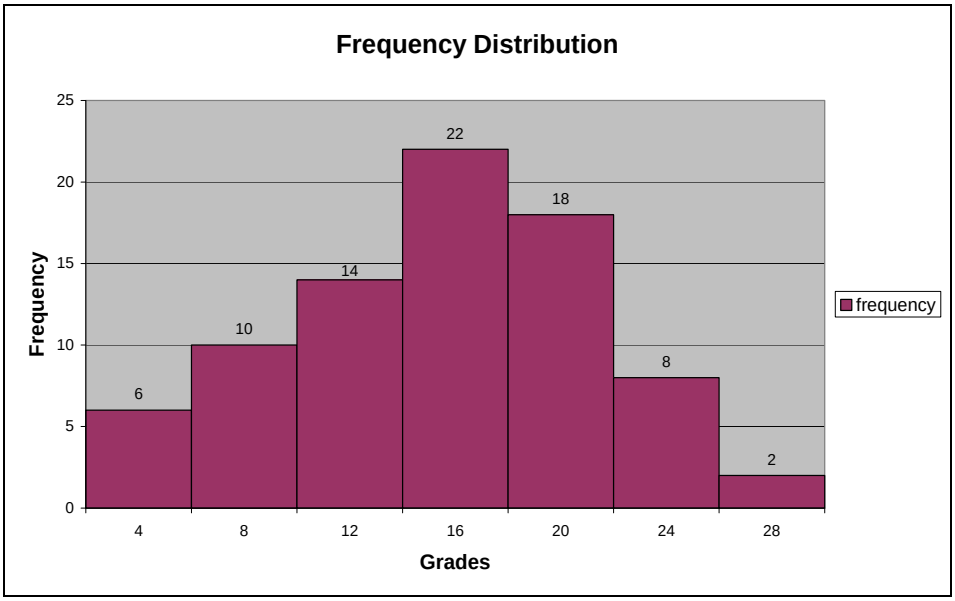
LECTURE 01

INTRODUCTION AND SAMPLING DISTRIBUTIONS

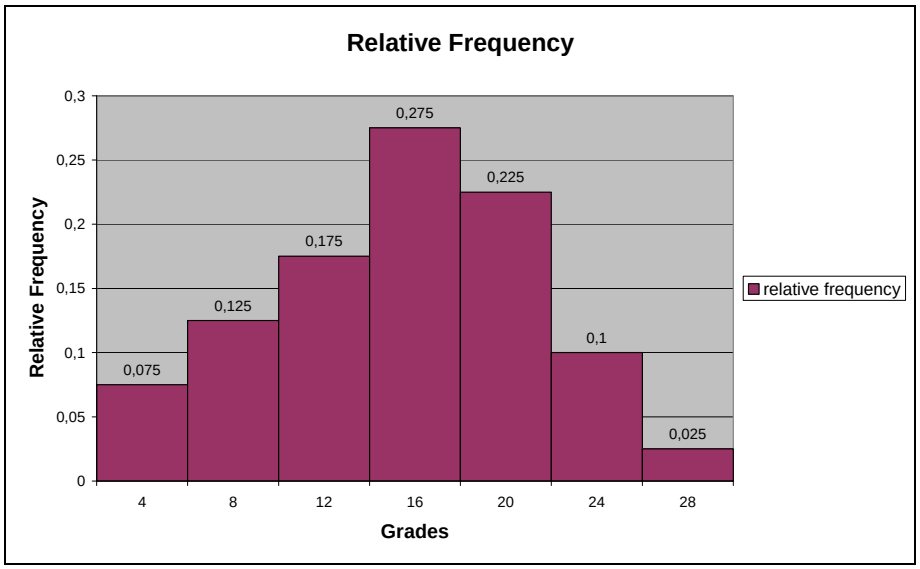
Outline of today's lecture:

I. Refreshment: Relative Frequency Density and Probability Distribution

A **frequency distribution** is a tabular summary of a set of data showing the frequency (or number) of items in each of several non-overlapping classes.

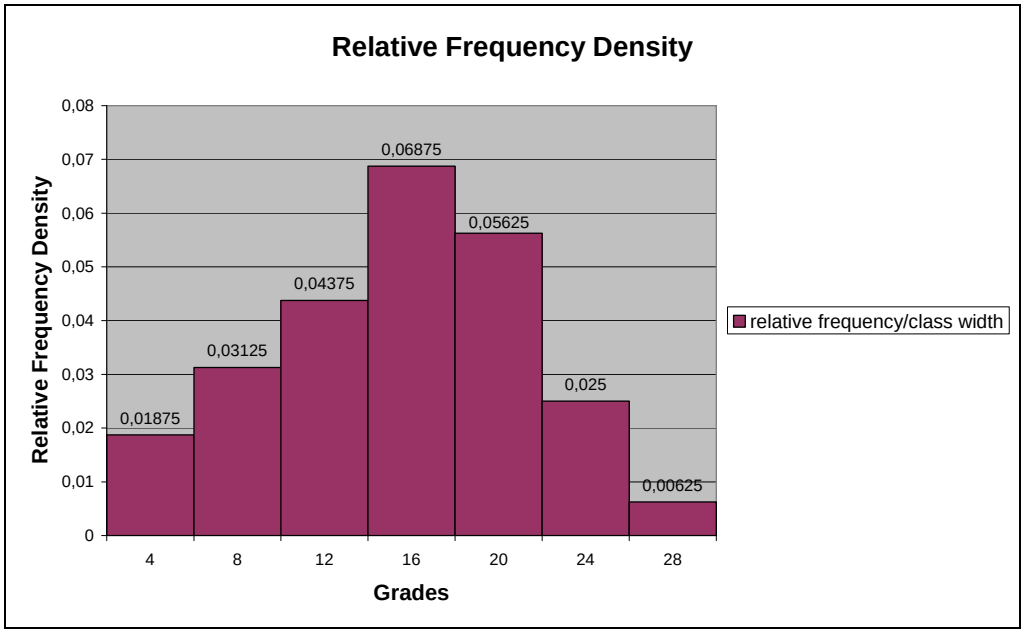


We can also graph *relative frequency* (*f/N*) using a bar graph, where *f* denotes *frequency* and *N* refers to the *sum of all frequencies*.



It is convenient to change the vertical scale to *relative frequency density*, which makes the total area (the sum of all the areas of the bars) equal to 1.

- We can call this the *probability distribution*.

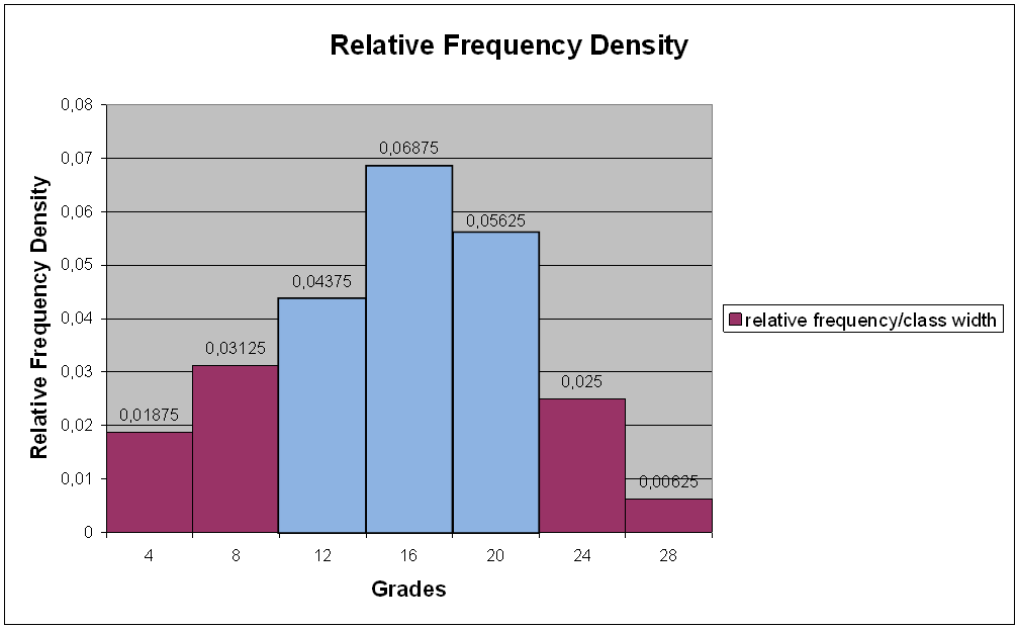


$$\text{Relative Frequency Density} = \frac{relative\ frequency}{class\ width} = \frac{f\ /\ N}{class\ width}$$

value	class-width	frequency	relative frequency	relative frequency/class width
4	4	6	0,075	0,01875
8	4	10	0,125	0,03125
12	4	14	0,175	0,04375
16	4	22	0,275	0,06875
20	4	18	0,225	0,05625
24	4	8	0,1	0,025
28	4	2	0,025	0,00625
			1	

The histogram can now be used to *find* other experimental *probabilities*.

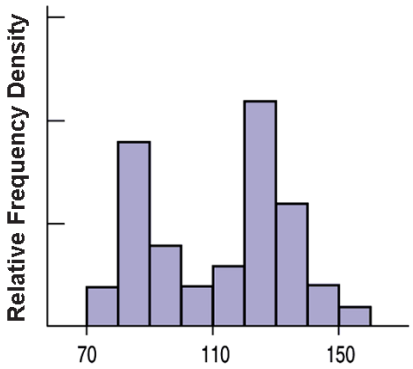
- Suppose we select a student from the class randomly. What would be the probability that this student has a grade between 12 and 20?



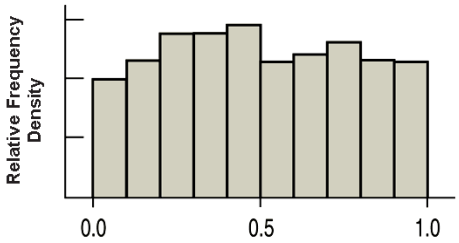
$$P(12 \le Grades < 20) = 4 \times 0.04375 + 4 \times 0.06875 + 4 \times 0.05625 = 0.675$$

Attention The distributions may have very different shapes; it <u>does not have to</u> be a *symmetric distribution* as we have sketched above.

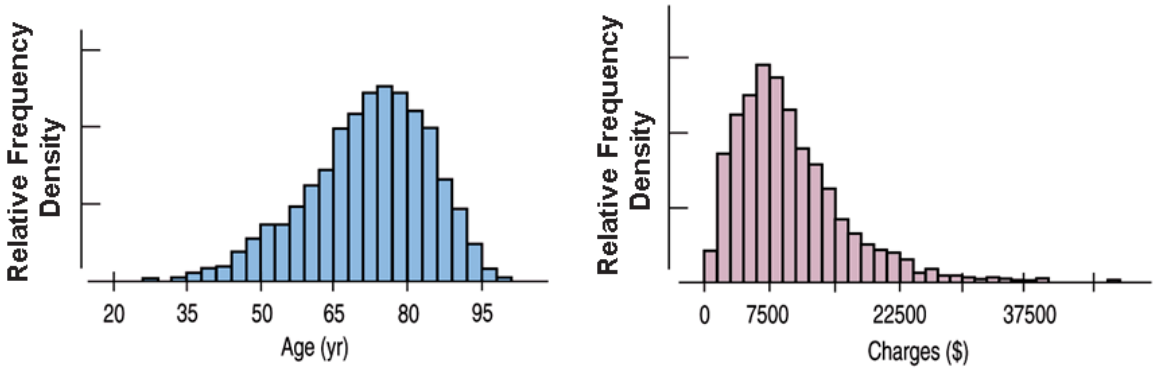
- A bimodal distribution has two different peaks:



- A histogram that doesn't appear to have any mode and in which all the bars are approximately the same height is called *uniform*:

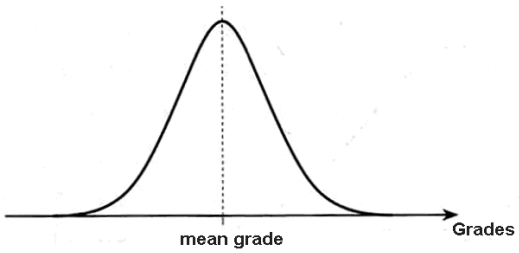


- The (usually) thinner ends of a distribution are called the *tails*. If one tail stretches out farther than the other, the histogram is said to be *skewed* to the side of the longer tail.
    - In the figure below, the histogram on the left is said to be *skewed left*, while the histogram on the right is said to be *skewed right*. They are *asymmetric distributions*.

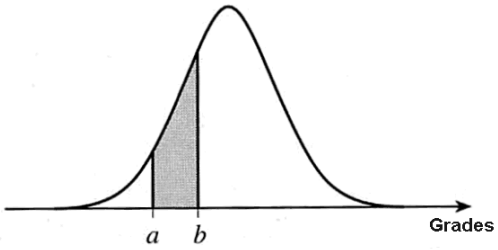


We can improve the ***grade example*** by collecting more data (*N* increases) and reducing the class width.

- What happens as a result of this process?
  - Because the area remains fixed at 1, the *relative frequency density* becomes approximately a smooth (continuous) curve.



  - We call the function of this ***continuous*** distribution, the *probability density function*.
  - In this case, the probability that the grade lies between *a* and *b* is given by the area under the curves between *a* and *b*.



- **Attention!** As we have stated before, the shape of the *resulted* distribution does not have to be symmetric; the distribution may have a bimodal, uniform, asymmetric (left-skewed or left-skewed), etc., shape.

II. Refreshment: Expected values (mathematical expectation)

Expected value expected value of a random variable is defined as its mean value.

- It could also be described as *'the value you would expect on average*'.
- The formula for calculating expected values follows directly from the formula calculating the mean value for grouped data.
- The expected value is a very important mathematical cool in the theory of statistics, it is used in many applications such as the calculation of expected profits and losses and risk analysis.

A. Formula for mean and variance in terms of expected values

1. Discrete random variables

The formula for calculating the mean or expected value for a random variable, the mean for grouped data

$$\mu = \frac{\sum f_i x_i}{N} \qquad where \ \sum f_i = N$$

$$\mu = \sum \left(\frac{f_i}{N} x_i \right)$$

$$\mu = \sum P(x_i)(x_i) \qquad \text{...since } P(x_i) = \frac{f_i}{N} \text{ for large } N$$

$$\mu = \sum x_i P(x_i)$$

The expected value (average value) of the random variable, *X*, is:

$$\mu = E(X) = \sum xP(x)$$

The expected value of any function (or formula) of the random variable is defined as

$$E(g(X)) = \sum g(x)P(x)$$

where $g(X)$ is the function or formula in X; for example $E(X^2) = \sum x^2 P(x)$

The variance of a random variable is a measure of variation of its values about the mean, *E(X)*. The formula for variance in terms of expected values, the variance for grouped data as follows:

$$\text{variance} = \sum \left[\left(x_i - \mu \right)^2 \frac{f_i}{N} \right]$$

$$\text{variance} = \sum \left[\left(x_i - \mu \right)^2 P(x_i) \right] \qquad \text{...since } P(x_i) = \frac{f_i}{N} \text{ when } N \text{ is large}$$

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

Hence the formula for the variance of a random variable is:

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

or, an alternative formula is:

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

Example

An experiment involves throwing a fair die.

- Calculate the expected value and variance for the random variable *X* where *X* is the number that shows when the die is thrown.

Solution

To calculate expected values for a random variable we first require its probability distribution.

Step 1: list every possible outcome of the experiment and the corresponding probability (columns one and two in the table below).

x	**1**	**2**	**3**	**4**	**5**	**6**
P(X=*x*)	1/6	1/6	1/6	1/6	1/6	1/6

Table Probability distribution of the outcome of a single throw of a dice.

Note: For notational simplicity P(X=x) is generally written as P(x)

The Expected value is calculated as follows:

$$E(X)=\sum xP(x)$$

$$E(X)=1\left(\frac{1}{6}\right)+2\left(\frac{1}{6}\right)+3\left(\frac{1}{6}\right)+4\left(\frac{1}{6}\right)+5\left(\frac{1}{6}\right)+6\left(\frac{1}{6}\right)=\frac{21}{6}$$

$$E(X)=3.5$$

In column 4, use the expected value, $E(X)=\mu=3.5$ in the calculation of variance by formula given above.

Calculation of expected values

x	$P(x)$	$xP(x)$	$(x-\mu)^2P(x)$	$x^2P(x)$
1	$\frac{1}{6}$	$1\times\frac{1}{6}=0.166667$	$(1-3.5)^2=1.041667$	$1^2\times\frac{1}{6}=0.166667$
2	$\frac{1}{6}$	$2\times\frac{1}{6}=0.333333$	$(2-3.5)^2=0.375000$	$2^2\times\frac{1}{6}=0.666667$
3	$\frac{1}{6}$	$3\times\frac{1}{6}=0.500000$	$(3-3.5)^2=0.041667$	$3^2\times\frac{1}{6}=1.500000$
4	$\frac{1}{6}$	$4\times\frac{1}{6}=0.666667$	$(4-3.5)^2=0.041667$	$4^2\times\frac{1}{6}=2.666667$
5	$\frac{1}{6}$	$5\times\frac{1}{6}=0.833333$	$(5-3.5)^2=0.375000$	$5^2\times\frac{1}{6}=4.166667$
6	$\frac{1}{6}$	$6\times\frac{1}{6}=1.000000$	$(6-3.5)^2=1.041667$	$6^2\times\frac{1}{6}=6.000000$
Totals	1	3.5	2.916667	15.16667

$E(X)=\sum xP(x)$ $V(X)=\sum(x-\mu)^2P(x)$ $E(X^2)=\sum x^2P(x)$

Alternatively, calculate variance formula $V(X)=E\left(X^2\right)-\left[E\left(X\right)\right]^2$. So, first calculate $E(X^2)=\sum x^2P(x)$, column five, then apply formula $V(X)=E\left(X^2\right)-\left[E\left(X\right)\right]^2$.

$$V(X) = E\left(X^2\right) - \left[E\left(X\right)\right]^2 \rightarrow V(X) = 15.16667 - \left[3.5\right]^2 = 2.91667$$

Verbal explanation One each throw of a die numbers from one to six may turn up in any random order such as 3, 4, 6, 3, 1, 5, 4, 2…Over a large number of trials, the average of these values is expected to be 3.5.

2. Continuous random variables

Expressions for the mean and variance of continuous distributions are given as follows:

The mean value: $E(X) = \sum xP(x) = \int_{-\infty}^{\infty} xf(x)dx$

The variance: $V(X) = E\left(X - \mu\right)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$

You notice that the *summation* ($\sum$) is replaced by the *integral* sign $\int$ and *P(x)* is replaced by the *probability density function*, *f(x)*.

III. Sampling Distribution of Mean

A. An Example for Sampling Distribution

Population:

A	B	C	D	E
3	1	5	6	2

Population mean, μ: (3+1+5+6+2)/5=17/5=3.4

Population Size, *N*: 5

From this population of five, list every possible sample of size 2 (n=2) and calculate the mean for each sample.

- Repeat this for samples of size 3 (n=3).

How many samples of size 2? → $\binom{5}{2} = \frac{5 \times 4}{1 \times 2} = 10$

How many samples of size 3? → $\binom{5}{3} = \frac{5 \times 4 \times 3}{1 \times 2 \times 3} = 10$

Samples	**Sample of Size 2 (n=2)**	**Sample Mean**
1	AB=3,1 (y_1=3, y_2=1) for Y_1 and Y_2	$\overline{Y}$ =2.0
2	AC=3,5	4.0
3	AD=3,6	4.5
4	AE=3,2	2.5
5	BC=1,5	3.0
6	BD=1,6	3.5
7	BE=1,2	1.5
8	CD=5,6	5.5
9	CE=5,2	3.5
10	DE=6,2	4.0
Average		3.4

Samples	**Sample of Size 3 (n=3)**	**Sample Mean**
Y_1	ABC=3,1,5 (y_1=3, y_2=1, y_3=5) for Y_1, Y_2 and Y_3	3.00
Y_2	ABD=3,5,6	4.67
Y_3	ABE=3,1,2	2.00
Y_4	ACD=3,5,6	3.67
Y_5	ACE=3,5,2	3.33
Y_6	ADE=3,6,2	3.67
Y_7	BCD=1,5,6	4.00
Y_8	BCE=1,5,2	2.67
Y_9	BDE=1,6,2	3.00
Y_{10}	CDE=5,6,2	4.33
		3.4

Means of all sample means $\mu_{\overline{Y}}$ $\mu_{\overline{Y}}$ =3.4 For n=2 $\mu_{\overline{Y}}$ =3.4 For n=3

Conclusion The mean of all the sample means is the same as the population mean: $\mu_{\overline{Y}} = \mu$

B. Process Going Into the Sampling Distribution Model

- We start with a population model, which can have any shape. It can even be bimodal or skewed (as this one is). We label the mean of this model μ and its standard deviation, σ.

- We draw one real sample (solid line) of size n and show its histogram and summary statistics. We imagine (or simulate) drawing many other samples (dotted lines), which have their own histograms and summary statistics.

- We (imagine) gathering all the means into a histogram

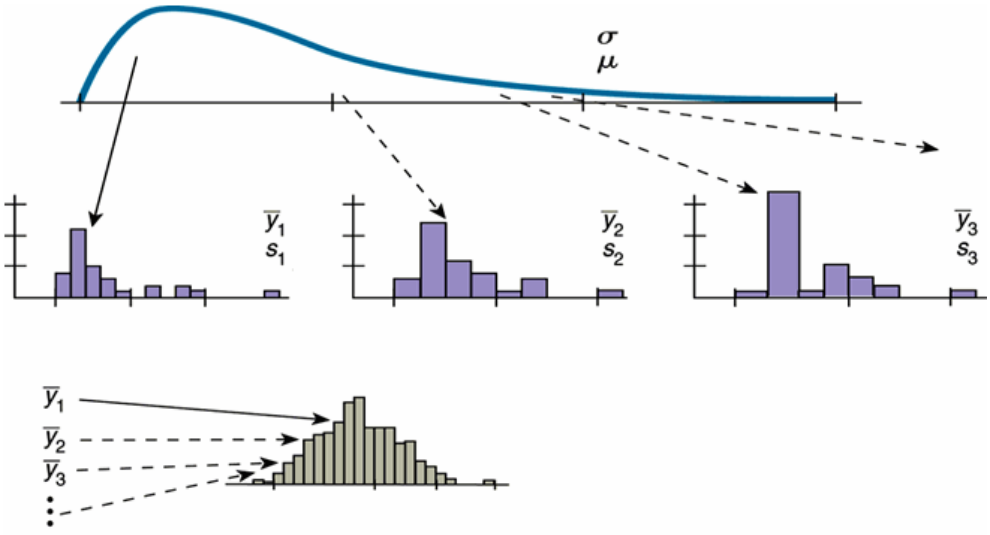


Figure 1 Process Going Into the Sampling Distribution Model