



# Group E

Cancer Patients Data Set

Francesca  
Guilherme  
Koketso  
Leonne  
Muhsin



# Agenda

- What is our data talking about?
- Cleaning Data
- Descriptive Analysis
- Crosstab



# The Data Base

The database shows a questionnaire made to patients who have cancer. The database computes age, gender, symptoms and external factors the patient was submitted on a scale of 1 to 9. For example, factors such as smoking, alcoholism and air pollution. And some symptoms like coughing up blood and chest pain.

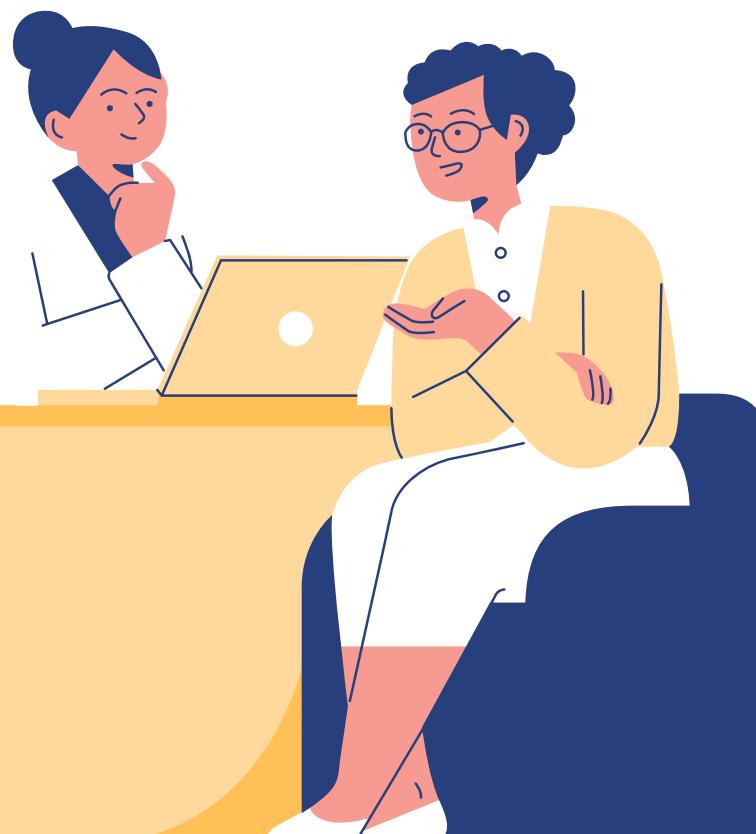




# Cleaning Data

Here we can see what variables were having missing data and how much missing values in each column. We also can see five columns had no missing data.

:	Coughing of Blood
	Alcohol use
	Shortness of Breath
	Air Pollution
	Obesity
	Chest Pain
	Smoking
	Dust Allergy
	Swallowing Difficulty
	Gender
	Occupational Hazards
	Genetic Risk
	Wheezing
	Fatigue
	Weight Loss
	Balanced Diet
	Clubbing of Finger Nails
	Passive Smoker
	Age
	Frequent Cold
	Dry Cough
	Snoring
	Patient Id
	chronic Lung Disease
	Level





# Cleaning Data

In the data base, most of the variables are categoricals and ordinals (scale from 1 to 9). And we could find some missing data and some wrong values (values that exceeds 9 or values with decimals, such as 3.5).

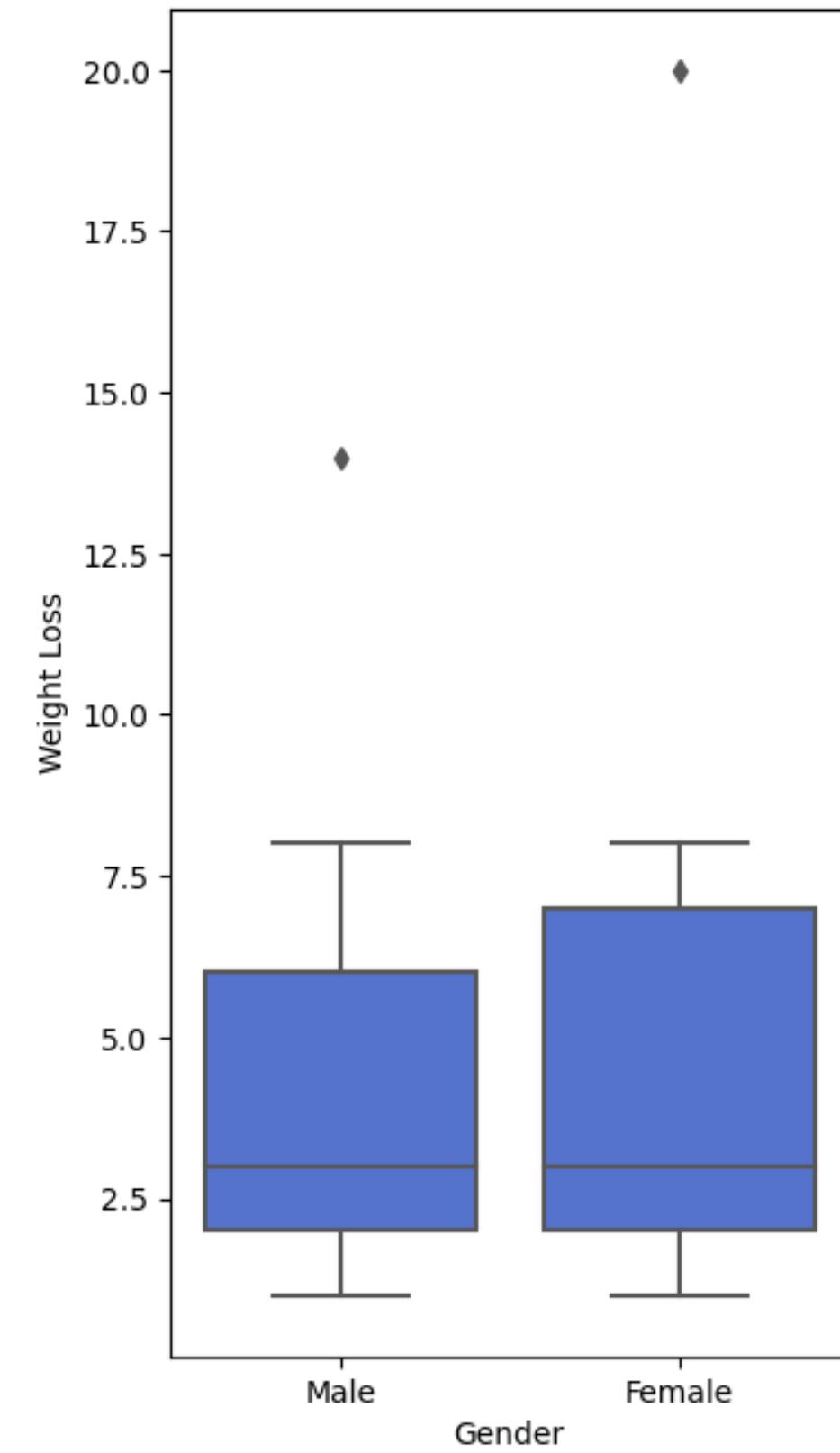
Therefore, for the wrong values, we deleted the values that exceeds 9, so they could be missing values. For the values with decimals, we rounded the values. If it was 3.7, we rounded to 4.





# Cleaning Data

Here is an example of the values exceeding 9 using the boxplot.

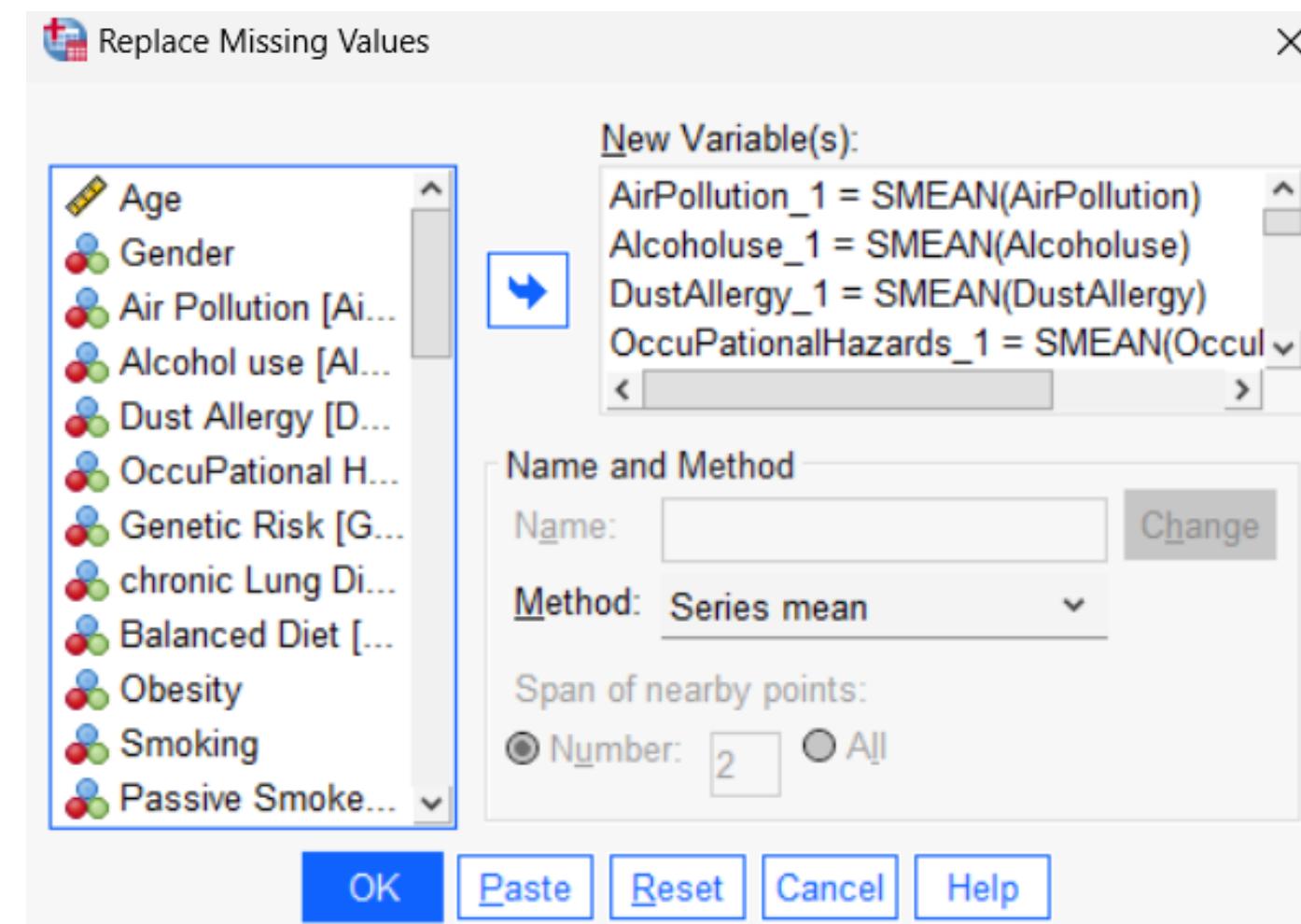




# Cleaning Data

After that, we replaced the missing values using the mean substitution method.

Transform -> Replace Missing Values -> Select all variables except the ID -> Method:  
Series Mean





# Cleaning Data

After replacing the missing data with mean substitution, we compared the new values with the old ones using the Paired-sample-t-test to guarantee the new and old values were alike.

Pair 14	Weight Loss	3,86 <sup>a</sup>	1000	2,200	,070
	SMEAN(WeightLoss)	3,862 <sup>a</sup>	1000	2,2004	,0696
Pair 15	Shortness of Breath	4,24 <sup>a</sup>	1000	2,277	,072
	SMEAN (ShortnessofBreath)	4,238 <sup>a</sup>	1000	2,2774	,0720
Pair 16	Wheezing	3,79 <sup>a</sup>	1000	2,033	,064
	SMEAN(Wheezing)	3,789 <sup>a</sup>	1000	2,0329	,0643
Pair 17	Swallowing Difficulty	3,75 <sup>a</sup>	1000	2,264	,072
	SMEAN (SwallowingDifficulty)	3,751 <sup>a</sup>	1000	2,2637	,0716
Pair 18	Clubbing of Finger Nails	3,93 <sup>a</sup>	1000	2,386	,075
	SMEAN (ClubbingofFingerNails)	3,929 <sup>a</sup>	1000	2,3857	,0754
Pair 19	Frequent Cold	3,54 <sup>a</sup>	1000	1,828	,058
	SMEAN(FrequentCold)	3,537 <sup>a</sup>	1000	1,8281	,0578
Pair 20	Dry Cough	3,85 <sup>a</sup>	1000	2,039	,064
	SMEAN(DryCough)	3,853 <sup>a</sup>	1000	2,0390	,0645
Pair 21	Snoring	2,93 <sup>a</sup>	1000	1,475	,047
	SMEAN(Snoring)	2,926 <sup>a</sup>	1000	1,4747	,0466
Pair 22	Age	37,19 <sup>a</sup>	1000	11,986	,379
	SMEAN(Age)	37,186 <sup>a</sup>	1000	11,9859	,3790

a. The correlation and t cannot be computed because the standard error of the difference is 0.





# Descriptive Analysis

(Frequency, central tendency, dispersion, position)





# Descriptive Analysis

Analyze -> Descriptive Statistics -> Frequencies

The image shows two overlapping SPSS dialog boxes. The top dialog is titled "Frequencies: Statistics" and the bottom one is "Frequencies".

**Frequencies: Statistics Dialog (Top Right):**

- Percentile Values:**  Quartiles,  Cut points for: 10 equal groups,  Percentile(s): [empty field], Add, Change, Remove.
- Central Tendency:**  Mean,  Median,  Mode,  Sum.
- Values are group midpoints.
- Dispersion:**  Std. deviation,  Variance,  Range,  Minimum,  Maximum,  S.E. mean.
- Distribution:**  Skewness,  Kurtosis.

**Frequencies Dialog (Bottom Left):**

- Variable(s):** Age, Gender, Air Pollution [Air...], Alcohol use [Alco...], Dust Allergy [Dus...], Occupational Ha... (partially visible), Genetic Risk [Ge...], chronic Lung Dis... (partially visible).
- Statistics... button** is highlighted.
- Display frequency tables** checkbox is checked.
- Create APA style tables** checkbox is unchecked.
- OK, Paste, Reset, Cancel, Help** buttons.

Below the dialogs, there are two data grids showing sample data:

Age	Gender	Air Pollution	Alcohol use	Dust Allergy	Occupational Health	Genetic Risk	Chronic Lung Disease
3	4	3	5	3	2	6	1
3	4	3	5	3	2	6	1
3	4	3	5	3	2	6	1
3	4	3	5	3	2	6	1
3	4	3	5	3	2	6	1

Age	Gender	Air Pollution	Alcohol use	Dust Allergy	Occupational Health	Genetic Risk	Chronic Lung Disease
2	1	2	1	2	3	4	2
2	1	2	1	2	3	4	2
2	1	2	1	2	3	4	2
2	1	2	1	2	3	4	2



# Descriptive Analysis

Screenshot of SPSS Statistics software interface showing the Frequencies, Charts, Format, and Multiple Variables dialog boxes.

**Frequencies Dialog:**

- Variable(s):** Age, Gender, Air Pollution [Air...], Alcohol use [Alco...], Dust Allergy [Dus...], Occupational Ha... (partially visible), Genetic Risk [Ge...], chronic Lung Dis... (partially visible).
- Statistics...**
- Charts...** (highlighted)
- Format...**
- Style...**
- Bootstrap...**

**Display frequency tables** (checkbox checked) **Create APA style tables** (checkbox uncheckable)

**OK**, **Paste**, **Reset**, **Cancel**, **Help**

**Chart Type:**

- None
- Bar charts
- Pie charts
- Histograms:  
 Show normal curve on histogram

**Chart Values:**

- Frequencies
- Percentages

**Continue**, **Cancel**, **Help**

**Format Dialog:**

**Order by:**

- Ascending values
- Descending values
- Ascending counts
- Descending counts

**Multiple Variables:**

- Compare variables
- Organize output by variables

Suppress tables with many categories  
Maximum number of categories: 10

**Continue**, **Cancel**, **Help**

Data grid (partial view):

	1	2	3	4	5	6	7	8	9	10
1	3	4	3	5	3	2	6	2	4	2
2	3	4	3	5	3	2	6	2	4	2
3	3	4	3	5	3	2	6	2	4	2
4	3	4	3	5	3	2	6	2	4	2
5	3	4	3	5	3	2	6	2	4	2
6	3	4	3	5	3	2	6	2	4	2
7	3	4	3	5	3	2	6	2	4	2
8	3	4	3	5	3	2	6	2	4	2
9	3	4	3	5	3	2	6	2	4	2
10	3	4	3	5	3	2	6	2	4	2



# Descriptive Analysis

\*Syntax1 - IBM SPSS Statistics Syntax Editor

File Edit View Data Transform Analyze Graphs Utilities Run Tools Extensions Window Help

Active DataSet: DataSet1 Search application

```
1 DATASET ACTIVATE
2 FREQUENCIES
3 DATASET ACTIVATE DataSet1.
4 FREQUENCIES VARIABLES=Age Gender AirPollution Alcoholuse DustAllergy OccupationalHazards
5 GeneticRisk chronicLungDisease BalancedDiet Obesity Smoking PassiveSmoker ChestPain CoughingofBlood
6 Fatigue WeightLoss ShortnessofBreath Wheezing SwallowingDifficulty ClubbingofFingerNails
7 FrequentCold DryCough Snoring Level
8 /NTILES=4
9 /STATISTICS=STDDEV MEAN MEDIAN MODE
10 /BARCHART FREQ
11 /ORDER=ANALYSIS.
```

Syntax



# Descriptive Analysis

Statistics										
	Air Pollution	Alcohol use	Dust Allergy	Occupational Hazards	Genetic Risk	Chronic Lung Disease	Balanced Diet	Obesity	Smoking	Passive Smoker
N	Valid	1000	1000	1000	1000	1000	1000	1000	1000	1000
	Missing	0	0	0	0	0	0	0	0	0
Mean	3,84	4,58	5,16	4,84	4,59	4,38	4,49	4,48	3,95	4,19
Median	3,00	5,00	6,00	5,00	5,00	4,00	4,00	4,00	3,00	4,00
Mode	6	2	7	7	7	6	7	7	2	2
Std. Deviation	2,023	2,604	1,975	2,101	2,120	1,849	2,130	2,120	2,486	2,306
Percentiles	25	2,00	2,00	4,00	3,00	3,00	2,00	3,00	2,00	2,00
	50	3,00	5,00	6,00	5,00	5,00	4,00	4,00	3,00	4,00
	75	6,00	7,00	7,00	7,00	7,00	6,00	7,00	7,00	7,00

a. Multiple modes exist shown

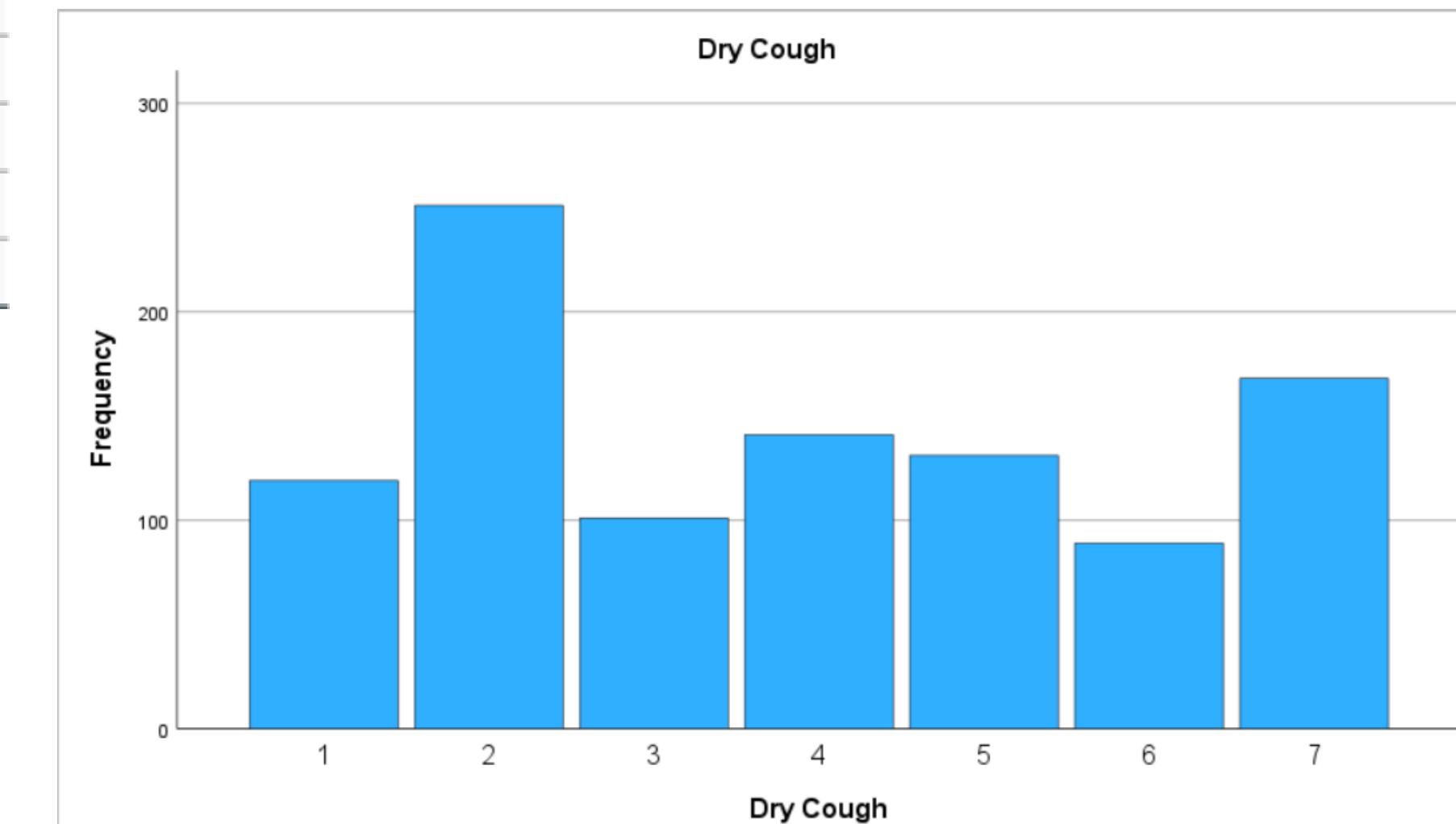
Statistics											
	Coughing of Blood	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring	
N	Valid	1000	1000	1000	1000	1000	1000	1000	1000	1000	
	Missing	0	0	0	0	0	0	0	0	0	
Mean	4,87	3,86	3,86	4,24	3,79	3,75	3,93	3,54	3,85	2,93	
Median	4,00	3,00	3,00	4,00	4,00	4,00	4,00	3,00	4,00	3,00	
Mode	7	2 <sup>a</sup>	2	2	2	1	2	3	2	2	
Std. Deviation	2,411	2,243	2,200	2,277	2,033	2,264	2,386	1,828	2,039	1,475	
Percentiles	25	3,00	2,00	2,00	2,00	2,00	2,00	2,00	2,00	2,00	
	50	4,00	3,00	3,00	4,00	4,00	4,00	3,00	4,00	3,00	
	75	7,00	5,00	6,00	6,00	5,00	5,00	5,00	6,00	4,00	



# Descriptive Analysis

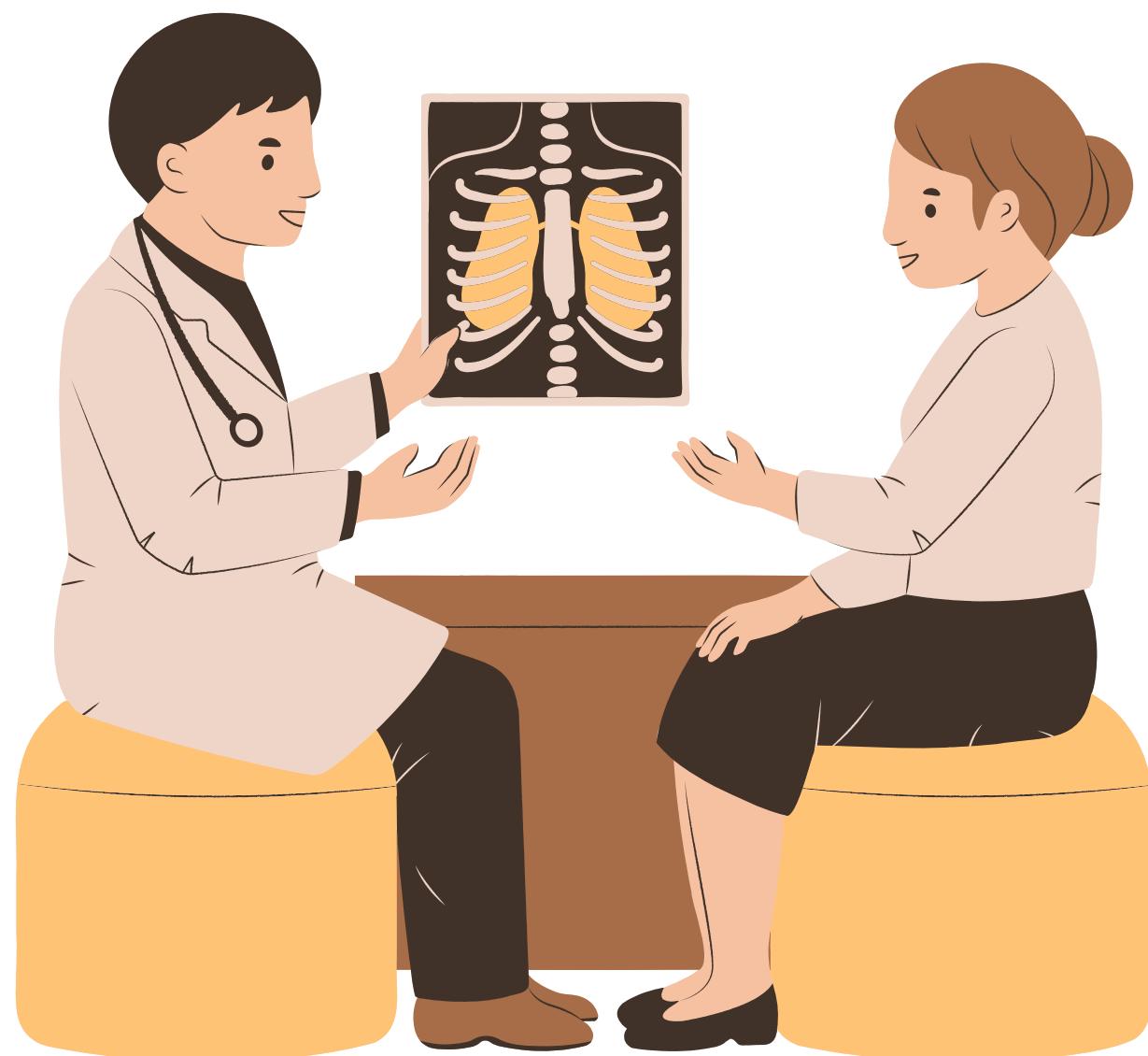
**Dry Cough**

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	119	11,9	11,9
	2	251	25,1	37,0
	3	101	10,1	47,1
	4	141	14,1	61,2
	5	131	13,1	74,3
	6	89	8,9	83,2
	7	168	16,8	100,0
Total	1000	100,0	100,0	





# Crosstab





# Crosstab

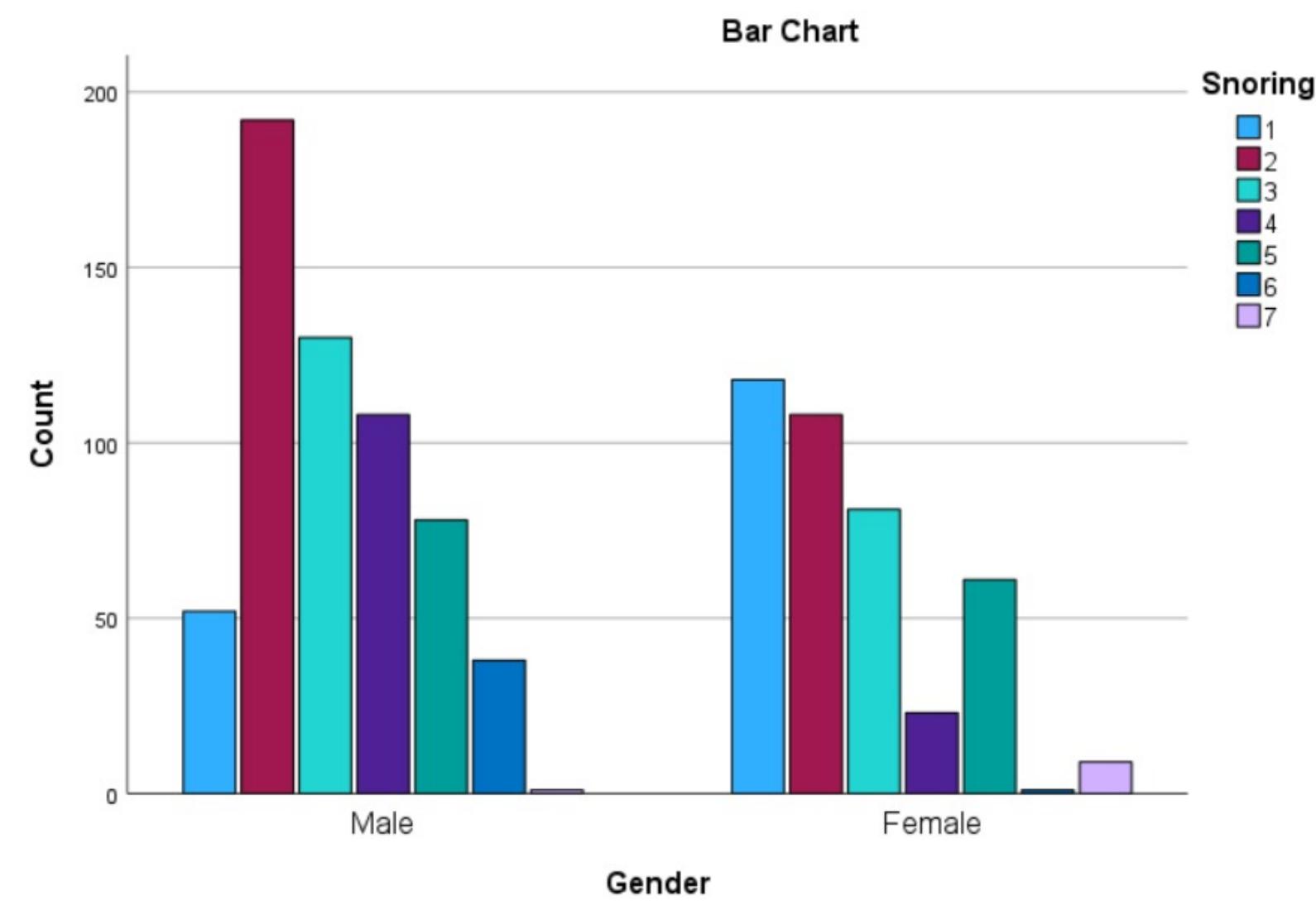
The image shows three overlapping dialog boxes from a statistical software interface:

- Crosstabs** Dialog:
  - Row(s):** Gender
  - Column(s):** Air Pollution [AirPollu...], Alcohol use [Alcohol...]
  - Layer 1 of 1**: Previous, Next
  - Display clustered bar charts
  - Suppress tables
- Crosstabs: Cell Display** Dialog:
  - Counts**:
    - Observed
    - Expected
    - Hide small counts
  - z-test**:
    - Compare column proportions
    - Adjust p-values (Bonferroni method)
  - Percentages**:
    - Row
    - Column
    - Total
  - Residuals**:
    - Unstandardized
    - Standardized
    - Adjusted standardized
  - Create APA style table
  - Noninteger Weights**:
    - Round cell counts
    - Truncate cell counts
    - No adjustments
    - Round case weights
    - Truncate case weights
- Crosstabs: Table Format** Dialog:
  - Row Order**:
    - Ascending
    - Descending



# Crosstab

		Snoring							Total	
		1	2	3	4	5	6	7		
Gender	Male	Count	52	192	130	108	78	38	1	599
	Male	% within Gender	8,7%	32,1%	21,7%	18,0%	13,0%	6,3%	0,2%	100,0%
	Female	% within Snoring	30,6%	64,0%	61,6%	82,4%	56,1%	97,4%	10,0%	59,9%
	Female	% of Total	5,2%	19,2%	13,0%	10,8%	7,8%	3,8%	0,1%	59,9%
Total	Male	Count	118	108	81	23	61	1	9	401
	Male	% within Gender	29,4%	26,9%	20,2%	5,7%	15,2%	0,2%	2,2%	100,0%
	Female	% within Snoring	69,4%	36,0%	38,4%	17,6%	43,9%	2,6%	90,0%	40,1%
	Female	% of Total	11,8%	10,8%	8,1%	2,3%	6,1%	0,1%	0,9%	40,1%
		Count	170	300	211	131	139	39	10	1000
		% within Gender	17,0%	30,0%	21,1%	13,1%	13,9%	3,9%	1,0%	100,0%
		% within Snoring	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% of Total	17,0%	30,0%	21,1%	13,1%	13,9%	3,9%	1,0%	100,0%





# Thank you!

