HUMBOLDT-UNIVERSITÄT ZU BERLIN

HUMBOLDT-UNIVERSITÄT ZU BERLIN
Resource Economics

Faculty of Life Sciences | Thaer-Institute

Introduction online course on Mathematics and Statistics

Preparatory Course for M.Sc. Integrated Natural Resource Management

3 Distributions

Department for Agricultural Economics | Resource Economics Group

HUMBOLDT-UNIVERSITÄT ZU BERLIN

HUMBOLDT-UNIVERSITÄT ZU BERLIN
Resource Economics

Syllabus

3 Distributions.

- 3.1 Organizing data. Frequency distribution
  - 3.1.1 Categorical frequency distribution
  - 3.1.2 Grouped frequency distribution
  - 3.1.3 Cumulative frequency distribution
  - 3.1.4 Graphical forms of distribution
- 3.2 Probability distribution
  - 3.2.1 Probability. Basic concepts
  - 3.2.2 Probability distributions
  - 3.2.3 Binomial distribution
  - 3.2.4 Mean, Variance, and Standard Deviation of a probability distribution
- 3.3 Normal distribution
- Questions of Repetition
- Answers to the exercises
- References

3.1 Organizing data. Frequency distribution

- The table on the right shows data on average annual relative humidity (%) for each state in the United States [3].
- Little information can be obtained from such raw data, which means that data are in original form, it makes sense to organize the data into so called a frequency distribution [2].
- One structure way of organizing such data is using a frequency distribution, which consists of ***classes*** and their corresponding ***frequencies***:

52	64	25	49	62
35	52	54	57	50
56	41	58	58	56
50	55	61	61	52
59	61	55	54	53
45	53	32	53	59
29	61	52	51	57
48	59	54	57	49
53	53	49	43	58
52	62	59	58	43

Class limits	Tally	Frequency
25-32	///	3
33-40	/	1
41-48	///	5
49-56	/// /// /// /// ////	24
57-64	/// /// /// //	17
		Total: 50

→ Each raw data value is placed into a quantitative or qualitative category called a **class.** The **frequency** of a class then is the number of data values contained in a specific class [2].

The classes in this distribution are 25–32, 33–40, etc. These values are called class limits. The data values 25, 32, 29 can be tallied in the first class; 35 in the second class; 41, 45, 48, 43, 43 in the third class and so on [2].

A ***frequency distribution*** is the organization of raw data in table form, using classes and frequencies [2].

3.1.1 Categorical frequency distribution

The ***categorical frequency distribution*** is used for data that can be placed in specific categories, such as nominal- or ordinal-level data [2].

Example: *Imagine you conduct a study on use of various soil conservation practices, such as no-tillage (NT), crop rotation (CR), mulching (M) and cross-slope farming (CSF). To get some actual data you carry out a survey and interview 30 farmers. Possible data set is represented next.*

NT CR CSF NT M
CR NT M NT NT
NT CR CSF CR CSF
CR CSF CR M NT
CSF M NT NT CSF
CR M CSF NT CR

Construct a frequency distribution for the data.

Step 1 Make a table as shown.

A Class	B Tally	C Frequency	D Percent
NT			
CR			
M			
CSF			

Step 2 Tally the data and place the results in column B.
Step 3 Count the tallies and place the results in column C.

3.1.1 Categorical frequency distribution

Step 4 Find the percentage of values in each class by using the formula [2]:

$\% = \frac{f}{n} \times 100\%$, where f - frequency of the class and n - total number of values.

For example, in the class of type NT practice , the percentage is $\% = \frac{10}{30} \times 100\% = 33\%$

Step 5 Find the totals for columns C (frequency) and D (percent) [2].

A Class	B Tally	C Frequency	D Percent
NT	𝍸 𝍸	10	33
CR	𝍸 ///	8	27
M	𝍸	5	17
CSF	𝍸 //	7	23
	Total	30	100

3.1.2 Grouped frequency distribution

When the range of the data is large, the data must be grouped into classes that are more than one unit in width, in what is called a ***grouped frequency distribution*** [2].

Example: *A distribution of average total yearly precipitation for 50 American states [3]:*

Class limits	Class boundaries	Tally	Frequency
240-412	239.5-412.5	𝍸 //	7
413-585	412.5-585.5	𝍸	5
586-758	585.5-758.5	𝍸	5
759-931	758.5-931.5	////	4
932-1104	931.5-1104.5	𝍸 𝍸	10
1105-1277	1104.5-1277.5	𝍸 ////	9
1278-1450	1277.5-1450.5	𝍸 /	6
1451-1623	1450.5-1623.5	////	4

- Values 240 and 412 of the first class are called *class limits.*
- The **lower class limit** is 240 and it represents the smallest data value that can be included in the class.
- The **upper class limit** is 412 and it represents the largest data value that can be included in the class [2].

➢ **How to find class boundaries?**

The class limits should have the same decimal place value as the data, but the class boundaries should have one additional place value and end in a 5 [2].

Lower limit - 0.5 = 240 - 0.5 = 239.5 = lower boundary
Upper limit + 0.5 = 412 + 0.5 = 412.5 = upper boundary

3.1.2 Grouped frequency distribution

- The **class width** for a class in a frequency distribution is found by subtracting the lower (or upper) class limit of one class from the lower (or upper) class limit of the next class [2]. For example, the class width in the preceding distribution on the average total yearly precipitation for American states is 173, found from 413 - 240 = 173.
- The class width can also be found by subtracting the lower boundary from the upper boundary for any given class [2]. In this case, 412.5 - 239.5 = 173.

Simple rules on how to construct a frequency distribution [2]**:**

1. *There should be between 5 and 20 classes.*
2. *It is preferable but not absolutely necessary that the class width be an odd number.*

- The **class midpoint** X_m is obtained by adding the lower and upper boundaries and dividing by 2, or adding the lower and upper limits and dividing by 2:

$$X_m = \frac{\text{lower boundary} + \text{upper boundary}}{2}$$

or

$$X_m = \frac{\text{lower limit} + \text{upper limit}}{2}$$

For example, the midpoint of the second class in the example with annual precipitation is

$$X_m = \frac{413 + 585}{2} = 499 \quad \text{or} \quad X_m = \frac{412.5 + 585.5}{2} = 499$$

3.1.2 Grouped frequency distribution

3. The classes must be mutually exclusive.

- Mutually exclusive classes should have non-overlapping class limits so that data cannot be placed into two classes [2]. Often one might come across frequency distributions such as:

Temperature
5 - 10
10 - 15
15 - 20
20 - 25

If the temperature is 15°C, into which class should it be placed? A better way to construct a frequency distribution is to use classes such as

Temperature
5 - 10
11 - 16
17 - 22
23 - 28

4. *The classes must be continuous.*

- Even if there are no values in a class, the class must be included in the frequency distribution. There should be no gaps in a frequency distribution [2].

3.1.2 Grouped frequency distribution

5. *The classes must be exhaustive.*

- There should be enough classes to accommodate all the data [2].

6. *The classes must be equal in width.*

- This avoids a distorted view of the data [2].
- Exception: open-ended distribution

Temperature	Frequency
5 - 10	5
11 - 16	2
17 - 22	7
23 - 28	4
29 and above	3

Precipitation	Frequency
Below 413	7
413 – 585	5
586 – 758	5
759 – 931	4
932 - 1104	10

- The frequency distribution for temperature is open-ended for the last class, which means that any temperature, which is equal to 29°C or is higher will be tallied in the last class. The distribution for precipitation is open-ended for the first class, meaning that any precipitation values below 413 mm will be tallied in that class[2].

3.1.2 Grouped frequency distribution

Constructing a Grouped Frequency Distribution [2]**:**

1) determining the classes, which includes

- finding the highest and lowest values → finding the range
- selecting the number of classes desired
- finding the width → the range / number of classes → rounding it up
- selecting a starting point (the lowest value or any convenient number less than the lowest value) → adding the width to get the lower limits
- finding the upper class limits
- finding the boundaries;

2) tallying the data;

3) finding the numerical frequencies from the tallies, and finding the cumulative frequencies [2].

Why to construct frequency distribution? [2]

1. To organize the data in a meaningful, intelligible way.
2. To determine the nature or shape of the distribution.
3. To facilitate computational procedures for measures of average and spread.
4. To enable the researcher to draw charts and graphs for the presentation of data.
5. To make comparisons among different data sets.

3.1.2 Grouped frequency distribution

Example: *The data below represent rice yields (kg/ha) in India over the period from 1950 till 2014 [1]. Construct a grouped frequency distribution for the data using 6 classes.*

668 714 764 902 820 874 900 790 930 937 1013 1028
931 1033 1078 862 863 1032 1076 1073 1123 1141 1070 1151
1045 1235 1089 1308 1328 1074 1336 1308 1231 1457 1417 1552
1471 1465 1689 1745 1740 1751 1744 1888 1911 1797 1882 1900
1921 1986 1901 2079 1744 2078 1984 2102 2131 2202 2178 2125
2239 2393 2462 2424

Step 1 Determine the classes [2].

- Find the highest and the lowest values: $H = 2462$ and $L = 668$.
- Calculate the range: R = highest value - lowest value = $H - L$, so $R = 2462 - 668 = 1794$
- Select the number of classes desired (usually between 5 and 20). In this case, 6 is arbitrarily chosen.
- Find the class width by dividing the range by the number of classes:

$$\text{Width} = \frac{\text{R}}{\text{number of classes}} = \frac{1794}{6} = 299$$

- Round the answer up to the nearest whole number if there is a remainder. But, in this case no remainder → adding an extra class to accommodate all the data.
- Select a starting point for the lowest class limit. This can be the smallest data value or any convenient number less than the smallest data value. In this case, 668 is used. Add the width to the lowest score taken as the starting point to get the lower limit of the next class. Keep adding until there are 6 classes.

3.1.2 Grouped frequency distribution

- Subtract one unit from the lower limit of the second class to get the upper limit of the first class. Then add the width to each upper limit to get all the upper limits.
  967 - 1 = 966
  →The first class is 668–966, the second class is 967–1265, etc.
- Find the class boundaries by subtracting 0.5 from each lower class limit and adding 0.5 to each upper class limit:

667.5–966.5, 104.5–109.5, etc.

Step 2

Tally the data [2].

Step 3

Find the numerical frequencies from the tallies [2].

The completed frequency distribution is:

Class limits	Class boundaries	Tally	Frequency
668-966	667.5-966.5	𝍸 𝍸 ///	13
967-1265	966.5-1265.5	𝍸 𝍸 𝍸 /	16
1266-1564	1265.5-1564.5	𝍸 ////	9
1565-1863	1564.5-1863.5	𝍸 //	7
1864-2162	1863.5-2162.5	𝍸 𝍸 ///	13
2163-2461	2162.5-2461.5	𝍸	5
2462-2760	2461.5-2760.5	/	1
			n = $\sum f$ = 64

3.1.3 Cumulative frequency distribution

A ***cumulative frequency distribution*** is a distribution that shows the number of data values less than or equal to a specific value (usually an upper boundary) [2].

- The values are found by adding the frequencies of the classes less than or equal to the upper class boundary of a specific class. This gives an ascending cumulative frequency. In the previous example, the cumulative frequency for the first class is 0 + 13 = 13; for the second class it is 0 + 13 + 16 = 29.
- The cumulative frequency distribution for the data of rice yields in India [1]:

	Cumulative frequency
Less than 667.5	0
Less than 966.5	13
Less than 1265.5	29
Less than 1564.5	38
Less than 1863.5	45
Less than 2162.5	58
Less than 2461.5	63
Less than 2760.5	64

When the range of the data values is relatively small, a frequency distribution can be constructed using single data values for each class. This type of distribution is called an ***ungrouped frequency distribution*** [2].

3.1. Organizing data. Frequency distribution

Exercise #1: Find the class boundaries, midpoints, and widths for the below classes.

a. 26–34
b. 4.56-7.81
c. 135.8–167.2

Exercise #2: U.S. Population (x 1000 inhabitants) of 40 biggest world cities are presented below [9]. Construct a grouped frequency distribution and a cumulative frequency distribution for the data using 8 classes.

37,843	20,365	12,084	7,217
20,597	14,667	7,246	4,889
4942	30,539	14,998	10,920
6,710	10,858	20,063	24,998
5,246	5,624	15,058	7,275
24,123	5,816	7,509	8,442
10,376	20,630	17,712	15,600
23,480	17,444	9,156	6,155
10,236	5,977	23,416	15,669
16,170	6,078	22,123	21,009

3.1.4 Graphical forms of distribution

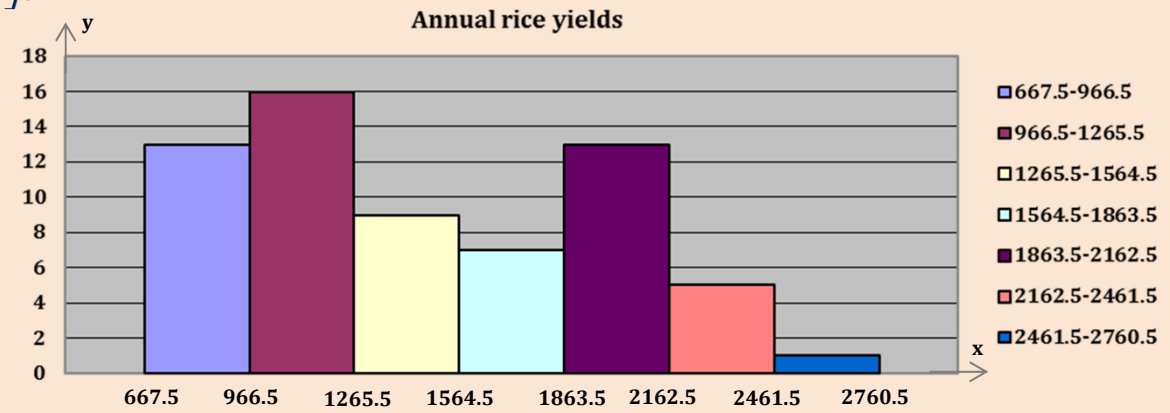
The three most commonly used graphs in research are:

- **Histograms**

The ***histogram*** is a graph that displays the data by using contiguous vertical bars (unless the frequency of a class is 0) of various heights to represent the frequencies of the classes [2].

Example: *Let us construct a histogram to represent the data shown for rice yields (kg/ha) in India over the period from 1950 till 2014 [1].*

Class boundaries	Frequency
667.5-966.5	13
966.5-1265.5	16
1265.5-1564.5	9
1564.5-1863.5	7
1863.5-2162.5	13
2162.5-2461.5	5
2461.5-2760.5	1



Step 1 *Draw and label the* ***x*** *and* ***y*** *axes.*
Step 2 *Represent the frequency on the* ***y*** *axis and the class boundaries on the* ***x*** *axis.*
Step 3 *Using the frequencies as the heights, draw vertical bars for each class.*
As the histogram shows, the class with the greatest number of data values (16) is 966.5–1265.5. The graph also has two peaks with the data clustering around it.

3.1.4 Graphical forms of distribution

- **Frequency poligons**

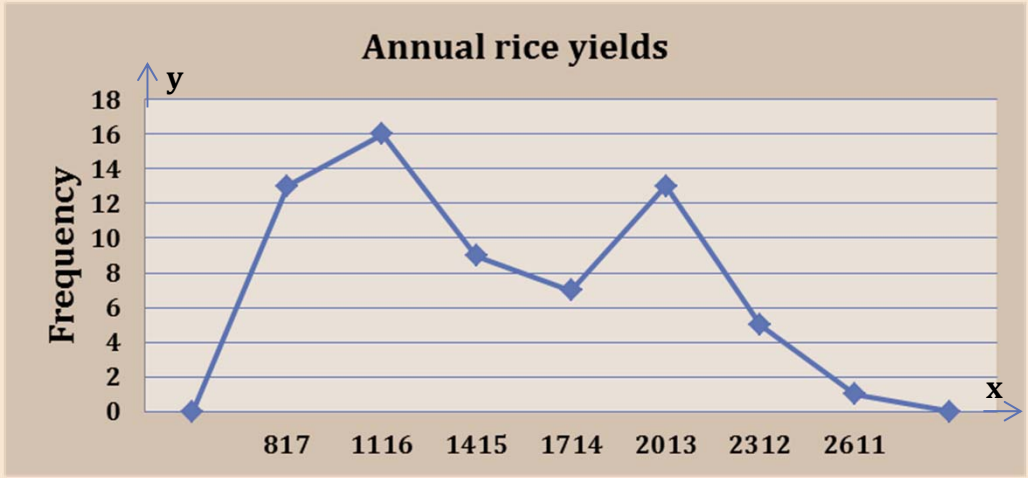
The ***frequency polygon*** is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoints of the classes. The frequencies are represented by the heights of the points [2].

Example: *Let us construct a frequency polygon using the same frequency distribution on rice yields (kg/ha) in India.*

Step 1 *Finding the midpoints of each class by adding the upper and lower boundaries and dividing by 2:*

$$\frac{667.5 + 966.5}{2} = 817$$

Class boundaries	Midpoints	Frequency
667.5-966.5	817	13
966.5-1265.5	1116	16
1265.5-1564.5	1415	9
1564.5-1863.5	1714	7
1863.5-2162.5	2013	13
2162.5-2461.5	2312	5
2461.5-2760.5	2611	1



Step 2 *Drawing and marking the axes: using the midpoints for the* ***x*** *values and the frequencies for the* ***y*** *values.*

Step 3 *Plot the points.*

Step 4 *Connect adjacent points with line segments. Draw a line back to the* ***x*** *axis at the beginning and end of the graph, at the same distance that the previous and next midpoints would be located [2].*

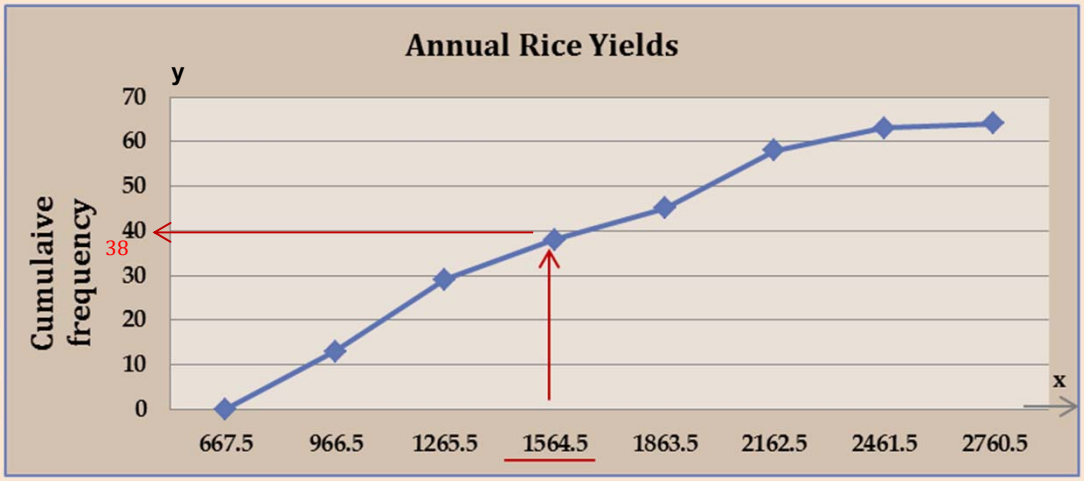
3.1.4 Graphical forms of distribution

- **Ogives**

The ***ogive*** is a graph that represents the cumulative frequencies for the classes in a frequency distribution [2].

Example: *Using the same frequency distribution of annual rice yields in India during the period of 1950-2014 [1], construct an ogive.*

	Cumulative frequency
Less than 667.5	0
Less than 966.5	13
Less than 1265.5	29
Less than 1564.5	38
Less than 1863.5	45
Less than 2162.5	58
Less than 2461.5	63
Less than 2760.5	64



Step 1 *Find the cumulative frequency for each class.*

Step 2 *Draw the x and y axes. Label the x axis with the class boundaries. Use an appropriate scale for the y axis to represent the cumulative frequencies. (Depending on the numbers in the cumulative frequency columns, scales such as 0, 1, 2, 3, . . . , or 5, 10, 15, 20, . . . , or 1000, 2000, 3000, . . . can be used. Do not label the y axis with the numbers in the cumulative frequency column.) In this example, a scale of 10, 20, 30, etc. is used.*

Step 3 *Plot the cumulative frequency at each upper class boundary, using spots as shown on a graph.*

Step 4 *Starting with the first upper class boundary, 966.5, connect adjacent points with line segments, as shown on a graph. Then extend the graph to the first lower class boundary, 667.5, on the x axis [2].*

3.1.4 Graphical forms of distribution

Exercise #3:

The data on methane and nitrous oxide emissions (kt of CO2 equivalent) stemming from agricultural and industrial production within the EU countries in 2010 are shown below [6]. Construct grouped frequency distributions and a histogram for each set of data, compare your results.

EU-28 countries	NO2 Emissions	CH4 emissions
Austria	3 759	8 391
Belgium	10 127	9 633
Bulgaria	4 479	12 011
Croatia	2 943	5 036
Cyprus	315	621
Czech Republic	7 291	12 033
Denmark	5 410	7 763
Estonia	912	2 329
Finland	5 818	8 896
France	38 668	83 753
Germany	42 432	57 230
Greece	5 118	8 417
Hungary	4 215	7 283
Ireland	7 716	13 896
Italy	19 632	37 548
Latvia	1 383	3 227
Lithuania	4 597	5 052
Luxembourg	476	1 236
Malta	61	235
Netherlands	9 205	20 269
Poland	26 758	65 453
Portugal	4 292	12 601
Romania	8 808	26 146
Slovakia	3 380	3 985
Slovenia	1 169	2 902
Spain	22 551	36 824
Sweden	5 629	10 845
United Kingdom	26 536	61 174

3.2 Probability distribution

Probability as a general concept can be defined as the chance of an event occurring [2].

Example: *Tossing a coin: when a coin is tossed, there are two possible outcomes: heads or tails. These two possible outcomes seem not to be distinguishable in any way that affects which side will land up or down.* → *The probability of heads is taken to be 1/2, as is the probability of tails.*

In general, if there are N symmetrical outcomes, the probability of any given one of them occurring is taken to be 1/N. Thus, if a six-sided die is rolled, the probability of any one of the six sides coming up is 1/6. What is the probability that either a two or a five will come up? The two outcomes about which we are concerned (a two or a five coming up) are called favorable outcomes. Given that all outcomes are equally likely, we can compute the probability of a two or a five using the formula [2], [5]:

$$\text{Probability} = \frac{\text{number of favorable outcomes}}{\text{number of possible outcomes}}$$

A ***probability experiment*** is a chance process that leads to well-defined results called outcomes [2].

An ***outcome*** is the result of a single trial of a probability experiment [2].

Set of all possible outcomes is called the ***sample space*** [2].

An ***event*** consists of a set of outcomes of a probability experiment [2].

3.2.1 Probability. Basic concepts

Basic probability axioms [2]**:**

1. The probability of any event E is a number (either a fraction or decimal) between and including 0 and 1 → $0 \leq P(E) \leq 1$.

2. If an event E cannot occur (i.e., the event contains no members in the sample space) → $E = 0$.

3. If an event E is certain → $P(E) = 1$.

4. The sum of the probabilities of all the outcomes in the sample space is 1.

Example: *When a single die is rolled the probability of getting a number less than 7 [P(7)] is equal to one since all outcomes—1, 2, 3, 4, 5, and 6—are less than 7. → The event of getting a number less than 7 is certain.*

However, the probability of getting a 9 [P(9)] equals to zero since the sample space is 1, 2, 3, 4, 5, and 6, therefore it is impossible to get a 9.

3.2.1 Probability. Basic concepts

Probability of two (or more) independent events [2]**:**
Events A and B are independent events if the probability of Event B occurring is the same whether or not Event A occurs.

Example: *If a coin is tossed two times, the probability that a head comes up on the second toss still remains ½ regardless of whether or not a head came up on the first toss. The two events are (1) first toss is a head and (2) second toss is a head and these events are independent.*

- Probability of events A and B: If events A and B are independent, then the probability of both A and B occurring is:

$$P(A \text{ and } B) = P(A) \times P(B),$$

where P(A and B) is the probability of events A and B both occurring, P(A) is the probability of event A occurring, and P(B) is the probability of event B occurring.

Example: *If one flips a coin and rolls a six-sided die, what is the probability that the coin comes up tails and the die comes up 3? Since the two events are independent, the probability is simply the probability of a tail (which is 1/2) times the probability of the die coming up 3 (which is 1/6). Therefore, the probability of both events occurring is 1/2 x 1/6 = 1/12.*

- Probability of A or B: If Events A and B are independent, the probability that either Event A or Event B occurs is: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

3.2.1 Probability. Basic concepts

Example: *If a coin is tossed twice, what is the probability that one gets a head on the first toss or a head on the second toss (or both)? If Event A is a head on the first toss and Event B is a head on the second toss, then P(A) = ½, P(B) = ½, and P(A and B) = ¼ → P(A or B) = ½ + ½ - ¼ = ¾.*

Conditional probability: what is the probability that two cards drawn at random from a deck of playing cards will both be kings [2]?

- Once the first card chosen is a king, the probability that the second card chosen is also a king is called the conditional probability of drawing a king. In this case, the "condition" is that the first card drawn is a king. Symbolically it can be written as: P(a king on second draw | a king on the first draw).

- Since after a king is drawn on the first draw, there are 3 kings out of 51 total cards left. This means that the probability that one of the kings will be drawn is 3/51 = 1/17.

- If Events A and B are not independent, then P(A and B) = P(A) x P(B|A). Applying this to the problem, the probability of drawing two kings from a deck is 4/52 x 3/51 = 1/221 [5].

3.2.2 Probability distribution

Example: *When four coins are tossed, the sample space is represented as HHHH, THHH, HHHT, HTHH, HHTH, ´TTHH, THTH, HHTT, HTHT, HTTT, TTTH, THTT, TTHT, TTTT; and if X is the random variable for the number of tails, then X assumes the value 0, 1, 2, 3 or 4. Probabilities for the values of X can be determined as follows:*

No tails	One tail	Two tails	There tails	Four tails
HHHH	THHH HHHT HTHH HHTH	HTTT TTTH THTT TTHT	TTHH THTH HHTT HTHT	TTTT
$\frac{1}{14}$	$\frac{4}{14}$	$\frac{4}{14}$	$\frac{4}{14}$	$\frac{1}{14}$

Number of tails	Probability
0	1/14
1	4/14
2	4/14
3	4/14
4	1/14

A ***discrete probability distribution*** consists of the values a random variable can take and the corresponding probabilities of the values. The probabilities are determined theoretically or by observation [2].

Requirements for probability distribution [2; 5]:

1. The sum of the probabilities of all the events in the sample space must equal 1; that is, $\sum P(X) = 1$.
2. The probability of each event in the sample space must be between or equal to 0 and 1. That is, $0 \leq P(X) \leq 1$.

3.2.3 Binomial distribution

A ***binomial experiment*** is a probability experiment that satisfies the following four conditions [2]:

1. There must be a fixed number of trials.
2. There are only two outcomes.
3. Each trial is independent of the others.
4. The probability of each outcome must remain the same for each trial.

The outcomes of a binomial experiment and the corresponding probabilities of these outcomes are called a ***binomial distribution*** [2].

- Notation for binomial distribution [2]:

$P(S)$: The symbol for the probability of success
$P(F)$: The symbol for the probability of failure
p: The numerical probability of a success
q: The numerical probability of a failure
$P(S) = p$ and $P(F) = 1 - p = q$
n The number of trials
X The number of successes in n trials
Note that $0 \leq X \leq n$ and $X = 0, 1, 2, 3, \ldots, n$.

- In a binomial experiment, the probability of exactly X successes in n trials [2]:

$$P(X) = \frac{n!}{(n-X)!X!} \times p^X \times q^{n-X}$$

3.2.3 Binomial distribution

Example: *According to the UN's report "World's Women 2015: Trends and Statistics" only 18 % of ministers are women, and are usually assigned to portfolios related to social issues [7]. If 4 ministers are randomly selected, find the probability that* ***exactly*** *2 of them are women.*

Solution: *In this case using the formula for binomial probability, $p = 0.18$, $q = 1 - 0.18 = 0.82$, $n = 4$, and $X = 2$. Hence,*

$$P(2) = \frac{4!}{(4-2)!\,2!}(0.18)^2(0.82)^2 = 0.13$$

Example: *According to the UN's report "World's Women 2015: Trends and Statistics" only 18 % of ministers are women, and are usually assigned to portfolios related to social issues [7]. If 4 ministers are randomly selected, find the probability that* ***at least*** *2 of them are women.*

Solution: *In this case in order to find the probability that at least 2 of 4 selected ministers are women, $X = 2, 3, 4$. Therefore, we have to find the probabilities of 2, 3 and 4 ministers being women and sum up them. Hence,*

$$P(3) = \frac{4!}{(4-3)!\,3!}(0.18)^3\,0.82 = 0.019 \qquad P(4) = \frac{4!}{(4-4)!\,4!}(0.18)^4(0.82)^0 = 0.001$$

→ $P(2) + P(3) + P(4) = 0.13 + 0.019 + 0.001 = 0.15$

3.2.3 Binomial distribution

Example: *According to the UN's report "World's Women 2015: Trends and Statistics" only 18 % of ministers are women, and are usually assigned to portfolios related to social issues [7]. If 4 ministers are randomly selected, find the probability that* ***at most*** *2 of them are women.*

Solution: *In this case at most 2 out of 4 selected ministers being women means 0, or 1, or 2 → X = 0, 1, 2. Therefore, we have to find the probabilities of 0, 1 and 2 and sum up them. Hence,*

$$P(0) = \frac{4!}{(4-0)!\,0!}(0.18)^0(0.82)^4 = 0.452 \qquad P(1) = \frac{4!}{(4-0)!\,0!}(0.18)^1(0.82)^3 = 0.397$$

$$P(0) + P(1) + P(2) = 0.452 + 0.397 + 0.13 = 0.979$$

Example: *According to the UN's report "World's Women 2015: Trends and Statistics" only 18 % of ministers are women, and are usually assigned to portfolios related to social issues [7]. If 4 ministers are randomly selected, find the probability that* ***fewer than*** *2 of them are women.*

Solution: *In this case we need simply to subtract from 1 the probability that at least 2 out of 4 chosen ministers are women.*

$$P(0) + P(1) = 1 - 0.15 = 0.85$$

3.2.3 Binomial distribution

Exercise #4:
A Master program "Integrated Natural Resource Management" has 30 students. The ages of these students are as follows: three students are 27 years old, four are 26, ten are 25, two are 30, eight are 24, and three are 28. Let ***x*** be the age of any student (randomly selected). Find the probability distribution for ***x***. Explain why the distribution represents a probability distribution.

Exercise #5: A global Synovate survey conducted in 2008 in association with BBC World, 68% of consumers are willing to use less packaging and bags [8]. If you choose 10 people at random, what will the probability that

a) all of them are willing to use less packaging and bags;
b) more than one-half are willing to use less packaging and bags;
c) exactly 4 of them are willing to use less packaging and bags.

3.2.4 Mean, Variance, and Standard Deviation of a probability distribution

- The mean for a sample or population: $\overline{X} = \frac{\sum X}{n}$ $\mu = \frac{\sum X}{N}$

Example: *Suppose three coins are tossed repeatedly, and the sample space is represented as HHH, HTH, THH, TTT, TTH, THT, HHT, HTT. What will be the mean of the number of tails?*
Initially, each outcome has a probability of 1/8. Over time, one might expect three tails (TTT) to occur approximately 1/8 of the time, two tails (TTH, THT, HTT) then 3/8 of the time, one tail (HTH, THH, HHT) also 3/8 of the time and no tails (HHH) to occur approximately 1/8 of the time. Thus, on average, the expected number of tails will be 1/8 × 3 + 3/8 × 2 + 3/8 × 1 + 1/8 × 0 = 1,5.
Therefore, if it were possible to flip the coins an infinite number of times, the average of the number of tails would be 1,5.

→ So in order to compute the mean for a probability distribution, each possible outcome must be multiplied by its corresponding probability and then the products must be constructively added [2].

Formula for the Mean of a Probability Distribution [2]**:**
The mean of a random variable with a discrete probability distribution is

$$\mu = X_1 \cdot P(X_1) + X_2 \cdot P(X_2) + X_3 \cdot P(X_3) + \ldots + X_n \cdot P(X_n) = \sum X \cdot P(X)$$

where $X_1, X_2, X_3, \ldots, X_n$ are the outcomes and $P(X_1), P(X_2), P(X_3), \ldots, P(X_n)$ are the corresponding probabilities.

3.2.4 Mean, Variance, and Standard Deviation of a probability distribution

- Variance and standard deviation for a sample or population [2]:

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N - 1} \quad \text{or} \quad \sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$$

- To find the variance for the random variable of a probability distribution [2]:
  - subtract the theoretical mean of the random variable from each outcome and square the difference;
  - then multiply each difference by its corresponding probability and add the products.

$$\sigma^2 = \sum[(X - \mu)^2 \times P(X)]$$

- Formula for the variance of a probability distribution [2]:
  - find the variance of a probability distribution by multiplying the square of each outcome by its corresponding probability, summing those products, and subtracting the square of the mean.
  - the formula for the variance of a probability distribution is:

$$\sigma^2 = \sum[X^2 \times P(X)] - \mu^2$$

- The standard deviation of a probability distribution is [2]

$$\sigma = \sqrt{\sigma^2} \quad \text{or} \quad \sqrt{\sum[X^2 \times P(X)] - \mu^2}$$

3.2.4 Mean, Variance, and Standard Deviation of a probability distribution

Example: *A bag contains 10 chips. 3 of the chips are numbered one, 5 of the chips are numbered two, and 2 of the chips are numbered 3. One chip is selected randomly, its number is recorded. Then it is replaced. If the action is repeated many times, find the mean, the variance and standard deviation of the numbers on the chips.*

Solution:

- *Let X be the number on each chip. Then the probability distribution is*

Number on chip (X)	1	2	3
Probability P(X)	0.3	0.5	0.2

- *The mean is* $\mu = \sum X \cdot P(X) = 1 \times 0{,}3 + 2 \times 0{,}5 + 3 \times 0{,}2 = 1{,}9$
- *The variance is* $\sigma^2 = \sum[X^2 \times P(X)] - \mu^2 = (1 \times 0{,}3 + 4 \times 0{,}5 + 9 \times 0{,}2) - 3{,}61 = 0{,}49$
- *The standard deviation is* $\sigma = \sqrt{\sigma^2} = \sqrt{0{,}49} = 0{,}7$

- *However, the mean, variance, and standard deviation can also be computed by using vertical columns:*

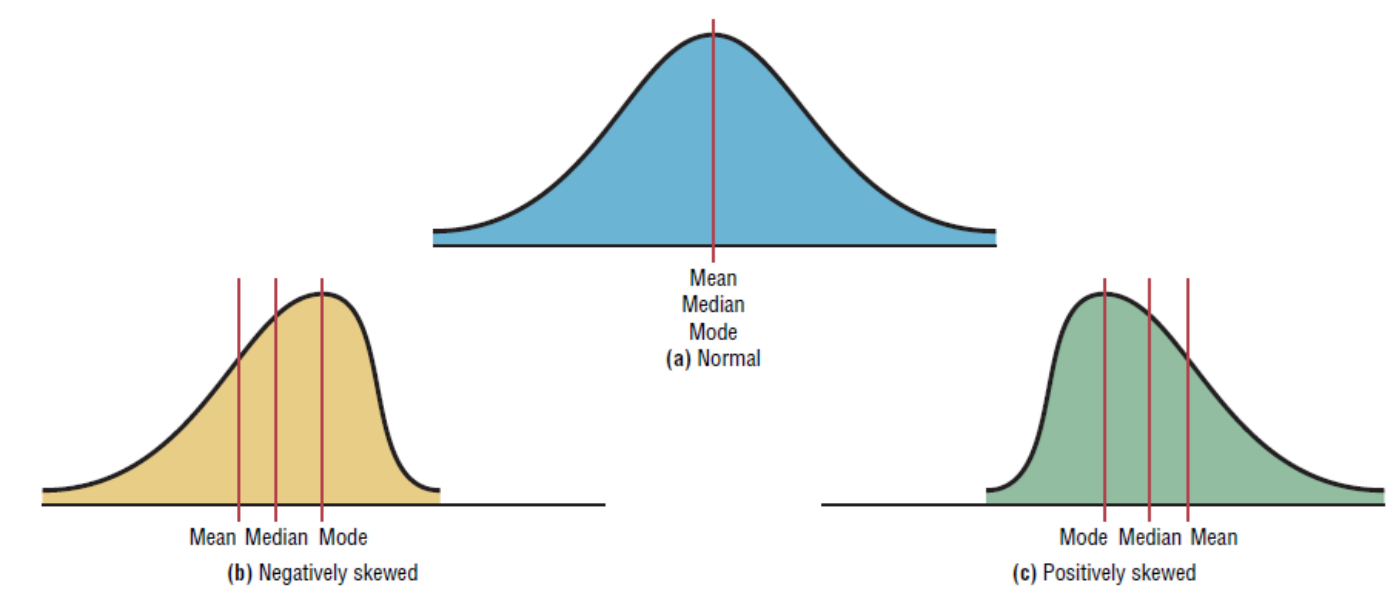
X	P(X)	X × P(X)	$X^2 \times P(X)$
1	0.3	0.3	0.3
2	0.5	1	2
3	0.2	0.6	1.8
		$\sum X \cdot P(X) = 1.9$	4.1

- *The mean can be found by summing the* $\sum X \times P(X)$ *column, and the variance – by summing the* $X^2 \times P(X)$ *column and subtracting the square of the mean [2].*

3.3 Normal distribution

- Continuous variables: assume all values between any two given values of the variables. → man's height or weight, body temperatures [2].
- Many continuous variables have distributions that are bell-shaped, and these are called approximately normally distributed variables [2; 5].

A ***normal distribution*** is a continuous, symmetric, bell-shaped distribution of a variable [2].



Normal and skewed distributions [2].

3.3 Normal distribution

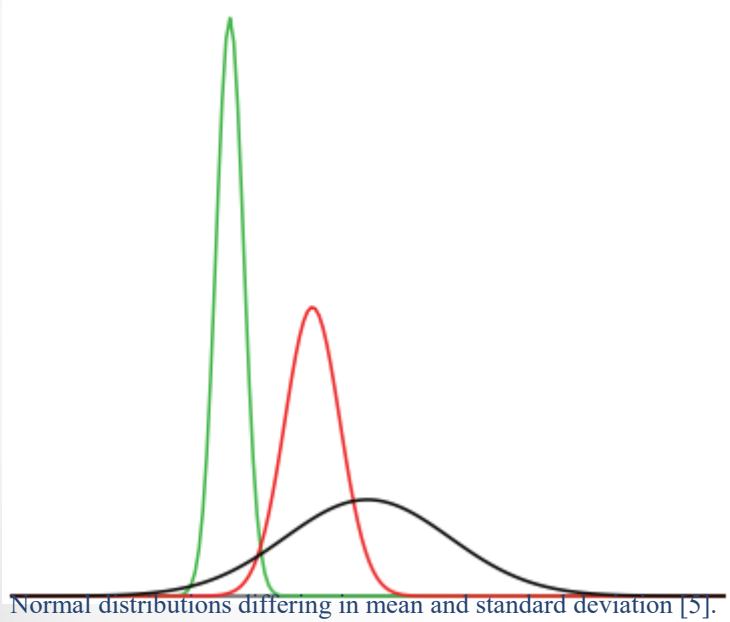
The mathematical equation for a normal distribution is [2,5]: $$y = \frac{e^{-(X-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}$$

where e ≈ 2.718 (is the base of the natural logarithm);

π ≈ 3.14;

μ - population mean;

σ – population standard deviation .

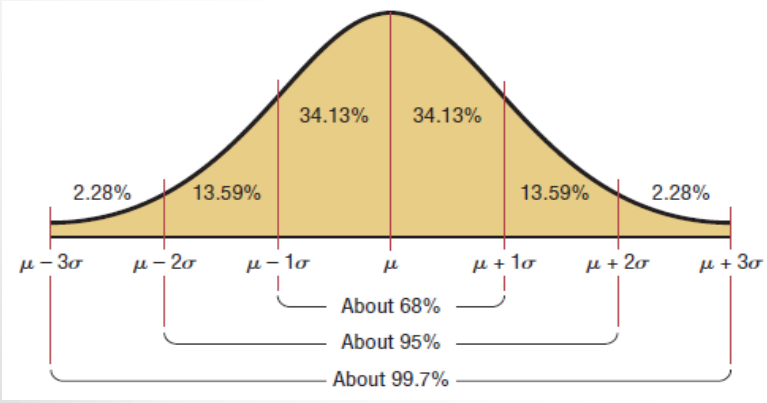


Normal distributions differing in mean and standard deviation [5].

3.3 Normal distribution

Properties of the Theoretical Normal Distribution [2, 5]:

- A normal distribution curve is bell-shaped.
- The mean, median, and mode of a normal distribution are equal.
- A normal distribution has only one mode.
- The curve is symmetric about the mean.
- The curve is continuous → there are no gaps or holes. For each value of X, there is a corresponding value of Y.
- The curve never touches the x axis, but it gets increasingly closer.
- The total area under a normal distribution curve is equal to 1.00, or 100%.
- Normal distributions are defined by two parameters: the mean (μ) and the standard deviation (σ).
- 68% of the area of a normal distribution is within one standard deviation of the mean; about 0.95, or 95% within two standard deviations.



Areas under a normal distribution curve [2].

3.3 Normal distribution

The standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1 [2].

- The formula for the standard normal distribution [2]:

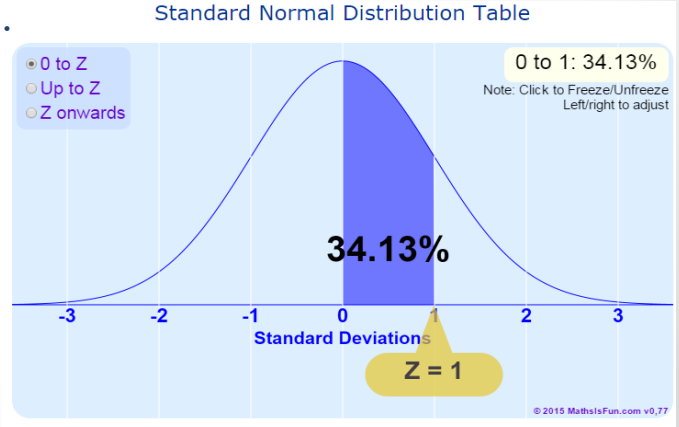
$$y = \frac{e^{-z^2/2}}{\sqrt{2\pi}}$$

- The formula for the conversion of the normally distributed variable into the standard normally distributed variable [2]:

$$z = \frac{\text{value - mean}}{\text{standard deviation}} \quad \text{or} \quad z = \frac{X - \mu}{\sigma}$$

- Tables of the standard normal distribution: https://www.mathsisfun.com/data/standard-normal-distribution-table.html.

- A normal distribution curve as a probability distribution curve: the area under the standard normal distribution curve can also be thought of as a probability [2, 5].
  → For example: P(0 < z < 1.00) = 34.13% (or 0.3413)



For example, there is a need to find the area between z = - 1.25 and z = 2.04. The area for z = - 1.25 is 10.56% (or 0.1056), and the area for z = 2.04 is 97.93% (or 0.9793). The area between two z values is 97.93% - 10.56% = 87.37%.

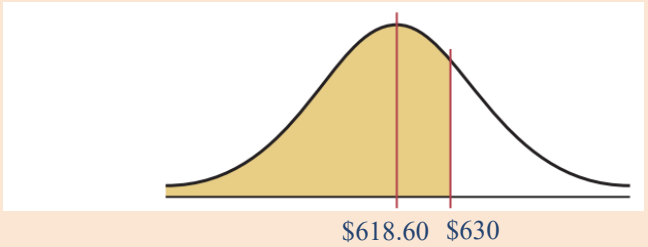
3.3 Normal distribution

Example: *Monthly food expenditure for families of two (19-50 years) in the US is on the average $618.60 and has a standard deviation $15 [4]. Assume that the monthly food expenditure are normally distributed and find the probability of the monthly food expenditure that are less than $630.*

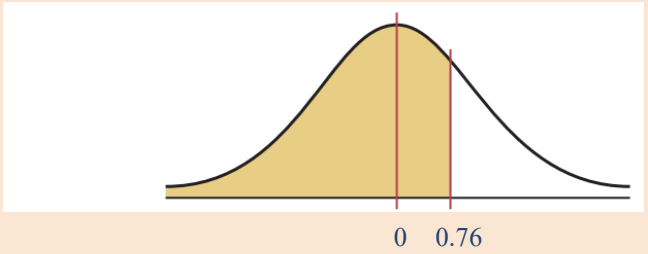
Solution:

- *First, we have to find z value corresponding to $630:*

$$z = \frac{X - \mu}{\sigma} = \frac{630 - 618.60}{15} = 0.76$$



→ *$630 is 0.76 of a standard deviation above the mean of $618.60, as shown in the z distribution.*



- *Now using the table of the standard normal distribution we can find the area under the curve: the area for z = 0.76 is 77.64%.*

→ *Therefore, monthly food expenditure of 77.64% of the families of two in the US are on average less than $630.*

3.3 Normal distribution

Exercise #6: The probability that a student answers correctly X number of questions at the written examination is presented below. Find the mean, variance, and standard deviation for this probability distribution.

X	0	2	4	6	8
P(X)	0,0	0,15	0,25	0,4	0,2

Exercise 7: Monthly food expenditure for families of two (19-50 years) in the US is on the average $618.60 and has a standard deviation $15 [4]. Assume that the monthly food expenditure are normally distributed and find the probability of the monthly food expenditure that are

a) between $600 and $650;

b) more than $630.

HUMBOLDT-UNIVERSITÄT ZU BERLIN

HUMBOLDT-UNIVERSITÄT ZU BERLIN
Resource Economics

Questions of repetition

- Which information does a frequency distribution provide?
- Which types of frequency distribution can be defined? When each should be used?
- What is a class? How to find class limits?
- How can large mass of data be best summarized graphically? Define the differences between various graphical forms.
- What is the difference between a discrete and a continuous random variable? Please give examples of each.
- What is probability? What are the basic probability rules?
- What is probability distribution? Which two requirements should be fulfilled in probability distribution?
- When is binomial distribution used?
- What is the difference between the mean for a sample or population and the mean of a probability distribution?
- How do standard deviation and variance correlate with one another?
- How to find the variance for the random variable of a probability distribution?
- What is standard deviation? What is it used for?
- What are the attributes of a normal distribution?

HUMBOLDT-UNIVERSITÄT ZU BERLIN

HUMBOLDT-UNIVERSITÄT ZU BERLIN
Resource Economics

Answers to the exercises

Exercise #1:

a. 25.5–34.5; 30; 9

b. 4.555–7.815; 6.185; 3.26

c. 135.75–167.25; 151.5; 31.5

Exercise #2:

- Population of the biggest world cities: Grouped frequency distribution

Class limits	Tally	Frequency
4800-9009	𝍸 𝍸 ////	14
9010-13129	𝍸 /	6
13130-17249	𝍸 /	6
17250-21369	𝍸 //	7
21370-25489	𝍸	5
25490-29609		0
29610-33729	/	1
33730-37849	/	1
		Total: 40

- Population of the biggest world cities: Cumulative frequency distribution

	Cumulative frequency
Less than 4799.5	0
Less than 9009.5	14
Less than 13129.5	20
Less than 17249.5	26
Less than 21369.5	33
Less than 25489.5	38
Less than 29609.5	38
Less than 33729.5	39
Less than 37849.5	40

HUMBOLDT-UNIVERSITÄT ZU BERLIN

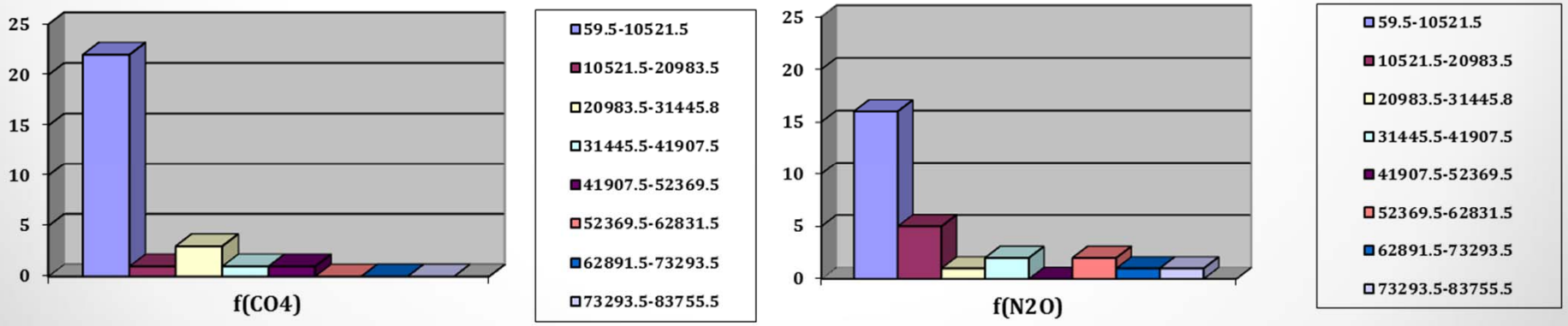
HUMBOLDT-UNIVERSITÄT ZU BERLIN
Resource Economics

Answers to the exercises

Exercise #3: Both distributions are positively skewed, but in case of nitrous oxide emissions the data are somewhat more spread out, while in case of methane emissions there is a distinct concentration of values within the first class boundaries.

- Methane and Nitrous Oxide Emissions: Grouped Frequency Distributions

Limits	Boundaries	f(CO4)	f(N2O)
60-10521	59.5-10521.5	22	16
10522-20983	10521.5-20983.5	1	5
20984-31445	20983.5-31445.8	3	1
31446-41907	31445.5-41907.5	1	2
41908-52369	41907.5-52369.5	1	0
52370-62831	52369.5-62831.5	0	2
62832-73293	62891.5-73293.5	0	1
73294-83755	73293.5-83755.5	0	1



Answers to the exercises

Exercise 4:

X	P(X)
24	8/30
25	10/30
26	4/30
27	3/30
28	3/30
30	2/30

$0 \leq P(x) \leq 1$

$\sum P(x) = 1$

Exercise 5:

a) $P(10) = 0.021$

b) $P(6) + P(7) + P(8) + P(9) + P(10) = 0.723$

c) $P(4) = 0.048$

Exercise 6:

$\mu = 5,3$

$\sigma^2 = 3,71$

$\sigma = 1,93$

Exercise 7:

a) 87.42%

b) 22.36% (100% - 77.64%)

HUMBOLDT-UNIVERSITÄT ZU BERLIN

HUMBOLDT-UNIVERSITÄT ZU BERLIN
Resource Economics

References

[1] Agricultural Statistics at a Glance 2014. Government of India Ministry of Agriculture Department of Agriculture & Cooperation Directorate of Economics & Statistics. Oxford University Press.

[2] Allan G. Bluman, 2009. *Elementary Statistics: A Brief Version*, 7th Edition, New York: McGraw-Hill.

[3] Current results. Weather and science facts. Retrieved from http://www.currentresults.com/Weather/US/annual-average-humidity-by-state.php and http://www.currentresults.com/Weather/US/average-annual-state-precipitation.php

[4] Official USDA Food Plans: Cost of Food at Home at Four Levels, U.S. Average, July 2014. Center for Nutrition Policy and Promotion. US Department of Agriculture. Retrieved from: http://www.cnpp.usda.gov/sites/default/files/usda_food_plans_cost_of_food/CostofFoodJul2014.pdf

[5] Online Statistics Education: A Multimedia Course of Study. Project Leader: David M. Lane, Rice University. Retrieved from http://onlinestatbook.com/

[6] The World Bank Data. Retrieved from http://data.worldbank.org/indicator/EN.ATM.METH.KT.CE/countries?page=13

[7] United Nations, 2015. The World's Women 2015: Trends and Statistics. New York: United Nations, Department of Economic and Social Affairs, Statistics Division. Sales No. E.15.XVII.8.

[8] World Business Council for Sustainable Development. Sustainable consumption facts and trends. From a business perspective. Retrieved from http://www.saiplatform.org/uploads/Modules/Library/WBCSD_Sustainable_Consumption_web.pdf

[9] 11th Annual Demographia International Housing Affordability Survey: 2015 Ratings for Metropolitan Markets. Retrieved from http://www.citymetric.com/horizons/england-and-wales-cities-oldest-populations-are-all-seaside-1497