Distributions

Outline

- Discrete random variables and probability distributions
- The Binomial probability distribution
- The Poisson probability distribution
- The normal probability distribution

Discrete Random Variables and Probability Distributions

- **Random variables**
  - value of which is the result of a random event, e.g. the number of laptops sold on a *randomly selected* day, the age of a student *randomly selected* on campus.

- **Discrete variables**
  - One type of ratio variable, can take only a limited number of possible values within a given range, e.g. number of laptops.
  - As opposed to **continuous variables** which can take unlimited number of possible values in the same range, e.g. the distance between two locations.

Discrete Random Variables and Probability Distributions

- **Probability distribution** for a discrete random variable:

$$P(A) = \lim_{n \to \infty} \frac{\text{Frequency}}{n}$$

  - Relative frequency distribution constructed for the entire population of measurements
  - Represents possible values of x and the probability associated with each value of x.
  - Possible values of x are mutually exclusive events.
  - $\sum_{i=1}^{n} p(x_i) = 1$
  - Example –The probability distribution of number of cups of coffee taken a day

x	0	1	2	3	4	5
$p(x)$	.28	.37	.17	.12	.05	.01

Discrete Random Variables and Probability Distributions

- **Probability distribution** for a discrete random variable:
  - Mean (aka the expected value): $\mu = \sum_{i=1}^{n} x_i p(x_i)$
  - Standard variation: $\sigma = \sqrt{\sum_{i=1}^{n} (x_i - \mu)^2 p(x_i)}$
- Example:

x	0	1	2	3	4	5
$p(x)$	.28	.37	.17	.12	.05	.01

  - What is the probability of a random coffee drinker taking no coffee a day?
  - What is the probability of a random coffee drinker taking more than 2 coffees a day?
  - What is the number of coffees a random coffee drinker expected to drink a day?
  - What is the probability of the number of coffees a day fall in between $\mu \pm 2\sigma$?

The Binomial Probability Distribution

- **Binomial random variable** – has only two possible values

- **Binomial experiment**
  - Contains n *identical* trials.
  - Each trial results in *one of two outcomes*, e.g. Success or Failure.
  - The probability *p* of an outcome, e.g. Success, *remains the same* for all trials.
  - Trials are *independent*.
  - We are interested in x, the *number of Successes* observed in n trials.

- Example of a binomial experiment:
  - Tossing a coin 1000 times and observing the number of heads
  - Test of a new drug and counting the number of successful cases
  - Purchase lottery tickets many (many) times and count the number of wins.

The Binomial Probability Distribution

- The **probability** of *x = k* successes (*p* is the probability of success) in *n* trials is

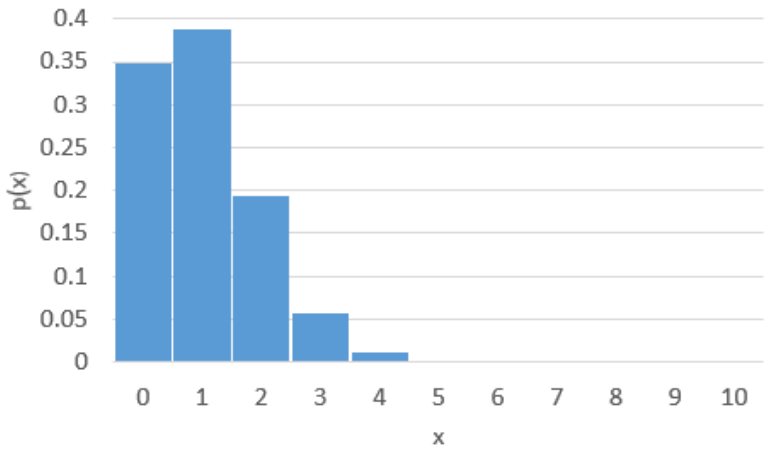
$$P(x = k) = C_k^n p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \qquad \text{for k = 0, 1, …, n}$$

where $n! = n(n-1)(n-2)\ldots(2)(1)$ and $0! = 1$

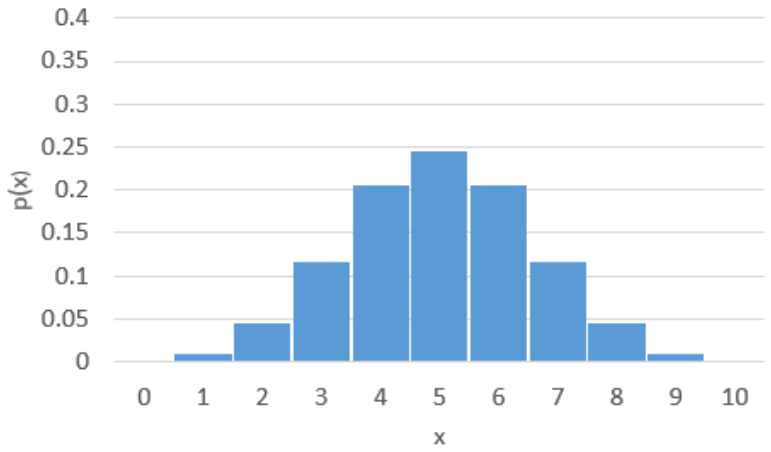
- Distribution of random variable *x* (the number of successes in *n* trials) has
  - **Mean** $\mu = np$
  - **Standard deviation** $\sigma = \sqrt{np(1-p)}$

The Binomial Probability Distribution

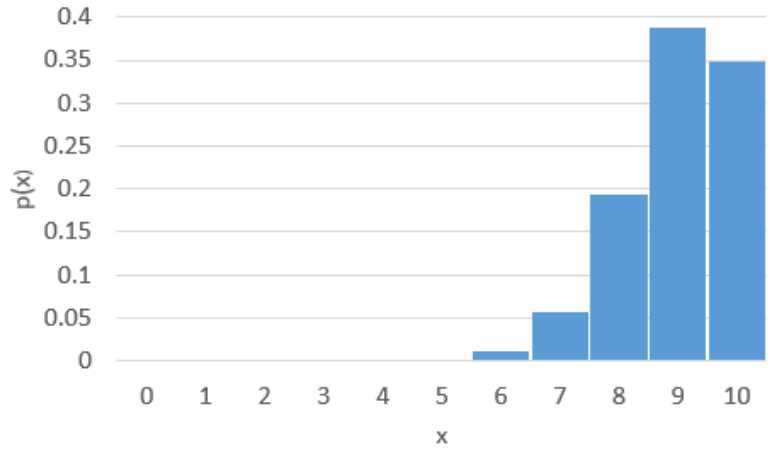
- Examples of binomial probability distribution



n = 10, p = .1
mean $\mu = np = 1$
std $\sigma = .95$



n = 10, p = .5
mean $\mu = np = 5$
std $\sigma = 1.58$



n = 10, p = .9
mean $\mu = np = 9$
std $\sigma = .95$

- Notice the shape of the distribution and expected value (mean) in each case.

The Binomial Probability Distribution - Examples

Example 1: What is the probability of tossing a coin 10 times and seeing 6 heads?

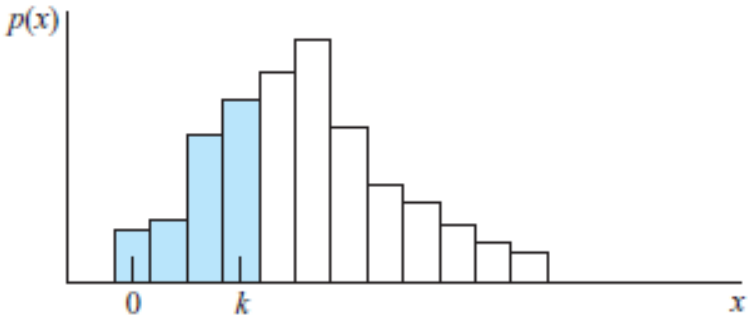
n = 10, p = 0.5, k = 6 => $P(6) = \frac{10!}{6!(4)!} 0.5^6 (0.5)^4 = 0.205$

Example 2: 80% of pet insurance owners are women. If we randomly pick 9 pet insurance owners, what is the probability of have 5 females?

Source: https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/binomial-theorem/binomial-distribution-formula/

The Binomial Probability Distribution

Table of Cumulative Binomial Probabilities provides value of $\mathrm{P}(\mathrm{x} \leq k)$



$n = 2$

					p								
k	.01	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95	.99	k
0	.980	.902	.810	.640	.490	.360	.250	.160	.090	.040	.010	.002	.000	0
1	1.000	.998	.990	.960	.910	.840	.750	.640	.510	.360	.190	.098	.020	1
2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	2

$n = 3$

					p								
k	.01	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95	.99	k
0	.970	.857	.729	.512	.343	.216	.125	.064	.027	.008	.001	.000	.000	0
1	1.000	.993	.972	.896	.784	.648	.500	.352	.216	.104	.028	.007	.000	1
2	1.000	1.000	.999	.992	.973	.936	.875	.784	.657	.488	.271	.143	.030	2
3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	3

$n = 4$

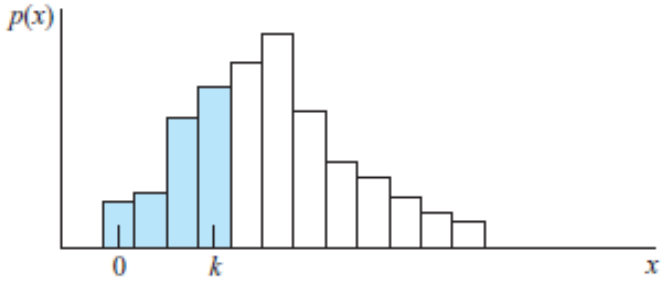
					p								
k	.01	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95	.99	k
0	.961	.815	.656	.410	.240	.130	.062	.026	.008	.002	.000	.000	.000	0
1	.999	.986	.948	.819	.652	.475	.312	.179	.084	.027	.004	.000	.000	1
2	1.000	1.000	.996	.973	.916	.821	.688	.525	.348	.181	.052	.014	.001	2
3	1.000	1.000	1.000	.998	.992	.974	.938	.870	.760	.590	.344	.185	.039	3
4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	4

The Binomial Probability Distribution - Examples

Example 3: 60% of sport car buyers are men. If we randomly pick 25 of sport car buyers, what is the probability of have 10 men?

Solution hints

Consult the cumulative binomial probability table for n = 25, p = 0.6



$n = 25$

	p													
k	.01	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95	.99	k
0	.778	.277	.072	.004	.000	.000	.000	.000	.000	.000	.000	.000	.000	0
1	.974	.642	.271	.027	.002	.000	.000	.000	.000	.000	.000	.000	.000	1
2	.998	.873	.537	.098	.009	.000	.000	.000	.000	.000	.000	.000	.000	2
3	1.000	.966	.764	.234	.033	.002	.000	.000	.000	.000	.000	.000	.000	3
4	1.000	.993	.902	.421	.090	.009	.000	.000	.000	.000	.000	.000	.000	4
5	1.000	.999	.967	.617	.193	.029	.002	.000	.000	.000	.000	.000	.000	5
6	1.000	1.000	.991	.780	.341	.074	.007	.000	.000	.000	.000	.000	.000	6
7	1.000	1.000	.998	.891	.512	.154	.022	.001	.000	.000	.000	.000	.000	7
8	1.000	1.000	1.000	.953	.677	.274	.054	.004	.000	.000	.000	.000	.000	8
9	1.000	1.000	1.000	.983	.811	.425	.115	.013	.000	.000	.000	.000	.000	9
10	1.000	1.000	1.000	.994	.902	.586	.212	.034	.002	.000	.000	.000	.000	10
11	1.000	1.000	1.000	.998	.956	.732	.345	.078	.006	.000	.000	.000	.000	11
12	1.000	1.000	1.000	1.000	.983	.846	.500	.154	.017	.000	.000	.000	.000	12
13	1.000	1.000	1.000	1.000	.994	.922	.655	.268	.044	.002	.000	.000	.000	13
14	1.000	1.000	1.000	1.000	.998	.966	.788	.414	.098	.006	.000	.000	.000	14
15	1.000	1.000	1.000	1.000	1.000	.987	.885	.575	.189	.017	.000	.000	.000	15
16	1.000	1.000	1.000	1.000	1.000	.996	.946	.726	.323	.047	.000	.000	.000	16
17	1.000	1.000	1.000	1.000	1.000	.999	.978	.846	.488	.109	.002	.000	.000	17
18	1.000	1.000	1.000	1.000	1.000	1.000	.993	.926	.659	.220	.009	.000	.000	18
19	1.000	1.000	1.000	1.000	1.000	1.000	.998	.971	.807	.383	.033	.001	.000	19
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.991	.910	.579	.098	.007	.000	20
21	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.998	.967	.766	.236	.034	.000	21
22	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.991	.902	.463	.127	.002	22
23	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.998	.973	.729	.358	.026	23
24	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.996	.928	.723	.222	24
25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	25

Source: https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/binomial-theorem/binomial-distribution-formula/

The Poisson Probability Distribution

- Poisson's probability distribution
  - a good model for representing *the number of events* over a unit of *time* or *space.*
  - The events must occur *randomly* and *independently* (i.e. not at the same time)
  - The average time (distance) between events is *known* but their exact timing (location) is *unknown*.

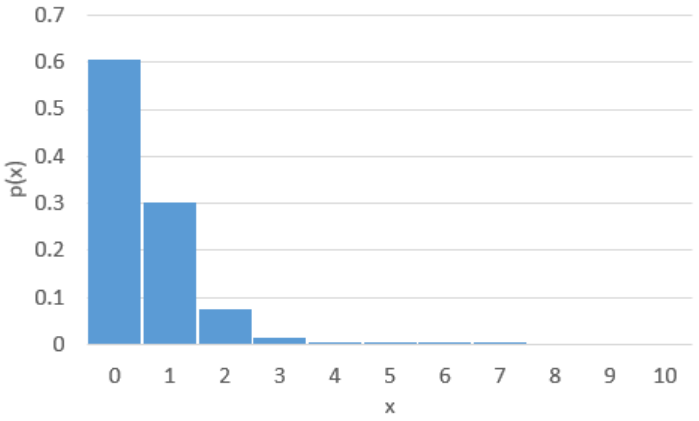
- Examples:
  - The number of machine breakdowns during a given day
  - The number of traffic accidents at a given intersection during a given time period
  - The number passengers arriving at a bus stop in a given time window.

The Poisson Probability Distribution

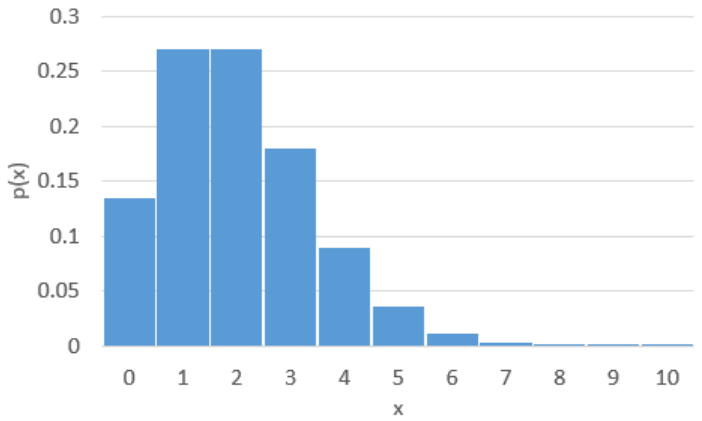
- The probability of k occurrences for the average number of occurrences μ

$$P(x = k) = \frac{\mu^k e^{-\mu}}{k!} \qquad \text{for k = 1, 2, 3, …}$$

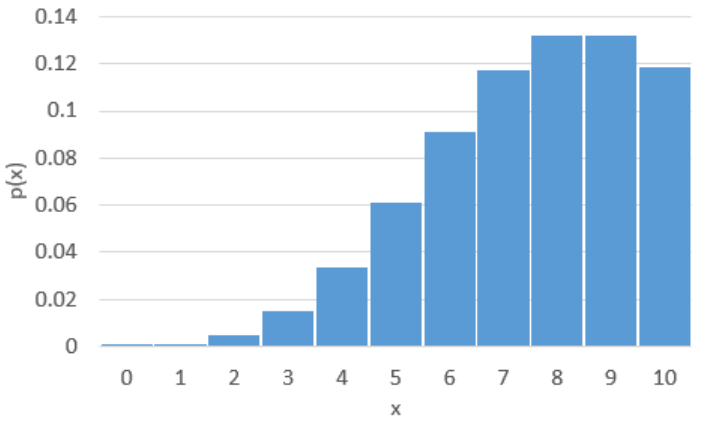
- Mean μ
- Standard deviation $\sigma = \sqrt{\mu}$



μ=0.5, σ=0.707



μ=2, σ=1.414



μ=9, σ=3

- Notice the mode = expected value (mean) and that p(x= μ) = p(x= μ-1)

The Poisson Probability Distribution

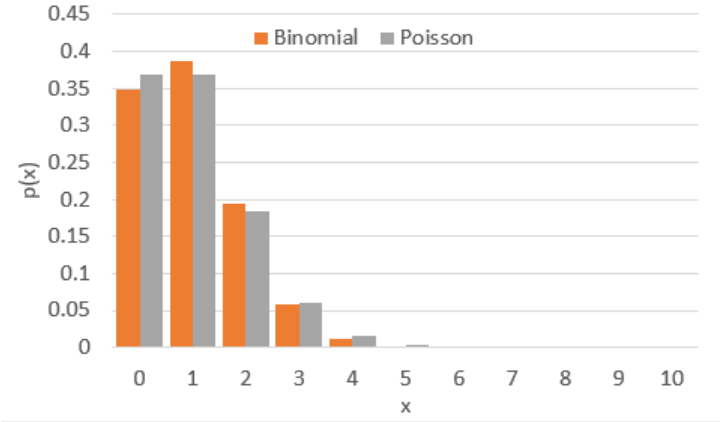
Example 1. The average number of traffic accidents on certain section of highway is 2 per week. What is (a) the probability of having at most 3 accidents and (b) the probability of having exactly 3 accidents a week?

	μ										
k	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0
0	.135	.082	.055	.033	.018	.011	.007	.004	.003	.002	.001
1	.406	.287	.199	.136	.092	.061	.040	.027	.017	.011	.007
2	.677	.544	.423	.321	.238	.174	.125	.088	.062	.043	.030
3	.857	.758	.647	.537	.433	.342	.265	.202	.151	.112	.082
4	.947	.891	.815	.725	.629	.532	.440	.358	.285	.224	.173
5	.983	.958	.916	.858	.785	.703	.616	.529	.446	.369	.301
6	.995	.986	.966	.935	.889	.831	.762	.686	.606	.563	.450
7	.999	.996	.988	.973	.949	.913	.867	.809	.744	.673	.599
8	1.000	.999	.996	.990	.979	.960	.932	.894	.847	.792	.729
9		1.000	.999	.997	.992	.983	.968	.946	.916	.877	.830
10			1.000	.999	.997	.993	.986	.975	.957	.933	.901
11				1.000	.999	.998	.995	.989	.980	.966	.947
12					1.000	.999	.998	.996	.991	.984	.973
13						1.000	.999	.998	.996	.993	.987
14							1.000	.999	.999	.997	.994
15								1.000	.999	.999	.998
16									1.000	1.000	.999
17											1.000

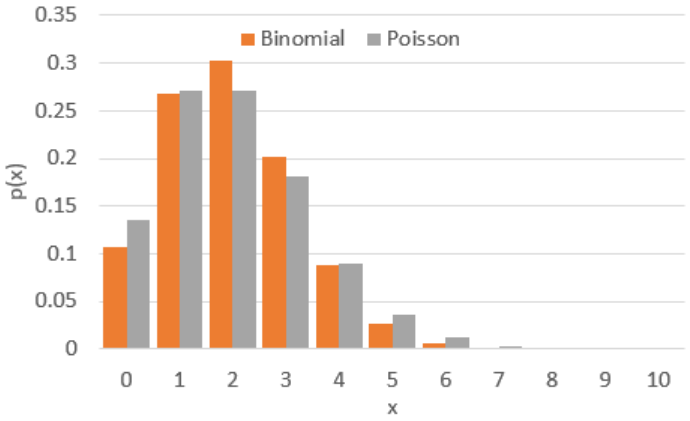
Solution hints. Assuming the number of weekly accidents follows a Poisson distribution. Consult the cumulative Poisson probability table with μ=2.

The Poisson Probability Distribution

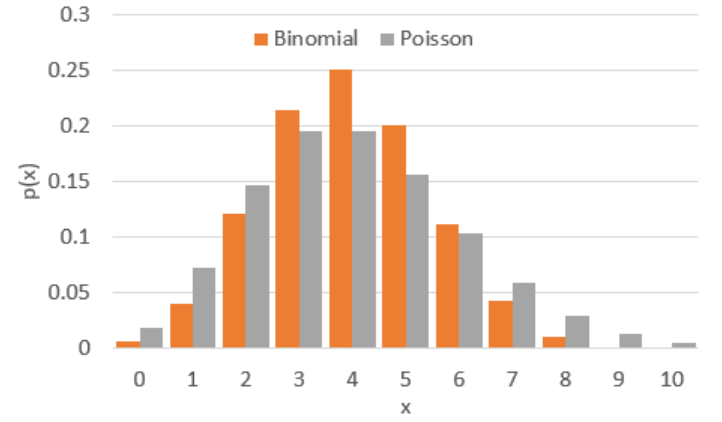
- Poisson distribution can be a good approximation of a binomial distribution that has small μ=np and preferably *large* n.



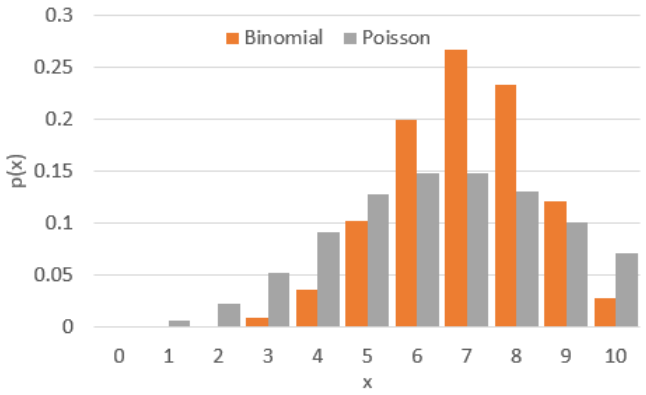
n = 10, p = 0.1, μ=1, MAD = 0.058



n = 10, p = 0.2, μ=2, MAD=0.104

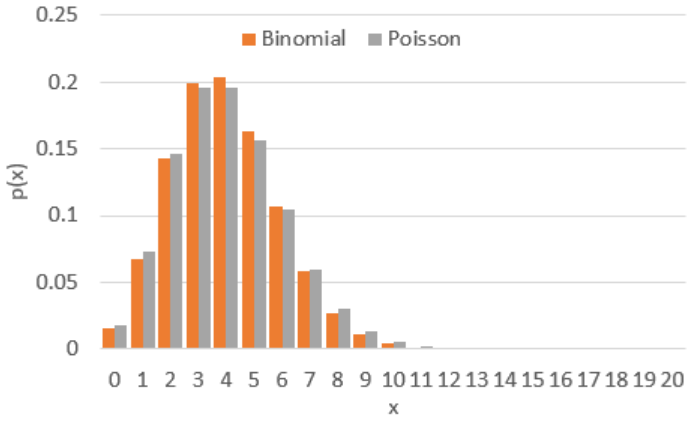


n=10, p=0.4, μ=4, MAD=0.251

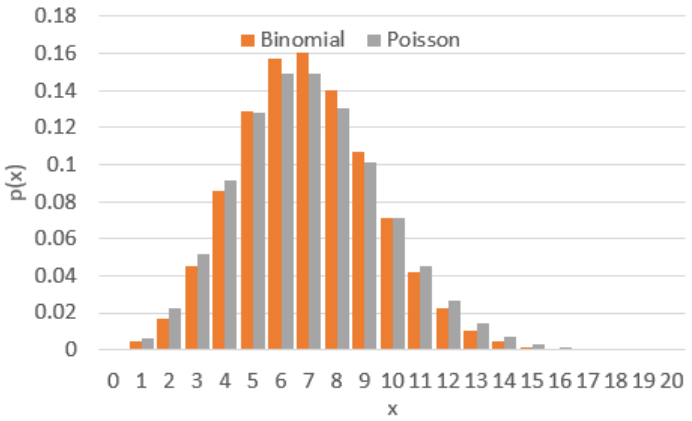


n = 10, p = 0.7, μ=7, MAD=0.485

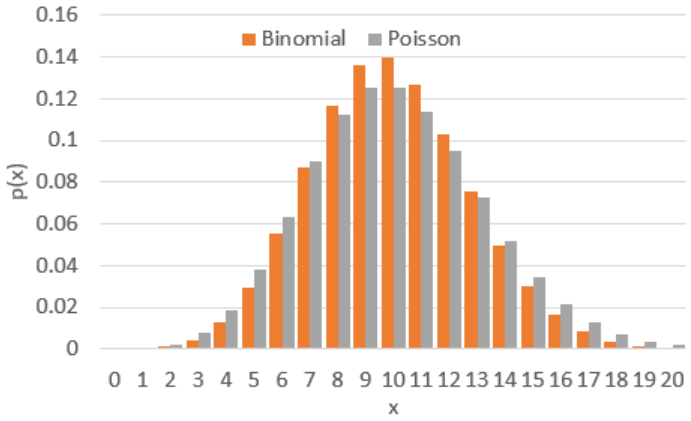
The Poisson Probability Distribution



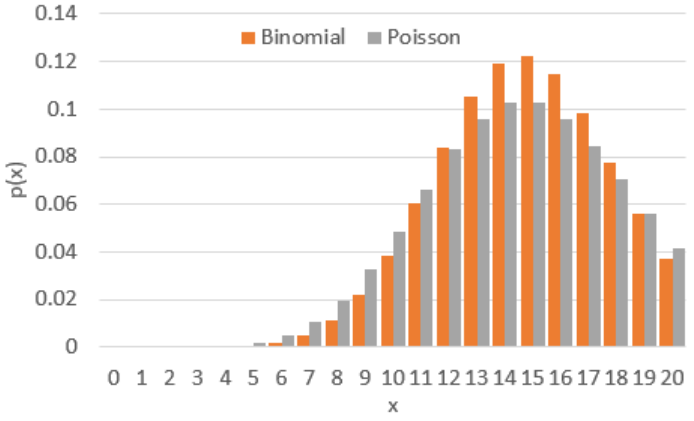
n = 50, p = 0.08, μ=4, MAD=0.042



n = 50, p = 0.14, μ=7, MAD=0.073



n = 50, p = 0.2, μ=10, MAD=0.108



n = 50, p = 0.3, μ=15, MAD=0.136

The Poisson Probability Distribution

Example 2. Assume the probability of a defective engine is p=0.001. Given a batch of 1000 engines, what is the probability of having 4 defective engines?

Solutions. This is a binomial experiment with n = 1000, p = 0.001, probability of having 4

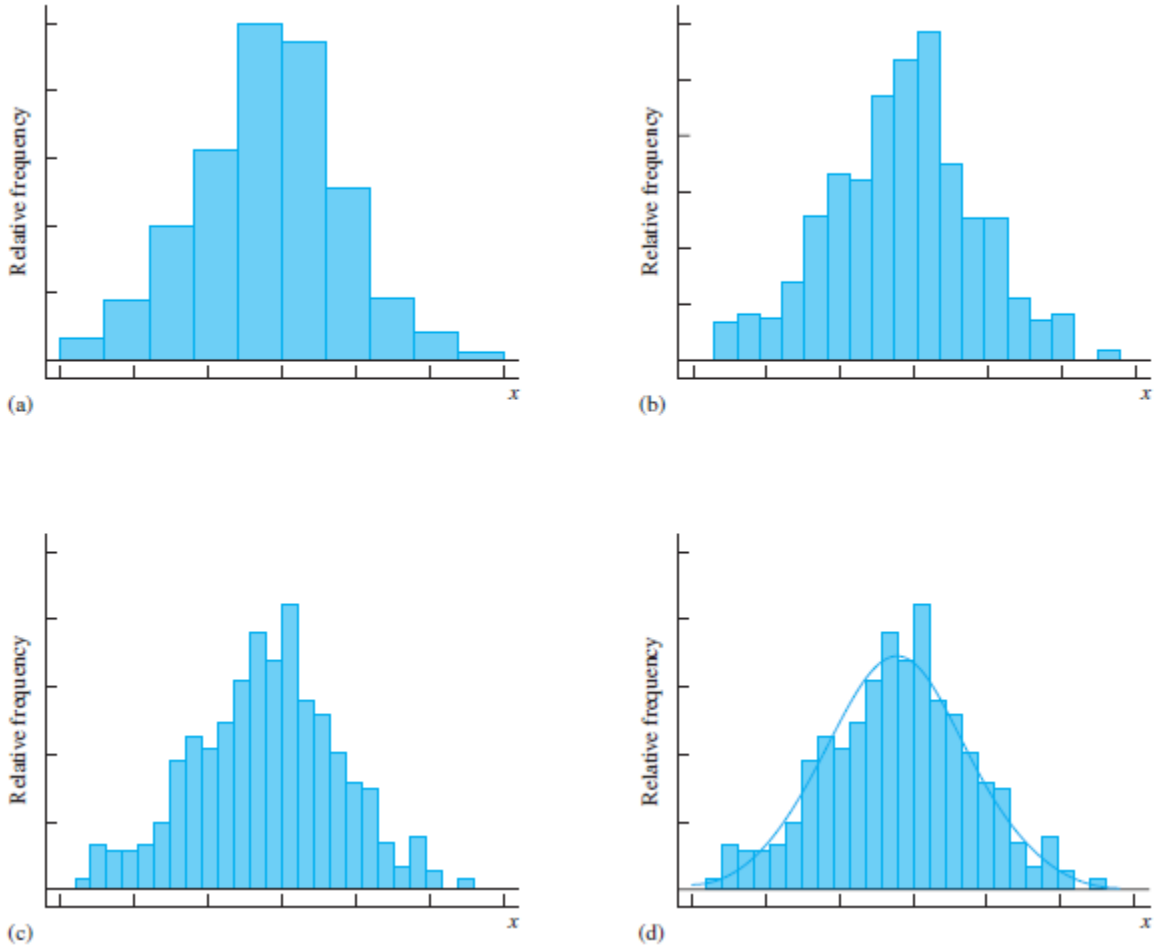
$$P(k=4)=\frac{1000!}{4!\,(996)!}(0.001)^4(1-0.001)^{996}=0.01529$$

Alternatively because mean of this binomial distribution is small $\mu = np = 1$, we can approximate it with a Poisson distribution with $\mu = 1$.

$$P(x=4)=\frac{1^4e^{-1}}{4!}=0.01533$$

Probability Distributions for Continuous Random Variables

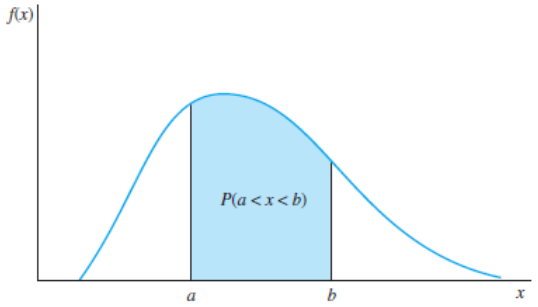
A **continuous variable** can take unlimited number of values in a given range.



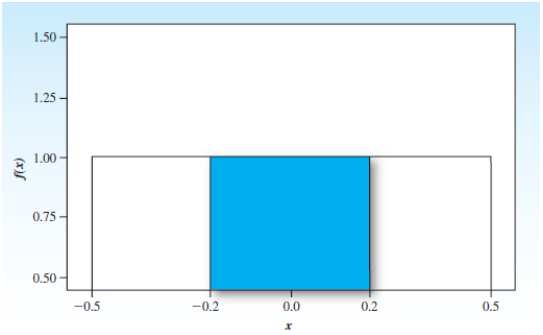
Relative frequency histograms for increasingly large number of samples of a continuous random variable

Probability Distributions for Continuous Random Variables

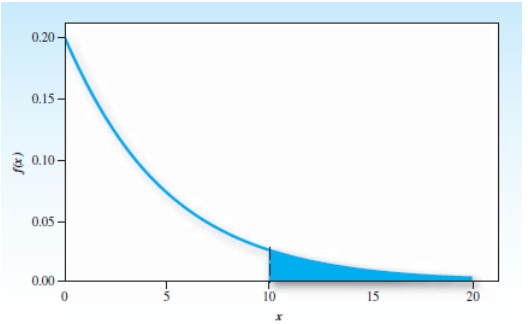
- Characteristics of a probability distribution f(x)
  - The area under the distribution equals to 1
  - $P(a < x < b)$ equals to the area between a and b
  - $P(x = a) = 0$ because there is no area above $x = a$
  - $P(x \geq a) = P(x > a)$ and $P(x \leq a) = P(x < a)$



- Examples of continuous random variables
  - Rounding error x to nearest integer of values between -0.5 and 0.5 has a uniform distribution (because they *all* become 0), f(x)=1



  - The wait time x at a supermarket checkout may follow an exponential distribution, $f(x) = .2e^{-.2x}$

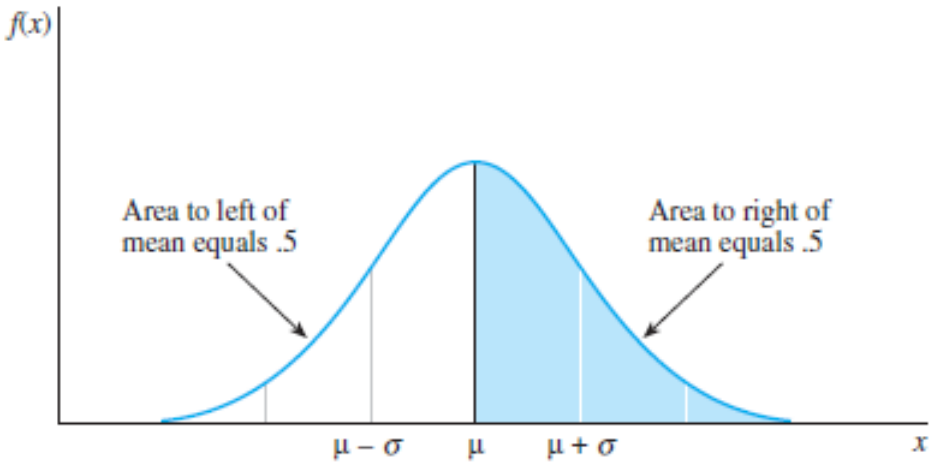


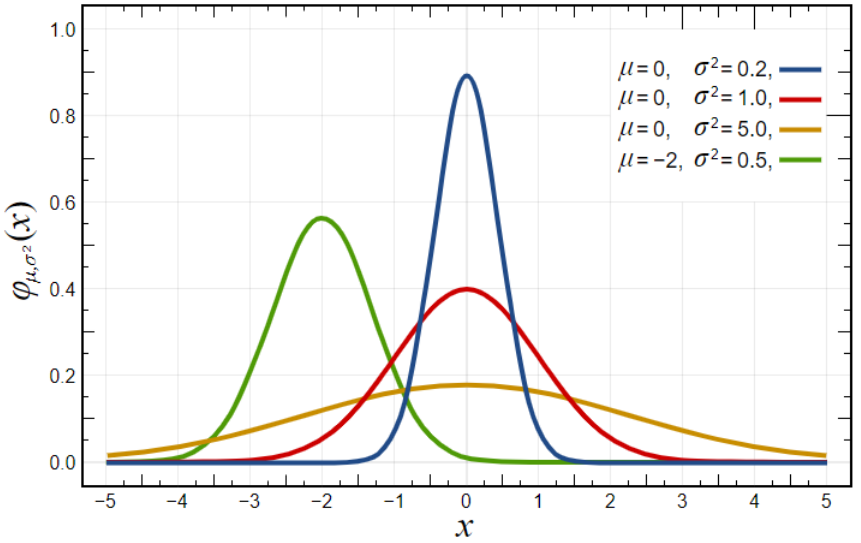
The Normal Probability Distribution

- Many continuous random variables in nature (weight, height, time) can be well described by **normal probability distribution** (thus the name *normal*)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \qquad \text{for } -\infty \le x \le \infty$$

- The distribution is symmetric about the mean μ, which is also the mode and median.
- The shape of the curve is determined by the population standard deviation σ.
- The Empirical Rule!

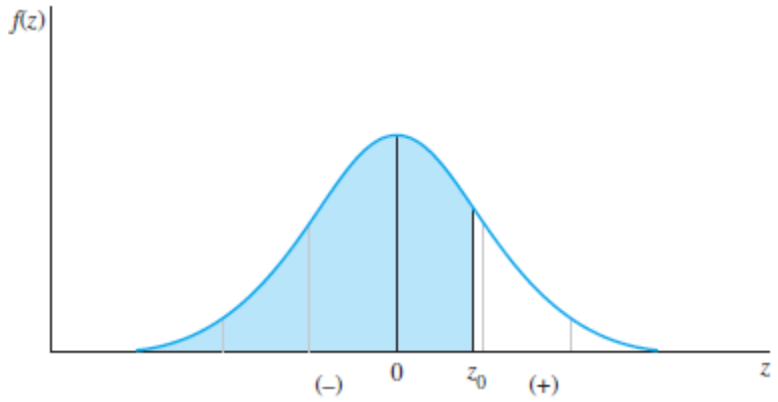




Source: https://en.wikipedia.org/wiki/Normal_distribution

Tabulated Areas of the Normal Probability Distribution

- **Standardized normal random variable** z is defined as $z = \frac{x-\mu}{\sigma}$, essentially the number of σ the variable x lies to the left or right of μ.
- The probability distribution of z is called the s**tandardized normal distribution** because the mean is 0 and standard deviation is 1.
- Value of $P(z<z_0)$ is the shaded area and is tabulated, an extract of which is given below.



z	.00	.01	.02	.03	.04	. . .	.09
0.0	.5000	.5040	.5080	.5120	.5160		
0.1	.5398	.5438	.5478	.5517	.5557		
0.2	.5793	.5832	.5871	.5910	.5948		
0.3	.6179	.6217	.6255	.6293	.6331		
0.4	.6554	.6591	.6628	.6664	.6700	. . .	.6879
0.5	.6915	.	.	.			
0.6	.7257	.	.	.			
0.7	.7580	.	.	.			
0.8	.7881						
0.9	.8159						
.	.						
.	.						
.	.						
2.0	.9772						

Example: P(z<0.41) = ?

Example. Let x be a normal distributed variable with $\mu = 10$ and $\sigma = 2$. Find the probability of x lies between 9.4 and 10.6.

Solution hints

Calculate the standardized normal random variable z_1 corresponding to $x_1 = 9.4$

Calculate the standardized normal random variable z_2 corresponding to $x_2 = 10.6$

$$P(x_1 < x < x_2) = P(z_1 < z < z_2)$$

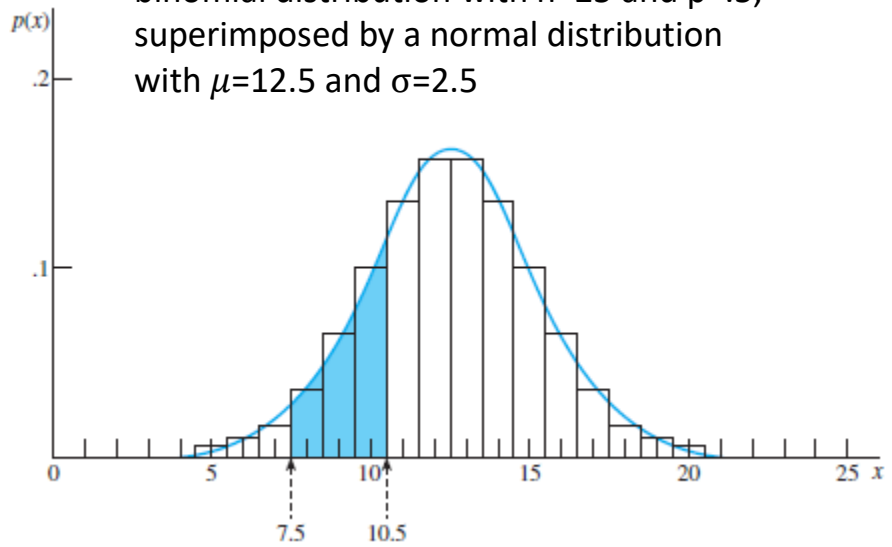
z	.00	.01
−3.4	.0003	.0003
−3.3	.0005	.0005
−3.2	.0007	.0007
−3.1	.0010	.0009
−3.0	.0013	.0013
−2.9	.0019	.0018
−2.8	.0026	.0025
−2.7	.0035	.0034
−2.6	.0047	.0045
−2.5	.0062	.0060
−2.4	.0082	.0080
−2.3	.0107	.0104
−2.2	.0139	.0136
−2.1	.0179	.0174
−2.0	.0228	.0222
−1.9	.0287	.0281
−1.8	.0359	.0351
−1.7	.0446	.0436
−1.6	.0548	.0537
−1.5	.0668	.0655
−1.4	.0808	.0793
−1.3	.0968	.0951
−1.2	.1151	.1131
−1.1	.1357	.1335
−1.0	.1587	.1562
−0.9	.1841	.1814
−0.8	.2119	.2090
−0.7	.2420	.2389
−0.6	.2743	.2709
−0.5	.3085	.3050
−0.4	.3446	.3409
−0.3	.3821	.3783
−0.2	.4207	.4168
−0.1	.4602	.4562
−0.0	.5000	.4960

z	.00	.01
0.0	.5000	.5040
0.1	.5398	.5438
0.2	.5793	.5832
0.3	.6179	.6217
0.4	.6554	.6591
0.5	.6915	.6950
0.6	.7257	.7291
0.7	.7580	.7611
0.8	.7881	.7910
0.9	.8159	.8186
1.0	.8413	.8438
1.1	.8643	.8665
1.2	.8849	.8869
1.3	.9032	.9049
1.4	.9192	.9207
1.5	.9332	.9345
1.6	.9452	.9463
1.7	.9554	.9564
1.8	.9641	.9649
1.9	.9713	.9719
2.0	.9772	.9778
2.1	.9821	.9826
2.2	.9861	.9864
2.3	.9893	.9896
2.4	.9918	.9920
2.5	.9938	.9940
2.6	.9953	.9955
2.7	.9965	.9966
2.8	.9974	.9975
2.9	.9981	.9982
3.0	.9987	.9987
3.1	.9990	.9991
3.2	.9993	.9993
3.3	.9995	.9995
3.4	.9997	.9997

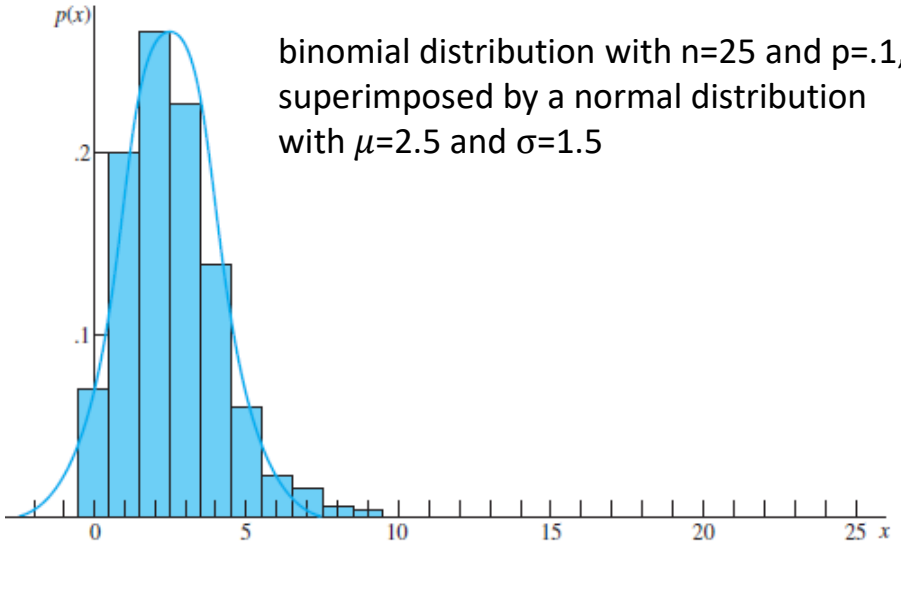
The Normal Approximation to the Binomial Distribution

- Probability of a binomial variable x can be calculated via
  - the binomial formula or corresponding binomial tables, or
  - the Poisson probabilities for $np < 7$.
- When $\mu = np$ of a binomial distribution is large, the normal probability with $\mu = np$ and standard deviation $\sigma = \sqrt{np(1-p)}$ can be used as an approximation.
- For this approximation to hold, *n* must be large and *p* is not too close to 0 or 1.

binomial distribution with n=25 and p=.5, superimposed by a normal distribution with μ=12.5 and σ=2.5



binomial distribution with n=25 and p=.1, superimposed by a normal distribution with μ=2.5 and σ=1.5



The Normal Approximation to the Binomial Distribution

Rule of thumb. A normal distribution approximates well a binomial distribution if both np>5 AND n(1-p)>5 (because the binomial distribution is *fairly symmetric*)

Example. A random sample of 1000 fuses were tested. Assuming defect probability is 0.02. What is the probability of having more than 27 fuses defected.

Solution. This is a binomial experiment with $\mu = np = 20$ and $\sigma = \sqrt{np(1-p)} = 4.43$.

However because both np and $n(1-p)$ is larger than 5, we can approximate it by a normal distribution with μ = 20 and σ = 4.43.

Because of **continuity correction**, the normal area corresponding to $P(x \geq 27)$ is the area to the right of x=26.5.

Standardised value of x=26.5 is $z_0 = \frac{26.5-20}{4.43} = 1.47$

Therefore $P(x \geq 27) \approx P(z > 1.47) = 1 - .9292 = .0708$