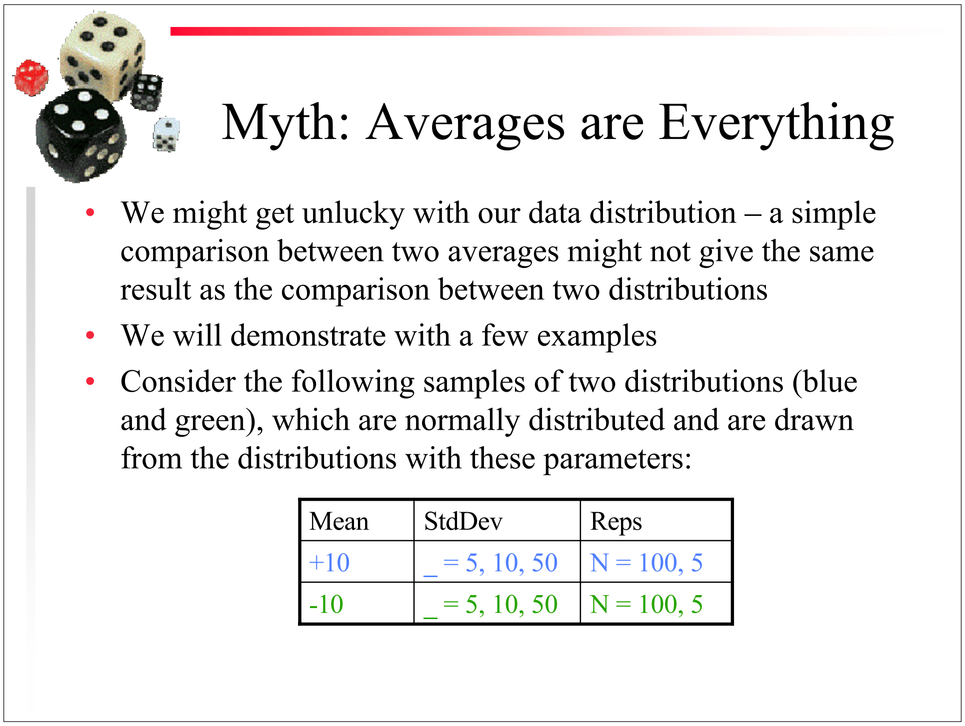
An Introduction to Statistics for EC Experimental Analysis

Authors: Mark Wineberg
Steffen Christensen

Top 5 Experimental Analysis Myths in EC

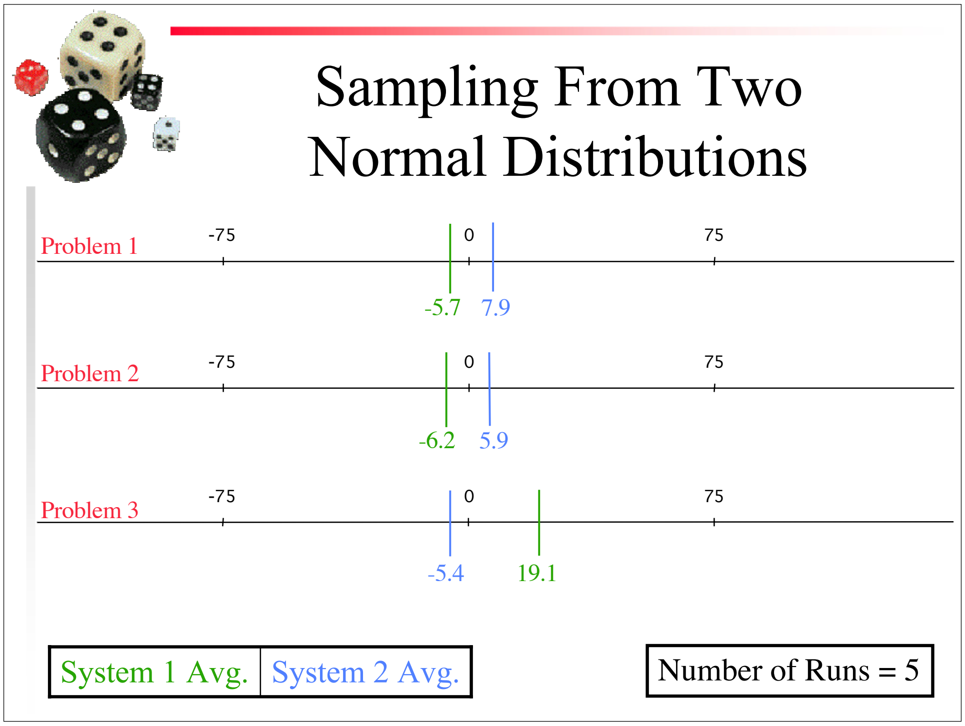
i. Results from 1 run is all that is needed
   - No, shows only proof of concept
ii. The best value achieved in a set of runs tells you something about the population distribution
   - No
iii. Using the same random number generator seed for both systems provides a fairer comparison
   - It doesn't - it’s the statistical properties of the system that we are looking for
iv. One system is obviously better than the other when looking at the data or graph - no statistics necessary
   - If it is so obvious, then will be easy to show statistically
     - might as well do the stats
     - shows that you are objectively confident in your conclusion
v. "My average is better than yours" means "my technique is better than yours"
   - In the best case you would need to take variance into account



Myth: Averages are Everything

- We might get unlucky with our data distribution – a simple comparison between two averages might not give the same result as the comparison between two distributions
- We will demonstrate with a few examples
- Consider the following samples of two distributions (blue and green), which are normally distributed and are drawn from the distributions with these parameters:

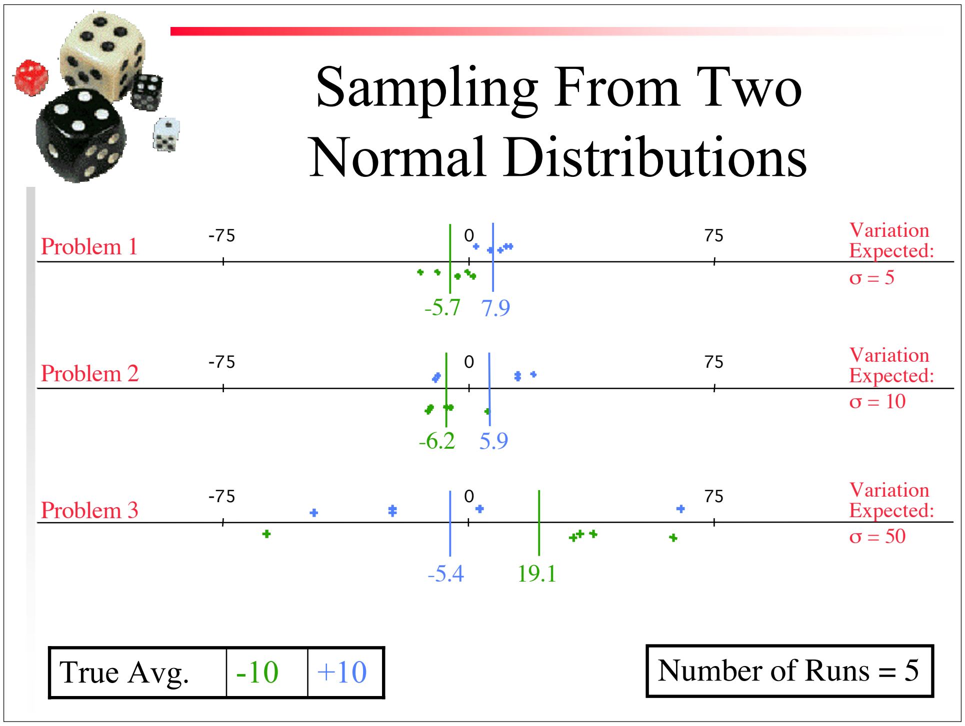
Mean	StdDev	Reps
+10	_ = 5, 10, 50	N = 100, 5
-10	_ = 5, 10, 50	N = 100, 5



Sampling From Two Normal Distributions

System 1 Avg.	System 2 Avg.

Number of Runs = 5



Sampling From Two Normal Distributions

Problem 1

Variation Expected: σ = 5

Problem 2

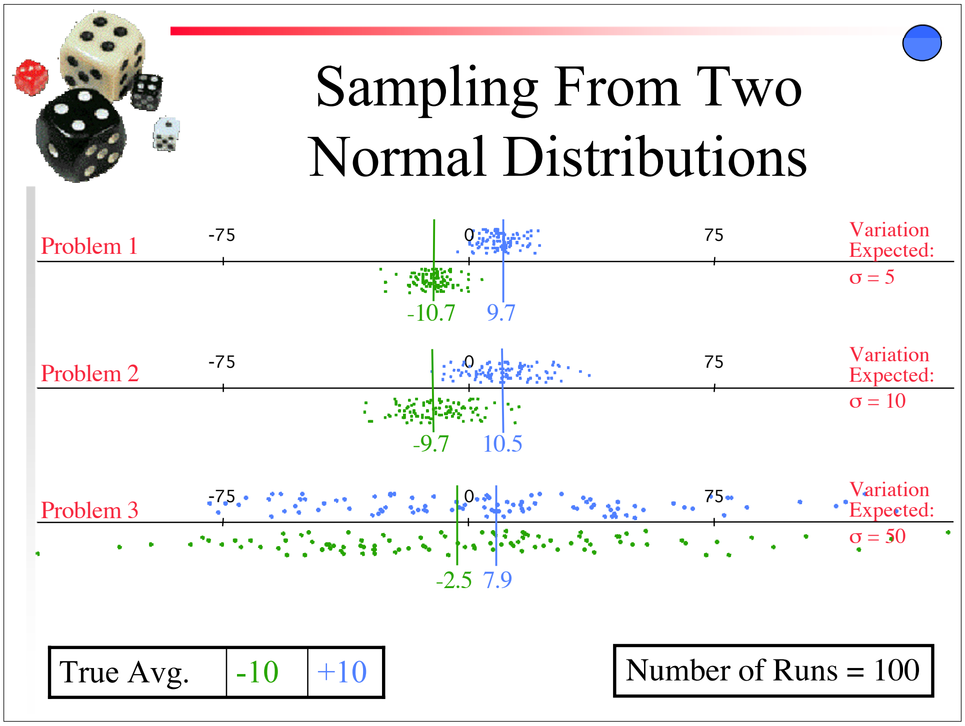
Variation Expected: σ = 10

Problem 3

Variation Expected: σ = 50

True Avg.	-10	+10

Number of Runs = 5



Sampling From Two Normal Distributions

Problem 1

Variation Expected: σ = 5

Problem 2

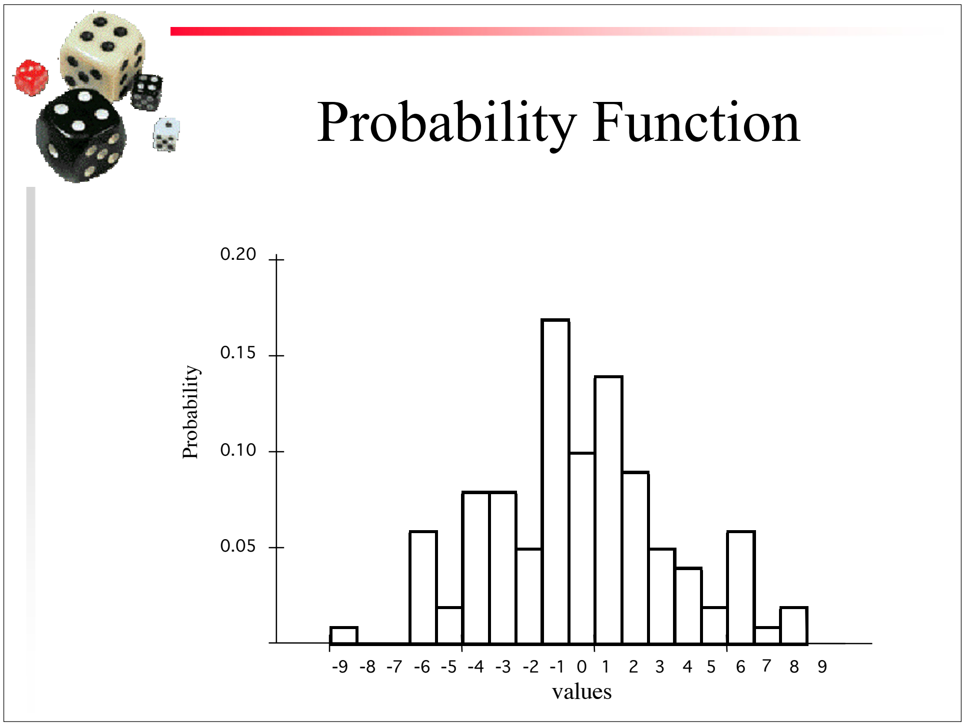
Variation Expected: σ = 10

Problem 3

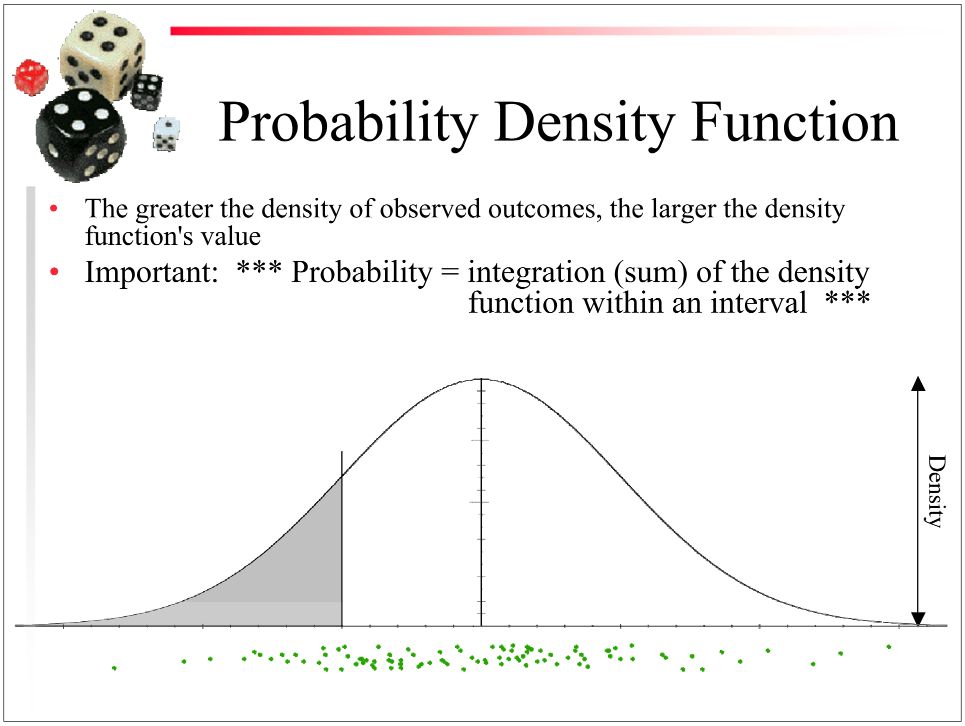
Variation Expected: σ = 50

True Avg.	-10	+10

Number of Runs = 100



Probability Function



Probability Density Function

- The greater the density of observed outcomes, the larger the density function's value
- Important: *** Probability = integration (sum) of the density function within an interval ***

What Are We Interested In?

- For most statistical analysis for EC the question is
  - Is my new way better than the old way?
  - Statistically this translates into a statement about the difference between means: "Is the difference between 'my mean' and 'the old 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?
  2. How can we compare the true means of *two* or more distributions?

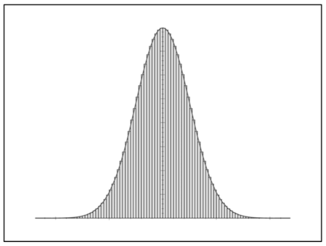
Distribution of the Mean

- 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 standard deviation of the mean of n samples is given by $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

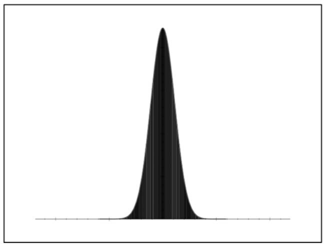
Distribution of the Mean

(Standard Normal Distribution)

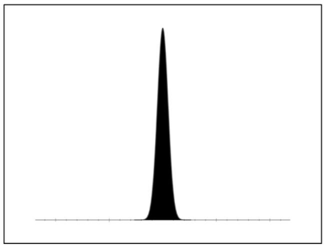
Mean of one sample



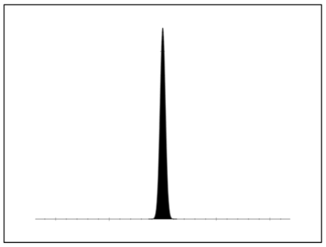
Mean of 5 samples



Mean of 25 samples



Mean of 100 samples



Confidence Intervals

- As the “finger” gets narrower, the mean of the samples approaches the true mean
- We’d like to say that in the overwhelming majority of all possible experiments, the true mean of this distribution will lie within a specified interval
  - Example: In 99% of cases, the true mean of the distribution, estimated from our 50 samples, lies within the interval [64 , 79] – called a *confidence interval* for the mean

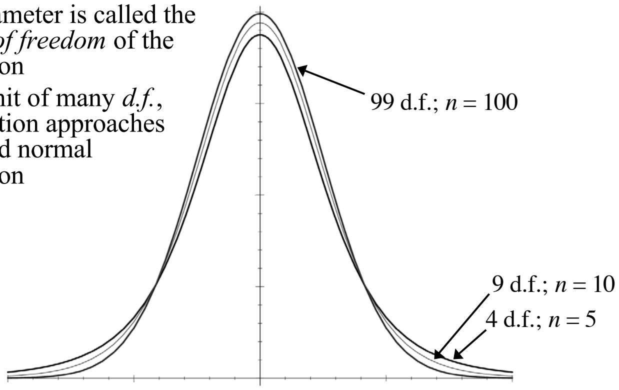
t Distribution

- Of course, we don't know the true mean, μ, or true standard deviation, σ
- We *do* know the mean of the samples, $\overline{X}$, 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 $\overline{X}$
  - The distribution of the sample average follows a t distribution with n - 1 degrees of freedom, where

$$T = \frac{(\overline{X} - \mu)}{s_{\overline{X}}} = \frac{(\overline{X} - \mu)}{s_X / \sqrt{n}}$$

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

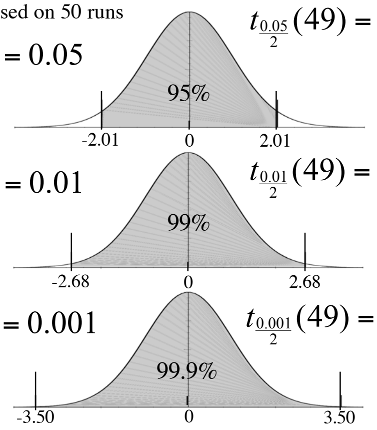
If samples taken from a *standard normal distribution* (μ = 0, σ = 1), the sample average has a *t* distribution.

Based on 50 runs

$\alpha = 0.05$ $t_{\frac{0.05}{2}}(49) = 2.01$

$\alpha = 0.01$ $t_{\frac{0.01}{2}}(49) = 2.68$

$\alpha = 0.001$ $t_{\frac{0.001}{2}}(49) = 3.50$



- For Confidence Intervals, we can use cutoff *t* values
- The wider the cutoff values, the more likely the true mean will fall between them
- α is the probability of obtaining values outside the cutoffs
  - Confidence Level is 1 – α
- Cut off *t* values can be computed using Excel: **=TINV(α, *n* - 1)**
  - Note: TINV() is already 2 sided

Estimating the Mean: Confidence Intervals Around the Average

- We know that

$$T = \frac{(\overline{X} - \mu_X)}{s_X / \sqrt{n}}$$

- 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

$$\pm t_{\frac{\alpha}{2}}(n-1) = \frac{(\overline{X} - \mu_{\overline{X}})}{s_{\overline{X}} / \sqrt{n}}$$

which can be rewritten as

$$\mu_X = \overline{X} \pm t_{\frac{\alpha}{2}}(n-1)\frac{s_X}{\sqrt{n}}$$

Estimating the Mean: Confidence Intervals Around the Average

- Confidence Intervals can be written in 3 equivalent ways

Error Bounds

$$\mu_X = \overline{X} \pm t_{\frac{\alpha}{2}}(n-1)\frac{s_X}{\sqrt{n}}$$

Confidence Intervals

$$\overline{X} - t_{\frac{\alpha}{2}}(n-1)\frac{s_X}{\sqrt{n}} \le \mu_X \le \overline{X} + t_{\frac{\alpha}{2}}(n-1)\frac{s_X}{\sqrt{n}}$$

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

Estimating the Mean: Confidence Intervals Around the Average

Example:

- An EC experimenter runs a GA on a TSP
- At the end of each run, the smallest length tour that had been found during the run was recorded
- The GA is run 50 times on the same TSP problem
- On average the GA 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 GA run produced tours of

  $\overline{X} = 272$ that had $s_X = 87$

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

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

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

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

  so $\overline{X} = \mu_X \pm 2.68\frac{s_X}{\sqrt{50}} = \mu_X \pm 0.38 s_X$

  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

---

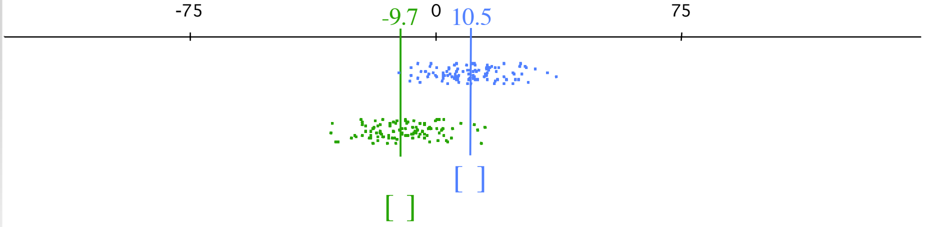
Using Confidence Intervals to Determine Whether My Way is Better

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

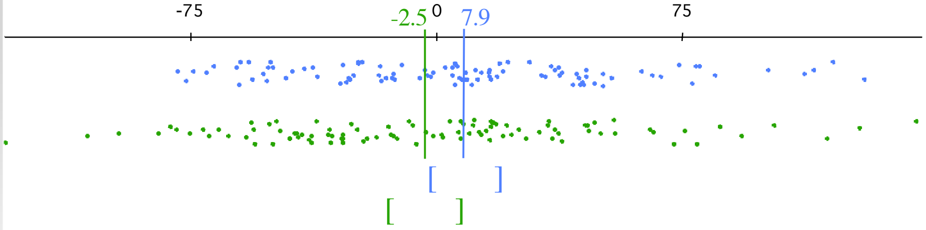


95% Confidence Level

μ	σ
+10	10
-10	10

n	X	s_X	$1.96\frac{s_X}{\sqrt{n}}$	Lower	Upper
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



95% Confidence Level

μ	σ
+10	50
-10	50

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

Assumptions, assumptions

- All we have said so far applies only if the source distribution is a normal distribution
- What if the source distribution is not a normal distribution?
  - In EC, the source distribution is essentially *never* normal!
- Fortunately, there is one nice property that can help us out
  - The *Central Limit Theorem*: the sum of many identically distributed random variables tends to a Gaussian
  - Equation of the mean:

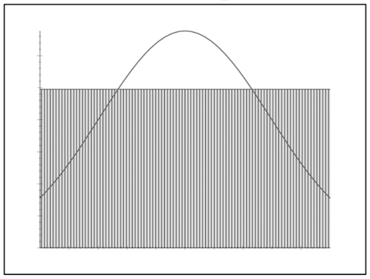
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - So the mean of any set of samples tends to a normal distribution

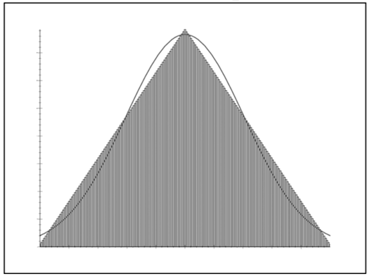
Central Limit Theorem

- The sum of many *independent*, *identically distributed (IID)* random variables approaches a Gaussian normal curve
- E.g. Uniform distribution on [0, 1]:

Mean of one sample



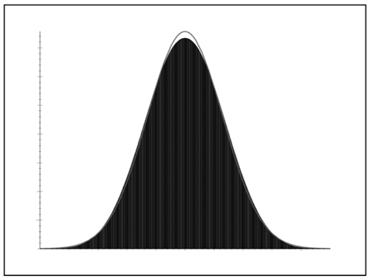
Mean of two samples



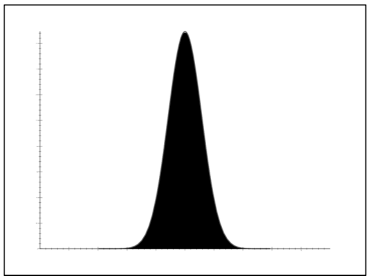
Central Limit Theorem

- E.g. Uniform distribution (continued):

Mean of five samples

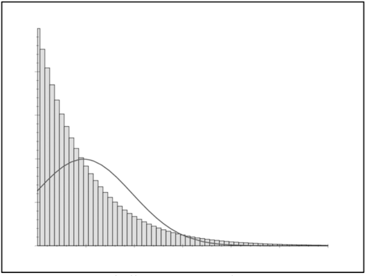


Mean of 25 samples

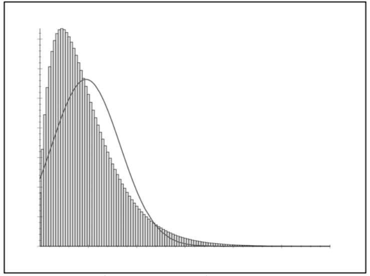


Exponential Distribution

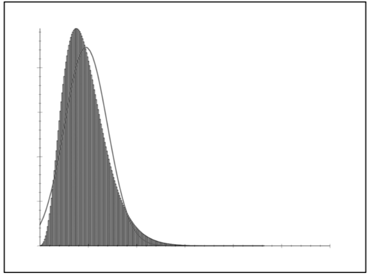
Mean of one sample



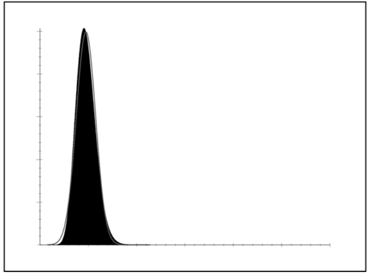
Mean of two samples



Mean of five samples



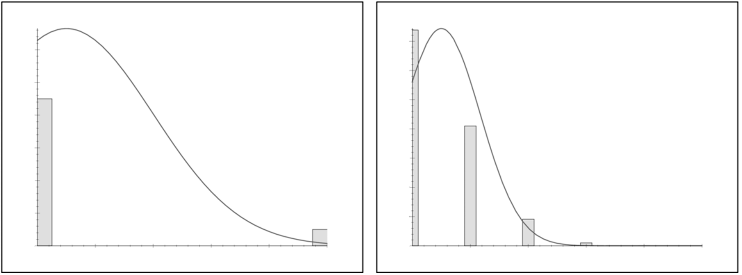
Mean of 25 samples



Binomial Distribution (p = 0.1)

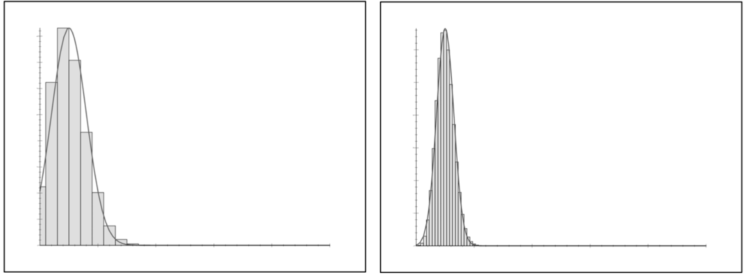
Mean of one sample

Mean of five samples



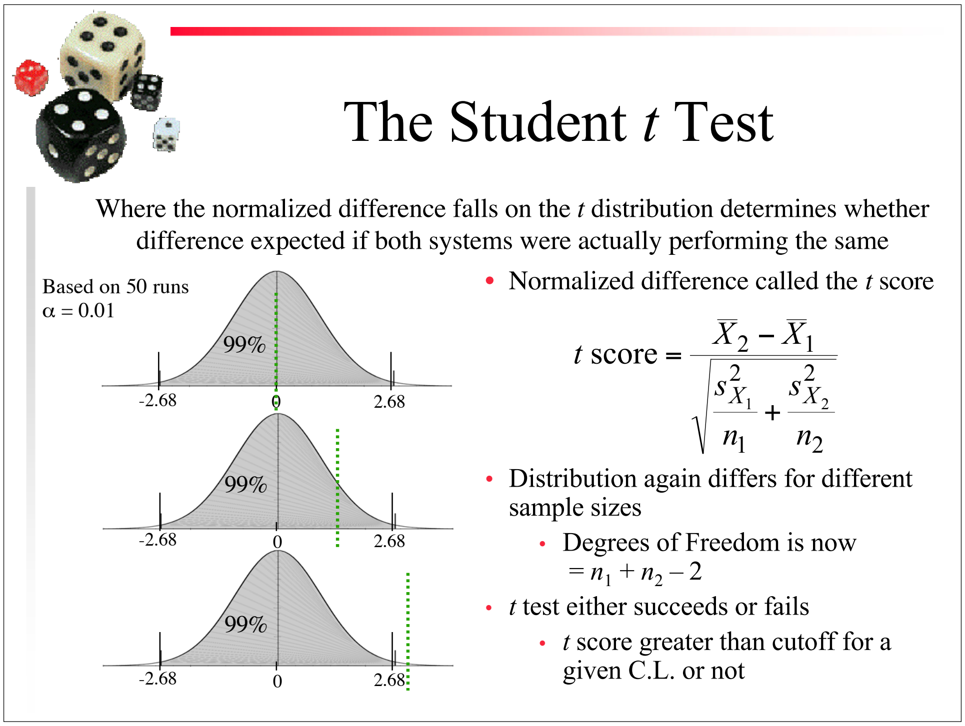
Mean of 25 samples

Mean of 100 samples



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
  - If there is no difference between the actual means of the 2 systems
    - then the difference between the sample averages should be 0, with some error that should follow the *t* distribution
    - this is because the difference btw 2 normal distributions is also normal
    - so the sample average should be a *t* distribution as usual
  - now we can see if the computed difference of the sample averages falls outside a confidence interval (for some α) for the *t* distribution



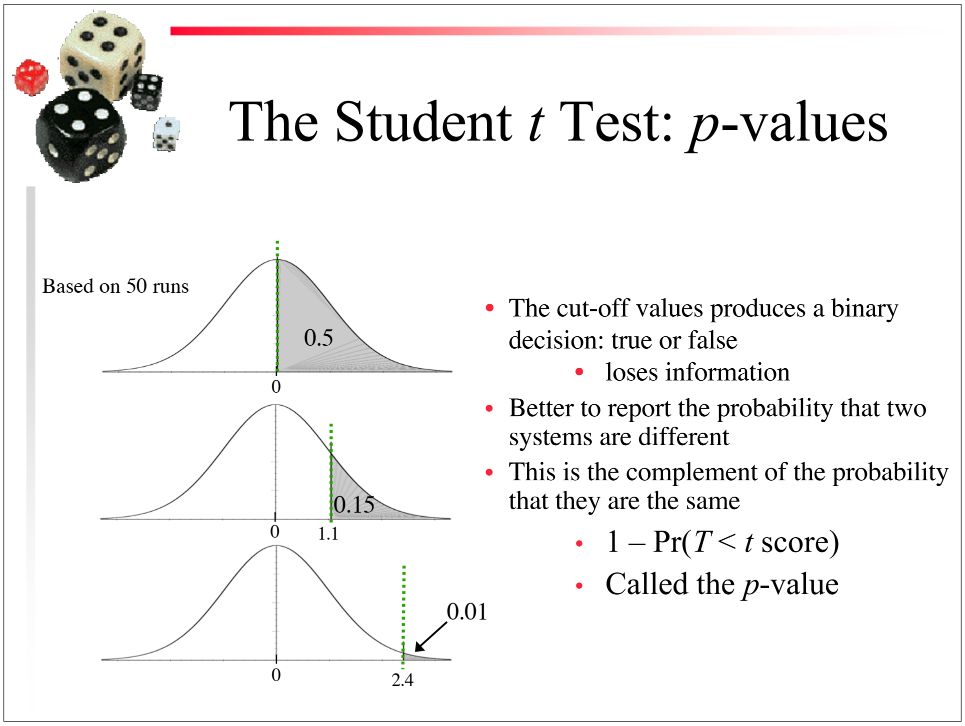
The Student *t* Test

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* score

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

- Distribution again differs for different sample sizes
  - Degrees of Freedom is now $= n_1 + n_2 - 2$
- *t* test either succeeds or fails
  - *t* score 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
  - $1 - \Pr(T < t\text{ score})$
  - 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: $n_1 + n_2 - 2$
4. Calculate T statistic: $T = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
5. Compute the p-value
   - p-value = the area under the t distribution outside [-T, T]
   - Use **=TDIST(T, $n_1 + n_2$ - 2, 2)** in Excel
     - The final "2" in Excel means "two-sided"

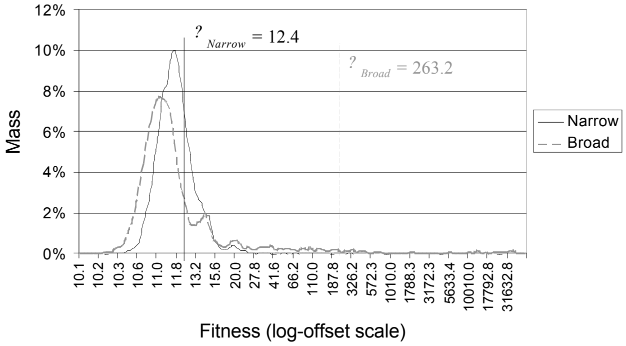
When The CLT Fails You

- Everything we have done so far depends on the Central Limit Theorem holding
  - But this is not always true
  - ***In fact for EC 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)
- EC alert!
  - Many data in evolutionary computing are often highly skewed because some local optima in the search space are very unfit
  - Example follows

When The CLT Fails You

From a node layout problem where fitness is absolute error (minimization)

- Here are the PDFs of 2 EC parameter settings, named Broad and Narrow for convenience



- Here Broad's mean is much *worse* than Narrow's because of its extended tail, even though Broad often beats Narrow in practice!
  - We don't really care about the 8% of trials where Broad performs badly

So what should we do?

- There are tests that use Ranks instead of actual values
  - These are called **Non-Parametric** Tests
  - They measure how interspersed the samples from the two treatments are
    - If the result is "alternating" it is assumed that there is no effective difference

Non-Parametric Tests

- Ranks are uniformly distributed
  (think of percentiles – uniform on [0%, 100%] = [0, 1])
- 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
  - This result is independent of the particular distributions of the 2 treatments
  - So we can perform a *t* test on the ranks
- 2 other techniques with similar results are commonly seen
  - Wilcoxon's Rank-Sum test
  - Mann-Whitney U test
- All are effectively equivalent, and the test is often called the "Mann-Whitney-Wilcoxon test" by statisticians

How To Rank the Data

- Augment each data point with a treatment identifier and an additional slot for its rank
- Sort the data sets together by value
  - record the ranks of all values in their rank slot
    - assign the average rank of tied values to each tied value
- Resort by the original order thus splitting the data sets back out
  - keep the combined ranking with each data point
- Apply your *t* test on the ranked values

Ranked Example

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

Two sets of Data

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

Ranked Example

Combine the data into a single array

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

Sort the combined data

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

Give each data element its corresponding rank

		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

Ranked Example

		ranks	
A	0.99	1	
A	0.91	2	t1
A	0.91	3	t1
A	0.64	4	t2
A	0.64	5	t2
B	0.64	6	t2
B	0.64	7	t2
A	0.27	8	t3
B	0.27	9	t3
A	0.16	10	t4
A	0.16	11	t4
A	0.16	12	t4
B	0.16	13	t4
B	0.16	14	t4
B	0.08	15	
A	0.03	16	t5
B	0.03	17	t5
B	0.03	18	t5
B	0.02	19	
B	0.01	20	

Identify ties

Average tied ranks together

t1	2.5
t2	5.5
t3	8.5
t4	12
t5	17

Replace tied ranks with average tied ranks

		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	

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
B	0.64	5.5
B	0.64	5.5
A	0.27	8.5
B	0.27	8.5
A	0.16	12
A	0.16	12
A	0.16	12
B	0.16	12
B	0.16	12
B	0.08	15
A	0.03	17
B	0.03	17
B	0.03	17
B	0.02	19
B	0.01	20

Resort by treatment

		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

		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

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?

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
  - 50^{th} 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. $value_l \le median \le value_u$
  - With a confidence level of $1-\alpha$

A Confidence Interval Around the Median: Thompson-Savur

- In Excel:
  - To calculate b use `CRITBINOM(`n`,`1/2`,`$\alpha/2$`)`
  - to compute the $value_u$ use the function `PERCENTILE(`dataArray, u`)`
  - to compute the $value_l$ use the function `PERCENTILE(`dataArray, l`)`

A Confidence Interval Alternative to the Ranked *t* Test

- Find the median confidence interval for the two data sets
- If the confidence intervals do not overlap
  - Data sets are taken from different distributions
  - With a confidence level of $1 - \alpha$ where α is the upper tail probability used in computing b
  - Advantages:
    - Gives better understanding of system
      - see median values with error bounds
    - Easy to draw and productive on a graph
  - Disadvantage:
    - Not as sensitive as the ranked *t* test

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 2.0×10^{-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 $d' = \dfrac{t}{\sqrt{(n_1+n_2)/2}}$
  - 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!

Perils of Computer Science

- Very easy to do a host of runs
  - 1 000 000 runs of each group not too hard on a PC
    - *t*-scores around 500
    - Bizarre-looking results
      - *p* scores show up as 0 or #NUM! in Excel
- We often know that there *will* be a difference between our method and the reference method
  - Enough runs will detect this difference
  - Difference might only be slight
    - Use care in choosing sample sizes, or (*better*) compute Cohen’s d'

Perils of Computer Science

- We can generate lots of data very quickly
  - Leads to over-complicated experimental designs
  - Many parameters in Evolutionary Computation
    - 10 levels of A, 10 levels of B, 10 levels of C = 1000 treatments!
    - Tough to analyze
    - A simple *t* test on the ranks is nearly always better (and see following)
- Always draw a scatter plot or histogram of your data!
  - This alerts you to strange things, like that the mean is very bad, but some individuals are very good
- Always record the performance of *ALL* the individuals
  - You’ll need this for doing the *t* test on the ranks
  - In EC, we mean ALL individuals of *interest*; i.e. best of run

Perils of Computer Science in Evolutionary Computation

- Don’t confuse Population averages with Best-of-Run averages!
  - In any GA or GP, the average of the population tells you almost nothing of interest
  - Use the median of the best-of-run, do the *WHOLE* experiment several times
  - Use the tree size of the best-of-run individuals as well!
    - They are the Heroes – hence they are of interest, unless you’re really looking to optimize average tree size during evolution
- Test 2 (or a few) parameter settings, one variable at a time
  - Fewer runs required
  - Get a feel for how the variation occurs

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
  - Perform a minimum of 30 repetitions for each system
  - Performing 50 to 100 repetitions is usually better

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

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 for Tests

- To correct, choose a smaller α

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

  - Where *m* is the number of comparisons
  - So for 95% CL use α = 0.025/6 = 0.004167
  - For a Z test the critical value changes from 1.96 to 2.64
- Called a Bonferroni post-hoc correction
  - Other post-hoc techniques such as Tukey and Scheffé that are more powerful than Bonferroni; also Holm's and Sidak's procedures can be useful
- You should apply the Bonferroni 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"

The Bonferroni Correction for Experiments

- The Bonferroni Correction is more widely applicable than just for multi-level comparisons
- We really need to control for the dilution of the confidence levels throughout the study, whether or not the CLs are applied to analyses of independent 'phenomena'
  - We must *divide* the α used for each CL test by the total number of CL tests in the study
- To apply the Bonferroni correction to *p*-values *multiply* the *p*-values by the number of CL tests performed
  - "Probabilities" bigger than 1 means "not significant"

The Bonferroni Correction for Experiments

- Example:
  - A robot dog has been created
    - Genetic Programming is used to control the ear wiggles of the robot
    - a Genetic Algorithm is used to optimize its tail wagging ability
  - A study is being done to improve both the ears and the tail independently, and we want to be 95% confident in our over all tests
    - For the ears the GP is tested with 3 different sets of terminal nodes
    - For the tail the GA is tested with 4 different fitness functions
    - There are $\binom{3}{2} + \binom{4}{2} = 3 + 6 = 9$ total CL inferences used in the study
    - Consequently the α used for any CL should be $\alpha = 0.025 / 9 = 0.0028$

Multiple Factors

- Most of the time, there are many different properties we are interested in studying
  - e.g. We may be trying out various kinds of crossovers, with and without mutation, under different selection pressures
  - Each of the above parameters has multiple levels
  - This is called a multiple factor analysis
    - with each factor having multiple levels
  - Use Analysis of Variance or General Linear Models to analyze
    - See text books on ANOVA and GLMs

Multiple Factors: Factorial Design

- When dealing with multiple factors with multiple levels
  - Important that all combinations of factor levels are tried
  - A given combination of factor levels is called a treatment
  - If you want accurate information about each possible interaction, each treatment should be repeated at least 30 times
    - If you interested largely in main effects, 10 repetitions is often fine, if you have enough levels

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)

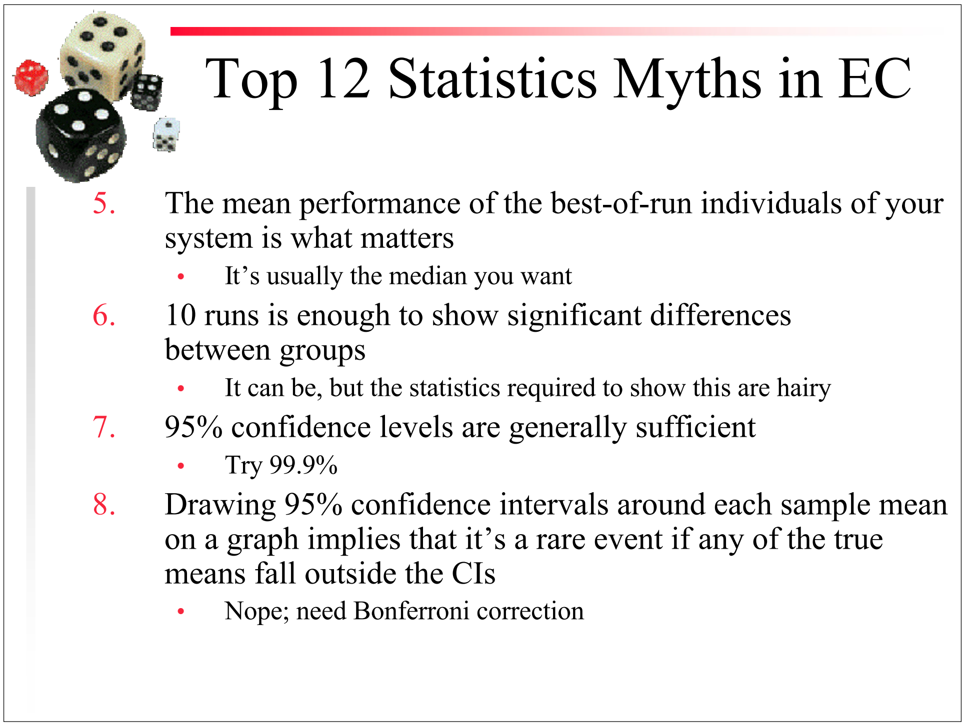
	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12
S	New	New	New	New	New	New	Std	Std	Std	Std	Std	Std
X	x	x	x	~~x~~	~~x~~	~~x~~	x	x	x	~~x~~	~~x~~	~~x~~
P	p1	p2	p3	p1	p2	p3	p1	p2	p3	p1	p2	p3

Multiple Factors: Factorial Design

- If we are performing 50 reps per treatment
  - In previous example we have $S \times X \times P \times 50 = 2 \times 2 \times 3 \times 50 = 12 \times 50 = 600$ experiments to perform
- The number of experiments goes up as the product of the number of levels in each factor
  - This is exponential in the number of factors
  - Consequently, carefully choose the factors and factor levels that you study in your experiments
  - Minimize what factors you vary
    (focus your experiments on the relevant factors)

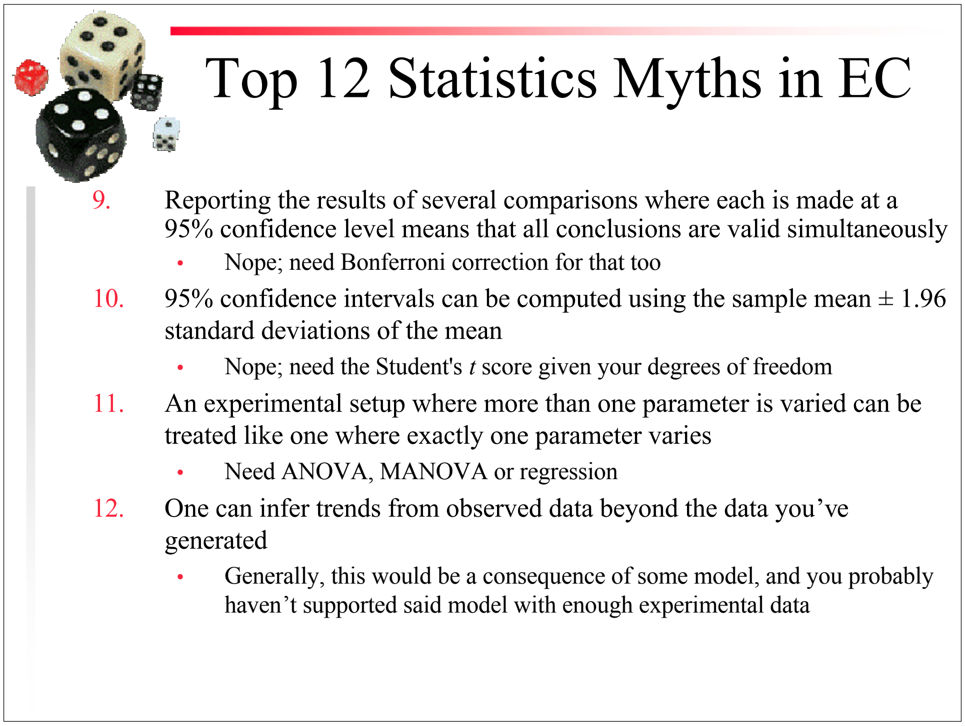
Top 12 Statistics Myths in EC

1. My mean result being better than yours means my technique is superior to yours
   - In the best case you need to perform a t test to assert this claim
2. Reporting the mean value of a statistic is good enough
   - You need some representative range
3. Reporting the mean and standard deviation of a statistic is good enough
   - Need number of runs
4. Your data are normally distributed
   - Not usually



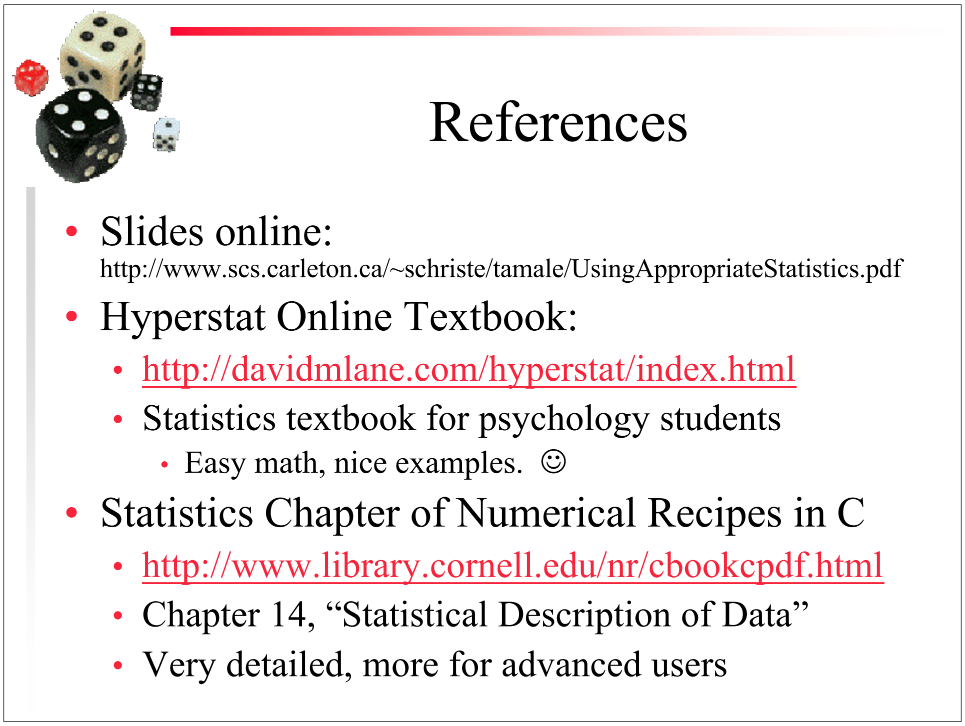
Top 12 Statistics Myths in EC

5. The mean performance of the best-of-run individuals of your system is what matters
    - It’s usually the median you want
6. 10 runs is enough to show significant differences between groups
    - It can be, but the statistics required to show this are hairy
7. 95% confidence levels are generally sufficient
    - Try 99.9%
8. Drawing 95% confidence intervals around each sample mean on a graph implies that it’s a rare event if any of the true means fall outside the CIs
    - Nope; need Bonferroni correction



Top 12 Statistics Myths in EC

9. Reporting the results of several comparisons where each is made at a 95% confidence level means that all conclusions are valid simultaneously
    - Nope; need Bonferroni correction for that too
10. 95% confidence intervals can be computed using the sample mean ± 1.96 standard deviations of the mean
    - Nope; need the Student's *t* score given your degrees of freedom
11. An experimental setup where more than one parameter is varied can be treated like one where exactly one parameter varies
    - Need ANOVA, MANOVA or regression
12. One can infer trends from observed data beyond the data you’ve generated
    - Generally, this would be a consequence of some model, and you probably haven’t supported said model with enough experimental data



References

- Slides online: http://www.scs.carleton.ca/~schriste/tamale/UsingAppropriateStatistics.pdf
- Hyperstat Online Textbook:
  - http://davidmlane.com/hyperstat/index.html
  - Statistics textbook for psychology students
    - Easy math, nice examples. ☺
- Statistics Chapter of Numerical Recipes in C
  - http://www.library.cornell.edu/nr/cbookcpdf.html
  - Chapter 14, “Statistical Description of Data”
  - Very detailed, more for advanced users