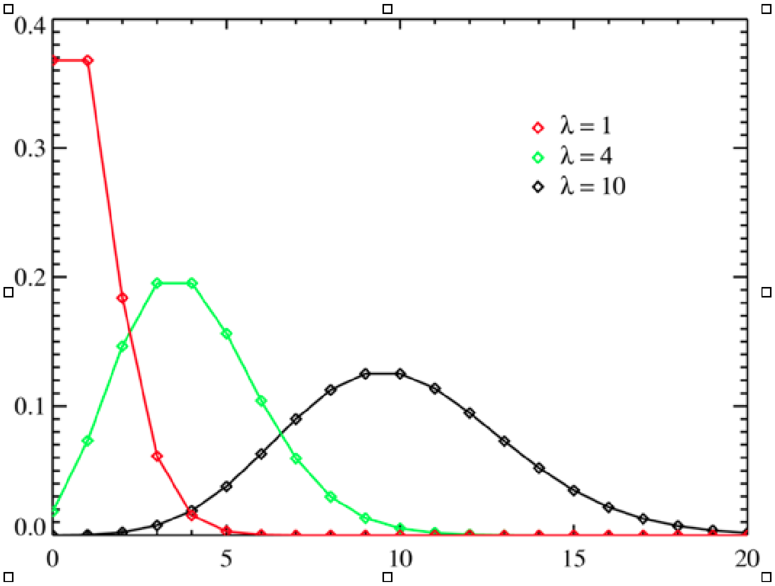
Discrete probability distributions - Class 5

January 20, 2014 Debdeep Pati

Poisson distribution

1. Poisson distribution is widely used in statistics for modeling rare events.
2. Ex. Infectious Disease The number of deaths attributed to typhoid fever over a long period of time, for example, 1 year, follow a Poisson distribution if:
   (a) The probability of a new death from typhoid fever in any one day is very small.
   (b) The number of cases reported in any two distinct periods of time are independent random variables.
3. Ex. Rare events occurring on a surface area. The distribution of number of bacterial colonies growing on an agar plate. The number of bacterial colonies over the entire agar plate follow a Poisson distribution if we assume:
   (a) The probability of finding any bacterial colonies in a small area is very small.
   (b) The events of finding bacterial colonies at any two areas are independent.
4. The probability of k events in a time period t for a Poisson random variable with parameter μ is
   $$P(X = k) = e^{-\mu}\frac{\mu^k}{k!},$$
   where $\mu = \lambda t$.
5. Parameter λ represents expected number of events per unit time.
6. Parameter μ represents expected number of events over time period t.
7. Difference between Binomial and Poisson distribution
   (a) There are a finite number of trials n in Binomial distribution
   (b) The number of events can be infinite for Poisson distribution
8. $E(X) = \mu, Var(X) = \mu$.

Poisson Distribution



Infectious disease example

1. Ex. Infectious Disease Consider the typhoid-fever example. Suppose the number of deaths from typhoid fever over a 1-year period is Poisson distributed with parameter $\mu = 4.6$. What is the probability distribution of the number of deaths over a 6-month period?

   Let X be the number of deaths in 6 months. Because $\mu = 4.6, t = 1\,\text{year}$, it follows that $\lambda = 4.6$ deaths per year. For a 6-month period we have $\lambda = 4.6$ deaths per year, t = .5 year. $\mu = 2.3$.

   $$P(X = k) = e^{-2.3}\frac{2.3^k}{k!}$$

   and $P(X > 5) = 1 - (P(X = 0 + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4 + P(X = 5)) = 0.030$

2. Ex. If $A = 100\,cm^2$ and $\lambda = .02$ colonies per cm^2, calculate the probability distribution of the number of bacterial colonies. We have $\mu = \lambda A = 100(0.2) = 2$. Let X = number of colonies. Then

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

$P(X > 5) = 0.053$.

3. Ex. Infectious Disease: The number of deaths attributable to polio during the years 1968-1976 is given in the following table. Based on this data set, can we use Poisson distribution to model the number of deaths from polio? The sample mean is 11.3 and the variance is 51.5. The Poisson distribution will not fit here because the mean and variance are too different.

Table 1: default

Year	68	69	70	71	72	73	74	75	76
# deaths	24	13	7	18	2	10	3	9	16

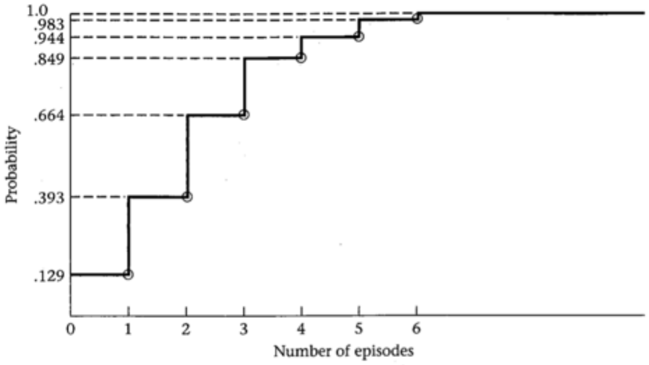
Cumulative distribution function

1. Every random variable X also has an associated cumulative distribution function defined on the real numbers by $F(x) = P(X \leq x)$. ?

2. Every cdf is an increasing function. Its limit at negative infinity (to the left) is 0 and its limit at positive infinity (to the right) is 1.

3. Once you know the cdf, you can easily find almost any probability that interests you. If the random variable X is discrete, then the cdf is a step function.

Figure 4.2 Cumulative-distribution function for the number of episodes of otitis media in the first 2 years of life



$F(x) = 0$ if $x < 0$
$F(x) = .129$ if $0 \le x < 1$
$F(x) = .393$ if $1 \le x < 2$
$F(x) = .664$ if $2 \le x < 3$
$F(x) = .849$ if $3 \le x < 4$
$F(x) = .944$ if $4 \le x < 5$
$F(x) = .983$ if $5 \le x < 6$
$F(x) = 1.0$ if $x \ge 6$

r	**0**	**1**	**2**	**3**	**4**	**5**	**6**
***P*(*X*=*r*)**	**.129**	**.264**	**.271**	**.185**	**.095**	**.039**	**.017**[38]

Normal Random Variables

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

- The distribution associated with Normal random variable is called Normal distribution.
- Carl Friedrich Gauss analyzed astronomical data using Normal distribution and defined the equation of its probability density function.
  - The distribution is also called Gaussian distribution.

Importance of Normal Distribution

- Describes many random processes of continuous phenomena
  - Height of a man, velocity of a molecule in gas, error made in measuring a physical quantity.
- Can be used to approximate discrete probability distributions
  - Binomial and Poisson
- Basis for classical statistical inference
  - Central limit theorem.

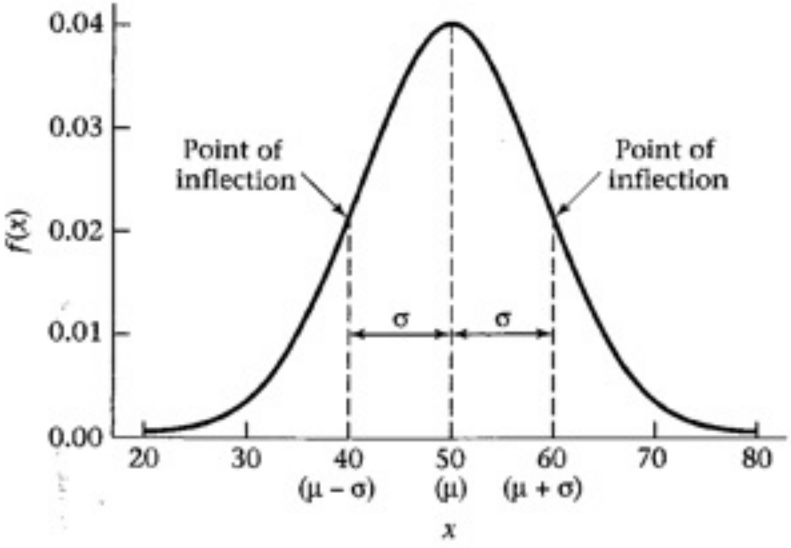
Importance of Normal Distribution

- Many distributions that are not themselves normal can be made approximately normal by transforming the data onto a different scale.
  - The distribution of serum-triglyceride concentrations is positively skewed. The log transformation of these measurements usually follows a normal distribution.
- Generally, any random variables that can be expressed as a sum of many other random variables can be well approximated by a normal distribution.
  - Many physiologic measures are determined by a combination of several genetic and environmental risk factors and can often be well approximated by a normal distribution.

Normal Distribution

- Bell-shaped and symmetrical
- Random variable has infinite range
- Mean measures the center of the distribution
- Standard deviation measures the spread of the distribution
- Normal distribution with parameter μ and σ is often written as $N(\mu, \sigma^2)$.

Figure 5.5 **Probability-density function for a normal distribution with mean μ (50) and variance σ² (100)**

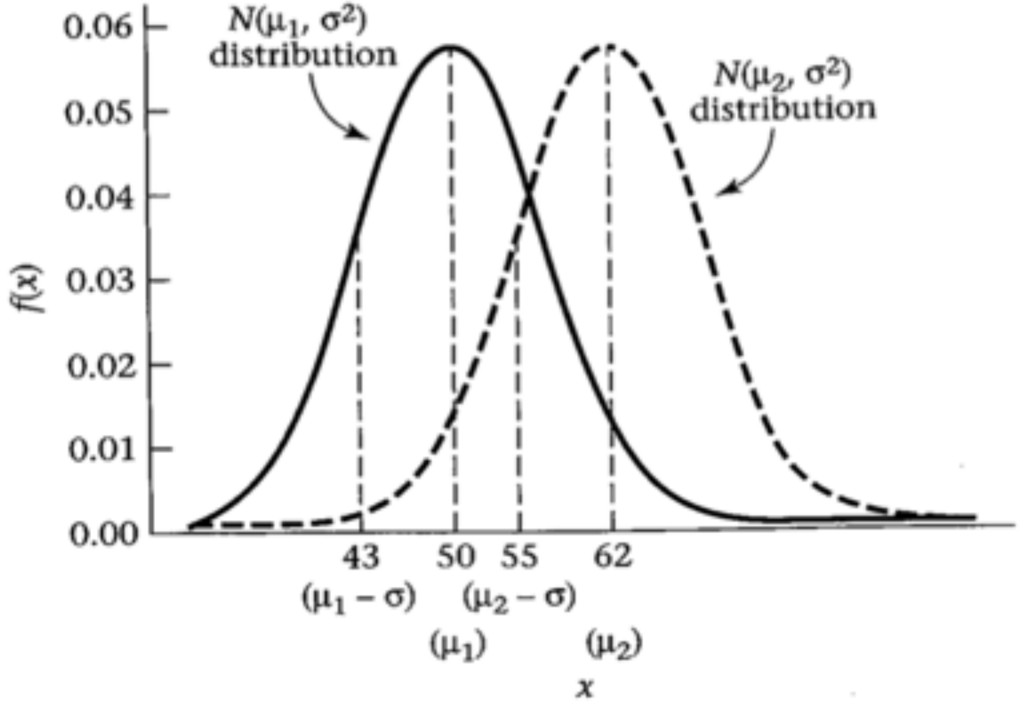


$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

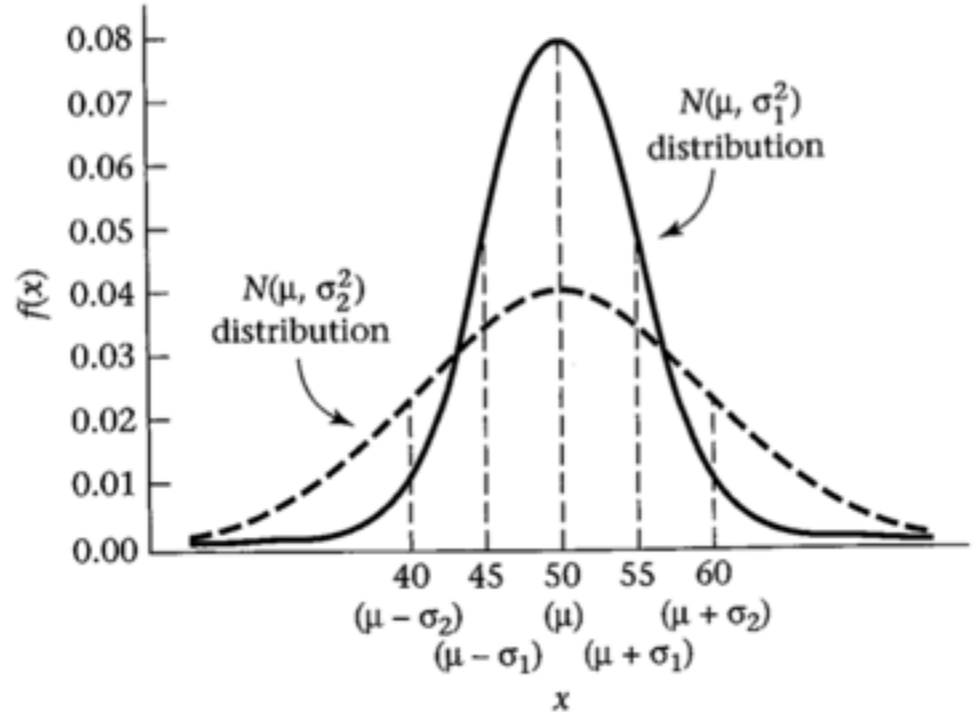
μ = Mean of random variable x

σ = Standard deviation

Comparison of two normal distributions with the same varian



Comparison of two normal distributions with the same mear

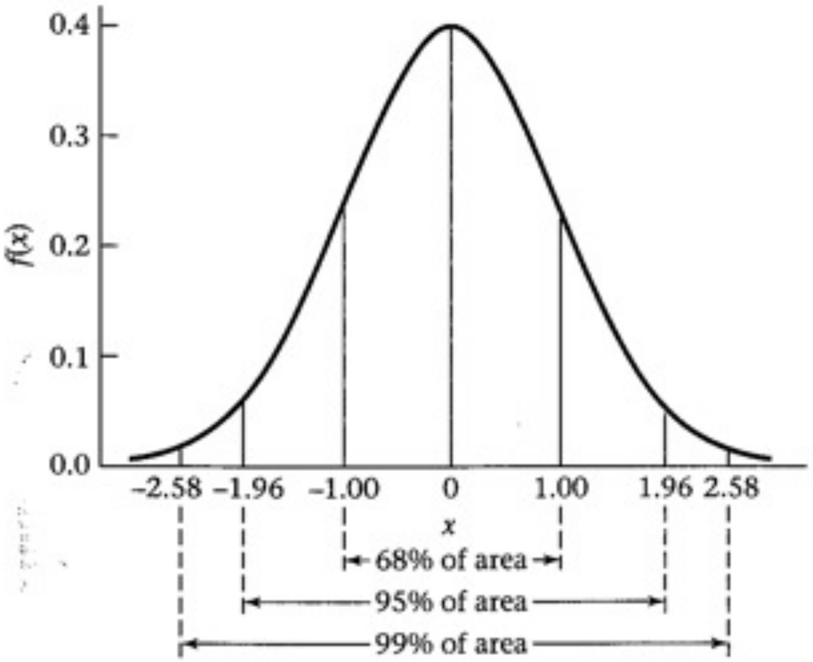


Standard Normal

- Normal distribution with parameter (0, 1), *N*(0,1), is also called standard normal.

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

Figure 5.9 **Empirical properties of the standard normal distribution**



- Ex. If $X \sim N(0,1)$, then find $P(X \leq 1.96)$ and $P(X \leq 1)$
- We can do it in R using pnorm().
- pnorm(1.96), pnorm(1)
- pnorm(1,1,1)

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE,
   log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE,
   log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

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

Compute $P(-1 \le X \le 1.5)$ if $X \sim N(0, 1)$

pnorm(1.5) – pnorm(-1)

- Ex. Pulmonary Disease
- Forced vital capacity (FVC)
  - a standard measure of pulmonary function based on the volume of air a person can expel in 6 seconds.
  - Current research looks at potential risk factors, such as cigarette smoking, air pollution, indoor allergies, or the type of stove used in the home, that may affect FVC in grade-school children.
  - It is known that age, sex, and height affect pulmonary function.
  - How can these variables be corrected for before looking at other risk factors?

- One way to make these adjustments for a particular child is to find the mean μ and standard deviation σ for children of the same age, sex, and height from large national surveys and compute a standardized FVC, which is defined as $(X- \mu)/\sigma$, where X is the original FVC. The standardized FVC then approximately follows an $N(0,1)$ distribution, if the distribution of the original FVC values was Normal.
- Suppose a child is considered in poor pulmonary health if his or her standardized FVC < -1.5. What percentage of children are in poor pulmonary health?
- $P(X < -1.5)$ = pnorm(-1.5) = .0668

- Ex. Pulmonary Disease
- Suppose a child is considered to have normal lung growth if his or her standard FVC is within 1.5 standard deviation of the mean. What proportion of children are within the normal range?

- Ex. Pulmonary Disease
- Suppose a child is considered to have normal lung growth if his or her standard FVC is within 1.5 standard deviation of the mean. What proportion of children are within the normal range?

- $P(-1.5 < X < 1.5) = ?$

- pnorm(1.5) - pnorm(-1.5) = 0.866

Conversion from an $N(\mu, \sigma^2)$ to $N(0, 1)$

- If X is $N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma$ is standard normal.

If $X \sim N(\mu, \sigma^2)$ and $Z = (X - \mu)/\sigma$

then $P(a < X < b) = P(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}) = \Phi[(b-\mu)/\sigma] - \Phi[(a-\mu)/\sigma]$

Because the Φ function, which is the cumulative distribution function for a standard normal distribution, is given in column A of Table 3 of the Appendix, probabilities for any normal distribution can be evaluated using the tables.

- Ex. Botany
- Suppose tree diameters of a certain species of tree from some defined forest area are assumed to be normally distributed with mean 8 in. and standard deviation 2 in. Find the probability of a tree having an unusually large diameter, which is defined as > 12 in.
- We have $X \sim N(8, 4)$ and require

  $P(X > 12) = 1 - P(X < 12) = 1 - P(Z < (12-8)/2)$

  $= 1 - P(Z < 2.0) = 1 - .977 = .023$

  In R: 1 - pnorm(12, mean=8, sd=2)

- Ex. Cerebrovascular Disease
- Diagnosing stroke strictly on the basis of clinical symptoms is difficult.
  - A standard diagnostic test used in clinical medicine to detect stroke in patients is the angiogram. This test has some risks for the patient, and researchers have developed several noninvasive techniques that they hope will be as effective as the angiogram.
  - One such method uses measurement of cerebral blood flow (CBF) in the brain, because stroke patients tend to have lower CBF levels than normal.
- Assume that in the general population, CBF is normally distributed with mean 75 and standard deviation 17. A patient is classified as being at risk for stroke if his or her CBF is lower than 40.
- What proportion of normal patients will be mistakenly classified as being at risk for stroke?

- Let X be the random variable representing CBF. Then $X \sim N(75, 17^2) = N(75, 289)$.
- $P(X < 40) = ?$

  $P(X < 40) = P(Z < (40\text{-}75)/17) = P(Z < -2.06) = \Phi(-2.06)$

  $= 1\text{-}\ \Phi(2.06) = 1\text{-}.9803 = .020$