CS 700 Reference Sheet

Kevin Andrea

March 25, 2015

Contents

1 Population Statistics

1.1 Fundamental Population Statistics

Mean (μ) AVERAGE(START:END)

SummarizingData [3]

$$= AVERAGE(A1 : A1001)$$

Variance (s^2) VAR.S(START:END)

SummarizingData [14]

$$= VAR(A1 : A1001)$$
$$= VAR.S(A1 : A1001)$$

This is not equal to VAR.P(A1:A1001). Only use VAR or VAR.S

Standard Deviation (s) STDEV.S(START:END)

SummarizingData [14]

$$= STDEV(A1 : A1001)$$
$$= STDEV.S(A1 : A1001)$$

This is not equal to STDEV.P(A1:A1001). Only use STDEV or STDEV.S

Coefficient of Variation (COV)

s/μ SummarizingData [16]

1.2 Confidence Interval - Mean [CONFIDENCE.NORM(α,s,n)]

At a 90% Confidence Level for a sample of more than 30.

SummarizingData [24]
EstimationProcedures [13]

α	0.1	For 90% Confidence Level
$1-\alpha$	0.9	
$1-\frac{\alpha}{2}$	0.95	
NORMSINV($1-\frac{\alpha}{2}$)	1.64485	NORM.S.INV($1-\frac{\alpha}{2}$) is equivalent
$s * \frac{NORMSINV(1-\frac{\alpha}{2})}{SQRT(n)}$	$\frac{1}{2}$CI	Half of the Confidence Interval

Is the population mean within the 90% CI?

So, the confidence interval for the mean is:

$$Lower = \mu - \frac{1}{2}CI$$
$$Upper = \mu + \frac{1}{2}CI$$

Alternatively, the following works in Excel, and are equivalent

$$CONFIDENCE(\alpha, s, n) = CONFIDENCE(0.1, 0.175, 50)$$
$$CONFIDENCE.NORM(\alpha, s, n) = CONFIDENCE.NORM(0.1, 0.175, 50)$$

1.3 Confidence Interval - Mean [CONFIDENCE.T(α,s,n)]

At a 90% Confidence Level for a sample of less than 30.

EstimationProcedures [15]

α	0.1	For 90% Confidence Level
$1-\alpha$	0.9	
$1-\frac{\alpha}{2}$	0.95	
$n-1$	DOF	Degrees of Freedom
TINV(α,DOF)	1.72913	T.INV($1-\frac{\alpha}{2}$,DOF) is equivalent
$s * \frac{TINV(\alpha,DOF)}{SQRT(n)}$	$\frac{1}{2}$CI	Half of the Confidence Interval

$t_{[1-\alpha/2;n-1]}$

Is the population mean within the 90% CI?

So, the confidence interval for the mean is:

$$Lower = \mu - \frac{1}{2}CI$$
$$Upper = \mu + \frac{1}{2}CI$$

Alternatively, the following works in Excel, and are equivalent

$$CONFIDENCE.T(\alpha, s, n) = CONFIDENCE.T(0.1, 0.175, 20)$$

1.4 Confidence Interval - Variance

EstimationProcedures [18]

At a 90% Confidence Level for a sample.

α	0.1	For 90% Confidence Level
$n-1$	DOF	Degrees of Freedom
CHIINV($1-\frac{\alpha}{2}$,DOF)	Lower Crit Value	CHISQ.INV($\frac{\alpha}{2}$.DOF)
CHIINV($\frac{\alpha}{2}$,DOF)	Upper Crit Value	CHISQ.INV.RT($\frac{\alpha}{2}$,DOF)

Is the variance within the 90% CI?

So, the confidence interval for the mean is:

$$LowerBoundCI = DOF * \frac{s^2}{UpperCritValue}CI$$

$$UpperBoundCI = DOF * \frac{s^2}{LowerCritValue}CI$$

1.5 Confidence Interval - Future Values

EstimationProcedures [22]

$$\bar{X} - t_{[1-\alpha/2;n-1]}s\sqrt{1+1/n} <= X_f <= \bar{X} + t_{[1-\alpha/2;n-1]}s\sqrt{1+1/n}$$

At a 90% Confidence Level for a sample.

$t_{[1-\alpha/2;n-1]}$

α	0.1	For 90% Confidence Level
$1-\alpha$	0.9	
$1-\frac{\alpha}{2}$	0.95	
$n-1$	DOF	Degrees of Freedom
TINV(α,DOF)	1.72913	T.INV($1-\frac{\alpha}{2}$,DOF) is equivalent

Obtain a 90% CI for a future value from the same population.

So, the confidence interval for a future value is:

$$LowerBoundCI = \mu - TINV(\alpha, DOF) * s * SQRT(1+\frac{1}{n})$$

$$UpperBoundCI = \mu + TINV(\alpha, DOF) * s * SQRT(1+\frac{1}{n})$$

1.6 Confidence Interval - Proportions (π)

EstimationProcedures [23]

The sample proportion(p) is determined from a random selection of n members of the population. The standard normal distribution is used to determine the confidence interval.

$$(p - z_{[1-\alpha/2]}\sqrt{\frac{p(1-p)}{n}}, p + z_{[1-\alpha/2]}\sqrt{\frac{p(1-p)}{n}})$$

1000 entries are randomly selected. 650 are type A. Find the 95% confidence level for the proportion that are type A.

At a 90% Confidence Level for a sample.

α	0.1	For 90% Confidence Level
$1-\alpha$	0.9	
$1-\frac{\alpha}{2}$	0.95	
NORMSINV(1 - α/2)	1.64	NORM.S.INV is equivalent

So, the confidence interval for a future value is:

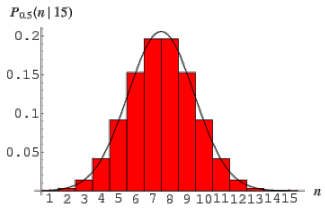
$$LowerBoundCI = p - NORMSINV(1-\alpha/2) * SQRT(\frac{p(1-p)}{n})$$

$$UpperBoundCI = p + NORMSINV(1-\alpha/2) * SQRT(\frac{p(1-p)}{n})$$

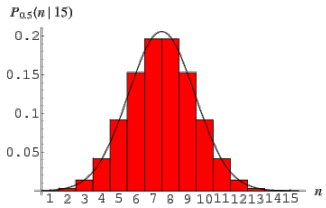
2 Comparing Datasets

Probability... [*]
Continuous... [*]

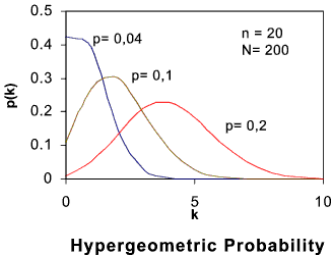
2.1 Create a Histogram and Visually Assess



(a) Normal

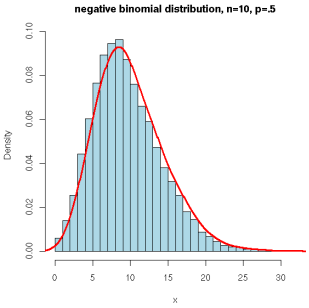


(b) Binomial

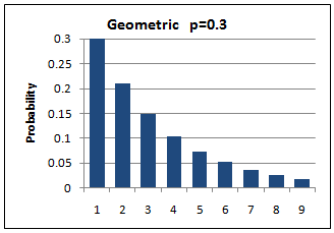


(c) Hyper Geometric

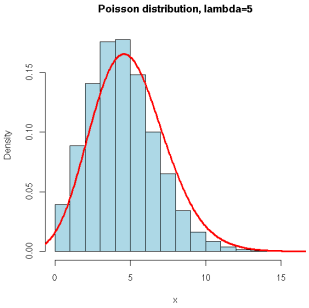
Do these two data sets come from the same distribution?



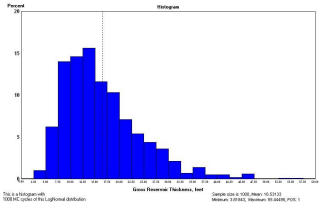
(d) Negative Binomial



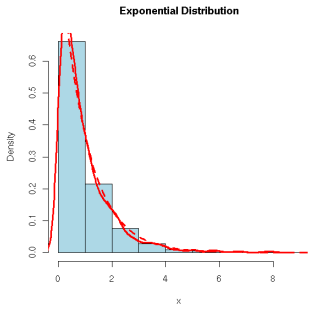
(e) Geometric



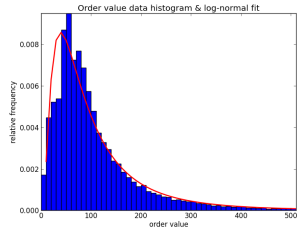
(f) Poisson



(g) Lognormal



(h) Exponential



(i) Pareto

2.2 Scatter Plot

Plot the data against each other to look for groupings. Fitting Dist... [6]

Not really a good tool. Just find a matching histogram and skip to data analysis – or Q-Q plot.

2.3 Q-Q Plot

This is a plot of $(Q_1(p),Q_2(p))$. $Q_1(p)$ is x_i of the first sorted dataset. Fitting Dist... [8]

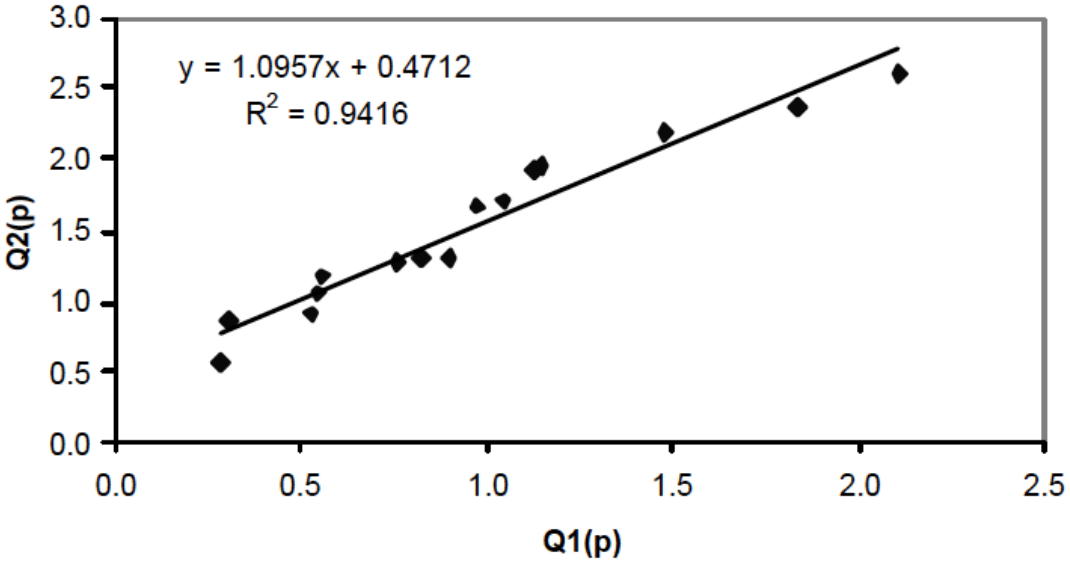


Figure 1: Q-Q Plot

Look for an R^2 value close to 1. This indicates both are from the same dataset.

2.4 Theoretical Q-Q

Plot $Q_1(x_i)$ against $Q_2(\frac{[i-0.5]}{n})$ to compare against a possible distribution, where x_i is the $\frac{[i-0.5]}{n}$th quantile of the theoretical distribution ($F^{-1}(\frac{[i-0.5]}{n})$) and $Q_2([\frac{[i-0.5]}{n}])$ is the ith ordered data point. Fitting Dist... [12]

Look for an R^2 value if 1 to confirm the distributions are the same.

1. Sort the data column in question to get Q_1
2. Create a column (i) of values from [1..n]
3. Create a column (p) of the values: $\frac{i-0.5}{n}$
4. Create a column (Q_2) with the F^{-1} values using p
5. Plot Q_1 vs Q_2 and look for R^2 close to 1

2.4.1 Inverse Function Reference

Standard Normal Inverse
$NORMSINV(i-0.5/n) = NORM.S.INV(i-0.5/n)$

LogNormal Inverse
$LOGNORM.INV(i-0.5/n, \mu, s)$
μ and s don't matter, since the distribution is the same. Use 1,1

Exponential Inverse
$-LN(1-p)/lambda$
Lambda doesn't matter, since the distribution is the same. Use 0.5

Pareto Inverse
$1/(1-p)^{1/a}$
a doesn't matter, since the distribution is the same. Use 0.5

Geometric Inverse
$LN(u)/LN(1-p)$

3 Comparing Alternatives

3.1 Paired Observations

Estimation... [29]

Workload	Cache Hit Ratio		
	Policy A	Policy B	A-B
1	0.35	0.28	0.07
2	0.46	0.37	0.09
3	0.29	0.34	-0.05
4	0.54	0.60	-0.06
5	0.32	0.22	0.10
6	0.15	0.18	-0.03
	Sample mean		0.02000
	Sample variance		0.00552
	Sample standard dev.		0.07430

Which policy is better?

Can we say with 90% confidence that A is better than B?

Figure 2: Paired Observations

1. Find the difference in measured values between the two items being compared, for the same input.
2. Compute $\bar{X}$ and s.
3. Compute the Confidence Interval at the desired percentage.
$$(\bar{X} - t_{[1-\alpha/2;n-1]}\frac{s}{\sqrt{n}}, \bar{X} + t_{[1-\alpha/2;n-1]}\frac{s}{\sqrt{n}})$$
For the above example, this would be, in Excel, for 90%...

Remember $t_{[1-\alpha/2;n-1]} = TINV(\alpha, n-1) = T.INV(1-\alpha/2, n-1)$

$$(AVERAGE(C1:C6) - TINV(0.1, 6-1) * STDEV(B1:B6)/SQRT(6),$$
$$AVERAGE(C1:C6) + TINV(0.1, 6-1) * STDEV(B1:B6)/SQRT(6))$$

If the resulting interval includes 0, then neither is statistically better than the other.

3.2 Unpaired Observations

Estimation... [33]

Workload	Cache Hit Ratio	
	Policy A	Policy B
1	0.35	0.49
2	0.23	0.33
3	0.29	0.33
4	0.21	0.55
5	0.21	0.65
6	0.15	0.18
7	0.42	0.29
8		0.35
9		0.44
Mean	0.2657	0.4011
St. Dev	0.0934	0.1447

Figure 3: Unpaired Observations

Which policy is better?

Can we say with 90% confidence that A is better than B?

1. Compute the two sample means $(\bar{x}_A, \bar{x}_B)$separately.
2. Compute the two sample standard deviations (s_A, s_B).
3. Compute the mean difference. $\bar{x}_A - \bar{x}_B$
4. Compute the standard deviation of the mean difference.
$$s = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$$
5. Compute the Effective Degrees of Freedom
$$\nu = \frac{(\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b})^2}{\frac{1}{n_a - 1}(\frac{s_a^2}{n_a})^2 + \frac{1}{n_b - 1}(\frac{s_b^2}{n_b})^2} - 2$$
6. Compute the Confidence Interval for the mean difference
$$(\bar{x}_a - \bar{x}_b) \pm t_{[1-\alpha/2;\nu]} * s$$
$$=(\bar{x}_a - \bar{x}_b) \pm TINV(\alpha, \nu) * s$$
7. Assess which is better based on whether or not this interval contains 0.

ie. The cache hit ratio for policy A is smaller than for policy B; therefore, with a 90% confidence level, policy B is better than policy A.

3.3 Determine Sample Size for Accurate Estimations

Estimation... [44]

Compute a sample size n, which will estimate the overall population parameter. This estimate will be with accuracy $r\%$ and a confidence level of $(1 - \alpha)$

$$n = (\frac{100 * z_{1-\alpha/2} * s}{r\bar{x}})^2$$

This is a bootstrapping estimate insofar as in an initial sample set is produced and used to provide the first sample mean ($\bar{x}$) to the calculation. Once this yields a value for n, a larger sample may be taken and used to refine the result.

How many repetitions are needed to get the response time within a 2% accuracy at a 95% confidence level?

For example, if 2% accuracy at a 95% confidence level is desired, then r would be 2, z would be calculated using $NORMSINV(1 - \alpha/2)$, which would be 1.96, and s and $\bar{x}$ would experimentally determined.

4 Hypothesis Testing

4.1 Fundamentals

Make inferences about a population parameter by observing differences between experimental data and expectations to prove an assumption.

HypothesisTesting [2]

Null Hypothesis:

$$H_0 : \mu = x$$

Alternative Hypothesis:

$$H1 : \mu \neq x$$

For this data, the sample mean is 12.3. Is the actual mean 12.5 with a 95% confidence?

The test statistic for the Null Hypothesis is defined below, where $\bar{X}$ is the sample mean, μ is the theorized population mean, σ is the given standard deviation, and n is the size of the sample:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

1. State the Null and Alternative Hypotheses
2. Choose the level of significance α (ie. $0.1 = 90\%$)
3. Choose the sample size to test (n)
4. Choose the appropriate statistical technique (Z or t)
5. Determine the critical values (ie. if $Z_{[1-\alpha/2]} = 1.96 \therefore [-1.96...1.96]$)
6. Collect the data and compute: mean (μ) and average ($\bar{X}$)
7. If the test statistic is within the critical values, do not reject H_0, else reject.

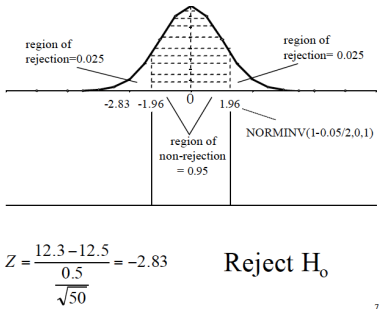


Figure 4: Hypothesis Testing with Critical Values

4.2 p-value Approach to Hypothesis Testing

This computes p and uses a simple comparison against the level of confidence desired (α). If $p < \alpha$, then it is less than the level of significance, therefore reject H_0. HypothesisTesting [11]

In the following, $-z$ refers to the negated value of the test statistic Z, as computed in the previous subsection. The mean here is 0, and the standard deviation is 1. This assumes standard normal distribution. A T-distribution would use $DOF = n - 1$

For this data, the sample mean is 12.3. Is the actual mean 12.5 with a 95% confidence?

$$\begin{aligned} \frac{p}{2} &= F(-z) \\ &= NORMDIST(-z, 0, 1, true) \end{aligned}$$

1. State the Null and Alternative Hypotheses
2. Choose the level of significance α (ie. $0.1 = 90\%$)
3. Choose the sample size to test (n)
4. Choose the appropriate statistical technique (Z or t)
5. Collect the data and compute: mean (μ), standard deviation (σ), average ($\bar{X}$)
6. Calculate the p value based on the test statistic
7. Compare p with α. If $p >= \alpha$, then do not reject H_0, else reject.

4.3 One-Tailed Test

The Null Hypothesis is an inequality, so compute based on an inequality using the test statistic. HypothesisTesting [18]

4.4 Hypothesis Testing with an Unknown Standard Deviation

When the population standard deviation σ is not known, use the sample s instead. HypothesisTesting [18]

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Always use the T-test with $DOF = n - 1$ when σ is unknown.

5 Experimental Design

5.1 One-Factor Comparisons - ANOVA

Analysis of Variations (ANOVA) looks at Total Variation (SST)

OneFactor... [3]

Within SST, there are variations among the group (SSA), which show differences between the groups, and variations within the group (SSW) that are attributable to experimental error.

Hypotheses (assuming c groups being compared within the same factor):

$$H_0 : \mu_1 = \mu_2 = ... = \mu_c$$
$$H_1 : \exists \mu_x s.t. \mu_x \neq \mu_y$$

A, B, C, and D are different page-replacement algorithms, each of which is run multiple times. Which is better?

Computations:
($\bar{\bar{X}}$ is the Grand Mean, $\bar{X}_j$ is the sample mean for group j, n_j is the count of observations in group j, and n is the total number of observations across all groups; MSA, MSW, MST are Mean Squares, which are Variances)

DOF Reference
SST = n - 1
SSA = c - 1
SSW = n - c

$$\bar{\bar{X}} = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n_j}(X_{ij})}{n}$$
$$SST = SSA + SSW$$
$$SST = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(X_{ij} - \bar{\bar{X}})^2$$
$$SSA = \sum_{j=1}^{c}[n_j(\bar{X}_j - \bar{\bar{X}})^2]$$
$$SSW = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_j)^2$$
$$MSA = \frac{SSA}{c-1}$$
$$MSW = \frac{SSW}{n-c}$$
$$MST = \frac{SST}{n-1}$$

The running times of various programs were measured on systems with different Cache sizes. What can you say about the impact of cache size on the running time?

The key statistic to compute is the one-way ANOVA F-Test statistic.

$$F = \frac{MSA}{MSW}$$

Based on this, the Null Hypothesis may be accepted or rejected, based on Fu, which is computed from a table using the two values:

$$DFNumerator = c - 1$$
$$DFDenominator = n - c$$

So, if there are $c = 4$ groups of the one-factor, and there are a grand total of $n = 40$ samples ($n_j = 10$ for each group), then the numerator will be 3, and the denominator will be 36. Looking up 3,36 in the table yields 2.87, which is the Fu critical value. Using Data Analysis, this will show up as 'F crit'

If $F > Fu$, then reject H_0, otherwise do not reject.

Using Data Analysis, a p value is also given. Reject H_0 if $p < \alpha$

5.2 Tukey-Kramer

Compute the differences $(\bar{X}_j - \bar{X}'_j)$ among all $c\frac{(c-1)}{2}$ pairs of means.

OneFactor... [26]

$$CriticalRange = q_u \sqrt{\frac{MSW}{2}(\frac{1}{n_j} + \frac{1}{n'_j})}$$

If H_0 is rejected, which groups are different?

Where q_u is the upper-tail critical value from a Studentized range (from a table) with c DOF in the numerator and $n - c$ DOF in the denominator. If the absolute difference in means exceeds this Critical Range, then the pair is statistically different.

$$|X_a - X_b| > CriticalRange$$

5.3 Randomized Block Model

Computations: OneFactor... [31]

($\bar{\bar{X}}$ is the Grand Mean, $\bar{X}_{.j}$ is the group mean for group j, $\bar{X}_{.i}$ is the group mean for group i,$\bar{X}_{i.}$ is the block mean for block i, n_j is the count of observations in group j, and n is the total number of observations across all groups; MSA, MSW, MST are Mean Squares, which are Variances). Furthermore r is the number of blocks.

DOF Reference
SST = n - 1
SSA = c - 1
SSBL = r - 1
SSE = (r-1)(c-1)

$$\bar{\bar{X}} = \frac{\sum_{j=1}^{c}\sum_{i=1}^{r}(X_{ij})}{rc}$$

$$SST = SSA + SSBL + SSE$$

$$SST = \sum_{j=1}^{c}\sum_{i=1}^{r}(X_{ij} - \bar{\bar{X}})^2$$

$$SSA = r\sum_{j=1}^{c}(\bar{X}_{.j} - \bar{\bar{X}})^2$$

Reject H_0 if F > Fu

$$SSBL = c\sum_{i=1}^{r}(\bar{X}_{i.} - \bar{\bar{X}})^2$$

$$SSE = \sum_{j=1}^{c}\sum_{i=1}^{r}(X_{ij} - X_{.j} - X_{i.} + \bar{\bar{X}})^2$$

$$MSA = \frac{SSA}{c-1}$$

$$MSBL = \frac{SSBL}{r-1}$$

$$MSE = \frac{SSE}{(r-1)(c-1)}$$

The key statistic to compute is the one-way ANOVA F-Test statistic.

$$F = \frac{MSA}{MSE}$$

$$DFNumerator = c - 1$$
$$DFDenominator = (r-1)(c-1)$$

5.4 Relative Efficiency

Relative Efficiency is computed as follows from the previous section's blocking analysis. OneFactor... [52]

$$RE - \frac{(r-1)MSBL + r(c-1)MSE}{(rc-1)MSE}$$

It is also provided as a statistic in Excel's Data Analysis as RE. If RE were 38.1, for instance, it would mean that if the test did not use blocking, it would take 38.1 times the number of observations to get the same level of precision in comparing the groups.

If blocking is not used, how many more times the number of observations would be needed to get the same precision?

5.5 Tukey-Kramer with Blocks

Like previously, Tukey-Kramer is based on the critical range. OneFactor... [53]

$$CriticalRange = q_u \sqrt{\frac{MSE}{r}}$$

Where q_u is determined by looking up the upper-tail critical value from the Studentized range distribution with c degrees of freedom in the numerator and $(r-1)(c-1)$ degrees of freedom in the denominator.

If the absolute difference between the sample means between each group is greater than the critical range, then the means are different.

5.6 Two-Factor Comparisons - ANOVA

There are three hypotheses here that need to be addressed.

FactorialExp... [4]

1. One, are there any differences due to factor A?
2. Two, are there any due to factor B?
3. Three, are any interactions between the factors?

Hypotheses (assuming c groups being compared within the same factor):

A website has the following response times when in 1 and 2 CPU configurations, as well as in 1 or 2 server clusters....

- Factor A (Reject H_0 if $F > Fu$)

$$H_0 : \mu_{1..} = \mu_{2..} = ... = \mu_{r..}$$
$$H_1 : Otherwise$$
$$F = \frac{MSA}{MSE}$$

- Factor B (Reject H_0 if $F > Fu$)

$$H_0 : \mu_{.1.} = \mu_{.2.} = ... = \mu_{.c.}$$
$$H_1 : Otherwise$$
$$F = \frac{MSB}{MSE}$$

DOF Reference
SST = n - 1
SSA = r - 1
SSB = c - 1
SSAB = (r - 1)(c - 1)
SSE = rc(n' - 1)

- Interaction Between Factors (Reject H_0 if $F > Fu$)

$$H_0 : None$$
$$H_1 : Otherwise$$
$$F = \frac{MSAB}{MSE}$$

In Excel, Sample is the Row factor

Variables:

r - Number of Levels of Factor A

c - Number of Levels of Factor B

n' - Number of Replications for Each Cell

n - Total number of Observations ($n = rcn'$)

X_{ijk} - The kth observation for level i of factor A and level j of factor B

The easiest way here is to just use Excel's Data Analysis toolset. Barring this, the full calculations are in FactorialExpDesign.pdf

6 Regression

SimpleRegression [8]

6.1 Linear Regression

The predicted value of Y for any observation i ($\hat{Y}_i$) is based on the value of the ith observation (X_i) and the two constants, b_0 and b_1, which are chosen to minimize SSE.

Subject to the Sum of Errors = 0

$$\hat{Y}_i = b_0 + b_1 X_i$$

$$b_1 = \frac{(\sum_{i=1}^{n} X_i Y_i) - n\bar{X}\bar{Y}}{(\sum_{i=1}^{n} X_i^2) - n(\bar{X})^2}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

1. Compute $\bar{X}$ by taking the AVERAGE of your X values
2. Compute $\bar{Y}$ by taking the AVERAGE of your Y values
3. Compute $(\sum_{i=1}^{n} X_i^2)$ by using SUMSQ(A2:A11)
4. Compute $(\sum_{i=1}^{n} X_i Y_i)$ by using SUMPRODUCT(A2:A11,B2:B11)
5. With n, compute b_1 and b_0
6. With b_1 and b_0, compute $\hat{Y}$ and put it in a column next to Y to compare.
7. Compute the error by $Y_i - \hat{Y}_i$ and put it in a column next to $\hat{Y}$ — Errors = Residuals
8. Square each error in the error column.
9. Sum the error squared column to get SSE.
10. Compute $Y_i - \bar{Y}$ and put it in a column next to the errors.
11. Sum the new column to get SST.
12. Subtract $SST - SSE$ to get SSR.
13. R^2 is now just SSR/SST
14. Verify errors are Normally Distributed, with no trends (independent) and constant standard deviations (no trend in spread) — SimpleRegression [23]

6.2 Degrees of Freedom

Each of the key values has its own level of degrees of freedom.

Value	DOF	Notes
SST	n-1	
SSY	n	
SS0	1	Can by computed from $\bar{Y}$
SSE	n-2	
SSR	1	SST - SSE

6.3 Standard Deviation of Errors (Simple Regression)

SimpleRegression [15]

$$s_e^2 = \frac{SSE}{n-2}$$

$$s_{b_0} = s_e \sqrt{\frac{1}{n} + \frac{(\bar{X})^2}{(\sum_{i=1}^{n} X_i^2) - n(\bar{X})^2}}$$

$$s_{b_1} = \frac{s_e}{\sqrt{(\sum_{i=1}^{n} X_i^2) - n(\bar{X})^2}}$$

Remember $t_{[1-\alpha/2;n-2]} = TINV(\alpha, n-2) = T.INV(1-\alpha/2, n-2)$

1. Compute Mean Squared Error (MSE), which is s_e^2
2. Compute s_e, which is $\sqrt{s_e^2}$
3. Compute s_{b_0} and s_{b_1}
4. Compute the confidence intervals for b_0 and b_1

$$b_0 \pm t_{[1-\alpha/2;n-2]} s_{b_0}$$
$$b_1 \pm t_{[1-\alpha/2;n-2]} s_{b_1}$$

5. List the ranges for lower and upper b_0 and b_1

6.4 Confidence Interval for Predicted Values (Simple Regression)

Here, for an existing sample size of n, a future sample size m is predicted.

SimpleRegression [20]

$$s_{\hat{y}_{mp}} = s_e \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(X_p - \bar{X})^2}{(\sum_{i=1}^{n} X_i^2) - n\bar{X}^2}}$$

$$\hat{y}_p \pm t_{[1-\alpha/2;n-2]} s_{\hat{y}_{mp}}$$

Compute 90% confidence intervals for a predicted value of 20 objects (X), assuming a future sample size of 40

1. Use previous s_e
2. Select m as your target future population, use n as existing sample size
3. Calculate $(\sum_{i=1}^{n} X_i^2)$ using SUMSQ(A2:A12)
4. Calculate $\bar{X}$, which is the average of your sample's X values
5. Calculate $(\sum_{i=1}^{n} X_i^2) - n\bar{X}^2$
6. Calculate $\hat{Y}(X_p)$, which is the predicted value under your existing $\hat{Y}$ for your predicted value (i.e... fore a predicted value of 20, $X_p = 20$)
7. Calculate $s_{\hat{y}_{mp}}$
8. Calculate and record the ranges for $\hat{y}_p$

Remember $t_{[1-\alpha/2;n-2]} = TINV(\alpha, n-2) = T.INV(1-\alpha/2, n-2)$

6.5 Multiple Linear Regression

Multiple linear regression is used when there are two factors, (X_{1_i}) and (X_{2_i}). MultipleRegression [3]

CPU Time (yi)	I/O Time (x1i)	Memory Requirement (x2i)
2	14	70
5	16	75
7	27	144
9	42	190
10	39	210
13	50	235
20	83	400

Figure 5: Multiple Regression Setup

$$CPUTime = b_0 + b_1 * IOTime + b_2 * MemoryRequirement$$

1. Just use Excel - Data Analysis, and look for Multiple R to be very close to 1

6.6 Curvilinear Regression

The objective here is to adjust the equation to achieve the standard linear form. MultipleRegression [5]

Non-linear	Linear
$y = a + b / x$	$y = a + b(1 / x)$
$y = 1/(a + bx)$	$(1 / y) = a + bx$
$y = x/(a + bx)$	$(x / y) = a + bx$
$y = a \times b^x$	$\ln y = \ln a + x \ln b$
$y = a + bx^n$	$y = a + b(x^n)$

Figure 6: Curvilinear Regression Approaches

You have data, see if the relationship of IO Rate = a * $MIPSRate^b$ is accurate.

Find y', x', a', and b' as needed and then perform a Linear Regression, looking for R^2