Comparing two different Evolutionary Algorithms

<table>
<thead>
<tr>
<th>Env 1</th>
<th>-75</th>
<th>0</th>
<th>75</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-5.7</td>
<td>7.9</td>
<td></td>
</tr>
<tr>
<td>Env 2</td>
<td>-75</td>
<td>0</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>-6.2</td>
<td>5.9</td>
<td></td>
</tr>
<tr>
<td>Env 3</td>
<td>-75</td>
<td>0</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>-5.4</td>
<td>19.1</td>
<td></td>
</tr>
</tbody>
</table>

EA1 Avg. | EA2 Avg. | Number of Reps = 5

Sampling From Two Normal Distributions

<table>
<thead>
<tr>
<th>Env 1</th>
<th>-75</th>
<th>0</th>
<th>75</th>
<th>Variation Expected: σ = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-5.7</td>
<td>7.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Env 2</td>
<td>-75</td>
<td>0</td>
<td>75</td>
<td>Variation Expected: σ = 10</td>
</tr>
<tr>
<td></td>
<td>-6.2</td>
<td>5.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Env 3</td>
<td>-75</td>
<td>0</td>
<td>75</td>
<td>Variation Expected: σ = 50</td>
</tr>
<tr>
<td></td>
<td>-5.4</td>
<td>19.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

True Avg. | -10 | +10 | Number of Reps = 5
Sampling From Two Normal Distributions

<table>
<thead>
<tr>
<th>Env 1</th>
<th>-75</th>
<th>0</th>
<th>75</th>
<th>Variation Expected: $\sigma = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-10.7</td>
<td>9.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Env 2</td>
<td>-75</td>
<td>0</td>
<td>75</td>
<td>Variation Expected: $\sigma = 10$</td>
</tr>
<tr>
<td></td>
<td>-9.7</td>
<td>10.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Env 3</td>
<td>-75</td>
<td>0</td>
<td>75</td>
<td>Variation Expected: $\sigma = 50$</td>
</tr>
<tr>
<td></td>
<td>-2.5</td>
<td>7.9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

True Avg. -10 +10  Number of Runs = 100

Basic Statistical Tests

Part 1 - Point Estimation:
Finding the Mean using Confidence Intervals

What Are We Interested In?

- For most statistical analysis for EC 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

- The system has a true mean ... but where is it?
- Show a range within which the true mean ($\mu$) likely lies
  - Called the *confidence interval*
- Also provide the probability that $\mu$ lies with the CI
  - Called the *confidence level*

Example:
There is a 99% chance (the confidence level) that $\mu$ lies in the interval $[-3, 47]$
Normal Distribution

- Most common distribution used is the normal distribution
- a.k.a. Gaussian Distribution
- a.k.a. Bell Curve

\[ f_{\mu, \sigma}(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \sim \frac{1}{e^{x^2}} \]

\[ X \sim N(\mu, \sigma^2) \text{ pdf in R: } \text{dnorm}(x, \mu, \sigma) \]
\[ \text{cdf in R: } \text{pnorm}(x, \mu, \sigma) \]

Most common distribution found in nature thanks to the Central Limit Theorem

- Most common distribution used is the normal distribution
- a.k.a. Gaussian Distribution
- a.k.a. Bell Curve

\[ f_{\mu, \sigma}(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \sim \frac{1}{e^{x^2}} \]

Standard Normal Distribution
\[ X \sim N(0, 1) \text{ pdf in R: } \text{dnorm}(x) \]
\[ \text{cdf in R: } \text{pnorm}(x) \]

Distribution of the Average
(of a normally distributed system)

The original distribution

Average of 5 samples

Average of 25 samples

Average of 100 samples

The mean
\[ \mu_X = \frac{1}{n} \sum_{i=1}^{n} p_i \cdot x_i \]

The Standard Deviation
\[ \sigma_X = \sqrt{E((X - \mu_X)^2)} \]

The average
\[ \bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

Variation around the average
\[ s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})^2 \]

The Sample Standard Deviation
\[ s_X = \frac{\sigma_X}{\sqrt{n}} \]

the variation of the ‘averages’ around the true mean
is less than
the variation of the original values around the true mean
Confidence Intervals

• Of course, we don’t know the true mean, \( \mu \), or true standard deviation, \( \sigma \)

• 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 as well as the size of the “finger” is known exactly!
  • We can determine the odds that the true mean lies within a specified range of \( \bar{X} \)

Confidence Intervals

\[
Z = \frac{\bar{X} - \mu_X}{\sigma_X} = \frac{\bar{X} - \mu_X}{\frac{s_X}{\sqrt{n}}}
\]

But the denominator is no longer a scaler!

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 dev to give it a std dev of 1
  • also gives a variance of 1

\[
Z = \frac{\bar{X} - \mu_X}{\sigma_X} = \frac{\bar{X} - \mu_X}{\frac{s_X}{\sqrt{n}}}
\]

Standard Deviation of the Normal Distribution: The Chi Distribution

\[
f_k(x) \propto \frac{x^{k-1}}{e^{x^2/2}} \quad k = \text{number of samples}
\]

from http://en.wikipedia.org/wiki/Chi_distribution
Confidence Intervals

- Want to find \( \mu \) the true mean in terms of the average
  - But we have not one but two unknowns - \( \sigma \) is also unknown
  - One equation - two unknowns - not good!!!
  - Trick - divide by the known sample standard deviation \( s \) instead of \( \sigma \)

\[
Z = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}
\]

\[
T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}
\]

So we have a normal divided by a chi distribution
This has a Student’s T distribution!

\[
f_{df}(x) \propto \left(1 + \frac{x^2}{df}\right)^{-\frac{df+1}{2}} \approx \frac{1}{\left(\frac{2}{df}\right)^{\frac{df}{2}}} \quad df = n - 1
\]


Student T distribution

\( CL = 0.95 \)
\( n = 5 \)
\( df = 4 \)

Confidence Interval

Student T vs Normal (Gaussian)

- Student T is “broader” than the Normal
- Student T goes to 0 much more slowly than the Normal
  (has substantial probability of very large values)
Confidence Intervals

Student T distribution

\[ T = \frac{X - \mu}{s} \]

\[ \frac{X - \mu}{s} \]

\[ \frac{X - \mu}{s} \]

\[ n = 5 \]

\[ df = 4 \]

\[ \alpha = 1 - CL \]

\[ CL = 0.95 \]

\[ 95\% \] chance a randomly generated value from a T distribution will fall inside the CI (grey area)

\[ \frac{\alpha}{2} \]

\[ \frac{\alpha}{2} \]

\[ -2.0 \]

\[ 2.0 \]

Confidence Level → \( 99\% \)

\[ \alpha = 0.01 \]

\[ CL = 0.99 \]

\[ n = 5 \]

\[ df = 4 \]

\[ 99\% \]

\[ \frac{\alpha}{2} \]

\[ \frac{\alpha}{2} \]

\[ -4.6 \]

\[ 4.6 \]

The confidence interval expands as the confidence level increases

i.e. it is more likely that a t-value (e.g. the true mean) will fall within a larger CI than a smaller one

\[ \alpha = 0.01 \]

\[ CL = 0.99 \]

\[ n = 55 \]

\[ df = 54 \]

\[ 99\% \]

\[ \alpha = 0.01 \]

\[ CL = 0.99 \]

\[ n = 55 \]

\[ df = 54 \]

\[ 99\% \]

\[ \frac{\alpha}{2} \]

\[ \frac{\alpha}{2} \]

\[ -2.7 \]

\[ 2.7 \]

The confidence interval expands as the confidence level increases

i.e. more samples produces a smaller CI at the same CL

Samples → 55

\[ \frac{\alpha}{2} \]

\[ \frac{\alpha}{2} \]

\[ -2.7 \]

\[ 2.7 \]
Confidence Intervals

But how do we know what the CI values are?

In general the CI can be represented as ... ±t_{α, df}.

Calculating the cut off t_{α, n−1} values using Excel: =TINV(α, n - 1)
using R: -qt(1-α/2, n - 1)

This creates CIs for the T distribution with a mean of 0

Confidence Intervals

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

\[ T = \frac{\bar{X} - μ_x}{s_x} \]

So we have a normal divided by a chi distribution
This has a Student’s T distribution!

Confidence Intervals

Student T distribution

α = 0.01
CL = 0.99
n = 55
df = 54

\[ -t_{α, df} \leq \frac{μ_x - \bar{X}}{s_x} \leq +t_{α, df} \]

Confidence Interval

\[ \bar{X} - t_{α, df} \frac{s_x}{\sqrt{n}} \leq μ_x \leq \bar{X} + t_{α, df} \frac{s_x}{\sqrt{n}} \]

Rearranging terms to isolate μ_x

Student T distribution

α = 0.01
CL = 0.99
n = 55
df = 54

Confidence Intervals
Estimating the Mean: Confidence Intervals Around the Average

- Confidence Intervals can be written in 3 equivalent ways

Error Bounds
\[ \mu_X = \bar{X} \pm t_{\alpha/2, n-1} \frac{s_X}{\sqrt{n}} \]

Confidence Intervals
\[ \bar{X} - t_{\alpha/2, n-1} \frac{s_X}{\sqrt{n}} \leq \mu_X \leq \bar{X} + t_{\alpha/2, n-1} \frac{s_X}{\sqrt{n}} \]
\[ \mu_X \in \left[ \bar{X} - t_{\alpha/2, n-1} \frac{s_X}{\sqrt{n}}, \bar{X} + t_{\alpha/2, n-1} \frac{s_X}{\sqrt{n}} \right] \]

Example:
- An experimenter runs a New Evolutionary Algorithm on a TSP
- At the end of each run, the smallest length tour that had been found during the run was recorded
- NEA is run 50 times on the same TSP problem
- On average NEA 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

- From the problem we know that the average NEA run produced tours of
  \[ \bar{X} = 272 \] that had \( s_X = 87 \)
  We know that
  \[ \mu_X = \bar{X} \pm t_{\alpha/2, n-1} \frac{s_X}{\sqrt{n}} \]
- Also from the problem \( n = 50 \) and \( \alpha = (1 - 0.99) = 0.01 \)
- so the \( \pm \) cutoff value is \( t_{0.005, 49} \)
- using Excel/R we see that TINV(0.01, 49) = qt(0.995, 49) = 2.68
  \[ \mu_X = 272 \pm 2.68 \frac{87}{\sqrt{50}} = 272 \pm 33 \]
  i.e. there is only a 1% chance that the true mean lies outside the confidence interval formed around average
- and so \( 239 \leq \mu_X \leq 305 \) with a 99% C.L.

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

<table>
<thead>
<tr>
<th>µ</th>
<th>σ</th>
<th>n</th>
<th>X</th>
<th>s_X</th>
<th>1.96 Xσ/√n</th>
<th>Lower</th>
<th>Uppe</th>
</tr>
</thead>
<tbody>
<tr>
<td>+10</td>
<td>10</td>
<td>100</td>
<td>10.5</td>
<td>10.0</td>
<td>3.3</td>
<td>7.2</td>
<td>13.8</td>
</tr>
<tr>
<td>-10</td>
<td>10</td>
<td>100</td>
<td>-9.7</td>
<td>10.1</td>
<td>3.3</td>
<td>-13.1</td>
<td>-6.4</td>
</tr>
</tbody>
</table>

Confidence Interval Example

Improving the Sensitivity: The Student $t$ Test

- The Student $t$ Test is the basic test used in statistics
- Idea: Gain sensitivity by looking at the difference between the means of the two systems
The Student $t$ Test

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

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

Distribution again differs for different sample sizes

- Degrees of Freedom is now \((n_1 - 1) + (n_2 - 1) = 2n - 2\)
- $t$ test either succeeds or fails
- $t$ value greater than cutoff for a given C.L. or not

Based on 50 runs $\alpha = 0.01$

$99\%$ $0 \quad 2.68 - 2.68$

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
- \(1 - \Pr(T < t \text{ score})\)
- Called the $p$-value

Based on 50 runs

$0 \quad 0 \quad 0.5 \quad 0.15 \quad 0.01 \quad 2.4$

$99\%$ $0 \quad 2.68 - 2.68$

$t$ 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:
   \[ T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2_{X_1}}{n_1} + \frac{s^2_{X_2}}{n_2}}} \]
5. Compute the $p$-value
   - $p$-value = the area under the $t$ distribution outside \([-T, T]\)
   - In Excel: =TDIST($T$, 2*$n$ - 2, 2)
     - The final “2” in Excel means “two-sided”
   - In R: > 2*pt(-$T$, 2*$n$ - 2)

Variance Assumptions and the T Test

\[ \sigma_1 = \sigma_2 = \sigma\] and \(n_1 = n_2 = n\)

\[ T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2_{X_1}}{n_1} + \frac{s^2_{X_2}}{n_2}} / n} \]

\[ \sigma_1 = \sigma_2 = \sigma\] but \(n_1 \neq n_2\)

\[ T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(n_1 - 1)s^2_{X_1} + (n_2 - 1)s^2_{X_2}} / (n_1 + n_2 - 2) / \sqrt{n_1 + n_2}} \]

In Excel: =ttest(A1:A50, B1:B50, 2, 2)
Variance Assumptions and the T Test

\[ T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \]

\( \Rightarrow \) Approximate variance not pooled

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

In Excel: \( = \text{ttest}(A1:A50, B1:B50, 2, 3) \)

\[ \text{t.test(): Welch’s vs Student’s} \]

\( n = 80 \) for both OEA and NEA

\[ > \text{t.test(OEA, NEA)} \]

Welch Two Sample t-test

data: OEA and NEA
t = -2.2549, df = 152.68, p-value = 0.02556
alternative hypothesis:
true difference in means is not equal to 0
95 percent confidence interval:
-4.7621535 to -0.3143734
average of OEA
5.119665
average of NEA
7.657929

\[ \text{t.test(): Welch’s vs Student’s} \]

\( n = 80 \) for both OEA and NEA

\[ > \text{t.test(OEA, NEA, var.equal=TRUE)} \]

Two Sample t-test

data: OEA and NEA
t = -2.2549, df = 158, p-value = 0.02551
alternative hypothesis:
true difference in means is not equal to 0
95 percent confidence interval:
-4.7621535 to -0.3143734
average of OEA
5.119665
average of NEA
7.657929

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 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)

Non-Parametric Statistics

- Basic Idea
  - Sort the data and then rank them
  - Use Ranks instead of actual values to perform statistics
- 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

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 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

Ranked Example

Two data sets combined into a single array

Give each data element its corresponding rank

Perform t test on Ranks

Resort by treatment

Perform t test on Ranks

Ranked Example

\[
s_T = \sqrt{\frac{s_A^2 + s_B^2}{n_A} \cdot \frac{1}{n_B}}
\]

\[
(t_k \text{ score}) = \frac{\text{avg}_A - \text{avg}_B}{s_T} \cdot t_k
\]

p-value

\[
\text{Ranked t Test}
\]

\[
\text{avg}_A = 7.85, \quad \text{avg}_B = 13.15
\]

\[
\text{stdDev}_A = 5.28, \quad \text{stdDev}_B = 5.33
\]

\[
n = 10
\]

\[
(t_k \text{ score}) = 2.23
\]

\[
p-value = 0.038
\]
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 \( u = 1 - l \)
  • Confidence Interval is \([value_l, value_u]\)
    • i.e. \( value_l \leq median \leq value_u \)
    • With a confidence level of \( 1 - \alpha \)

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 \( u = 1 - l \)
  • Confidence Interval is \([value_l, value_u]\)
    • i.e. \( value_l \leq median \leq value_u \)
    • With a confidence level of \( 1 - \alpha \)

Box Plot: Example

<table>
<thead>
<tr>
<th>Value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.03</td>
</tr>
<tr>
<td>2</td>
<td>0.03</td>
</tr>
<tr>
<td>3</td>
<td>0.03</td>
</tr>
<tr>
<td>4</td>
<td>0.08</td>
</tr>
<tr>
<td>5</td>
<td>0.16</td>
</tr>
<tr>
<td>6</td>
<td>0.16</td>
</tr>
<tr>
<td>7</td>
<td>0.16</td>
</tr>
<tr>
<td>8</td>
<td>0.16</td>
</tr>
<tr>
<td>9</td>
<td>0.16</td>
</tr>
<tr>
<td>10</td>
<td>0.27</td>
</tr>
<tr>
<td>11</td>
<td>0.27</td>
</tr>
<tr>
<td>12</td>
<td>0.64</td>
</tr>
<tr>
<td>13</td>
<td>0.64</td>
</tr>
<tr>
<td>14</td>
<td>0.64</td>
</tr>
<tr>
<td>15</td>
<td>0.64</td>
</tr>
<tr>
<td>16</td>
<td>0.91</td>
</tr>
<tr>
<td>17</td>
<td>0.91</td>
</tr>
<tr>
<td>18</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Sort
Data
Effect Size and Repetitions
Does My Difference Matter?

• Okay, so your results are significantly better than the published results. So what?
  • Statistics can answer, “is it better?”, but not “does it matter?”
• You perform 100,000 runs of your classifier and 100,000 runs of the reference classifier
  • You get a t-score of 31.6! 😊
  • The p-score is reported by Excel as 0! (Actually 2.0 x 10^-219)
  • But…your way classifies data at 91.0% accuracy, whereas the reference technique classifies at 90.8% accuracy.
  • Not much difference!
    • Especially if your technique is much slower than the reference way

Measuring Effect Size

• One statistic for effect size: Cohen’s $d'$
  • $d'$ is computed by $d' = \frac{t}{\sqrt{(n_1 + n_2)/2}}$
  • Measures the difference between means in terms of the pooled standard deviation
  • Cohen suggests that 0.25 is a small difference; 0.50 is a medium-sized difference; 0.75 is a large difference
  • For our example, $d'$ is 0.10
    • Essentially an insignificant difference
• Problem: we did too many runs!

Repetitions

• What is the number of repetitions needed to see if there is a difference between two means or between two medians?
  • Depends on the underlying distributions
  • But underlying distributions are unknown
• Rule of thumb for t-tests…
  • Perform a minimum of 30 repetitions for each system
  • Performing 50 to 100 repetitions is usually better

ANOVA: Analysis of Variance

• Part 1a: Multi-Level Analysis
  Basic Concept
More Than 2 Levels

- Preceding stats to be used for simple experiment designs
- More sophisticated stats needs to be done if:
  - Comparing multiple systems instead of just 2 systems
    - 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 levels
  - Called a ‘multi-level’ analysis

Analysis of Variance (ANOVA)

<table>
<thead>
<tr>
<th></th>
<th>no xover</th>
<th>xover = 1pt</th>
<th>xover = 2pt</th>
<th>xover = 3pt</th>
<th>xover = 4pt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fitness Values</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>avg fitness</td>
<td>4.02</td>
<td>8.13</td>
<td>5.09</td>
<td>7.02</td>
<td>5.76</td>
</tr>
<tr>
<td>std dev</td>
<td>0.451</td>
<td>0.313</td>
<td>0.424</td>
<td>0.478</td>
<td>0.471</td>
</tr>
</tbody>
</table>

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 means)
- From here on we will be comparing variances
  - This is why it is called “Analysis of Variance”
  - Remember - compare the ratio of variances
    - see if it equals 1
    - distribution known: F distribution

The F Test

$$F^* = \frac{S^2_x}{S^2_y}$$

ANOVA: Discrete Levels

Average of $Y$ (no model)

Variance of $Y$ (no model) represented as a std deviation

Add average for each level, a model of the behavior of the system

Subtract the level average from each level, leaving the residuals (errors)

$$s^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2$$

levels: $r = 5$
reps per level: $n = 20$

total reps: $n_T = 100$
Compute the Variance of the Residuals

\[ MS_{\text{total}} \]
\[ MS_{\text{error}} \]

levels: \( r = 5 \)
reps per level: \( n = 20 \)

\[ F^* = \frac{MS_{\text{total}}}{MS_{\text{error}}} \]
ANOVA: Discrete Levels

Problem: Variances must be independent for the $F$ test

$$F^* = \frac{SS_{\text{total}} - SS_{\text{error}}}{df_{\text{total}} - df_{\text{error}}} / MS_{\text{error}}$$

levels: $r = 5$
reps per level: $n = 20$
total reps: $n_T = 100$

MSerror

ANOVA: Discrete Levels

Assumption:
- variance for every level is the same and equals $\sigma^2$

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 for the $i$th level
  - Generalized ANOVA using regression

ANOVA table for example

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F-ratio</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>1</td>
<td>3592.9</td>
<td>3592.9</td>
<td>13967</td>
<td>$\leq 0.0001$</td>
</tr>
<tr>
<td>xover</td>
<td>4</td>
<td>210.9</td>
<td>52.7</td>
<td>204.94</td>
<td>$\leq 0.0001$</td>
</tr>
<tr>
<td>Error</td>
<td>95</td>
<td>24.4</td>
<td>0.257</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>99</td>
<td>235.3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

$F$ test (From Excel)

$fdist(204.94, 4, 95) = 8.19E-46$

Non-parametric ANOVA

- Again, what happens if $Y$ (or actually $\varepsilon$) 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 ($DoF = r - 1$) for large $n$ or $r$ ($n > 15$ or $r > 4$)
    - a special purpose distribution for small $n$ or $r$
**ANOVA: Analysis of Variance**

Part 1b: 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*

  
  \[
  t\text{ value} = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{s^2_{x_i}}{n_1} + \frac{s^2_{x_j}}{n_2}}}
  \]

  original T test comparison

  \[
  t\text{ value} = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}}
  \]

  comparing level \(i\) with level \(j\) across the ANOVA model

---

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

\[
\sigma^2 = \frac{\text{MSE}}{n}
\]
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 $0.95 \times 0.95 \times 0.95 \times \ldots \times 0.95 = (0.95)^6 = 0.735$
  - 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 $\alpha$
  \[ \alpha' = \frac{\alpha}{m} \]
  - 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)

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

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

<table>
<thead>
<tr>
<th></th>
<th>Diff</th>
<th>std. err</th>
<th>t-value</th>
<th>df</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>n - 1</td>
<td>-4.04</td>
<td>0.16</td>
<td>-25.2</td>
<td>95</td>
<td>7.7E-43</td>
</tr>
<tr>
<td>n - 3</td>
<td>-3.18</td>
<td>0.16</td>
<td>-19.8</td>
<td>95</td>
<td>1.7E-34</td>
</tr>
<tr>
<td>2 - 1</td>
<td>-3.04</td>
<td>0.16</td>
<td>-19.0</td>
<td>95</td>
<td>4.8E-33</td>
</tr>
<tr>
<td>3 - 2</td>
<td>2.16</td>
<td>0.16</td>
<td>13.6</td>
<td>95</td>
<td>6.0E-23</td>
</tr>
<tr>
<td>4 - 1</td>
<td>-2.09</td>
<td>0.16</td>
<td>-13.0</td>
<td>95</td>
<td>7.5E-22</td>
</tr>
<tr>
<td>n - 4</td>
<td>-1.95</td>
<td>0.16</td>
<td>-12.2</td>
<td>95</td>
<td>4.4E-20</td>
</tr>
<tr>
<td>4 - 3</td>
<td>-1.22</td>
<td>0.16</td>
<td>-7.6</td>
<td>95</td>
<td>1.8E-10</td>
</tr>
<tr>
<td>n - 2</td>
<td>-1.00</td>
<td>0.16</td>
<td>-6.2</td>
<td>95</td>
<td>1.2E-07</td>
</tr>
<tr>
<td>4 - 2</td>
<td>0.95</td>
<td>0.16</td>
<td>5.9</td>
<td>95</td>
<td>4.8E-07</td>
</tr>
<tr>
<td>3 - 1</td>
<td>-0.86</td>
<td>0.16</td>
<td>-5.4</td>
<td>95</td>
<td>5.1E-06</td>
</tr>
</tbody>
</table>
## Pairwise Comparisons between Factor-Level Means

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

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

**Student-T with Bonf. Correction**

\[
\text{t-value} = \frac{\text{Diff}}{\text{stdError}} = \frac{Y_i \cdot - Y_j \cdot}{\sqrt{\frac{MS_{error}}{n_i} + \frac{MS_{error}}{n_j}}} = \frac{2 \cdot MS_{error}}{n} = \sqrt{\frac{2 \times 0.257}{20}} = 0.1604
\]

\[
\text{df} = n_{T} - r = r \cdot n - r = 5 \times 20 - 5 = 95
\]

\[
\text{p-value} = m \times \text{tdist(t-value, df, two-sided)} = 10 \times \text{tdist(t-value, 95, 2)}
\]

### Other Post-Hoc Corrections

- **Tukey**
  - Used when comparing all pair-wise differences
  - produces narrower confidence intervals than Bonferroni in this situation
  - usual situation when trying to order results
  - e.g. comparing 5 different EC systems
  - Found out that $EC_3 > EC_2 | EC_3 > EC_1 > EC_4$
  - Note: Although there are 4 comparison symbols above, there are really 6 comparisons
  - actually there are $5C2 = 10$ implicit comparisons
    - because we did not know how many comparisons there would be apriori

- **Holm - Sidak (really Bonferroni done “right”)**
  - Order the p-values from smallest to largest
  - Compare the smallest p-value to $\alpha/k$ (regular Bonferroni)
  - If that p-value is less than $\alpha/k$, then accept that alternative hypothesis
  - Now look at the next smallest p-value at $\alpha / (k - 1)$
  - 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 Bonferroni in this situation
    - usual situation when trying to order results
    - e.g. comparing 5 different EC systems
    - Found out that $EC_3 > EC_2 | EC_3 > EC_1 > EC_4$
    - Note: Although there are 4 comparison symbols above, there are really 6 comparisons
    - actually there are $5C2 = 10$ implicit comparisons
      - because we did not know how many comparisons there would be apriori
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 \( N(0,1) \)
  - \( \text{DofF} = n_T - r \)
  - \( q \) 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

- 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
ANOVA: Analysis of Variance

Part 2: Multi-Factor ANOVA

Main Effects

Interaction Effects

Multiple Factors: Factorial Design

E.g. if we have 2 EC systems, new and standard (New and Std) and we want to see their behavior under

- crossover and no crossover (x and x)
- 3 different selection pressures (p1, p2 and p3)

<table>
<thead>
<tr>
<th></th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>t6</th>
<th>t7</th>
<th>t8</th>
<th>t9</th>
<th>t10</th>
<th>t11</th>
<th>t12</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>new</td>
<td>new</td>
<td>new</td>
<td>new</td>
<td>new</td>
<td>new</td>
<td>std</td>
<td>std</td>
<td>std</td>
<td>std</td>
<td>std</td>
<td>std</td>
</tr>
<tr>
<td>X</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>P</td>
<td>p1</td>
<td>p2</td>
<td>p3</td>
<td>p1</td>
<td>p2</td>
<td>p3</td>
<td>p1</td>
<td>p2</td>
<td>p3</td>
<td>p1</td>
<td>p2</td>
<td>p3</td>
</tr>
</tbody>
</table>

Statistical Terminology

factor: dependent variable (not-stochastic)

levels: values that the factors can equal

- S has 2 levels: new, std
- P has 3 levels: p1, p2, p3

treatment: an instantiation where each factor is set to a particular level

- S = std; X = x; P = p2

Two Factor Analysis

- What do we want to know?
  - Whether the new system is better than the old system overall?
  - Whether the performance is better using crossover or without?
  - But probably also…
    - The new system is better than the old system given that crossover is used
    - The old system is better than the new system given that crossover is not used
    - This is called an interaction
Two Factor Analysis

• What do we want to know?
  • Factor A main effect
  • Factor B main effect
  • But probably also…
    • Factor A and Factor B levels interact
    • Called an interaction term

• Linear Model
  • \( Y = A + B + AB + \varepsilon \)

Multi-Factor ANOVA: Results Report

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Const</td>
<td>1</td>
<td>16970</td>
<td>16970</td>
<td>12930</td>
<td>≤ 0.0001</td>
</tr>
<tr>
<td>S</td>
<td>1</td>
<td>113</td>
<td>113</td>
<td>86.5</td>
<td>≤ 0.0001</td>
</tr>
<tr>
<td>X</td>
<td>1</td>
<td>775</td>
<td>775</td>
<td>591.0</td>
<td>≤ 0.0001</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>939</td>
<td>469.5</td>
<td>357.7</td>
<td>≤ 0.0001</td>
</tr>
<tr>
<td>S*X</td>
<td>1</td>
<td>4.05</td>
<td>4.05</td>
<td>3.1</td>
<td>0.0809</td>
</tr>
<tr>
<td>S*P</td>
<td>2</td>
<td>307</td>
<td>153.5</td>
<td>116.8</td>
<td>≤ 0.0001</td>
</tr>
<tr>
<td>X*P</td>
<td>2</td>
<td>0.570</td>
<td>0.285</td>
<td>0.217</td>
<td>0.8049</td>
</tr>
<tr>
<td>S<em>X</em>P</td>
<td>2</td>
<td>0.308</td>
<td>0.154</td>
<td>0.117</td>
<td>0.8892</td>
</tr>
<tr>
<td>Error</td>
<td>168</td>
<td>220.5</td>
<td>1.312</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>179</td>
<td>2360.12</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( n_T = 180 \)
\( a = 2 \)
\( b = 2 \)
\( c = 3 \)
\( n = 15 \)

Part 3

Regression by means of Least Squares

Fitness (\( F \))

Factor in Population Size

\[ E(\varepsilon_F) = 0 \]
\[ V(\varepsilon_F) = \sigma_F^2 \]

\[ F_i = 72 + \varepsilon_F \]
Linear Regression

**Population Size (p)**

Fitness ($F$)

- $F_i = 0.12 p_i + \epsilon$

**Factor in Population Size**

- $E(\epsilon) = 0$
- $V(\epsilon) = \sigma^2$

Modeling Response Behavior: Treating $X$ as a factor

- Simplest model - linear relationship
  
  $Y_i = f(x_i) + \epsilon$  with  $f(x_i) = \beta_0 + \beta_1 x_i$

  $Y_i = \beta_0 + \beta_1 x_i + \epsilon$

  Two parameters $\beta_0$ and $\beta_1$
  define the function

Linear Regression by Means of Least Squares

- Idea:
  - From sample pairs \{(\(Y_1, x_1\), (\(Y_2, x_2\), \ldots, (\(Y_n, x_n\))\}
  determine $b_0$, $b_1$

  - Estimates of the two unknowns $\beta_0$, $\beta_1$

  $Y_i = \beta_0 + \beta_1 x_i + \epsilon$  $\iff$  $\hat{Y}_i = b_0 + b_1 x_i$

  - chosen such that the sum of squared errors is minimized
  - i.e. find the model that has the smallest (least) total squared error
Linear Regression by Means of Least Squares

- Idea:
  - From sample pairs \( \{(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)\} \)
  - Estimate unknowns \( \beta_0 \) and \( \beta_1 \)

\[ Y_i = \beta_0 + \beta_1 x_i \]

- chosen such that the sum of squared errors is minimized

\[ \text{Error} = e_i = Y_i - \hat{Y} \]
\[ \text{Error} = e_i = Y_i - b_0 - b_1 x_i \]

\[ \text{Squared Error} = e_i^2 = (Y_i - b_0 - b_1 x_i)^2 \]

\[ \text{Sum of Squared Errors} = SSE = \sum_{i=1}^{n} e_i^2 \]

Find the linear function

\[ \hat{Y}_i = b_0 + b_1 x_i \]

Poor choice

Sum of squared error is large
Linear Regression by Means of Least Squares

Better

Sum of squared error reduced

Best

Minimized

Sum of squared error
Linear Regression by Means of Least Squares

- Determine \( \hat{Y}_i = b_0 + b_1 X_i \)
- Find \( b_0, b_1 \) such that
  \[
  \min \sum_{i=1}^{n} e_i^2 = \min \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2
  \]
- Use calculus (minimum finding)
  - Take partial derivatives wrt \( b_0 \) and \( b_1 \)
  - set to zero
  - two equations, two unknowns ... solve

\[
\hat{Y}_i = b_0 + b_1 X_i
\]

Linear Regression by Means of Least Squares

- Determine \( \hat{Y}_i = b_0 + b_1 X_i \)
- Solution
  \[
  b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
  = \frac{\text{cov}(x,Y)}{\text{var}(x)} = \frac{S_{xy}}{S_x^2}
  \]

\[
b_0 = \bar{Y} - b_1 \bar{x}
\]

What are the distributions of \( b_1 \) and \( b_0 \)?

- \( b_1 \) can be rewritten as
  \[
b_1 = \sum_{i=1}^{n} k_i Y_i \quad \text{where} \quad k_i = \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}
  \]

and \( b_0 = \bar{Y} - b_1 \bar{x} \)

- since the \( x_i \) are constant
  \( b_1 \) is a linear combination of \( Y_i \)'s
- linear combinations of normally distributed random variables are normally distributed
- SO ...

What are the distributions of \( b_1 \) and \( b_0 \)?

- \( b_1 \) can be rewritten as
  \[
b_1 = \sum_{i=1}^{n} k_i Y_i
  \]
  \( \text{if } Y \text{ is normally distributed, } b_1 \text{ is normally distributed} \)

and \( b_0 = \bar{Y} - b_1 \bar{x} \)

- since the \( x_i \) are constant
  \( b_1 \) is a linear combination of \( Y_i \)'s
- linear combinations of normally distributed random variables are normally distributed
- SO ...

\( b_1 \) (Slope) \( \text{is a random variable } \)
\( \text{i.e has a probability distribution} \)
\( b_0 \) (Y intercept) \( \text{is also a random variable} \)
Expectation and Variance of \( b_1 \) and \( b_0 \)

\( b_1 \) and \( b_0 \) can be thought of as sample means

\[
E(b_0) = \beta_0 \quad E(b_1) = \beta_1
\]

and they have associated variances

\[
V(b_1) = \frac{\sigma_Y^2}{nS_x^2} \quad \Rightarrow \quad s^2_{b_1} = \frac{MS_{\text{error}}}{nS_x^2}
\]

\[
V(b_0) = \left( 1 + \frac{x^2}{S_x^2} \right) \frac{\sigma_Y^2}{n} \quad \Rightarrow \quad s^2_{b_0} = \left( 1 + \frac{x^2}{S_x^2} \right) \frac{MS_{\text{error}}}{n}
\]

Confidence Interval around the Slope

\[
\beta_1 = b_1 \pm t_{\alpha, n-2} s_{b_1}
\]

Confidence Interval around the Intercept

\[
\beta_0 = b_0 \pm t_{\alpha, n-2} s_{b_0}
\]

Confidence Bands

\[
\hat{Y} = b_1 x + b_0 \pm k_{\alpha, n, X} \left[ S_X^2 + (x - \bar{X})^2 \right]^{1/2}
\]

\[
k_{\alpha, n, X} = \sqrt{2 \frac{F_{\alpha/2, n-2}}{nS_X^2}}
\]
T test to see if a the slope is statistically significant
- To see if the slope $b_1$ is statistically different from 0
  - use the T test
    \[ T = \frac{(b_1 - 0)}{S_{b_1}} = \frac{b_1}{S_{b_1}} \]
  - and find the corresponding p-value
  - because we originally estimated 2 parameters use
    \[ df = n - 2 - 1 = n - 3 \]

T test to see if a y intercept is statistically significant
- To see if the regression line goes through the origin check if $b_0$ is statistically different from 0
  - use the T test
    \[ T = \frac{(b_0 - 0)}{S_{b_0}} = \frac{b_0}{S_{b_0}} \]
  - and find the corresponding p-value
  - again because we originally estimated 2 parameters use
    \[ df = n - 2 - 1 = n - 3 \]

These confidence intervals and tests are very important to perform.
Yet they are not commonly done!

Part 4
Multi-factor and Polynomial Regression
Multifactor Regression

- General model for one factor
  \[ Y_i = f(x_i) + \epsilon \]

- General model for multiple factors
  \[ Y_i = f(x_{1,i}, x_{2,i}, \ldots, x_{k,i}) + \epsilon \]
  - Note: still not a multivariate analysis – error term still additive to the (now multiple) factors – factors themselves not stochastic

Multifactor Regression

- Assume linear combination of factors … simplest \( f \)
  \[ Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + \epsilon \]
  \[ \Rightarrow \hat{Y}_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + \cdots + b_k x_{k,i} \]

- Just
  - take the partial derivative of the squared error function for each parameter
  - Set each derivative to zero to find the maximum
  - Solve the set of linear equations
    - \( k \) unknown parameters, \( k \) equations

T test to see if a factor is statistically significant

- Each factor \( b_i \) has known estimated variance
  - Found analogously to \( b_1 \) and \( b_0 \)
  - To see if the factor is meaningful, see if \( b_i \) is statistically different from 0
    - using the T test
      \[ T = \frac{(b_i - 0)}{S_{b_i}} = \frac{b_i}{S_{b_i}} \]
      - find the corresponding p-value
      - because we are estimating \( k \) parameters use \( df = n - k - 1 \)

This is very important to compute!!! Yet not commonly provided.

Polynomial Regression

- One trick is to set \( x_2 = x^2, x_3 = x^3, \) etc.
  - This can be done since each factor is not a random variable, just a regular variable
  - Since it is known that any function can be formed through a linear combination of polynomial variables (a power series), we can now regress against any function!!
    - We must know the function to regress against
      - Again called the model
    - Must check to see if each term is statistically significant
      - Use T test from previous slide
      - If a term is not significant, eliminate it from the model and apply least squares again on simpler model
Polynomial Regression E.g.

R squared = 70.2%  R squared (adjusted) = 70.1%
s = 0.1466 with 1000 - 5 = 995 degrees of freedom

**Source**  **Sum of Squares**  **df**  **Mean Square**  **F-ratio**
---  ---  ---  ---  ---
Regression 50.4708 4 12.6177 587
Residual 21.3783 995 0.0215

**Variable**  **Coefficient**  **s.e. of Coeff**  **t-ratio**  **p-value**
---  ---  ---  ---  ---
Constant 0.515460 0.0236 21.9 ≤ 0.0001
X -2.27114 0.3210 -7.07 ≤ 0.0001
X^2 8.87396 1.303 6.81 ≤ 0.0001
X^3 -6.94563 1.968 -3.53 0.0004
X^4 0.331472 0.9828 0.337 0.7360

**Polynomial Regression E.g.**

R squared = 70.2%  R squared (adjusted) = 70.1%
s = 0.1466 with 1000 - 4 = 996 degrees of freedom

**Source**  **Sum of Squares**  **df**  **Mean Square**  **F-ratio**
---  ---  ---  ---  ---
Regression 50.4684 3 16.8228 784
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
X^4 0.331472 0.9828 0.337 0.7360

Polynomial Regression E.g.

Use multiple factor least squares using const, x, x^2, x^3, x^4 as factors

**Polynomial Regression E.g.**

R squared = 70.2%  R squared (adjusted) = 70.1%
s = 0.1466 with 1000 - 5 = 995 degrees of freedom

**Source**  **Sum of Squares**  **df**  **Mean Square**  **F-ratio**
---  ---  ---  ---  ---
Regression 50.4708 4 12.6177 587
Residual 21.3783 995 0.0215

**Variable**  **Coefficient**  **s.e. of Coeff**  **t-ratio**  **p-value**
---  ---  ---  ---  ---
Constant 0.515460 0.0236 21.9 ≤ 0.0001
X -2.27114 0.3210 -7.07 ≤ 0.0001
X^2 8.87396 1.303 6.81 ≤ 0.0001
X^3 -6.94563 1.968 -3.53 0.0004
X^4 0.331472 0.9828 0.337 0.7360

**Polynomial Regression E.g.**

X^4 is not statistically significant

… reduce the number of terms by one
Polynomial Regression E.g.

\[
R^2 = 70.2\% \quad R^2 \text{ (adjusted)} = 70.2\%
\]

\[
s = 0.1465 \quad \text{with} \quad 1000 - 4 = 996 \quad \text{degrees of freedom}
\]

Source of variation | Sum of Squares | df | Mean Square | F-ratio | p-value
--- | --- | --- | --- | --- | ---
Regression | 50.4684 | 3 | 16.8228 | 784 | All factors statistically significant
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

Polynomial Regression E.g.

\[
Y = -6.29x^3 + 8.45x^2 - 2.18x + 0.51
\]

Actual model used to generate the data: \(Y = -6x^3 + 8x^2 - 2x + 0.5 + \epsilon\)

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
  - 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/