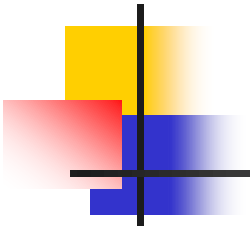
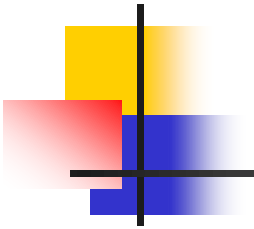


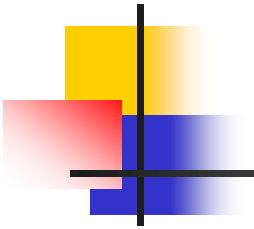
Data Modeling & Analysis Techniques

Probability Distributions

Experiment and Sample Space



- A (random) experiment is a procedure that has a number of possible outcomes and it is not certain which one will occur
- The sample space is the set of all possible outcomes of an experiment (often denoted by S).
  - Examples:
    - Coin : S={H, T}
    - Two coins: S={HH, HT, TH, TT}
    - Lifetime of a system: S={0..∞}

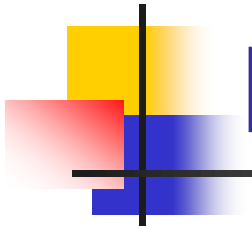


Probability Distributions

- Probability distributions represent the likelihood of certain events
  - Probability "mass" (or density for continuous variables) represents the amount of likelihood attributed to a particular point
  - Cumulative distribution represents the accumulated probability "mass" at a particular point
    - Distributions in probability are usually given and their results are computed
    - Distributions (or their parameters) are usually the items to be estimated in statistics

PMF, PDF and CDF

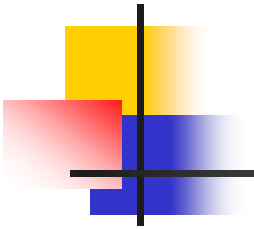
	Discrete Data	**Continuous Data**
Distributions	Probability Mass Function Cumulative Distribution Function	Probability Density Function Cumulative Distribution Function



PMF, PDF and CDF Discussion

Probability Distributions

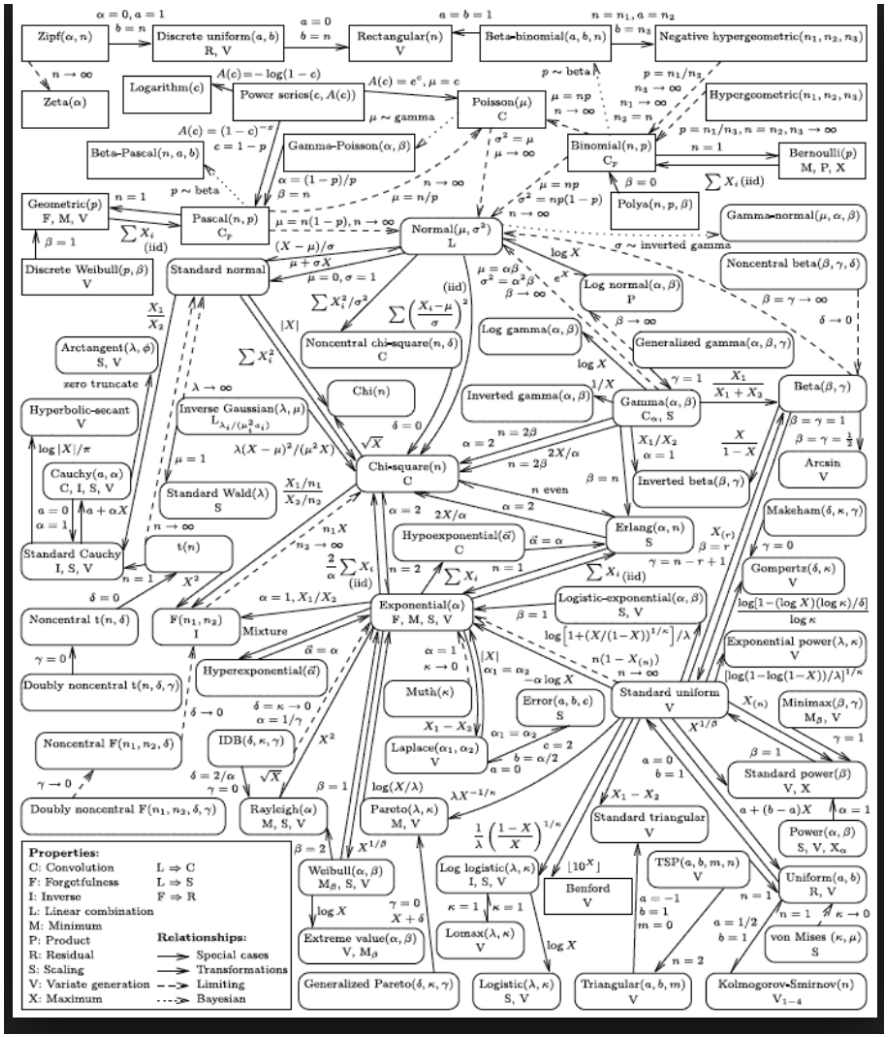
- Distributions can be characterized by their moments
  - r^{th} moment: $E\left[\left(x - a\right)^r\right]$
  - Important moments:
    - Mean: $E\left[\left(x - 0\right)^1\right]$
    - Variance: $E\left[\left(x - m\right)^2\right]$
    - Skewness: $E\left[\left(x - m\right)^3\right]$



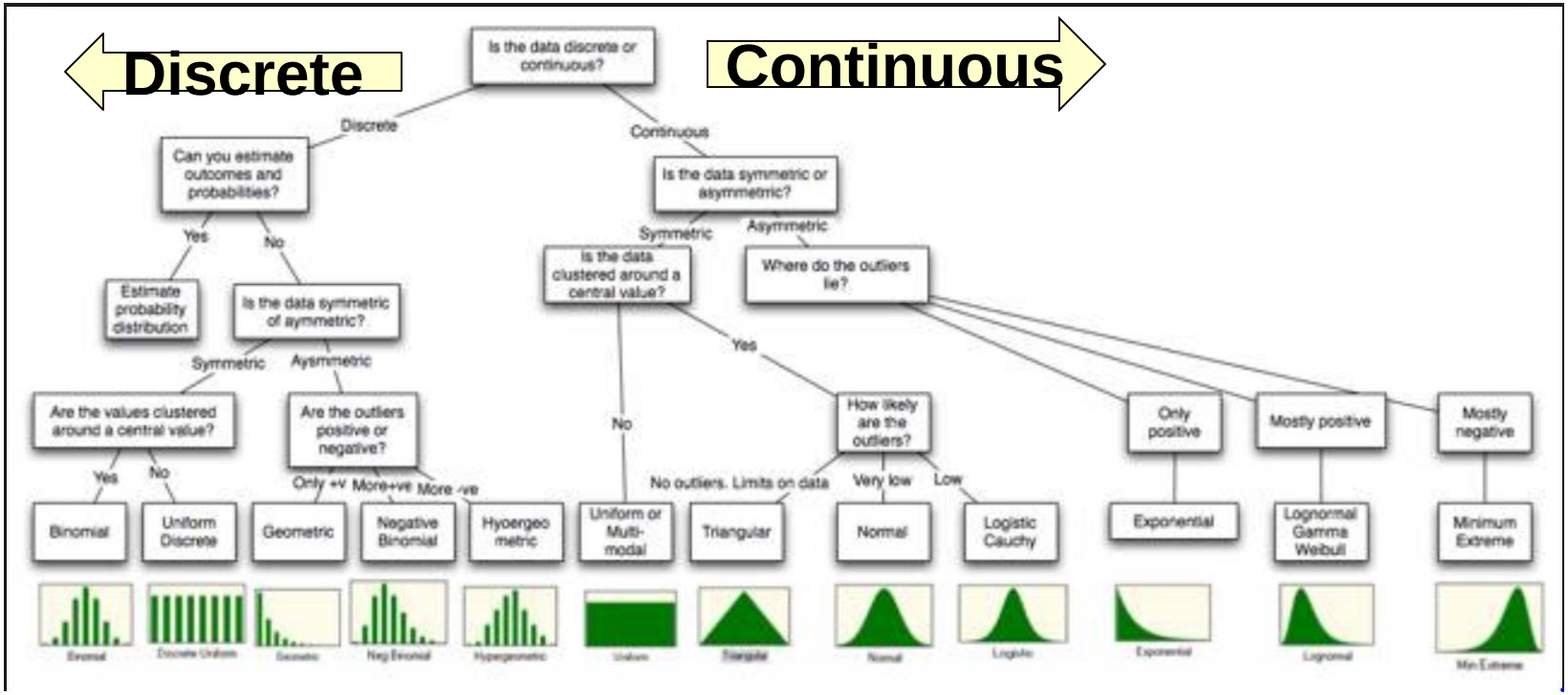
Distributions

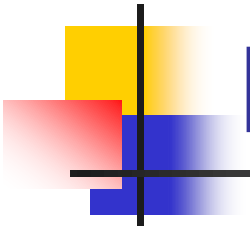
- There are families of important distributions that are useful to model or analyze events
  - Families of distributions are parameterized
  - Different distributions are used to answer different questions about events
    - What is the probability of an individual event
    - How many times would an event happen in a repeated experiment
    - How long will it take until an event happens

Statistical Distributions



Statistical Distbriutions

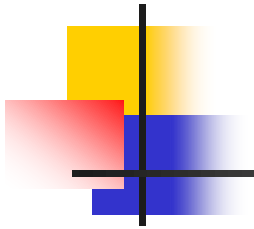




Discrete Distributions

- Bernoulli Distribution
- Binomial Distribution
- Geometric Distribution
- Poisson Distribution
- Hypergeometric
- Uniform Distribution

Distributions-Bernoulli Distribution



- Bernoulli distribution
  - Models the likelihood of one of two possible events happening
    - Ex: good or bad/defective, pass or fail, transmitted or lost signal, benign or malicious attachments, boys or girls, heads or tails, etc..
  - Two outcomes: Successes and Failures
    - Successes do not have to be good or failures do not have to be bad
  - Parameterized by the likelihood, *p*, of event *1*

Bernoulli Distribution Contd..

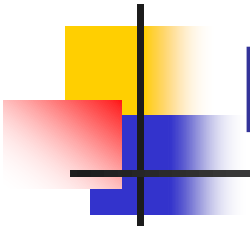
- Bernoulli distribution
  - Probability function:

PMF : $P(x;p) = p^x (1-p)^{1-x}$

$$P(x;p) = P(X = x) = \begin{cases} p & if\ x = 1 \\ 1 - p & otherwise \end{cases}$$

  - Can be easily extended to represent more than two possible events
  - Mean: $\mu = p$ Variance: $\sigma^2 = p*(1-p)$
  - *Example*

Ex. Flip a fair coin. Let X = number of heads. Then X is a Bernoulli random variable with p=1/2.
E(X) = 1/2
Var(X) = 1/4



Problem Review

Binomial Distribution

- Discrete distributions for event frequency
  - Binomial distribution
    - Models the likelihood that an event will occur a certain number of times in *n* Bernoulli experiments
    - 1. Parameterized by the likelihood, *p*, of event *1* in the Bernoulli experiment and the number of experiments, *n*
    - *2. Two Outcomes (Success/Failure)*
      - *i.e yes/no, Dead/live, treated/untreated, sick/well*
    - *3. P(Success)=p, P(Failure)=q or 1-p*
      - *p+q=1*

Binomial Distribution Contd..

- Probability function: The p(x), the probability that there will be exactly x success in n trials

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

where
n = the number of trials (or the number being sampled)
x = the number of successes desired
p = probability of getting a success in one trial
$q = 1 - p$ = the probability of getting a failure in one trial

- Mean: $\mu = np$ Variance: $\sigma^2 = np(1-p)$

Other Characteristics of Binomial

Cumulative distribution function [edit]

The cumulative distribution function can be expressed as:

$$F(k; n, p) = \Pr(X \leq k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$$

$$\text{mode} = \begin{cases} \lfloor (n+1)\,p \rfloor & \text{if } (n+1)p \text{ is 0 or a noninteger,} \\ (n+1)\,p \text{ and } (n+1)\,p - 1 & \text{if } (n+1)p \in \{1, \ldots, n\}, \\ n & \text{if } (n+1)p = n+1. \end{cases}$$

https://en.wikipedia.org/wiki/Binomial_distribution

Binomial Distribution Contd..Calculating from Tables

n always compute it from the table as

$$P(x) = F(x) - F(x-1).$$

Example 3.16. As part of a business strategy, randomly selected 20% of new internet service subscribers receive a special promotion from the provider. A group of 10 neighbors signs for the service. What is the probability that at least 4 of them get a special promotion?

Solution. We need to find the probability $P\{X \geq 4\}$, where X is the number of people, out of 10, who receive a special promotion. This is the number of successes in 10 Bernoulli trials, therefore, X has Binomial distribution with parameters $n = 10$ and $p = 0.2$. From Table A2,

$$P\{X \geq 4\} = 1 - F(3) = 1 - 0.8791 = \underline{0.1209}.$$

◊

Binomial Table A2

$p=$		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
$n=7$	$x=0$	0.9321	0.8681	0.8080	0.7514	0.6983	0.6485	0.6017	0.5578	0.5168	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078
	1	0.9980	0.9921	0.9829	0.9706	0.9556	0.9382	0.9187	0.8974	0.8745	0.8503	0.7166	0.5767	0.4449	0.3294	0.2338	0.1586	0.1024	0.0625
	2	1.0000	0.9997	0.9991	0.9980	0.9962	0.9937	0.9903	0.9860	0.9807	0.9743	0.9262	0.8520	0.7564	0.6471	0.5323	0.4199	0.3164	0.2266
	3	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9993	0.9988	0.9982	0.9973	0.9879	0.9667	0.9294	0.8740	0.8002	0.7102	0.6083	0.5000
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998	0.9988	0.9953	0.9871	0.9712	0.9444	0.9037	0.8471	0.7734
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9987	0.9962	0.9910	0.9812	0.9643	0.9375
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9994	0.9984	0.9963	0.9922
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n=8$	$x=0$	0.9227	0.8508	0.7837	0.7214	0.6634	0.6096	0.5596	0.5132	0.4703	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039
	1	0.9973	0.9897	0.9777	0.9619	0.9428	0.9208	0.8965	0.8702	0.8423	0.8131	0.6572	0.5033	0.3671	0.2553	0.1691	0.1064	0.0632	0.0352
	2	0.9999	0.9996	0.9987	0.9969	0.9942	0.9904	0.9853	0.9789	0.9711	0.9619	0.8948	0.7969	0.6785	0.5518	0.4278	0.3154	0.2201	0.1445
	3	1.0000	1.0000	0.9999	0.9998	0.9996	0.9993	0.9987	0.9978	0.9966	0.9950	0.9786	0.9437	0.8862	0.8059	0.7064	0.5941	0.4770	0.3633
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9997	0.9996	0.9971	0.9896	0.9727	0.9420	0.8939	0.8263	0.7396	0.6367
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9988	0.9958	0.9887	0.9747	0.9502	0.9115	0.8555
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9987	0.9964	0.9915	0.9819	0.9648
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9993	0.9983	0.9961
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n=9$	$x=0$	0.9135	0.8337	0.7602	0.6925	0.6302	0.5730	0.5204	0.4722	0.4279	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020
	1	0.9966	0.9869	0.9718	0.9522	0.9288	0.9022	0.8729	0.8417	0.8088	0.7748	0.5995	0.4362	0.3003	0.1960	0.1211	0.0705	0.0385	0.0195
	2	0.9999	0.9994	0.9980	0.9955	0.9916	0.9862	0.9791	0.9702	0.9595	0.9470	0.8591	0.7382	0.6007	0.4628	0.3373	0.2318	0.1495	0.0898
	3	1.0000	1.0000	0.9999	0.9997	0.9994	0.9987	0.9977	0.9963	0.9943	0.9917	0.9661	0.9144	0.8343	0.7297	0.6089	0.4826	0.3614	0.2539
	4	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9995	0.9991	0.9944	0.9804	0.9511	0.9012	0.8283	0.7334	0.6214	0.5000
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9969	0.9900	0.9747	0.9464	0.9006	0.8342	0.7461
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987	0.9957	0.9888	0.9750	0.9502	0.9102
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9986	0.9962	0.9909	0.9805
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9992	0.9980
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n=10$	$x=0$	0.9044	0.8171	0.7374	0.6648	0.5987	0.5386	0.4840	0.4344	0.3894	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010
	1	0.9957	0.9838	0.9655	0.9418	0.9139	0.8824	0.8483	0.8121	0.7746	0.7361	0.5443	0.3758	0.2440	0.1493	0.0860	0.0464	0.0233	0.0107
	2	0.9999	0.9991	0.9972	0.9938	0.9885	0.9812	0.9717	0.9599	0.9460	0.9298	0.8202	0.6778	0.5256	0.3828	0.2616	0.1673	0.0996	0.0547
	3	1.0000	1.0000	0.9999	0.9996	0.9990	0.9980	0.9964	0.9942	0.9912	0.9872	0.9500	0.8791	0.7759	0.6496	0.5138	0.3823	0.2660	0.1719
	4	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9994	0.9990	0.9984	0.9901	0.9672	0.9219	0.8497	0.7515	0.6331	0.5044	0.3770
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9986	0.9936	0.9803	0.9527	0.9051	0.8338	0.7384	0.6230
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9991	0.9965	0.9894	0.9740	0.9452	0.8980	0.8281
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9984	0.9952	0.9877	0.9726	0.9453
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9983	0.9955	0.9893
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Binomial Distribution Example

Example 5.2: The probability that a patient recovers from a rare blood disease is 0.4. If 15 people are known to have contracted this disease, what is the probability that (a) at least 10 survive, (b) from 3 to 8 survive, and (c) exactly 5 survive?

Solution: Let X be the number of people who survive.

(a) $$P(X \geq 10) = 1 - P(X < 10) = 1 - \sum_{x=0}^{9} b(x; 15, 0.4) = 1 - 0.9662$$
$$= 0.0338$$

(b) $$P(3 \leq X \leq 8) = \sum_{x=3}^{8} b(x; 15, 0.4) = \sum_{x=0}^{8} b(x; 15, 0.4) - \sum_{x=0}^{2} b(x; 15, 0.4)$$
$$= 0.9050 - 0.0271 = 0.8779$$

(c) $$P(X = 5) = b(5; 15, 0.4) = \sum_{x=0}^{5} b(x; 15, 0.4) - \sum_{x=0}^{4} b(x; 15, 0.4)$$
$$= 0.4032 - 0.2173 = 0.1859$$

Distributions

- Discrete distributions for inter-event timing
  - Geometric distribution
    - Models the likelihood that an event will occur for the first time in the x^{th} Bernoulli experiment
    - Parameterized by the probability, p, of the event in each Bernoulli experiment
    - Probability function:

$$P(x;p) = (1 - p)^{x-1} p$$

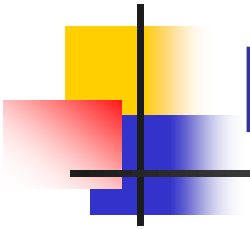
    - Mean: μ=1/p Variance: $\sigma^2=(1-p)/p^2$

Distributions

- Hypergeometric distribution
  - Models the likelihood that an event type will occur a certain number of times in *n* experiments if no specific event can occur twice and they are all equally likely
  - Parameterized by the total number of events, *N*, the number of events of the event type, *M*, and the number of experiments, *n*
  - Probability function:

$$P(x;M,N,n)=\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

  - Mean: $\mu=nM/N$ Variance: $\sigma^2=n(M(n-M)(N-n)/(N^2(N-1))$



Example*

we have a “Lot” of 100 items of which 12 are defective. What is the probability that in a sample of 10, 3 are defective?

Hypergeometric Distribution Example

- Lots of 40 components each are deemed unacceptable if they contain 3 or more defectives. The procedure for sampling a lot is to select 5 components at random and to reject the lot if a defective is found. What is the probability that exactly 1 defective is found in the sample if there are 3 defectives in the entire lot? *

Distributions

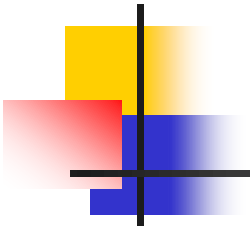
- Poisson distribution
  - Models the likelihood that an event will occur a given number of times in a continuous experiment with constant likelihood that does not depend on the time since the last occurrence
  - Parameterized by the expected number of occurrences, λ, of the event within one time period
  - Probability function:

$$P(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

  - Mean: $E[x]=\mu=\lambda$ Variance: $\sigma^2=\lambda$

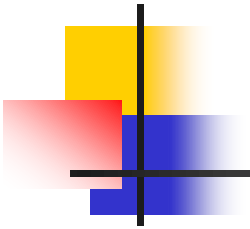
Poisson Distribution Example

- A radioactive source emits 4 particles on average during a 5 second period
  - A. Calculate the probability that it emits 3 particles during a 5-second period
  - B. Calculate the probability that it emits "at least" one particle during a 5-second period



Poisson Distribution Example

- UTA has a printing system and the print jobs arrival follows a Poisson distribution at an average of 2 jobs/min. Determine the probability that in any one minute interval there will be
- A. 0 jobs
- B. exactly two jobs
- C. at most 3 jobs



Poisson Distribution Example

- In a news paper (not UTA communication paper), an average of 3 typos were found.
- A. What is the probability that a randomly selected page has at leas one typo
- B. What is the probability at most one typo per page
- C. Exactly 4 typos on it

Poisson Distribution Example

Calculating from Tables

Example 3.22 (NEW ACCOUNTS). Customers of an internet service provider initiate new accounts at the average rate of 10 accounts per day.

(a) What is the probability that more than 8 new accounts will be initiated today?
(b) What is the probability that more than 16 accounts will be initiated within 2 days?

Solution. (a) New account initiations qualify as rare events because no two customers open accounts simultaneously. Then the number X of today's new accounts has Poisson distribution with parameter $\lambda = 10$. From Table A3,

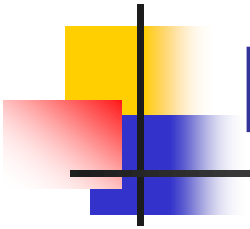
$$P\{X > 8\} = 1 - F_X(8) = 1 - 0.333 = \underline{0.667}.$$

Poisson Distribution Example

- A radioactive source emits 4 particles on average during a 5 second period
  - A. Calculate the probability that it emits 3 particles during a 5-second period
  - B. Calculate the probability that it emits at least one particle during a 5-second period

Distributions-Uniform Distributions

- Discrete distributions for event probability
- Uniform distribution
  - Models the likelihood of a set of events assuming they are all equally likely
  - Parameterized by the number of discrete events, *N*
  - Probability function: $$P(x;N) = P(X = x) = \frac{1}{N}$$
  - If the events are integers in the interval [*a..b*] (with *N=b-a+1*) we can compute a mean and variance
  - Mean: E(X)=$\mu=(b+a)/2$ Variance: $\sigma^2=(N^2-1)/12$



Example

- Let the RV X denote the number of 48 voice lines that are used at a particular time. Assume that X is a discrete uniform RV with a range of 0 to 48

What is the mean and variance?

Example *

- Let the continuous variable X denote the current measured in a thin copper wire in mill amperes. Assume that the range of X is [4.9,5.1]mA, and assume that the pdf of X is f(x)= 5 for 4.9 <= x<= 5.1. What is the probability that a measurement of current is bet. 4.95 and 5.0 milli amperes