



Descriptive Statistics in Analysis of Survey Data

March 2013

Kenneth M Coleman

Mohammad Nizamuddiin Khan



Survey: Definition

“A survey is a systematic method for gathering information from (a sample of) entities for the purposes of constructing quantitative descriptors of the attributes of the larger population of which the entities are members.” *Groves et al. Survey Methodology 2009*

1. Note the reference to “quantitative descriptors” of a larger population from which a sample is drawn. One thing survey research can yield is quantitative estimates of the amount and nature of variation in variables which interest us in larger populations. These are descriptive statistics.

2. At the heart of descriptive statistics are:

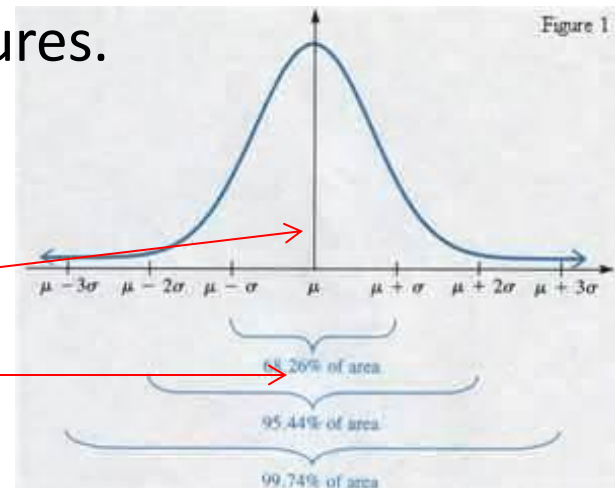
Measures of central tendency in a sample variable

Measures of dispersion in the same variable

3. In order to compare a given variable to a normal distribution, one needs both types of measures.

Measure of central tendency:

Measure of dispersion:



Measures of Central Tendency

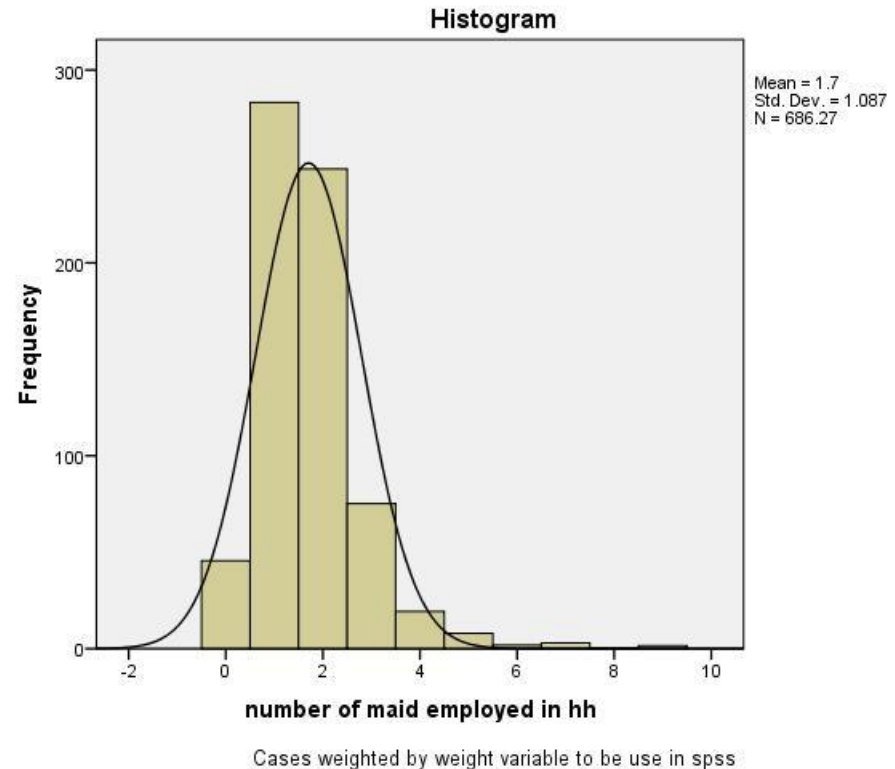
- Mean – The sum of individual values on a variable divided by the number of cases. Appropriate only for interval or scale level data, although sometimes used with ordinal data. Means can be affected outliers.
- Median – The value where half of the cases fall above and half of the cases fall below. Can be used with interval or ordinal data. Medians are less affected by outliers.
- Mode – The most frequently encountered value in a distribution. Can be used with interval, ordinal or nominal data. Not affected by outliers.

A note of caution: SPSS will calculate each of these measures for any variable. Be sure to pick a measure of central tendency appropriate to the level of measurement.

Number of Maids in Qatari Households

Mean:	1.7
Median:	2.0
Mode:	1.0

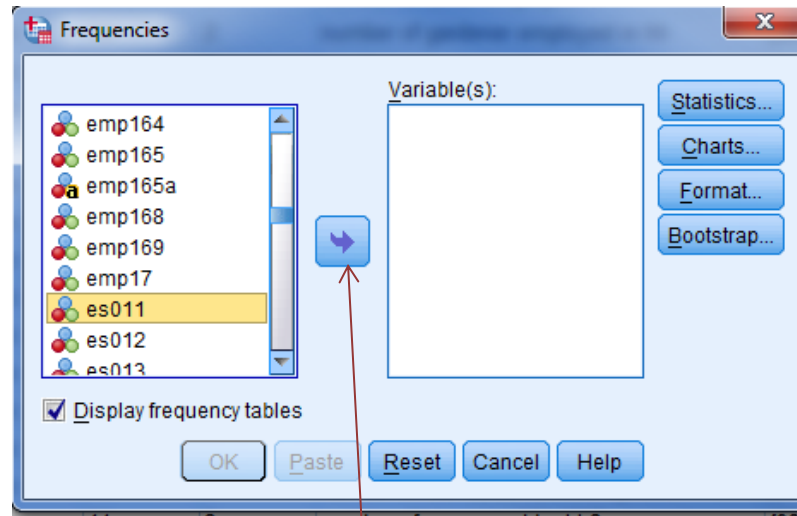
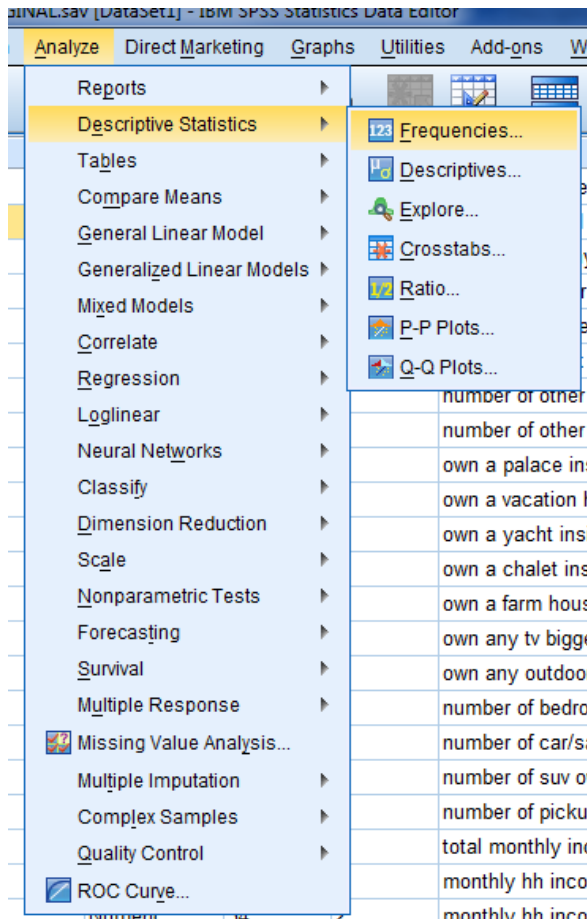
On this variable from the 2010 SESRI Omnibus survey, the mean and median are rather close, but the mode differs more substantially, as it picks up only the most frequent answer – *but there are a lot of other answers.*



As a next step, let's have you open up the **Data Set 2**, and do an exercise. Handouts will be distributed.

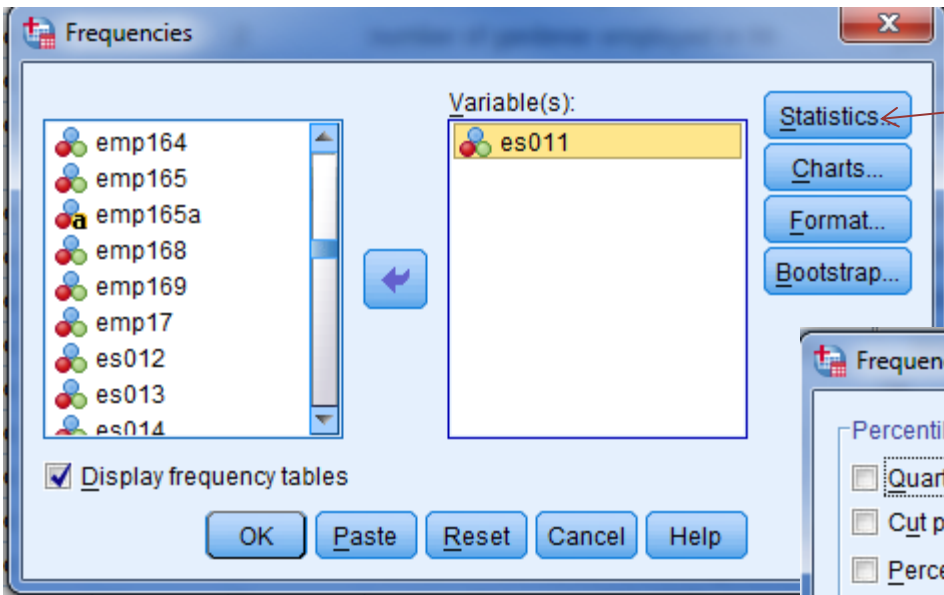
Finding the Mean, Median, & Mode of a Variable in SPSS

- The *Frequencies* command allows you to find the mean, median, and mode of a variable, in addition to listing the number of cases for each unique value of a variable. The *Frequencies* command is found under the *Analyze / Descriptive Statistics* menu.

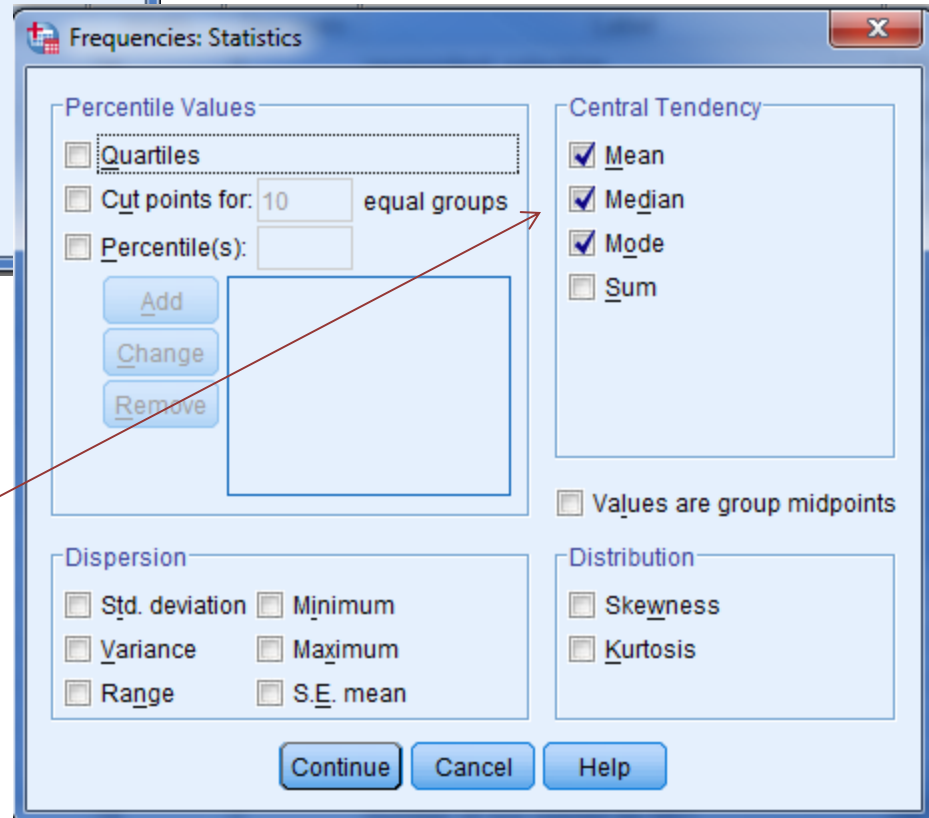


Select the variable for which you want to run the frequencies command, and use the arrow to send the variable over to the "Variable(s):" box.

Finding the Mean, Median, & Mode of a Variable in SPSS



Once you've selected your variable, select the "Statistics..." button.



Select "Mean", "Median" and "Mode."

Finding the Mean, Median, & Mode of a Variable in SPSS

Statistics

number of maid employed in hh

N	Valid	686
	Missing	1453
Mean		1.7040
Median		2.0000
Mode		1.00

The mean, median, and mode, are displayed first in the SPSS output window. Below, the frequencies for the variable **es011** are displayed.

number of maid employed in hh

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	46	2.1	6.6	6.6
	1.00	283	13.2	41.3	47.9
	2.00	249	11.6	36.2	84.1
	3.00	75	3.5	11.0	95.1
	4.00	19	.9	2.8	97.9
	5.00	8	.4	1.2	99.1
	6.00	2	.1	.3	99.3
	7.00	3	.1	.4	99.8
	8.00	0	.0	.0	99.8
	9.00	1	.1	.2	100.0
	Total	686	32.1	100.0	
Missing	98. don't know	2	.1		
	99. refused	1	.0		
	System	1450	67.8		
	Total	1453	67.9		
Total		2139	100.0		

Mean, Median and Mode Exercise [Distribute]

There are several measures of wealth or family Affluence in the 2010 Omnibus dataset, including the following:

Variable Name	Variable Label
es011	Number of maids employed in household
es012	Number of nannies employed in household
es013	Number of drivers employed in household
es014	Number of gardeners employed in household
es015	Number of cooks employed in household
es04	Number of bedrooms in household
es05	Total monthly income of all household members

For each of the above variables, find their mean, median, and mode. Use the *Analyze/Frequencies/Statistics* command.

Mean, Median and Mode Answers

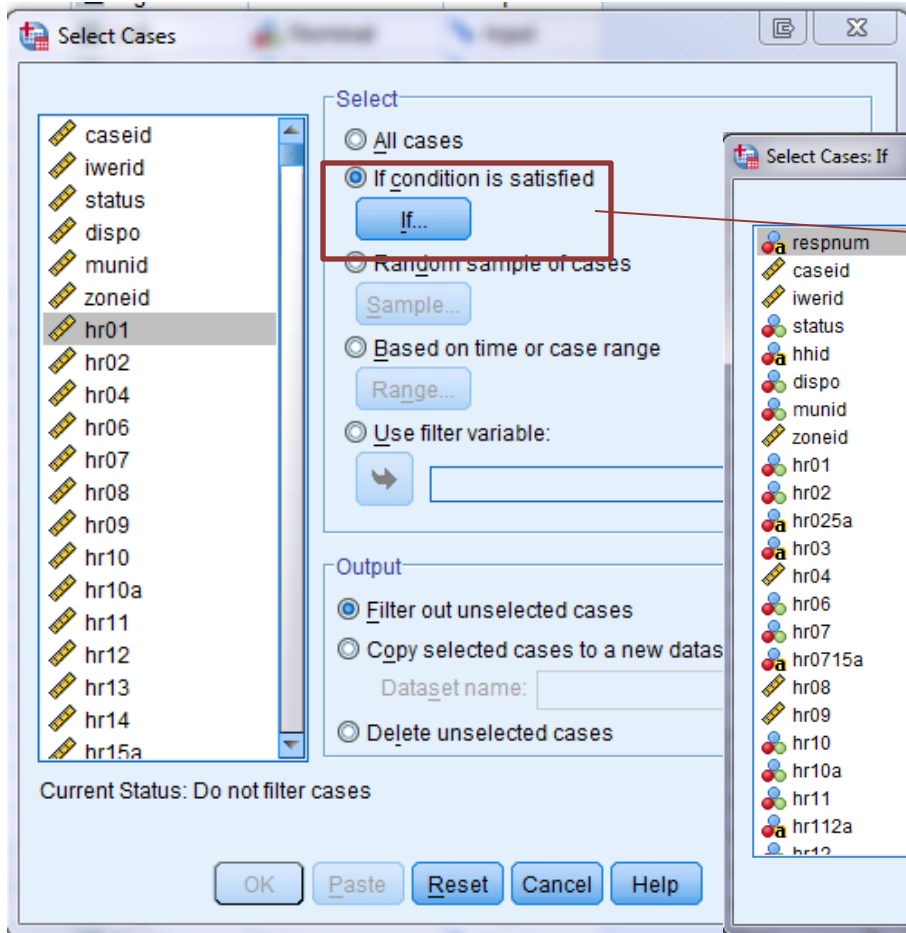
Statistics

		number of maid employed in hh	number of nanny employed in hh	number of driver employed in hh	number of gardener employed in hh	number of cook employed in hh	number of bedrooms in hh	total monthly income of all hh members
N	Valid	686	669	681	666	668	689	1359
	Missing	1453	1470	1458	1473	1471	1450	780
Mean		1.7040	.2241	.8743	.1477	.1170	5.1557	1.1188
Median		2.0000	.0000	1.0000	.0000	.0000	5.0000	1.0000
Mode		1.00	.00	1.00	.00	.00	5.00	1.00

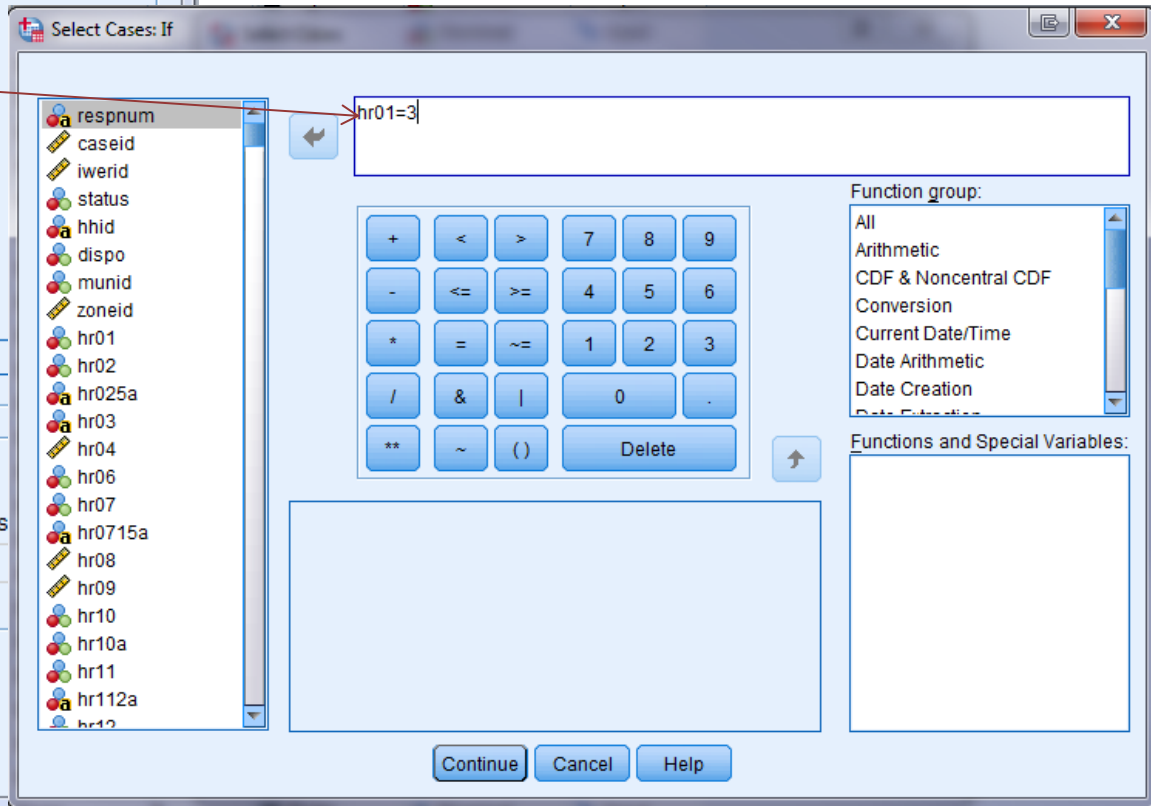
Measures of Dispersion: Range

- Range (minimum & maximum) –
 - The range is the distance between the lowest and the highest values encountered in a distribution. While rudimentary, knowledge of minimum and maximum values, plus range, gives us some useful information in certain cases.
 - Certain measures of dispersion, such as variance and standard deviation, can be properly calculated only for interval or scale measures, so minimum, maximum and range can be helpful for ordinal data.
 - Illustratively, we will look at the range of income categories we found in the 2010 Omnibus Survey among blue collar guest workers residing in Qatar.

To find the range of income categories among blue collar workers in Qatar, we first need to limit our analysis to this group. “hr01” is the variable in our dataset that indicates whether a respondent is a Qatari citizen, a white collar worker, or a blue collar worker. To limit our analysis to blue collar workers, we use the *Data/Select Cases* menu option.

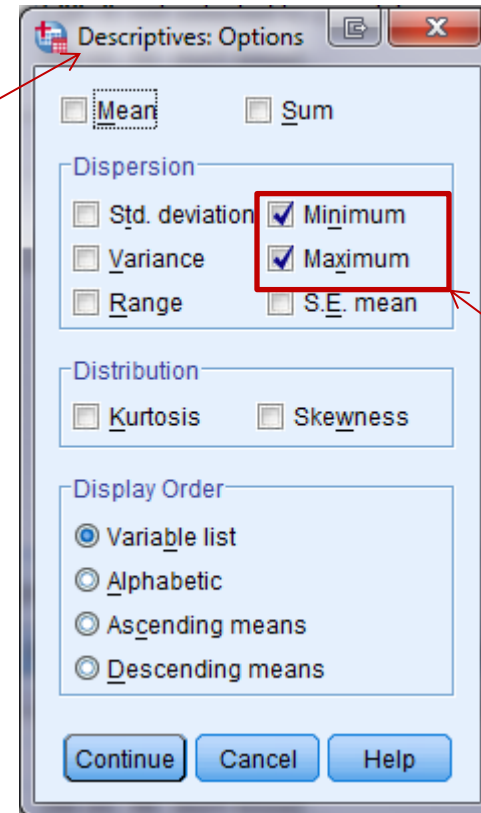
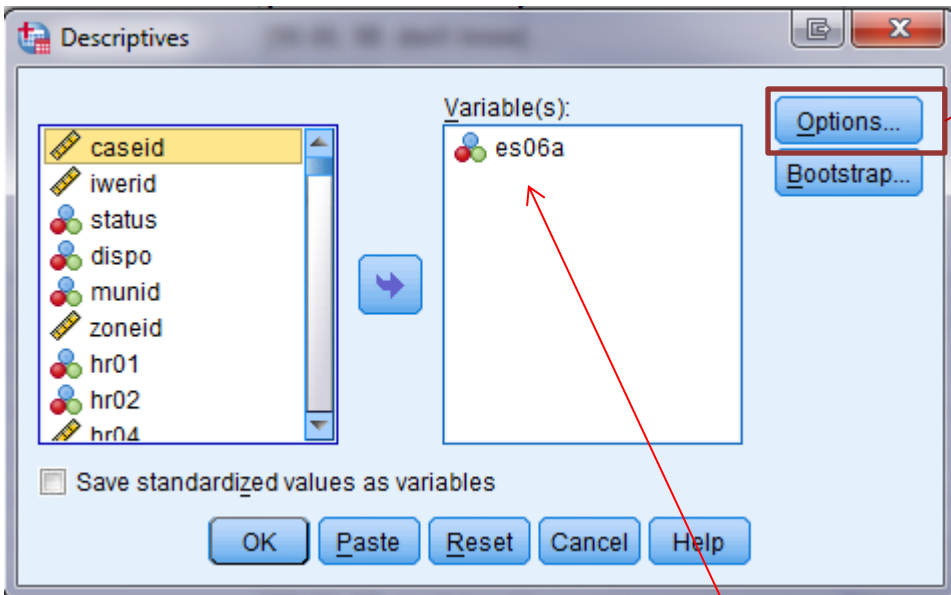


Here, we tell SPSS to limit our analysis to cases where **hr01=3** (blue collar workers).



Finding the Range

We can use the options under the *Analyze / Descriptive Statistics / Descriptives* menu option to find the range (the minimum and maximum values of the variables).



Check the Minimum and Maximum boxes.

The variable that gives us the income categories for Blue Collar Guest workers is **es06a**.

Minimum & Maximum

- The first output of minimum and maximum is not very helpful because it doesn't tell us what a 1 or 7 means on **es06a**.

Descriptive Statistics

	N	Minimum	Maximum
monthly income from all sources in qatar	674	1.00	7.00
Valid N (listwise)	674		

- But on the next page once we can see the value labels, we can learn something from the minimum and maximum.

Inferences from the Range of Responses: Does This Tell Us Anything?

Guest Workers, 2010 Omnibus Survey

monthly income from all sources in qatar

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1. less than qr600	52	7.7	7.8	7.8
	2. qr600 to lt qr800	144	21.1	21.4	29.1
	3. qr800 to lt qr1,000	100	14.7	14.9	44.0
	4. qr1,000 qr to lt qr2,000	278	40.8	41.2	85.2
	5. qr2,000 to less than qr3,000	73	10.7	10.8	96.0
	6. qr3,000 qr to less than qr4,000	18	2.7	2.7	98.8
	7. qr4,000 qr to less than qr5,000	8	1.2	1.2	100.0
	Total	674	98.9	100.0	
Missing	8. don't know	2	.2		
	9. refused	1	.2		
	System	5	.7		
	Total	8	1.1		
Total		682	100.0		

Interpreting A Range

- If one knows other data even a range of values can be useful.
 - How do these values compare to household incomes of Qatari families?
 - How do these values compare to incomes of white collar ex-patriot workers?

Measures of Dispersion: Variance

- The variance in a distribution of sample data is the sum of the squared differences between each individual value and the mean of all values, divided by the number of cases (minus one), or mathematically:

$$s^2 = \sum (x_i - \bar{x})^2 / (n - 1)$$

- Note that this quantity involves the square of those differences, not just the differences.

Measures of Dispersion:

Standard Deviation

- The standard deviation is commonly used to characterize the dispersion of an array of cases around a measure of central tendency, the mean. Mathematically, it can be seen as:
- The **sample standard deviation formula** is:

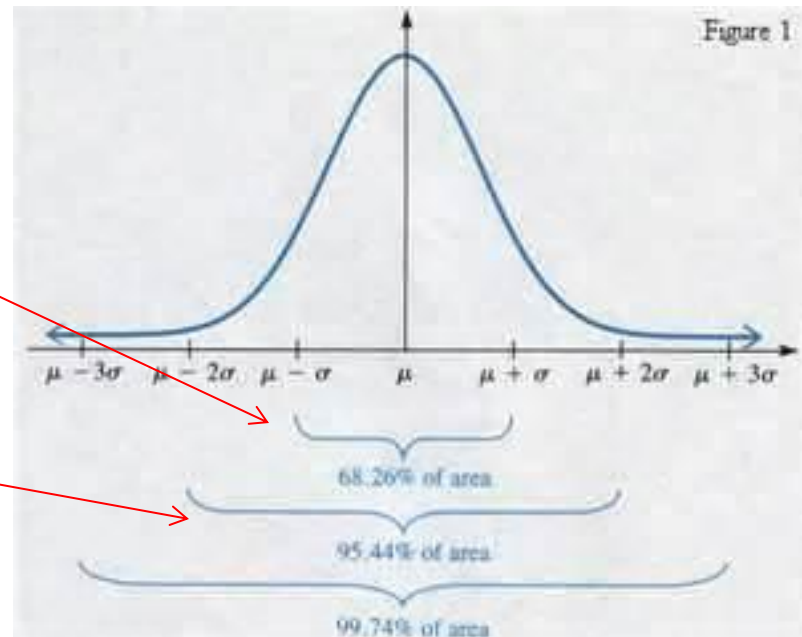
$$s = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}}$$

- Note that this formula takes the square root of the variance.

Normal Distribution

- One important thing to know about a normal distribution is that under a normal curve there will be a constant area (or proportion of cases) between the mean and an ordinate which is a given distance from the mean in terms of standard deviation units. This was seen in my introductory graphic.

Note that roughly 2/3 of cases [68.26%] fall within ± 1 standard deviation of the mean, with 95.44% fall within 2 standard deviations of the mean.



Comparison to a Normal Distribution

- There are various measures that help one to discern how far a given distribution of values deviates from a normal distribution. We will consider two such measures, but first let us characterize a normal distribution.
 - The first thing to note is that there are many different normal distributions, one for every combination of mean and standard deviation.
- The following graph illustrates that point well.

Differing Normal Distributions

- Normal distributions with differing means

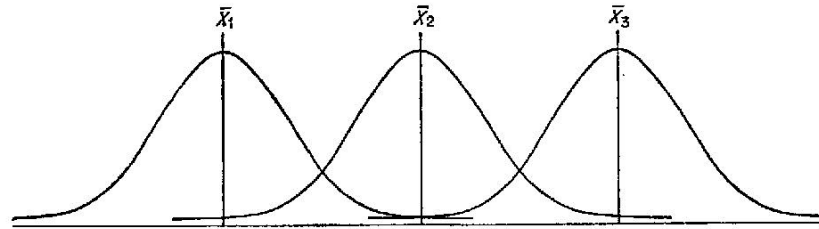


Fig. 7.4. Comparison of normal curves with the same standard deviations but different means.

- Normal distributions with differing standard deviations:

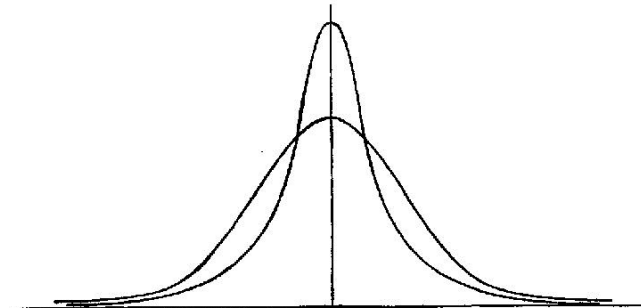


Fig. 7.5. Comparison of two normal curves with the same means but different standard deviations.

- Normal distributions with same standard deviations but differing levels of peakedness (kurtosis)

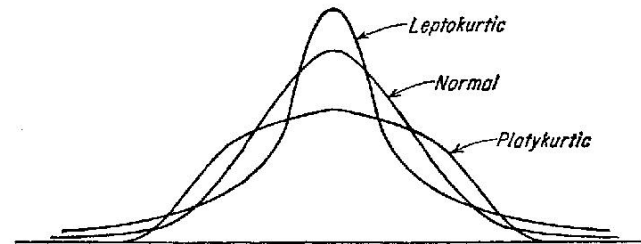


Fig. 7.6. Comparison of normal curve with curves having the same standard deviation but which differ with respect to peakedness.

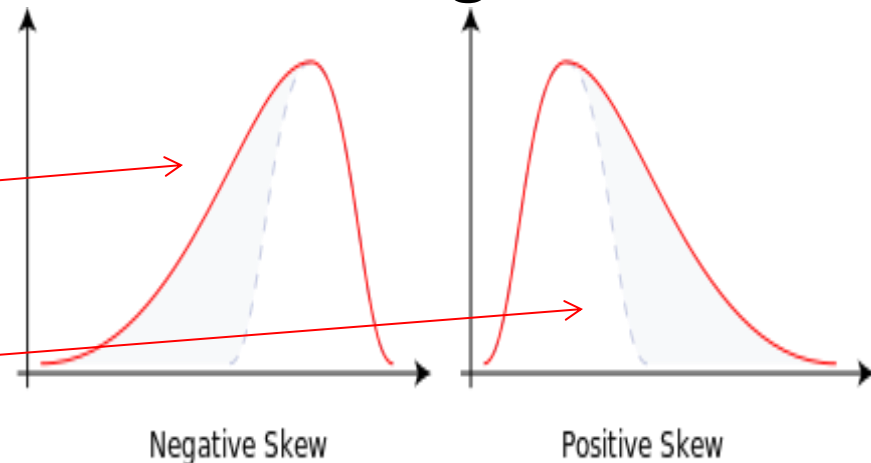
Measures of Dispersion:

Skewness

- Skewness is calculated by this formula:
 - $\frac{3(\text{mean} - \text{median})}{s}$ [or the standard deviation, as defined above]
- The essential point of Skewness is an assessment of the difference between the mean and the median. The greater that distance, the greater a distribution is skewed away from the median.
- Note the difference between a negative skew and a positive skew.

Mean lower than median

Mean higher than median



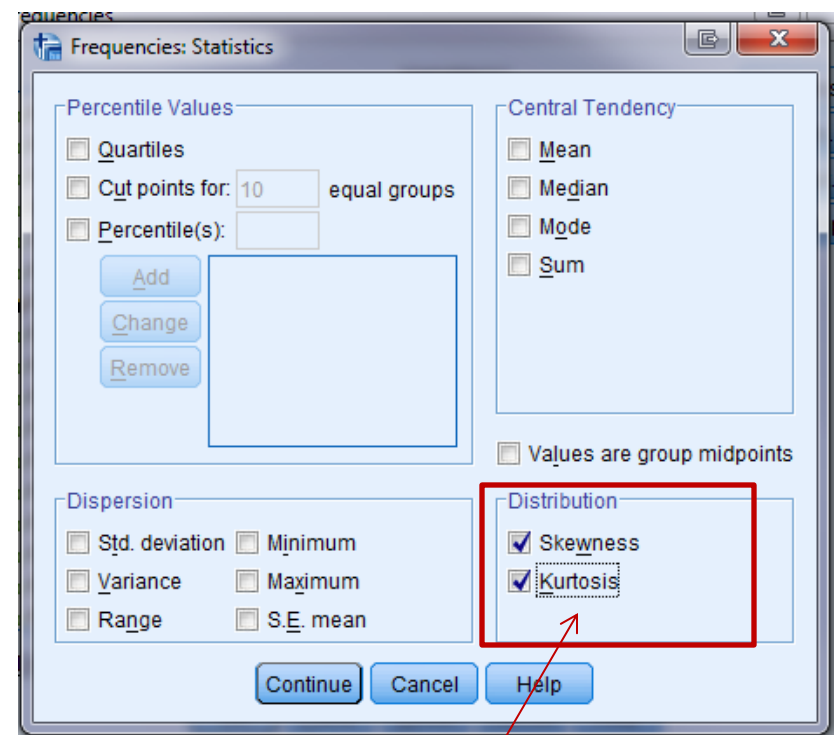
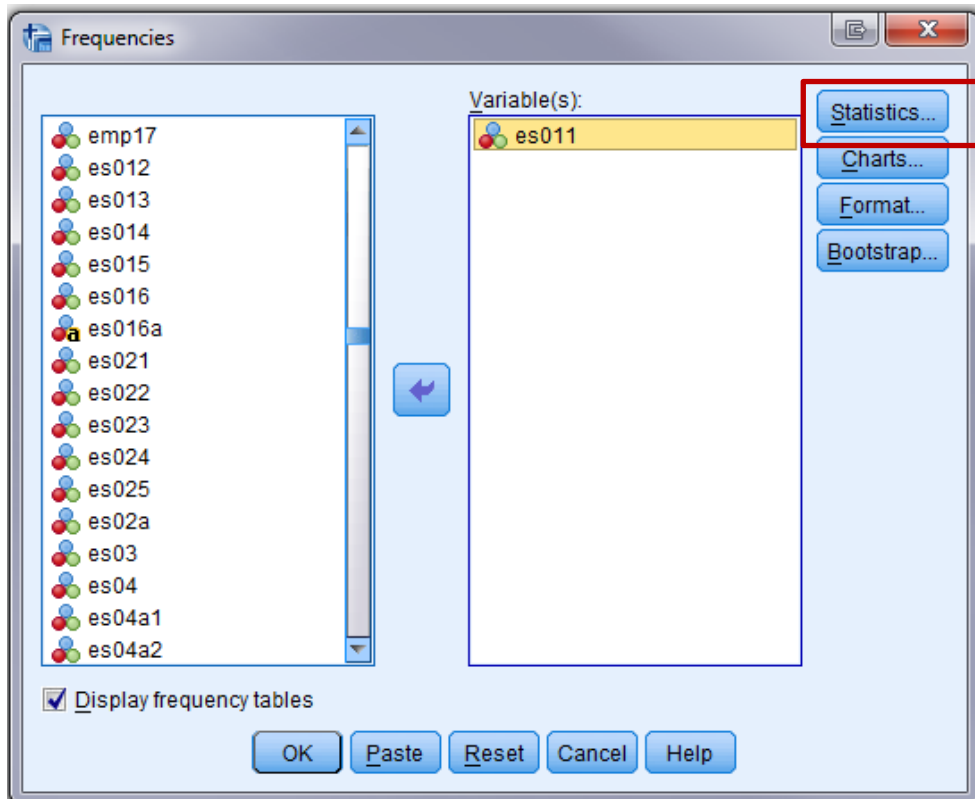
Measures of Dispersion:

Kurtosis

- Kurtosis refers to the peakedness of a distribution. Normal distributions can be more or less peaked, but any two distributions – normal or not – can also vary in the extent to which a peak occurs, ***meaning that a large number of cases share the same value [or range of values]***.
- We won't go into the definition, but will note that the higher the value of kurtosis, the more peaked the distribution.

Finding the Skewness and Kurtosis Values

SPSS can calculate the skewness and kurtosis of a variable under the *Analyze / Descriptives / Frequencies* menu option, illustrated below with the case of maids.



Select the Skewness
& Kurtosis options

Skewness & Kurtosis: Exercise

- Short Exercise: Using the same variables (**es011 [maids]** and **es014 [gardeners]**), examine Skewness and Kurtosis.
- One of these variables approaches a normal distribution in which Skewness and Kurtosis are relatively low, while the other represents a distribution in which both Skewness and Kurtosis are much larger. Get a sense of the size of the values for skewness and kurtosis in a highly skewed and peaked distribution.

Skewness & Kurtosis: Results

Examine Skewness and Kurtosis. One of these variables [number of maids employed in the hh, es011] approaches a normal distribution in which Skewness and Kurtosis are relatively low, while the other [number of gardeners, es014] represents a distribution in which both Skewness and Kurtosis are much larger, and normality is not approached.

Statistics

number of maid employed in hh

N	Valid	686
	Missing	7743
Skewness		1.776
Std. Error of Skewness		.093
Kurtosis		7.004
Std. Error of Kurtosis		.186

number of maid employed in hh

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid				
.00	46	.5	6.6	6.6
1.00	283	3.4	41.3	47.9
2.00	249	3.0	36.2	84.1
3.00	75	.9	11.0	95.1
4.00	19	.2	2.8	97.9
5.00	8	.1	1.2	99.1
6.00	2	.0	.3	99.3
7.00	3	.0	.4	99.8
8.00	0	.0	.0	99.8
9.00	1	.0	.2	100.0
Total	686	8.1	100.0	
Missing				
98. don't know	2	.0		
99. refused	1	.0		
System	7740	91.8		
Total	7743	91.9		
Total	8429	100.0		

Statistics

number of gardener employed in hh

N	Valid	666
	Missing	7763
Skewness		3.731
Std. Error of Skewness		.095
Kurtosis		20.967
Std. Error of Kurtosis		.189

number of gardener employed in hh

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid				
.00	578	6.9	86.7	86.7
1.00	82	1.0	12.4	99.1
2.00	4	.0	.6	99.7
3.00	1	.0	.1	99.8
4.00	1	.0	.2	100.0
Total	666	7.9	100.0	
Missing				
98. don't know	18	.2		
99. refused	5	.1		
System	7740	91.8		
Total	7763	92.1		
Total	8429	100.0		

Implications of Non-Normal Distributions, I

- As data analysts, we typically want variables that vary. A very high degree of kurtosis may mean that we are not capturing variation. That may imply a defect in our measuring procedure.
 - To take but one example, we would probably not find these categories useful, because most respondents would fall into one category:
 - Three categories of height: under 2 meters, 2.01 meters to 2.5 meters, 2.51 meters and up.
 - Implications:
 - Need to pretest certain categories before a final survey is launched to see if variation is captured by our measurement procedures.
 - If the final survey does not generate variation, we may need to adjust measurement procedures in future studies.

Implications of Non-Normal Distributions, II

- Most statistical procedures of inference are based on the assumption of a normal distribution. If not present, there are various possible solutions:
 - The Law of Large Numbers for sampling distributions, to be discussed below.
 - Non-parametric statistical procedures are sometimes used. One can learn about them in this volume:
 - Marjorie A. Pett, Nonparametric Statistics in Healthcare Research: Statistics for Small Samples and Unusual Distributions. Sage Publications: 1997.

Statistical Significance: The Basic Notion, I

- At various points in subsequent presentations we will refer to statistical significance.
- The notion of statistical significance addresses the probability that an outcome would be attained by chance alone. To do so, one invokes the notion of a normal sampling distribution, seeking an outcome that would occur “normally” less than one time in twenty, denoted as a probability of $p < .05$. But other, more demanding, standards may also be used, such a $p < .01$ or $p < .001$.

Statistical Significance: The Basic Notion, II

- Recall that in any normal distribution 95.44% of cases fall within ± 2 standard deviations of the mean. That trait is invoked in what is known as a two-tailed test of statistical significance, in which one considers the probability that a given statistic would occur by chance alone – either above or below the mean of such statistics that would be attained in a large number of random samples.

Statistical Significance: The Basic Notion, III

- The curve to which one compares results *when testing for statistical significance* is called a **sampling distribution**. The Law of Large Numbers holds for sampling distributions. That states that *if repeated random samples of size N are drawn from any population (of whatever form), then as N becomes large, the sampling distribution of sample means approaches normality.*

Statistical Significance: The Basic Notion, IV

- So making the assumption of a normal sampling distribution, one seeks results sufficiently different from the mean of a normal sampling distribution that those results would have occurred by chance fewer than one time in twenty trials ($p < .05$).
- Technically, one is referring to *the odds of making an error in rejecting a null hypothesis of no difference* between this result and the mean of a very large number of results, an error known as **Type I error**. So the probability sought is that a difference from the sampling mean of this magnitude would occur by chance fewer than one time in twenty.

Statistical Significance: The Basic Notion, V

- In survey research, we seek LARGE random samples, allowing the Law of Large Numbers to operate. Typically, the Law of Large Numbers kicks in at Ns of 100 or more, although for some purposes even at a lower threshold.
- In the SESRI Omnibus surveys, we typically have subsamples (Qataris, Ex-Patriot White Collar Workers, Blue Collar Guest Workers) of 600 or more. The Law of Large Numbers helps with *random samples* of this size by assuming that any given sample is one of many that could be taken, and that, if samples are of this size, the results that occur will resemble a normal distribution.

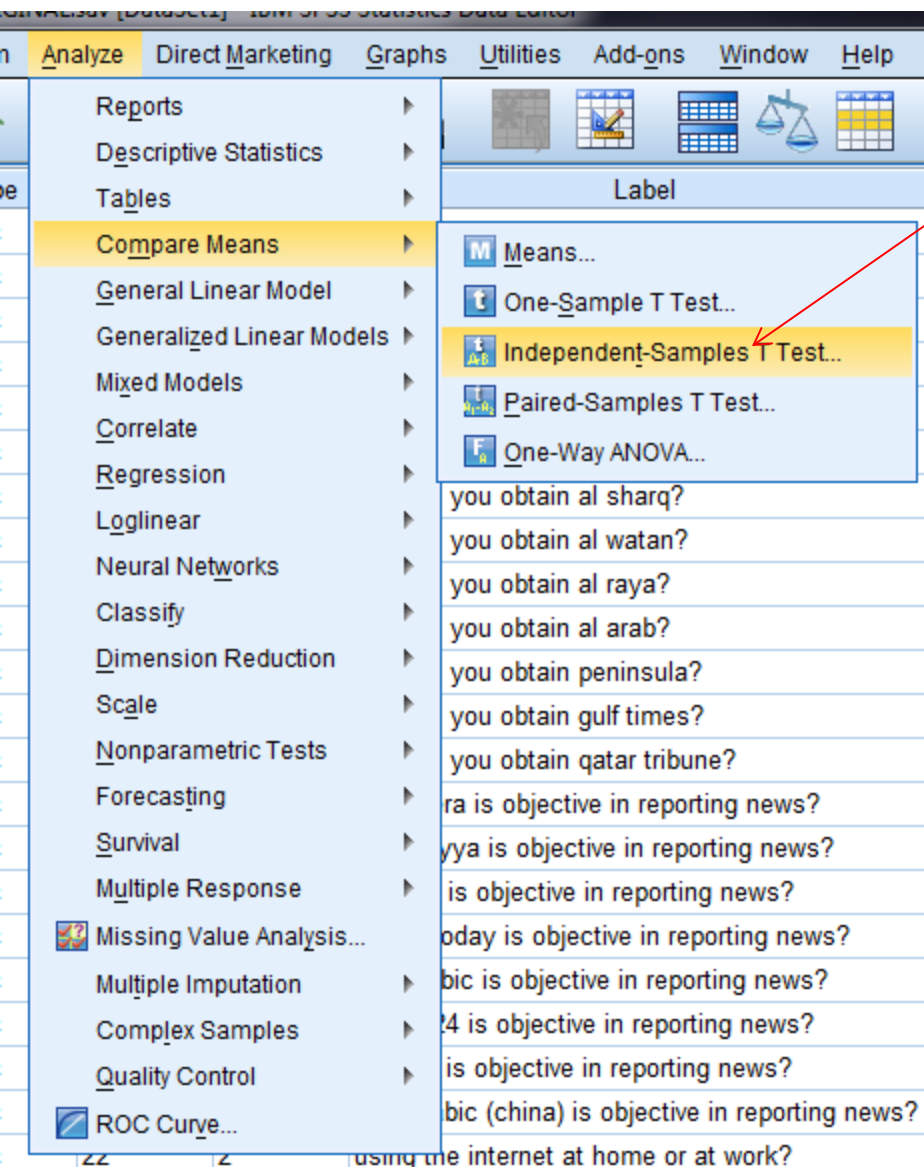
Statistical Significance: The Basic Notion, VI

- One-tailed tests of statistical significance can be conducted. They refer to situations where one has a clear hypothesis regarding the direction in which one's finding should differ from that to be found *in many random samples*.
- But, actually, it is “easier” to attain statistical significance in a one-tailed test, so most statistical programs, including SPSS, default to the two-tailed test. This is the statistically “more conservative” procedure, i.e. a more demanding standard to meet.

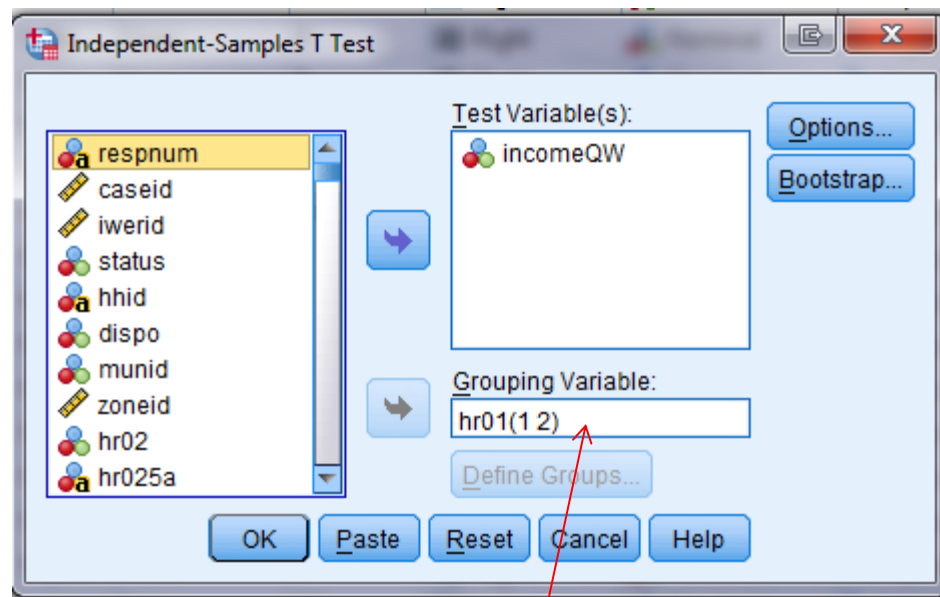
Hypothesis Testing with Descriptive Statistics

- In T-Tests, we compare means from two or more independent samples, as is the case with the samples of Qataris and White Collar Ex-Pats, testing the notion that a difference of two means is sufficiently large that it would have occurred by chance alone fewer than 5 times in 100 *if one took a large number of independent samples and compared their means.*
- Assume two samples defined by nominal criteria, such as samples of Qataris and White Collar Ex-Patriots, and that you want to compare their average (or mean) income (**incomeQW**).

Performing an Independent Samples T-Test



SPSS treats Difference of Means tests under the Compare Means tab. Here we compare Independent Samples.



Remember, **hr01** is our variable indicating the household type of the respondent. 1=Qatari Citizens, and 2=White Collar Ex-Patriots. Specifying **hr01** as the grouping variable and selecting these values tells SPSS to compare these two groups.

Monthly household income for Qatari Citizens & White Collar Workers ^ household type
Crosstabulation

One could use cross-tabulations, which we discuss in the next session, to get an initial feel for the data.

Here are the raw data that you would see from a cross tabulation. For example, among Qataris there were only 94 respondents who report a family income of under 10,000 Qr, while among white collar expatriots there were 294 such family incomes in 2010.

Count		household type		Total
		1. qatari	2. white collar	
Monthly household income for Qatari Citizens & White Collar Workers	Less than QR10,000	94	294	388
	QR10,000 to less than QR20,000	184	279	463
	QR20,000 to less than QR30,000	127	79	206
	QR30,000 to less than QR40,000	55	39	94
	QR40,000 to less than QR50,000	45	11	56
	QR50,000 to less than QR60,000	27	11	38
	QR60,000 to less than QR70,000	27	3	30
	QR70,000 to less than QR80,000	17	3	20
	QR80,000 to less than QR90,000	5	1	6
	QR90,000 to less than QR100,000	6	0	6
	QR100,000 to less than QR110,000	3	0	3
	QR110,000 to less than QR120,000	2	0	2
	QR130,000 to less than QR140,000	0	1	1
	QR140,000 to less than QR150,000	2	0	2
	QR150,000 to less than QR160,000	1	0	1
	QR160,000 to less than QR170,000	1	0	1
	QR170,000 to less than QR180,000	0	1	1
	QR180,000 to less than QR190,000	0	1	1
	QR200,000 or more	2	1	3
	Total	598	724	1322

T-Test: Output from SPSS

Group Statistics

	household type	N	Mean	Std. Deviation	Std. Error Mean
Qatari citizens & non-Qatara white collar workers income	1. qatari	596	3.5747	2.68417	.10991
	2. white collar	727	2.0969	1.66130	.06159

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Qatari citizens & non-Qatara white collar workers income	Equal variances assumed	110.146	.000	12.260	1322	.000	1.47778	.12054	1.24131	1.71425
	Equal variances not assumed			11.729	951.177	.000	1.47778	.12599	1.23053	1.72504

But these results from SPSS show that the T value is 12.260, with degrees of freedom equal to the total number of cases minus one, and that with these means and this number of cases, the probability of attaining means that differ by this amount by chance alone is less than $p=.001$, presented in the table as a two-tailed probability of $p = .000$.

Add Slides from Nizam Khan

- Probably 10-12 slides would go here.

Two Sample T-Test Exercise

Variable QOL01 in your dataset contains survey respondents' evaluations of Qatar as a place to live:

QOL01

Using a scale from 1 to 10, where 1 represents the worst possible place in which to live and 10 represents the best possible place, where on that scale would rate Qatar as a place to live?

1 2 3 4 5 6 7 8 9 10

Worst

Best

98 DON'T KNOW/UNABLE TO RATE

99 REFUSED

We would like to know whether the average rating of Qatar as a place to live is significantly different for Qatari citizens, White collar workers, and Blue collar workers.

Two Sample T-Test Exercise, continued

First, run an independent samples t-test comparing Qatari citizens (hr01=1) to white collar workers (hr01=2). The independent samples T-test option is under the *Analyze/Compare Means/Independent Samples T-test* menu. Your grouping variable should be hr01 (1 2).

What is the mean rating among Qatari citizens?

What is the mean rating among White collar workers?

Are the means significantly different from one another? How do you know whether they are different?

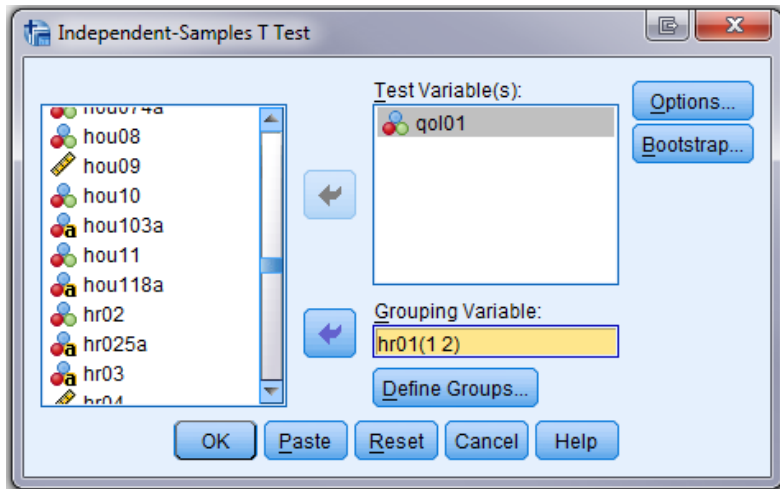
Next, run an independent samples T-test comparing Qatari citizens (hr01=1) to blue collar workers (hr01=3). Your grouping variable should be hr01 (1 3).

What is the mean rating among blue collar workers?

Are the means of these two groups significantly different from one another?
How do you know whether they are different?

Finally, run an independent samples T-test comparing white collar workers (hr01=2) to blue collar workers (hr01=3). Your grouping variable should be hr01 (2 3).

Comparing Evaluations of Life in Qatar Between Qatari Citizens and White Collar Workers



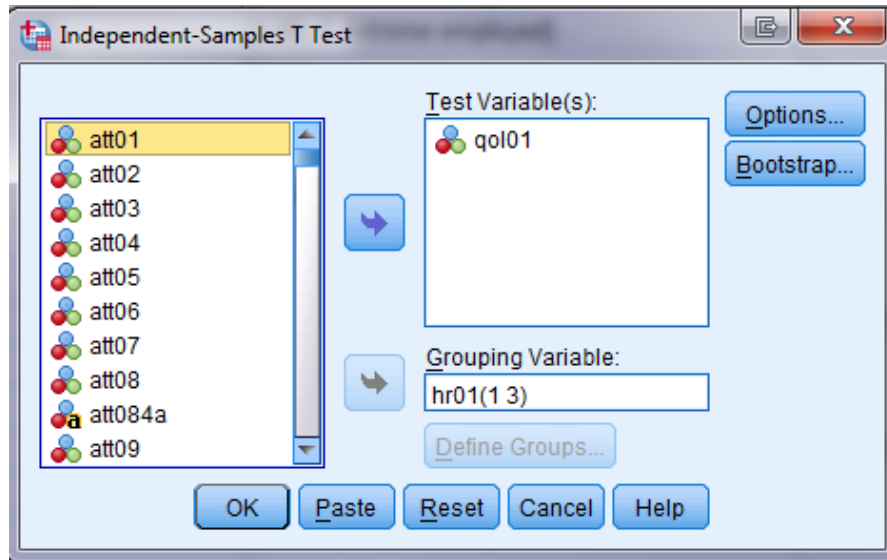
Group Statistics

	household type	N	Mean	Std. Deviation	Std. Error Mean
rating qatar as a place to live	1. qatari	661	8.6766	1.72910	.06726
	2. white collar	743	7.9029	1.83102	.06717

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
rating qatar as a place to live	Equal variances assumed	.343	.558	8.112	1402	.000	.77368	.09538	.58658	.96078
	Equal variances not assumed			8.139	1396.852	.000	.77368	.09506	.58720	.96015

Comparing Evaluations of Life in Qatar Between Qatari Citizens and Blue Collar Workers



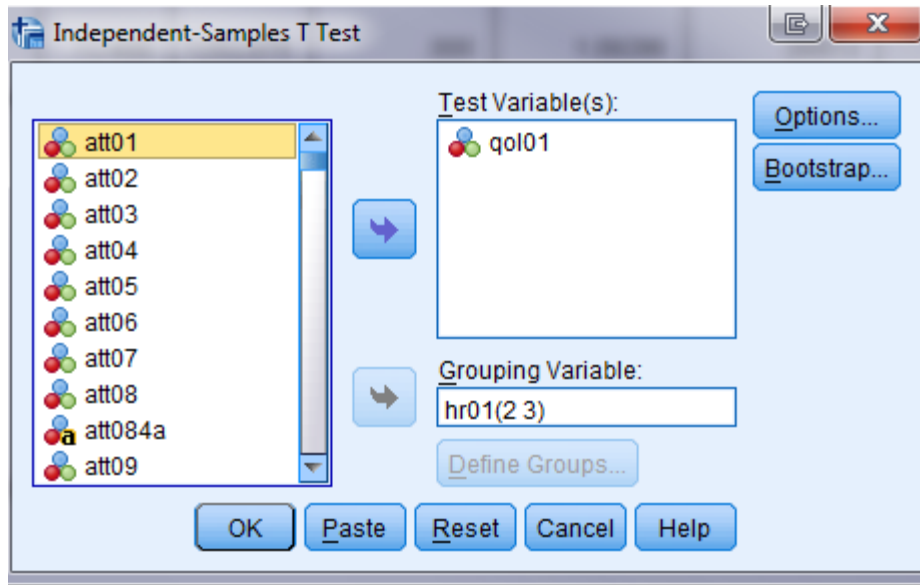
Group Statistics

	household type	N	Mean	Std. Deviation	Std. Error Mean
rating qatar as a place to live	1. qatari	661	8.6766	1.72910	.06726
	3. blue collar	614	7.5838	1.41088	.05695

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
rating qatar as a place to live	Equal variances assumed	24.207	.000	12.308	1273	.000	1.09286	.08879	.91867	1.26705
	Equal variances not assumed			12.400	1252.076	.000	1.09286	.08813	.91996	1.26577

Comparing Evaluations of Life in Qatar Between White Collar Workers and Blue Collar Workers



Group Statistics

	household type	N	Mean	Std. Deviation	Std. Error Mean
rating qatar as a place to live	2. white collar	743	7.9029	1.83102	.06717
	3. blue collar	614	7.5838	1.41088	.05695

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
rating qatar as a place to live	Equal variances assumed	28.148	.000	3.537	1355	.000	.31918	.09023	.14218	.49619
	Equal variances not assumed			3.625	1348.524	.000	.31918	.08806	.14643	.49194

T-Test Results

Qataris vs. White Collar Expatriates

What is the mean rating among Qatari citizens? **8.6766**

What is the mean rating among White collar workers? **7.9029**

Are the means significantly different from one another? How do you know whether they are different? **$t = 8.112, p = .000$**

Qataris vs. Blue Collar Guest Workers

What is the mean rating among blue collar workers? **7.5838**

Are the means of these two groups significantly different from one another?
How do you know whether they are different? **$t = 12.308, p = .000$**

White Collar Ex-Pats vs. Blue Collar Guest Workers

Are the two means significantly different from one another? How do you know whether they are different? **$t=3.537$, assuming equal variances, $p = .001$, but $t=3.625$, assuming unequal variances, which may be the better assumption, and $p = .000$.**

Appendix on T-Tests

There are three types of T-tests, as will be seen in SPSS.

- One sample T-test. Used when comparing a sample statistic (a mean) with a known population parameter (the mean), but without knowing the standard deviation of the population. Might be used to compare a given sample to a target population or to a normal distribution.
- Two sample T-test. Used if two samples are independently selected from differing populations - as in Qataris vs. Blue Collar Guest Workers. **What we used (or will use) in our exercise.**
- Paired sample T-test. Might be used if subjects have been measured before and after exposure to a stimulus, or if research subjects have been matched on one or more attributes.
- Analysis of Variance (ANOVA) a more sophisticated procedure that allows us to compare three or more means at once (Qataris vs. White Collar Expatriates vs. Blue Collar Guest Workers) *instead of using a series of sequential comparisons* (Qataris vs. White Collar Ex-Patriots, then White Collar Ex-Patriots vs. Blue Collar Guest Workers, etc.) *as is necessary with the Two-Sample T-Test.*

Additional References

- Shively, W. Phillips. The Craft of Political Research, Sixth Edition, Upper Saddle River, New Jersey, 2005, especially chapters on accuracy, precision and causal thinking.
- Steagall, Jeffrey W, and Robert L Hale. MYSTAT for Windows. Cambridge, MA: Course Technology, Inc., 1994, especially chapters on descriptive statistics, one-sample statistical tests, two-sample statistical tests and analysis of variance (ANOVA).
- Weisberg, H.F., Krosnick, J, and Bowen, B. An Introduction to Survey Research, Polling and Data Analysis, Third Edition, Beverly Hills, CA, 1996, especially chapters on single variable statistics and statistical inference for means.

Why T-Test

It assesses whether a sample mean we calculated from our data *statistically* differs from a hypothesized value

It assesses whether the means of two groups are *statistically* different

Three types of T-Test

- **One sample t-test**
- **Two independent samples t-test**
- **Paired Sampled t-test**

**How to chose which test is appropriate for your
research questions?**

Example-1

es011. How many maids are currently employed in this household?

Means number of maids employed : 1.7

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
number of maid employed in hh	686	32.1%	1453	67.9%	2139	100.0%

Report

number of maid employed in hh

Mean	N	Std. Deviation
1.7040	686	1.08733

Suppose somebody told you that he believed
that actually it was actually 1.8 not 1.7.

Research question-1

Is the average number of maids employed in
Qatari households 1.8?

One sample t-test

- One sample t-test allows to test whether a sample mean significantly differs from a hypothesized value.
- Whether the average number of maids employed estimated differs significantly from 1.8

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
number of maid employed in hh	686	1.7040	1.08733	.04151

One-Sample Test

	Test Value = 1.8					
	t	Df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
number of maid employed in hh	-2.314	685	.021	-.09604	-.1775	-.0145

Example-2

es011. How many maids are currently employed in this household?

	Gender of Respondent	N	Mean	Std. Deviation	Std. Error Mean
number of maid employed in hh	Male	319	1.8405	1.18366	.06625
	Female	367	1.5852	.98230	.05127

Research question-2

Is the average numbers reported by males (1.8) and females (1.6) are the same?

Two independent samples t-test

- An independent samples t-test is used when you want to compare the means of a normally distributed interval dependent variable for two independent groups.
- Whether the average number of maids reported by males and females are significantly different?

Example-3

- Es04a1. How many SUVs are owned by this household for personal use?
- Es04a2. How many CARs are owned by this household for personal use?

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
number of car/saloon owned by hh?	683	.00	34.00	1.2926	1.98742
number of suv owned by hh?	686	.00	11.00	1.8592	1.39772
Valid N (listwise)	681				

Research question-3

Is the mean number of cars owned equal to the mean number of SUVs owned?

Paired Sampled t-test

- A paired (samples) t-test is used when you have two related observations (i.e., two observations per subject) and you want to see if the means on these two normally distributed interval variables differ from one another.