Statistical Analysis for Evolutionary Computation: An Introduction

Mark Wineberg

School of Computer Science
University of Guelph, Canada
wineber@uoguelph.ca

GECCO'12 Companion, July 7–11, 2012,
Philadelphia, PA, USA.
ACM 978-1-4503-1178-6/12/07.

Probability Theory vs. Statistics

Stochastic (Random) Model

↓ *Probability Theory*

Results

Probability Theory vs. Statistics

Samples

↓ *Statistics*

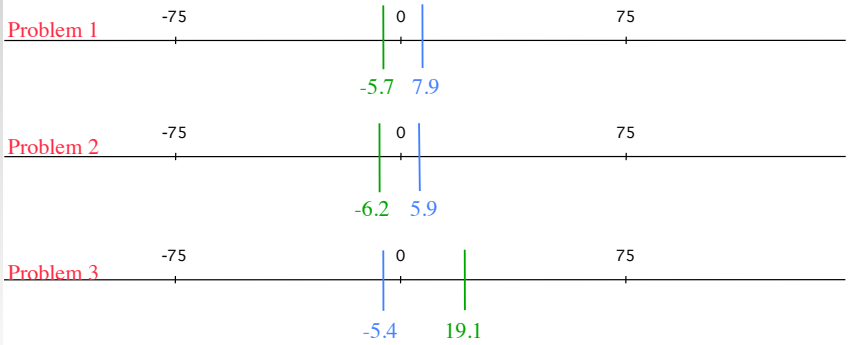
Stochastic (Random) Model

Basic Statistical Tests

Part 1 - Point Estimation:

Finding the Mean using Confidence Intervals

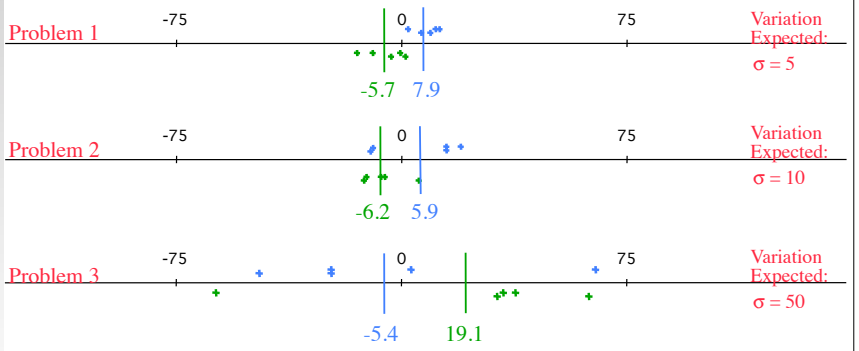
Comparing two different simulated strategies



Strategy 1 Avg.	Strategy 2 Avg.

Number of Reps = 5

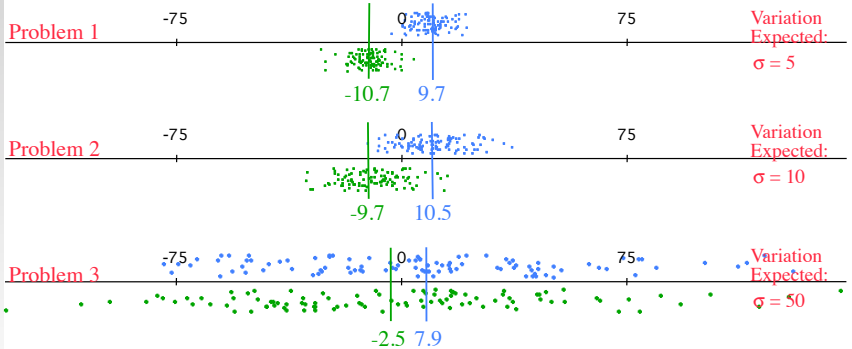
Sampling From Two Normal Distributions



True Avg.	-10	+10

Number of Reps = 5

Sampling From Two Normal Distributions



True Avg.	-10	+10

Number of Runs = 100

What Are We Interested In?

- For most statistical analysis for CS the question is
  - Is one way better than another way?
  - Statistically this translates into a statement about the difference between means: "Is the difference between 'my mean' and 'the other mean' greater than zero?"
- We will approach this question in 2 steps:
  1. What can we say about the true mean of a *single* distribution?
     - Called *point estimation*
  2. How can we compare the true means of *two* or more distributions?

Confidence Intervals

- Where is the true mean?
  - Sample averages bounce around so can't give exact point
- Can give a range within which it is very likely the true mean lies
  - Called the *confidence interval*
- Also should consider the probability that the true mean lies with the range
  - Called the *confidence level*
- Example:
  - In 99% of cases, the true mean of the distribution, estimated from our 50 samples, lies within the interval [64 , 79]

Normal (Gaussian) Distribution

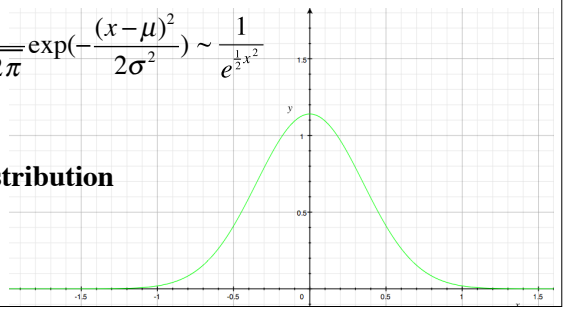
- For now we will assume that we are sampling from (the results are distributed as) a *normal distribution*
  - AKA Bell Curve
  - Most common distribution found in nature thanks to the *Central Limit Theorem*

$$P(X = x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \sim \frac{1}{e^{\frac{1}{2}x^2}}$$

$$X \sim N(\mu, \sigma^2)$$

Standard Normal Distribution

$$X \sim N(0, 1)$$



Distribution of the Mean

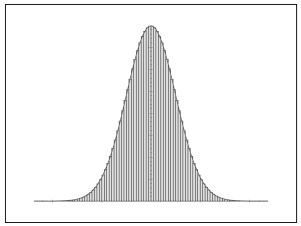
- Consider the distribution of the average of a set of *n* independent samples
  - If *n* = 1, the distribution of the average is just the distribution itself, since we have only the single data point
  - If *n* is larger than one, the distribution of the mean must be narrower than the distribution of the population
    - i.e. the variance and standard deviation must be smaller
  - In fact, the mean & variance of the mean of n samples is

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

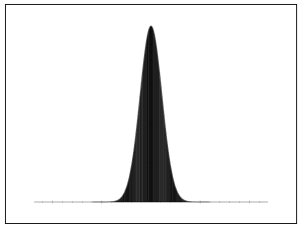
Distribution of the Mean

(Standard Normal Distribution)

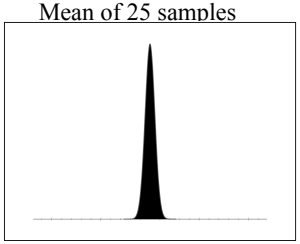
Mean of one sample



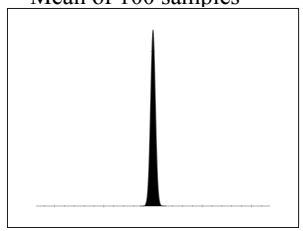
Mean of 5 samples



Mean of 25 samples



Mean of 100 samples



Confidence Intervals

- Of course, we don't know the true mean, μ, or true standard deviation, σ
- We *do* know the mean of the samples, $\bar{X}$, the sample size, n, and the sample standard deviation, s_X
- If the source distribution is *normally distributed*, the shape and size of the "finger" is known exactly!
  - We can determine the odds that the true mean lies within a specified range of $\bar{X}$
- First since $\bar{X}$ is normally distributed, we can turn it into a standard normal distribution

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_X}{\sigma_X / \sqrt{n}}$$

Confidence Intervals

- First since $\bar{X}$ is normally distributed, we can turn it into a standard normal distribution
  - subtract off the mean to zero it
  - divide by the std deviation to give it a std deviation of 1
    - also gives a variance of 1

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_X}{\sigma_X / \sqrt{n}}$$

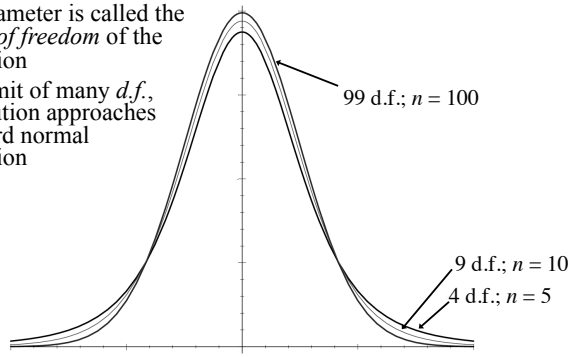
t Distribution

- Want to find μ the true mean in terms of the average
  - But we have not one but two unknowns - σ is also unknown
  - One equation - two unknowns - not good!!!
  - Trick - divide by the known sample standard deviation s instead of σ
- But now we have a normally distributed numerator divided by a non-normally distributed denominator
  - Denominator has a chi distribution
  - A normal distribution over a chi distribution has a Student's *t* distribution

$$T = \frac{\bar{X} - \mu_{\bar{X}}}{s_{\bar{X}}} = \frac{\bar{X} - \mu_X}{s_X / \sqrt{n}}$$

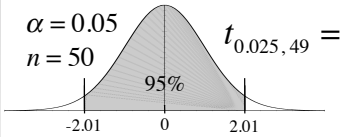
t Distribution

- The *t* "distribution" is really a family of distributions – the shape of the distribution changes as the number of samples, n, changes
  - This parameter is called the *degrees of freedom* of the distribution
  - In the limit of many *d.f.*, *t* distribution approaches a standard normal distribution



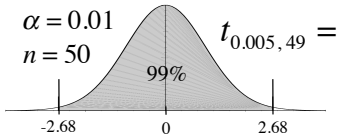
Estimating the Mean: Confidence Intervals Around the Average

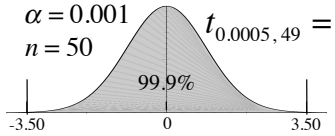
If samples taken from a *standard normal distribution* ($\mu = 0, \sigma = 1$),
Based on $n = 50$ runs the sample average has a t distribution.



- α is the probability of seeing values outside the cutoffs
  - Confidence Level is 1 – α
- Cut off $t_{\frac{\alpha}{2}, n-1}$ values can be computed using Excel: **=TINV(α, *n* - 1)**
  - Note: TINV() is already 2 sided

- For CI, we can use cutoff t values
- The wider the cutoff values, the more likely the true mean lies between them





Estimating the Mean: Confidence Intervals Around the Average

- We know that $$T = \frac{(\bar{X} - \mu_X)}{s_X / \sqrt{n}}$$
- Using the $\pm t_{\frac{\alpha}{2}, n-1}$ cutoff t-values we can form a Confidence Interval that has a 1 - α C.L with n - 1 degrees of freedom
- Substituting the cutoff values from the C.I. into the above equation produces

$$\pm t_{\frac{\alpha}{2}, n-1} = \frac{\bar{X} - \mu_X}{s_X / \sqrt{n}}$$

which can be rewritten as $$\mu_X = \bar{X} \pm t_{\frac{\alpha}{2}, n-1} \frac{s_X}{\sqrt{n}}$$

Estimating the Mean: Confidence Intervals Around the Average

- Confidence Intervals can be written in 3 equivalent ways

Error Bounds

$$\mu_X = \bar{X} \pm t_{\frac{\alpha}{2}, n-1} \frac{s_X}{\sqrt{n}}$$

Confidence Intervals

$$\bar{X} - t_{\frac{\alpha}{2}, n-1} \frac{s_X}{\sqrt{n}} \le \mu_X \le \bar{X} + t_{\frac{\alpha}{2}, n-1} \frac{s_X}{\sqrt{n}}$$

$$\mu_X \in \left[\bar{X} - t_{\frac{\alpha}{2}, n-1} \frac{s_X}{\sqrt{n}}, \quad \bar{X} + t_{\frac{\alpha}{2}, n-1} \frac{s_X}{\sqrt{n}} \right]$$

Estimating the Mean: Confidence Intervals Around the Average

Example:

- An experimenter runs a New Search Algorithm on a TSP
- At the end of each run, the smallest length tour that had been found during the run was recorded
- NSA is run 50 times on the same TSP problem
- On average NSA found solutions with a tour length of 272
- The standard deviation of these tours is 87
- We want to compute a Confidence Interval using a 99% Confidence level

Estimating the Mean: Confidence Intervals Around the Average

- From the problem we know that the average NEA run produced tours of

$$\bar{X} = 272 \text{ that had } s_X = 87$$

We know that $\mu_X = \bar{X} \pm t_{\frac{\alpha}{2}, n-1} \frac{s_X}{\sqrt{n}}$

- Also from the problem $n = 50$ and $\alpha = (1 - 0.99) = 0.01$

so the $\pm t$ cutoff value is $t_{\frac{0.01}{2}, 50-1} = t_{0.005, 49}$

using Excel we see that TINV(0.01,49) is 2.68

$$\mu_X = \bar{X} \pm 2.68 \frac{s_X}{\sqrt{50}} = \bar{X} \pm 0.38 s_X = \bar{X} \pm 33$$

and so $239 \le \mu_X \le 305$ with a 99% C.L.

i.e. there is only a 1% chance that the true mean lies outside the confidence interval formed around average

Basic Statistical Tests

Part 2 - Comparisons:
Non-Overlapping Confidence Intervals and the Student's T Test

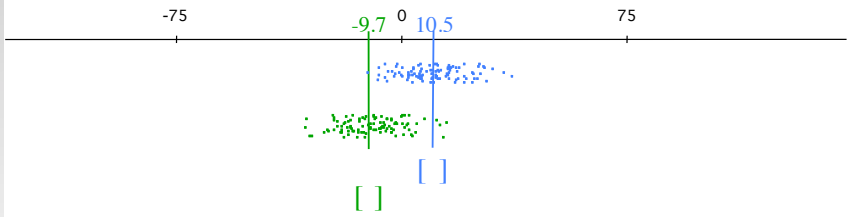
Using Confidence Intervals to Determine Whether My Way is Better

If we have two different EC systems how can we tell if one is better than the other?

Trivial method: Find confidence intervals around both means

- If the CIs don't overlap
  - Then it is a rare occurrence when the two systems do have identical means
  - The system with the better mean can be said to be better on average with a probability better than the Confidence Level
- If the CIs do overlap
  - Can't say that the two systems are different with this technique
  - Either:
    1. The two systems are equivalent
    2. We haven't sampled enough to discriminate between the two

Confidence Interval Example

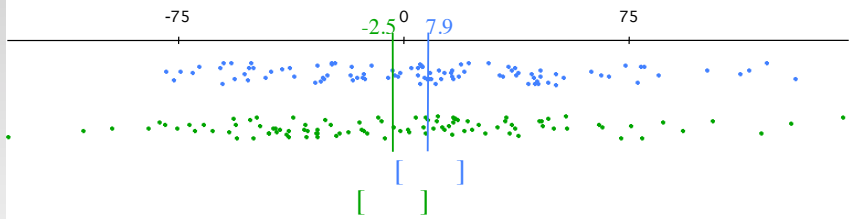


μ	σ
+10	10
-10	10

95% Confidence Level

n	X	s_X	$1.96\frac{s_X}{\sqrt{n}}$	Lower	Uppe
100	10.5	10.0	3.3	7.2	13.8
100	-9.7	10.1	3.3	-13.1	-6.4

Confidence Interval Example



μ	σ
+10	50
-10	50

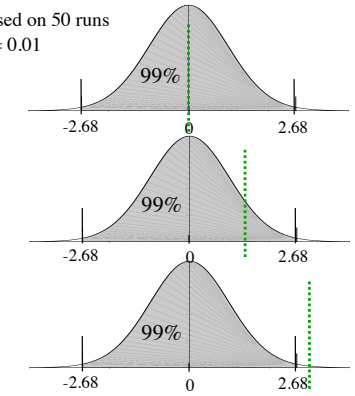
95% Confidence Level

n	X	s_X	$1.96\frac{s_X}{\sqrt{n}}$	Lower	Uppe
100	7.9	47.1	9.2	-1.3	17.1
100	-2.5	52.1	10.2	-12.7	7.7

The Student *t* Test

- Idea: Gain sensitivity by looking at the difference between the means of the two systems

Where the normalized difference falls on the *t* distribution determines whether difference expected if both systems were actually performing the same

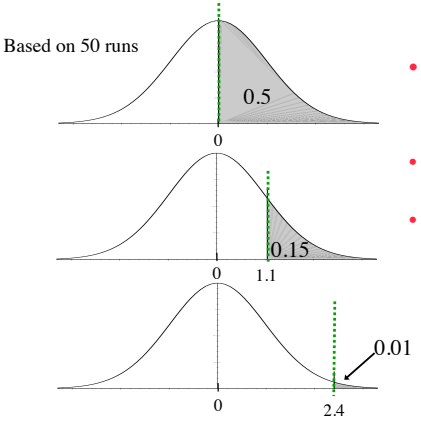


- Normalized difference called the *t* value

$$t\ value = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{n}}}$$

- Distribution again differs for different sample sizes
  - Degrees of Freedom is now $= (n-1) + (n-1) = 2n-2$
- *t* test either succeeds or fails
  - *t* value greater than cutoff for a given C.L. or not

The Student *t* Test: *p*-values



- The cut-off values produces a binary decision: true or false
  - loses information
- Better to report the probability that two systems are different
- This is the complement of the probability that they are the same
  - 1 – Pr($T < t$ score)
  - Called the *p*-value

t Test Step by Step

1. Compute the 2 averages X_1 and X_2
2. Compute standard deviations s_1 and s_2
3. Compute degrees of freedom: $n_1 + n_2 - 2 = 2n - 2$
4. Calculate *T* statistic: $T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{n}}}$
5. Compute the *p*-value
   1. *p*-value = the area under the *t* distribution outside [-*T*, *T*]
   2. Use **=TDIST(*T*, $n_1 + n_2$ - 2, 2)** in Excel
      1. The final "2" in Excel means "two-sided"

Variance Assumptions and the T Test

$\sigma_1 = \sigma_2 = \sigma$ and $n_1 = n_2 = n$

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{n}}}$$

$\sigma_1 = \sigma_2 = \sigma$ but $n_1 \neq n_2$

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)s_{X_1}^2 + (n_2-1)s_{X_2}^2}{(n_1+n_2-2)}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Variance Assumptions and the T Test

$\sigma_1 \neq \sigma_2$ and $n_1 \neq n_2$

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_{X_1}^2}{n_1} + \frac{s_{X_2}^2}{n_2}}}$$

← Approximate variance not pooled

$$\text{D.F.} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$$

called the Welch's T test

t.test(): Everything in one simple R function

- R comes with a function that produces all of the above information in one function
  - t.test(X,Y)
- See R help for details

Tests on Non-Normally Distributed Random Variables

Non-Parametric Statistics

When The Normality Fails

- Everything so far has depended on the assumption of normality which in turn depends on the Central Limit Theorem holding
  - But this is not always true
  - ***In in many areas of CS it rarely holds***
- Problems occur when
  - …you have a non-zero probability of obtaining infinity
    - Mean and standard deviation are infinite!
  - …the sample average depends highly on a few scores
    - When the mean of your distribution is not measuring what you want, consider using the median instead (rank-based statistics)
  - …you don't know how fast your sample series converges to normal
    - if your sample average distribution converges very slowly than the number of samples may be *insufficient to assume normality*

So what should we do?

First test for normality

- Many such tests
- Recommended
  - Normal Probability Plot
    (QQ plot: sorted data vs Normal quantiles)
  - Lilliefors test (variant of the KS test)

So what should we do?

There are 3 basic remedial measures:

1. Transforming data to make them normally distributed
   - also called *data re-expression*
   - traditional approach (required before the advent of fast computers)
2. Resampling techniques
3. Non-parametric statistics

Non-Parametric Statistics

- Basic Idea
  - Sort the data and then rank them
  - Use Ranks instead of actual values to perform statstics
- Also known as
  - *order statistics*,
  - *ordinal statistics*
  - *rank statistics*
- Measures how interspersed the samples are from the 2 treatments
  - If the result is "alternating" it is assumed that there is no difference
- Can't be affected by outliers (extrememly large or small values)
  - Just the highest or lowest rank

Non-Parametric Tests

- Reason behind the appropriateness of non-parametric tests
  - Both the sum of ranks and average of ranks will be approximately normally distributed
    - because of the Central Limit Theorem,
    - as long as we have 5 or more samples
  - result is independent of the underlying distribution
- Ranked T-test
  - Perform a *t* test on the ranks of the values
    - instead of the values themselves
- 2 other techniques with similar results are commonly seen
  - Wilcoxon's Rank-Sum test
  - Mann-Whitney U test
  - All are effectively equivalent

A	0.03
A	0.91
A	0.64
A	0.99
A	0.64
A	0.16
A	0.16
A	0.91
A	0.16
A	0.27
B	0.64
B	0.08
B	0.16
B	0.27
B	0.02
B	0.01
B	0.16
B	0.03
B	0.03
B	0.64

Two sets of Data

Combine the data into a single array

Ranked Example

		ranks
A	0.99	1
A	0.91	2
A	0.91	3
A	0.64	4
A	0.64	5
B	0.64	6
B	0.64	7
A	0.27	8
B	0.27	9
A	0.16	10
A	0.16	11
A	0.16	12
B	0.16	13
B	0.16	14
B	0.08	15
A	0.03	16
B	0.03	17
B	0.03	18
B	0.02	19
B	0.01	20

Sort the combined data

Give each data element its corresponding rank

Ranked Example

		ranks	
A	0.99	1	
A	0.91	2.5	t1
A	0.91	2.5	t1
A	0.64	5.5	t2
A	0.64	5.5	t2
B	0.64	5.5	t2
B	0.64	5.5	t2
A	0.27	8.5	t3
B	0.27	8.5	t3
A	0.16	12	t4
A	0.16	12	t4
A	0.16	12	t4
B	0.16	12	t4
B	0.16	12	t4
B	0.08	15	
A	0.03	17	t5
B	0.03	17	t5
B	0.03	17	t5
B	0.02	19	
B	0.01	20	

Average tied ranks together

t1	2.5
t2	5.5
t3	8.5
t4	12
t5	17

Replace tied ranks with average tied ranks

Ranked Example

		ranks
A	0.99	1
A	0.91	2.5
A	0.91	2.5
A	0.64	5.5
A	0.64	5.5
A	0.27	8.5
A	0.16	12
A	0.16	12
A	0.16	12
A	0.03	17
B	0.64	5.5
B	0.64	5.5
B	0.27	8.5
B	0.16	12
B	0.16	12
B	0.08	15
B	0.03	17
B	0.03	17
B	0.02	19
B	0.01	20

Resort by treatment

Perform *t* test on Ranks

	A_{rank}	B_{rank}
avg	7.85	13.15
stdDev	5.28	5.33

	Ranked *t* Test	
$s_T = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$	2.37	$n = 10$
$(avg_A - avg_B)/s_T$	2.23	t_R score
p-value	0.038	

Ranked Example

Ranked *t* Test: What do we pay?

- *t* Test is optimized for the normal distribution
- *t* Test on the ranks is not
  - How much do we pay?

Table created by Steffen Christensen

Distribution	# Samples for *t* Test	# Samples for *t* Test on Ranks	# Samples of t_R, normalized to 50 runs of *t*
Normal	31	32	52
Exponential	29	16	27
Uniform	31	34	55
Bimodal	31	34	54
Chubby Tails	40	12	15

A Non-Parametric 'Mean': The Median

- Average of a data set that is not normally distributed produces a value that behaves non-intuitively
  - Especially if the probability distribution is skewed
    - Large values in 'tail' can dominate
    - Average tends to reflect the typical value of the "worst" data not the typical value of the data in general
- Instead use the Median
  - 50th percentile
  - Counting from 1, it is the value in the $\frac{n+1}{2}$ position
    - If *n* is even, (*n*+1)/2 will be between 2 positions, average the values at that position

A Confidence Interval Around the Median: Thompson-Savur

- Find the *b* the binomial value that has a cumulative upper tail probability of $\alpha/2$
  - *b* will have a value near *n*/2

The lower percentile $l = \frac{b}{n-1}$

The upper percentile $u = 1 - l$

In Excel:

To calculate *b* use
`CRITBINOM(n, 1/2, α/2)`
to compute the $value_u$ use the function
`PERCENTILE(dataArray, u)`
to compute the $value_l$ use the function
`PERCENTILE(dataArray, l)`

- Confidence Interval is $[value_l, value_u]$
  - i.e. $value_l \le median \le value_u$
  - With a confidence level of $1-\alpha$

Thompson-Savur: Example

0.27
1.06
0.23
0.70
3.42
0.87
0.42
0.33
0.48
0.23
1.04
0.64
1.04
0.31
0.06
1.58
2.79
0.09
0.49

Sort Data

18	3.42
17	2.79
16	1.58
15	1.06
14	1.04
13	1.04
12	0.87
11	0.70
10	0.64
9	0.49
8	0.48
7	0.42
6	0.33
5	0.31
4	0.27
3	0.23
2	0.23
1	0.09
0	0.06

$n = 19$ $CL = 99\%$

$\alpha = 0.01$

$b = \texttt{CRITBINOM}(n, 1/2, \alpha/2)$

$= \texttt{CRITBINOM}(19, 0.5, 0.005)$

$= 4$

$l = b/(n-1) = 4/(19-1) = 2/9 = 0.222$

$u = 1 - l = 0.778$

$rank_l = b = 4$ $value_u = 1.04$

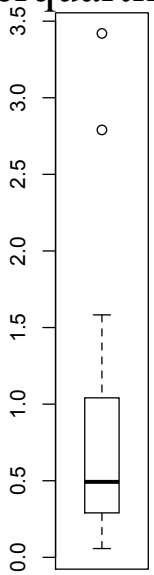
$rank_{med} = (n-1)/2 = 9$ median = 0.49

$rank_u = (n-1) - b = 14$ $value_l = 0.27$

Interquartile Range & Box Plots

18	3.42
17	2.79
16	1.58
15	1.06
14	1.04
13	1.04
12	0.87
11	0.70
10	0.64
9	0.49
8	0.48
7	0.42
6	0.33
5	0.31
4	0.27
3	0.23
2	0.23
1	0.09
0	0.06

Sorted Data



$rank_{uq} = 0.75 * (n-1) = 13.5$

$rank_{med} = 0.5 * (n-1) = 9$

$rank_{lq} = 0.25 * (n-1) = 4.5$

$box_{uq} = (1.04 + 1.04)/2 = 1.04$

$median = 0.49$

$box_{lq} = (0.27 + 0.31)/2 = 0.29$

IQR = 1.04 - 0.29 = 0.75

$wisker_u = 1.58 <= 1.5 * IQR + box_{uq} = 2.165$

$wisker_l = 0.06 >= box_{lq} - 1.5 * IQR = -0.835$

Summary

$wisker_u = 1.58$

$box_{uq} = 1.04$

$median = 0.49$

$box_{lq} = 0.29$

$wisker_l = 0.06$

IQR ~ standard deviation
wiskers ~ error bars * $\sqrt{n}$

ANOVA: Analysis of Variance

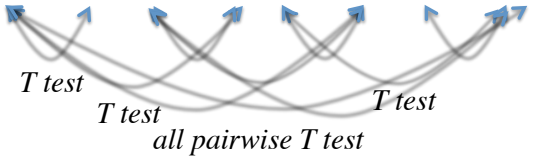
Part 1: Multi-Level Analysis Basic Concept

More Than 2 Treatments

- Preceding stats to be used for simple experiment designs
- More sophisticated stats needs to be done if:
  - Comparing multiple systems instead of just 2 treatments
    - E.g. comparing the effect on a Genetic Algorithm of using no mutation, low, medium and high levels of mutation
      - We say there are 4 *levels* of the mutation variable
      - Need $\binom{4}{2} = 6$ possible comparisons to test all pairs of treatments
  - Called a 'multi-level' analysis

Analysis of Variance (ANOVA)

	no xover	*xover* = 1pt	*xover* = 2pt	*xover* = 3pt	*xover* = 4pt
Fitness Values	4.3	8.8	5.0	6.3	5.4
	3.7	7.7	5.3	6.6	5.9
	4.7	8.3	5.1	7.2	5.4
	3.7	8.1	5.2	7.4	5.4
	4.2	8.1	5.5	7.4	6.2
	3.6	8.0	4.9	7.3	6.7
avg fitness	4.02	8.13	5.09	7.02	5.76
std dev	0.451	0.313	0.424	0.478	0.471



Question: Do crossover settings make a difference at all?

Comparing Variances

- Up to now we have been comparing means
  - Student's T test
  - Difference between averages (after normalization)
    - see if it equals 0
- From here on we will be comparing variances
  - Won't take the difference between variances
    - Difference between variances not a nice distribution
  - Rather will take the ratio of variances
    - see if it equals 1
    - distribution known: F distribution

The F Distribution

- Remember the sample variance over the true variance has a χ^2 distribution with a n - 1 degrees of freedom
- But what about the ratio of two variances?
  - With degrees of freedom of d_1 and d_2
- Answer: It has a F(d_1,d_2) distribution
  - F distribution is the ratio of two χ^2 distribution over their degrees of freedom

$$F(d_1,d_2) = \frac{\chi_1^2 / d_1}{\chi_2^2 / d_2} = \frac{d_2 \chi_1^2}{d_1 \chi_2^2}$$

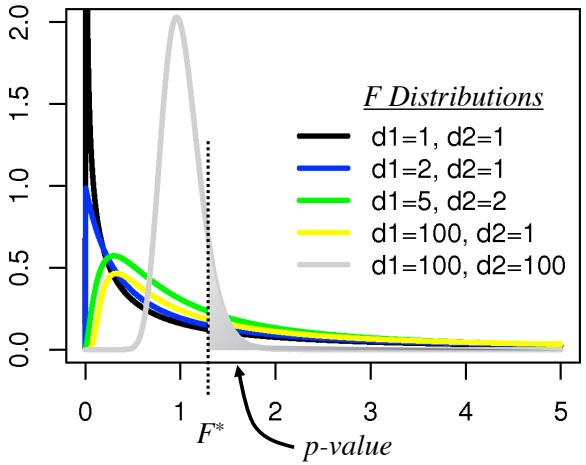
The F test

- H_0: $V(X_1) = V(X_2)$
- H_a: $V(X_1) \neq V(X_2)$
- Test Statistic $F^* = \frac{V(X_1)}{V(X_2)}$ *has a* $F(df_1, df_2)$ *distribution*

If X_1 is sampled n_1 times and X_2 is sampled n_2 times,

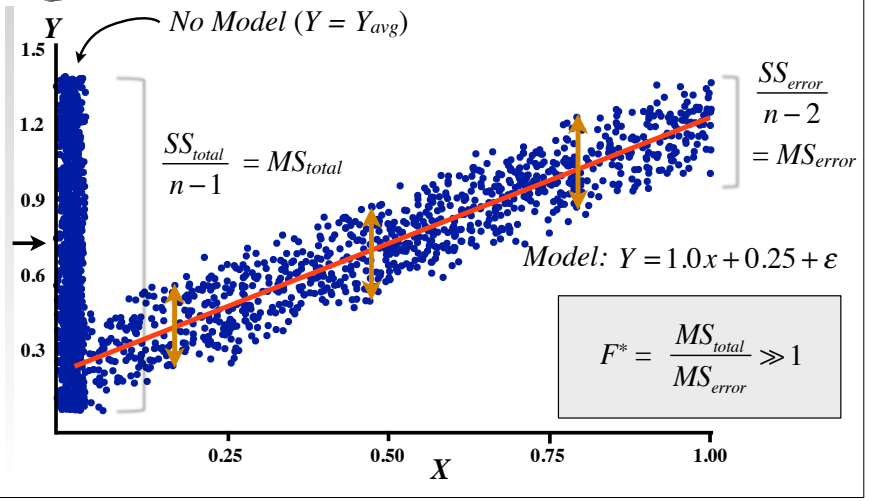
then F has a $F(n_1 - 1, n_2 - 1)$ distribution*

The F Test



From Wikipedia: http://en.wikipedia.org/wiki/F_distribution

ANOVA: Graphical Intuition



ANOVA: Graphical Intuition

Linear regression can be solved by means of least squared error

Model $Y = \beta_0 - \beta_1 x + \varepsilon$

Estimate $\hat{Y}_i = b_0 - b_1 x_i \qquad b_0 = \bar{Y} - b_1 \bar{x}$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\text{cov}(x,Y)}{\text{var}(x)} = \frac{S_{xy}}{S_x^2}$$

ANOVA: Graphical Intuition

Linear regression can be solved by means of least squared error

Mo

Esti

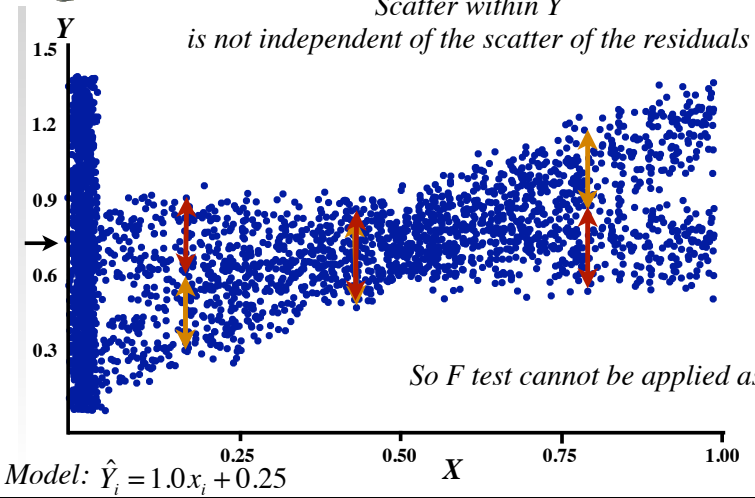
Note:

b_0 and b_1 vary depending on your samples
So they have computable variance
and an unknown true mean

Consequently you must compute
confidence intervals for each
T tests to see if they are statistically different than 0

ANOVA: Graphical Intuition

Scatter within Y
is not independent of the scatter of the residuals

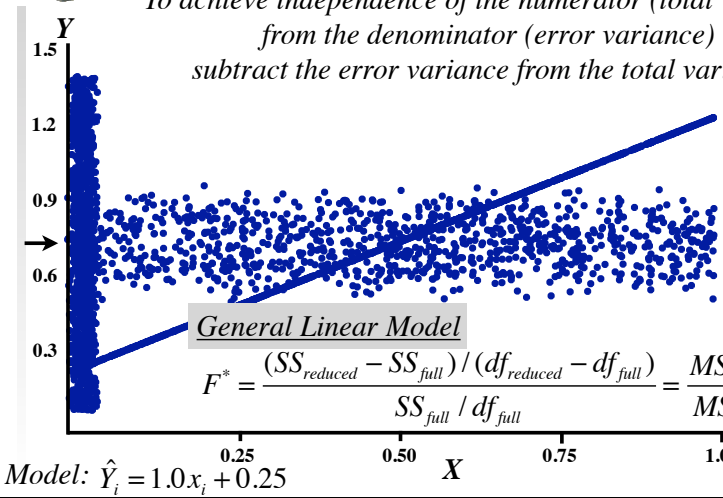


So F test cannot be applied as is

Model: $\hat{Y}_i = 1.0x_i + 0.25$

ANOVA: Graphical Intuition

To achieve independence of the numerator (total variance)
from the denominator (error variance)
subtract the error variance from the total variance



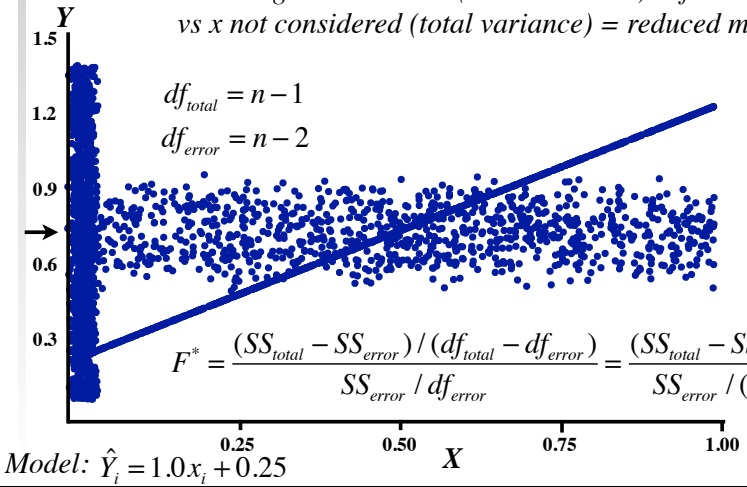
General Linear Model

$$F^* = \frac{(SS_{reduced} - SS_{full})/(df_{reduced} - df_{full})}{SS_{full}/df_{full}} = \frac{MS_{model}}{MS_{error}}$$

Model: $\hat{Y}_i = 1.0x_i + 0.25$

ANOVA: Graphical Intuition

For a regression model (error variance) = full model
vs x not considered (total variance) = reduced model



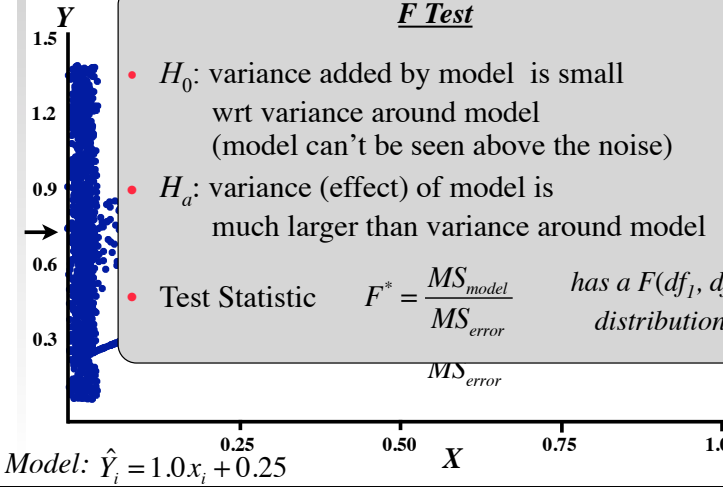
$df_{total} = n - 1$

$df_{error} = n - 2$

$$F^* = \frac{(SS_{total} - SS_{error})/(df_{total} - df_{error})}{SS_{error}/df_{error}} = \frac{(SS_{total} - SS_{error})/1}{SS_{error}/(n-2)}$$

Model: $\hat{Y}_i = 1.0x_i + 0.25$

ANOVA: Graphical Intuition

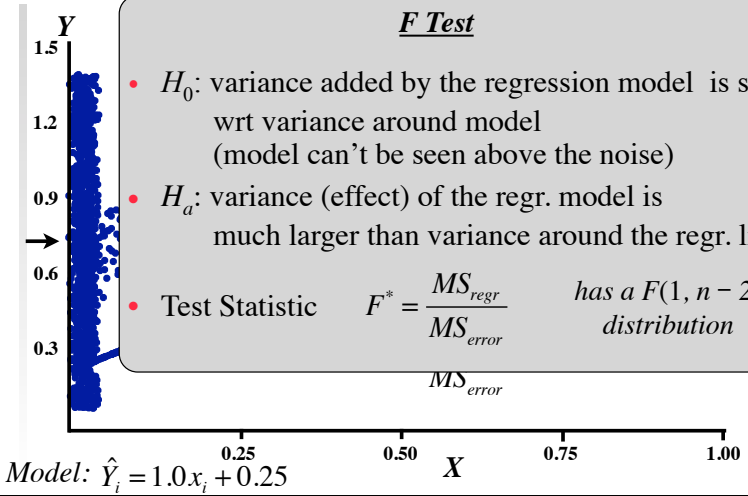


F Test

- H_0: variance added by model is small wrt variance around model (model can't be seen above the noise)
- H_a: variance (effect) of model is much larger than variance around model
- Test Statistic $F^* = \frac{MS_{model}}{MS_{error}}$ *has a* $F(df_1, df_2)$ *distribution*

Model: $\hat{Y}_i = 1.0x_i + 0.25$

ANOVA: Graphical Intuition



F Test

- H_0: variance added by the regression model is small wrt variance around model (model can't be seen above the noise)
- H_a: variance (effect) of the regr. model is much larger than variance around the regr. line
- Test Statistic $F^* = \frac{MS_{regr}}{MS_{error}}$ *has a* $F(1, n-2)$ *distribution*

MS_{error}

Model: $\hat{Y}_i = 1.0x_i + 0.25$

Polynomial Regression E.g.

R squared = 70.2% R squared (adjusted) = 70.2%
s = 0.1465 with 1000 - 4 = 996 degrees of freedom

Source	*Sum of Squares*	*df*	*Mean Square*	*F-ratio*	*p-value*
Regression	*50.4684*	*3*	*16.8228*	*784*	*≤ 0.0001*
Residual	*21.3807*	*996*	*0.021467*		

Variable	*Coefficient*	*s.e. of Coeff*	*t-ratio*	*p-value*
Constant	*0.510755*	*0.0190*	*26.9*	*≤ 0.0001*
X	*-2.17801*	*0.1636*	*-13.3*	*≤ 0.0001*
X^2	*8.45358*	*0.3813*	*22.2*	*≤ 0.0001*
X^3	*-6.28741*	*0.2515*	*-25.0*	*≤ 0.0001*

Regression model is statistically significant
F-ratio = 784 >> 1

ANOVA: Back to Discrete Levels

	no xover	*xover* = 1pt	*xover* = 2pt	*xover* = 3pt	*xover* = 4pt
Fitness Values	4.3	8.8	5.0	6.3	5.4
	3.7	7.7	5.3	6.6	5.9
	4.7	8.3	5.1	7.2	5.4
	3.7	8.1	5.2	7.4	5.4
	4.2	8.1	5.5	7.4	6.2
	3.6	8.0	4.9	7.3	6.7
avg fitness	4.02	8.13	5.09	7.02	5.76
std dev	0.451	0.313	0.424	0.478	0.471

T test
T test
T test
all pairwise T test

Question: Do crossover settings make a difference at all?

ANOVA: Discrete

fitness	*no xover*	*xover* = 1pt	*xover* = 2pt
	4.3	8.8	5.0
	3.7	7.7	5.3
	4.7	8.3	5.1
	3.7	8.1	5.2
	4.2	8.1	5.5
	3.6	8.0	4.9

→

fitness	xover
4.3	*no*
3.7	*no*
4.7	*no*
3.7	*no*
4.2	*no*
3.6	*no*
8.8	1pt
7.7	1pt
8.3	1pt
8.1	1pt
8.1	1pt
8.0	1pt
5.0	2pt
5.3	2pt
5.1	2pt
5.2	2pt
5.5	2pt
4.9	2pt

Most statistic packages work with tables formatted as the chart to the right

each column is a variable
(fitness is a response variable, xover is a factor)
each row is a treatment
(i.e the settings and results of a single run)

ANOVA Notation

i subscript: levels - there are r levels
j subscript: repetitions - there are n_i repetitions for level i

$$Y_{i\bullet} = \sum_{j=1}^{n_i} Y_{ij} \qquad \bar{Y}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$

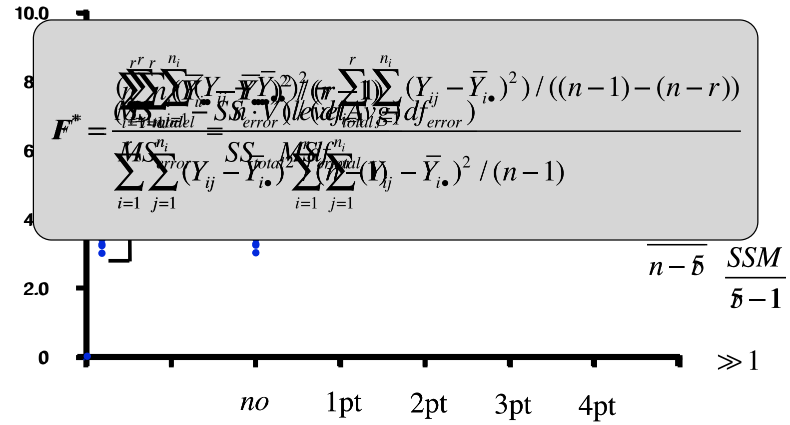
$$Y_{\bullet\bullet} = \sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{ij} \qquad \bar{Y}_{\bullet\bullet} = \frac{1}{n_T}\sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{ij}$$

Generalizes to higher dimensions

ANOVA: Discrete Levels

Perform the F test

$$F^* = \frac{MS_{model}}{MS_{error}} = \frac{SS_{model}/df_{model}}{SS_{error}/df_{error}} = \frac{\left(\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{\bullet\bullet})^2 - \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2\right)/((n-1)-(n-r))}{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2/(n-1)}$$



ANOVA: Discrete Levels

If all levels are sampled equally

$$E(MS_{model}) = \frac{\sigma^2 + 4\sum_{i=1}^{r}(\mu_i - \bar{\mu})^2}{r-1}$$

Assumption:
variance for every level is the same and equals σ^2

Test for equivalent variances:
modified Levene's test (more powerful F test)

If test fail: *(advanced technique)*
use weighted least squares regression using
- *indicator variables for the different levels as the weight as the weight for the i^{th} level*
- *Generalized ANOVA using regression*

ANOVA table for example

from DataDesk

Source	*df*	*SS*	*MS*	*F-ratio*	*Prob*
const	1	3592.9	3592.9	13967	≤ 0.0001
xover	4	210.9	52.7	204.94	≤ 0.0001
Error	95	24.4	0.257		
Total	99	235.3			

$$F^* = \frac{MS_{model}}{MS_{error}} = \frac{52.7}{0.257} = 204.94$$

F test (From Excel)

$fdist(204.94, 4, 95) = 8.19\text{E-}46$

Non-parametric ANOVA

- Again, what happens if Y (or actually ε) is not normally distributed?
- Various non-parametric techniques
  - Kruskal-Wallis first such test
- However, even simpler technique
  - Like Spearman's correlation coefficient and non-parametric regression, replace the Y_i values with their corresponding ranks
  - Perform ANOVA on ranked values as usual
- A slightly more accurate version is called the Friedman test
  - Same as above, except
    - the F distribution is replaced by the Chi-Squared distribution ($DofF = r - 1$) for large n or r ($n > 15$ or $r > 4$)
    - a special purpose distribution for small n or r

ANOVA: Analysis of Variance

Part 2: Multi-Level Analysis
Pairwise Comparisons
Post-Hoc Analysis

Pairwise Comparisons between Factor-Level Means

- What if we want to know more detailed information?
  - Which of the means is the significantly different one?
  - Are there more than one significantly different mean?
  - If so, what are the pair-wise differences and are they statistically significant?

Pairwise Comparisons between Factor-Level Means

- This is determined by a series of pair-wise T tests
- However, commonly uses pooled information from the model for the variance to provide greater accuracy
  - Called *standard error*

original T test comparison

$$t\ value = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2_{X_1}}{n_1} + \frac{s^2_{X_2}}{n_2}}}$$

comparing level i with level j across the ANOVA model

$$\rightarrow \quad t\ value = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{2 \cdot MSE}{n}}}$$

when $n_i = n_j = n$

Assumption: variances for each factor level is the same (σ^2) which is best estimated by the *MSE*

Multiple Levels: Post-hoc Analysis

- For 4 levels of mutation there are 6 comparisons possible
  - *Each one* of the comparison holds at a 95% C.L. independent of the other comparisons
  - If *all* comparisons are to hold at once the odds are 0.95 x 0.95 x 0.95 x … x 0.95 = $(0.95)^6 = 0.735$
  - So in practice we only have 73.5% C.L
    - Wrong 1/4 of the time
- For 7 levels of mutation there are 21 comparisons possible
  - C.L. = $(0.95)^{21} = 0.341$
    - Chances are better than half that at least one of the decisions may be wrong!

The Bonferroni Correction

- To correct, choose a smaller α

$$\alpha' = \frac{\alpha}{m}$$

  - Where *m* is the number of comparisons
  - So for 95% CL use $\alpha = 0.025/6 = 0.004167$
  - For a Z test the critical value changes from 1.96 to 2.64
- You should apply the Bonferroni (etc.) correction:
  - To *t* tests (*t* tests and ranked *t* tests)
  - To Confidence Intervals and Error Bounds
  - Whenever you mean "all the significant results we found hold at once"

Pairwise Comparisons between Factor-Level Means

Regular Pair-wise T test (with Bonf. Correction)

	Diff	*std. err.*	*t-value*	*df*	*p-value*
n - 1	-4.04	0.15	-27.5	18	3.6E-15
n - 3	-3.18	0.16	-20.5	18	6.3E-13
2 - 1	-3.04	0.16	-20.2	18	8.4E-13
3 - 2	2.16	0.17	13.7	18	5.5E-10
4 - 1	-2.09	0.17	-12.7	18	2.0E-09
n - 4	-1.95	0.17	-11.4	18	1.1E-08
4 - 3	-1.22	0.18	-7.1	18	1.3E-05
n - 2	-1.00	0.16	-6.3	18	5.8E-05
4 - 2	0.95	0.16	5.6	18	2.6E-04
3 - 1	-0.86	0.15	-5.6	18	2.6E-04

Pairwise Comparisons between Factor-Level Means

ANOVA Pair-wise T test (with Bonf. Correction)

	Diff	*std. err.*	*t-value*	*df*	*p-value*
n - 1	-4.04	0.16	-25.2	95	7.7E-43
n - 3	-3.18	0.16	-19.8	95	1.7E-34
2 - 1	-3.04	0.16	-19.0	95	4.8E-33
3 - 2	2.16	0.16	13.6	95	6.0E-23
4 - 1	-2.09	0.16	-13.0	95	7.5E-22
n - 4	-1.95	0.16	-12.2	95	4.4E-20
4 - 3	-1.22	0.16	-7.6	95	1.8E-10
n - 2	-1.00	0.16	-6.2	95	1.2E-07
4 - 2	0.95	0.16	5.9	95	4.8E-07
3 - 1	-0.86	0.16	-5.4	95	5.1E-06

Pairwise Comparisons between Factor-Level Means

$$Diff = \bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}$$

$$df = n_T - r = rn - r = 5*20 - 5 = 95$$

$$stdError = \sqrt{\frac{MS_{error}}{n_i} + \frac{MS_{error}}{n_j}} = \sqrt{\frac{2 \cdot MS_{error}}{n}} = \sqrt{\frac{2*0.257}{20}} = 0.1604$$

$$t\text{-}value = \frac{Diff}{stdError}$$

Student-T with Bonf. Correction

p-value = m * tdist(t-value, df, two-sided)

= 10 * tdist(t-value, 95, 2)

Other Post-Hoc Corrections

- Holm -Sidak (really Bonferroni done "right")
  - Order the p-values from smallest to largest
  - Compare the smallest p-value to α/k (regular Bonferroni)
  - If that p-value is less than α/k, then accept that alternative hypothesis
  - Now look at the next smallest p-value at α / (k − 1)
  - Continue until the p-value is not smaller than the modified value
  - At that point, stop and accept all the rest as null hypotheses

Other Post-Hoc Corrections

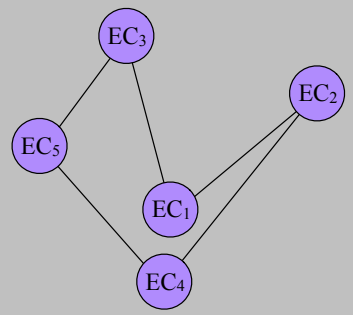
- Tukey
  - Used when comparing ***all*** pair-wise differences
    - produces narrower confidence intervals than Bonferonni in this situation
    - usual situation when trying to order results
      - e.g. comparing 5 different EC systems
      - Found out that $EC_3 > EC_2 \mid EC_5 > EC_1 > EC_4$
      - Note: Although there are 4 comparison symbols above, there are really 6 comparisons (3v2, 3v5, 2v5, 2v1, 5v1, 1v4)
      - actually there are 5C2 = 10 implicit comparisons
        - because we did not know how many comparisons there would be a priori

Note:

Pair-wise statistical comparisons form a partial order
Consequently best represented as a DAG not a list

- T

E.g.: **$EC_3 \mid EC_2 \mid EC_5 \mid EC_1 \mid EC_4$**
Yet **($EC_3 > EC_5$, EC_1 , EC_4)**, **($EC_2 > EC_1$, EC_4)** and **($EC_5 > EC_4$)**

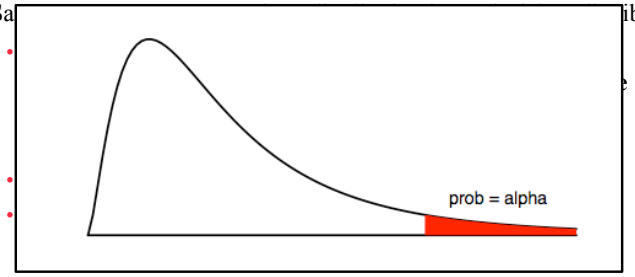


Other Post-Hoc Corrections

- Tukey
  - Same as T test except uses the *q* distribution instead of the *t* distribution
    - $q(1 - \alpha, r, n_T - r)$ value is the cut off value
      where the difference observed would be **less than this value**
      with a probability of $1 - \alpha$
      if *r* values are sampled from a normal distribution $N(0,1)$
    - $DofF = n_T - r$
    - *q* distribution is called the studentized range distribution
      - *q* "broader" than *t,*
      - *q* is not as "broad" as *t* after Bonferroni correction
    - *q* distribution is not in Excel,
      but it is in most other stats packages including R

Other Post-Hoc Corrections

- Tukey
  - Sa… …ibution
    - …
    - …
    - …



If the computed standardized difference is larger than *q,*
where *q* is the largest distance one would expect from a normal distribution,
then the difference is statistically real (with confidence level $1 - \alpha$)

Other Post-Hoc Corrections

- Many others
  - Scheffé
    - used when comparing pairs, and triples and quadruples etc., not just pairs
  - many many others
    - Duncan's multiple range test
    - The Nemenyi test
    - The Bonferroni–Dunn test
    - Newman-Keuls post-hoc analysis

Important Topics Not Covered

No time

Important Topics Not Covered

- Multifactor ANOVA (MANOVA)
  - Main Effects vs Interaction Terms
  - F tests to determine which factors are statistically significant (validating the model
  - T tests to compare treatments
- Regression
  - Multivariate regression, Polynomial Regression
  - Confidence Intervals around model parameters
  - Statistical Testing for factor relevance
  - Correlation Coefficients: r, r^2, adjusted r^2
- How to perform ANOVA as a multivariate regression
  - Indicator Variables

Important Topics Not Covered

- Testing for equality (homogeneity) of variance across different factor-levels / treatments
  - Levene's Test
- Correcting for inequality of variance
  - Convert to multivariate regression using indicator variables
  - Perform Weighted Least Squares
- How to perform ANOVA when using different test functions
  - Test functions as *blocking variables*
  - Non-parametric blocking
- What if one EC system has parameters the other EC system doesn't?
  - Nesting factor analysis

References: Books

- Mathematical statistics with applications
  - *Dennis D. Wackerly, William Mendenhall, Richard L. Scheaffer.*
  - *Boston : Duxbury Press, (6th Ed.)*
  - Introductory material - probability distributions, simple sample statistics
  - Easy to understand concrete proofs and examples - good exercises
- Applied linear statistical models
  - *Michael H. Kutner, Christopher J. Nachtsheim, John Neter, William Li*
  - *Boston: McGraw-Hill Irwin, 2005. (5th Ed.)*
  - Advanced Regression techniques, ANOVA, and GLM
- Nonparametric statistical methods
  - *Myles Hollander and Douglas A. Wolfe.*
  - *New York: Wiley, 1973*
  - Classic nonparametric statistics textbook (very practical)

Online Resources

- **Websites**
  - Wikipedia (various pages)
    - http://en.wikipedia.com
  - HyperStat Online
    - http://davidmlane.com/hyperstat
  - Mathworld
    - http://mathworld.wolfram.com/