SPSS Session 1

Data can be either numeric, or categorical.

Examples of Numeric variables:

- What is your GPA? (1.0 to 4.0)
- About how many hours per week do you expect to work at an outside job this semester?
- What do you think is the ideal number of children for a married couple?
- Counting yourself, how many children did your parents have?
- How far from campus do you live, in miles? (0 if on campus)
- What is the approximate population of Indonesia? (in millions)
- What is the age of the car you usually drive? (years)

The values for numeric variables are numbers representing scale.
Numeric variables can be discrete or continuous.
Numeric variables always have meaningful units of measurement.

Examples of categorical variables:

- What is your gender? (Male=0, Female=1)
- Who is your statistics instructor? (1=Wharton, 2=Sase, 3=Murphy, 4=Tountas)
- Which political orientation most nearly fits you? (1=Liberal, 2=Middle-of-Road, 3=Conservative)
- How often do you read a daily newspaper?(0=Never, 1 =Occasionally, 2=Regularly)
- How often do you exercise (aerobics, running, etc)? (0=Not At All, 1=Sometimes, 2=Regularly)
- Do you like cats? (0=No, 1=Yes)

The values for categorical variables are the options that can be represented either by text, or by numeric code.
If the categorical variables are presented by numeric code, they do not turn into numeric variables! Consider, the zipcode (numbers are meaningless, they refer to location code); the telephone number, the bank card number and etc.

Numeric and categorical variables must be treated differently in SPSS.

1) **SPSS Interface**: in the left lower corner there are tabs **"Data View"** and **"Variable View"**. The Variable View tab shows the spreadsheet of variables, the data set itself. The Data View tab shows the description of variables in the data set.

2) **For numeric variables**, the basic statistical analysis will include **the descriptive statistics** (summary of measures such as mean, min, max, standard deviation, and etc.).

Analyze -> Descriptive Statistics -> Descriptives -> Choose one or more variables -> Ok

Additionally, when choosing variables you may click on **Options**, and add or reduce the measures that will be reported. For example, you may add variance and range.

This is the output that you get if you request descriptives for variables "Height" and "cupsofcoffee":

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
height in inches	270	40	39	79	68.93	4.479	20.059
cups of coffee per day	277	7	0	7	.45	1.036	1.073
Valid N (listwise)	268						

270 people answered the question "height" (sample size); the range of their heights is 40 inches with the lowest of 39 and the highest of 79. The mean height is 68.93 inches with the standard deviation of 4.479 inches and the variance of 20.059 squared inches.
Also, for the numeric variables, we may find the frequency distribution. This will be a valid procedure, although usually not very useful.

3) **For categorical variables**, the descriptive statistics makes no sense! Consider the "class" variable, where 1 = Freshman, 2 = Sophomore, 3 = Junior, 4 = Senior.

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
class	268	3.00	1.00	4.00	2.1940	1.06688	1.138
Valid N (listwise)	268						

The average, standard deviation and other kinds of arithmetic measures do not represent anything. 2.194 does not mean that the average student is approximately a sophomore student!

For categorical variables, the most important and useful analysis is **the frequency distribution**.

Analyze -> Descriptive Statistics -> Frequencies -> Choose one or more variables -> Ok

class

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Freshman	84	30.1	31.3	31.3
	Sophomore	95	34.1	35.4	66.8
	Junior	42	15.1	15.7	82.5
	Senior	47	16.8	17.5	100.0
	Total	268	96.1	100.0	
Missing	System	11	3.9		
Total		279	100.0		

The first table shows the absolute frequency of values. It reports 11 missing values. The second and the third columns show the relative frequency including and excluding missing values respectively. The fourth column shows the cumulative percent (for example 66.8 = 31.3 + 35.4; 82.5 =31.3 + 35.4 + 15.7, and etc.)

Summarizing the information of this session

	Numeric Variables	Categorical Variables
Descriptives	**Valid, useful**	**Not valid**
Frequencies	**Valid, not useful**	**Valid, useful**

SPSS Session 2

1) **Selecting cases.**
You know now that you can find descriptives for numeric variables and frequencies for categorical variables. So far, we were considering the whole sample, i.e. the analysis covered each and every observation.
Sometimes, you may be interested in analyzing some part of the sample. For example, you may limit your interest to people of some particular level of income, or particular city, or particular race. In fact, you may limit your sample according to any possible rule. You are not dropping some observations from your data set, you just temporarily exclude them from the analysis.
To **select cases:**

Data -> Select Cases -> If condition is satisfied -> If -> Specify the condition -> Ok

The examples of condition:

- HI = 1;
- Population <= 5,000,000;
- Income > 30000 & CollegeGrad <= 27
  ("&" is AND operator);
- Income > 30000 | CollegeGrad <= 27
  ("|" is OR operator);
- HI ~= 4
  ("~=" is NOT EQUAL);
- (Unemp+10) / 6 > 2,
  and so on.

Once you selected a specific part of the sample, you can go ahead and calculate Descriptives or Frequencies for any variables, the same way you would do it with the whole sample.

For example, I have selected only states with per capita income strictly greater than $40,000. Now I can find descriptives for "Unemp" and "Income".

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Percent Unemployment Rate	19	3.8	12.4	8.332	2.0287
Per Capita Yearly Income	19	40283	56001	45166.00	4307.810
Valid N (listwise)	19				

You can see that the sample size reduced to only 19 states. For these states, average unemployment rate is 8.332% and average per capita income is $45,166.

To come back to the whole sample, go
Data -> Select Cases -> Reset.

2) **Correlation and Covariance**.

To calculate a **correlation** coefficient for two vairables:

Analyze -> Correlate -> Bivariate -> Choose Variables -> Ok

Be careful, correlation and covariance can be applied only to NUMERIC variables!

For example, we choose variables “Income” and “CollegeGrad”.
You may choose more than two variables - in this case you will be given all the pairwise correlations.

Correlations

		Per Capita Yearly Income	Percent with college degree or higher
Per Capita Yearly Income	Pearson Correlation	1	.782**
	Sig. (2-tailed)		.000
	N	50	50
Percent with college degree or higher	Pearson Correlation	.782**	1
	Sig. (2-tailed)	.000	
	N	50	50

**. Correlation is significant at the 0.01 level (2-tailed).

The correlations table is usually presented as matrix. If you choose 3 variables, you will be getting 3*3 table, and so on.
In every result cell, you are given the correlation coefficient (first), the significance level (second), the number of observations (third). We don’t need the last two measures; pay attention to the correlation coefficient. Obviously, the variable correlation with itself equals 1. The correlation between Income and CollegeGrad is 0.782, as you can see from two cells.

To find the **covariance**, when choosing variables click **Options** and add Cross-product deviations and covariances.

Correlations

		Per Capita Yearly Income	Percent with college degree or higher
Per Capita Yearly Income	Pearson Correlation	1	.782**
	Sig. (2-tailed)		.000
	Sum of Squares and Cross-products	1.541E9	1035234.650
	Covariance	31456703.480	21127.238
	N	50	50
Percent with college degree or higher	Pearson Correlation	.782**	1
	Sig. (2-tailed)	.000	
	Sum of Squares and Cross-products	1035234.650	1136.625
	Covariance	21127.238	23.196
	N	50	50

**. Correlation is significant at the 0.01 level (2-tailed).

Now in every cell you see two more measures. You are interested in the covariance.
The covariance of variable with itself is simply a variance of that variable. You can see that the variance of "Income" is 31,456,703.480 and the variance of "CollegeGrad" is 23,196.
Finally, the covariance of two variables is 21,127.238 as can be seen in both intersection cells.

SPSS Session 3

1) **Binary Variables**.
In addition to numeric and categorical types of variables, there is a third type of variables, a special one – **binary** variables.
A binary variable is a variable that takes only two values – **0** and **1**. Binary variable usually represents the presence or the absence of some property or quality. 1 stands for "yes" or "true", and 0 stands for "no" or "false".

For example,

- *Do you like cats?* 1 = yes, 0 =no
- *Are you a full-time worker?* 1 = yes, 0 = no
- *Male*. 1 = yes, 0 = no

Binary variables are special because they have properties of both numeric and categorical variables. Thus, for binary variables both descriptive statistics and frequencies make sense.

Consider, the variable "w*ebsite*".
Descriptives:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
The firm has website	5914	0	1	.43	.495
Valid N (listwise)	5914				

We see that the sample covers 5,914 firms. The minimum value is 0, the maximum is 1 (these are the only two values this variable takes). The mean value is 0.43 which means that there is 43% of ones and 57% of zeroes. In other words, 43% of firms have website and 57% don't. The standard deviation is 0.495.

Frequencies:

The firm has website

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	3388	57.3	57.3	57.3
	yes	2526	42.7	42.7	100.0
	Total	5914	100.0	100.0	

We have 3,388 zeroes and 2,526 ones. This means 3,388 firms don't have website, 2526 firms do have. The percentage of firms that have website is 42.7% which is the same number that we have seen in the descriptives table.

This is an updated summary of the variables types.

	Numeric Variables	Binary Variables	Categorical Variables
Descriptives	**Valid, useful**	**Valid, useful**	**Not valid**
Frequencies	**Valid, not useful**	**Valid, useful**	**Valid, useful**

2) **Creating new variables**.
In SPSS, you can create new variables by going

Transform -> Compute Variable -> Enter target variable -> Enter Numeric Expression -> Enter the "If" condition (*optional*) -> Ok

For example:

- *productivity = sales/labor* for the whole sample

Target variable: "*productivity"* (name of the new variable)
Expression: *sales/labor*

Other examples:

- *lnlabor* = log(labor)
- *senseless_variable* = (labor + capital)*5 – innov if sales=70,000,000

- *x* = 10

This variable will take a value of 10 for every observation, i.e. this will make a column of tens.

You can create a categorical variable out of a numeric variable:

- *size* = 1 if labor <= 10
  = 2 if 10 < labor <= 100
  = 3 if labor > 100
  1 will represent small, 2 – medium, 3 – large size of a firm.

You can create a binary variable out of a numeric variable:

- *huge_size* = 1 if labor > 1000
  = 0 if labor <= 1000.

You can create a binary variable out of a categorical variable:

- *Armenia* = 1 if country = 1
  = 0 if country ~= 1.

SPSS Session 4

Contingency Tables

When analyzing a data set we may be interested in cross-tabulations of frequencies for two or more variables. Using contingency table, we are able to answer all kinds of questions about marginal, joint, and conditional probabilities and about other probability properties, such as independence.

To create a contingency table we do

Analyze -> Descriptive Statistics -> Crosstabs -> Choose Row Variable and Column Variable -> Ok

For example, this is the table we can get for "eye_color" and "hair_color" tabulation:

eye_color * hair_color Crosstabulation

Count

		hair_color				
		Black	Blond	Brown	Red	Total
eye_color	Blue	20	94	84	17	215
	Brown	68	7	86	26	187
	Green	5	16	29	14	64
	Hazel	15	10	54	14	93
Total		108	127	253	71	559

Every number represents absolute two-way frequency. For example, there are 20 people with blue eyes and black hair, 94 people with blue eyes and blond hair, and etc. Out of this table, we can calculate any necessary indices or probabilities.

For a contingency table we may choose any type of variable – numeric, categorical, or binary. The main limitation is the number of values. If there are too many values for a numeric variable, or too many options for a categorical variable, the contingency table is still valid, but gets not useful.

This last step is not necessary, but may be useful.

You may add to this table the relative frequencies (percentages). When choosing variables, you can click on **Cells** and add Percentages (row, column, or total, or all of them).

Total percentages will be the ratios of every number in the table to the grand total.
Row percentages will be the ratios of every number to the row total of that number.
Column percentages will be the ratios of every number to the column total of that number.

SPSS Session 5-6

For the goals of inferential analysis, we only need to gather descriptive statistics for key variables and use it to construct confidence intervals or to test statistical hypotheses. Below are examples.

Confidence Intervals

Question: What is the 99% confidence interval for the population average price of a house?
Solution: To build a confidence interval for the population mean with unknown population standard deviation, we need the sample average, the sample size and the sample standard deviation for the variable "price". Obviously, we can get them using the Descriptives menu.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
house price, in $1000	1000	134.32	345.20	247.6557	42.19273
Valid N (listwise)	1000				

From this table we get $\bar{x} = 247.6557,\ n = 1000,\ s = 42.19273$.
The value from the distribution $t_{0.5\%}(999) = z = 2.58$.
Using the formula $\bar{x} \pm t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$, we get the confidence interval $244.213 \le \mu \le 251.098$.

One-Sample Hypothesis Tests

Question: At the 5% significance level, test whether for the houses that have pool, the population standard deviation of house age is greater than 9 years.
Solution: At first, we select only houses that have pool. The "if" condition would be "pool"=1. Then, to test the hypothesis about the population standard deviation of age, we need the sample size and the sample standard deviation for the variable "age".

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
house age, in years	204	0	49	9.91	9.936
Valid N (listwise)	204				

From this table we get $n = 204,\ s = 9.936$.
The hypotheses are $H_0: \sigma^2 \le 81$ vs $H_0: \sigma^2 > 81$.
Using the formula $\chi^2_{calc} = \frac{(n-1)s^2}{\sigma_0^2}$, we find the test statistic to be 247.42.
In the right-tailed test the critical value is $\chi^2_{crit} = \chi^2_{5\%}(203) \approx \chi^2_{5\%}(200) = 234.0$.
$247.42 > 234.0$, so the decision is to reject H_0.

The conclusion is that the standard deviation of house age indeed is greater than 9 years.

Two-Sample Hypothesis Tests

Question: At the 10% significance level, is the population proportion of houses close to university among the houses with more than 25 square feet of living area smaller than the population proportion of houses close to university among all houses?
Solution: For this test we need to compare parameters of two populations, so we need descriptive statistics for two samples.

Sample 1: houses with more than 25 square feet of living area.
We select the indicated cases and find the descriptive statistics for the binary variable "utown".

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
utown	534	0	1	.51	.500
Valid N (listwise)	534				

From this table we get $n_1 = 534,\ p_1 = 0.51$. The mean of the binary variable is nothing but the proportion of ones.

Sample 2: all houses.
We reset the selection of cases and get back to the whole sample. We get the descriptive statistics for the binary variable "utown".

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
utown	1000	0	1	.52	.500
Valid N (listwise)	1000				

From this table we get $n_2 = 1000,\ p_2 = 0.52$.

The hypotheses are $H_0: \pi_1 - \pi_2 \geq 0$ vs $H_1: \pi_1 - \pi_2 < 0$.
$x_1 = p_1 n_1 = 0.51 * 534 = 272.34, x_2 = p_2 n_2 = 0.52 * 1000 = 520.$
$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{272.34 + 520}{534 + 1000} = 0.5165.$
Using the formula $z_{calc} = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)_0}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, we find the test statistic $z_{calc} = -0.373$.
In the left-tailed test, the 10% critical value of the z distribution is -1.28.
$-1.28 < -0.373$, so the decision is to fail to reject H_0.
The population proportion of houses close to university for houses with more than 25 square feet of living area is not smaller than that for all houses.

SPSS Session 7

1) **Visual Display: Scatter Plot.**
Scatter plot is a useful tool to examine the relationship between any two variables.
To create a **scatter plot**:

Graphs -> Chart Builder -> Choose Scatter/Dot in the Gallery Tab ->
Drag the first picture to the chart preview ->
Choose an X-variable, Drag it to the chart preview ->
Choose a Y-variable, Drag it to the chart preview ->
Ok

When analyzing the visual information, pay an attention to the strength of relationship, to its sign, and to linearity/nonlinearity. This may give you a clue to a better model. Besides, you may confirm you visual findings with numeric characteristic, such as correlation coefficient.
Also, scatter plot may reveal outliers in your data. If there are ones, you should make a decision how to treat them – whether to exclude them from analysis or not. One solution would be to analyze data first including outliers - then excluding them, and to compare results.

2) **Regressions**.

To run a regression:

Analyze -> Regression -> Linear -> Choose Dependent Variable -> Choose Independent Variable(s) -> Ok

Regressions use only numeric and binary variables! Categorical variables in regression will make **no sense**, though you technically may put in a regression any variables.
This fact does not mean that qualitative information that categorical variables carry is excluded from the analysis. It means that you cannot use categorical variables as they are. But you may (and should) create binary variables representing categories and include them in the model if necessary.

You may incorporate in your model variables in different functional forms. They may be polynomial, inverse, logarithmic, power, exponential, interaction, and any other forms if it makes sense to explain the behavior of dependent variable.

For example, to estimate a model
$\ln(pizza) = \beta_0 + \beta_1 \cdot \ln(income) + \beta_2 \cdot age + \beta_3 \cdot female + \beta_5 \cdot college + \varepsilon,$

we need to create two new variables *lnpizza* = ln(pizza) and *lnincome*=ln(income). Now the dependent variable is *lnpizza* and the independent variables are *lnincome*, *age*, *female*, and *college*.

The regression yields the following output:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
	college, female, age in years, lnincome	.	Enter

a. All requested variables entered.

b. Dependent Variable: lnpizze

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
	.696[a]	.484	.418	.68348

a. Predictors: (Constant), college, female, age in years, lnincome

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	13.610	4	3.403	7.284	.000[a]
	Residual	14.482	31	.467		
	Total	28.092	35			

a. Predictors: (Constant), college, female, age in years, lnincome

b. Dependent Variable: lnpizze

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
	(Constant)	.757	1.894		.400	.692
	lnincome	.630	.212	.527	2.971	.006
	age in years	-.048	.015	-.524	-3.128	.004
	female	-.967	.232	-.546	-4.173	.000
	college	-.498	.276	-.275	-1.802	.081

a. Dependent Variable: lnpizze

The interpretation of SPSS output of a regression is considered deeply in the course and is out of the scope of SPSS Session 7.

P.S. I hope everything you've learnt from the SPSS sessions will be useful in your future career.