Session 10: Probability Distributions

Module BUSN9690

Business Statistics with Python

Outline: Probability Distributions

- Random Variables
- Expected Value and Variance
- Binomial Distribution
- Poisson Distribution
- Normal Distribution
- Exponential Distribution

Example, Random Variables

Question	Random Variable X	Type
Family size	X = Number of dependents reported on tax return	Discrete
Distance from home to store	X = Distance in miles from home to the store site	Continuous
Own dogs or cats	X = 1 if own no pet; = 2 if own dog(s) only; = 3 if own cat(s) only; = 4 if own dog(s) and cat(s)	Discrete

- A random variable is a quantitative variable whose value depends on chance
- A discrete random variable assumes a *countable* number of values, which can be finite or infinite
- A continuous random variable takes an infinite number of possible values

The probability distribution for a random variable describes how probabilities are distributed over the values of the random variable.

Discrete Probability Distributions

Properties for Discrete Probability Distributions

The probability distribution is defined by a <u>probability function</u>, denoted by $P(X{=}x)$, which provides the probability for each value of the random variable

The required conditions for a discrete probability function are

- $P(X = x) \geq 0;$
- $\sum_{k=1}^{n} P(X = x_k) = 1$

Example 1

Let X =the number of TVs sold to the first 4 customers who enter the store, where X can take on 5 values (0, 1, 2, 3, 4)

x	$P(X{=}x)$
0	.40
1	.25
2	.20
3	.05
4	.10

Expected Value

Mean of a Discrete Random Variable

The **mean of a discrete random variable** X is denoted μ. It is defined by

$$\mu = \sum_{k=1}^{n} x_k P(X = x_k)$$

The terms **expected value** and **expectation** are commonly used in place of the term mean.

x	$P(X=x)$	$xP(X=x)$
0	.40	.00
1	.25	.25
2	.20	.40
3	.05	.15
4	.10	.40
	$E(X) = \mu =$	1.20

expected number of TVs sold in a day

Variance and Standard Deviation

Variance of a **Discrete Random Variable**

The **variance** summarizes the variability in the values of a random variable and is defined as

$$\mathrm{Var}(X) = \sigma^2 = \sum_{k=1}^{n} (x_k - \mu)^2 P(X = x_k)$$

The **standard deviation** of a discrete random variable is

$$\sigma = \sqrt{\sum_{k=1}^{n} (x_k - \mu)^2 P(X = x_k)}$$

Remember: μ = 1.20 in the TV example

x	$x-\mu$	$(x-\mu)^2$	$P(X=x)$	$(x-\mu)^2$ $P(X=x)$
0	-1.2	1.44	.40	.576
1	-0.2	0.04	.25	.01
2	0.8	0.64	.20	.128
3	1.8	3.24	.05	.162
4	2.8	7.84	.10	.784

$Var(X) = \sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 P(X = x_i) = 1.66$

Variance of TVs sold in a day

Standard deviation σ of daily sales = 1.2884 TVs

Roughly speaking, on average, the number of TV's sold to the first 4 customers is 1.2884 from the mean of 1.2 TV's

The Binomial Distribution

Bernoulli Trials

- A Bernoulli trial is a random experiment in which there are only two possible outcomes - *success* and *failure*.

- **Examples**
  - Tossing a coin and considering heads as success and tails as failure.
  - Checking items from a production line: success = not defective, failure = defective.
  - Phoning a call centre: success = operator free; failure = no operator free.

Binomial Distribution

- **Four Properties of a Binomial Experiment**

1. The experiment consists of a sequence of n identical trials.
2. Two outcomes, success and failure, are possible on each trial.
3. The probability of a success, denoted by p, does not change from trial to trial.
4. The trials are independent.

The Binomial Distribution

Our interest is in the number of successes occurring in the n trials and the probability that a given number of successes occurs.

We let X denote the number of successes occurring in the n trials.

- A Bernoulli random variable X takes the values 0 and 1 and

  $P(X = 1) = p$

  $P(X = 0) = 1 - p$

Recall: combinations

Find the number of different subsets of size 3 in the set $\{m, a, t, h, r\}$.

Solution

A subset of size 3 must have 3 distinct elements, so repetitions are not allowed. Order is not important.

$$C_5^3 = \frac{5!}{3!\,(5-3)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1) \times (2 \times 1)} = 10$$

mat, mah, mar, mth, mtr, mhr, ath, atr, ahr, thr

$5! = 5 \times 4 \times 3 \times 2 \times 1$

Binomial Distribution, expected value, and variance

- Binomial Probability Function

$$P(X=x)=\frac{n!}{x!(n-x)!}p^x(1-p)^{(n-x)}$$

where:

$P(X=x)$ = the probability of x successes in n trials

n = the number of trials

p = the probability of success on any one trial

- The expected value

$$E(X)=np$$

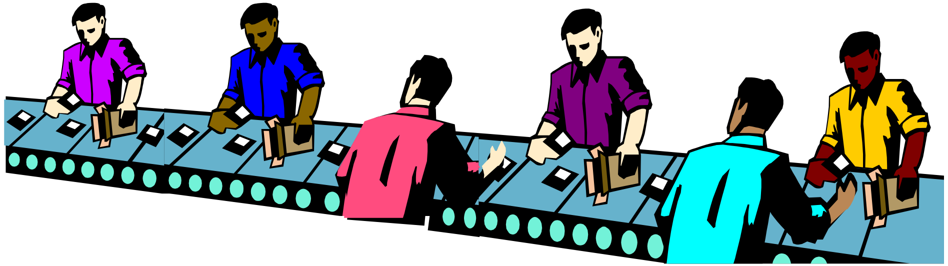
- The variance

$$Var(X)=np(1-p)$$

Example

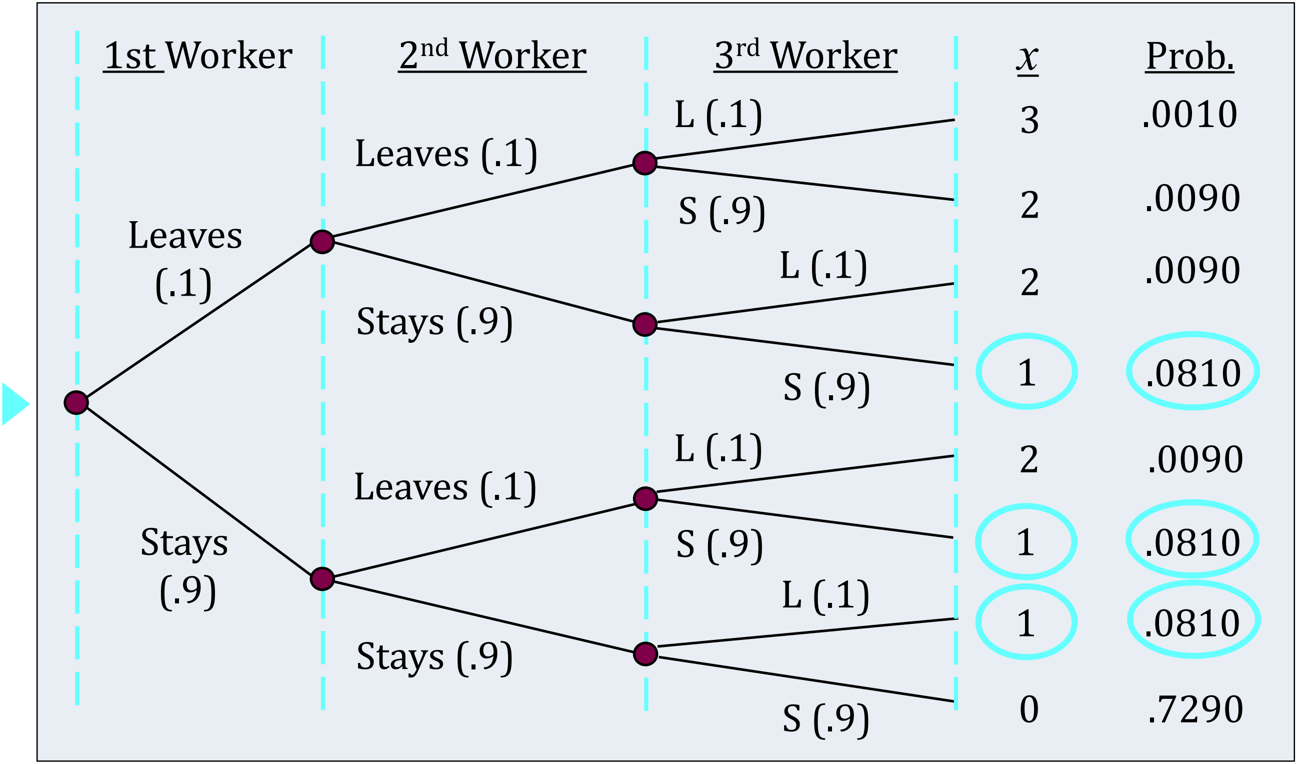
- Example: Bob Electronics

  Bob is concerned about a low retention rate for employees. In recent years, management has seen a turnover of 10% of the hourly employees annually. Thus, for any hourly employee chosen at random, management estimates a probability of 0.1 that the person will not be with the company next year.



Binomial Distribution

- Tree Diagram



Binomial Distribution

- **Using the Binomial Probability Function**

  Choosing 3 hourly employees at random, what is the probability that 1 of them will leave the company this year?

$$\text{Let: } p = .10,\ n = 3,\ x = 1$$

$$P(X = 1) = \frac{3!}{1!(3-1)!} \times 0.1^1 \times 0.9^2 = 3 \times 0.1 \times 0.81 = 0.243$$

- **Expected Value**

$$E(X) = \mu = 3 \times 0.1 = .3 \text{ employees out of } 3$$

- **Variance**

$$\mathrm{Var}(X) = \sigma^2 = 3 \times 0.1 \times 0.9 = .27$$

The Poisson distribution

Poisson Distribution

A Poisson distributed random variable is often useful in estimating the number of events over a <u>specified interval of time or space</u>

It is a discrete random variable that may assume an <u>infinite sequence of values</u> (X= 0, 1, 2, . . .).

Poisson distributions: Examples

- In many practical situations we are interested in measuring how many times a certain event occurs in a specific time interval or in a specific length or area. For instance:
  1) the number of phone calls received at an exchange or call centre in an hour;
  2) the number of customers arriving at a toll booth per day;
  3) the number of flaws on a length of cable;
  4) the number of cars passing using a stretch of road during a day.

- Two Properties of a Poisson Experiment

1. The probability of an occurrence is the same for any two intervals of equal length.

2. The occurrence or nonoccurrence in any interval is independent of the occurrence or nonoccurrence in any other interval.

Poisson Distribution

- Poisson Probability Function

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

where:

$P(X=x)$ = probability of x events in an interval

λ = mean value of events in an interval

e = 2.71828

A property of the Poisson distribution is that the mean and variance are equal.

$\mu = \sigma^2$

Example: Poisson Distribution

- Example: Misericord Hospital

Patients arrive at the emergency room of Misericord Hospital at the average rate of 6 per hour on weekend evenings.

What is the probability of 4 arrivals in 30 minutes on a weekend evening?

Poisson Distribution

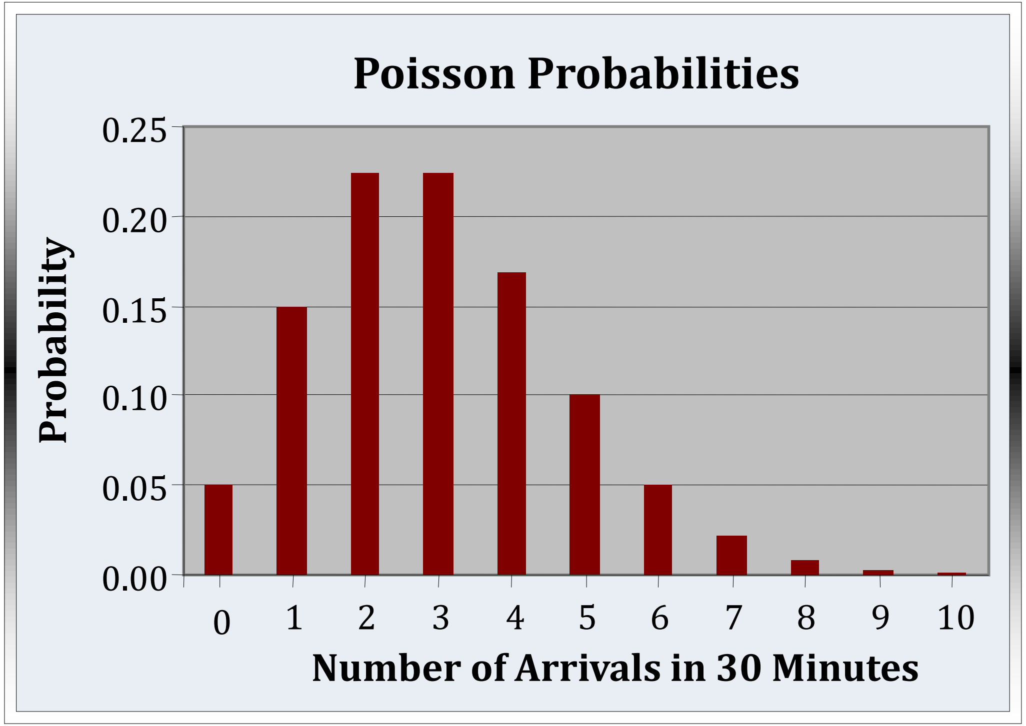
- Using the Poisson Probability Function

$\lambda = 6/\text{hour} = 3/\text{half-hour},\ X = 4$

$$P(X=4) = \frac{3^4 \times 2.71828^{-3}}{4!} = 0.168$$

- Variance for Number of Arrivals During 30-Minute Periods

$$\mu = \sigma^2 = 3$$



Continuous Probability Distributions

Normal Probability Distribution

Example 1: Heights of Female College Students

A midwestern college has an enrollment of 3264 female students. Frequency and relative-frequency distributions for these heights appear in the left Table.

Q1: What is the relative frequency that the students are between 60 and 68 inches tall?

Q2: What is the relative frequency that the students are shorter than 68 inches?

Q3: What is the relative frequency that the students are taller than 68 inches?

Height	Frequency	Relative frequency
56--under_57	3	0.0009
57--under_58	6	0.0018
58--under_59	26	0.008
59--under_60	74	0.0227
60--under_61	147	0.045
61--under_62	247	0.0757
62--under_63	382	0.117
63--under_64	483	0.148
64--under_65	559	0.1713
65--under_66	514	0.1575
66--under_67	359	0.11
67--under_68	240	0.0735
68--under_69	122	0.0374
69--under_70	65	0.0199
70--under_71	24	0.0074
71--under_72	7	0.0021
72--under_73	5	0.0015
73--under_74	1	0.0003
0	3264	1

Question 1

Let X=height

Q1: What is the relative frequency that the students are between 60 and 68 inches tall?

The question asks $P(60 \leq X < 68)$?

A: The relative frequency of heights from 60 to 68 is **0.898**

Height	Frequency	Relative frequency
56--under_57	3	0.0009
57--under_58	6	0.0018
58--under_59	26	0.008
59--under_60	74	0.0227
60--under_61	147	[illegible]
61--under_62	247	[illegible]
62--under_63	382	[illegible]
63--under_64	483	[illegible]
64--under_65	559	[illegible]
65--under_66	514	[illegible]
66--under_67	359	[illegible]
67--under_68	240	[illegible]
68--under_69	122	0.0374
69--under_70	65	0.0199
70--under_71	24	0.0074
71--under_72	7	0.0021
72--under_73	5	0.0015
73--under_74	1	0.0003
0	3264	1

Question 2

Q2: What is the relative frequency that the students are shorter than 68 inches?

The question asks $P(X < 68)$?

A1: The relative frequency of heights smaller than 68

=0.9314

A2: 1- the relative frequency of heights taller than 68

=1-**0.0686**

=0.9314

Height	Frequency	Relative frequency
56--under_57	3	
57--under_58	6	
58--under_59	26	
59--under_60	74	
60--under_61	147	
61--under_62	247	
62--under_63	382	
63--under_64	483	
64--under_65	559	
65--under_66	514	
66--under_67	359	
67--under_68	240	
68--under_69	122	0.0374
69--under_70	65	0.0199
70--under_71	24	0.0074
71--under_72	7	0.0021
72--under_73	5	0.0015
73--under_74	1	0.0003
0	3264	1

Question 3

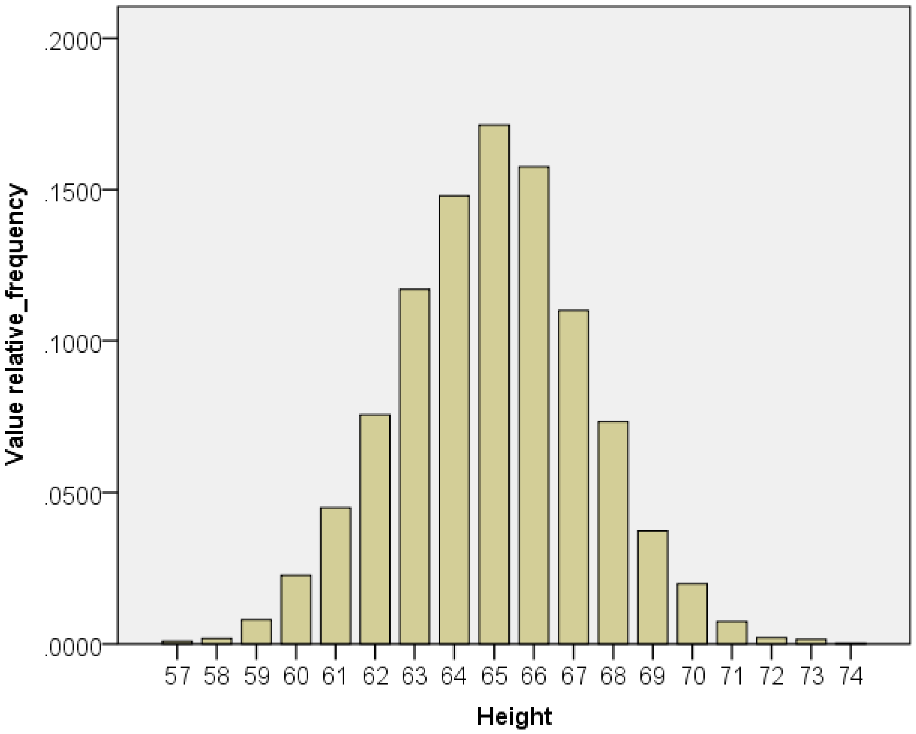
Q3: What is the relative frequency that the students are taller than 68 inches?

The question asks $P(X \geq 68)$?

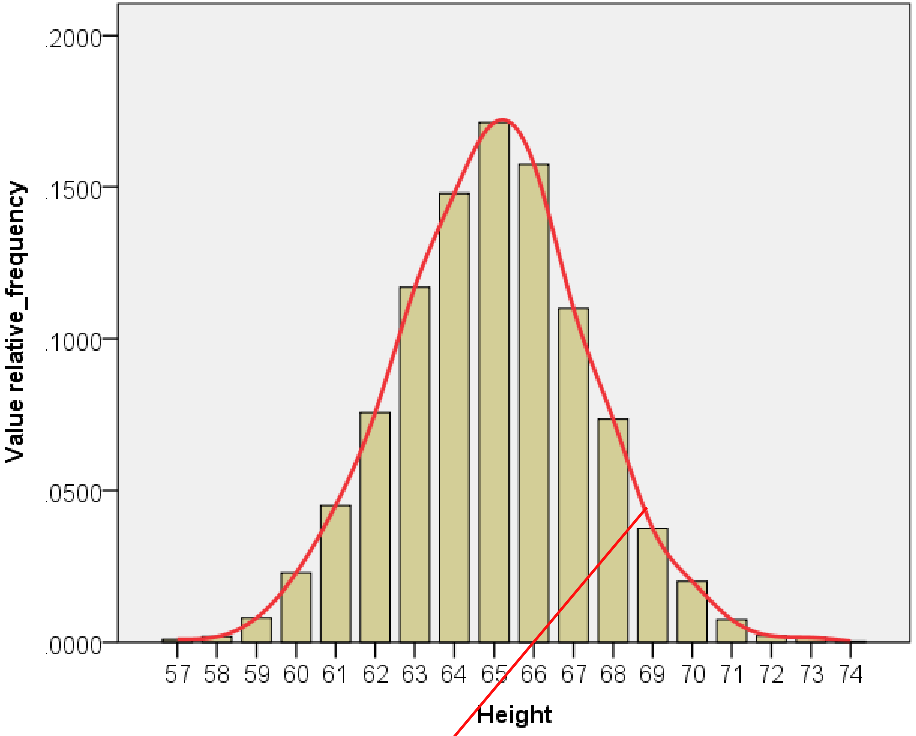
A: 0.0686

Height	Frequency	Relative frequency
56--under_57	3	0.0009
57--under_58	6	0.0018
58--under_59	26	0.008
59--under_60	74	0.0227
60--under_61	147	0.045
61--under_62	247	0.0757
62--under_63	382	0.117
63--under_64	483	0.148
64--under_65	559	0.1713
65--under_66	514	0.1575
66--under_67	359	0.11
67--under_68	240	0.0735
68--under_69	122	[illegible]
69--under_70	65	[illegible]
70--under_71	24	[illegible]
71--under_72	7	[illegible]
72--under_73	5	[illegible]
73--under_74	1	[illegible]
0	3264	1

Histogram


.2000
.1500
.1000
.0500
.0000
Value relative_frequency
57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74
Height

Histogram---probability density function



probability density function

Cumulative relative frequency

Height	frequency	Relative frequency	Cumulative relative frequency
56--under_57	3	0.0009	0.0009
57--under_58	6	0.0018	0.0027
58--under_59	26	0.008	0.0107
59--under_60	74	0.0227	0.0334
60--under_61	147	0.045	0.0784
61--under_62	247	0.0757	0.1541
62--under_63	382	0.117	0.2711
63--under_64	483	0.148	0.4191
64--under_65	559	0.1713	0.5904
65--under_66	514	0.1575	0.7479
66--under_67	359	0.11	0.8579
67--under_68	240	0.0735	0.9314
68--under_69	122	0.0374	0.9688
69--under_70	65	0.0199	0.9887
70--under_71	24	0.0074	0.9961
71--under_72	7	0.0021	0.9982
72--under_73	5	0.0015	0.9997
73--under_74	1	0.0003	1

Q1: What is the relative frequency that the students are between 60 and 68 inches tall?

A: 0.9314 – 0.0334=0.898

Cumulative relative frequency

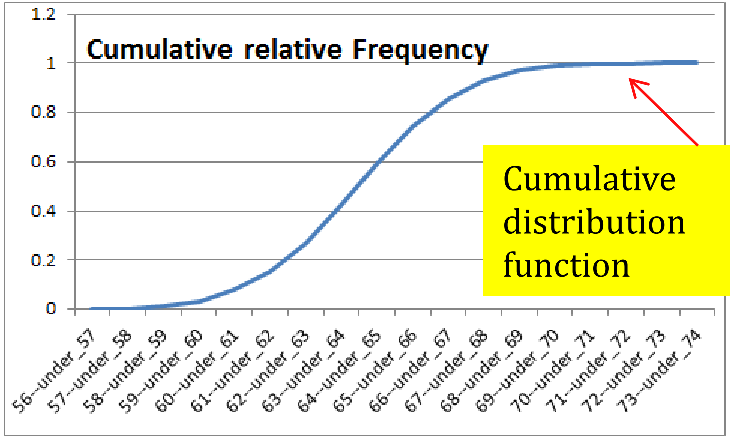
Height	frequency	Relative frequency	Cumulative Relative frequency
56--under_57	3	0.0009	0.0009
57--under_58	6	0.0018	0.0027
58--under_59	26	0.008	0.0107
59--under_60	74	0.0227	0.0334
60--under_61	147	0.045	0.0784
61--under_62	247	0.0757	0.1541
62--under_63	382	0.117	0.2711
63--under_64	483	0.148	0.4191
64--under_65	559	0.1713	0.5904
65--under_66	514	0.1575	0.7479
66--under_67	359	0.11	0.8579
67--under_68	240	0.0735	0.9314
68--under_69	122	0.0374	0.9688
69--under_70	65	0.0199	0.9887
70--under_71	24	0.0074	0.9961
71--under_72	7	0.0021	0.9982
72--under_73	5	0.0015	0.9997
73--under_74	1	0.0003	1
0	3264	1	

Q2: What is the relative frequency that the students are shorter than 68 inches?

A: 0.9314

Q3: What is the relative frequency that the students are taller than 68 inches?

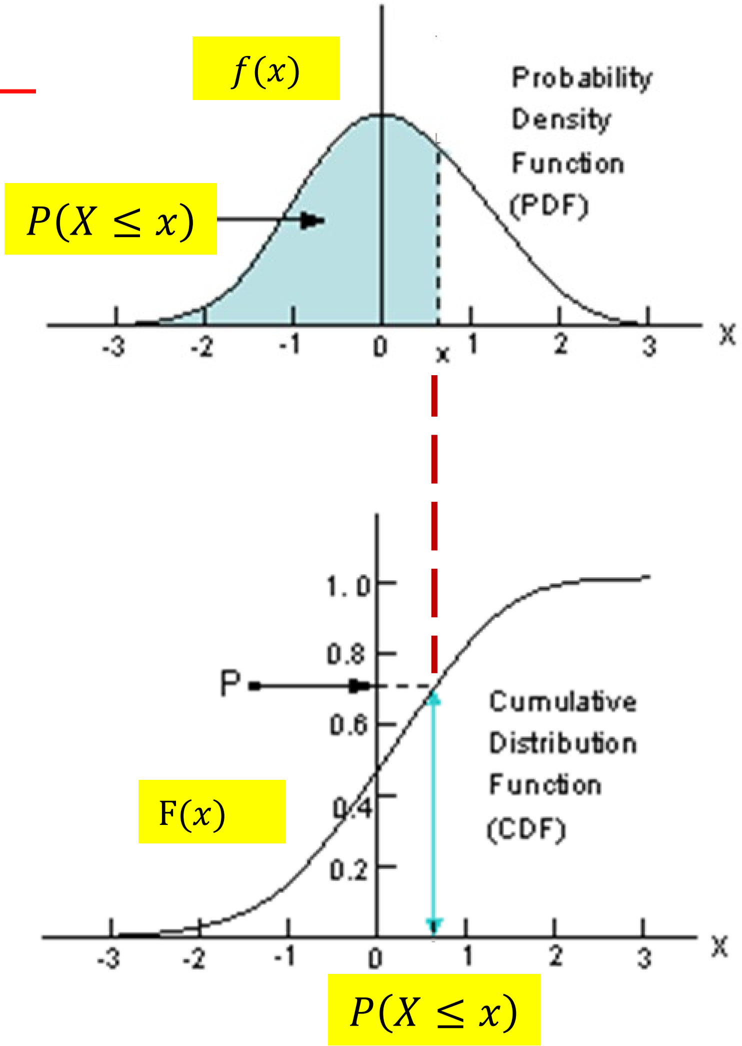
Cumulative relative frequency



1. **The cumulative relative frequency has an increasing trend**
2. **It eventually approaches to 1**

Height	frequency	Relative frequency	Cumulative relative frequency
56--under_57	3	0.0009	0.0009
57--under_58	6	0.0018	0.0027
58--under_59	26	0.008	0.0107
59--under_60	74	0.0227	0.0334
60--under_61	147	0.045	0.0784
61--under_62	247	0.0757	0.1541
62--under_63	382	0.117	0.2711
63--under_64	483	0.148	0.4191
64--under_65	559	0.1713	0.5904
65--under_66	514	0.1575	0.7479
66--under_67	359	0.11	0.8579
67--under_68	240	0.0735	0.9314
68--under_69	122	0.0374	0.9688
69--under_70	65	0.0199	0.9887
70--under_71	24	0.0074	0.9961
71--under_72	7	0.0021	0.9982
72--under_73	5	0.0015	0.9997
73--under_74	1	0.0003	1
0	3264	1	

PDF and CDF



- Probability density function (pdf): $f(x)$

$$f(x) \geq 0, \int_{-\infty}^{+\infty} f(x)\,dx = 1$$

- Cumulative distribution function (cdf), $F(x)$:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

$$f(x) = \frac{dF(x)}{dx}$$

Continuous Probability Distributions

- A <u>continuous random variable</u> can assume any value in an interval on the real line or in a collection of intervals.
- It is not possible to talk about the probability of the random variable assuming a particular value. That is:

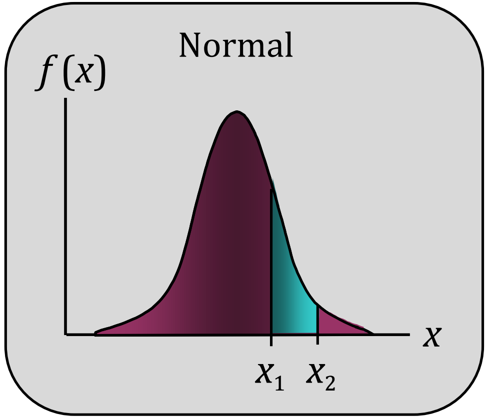
$$\boldsymbol{P(X = x) = 0}$$

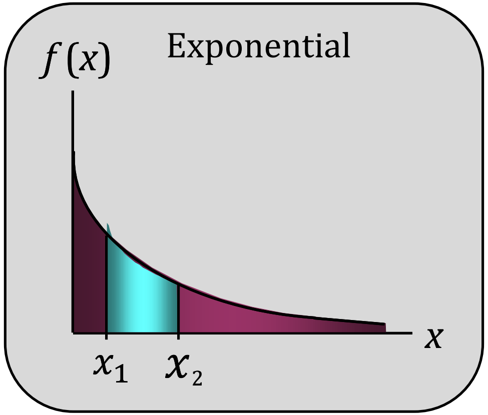
- Instead, we talk about the probability of the random variable assuming a value within a given interval.

  - Expected value
    - $E[X] = \int_{-\infty}^{+\infty} xf(x)dx$
  - Variance
    - $E[X] = \int_{-\infty}^{+\infty} x^2 f(x)dx - (E[X])^2$

Continuous Probability Distributions

- The probability of the random variable assuming a value within some given interval from x_1 to x_2 is defined to be the area under the graph of the probability density function between x_1 and x_2.





Normal Probability Distribution

- The <u>normal probability distribution</u> is an *extremely* important distribution for describing a continuous random variable.
  - It is WIDELY used in statistical inference!!!

- Normal Probability Density Function

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

where: μ = mean

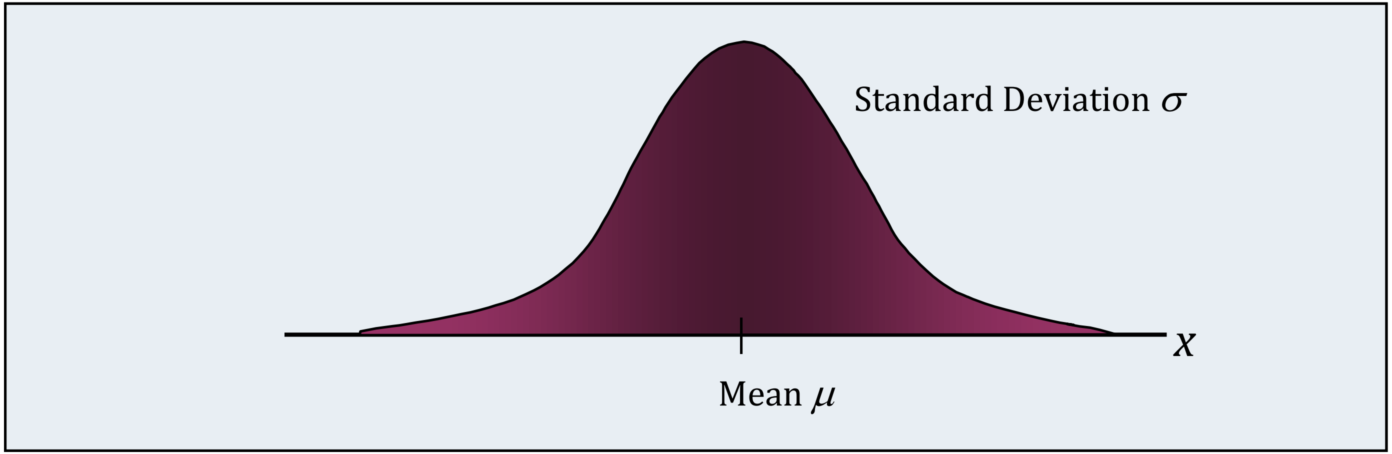
σ = standard deviation

π = 3.14159

e = 2.71828

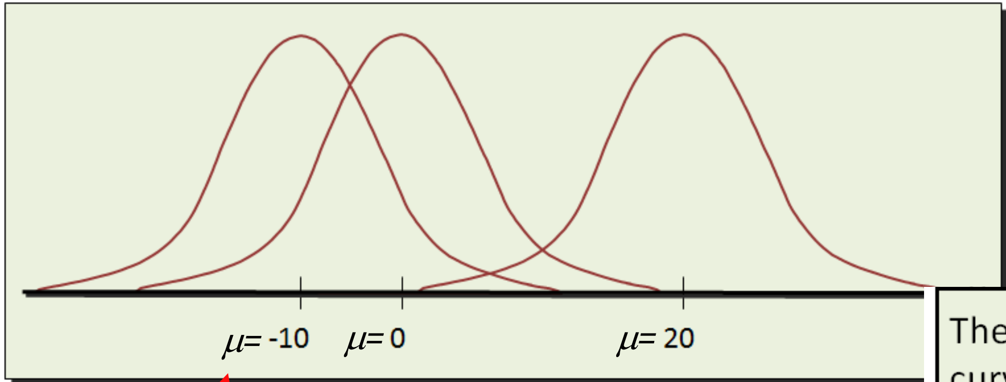
Normal Probability Distribution: Characteristics

- The distribution is symmetric
- The entire family of normal probability distributions is defined by its mean μ and its standard deviation σ.
- The highest point on the normal curve is at the mean, which is also the median and mode.



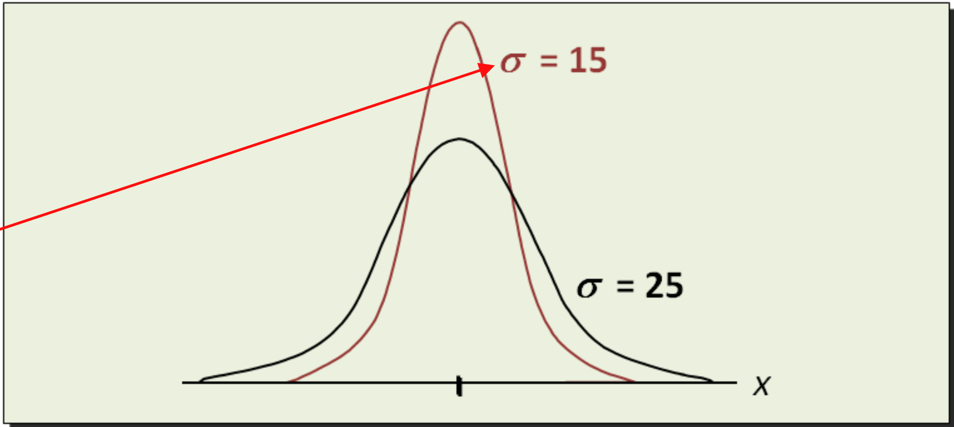
Normal Probability Distribution: Characteristics

The mean can be any numerical value: negative, zero, or positive.



Location parameter

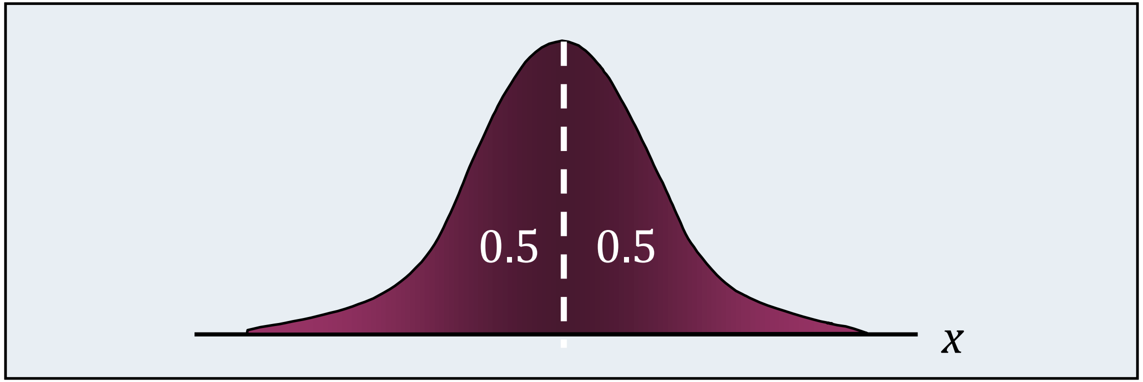
The standard deviation determines the width of the curve: larger values result in wider, flatter curves.



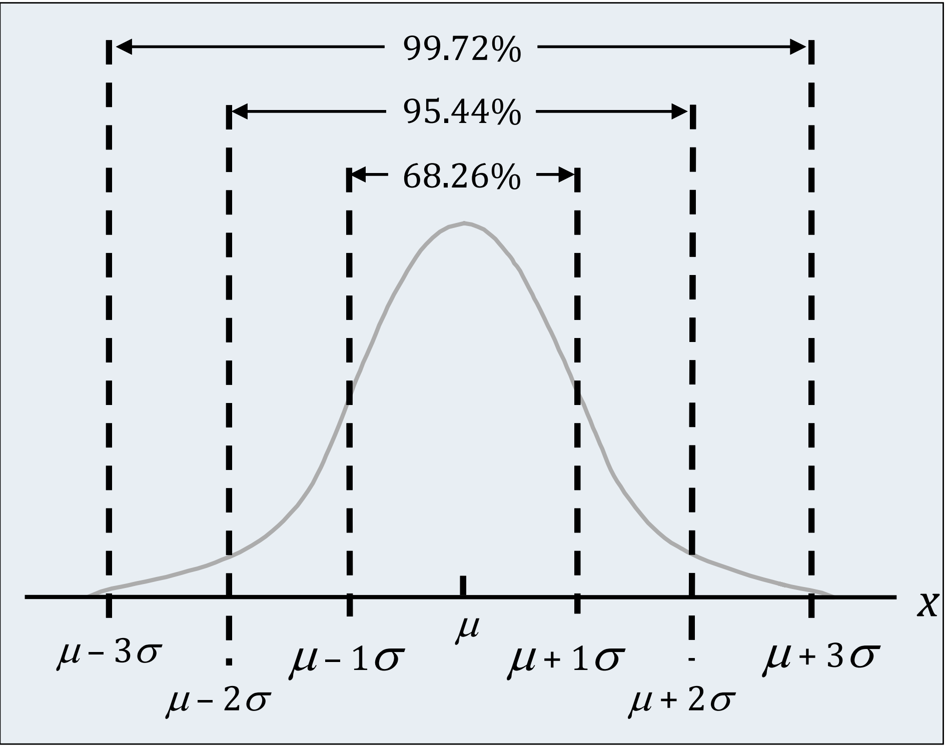
Scale parameter

Normal Probability Distribution: Characteristics

Probabilities for the normal random variable are given by <u>areas under the curve</u>. The total area under the curve is 1 (0.5 to the left of the mean and 0.5 to the right).



Normal Probability Distribution: Characteristics



68.26% of values of a normal random variable are within +/- 1 standard deviation of its mean.

95.44% of values of a normal random variable are within +/- 2 standard deviation of its mean.

99.72% of values of a normal random variable are within +/- 3 standard deviation of its mean.

Normal Probability Distribution

- Example: Pep Zone

Pep Zone sells auto parts and supplies including a popular multi-grade motor oil. When the stock of this oil drops to 80 litres, a replenishment order is placed. The store manager is concerned that sales are being lost due to stockouts while waiting for an order. It has been determined that demand during replenishment lead-time is normally distributed with a mean of 60 litres and a standard deviation of 24 litres.

The manager would like to know the probability of a stockout, $P(X > 80)$.

Normal Probability Distribution

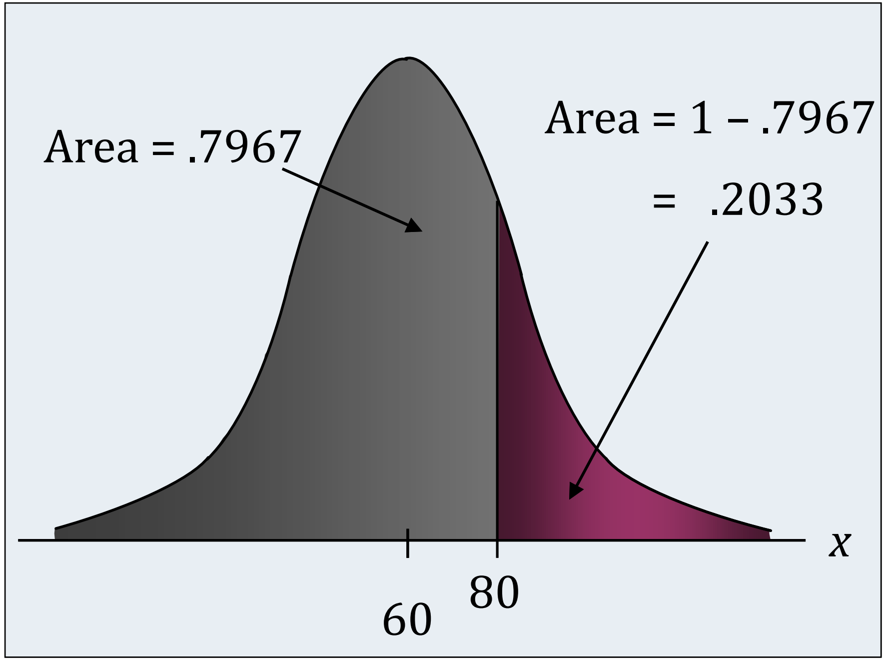
Solving for the Stockout Probability

Compute the area under the normal curve to the right of X= 80, that is, $P(X < 80)$.

$$P(X > 80) = 1 - P(X \leq 80)$$
$$= 1 - 0.7967$$
$$= \ .2033$$

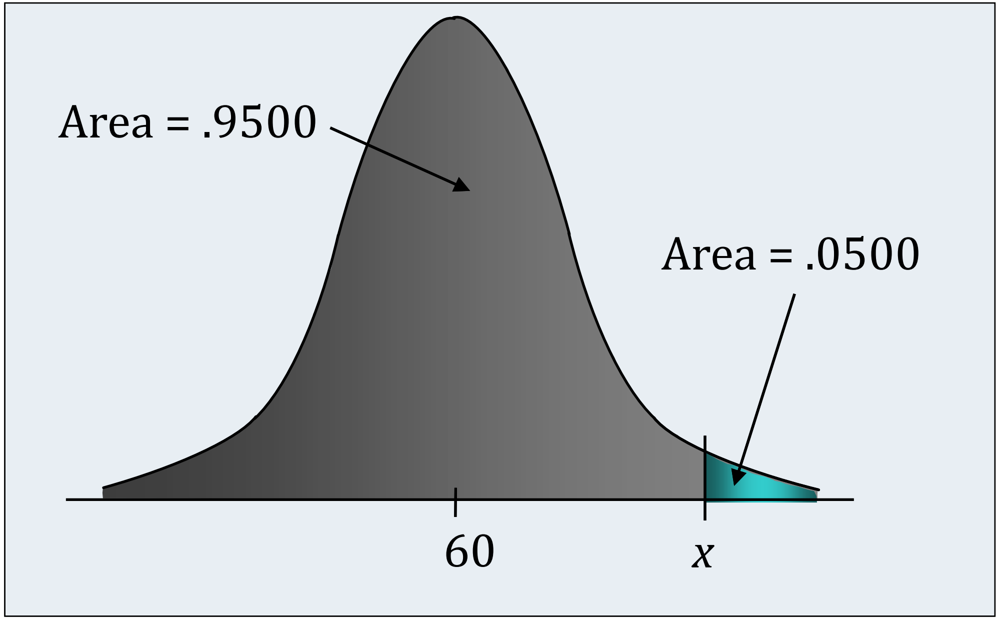
Probability of a stockout

$P(x > 80)$



If the manager of Pep Zone wants the probability of a stockout to be no more than 0.05, what should the reorder point be?

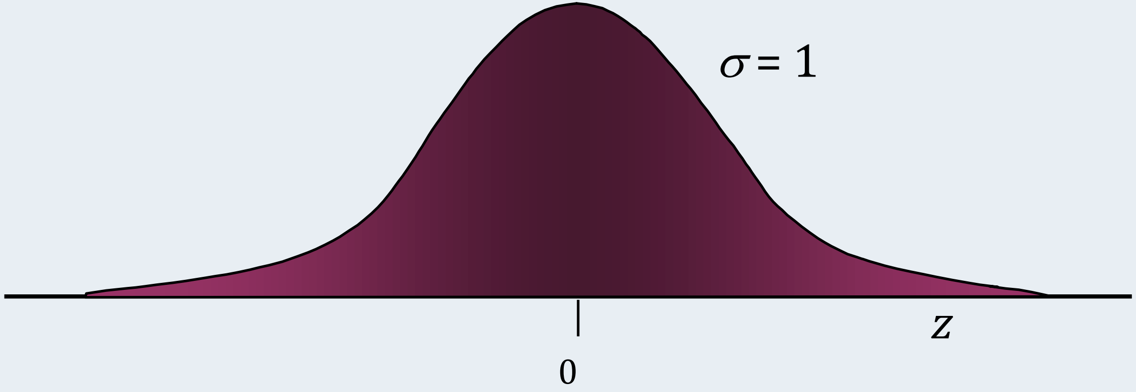
- Solving for the Reorder Point



The question asks you to give x, which satisfies $P(X > x) = 0.05$

A reorder point of 99.48 litres will place the probability of a stockout during lead time at (slightly less than) .05.

A random variable having a normal distribution with a mean of 0 and a standard deviation of 1 is said to have a standard normal probability distribution.



Standard Normal Probability Distribution

- The letter z is used to designate the standard normal random variable.
- Converting to the Standard Normal Distribution

$$\boxed{z = \frac{x - \mu}{\sigma}}$$

- We can think of z as a measure of the number of standard deviations x is from μ.

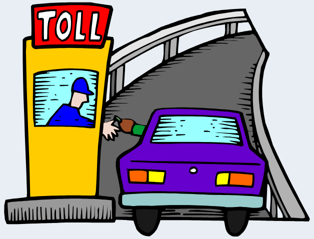
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \Rightarrow \quad \boxed{f(x) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}}$$

Exponential Probability Distribution

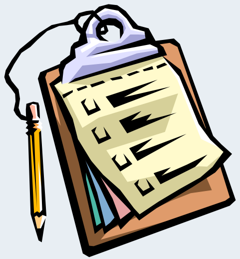
Exponential Probability Distribution

- The exponential probability distribution is useful in describing the time it takes to complete a task.
- The exponential random variables can be used to describe:

Time between vehicle arrivals at a toll booth



Time required to complete a questionnaire



Distance between major defects in a highway

Exponential Probability Distribution

- Probability Density Function

$$f(x) = \frac{1}{\mu} e^{-x/\mu} \quad \text{for } x \geq 0,\ \mu > 0$$

where: μ = mean ← Scale parameter

e = 2.71828

- Cumulative distribution function

$$F(x) = 1 - e^{-x/\mu}$$

A property of the exponential distribution is that the mean, μ, and standard deviation, σ, are equal.

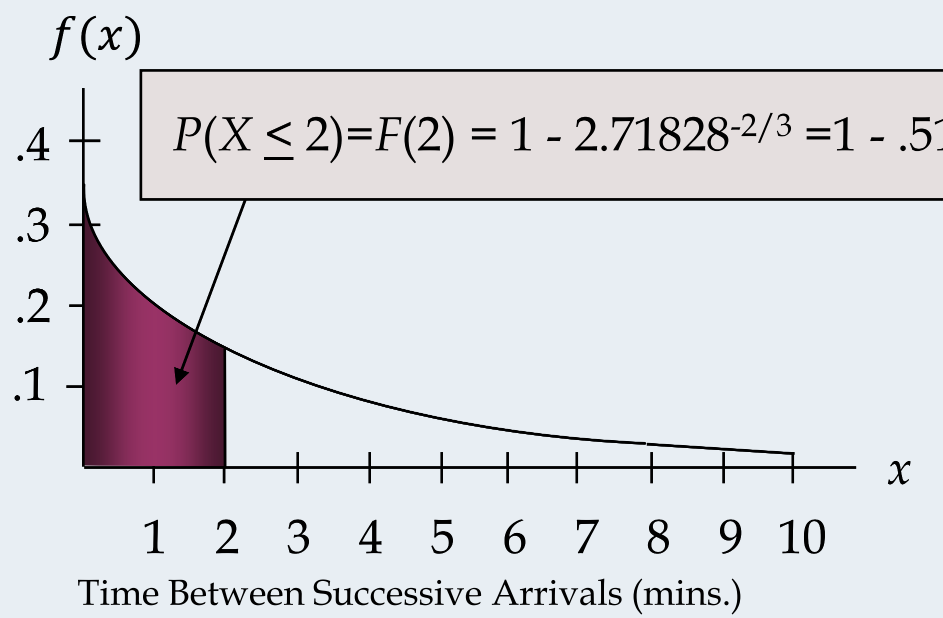
Exponential Probability Distribution

- Example: Al's Full-Service Pump

The time between arrivals of cars at Al's full-service petrol pump follows an exponential probability distribution with a mean time between arrivals of 3 minutes. Al would like to know the probability that the time between two successive arrivals will be 2 minutes or less.

Exponential Probability Distribution

$$P(X \leq 2) = F(2) = 1 - 2.71828^{-2/3} = 1 - .5134 = .4866$$



Thus, the standard deviation, σ, and variance, σ^2, for the time between arrivals at Al's full-service pump are:

$\sigma = \mu = 3$ minutes

$\sigma^2 = (3)^2 = 9$

Relationship between the Poisson and Exponential Distributions

The Poisson distribution provides an appropriate description of the number of events per interval

The exponential distribution provides an appropriate description of the length of the interval between events

Summary

- Expected value, variance
- Binomial Distribution
- Poisson Distribution
- Normal Probability Distribution
- Exponential Probability Distribution