Common Probability Distributions

林 建 甫
C.F. Jeff Lin, MD. PhD.
台北大學統計系助理教授

Probability Distribution

Definition: A **random variable** is a characteristic whose obtained values arise as a result of chance factors.

Definition: A **probability distribution** gives the probability of obtaining all possible (sets of) values of a random variable. It gives the probability of the outcomes of an experiment. Note that a probability distribution is an example of the "classical" definition of probability. Think of it as a histogram of the entire population.

Population ⟷ Sample

Random variable ⟷ Measurement

Probability dist. ⟷ Frequency dist.

Parameters ⟷ Statistics (Estimates)

Probability Distributions

- norm, binom, beta, cauchy, chisq, exp, f, gamma, geom, hyper, lnorm, logis, nbinom, t, unif, weibull, wilcox
- Four prefixes:
  - 'd' for density (PDF)
  - 'p' for distribution (CDF)
  - 'q' for quantile (percentiles)
  - 'r' for random generation (simulation)
- Each distribution has arguments that need to be specified

Probability Distributions

- Cumulative distribution function $P(X \le x)$: 'p' for the CDF
- Probability density function: 'd' for the density,,
- Quantile function (given q, the smallest x such that $P(X \le x) > q$): 'q' for the quantile
- simulate from the distribution: 'r'

Probability Distributions

Distribution	R name	additional arguments
beta	beta	shape1, shape2, ncp
binomial	binom	size, prob
Cauchy	cauchy	location, scale
chi-squared	chisq	df, ncp
exponential	exp	rate
F	f	df1, df1, ncp
gamma	gamma	shape, scale
geometric	geom	prob
hypergeometric	hyper	m, n, k
log-normal	lnorm	meanlog, sdlog

logistic logis; negative binomial nbinom; normal norm; Poisson pois; Student's t t ; uniform unif; Weibull weibull; Wilcoxon wilcox

Common Discrete Distributions

- Often, the observations generated by different statistical experiments have the same general type of behavior.
- Discrete random variables associated with these experiments can be described by the same probability distribution. They will share a single formula.
- A hand full of distributions describe many of the random phenomena encountered in real life.

Some Common Discrete Distributions

- Discrete Uniform
- Bernoulli
- Binomial
- Geometric
- Hypergeometric
- Poisson

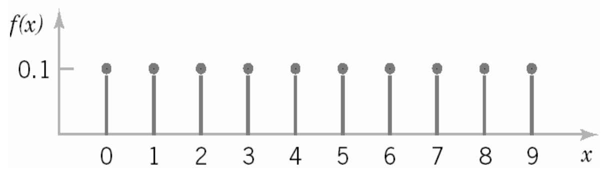
Sec 5.2.1

Discrete Uniform Distribution

Discrete Uniform Distribution

- Simplest random variable assumes only a finite number of possible values, each with equal probability.
- A random variable X has a **discrete uniform distribution** if each of the N values in its range, say, x_1, x_2, …, x_N, has **equal probability**. Then,

$$f(x_i) = 1/N$$



Discrete Uniform Distribution

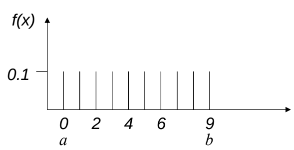
- Suppose X is a discrete uniform random variable on the consecutive integers a, a+1, a+2, …, b, for a < b. The mean of X is

$$\mu = E(X) = \frac{b+a}{2}$$

- The variance of X is

$$\sigma^2 = \frac{(b-a+1)^2 - 1}{12}$$

Discrete Uniform Distribution



$$f(x_i) = \frac{1}{n}$$

$$\mu = E(X) = \frac{b+a}{2}$$

$$\sigma^2 = \frac{(b-a+1)^2 - 1}{12}$$

- P(X=0)=0.1
- P(X=1)=0.1
- .
- .
- P(X=9)=0.1

Discrete Uniform Distribution REVISION

$$f(x \mid N_0, N_1) = \frac{1}{(N_0 - N_1 + 1)}$$

$$\mu = E(X) = \frac{N_0 + N_1}{2} \qquad \sigma^2 = \frac{(N_0 + N_1 + 1)^2 - 1}{12}$$

Sec 5.2.2

Bernoulli Probability Distribution

Bernoulli Random Variable

Any random variable whose only possible values are 0 and 1 is called a *Bernoulli random variable.*

Bernoulli Distribution

- A random experiment whose outcomes have been classified into two categories, called "**success**" and "**failure**", is called a **Bernoulli trial**.
- If a random variable X is defined as 1 if a Bernoulli trial results in success and 0 if the same Bernoulli trial results in failure.
- The value X=1 is often termed a "success"
- **p** is referred to as the **success probability**
- The value X=0 is often termed a "failure"

Bernoulli Random Variables

- The Bernoulli distribution characterizes the coin toss. Specifically, there are two events X=0,1 with X=1 occurring with probability p. The probability distribution function P[X] can be written as:

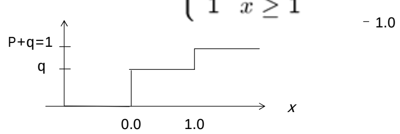
$$P[X] = p^x (1 - p)^{1-x}$$

Bernoulli Distribution

- Bernoulli distribution → RV generated by a single Bernoulli trial that has a binary valued outcome {0,1}

$$p_0 = p_X(0) = q;\ p_1 = p_X(1) = p;\ (p + q = 1)$$

$$F(x) = \begin{cases} 0 & x < p \\ q & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases}$$



Bernoulli Distribution

- Parameter: **p**

$$\mu = E(X) = p$$

$$\sigma^2 = p(1-p)$$

Sec 5.2.3

Binomial Probability Distribution

Binomial Experiment

An experiment for which the following four conditions are satisfied is called a *binomial experiment*.

Binomial Experiment

1. The experiment consists of a sequence of ***n*** **trials**, where *n* is fixed in advance of the experiment.
2. The **trials are identical**, and each trial can result in one of the same **two possible outcomes**, which are denoted by **success (*S*)** or **failure (*F*).**
3. The trials are independent.
4. The probability of success is constant from trial to trial: denoted by *p*.

Binomial Dist.: Sampling with Replacement

- Binomial distribution is also used in acceptance sampling.
  - Sampling is from an infinite population or **with replacement**.
- EX: A supplier of bolts claims that the defective rate is 5%, which is acceptable to us. So we randomly sample 5 bolts from a shipment and if we find 1 or more bad ones, we reject the shipment.

Binomial Dist: sampling w/o Replacement

Binomial Experiment

Suppose each trial of an experiment can result in *S* or *F*, but the **sampling is without replacement** from a population of size *N*. If the **sample size *n* is at most 5% of the population size**, the experiment can be analyzed as though it were exactly a binomial experiment.

Binomial Random Variable

Given a binomial experiment consisting of n trials, the **binomial random variable X** associated with this experiment is defined as

X = the number of S's among n trials

Example: Infectious disease

Suppose we examine five consecutive white cells.

1. **Let x = success = {neutrophil}**
2. **Let o=failure = {not neutrophil}**
3. **Let p = Pr{neutrophil},**
4. **Let q = 1-p.**

Then

Pr{oxoox} = (q)(p)(q)(q)(p) = p^2 q^3.

- **Note that any other outcome with exactly 2 x's and 3 o's will have the same probability.**
- **Thus the probability of any event with exactly 2 successes and 3 failures will be exactly p^2q^3.**

- **What is the probability that <u>any</u> 2 cells out of 5 will be neutrophils?**

All possible outcomes with exactly 2 neutrophils:

xxooo	**oxxoo**	**ooxox**
xoxoo	**oxoxo**	**oooxx**
xooxo	**oxoox**	
xooox	**ooxxo**	

Pr{exactly 2 neutrophils) =

(number of outcomes with exactly 2 neutrophils)

times

Pr{any outcome with exactly 2 neutrophils}

$$= \binom{5}{2} \text{ (times) } p^2 (1-p)^3$$
$$= \binom{5}{2} p^2 (1-p)^3.$$

If the probability of a success is p = 0.6 then

Pr{exactly 2 netruphils}

$$= \binom{5}{2} (0.6)^2 (1-0.6)^3$$
$$= \frac{5!}{2!3!}(0.6)^2(0.4)^3 = \frac{(5)(4)}{2}(0.6)^2(0.4)^3$$
$$= 0.230.$$

 # In R

```
> dbinom(2,5,0.6)
[1] 0.2304
```

Suppose in general we have n trials.

- **Pr{any sequence with exactly k success in n trials} =**

  p^k (1-p)$^{n-k}$

- **The number of possible sequences with exactly k successes is**

$$\binom{n}{k} = \frac{(n!)}{k!(n-k)!}$$

- **Pr{exactly k success in n trials} =**

$$\binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} p^k q^{n-k}.$$

k = 0, 1, 2, . . . , n.

Notation for the pmf of a Binomial rv

Because the pmf of a binomial rv *X* depends on the **two parameters** *n* and *p*, we denote the pmf by $b(x;n,p)$.

Binomial Distribution

- The random variable X that equals the number of trials that result in a success has a binomial random variable with parameters 0 < p < 1 and n = 1, 2, …
- The probability mass function of X … is:
  (X = 0, 1, 2, …, n)

$$f(x) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

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

Example: Evans Electronics

- Using the Binomial Probability Function

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

$$f(1) = \frac{3!}{1!(3-1)!} (0.1)^1 (0.9)^2$$

$$= (3)(0.1)(0.81)$$

$$= .243$$

```
> dbinom(1,3,0.1)
[1] 0.243
```


Ex. A card is drawn from a standard 52-card deck. If drawing a club is considered a success, find the probability (before experiment of population) of

a. exactly one success in 4 draws (with replacement).

$\boxed{p = 1/4;\ q = 1 - 1/4 = 3/4}$

$$\binom{4}{1}\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^3 \quad \boxed{\approx 0.422}$$

```
# in R
> dbinom(1,4,0.25)
[1] 0.421875
```

b. no successes in 5 draws (with replacement).

$$\binom{5}{0}\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^5 \quad \boxed{\approx 0.237}$$

Notation for cdf

For $X \sim \text{Bin}(n, p)$, the cdf will be denoted by

$$P(X \le x) = B(x; n, p) = \sum_{y=0}^{x} b(y; n, p)$$

$$x = 0, 1, 2, \ldots n$$

Binomial Distribution
Mean and Variance

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

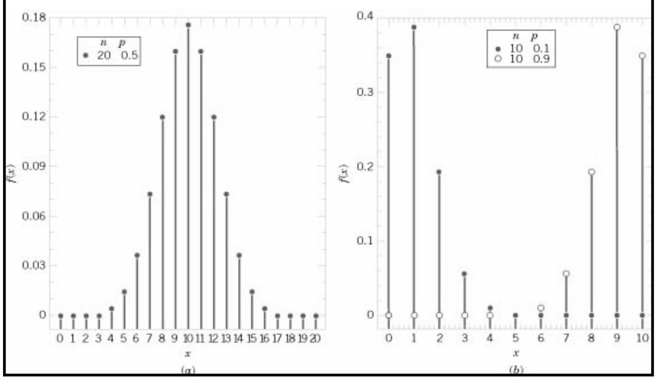
$E(X) = np$,

$V(X) = np(1-p) = npq,\ \sigma_X = \sqrt{npq}$
(where $q = 1 - p$).

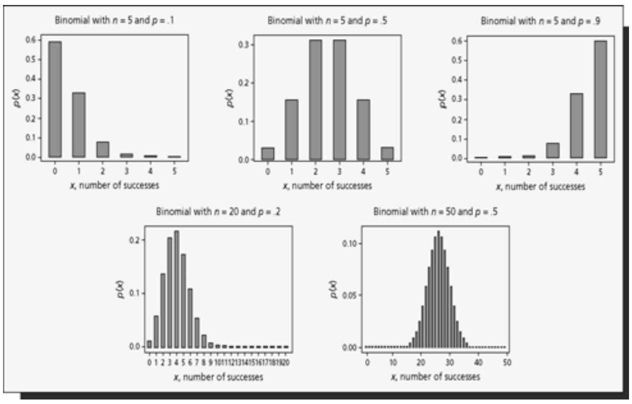
Binomial Distribution

- For a fixed n, the distribution becomes more **symmetric as p increases from 0 to 0.5 or decreases from 1 to 0.5.**
- For a fixed p, the distribution becomes more **symmetric as n increases.**

Binomial Distribution



Several Binomial Distributions



Binomial distribution

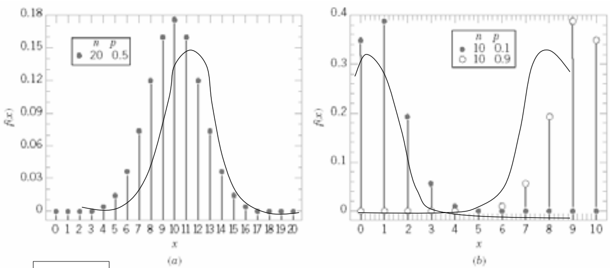


Figure 3-8

Figu… mial distribution for selected values of n and p.

Ex. 5 cards are drawn, with replacement, from a standard 52-card deck. If drawing a club is considered a success, find the (population) mean, variance, and standard deviation of *X* (where *X* is the number of successes).

$\boxed{p = ¼; q = 1 - ¼ = ¾}$

$$\boxed{\mu = np = 5\left(\frac{1}{4}\right) = 1.25}$$

$$\boxed{V(X) = npq = 5\left(\frac{1}{4}\right)\left(\frac{3}{4}\right) = 0.9375}$$

$$\boxed{\sigma_X = \sqrt{npq} = \sqrt{0.9375} \approx 0.968}$$

Ex. If the probability of a student successfully passing this course (C or better) is p=0.82, find the probability that given total 8 students

a. all 8 pass. $\binom{8}{8}(0.82)^8(0.18)^0$ $\boxed{\approx 0.2044}$

b. none pass. $\binom{8}{0}(0.82)^0(0.18)^8$ $\boxed{\approx 0.0000011}$

c. at least 6 pass.

> 1-pbinom(5,8,0.82) # 1-CDF
[1] 0.83918

$$\binom{8}{6}(0.82)^6(0.18)^2 + \binom{8}{7}(0.82)^7(0.18)^1 + \binom{8}{8}(0.82)^8(0.18)^0$$

$$\approx 0.2758 + 0.3590 + 0.2044 \quad \boxed{= 0.8392}$$

(y=x) Binomial Cumulative Probability Table

n	y	0.01	0.02	0.03	0.04	0.05
2	0	0.9801	0.9604	0.9409	0.9216	0.9025
	1	0.9999	0.9996	0.9991	0.9984	0.9975
	2	1.0000	1.0000	1.0000	1.0000	1.0000
3	0	0.9703	0.9412	0.9127	0.8847	0.8574
	1	0.9997	0.9988	0.9974	0.9953	0.9928
	2	1.0000	1.0000	1.0000	0.9999	0.9999
	3	1.0000	1.0000	1.0000	1.0000	1.0000
4	0	0.9606	0.9224	0.8853	0.8493	0.8145
	1	0.9994	0.9977	0.9948	0.9909	0.9860

Binomial Distribution

If the success (bad bolt) rate is 5%, what is the probability a random sample of 5 bolts will contain at least 1 defective bolt?

Let X = # of defective bolts in sample of 5

$P(X \geq 1) = 1 - P(X = 0)$

Airline Flight Booking

- Not all passengers show up for their reserved seats.
- There is 10% chance that a passenger would not show up.
- Overbooking, 120 seats, sell 125 tickets.
- What is the probability that every passenger who shows up can take the flight?
  - $P(X \leq 120)$
- What is the probability that flight takes off with all empty seats?
  - $P(x=0)$

Airline Flight Booking

- Not all passengers show up for their reserved seats.
- There is 10% chance that a passenger would not show up.
- On the average, how many passengers will show up? Standard deviation?
  - $\mu = n(p) = 125(0.9) = 112.5$
  - $\sigma^2 = Np(1-p) = 125(0.9)(0.1) = 11.25$, $\sigma = ?$

Sec 5.2.4

Geometric Probability Distribution

Geometric Distribution

- A series of Bernoulli trials with
- Probability of success = p
- X: random variable,
  = **the number of trials until the first success.**

$$f(x) = p^1 (1-p)^{x-1}, x = 1, 2, \ldots$$

$$\mu = E(X) = \frac{1}{p}$$

$$\sigma^2 = V(X) = \frac{(1-p)}{p^2}$$

Geometric Distribution

- **Multiple Bernoulli trials → occurrence of 1st success**.

$$S = \{\underbrace{0,0\ldots,0}_{i-times},1\} = \{0^{i-1}1 | i = 1,2,3,\ldots\}$$

- In general, S may have countably infinite size
- Define RV $Z(\in S)$: sample point $0^{i-1}1 = i$
- Z has image {1,2,3,….}. Assuming independent trials,

$$p_Z(i) = q^{i-1}p = p(1-p)^{i-1},\ i = 1,2,3,\ldots$$

$$F_Z(t) = \sum_{i=1}^{\lfloor t \rfloor} p(1-p)^{i-1} = 1-(1-p)^{\lfloor t \rfloor},\ t \geq 0$$

Lack of Memory

- Flip a coin, P(head)=0.4
- You have flipped a coin and see a tail.
- What's the probability that you will first see a head only at the end of next 10 flips.
- $P=(1-0.4)^{10-1}(0.4)$
- You have done 1000 trials, and see no head.
- What's the probability that you will first see a head only at the end of next 10 flips.
- $P=(1-0.4)^{10-1}(0.4)$

Geometric Random Variables

- For example, suppose once again that a basketball player hits 70% of his free throws. What is the probability that if he starts shooting free throws, he gets his first basket on the fifth shot? Here we have a random variable X~geometric(.7) and we want to know f(5). By the formula

$$f(5) = .7 \cdot .3^4 = 0.00567$$

```
> dgeom(4,0.7) #PDF
[1] 0.00567
```

- Not very likely!

Sec 5.2.5:

Hypergeometric Distributions

Negative Binomial Distributions

Hypergeometric Distribution

Three Assumptions

- The population or set to be sampled consists of ***N*** individuals, objects, or elements (a **finite population**).
- Each individual can be characterized as a success (*S*) or failure (*F*), and there are ***M*** successes in the population.
- A sample of ***n* individuals is selected without replacement** in such a way that each subset of size *n* is equally likely to be chosen.

Hypergeometric Distribution

If *X* is the number of *S*'s in a completely random sample of size *n* drawn from a population consisting of ***M*** *S*'s and (***M– K***) *F*'s, then the probability distribution of *X*, called the hypergeometric distribution, is given by

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

$\max(0, n-N+M)\le x\le \min(n,M)$

Hypergeometric Distribution

Mean and Variance

$$E(X)=n\cdot\frac{M}{N}\qquad V(X)=\left(\frac{N-n}{N-1}\right)\cdot n\cdot\frac{M}{N}\left(1-\frac{M}{N}\right)$$

Negative Binomial Distribution

The *negative binomial rv* and *distribution* are based on an experiment satisfying the following four conditions:

Negative Binomial Distribution

1. The experiment consists of a sequence of independent trials.
2. Each trial can result in a success (*S*) or a failure (*F*).
3. The probability of success is constant from trial to trial, so $P(S \text{ on trial } i) = p$ for i = 1, 2, 3, …
4. The experiment continues **until a total of *r* successes** have been observed, where *r* is a specified positive integer.

pmf of a Negative Binomial

The pmf of the negative binomial rv *X* with parameters *r* = number of *S*'s and $p = P(S)$ is

$$nb(x;r,p)=\binom{x+r+1}{r-1}p^r(1-p)^x$$

x = 0, 1, 2, …

Negative Binomial

Mean and Variance

$$E(X) = \frac{r(1-p)}{p} \qquad V(X) = \frac{r(1-p)}{p^2}$$

Sec 5.2.6

Poisson Probability Distribution

Poisson Distribution

A random variable X is said to have a *Poisson distribution* with parameter $\lambda \left(\lambda > 0\right)$, if the pmf of X is

$$p(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad x = 0,1,2...$$

λ represents the expected number of events per unit time

Poisson Distribution

Mean and Variance

If X has a Poisson distribution with parameter λ, then

$$E(X) = V(X) = \lambda$$

Poisson Distribution

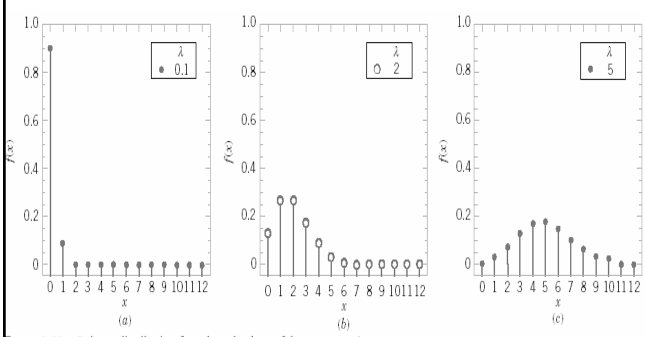
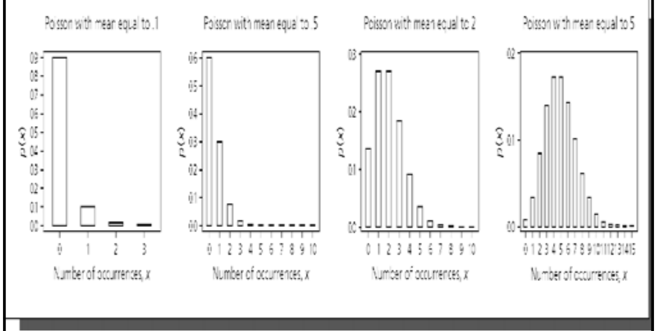


Figure 3-33 Poisson distribution for selected values of the parameter **λ**.

Several Poisson Distributions



- Note that
  - λ represents the expected number of events per unit time.
  - $\mu = \lambda t$ represents the expected number of events over the time period t.
  - For the binomial distribution there are a finite number of events possible.
  - For the Poisson distribution the number of events can be indefinitely large, although the probability of k events will get very small as k gets large.

The Poisson Distribution

Example:

- Consider the distribution of the number of deaths attributed to typhoid fever over a long period of time.

Consider the typhoid example:

- Assume that the expected number of events in one year is λ =4.6.
- For a 6-month period we have that λ = 4.6 and t=0.5.
- Thus $\mu = \lambda t = 2.3$.

$$\Pr(X = k) = \frac{\mu^k e^{-\mu}}{k!} \quad k = 0, 1, 2, \ldots$$

$$\Pr(X = 0) = \frac{2.3^0 e^{-2.3}}{0!} = 0.100$$

$$\Pr(X = 1) = \frac{2.3^1}{1!} e^{-2.3} = 0.231$$

$$\Pr(X = 2) = \frac{2.3^2}{2!} e^{-2.3} = 0.265$$

$$\Pr(X = 3) = \frac{2.3^3}{3!} e^{-2.3} = 0.203$$

$$\Pr(X = 4) = \frac{2.3^4}{4!} e^{-2.3} = 0.117$$

$$\Pr(X = 5) = \frac{2.3^5}{5!} e^{-2.3} = 0.054$$

```
> dpois(0:5,2.3)
[1] 0.10025884
0.23059534
0.26518464
0.20330823
0.11690223
0.05377503
>
> 1-ppois(5,2.3)
[1] 0.02997569
```

$$\Pr(X \geq 6) = 1 - \Pr(X \leq 5)$$
$$= 1 - (.100+.231+.265+.203+.117+.054) = 0.030.$$

Let Y = the number of deaths in 3 months.

For a 3-month period we have that λ = 4.6 and t=0.25.

Thus $\mu = \lambda t = 1.15$.

$$\Pr(X = 0) = e^{-1.15} = 0.317$$

$$\Pr(X = 1) = \frac{1.15^1}{1!} e^{-1.15} = 0.364$$

$$\Pr(X = 2) = \frac{1.15^2}{2!} e^{-1.15} = 0.209$$

$$\Pr(X = 3) = \frac{1.15^3}{3!} e^{-1.15} = 0.080$$

$$\Pr(X \geq 4) = 1 - (0.317 + 0.364 + .209 + .080) = 0.03$$

```
> dpois(0:3,1.15)
[1] 0.31663677
0.36413228
0.20937606
0.08026082
>
> 1-ppois(3,1.15)
[1] 0.02959406
```


Poisson Distribution

$$p(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad x = 0,1,2\ldots$$

Recursion Rule for Poisson Probabilities

If Pr(X = k) is the Poisson probability of observing k events with underlying parameter :, then

Pr(X = k+1) = [λ /(k+1)]Pr(X = k).

Example: Mercy Hospital

- Using the Poisson Probability Function

  Patients arrive at the emergency room of Mercy 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?

$$\mu = 6/\text{hour} = 3/\text{half-hour},\ x = 4$$

$$f(4) = \frac{3^4(2.71828)^{-3}}{4!} = .1680$$

Poisson Distribution

- If the Poisson random variable represents the number of counts in some interval, the mean of the random variable must equal the expected number of counts in the same length of interval.
- **It is important to use consistent units in the calculation of probabilities, means, and variances involving Poisson random variables.**
- **Example:**
  If average number of flaws per millimeter of wire is 3.4, then the average number of flaws in 10 millimeters of wire is 34.

Consistent Units

- If RV, X, represents the number of counts in some interval, the mean must be the expected number of counts in the same length of interval
  - Average number of cars arriving at a toll booth is 20 cars/hour
  - Average number of cars arriving at a toll booth is 1/3 cars/minutes
  - Average number of cars arriving at a toll booth is 480 cars/day

Highway Crack Repair

- The number of cracks in a section of interstate highway that are significant enough to require repair is assumed to follow a Poisson distribution with a **mean of 2/mile**.
- The probability of no crack repairs in 5 miles.

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!};\ \lambda = 5(2)/mile\ ,\ x = 0$$

$$P(X = 0) = \frac{e^{-10}10^0}{0!} = 0.0000454$$

- What is the probability that at least one crack requires repair in ½ mile of highway.
- $$\lambda = 0.5(2)/mile, x = 1$$

$$P(X \geq 1) = 1 - \frac{e^{-1}1^0}{0!} = 0.632$$

Binomial with Poisson Approximation

Poisson Distribution as a Limit

Suppose that in the binomial pmf $bin(x;n,p)$, we let $n \to \infty$ and $p \to 0$ in such a way that np approaches a value $\lambda > 0$.
Then $b(x;n,p) \to p(x;\lambda)$.

Relationship between Poisson and Binomial Distributions

- How large does n have to be and how small does p have to be?
  - Rule of thumb: $n \geq 100$ and $p \leq 0.01$
- EX: We are assembling circuit boards. History tells us that 1% of the connections are defective. What is the probability that on a board with 100 connections exactly 4 will be defective?

Common Continuous Distribution

Common Continuous Distributions

- Uniform
- Exponential
- Gamma
- Gaussian (Normal)

Continuous Random Variables

A random variable X is *continuous* if its set of possible values is an entire interval of numbers (If $A < B$, then any number x between A and B is possible).

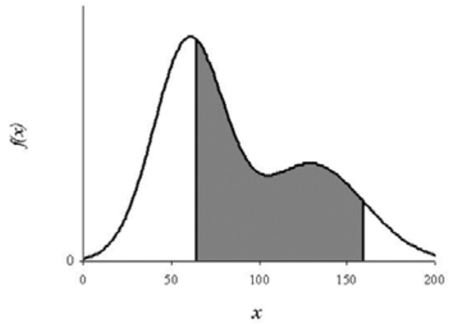
Probability Distribution

Let X be a continuous rv. Then a *probability distribution or probability density function (pdf)* of X is a function $f(x)$ such that for any two numbers a and b,

$$P\left(a \le X \le b\right) = \int_a^b f(x)dx$$

The graph of f is the *density curve*.

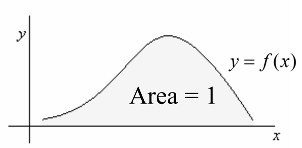
- The Total Area under the probability density curve is 1.
- The Area under the probability density curve is from a to b is $P[a < X < b]$.



Probability Density Function

For $f(x)$ to be a pdf

1. $f(x) > 0$ for all values of x.
2. The area of the region between the graph of f and the x – axis is equal to 1.



Def 5.3.1

Uniform Distribution

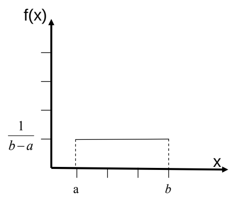
(Continuous) Uniform Distribution

A continuous rv X is said to have a *uniform distribution* on the interval $[A, B]$ if the pdf of X is

$$f\left(x; A, B\right) = \begin{cases} \dfrac{1}{B-A} & A \le x \le B \\ 0 & \text{otherwise} \end{cases}$$

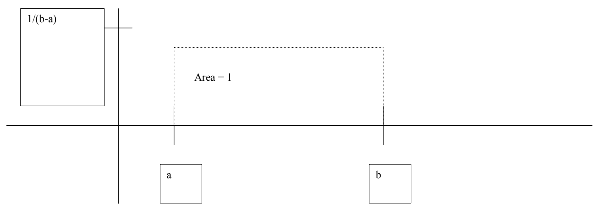
(Continuous) Uniform Distribution

- Prob. Density Function $f(x) = \dfrac{1}{b-a};\ a < x < b$
- Mean
$$\mu = \frac{(a+b)}{2}$$
- Variance
$$\sigma^2 = \frac{(b-a)^2}{12}$$



Uniform Random Variables: pdf

- The graph of the pdf satisfies our intuition about "equal likelihood" of all intervals of a given length within (a,b). It also clearly has total area 1 under the pdf curve.



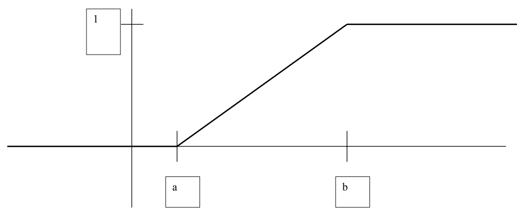
Uniform Random Variables: cdf

- The cumulative distribution function F given below is easy to compute either by integration of the pdf or by finding the area of rectangles. Note that it has all the usual properties of a cdf: 0 to the left, 1 to the right, increasing and continuous inbetween.

$$F(x) = \begin{cases} 0, \text{ if } x \le a \\ \frac{x-a}{b-a}, \text{ if } a < x < b \\ 1, \text{ if } x \ge b \end{cases}$$

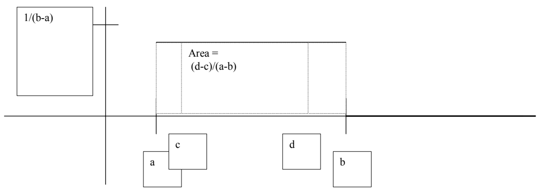
Uniform Random Variables: cdf

- Here we see F and its properties graphically.



Uniform Random Variables: cdf

- If we want to find the probability P(c<X<d) where a<c<d<b, then we can integrate formally, but it is easier to note that the probability is simply the ratio of the length of (c,d) to the length of (a,b).



Uniform Random Variables: Expectation

- Intuitively we anticipate E(X)=(a+b)/2, the midpoint of the interval. This turns out to be correct.
  - If X~uniform(a,b) we calculate

$$E(X)=\int_{-\infty}^{\infty} xf(x)dx =$$

$$\int_a^b \frac{1}{b-a} x dx = \frac{1}{2(b-a)} x^2 \Big|_a^b = \frac{b^2-a^2}{2(b-a)}$$

$$= \frac{(b+a)(b-a)}{2(b-a)} = \frac{b+a}{2}$$

Def 5.3.2

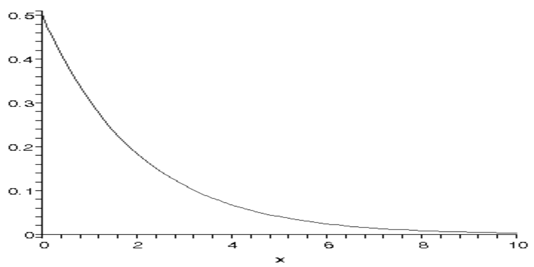
Exponential Distribution

Exponential Distribution

A continuous rv X has an *exponential distribution* with parameter λ if the pdf is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

Exponential Random Variables: Pdf



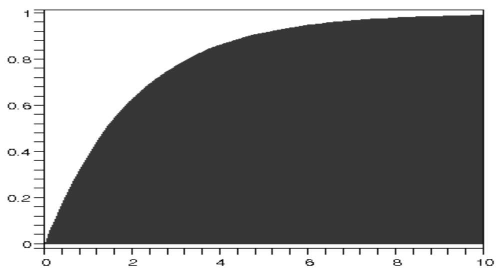
Exponential Distribution

Let X have a exponential distribution
Then the cdf of X is given by

$$F(x;\lambda) = \begin{cases} 0 & x < 0 \\ 1-e^{-\lambda x} & x \ge 0 \end{cases}$$

Exponential Random Variables: Cdf

- Essentially the same computation as above shows that . Here is the graph of the cdf for X~exponential(0.5).



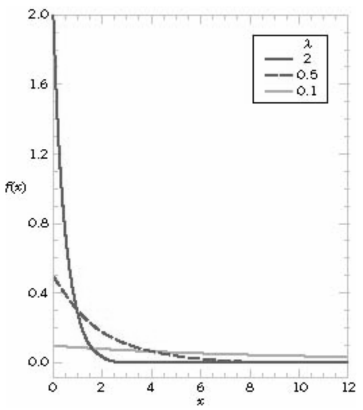
Exponential Distribution

Mean and Variance

The mean and variance of a random variable *X* having the exponential distribution

$$\mu = \frac{1}{\lambda} \quad \sigma^2 = \frac{1}{\lambda^2}$$

Probability distribution function of exponential random variable. Note how the distribution is skewed.



Exponential Random Variables: Applications

- Exponential distributions are sometimes used to **model waiting times** or **lifetimes**. That is, they **model the time until some event happens** or something quits working. Of course mathematics cannot tell us that exponentials are right to describe such situation. That conclusion depends on finding data from such real-world situations and fitting it to an exponential distribution.

.

Exponential Random Variables: Applications

- Suppose the wait time X for service at the post office has an exponential distribution with mean 3 minutes. If you enter the post office immediately behind another customer, what is the probability you wait over 5 minutes? Since E(X)=1/λ=3 minutes, then λ=1/3, so X~exponential(1/3). We want .

$$P(X > 5) = 1 - P(X \le 5) = 1 - F(5)$$

$$= 1 - \left(1 - e^{-\frac{1}{3}5}\right) = e^{-\frac{5}{3}} \approx 0.189$$

```
> 1-pexp(5,1/3)
[1] 0.1888756
```


Exponential Random Variables: Applications

- Under the same conditions, what is the probability of waiting between 2 and 4 minutes? Here we calculate .

$$P(2 \le X \le 4) = F(4) - F(2) = \left(1 - e^{-\frac{4}{3}}\right) - \left(1 - e^{-\frac{2}{3}}\right)$$

$$= e^{-\frac{2}{3}} - e^{-\frac{4}{3}} \approx 0.250$$

```
> pexp(4,1/3)-pexp(2,1/3)
[1] 0.24982
```

Exponential Random Variables: Applications

- The trick in the previous example of calculating

$$P(a \le X \le b) = F(b) - F(a)$$

is quite common. It is the reason the cdf is so useful in computing probabilities of continuous random variables. .

Applications of the Exponential Distribution

Suppose that the number of events occurring in any time interval of length t has a Poisson distribution with parameter αt and that the numbers of occurrences in nonoverlapping intervals are independent of one another. Then the distribution of elapsed time between the occurrences of two successive events is exponential with parameter $\lambda = \alpha$.

Exponential Distribution

- The number of events during an interval, X~Poisson with mean rate of λ.

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0,1,2,3,...; \mu = \lambda; \sigma^2 = \lambda$$

- 1/rate= average (time) length between two events.
- The length to next event has a exponential distribution with (length) mean of 1/ λ

$$f(x) = \lambda e^{-\lambda x}; x \ge 0; \mu = \frac{1}{\lambda}; \sigma^2 = \frac{1}{\lambda^2}$$

- Mean Time To Failure (MTTF)

Relationship Between Exponential & Poisson

Recall:

$$p(x;\lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}, x = 0,1,2,...$$

where λ is mean number of events per base unit time or space and t is the number of base units being inspected.

The probability that no events occur in the span of time (or space) t is:

$$p(0;\lambda t) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t}$$

Relationship Between Exponential & Poisson

Let X = the time (or space) to the first Poisson event.

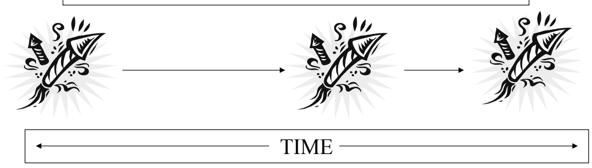
Note, the probability that the length of time (or space) until the first event > some time (or space), x is the same as the probability that no events will occur in x, which = $e^{-\lambda x}$.

So, $P(X > x) = e^{-\lambda x}$ and $P(X < x) = 1 - e^{-\lambda x}$

$1 - e^{-\lambda x}$ is the cumulative distribution function for an exponential random variable with λ.

Relationship between Exponential & Poisson

Exponential distribution models time (or space) between Poisson events.

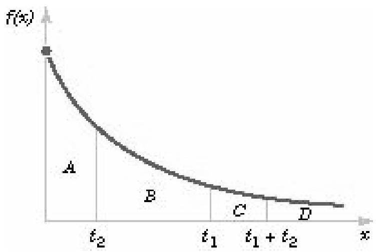


User login Example

- Users log on to computer network is modeled as a Poisson process with mean of 25 logons/hr.
- Find the probability that no logins with a 6-minute interval.
- Poisson Distribution
  - X ~ # of logins in 6-min interval
  - 25/hr=25(6)/60=2.5 longins/6-min
  - P(X=0)=Poisson(0,2.5,true)=0.0820
- Exponential Distribution
  - X~ time (min) until next login.
  - 25/hr→1/λ=60min/25=2.4 min; λ= 0.4167
  - P(X>6min)=1-P(X≤6min) = 1 - Expondist(6,0.4167,true) =0.0820

Exp. Distribution: Memoryless

- X~ time to next login
  X~ exponential distribution with mean of 2.4 min.
- We know that there are no logins from 12:00 to 12:15. What is the probability that there are no logins from 12:15 to 12:21?
  - P(X>6min)=1-P(X≤6min) =0.0820
- We know that there are no logins from 1:00-2:00PM. It is 2:00PM now. What is the probability that next login will occur at 2:06PM?
  - P(X=6min)=0.0342



$P(X<t_1+t_2|X>t_1)=P(X<t_2)$

Exponential Distribution

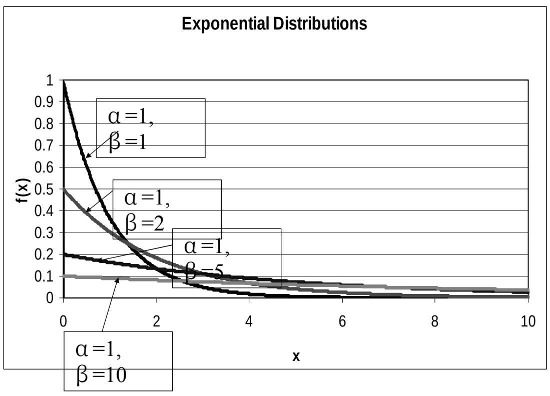
A gamma distribution with α = 1 is called the exponential distribution.

$$f(x) = \frac{1}{\beta} e^{-x/\beta}, x > 0 \qquad F(x) = 1 - e^{-x/\beta}$$

where $\beta > 0$

$$\mu = \beta \quad \text{and} \quad \sigma^2 = \beta^2$$

Exponential Distributions



Def 5.3.3

Gamma Distribution

Gamma Function

For $\alpha > 0$, the *gamma function* $\Gamma(\alpha)$ is defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

Gamma Distribution

X is a Gamma RV with parameters λ and r.

$$f(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)}, x > 0$$

$$\Gamma(r) = \int_{-\infty}^{\infty} x^{r-1} e^{-x} dx, r > 0$$

$$\Gamma(r) = (r-1)\Gamma(r-1)$$

$$\mu = \frac{r}{\lambda}$$

$$\sigma^2 = \frac{r}{\lambda^2}$$

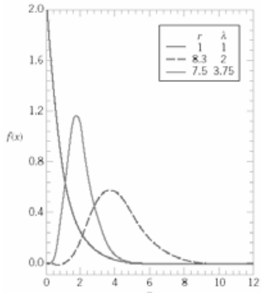


Figure 3-21 Gamma probability density functions for selected values of λ and r.

Gamma Distribution (Special Cases)

A Gamma RV with parameters λ and $r=1$, it is exponential distribution

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx = \int_0^\infty x^{1-1} e^{-x} dx = -e^{-x}\Big|_0^\infty = 1$$

$$f(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)} = \frac{\lambda^1 x^{1-1} e^{-\lambda x}}{1} = \lambda e^{-\lambda x}$$

Chi-squared Distribution

If $\lambda = 1/2$, $r=(1/2)n$, $n=1,2,3,\ldots,$

Erlang Distribution

If r is an integer

Gamma Distribution

A continuous rv X has a *gamma distribution* if the pdf is

$$f(x;\alpha,\beta) = \begin{cases} \dfrac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

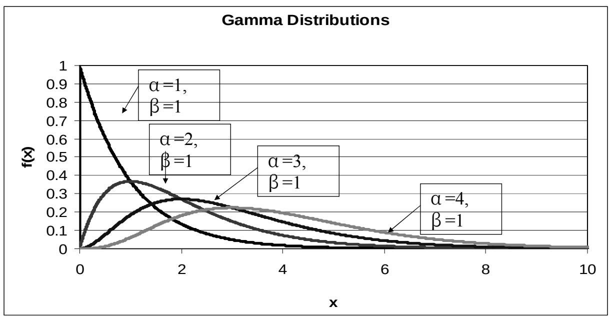
where the parameters satisfy $\alpha > 0, \beta > 0$.
The standard gamma distribution has $\beta = 1$.

Mean and Variance

The mean and variance of a random variable X having the gamma distribution $f(x;\alpha,\beta)$ are

$$E(X) = \mu = \alpha\beta \quad V(X) = \sigma^2 = \alpha\beta^2$$

Gamma Distributions



Gamma & Exponential Distributions

- Exponential and gamma distributions find application in queuing theory and reliability studies.
  - Time between customer arrivals at a terminal


  - Time to failure of electrical components

The exponential distribution is a special case of the gamma distribution.

Def 5.3.4

Normal Distribution

Normal Distribution

- **The most widely used model for the distribution of random variables.**
- Whenever a random experiment is replicated, the random variable that equals **the average (or total)** result over the replicates tends to have a normal distribution **as the number of replicates becomes larger.**
- Normal distribution arises in the study of numerous basic physical phenomena (e.g., velocity of molecules in a gas)

Normal Distributions

A continuous rv X is said to have a normal distribution with parameters

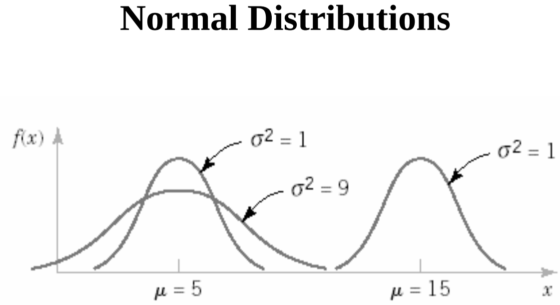
μ and σ, where $-\infty < \mu < \infty$ and $0 < \sigma$, if the pdf of X is

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

Normal Distribution

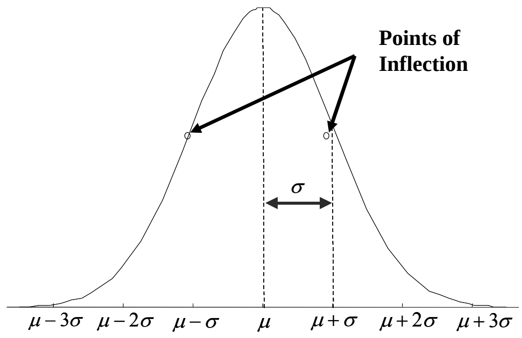
- **Random variables with different means and variances can be modeled by normal probability density function with appropriate choice of the center and width of the curve.**
- The value of **E(X) = μ** determines the center of the probability density function.
- The value of **V(X) = σ²** determines the width.

Normal Distributions



Normal probability density functions for selected values of the parameters μ and σ2

The Normal Probability Distribution

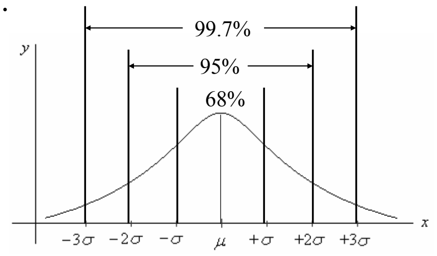


Characteristics of the Normal Distribution

- Bell Shaped, symmetric
- Points of inflection on the bell shaped curve are at $\mu - \sigma$ and $\mu + \sigma$. That is one standard deviation from the mean
- Area under the bell shaped curve between $\mu - \sigma$ and $\mu + \sigma$ is approximately 2/3.
- Area under the bell shaped curve between $\mu - 2\sigma$ and $\mu + 2\sigma$ is approximately 95%.

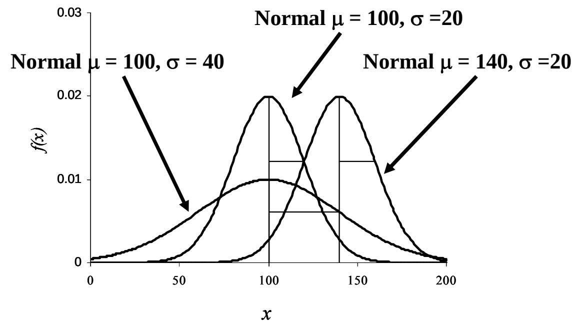
Normal Curve (Bell Shape Curve)

Approximate percentage of area within given standard deviations (empirical rule).

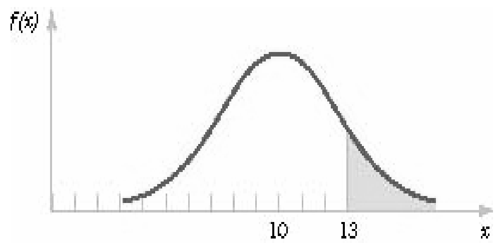


There are many Normal distributions

depending on by μ and σ



Normal Distribution



Probability that a measurement exceed 13

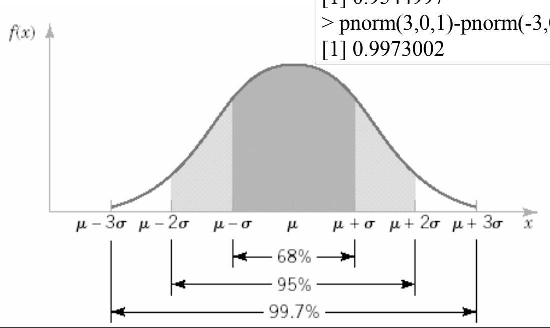
P (X > 13)

Normal Distribution

- There is no closed-form expression for the integral of a normal provability density function.
- **Probabilities based on the normal distribution are typically found numerically or from a table.**

Width of a Normal Distribution

```
> pnorm(1,0,1)-pnorm(-1,0,1)
[1] 0.6826895
> pnorm(2,0,1)-pnorm(-2,0,1)
[1] 0.9544997
> pnorm(3,0,1)-pnorm(-3,0,1)
[1] 0.9973002
```



Normal Distribution

- Due to symmetry, $P(X > \mu) = P(X < \mu) = 0.5$
- Probability that a measurement falls far from μ is small, and at some distance from μ the probability of an interval can be approximated as zero.
- **Width of the normal distribution = 6 σ**
- Area under the normal probability density function from $-\infty < \mu < \infty$ is 1.

Standard Normal pdf

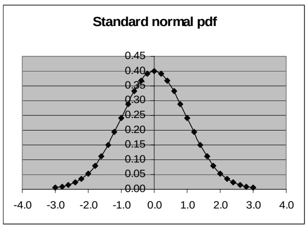
If μ=0, σ2 =1, It is standard normal distribution ~Z=N(0, 1)

Symmetrical

P(X<μ)=0.5

A normal RV, X, with parameters of μ and σ can be converted into a N(0,1) standard normal RV by:

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



Standard Normal Distributions

The normal distribution with parameter values $\mu = 0$ and $\sigma = 1$ is called a *standard normal distribution*. The random variable is denoted by *Z*. The pdf is

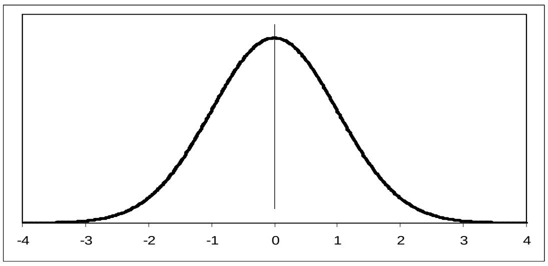
$$f(z;0,1) = \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/2} \qquad -\infty < z < \infty$$

The cdf is

$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} f(y;0,1)dy$$

Standard Normal Distribution

The distribution of a normal random variable with mean 0 and variance 1 is called a **standard normal distribution**.



Standard Normal Distribution

- The letter ***Z*** is traditionally used to represent a standard normal **random variable**.
- ***z*** is used to represent a **particular value** of *Z*.
- The standard normal distribution has been tabularized.

Nonstandard Normal Distributions

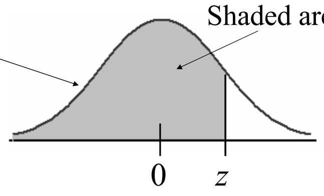
To Standard Normal Distributions

If X has a normal distribution with mean μ and standard deviation σ, then

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

has a standard normal distribution.

Standard Normal Cumulative Areas

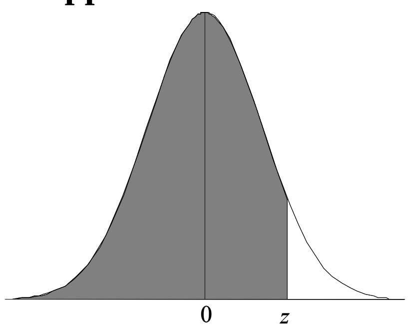


Standard Normal Distribution

Properties:

- The total area under the normal curve is equal to 1
- The distribution is bell-shaped and symmetric; it extends indefinitely in both directions, approaching but never touching the horizontal axis
- The distribution has a mean of 0 and a standard deviation of 1
- The mean divides the area in half, 0.50 on each side
- Nearly all the area is between z = -3.00 and z = 3.00

Appendix A : Table I

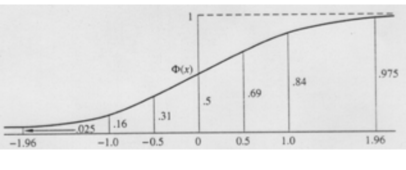


- The table contains the area under the standard normal curve between -∞ and a specific value of z

You can easily calculate probabilities for normal distributions if you know the values for the standard normal cdf.

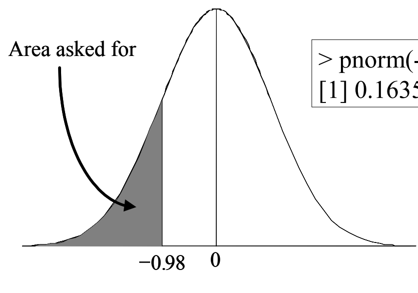
- **A simple rule is that**

Pr(a < X < b) = Pr{X<b} - Pr{X < a}.



Example

Find the area to the left of -0.98; $P(z < -0.98)$



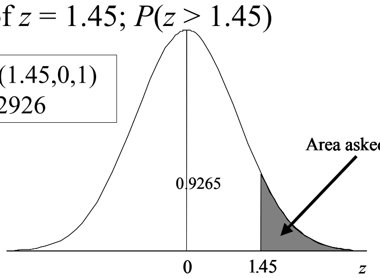

```
> pnorm(-0.98,0,1)
[1] 0.1635431
```

$P(z < -0.98) = 0.1635$

Example

Find the area under the normal curve to the right of z = 1.45; $P(z > 1.45)$

```
> 1-pnorm(1.45,0,1)
[1] 0.07352926
```

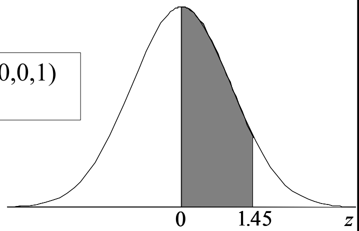


$P(z > 1.45) = 1.0000 - 0.9265 = 0.0735$

Example

Find the area to the between z = 0 and of z = 1.45; P(0 < z < 1.45)

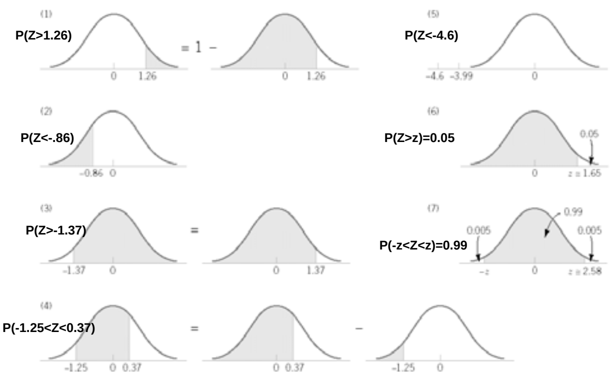
```
> pnorm(1.45,0,1)-pnorm(0,0,1)
[1] 0.4264707
```



$P(z < 1.45) = 0.9265 - 0.5000 = 0.4265$

- **Area between two points = differences in two tabled areas**

Some More Practices

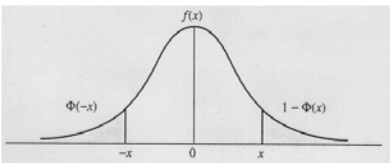


Result: (Symmetry properties of the standard normal distribution).

From the symmetry properties of the standard normal distribution,

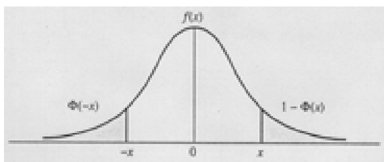
Φ(-x) = Pr(X ≤ -x) = Pr(X ≥ x) = 1 - Pr(X ≤ x) = 1 - Φ(x).

This property is depicted in the following figure.



- **Example**

Pr(X ≤ -1.96) = Pr (X ≥ 1.96) = 0.025



Standard Normal Distribution

Let Z be the standard normal variable.
Find (from table)

a. $P(Z \le 0.85)$

Area to the left of 0.85 = 0.8023

b. $P(Z > 1.32)$

$1 - P(Z \le 1.32) = 0.0934$

c. $P(-2.1 \le Z \le 1.78)$

Find the area to the left of 1.78 then subtract the area to the left of –2.1.

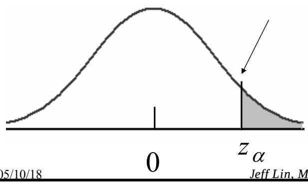
$= P(Z \le 1.78) - P(Z \le -2.1)$

= 0.9625 – 0.0179

= 0.9446

z_α and $z_{1-\alpha}$ Notation

z_α or $z_{1-\alpha}$ will denote the value on the measurement axis for which the area under the z curve lies to the right of z_α or $z_{1-\alpha}$



Shaded area

$= P(Z \ge z_\alpha) = \alpha$

$= P(Z \ge z_{1-\alpha}) = \alpha$

Ex. Let Z be the standard normal variable. Find z if a. $P(Z < z) = 0.9278$.

Look at the table and find an entry = 0.9278 then read back to find

$\boxed{z = 1.46.}$

b. $P(-z < Z < z) = 0.8132$

$P(z < Z < -z) = 2P(0 < Z < z)$

$= 2[P(z < Z) - ½]$

$= 2P(z < Z) - 1 = 0.8132$

$P(z < Z) = 0.9066$

$\boxed{z = 1.32}$

Normal Distribution

- If X is a normal variable with E(X) = μ and V(X) = σ^2, the random variable

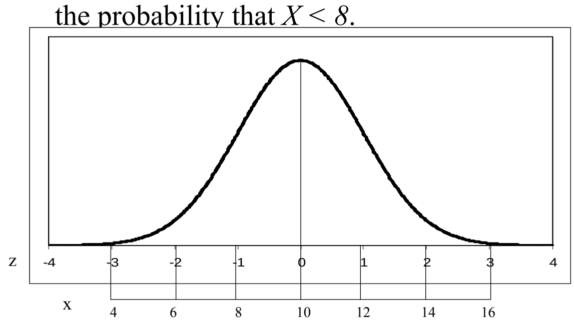
$$Z = \frac{X - \mu}{\sigma}$$

is a normal random variable with E(X) = 0 and V(X) = 1.

Z is a standard normal random variable.

Normal Distribution

- Given a random Variable X having a normal distribution with $\mu_x = 10$ and $\sigma_x = 2$, find the probability that $X < 8$.



Normal Distribution

- Suppose X is a normal random variable with mean μ and variance σ^2. Then,
- $P(X \le x) = P(\dfrac{X-\mu}{\sigma} \le \dfrac{x-\mu}{\sigma}) = P(Z \le z)$

  where Z is a standard normal random variable, and $z = (x-\mu)/\sigma$ is the z-value obtained by standardization X
- The probability is obtained by entering Appendix Table-I with $z = (x-\mu)/\sigma$.

To get this we must use some results:

1. $E(X + c) = E(X) + c$.
2. $E(cX) = c\,E(X)$.
3. $Var(X + c) = Var(X)$
4. $Var(cX) = c^2\ Var(X)$.

Using these results we have:

$$E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}E(X-\mu) = \frac{1}{\sigma}(E(X)-\mu) = \frac{1}{\sigma}(\mu-\mu) = 0.$$

$$Var(Z) = Var\left(\frac{X-\mu}{\sigma}\right) = \left(\frac{1}{\sigma}\right)^2 Var(X-\mu) = \left(\frac{1}{\sigma}\right)^2 (Var(X)) = \frac{\sigma^2}{\sigma^2} = 1.$$

Evaluation of Probabilities for Any Normal Distribution via Standardization

If X is N(μ, σ^2) and

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

$$\Pr(a < X < b) = \Pr\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Pr\left(Z < \frac{b-\mu}{\sigma}\right) - \Pr\left(Z < \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

This follows because

$a < X < b$ **is equivalent to**

$a-\mu < X-\mu < b-\mu$ **is equivalent to**

$\dfrac{a-\mu}{\sigma} < \dfrac{X-\mu}{\sigma} < \dfrac{b-\mu}{\sigma}$ **is equivalent to**

$\dfrac{a-\mu}{\sigma} < Z < \dfrac{b-\mu}{\sigma}$.

Therefore

$$\Pr(a < X < b) = \Pr\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right).$$

However, because Z is N(0, 1)

$$\Pr(a < X < b) = \Pr\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Pr\left(Z < \frac{b-\mu}{\sigma}\right) - \Pr\left(Z < \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

This procedure is known as standardization of a normal variable.

Ex. Let X be a normal random variable with $\mu = 80$ and $\sigma = 20$.
Find $P(X \le 65)$.

$$P(X \le 65) = P\left(Z \le \frac{65-80}{20}\right)$$
$$= P(Z \le -.75)$$
$$= 0.2266$$

```
> pnorm(65,80,20)
[1] 0.2266274
```

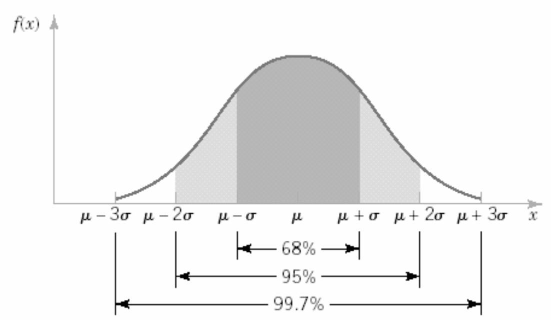

For example:

The probability of being borderline hypertensive when Diastolic Blood Pressure (X) is N(80, 144) is

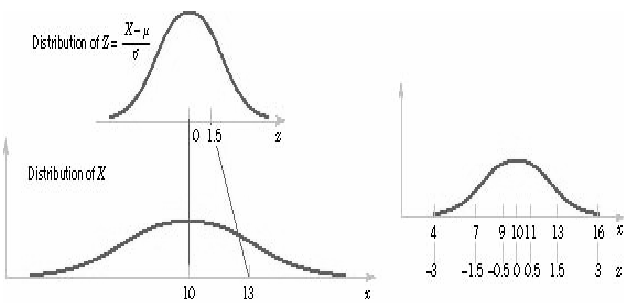
$$\Pr(90 < X < 95) = \Pr\left(\frac{90-80}{12} < Z < \frac{95-80}{12}\right)$$
$$= \Pr(0.83 < Z < 1.25)$$
$$= F(1.25) - F(0.83)$$
$$= 0.8944 - 0.7967 = 0.0977.$$

```
> pnorm(95,80,sqrt(144))-
pnorm(90,80,sqrt(144))
[1] 0.0966786
```

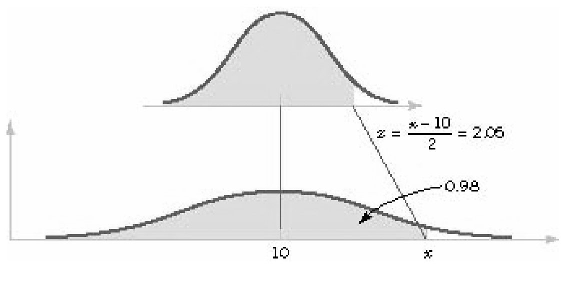

Width of a Normal Distribution



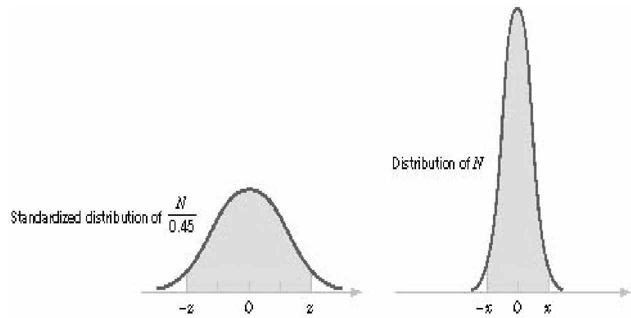
Normal Distribution



Normal Distribution

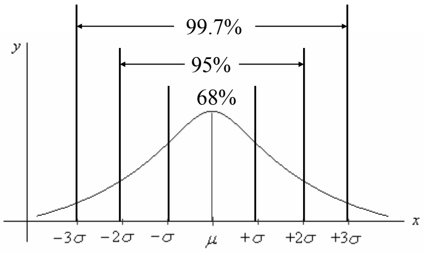


Normal distribution (Cont.)



Normal Curve (Bell Shape Curve)

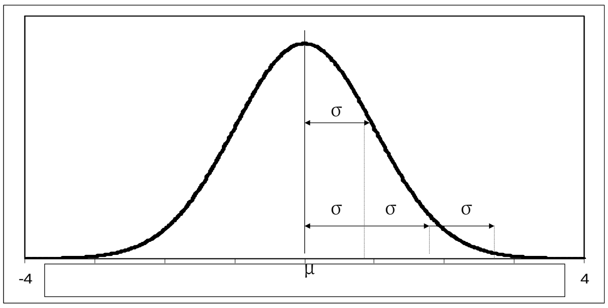
Approximate percentage of area within given standard deviations (empirical rule).



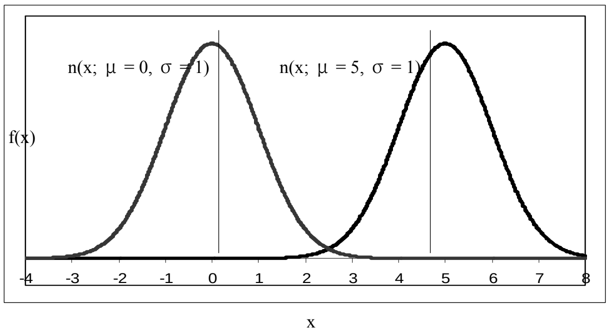
Normal Distribution

- Characteristics
  - Bell-shaped curve
  - $-\infty \leq x \leq +\infty$
  - μ determines distribution location and is the highest point on curve
  - Curve is symmetric about μ
  - σ determines distribution spread
  - Curve has its points of inflection at $\mu \pm \sigma$
  - $\mu \pm 1\sigma$ covers 68% of the distribution
  - $\mu \pm 2\sigma$ covers 95% of the distribution
  - $\mu \pm 3\sigma$ covers 99.7% of the distribution

Normal Distribution

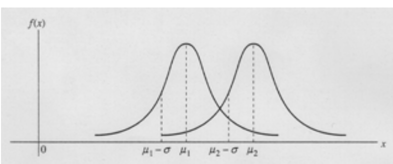


Normal Distribution

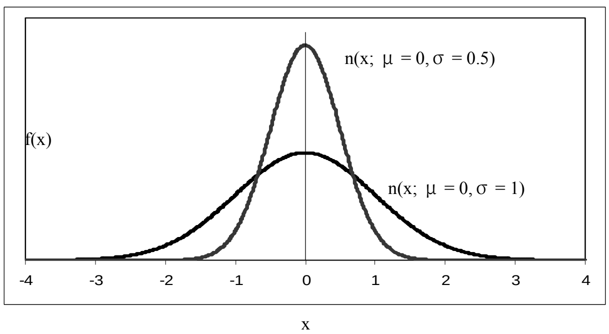


- **The entire shape of the normal distribution is determined by the two parameters μ and σ.**
- **If two normal distributions**
  1. **have the same variance σ^2 ,**
  2. **have different means μ_1 and μ_2, and**
  3. **$\mu_1 < \mu_2$ then**

  their density functions look like the following:



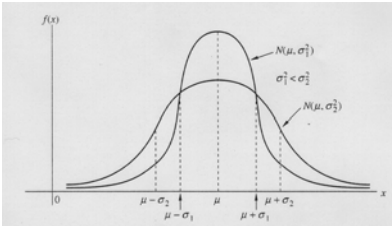
Normal Distribution



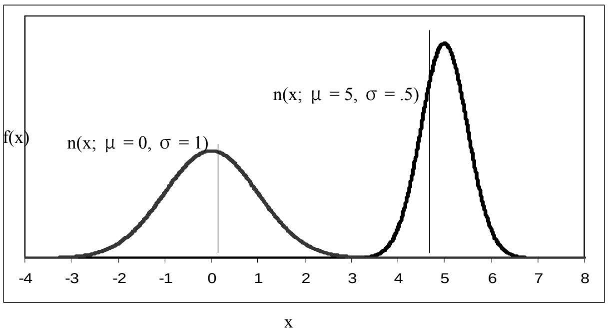
Similarly, two normal curves with

1. **the same means but**
2. **different variances ($\sigma_2^2 > \sigma_1^2$)**

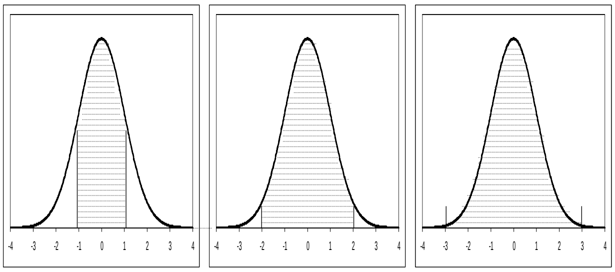
then their probability distribution functions are represented in the following figure:



Normal Distribution



Normal Distribution



$\mu \pm 1\sigma$ covers 68%	$\mu \pm 2\sigma$ covers 95%	$\mu \pm 3\sigma$ covers 99.7%

Ex. A particular rash shown up at an elementary school. It has been determined that the length of time that the rash will last is normally distributed with $\mu = 6$ days and $\sigma = 1.5$ days.

Find the probability that for a student selected at random, the rash will last for between 3.75 and 9 days.

$$P(3.75 \le X \le 9) = P\left(\frac{3.75-6}{1.5} \le Z \le \frac{9-6}{1.5}\right)$$

$$= P(-1.5 \le Z \le 2)$$

$$= 0.9772 - 0.0668$$

$$= 0.9104$$

```
> pnorm(9,6,1.5)-pnorm(3.75,6,1.5)
[1] 0.9104427
```


Definition:

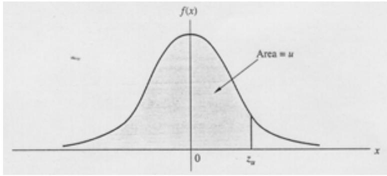
The (100 x u)th percentile of the standard normal distribution is denoted by z_u.

It is defined by the relationship

$Pr(X < Z_u) = u$, where X is N(0, 1).

Percentiles of an Arbitrary Normal Distribution

$$\begin{array}{l}(100p)\text{th percentile}\\ \text{for normal}\left(\mu,\sigma\right)\end{array} = \mu + \left[\begin{array}{l}(100p)\text{th for}\\ \text{standard normal}\end{array}\right]\cdot\sigma$$



- **The function Z_u is sometimes referred to as the <u>Inverse Normal Function</u>.**

Normal Approximation to the Binomial Distribution

- For many physical systems, the binomial model is appropriate with an extremely large values for n.
- In these cases, it is difficult to calculate probabilities by using the binomial distribution.
- Normal approximation is most effective in these cases.

Normal Approximation to the Binomial Distribution

- If X is a binomial random variable,

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

is approximately a standard normal random variable.
- The approximation is good for

np > 5 and n(1-p) > 5
- **The approximation is good when n is large relative to p.**

Normal Approximation to the Binomial Distribution

Let X be a binomial rv based on n trials, each with probability of success p. If the binomial probability histogram is not too skewed, X may be approximated by a normal distribution with $\mu = np$ and $\sigma = \sqrt{npq}$.

$$P(X \le x) \approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{npq}}\right)$$

Ex. At a particular small college the pass rate of Intermediate Algebra is 72%. If 500 students enroll in a semester determine the probability that at least 375 students pass.

$$\mu = np = 500(.72) = 360$$

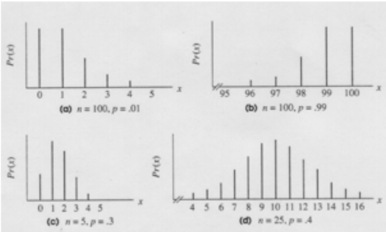
$$\sigma = \sqrt{npq} = \sqrt{500(.72)(.28)} \approx 10$$

$$P(X \le 375) \approx \Phi\left(\frac{375.5 - 360}{10}\right) = \Phi(1.55)$$

```
> pnorm(375.5, 360, 10)
[1] 0.9394292
```

$= 0.9394$

- Consider the following graphs.

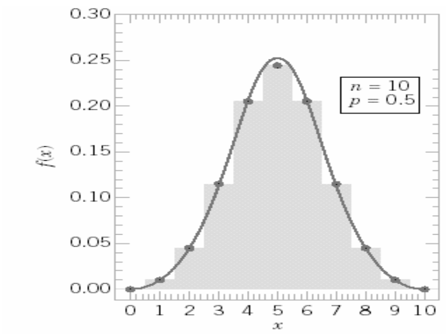


Normal Approximation to the Binomial Distribution

- **Note that as n gets large, the binomial distribution becomes more symmetric.**
- **Indeed if n is large relative to p, the binomial is well approximated by the normal distribution.**

Normal Approximation to Binomial

- When np > 5 and n(1-p) > 5, the approximation is good.

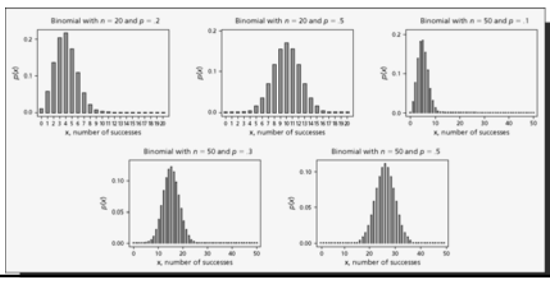


- **Note that as n gets large, the binomial distribution becomes more symmetric.**
- **Indeed if n is large relative to p, the binomial is well approximated by the normal distribution.**

Normal Approximation to the Binomial

If x is binomial, n trials each with probability of success p and n and p are such that np ≥ 5 and n(1-p) ≥ 5, then

x is approximately normal with



Binomial Approximation Example

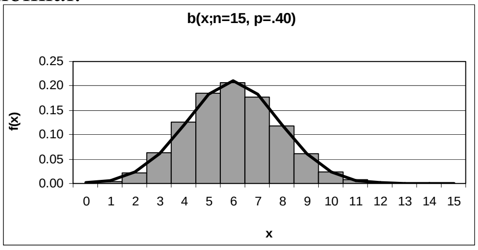
- A lot of chips contains 1000 chips. The process produces 2% defects.
- X: #of defects, follows Binomial
  $P(X > 25) = 1 - P(X < 25) = 0.1099 = 11\%$
- Approx. Normal

$$Z = \frac{25 - 1000\,(0.02)}{\sqrt{1000\,(0.02)(1 - 0.02)}} = 1.129385$$

$$P(X > 25) = 1 - Z(X < 1.129385) = 0.1294 = 13\%$$

Relationship between the Normal and Binomial Distributions

- Consider b(x;15,0.4). Bars are calculated from binomial. Curve is normal approximation to binomial.



Normal Approximation to the Binomial distribution

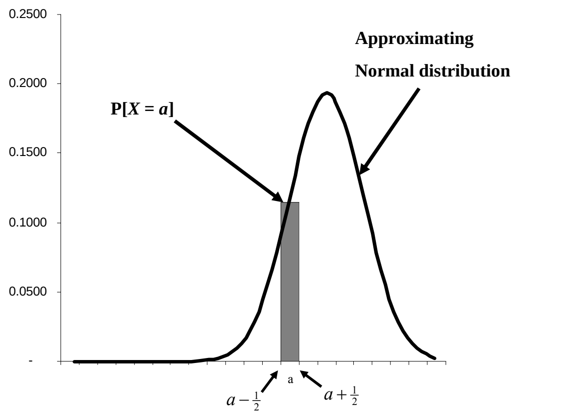
$$P[X=a] \approx P\left[a-\tfrac{1}{2} \le Y \le a+\tfrac{1}{2}\right]$$

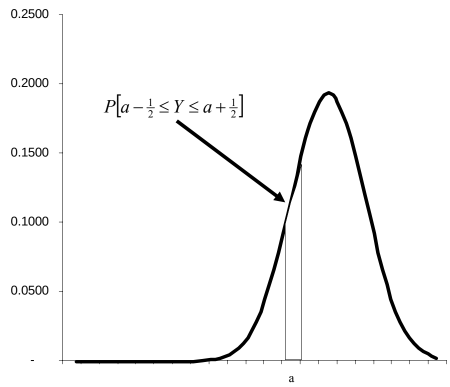
- X has a Binomial distribution with parameters n and p
- Y has a Normal distribution

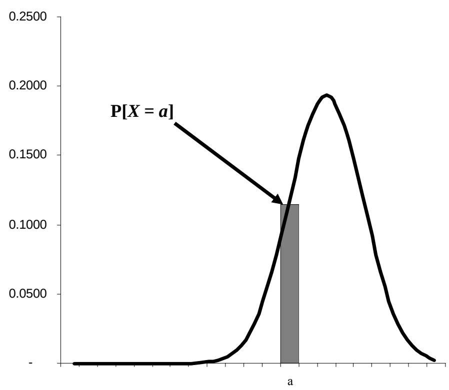
$$\mu = np$$

$$\sigma = \sqrt{npq}$$

$\frac{1}{2}$ = continuity correction







Using the Normal approximation to the Binomial distribution

$$P[X=13] \approx P\left[12\tfrac{1}{2} \le Y \le 13\tfrac{1}{2}\right]$$

Where Y has a Normal distribution with:

$$\mu = np = 20(0.70) = 14$$

$$\sigma = \sqrt{npq} = \sqrt{20(.70)(.30)} = 2.049$$

Hence

$P[12.5 \le Y \le 13.5]$

$$= P\left[\frac{12.5-14}{2.049} \le \frac{Y-14}{2.049} \le \frac{13.5-14}{2.049}\right]$$

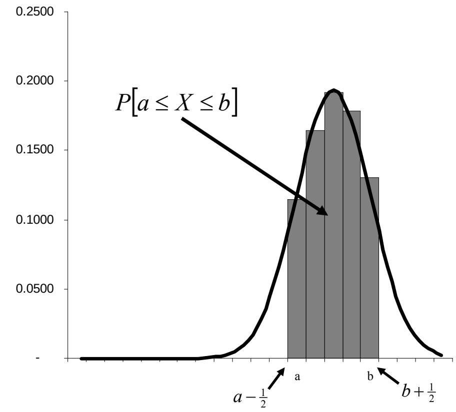
$= P[-0.73 \le Z \le -0.24]$

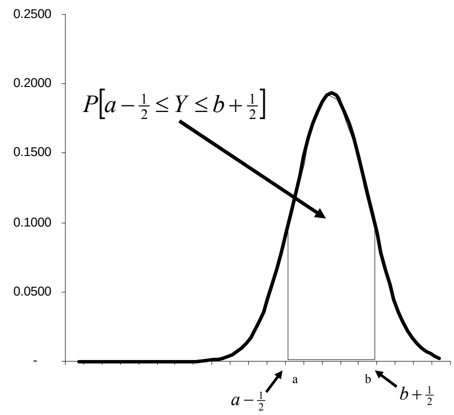
= 0.4052 - 0.2327 = 0.1725

```
> dbinom(13,20,0.7)
[1] 0.164262
```

Compare with 0.1643

```
> pnorm(13+1/2,20*0.7,sqrt(20*0.7*(1-0.7)))-
pnorm(13-1/2,20*0.7,sqrt(20*0.7*(1-0.7)))
[1] 0.1715179
```



Example

- X has a Binomial distribution with parameters $n = 20$ and $p = 0.70$

We want $P[11 \le X \le 14]$

The exact value $P[11 \le X \le 14]$

$= p(11) + p(12) + p(13) + p(14)$

$$= \binom{20}{11}(0.70)^{11}(0.30)^{9} + \cdots + \binom{20}{14}(0.70)^{14}(0.30)^{6}$$

$= 0.0654 + 0.1144 + 0.1643 + 0.1916 = 0.5357$

Using the Normal approximation to the Binomial distribution

$$P[11 \le X \le 14] \approx P[10\tfrac{1}{2} \le Y \le 14\tfrac{1}{2}]$$

Where Y has a Normal distribution with:

$\mu = np = 20(0.70) = 14$

$\sigma = \sqrt{npq} = \sqrt{20(.70)(.30)} = 2.049$


```
>pnorm(14+1/2,20*0.7,sqrt(20*0.7*(1-0.7)))-
pnorm(11-1/2,20*0.7,sqrt(20*0.7*(1-0.7)))
[1] 0.5525405
```

Hence

$P[10.5 \le Y \le 14.5]$

$$= P\left[\frac{10.5-14}{2.049} \le \frac{Y-14}{2.049} \le \frac{14.5-14}{2.049}\right]$$

$= P[-1.71 \le Z \le 0.24]$

= 0.5948 - 0.0436 = 0.5512

Compare with 0.5357

Comment:

- **The accuracy of the normal appoximation to the binomial increases with increasing values of *n***

Normal Approximation to the Poisson Distributions

- If X is a Poisson random variable with E(X) = λ and V(X) = λ,

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

is approximately a standard normal random variable. The approximation is good for λ > 5.

Normal Approximation to the Poisson Distribution

A Poisson distribution with parameter : is approximated by a normal distribution with mean and variance both equal to μ.

Pr(X=x) is approximated by

1) **the area under an N(μ, μ) density from x-1/2 to x+1/2 for x>0**
2) **the area to the left of 1/2 for x = 0.**

This approximation is used for μ ≥ 10.

Compute $\Pr(X \geq 20) = \Pr(Y \geq 19.5)$
where Y is distributed N(λA, λA) = N(10, 10).

$$\Pr(Y \geq 19.5) = 1 - \Pr(Y \leq 19.5) = 1 - \Phi\left[\frac{19.5 - 10}{\sqrt{10}}\right]$$

$$= 1 - \Phi\left[\frac{9.5}{\sqrt{10}}\right] = 1 - \Phi(3.00)$$

$$= 1 - .9987 = 0.0013$$

```
> 1-pnorm(19.5,10,sqrt(10))
[1] 0.001331560
```


Example: Bacteriology

- **Consider the distribution of the number of bacteria in a petri plate of area A.**
- **Assume that the probability of observing x bacteria is given exactly by a Poisson distribution with parameter μ = λA, where λ = 0.1 and A = 100 cm^2 [μ = σ^2 = .1 (100) = 10].**
- **Suppose 20 bacteria are observed in this area. How unusual is that event.**

Linear Combination of Random Variables

Linear Combinations
of Random Variables

A <u>linear combination</u> L of the random variables $X_1, X_2, \dots, X_n$ is defined as any function of the form $L = c_1X_1 + c_2X_2 + \dots + c_nX_n$ where $c_1, c_2, \dots, c_n$ are any fixed constants.

Results:
Suppose that L is the linear combination of random variables $X_1, X_2, \dots, X_n$ given by

$$L = \sum_{i=1}^{n} c_i X_i$$

Then the expected value of L is given by

$$E(L) = \sum_{i=1}^{n} c_i E(X_i),$$

and the variance of L is given by

$$Var(L) = \sum_{i=1}^{n} c_i^2 Var(X_i).$$

Result:

If $X_1, X_2, \dots, X_n$ are independent normal random variables with expected values $\mu_1, \mu_2, \dots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ and L is any linear combination

$$L = \sum_{i=1}^{n} c_i X_i \text{ then}$$

$$E(L) = \sum_{i=1}^{n} c_i \mu_i \text{ and}$$

$$Var(L) = \sum_{i=1}^{n} c_i^2 \sigma_i^2.$$

Example:

Suppose X_1 and X_2 represent serum-creatinine levels for two different individuals with end-stage renal disease. Then

1) The sum $L = X_1 + X_2$ is a linear combination with $c_1 = 1$ and $c_2 = 1$.
   a) The sum $X_1 + X_2$ will be normally distributed.
   b) The mean of the sum will be $\mu_1 + \mu_2$ and
   c) the variance will be $\sigma_1^2 + \sigma_2^2$

2) The difference $L = X_1 - X_2$ is also a linear combination with $c_1 = 1$ and $c_2 = -1$.
   a) The difference will be normally distributed.
   b) The mean of the difference will be $\mu_1 - \mu_2$ and
   c) the variance will be $\sigma_1^2 + \sigma_2^2$.

3) The average $L = (X_1 + X_2)/2$ is also a linear combination with $c_1 = 0.5$ and $c_2 = 0.5$.
   a) The average will be normally distributed.
   b) The mean of the average will be

      $0.5\mu_1 + 0.5\mu_2$.

   c) the variance will be

      $.25\sigma_1^2 + .25\sigma_2^2$.

4) **If the mean for both measures is 1.4 and the standard deviation is 0.5 then the average of the two serum-creatinine levels will be**

   i) **normally distributed**

   ii) **the mean will be (1.4 + 1.4)/2 = 1.4**

   iii) **and the variance of the mean will be**

   $$\text{var(average)} = 0.5^2\ \text{Var}(X_1) + 0.5^2\ \text{Var}(X_2)$$
   $$= 0.25(0.25) + 0.25(0.25)$$
   $$= 0.125$$

   iv) **Therefore the average is N(1.4, 0.125).**

5) **the mean of a sample has the same mean as the population but it has a smaller variance.**