

KSE

Kyiv
School of
Economics

Statistical distributions

Ass. Prof. Andriy Stavytskyy

Outline

1. Expected Value, Variance, Covariance, Correlation
2. Binominal Distribution
3. Poisson Distribution
4. Normal Distribution

Random variables

- **Definition.** A random variable, X , is a numerical measure of the outcomes of an experiment

Example – 1

Experiment: Two cards randomly selected

Let X be the number of diamonds selected

$$\Omega = \left\{ \begin{array}{cccc} CC & CD & CH & CS \\ DC & DD & DH & DS \\ HC & HD & HH & HS \\ SC & SD & SH & SS \end{array} \right\}$$

Example – 2

- Events can be described in terms of random variables:
- If the event that exactly one diamond is selected

$$X = 1$$

- if the event that at most one diamond is selected

$$X \leq 1$$

Example – 3

Probabilities of events can be stated as probabilities of the corresponding values of X

$$P(X = 1) = \frac{6}{16} = \frac{3}{8}$$

$$P(X \leq 1) = P \left(\left\{ \begin{array}{ccccc} CC & CD & CH & CS & DC \\ DH & DS & HC & HD & HH \\ HS & SC & SD & SH & SS \end{array} \right\} \right) = \frac{15}{16}$$

$$\Omega = \left\{ \begin{array}{cccc} CC & CD & CH & CS \\ DC & DD & DH & DS \\ HC & HD & HH & HS \\ SC & SD & SH & SS \end{array} \right\}$$

In general

In general,

$P(X = x)$ is the probability that X takes on the value x

$P(X \leq x)$ is the probability that X takes on a value that
is less than or equal to x

Unity of probabilities

Suppose that X can only assume the values x_1, x_2, \dots, x_n .


Then

$$\sum_{i=1}^n P(X = x_i) = 1$$

The expected value

Definition. Expected value of X gives the value that we would expect to observe on average in a large number of repetitions of the experiment.

$$\mu_X = E(X) = \sum_{i=1}^n x_i * P(X = x_i)$$



Sum of the values,
weighted by their
respected probabilities

Example

- An investment in Project A will result in a **loss** of \$26,000 with probability 0.30, break even with probability 0.50, or result in a profit of \$68,000 with probability 0.20.
- An investment in Project B will result in a **loss** of \$71,000 with probability 0.20, break even with probability 0.65, or result in a profit of \$143,000 with probability 0.15.
- Which investment is better?

Tools to calculate $E(X)$ -Project A

Random Variable (X) - The amount of money received from the investment in Project A. X can assume only x_1, x_2, x_3

- $X = x_1$ is the event that we have Loss
- $X = x_2$ is the event that we are breaking even
- $X = x_3$ is the event that we have a Profit

$$x_1 = \$-26,000$$

$$x_2 = \$0$$

$$x_3 = \$68,000$$

$$P(X = x_1) = 0.3$$

$$P(X = x_2) = 0.5$$

$$P(X = x_3) = 0.2$$

Tools to calculate $E(X)$ -Project B

Random Variable (X) - The amount of money received from the investment in Project B. X can assume only x_1, x_2, x_3

- $X = x_1$ is the event that we have Loss
- $X = x_2$ is the event that we are breaking even
- $X = x_3$ is the event that we have a Profit

$$x_1 = \$-71,000$$

$$x_2 = \$0$$

$$x_3 = \$143,000$$

$$P(X = x_1) = 0.2$$

$$P(X = x_2) = 0.65$$

$$P(X = x_3) = 0.15$$

Solution

Project A :

$$\begin{aligned} E(X) &= 0.30 \cdot (-\$26,000) + 0.50 \cdot \$0 + 0.20 \cdot \$68,000 \\ &= \$5800 \end{aligned}$$

Project B :

$$\begin{aligned} E(X) &= 0.20 \cdot (-\$71,000) + 0.65 \cdot \$0 + 0.15 \cdot \$143,000 \\ &= \$7250 \end{aligned}$$

Variance

- The variance of a random value is the sum of the squared deviations from the expected value weighted by their associated probabilities.

$$\sigma^2(X) = \sum_{i=1}^n [X_i - E(X)]^2 P(X_i) = E\left(X - E(X)\right)^2$$

- This value is a measure of the dispersion of possible values.
- Because it has units that are squared, it is not easy to interpret. Accordingly, we use its positive square root, standard deviation, more often because it also measures dispersion but has the same units as expected value.

Note

- It's easier to calculate variance in such way:

$$\sigma^2(X) = EX^2 - (EX)^2$$

where

$$EX^2 = \sum_{i=1}^n X_i^2 P_i$$

Example

X	1	2	3	4	5	6		Sum
P	1/6	1/6	1/6	1/6	1/6	1/6		
X*P	0.17	0.33	0.50	0.67	0.83	1.00		3.50
X²	1	4	9	16	25	36		
X²*P	0.17	0.67	1.50	2.67	4.17	6.00		15.17

$$\sigma^2(X) = EX^2 - (EX)^2 = 15.17 - 3.50^2 = 2.92$$

Statistical Distributions

1. Binominal Distribution



2. Poisson Distribution

3. Normal Distribution

Binomial Probability Distribution

- A fixed number of observations (trials), n
 - e.g., 15 tosses of a coin; 20 patients; 1000 people surveyed
- A binary outcome
 - e.g., head or tail in each toss of a coin; disease or no disease
 - Generally called “success” and “failure”
 - Probability of success is p , probability of failure is $1 - p$
- Constant probability for each observation
 - e.g., Probability of getting a tail is the same each time we toss the coin

Binomial distribution

- Take the example of 5 coin tosses. What's the probability that you flip exactly 3 heads in 5 coin tosses?

Binomial distribution: solution – 1

- One way to get exactly 3 heads: HHHTT
- What's the probability of this exact arrangement?
$$P(\text{heads}) \times P(\text{heads}) \times P(\text{heads}) \times P(\text{tails}) \times P(\text{tails})$$
$$= (1/2)^3 \times (1/2)^2$$
- Another way to get exactly 3 heads: THHHT
Probability of this exact outcome = $(1/2)^1 \times (1/2)^3 \times (1/2)^1$
= $(1/2)^3 \times (1/2)^2$

Binomial distribution: solution – 2

- In fact, $(1/2)^3 \times (1/2)^2$ is the probability of each unique outcome that has exactly 3 heads and 2 tails.

- So, the overall probability of 3 heads and 2 tails is:

$$(1/2)^3 \times (1/2)^2 + (1/2)^3 \times (1/2)^2 + (1/2)^3 \times (1/2)^2 + \dots$$

for as many unique arrangements as there are - but how many are there??

Binomial distribution: solution – 3

<u>Outcome</u>	<u>Probability</u>
THHHT	$(1/2)^3 \cdot (1/2)^2$
HHHTT	$(1/2)^3 \cdot (1/2)^2$
TTHHH	$(1/2)^3 \cdot (1/2)^2$
HTTHH	$(1/2)^3 \cdot (1/2)^2$
HHTTH	$(1/2)^3 \cdot (1/2)^2$
HTHHT	$(1/2)^3 \cdot (1/2)^2$
THTHH	$(1/2)^3 \cdot (1/2)^2$
HTHTH	$(1/2)^3 \cdot (1/2)^2$
HHTHT	$(1/2)^3 \cdot (1/2)^2$
THHTH	$(1/2)^3 \cdot (1/2)^2$
10 arrangements	$\cdot (1/2)^3 \cdot (1/2)^2$

The probability of each unique outcome (note: they are all equal)

ways to arrange 3 heads in 5 trials

$$\binom{5}{3}$$

$${}_5C_3 = 5!/(3!2!) = 10$$

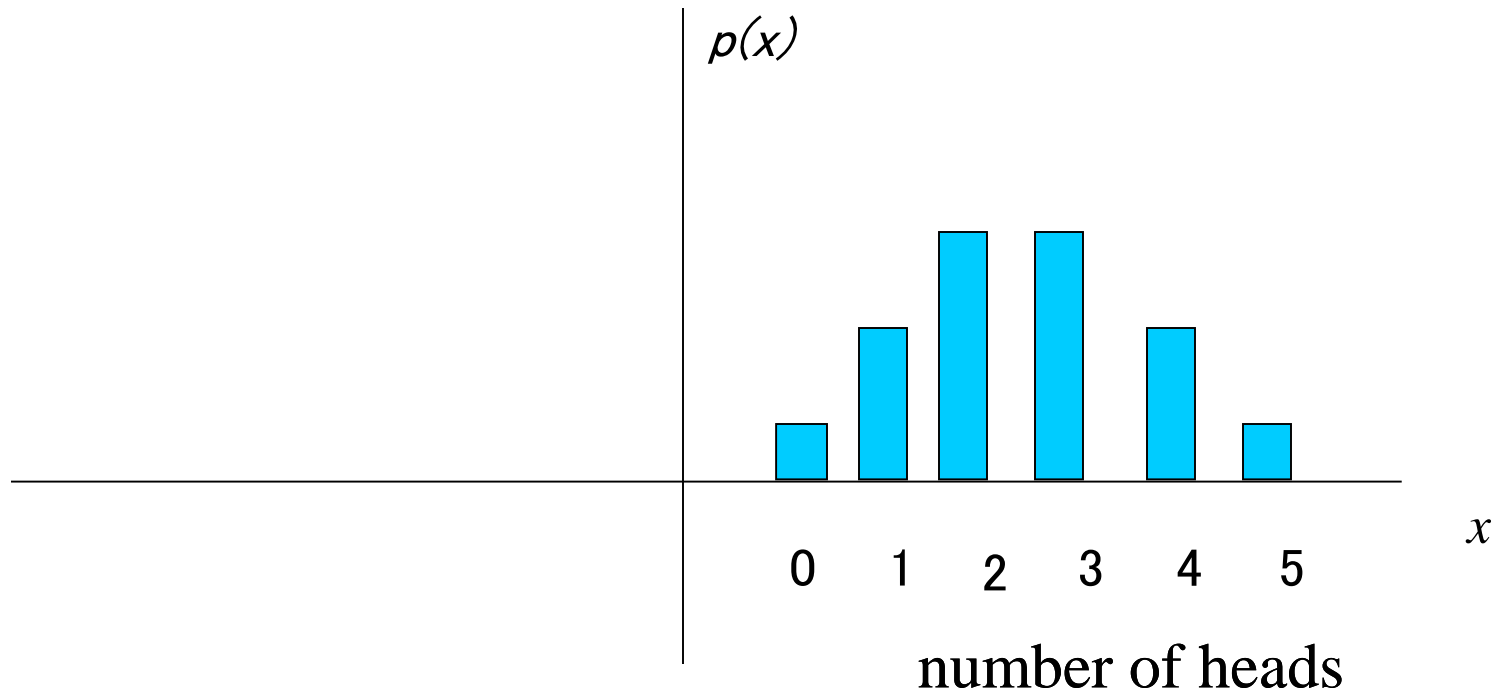
Factorial review: $n! = n(n-1)(n-2)\dots$

Binomial distribution: solution – 4

$$\begin{aligned} P(3 \text{ heads and } 2 \text{ tails}) &= \binom{5}{3} \times P(\text{heads})^3 \times P(\text{tails})^2 = \\ &= 10 \times (1/2)^5 = 31.25\% \end{aligned}$$

Binomial distribution function

- X = the number of heads tossed in 5 coin tosses



Binomial distribution, generally

Note the general pattern emerging \rightarrow if you have only two possible outcomes (call them 1/0 or yes/no or success/failure) in n independent trials, then the probability of exactly X “successes” =

The diagram shows the binomial distribution formula $\binom{n}{X} p^X (1-p)^{n-X}$ enclosed in a purple box. Arrows point from descriptive text to parts of the formula: n is the number of trials, X is the number of successes, p is the probability of success, and $1-p$ is the probability of failure.

$$\binom{n}{X} p^X (1-p)^{n-X}$$

$n =$ number of trials

$X = \#$
successes
out of n
trials

$p =$
probability of
success

$1-p =$ probability
of failure

All probability distributions are characterized by an expected value and a variance

If X follows a binomial distribution with parameters n and p:

$$X \sim \text{Bin}(n, p)$$

Then:

- $E(X) = np$
- $\text{Var}(X) = np(1-p)$
- $\text{SD}(X) = \sqrt{np(1-p)}$



Note: the variance will always lie between $0 \cdot N$ and $.25 \cdot N$
 $p(1-p)$ reaches maximum at $p=.5$
 $P(1-p) = .25$

Practice Problem

1. You are performing a cohort study. If the probability of developing disease in the exposed group is .05 for the study duration, then if you (randomly) sample 500 exposed people, how many do you expect to develop the disease? Give a margin of error (+/- 1 standard deviation) for your estimate.
2. What's the probability that at most 10 exposed people develop the disease?

Answer – 1

1. How many do you expect to develop the disease? Give a margin of error (+/- 1 standard deviation) for your estimate.

$$X \sim \text{binomial}(500, .05)$$

$$E(X) = 500 (.05) = 25$$

$$\text{Var}(X) = 500 (.05) (.95) = 23.75$$

$$\text{StdDev}(X) = \sqrt{23.75} = 4.87$$

$$\text{Interval: } 25 \pm 4.87$$

Answer – 2

2. What's the probability that at most 10 exposed subjects develop the disease?

- This is asking for a CUMULATIVE PROBABILITY: the probability of 0 getting the disease or 1 or 2 or 3 or 4 or up to 10.
- $P(X \leq 10) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) + \dots + P(X=10) =$

$$\binom{500}{0} (.05)^0 (.95)^{500} + \binom{500}{1} (.05)^1 (.95)^{499} + \binom{500}{2} (.05)^2 (.95)^{498} + \dots + \binom{500}{10} (.05)^{10} (.95)^{490} < .01$$

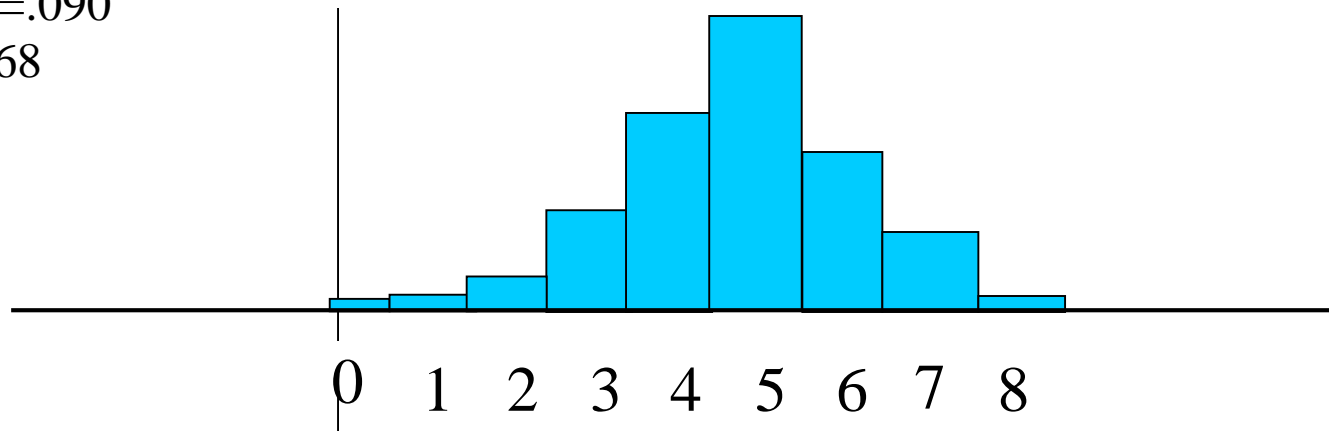
Practice Problem

You are conducting a case-control study of smoking and lung cancer. If the probability of being a smoker among lung cancer cases is 0.6, what's the probability that in a group of 8 cases you have:

1. Less than 2 smokers?
2. More than 5?
3. What are the expected value and variance of the number of smokers?

Answer – 1

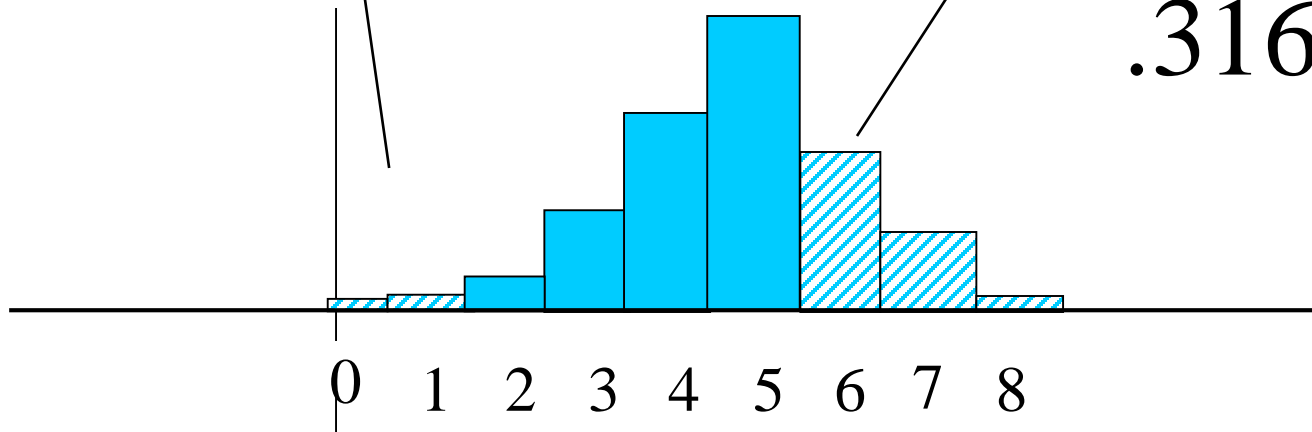
X	P(X)
0	$1(.4)^8 = .00065$
1	$8(.6)^1 (.4)^7 = .008$
2	$28(.6)^2 (.4)^6 = .04$
3	$56(.6)^3 (.4)^5 = .12$
4	$70(.6)^4 (.4)^4 = .23$
5	$56(.6)^5 (.4)^3 = .28$
6	$28(.6)^6 (.4)^2 = .21$
7	$8(.6)^7 (.4)^1 = .090$
8	$1(.6)^8 = .0168$



Answer – 2

$$P(<2) = .00065 + .008 = .00865$$

$$P(>5) = .21 + .09 + .0168 = .3168$$



$$E(X) = 8 (0.6) = 4.8$$

$$\text{Var}(X) = 8 (0.6) (0.4) = 1.92$$

$$\text{StdDev}(X) = 1.38$$

Binomial distribution: example

- If I toss a coin 20 times, what's the probability of getting exactly 10 heads?

$$\binom{20}{10} (0.5)^{10} (0.5)^{10} = 0.176$$

Binomial distribution: example

- If I toss a coin 20 times, what's the probability of getting 2 or fewer heads?

$$\binom{20}{0} (0.5)^0 (0.5)^{20} = \frac{20!}{20!0!} (.5)^{20} = 9.5 \cdot 10^{-7} +$$

$$\binom{20}{1} (0.5)^1 (0.5)^{19} = \frac{20!}{19!1!} (.5)^{20} = 20 \cdot 9.5 \cdot 10^{-7} = 1.9 \cdot 10^{-5} +$$

$$\binom{20}{2} (0.5)^2 (0.5)^{18} = \frac{20!}{18!2!} (.5)^{20} = 190 \cdot 9.5 \cdot 10^{-7} = 1.8 \cdot 10^{-4}$$

$$= 1.8 \cdot 10^{-4}$$

Poisson Distribution

- A random variable X is exponentially distributed with parameter $\lambda > 0$ if probability function

$$P(x = k) = \begin{cases} \frac{\lambda^k}{k!} e^{-\lambda}, & \text{for } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

- Mean = λ
- Variance = λ

- Note: $e = 2.71828182$

Table of Poisson Probabilities

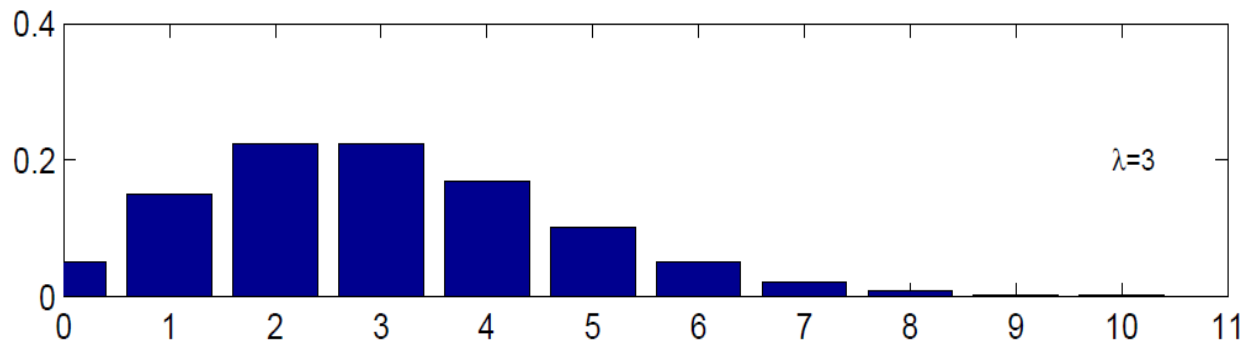
		λ									
X	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679	
1	0.0905	0.1637	0.2222	0.2681	0.3033	0.3293	0.3476	0.3595	0.3659	0.3679	
2	0.0045	0.0164	0.0333	0.0536	0.0758	0.0988	0.1217	0.1438	0.1647	0.1839	
3	0.0002	0.0011	0.0033	0.0072	0.0126	0.0198	0.0284	0.0383	0.0494	0.0613	
4	0.0000	0.0001	0.0003	0.0007	0.0016	0.0030	0.0050	0.0077	0.0111	0.0153	
5	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0012	0.0020	0.0031	
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0005	
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	
		λ									
X	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	
0	0.3329	0.3012	0.2725	0.2466	0.2231	0.2019	0.1827	0.1653	0.1496	0.1353	
1	0.3662	0.3614	0.3543	0.3452	0.3347	0.3230	0.3106	0.2975	0.2842	0.2707	
2	0.2014	0.2169	0.2303	0.2417	0.2510	0.2584	0.2640	0.2678	0.2700	0.2707	
3	0.0738	0.0867	0.0998	0.1128	0.1255	0.1378	0.1496	0.1607	0.1710	0.1804	
4	0.0203	0.0260	0.0324	0.0395	0.0471	0.0551	0.0636	0.0723	0.0812	0.0902	
5	0.0045	0.0062	0.0084	0.0111	0.0141	0.0176	0.0216	0.0260	0.0309	0.0361	
6	0.0008	0.0012	0.0018	0.0026	0.0035	0.0047	0.0061	0.0078	0.0098	0.0120	
7	0.0001	0.0002	0.0003	0.0005	0.0008	0.0011	0.0015	0.0020	0.0027	0.0034	
8	0.0000	0.0000	0.0001	0.0001	0.0001	0.0002	0.0003	0.0005	0.0006	0.0009	
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0002	

Examples of possible Poisson distributions

- 1) Number of messages arriving at a telecommunications system in a day
- 2) Number of flaws in a metre of fibre optic cable
- 3) Number of radio-active particles detected in a given time
- 4) Number of photons arriving at a CCD pixel in some exposure time (e.g. astronomy observations)

Example

- On average lightning kills three people each year in the UK, $\lambda = 3$. What is the probability that only one person is killed this year?
- Assuming these are independent random events, the number of people killed in a given year therefore has a Poisson distribution:



Let the random variable X be the number of people killed in a year.

Poisson distribution $P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$ with $\lambda = 3 \Rightarrow P(X = 1) = \frac{e^{-3} 3^1}{1!} \approx 0.15$

Radioactive decay – 1

- x =number of particles/min
- $\lambda=2$ particles per minutes

$$P(x = 3) = \frac{2^3 e^{-2}}{3!}, \quad x=0, 1, 2, \dots$$

Self-task

Suppose that trucks arrive at a receiving dock with an average arrival rate of 3 per hour. What is the probability exactly 5 trucks will arrive in a two-hour period?

Answer: In two hours mean number is $\lambda = 2 \times 3 = 6$.

$$P(X = k = 5) = \frac{e^{-\lambda} \lambda^k}{k!} = \frac{e^{-6} 6^5}{5!}$$

Radioactive decay – 2

- Radioactive decay
- X =number of particles/hour
- $\lambda = 2$ particles/min * 60min/hour=120 particles/hr

$$P(x = 125) = \frac{120^{125} e^{-120}}{125!}, \quad x = 0, 1, 2, \dots$$

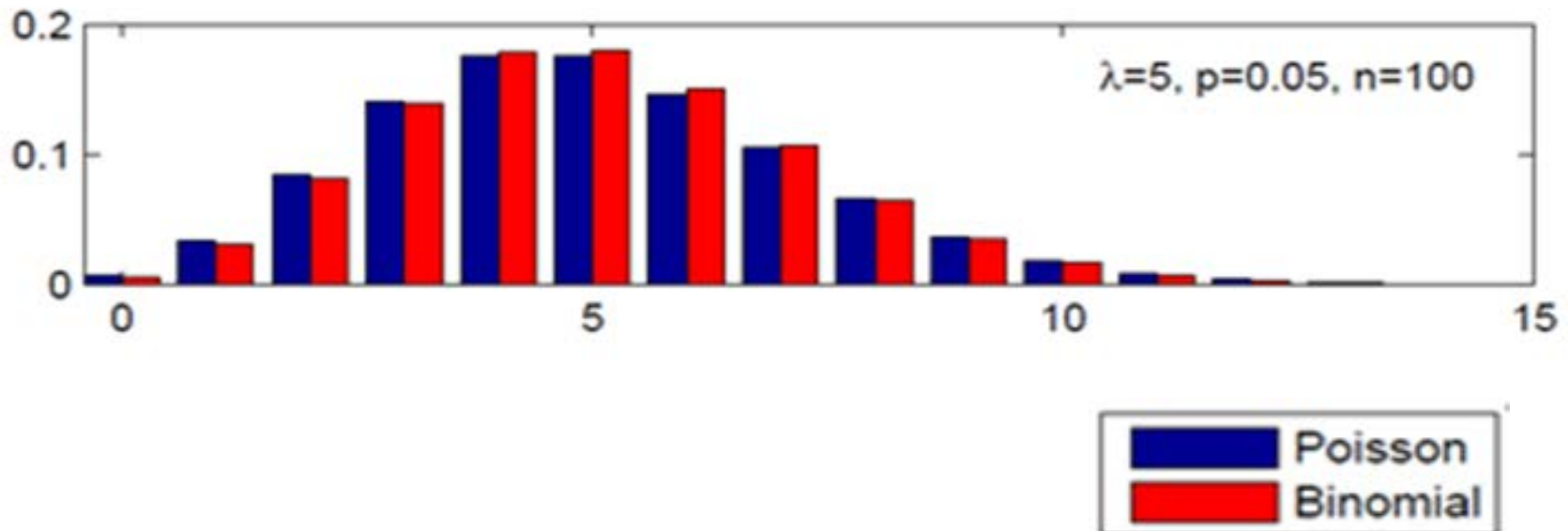
Approximation to the Binomial distribution

- The Poisson distribution is an approximation to $B(n, p)$, when n is large and p is small (e.g. if $np < 7$, say).

- In that case, if $X \sim B(n, p)$ then $P(X = k) \approx \frac{e^{-\lambda} \lambda^k}{k!}$

where $\lambda = np$

- i.e. X is approximately Poisson, with mean $\lambda = np$.



Poisson or not?

Which of the following are likely to be well modelled by a Poisson distribution?

- | | |
|---|---|
| <ol style="list-style-type: none"> 1. Number of duds found when I test four components 2. The number of heart attacks in Brighton each year 3. The number of planes landing at Heathrow between 8 and 9 am 4. The number of cars getting punctures on the M1 each year 5. Number of people in the UK flooded out of their home in July | <ol style="list-style-type: none"> 1. NO: this is Binomial (it is not the number of independent random events in a continuous interval) 2. YES: large population, no obvious correlation between heart attacks in different people 3. NO: 8-9am is rush hour, planes land regularly to land as many as possible (1-2 a minute) – they do not land at random times or they would hit each other! 4. YES (roughly): If punctures are due to tires randomly wearing thin, then expect punctures to happen independently at random
<i>But: may not all be independent, e.g. if there is broken glass in one lane</i> 5. NO: floodings of different homes not at all independent; usually a small number of floods each flood many homes at once,
$P(\text{flooded} \text{next door flooded}) \gg P(\text{flooded})$ |
|---|---|

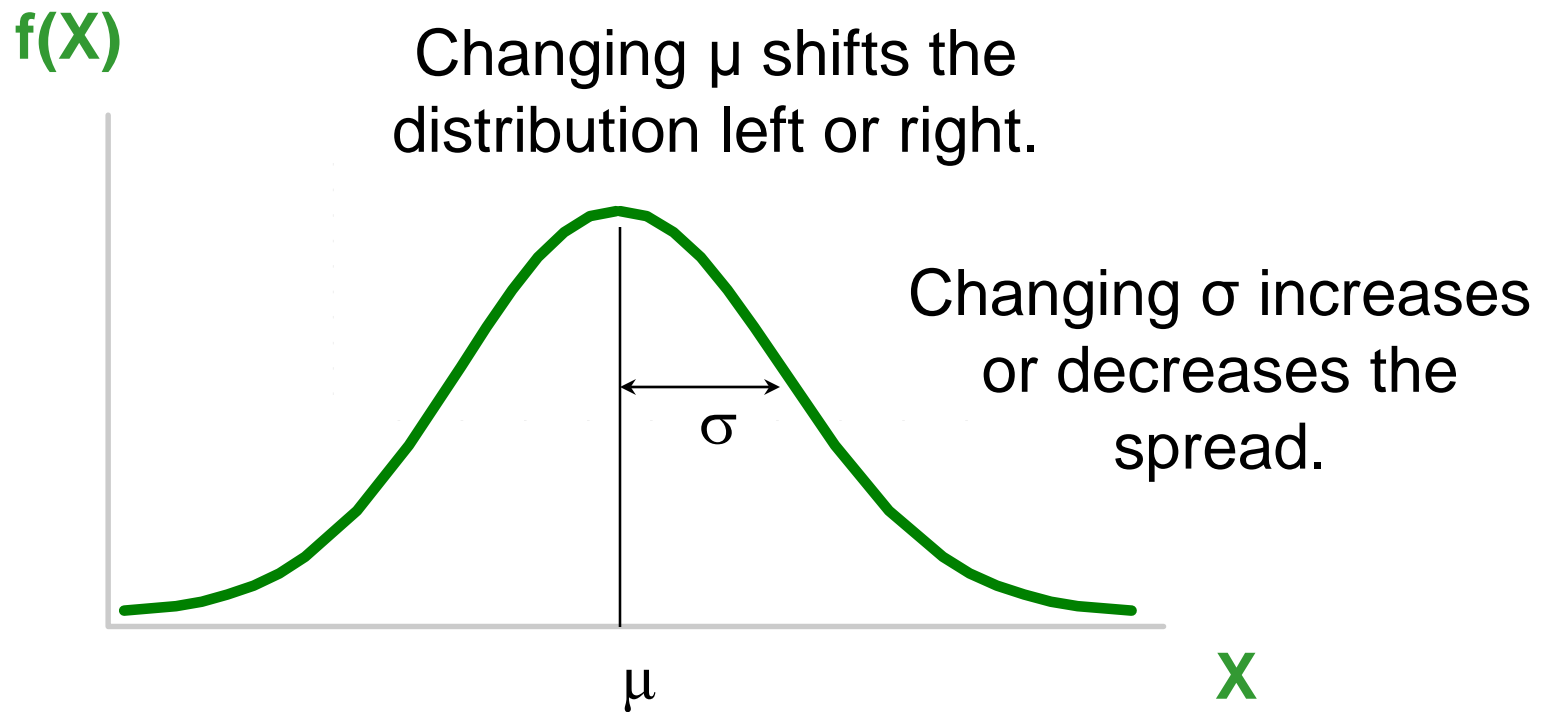
Exercise

A mailroom clerk is supposed to send 6 of 15 packages to Europe by airmail, but he gets them all mixed up and randomly puts airmail postage on 6 of the packages. What is the probability that only three of the packages that are supposed to go by air get airmail postage?

Answer:

$$\lambda = \frac{6}{15} = \frac{2}{5} = 0.4.$$
$$P\{k = 3\} = \frac{0.4^3}{3!} e^{-0.4} \approx 0.00715$$

The Normal Distribution



The Normal Distribution: as mathematical function (pdf)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Note constants:

$\pi=3.14159$

$e=2.71828$

This is a bell shaped curve with different centers and spreads depending on μ and σ

Normal distribution is defined by its mean and standard dev.

- $E(X)=\mu = \int_{-\infty}^{+\infty} x \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$

- $\text{Var}(X)=\sigma^2 = \int_{-\infty}^{+\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx - \mu^2$

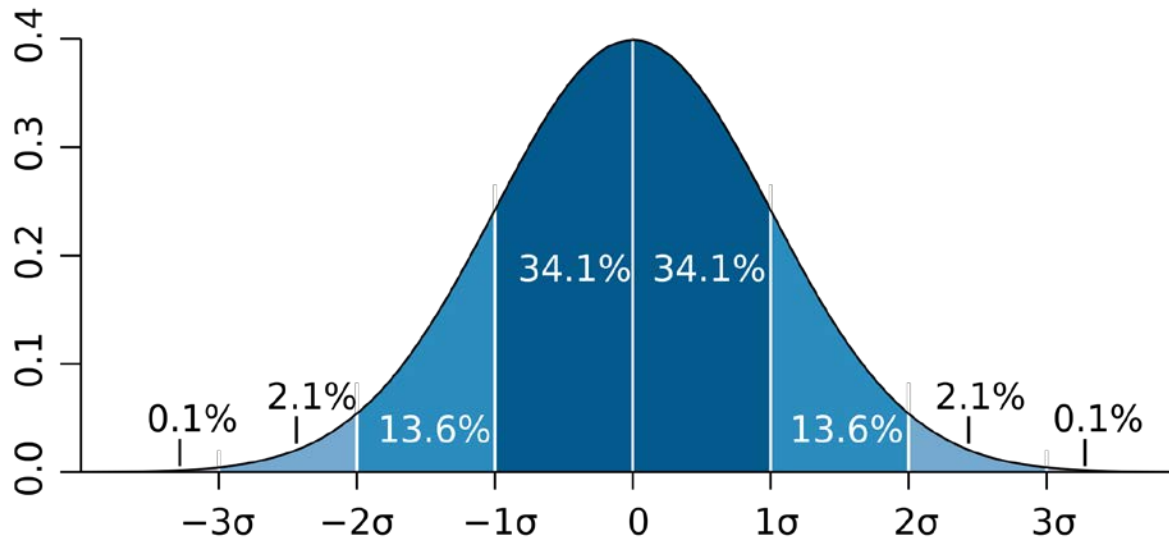
- Standard Deviation(X)= σ

The beauty of the normal curve:

No matter what μ and σ are,

- the area between $\mu - \sigma$ and $\mu + \sigma$ is about 68%;
- the area between $\mu - 2\sigma$ and $\mu + 2\sigma$ is about 95%;
- and the area between $\mu - 3\sigma$ and $\mu + 3\sigma$ is about 99.7%.

Almost all values fall within 3 standard deviations!!!



Example

- Suppose SAT scores roughly follows a normal distribution in the U.S. population of college-bound students (with range restricted to 200-800), and the average math SAT is 500 with a standard deviation of 50, then:
 - 68% of students will have scores between 450 and 550
 - 95% will be between 400 and 600
 - 99.7% will be between 350 and 650

Example

- BUT...
- What if you wanted to know the math SAT score corresponding to the 90th percentile (=90% of students are lower)?
- $P(X \leq Q) = .90 \rightarrow$

$$\int_{200}^Q \frac{1}{(50)\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-500}{50}\right)^2} dx = 0.90$$

Solve for Q?....Yikes!

The Standard Normal (Z): “Universal Currency”

The formula for the standardized normal probability density function is

$$p(Z) = \frac{1}{(1)\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{Z-0}{1}\right)^2} = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(Z)^2}$$

The Standard Normal Distribution (Z)

All normal distributions can be converted into the standard normal curve by subtracting the mean and dividing by the standard deviation:

$$Z = \frac{X - \mu}{\sigma}$$

Somebody calculated all the integrals for the standard normal and put them in a table! So we never have to integrate!

Even better, computers now do all the integration.

Standard normal distribution

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831

Example

What's the probability of getting a math SAT score of 575 or less, $\mu=500$ and $\sigma=50$?

$$Z = \frac{575 - 500}{50} = 1.5$$

i.e., a score of 575 is 1.5 standard deviations above the mean

$$P(X \leq 575) = \int_{200}^{575} \frac{1}{(50)\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-500}{50}\right)^2} dx \longrightarrow \int_{-\infty}^{1.5} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}z^2} dz$$

Look up $Z= 1.5$ in standard normal chart $\rightarrow = 0.9332$

Example

The scores on a college entrance exam are normally distributed with a mean of 75 and a standard deviation of 10.

What % of scores fall between 70 and 90?

$$Z(70) = (70 - 75)/10 = -0.5$$

$$Z(90) = (90 - 75)/10 = 1.5$$

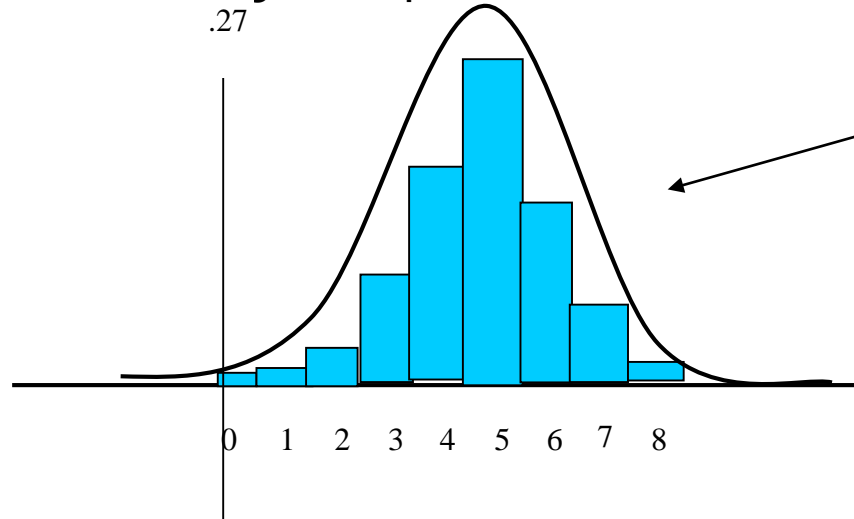
$$Z(1.5) - Z(-0.5) = 0.93319 - 0.30854 = 0.6247 = 62.47\%$$

Normal approximation to the binomial

- When you have a binomial distribution where n is large and p is middle-of-the road (not too small, not too big, closer to 0.5), then the binomial starts to look like a normal distribution in fact, this doesn't even take a particularly large n
- Recall: What is the probability of being a smoker among a group of cases with lung cancer is .6, what's the probability that in a group of 8 cases you have less than 2 smokers?

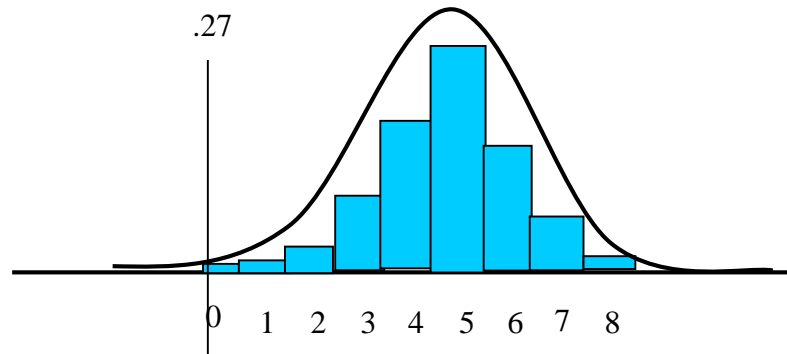
Normal approximation to the binomial

- When you have a binomial distribution where n is large and p isn't too small (rule of thumb: $np > 5$), then the binomial starts to look like a normal distribution
- Recall: smoking example...



Starting to have a normal shape even with fairly small n . You can imagine that if n got larger, the bars would get thinner and thinner and this would look more and more like a continuous function, with a bell curve shape. Here $np=4.8$.

Normal approximation to binomial – 1



What is the probability of fewer than 2 smokers?

Exact binomial probability (from before) = .00065 + .008 = [.00865](#)

Normal approximation probability:

$$\mu=4.8$$

$$\sigma=1.39$$

$$Z \approx \frac{2 - (4.8)}{1.39} = \frac{-2.8}{1.39} = -2$$

$$P(Z < 2) = Z(-2) = [.02275](#)$$

Normal approximation to binomial – 2

- A little off, but in the right ballpark... we could also use the value to the left of 1.5 (as we really wanted to know less than but not including 2; called the “continuity correction”)...

$$Z \approx \frac{1.5 - (4.8)}{1.39} = \frac{-3.3}{1.39} = -2.37$$

$$P(Z \leq -2.37) = Z(-2.37) = 0.0089$$


A fairly good approximation of the exact probability, 0.00865.

REVIEW

The expected value

Definition. Expected value of X gives the value that we would expect to observe on average in a large number of repetitions of the experiment.

$$\mu_X = E(X) = \sum_{i=1}^n x_i * P(X = x_i)$$



Sum of the values,
weighted by their
respected probabilities

Variance

- The variance of a random value is the sum of the squared deviations from the expected value weighted by their associated probabilities.

$$\sigma^2(X) = \sum_{i=1}^n [X_i - E(X)]^2 P(X_i) = E\left(X - E(X)\right)^2$$

- This value is a measure of the dispersion of possible values.
- Because it has units that are squared, it is not easy to interpret. Accordingly, we use its positive square root, standard deviation, more often because it also measures dispersion but has the same units as expected value.

Statistical Distributions

1. Binominal Distribution



2. Poisson Distribution

3. Normal Distribution

Binomial distribution, generally

Note the general pattern emerging \rightarrow if you have only two possible outcomes (call them 1/0 or yes/no or success/failure) in n independent trials, then the probability of exactly X “successes” =

The diagram shows the binomial distribution formula $\binom{n}{X} p^X (1-p)^{n-X}$ enclosed in a purple box. Arrows point from descriptive text to parts of the formula: n is the number of trials, X is the number of successes, p is the probability of success, and $1-p$ is the probability of failure.

$$\binom{n}{X} p^X (1-p)^{n-X}$$

$n =$ number of trials

$X = \#$
successes
out of n
trials

$p =$
probability of
success

$1-p =$ probability
of failure

Poisson Distribution

- A random variable X is exponentially distributed with parameter $\lambda > 0$ if probability function

$$P(x = k) = \begin{cases} \frac{\lambda^k}{k!} e^{-\lambda}, & \text{for } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

- Mean = λ
- Variance = λ

- Note: $e = 2.71828182$

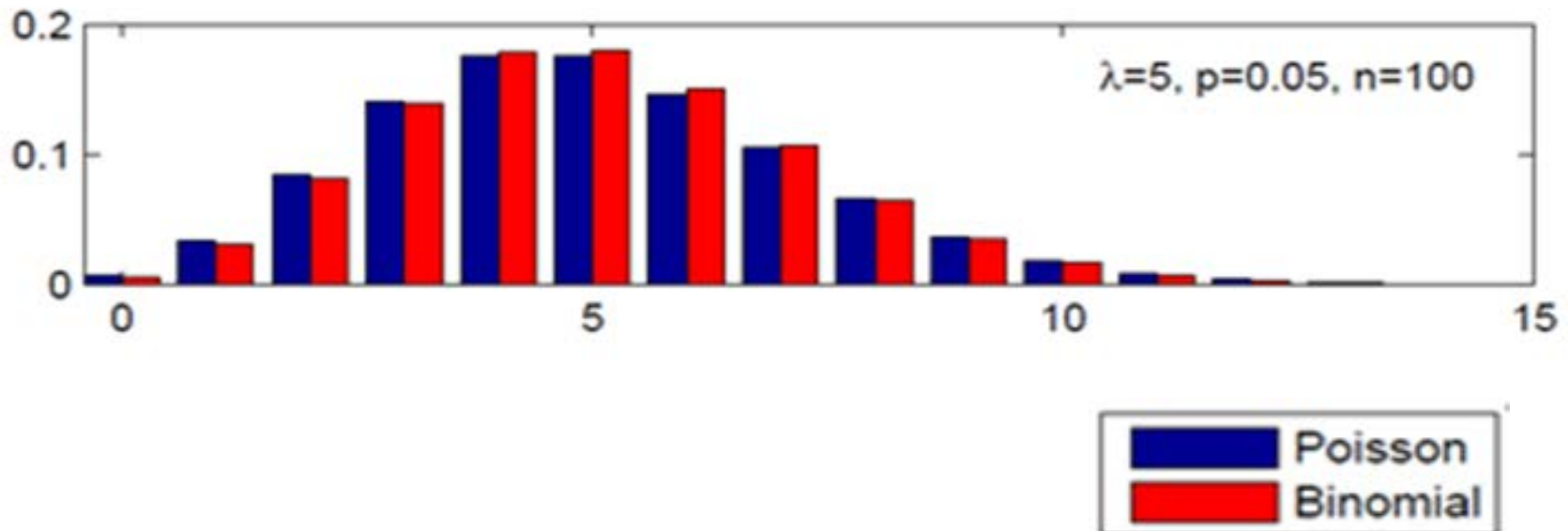
Approximation to the Binomial distribution

- The Poisson distribution is an approximation to $B(n, p)$, when n is large and p is small (e.g. if $np < 7$, say).

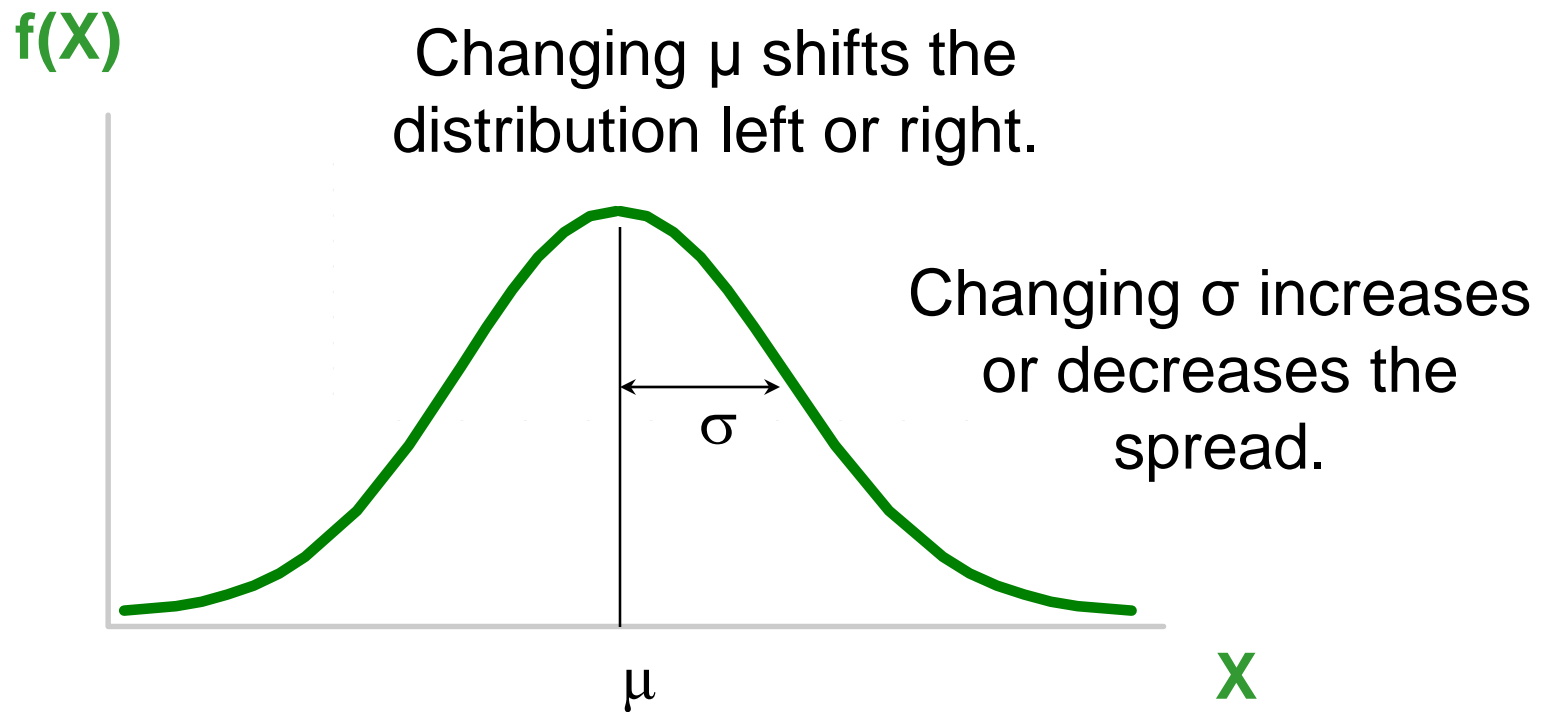
- In that case, if $X \sim B(n, p)$ then $P(X = k) \approx \frac{e^{-\lambda} \lambda^k}{k!}$

where $\lambda = np$

- i.e. X is approximately Poisson, with mean $\lambda = np$.



The Normal Distribution



Normal distribution is defined by its mean and standard dev.

- $E(X)=\mu = \int_{-\infty}^{+\infty} x \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$

- $\text{Var}(X)=\sigma^2 = \int_{-\infty}^{+\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx - \mu^2$

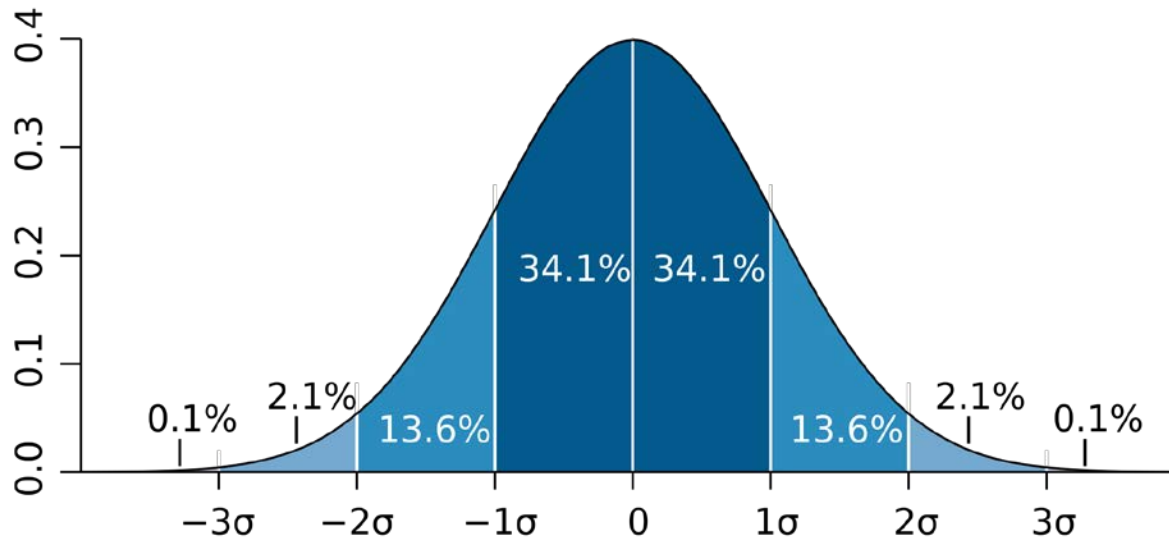
- Standard Deviation(X)= σ

The beauty of the normal curve:

No matter what μ and σ are,

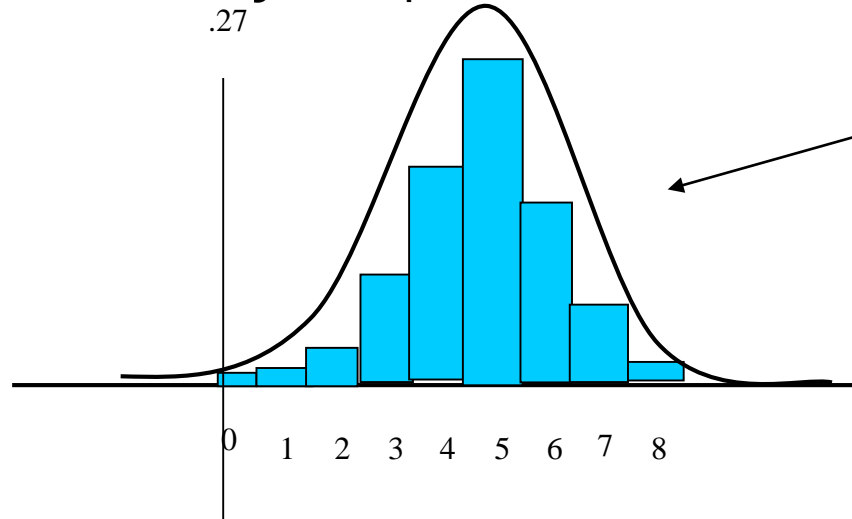
- the area between $\mu - \sigma$ and $\mu + \sigma$ is about 68%;
- the area between $\mu - 2\sigma$ and $\mu + 2\sigma$ is about 95%;
- and the area between $\mu - 3\sigma$ and $\mu + 3\sigma$ is about 99.7%.

Almost all values fall within 3 standard deviations!!!



Normal approximation to the binomial

- When you have a binomial distribution where n is large and p isn't too small (rule of thumb: $np > 5$), then the binomial starts to look like a normal distribution
- Recall: smoking example...



Starting to have a normal shape even with fairly small n . You can imagine that if n got larger, the bars would get thinner and thinner and this would look more and more like a continuous function, with a bell curve shape. Here $np=4.8$.

Thank you for attention!