Exploring Relationships and Testing Hypotheses in CDC Cancer Data

NURS 60N

Table of contents

1 Objectives

1. To investigate the relationships between:
    1. Cancer site,
    2. Patient's race, age, & sex assigned at birth, and
    3. The number of patients diagnosed with cancer using descriptive statistics in SPSS.
2. Conduct *t*-tests on comparisons of your choice and reflect on potential confounding variables.

2 Background

The Centers for Disease Control and Prevention monitors the rates of cancers across the US and makes portions of those data available to the public. This activity will use one set of those data to guide you through common steps in exploring data. I have already downloaded a subset of these data and prepared them for you. I have not included the place where the participants live (to reduce the already-large size of the data), so these include the number of people diagnosed with any type of malignant neoplasm at a given body site.

3 Using Excel

- This cancer incidence dataset is an Excel-formatted version of the data.

3.1 Descriptive Statistics

3.1.1 Generate Descriptive Statistics

3.1.1.1 Descriptives of `Count`

We will use Excel to generate descriptive statistics for the `Count` variable. This variable represents the number of people diagnosed with cancer at a given site, broken down by race and 10-year age group.

1. Open your data in Excel and locate the column labeled `Count`. This column contains the number of cancer diagnoses.
2. Click on the Data tab at the top and then select Data Analysis (if it's not visible, you may need to enable the "Analysis ToolPak" add-in).
3. In the Data Analysis dialog, select Descriptive Statistics and click OK.
4. In the Descriptive Statistics dialog box:
    - Input the range for the `Count` data.
    - Select the checkbox for Summary Statistics.
    - Choose to display the output in a new worksheet or in the same worksheet.
5. Click OK to generate the descriptive statistics.

3.1.2 Review the Output

1. After clicking OK, Excel will generate a table of descriptive statistics for the `Count` variable. This table will look something like this:

Statistic	Value
Mean	X.XX
Standard Deviation	X.XX
Median	X.XX
Minimum	X.XX
Maximum	X.XX
Range	X.XX
Skewness	X.XX
Kurtosis	X.XX

3.1.3 Interpret the Descriptive Statistics

1. The mean gives an idea of the average number of cancer diagnoses.
2. If the mean is much higher than the median, it suggests a right-skewed distribution with some high values pulling the mean up.
3. The range and standard deviation help in understanding how spread out the data are.
4. Skewness and kurtosis provide additional details about the data distribution. A high skew indicates more extreme values on one side, while high kurtosis suggests the presence of outliers.

3.1.4 Confidence Intervals in Excel

To calculate 95% confidence intervals for the `Count` variable, we need to create our own formula:

1. Compute the **standard error (SE)**:

$$SE = \frac{\text{Standard Deviation}}{\sqrt{N}}$$

   Use the formula in Excel:

```
=STDEV.S(B2:B100)/SQRT(COUNT(B2:B100))
```

2. Calculate the **margin of error** for a 95% confidence interval:

$$\text{Margin of Error} = 1.96 \times SE$$

   Excel formula:

```
=1.96 * (STDEV.S(B2:B100)/SQRT(COUNT(B2:B100)))
```

3. Compute the **confidence interval bounds**:

   - **Lower Bound**: `= AVERAGE(B2:B100) - Margin_of_Error`
   - **Upper Bound**: `= AVERAGE(B2:B100) + Margin_of_Error`

3.1.5 Correlations

1. Click on the `Data` tab and select `Data Analysis`.
2. Choose `Correlation` and click `OK`.
3. Select the input range (e.g., `Count` and `Year` columns).
4. Check `Labels in First Row` if applicable.

5. Click `OK` to generate a correlation matrix.

If the correlation coefficient is close to `1`, it indicates a strong positive relationship; if close to `-1`, it suggests a strong negative relationship.

3.1.6 *t*-Test

To compare mean cancer counts by sex (female vs. male): 1. Click on `Data` > `Data Analysis`. 2. Select `t-Test: Two-Sample Assuming Equal Variances` (or `Unequal Variances` if Levene's test suggests so). 3. For `Variable 1 Range`, select all values in `Count` for females. 4. For `Variable 2 Range`, select all values in `Count` for males. 5. Enter `0.05` as the significance level (alpha). 6. Click `OK`.

Excel will output: - `t-statistic` and `p-value` (compare with `0.05` for significance) - `Mean difference` between the two groups

3.2 Descriptives of Two Variables

3.2.1 Using the PivotTable Function in Excel

1. Select your data and go to the Insert tab, then click on PivotTable.
2. In the PivotTable Field List, drag `Count` into the Values box and set the Summarize by option to Average.
3. Drag any categorical variable (like `Race`, `Sex`, or `Age Group`) into the Rows box to segment the data by that variable.
4. To get additional descriptive statistics such as Standard Deviation, click the drop-down next to `Count` in the PivotTable and select Value Field Settings, then choose Standard Deviation.
5. This will generate a table summarizing the average and standard deviation of `Count` by the selected categorical variable.

3.2.2 Review the Output

1. The PivotTable will show descriptive statistics for `Count` segmented by the chosen categorical variable (e.g., `Race` or `Sex`).
2. It will display values such as mean, standard deviation, and count, showing the distribution of `Count` across different categories.

3.2.3 Interpret the Output

1. Compare the mean and standard deviation for different groups. If there is high variability within a group, the standard deviation will be large.
2. Consider skewness and kurtosis for additional insights into the distribution of the data. These calculations may need to be done separately outside of PivotTables.

4 Using SPSS

4.1 Materials Needed

- SPSS software
- Cancer incidence dataset

4.2 Steps

4.2.1 Data Preparation

1. Please follow these instructions to open SPSS online through Apporto, the method made available to us.
2. The data are available as a `.sav` file from here. This is SPSS's native format for a data file.
3. Load the cancer incidence dataset into SPSS. Note that although I've not included all of the possible variables in these data, it's still a pretty large file—about 4.5 MB and 50k rows—so it may take a couple of seconds to load.

4.2.2 Orientation to SPSS

SPSS is organized into a set of windows with drop-down menus at the top of each. The two windows you will work with most (if not exclusively) are the `Data Editor` and the `Output` window.

The `Data Editor` window shows the actual data. It has two tabs which show the data in two ways:

- The `Data View` tab shows the data like it would be shown in a spreadsheet program like MS Excel, LO Calc, or Google Sheets. Here, each column is a variable and each row is one instance of data, here a group of people in a given year who are all of the same race and age group.

- The `Variable View` tab instead shows a listing of the variables—not the actual data—and the characteristics of that variable, such as whether it is a "nominal" or "scale" variable. A nominal variable separates people out into groups that only differ by the name they're given; no one group is "more" than any other. A scale variable assigns a numerical value, like how tall they are or—like here—how many people were diagnosed with a given type off cancer.

More about the SPSS interface can be found here.

4.2.3 Descriptive Statistics

4.2.3.1 Generate Descriptive Statistics

4.2.3.1.1 Descriptives of `Count`

We will first use SPSS to look at one variable alone, in this case, the number of incidents of cancer per age group. More specifically, the `Count` variable is the number of people of a given race and in a given 10-year age group who were diagnosed with cancer at a given site.

1. In the `Data View` or `Variable View` in SPSS's `Data Editor` window, please locate the variable labeled `Count`. This variable represents the number of people diagnosed with cancer at the given site (and thus likely the general type of cancer, although there is no indication here that this is the site of origin).
2. Go to the `Analyze` menu at the top of SPSS and select `Descriptive Statistics`.
3. In the `Descriptives` dialog box, you will see a list of variables. Select `Count` and move it into the `Variables` box.
4. Under `Options,` you can choose various statistics you want to compute. Common choices include:
    - Mean: The average number of cancer diagnoses.
    - Median: The middle value in the data, which is less affected by outliers.
    - Standard Deviation: A measure of the spread or variability in the data, it's also the average distance of scores from the mean.

4.2.3.2 Review the Output

1. After configuring the settings, click `OK`. SPSS will generate the descriptive statistics for the `Count` variable.
2. The output will include a table with the statistics you selected. It will look something like this: