4/18/12

Poisson Distribution

Used to model:
1. Number of deaths by horse kicks in a year in Prussian cavalry
2. Number of phone calls per minute to a call center
3. Number of mutations of a stretch of DNA after exposure to radiation
4. Clustering of galaxies in the Universe

A Poisson random variable models the number of events in a specified period of time.

Let jobs arrive at a server with rate λ jobs/second. Let the period of observation be **t**. Let α = λ**t**.

Let X be a Poisson random variable that denotes the number jobs arrived at the server in time **(0,t]**.

Then X ~ Poisson(α) and its pmf and CDF are as follows.

pmf: $$\underbrace{P(X=k)}_{p(k)} = \frac{\alpha^k}{k!} e^{-\alpha}, \quad k=0,1,\ldots$$

$$P(X=0) = \frac{\alpha^0}{0!} e^{-\alpha} = e^{-\alpha}$$

$$P(X=1) = \frac{\alpha^1}{1!} e^{-\alpha} = \alpha e^{-\alpha}$$

$$P(X=2) = \frac{\alpha^2}{2!} e^{-\alpha} = \frac{\alpha^2 e^{-\alpha}}{2}$$

$$E(X) = \alpha$$

$$V(X) = \alpha$$

CDF: $$F(k) = P(X \le k) = \sum_{i=0}^{k} p(k), \quad k=0,1,\ldots$$

Connection with the binomial distribution

Consider a binomial distribution with n trials and p as the probability of success.
Then np= α, where α is parameter of the corresponding Poisson distribution.

Problem: Ten percent of tools produced by a manufacturer tend to be defective. Find the probabilty that in a sample of 10 tools, exactly two tools are defective.

$p = 0.1, \quad n = 10 \qquad \alpha = np = 0.1 * 10 = 1$

<u>Binomial</u> (n, p)

$b(2; 10, 0.1) = \binom{10}{2} 0.1^2 \, 0.9^8$

$= 0.19$

<u>Poisson</u> (α), $\alpha = np$

$P(X=2) = \frac{\alpha^2}{2!} \cdot e^{-\alpha} = \frac{1^2}{2!} e^{-1} = 0.18$

problem:

Let X ~ Poisson(3t) denote the number of packets transmitted on a communication link in t seconds. Calculate' the probability of transmitting exactly three packets in 10 seconds.

$\alpha = 3t, \quad t = 10s \quad \Rightarrow \alpha = 3 * 10 = 30$

$P(X=3) = p(3) = \frac{\alpha^3}{3!} e^{-\alpha} = \frac{30^3}{6} \cdot e^{-30} = 0.42 \times 10^{-9} \approx 0$

Calculate the prob. that <u>at most 20 packets</u> are transmitted in a 20s interval.

Now, $\alpha = 3 * 20 = 60$

$P(X \le 20) = F(20) = \sum_{i=0}^{20} \frac{\alpha^i}{i!} e^{-\alpha} = \sum_{i=0}^{20} \frac{60^i}{i!} e^{-60}$

Continuous Random Variables

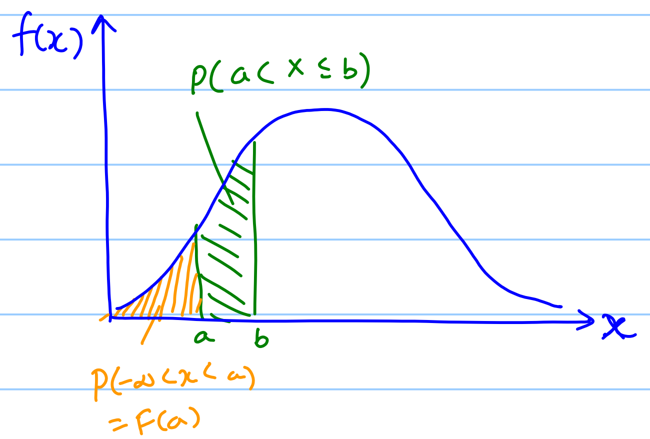
Let X be a RV.

$F(x) = P(X \le x), \quad -\infty < x < \infty$

If $F(x)$ is continuous instead of discrete, then X is a continuous RV.

Then, X does not have a pmf. – instead it has the prob. density function (pdf), denoted $f(x)$, and is given by

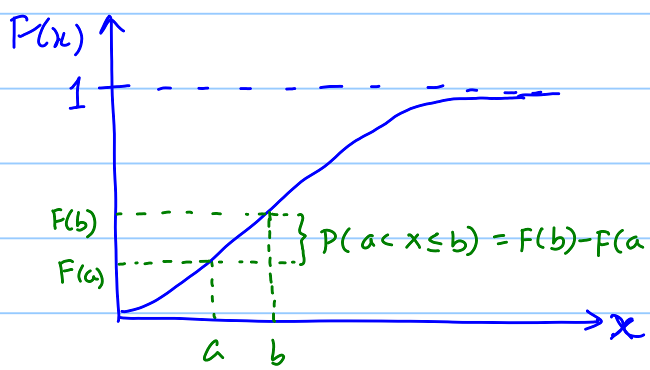
$$f(x) = \frac{d\,F(x)}{dx}$$



$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

$$P(a \le X \le b) = \int_{a}^{b} f(x)\,dx$$

$$P(X = c) = \int_{c}^{c} f(x)\,dx = 0$$



$P(a \le X \le b)$ {

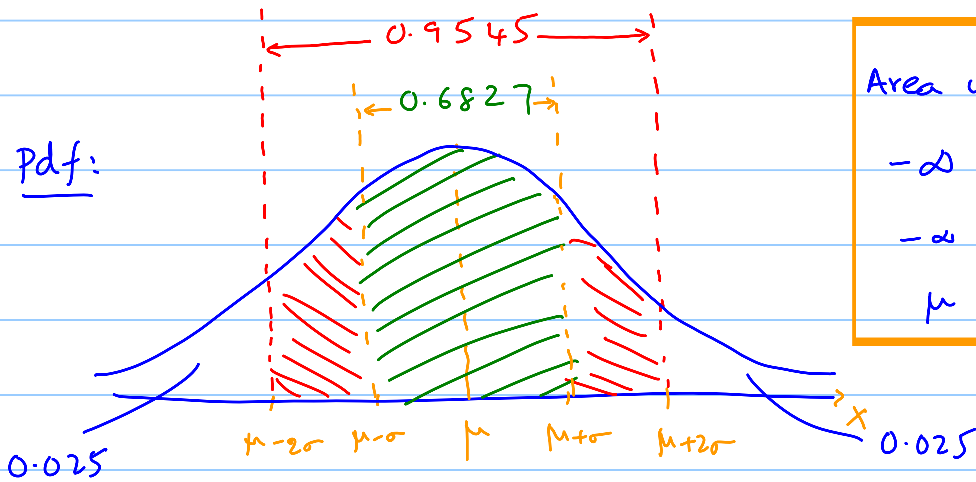
Normal Distribution

Let X be a normally distributed RV with mean μ and variance σ^2, then denote $X \sim N(\mu, \sigma^2)$.

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\left(\frac{x-\mu}{\sigma}\right)^2/2}, \quad -\infty < x < \infty$$

pdf and CDF are too complex

pdf:



Area under the curve		
$-\infty$ to $+\infty$		1.0
$-\infty$ to μ		0.5
μ to $+\infty$		0.5

Since CDF has no closed form solution, pre-calculated tables are used to estimate probabilities.

Standard Normal Distribution

A standard (or unit) normal random variable has mean 0 and variance 1.

The standard normal random variable is commonly denoted as Z. That is, Z ~ Normal (0,1).

Given a normal RV X, we can construct the corresponding standard normal RV Z using the following

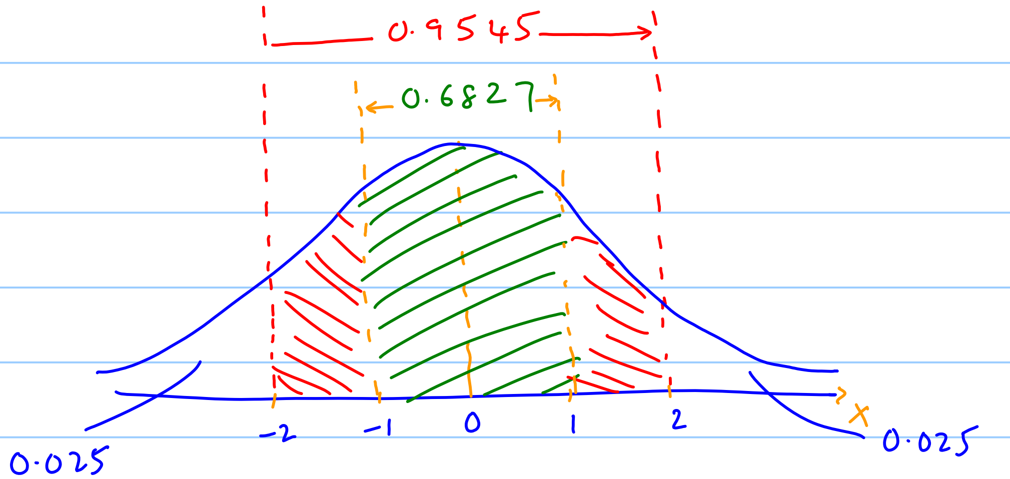
If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma}$.

Z shifts the midpoint of pdf of X to '0' and scales the x-axis by a factor of $1/\sigma$.

CDF of Z is often denoted as $\Phi(\ )$.

Tables are used estimate CDF of Z.

Pdf of z:



$$X \sim N(\mu, \sigma^2), \quad Z = \frac{X-\mu}{\sigma} \Rightarrow \underline{X = \mu + Z\sigma}$$

$$F(a) = P(X \le a) = P\left(Z \le \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right)$$

$$\Phi(b) = P(Z \le b) = P(X \le \mu + b\sigma)$$

$$\Phi(-c) = 1 - \Phi(c) \qquad \text{for any } c \geq 0$$

The table provides $P(0 \le Z \le c)$ for $c \geq 0$.

$$\Phi(c) = P(Z \le c) = \underbrace{P(Z \le 0)} + P(0 \le Z \le c)$$

$$= 0.5 + \text{Table value for } C$$

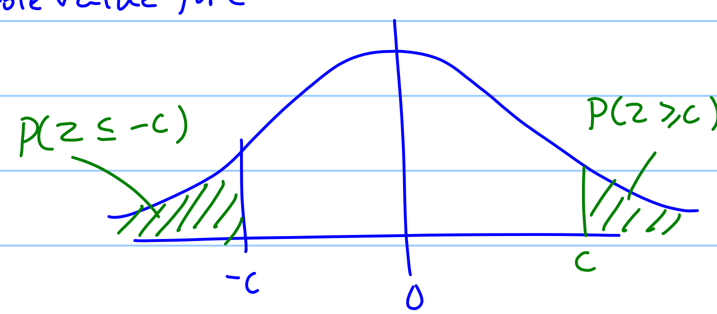
$$P(-c \le Z \le c) = 2 * P(0 \le Z \le c)$$

$$= 2 * \text{Table value for } c$$

$P(Z \le -c)$

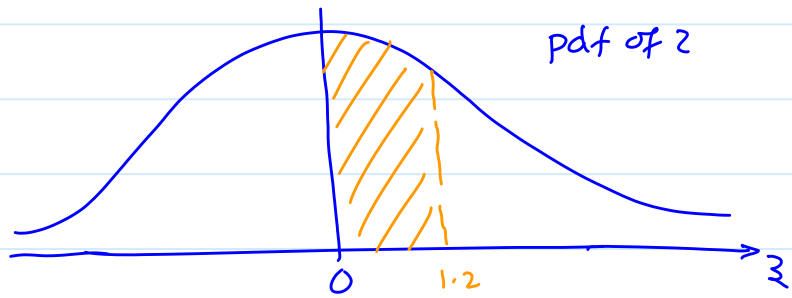
$= \Phi(-c) = 1 - \Phi(c)$

$= 0.5 - P(0 \le Z \le c)$



4.12, S3
Find the area under the standard normal curve for various values of z.

(a) 0<= Z <= 1.2

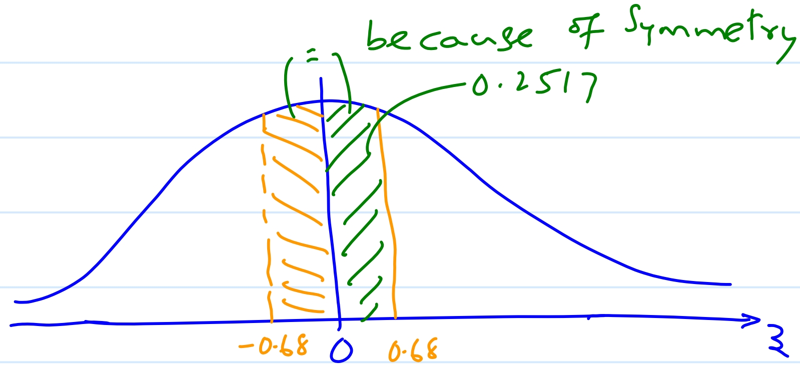


From the Table, $P(0 \le z \le 1.2) = 0.3849$

(b) $P(-0.68 \le z \le 0)$

$= P(0 \le z \le 0.68)$

$= 0.2517$



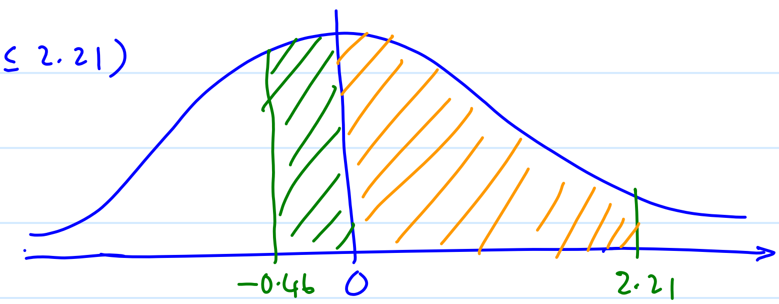
(c) $P(-0.46 \le z \le 2.21)$

$= P(-0.46 \le z \le 0) + P(0 \le z \le 2.21)$

$= P(-0.46 \le z \le 0) + 0.4864$

$= P(0 \le z \le 0.46) + 0.4864$

$= 0.1772 + 0.4864 = 0.6636$



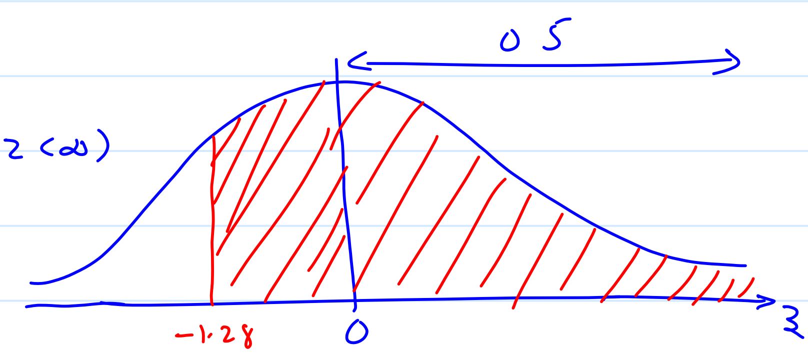
(e) $P(z \geq -1.28)$

$= P(-1.28 \le z \le 0) + P(0 \le z < \infty)$

$= \underbrace{P(-1.28 \le z \le 0)} + 0.5$

$= P(0 \le z \le 1.28) + 0.5$

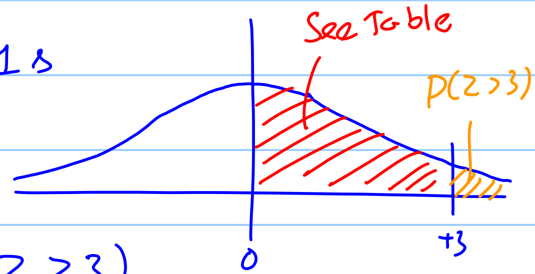
$= 0.3997 + 0.5 = 0.8997$



Problem. The execution times of queries on a database are normally distributed with a mean of 6 seconds and standard deviation of 1s. Let X be The RV denoting the execution times.

(a) $P(X > 9s)$ $\quad \mu = 6s \quad \sigma = 1s$

$$Z = \frac{X-\mu}{\sigma} = \frac{X-6}{1} = X-6$$



$$P(X > 9s) = P(Z > 9-6) = P(Z > 3)$$

$$\Phi(a) = P(Z \le a)$$

$$P(Z > 3) = 1 - P(Z \le 3)$$

$$P(Z \le 3) = \underbrace{P(Z \le 0)} + P(0 \le Z \le 3)$$

$= 0.5 +$ Table value for 3

$$= 0.5 + 0.4987 = 0.9987$$

$$P(Z > 3) = 1 - 0.9987 = 0.0013 = P(X > 9s)$$

(b) $P(X < 7s) = P\left(Z < \frac{7-6}{1}\right) = P(Z < 1)$

$$= 0.5 + P(0 \le Z \le 1)$$

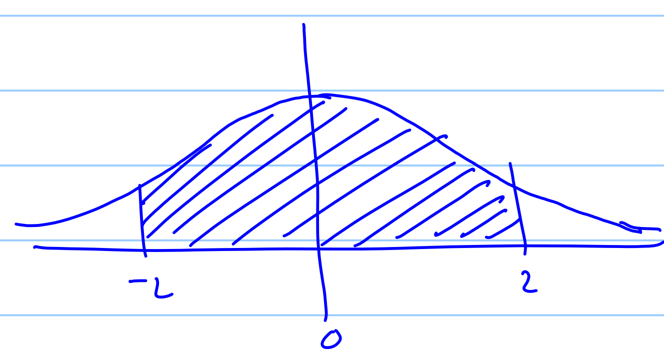
$$= 0.5 + 0.3413 = 0.8413$$

(c) $P(4 < X < 8) = P\left(\frac{4-6}{1} < Z < \frac{8-6}{1}\right) = P(-2 < Z < 2)$

$$= 2 * P(0 \le Z \le 2)$$

$$= 2 * 0.4772$$

$$= 0.9544$$



(d) what is the $\underbrace{\text{90th percentile execution time}}$?

90% of all queries take less time than this

find the value 'a' such that

$\underbrace{P(z \le a)}_{\downarrow} = 0.9$

$0.5 + P(0 \le z \le a) = 0.9$

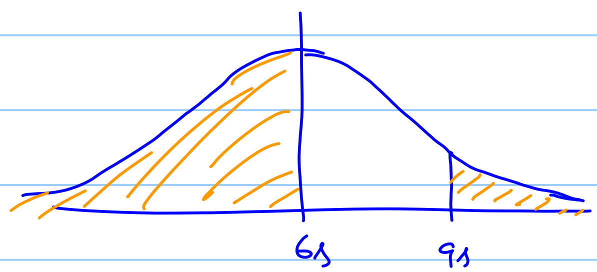
$\Rightarrow \underbrace{P(0 \le z \le a)}_{\text{use the Table}} = 0.4$ $\therefore$ $a = 1.28$

$P(z < 1.28) = P(x < \mu + 1.28\sigma)$

$= P(x < 6 + 1.28 \times 1) = P(x < 7.28)$.

$\therefore$ 90% of all queries take 7.28 s or less.

(c') What is the prob. that a query takes less than 6 seconds or more than 9 seconds.



$P(x < 6) + P(x > 9s)$

$= 1 - P(6 < x < 9)$

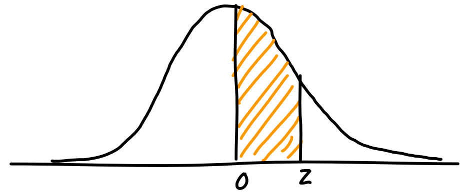
$= 1 - P\left(\frac{6-6}{1} < z < \frac{9-6}{1}\right)$

$= 1 - P(0 < z < 3) = 1 - \text{Table}(3)$

$= 1 - 0.4997 = 0.5003$.

Standard Normal RV Distribution Function Table

Area between 0 and z



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

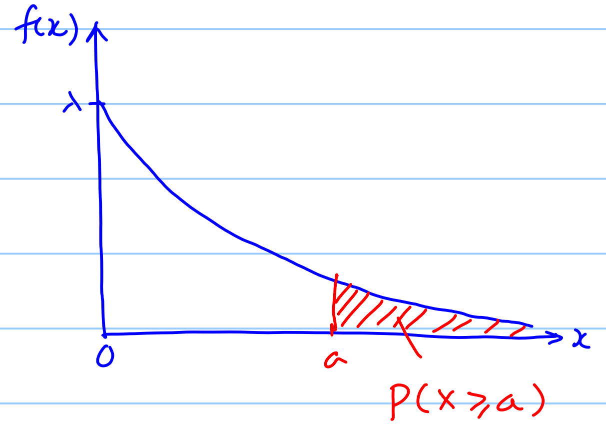
Not done in class.

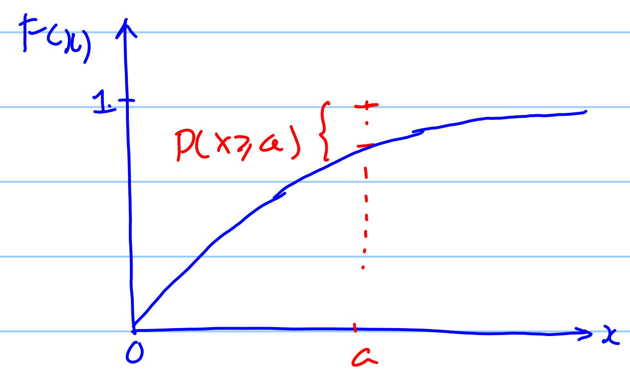
Exponential Distribution

$X \sim Exp(\lambda) \Rightarrow X$ is an exponential RV with parameter λ

pdf: $f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$

CDF: $F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$





If events occur such that the time elapsed between consecutive events is exponentially distributed, then the number of events is Poisson distributed.

Example:
A web server receives requests at an average rate of l= 0.1 jobs/sec. The number of arrivals is Poisson distributed with parameter Nt. What is the prob. that the time elapsed between two consecutive job arrivals is at least 10 seconds.

Let X denote the interarrival time of jobs.

$\therefore X \sim Exp(\lambda)$, $\underline{\lambda = 0.1}$ jobs/sec.

$$P(X \geq 10) = 1 - \underbrace{P(X \leq 10)}_{F(10)} = 1 - \left[1 - e^{-\lambda \cdot 10}\right] = e^{-\lambda \cdot 10} = e^{-1} = 0.368$$