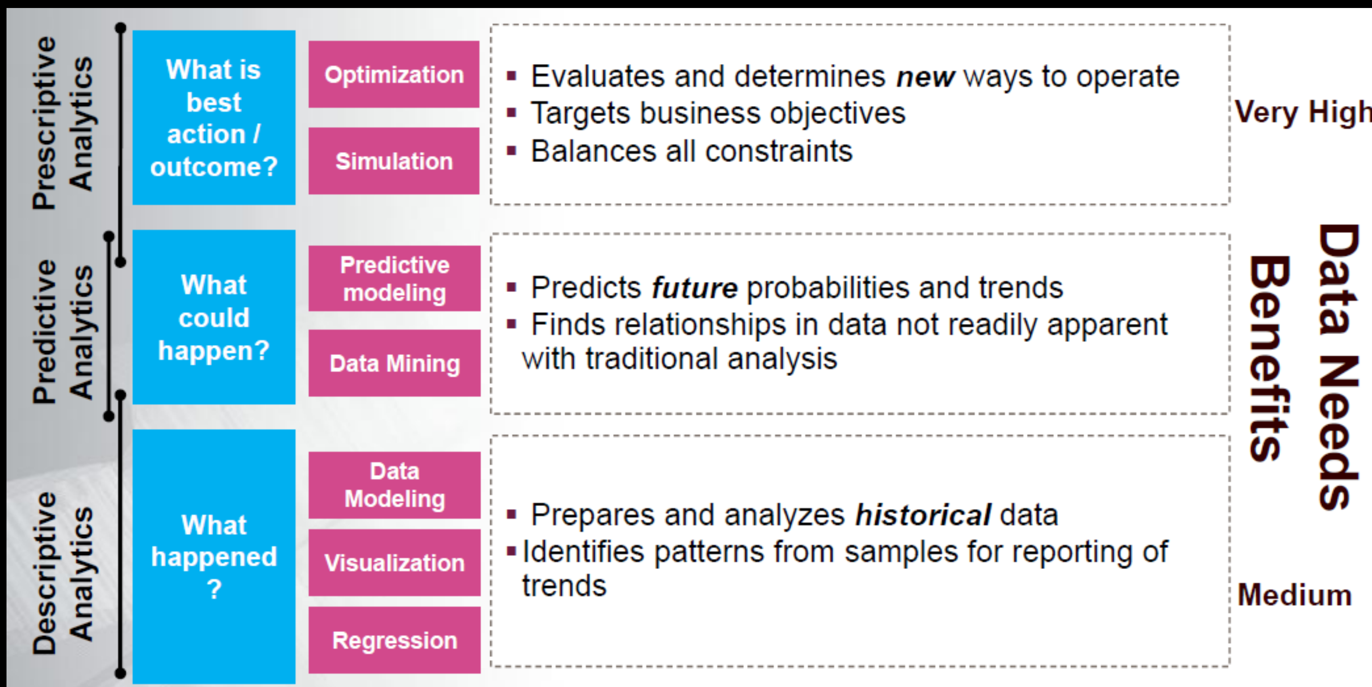
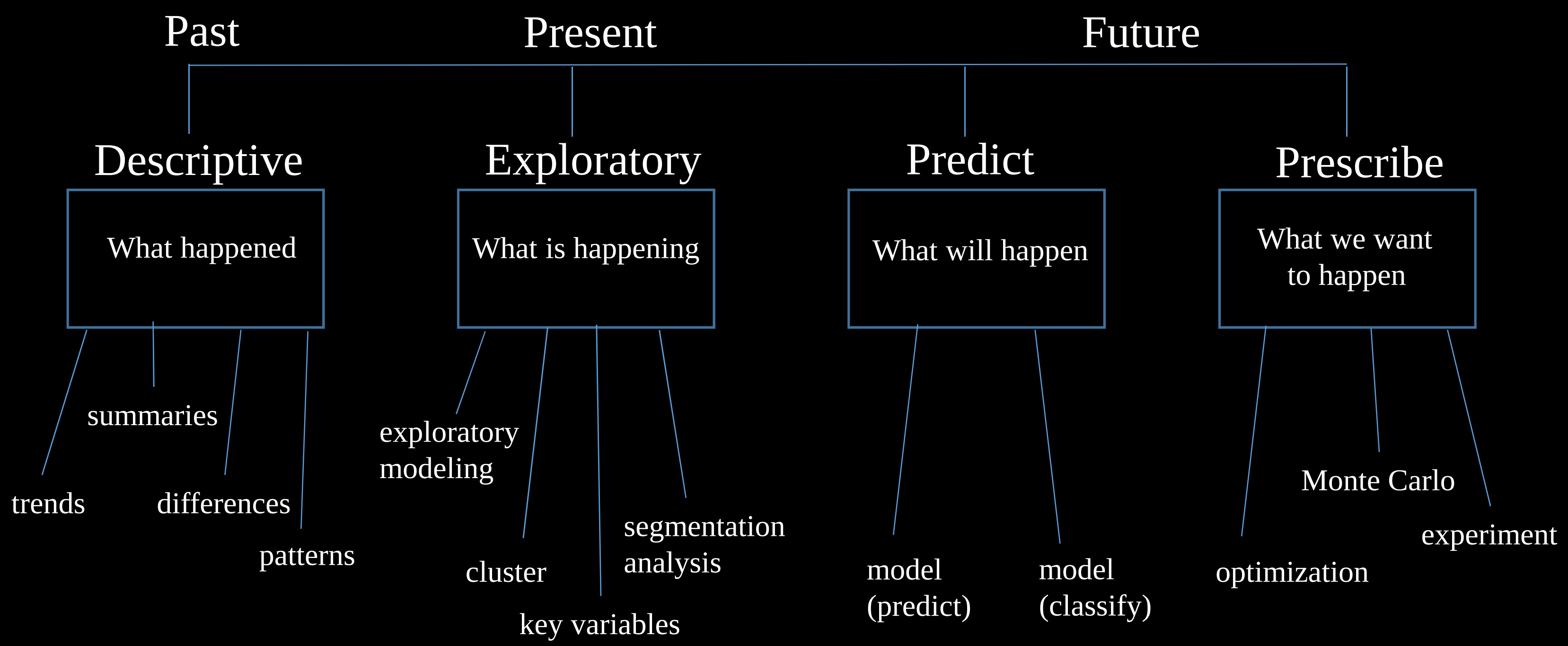
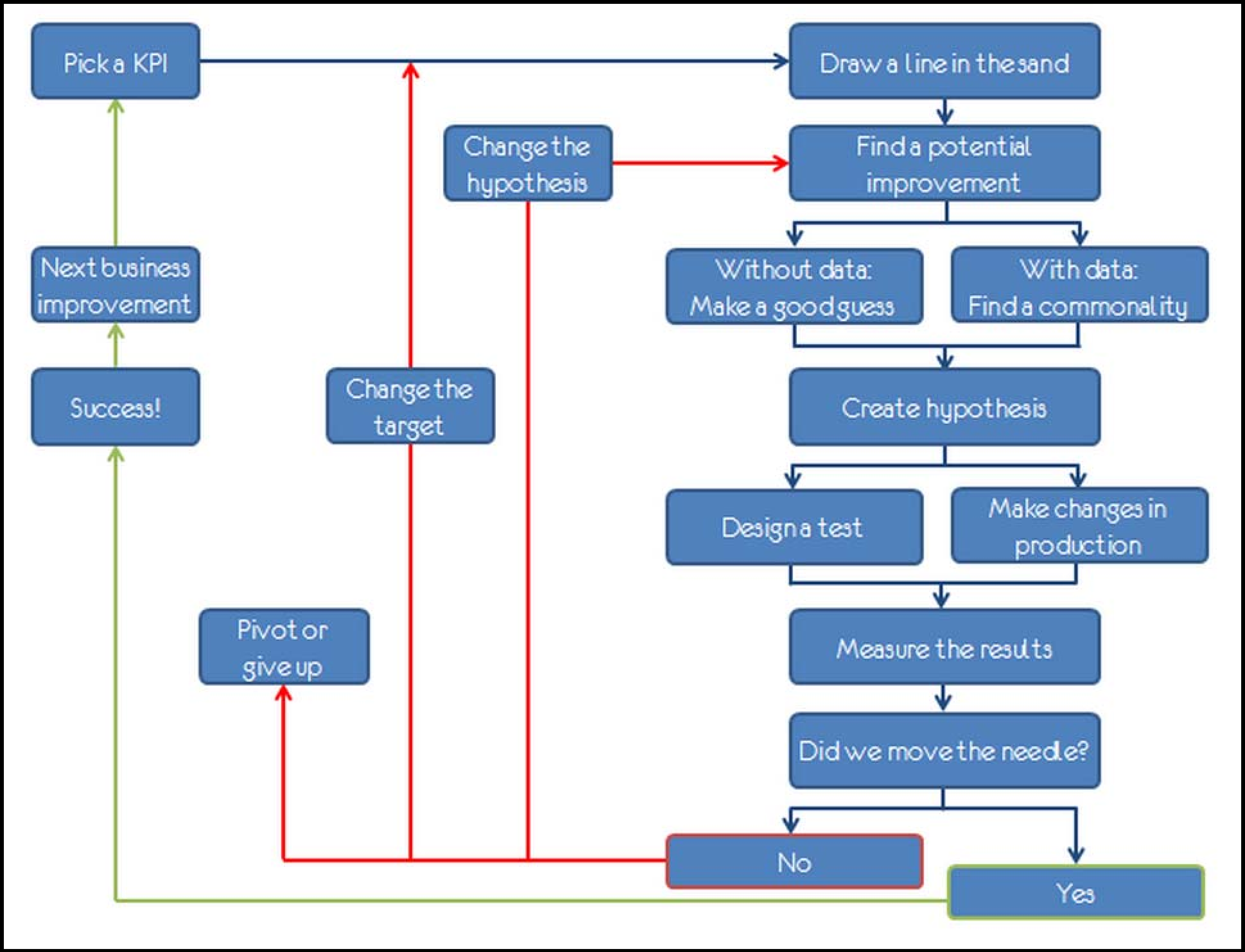
Differences


Prescriptive Analytics
What is best action / outcome?
Optimization
Simulation
▪ Evaluates and determines ***new*** ways to operate
▪ Targets business objectives
▪ Balances all constraints
Very High
Predictive Analytics
What could happen?
Predictive modeling
Data Mining
▪ Predicts ***future*** probabilities and trends
▪ Finds relationships in data not readily apparent with traditional analysis
Data Needs
Benefits
Descriptive Analytics
What happened ?
Data Modeling
Visualization
Regression
▪ Prepares and analyzes ***historical*** data
▪ Identifies patterns from samples for reporting of trends
Medium



Source: Jim Grayson, Ph.D. and Mia Stephens



Avinash Kaushik on Lean Analytics | http://www.kaushik.net/avinash/lean-analytics-cycle-metrics-hypothesis-experiment-act/

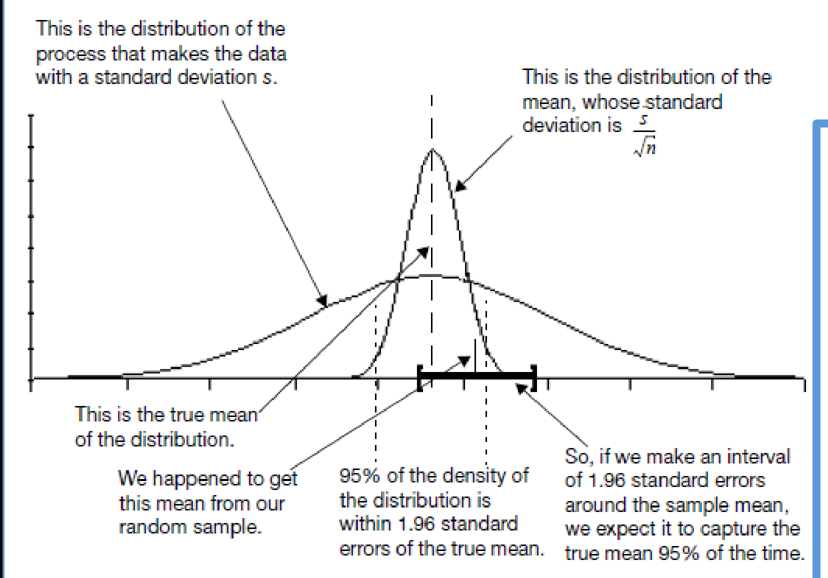
One Variable

Mean: Hypothesis Testing

Mean: Confidence Interval

Understanding CIs

Figure 7.10 Illustration of Confidence Interval



JMP Start Statistics, 4e (SAS Press)

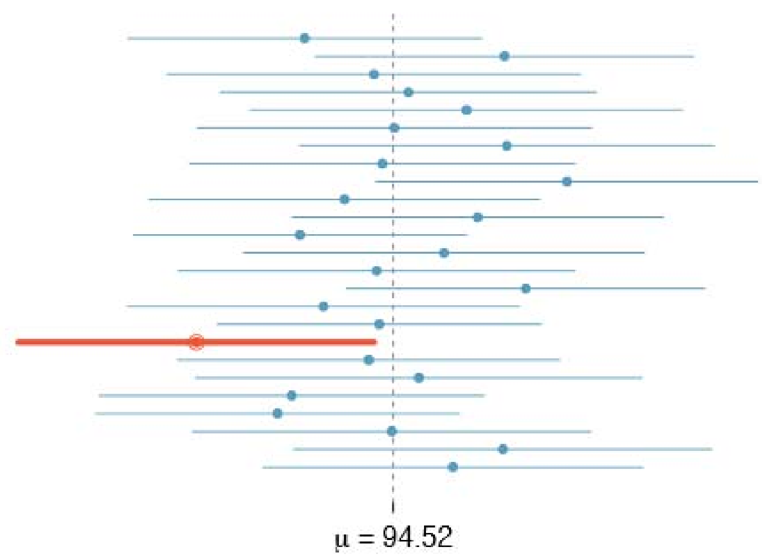


Figure 4.8: Twenty-five samples of size $n = 100$ were taken from the `run10` data set. For each sample, a confidence interval was created to try to capture the average 10 mile time for the population. Only 1 of these 25 intervals did not capture the true mean, $\mu = 94.52$ minutes.

OpenIntro Statistics, 2e

Confidence Interval Principle

Every confidence interval is constructed as:

Point Estimate ± Confidence Level Multiplier * Standard Error

Every confidence interval is attempting to “capture” the population parameter and is interpreted as:

We are X% confident [insert confidence level] the true (but unknown) population parameter is between [lower CI limit] and [upper CI limit]

jmp.com/learn rev 02/2014

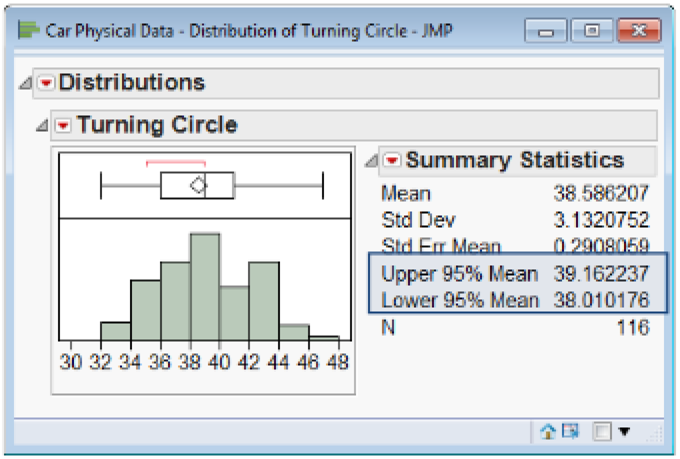
One Sample t-Test and CI

Confidence Interval for the Mean

1. From an open JMP® data table, select **Analyze > Distribution**.
2. Select one or more continuous variables from **Select Columns**, click **Y, Columns** (continuous variables have blue triangles), and click **OK**.

   The **Upper 95% Mean** and **Lower 95% Mean** give the 95% confidence interval for the true mean (39.163 and 38.01).

Example: Car Physical Data.jmp (Help > Sample Data)

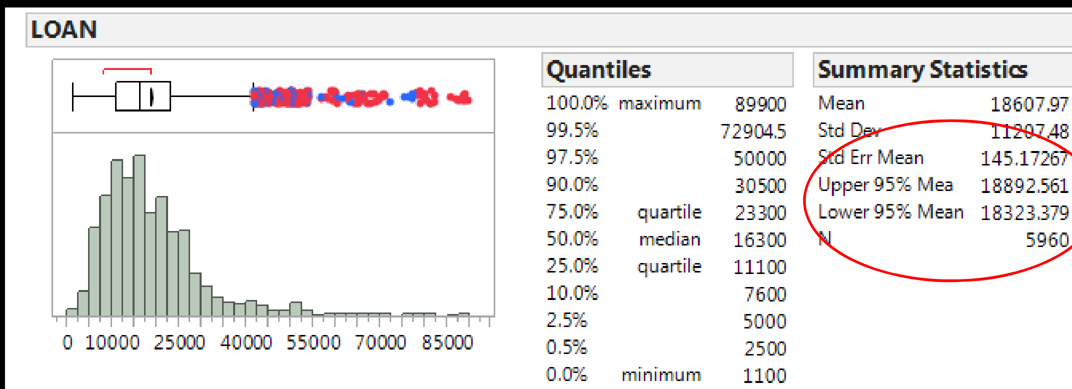


Tips:

- To change the display from vertical to horizontal (as shown), click on the **top red triangle** and select **Stack**.
- To change the confidence level, request a one-sided confidence limit or specify sigma, click on the **red triangle** for the variable, select **Confidence Interval**, and select the confidence level or click **Other**.

Confidence Interval

Source: Equity Data | Distribution Platform



Red Triangle: Confidence Interval | 0.95

$$(1-\alpha)\ \text{C.I. for the mean} = \bar{x} \pm \left(t_{1-\frac{\alpha}{2}} \cdot s_{\bar{y}}\right)$$

JMP Start Statistics, 4e (SAS Press)

One Variable

$\boxed{\text{Mean: Hypothesis Testing}}$

Mean: Confidence Interval

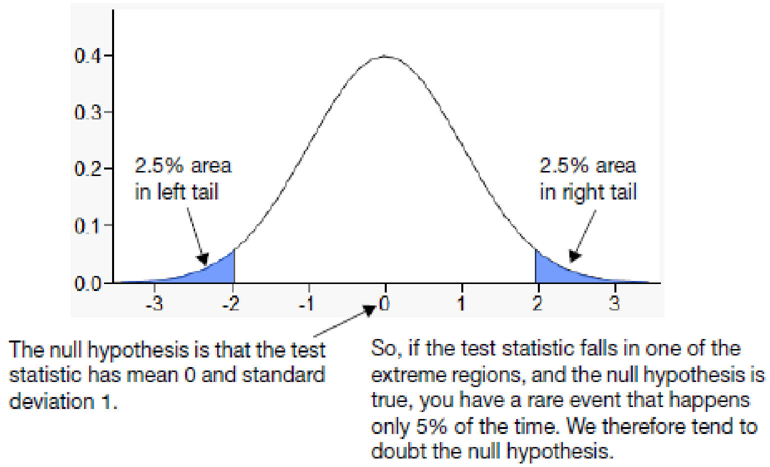
Hypothesis Test Principles

Every hypothesis test has these elements:

1. There is a null (Ho) and alternate (Ha) hypothesis – since the null cannot be proved we put what we hope to find in the alternate hypothesis. The null hypothesis is always a hypothesis of “no difference”
2. There is a test statistic (the particular form depends on our question and data) – this is needed because we need a way of deciding between our hypotheses
3. We construct (compute) the test statistic
4. We compare the value of the test statistic against a reference distribution (this reference distribution depends on our data type and our test statistic)
5. This comparison of the test statistic to the reference distribution gives us a probability (called a p-value) of seeing this value for the test statistic when the null hypothesis is indeed true
6. We compare this p-value against some predetermined threshold called alpha (or cut off) and if the p-value is less than alpha we state our conclusion of rejecting the null and accepting the alternate
7. Stated conclusions will be one of two statements:
   a. We reject the null and accept the alternate recognizing we have a “p-value” chance of being wrong in our conclusion – this is called a Type 1 error
   b. We fail to reject the null

Test Distribution

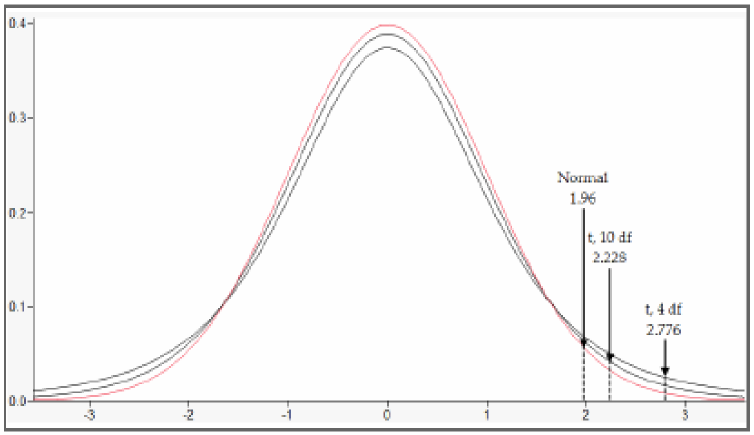
Figure 7.12 Illustration of the Two-Tailed *z*-test



$$\text{z-statistic} = \frac{\text{estimate} - \text{hypothesized estimate}}{\text{standard deviation}}$$

JMP Start Statistics, 4e (SAS Press)

Figure 7.14 Comparison of Normal and Student's *t* Distributions



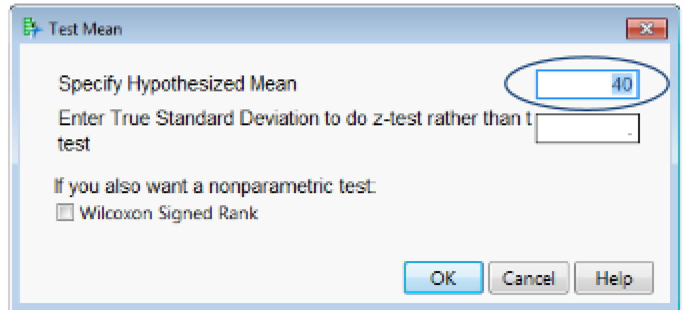
$$t\text{-statistic} = \frac{\text{sample mean} - \text{hypothesized value}}{\text{standard error of the mean}}$$

JMP Start Statistics, 4e (SAS Press)

One Sample t-Test and CI

One Sample t-Test for the Mean

1. From the Distributions report window (shown above), click on the **red triangle** for the variable and select **Test Mean**.
2. Enter the hypothesized value under **Specify Hypothesized Mean**, and click **OK**.



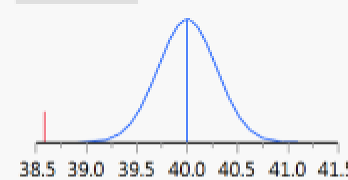
JMP will generate:

- The **t-Ratio** (next to **Test Statistic**).
- **P-values** for the two-tailed and one-tailed tests.
- A graph to aid in interpreting the p-values, showing the hypothesized mean (center of the curve) and the sample mean (red line).

Test Mean

Hypothesized Value	40
Actual Estimate	38.5862
DF	115
Std Dev	3.13208

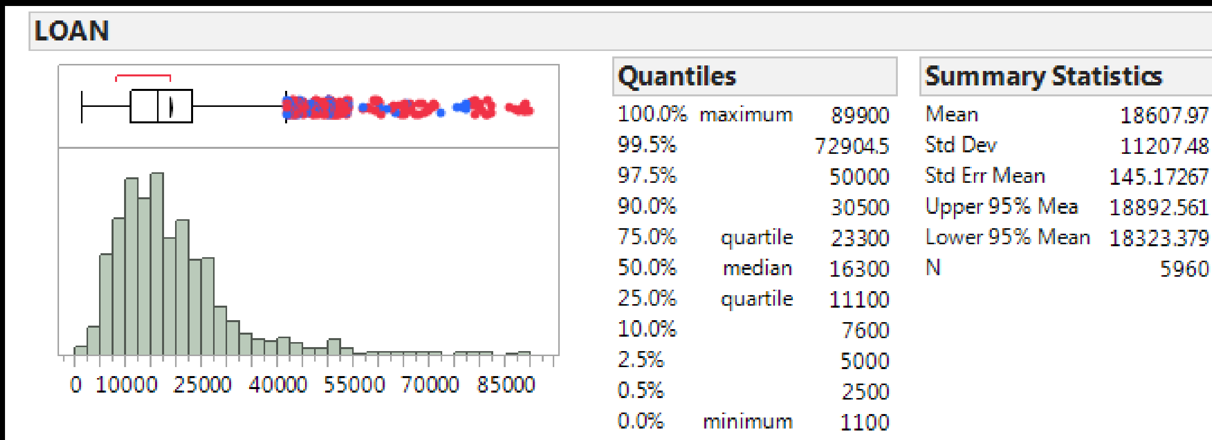
	t Test
Test Statistic	-4.8616
Prob > |t|	<.0001*
Prob > t	1.0000
Prob < t	<.0001*

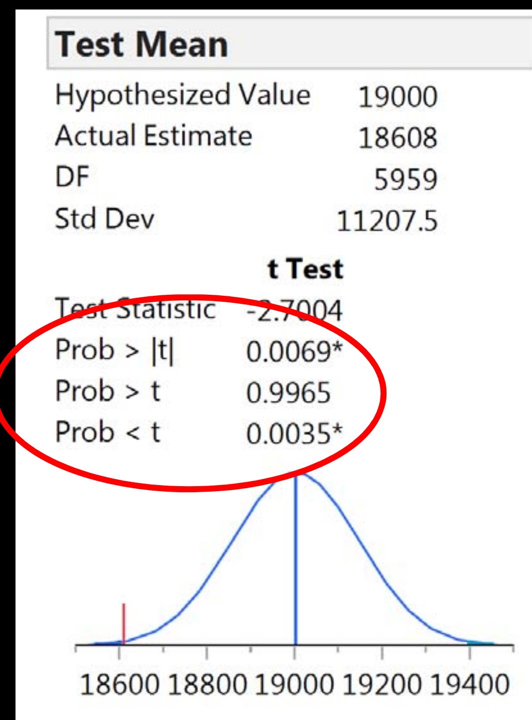


Interpretation of p-values for this example (using a significance level of 0.05):

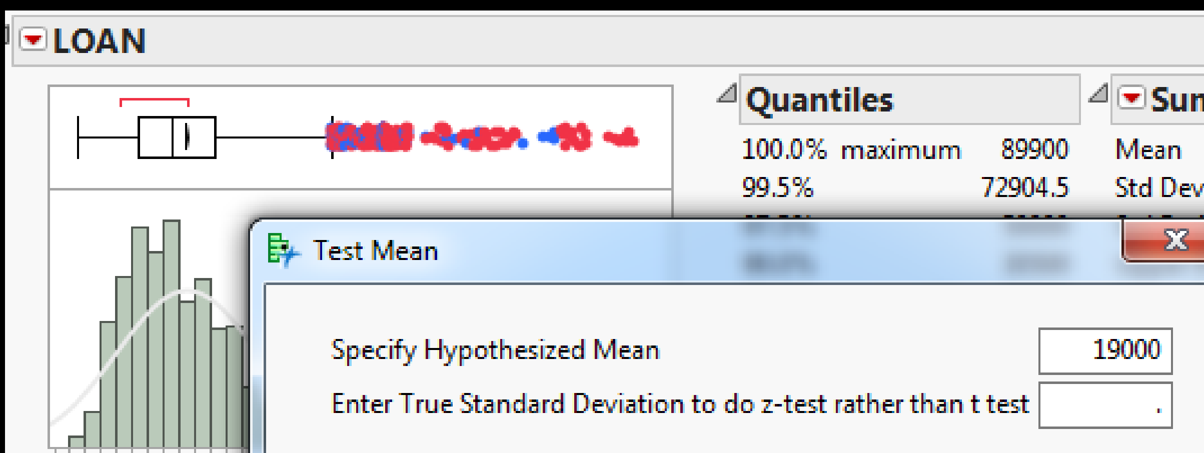
1. **Prob > |t| is less than 0.05 - reject the null hypothesis** that the true mean is 40. This is the two-tailed test. Conclude that the true mean is not 40.
2. **Prob > t is greater than 0.05 - fail to reject the null hypothesis** that the true mean is <= 40. This is a one-tailed test. There is insufficient evidence to reject the null hypothesis.
3. **Prob < t is less than 0.05 - reject the null hypothesis** that the true mean is >= 40. Conclude that the true mean is less than 40.

Testing Mean (HT)





Red Triangle: Test Mean | supply value



Insurance Profits

Data set: Insurance Profits_GE11 | Source: Business Statistics 3e

Profit
222.800
1756.23
1100.85
3340.66
1006.50
445.500
3255.60
3701.85
-803.35
3865.90
463.350
-66.200
57.9000
833.950
1390.70

2447.50
1847.50
865.400
1415.65
2756.94
2089.40
2692.75
2495.70
2172.70
3249.65
-397.10
-397.31
186.250
590.850
578.950

A manager wanted to see how well one of her sales representatives was doing, selected a sample of 30 policies and wants you to construct a 95% CI for the mean profit of the policies sold by this sales rep

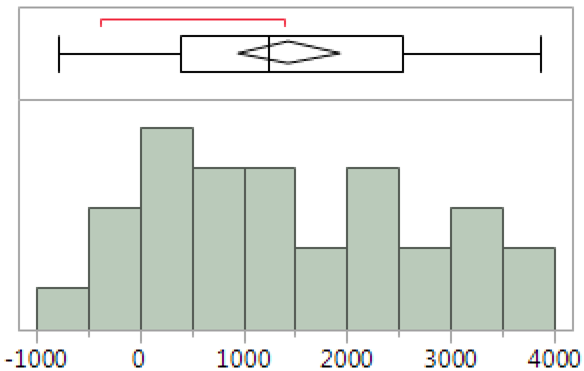
Insurance Profits

Data set: Insurance Profits_GE11 | Source: Business Statistics 3e

Profit	
222.800	2447.50
1756.23	1847.50
1100.85	865.400
3340.66	1415.65
1006.50	2756.94
445.500	2089.40
3255.60	2692.75
3701.85	2495.70
-803.35	2172.70
3865.90	3249.65
463.350	-397.10
-66.200	-397.31
57.9000	186.250
833.950	590.850
1390.70	578.950

Distributions

Profit



Quantiles

100.0%	maximum	3865.9
99.5%		3865.9
97.5%		3865.9
90.0%		3332.154
75.0%	quartile	2544.9625
50.0%	median	1245.775
25.0%	quartile	389.825
10.0%		-364.01
2.5%		-803.35
0.5%		-803.35
0.0%	minimum	-803.35

Summary Statistics

Mean	1438.9023
Std Dev	1329.5955
Std Err Mean	242.74981
Upper 95% Mean	1935.3814
Lower 95% Mean	942.42322
N	30

Insurance Profits

Data set: Insurance Profits_GE11 | Source: Business Statistics 3e

Profit
222.800
1756.23
1100.85
3340.66
1006.50
445.500
3255.60
3701.85
-803.35
3865.90
463.350
-66.200
57.9000
833.950
1390.70

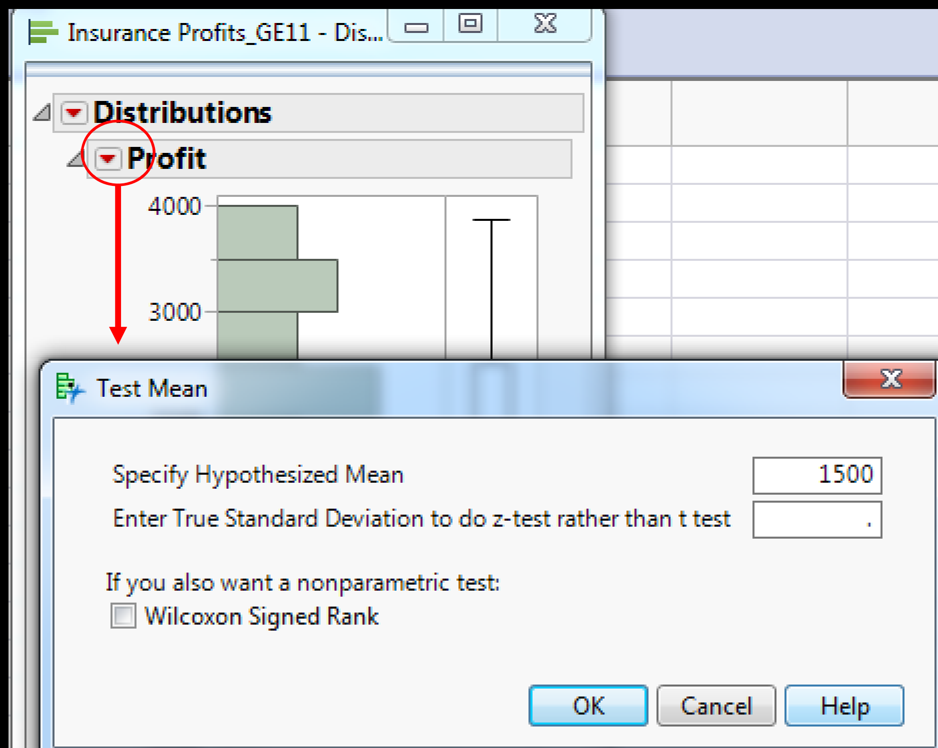
2447.50
1847.50
865.400
1415.65
2756.94
2089.40
2692.75
2495.70
2172.70
3249.65
-397.10
-397.31
186.250
590.850
578.950

A manager wanted to know if there's evidence that the mean profit of policies sold by this sales rep is less than $1500

Insurance Profits

Data set: Insurance Profits_GE11 | Source: Business Statistics 3e

Profit
222.800
1756.23
1100.85
3340.66
1006.50
445.500
3255.60
3701.85
-803.35
3865.90
463.350
-66.200
57.9000
833.950
1390.70
2447.50
1847.50
865.400
1415.65
2756.94
2089.40
2692.75
2495.70
2172.70
3249.65
-397.10
-397.31
186.250
590.850
578.950



UST1

Insurance Profits

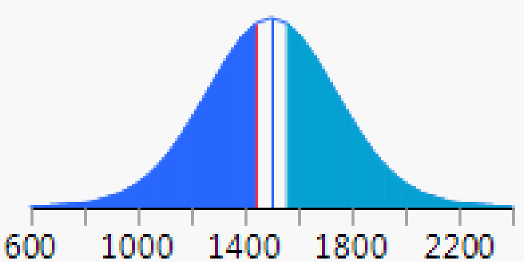
Data set: Insurance Profits_GE11 | Source: Business Statistics 3e

Profit	
222.800	2447.50
1756.23	1847.50
1100.85	865.400
3340.66	1415.65
1006.50	2756.94
445.500	2089.40
3255.60	2692.75
3701.85	2495.70
-803.35	2172.70
3865.90	3249.65
463.350	-397.10
-66.200	-397.31
57.9000	186.250
833.950	590.850
1390.70	578.950

Test Mean

Hypothesized Value	1500
Actual Estimate	1438.9
DF	29
Std Dev	1329.6

	t Test
Test Statistic	-0.2517
Prob > |t|	0.8031
Prob > t	0.5985
Prob < t	0.4015



There is not enough evidence in this sample of policies to indicate that the true mean is below $1500

Slide 19

UST1 User Support Test, 5/21/2015

Two Variable: Continuous Response and Categorical Factor

Difference Between Two Means

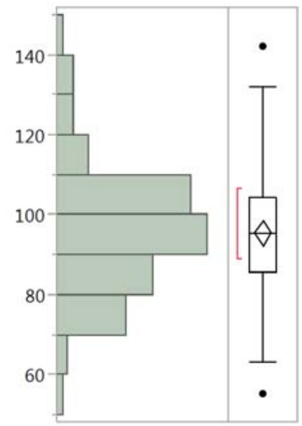
- Hypothesis Testing
- Confidence Interval

We would like to estimate the average difference in run times for men and women using the run10Samp data set, which was a simple random sample of 45 men and 55 women from all runners in the 2012 Cherry Blossom Run.

Data source: run10Samp (openintro stats, 2e)

Distributions

time



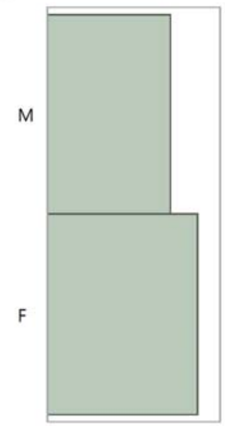
Quantiles

100.0%	maximum	141.94
99.5%		141.94
97.5%		132.061
90.0%		114.051
75.0%	quartile	104.205
50.0%	median	95.46
25.0%	quartile	85.555
10.0%		74.352
2.5%		65.9948
0.5%		55.31
0.0%	minimum	55.31

Summary Statistics

Mean	95.6146
Std Dev	15.776914
Std Err Mean	1.5776914
Upper 95% Mea	98.745082
Lower 95% Mean	92.484118
N	100

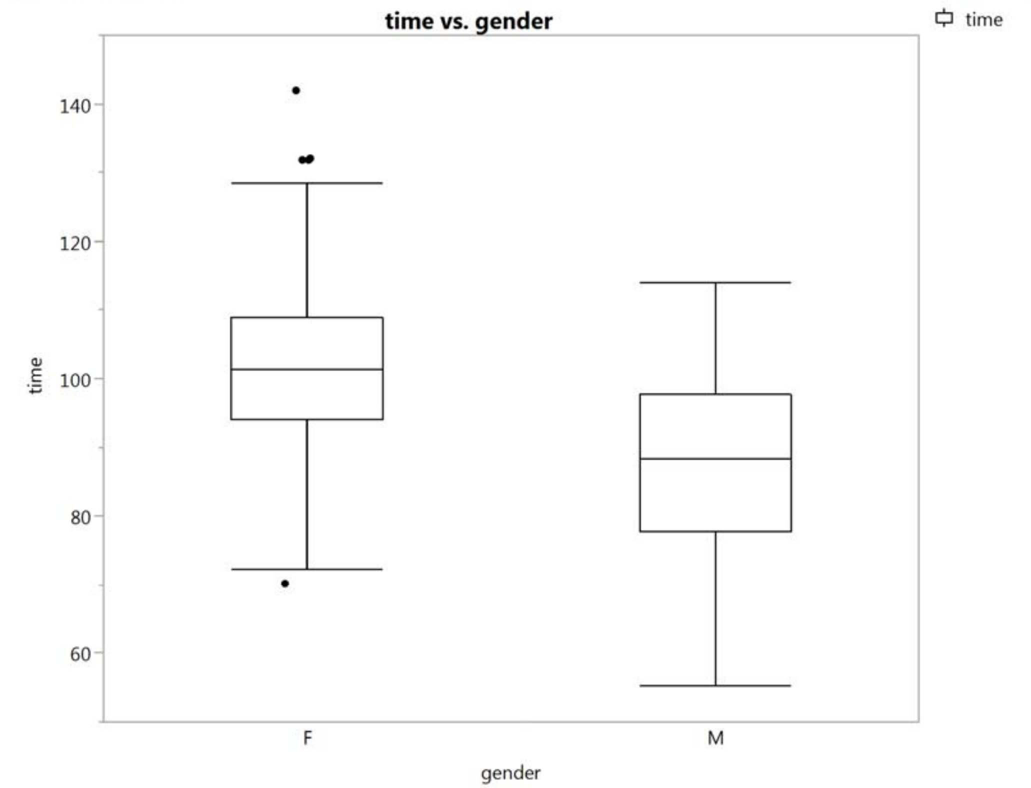
gender



Frequencies

Level	Count	Prob
F	55	0.55000
M	45	0.45000
Total	100	1.00000
N Missing	0	
2 Levels		

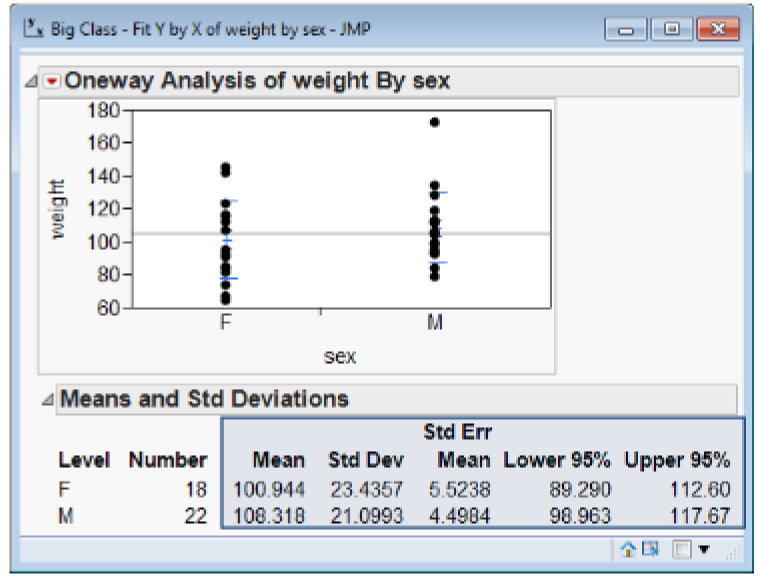
Graph Builder



Two Sample t-Test and CIs

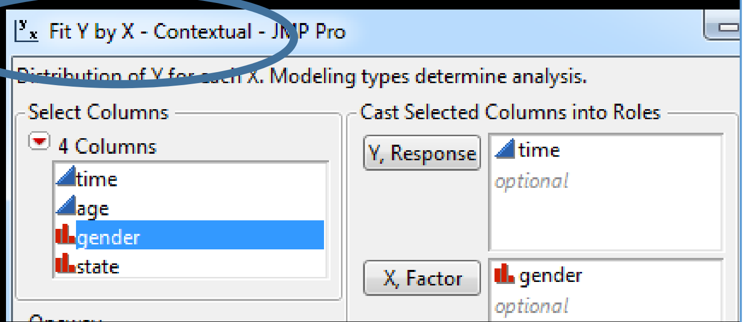
Confidence Intervals

Example: Big Class.jmp (Help > Sample Data)

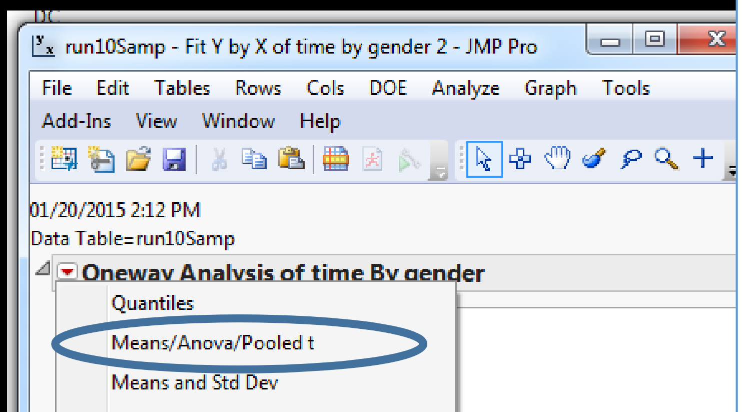


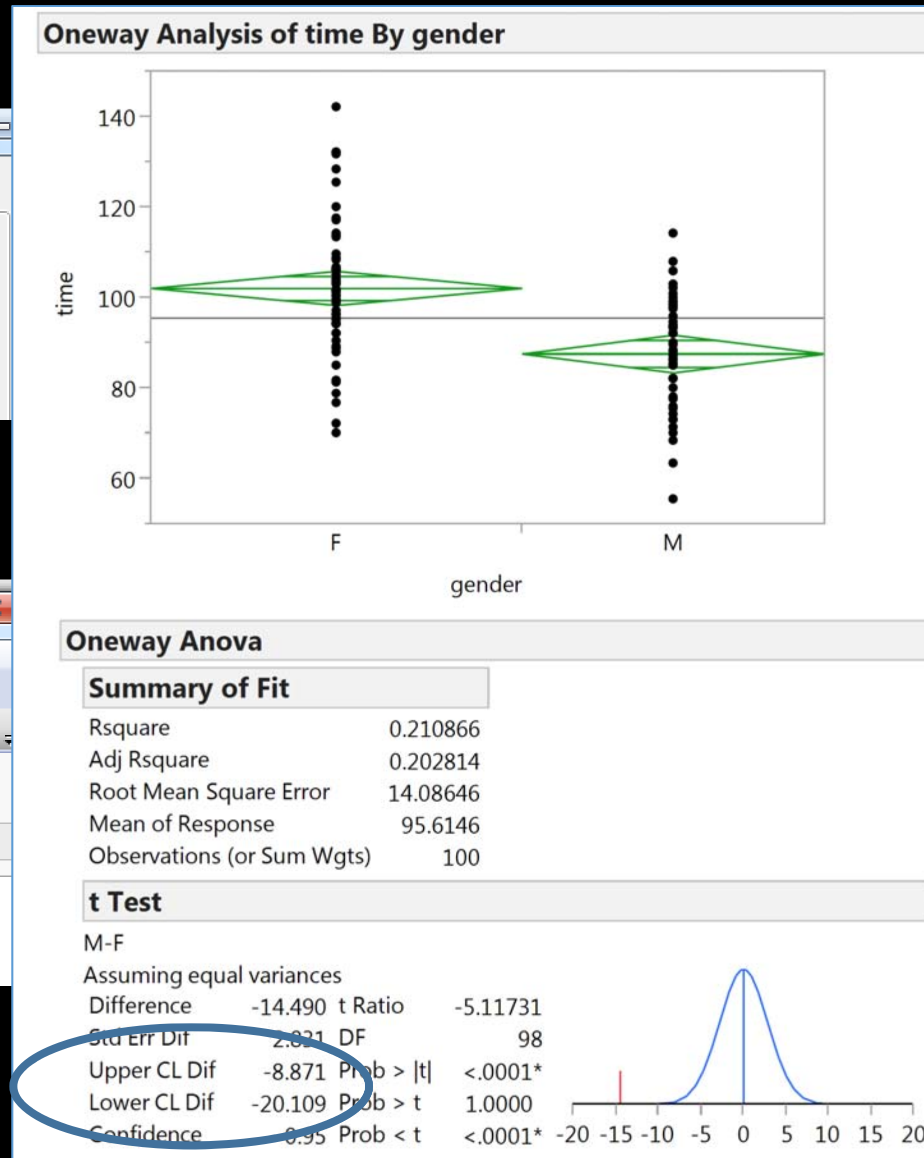
1. Select **Analyze > Fit Y by X**.
2. Click on a continuous variable from **Select Columns**, and click **Y, Response** (continuous variables have blue triangles).
3. Click on a two-level categorical variable and click **X, Factor** (categorical variables have red or green bars).
4. Click **OK**. The Oneway Analysis output window will display.
5. Click on the **red triangle**, and select **Means and Std Dev** to produce summary statistics and individual confidence intervals for each mean (Lower 95% and Upper 95%).

Analyze | Fit Y by X



Means/Anova/Pooled t





Two Variable: Continuous Response and Categorical Factor

Difference Between Two Means

- Hypothesis Testing
- Confidence Interval

Two Sample t-Test and CIs

Two Sample t-Test

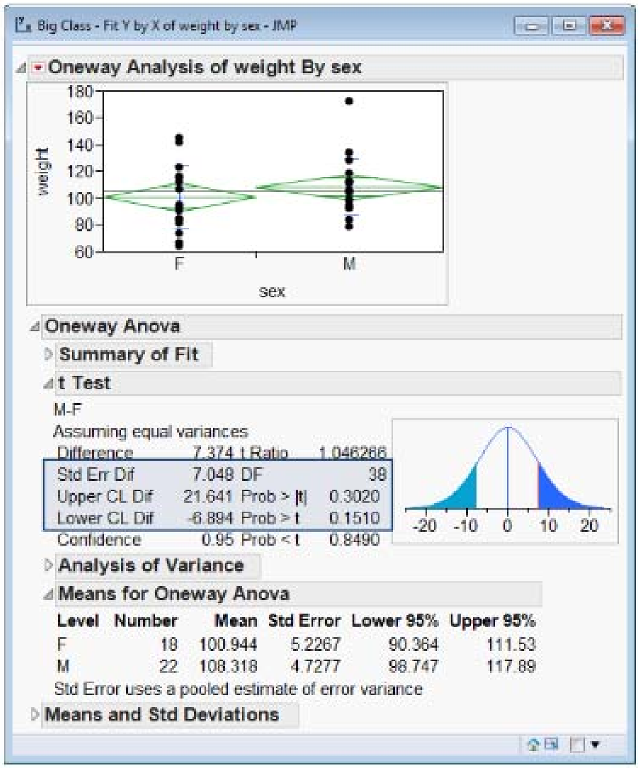
From the Oneway Analysis output window (shown above), click on the **red triangle** and select **Means/Anova/Pooled t**.

JMP® will plot **means diamonds (95% confidence intervals for each mean)**, and will generate:

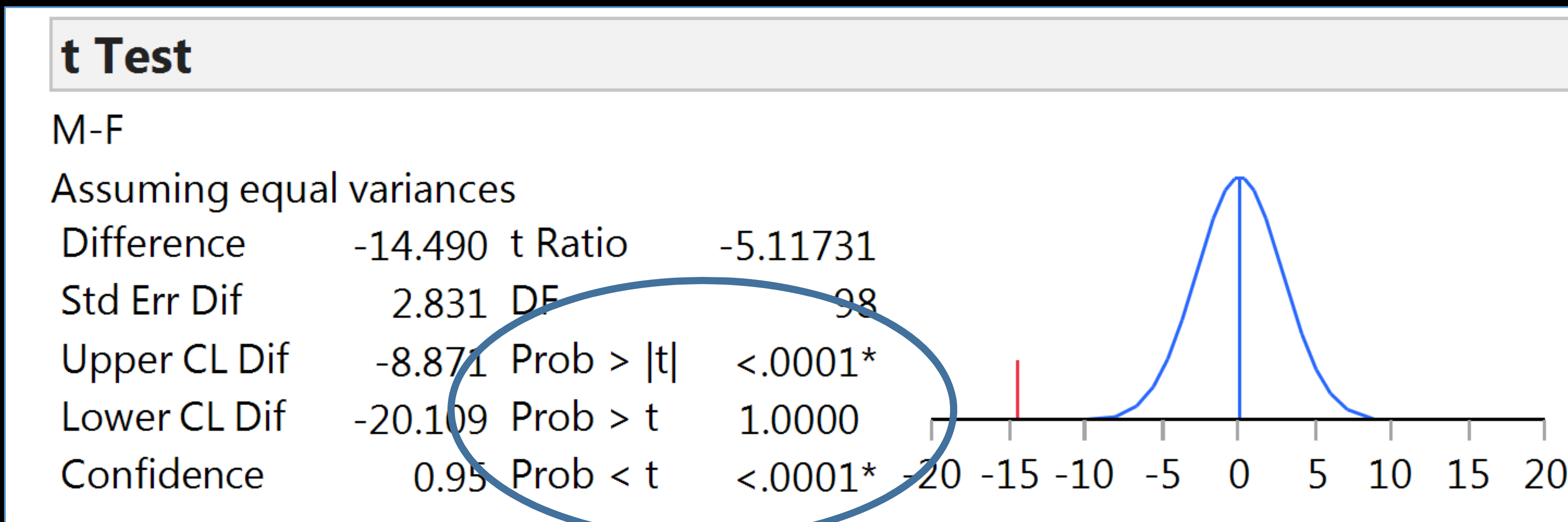
- The Summary of Fit (not shown).
- The t-test report, with a graph to aid in interpreting the results.
- The Analysis of Variance table.
- Means for Oneway Anova (not shown), which includes confidence intervals based on the pooled estimate of the standard error.

Interpretation of the results (using a significance level of 0.05 - click **the red triangle, Set α Level** to change significance level):

1. **Upper CL Dif** and **Lower CL Dif** give the 95% CI for the true difference. Since the **95% CI contains zero, conclude that there is not a significant difference** between the means.
2. **Prob > |t|** is the p-value for the two-tailed test. The null hypothesis is that means are equal (the mean difference is zero). Since the **Prob > |t| is greater than 0.05, cannot reject the null hypothesis** (i.e., we cannot conclude that there is a significant difference).



Notes: **Means/Anova/Pooled t** is the test under the assumption of equal variances. For a test without the assumption of equal variances, select **t Test** instead. See the ***Basic Analysis*** book (under **Help > Books**) for more details.



t Test

M-F

Assuming equal variances

Difference	-14.490	t Ratio	-5.11731
Std Err Dif	2.831	DF	98
Upper CL Dif	-8.871	Prob > |t|	<.0001*
Lower CL Dif	-20.109	Prob > t	1.0000
Confidence	0.95	Prob < t	<.0001*

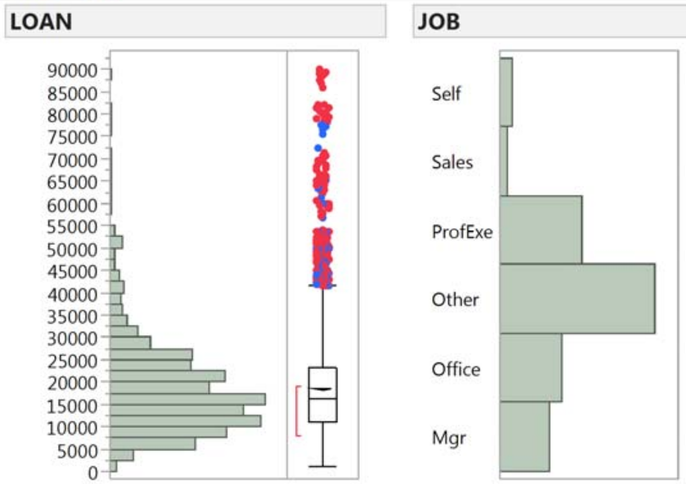
Multiple Variable: Continuous Response with Categorical Variables

ANOVA

F Test

Source: Equity Data

Distributions



Quantiles

100.0%	maximum	89900
99.5%		72904.5
97.5%		50000
90.0%		30500
75.0%	quartile	23300
50.0%	median	16300
25.0%	quartile	11100
10.0%		7600
2.5%		5000
0.5%		2500
0.0%	minimum	1100

Summary Statistics

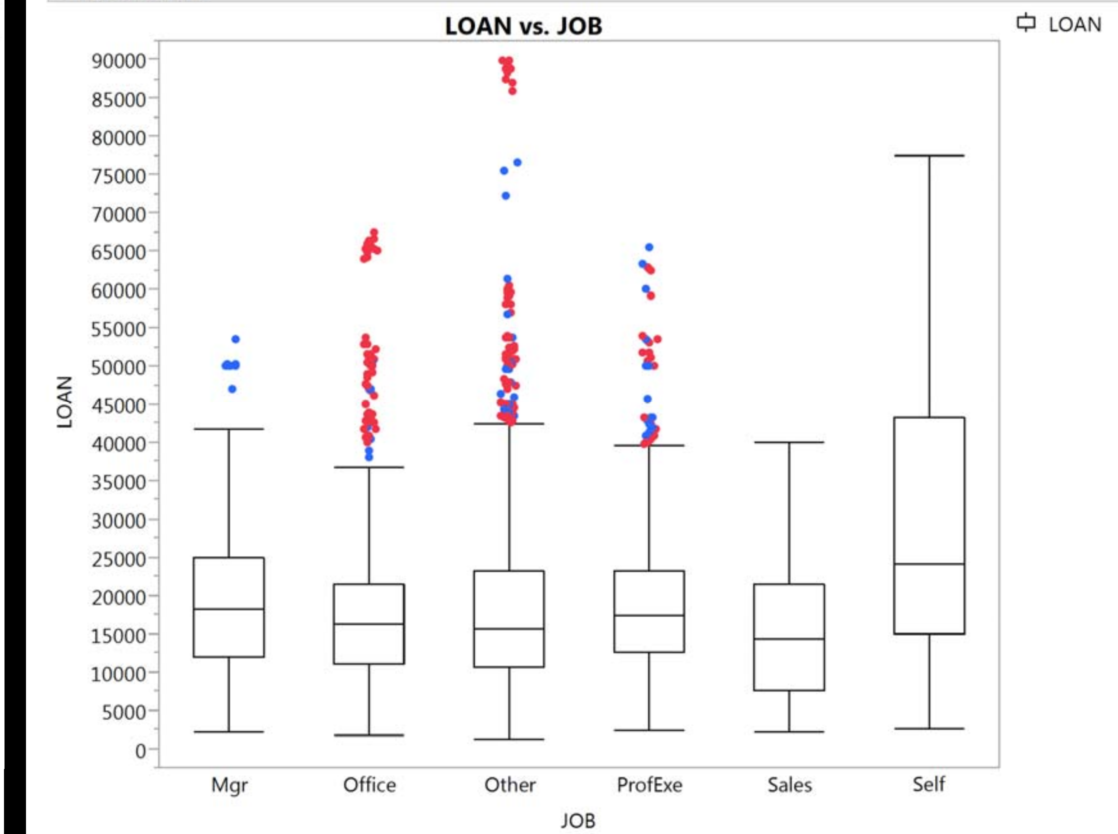
Mean	18607.97
Std Dev	11207.48
Std Err Mean	145.17267
Upper 95% Mea	18892.561
Lower 95% Mean	18323.379
N	5960

Frequencies

Level	Count	Prob
Mgr	767	0.13501
Office	948	0.16687
Other	2388	0.42035
ProfExe	1276	0.22461
Sales	109	0.01919
Self	193	0.03397
Total	5681	1.00000

N Missing 279
6 Levels

Graph Builder



One-Way ANOVA

One-Way Analysis of Variance

1. Select **Analyze > Fit Y by X**.
2. Click on a continuous variable from **Select Columns**, and Click **Y, Response** (continuous variables have blue triangles).
3. Click on a categorical variable and click **X, Factor** (categorical variables have red or green bars).
4. Click **OK**. The Oneway Analysis output window will display.
5. Click on the **red triangle**, and select **Means/Anova**.

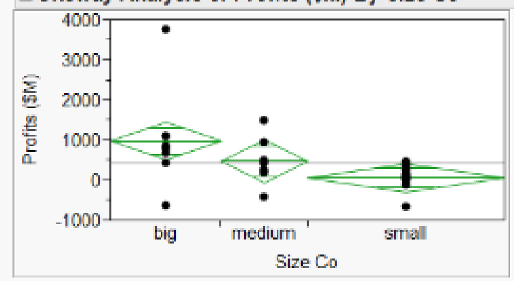
JMP® will plot means diamonds (95% confidence intervals for each mean), and will generate:

- The Summary of Fit.
- The Analysis of Variance (Anova) table.
- Means for Oneway Anova, containing summary statistics and confidence intervals for each mean (based on the pooled estimate of the standard error).

Example: Companies.jmp (Help > Sample Data)

Companies - Fit Y by X of Profits ($M) by Size Co - JMP

Oneway Analysis of Profits ($M) By Size Co



Oneway Anova

Summary of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Size Co	2	4938838	2469419	5.1389	0.0123*
Error	29	13935447	480533		
C. Total	31	18874286			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
big	9	981.311	231.07	508.7	1453.9
medium	7	474.771	262.01	-61.1	1010.6
small	16	58.938	173.30	-295.5	413.4

Std Error uses a pooled estimate of error variance

Interpretation of the results in the Anova table (using a significance level of 0.05 – click **the red triangle, Set α Level** to change significance level):

- The null hypothesis is that there are no differences between the population means (i.e., all means are equal).
- **Prob > F** is the p-value for the whole model test. Since the **Prob > F is less than than 0.05, reject the null hypothesis.** Conclude that there are differences between at least two of the means.
- To determine which means are different, a post hoc multiple comparison technique can be used.

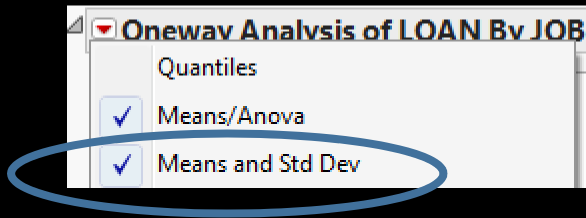
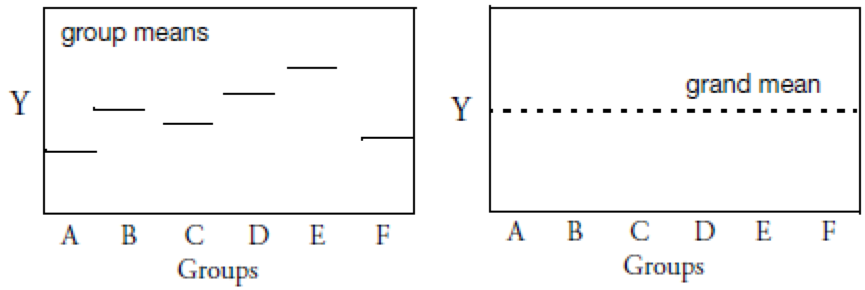


Figure 9.1 Different Mean for Each Group Versus a Single Overall Mean

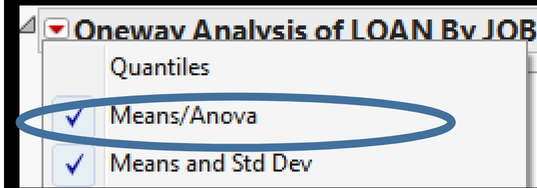


JMP Start Statistics, 4e (SAS Press)

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Mgr	767	19155.3	389.2	18392	19918
Office	948	18142.6	350.1	17456	18829
Other	2388	18061.7	220.6	17629	18494
ProfExe	1276	18983.5	301.7	18392	19575
Sales	109	14913.8	1032.4	12890	16938
Self	193	28314.5	775.9	26793	29836

Std Error uses a pooled estimate of error variance



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
JOB	5	2.093e+10	4.1861e+9	36.0300	<.0001*
Error	5675	6.5934e+11	116183575		
C. Total	5680	6.8027e+11			

H_0: The mean outcome is the same across all groups. In statistical notation, $\mu_1 = \mu_2 = \cdots = \mu_k$ where μ_i represents the mean of the outcome for observations in category i.

H_A: At least one mean is different.

OpenIntro Statistics, 2e

OpenIntro Statistics, 2e

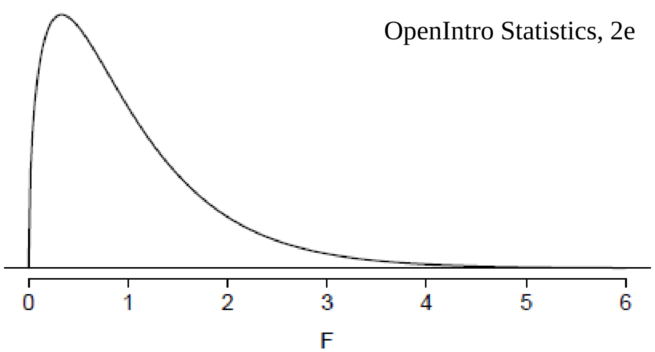
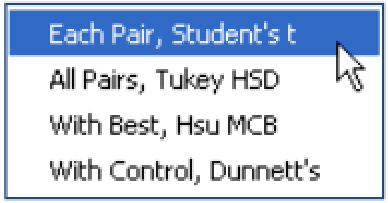


Figure 5.29: An F distribution with $df_1 = 3$ and $df_2 = 323$.

One-Way ANOVA

Multiple Comparison Procedures

From the Oneway Analysis output window (shown above), click on the **red triangle**, select **Compare Means**, and select one of the four methods (described in JMP Help).

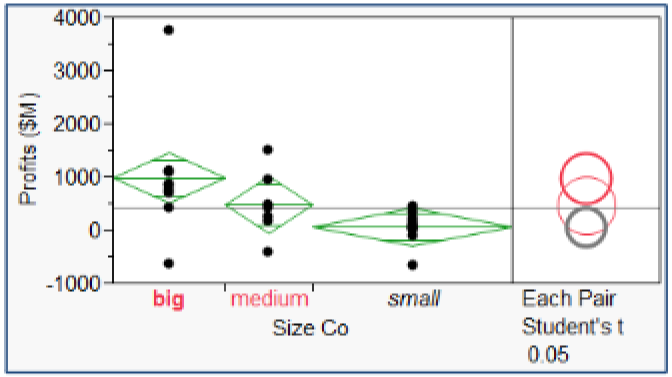


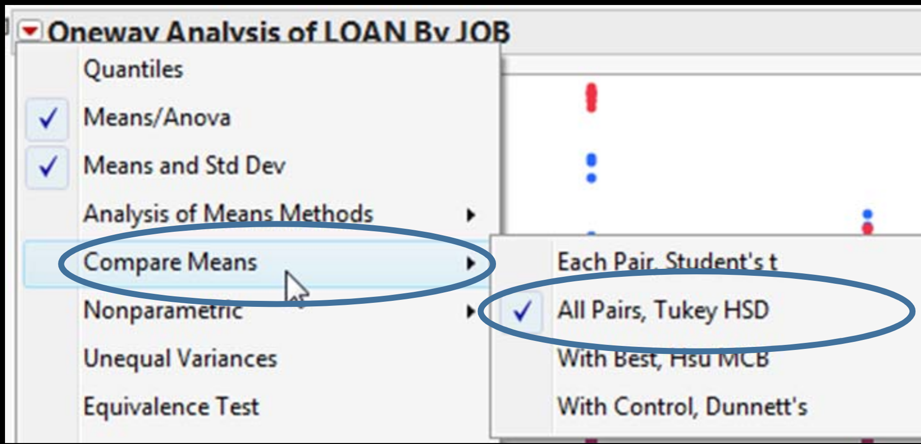
Each Pair, Student's t has been selected. This produces comparison circles (shown), along with statistical output (not shown).

Click on a circle for a mean to test for paired differences.

- The **selected mean** will have a **bold, red circle and variable label.**
- Means that **are not significantly different** from the selected mean will have **unbolded, red circles and variable labels.**
- Means that **are significantly different** from the selected mean will have **gray circles** and **gray italicized variable labels.**

In this example, the mean for **big** is significantly different from the mean for **small**, but is not significantly different from the mean for **medium**.





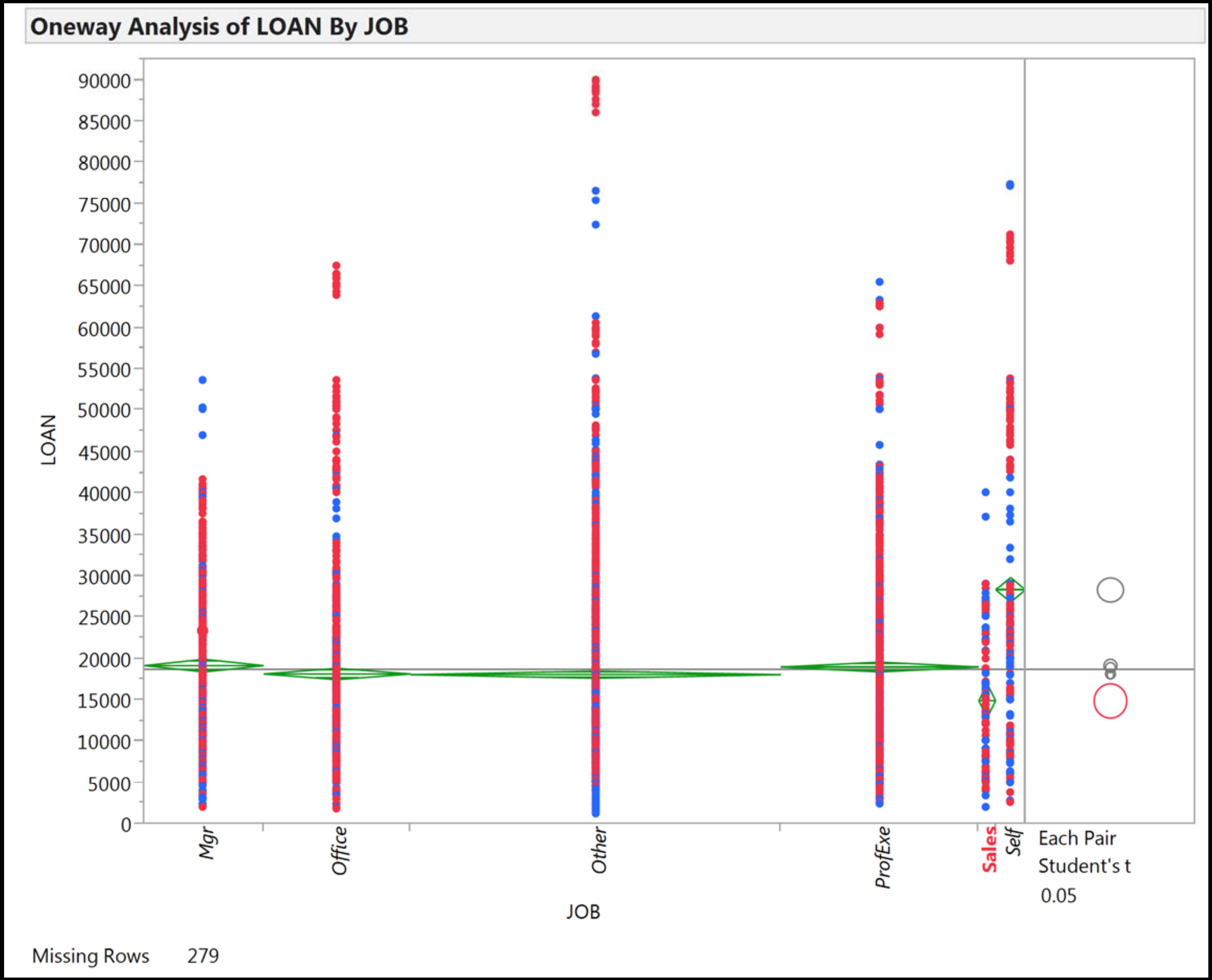
Connecting Letters Report

Level				Mean
Self	A			28314.508
Mgr		B		19155.280
ProfExe		B		18983.464
Office		B		18142.616
Other		B		18061.683
Sales			C	14913.761

Levels not connected by same letter are significantly different.

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
Self	Sales	13400.75	1291.469	9719.13	17082.37	<.0001*
Self	Other	10252.82	806.623	7953.37	12552.28	<.0001*
Self	Office	10171.89	851.201	7745.35	12598.43	<.0001*
Self	ProfExe	9331.04	832.490	6957.84	11704.24	<.0001*
Self	Mgr	9159.23	868.024	6684.73	11633.72	<.0001*
Mgr	Sales	4241.52	1103.350	1096.17	7386.86	0.0017*
ProfExe	Sales	4069.70	1075.620	1003.41	7135.99	0.0022*
Office	Sales	3228.85	1090.166	121.10	6336.61	0.0363*
Other	Sales	3147.92	1055.726	138.34	6157.50	0.0342*
Mgr	Other	1093.60	447.360	-181.70	2368.90	0.1413
Mgr	Office	1012.66	523.483	-479.64	2504.97	0.3810
ProfExe	Other	921.78	373.773	-143.74	1987.30	0.1344
ProfExe	Office	840.85	462.179	-476.70	2158.39	0.4532
Mgr	ProfExe	171.82	492.474	-1232.09	1575.72	0.9993
Office	Other	80.93	413.775	-1098.62	1260.49	1.0000



Oneway Analysis of LOAN By JOB
LOAN
90000
85000
80000
75000
70000
65000
60000
55000
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
Mgr
Office
Other
ProfExe
Sales
Self
JOB
Each Pair
Student's t
0.05
Missing Rows 279

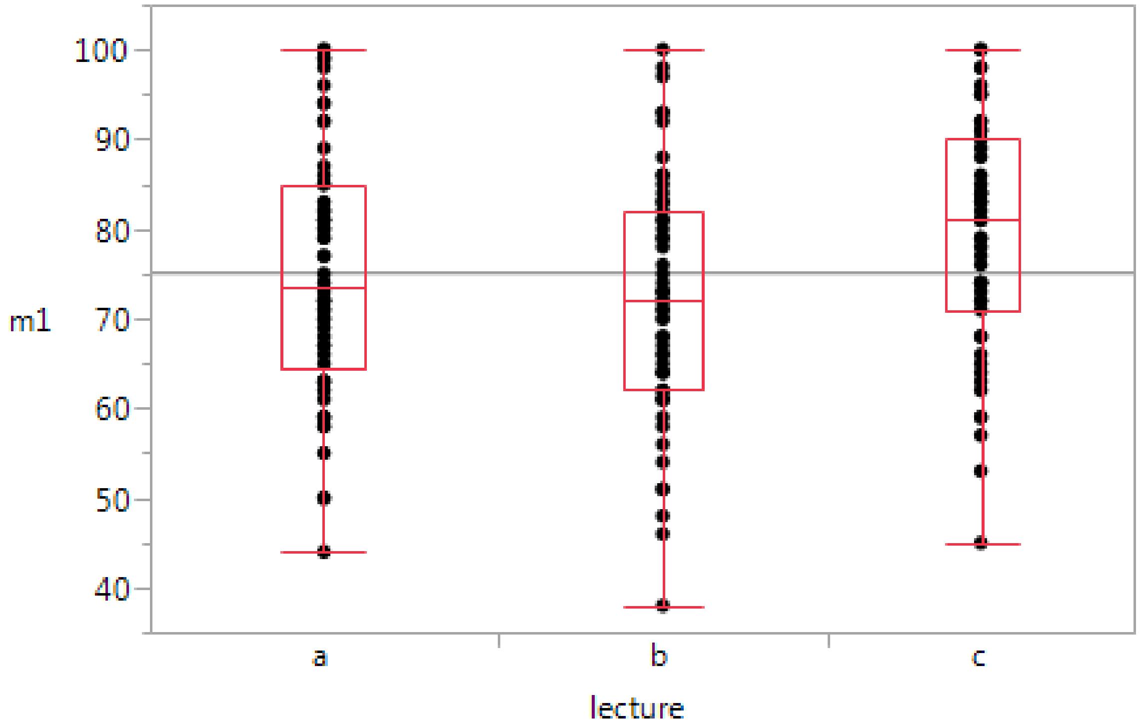
College departments commonly run multiple lectures of the same introductory course each semester because of high demand. Consider a statistics department that runs three lectures of an introductory statistics course. We might like to determine whether there are statistically significant differences in first exam scores in these three classes (A, B, and C). Describe appropriate hypotheses to determine whether there are any differences between the three classes.

The hypotheses may be written in the following form:

H0: The average score is identical in all lectures. Any observed difference is due to chance.
HA: The average score varies by class. We would reject the null hypothesis in favor of the alternative hypothesis if there were larger differences among the class averages than what we might expect from chance alone.

Example 5.34, Open Intro Statistics | Data set: classData

Oneway Analysis of m1 By lecture



Quantiles

Level	Minimum	10%	25%	Median	75%	90%	Maximum
a	44	58	64.5	73.5	85	96.2	100
b	38	52.8	62	72	82	92.4	100
c	45	59.6	71	81	90	95.8	100

Oneway Anova

Summary of Fit

Rsquare	0.041482
Adj Rsquare	0.029575
Root Mean Square Error	13.60721
Mean of Response	75.2439
Observations (or Sum Wgts)	164

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
lecture	2	1290.114	645.057	3.4839	0.0330*
Error	161	29810.130	185.156		
C. Total	163	31100.244			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
a	58	75.1034	1.7867	71.575	78.632
b	55	71.9636	1.8348	68.340	75.587
c	51	78.9412	1.9054	75.178	82.704

Std Error uses a pooled estimate of error variance

Means Comparisons

Comparisons for all pairs using Tukey-Kramer HSD

Confidence Quantile

q*	Alpha
2.36563	0.05

HSD Threshold Matrix

Abs(Dif)-HSD

	c	a	b
c	-6.3745	-2.3414	0.7200
a	-2.3414	-5.9775	-2.9186
b	0.7200	-2.9186	-6.1383

Positive values show pairs of means that are significantly different.

Connecting Letters Report

Level			Mean
c	A		78.941176
a	A	B	75.103448
b		B	71.963636

Levels not connected by same letter are significantly different.

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
c	b	6.977540	2.645182	0.72001	13.23507	0.0247*
c	a	3.837728	2.612060	-2.34145	10.01690	0.3084
a	b	3.139812	2.561019	-2.91862	9.19824	0.4396