Biostatistics

Week #5 3/22/2022

Chapter 7 Theoretical Probability Distributions – Part 1

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

7.1 Probability Distribution

We know what is probability. But what is distribution? What is probability distribution?

Random Variable

- Any characteristic that can be measured or categorized is called a ***variable***.
- If a variable can assume **different values** such that any particular outcome is determined **by chance**, it is called a ***random variable***.
- A ***probability distribution*** applies the theory of probability to ***describe*** the random variable.

Discrete and Continuous Random Variables

- A random variable is ***discrete*** if it can assume a **countable** number of values. For example, the “coin” example assumes only 2 values – 1 and 0.
- A random variable is ***continuous*** if it can assume an uncountable number of values. For example, a height or a weight, which can take on any value within a specified interval or continuum.

Probability Distribution

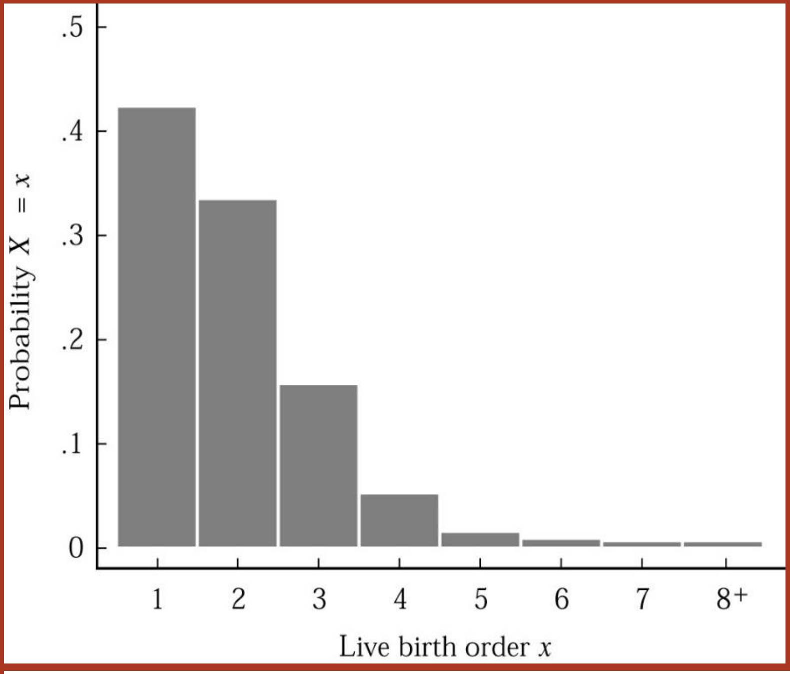
- In probability theory and statistics, a **probability distribution** identifies either
  - the probability of ***each value*** of an unidentified random variable (when the variable is ***discrete***), or
  - the probability of ***the value falling within a particular interval*** (when the variable is ***continuous***).
- Every random variable has a corresponding probability distribution.

Example

A discrete probability distribution of the **birth order of children born to women** in US (based on the experience of the US population in 1986).

TABLE 7.1
Probability distribution of a random variable X representing the birth order of children born in the United States

x	$P(X = x)$
1	0.416
2	0.330
3	0.158
4	0.058
5	0.021
6	0.009
7	0.004
8^+	0.004
Total	1.000



$P(X=4)=0.058$
$P(X=1 \text{ or } X=2)=P(X=1)+P(X=2)=0.746$

Additive rule of probability for mutually exclusive events.

Comments

- In previous example, it is possible to **tabulate** the distribution because of limited count for this random variable.
- If a random variable can take on a large number of values, a probability distribution may not be a useful way to summarize its behavior.
- In this case, a number of summarization can help – population ***mean***, population ***variance*** and population ***standard deviation***.

Population Mean (Expected Value期望值)

- Given a discrete random variable X with values x_i, that occur with probabilities $p(x_i)$, the population mean of X is

$$E(X) = \mu = \sum_{all\ x_i} x_i \cdot p(x_i)$$

For the case of rolling a dice, for example, we have

$$E(X) = \mu = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \ldots + 6 \cdot \frac{1}{6}$$

$$= \frac{21}{6} = 3.5$$

Population Variance

- Let X be a discrete random variable with possible values x_i that occur with probabilities $p(x_i)$, and let $E(X) = \mu$.

  The variance of X is defined by

$$V(X) = \sigma^2 = E\left[(X-\mu)^2\right] = \sum_{all\ x_i} (x_i - \mu)^2 p(x_i)$$

The standard *deviation is*

$$\sigma = \sqrt{\sigma^2}$$

For the dice-rolling example

$$V(X) = \sigma^2 = E\left[(X-\mu)^2\right] =$$

$$(1-3.5)^2 \cdot \frac{1}{6} + (2-3.5)^2 \cdot \frac{1}{6} + \ldots + (6-3.5)^2 \cdot \frac{1}{6}$$

$$= (6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25) \cdot \frac{1}{6}$$

$$= 2.916667$$

The standard deviation is

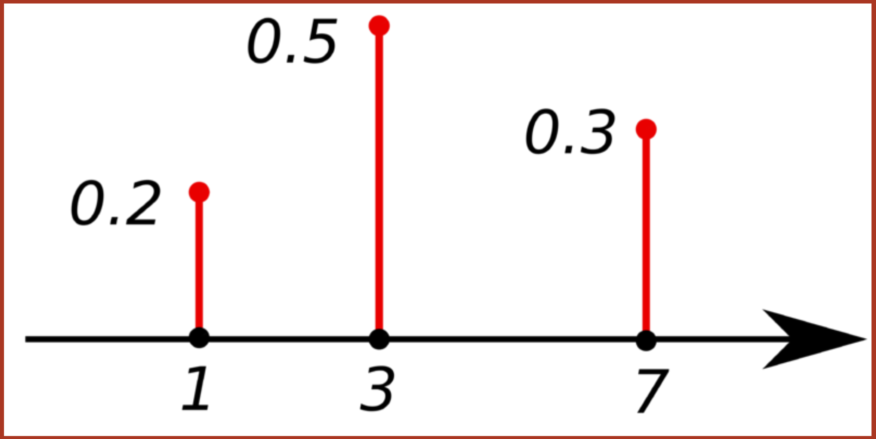
$$\sigma = \sqrt{\sigma^2} = \sqrt{2.916667} = 1.707825$$

A brief summary

- This example tells you that, if you roll the dice many times, the ***average*** you may get is 3.5 points.
- It is likely that the average may ‘mostly’ be within the range 3.5 ± 1.7 points.

pmf and pdf

- In probability theory, a **probability mass function** (abbreviated **pmf**) is a function that gives the probability that **a discrete random variable** is exactly equal to some value.

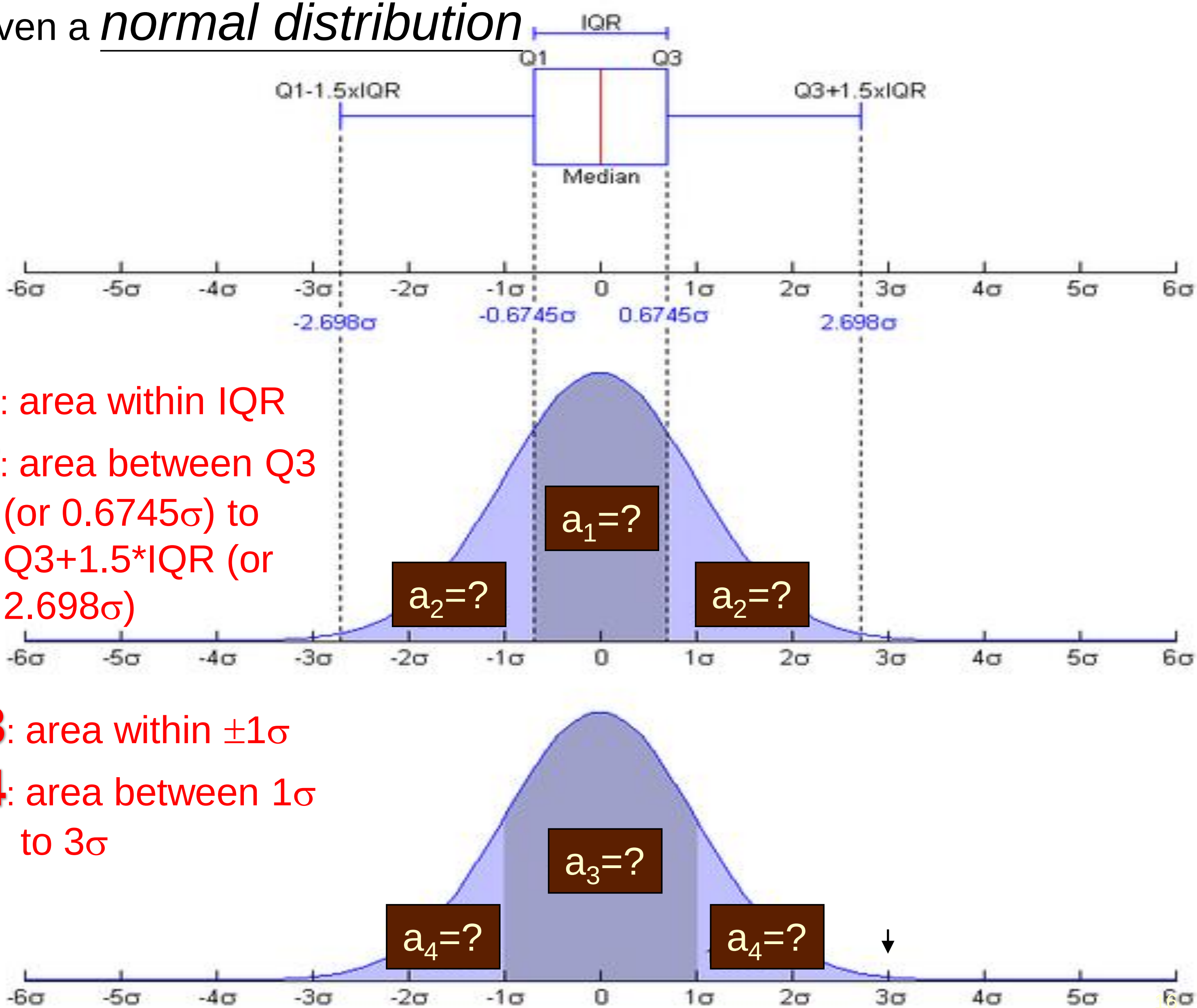


The graph of a probability mass function. All the values of this function must be non-negative and sum up to 1.

Cont'd

- A pmf differs from a **probability density function** (abbreviated **pdf**) in that the values of a pdf are defined only for **continuous** random variables.
- It is the **integral** of a pdf over a range of possible values that gives the probability of the random variable **falling within that range**.

Given a *normal distribution*



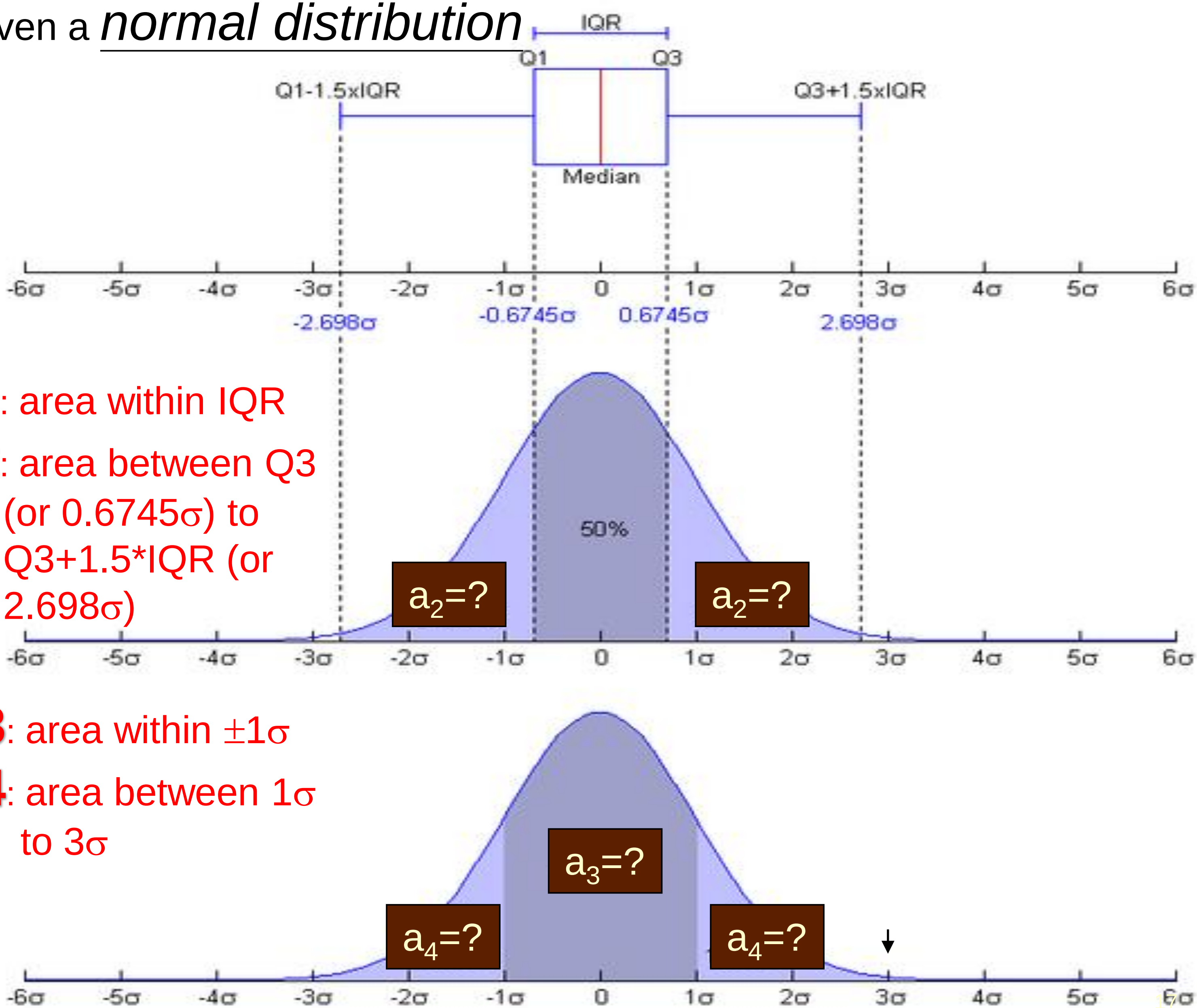
1: area within IQR

2: area between Q3 (or 0.6745σ) to Q3+1.5*IQR (or 2.698σ)

3: area within $\pm 1\sigma$

4: area between 1σ to 3σ

Given a *normal distribution*



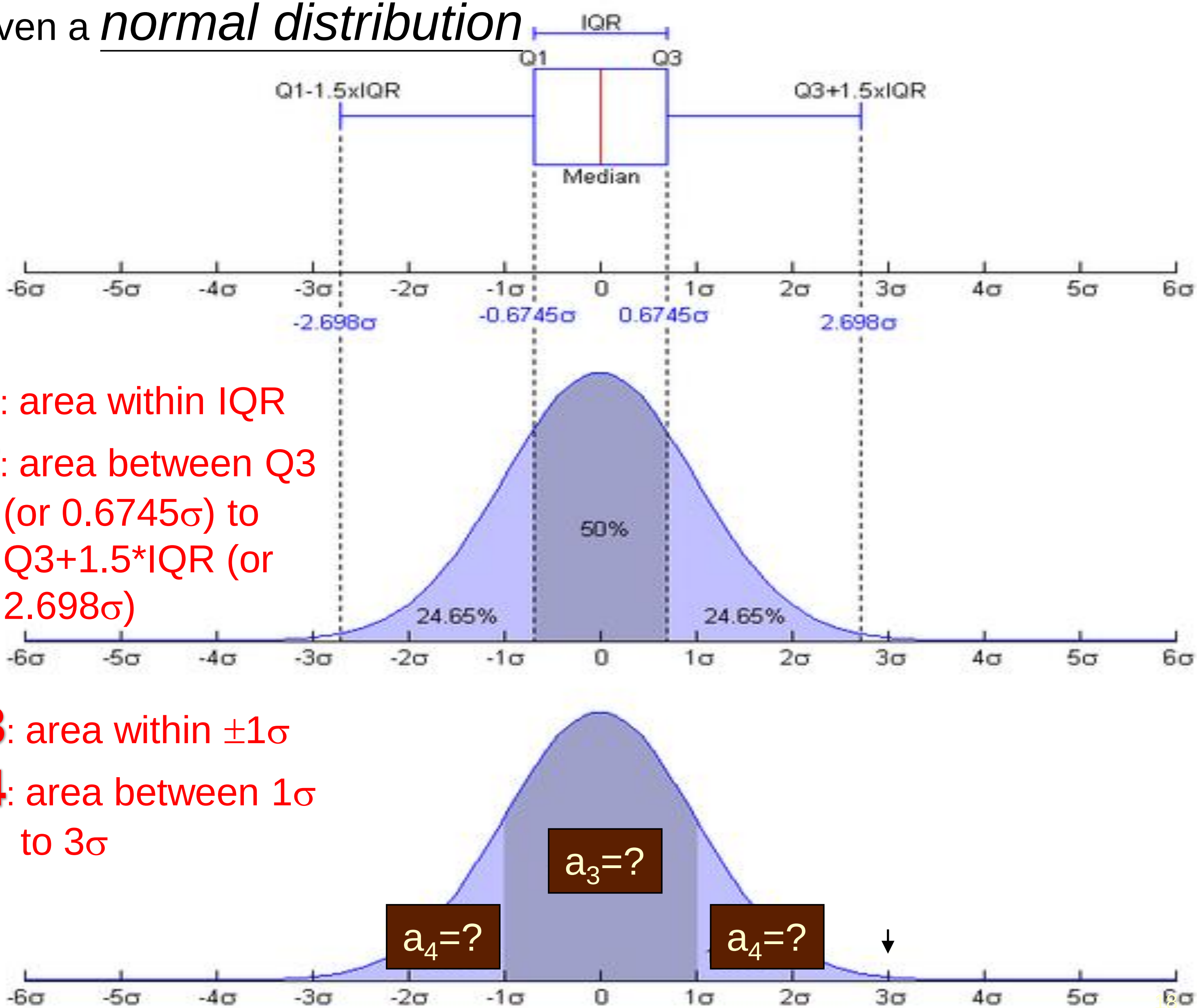
1: area within IQR

2: area between Q3 (or 0.6745σ) to Q3+1.5*IQR (or 2.698σ)

3: area within ±1σ

4: area between 1σ to 3σ

Given a *normal distribution*



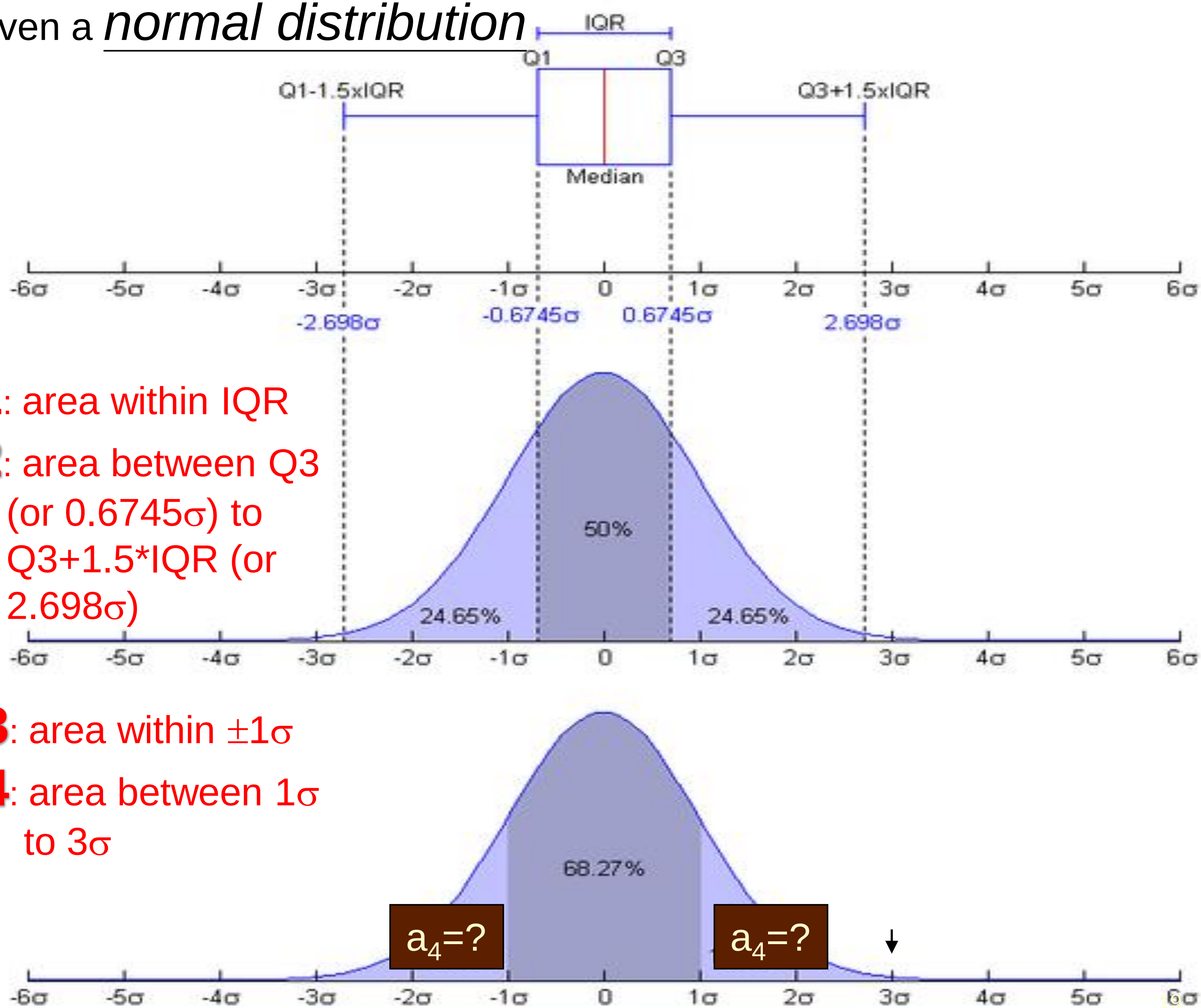
1: area within IQR

2: area between Q3 (or 0.6745σ) to Q3+1.5*IQR (or 2.698σ)

3: area within ±1σ

4: area between 1σ to 3σ

Given a *normal distribution*



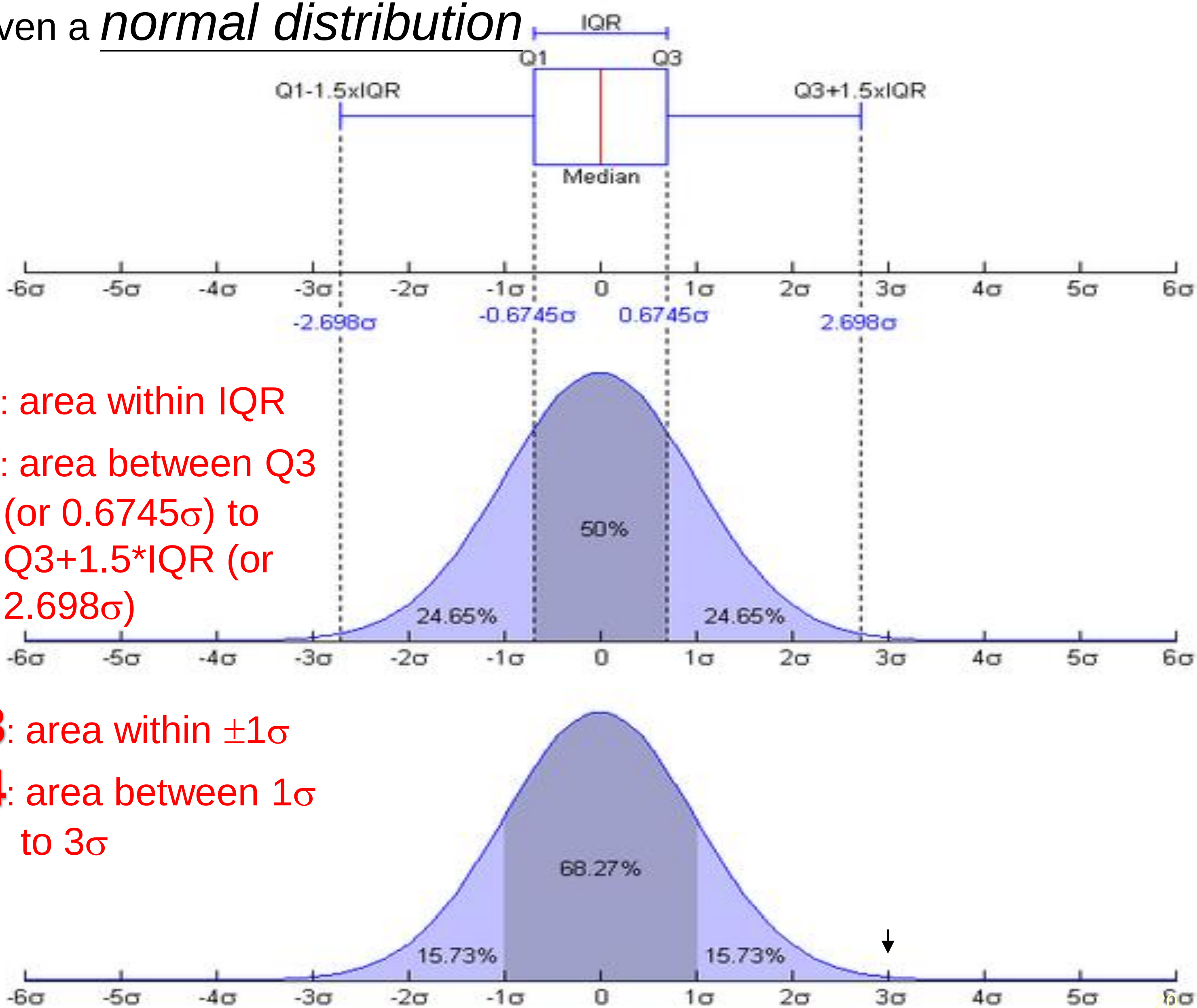
1: area within IQR

2: area between Q3 (or 0.6745σ) to Q3+1.5*IQR (or 2.698σ)

3: area within ±1σ

4: area between 1σ to 3σ

Given a *normal distribution*



1: area within IQR

2: area between Q3 (or 0.6745σ) to Q3+1.5*IQR (or 2.698σ)

3: area within ±1σ

4: area between 1σ to 3σ

Summary

- Probabilities calculated based on a **finite** amount of data (such as the birth order example mentioned previously) are called ***empirical probability***.
- The probability distributions for many other random variables of interest, however, can be determined (or approximated) based on *theoretical (or mathematical)* consideration.
- These are called ***theoretical probability distributions***.

7.2 The Binomial Distribution

From Wikipedia

- The **binomial distribution** (with **parameters *n* and *p***) is the **discrete probability distribution** of the number of successes in **a sequence of *n* independent yes/no experiments**, each of which yields **success** with probability *p*.
- A success/failure experiment is also called a ***Bernoulli experiment*** or ***Bernoulli trial***.

Introduction

- **Bernoulli** random variable
  - A **dichotomous** (二分的) random variable Y can result in only one of two possible outcomes, referred to as “failure” and “success”. (or yes vs no, male vs female, life vs death, sickness vs health, etc.)
- Typical cases where the binomial experiment applies:
  - A coin flipped results in heads or tails
  - An election candidate wins or loses

Example 1

- Let Y be a random variable that represents smoking status; **Y=1 if an adult is currently a smoker ; Y=0 if not**.
- In 1987, 29% of the adults in US smoked; the probabilities associated with the outcomes of Y are **$P(Y=1) = p = 0.29$**; and **$P(Y=0) = 1–p = 0.71$**.
- These are the probability distribution of the random variable Y.

Cont'd

- We randomly select two adults from the population, Y1 and Y2.
- We now **introduce a new random variable X** that represents the number of smokers in the pair (2 persons).
- **X=Y1+Y2** , the possible outcomes of X are {0, 1, 2}
  - 0: both non-smokers
  - 1: one smokes & one does not
  - 2: both smokers
- What is the probability distribution of X ?

Cont'd

Outcomes of Y's Y1	Y2	Probabilities of these outcomes	Outcomes of X=Y1+Y2
0	0	(1–p)*(1–p)	0
1	0	p*(1–p)	1
0	1	(1–p)*p	1
1	1	p*p	2

$P(X=0) = (1 - p)^2 = (0.71)^2 = 0.504$

$P(X=1) = p(1 - p) + (1 - p)p = 2p(1 - p) = 2(0.29)(0.71) = 0.412$

$P(X=2) = p^2 = (0.29)^2 = 0.084$

Note:

$P(X=0) + P(X=1) + P(X=2) = 0.504 + 0.084 + 0.412 =$ **1.000**

(all mutually exclusive)

We call X a special case of the *Binomial distribution*

Binomial Probability Distribution

- There are ***n* independent Bernoulli trials** (*n* is finite and fixed).
- Each trial can result in a success or a failure (one of two mutually exclusive outcomes).
- The probability *p* of success is ***the same*** for all the trials.
- All the trials of the experiment are ***independent***.

Introduce a new random variable X that represents the number of smokers in 3 persons.

Outcomes of Y's Y1	 Y2	 Y3	Probabilities of these outcomes	Outcomes of X=Y1+Y2+Y3
0	0	0	(1-p)(1-p) (1-p)	0
1	0	0	p(1-p) (1-p)	1
0	1	0	(1-p)p(1-p)	1
0	0	1	(1-p) (1-p)p	1
1	1	0	pp(1-p)	2
1	0	1	p(1-p)p	2
0	1	1	(1-p)pp	2
1	1	1	ppp	3

$P(X=0) = (1-p)^3 = (0.71)^3 = 0.358$

$P(X=1) = 3p(1-p)^2 = 3(0.29)(0.71)^2 = 0.439$

$P(X=2) = 3p^2(1-p) = 3(0.29)^2(0.71) = 0.179$

$P(X=3) = p^3 = (0.29)^2 = 0.024$

What if we continue?

- N=2 : X=0,1,2 (frequency=1,2,1)
- N=3 : X=0,1,2,3 (frequency=1,3,3,1)
- N=4 : X=0,1,2,3,4 (frequency=1,4,6,4,1)
- N=5 : X=0,1,2,3,4,5,6 (frequency=1,5,10,10,5,1)
- N=6…
- The coefficients of these expansions $(a+b)^2$, $(a+b)^3$, $(a+b)^4$, $(a+b)^5$,…

Calculating the Binomial Probability

- In general, the binomial probability is calculated by：

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

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

N = 1, 2, 3, … and x = 0, 1, 2, …, n

For example, with n = 3 and x = 0 (three non-smokers) with p = 0.29, we have

$$P(X=0)=p(0)=C_0^3 0.29^0(1-0.29)^{3-0}$$

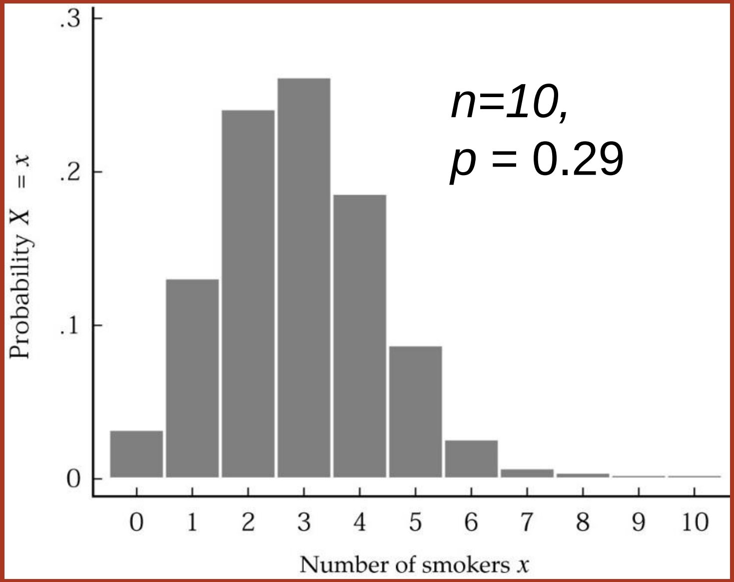
$$=\frac{3!}{0!(3-0)!}0.29^0(0.71)^3=0.71^3=0.357911$$

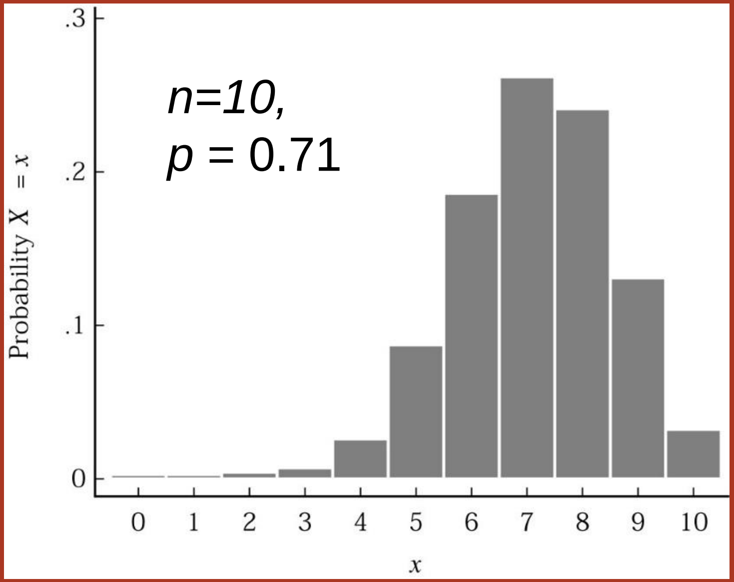
Previously we had :

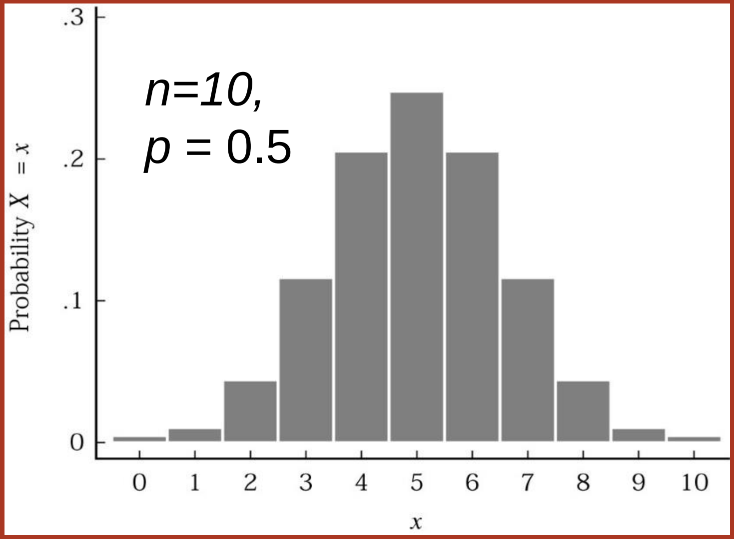
$P(X=0) = (1-p)^3 = (0.71)^3 = 0.358$

Some simulations

p is the percent of US people smoked.







☑ **Can you explain why some of these graphs is left, right or center-peaked?**
☑ **What is the area sum for each of these graphs? What is the meaning of it?**

Mean and STD for Binomial Distribution

- **Mean value = *np***
- **Variance = σ^2 = *np*(1–*p*)**, where σ is the standard deviation of this binomial random variable *X* with repeated samples of size *n*.
- For previous example, we have (for n = 10 and p = 0.29)

$$\mu = np = 10(0.29) = 2.9$$

$$\sigma = \sqrt{10(0.29)(1-0.29)} = 1.435$$

Example 2

- Given 14 newborns, and knowing that there are ***X baby girls*** among these 14.
- We'd like to build the probability distribution of X (=1 to 14) and know the mean and standard deviation of it.

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

$$= \frac{n!}{(n-x)!\,x!} 0.5^n$$

$$P(X=0)=\frac{14!}{(14-0)!0!}0.5^{14}=0.5^{14}$$

$$P(X=1)=\frac{14!}{(14-1)!1!}0.5^{14}=14\times 0.5^{14}$$

$$P(X=2)=\frac{14!}{(14-2)!2!}0.5^{14}=\frac{14\times 13}{2}\times 0.5^{14}$$

$$P(X=3)=\frac{14!}{(14-3)!3!}0.5^{14}=\frac{14\times 13\times 12}{3!}\times 0.5^{14}$$

Using MATLAB:

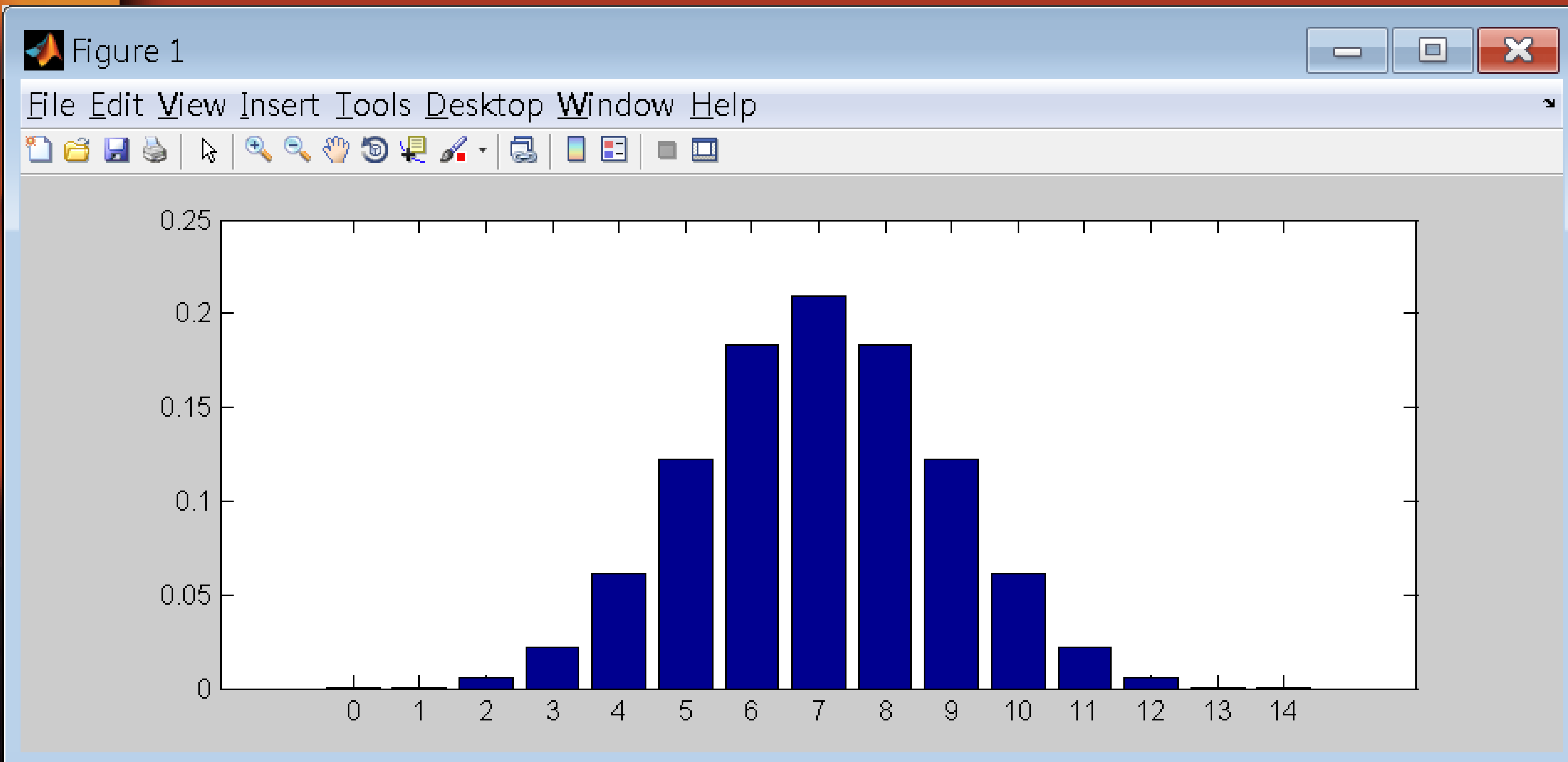
$$P(X=x)=p(x)=C_x^n p^x(1-p)^{n-x}$$

$$=\frac{n!}{(n-x)!\,x!}0.5^n$$

```
>> X=0:14;
>> P=factorial(14)/(factorial(14-X)*factorial(X))*0.5^14;
??? Error using ==> mtimes
Inner matrix dimensions must agree.
```

```
>> X=0:14;
>> P=factorial(14)./(factorial(14-X).*factorial(X))*0.5^14;
>> bar(X, P)
>>
```

Observing the two newly added dots.



Population Mean & Standard Deviation (Expected Value期望值)

- Recall that, given a discrete random variable *X* with values x_i, that occur with probabilities $p(x_i)$, the population mean of *X* is

$$E(X) = \mu = \sum_{all\ x_i} x_i \cdot p(x_i)$$

- The variance of *X* is defined by

$$V(X) = \sigma^2 = E\left[(X-\mu)^2\right] = \sum_{all\ x_i} (x_i - \mu)^2 p(x_i)$$

$$The \text{ standard } deviation\ is$$

$$\sigma = \sqrt{\sigma^2}$$

```
>> X
X =
  Columns 1 through 15
     0     1     2     3     4     5     6     7     8     9    10    11    12    13    14
>> P
P =
  Columns 1 through 8
    0.0001    0.0009    0.0056    0.0222    0.0611    0.1222    0.1833    0.2095
  Columns 9 through 15
    0.1833    0.1222    0.0611    0.0222    0.0056    0.0009    0.0001
```

>> N=X.*P

$$E(X)=\mu=\sum_{all\ x_i} x_i \cdot p(x_i)$$

```
N =
  Columns 1 through 8
         0    0.0009    0.0111    0.0667    0.2444    0.6110    1.0997    1.4663
  Columns 9 through 15
    1.4663    1.0997    0.6110    0.2444    0.0667    0.0111    0.0009
```

>> sum(N)

```
ans =
     7
>>
```

This can also be obtained conveniently by:

*Mean value = np = 14*0.5 = 7*

$$V(X) = \sigma^2 = E\left[(X-\mu)^2\right] = \sum_{all\ x_i} (x_i - \mu)^2 p(x_i)$$

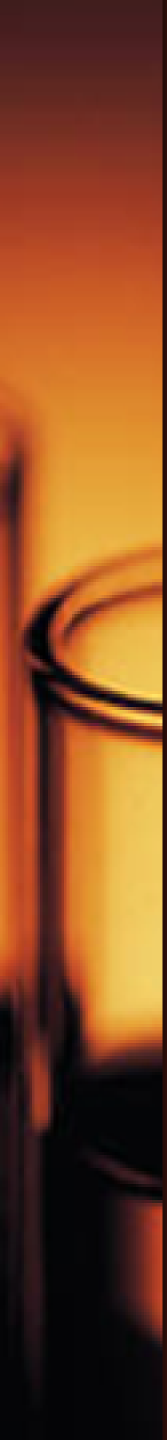
```
>> VAR=P.*(X-7).^2
VAR =
  Columns 1 through 11
    0.0030    0.0308    0.1389    0.3555    0.5499    0.4888    0.1833         0
0.1833    0.4888    0.5499
  Columns 12 through 15
    0.3555    0.1389    0.0308    0.0030
>> sum(VAR)
ans =
    3.5000
>> sqrt(ans)
ans =
    1.8708
>>
```

This can also be obtained conveniently by:

*Standard deviation = sqrt(n*p*(1-p)) = sqrt(14*0.5*0.5) =sqrt(3.5) = 1.8708*

Conclusion

- Among randomly chosen 14 newborns, the number of baby girls would range from 0 to 14. They follow binomial distribution.
- The average number of baby girls would be 7, with standard deviation of 1.8708.
- It is this “theoretical” feature of binomial distribution that makes those formulas available for easier conputation.



Statistics - parametric vs nonparametric

- Parametric statistics are based on assumptions about the ***distribution*** of population from which the sample was taken.
- Nonparametric statistics are not based on assumptions, that is, the data can be collected from a sample that ***does not*** follow a specific distribution.

Cont'd

- Common parametric statistics are, for example, the Student's t-tests.
- Common nonparametric statistics are, for example, the Mann-Whitney-Wilcoxon (MWW) test or the Wilcoxon test.
- We will cover only parametric statistics in this course.

MATLAB Supported Distributions

- MATLAB's "Statistics and Machine Learning Toolbox™" supports more than 30 probability distributions, including parametric, nonparametric, continuous, and discrete distributions.

A couple of discrete probability distributions we will cover in this course:

Binomial	**binopdf binocdf binoinv** binostat binofit binornd
Poisson	**poisspdf poisscdf poissinv** poisstat poissfit poissrnd

A number of continuous probability distributions we will cover in this course:

Normal (Gaussian)	**normpdf** **normcdf** **norminv** normstat normfit normlike normrnd
Chi-square	**chi2pdf** **chi2cdf** **chi2inv** chi2stat chi2rnd
F	**fpdf** **fcdf** **finv** fstat frnd
Student's *t*	**tpdf** **tcdf** **tinv** tstat trnd

Cont'd

- **pdf** — ***Probability*** density functions
- **cdf** — ***Cumulative*** distribution functions
- **inv** — **Inverse** cumulative distribution functions

>> help **binopdf**

BINOPDF Binomial probability density function.

Y = BINOPDF(X,N,P) returns the binomial probability density function with parameters N and P at the values in X.

Note that the density function is zero **unless X is an integer**.

>> X=0:14;

>> P=binopdf(X, 14, 0.5)

P =

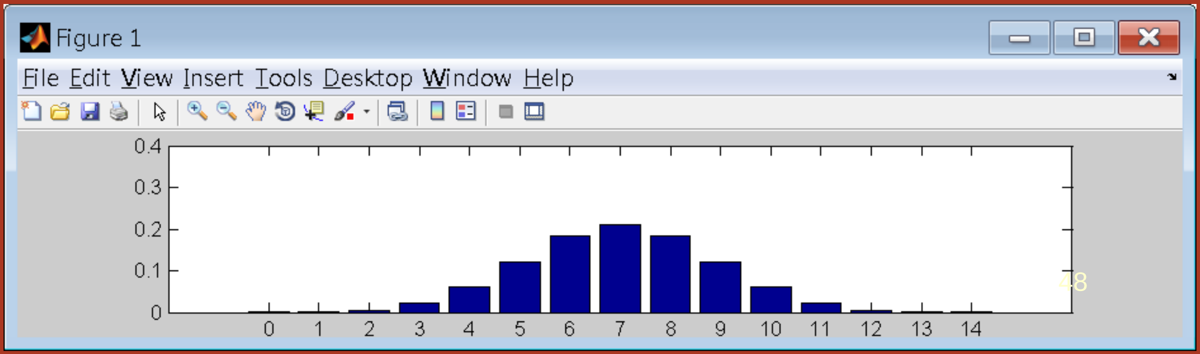
Columns 1 through 11

0.0001 0.0009 0.0056 0.0222 0.0611 0.1222 0.1833 0.2095 0.1833 0.1222 0.0611

Columns 12 through 15

0.0222 0.0056 0.0009 0.0001

>> **bar(X, P)**



>> **format long**

>> P

P =

Columns 1 through 5

0.000061035156250   0.000854492187500   0.005554199218750
0.022216796875000   0.061096191406250

Columns 6 through 10

0.122192382812500   0.183288574218750   0.209472656250000
0.183288574218750   0.122192382812500

Columns 11 through 15

0.061096191406250   0.022216796875000   0.005554199218750
0.000854492187500   0.000061035156250

>> **sum(P(1:1)), sum(P(1:2)), sum(P(1:3)), sum(P(1:4)), <u>sum(P(1:5))</u>**

ans = 6.103515625000003e-005
ans = 9.155273437499991e-004
ans = 0.006469726562500
ans = 0.028686523437500
<u>ans = 0.089782714843750</u>

These are cumulative probabilities (cumulating from 0 to 0, 0 to 1, 0 to 2, 0 to 3, and 0 to 4)

Probability of having 0 to 4 baby girls from randomly choosing 14 newborns.

>> help **binocdf**

BINOCDF Binomial ***cumulative*** distribution function.

Y=BINOCDF(X,N,P) returns the binomial cumulative distribution function with parameters N and P at the values in X.

>> binocdf(0,14,0.5), binocdf(1,14,0.5), binocdf(2,14,0.5), binocdf(3,14,0.5), **binocdf(4,14,0.5)**

ans =
6.103515625000003e-005 *P(0)*

ans =
9.155273437499991e-004 *P(0) + P(1)*

ans =
0.006469726562500 *P(0) + P(1) + P(2)*

ans =
0.028686523437500 *P(0) + P(1) + P(2) + P(3)*

ans =
0.089782714843750 *P(0) + P(1) + P(2) + P(3) + P(4)*

>> help **binoinv**

BINOINV ***Inverse*** of the binomial cumulative distribution function (cdf).

X = BINOINV(Y,N,P) returns the inverse of the binomial cdf with parameters N and P. Since the binomial distribution is discrete, BINOINV returns **the least integer X** such that the binomial cdf evaluated at X, ***equals or exceeds Y***.

```
>> binoinv(0.01, 14, 0.5)
ans =
    3
>> binoinv(0.05, 14, 0.5)
ans =
    4
```

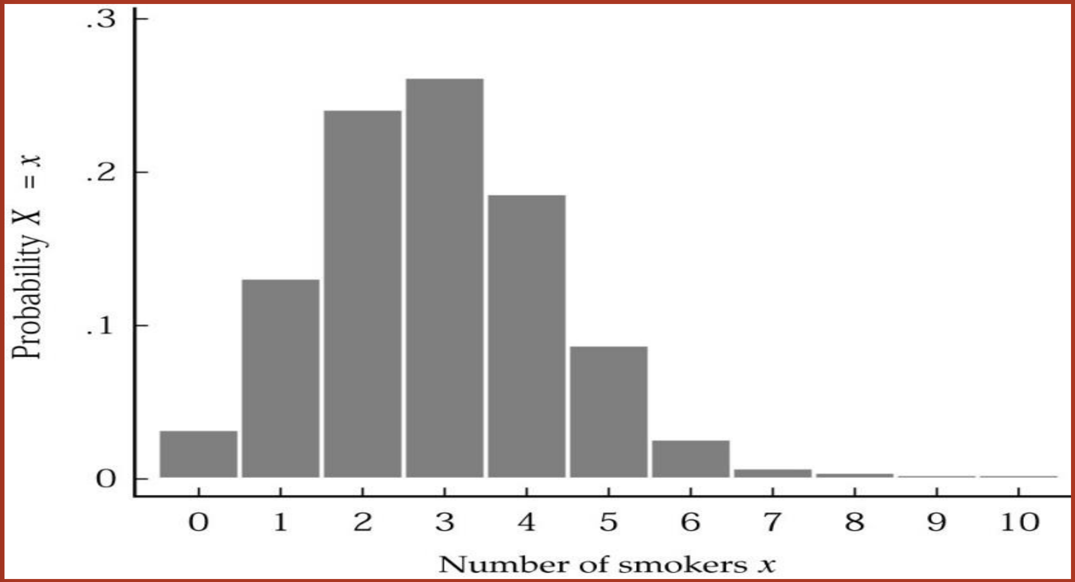
P(0) + P(1) + P(2) = 0.0065 will not equal or exceed 0.01. Adding one more (P(0) + P(1) + P(2) + P(3)=0.0287) will.

```
>> binoinv(0.1, 14, 0.5)
ans =
    5
```

Cumulating up to P(X=5) is needed to get probability equal or exceed 0.1.

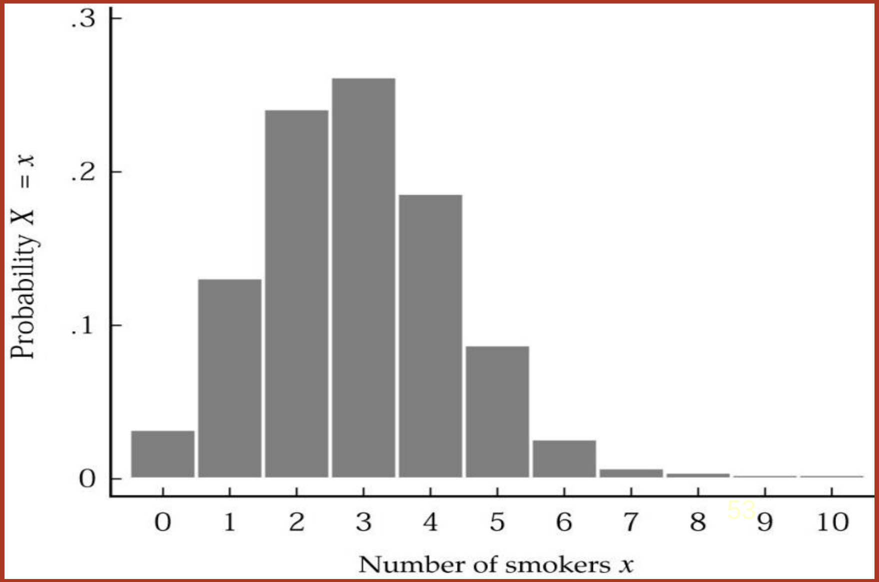
Summary

- Y = binopdf(X,N,P) = bino + pdf
- Y = binocdf(X,N,P) = bino + cdf
- X = binoinv(Y,N,P) = bino + inv



Cont'd

- Probability density function for Binomial distribution.
- Horizontal axis is the discrete X values.
- Vertical axis is the probability density for each X.
- Note that "probability density" is "probability" for discrete distribution.



Example 3

- Suppose, according to the latest police reports, 80% of all petty crimes (輕罪) are **unresolved**.
- In your town, at least three of such petty crimes are committed.
- The three crimes are all independent of each other.
- From the given data, what is the probability that **one** of the three crimes will be **resolved**?

Solution

The first step in finding the binomial probability is to verify that the situation satisfies the four rules of binomial distribution:

- Number of fixed trials (n): 3 (Number of petty crimes)
- Number of mutually exclusive outcomes: 2 (solved and unsolved)
- The probability of success (p): 0.2 (20% of cases are solved)
- Independent trials: Yes

We find the probability that one of the crimes will be solved in the three independent trials. It is shown as follows:

Trial 1 = Solved 1st, unsolved 2nd, and unsolved 3rd
= 0.2 x 0. 8 x 0.8
= 0.128

Trial 2 = Unsolved 1st, solved 2nd, and unsolved 3rd
= 0.8 x 0.2 x 0.8
= 0.128

Trial 3 = Unsolved 1st, unsolved 2nd, and solved 3rd
= 0.8 x 0.8 x 0.2
= 0.128

Total (for the three trials):

= 0.128 + 0.128 + 0.128

= 0.384

Alternatively, we can apply the information in the binomial probability formula, as follows:

$$P = \binom{N}{x} p^x (1-p)^{N-x}$$

where:

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

Or

```
>> binopdf(1,3,0.2)
ans =
   0.3840
>>
```

Reminder

- We will have our 2nd quiz next Tuesday after Lecture 6.
- This is what our first mid-term exam will cover (Lectures 1~6).