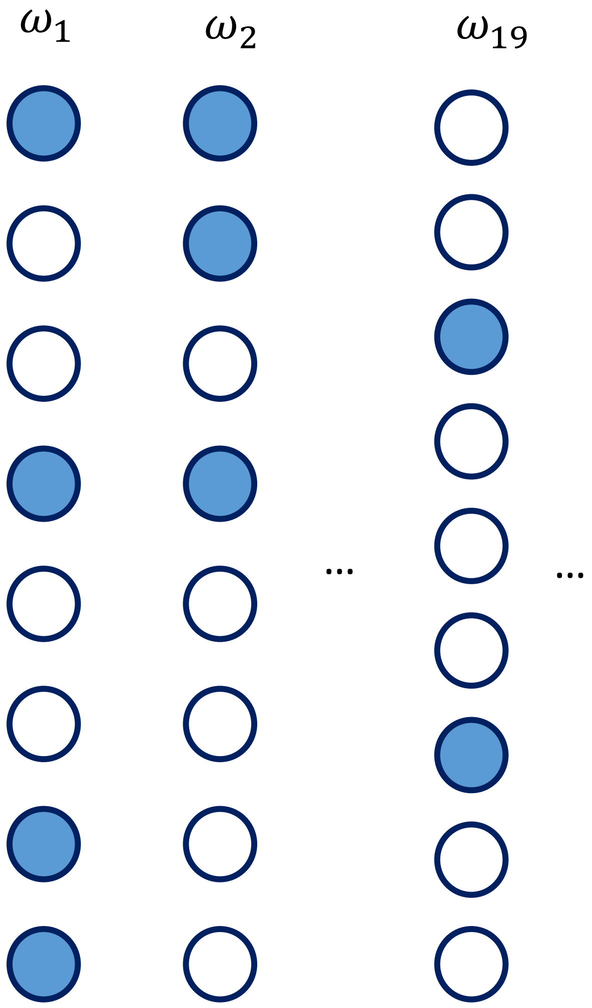
$$E\left(Y^2\right) = 0^2(1-p) + 1^2 p = p$$

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

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

IDC HERZLIYA

Binomial Distribution



- A binomial experiment consists of a series of *n* identical trials
- Consider all possible tossing trajectories. This is our probability space, Ω .
- Each trial is Bernoulli as above
- Trials are independent (outcome of one has no bearing on outcomes of others – formal definition next week)
- Probability of Success, p, is constant for all trials
- The random variable Y which counts the number of Successes in the n trials is said to follow a **<u>Binomial Distribution with parameters n and p</u>**
- Y can take on the values $y = 0,1, \dots , n$
- Notation: $Y \sim Binom(n, p)$

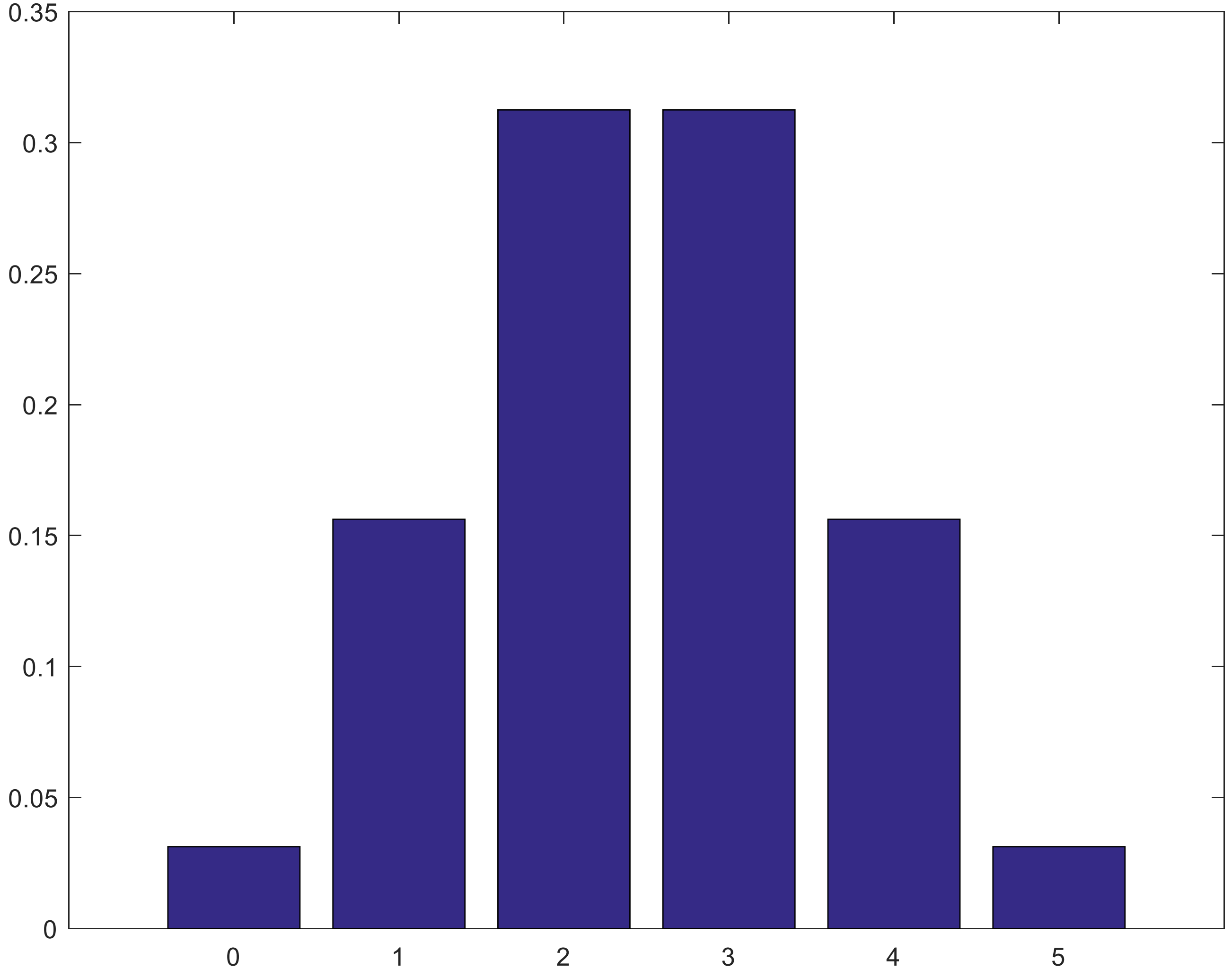
Binomial Distribution

- $P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- $\sum_{k=0}^{n} P(Y = k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + 1 - p)^n = 1$

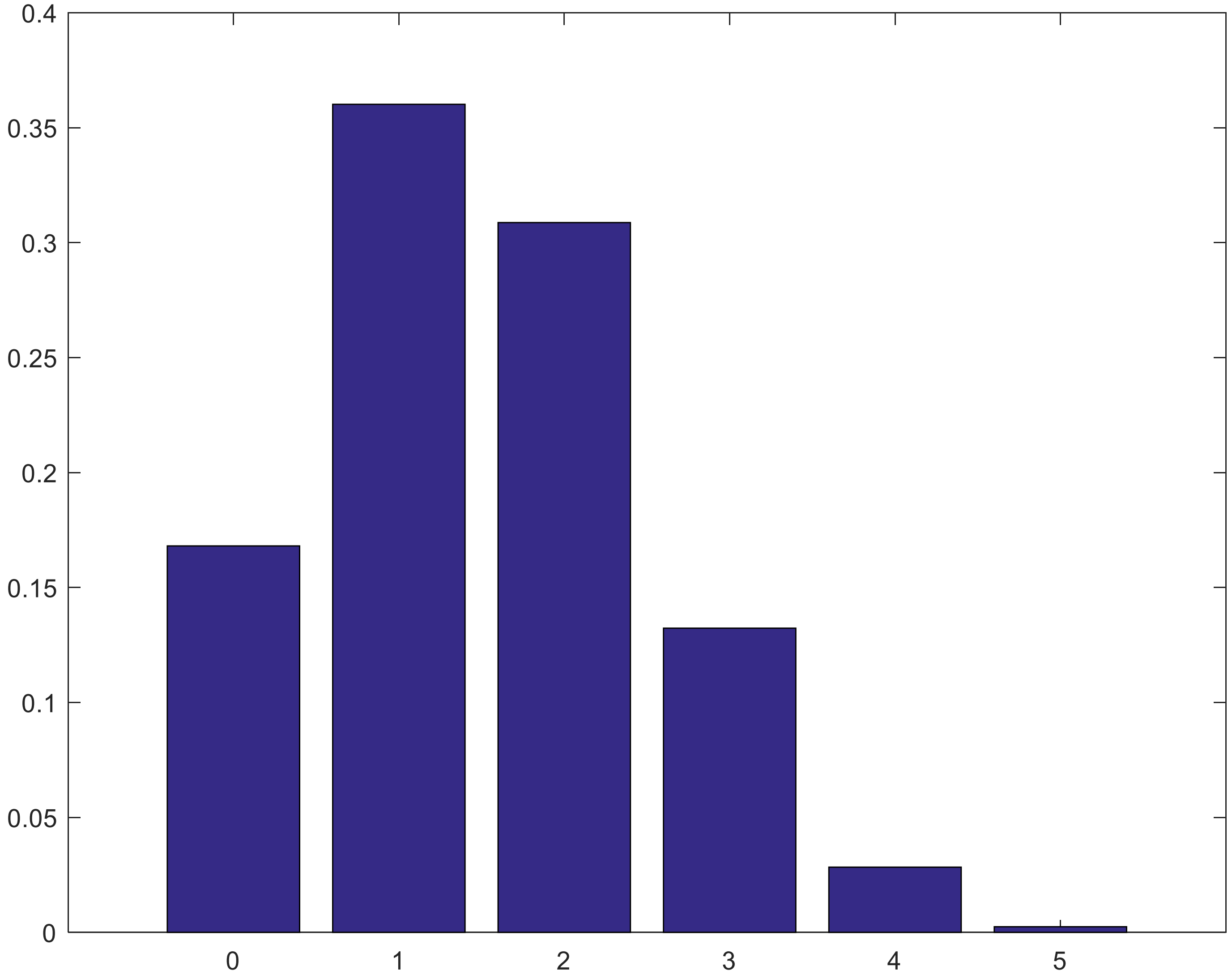
$\omega \epsilon \Omega$:

IDC HERZLIYA

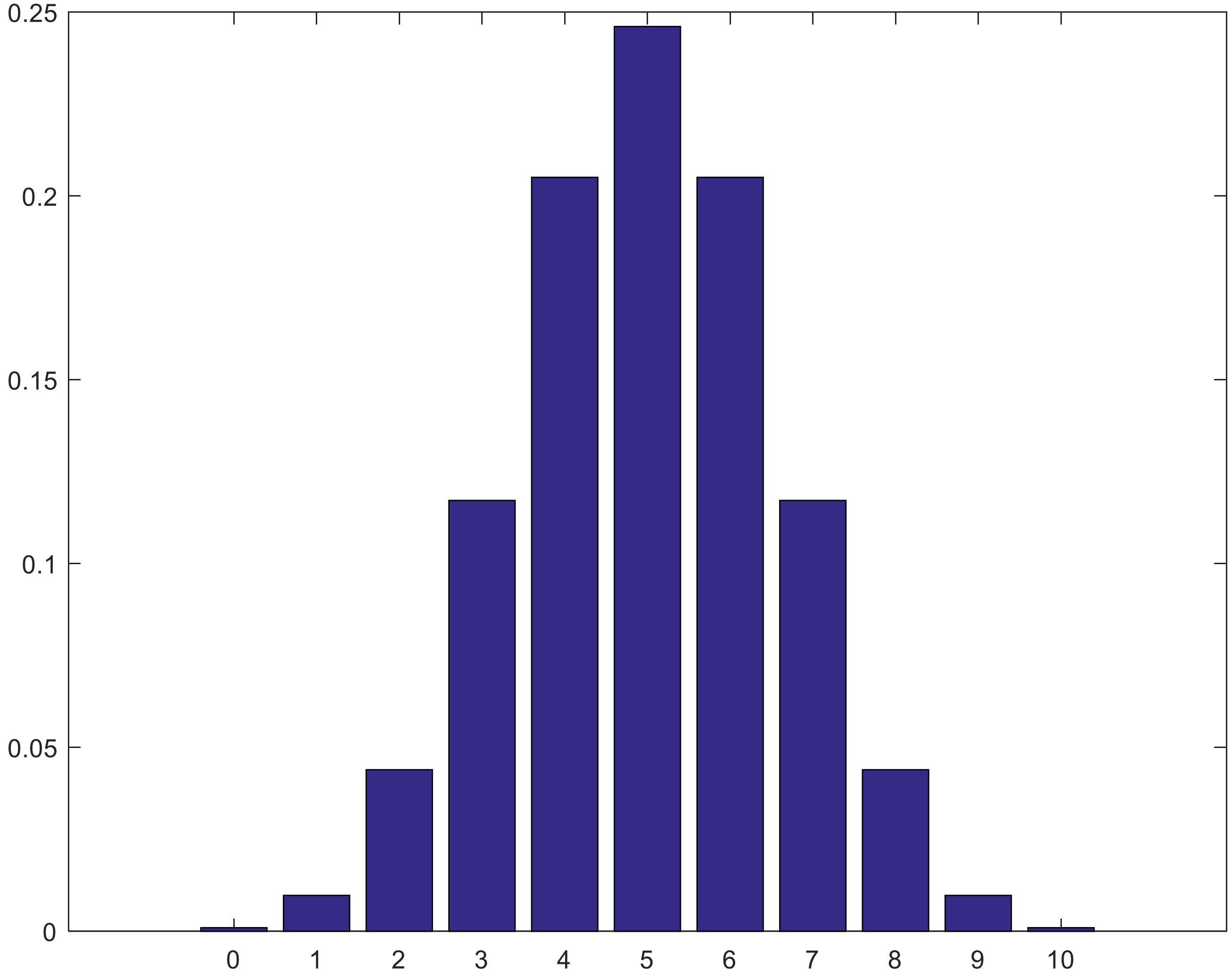
Binom(5,0.5) –
tossing a fair coin 5 times, counting successes



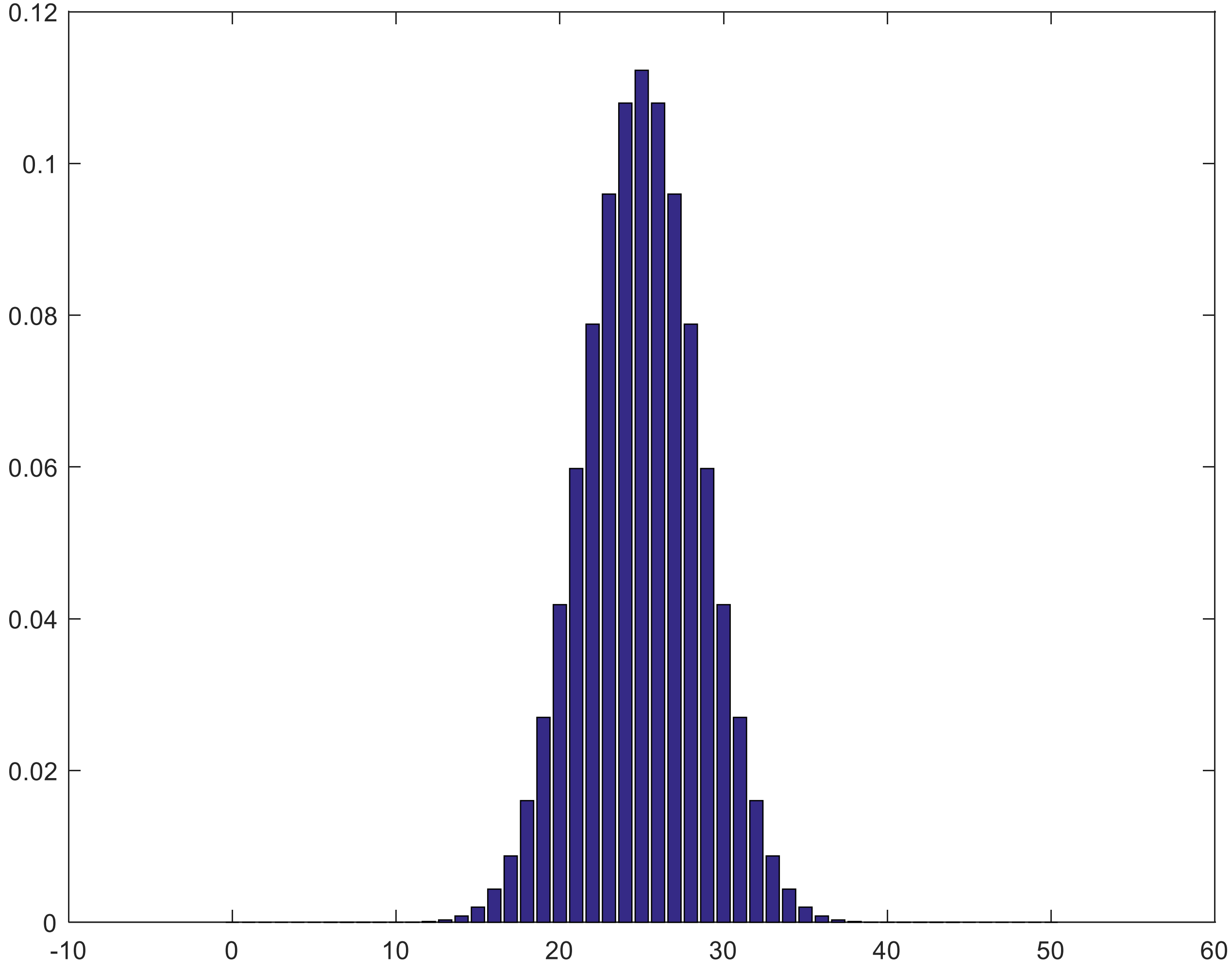
Binom(5,0.3)



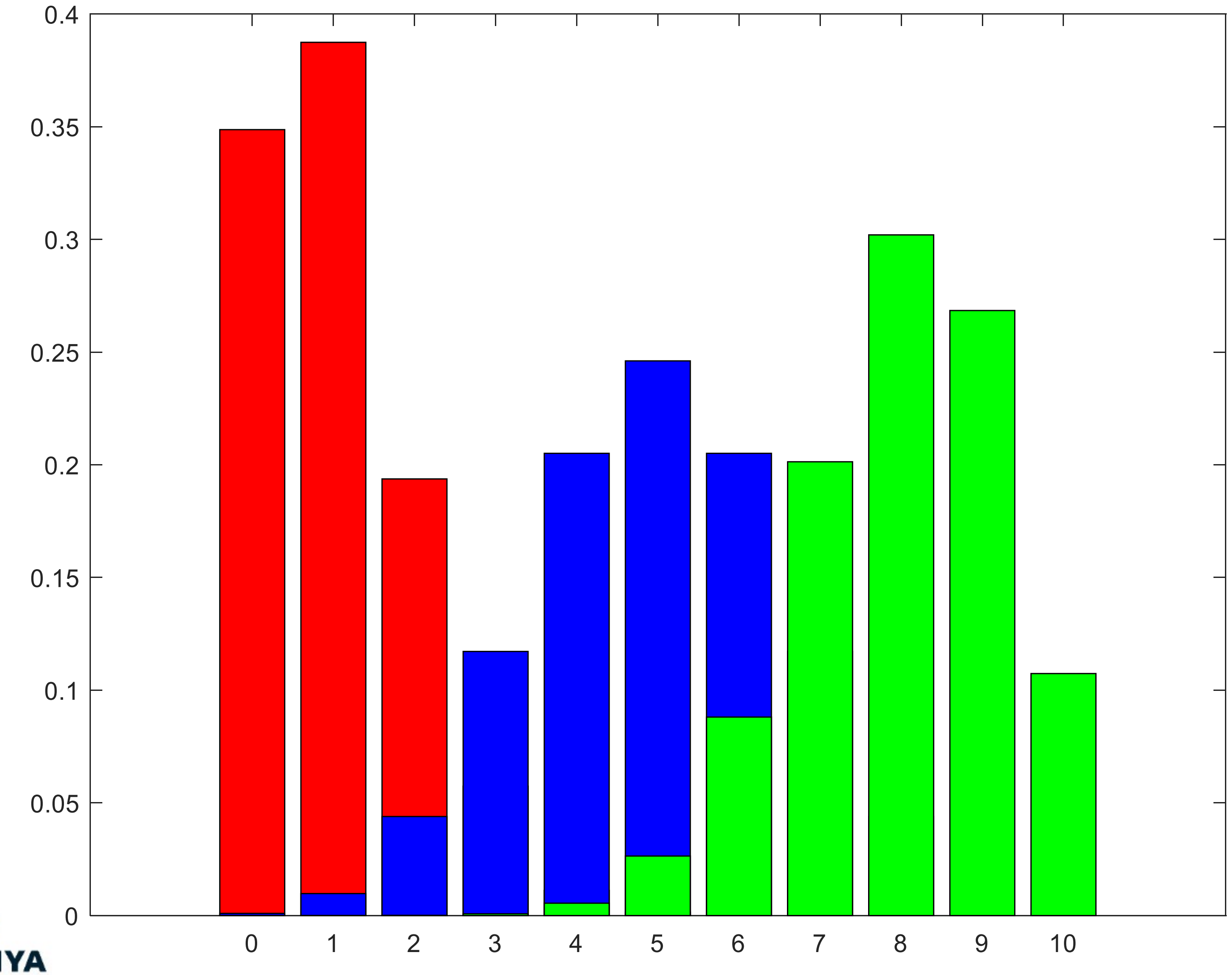
Binom(10,0.5) –
tossing a fair coin 10 times, counting successes



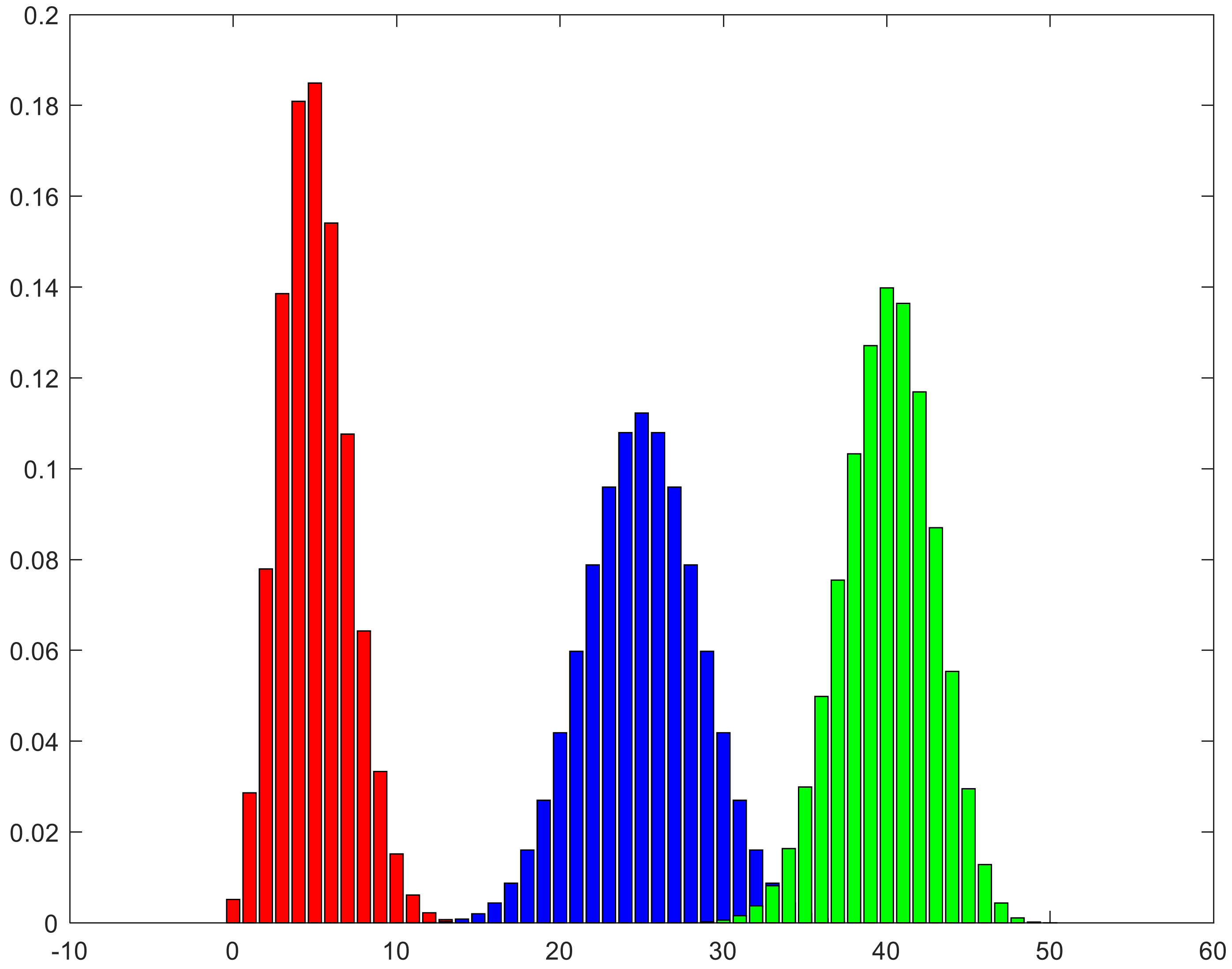
Binom(50,0.5) –
tossing a fair coin 50 times, counting successes



Tossing 10 coins w p = 0.1, 0.5, 0.8



Tossing 50 coins w p = 0.1, 0.5, 0.8



Binomial Distribution – Expected Value

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

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

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

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

Let $y^*=y-1 \Rightarrow y=y^*+1$ Note: $y=1,\ldots,n \Rightarrow y^*=0,\ldots,n-1$

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

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

Better way: linearity of expectations ....

$\omega \epsilon \Omega$:

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

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

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

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

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

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

Again: linearity of variance for independent variables

Using the binomial distribution. Example: Experimental treatment for Kidney Cancer

- Suppose we have $n = 40$ patients who will be receiving an experimental therapy (Tx) which is believed to be better than current treatments (standard of care = SoC).
  The latter has a historically derived 5-year survival rate of 20%. That is, under the SoC the probability of 5-year survival is $p = 0.2$
- We will now count 5-year survival under Tx and will then need to decide if we can confidently say that the new experimental treatment is better.

IDC HERZLIYA

Results and "The Question"

- Suppose that using the new treatment we find that 16 out of the 40 patients survive at least 5 years past diagnosis.
- Q: Does this result suggest that the new therapy, Tx, has a better 5-year survival rate than that of the SoC?
  That is:
  is the probability that a patient survives at least 5 years greater than 0.2 when treated using the new therapy?

IDC HERZLIYA

What do we consider in answering the question of interest?

We essentially ask ourselves the following:

- If we assume that new therapy is **no better** than the current then what is the probability of seeing the observed numbers? That is – how likely are they to occur, in such case, by chance alone?
- More specifically:
  What is the probability of seeing 16 <u>or more</u> successes out of 40 if the success rate of the new therapy is also 0.2?
- This is called estimating the **<u>p-value</u>** of the **<u>OBSERVED RESULT</u>** under the **<u>NULL model</u>**

IDC HERZLIYA

Binomial ….

- This is a binomial experiment situation…
  - There are n = 40 patients and we are counting the number of patients that survive 5 or more years.
  - The individual patient outcomes are independent and under the NULL MODEL the probability of success is *p = 0.2* for all patients.
    (that is: we assume that Tx is NOT better than the standard of care)

- So the random variable *X = # of "successes" in the clinical trial* is, under the NULL model, Binomial with *n = 40* and *p = 0.2,*

i.e., under the null: $\underline{X \sim Binomial(40{,}0.2)}$

Example: Treatment of Kidney Cancer - cont

- $X \sim BIN(40{,}0.2)$, find the probability that exactly 16 patients survive at least 5 years.

$$P(X=16)=\binom{40}{16}.20^{16}.80^{24}=.001945$$

- This requires some calculator gymnastics and some scratchwork (or a Matlab command …)
- But - keep in mind that we need to find the probability of having **16 or more** patients surviving at least 5 yrs.

Example: Treatment of Kidney Cancer

- So we actually need to find:

$P(X \geq 16) = P(X = 16) + P(X = 17) + \ldots + P(X = 40)$

$$P(X=16)=\binom{40}{16}.20^{16}.80^{24}=.001945$$

$$+\quad P(X=17)=\binom{40}{17}.20^{17}.80^{23}=.000686$$

$$\ldots$$

$$+\quad P(X=40)=\binom{40}{40}.20^{40}.80^{0}\approx 0$$

= .002936 ***Yupp!***

When using commands in a statistical language we will use the CDF

```python
from scipy.stats import binom
rv = binom(40, 0.2)
x_16_and_up = 1 - rv.cdf(15)
print("{:.4f}".format(x_16_and_up))
```

```
0.0029
```

IDC HERZLIYA

Conclusion (statistics helps decision making …)

Because it is highly unlikely (p = 0.0029) that we would see this many successes in a group of *40* patients if the new Tx had the same probability of success as the SoC we have to make a choice, either …

A) Tx's survival rate is less than 0.2 and we have obtained a very rare result by chance.

OR

B) our assumption about the success rate of the new Tx is wrong and in actuality it has a better than 20% 5-year survival rate making the observed result more plausible.

Caveat: other aspects of the null model can also be wrong …

IDC HERZLIYA

Tx is better than the SoC treatment with <u>p-value <0.003 under a binomial null model</u>

Next week we will start with two waiting time distributions: the geometric distribution and the negative binomial (Polya) distribution

$\omega\epsilon\Omega$:

Continue to infinity …

A Geometric random variable:

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

Summary and what's next ….

- Statistics provides tools and frameworks for the rigorous interpretation of data, for effective (and efficient) inference and for clearly presenting and stating conclusions.
- Data analysis uses computational approaches to implement statistical principles in practically analyzing data.
- In this course we will address theoretical and practical aspects of both.
- We will emphasize computer age aspects: efficiency, volume etc
- We learned about Bernoulli random variables and about the Binomial distribution.
- Next time: Geometric, NegB, Poisson distributions and related aspects
- During the course we will present and investigate more distributions.
- We proved Tchebychef's Thm and saw how it yields a bound on large deviations from the mean
- In following weeks we will derive more efficient approaches/bounds and see how to use them in practice.
- Next week: independence or not? And the consequences …

IDC HERZLIYA