

Biostatistics

Week #6 3/29/2022



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

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

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

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

- 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.*

Cont'd

- 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).

Example #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 $P(X=7+)=1-P(X<7)=0.012$

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

Using MATLAB:

```
>> x=[0 1 2 3 4 5 6 7];
```

```
>> y=(exp(-2.4)*2.4.^x)./factorial(x)
```

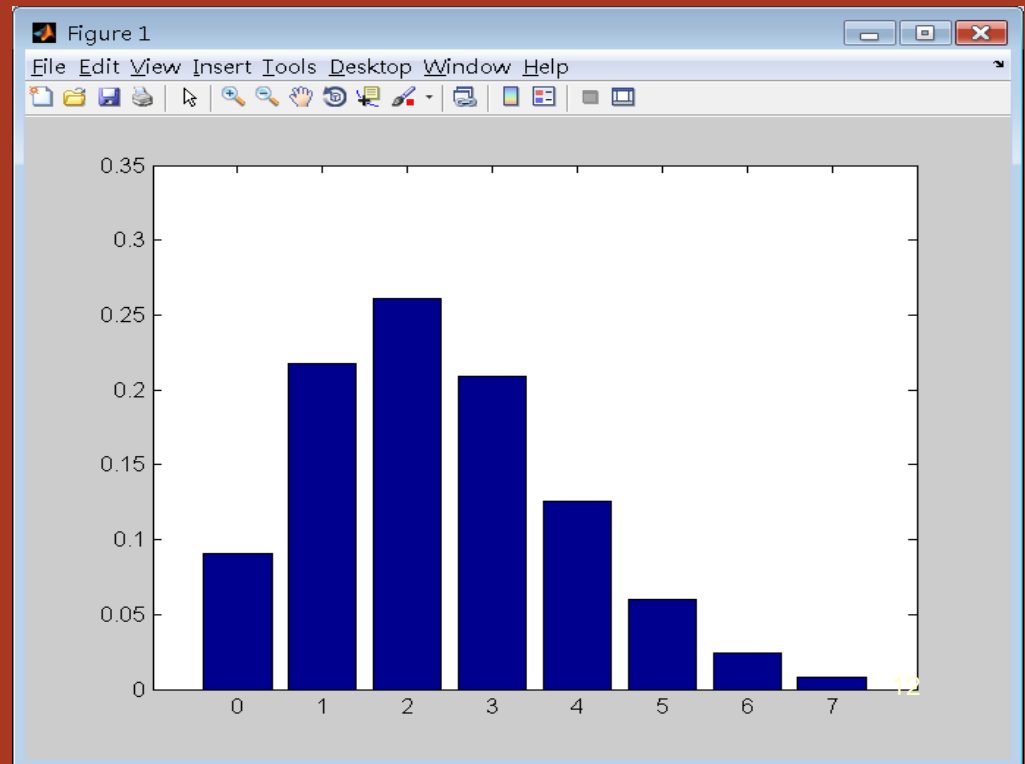
```
ans = 0.0907    0.2177    0.2613    0.2090    0.1254  
       0.0602    0.0241    0.0083
```

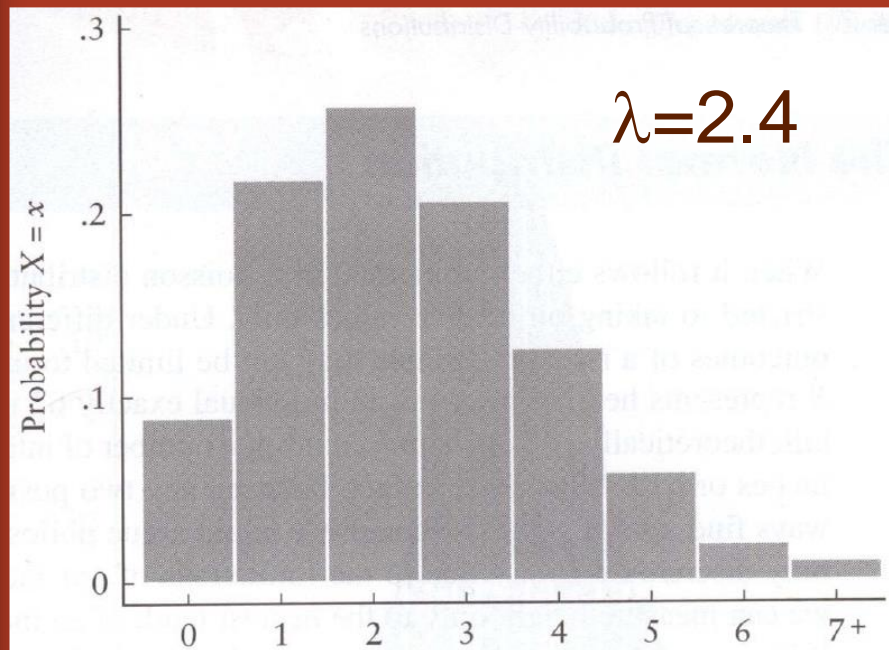
```
>> sum(y)
```

```
ans = 0.9967
```

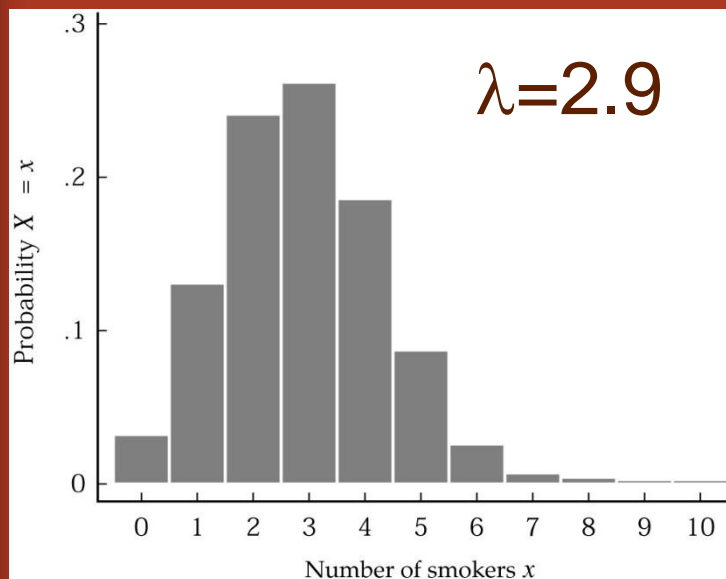
```
>> bar(x,y)
```

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

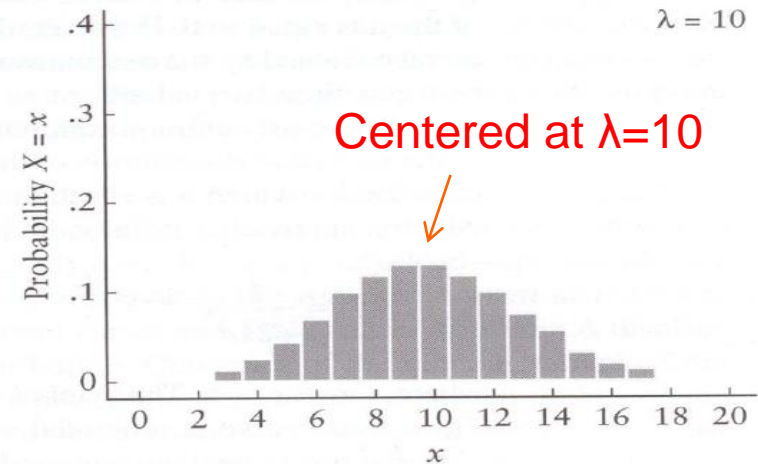
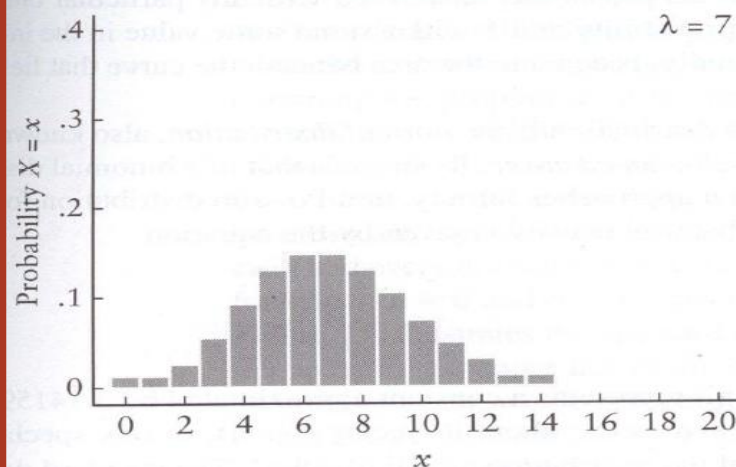
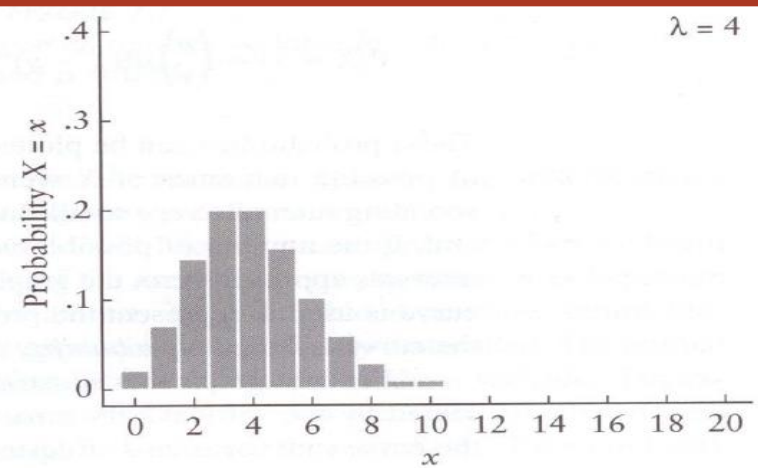
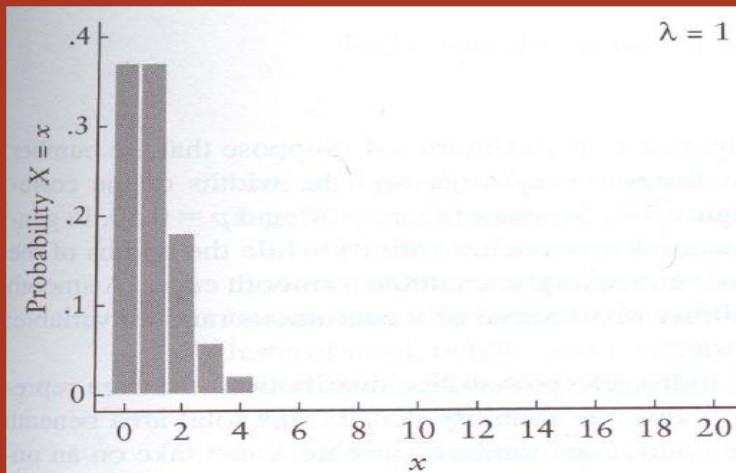




Number of persons x involved in an accident using Poisson distribution.



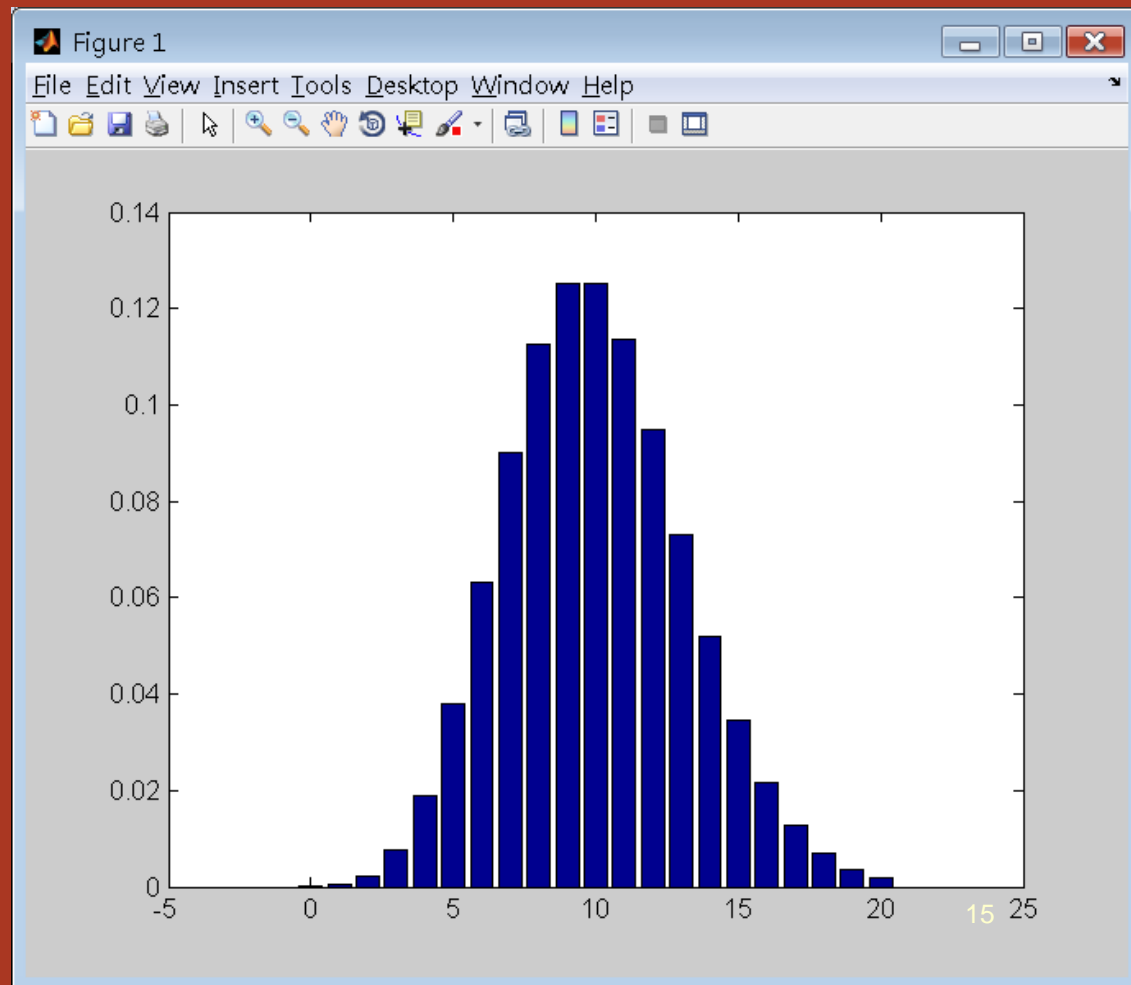
Compare the above figure with the binomial case (number of smokers x) in previous lecture, where we have $\lambda = np = 10 \times 0.29 = 2.9$. Although $n=10$ is not large, and $p=0.29$ is not small, you can see the resemblance of these two distributions.



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.

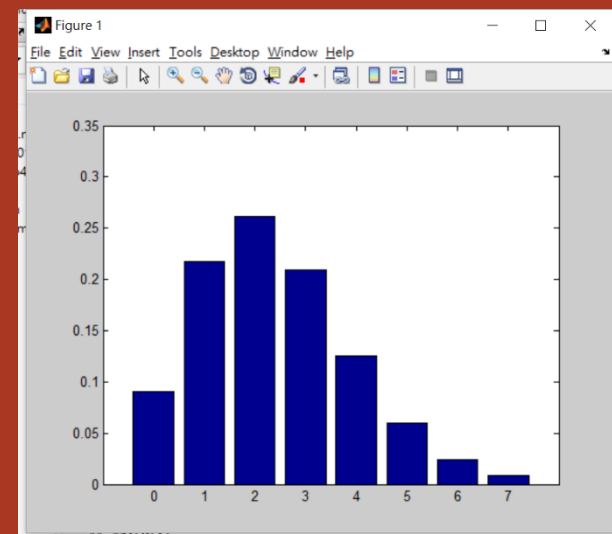
```
>> X=0:7;
```

```
>> 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
```

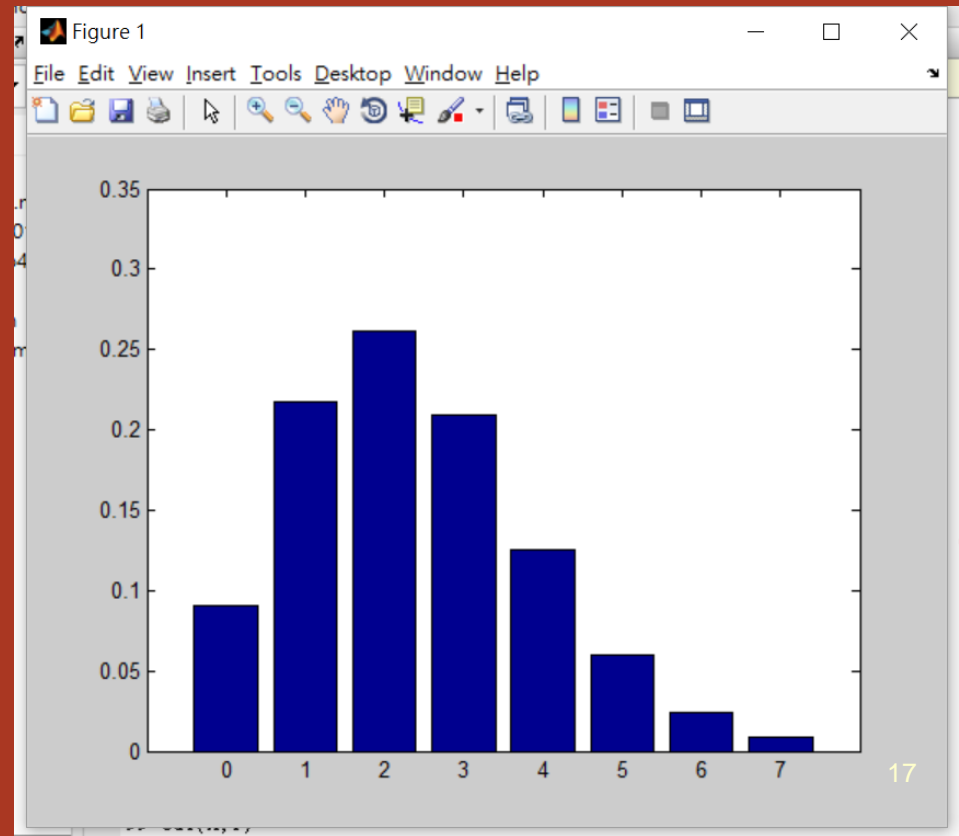
```
>> poisscdf(2,2.4)
```

```
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
```

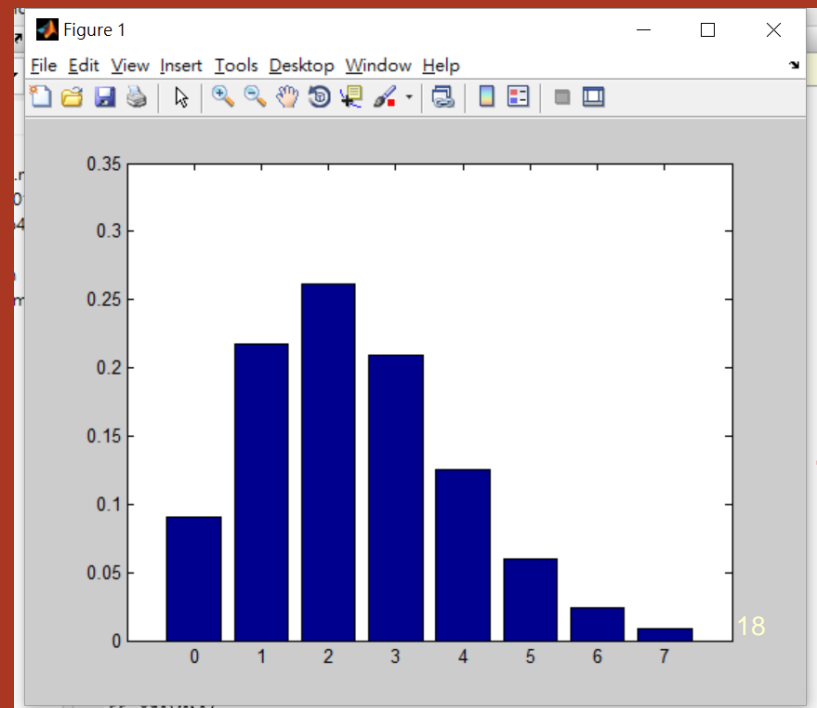
```
>> poissinv(0.75,2.4)
```

```
ans =    3
```

```
>> poissinv(0.9,2.4)
```

```
ans =    4
```

```
>>
```



Example #2

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

- Patients arrive at a hospital at random at a rate of 6/hr. Find the probability that, during any 90 min period, the number of patients arriving in the hospital is (a) exactly 7, (b) at least 10.

```
>> lambda=6*(90/60)
lambda = 9
```

First find a correct lambda to use.

```
>> poisspdf(7, lambda)
ans = 0.1171
```

For problem (a), we need the probability at $X=x=7$. This can be conveniently obtained with `poisspdf(7, lambda)`.

```
>> 1-poisscdf(9, lambda)
ans = 0.4126
>>
```

For problem (b), we need to exclude the accumulated probability of 9 or less patients, to get at least 10 patients. This can be conveniently obtained with `1.0 minus poisscdf(9, lambda)`.

Example #3

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

- Following example #2. Assuming a patient arrives at 11:30 am. Find the probability that the next patient arrives before 11:45 am.

```
>> lambda=6*(15/60)
lambda = 1.5000
```

```
>> 1-poisscdf(0, lambda)
ans = 0.7769
>>
```

First find a correct lambda to use.

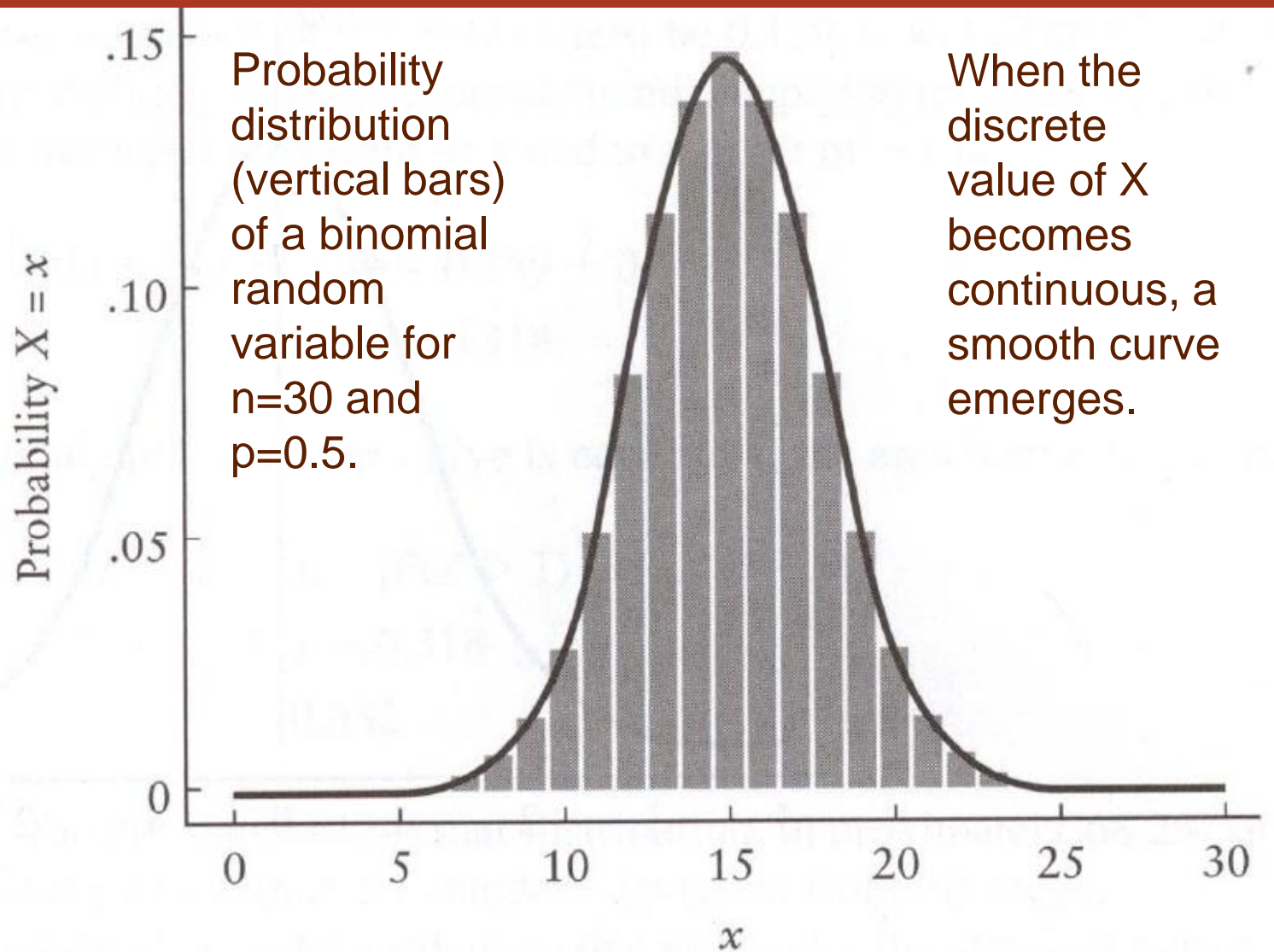
In this example, we'd like to expect at least one patient would arrive. So we use 1.0 minus poisscdf(0, lambda) to obtain probability for $x \geq 1$.

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.



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 $\lambda=np$, 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:

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

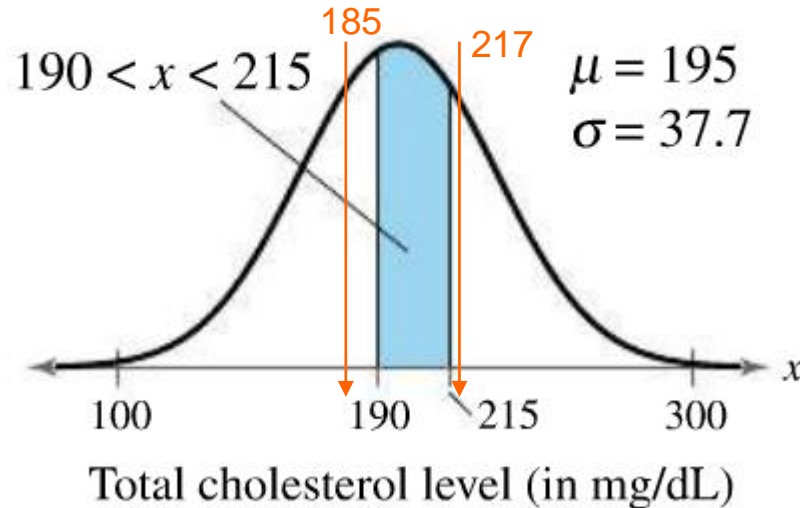
Cont'd

- 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.
- More specifically, the shape of the distribution is determined by its **mean** and the spread is determined by its standard deviation.

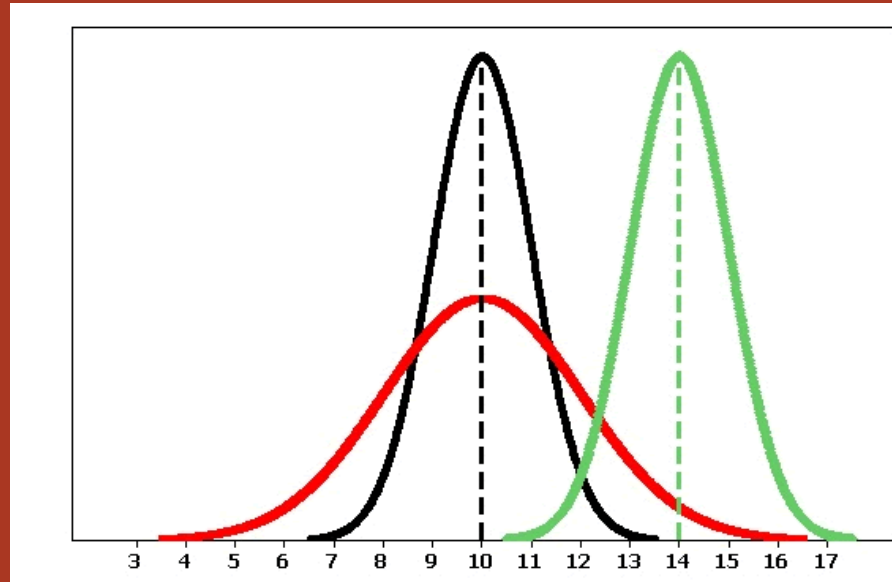
**16. U.S. Women Ages 35–44:
Total Cholesterol**



20. Cholesterol Use the normal distribution in Exercise 16.

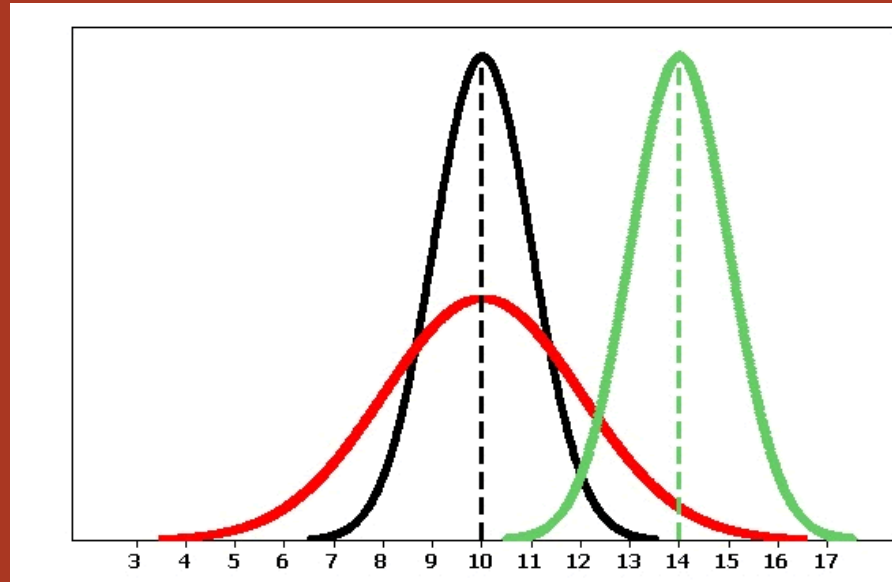
- (a) What percent of the women have a total cholesterol level less than 217 milligrams per deciliter of blood?
- (b) Out of 200 randomly selected U.S. women in the 35–44 age group, about how many would you expect to have a total cholesterol level greater than 185 milligrams per deciliter of blood?

Putting multiple normal distributions together



- The black and the red normal curves have means or centers at $\mu = 10$. However, the red curve is **more spread out** and thus has a larger standard deviation.
- As you look at these two normal curves, notice that as the red graph is **squished (stepped)** down, the spread gets larger, thus allowing the area under the curve to remain the same (which is 1.0).

Putting multiple normal distributions together



- Black and the green normal curves have the same STD or spread (the range of the black curve is 6.5-13.5, and the green curve's range is 10.5-17.5).
- Are the two curves “separated”?
- Can this separation be used to indicate that the two curves each represents different group of people (e.g., healthy vs sick, with or without drug treatment).

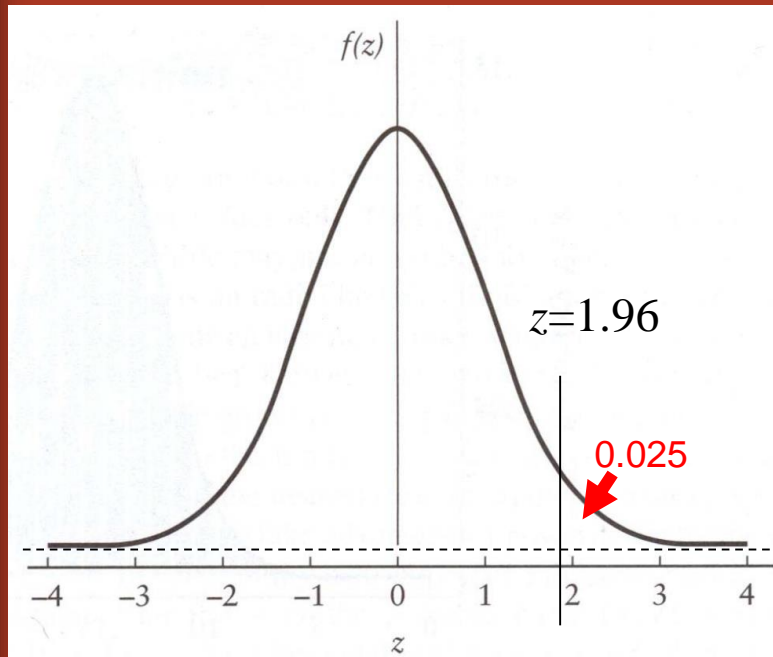
Standard Normal Distribution

- 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 $\mu=0$ and $\sigma=1$, which is called the standard normal distribution (see next slide).

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



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



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

z	Area in the <u>right tail</u>
0.00	0.500
1.65	0.049
1.96	¹0.025
2.58	²0.005
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.

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

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

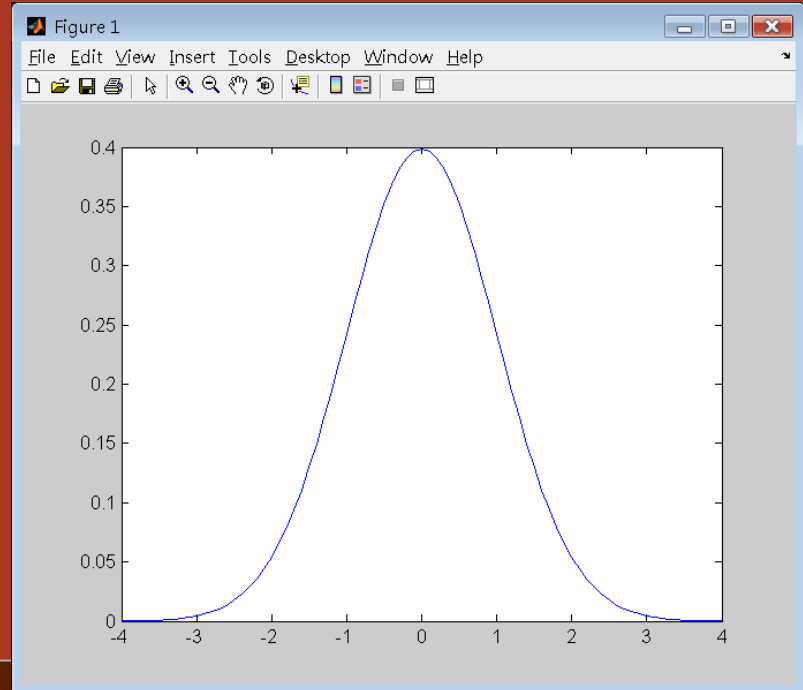
**Using MATLAB's symbolic form
and finite integration function 'int':**

```
>> syms f
>> f=1/(sqrt(2*pi))*exp(-0.5*x*x)
f =
(7186705221432913*exp(-x^2/2))/18014398509481984

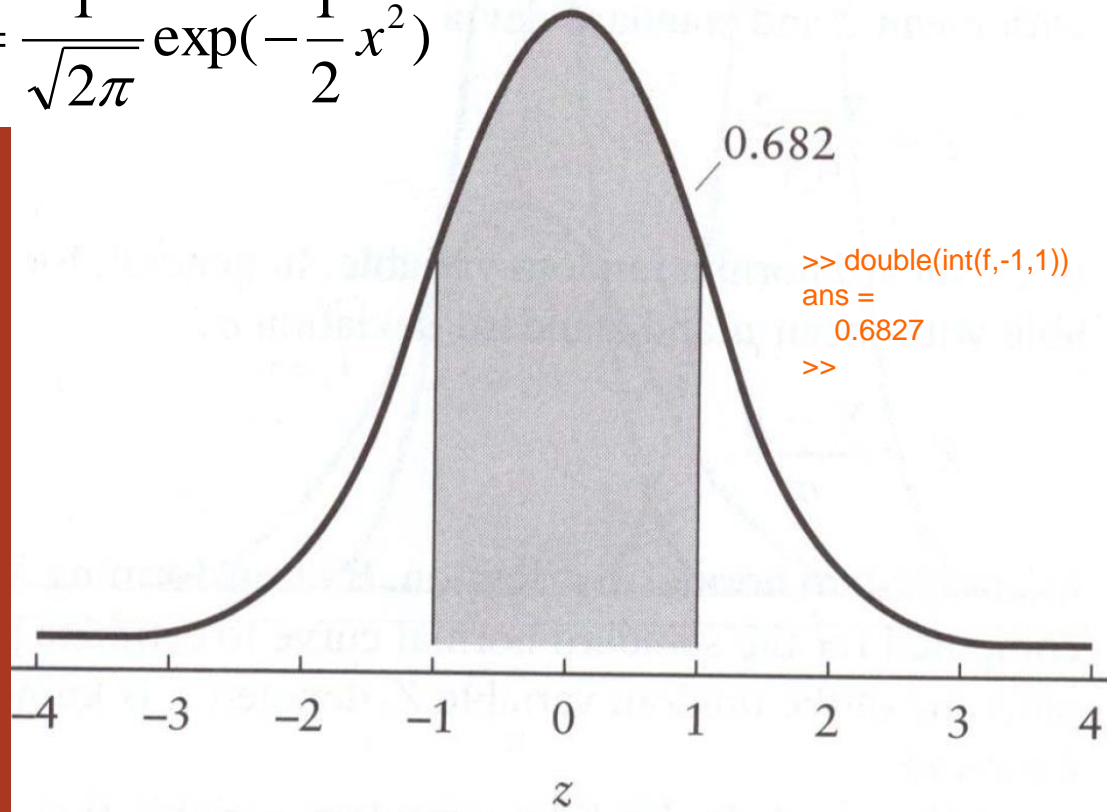
>> int(f,-4,4)
ans =
(7186705221432913*2^(1/2)*pi^(1/2)*erf(2*2^(1/2)))/18014398509481984

>> double(ans)
ans = 0.9999
```

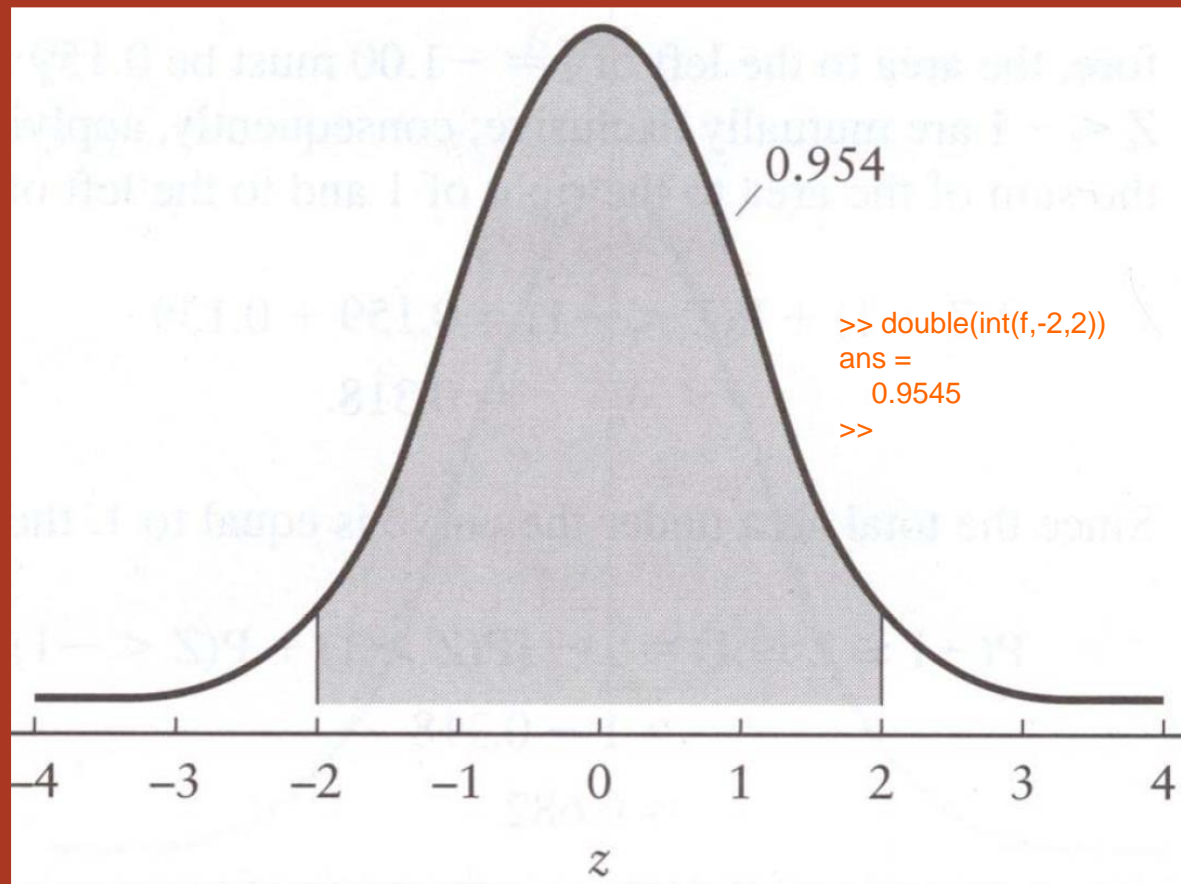
*An integration from -4 to +4 gives almost
99.99% of the area under curve.*



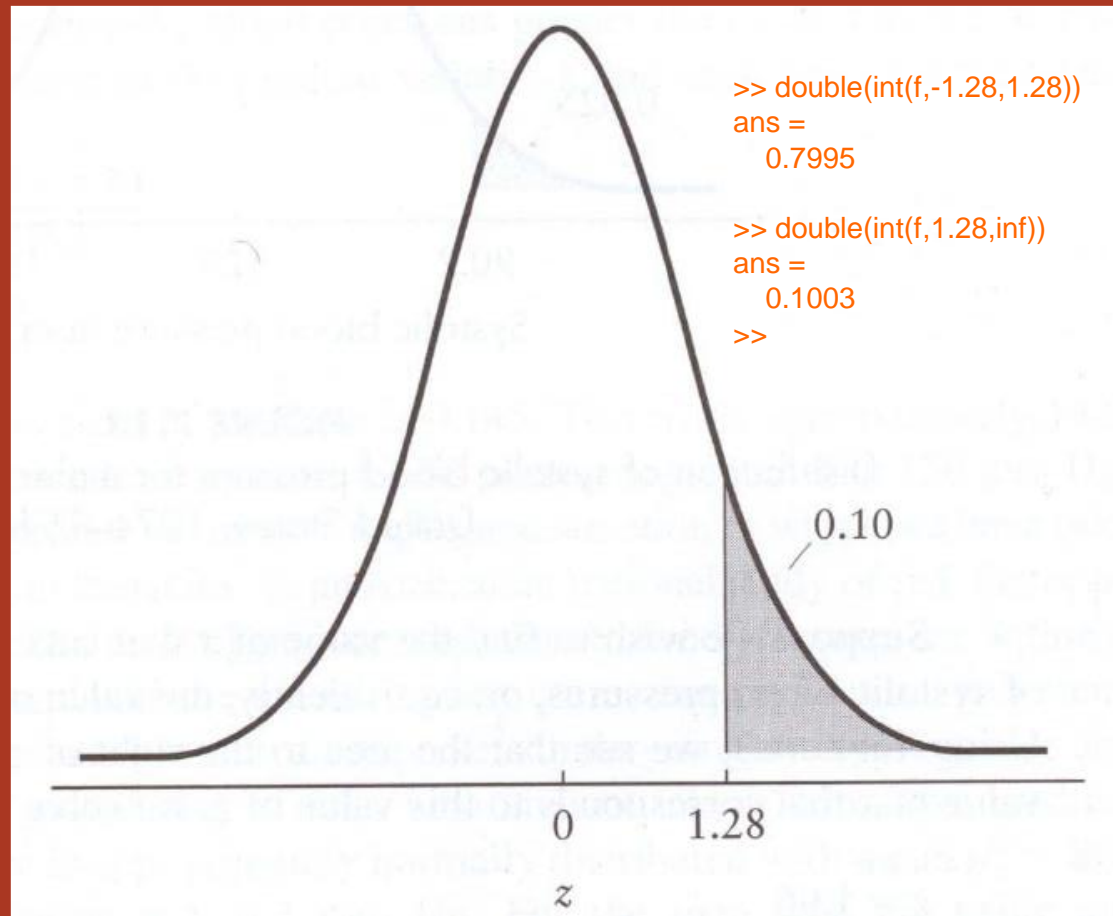
$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$



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).

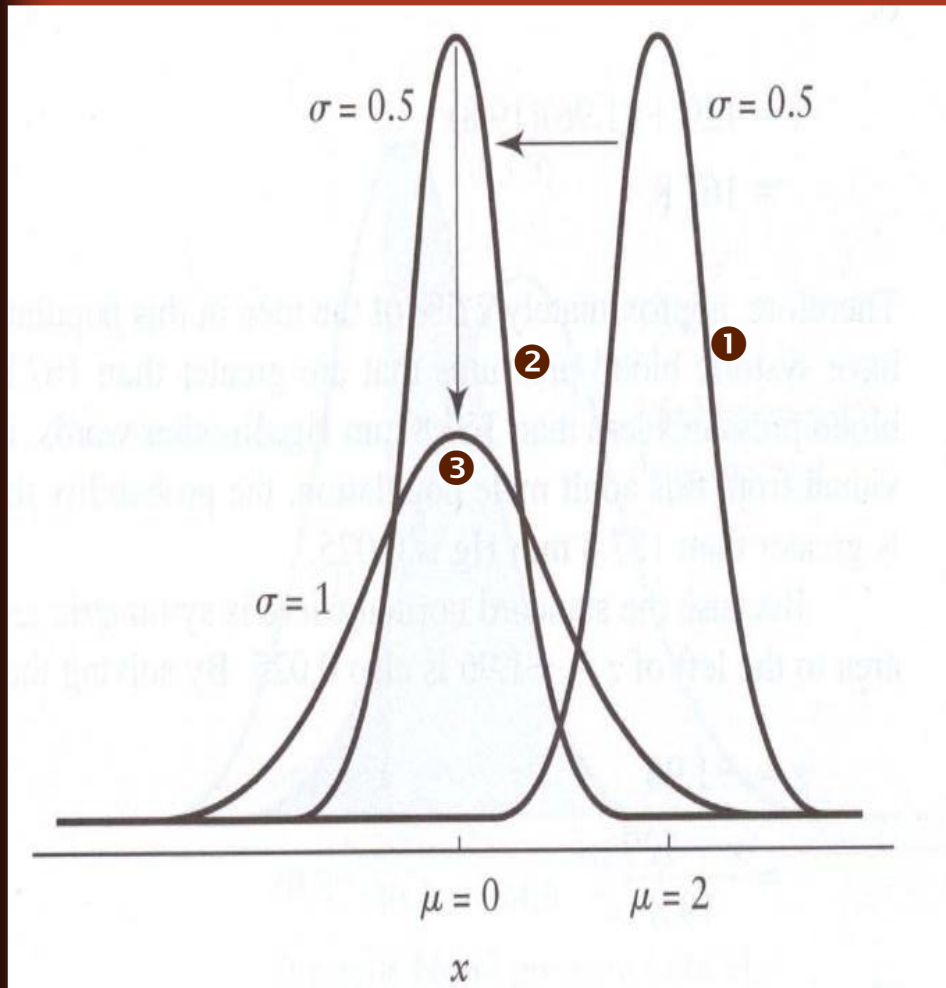
Z-score

- 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.

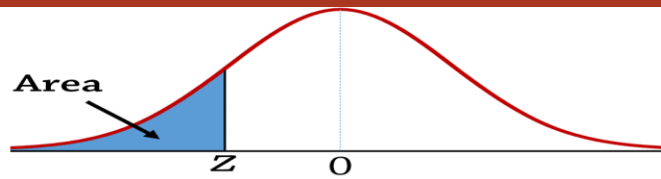
$$Z = \frac{X - \mu}{\sigma}$$



The effect is apparent by the following example. Suppose curve ① has mean at $x=2$ and $\sigma=0.5$. Subtracting the mean from x would shift graph ① to ②. Further divided by σ would compress ② into ③, 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.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

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

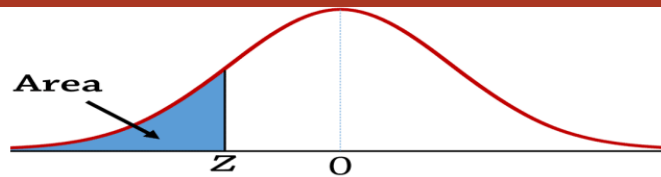
$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right)$$

- 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
```



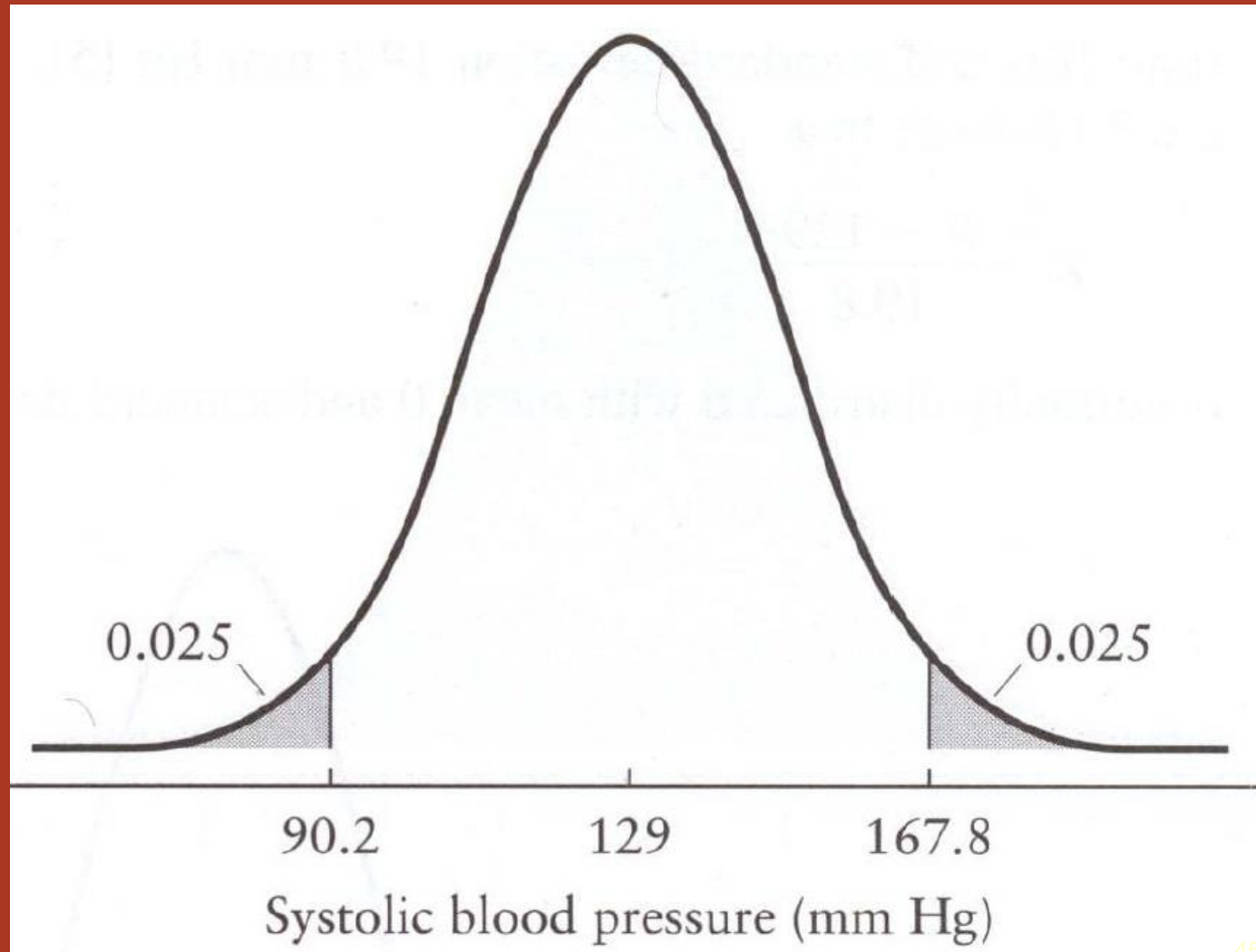
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	One can also use this table to look for the correct z-value (to be 1.96).						30	0.0029	0.0028	0.0027
-2.6							40	0.0039	0.0038	0.0037
-2.5							54	0.0052	0.0051	0.0049
-2.4							71	0.0069	0.0068	0.0066
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Cont'd

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

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

Similarly we may get the lower cut-off $X = 129 + \underline{1.96}*(-19.8) = 90.2$



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.

Reminder

- Next Tuesday (4/5) is a holiday. No class meeting.
- Two weeks later (4/12) is our first midterm exam, a close-book test. You are only allowed to use hand-calculators (including Microsoft's 小算盤), Excel, MATLAB.
- No mobile phone is allowed at all time.