Centre Universitaire Abdelhafid Boussouf-Mila

Matière: Applied Biostatistics

Dr. Khadidja Daas

khadidja.daas@centre-univ-mila.dz

Chapter 1

Review of descriptive statistics

1.1 Introduction

Descriptive statistics are methods used to summarize, organize, and present data in a meaningful way. In this chapter, we review the main concepts that will be useful in hypothesis testing and advanced topics later.

1.2 Types of Variables

- **Qualitative variables:** categories (e.g., gender, blood type).

- **Quantitative variables:** numerical values.

  - Discrete: countable (e.g., number of children).

  - Continuous: measurable (e.g., height, weight).

1.3 Measures of Central Tendency

1.3.1 Mean (Average)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example: Data: 4, 6, 8, 10, 12.

$$\bar{x} = \frac{4 + 6 + 8 + 10 + 12}{5} = \frac{40}{5} = 8$$

1.3.2 Median

The median is the value that divides the dataset into two equal halves.

- If n is odd: the middle value.
- If n is even: average of the two middle values.

Example: Data: 3, 7, 9, 15, 20. Median = 9 (middle value).

Data: 3, 7, 9, 15. Median = $\frac{7+9}{2} = 8$.

3. Mode

The most frequent value.

Example: Data: 2, 4, 4, 6, 8. Mode = 4.

1.4 Measures of Dispersion

1. Range

$$\text{Range} = \text{Max} - \text{Min}$$

Example: Data 4, 6, 8, 10, 12. Range $= 12 - 4 = 8$.

1.4.1 Variance and Standard Deviation

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}, \quad s = \sqrt{s^2}$$

Example: Data: 2, 4, 6. Mean = $\bar{x} = 4$.

$$s^2 = \frac{(2-4)^2 + (4-4)^2 + (6-4)^2}{3-1} = \frac{4+0+4}{2} = \frac{8}{2} = 4$$

$$s = \sqrt{4} = 2$$

1.5 Quartiles

Consider the ordered dataset:

$$2,\ 5,\ 7,\ 10,\ 12,\ 15,\ 18,\ 20,\ 25,\ 30$$

Here, the number of observations is $n = 10$.

Step 1: Median (Q_2) Since $n = 10$ (even), the median is the average of the two middle values (5th and 6th):

$$Q_2 = \frac{12+15}{2} = \frac{27}{2} = 13.5$$

Step 2: First Quartile Q_1 is the median of the lower half:

$$2,\ 5,\ 7,\ 10,\ 12$$

There are 5 values, so the median is the 3rd value:

$$Q_1 = 7$$

Step 3: Third Quartile Q_3 is the median of the upper half:

$$15, \ 18, \ 20, \ 25, \ 30$$

There are 5 values, so the median is the 3rd value:

$$Q_3 = 20$$

Final Results

$$Q_1 = 7, \quad Q_2 = 13.5, \quad Q_3 = 20$$

Thus:

- 25% of the data are below $Q_1 = 7$,
- 50% of the data are below $Q_2 = 13.5$,
- 75% of the data are below $Q_3 = 20$.

1.6 Summary Table

c

Concept	Formula / Definition
Mean	$\bar{x} = \frac{\sum x_i}{n}$
Median	Middle value (or average of 2 middle values)
Mode	Most frequent value
Range	Max - Min
Variance	$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$
Standard deviation	$s = \sqrt{s^2}$

1.7 Practice Problem

The weights (in kg) of 6 students are: 50, 55, 60, 65, 70, 80.

Tasks:

1. Compute the mean, median, and mode.
2. Compute the range.
3. Compute the variance and standard deviation.

Solution:

- Mean $= \frac{50+55+60+65+70+80}{6} = \frac{380}{6} \approx 63.33$
- Median = average of 3rd and 4th values $= \frac{60+65}{2} = 62.5$
- Mode = none (all values occur once).
- Range $= 80 - 50 = 30$
- Variance:

$$s^2 = \frac{(50-63.33)^2 + (55-63.33)^2 + (60-63.33)^2 + (65-63.33)^2 + (70-63.33)^2 + (80-63.33)^2}{6-1}$$

$$= \frac{177.78 + 69.44 + 11.11 + 2.78 + 44.44 + 277.78}{5} = \frac{583.33}{5} = 116.67$$

- Standard deviation: $s = \sqrt{116.67} \approx 10.8$

Chapter 2

Generalities on Tests / Comparison of Two or More Proportions

Chapter 2: Generalities on Tests – Comparison of Two or More Proportions

Types of Comparisons

1. **One proportion test:** Does the sample proportion differ from a theoretical proportion?

   Example: A book claims that 30% of young people smoke. We take a sample and find 35%. Is the difference significant?

   ⇒ Use the one-proportion Z-test.

2. **Two proportions test:** Does the proportion in sample A differ from the proportion in sample B?

   Example: 20% of women vs. 30% of men have hypertension. Is the difference significant?

   ⇒ Use the two-proportions Z-test.

3. **More than two proportions:** When there are more than two groups.

Example: Comparing the cure rate in patients treated with 3 drugs.

⇒ Use the Chi-square (χ^2) test with contingency tables.

Principle of Hypothesis Testing

1. State the null hypothesis (H_0): there is no difference between proportions.
2. Compute the test statistic (Z or χ^2).
3. Compare it with a critical value (from Z or χ^2 tables), or compute the p-value.
4. If $p < 0.05$ (commonly), reject H_0 and conclude: the difference between proportions is statistically significant.

Example 1: Two Proportions

In a study:

- 200 men: 60 have diabetes (30%).
- 150 women: 30 have diabetes (20%).

Hypotheses:

$$H_0 : p_1 = p_2 \qquad H_1 : p_1 \neq p_2$$

Test statistic:

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Substitution:

$$p_1 = 0.3, \quad p_2 = 0.2, \quad p = \frac{90}{350} = 0.257$$

$$Z \approx 1.82 \quad \Rightarrow \quad p \approx 0.07 > 0.05$$

From the standard normal (Z) table: at $Z = 1.82$, the p-value is about 0.07.

Decision: The difference is not statistically significant (may be due to chance).

Exercise

In village A: 50/200 children have caries (25%). In village B: 45/150 children have caries (30%). Compare the two proportions.

More than Two Proportions (Chi-square Test)

When more than two groups are compared in terms of proportions (success/failure, yes/no, etc.):

- Example: Comparing success rates of 3 drugs.
- Example: Comparing infection rates across 4 regions.

Here, the Z-test is not sufficient. We use the Chi-square (χ^2) test.

Steps of the χ^2 Test

1. Build a contingency table: Rows = categories (success/failure), Columns = groups (Drug A, B, C...).
2. Compute the expected frequencies (E):

$$E = \frac{(\text{row total}) \times (\text{column total})}{\text{grand total}}$$

3. Compute the Chi-square statistic:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

4. Compute degrees of freedom:

$$df = (r-1)(c-1)$$

5. Compare χ^2 calculated with the critical value from the χ^2 table at $\alpha = 0.05$.

Worked Example: Effectiveness of 3 Drugs

Drug	Patients	Success	Failure
A	50	20	30
B	60	30	30
C	40	25	15
Total	150	75	75

Step 1: Expected frequencies (E)

For Drug A, success:

$$E = \frac{75 \times 50}{150} = 25$$

For Drug A, failure: 25. For Drug B, success: 30; failure: 30. For Drug C, success: 20; failure: 20.

Step 2: Table of O and E

Drug	O Success	E Success	O Failure	E Failure
A	20	25	30	25
B	30	30	30	30
C	25	20	15	20

Step 3: Compute χ^2

$$\chi^2 = \frac{(20-25)^2}{25} + \frac{(30-25)^2}{25} + \frac{(30-30)^2}{30} + \frac{(25-20)^2}{20} + \frac{(15-20)^2}{20}$$

$$= 1 + 1 + 0 + 1.25 + 1.25 = 4.5$$

Step 4: Degrees of freedom

$$df = (2 - 1) \times (3 - 1) = 2$$

Step 5: Decision

From the χ^2 table at $df = 2$ and $\alpha = 0.05$, the critical value = 5.99.

Since $\chi^2_{calc} = 4.5 < 5.99$, the difference is not significant.

Conclusion: There is no statistically significant difference between the cure rates of the three drugs.

Standard Normal Distribution Table (up to $z = 2.0$)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

chi square table

	P										
DF	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	.0004	.00016	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.824	9.21	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.86	16.924	18.467
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086	16.75	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.69	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.18	11.03	13.362	15.507	17.535	18.168	20.09	21.955	24.352	26.124
9	1.735	2.7	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588