Biostatistics

Week #6 4/07/2020

Chapter 7 Theoretical Probability Distributions – part 2

Outline

- 7.1 Probability Distribution
- 7.2 The Binomial Distribution
- 7.3 The Poisson Distribution
- 7.4 The Normal Distribution
- 7.5 Z-score and Applications

7.3 The Poisson Distribution

Introduction

- Recall that in binomial distribution, we can compute the probability using the formula

$$P(X = x) = p(x) = C_x^n p^x (1 - p)^{n-x}$$

$$\text{where } C_x^n = \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

$N = 1, 2, 3, \ldots$ and $x = 0, 1, 2, \ldots, n$

- **Additionally, the mean value of *X* is *np*, and the variance of *X* is *np*(1–*p*).**

Cont’d

- When *n* is large, the computation for C(*n*,*x*) would be tedious.
- For the case that ***n* is large** and ***p* is small** (for example, we’d like to know among all the drivers in the US, the probability of one particular individual involves in a motor accident), the binomial distribution is well *approximated* by another theoretical probability distribution called the **Poisson distribution**.

Poisson Distribution

- The Poisson distribution is used to model discrete events that occur ***infrequently*** in time or space; hence is sometimes called **the distribution of *rare* events**.
- *Many disease or disaster-related events belong to this category.*
- Here *X* is said to have a Poisson distribution with parameter λ.

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

Here the parameter **λ is a constant that denotes the average number of occurrences of the event in an interval.**

Cont’d

- When *n* is large and *p* is small, **the mean and variance are approximately equal ($np(1-p)\approx np$), which is characteristic for Poisson distribution**.
- For example:
  - – The number of ambulances needed in a city for a given night.
  - – Number of particles emitted from a specified amount of radioactive material.
  - – The number of bacterial colonies (菌落) growing in a Petri dish (培養皿).

Case #1

- Assume the probability that a particular individual involved in a car accident each year is p=0.00024.
- For a population of n=10,000 people, the mean number of persons involved is **np=2.4**. Let’s use it to represent the parameter λ in the Poisson distribution formula.

$$P(X = x) = \frac{e^{-2.4} 2.4^{x}}{x!}$$

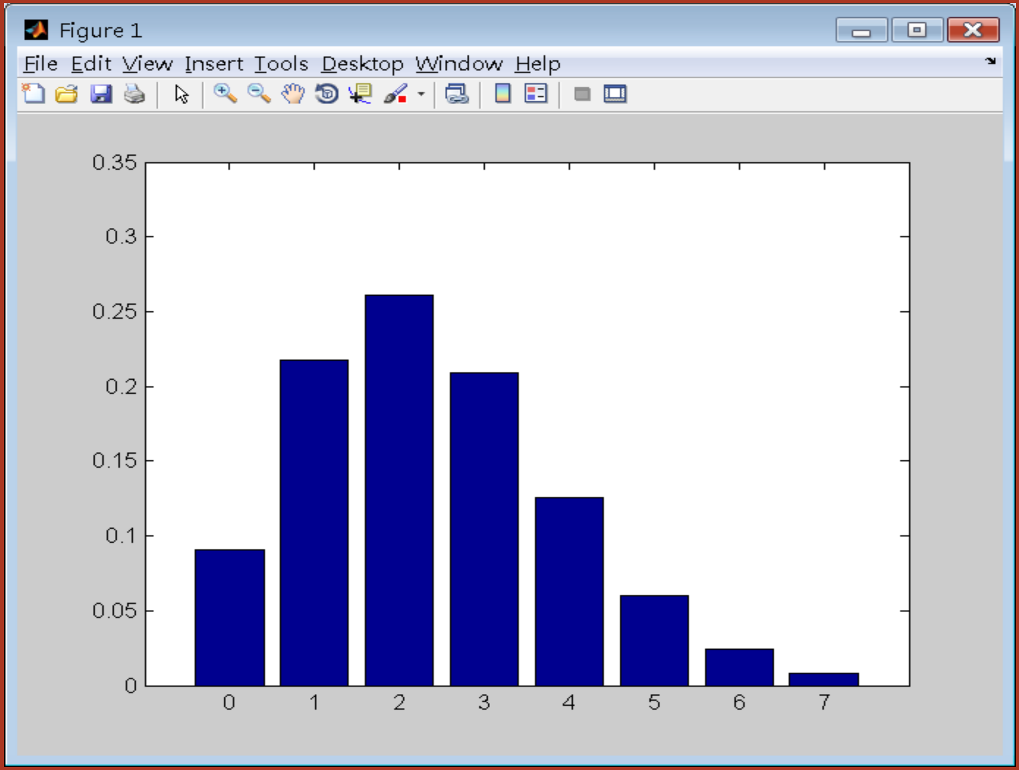
- P(X=0)=0.091. This is the probability that **no one** in this population will be involved in an accident.
- P(X=1)=0.218
- P(X=2)=0.261
- P(X=3)=0.209
- P(X=4)=0.125
- P(X=5)=0.060
- P(X=6)=0.024
- Since the outcomes of X are **mutually exclusive**, we may compute P(X=7+)=1–P(X<7)=0.012

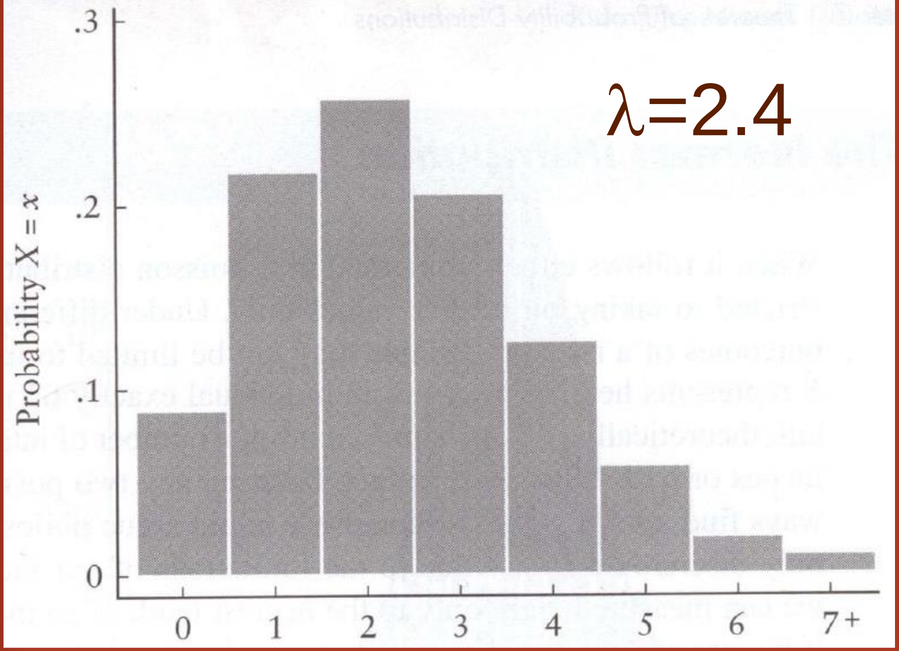
$$\mathrm{P}(X = x) = \frac{e^{-2.4}\,2.4^{x}}{x!}$$

Using MATLAB:

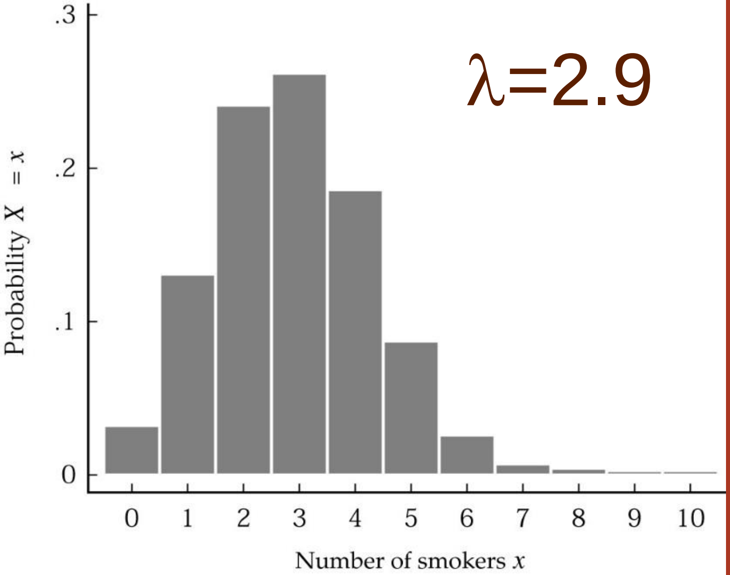
```
>> x=[0 1 2 3 4 5 6 7];
>> y=(exp(-2.4)*2.4.^x)./factorial(x)
ans = 0.0907    0.2177    0.2613    0.2090    0.1254
         0.0602    0.0241 0.0083
>> sum(y)
ans = 0.9967
>> bar(x,y)
```

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

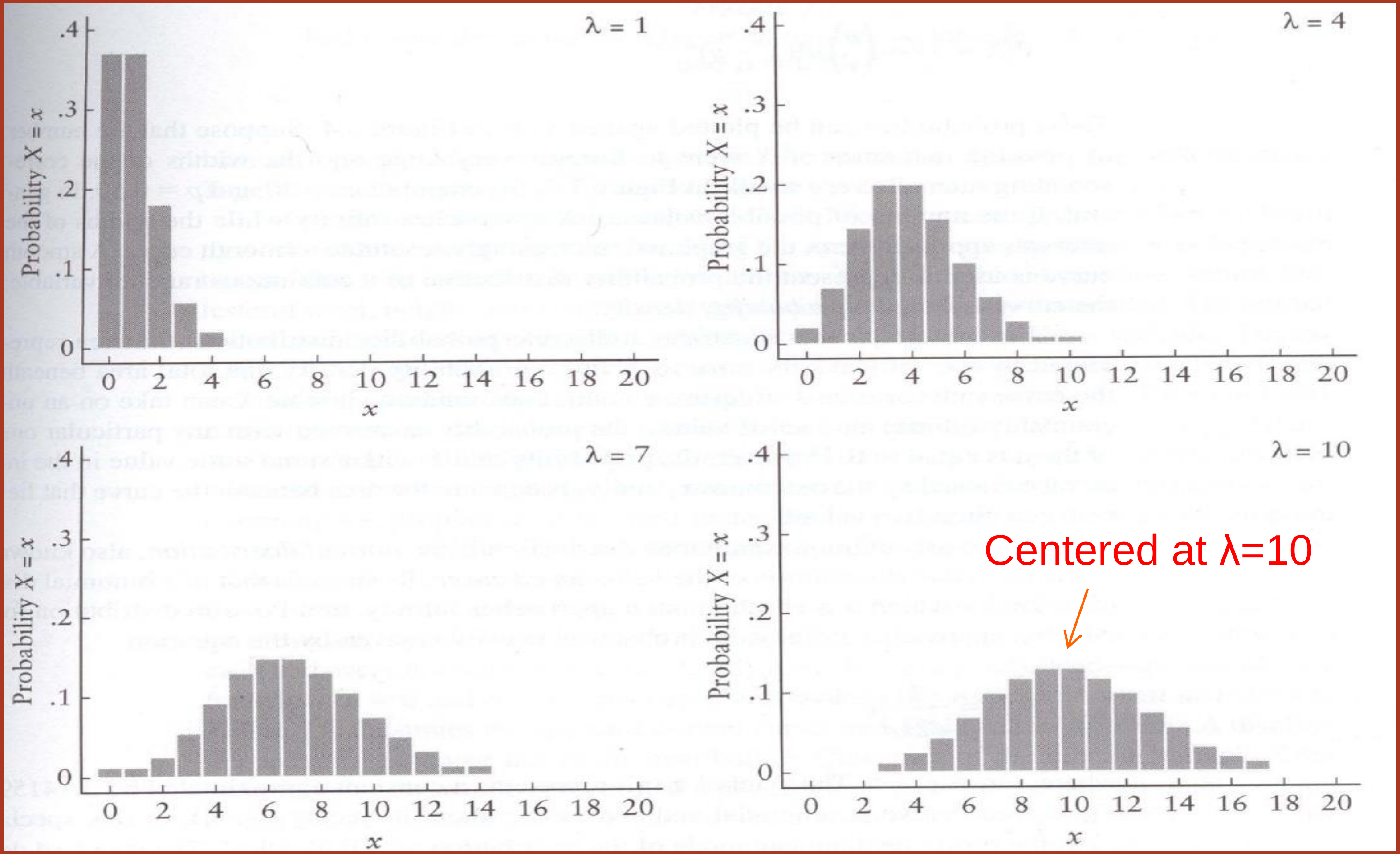




Number of persons *x* involved in an accident using Poisson distribution.



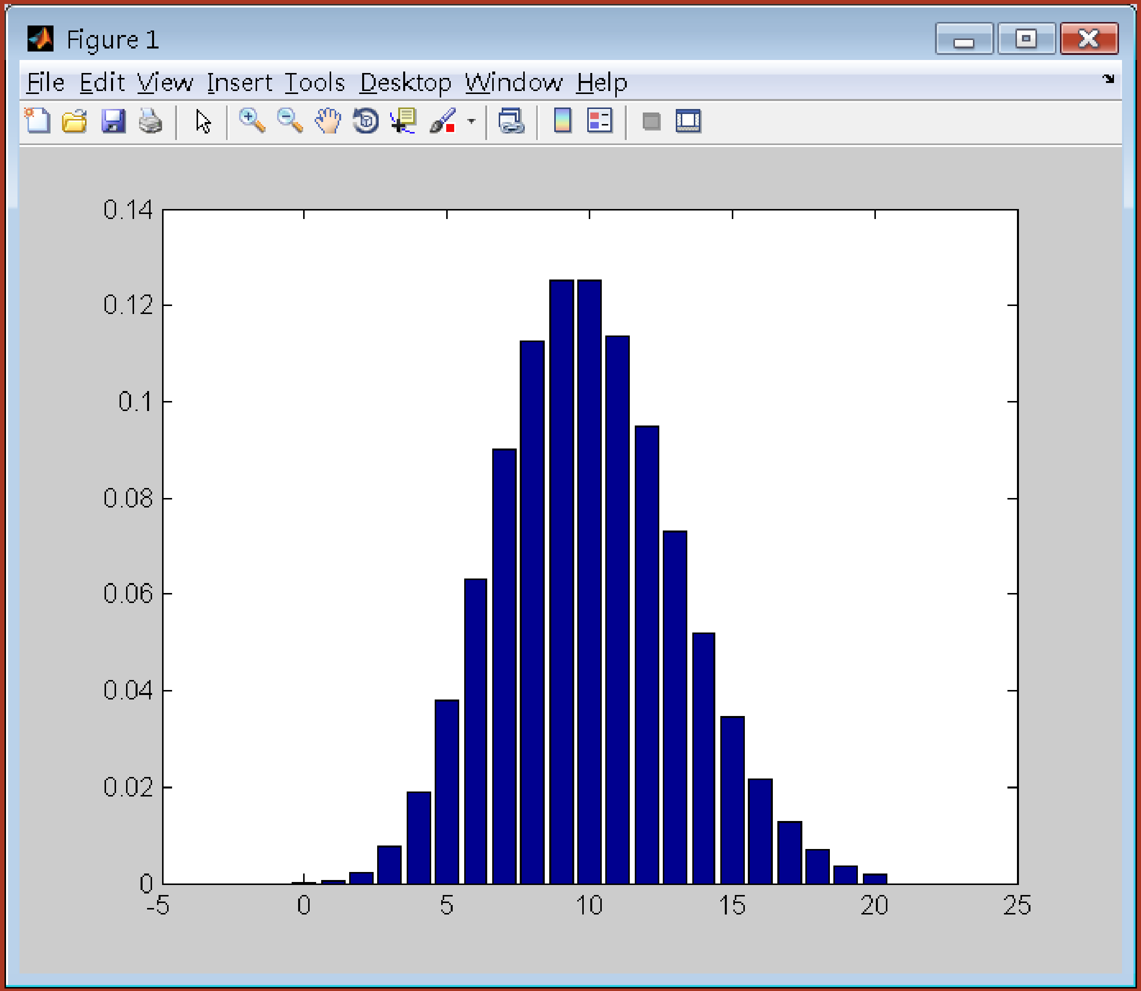
Compare the above figure with the binomial case (number of smokers x) in previous lecture, where we have $\lambda = np = 10\times0.29=2.9$. Although n=10 is not large, and p=0.29 is not small, you can see the resemblance of these two distributions.



Centered at λ=10

One can see that, for a Poisson distribution, the graph is highly skewed for small λ; as λ increases, the graph becomes more symmetric (advancing into larger x values, with top probability declines). **[Will it further skew to the right if we continue adding up the λ value?]**

```
>> x=[0:1:20];
>> y=(exp(-10)*10.^x)./factorial(x);
>> sum(y)
ans = 0.9984
>> bar(x,y)
>>
```



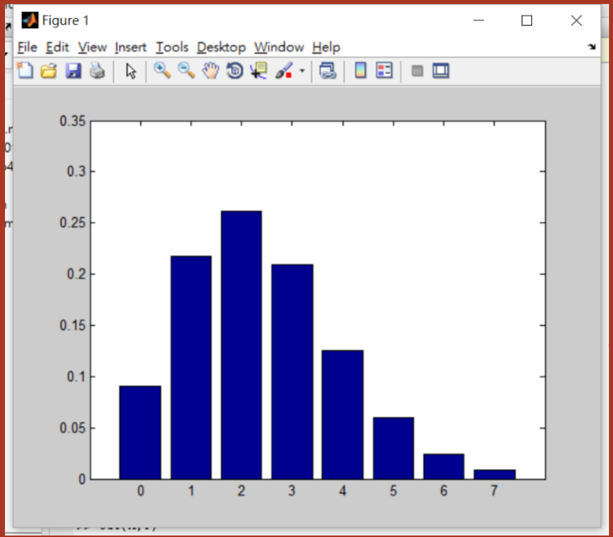
$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

>> help **poisspdf**

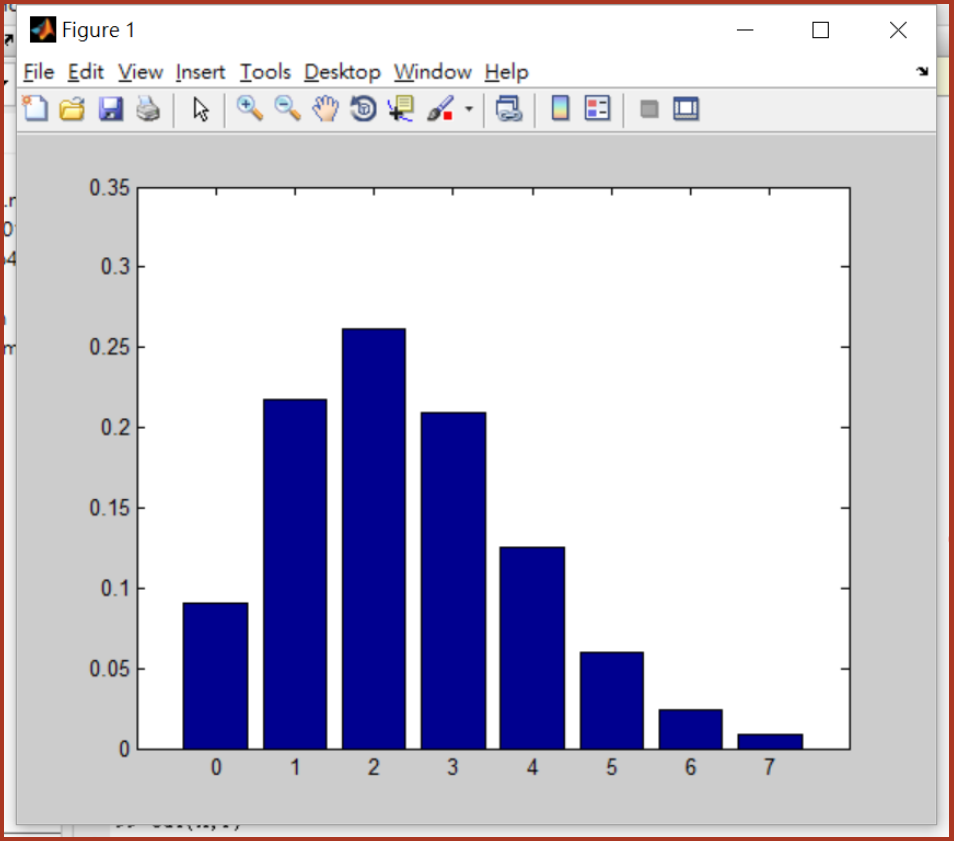
POISSPDF Poisson probability density function.

Y = POISSPDF(X,LAMBDA) returns the Poisson probability density function with parameter LAMBDA at the values in X.

```
>> X=0:7;
>> Y=poisspdf(X, 2.4)
Y =
   0.0907    0.2177    0.2613    0.2090    0.1254    0.0602
0.0241    0.0083
>> bar(X,Y)
```



```
>> help poisscdf
 POISSCDF Poisson cumulative distribution function.
   P = POISSCDF(X,LAMBDA) computes the Poisson
cumulative distribution function with parameter LAMBDA at the values
in X.
```



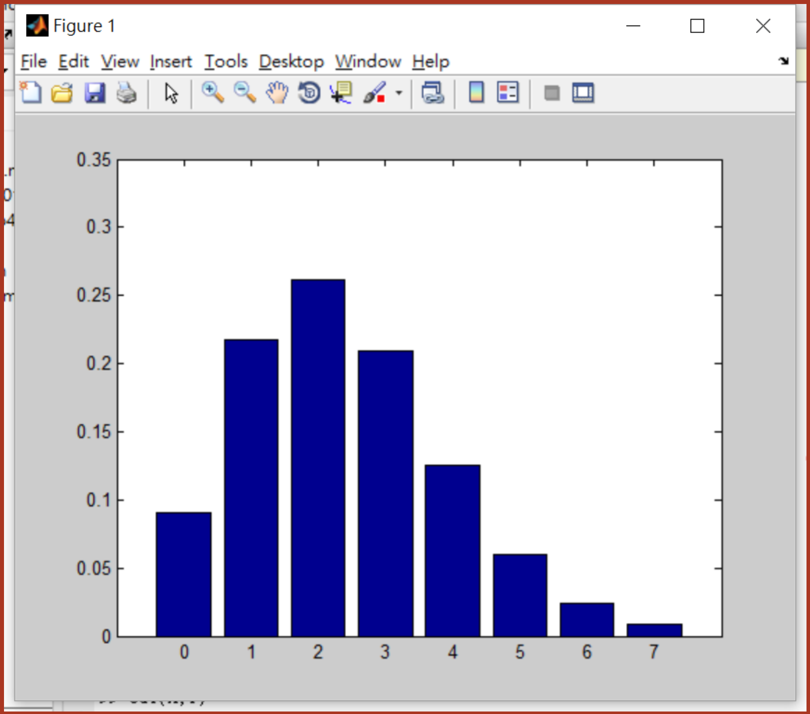
```
>> poisscdf(0,2.4)
ans =    0.0907
>> poisscdf(1,2.4)
ans =    0.3084
>> poisscdf(2,2.4)
ans =    0.5697
>> poisscdf(3,2.4)
ans =    0.7787
>>
```

>> help **poissinv**

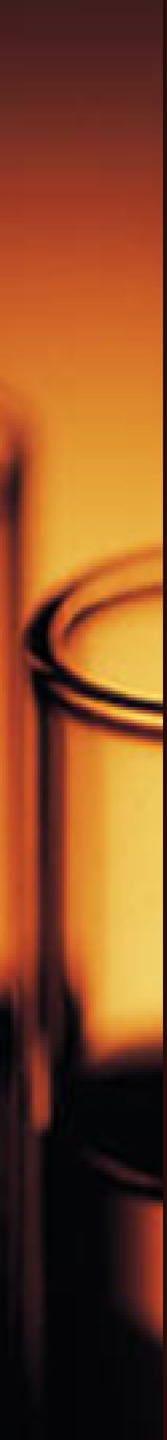
POISSINV Inverse of the Poisson cumulative distribution function (cdf).

X = POISSINV(P,LAMBDA) returns the inverse of the Poisson cdf with parameter lambda. Since the Poisson distribution is discrete, POISSINV returns the smallest value of X, such that the poisson cdf evaluated, at X, equals or exceeds P.

```
>> poissinv(0.5,2.4)
ans =     2
>> poissinv(0.75,2.4)
ans =     3
>> poissinv(0.9,2.4)
ans =     4
>>
```

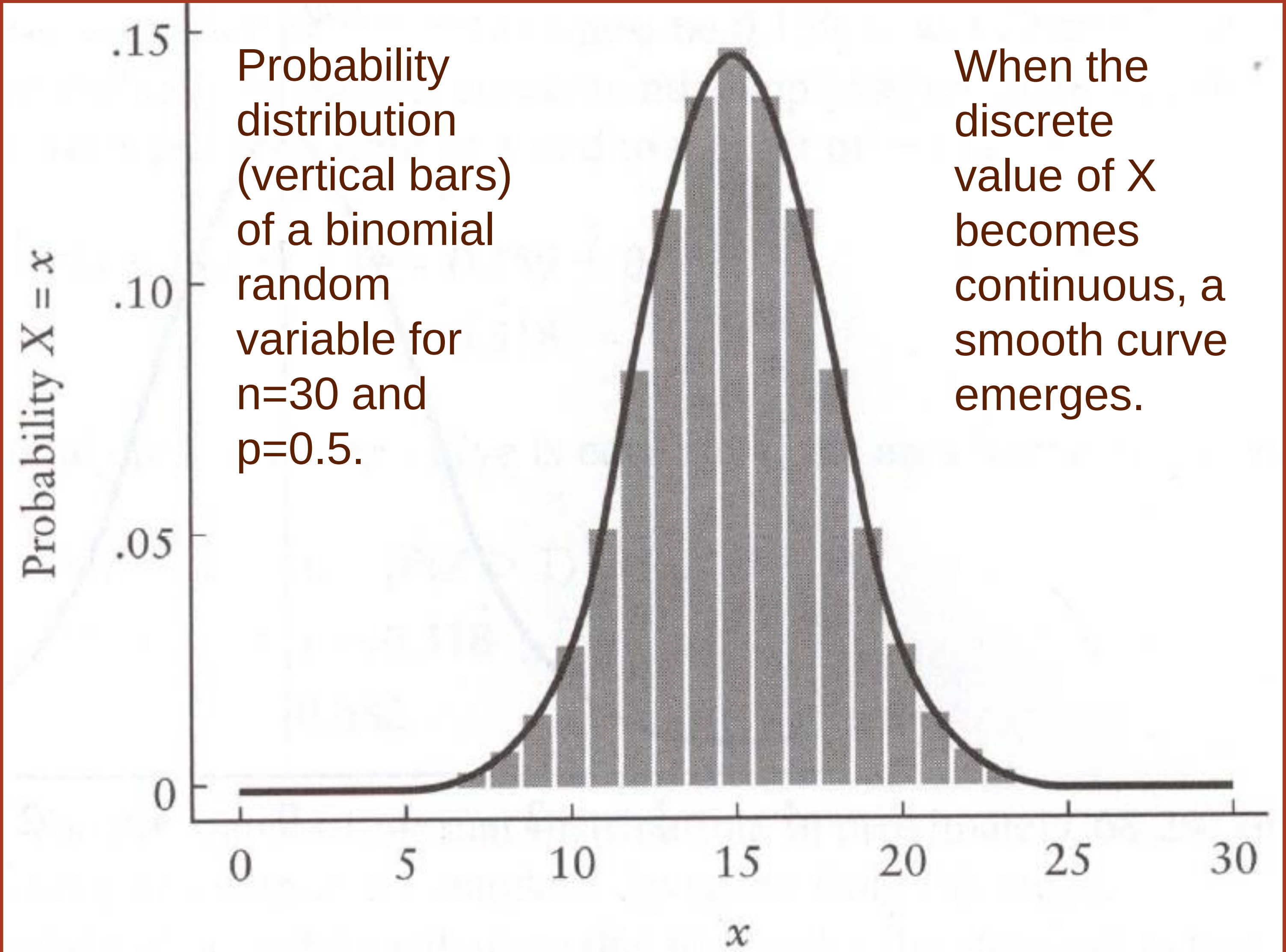


7.4 The Normal Distribution



Introduction

- In either binomial or Poisson distribution, the random variable X is **discrete**.
- When X becomes **continuous**, we then have **a smooth curve** to represent the probability distribution of this random variable; **this curve is called a probability density**.
- “Probability density” **IS NOT** “probability” for continuous random variables.



Probability distribution (vertical bars) of a binomial random variable for n=30 and p=0.5.

When the discrete value of X becomes continuous, a smooth curve emerges.

Normal Distribution

- Normal distribution (also known as **Gaussian distribution** or **bell-shaped** curve) is the most common continuous distribution.
- This is a binomial distribution of constant p and **infinite n**, or a Poisson distribution when λ approaches infinity. (Since λ=np, these two statements are practically the same.)

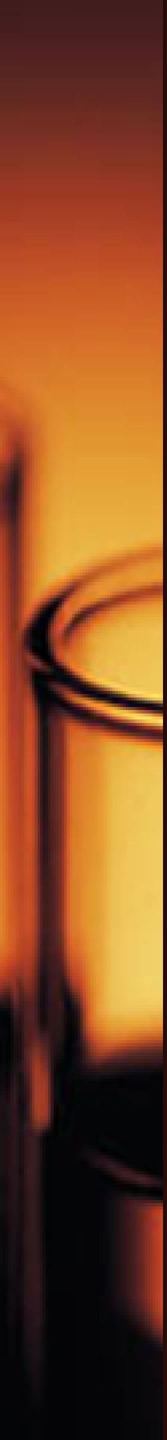
Cont’d

- Once the **mean value μ** (which is the same as the median and the mode of this graph) and the **standard deviation σ** is specified, the distribution can be represented by the following formula:

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

Cont'd

- Many random variables of interest can be approximated by normal distribution, which makes the estimation of probabilities associated with these variables relatively easy.



Cont'd

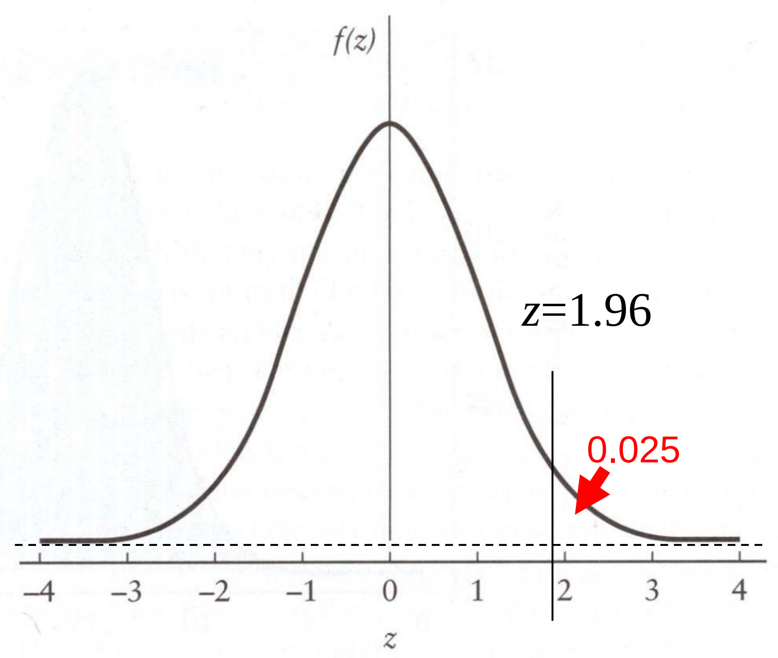
- For example, **once μ and σ is known for a cholesterol (膽固醇) level survey** (which is of normal distribution, likely), we may easily **estimate the probability of a random individual whose cholesterol is over 250**.
- Such knowledge may help us to plan for future cardiac (心臟科) services.

Standard Normal Distribution

- Since a normal distribution could have infinite number of possible values for μ and σ, it is impossible to draw graphs for all scenarios.
- Instead, only a single curve is kept (tabulated) – **the special case for which μ=0 and σ=1**, which is called the **standard normal distribution** (see next slide).

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

$$f(z)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}z^2)$$



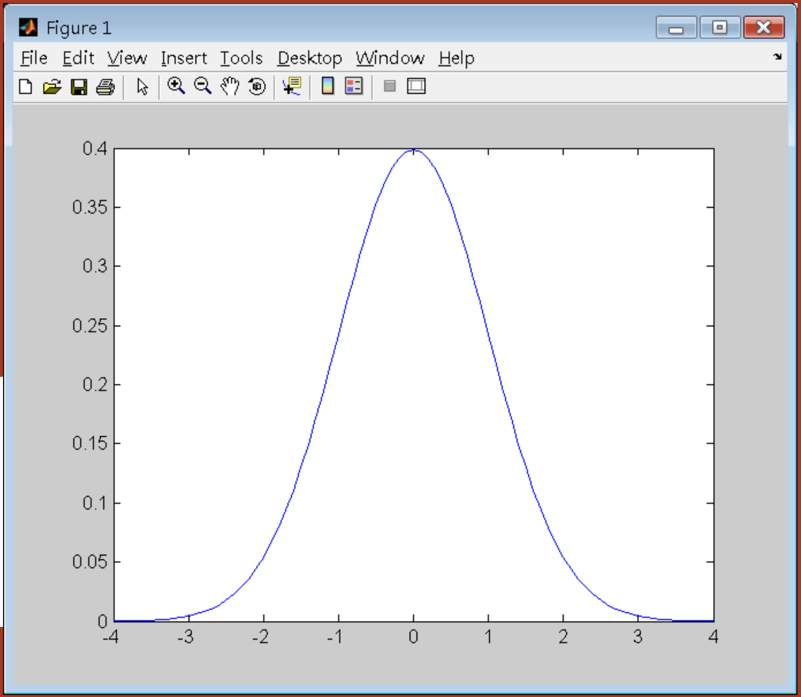
- With this, we can easily compute, for example, the area (recall that this area represent the probability for a specific random variable z, or a range of z) under this curve.

z	Area in the **right tail**
0.00	0.500
1.65	0.049
1.96	[1]**0.025**
2.58	[2]**0.005**
3.00	0.001

[1] A standard deviation of 1.96 would cover 95% of cases.

[2] A standard deviation of 2.58 would cover 99% of cases.

```
>> x=[-4:0.1:4];
>> y=1/(sqrt(2*pi))*exp(-0.5.*x.*x);
>> plot(x,y)
>>
```

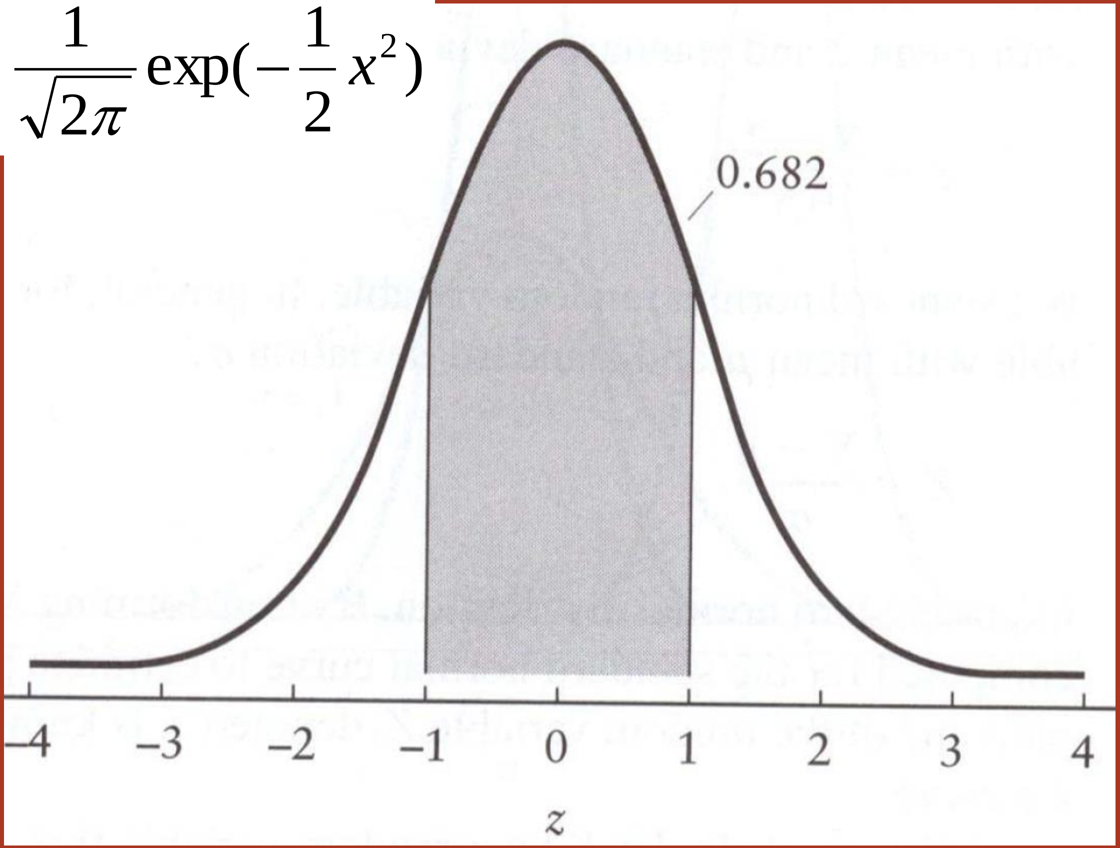


$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} x^2)$$

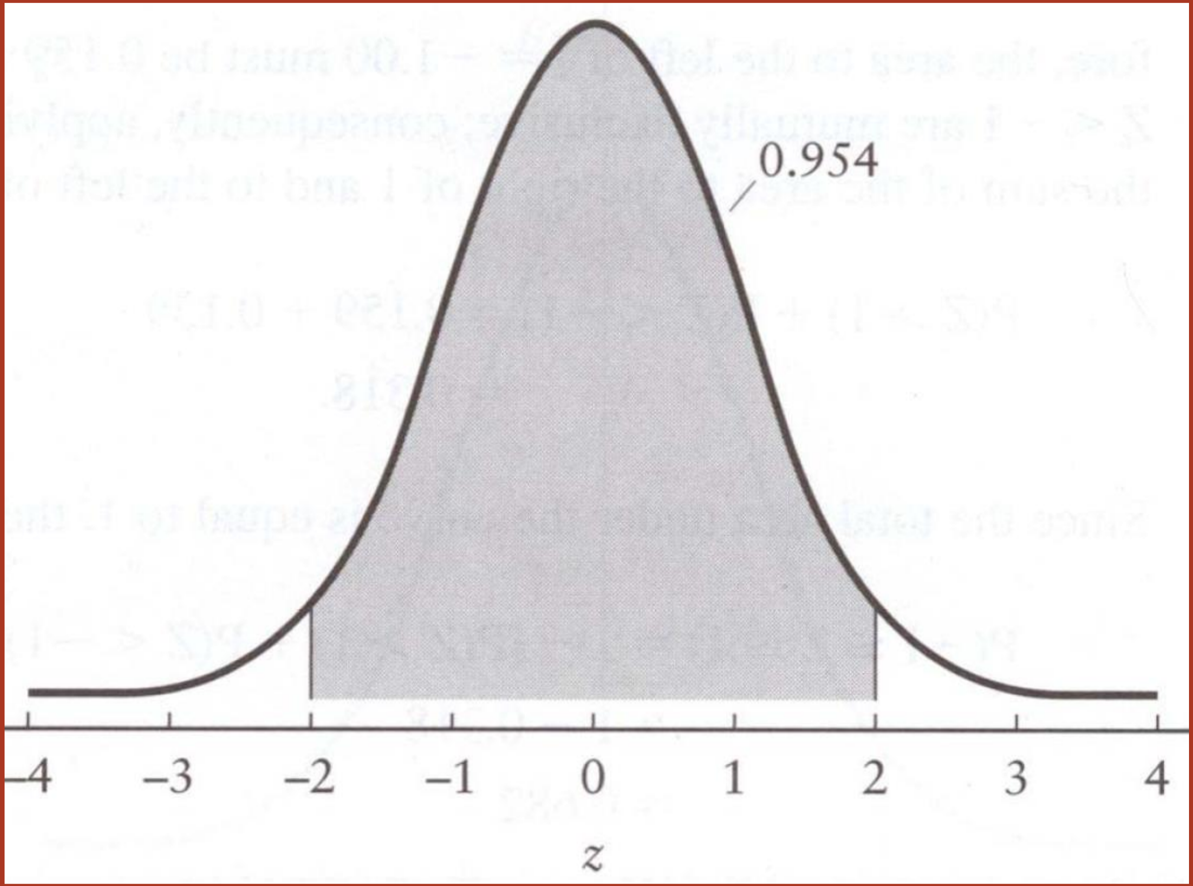
```
>> syms F(x);
>> F(x)='1/(sqrt(2*pi))*exp(-0.5*x*x)’
F(x) = (2^(1/2)*exp(-0.5*x^2))/(2*pi^(1/2))
>> int(F,-4,4)
ans = .99993665751633376015749245848656
>>
```

An integration from -4 to +4 gives almost 100% of the area under curve.

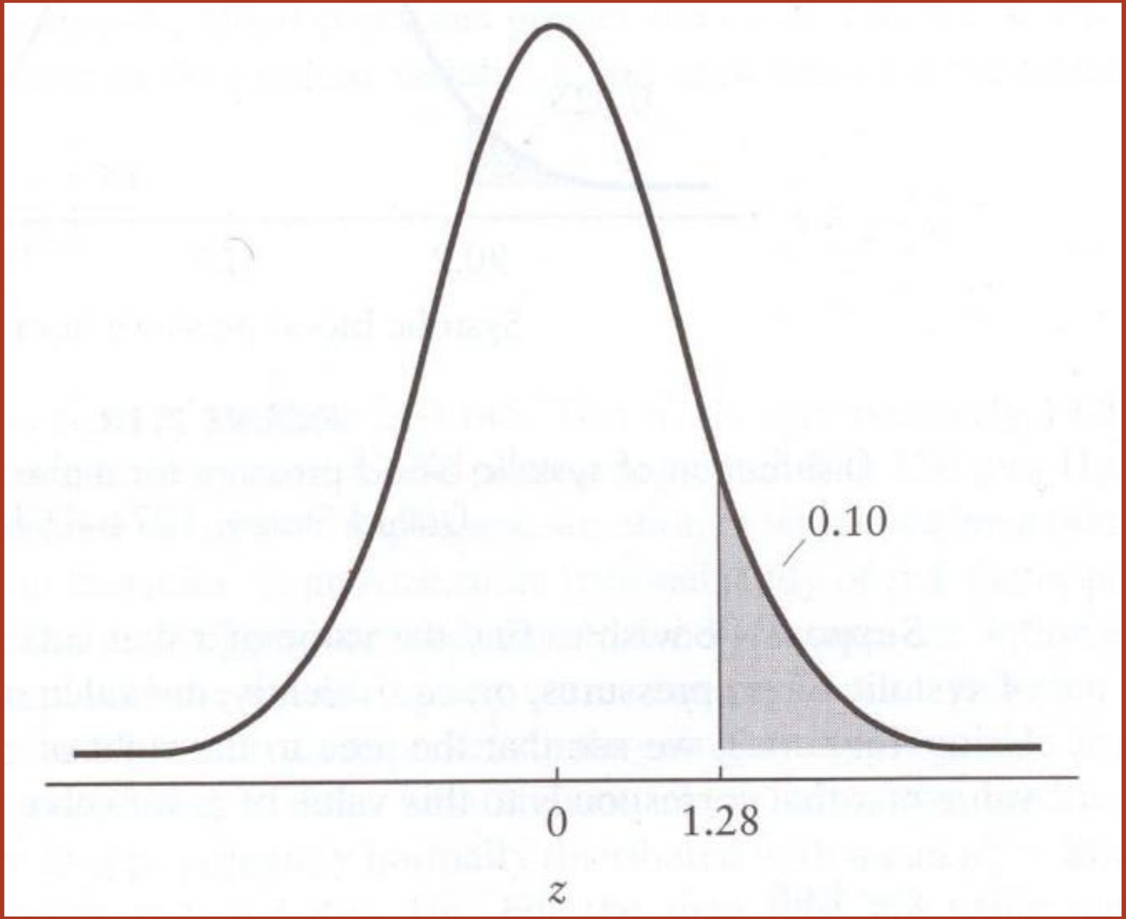
$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} x^2)$$



Or, performing an integration of this function f(x) from –1 to +1 would give you 0.682, meaning that **one STD away from mean would cover you 68.2% of cases.**



Or two STD’s away from mean would cover you 95.4% of cases.



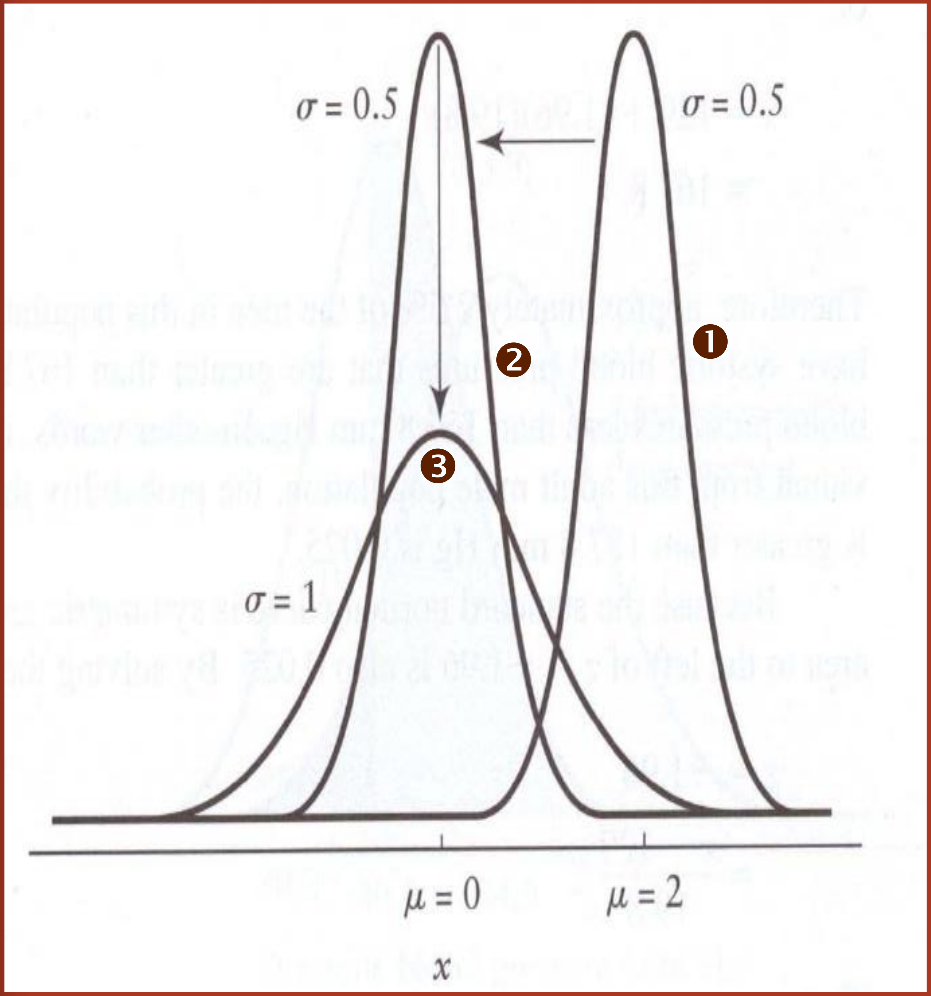
Or 1.28 STD’s away from mean would cover you 80% of cases (so you would have 10% on each of the two tails).

Z-score

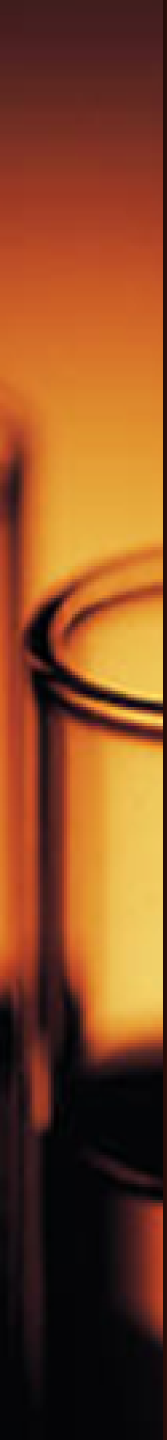
- A Z-score (also known as a standard normal **deviate**) is defined by the following formula.

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

- That is, we may convert the random variable of interest, X, to a new random variable Z, by subtracting each value of X from their mean value, and have it normalized (divided) by their standard deviation.



The effect is apparent by the following example. Suppose curve ❶ has mean at x=2 and σ=0.5. Subtracting the mean from x would shift graph ❶ to ❷. Further divided by σ would compress ❷ into ❸, which is a standard normal distribution.



Comments

- In old days when calculators were not that advanced as we see today, such conversion to a z-score distribution (thus a normalized standard distribution) is extremely important, since only the probabilities of a normalized standard distribution can be tabulated for hand calculation.

Example #1

- Let X be a random variable representing systolic blood pressure (收縮壓).
- From 18- to 74-yr-old males in the US, X is approximately normally distributed with mean 129 mm Hg and STD 19.8 mm Hg.
- From previous discussion, we know that Z will follow a standard normal distribution with the conversion:

$$Z = \frac{X - 129}{19.8}$$

Cont’d

$$f(z)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}z^2)$$

- Suppose we wish to find the value of x that cuts off the *upper* 2.5% of the curve.
- This is equivalent of considering the value of x for which P(X>x)=0.025.
- Assume this cut-off value is r in terms of Z-score. Then:

$$\int_0^r f(z)dz=0.5-0.025=0.475$$

Try taking z = 1.96:

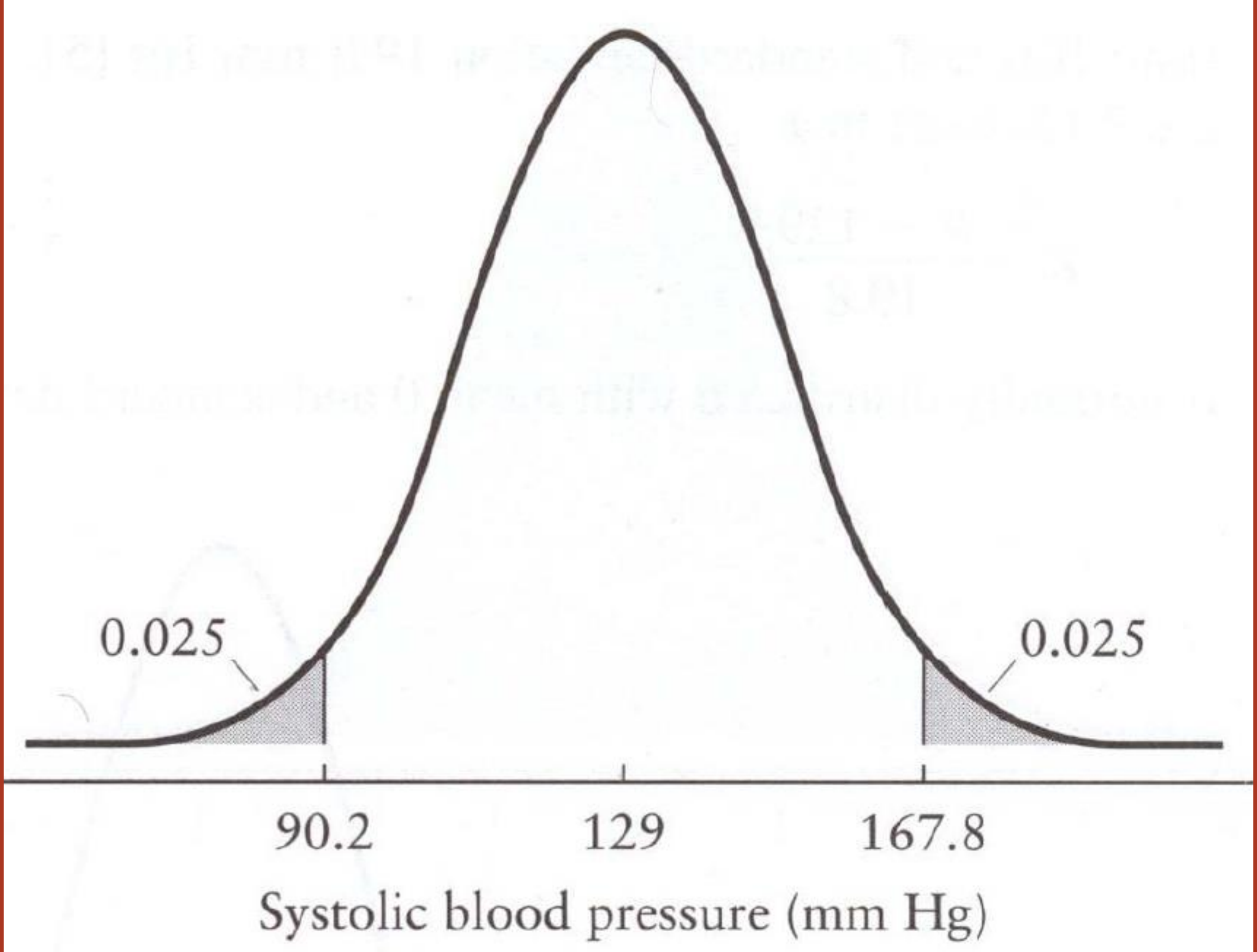
```
>> int(F,0,1.96)
ans =
.475002104
```

Cont’d

- So Z=**1.96** will give the area under the right tail be 0.025.
- Using the formula we had earlier, we may get the corresponding X = 129 + Z*19.8 = 129 + **1.96***19.8 =167.8
- That is, if we randomly select an adult male, the probability for his systolic pressure to go over 167.8 would be 0.025.

$$Z = \frac{X - 129}{19.8}$$

Similarly we may get the lower cut-off X = 129 + **1.96***(-19.8) = 90.2



Comments

- More applications using normal distribution will be discussed later.
- Similarly, using MATLAB functions for normal distribution (normpdf, normcdf, norminv, etc) will be discussed later.