**1. Frequency Tables and Graphs**

**Introduction**

* What is statistics

Statistics is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions.

# 

* **Types of Statistics**

***Descriptive Statistics*** involves organizing, summarizing, and displaying data. For example, we can use tables, charts, averages, etc.

**C:\Documents and Settings\Lyn\Local Settings\Temporary Internet Files\Content.IE5\O7HXM29P\MCj02331070000[1].wmf**

***Inferential Statistics*** uses the *sample data* to make inferences about the underlying *population.*

* Population and sample

**Population:** The collection of ***all***outcomes, responses, measurements, or counts that are of interest.

:

**Sample:** A subset of the population.

* + Data Types
    1. ***Quantitative (also known as “numerical”)***



Age



Weight of a letter

Temperature



* Continuous

**Examples**: Weight, Height

* Discrete (a subset of which is “counting”)

**Examples**: Number of children in a family (0,1,2, etc.) Shoe Size (6, 6 ½, 7, 7 ½, etc.)

**Comment**: The distinction between “continuous” and “discrete” is somewhat blurry. In some situations, the distinction is important, while in others it is not.

* + 1. ***Qualitative (also known as “categorical”)***
       - Unordered (also known as “nominal”)

**Examples**: the place of birth, major, eye color

Eye color



Major

Place of birth

* Ordered (also known as “ordinal”)

**Examples**: Military Rank (private, corporal, etc.), Course Grade (A, B, C, D, F)

**Sampling and Experimental Design**

Reading assignment: how to collect statistically valid data for statistical analysis?

1. Sampling plans: pages 5-8.

Focus on Simple Random Sampling and Stratified Sampling.

1. Experimental designs: Section 1.3.

Focus on the high-level understanding of a few observational study designs.

**Summarizing Qualitative Data**

* **Example 1**: **Grade Data.** In a class of 20 students, 3 students received a grade of “A”, 6 students received a “B”, 7 students received a “C”, 3 students received a “D”, and 1 student received an “F”. These results are summarized, in a variety of ways, in the following table: (Note that for the ordered categorical variable “Grade” we also create the discrete quantitative variable “Grade-point”.)
* The raw data might be in the following form:

A, C, B, F, D, B, C, B, C, C,

A, D, C, B, C, B, A, C, B, D

* **Frequency tables**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Grade** | **Freq** | **Rel. Freq** | **Cum. Freq** | **Cum. Rel. Freq** |
| **F** | 1 | .05 | 1 | .05 |
| **D** | 3 | .15 | 4 | .20 |
| **C** | 7 | .35 | 11 | .55 |
| **B** | 6 | .30 | 17 | .85 |
| **A** | 3 | .15 | 20 | 1.00 |
| **Total** | **20** | **1.00** |  |  |

* **Comments**

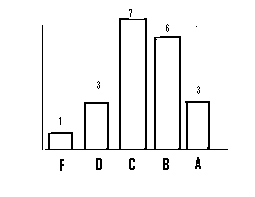
**Terminology**: Frequency, Relative Frequency (Proportion), and Cumulative Relative Frequency tables

**In Calculations** of the values in the “Relative Frequency”, “Cumulative Relative Frequency”, and “Pie Chart Angle” columns.

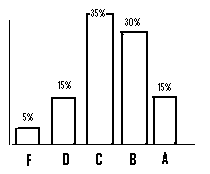
Percent = Relative Frequency × 100

Another term for “Relative Frequency” is “Proportion”.

* To visualize the information in the table we consider the following graphs:
  1. **Bar chart based on frequencies**

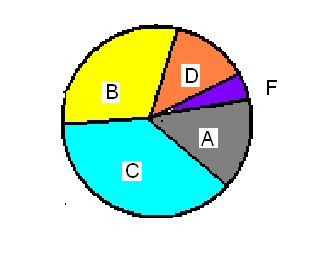


* 1. **Bar chart based on relative frequencies**



* 1. **Pie chart**

|  |  |  |
| --- | --- | --- |
| Grade | Relative Freq | Pie Chart Angle |
| F | .05 | 0.05 x 360 O = 18 O |
| D | .15 | 0.15 x 360 O = 54 O |
| C | .35 | 126 O |
| B | .30 | 108 O |
| A | .15 | 54 O |
| Total | 1.00 | 360 O |



**Summarizing Quantitative Data**

* **Example 2**: (**Lengths of CDs**) Listed below are the lengths (in minutes) of randomly selected CDs of country, rock, and movie soundtracks.

20.5, 29, 32, 32, 32, 33, 36, 37, 38, 39, 39, 43, 47, 48, 49, 49, 49, 50, 50, 51, 51, 52, 52, 52, 53, 54, 54, 54, 56, 56, 57, 58, 60, 61, 62, 62, 69, 73, 74, 74.5

* **Question:** The data set contains 40 distinct values 🡪 What would the “ungrouped” frequency and relative frequency tables, and their associated histograms, look like?
* ***The following issues need to be considered when constructing a frequency table.***
  1. How many groups (“classes”) should there be?

**Answer**: *The number of classes depends on the size of the data. Usually, the number of classes is between 5 and 20. As a convention, we choose 5 classes for a small data set.*

* 1. How wide should each (class) interval be?

**Answer**: *When the number of classes is given, the class width can be calculated by*

.

If the above formula gives an “inconvenient” class width (i.e., decimal), we can choose a convenient width that is ***slightly bigger than*** the actual width.

*In this example, Class.Widthh = (74.5-20.5)/5 ≈ 10.8, we a convenient width w = 11 > 10.8.*

* 1. How to determine the upper and lower limits of the intervals?

You may want to choose a convenient starting value that is ***slightly smaller than*** the minimum value if the minimum value is inconvenient.We

Original data window

20.5 74.5

20 31 42 53 64 75

New data window based on convenient width and starting value.

**Caution:** The original data window must be in the new data window.

**Answer**: This step is broken into two sub-steps as follows:

*(1). Lower limits: 20, 31, 42, 53, 64*

*(2) Upper limits: 31, 42, 53, 64, 75*

***Class Boundaries****:* 20, 31, 42, 53, 64, 75.

***Caution****: Each boundary value can only be in one and only one interval. We make a convention,*

[20,31], (31,42], (42,53], (53, 64], (64,75].

Include the upper limit(boundary) in the current class and the lower limit (boundary) in the preceding class!

* 1. Should all intervals be of the same width?

**Answer: *Yes. In practice, we can get equal width intervals by choosing an appropriate number of intervals and the width of each interval.***

* ***The frequency table based on the Length of CD’s data***

Next, we apply the above steps to the lengths of CDs to construct a frequency table with different frequencies.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classes** | **Freq** | **Cum. Freq** | **Rel. Freq** | **Cum. Rel. Freq.** |
| [20, 31] | 2 | 2 | 0.05 | 0.05 |
| (31, 42] | 9 | 11 | 0.225 | 0.275 |
| (42, 53] | 14 | 25 | 0.35 | 0.625 |
| (53, 64] | 11 | 36 | 0.275 | 0.9 |
| (64, 75] | 4 | 40 | 0.1 | 1.0 |

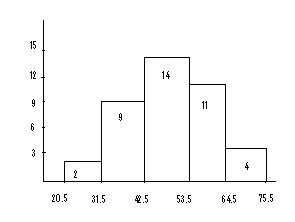
* In constructing a ***histogram*** associated with a particular grouped frequency or relative frequency table, many questions arise:
  1. Should the horizontal axis be labeled at the interval boundaries or the midpoints of the intervals?

**Answer**: *The horizontal axis should be labeled at the interval boundaries.*

* 1. For a relative frequency histogram with intervals of equal length, should we label the vertical axis with “relative frequency per interval” or with “relative frequency per unit”?

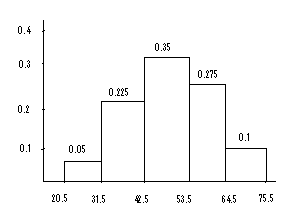
**Answer**: *we should label the vertical axis with “relative frequency per unit”*

*In this example, we should the vertical axis with 10%, 20%, 30%, 40%, 50%.*



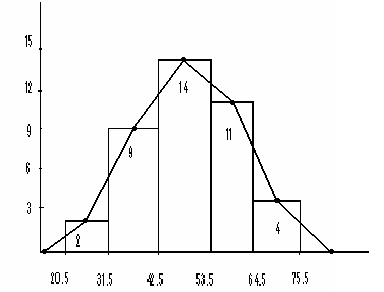
20 31 42 53 64 75

*Histogram Based on the frequency table*



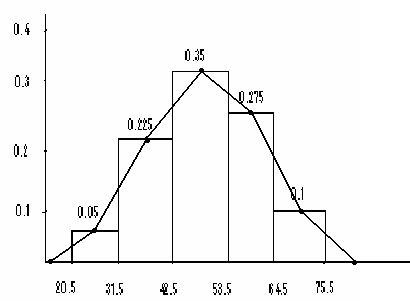
20 31 42 53 64 75

*Histogram based on rel. freq. table*



14.5 20 31 42 53 64 75 80.5

Frequency Polygon



14.5 20 31 42 53 64 75 80.5

Relative Frequency Polygon

**Example 2** The following sample data set lists the number of minutes 50 Internet subscribers spent on the Internet during their most recent session. Construct a frequency distribution that has seven classes.

50, 40, 41, 17, 11, 7, 22, 44, 28, 21, 19, 23, 37, 51, 54, 42, 86, 41, 78, 56, 72, 56, 17, 7, 69, 30, 80, 56, 29, 33, 46, 31, 39, 20, 18, 29, 34, 59, 73, 77, 36, 39, 30, 62, 54, 67, 39, 31, 53, 44

Following the steps used in the previous example, we have the following frequency table:

1. #class = 7
2. Width = (max-min)/#class = 11.29, we use 12.
3. Starting point: 7
4. Boundary values: 7, 19, 31, 43, 55, 67, 79, 91

Class: [7,19], (19,31], (31,43], (43,55], (55,67], (67,79], (79, 91]

|  |  |  |  |
| --- | --- | --- | --- |
| Class | Frequency  *f* | Relative  frequency | Cumulative frequency |
| [7, 19] | 7 | 0.14 | 7 |
| (19, 31] | 11 | 0.22 | 18 |
| (31, 43] | 11 | 0.22 | 29 |
| (43, 55] | 8 | 0.16 | 37 |
| (55, 67] | 6 | 0.12 | 43 |
| (67, 79] | 5 | 0.10 | 48 |
| (79, 91] | 2 | 0.04 | 50 |

**Practice Exercises**

Summarize the following data sets by using frequency tables (relative frequency, cumulative frequency, etc.) and graphs (histogram, frequency polygon).

**Exercise 1**. Weights of 18- to 24- Year- Old Males. The U. S. National Center for Health Statistics publishes data on weights and heights by age and sex in the document Vital and Health Statistics. The weights shown in the following, given to the nearest tenth of a pound, were obtained from a sample of 18- to 24- year- old males. Use the cut-point grouping to organize these data into frequency and relative- frequency distributions. Use a class width of 20 and a first cut-point of 120.

129.2, 132.1, 136.7, 142.8, 145.6, 146.4, 149.9, 150.7, 151.3, 155.2, 158.5, 158.6, 161.0, 161.7, 165.0, 165.8, 167.3, 170.0, 170.1, 172.5, 173.6, 173.7, 175.4, 175.6, 178.2, 178.7, 182.0, 182.5, 185.3, 187.0, 187.5, 188.7, 191.1, 209.1, 214.6, 218.1, 278.8

**Exercise 2**. The following are the miles per gallon.

22.8, 22.9, 23.3, 23.4, 23.6, 23.7, 23.8, 23.9, 23.9, 24.1, 24.1, 24.2, 24.3, 24.4, 24.5, 24.5, 24.6, 24.6, 24.7, 24.7, 24.7, 24.8, 24.8, 24.9, 24.9, 25.0, 25.0, 25.1, 25.2, 25.3

**Exercise 3**. Following are 80 measurements of the iron-solution index of tin-plate specimens, designed to measure the corrosion resistance of tin-plated steel. The original data set has been sorted in an ascending order as:

14, 26, 28, 28, 28, 28, 30, 32, 34, 35, 36, 36, 37, 37, 40, 40, 40, 41, 41, 41, 42, 42, 42, 43, 43, 43, 44, 44, 44, 44, 45, 45, 45, 45, 45, 45, 46, 46, 46, 46, 47, 47, 47, 48, 49, 49, 49, 50, 50, 50, 51, 52, 52, 52, 52, 52, 53, 53, 54, 54, 54, 54, 55, 55, 55, 55, 55, 55, 55, 56, 56, 56, 56, 56, 56, 57, 57, 57, 57, 57, 58, 58, 58, 58, 58, 59, 59, 60, 60, 60, 60, 61, 61, 61, 61, 61, 62, 62, 62, 62, 62, 62, 63, 63, 63, 63, 63, 63, 64, 65, 66, 66, 67, 68, 68, 69, 69, 70, 70, 70, 70, 70, 70, 71, 71, 72, 72, 72, 73, 74, 74, 74, 76, 76, 77, 77, 79, 80, 81, 81, 83, 83, 84, 86, 86, 86, 87, 89, 92, 95

**Exercises 4**. From the 140 children whose urinary concentration of lead were investigated 40 were chosen who were aged at least 1 year but under 5 years. The following concentrations of copper were found.

0.70, 0.45, 0.72, 0.30, 1.16, 0.69, 0.83, 0.74, 1.24, 0.77,

0.65, 0.76, 0.42, 0.94, 0.36, 0.98, 0.64, 0.90, 0.63, 0.55,

0.78, 0.10, 0.52, 0.42, 0.58, 0.62, 1.12, 0.86, 0.74, 1.04,

0.65, 0.66, 0.81, 0.48, 0.85, 0.75, 0.73, 0.50, 0.34, 0.88