StatCrunch Data analysis on the Web

About This Study Card

StatCrunch is a Web-based statistical software package for analyzing data. To access StatCrunch, visit <u>http://www.statcrunch.com</u>. This study card is intended to serve as a brief introduction to the use of StatCrunch covering the procedures that most students will encounter in an introductory statistics course. Follow the help links at the statcrunch.com site for more extensive documentation.

Data > Load data > from file

- 1. Choose **My computer** to load a data file from the local system or **WWW** to load a data file from the Web.
- 2. Specify the location of a text file (.txt, .csv, etc.) or Microsoft Excel file.
- If the first line in your file does not contain column names, deselect the use first line as column names option.
- 4. For text files (not Excel files), specify the delimiter for the data values. For example, the delimiter for a .csv file is a comma.
- 5. Click the Load File button to upload the data file.

About this data set: The data set contains information from a random sample of 30 four-bedroom homes listed for sale in the Bryan–College Station, Texas area in December 2008. For each home, the data set contains the list price in thousands of dollars (Price), square footage (Sqft), number of bathrooms (Baths), and location (Bryan, TX, or College Station, TX). It is currently being shared on the StatCrunch site at http://www.statcrunch.com/5.0/index.php?dataid=359673.

Data > Compute Expression

- 1. Enter a mathematical or Boolean expression in the **Expression** input box.
- 2. Alternatively, use the menus for **Y**, **X**, **a**, **b**, and **Function** to construct the expression. Then, click **Set Expression**.
- 3. Click **Compute** and the results of the expression will be added as a new column to the StatCrunch data table.

Note: A Boolean (true/false) expression can be used as a **Where** statement in many StatCrunch procedures to exclude outliers or to focus an analysis on a subset of the data. Some of the following examples illustrate this feature.

	Loud data from my compared	Location	Baths	Sqft	Price	Row
		Bryan, TX	4.5	4918	2400	1
Bernan	File: Nc2\users\west\My Documents\bes vis	Bryan, TX	3	3645	680	2
-	The best states at account as and one	Bryan, TX	2	1900	570	3
	Use first line as column names:	Bryan, TX	3.5	3123	400	4
		Bryan, TX	3	2365	.320	5
	Delimiter	Bryan, TX	- 3	3361	280	6
	whitespace	Bryan, TX	2.5	2383	260	7
	(required for text files only) Additional onlines amilted	Bryan, TX	4.5	3770	229	8
	Province of the office of the	Bryan, TX	3.5	2979	215	9
	Load File. Cancel	Bryan, TX	2	2088	200	10
		Bryan, TX	2.5	2343	183	11
	Load data from WWW address	Bryan, TX	2	2526	170	12
		Bryan, TX	2	1812	145	13
	Laugh Care and Care	Bryan, TX	2	2432	100	14
-	WWW Address:	Bryan, TX	2.5	2988	163	15
	The pay in the decision in the bar section					
	Use first line as column names:					
	Delimiter					
	whitespace					
	(required for test files oney)					
	Additional options onlined					
	Additional options omitted					

Example: Loading an Excel file from the local file system or from a Web address

0w/ 1 2	Price 2400 680	Sqft 0 4918 0 3645	Baths 4.5	Location Bryan, TX Bryan, TX	Price/Sqt 488.00327 186.55693	can be edit the express	ed after ion has	var8
3	Compute ex	pression	4	Diffunction)	1	Concertainty	UREO,	-
5	Expression	1000*(Price/SqR)				ī.		
7	Enter express	sion above or use i	menus below to	o create it.				
910 11 2 13 4	* X: 8	Select Column Select Column 0 1		Functions: log(Y) log(Y) sqft(Y) 1/Y Y*Y Y+X		8		
5	-	Set Expression	1	4/X 4-X			C -	

Example: Computing the mathematical expression, 1000*(Price/Sqft)



Graphics > Bar plot

- 1. Choose the **with data** option to use data consisting of individual outcomes in the data table.
 - a. Select the column(s) to be displayed.
 - b. Use an optional Where statement to specify the data included.
 - c. Select an optional **Group by** column for a side-by-side bar plot.

Choose the **with summary** option to use summary information consisting of categories and counts.

- a. Select the column containing the categories.
- b. Select the column containing the counts.
- 2. Click **Next** to set additional options such as **Type** (Frequency or Relative Frequency).
- 3. Click Create Graph! to construct the bar plot(s).

Graphics > Histogram

- 1. Select the column(s) to be displayed.
- 2. Use an optional **Where** statement to specify the data included.
- 3. Select an optional Group by column to color code the bars.
- 4. Click **Next** to set additional options such as the starting value of the bins and the bin width.
- 5. Click **Next** again to specify an optional density to overlay on the histogram(s).
- 6. Click Create Graph! to construct the histogram(s).

Graphics > Boxplot

- 1. Select the column(s) to be displayed. By default, a boxplot for each column will be included on a single graph.
- 2. Use an optional **Where** statement to specify the data included.
- 3. Select an optional **Group by** column to compare boxplots across groups on a single graph.
- 4. Click **Next** to indicate whether or not to use fences for the boxplots. By default, the five-number summary is used.
- 5. Click **Create Graph!** to construct the boxplot(s).

Graphics > Scatter Plot

- 1. Select the X variable and Y variable for the plot.
- 2. Use an optional **Where** statement to specify the data included.
- 3. Color code points with an optional Group by column.
- 4. Click Create Graph! to produce the plot.



Example: A bar plot showing the number of homes in each location with three or more baths



Example: A histogram of the Sqft column





Date D	Data Stat	Coraphics	Scatter Plot	
1 Scatter Plot	2400	4918 2046	Price	
X variable: Y variable: Where: Group by:	Sqt Price optional- Location C Separate	graph for eac	2000 - 1500 - 1800 -	Location Bran, TK College Station, TX

Example: Scatter plot of Price vs. Sqft color coded by location

Graphics > QQ Plot

- 1. Select the column(s) to be displayed.
- Use an optional Where statement to specify the data included.
- 3. Select an optional **Group by** column to produce separate plots for different groups.
- 4. Click Create Graph! to construct the plot(s).



Example: QQ plot of the Sqft column grouped by Location

Interacting with Graphics

- 1. Click and drag the mouse around graph objects to highlight them.
- 2. The corresponding rows will be highlighted in the data table and in all other graphics.
- 3. Toggle highlighting on and off by clicking on the row number in the data table.
- 4. To clear all highlighted rows, click on the **Row** heading atop the first column in the data table.
- To highlight rows based on categories or numeric ranges, use the Data > Row selection > Interactive tools option.

Stat > Summary Stats > Columns

- 1. Select the column(s) for which summary statistics are to be computed.
- 2. Use an optional **Where** statement to specify the data included.
- 3. Compare statistics across groups using an optional **Group by** column.
- 4. Click **Next** for additional options such as the statistics to be computed.
- 5. Click Calculate to compute the summary statistics.



Example: Comparing prices of homes listed in Bryan to those listed in College Station with the potential outlier removed

Price	8	Fait	Baths	Location	Price/Saft	varfi	var7	var8
	2408	4918	4.5	Bryan, TX	🛃 Fre	quency Tal	ble 💶 🗖	×
	690	3845	3	Bryan, TX	Della	1746		
	570	1900	2	Bryan, TK	- SADAR	ITAL .		
		100			Freque	ncy table res	suits for Baths:	
10.00	iquency Te	ble			Baths	Frequency	Relative Frequency	
Sele	ect Columns				2	7	0.23333333	-
Pric	8		Baths		2.5	6	0.2	
Sqt		_			3	8	0.26666668	
Loc	ation				3.5	5	0.18666667	
Pric	e/Sqft				4	2	0.056666667	
Whe	in:	wal-			4.5	2	0.06666667	_
								1 m .

Stat > Tables > Frequency

- Select the column(s) for which a frequency table is to be computed.
- 2. Use an optional **Where** statement to specify the data included.
- 3. Click **Calculate** to compute the frequency table(s).

Example: A frequency table for the number of bathrooms

Stat > Proportions > One sample

- 1. Choose the **with data** option to use data from the data table.
 - a. Select the column containing the sample values.
 - b. Specify the outcome that denotes a success.
 - c. Use an optional Where statement to specify the data included.

Choose the **with summary** option to enter the **number of successes** and **number of observations**.

- Click Next and select the hypothesis test or confidence interval option.
 - a. For a hypothesis test, enter the **Null** proportion and choose \neq , <, or > for **Alternative**.
 - b. For a confidence interval, enter a value between 0 and 1 for Level (0.95 provides a 95% confidence interval). For Method, choose Standard-Wald or Agresti-Coull.
- 3. Click Calculate to view the results.

Stat > Proportions > Two sample

- 1. Choose the **with data** option to use sample data from the StatCrunch data table.
 - a. Select the columns containing the first and second samples.
 - b. Specify the sample outcomes that denote a success for both samples.
 - c. Enter optional **Where** statements to specify the data rows to be included in both samples. If the two samples are in separate columns, this step is typically not required.

Choose the **with summary** option to enter the **number of successes** and **number of observations** for both samples.

- 2. Click **Next** and select the **hypothesis test** or **confidence interval** option.
 - a. For a hypothesis test, enter the **Null** proportion difference and choose \neq , <, or > for **Alternative**.
 - b. For a confidence interval, enter a value between 0 and 1 for **Level** (0.95 provides a 95% confidence interval).
- 3. Click Calculate to view the results.

Stat > T statistics > One sample

- 1. Choose the **with data** option to use sample data from the StatCrunch data table.
 - a. Select the column containing the sample data values.
 - b. Use an optional **Where** statement to specify the data included.

Choose the with summary option to enter the sample mean, sample standard deviation, and sample size.

- 2. Click **Next** and select the **hypothesis test** or **confidence interval** option.
 - a. For a hypothesis test, enter the **Null** mean and choose \neq , <, or > for **Alternative**.
 - b. For a confidence interval, enter a value between 0 and 1 for **Level** (0.95 provides a 95% confidence interval).
- 3. Click Calculate to view the results.

1 no Cptions	
2 10	
3 no 95% confidence interval results:	
4 np Outcomes in : premature	
5 no Success no	
6 yes Cone sample p' proportion of successes	
One sample Proportion with d	
Variable Count Total Sample Prop. Std.E	L.Limit U.Limit
nonmature 200 000 00000000000000000000000000000	1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
premature ever abe dievrosse contar.	140 0.02040200 0.00000075
Success: no. Confidence interval	
Level 0.95	
Wiete:	
ireup by:optional v Method: Standard-Wald 💌	3

Example: For each of 998 North Carolina births, this data set indicates whether or not the birth was premature. A 95% confidence interval for the proportion of all North Carolina births that are *not* premature is shown above.

Sam	ple 1 - Sample 2:						(+ Hypoth	eers lest		
Sar	mple1: Number of	f success	805	181	-		Null: p	rop. diff. =	8	
	Number of	l observa	Bons:	336			Altern	ative:		
Sal	mple2: Number of	sucesse		71	-		C Confid	ence Intern	len	
	Number of	Icheener	tioner	168	-		Lower		D.R.S.	1
					-					
7 Sn	apshot Cance	Bar	R Neo	xt> C	alculate	2 Sn	apshot Car	ncei <b< th=""><th>ack lines</th><th>Calcula</th></b<>	ack lines	Calcula
7 _ Sn	epshot Cance	nple Prop	e Neo	xt> C with su	alculate	2 Sn	apshot Ca	ncel <b< td=""><td>ack Just</td><td>Calcula</td></b<>	ack Just	Calcula
? Sn	apshot Cance	npte Prop	portion	xt> C with su	alculate	2 Sn	apshot Car	ncel <b< td=""><td>ack Jikes</td><td>Calcula</td></b<>	ack Jikes	Calcula
? <u></u> Sn	Append Cancel Two ser Options Hypothesis P1 proportion P2 proportion P3 proportion P4 proportion P4 proportion P4 proportion	test result on of succ on of succ orence in p : 0	Nes pertion Its: resses for respective	xt> C with su or popula or popula ns	alculate another y ation 1 ation 2	2 <u>3n</u>	apshot Ca	ncel <b< td=""><td>ack lines</td><td>Calcula</td></b<>	ack lines	Calcula
? <u></u> Sn	Image: Second state Cancel Image: Second state Cptons Hypothesis P1 P2 proportion P3 p1 P4 p1	test resul on of succ on of succ rence in p : 0 • 0 Count1	Nex portion Its: cesses for cesses for cesse	xt> C with su or popula or popula ns Count2	alculate ation 1 ation 2 Total2	Sample Diff.	apshot Can	Z.Stat	P-volue	Calcula

Example: In a survey, 181 of 336 people who attended church services at least once a week said the use of torture against suspected terrorists is "often" or "sometimes" justified. Only 71 of 168 people who "seldom or never" went to services agreed. Using this **summary** information, a two-sided hypothesis test is performed below to compare the proportion with this opinion for these two populations.

StatCrun	ch Data St	a Gr				🖪 Qne s	ample T statis	tics with d	ata		
Row	Price	Saft	One sample 1 stat	stics with e	464	Options					
1 2 3	2400 680 570	4 3 1	(* Hypot	Diesis Test	200	Hypothesi Where: Lo	is test results: istation = "Colleg	e Station, T	ĸ		
🚯 One si	anple T statistic	a s with c	Alter	native:	~	Hg: µ = 20 Hg: µ > 20	000 000			-	
Select 0	Columns:		C Conti	dence Intervi	4	Variable	Sample Mean	Std. Err.	DF	T-Stat	P-value
Price Soft Baths Price/So	গ	Sat	? Snapshot Can	cei «Baci	42. 	Sqf	2750.6	199.84525	14	3.7559063	0.0011
Where: Group b	Location = "C ny:opbonal- gistiot Cencel	College Stat	bon, TX Next> Calculate	han, TX han, TX han, TX han, TX	67 80 41 56	30008 92207 11842 52249					

Example: Testing to see if the average home listed in College Station, Texas, is larger than 2000 square feet.

Stat > T statistics > Two sample

- 1. Choose the **with data** option to use sample data from the data table.
 - a. Select the columns containing the first and second samples.
 - b. Enter optional Where statements to specify the data rows to be included in both samples. If the two samples are in separate columns, this step is typically not required.

Choose the **with summary** option to enter the **sample mean**, **sample standard deviation**, and **sample size** for both samples.

- 2. Deselect the **Pool variance** option if desired.
- 3. Click **Next** and select the **hypothesis test** or **confidence interval** option.
 - a. For a hypothesis test, enter the **Null** mean difference and choose \neq , <, or > for **Alternative**.

b. For a confidence interval, enter a value between 0 and 1 for **Level** (0.95 provides a 95% confidence interval).

4. Click **Calculate** to view the results.

Stat > T statistics > Paired

- 1. Select the columns containing the first and second samples.
- 2. Use an optional Where statement to specify the data included.
- 3. Click **Next** and select the **hypothesis test** or **confidence interval** option.
 - a. For a hypothesis test, enter the **Null** mean difference and choose \neq , <, or > for **Alternative**.
 - b. For a confidence interval, enter a value between 0 and 1 for **Level** (0.95 provides a 95% confidence interval).
- 4. Click Calculate to view the results.

Row	Price	Suft	Baths	Lot	
1	2400	4918	4.5		a service and a service and a service a s
2	690	3645	3		Hypothesis Test
3	570	1900	2		Null: mean ditt. = 0
4	400	3123	3.5	- 4	
5	320	2365	3	1	Alternative:
6	280	3361	.3	1	C Confidence Interval
Sample	1 - Sample 2:	_			7. Snapshot Cancel <back calculate<="" january="" th=""></back>
Sample Sample Sample 7 (Pao)	1 - Sample 2: 1 III Price Where : 2 III Price Where: variances Snapshot Ca	Location = "Bri M Location = "Ga	ian, TX*		Smacshiel Cancel +Black Inter Calculate Dictions Hypothesis test results: y: mean of Price where Location = "College Station, Tc" y: y: mean of Price where Location = "College Station, Tc" y: y: g: mean of Price where Location = "College Station, Tc" y: y: g: mean of Price where Location = "College Station, Tc" y: y: g: mean of Price where Location = "College Station, Tc" y: y: g: g: g: g Ho, y: g: g: g Ho, y: g: g: g
Sample Sample Sample 2 Pool	1 - Sample 2: 1 m Price Where : 2 m Price Where: variances Snapshot Ca	Location = "Br Cocation = "Ca	van. TX Niege Station, T		Unset Unset Cancel +Elack Unset Cancel Cancel

Example: Hypothesis test comparing the prices of homes listed in Bryan to those in College Station, with potential outliers included.



Example: A 95% confidence interval for the average difference between the graduation rate of full- and part-time students across all colleges based on a random sample of four colleges.

Stat > Regression > Simple Linear

- 1. Select the X variable (independent variable) and Y variable (dependent variable) for the regression.
- 2. Use an optional Where statement to specify the data included.
- 3. Compare results across groups by selecting an optional Group by column.
- 4. Click Next to specify optional X values for predictions and to indicate whether or not to save residuals/fitted values.
- 5. Click Next again to choose from a list of optional graphics for plotting the fitted line and examining residuals.
- 6. Click Calculate to view the results.

StatCrunich	Data St	at Graphics	Help													
Row	Price	Sat	Baths	Location	Price/Sqft	var6	var7	varā								
1	2400	4918	4.5	Bryan, TX	488.00327	Simple I	Linear Rem	assion -								
2	680	3645	3	Bryan, TX	186.55693	Contraction of the second	annear magn	Contraction -							-	-
3	570	1900	2	Bryan, TX	300	Opbons						🗈 Simple Linear Reg				- 8
4	400	3123	3.5	Bryan, TX	128.08197	Simple linea	r regression	results:				Ontions				
5	320	2385	3	Bryan, TX	135.30655	Where: Loca	don = "Colle	pe Station, TX				E-people				
6	280	3361	3	Bryan, TX	\$3,30854	Dependent	/ariable: Pric	0				Dates	Fitted line plu	M		
7	260	2383	📕 Sin	ple Linuar Re	gression	Independent	tVariable: Sq	n				+nce				1
	2001	2220				Price = -65.0	7215 • 0.13	972908 Sqft							1	
Simple Li	imple Linear Regression Graphics:				157	Sample size	:15					500		100	/	
			-			P (cometabol	n coefficient)	= 0.9002							/	
				N PI	of the littled line	Fetimate of a	14137	a designation in	1 2026					/		
				FR	esidual indes plo	Parameter e	estimates:	a neverus 2-	4.2030			400 -		/		
X-Variable	e Sqit	~		Гн	stagram of resid	Parameter	Estimate	Std. Err.	DF	T-Stat	P-Value		./			
	-			En	inter of south inter	Intercept	-65.07218	53.4287	6 13	-1.2179236	0.2449	300 +	1			
Y-Variable	Price	~		1 4	a prot of residual	Slope	0.13972908	0.01874412	26 13	7.4545538	*8.0001					
Where:	Location =	College Station	TX"	L B	rsidualo vs. X-va	Analysis of v	rariance tabl	e for regress	ion mo	del:		200				
Group by:	-optional-	¥		LTY.	values vs. reside	Source DF	SS	MS	F-stat	P-value		1				
		-		17 P	ned values vs. re		1		Net-	-	-	1500 2000	2500 3000 Sqft	3500	4000	4500
			Asta	and a	and drived	Building ou	and and				-		- Bach	-		
				stration (Ca	ACAR ABACK	HEXT > La	HC UNITED					-				-
2 Snapat	al causel	I Share I Filler	The Colour	to I								FD		f Duiro		ford

Example: Regression of Price on Sqft for College Station

Stat > ANOVA > One way

- 1. Select one of the following options:
 - a. If the samples are in separate columns, select the **Compare** selected columns option and then select the columns containing the samples.
 - b. If the samples are in a single column, select the **Compare** values in a single column option. Then specify the column containing the samples (**Responses In**) and the column containing the population labels (Factors In).
- Use an optional Where statement to specify the data to be included.
- Select the Tukey HSD option and specify a confidence level to perform a post hoc means analysis. The default value of 0.95 provides 95% confidence intervals for all pairwise mean differences.
- 4. Click Calculate to view the results.

Stat > Tables > Contingency

- 1. Select one of the following options:
 - a. Choose the **with data** option to cross tabulate two columns of raw data from the data table. Then,
 - A. Select the column to be tabulated across the rows.
 - B. Select the column to be tabulated across the columns.
 - C. Use an optional Where statement to specify the data included.
 - D. Select an optional **Group by** column to compute separate tables across groups.
 - b. Choose the **with summary** option to use a two-way cross classification already entered in the data table. Then,
 - A. Select the columns that contain the summary counts.
 - B. Select the column that contains the row labels.
 - C. Enter a name for the column variable.
- Click Next to specify additional information to be displayed in each table cell.
- 3. Click Calculate to view the results.

Stat > Calculators

- Select the name of the desired distribution from the menu listing (e.g., Binomial, Normal, etc.).
- 2. In the first line below the plot in the calculator window, specify the distribution parameters. As examples, with the normal distribution, specify the mean and standard deviation or with the binomial distribution, specify *n* and *p*.
- 3. In the second line below the plot, specify the direction of the desired probability.
 - a. To compute a probability, enter a value to the right of the direction selector and leave the remaining field empty (e.g., P(X < 3) =___).
 - b. To determine the point that will provide a specified probability, enter the probability to the right of the direction selector and leave the other field empty (e.g., $P(X \ge 0.25)$). This option is available only for continuous distributions.

6

4. Click **Compute** to fill in the empty fields and to update the graph of the distribution.



Example: ANOVA for the time, in minutes, a caller stays on hold before hanging up under three different treatments.

Row	gender	no	yes	yar4	varfi			var6	var7	yar8	-
10	female	31	24		(I I I I I I I I I I I I I I I I I I I			and in		1	
2	malé	37	8	_	En Cont	ung	ency	Table	<u></u>		
3		_		-	Option	15					
Se	lect columns for	table:	ау <u>(</u> =		Rows g Column	end s; B;	er ack p	e results:			
1 🔤		no.			-	rio.	yes	Total			
11		Tas			female	31	24	55			
1				44	male.	37	18	45			
	internet in		-	-	Total	68	32	100		-	
H	W LODGES MT.	Geiliger	m		Statist	tic	DF	Value	P-value		
Co	lumn variable:	Back problems			Chi-sou	are	Ŧ	7 6054864	0.0058	-	



