The Office of the National Coordinator for Health Information Technology

Activity

Health Care Data Analytics
Working with Data

Health IT Workforce Curriculum
Version 4.0/Spring 2016

This material (Comp 24 Unit 2) was developed by The University of Texas Health Science Center at Houston, funded by the Department of Health and Human Services, Office of the National Coordinator for Health Information Technology under Award Number 90WT0006.

Downloading Excel Analysis ToolPak

The Analysis ToolPak is a Microsoft Office Excel add-in program that is available when you install Microsoft Office or Excel.

To use the Analysis ToolPak in Excel, however, you need to load it first.

For Excel 2013, 2016 on PC:

1. Click the **File** tab, and then click **Options**.
2. Click **Add-Ins**, and then in the **Manage** box, select **Excel Add-ins**.
3. Click **Go**.
4. In the **Add-Ins available** box, select the **Analysis ToolPak** check box, and then click **OK**.
   - If **Analysis ToolPak** is not listed in the **Add-Ins available** box, click **Browse** to locate it.
   - If you get prompted that the Analysis ToolPak is not currently installed on your computer, click **Yes** to install it.
5. After you load the Analysis ToolPak, the **Data Analysis** command is available in the **Analysis** group on the **Data** tab.

For Excel 2007, 2010 on PC:

1. Click the **Microsoft Office Button**  , and then click **Excel Options**.
2. Click **Add-Ins**, and then in the **Manage** box, select **Excel Add-ins**.
3. Click **Go**.
4. In the **Add-Ins available** box, select the **Analysis ToolPak** check box, and then click **OK**.
   - **Tip** If **Analysis ToolPak** is not listed in the **Add-Ins available** box, click **Browse** to locate it.
   - If you get prompted that the Analysis ToolPak is not currently installed on your computer, click **Yes** to install it.
5. After you load the Analysis ToolPak, the **Data Analysis** command is available in the **Analysis** group on the **Data** tab.

For Excel 2016 on Mac:

1. Click the **Tools** menu, and then click **Add-Ins**.
2. In the **Add-Ins available** box, select the **Analysis ToolPak** check box, and then click **OK**.

- If **Analysis ToolPak** is not listed in the **Add-Ins available** box, click **Browse** to locate it.
- If you get prompted that the Analysis ToolPak is not currently installed on your computer, click **Yes** to install it.

3. Quit and restart Excel.

Now the **Data Analysis** command is available on the **Data** tab.

For Excel 2011 or earlier on Mac:

Analysis Toolpak is not available. You must install a third-party Data Analysis tool such as **StatPlus:mac LE free download (Links to an external site.)** to perform descriptive statistics and Chi Square tests. **StatPlus:mac (Links to an external site.)** runs alongside Excel and offers extra menu options, which run statistical tests on data in an open Excel sheet. If you cannot download StatPlus, cannot access Excel 2016, or cannot access a PC with a previous version of Excel, you can still download and follow instructions below to create a pivot table (Activity Part 2) on Excel 2011 for Macs without downloading additional plugins or software.

Activity 1: Descriptive Statistics

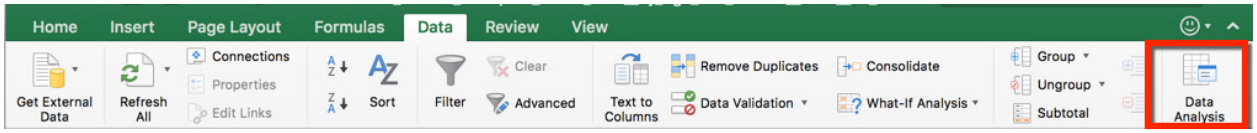
Descriptive statistics are an essential step in understanding your data.

In this exercise, we will look at basic statistics to answer the question, "In 2012, did males or females have a higher death rate due to motor vehicle crashes?"

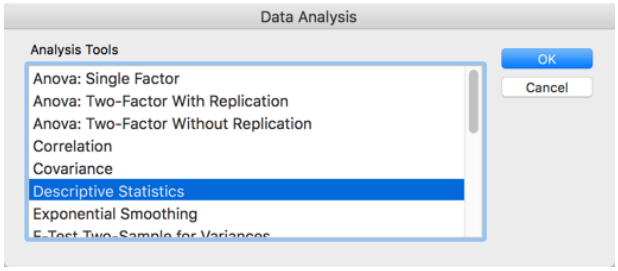
1. Open the file **comp24_unit2_dataset_motor_vehicle_occupant_death_rate_by_age_and_gender.xlsx.** This dataset is derived from the Centers for Disease Control and Prevention at the address given in your handout (https://data.cdc.gov/Motor-Vehicle/Motor-Vehicle-Occupant-Death-Rate-by-Age-and-Gende/rgg5-mkef) and gives the rate of deaths by age/gender (per 100,000 population) for motor vehicle occupants killed in crashes in 2012.
2. Take a few moments to look at the data. You'll see the state names along the left side, then age ranges in the column headings across the top, as well as columns for males and females
3. Notice that not all states have complete data. For example, Alaska only has data listed under "all ages" and nothing for females

Run Descriptive Statistics for the Males

1. On the **Data** tab, click **Data Analysis**

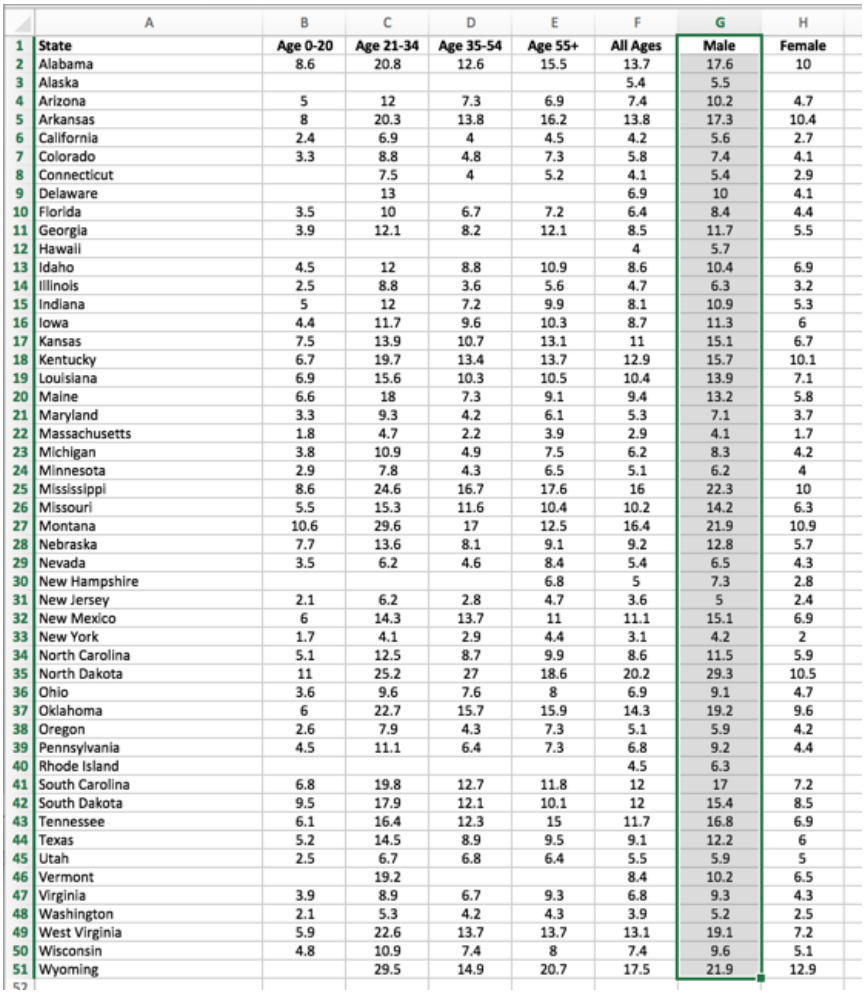


2. The **Data Analysis** tools dialog box will display. Click **Descriptive Statistics** and then click **OK**.



3. The **Descriptive Statistics** dialog box will display.
   - **Input Range**: click the drop-down at the right of the Input Range box. This will collapse the Descriptive Statistics dialog box so that you can see your data. Click the heading for the Males column and drag down until all the entries for Males, through the state of Wyoming, are highlighted. Your

screen should now look like this, and you should now see the entry $G$1:$G$51 in the Input Range box.



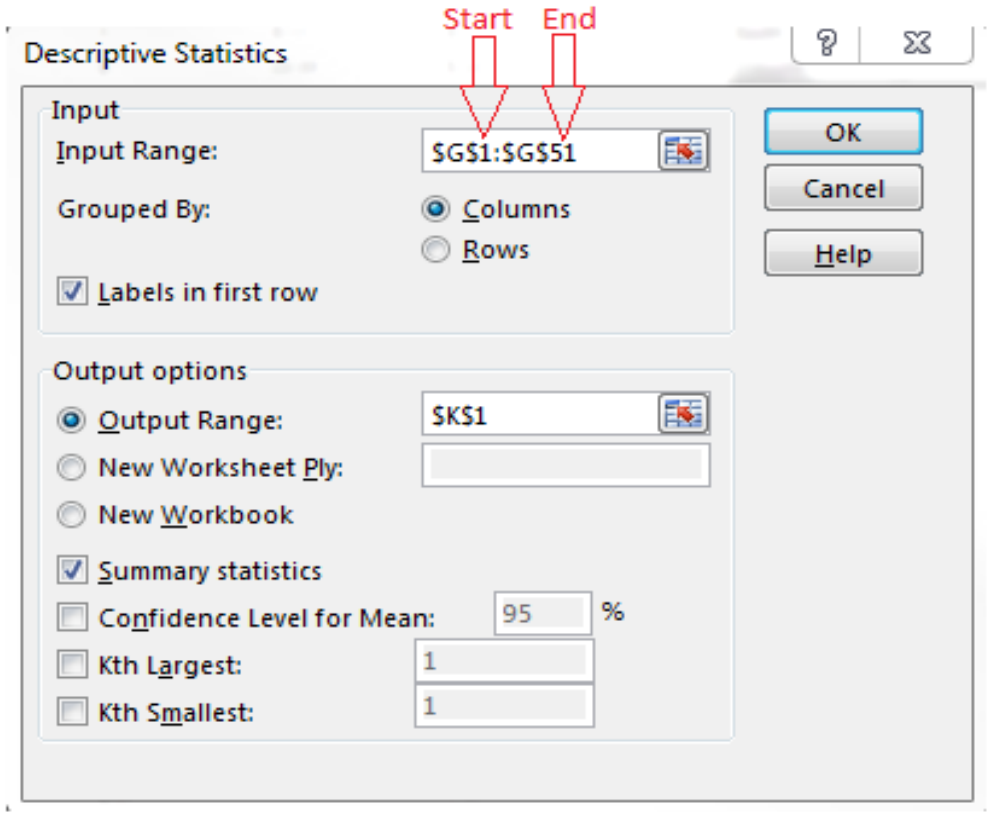
	A	B	C	D	E	F	G	H
1	State	Age 0-20	Age 21-34	Age 35-54	Age 55+	All Ages	Male	Female
2	Alabama	8.6	20.8	12.6	15.5	13.7	17.6	10
3	Alaska					5.4	5.5	
4	Arizona	5	12	7.3	6.9	7.4	10.2	4.7
5	Arkansas	8	20.3	13.8	16.2	13.8	17.3	10.4
6	California	2.4	6.9	4	4.5	4.2	5.6	2.7
7	Colorado	3.3	8.8	4.8	7.3	5.8	7.4	4.1
8	Connecticut		7.5	4	5.2	4.1	5.4	2.9
9	Delaware		13			6.9	10	4.1
10	Florida	3.5	10	6.7	7.2	6.4	8.4	4.4
11	Georgia	3.9	12.1	8.2	12.1	8.5	11.7	5.5
12	Hawaii					4	5.7	
13	Idaho	4.5	12	8.8	10.9	8.6	10.4	6.9
14	Illinois	2.5	8.8	3.6	5.6	4.7	6.3	3.2
15	Indiana	5	12	7.2	9.9	8.1	10.9	5.3
16	Iowa	4.4	11.7	9.6	10.3	8.7	11.3	6
17	Kansas	7.5	13.9	10.7	13.1	11	15.1	6.7
18	Kentucky	6.7	19.7	13.4	13.7	12.9	15.7	10.1
19	Louisiana	6.9	15.6	10.3	10.5	10.4	13.9	7.1
20	Maine	6.6	18	7.3	9.1	9.4	13.2	5.8
21	Maryland	3.3	9.3	4.2	6.1	5.3	7.1	3.7
22	Massachusetts	1.8	4.7	2.2	3.9	2.9	4.1	1.7
23	Michigan	3.8	10.9	4.9	7.5	6.2	8.3	4.2
24	Minnesota	2.9	7.8	4.3	6.5	5.1	6.2	4
25	Mississippi	8.6	24.6	16.7	17.6	16	22.3	10
26	Missouri	5.5	15.3	11.6	10.4	10.2	14.2	6.3
27	Montana	10.6	29.6	17	12.5	16.4	21.9	10.9
28	Nebraska	7.7	13.6	8.1	9.1	9.2	12.8	5.7
29	Nevada	3.5	6.2	4.6	8.4	5.4	6.5	4.3
30	New Hampshire				6.8	5	7.3	2.8
31	New Jersey	2.1	6.2	2.8	4.7	3.6	5	2.4
32	New Mexico	6	14.3	13.7	11	11.1	15.1	6.9
33	New York	1.7	4.1	2.9	4.4	3.1	4.2	2
34	North Carolina	5.1	12.5	8.7	9.9	8.6	11.5	5.9
35	North Dakota	11	25.2	27	18.6	20.2	29.3	10.5
36	Ohio	3.6	9.6	7.6	8	6.9	9.1	4.7
37	Oklahoma	6	22.7	15.7	15.9	14.3	19.2	9.6
38	Oregon	2.6	7.9	4.3	7.3	5.1	5.9	4.2
39	Pennsylvania	4.5	11.1	6.4	7.3	6.8	9.2	4.4
40	Rhode Island					4.5	6.3	
41	South Carolina	6.8	19.8	12.7	11.8	12	17	7.2
42	South Dakota	9.5	17.9	12.1	10.1	12	15.4	8.5
43	Tennessee	6.1	16.4	12.3	15	11.7	16.8	6.9
44	Texas	5.2	14.5	8.9	9.5	9.1	12.2	6
45	Utah	2.5	6.7	6.8	6.4	5.5	5.9	5
46	Vermont		19.2			8.4	10.2	6.5
47	Virginia	3.9	8.9	6.7	9.3	6.8	9.3	4.3
48	Washington	2.1	5.3	4.2	4.3	3.9	5.2	2.5
49	West Virginia	5.9	22.6	13.7	13.7	13.1	19.1	7.2
50	Wisconsin	4.8	10.9	7.4	8	7.4	9.6	5.1
51	Wyoming		29.5	14.9	20.7	17.5	21.9	12.9
52								

Click the drop-down arrow at the right again to redisplay the entire Descriptive Statistics dialog box.

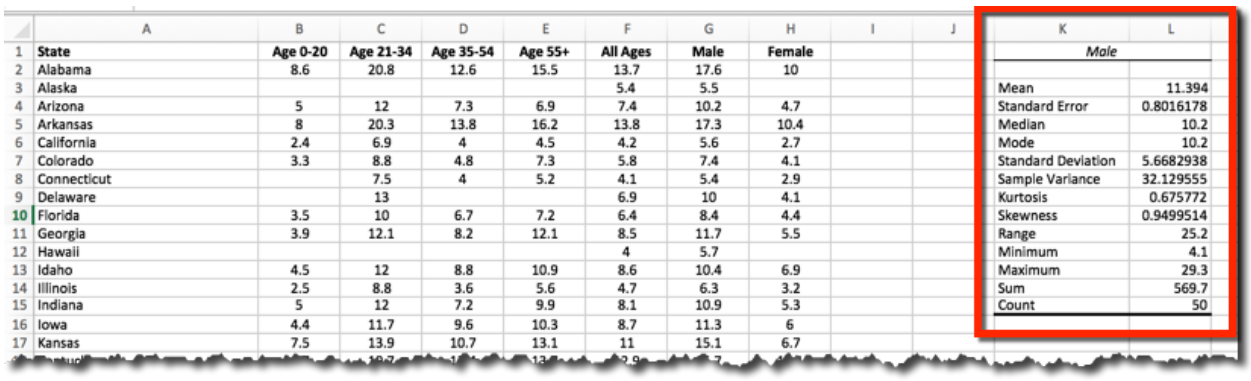
Now set these remaining options for the Descriptive Statistics:

- **Grouped By:** Click “columns”
- **Labels in first row**: This means that row 1 has titles, such as “male” or “female”. Click this option.
- **Output range**: Where do you want the statistics to be placed? You can click the icon at the right of the Output Range field and then click an empty cell on your worksheet. In this example, I clicked the first cell in column K.
- **Summary statistics**: We want summary statistics, so click this option.

4. When your screen is similar to the one below, click **OK**.



5. Your worksheet will redisplay, and will now have descriptive statistics for the males to the right of your original data. Notice that it is placed starting under column K.



	State	Age 0-20	Age 21-34	Age 35-54	Age 55+	All Ages	Male	Female
2	Alabama	8.6	20.8	12.6	15.5	13.7	17.6	10
3	Alaska					5.4	5.5	
4	Arizona	5	12	7.3	6.9	7.4	10.2	4.7
5	Arkansas	8	20.3	13.8	16.2	13.8	17.3	10.4
6	California	2.4	6.9	4	4.5	4.2	5.6	2.7
7	Colorado	3.3	8.8	4.8	7.3	5.8	7.4	4.1
8	Connecticut		7.5	4	5.2	4.1	5.4	2.9
9	Delaware		13			6.9	10	4.1
10	Florida	3.5	10	6.7	7.2	6.4	8.4	4.4
11	Georgia	3.9	12.1	8.2	12.1	8.5	11.7	5.5
12	Hawaii					4	5.7	
13	Idaho	4.5	12	8.8	10.9	8.6	10.4	6.9
14	Illinois	2.5	8.8	3.6	5.6	4.7	6.3	3.2
15	Indiana	5	12	7.2	9.9	8.1	10.9	5.3
16	Iowa	4.4	11.7	9.6	10.3	8.7	11.3	6
17	Kansas	7.5	13.9	10.7	13.1	11	15.1	6.7

	Male
Mean	11.394
Standard Error	0.8016178
Median	10.2
Mode	10.2
Standard Deviation	5.6682938
Sample Variance	32.129555
Kurtosis	0.675772
Skewness	0.9499514
Range	25.2
Minimum	4.1
Maximum	29.3
Sum	569.7
Count	50

Interpreting the Data for the Males

K	L
Male	
Mean	11.394
Standard Error	0.8016178
Median	10.2
Mode	10.2
Standard Deviation	5.6682938
Sample Variance	32.129555
Kurtosis	0.675772
Skewness	0.9499514
Range	25.2
Minimum	4.1
Maximum	29.3
Sum	569.7
Count	50

6. First, look at the entry for **Count** at the bottom of the report. Go back to the original spreadsheet and scroll through the list. Do the counts match?
7. Now look at the **Minimum** and **Maximum** values. These are the lowest and highest numbers in the Male column. Again, go back to the original spreadsheet and scroll through the list. Are these numbers correct?
8. **Range** is the spread or distance between the **Minimum** and **Maximum** values. This should equal 29.3 – 4.1.
9. The next item we want to look at is the **Mean**. This is the average rate for all the Males.
10. The **Median** is the middle point in the list of numbers, and the **Mode** is the most frequently occurring number.
11. Now repeat this exercise for the Females, using **$N$1** for the **Output Range** so that you can have the results next to your results for the males

Answer the following questions

1. What was the average death rate (number of deaths per 100,000) for males in 2012?
2. What was the average death rate (number of deaths per 100,000) for females in 2012?
3. How would you compare the rate for males against the rate for females?

Other things you can do

Run descriptive statistics on the four Age columns (Age 0-20, Age 21-34, Age 35-54, Age 55+). Which has the highest death rate? Does that make sense to you?

Activity 2: Using Filters and Creating a Pivot Table Report

Setup

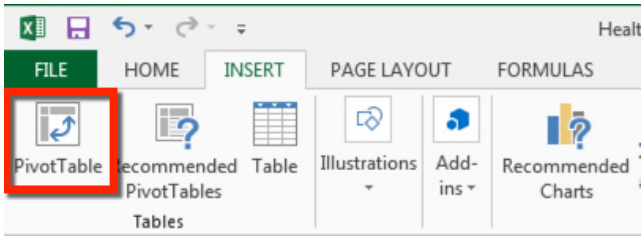
Open the file comp24_unit2_dataset_healthcare_associated_infections_state.xlsx.

[This dataset is from http://www.healthdata.gov/dataset/healthcare-associated-infections-state]

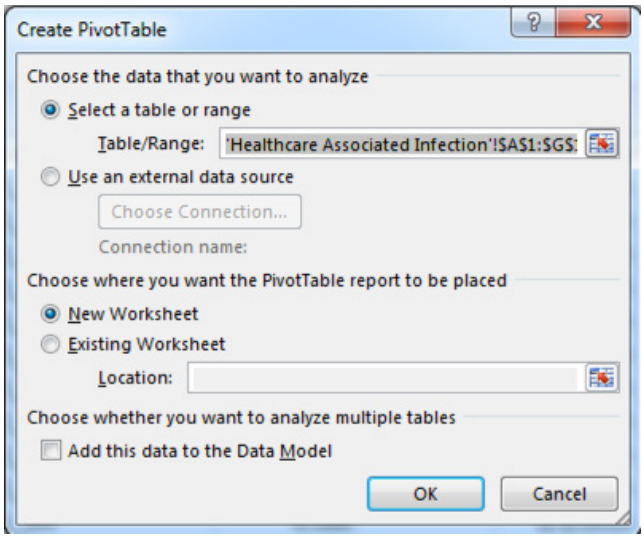
Take a moment to scroll through the file. This is a very large dataset, with over 1300 rows. We want to know about certain types of infections that are occurring in Louisiana, Oklahoma, New Mexico, and Texas. For example, some of the questions we are asking are: which state has the highest number of surgical site infections from colon surgery? Which state has the lowest number of methicillin-resistant *Staphylococcus aureus* bloodstream infections? With such a large file, it is impossible to review the data file and answer these questions without some work.

Creating a Pivot Table

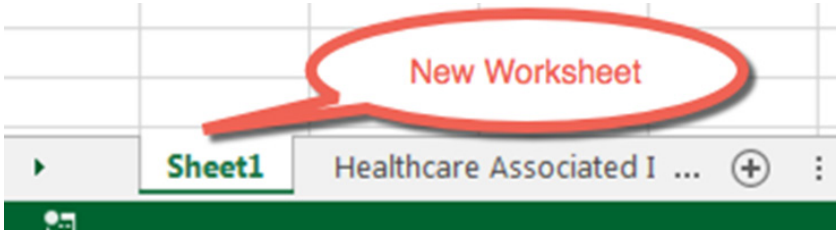
1. On the **Insert** tab, in the **Tables** group, click **PivotTable**.



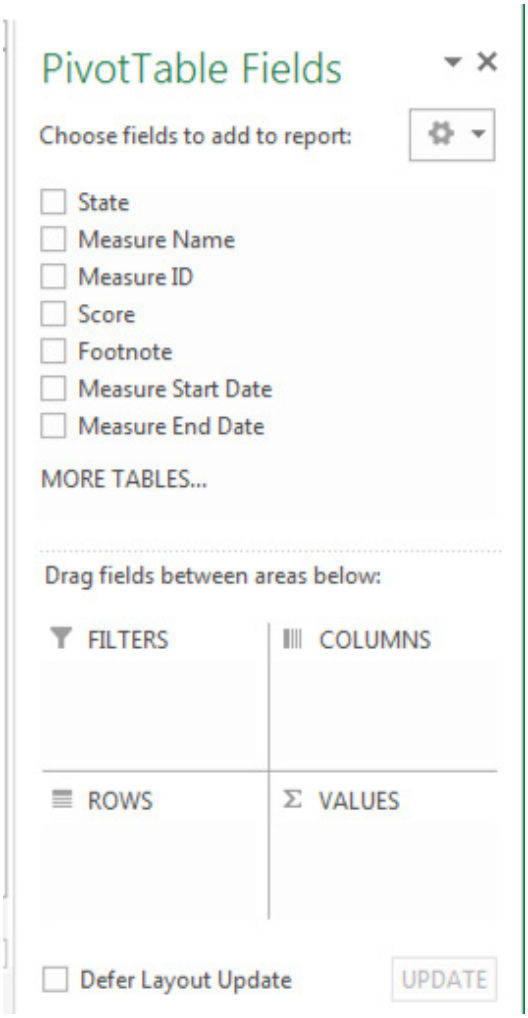
2. The Create PivotTable dialogue box will display. Make sure that **Select a table or range** is selected, and verify the range of cells in the Table/Range box. (If you can't see the entire entry in the field, drag the corner to enlarge the window.) The Pivot Table will be placed on a new worksheet in your file.



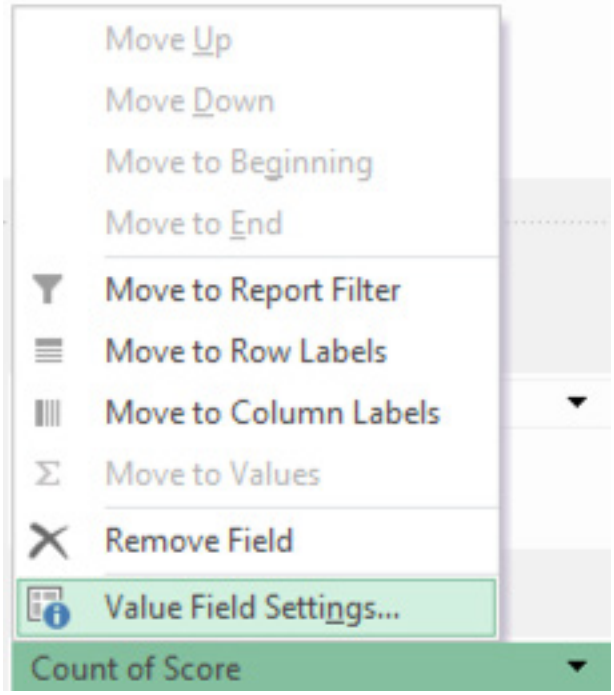
Click **OK**. Excel will create and display a new worksheet called Sheet1, and will add an empty PivotTable report to this worksheet.

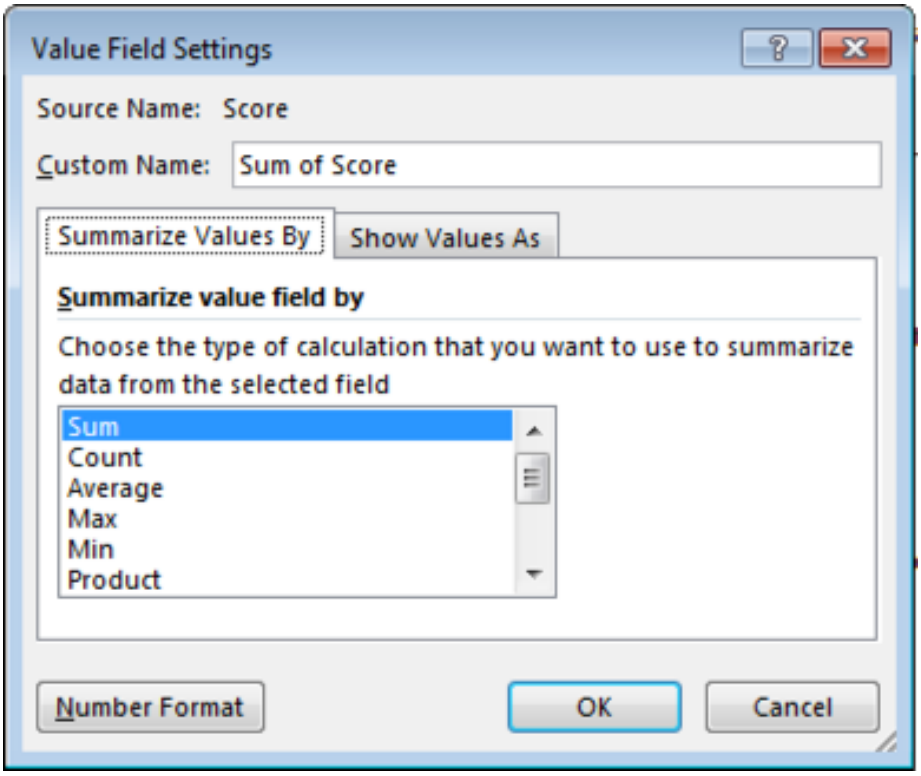


3. Excel will also display the PivotTable Field List so that you can add fields, create a layout, and customize the PivotTable report.



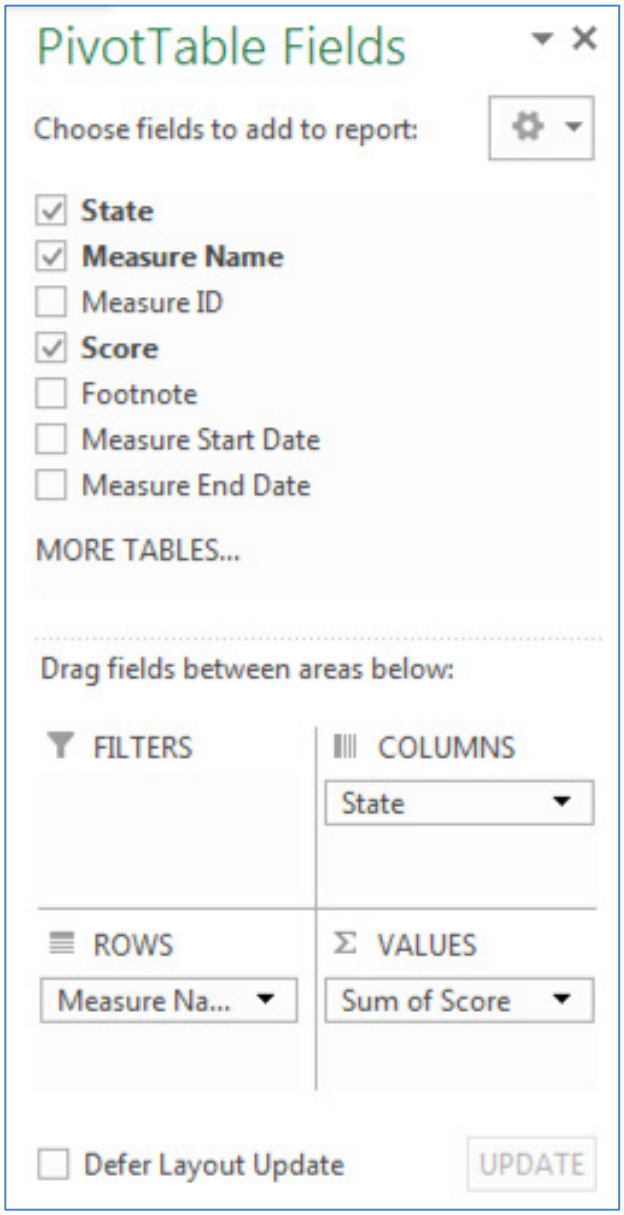
4. Click **State**. Excel will place this under **ROWS.** Notice that the states appear along the left side of the screen. However, we want the states to be column headings, so drag State to the **COLUMNS** area.
5. Click **Measure Name.** Excel will place this under **ROWS**.
6. Click **Score**. Excel will place this under **ROWS**. However, this is the actual value we are interested in, so drag Score to the **VALUES** area. Notice that Excel changes this to Count of Score, but we want totals for each state. To do this, click **Count of Score** under **VALUES** area, choose **Value Field Settings**, and change the setting from **Count of Score** to **Sum of Score**, and then click OK.



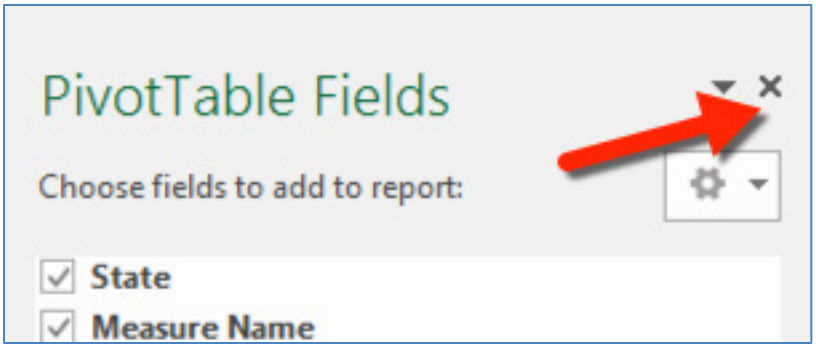


7. If extra fields appear in the Pivot Table fields, they may affect the final display. To remove a field from a section, click the drop-down arrow next to the field's name, and select **Remove Field**.

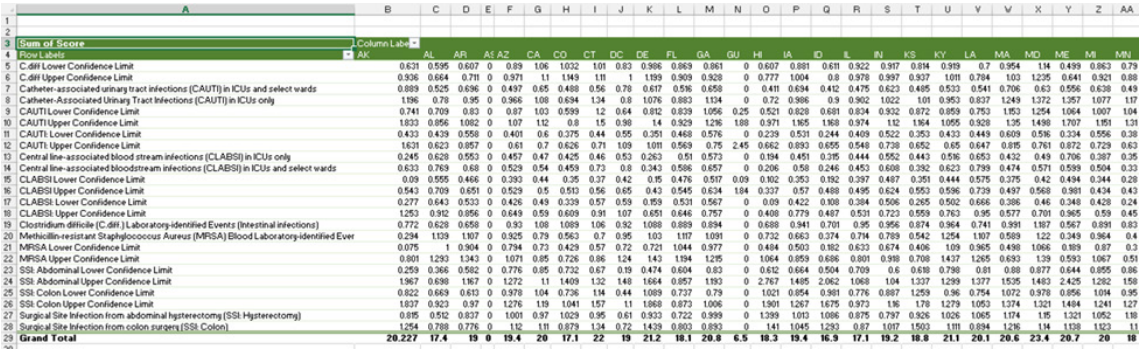
8. Your final Pivot Table should be similar to this one:



9. Now we have a full Pivot table. To see more of it, close the Pivot Table Field List by clicking the X in the top right corner of the panel.

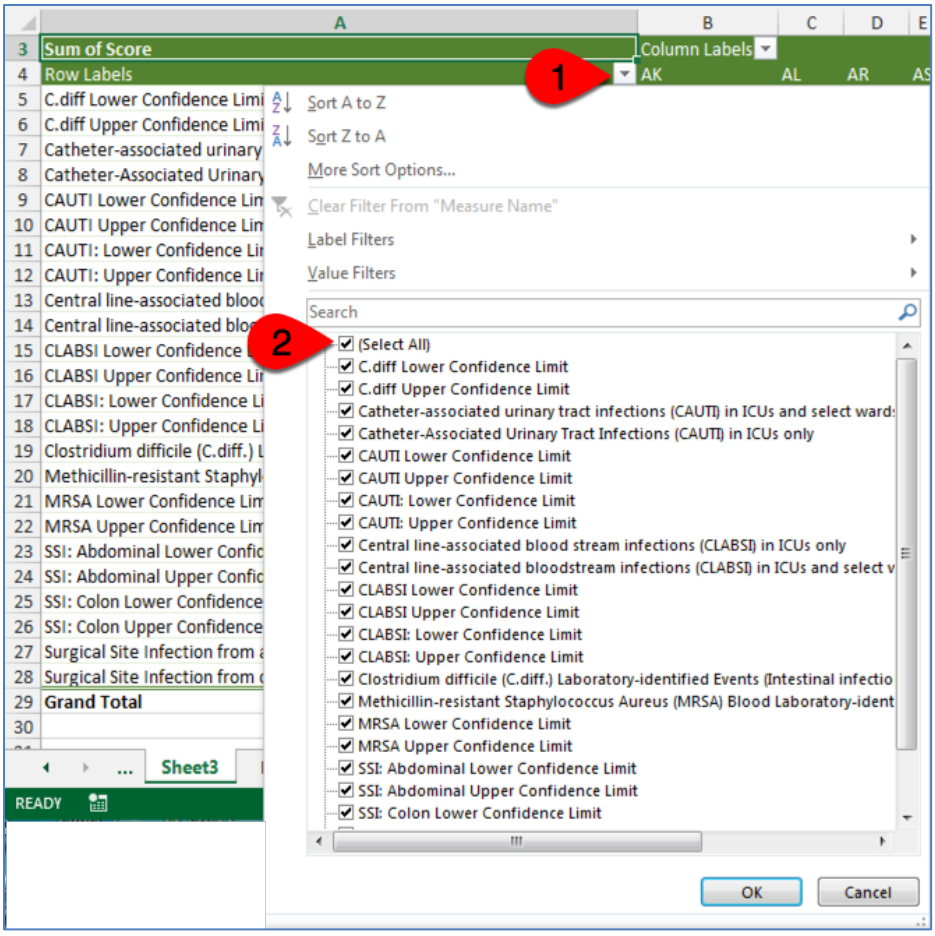


10. But notice that we have data for every state and we are only interested in four states. We also have lots of entries for confidence limits that we're not interested in.

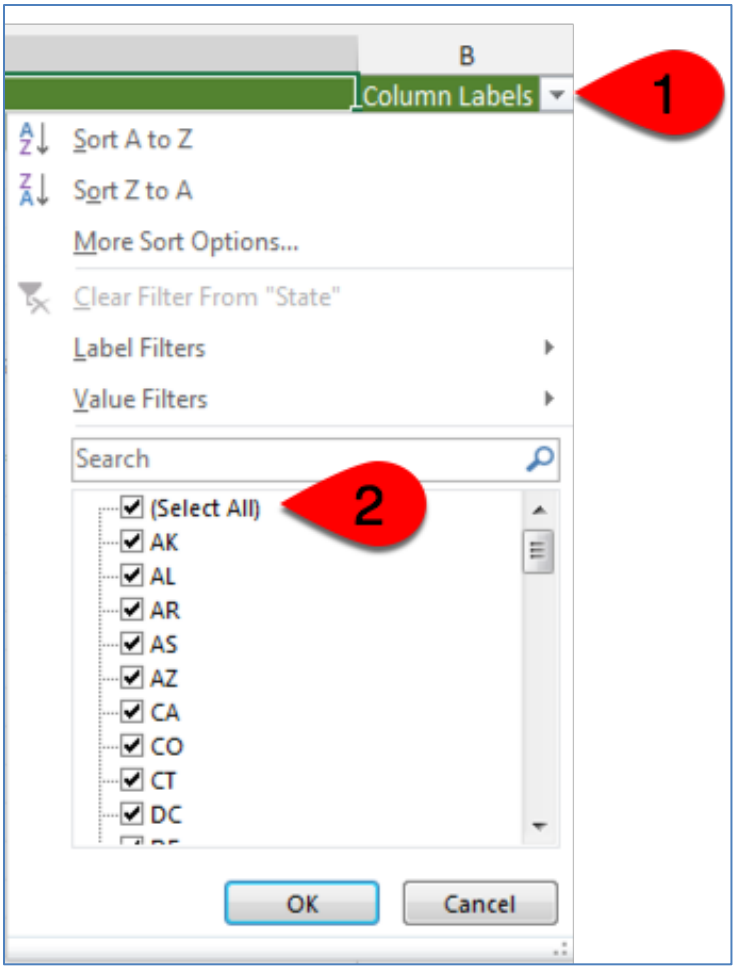


Sum of Score	Column Labels																														
Row Labels	AK	AL	AR	AS	AZ	CA	CO	CT	DC	DE	FL	GA	GU	HI	IA	ID	IL	IN	KS	KY	LA	MA	MD	ME	MI	MN					
C.diff Lower Confidence Limit	0.631	0.595	0.607	0	0.89	1.06	1.032	1.01	0.83	0.986	0.869	0.861	0	0.607	0.881	0.611	0.922	0.917	0.814	0.919	0.7	0.954	1.14	0.499	0.863	0.79					
C.diff Upper Confidence Limit	0.936	0.664	0.711	0	0.971	1.1	1.149	1.11	1	1.199	0.909	0.928	0	0.777	1.004	0.8	0.978	0.997	0.937	1.011	0.784	1.03	1.235	0.641	0.921	0.88					
Catheter-associated urinary tract infections (CAUTI) in ICUs and select wards	0.889	0.525	0.696	0	0.497	0.65	0.488	0.56	0.78	0.617	0.516	0.658	0	0.411	0.694	0.412	0.475	0.623	0.485	0.533	0.541	0.706	0.63	0.556	0.638	0.49					
Catheter-Associated Urinary Tract Infections (CAUTI) in ICUs only	1.196	0.78	0.95	0	0.966	1.08	0.694	1.34	0.8	1.076	0.883	1.114	0	0.72	0.986	0.9	0.902	1.022	1.01	0.953	0.837	1.249	1.372	1.357	1.077	1.17					
CAUTI Lower Confidence Limit	0.741	0.709	0.83	0	0.87	1.03	0.599	1.2	0.64	0.812	0.839	1.056	0.25	0.521	0.828	0.681	0.834	0.932	0.872	0.859	0.753	1.153	1.254	1.064	1.007	1.04					
CAUTI Upper Confidence Limit	1.833	0.856	1.082	0	1.07	1.12	0.8	1.5	0.98	1.4	0.929	1.216	1.88	0.971	1.165	1.168	0.974	1.12	1.164	1.055	0.928	1.35	1.498	1.707	1.151	1.31					
CAUTI: Lower Confidence Limit	0.433	0.439	0.558	0	0.401	0.6	0.375	0.44	0.55	0.351	0.468	0.576	0	0.239	0.531	0.244	0.409	0.522	0.353	0.433	0.449	0.609	0.516	0.334	0.556	0.38					
CAUTI: Upper Confidence Limit	1.631	0.623	0.857	0	0.61	0.7	0.626	0.71	1.09	1.011	0.569	0.75	2.45	0.662	0.893	0.655	0.548	0.738	0.652	0.65	0.647	0.815	0.761	0.872	0.729	0.63					
Central line-associated blood stream infections (CLABSI) in ICUs only	0.245	0.628	0.553	0	0.457	0.47	0.425	0.46	0.53	0.263	0.51	0.573	0	0.194	0.451	0.315	0.444	0.552	0.443	0.536	0.653	0.432	0.49	0.706	0.387	0.35					
Central line-associated bloodstream infections (CLABSI) in ICUs and select wards	0.633	0.769	0.68	0	0.529	0.54	0.459	0.73	0.8	0.343	0.586	0.657	0	0.206	0.58	0.246	0.453	0.608	0.392	0.623	0.799	0.474	0.571	0.599	0.504	0.33					
CLABSI Lower Confidence Limit	0.09	0.555	0.466	0	0.393	0.44	0.35	0.37	0.42	0.15	0.476	0.517	0.09	0.102	0.353	0.192	0.397	0.487	0.351	0.444	0.575	0.375	0.42	0.494	0.344	0.28					
CLABSI Upper Confidence Limit	0.543	0.709	0.651	0	0.529	0.5	0.513	0.56	0.65	0.43	0.545	0.634	1.84	0.337	0.57	0.488	0.495	0.624	0.553	0.596	0.739	0.497	0.568	0.981	0.434	0.43					
CLABSI: Lower Confidence Limit	0.277	0.643	0.533	0	0.426	0.49	0.339	0.57	0.59	0.159	0.531	0.567	0	0.09	0.422	0.108	0.384	0.506	0.265	0.502	0.666	0.386	0.46	0.348	0.428	0.24					
CLABSI: Upper Confidence Limit	1.253	0.912	0.856	0	0.649	0.59	0.609	0.91	1.07	0.651	0.646	0.757	0	0.408	0.779	0.467	0.531	0.723	0.559	0.763	0.95	0.577	0.701	0.965	0.59	0.45					
Clostridium difficile (C.diff.) Laboratory-identified Events (Intestinal infections)	0.772	0.628	0.658	0	0.93	1.08	1.089	1.06	0.92	1.088	0.889	0.894	0	0.688	0.941	0.701	0.95	0.956	0.874	0.964	0.741	0.991	1.187	0.567	0.891	0.83					
Methicillin-resistant Staphylococcus Aureus (MRSA) Blood Laboratory-identified Ever	0.294	1.139	1.107	0	0.925	0.79	0.563	0.7	0.95	1.03	1.117	1.091	0	0.732	0.663	0.374	0.714	0.789	0.542	1.254	1.107	0.589	1.22	0.349	0.964	0.4					
MRSA Lower Confidence Limit	0.075	1	0.904	0	0.794	0.73	0.429	0.57	0.72	0.721	1.044	0.977	0	0.484	0.503	0.192	0.633	0.674	0.406	1.09	0.965	0.498	1.066	0.189	0.87	0.3					
MRSA Upper Confidence Limit	0.801	1.293	1.343	0	1.071	0.85	0.726	0.86	1.24	1.43	1.194	1.215	0	1.064	0.859	0.686	0.801	0.918	0.708	1.437	1.265	0.693	1.39	0.593	1.067	0.51					
SSI: Abdominal Lower Confidence Limit	0.259	0.366	0.582	0	0.776	0.85	0.732	0.67	0.19	0.474	0.604	0.83	0	0.612	0.664	0.504	0.709	0.6	0.618	0.798	0.81	0.88	0.877	0.644	0.855	0.86					
SSI: Abdominal Upper Confidence Limit	1.967	0.698	1.167	0	1.272	1.1	1.409	1.32	1.48	1.664	0.857	1.193	0	2.767	1.485	2.062	1.068	1.04	1.337	1.299	1.277	1.535	1.483	2.425	1.282	1.58					
SSI: Colon Lower Confidence Limit	0.822	0.669	0.613	0	0.978	1.04	0.736	1.14	0.44	1.089	0.737	0.79	0	1.021	0.854	0.981	0.776	0.887	1.259	0.96	0.754	1.072	0.978	0.856	1.014	0.95					
SSI: Colon Upper Confidence Limit	1.837	0.923	0.97	0	1.276	1.19	1.041	1.57	1.1	1.968	0.873	1.006	0	1.901	1.267	1.675	0.973	1.16	1.78	1.279	1.053	1.374	1.321	1.484	1.241	1.27					
Surgical Site Infection from abdominal hysterectomy (SSI: Hysterectomy)	0.815	0.512	0.837	0	1.001	0.97	1.029	0.95	0.61	0.933	0.722	0.999	0	1.399	1.013	1.086	0.875	0.797	0.926	1.026	1.065	1.174	1.15	1.321	1.052	1.18					
Surgical Site Infection from colon surgery (SSI: Colon)	1.254	0.788	0.776	0	1.12	1.11	0.879	1.34	0.72	1.439	0.803	0.893	0	1.41	1.045	1.293	0.87	1.017	1.503	1.111	0.894	1.216	1.14	1.138	1.123	1.1					
Grand Total	**20.227**	**17.4**	**19**	**0**	**19.4**	**20**	**17.1**	**22**	**19**	**21.2**	**18.1**	**20.8**	**6.5**	**18.3**	**19.4**	**16.9**	**17.1**	**19.2**	**18.8**	**21.1**	**20.1**	**20.6**	**23.4**	**20.7**	**20**	**18**					

11. So now we need to *filter* this report.
    a. In the column heading for Column A, Row Labels, click the drop-down arrow at the right.
    b. In the dialog box that appears, uncheck the **Select All** box, and then check only the entries that do not contain "Confidence Limit". (If you can't see the entire label, drag the bottom right corner to enlarge the window.) You should have 8 types of infections checked.
    c. Click **OK**.



12. Now we want to only display infections reported from Louisiana, Oklahoma, New Mexico, and Texas.
    a. In the column heading for Column B, Column Labels, click the drop-down arrow at the right.
    b. In the dialog box that appears, uncheck the **Select All** box, and then check only the entries for LA, NM, OK, and TX.
    c. Click **OK.**

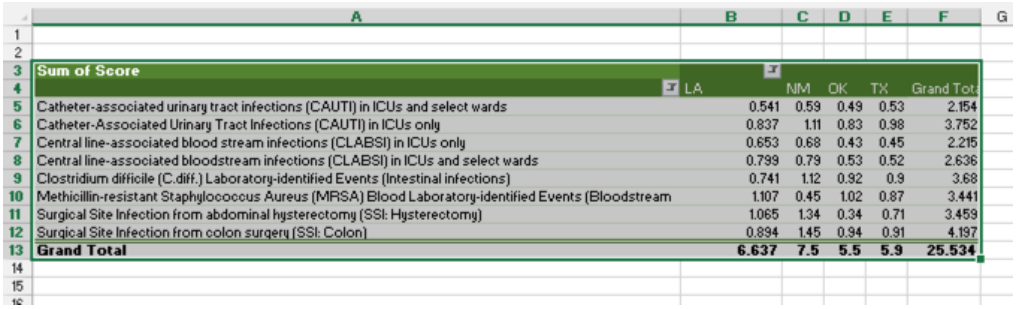


13. Your PivotTable should now look like this.

	A	B	C	D	E	F
1						
2						
3	Sum of Score	Column Labels				
4	Row Labels	LA	NM	OK	TX	Grand Total
5	Catheter-associated urinary tract infections (CAUTI) in ICUs and select wards	0.541	0.589	0.492	0.532	2.154
6	Catheter-Associated Urinary Tract Infections (CAUTI) in ICUs only	0.837	1.109	0.829	0.977	3.752
7	Central line-associated blood stream infections (CLABSI) in ICUs only	0.653	0.684	0.429	0.449	2.215
8	Central line-associated bloodstream infections (CLABSI) in ICUs and select wards	0.799	0.79	0.531	0.516	2.636
9	Clostridium difficile (C.diff.) Laboratory-identified Events (Intestinal infections)	0.741	1.119	0.923	0.897	3.68
10	Methicillin-resistant Staphylococcus Aureus (MRSA) Blood Laboratory-identified Events (Bloodstream infections)	1.107	0.446	1.023	0.865	3.441
11	Surgical Site Infection from abdominal hysterectomy (SSI: Hysterectomy)	1.065	1.338	0.344	0.712	3.459
12	Surgical Site Infection from colon surgery (SSI: Colon)	0.894	1.454	0.944	0.905	4.197
13	**Grand Total**	**6.637**	**7.529**	**5.515**	**5.853**	**25.534**

Graphing the results

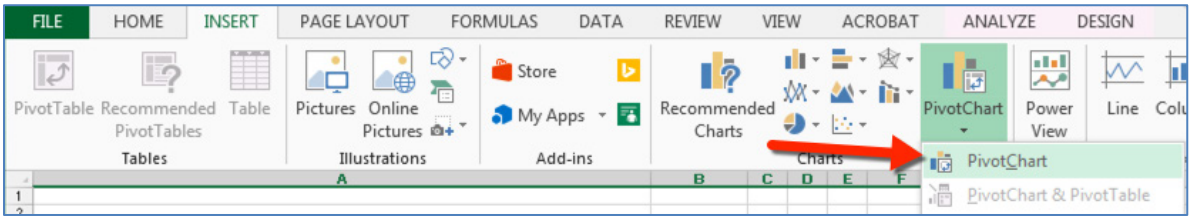
14. Now let's create a visualization – a chart – of the results to graphically present the data. Starting in the top corner of the data, click and drag until the entire table is selected.



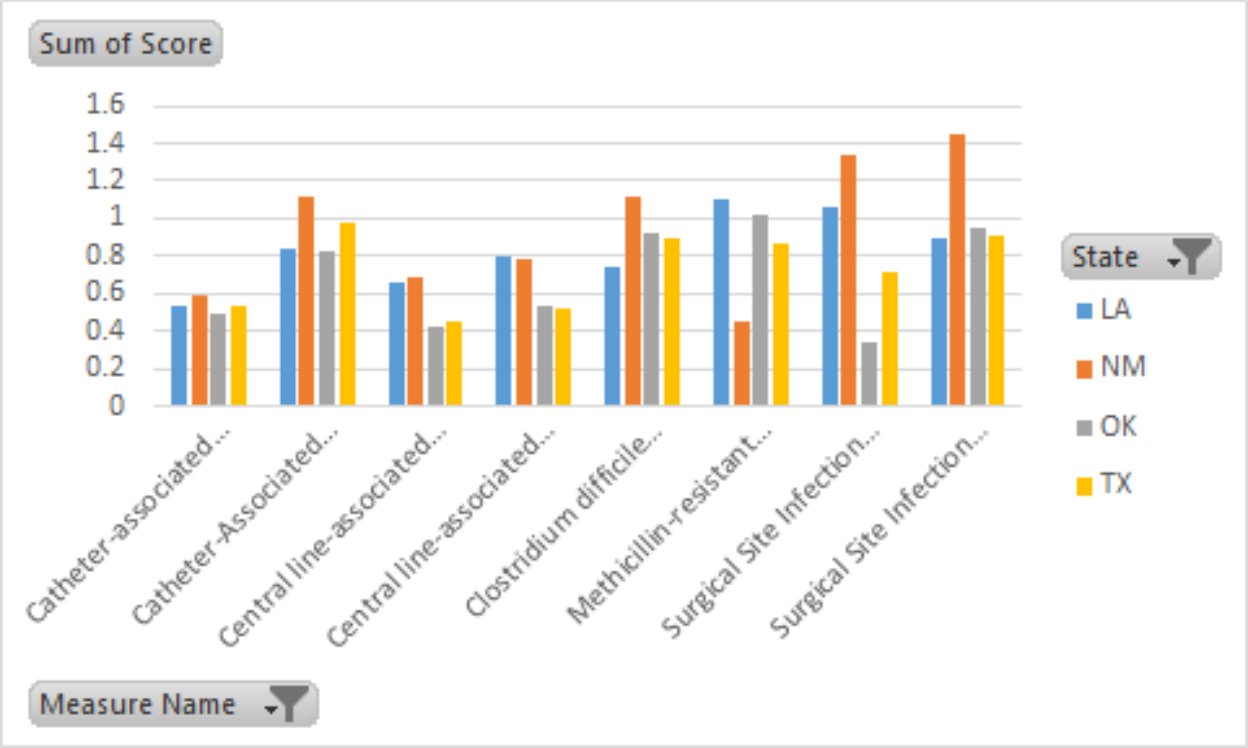
Sum of Score	LA	NM	OK	TX	Grand Total
Catheter-associated urinary tract infections (CAUTI) in ICUs and select wards	0.541	0.59	0.49	0.53	2.154
Catheter-Associated Urinary Tract Infections (CAUTI) in ICUs only	0.837	1.11	0.83	0.98	3.752
Central line-associated blood stream infections (CLABSI) in ICUs only	0.653	0.68	0.43	0.45	2.215
Central line-associated bloodstream infections (CLABSI) in ICUs and select wards	0.799	0.79	0.53	0.52	2.636
Clostridium difficile (C.diff.) Laboratory-identified Events (Intestinal infections)	0.741	1.12	0.92	0.9	3.68
Methicillin-resistant Staphylococcus Aureus (MRSA) Blood Laboratory-identified Events (Bloodstream	1.107	0.45	1.02	0.87	3.441
Surgical Site Infection from abdominal hysterectomy (SSI: Hysterectomy)	1.065	1.34	0.34	0.71	3.459
Surgical Site Infection from colon surgery (SSI: Colon)	0.894	1.45	0.94	0.91	4.197
Grand Total	**6.637**	**7.5**	**5.5**	**5.9**	**25.534**

(Here's a tip: You can click the first cell, then hold down the Shift key and then click the bottom right cell. Sometimes that is easier than clicking and dragging.)

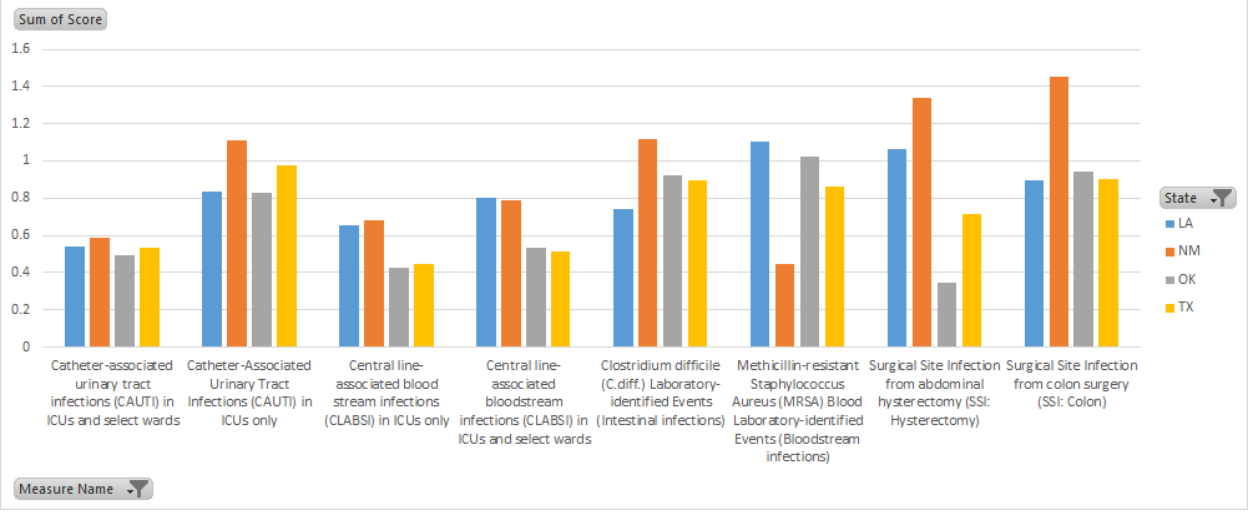
15. Click **Insert → PivotChart → PivotChart** (Note: this may look different on different versions of Excel)



16. Choose **Clustered Column** and click **OK**. You should now see a column chart similar to this one:



To see the complete label for each set of columns, grab each side of the chart and drag to enlarge the chart. You can also drag the chart vertically to adjust the height of the columns. The chart should now look like this:



Discussion Questions:

1. Which state has the highest number of surgical site infections from colon surgery?
2. Which state has the highest number of methicillin-resistant *Staphylococcus aureus* bloodstream infections?
3. Which state has the highest number of surgical site infections from abdominal hysterectomy?
4. For all the infections reported in the PivotTable, which state has the lowest number of infections?
5. For all the infections reported in the PivotTable, which state has the highest number of infections?

Other things you can do:

- You can use the PivotTable Field List to rearrange the fields later as needed by right-clicking the fields in the layout section, and then selecting the area that you want, or by dragging the fields between the areas in the layout section.
- Click the Options and Design tabs of the PivotTable Tools that become available when you click anywhere in a PivotTable, and then explore the groups and options that are provided on each tab.
- You can also access options and features that are available for specific PivotTable elements by right-clicking those elements.
- For detailed information about how to work with PivotTable reports and PivotChart reports, see Overview of PivotTable and PivotChart reports, Create or delete a PivotTable or PivotChart report, and Pivot data in a PivotTable or PivotChart report on Office.com.

Activity 3: Column Charts and Histograms

In this tutorial you will learn the difference between a column chart and a histogram and learn how to create them.

Column charts and histograms initially look very similar. However, they are quite different.

- A column chart plots each value in a data set as a vertical column
- A histogram is a graph that shows the *frequency* of values in a data set – in other words, how many times a particular value occurs -- and so histograms are very useful for showing how the data are distributed.

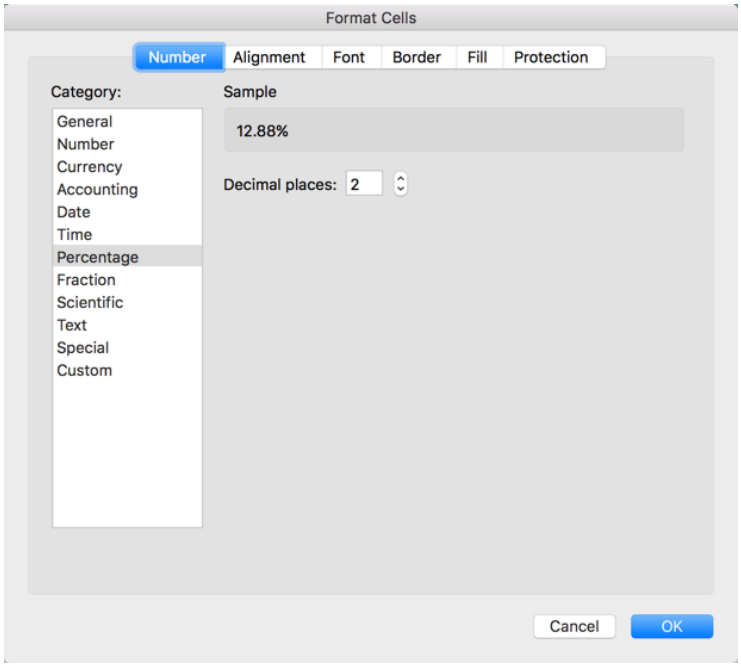
1. Open the file **comp24_unit2_dataset_prevalence_and_trends_data_tobacco_use.xlsx** [This dataset is derived from https://data.cdc.gov/Smoking-Tobacco-Use/BRFSS-Prevalence-and-Trends-Data-Tobacco-Use-Four-/8zak-ewtm]
2. Take a few moments to look at the data. You'll see the year and state names along the left side, then four categories of smoking status in the column headings across the top. Notice that the year goes from 2010 down to 1995, and that there are over 800 rows in the file.

	A	B	C	D	E	F
1	Year	State	Smoke everyday	Smoke some days	Former smoker	Never smoked
2	2010	Alabama	15.60%	6.30%	23.90%	54.20%
3	2010	Alaska	13.50%	6.80%	26.10%	53.60%
4	2010	Arizona	10.70%	4.40%	27.90%	57.10%
5	2010	Arkansas	17.30%	5.60%	24.10%	53%
6	2010	California	7.50%	4.60%	23.10%	64.80%
7	2010	Colorado	11.40%	4.60%	24.70%	59.30%
8	2010	Connecticut	9.20%	4%	29.20%	57.60%
9	2010	Delaware	12.80%	4.50%	26.80%	56%
10	2010	District of Columbia	10%	5.70%	23.40%	61%
11	2010	Florida	12%	5.20%	29.80%	53%
12	2010	Georgia	12.80%	4.80%	23.10%	59.30%
13	2010	Guam	19.70%	6.10%	16.60%	57.60%
14	2010	Hawaii	10.70%	3.80%	25.30%	60.20%
15	2010	Idaho	11.30%	4.40%	22.90%	61.50%
16	2010	Illinois	11.50%	5.40%	23.60%	59.50%
17	2010	Indiana	16.30%	5%	25.10%	53.70%
18	2010	Iowa	12.10%	4.10%	23.40%	60.40%
19	2010	Kansas	11.90%	5.10%	24.20%	58.80%
20	2010	Kentucky	19.30%	5.50%	26%	49.20%
21	2010	Louisiana	15.90%	6.20%	22%	56%

Run Descriptive Statistics

3. Refer to the instructions for the tutorial on descriptive statistics. Run **Descriptive Statistics** on the 2010 data. For this video, I have already created the descriptive statistics. You will need to format the output for each Mean

value so that they display as percentages. To do this, right click each Mean, then choose **Format Cells**, then **Percentages,** and then click **OK**.



4. Your descriptive statistics for 2010 should look like this:

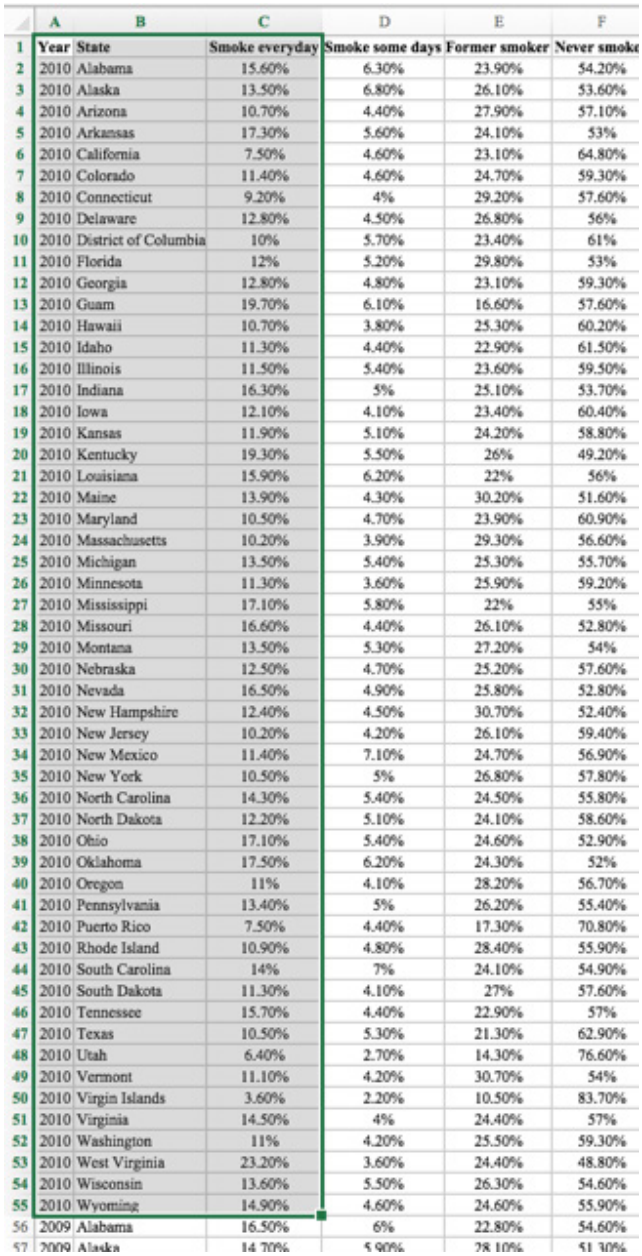
Smoke everyday		*Smoke some days*		*Former smoker*		*Never smoked*	
Mean	12.88%	Mean	4.85%	Mean	24.70%	Mean	57.57%
Standard Error	0.004679149	Standard Error	0.001304915	Standard Error	0.005010538	Standard Error	0.00801738
Median	0.123	Median	0.0475	Median	0.247	Median	0.568
Mode	0.135	Mode	0.044	Mode	0.261	Mode	0.576
Standard Deviatior	0.034384581	Standard Deviation	0.009589131	Standard Deviation	0.036819787	Standard Deviation	0.05891549
Sample Variance	0.001182299	Sample Variance	9.19514E-05	Sample Variance	0.001355697	Sample Variance	0.00347103
Kurtosis	1.184362911	Kurtosis	0.76750603	Kurtosis	4.366819264	Kurtosis	8.20978174
Skewness	0.311161025	Skewness	0.065576286	Skewness	-1.540589324	Skewness	2.40798844
Range	0.196	Range	0.049	Range	0.202	Range	0.349
Minimum	0.036	Minimum	0.022	Minimum	0.105	Minimum	0.488
Maximum	0.232	Maximum	0.071	Maximum	0.307	Maximum	0.837
Sum	6.953	Sum	2.621	Sum	13.34	Sum	31.089
Count	54	Count	54	Count	54	Count	54

5. Now run Descriptive Statistics on the 1995 data, formatting the Means as above.
6. Compare the two sets of data. Can you draw any conclusions yet?

Creating a Column Chart

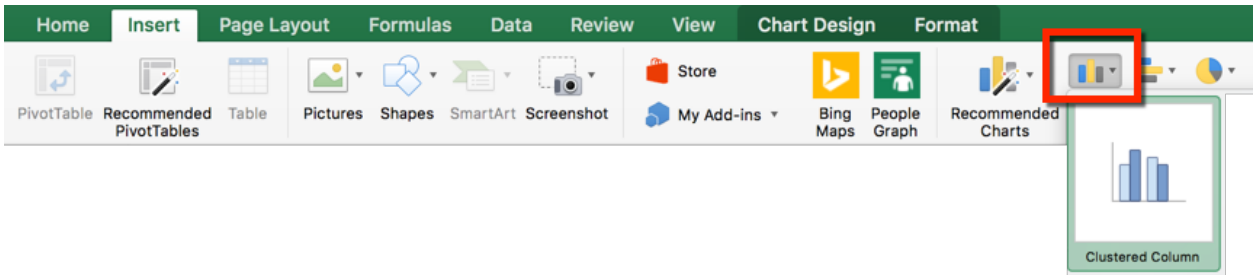
In this step, you will graph the “smoke every day” data for all states for 2010. This is so that you will be able to see the difference between a column chart and a histogram.

1. Click in the cell labeled **Year** (cell A1) and drag down and to the right until Wyoming’s 2010 value of 14.90% is highlighted. Your screen should look like this:

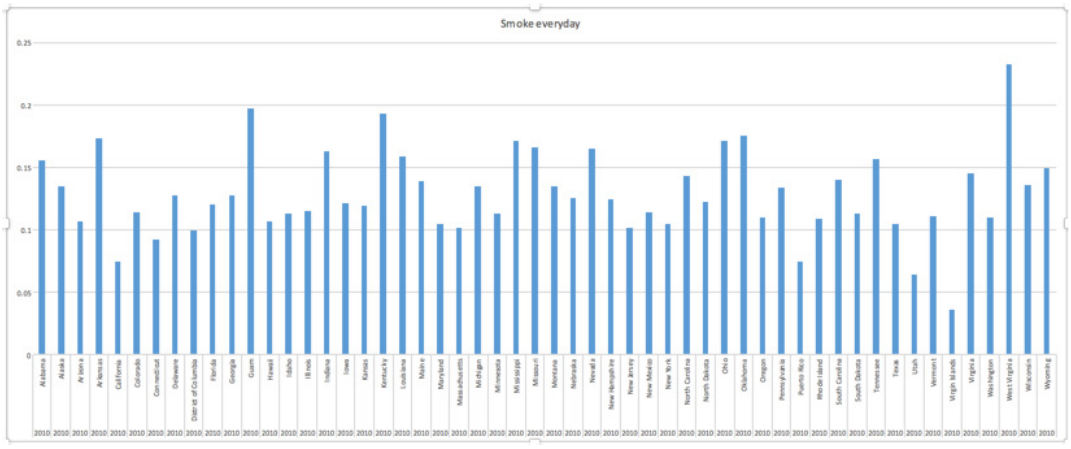


	A	B	C	D	E	F
1	Year	State	Smoke everyday	Smoke some days	Former smoker	Never smoke
2	2010	Alabama	15.60%	6.30%	23.90%	54.20%
3	2010	Alaska	13.50%	6.80%	26.10%	53.60%
4	2010	Arizona	10.70%	4.40%	27.90%	57.10%
5	2010	Arkansas	17.30%	5.60%	24.10%	53%
6	2010	California	7.50%	4.60%	23.10%	64.80%
7	2010	Colorado	11.40%	4.60%	24.70%	59.30%
8	2010	Connecticut	9.20%	4%	29.20%	57.60%
9	2010	Delaware	12.80%	4.50%	26.80%	56%
10	2010	District of Columbia	10%	5.70%	23.40%	61%
11	2010	Florida	12%	5.20%	29.80%	53%
12	2010	Georgia	12.80%	4.80%	23.10%	59.30%
13	2010	Guam	19.70%	6.10%	16.60%	57.60%
14	2010	Hawaii	10.70%	3.80%	25.30%	60.20%
15	2010	Idaho	11.30%	4.40%	22.90%	61.50%
16	2010	Illinois	11.50%	5.40%	23.60%	59.50%
17	2010	Indiana	16.30%	5%	25.10%	53.70%
18	2010	Iowa	12.10%	4.10%	23.40%	60.40%
19	2010	Kansas	11.90%	5.10%	24.20%	58.80%
20	2010	Kentucky	19.30%	5.50%	26%	49.20%
21	2010	Louisiana	15.90%	6.20%	22%	56%
22	2010	Maine	13.90%	4.30%	30.20%	51.60%
23	2010	Maryland	10.50%	4.70%	23.90%	60.90%
24	2010	Massachusetts	10.20%	3.90%	29.30%	56.60%
25	2010	Michigan	13.50%	5.40%	25.30%	55.70%
26	2010	Minnesota	11.30%	3.60%	25.90%	59.20%
27	2010	Mississippi	17.10%	5.80%	22%	55%
28	2010	Missouri	16.60%	4.40%	26.10%	52.80%
29	2010	Montana	13.50%	5.30%	27.20%	54%
30	2010	Nebraska	12.50%	4.70%	25.20%	57.60%
31	2010	Nevada	16.50%	4.90%	25.80%	52.80%
32	2010	New Hampshire	12.40%	4.50%	30.70%	52.40%
33	2010	New Jersey	10.20%	4.20%	26.10%	59.40%
34	2010	New Mexico	11.40%	7.10%	24.70%	56.90%
35	2010	New York	10.50%	5%	26.80%	57.80%
36	2010	North Carolina	14.30%	5.40%	24.50%	55.80%
37	2010	North Dakota	12.20%	5.10%	24.10%	58.60%
38	2010	Ohio	17.10%	5.40%	24.60%	52.90%
39	2010	Oklahoma	17.50%	6.20%	24.30%	52%
40	2010	Oregon	11%	4.10%	28.20%	56.70%
41	2010	Pennsylvania	13.40%	5%	26.20%	55.40%
42	2010	Puerto Rico	7.50%	4.40%	17.30%	70.80%
43	2010	Rhode Island	10.90%	4.80%	28.40%	55.90%
44	2010	South Carolina	14%	7%	24.10%	54.90%
45	2010	South Dakota	11.30%	4.10%	27%	57.60%
46	2010	Tennessee	15.70%	4.40%	22.90%	57%
47	2010	Texas	10.50%	5.30%	21.30%	62.90%
48	2010	Utah	6.40%	2.70%	14.30%	76.60%
49	2010	Vermont	11.10%	4.20%	30.70%	54%
50	2010	Virgin Islands	3.60%	2.20%	10.50%	83.70%
51	2010	Virginia	14.50%	4%	24.40%	57%
52	2010	Washington	11%	4.20%	25.50%	59.30%
53	2010	West Virginia	23.20%	3.60%	24.40%	48.80%
54	2010	Wisconsin	13.60%	5.50%	26.30%	54.60%
55	2010	Wyoming	14.90%	4.60%	24.60%	55.90%
56	2009	Alabama	16.50%	6%	22.80%	54.60%
57	2009	Alaska	14.70%	5.90%	28.10%	51.30%

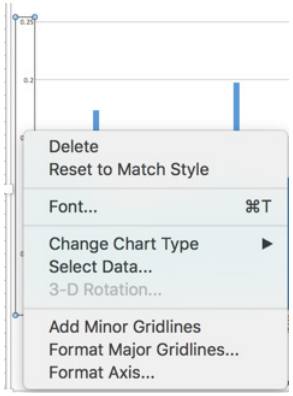
2. On the **Insert** menu, click the column chart icon and then choose **Clustered Column** (Note: this may look different on different versions of Excel)



3. Excel should now create a column chart similar to this one. Note that to improve the readability of the chart, you may need to grab the right or left side and drag to enlarge the chart.



4. If the Y axis values are not displaying as percentages, right-click one of the numbers on the Y axis and choose **Format Axis.** Under **Number**, change the category to **Percentage**. The Y axis will now show the percentages.



5. Study the column chart. What can you tell from this chart? How difficult do you think it would be to interpret this chart if it included all four smoking statuses for all the years in the spreadsheet?

Creating a Frequency Table and Histogram

A frequency table shows how many times a value occurs in a data set, and a histogram is a graph of that data. Take a look again at the data for the "Smoke everyday" category. What we want to do is to set up groupings or "bins" for the values to fall into, such as 0-5, 6-10, 11-15, and so on. So how do we decide what those categories should be?

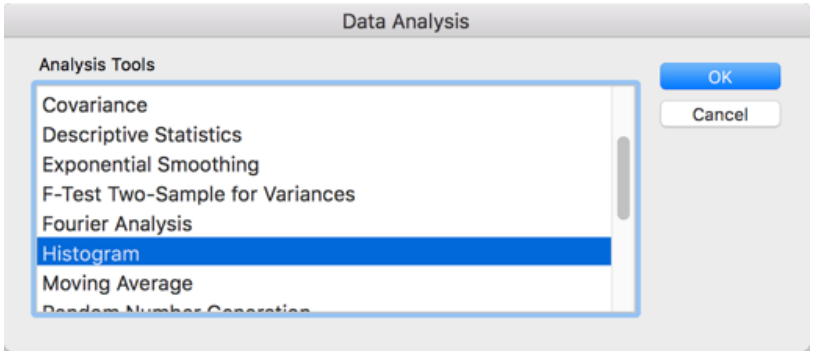
1. First, let's look again at the descriptive statistics for the "Smoke everyday" category. The Minimum value is .036 or 3.6%, and the Maximum is 0.232 or 23.2%.

Smoke everyday	
Mean	12.88%
Standard Error	0.004679149
Median	0.123
Mode	0.135
Standard Deviation	0.034384581
Sample Variance	0.001182299
Kurtosis	1.184362911
Skewness	0.311161025
Range	0.196
Minimum	0.036
Maximum	0.232
Sum	6.953
Count	54

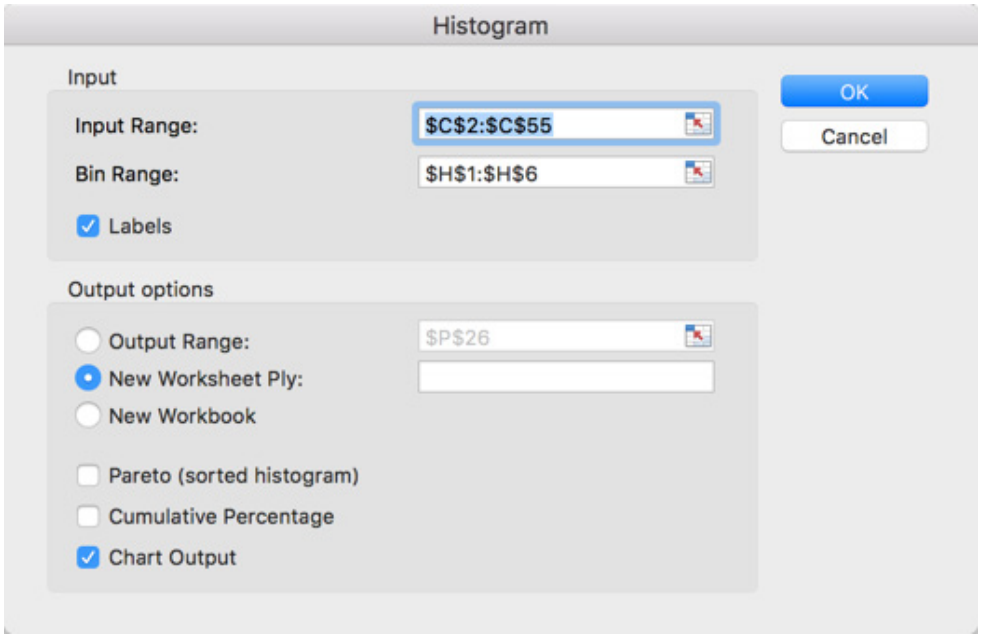
2. So, we could create our categories, or "bins", as 0-5, 6-10, 11-15, 16-20, and 21-25. To do this, click in the first row under column H and enter the following values: 0, 5, 10, 15, 20, 25. Your entries should look like this:

H
0
5.00%
10.00%
15.00%
20.00%
25.00%

3. Now you are ready to create the frequency table and the histogram. Choose **Data → Data Analysis** and then choose **Histogram**. Click **OK**.

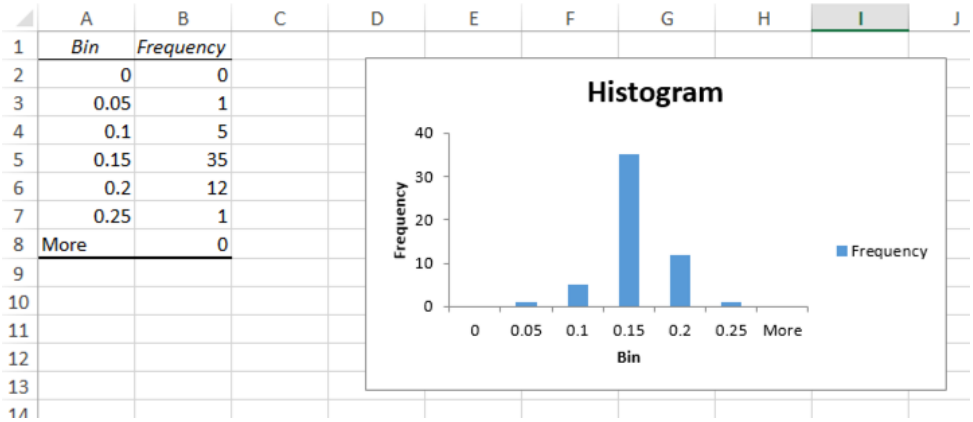


4. The **Histogram** dialog box will display.
   - For **Input Range**: enter the range of cells for 2010 Smoke Every Day (should be $C$2:$C$55)
   - For **Bin Range:** enter the cells where you put your categories (should be $H$1:$H$6) (Remember that you can click the drop-down at the right side of the field, then highlight the desired cells, and then click the drop-down again to redisplay the entire dialog box.)
   - For **Labels**: we don't need to check this option because we didn't add a column label for the Bin column, and we didn't use cell C1, which was the label for the Smoke Every Day column.
   - For **Output Options**: click **New Worksheet Ply** to have the results show up on a new worksheet. (You may or may not see a value in Output range depending on the version of Microsoft Excel).
   - Check **Chart Output** to have Excel automatically graph the results as a histogram. Your screen should look like this:



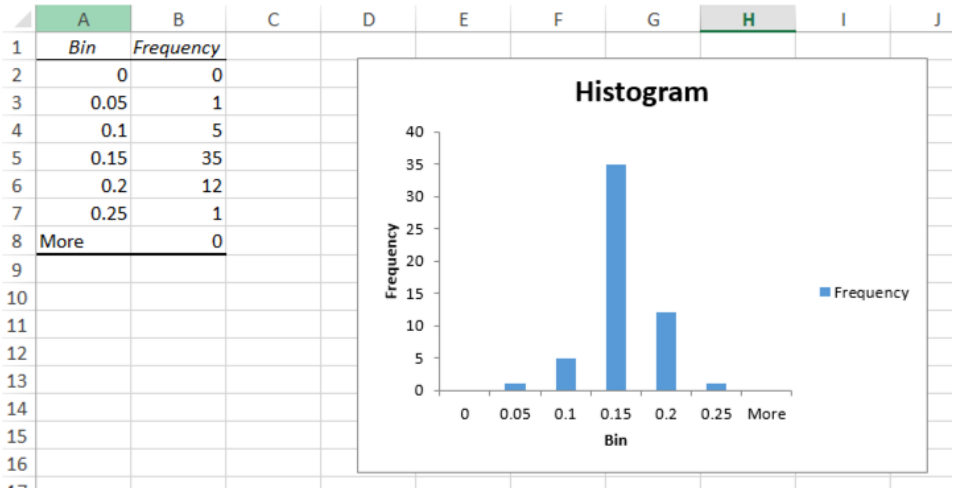
   - Click **OK**.

5. Excel will create a new worksheet called Sheet1 and will display this worksheet. You should see the following output (recall that you can reformat the axes to display as percentages)



	A	B
1	*Bin*	*Frequency*
2	0	0
3	0.05	1
4	0.1	5
5	0.15	35
6	0.2	12
7	0.25	1
8	More	0

To see more levels in the Frequency axis, such as 25 and 35, drag the histogram down to expand it.



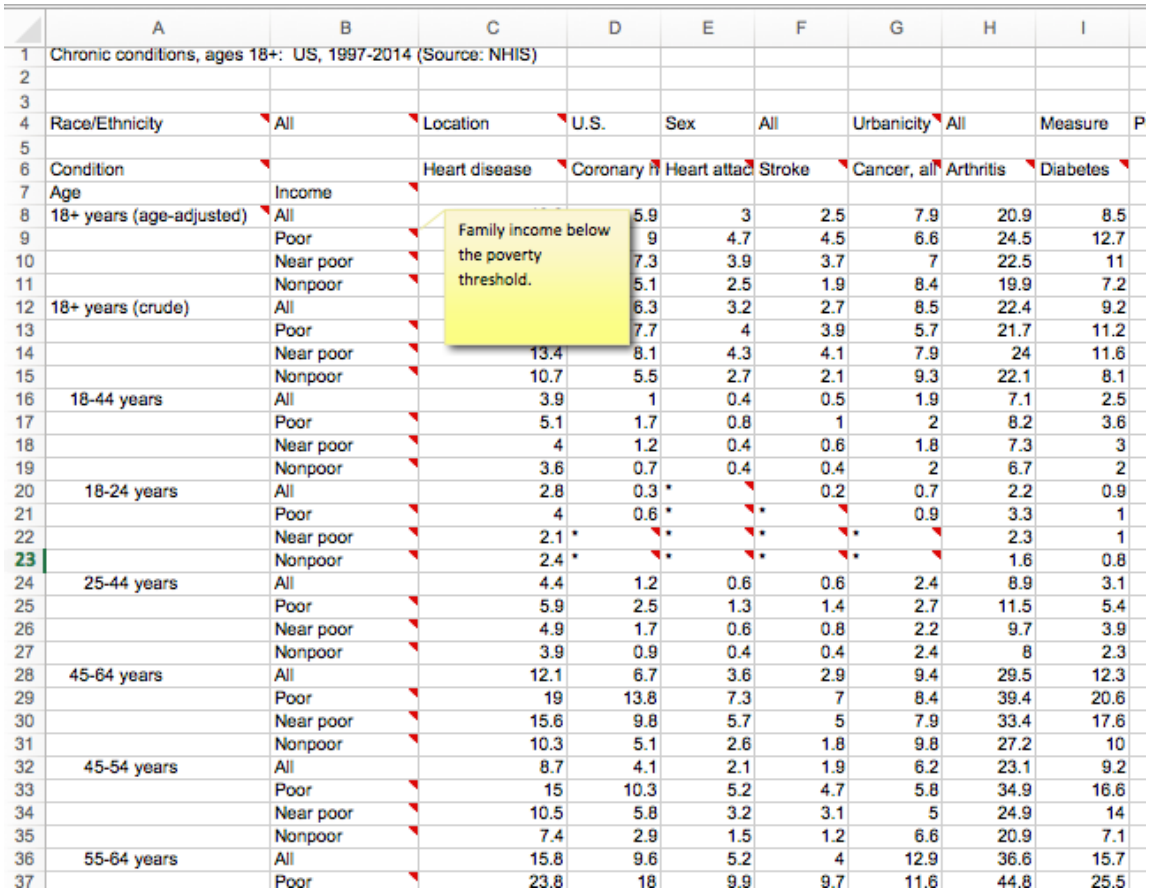
	A	B
1	*Bin*	*Frequency*
2	0	0
3	0.05	1
4	0.1	5
5	0.15	35
6	0.2	12
7	0.25	1
8	More	0

6. Now create a frequency table and histogram for the following sets of data (*Tip: you will need to run Descriptive Statistics and adjust your bins*):
   - Smokes Every Day data for 1995
   - Never Smoked data for 2010
   - Never Smoked data for 1995

7. Compare the four histograms. Comparing 1995 data to 2010 data, what conclusions can you draw about smoking in the United States?

Activity 4: Testing for Independence with Chi Square

1. A Chi square test for independence tests to see if there is a relationship between two categorical (nominal) variables. Another way to state this is whether the two variables are independent of each other. In this exercise, we will answer the question, "From 1997-2014, is there a relationship between income and heart disease for people 55-64 years old?"
2. Open the file **comp24_unit2_dataset_chonic_conditions.xlsx**
   This dataset gives data on persons reporting specific chronic diseases on the National Health Interview Survey. [This dataset is derived from Health Conditions → Chronic conditions at http://205.207.175.93/hdi/ReportFolders/reportFolders.aspx].
3. Take a few moments to look at the data. You'll see the parameters that were used to generate the report on line 4, and then on line 6, the different chronic diseases, such as heart disease, stroke, and arthritis. Along the left side are age categories, subdivided into income levels of Poor, Near poor, and non-poor. There are also numerous comments, indicated by the red triangles, that display messages when the cursor is hovered over them.



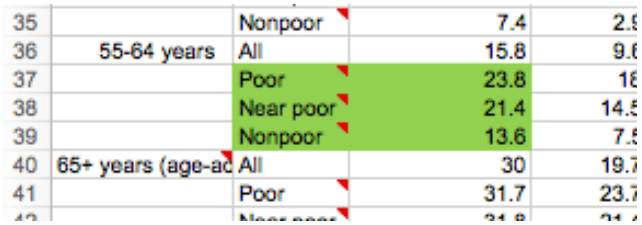
	A	B	C	D	E	F	G	H	I
1	Chronic conditions, ages 18+: US, 1997-2014 (Source: NHIS)								
2									
3									
4	Race/Ethnicity	All	Location	U.S.	Sex	All	Urbanicity	All	Measure	P
5									
6	Condition		Heart disease	Coronary h	Heart attac	Stroke	Cancer, all	Arthritis	Diabetes	
7	Age	Income								
8	18+ years (age-adjusted)	All		5.9	3	2.5	7.9	20.9	8.5	
9		Poor		9	4.7	4.5	6.6	24.5	12.7	
10		Near poor		7.3	3.9	3.7	7	22.5	11	
11		Nonpoor		5.1	2.5	1.9	8.4	19.9	7.2	
12	18+ years (crude)	All		6.3	3.2	2.7	8.5	22.4	9.2	
13		Poor		7.7	4	3.9	5.7	21.7	11.2	
14		Near poor	13.4	8.1	4.3	4.1	7.9	24	11.6	
15		Nonpoor	10.7	5.5	2.7	2.1	9.3	22.1	8.1	
16	18-44 years	All	3.9	1	0.4	0.5	1.9	7.1	2.5	
17		Poor	5.1	1.7	0.8	1	2	8.2	3.6	
18		Near poor	4	1.2	0.4	0.6	1.8	7.3	3	
19		Nonpoor	3.6	0.7	0.4	0.4	2	6.7	2	
20	18-24 years	All	2.8	0.3	*	0.2	0.7	2.2	0.9	
21		Poor	4	0.6	*	*	0.9	3.3	1	
22		Near poor	2.1	*	*	*	*	2.3	1	
23		Nonpoor	2.4	*	*	*	*	1.6	0.8	
24	25-44 years	All	4.4	1.2	0.6	0.6	2.4	8.9	3.1	
25		Poor	5.9	2.5	1.3	1.4	2.7	11.5	5.4	
26		Near poor	4.9	1.7	0.6	0.8	2.2	9.7	3.9	
27		Nonpoor	3.9	0.9	0.4	0.4	2.4	8	2.3	
28	45-64 years	All	12.1	6.7	3.6	2.9	9.4	29.5	12.3	
29		Poor	19	13.8	7.3	7	8.4	39.4	20.6	
30		Near poor	15.6	9.8	5.7	5	7.9	33.4	17.6	
31		Nonpoor	10.3	5.1	2.6	1.8	9.8	27.2	10	
32	45-54 years	All	8.7	4.1	2.1	1.9	6.2	23.1	9.2	
33		Poor	15	10.3	5.2	4.7	5.8	34.9	16.6	
34		Near poor	10.5	5.8	3.2	3.1	5	24.9	14	
35		Nonpoor	7.4	2.9	1.5	1.2	6.6	20.9	7.1	
36	55-64 years	All	15.8	9.6	5.2	4	12.9	36.6	15.7	
37		Poor	23.8	18	9.9	9.7	11.6	44.8	25.5	

Family income below the poverty threshold.

Set up a table of observed values

Since we want to look at the relationship between income level and heart disease for people between 55-64 years old, we need to set up a table that has columns for Income Level, Has Heart Disease, and No Heart Disease. This will be our Observed Values.

1. Copy the values for Income and Heart Disease for rows 37, 38, and 39 to a blank area of your worksheet. This gives us the number per 100 patients who reported that yes, they had heart disease:



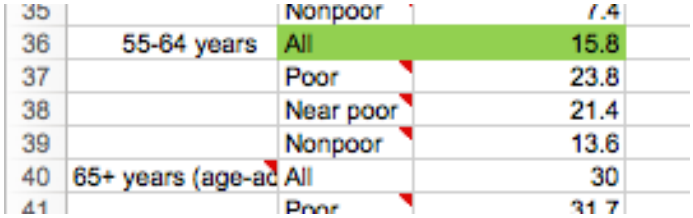
35		Nonpoor	7.4	2.5
36	55-64 years	All	15.8	9.6
37		Poor	23.8	18
38		Near poor	21.4	14.5
39		Nonpoor	13.6	7.5
40	65+ years (age-ad	All	30	19.7
41		Poor	31.7	23.7

2. Now we need to manually calculate how many patients per 100 who did not have heart disease. In the blank column to the right, subtract each value from 100.
3. Add a title above each column; the first column should be Income, the second “Heart Disease per 100”, and the third column “No Heart Disease”.
4. Above the Income column, add the title “Observed Values”.
5. Your work area should now look like this:

Observed Values		
Income	Heart Disease per 100	No Heart Disease
Poor	23.8	76.2
Near poor	21.4	78.6
Non-poor	13.6	86.4

Set up a table of expected values

The Chi square test needs to have a range of expected values to compare the observed values against. For the purposes of this exercise, we will use the data for All patients between 55-64 years old on row 36, which has a value of 15.8.



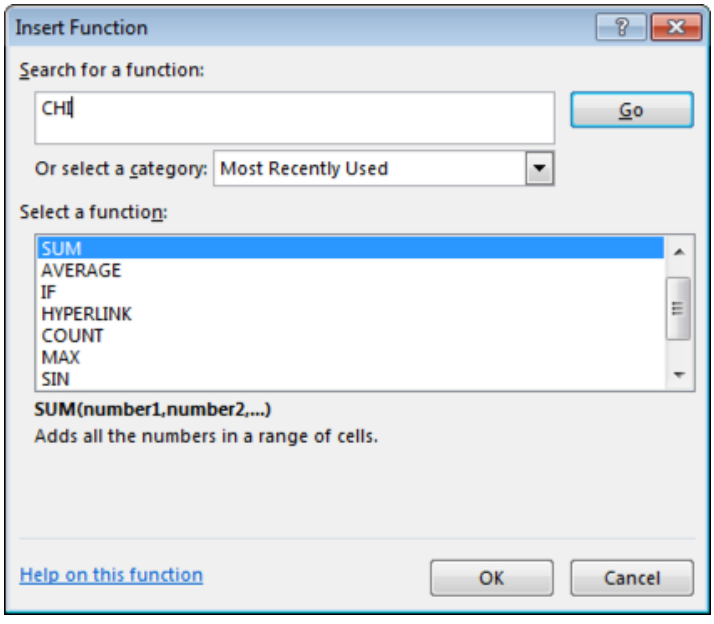
35		Nonpoor	7.4
36	55-64 years	All	15.8
37		Poor	23.8
38		Near poor	21.4
39		Nonpoor	13.6
40	65+ years (age-ad	All	30
41		Poor	31.7

6. Copy the cell with the value All and the adjacent cell with the value 15.8. This gives us the number per 100 patients, of all income levels, who reported that yes, they had heart disease.
7. Paste this section a few lines below your new Observed Values area of your worksheet.
8. Now we need to manually calculate how many patients per 100 who did not have heart disease. In the blank column to the right, subtract 15.8 from 100.
9. Since we have three rows of data in our Observed Values area, we have to have three rows in our Expected Values area. Copy the three cells "All", "15.8", and "84.2" to two additional rows.
10. Add the title "Expected Values".
11. Your work area should now look similar to this:

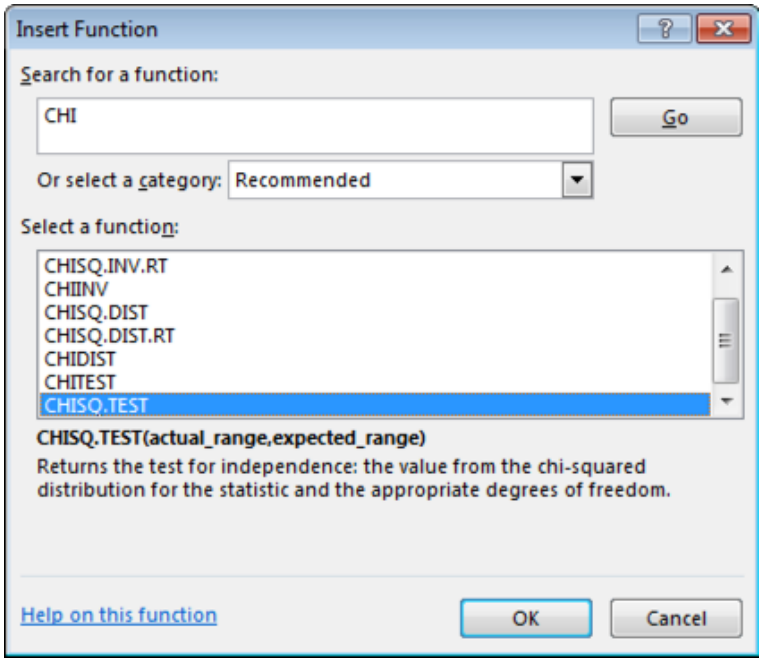
Observed Values		
Income	Heart Disease per 100	No Heart Disease
Poor	23.8	76.2
Near poor	21.4	78.6
Non-poor	13.6	86.4
Expected values		
All	15.8	84.2
All	15.8	84.2
All	15.8	84.2

Run the Chi square test for independence

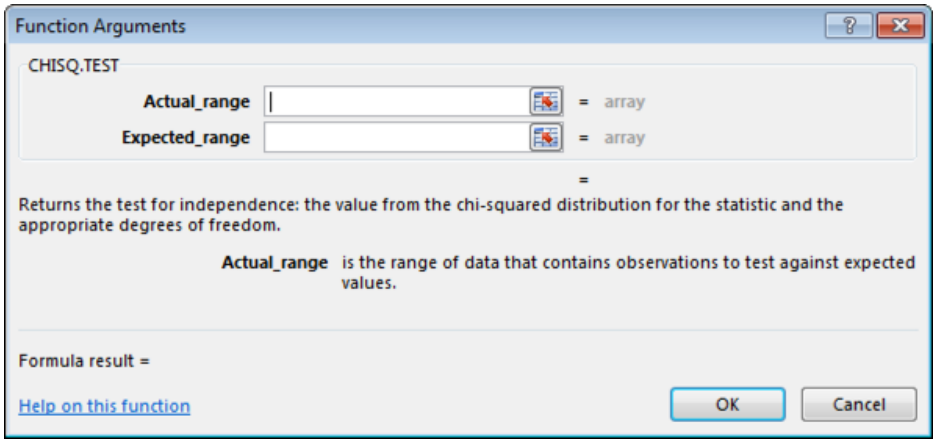
12. Click in an empty cell underneath your table of expected values.
13. On the menu bar, choose **Formulas → Insert Function**. The Insert Function dialog box will display. Type in **CHI** in the search box and click **Go.**



14. A list of functions whose names begin with CHI will be displayed. Select **CHISQ.TEST** and click **OK**.



15. You will now see the **Function Arguments** dialog box for the CHISQ.TEST function.



- **Actual_range**: click in the Actual Range field, then click the cell with the 23.8 value and drag down and to the right until all six data values for the Observed Values are highlighted. The cell names will be entered in the field. (Note that your cell names may be different, depending on where you placed your data in your spreadsheet.)

- **Expected_range:** click the Expected Range field, then click the cell with the first 15.8 and drag down and to the right until all six data values for the Expected Values are highlighted. The cell names will be entered in the field. Again, your cell names may be different.
- Click **OK.** The number 0.023147 should appear in the cell.

Interpreting the Data

In this case, the result reported as 0.023147.

- If the value is less than or equal to (< =) 0.05, then there is a statistically significant relationship between income level and heart disease.
- If the value is greater than (>) 0.05, then there is not a statistically significant relationship between income level and heart disease

Answer the following questions

- Is there a statistically significant relationship between income level and heart disease?
- Are heart disease and income level independent of each other?