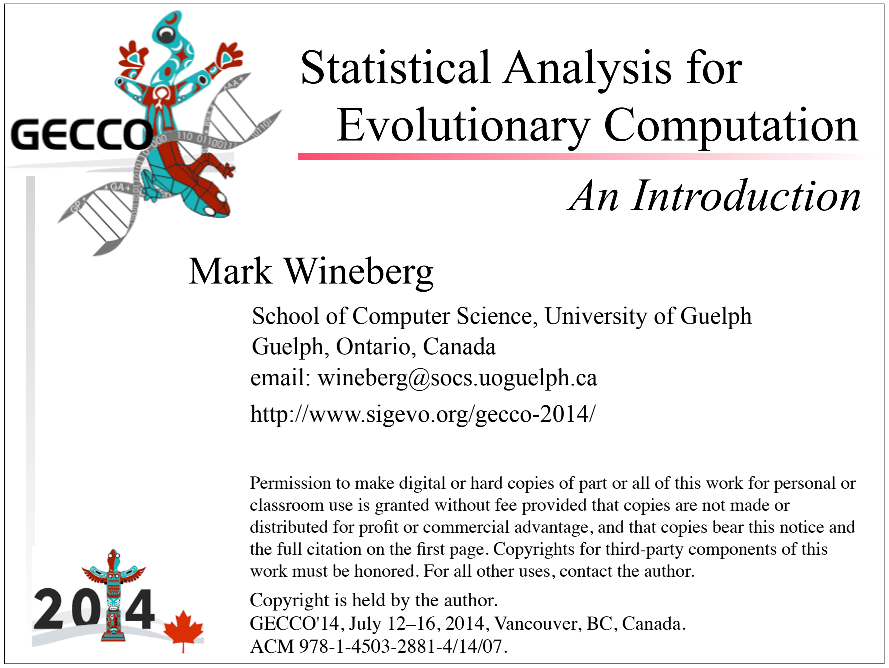


Statistical Analysis for Evolutionary Computation

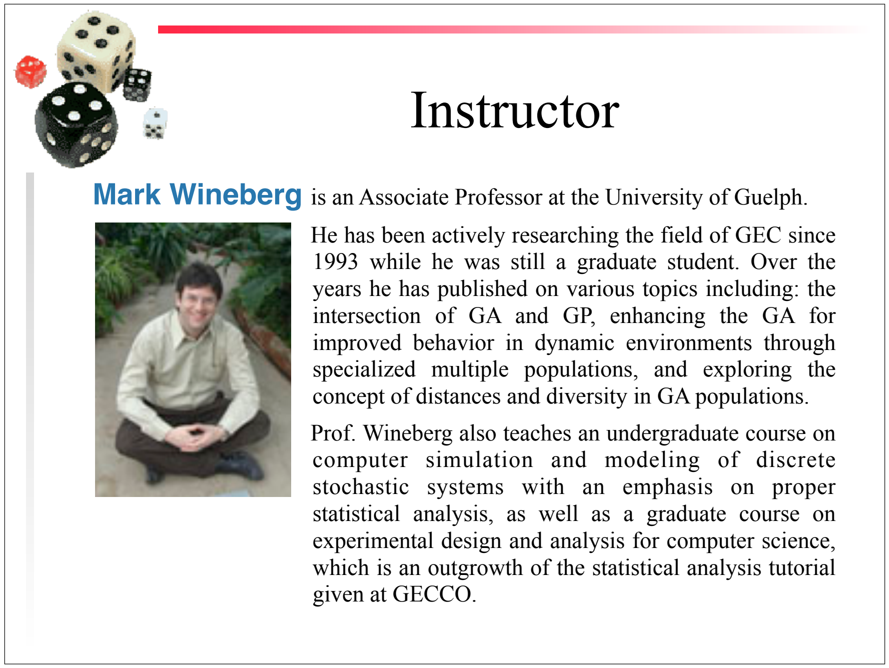
An Introduction

Mark Wineberg

School of Computer Science, University of Guelph
Guelph, Ontario, Canada
email: wineberg@socs.uoguelph.ca

http://www.sigevo.org/gecco-2014/

GECCO'14, July 12–16, 2014, Vancouver, BC, Canada.
ACM 978-1-4503-2881-4/14/07.

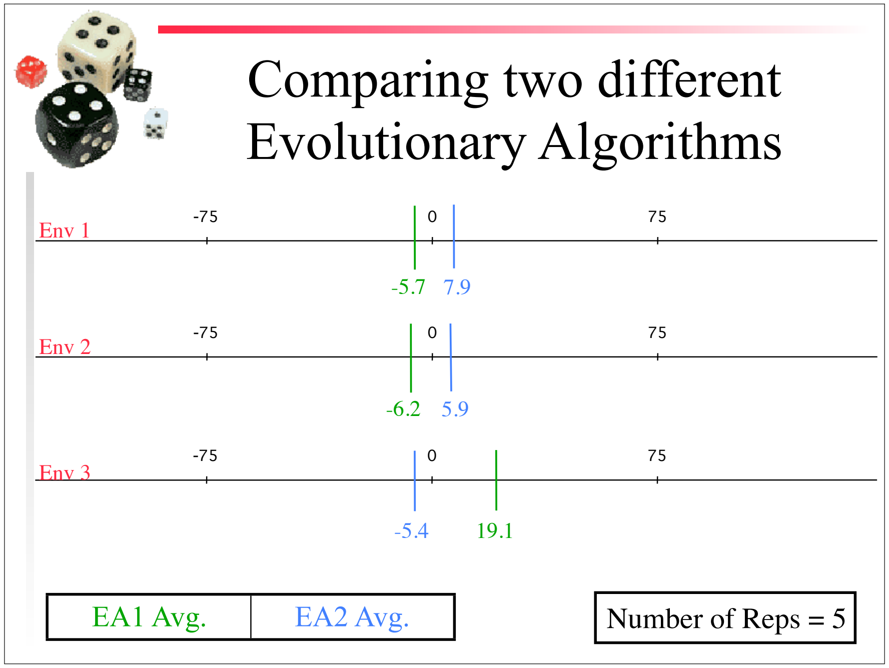


Instructor

Mark Wineberg is an Associate Professor at the University of Guelph.

He has been actively researching the field of GEC since 1993 while he was still a graduate student. Over the years he has published on various topics including: the intersection of GA and GP, enhancing the GA for improved behavior in dynamic environments through specialized multiple populations, and exploring the concept of distances and diversity in GA populations.

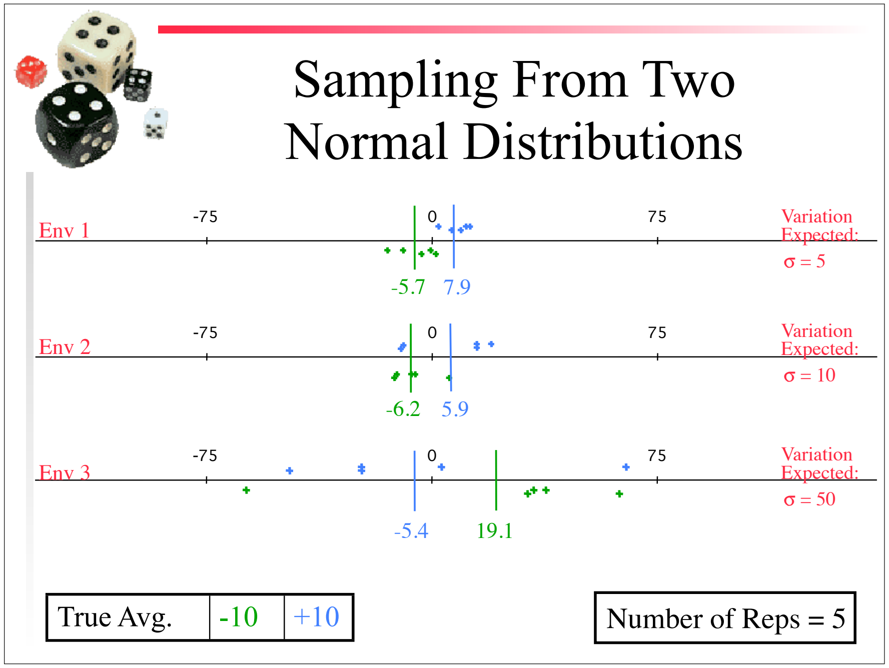
Prof. Wineberg also teaches an undergraduate course on computer simulation and modeling of discrete stochastic systems with an emphasis on proper statistical analysis, as well as a graduate course on experimental design and analysis for computer science, which is an outgrowth of the statistical analysis tutorial given at GECCO.



Comparing two different Evolutionary Algorithms

EA1 Avg.	EA2 Avg.

Number of Reps = 5

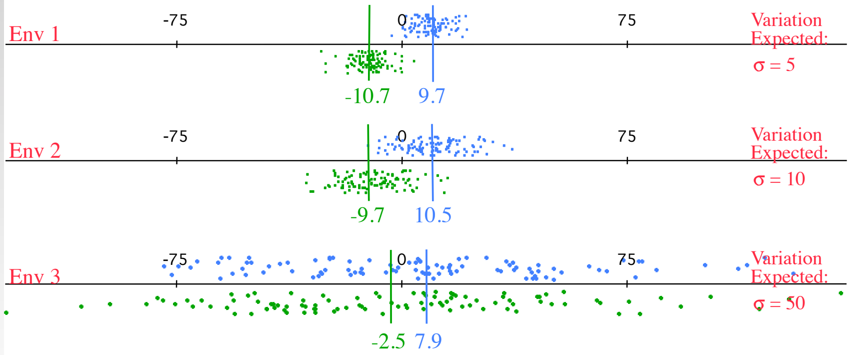


Sampling From Two Normal Distributions

True Avg.	-10	+10

Number of Reps = 5

Sampling From Two Normal Distributions



True Avg.	-10	+10

Number of Runs = 100

Basic Statistical Tests

Part 1 - Point Estimation:

Finding the Mean using Confidence Intervals

What Are We Interested In?

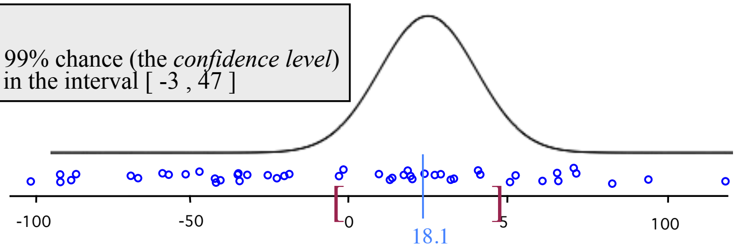
- For most statistical analysis for EC the question is
  - Is one way better than another way?
  - Statistically this translates into a statement about the difference between means: “Is the difference between ‘my mean’ and ‘the other mean’ greater than zero?”
- We will approach this question in 2 steps:
  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 (μ) likely lies
  - Called the *confidence interval*
- Also provide the probability that μ lies with the CI
  - Called the *confidence level*

Example:

There is a 99% chance (the *confidence level*) that μ lies in the interval [-3 , 47]

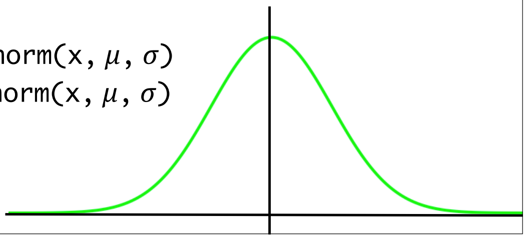


Normal Distribution

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

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \propto \frac{1}{e^{x^2}}$$

$X \sim N(\mu, \sigma^2)$ *pdf* in R: `dnorm(x, μ, σ)`
cdf in R: `pnorm(x, μ, σ)`



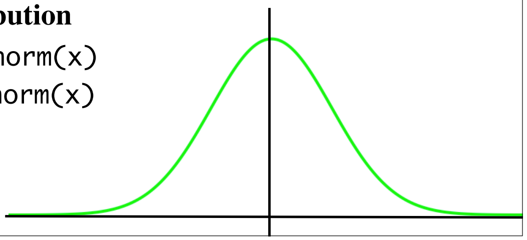
Most common distribution found in nature thanks to the *Central Limit Theorem*

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

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \propto \frac{1}{e^{x^2}}$$

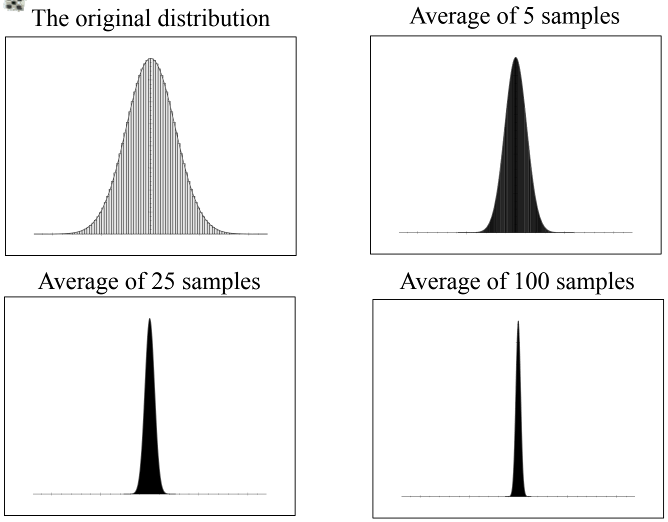
Standard Normal Distribution

$X \sim N(0, 1)$ *pdf* in R: `dnorm(x)`
cdf in R: `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

$$\mu_X = \sum_{i=1}^{n} p_i \cdot x_i$$

The Standard Deviation

$$\sigma_X = \sqrt{E((X-\mu_X)^2}$$

The average

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

Variation around the average

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

The Sample Standard Deviation

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\mu_{\overline{X}} = \mu_X$$

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

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, 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 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 $\overline{X}$

Confidence Intervals

- First since $\overline{X}$ 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

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

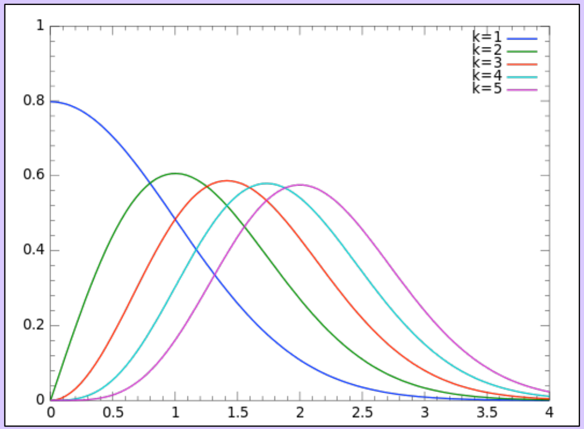
Confidence Intervals

- 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

$$Z = \frac{\overline{X}-\mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X}-\mu_X}{\frac{\sigma_X}{\sqrt{n}}} \longrightarrow Z \simeq \frac{\overline{X}-\mu_{\overline{X}}}{s_{\overline{X}}} = \frac{\overline{X}-\mu_X}{\frac{s_X}{\sqrt{n}}}$$

But the denominator is no longer a scaler!

Standard Deviation of the Normal Distribution:
The Chi Distribution



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

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

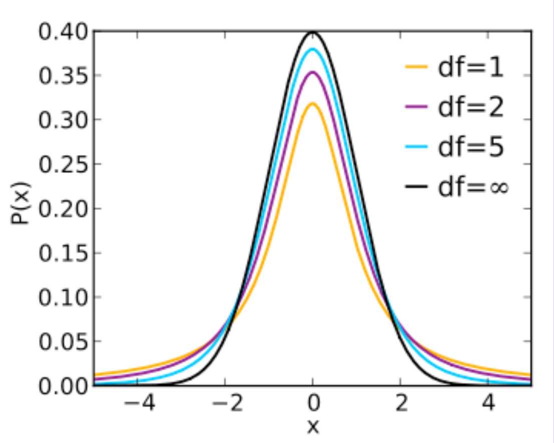
Confidence Intervals

- 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

$$Z \simeq \frac{\overline{X}-\mu_X}{\frac{s_X}{\sqrt{n}}} \longrightarrow T = \frac{\overline{X}-\mu_X}{\frac{s_X}{\sqrt{n}}}$$

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

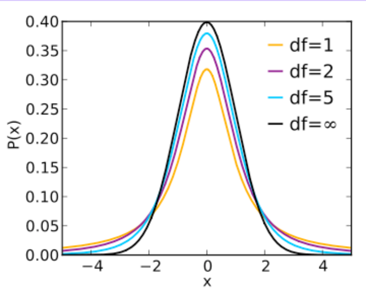
Student's T Distribution



$$f_{df}(x) \propto \left(1+\frac{x^2}{df}\right)^{-\frac{df+1}{2}} \underset{\text{large } x, df}{\approx} \frac{1}{x^{df}} \qquad df = n - 1$$

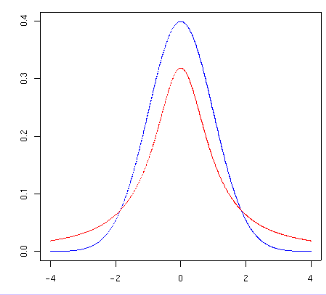
from http://en.wikipedia.org/wiki/T-distribution

Student's T vs Normal (Gaussian)



Student's T with various *dfs*

$$f_{df}(x) \underset{\text{large } x, df}{\approx} \frac{1}{x^{df}}$$

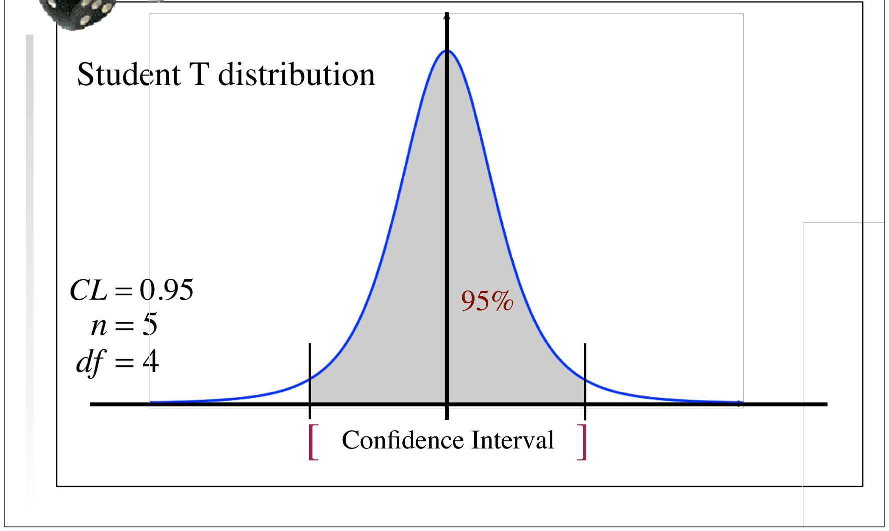


Blue: Gaussian
Red: T with *df* = 1

$$f(x) \underset{\text{large } x}{\propto} \frac{1}{e^{x^2}}$$

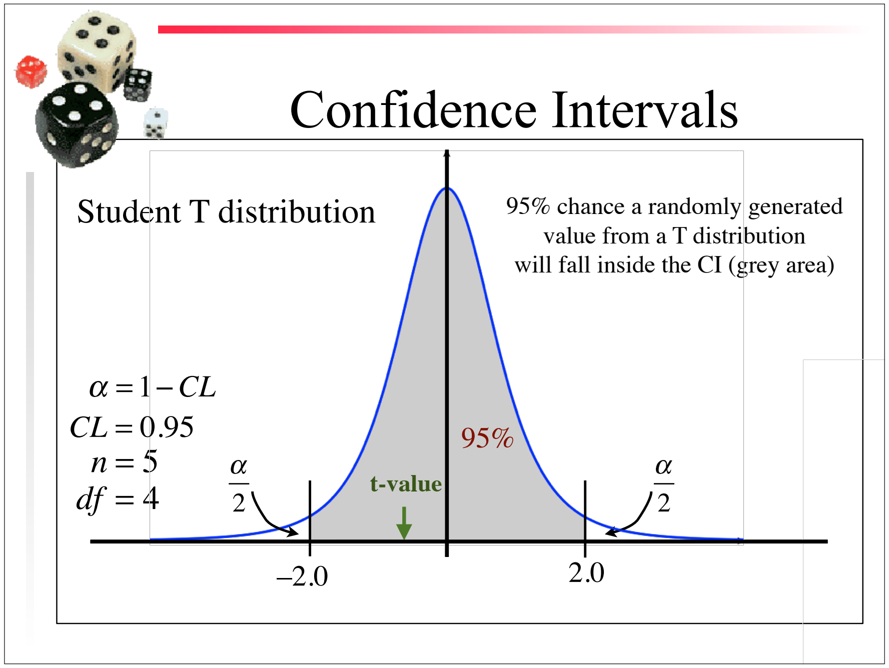
- Student T is "broader" than the Normal
- Student T goes to 0 much more slowly than the Normal
  (*has substantial probability of very large values*)

Confidence Intervals



Student T distribution

$CL = 0.95$
$n = 5$
$df = 4$

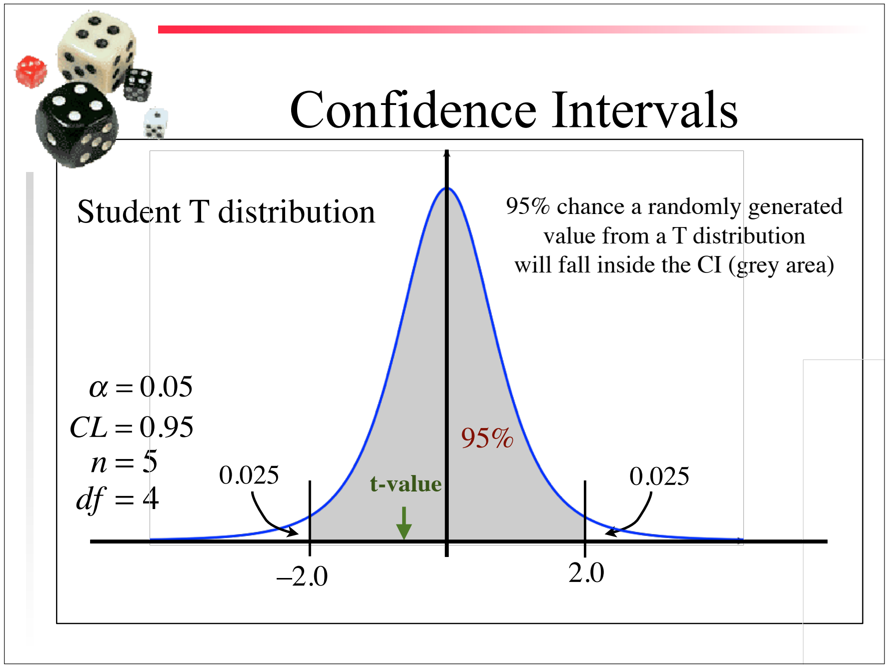


Confidence Intervals

Student T distribution

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

$\alpha = 1 - CL$
$CL = 0.95$
$n = 5$
$df = 4$

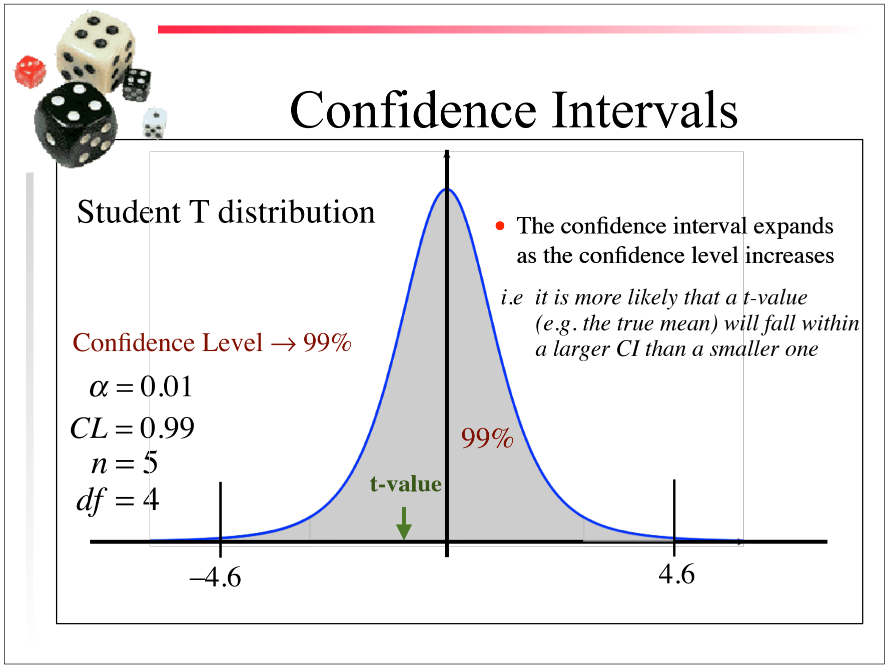


Confidence Intervals

Student T distribution

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

$\alpha = 0.05$
$CL = 0.95$
$n = 5$
$df = 4$



Confidence Intervals

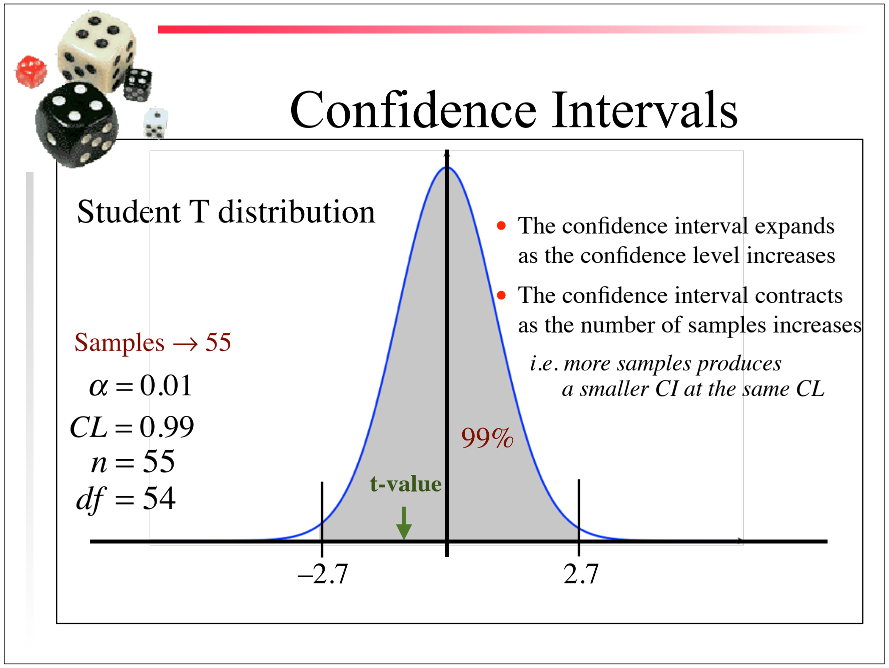
Student T distribution

- 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

Confidence Level → 99%

$\alpha = 0.01$
$CL = 0.99$
$n = 5$
$df = 4$



Confidence Intervals

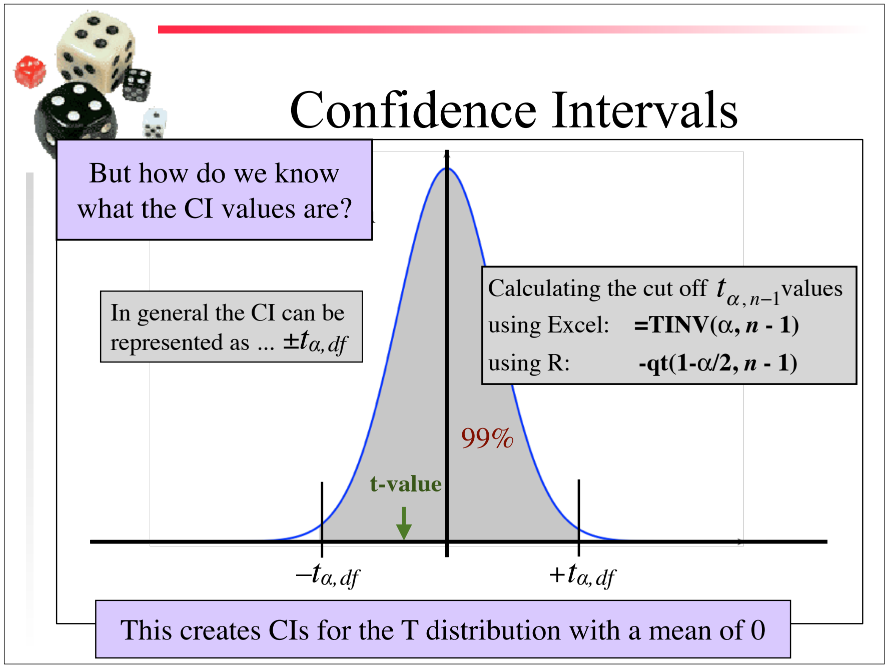
Student T distribution

- The confidence interval expands as the confidence level increases
- The confidence interval contracts as the number of samples increases

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

Samples → 55

$\alpha = 0.01$
$CL = 0.99$
$n = 55$
$df = 54$



Confidence Intervals

But how do we know what the CI values are?

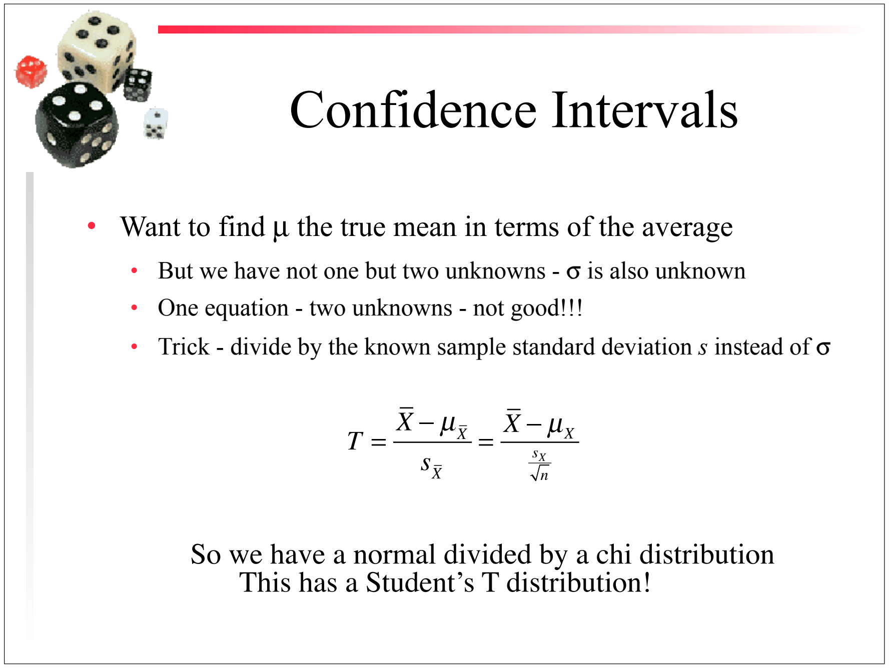
In general the CI can be represented as ... $\pm t_{\alpha, df}$

Calculating the cut off $t_{\alpha, n-1}$ 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

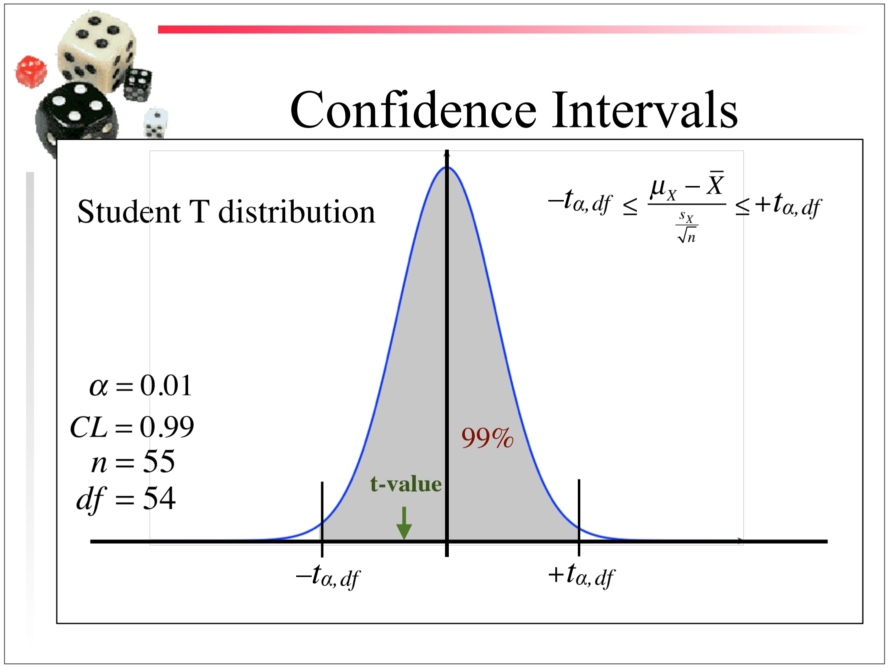


Confidence Intervals

- 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

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

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

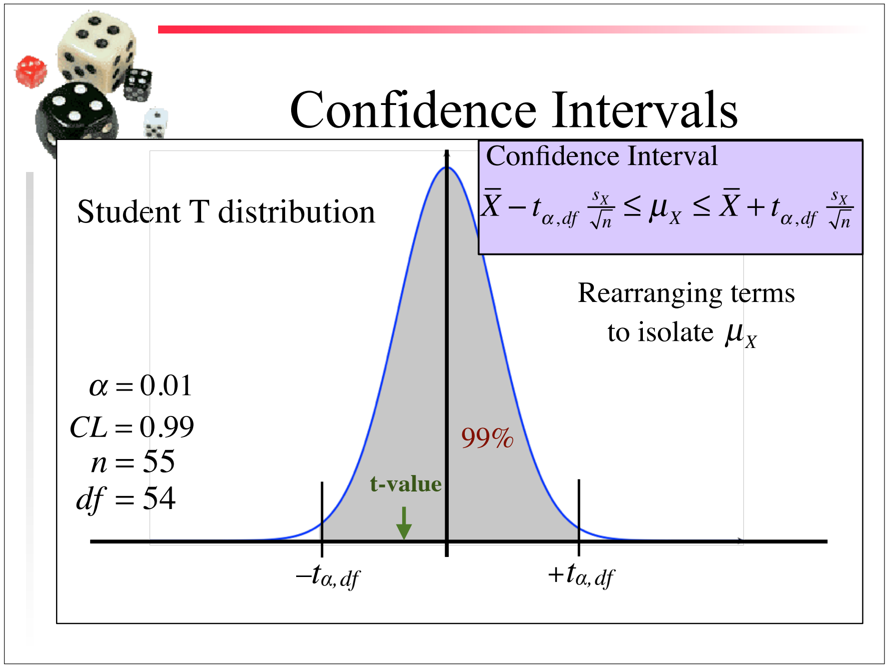


Confidence Intervals

Student T distribution

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

$\alpha = 0.01$
$CL = 0.99$
$n = 55$
$df = 54$



Confidence Intervals

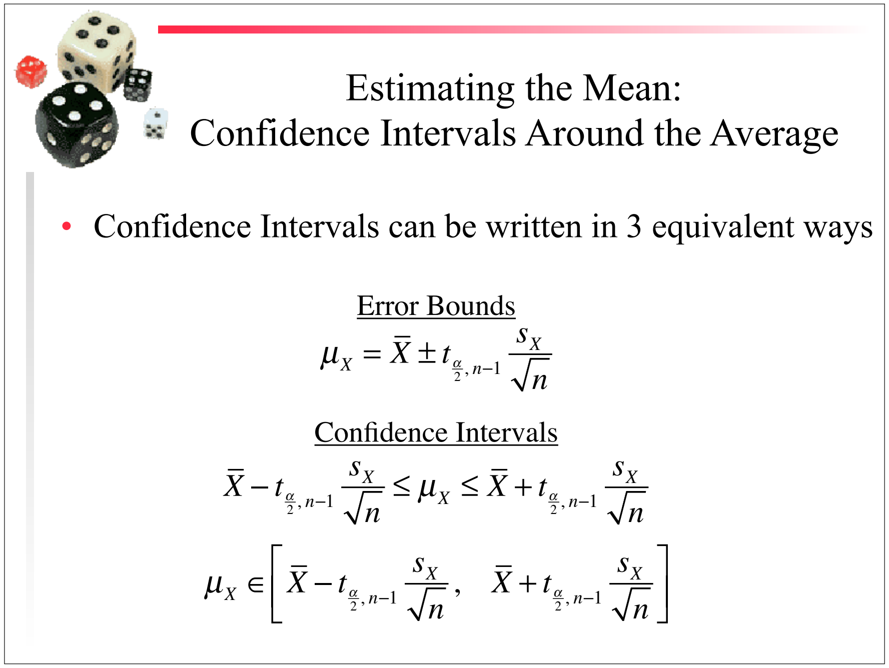
Confidence Interval

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

Student T distribution

Rearranging terms to isolate μ_X

$\alpha = 0.01$
$CL = 0.99$
$n = 55$
$df = 54$



Estimating the Mean: Confidence Intervals Around the Average

- Confidence Intervals can be written in 3 equivalent ways

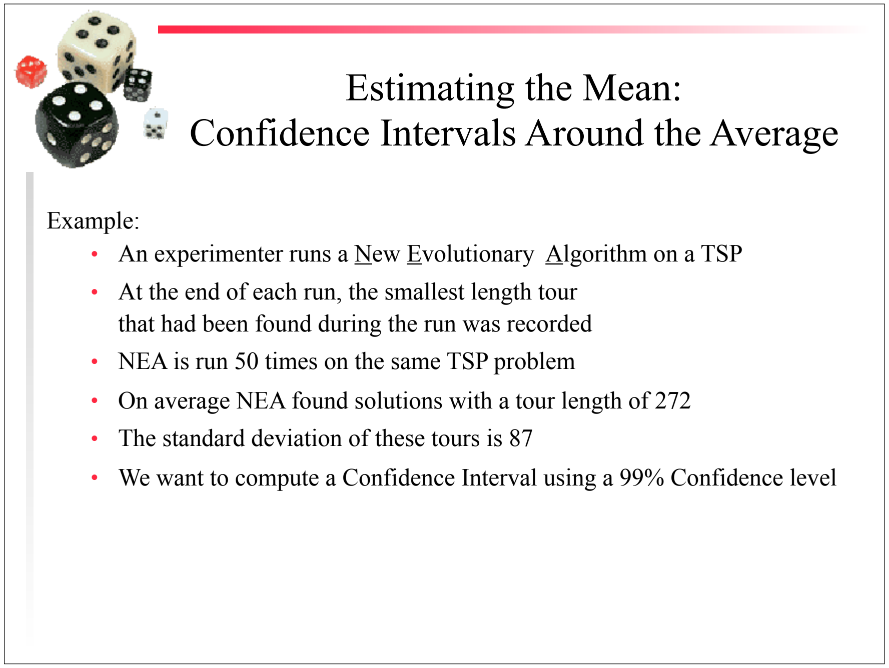
Error Bounds

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

Confidence Intervals

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

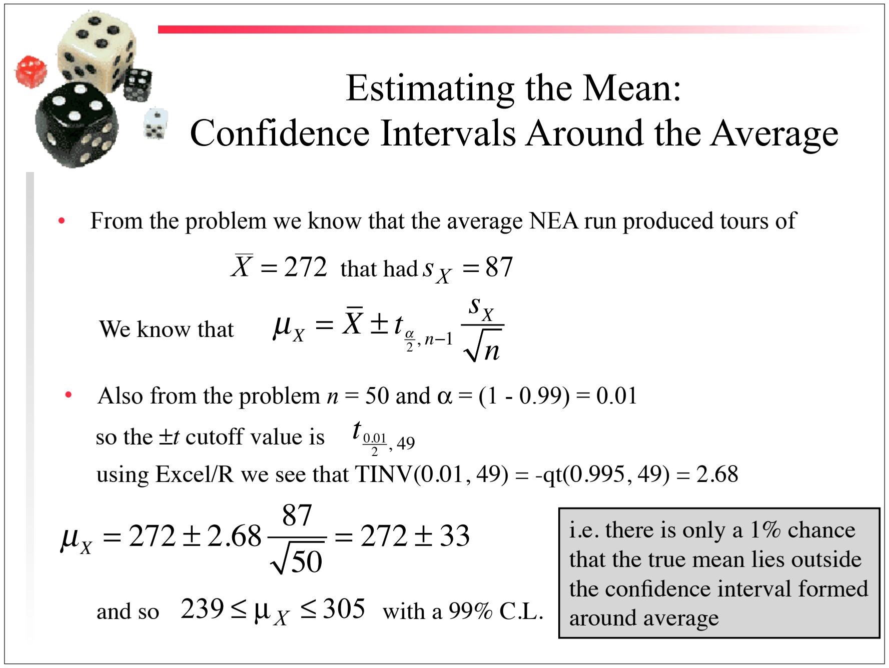
$$\mu_X \in \left[\bar{X} - t_{\frac{\alpha}{2},n-1} \frac{s_X}{\sqrt{n}}, \quad \bar{X} + t_{\frac{\alpha}{2},n-1} \frac{s_X}{\sqrt{n}} \right]$$



Estimating the Mean: Confidence Intervals Around the Average

Example:

- An experimenter runs a New 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



Estimating the Mean: Confidence Intervals Around the Average

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

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

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

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

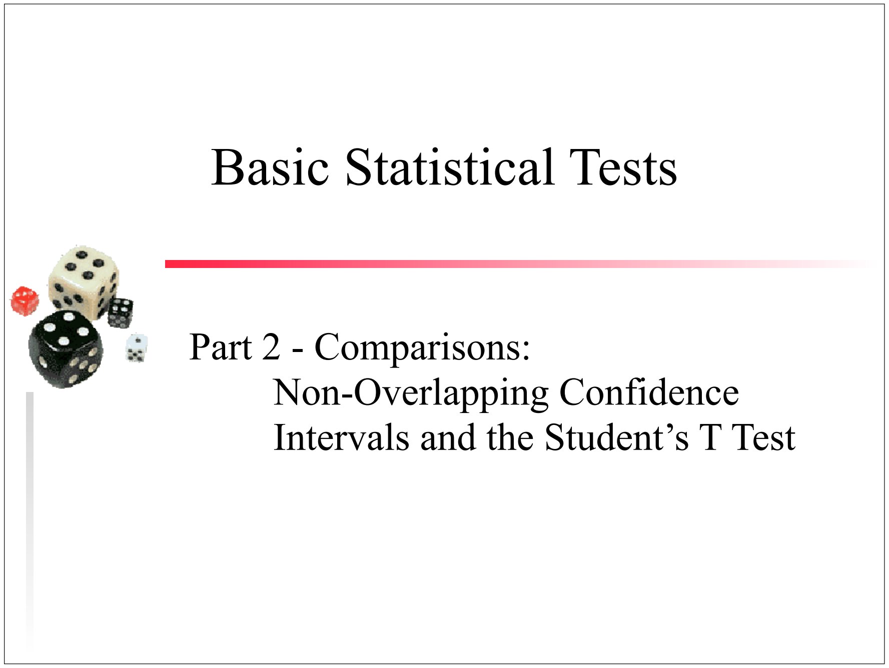
  so the ±*t* cutoff value is $t_{\frac{0.01}{2},49}$

  using Excel/R we see that TINV(0.01, 49) = -qt(0.995, 49) = 2.68

$$\mu_X = 272 \pm 2.68 \frac{87}{\sqrt{50}} = 272 \pm 33$$

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

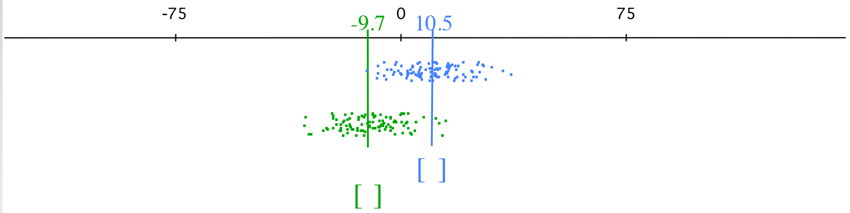
Using Confidence Intervals to Determine Whether My Way is Better

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

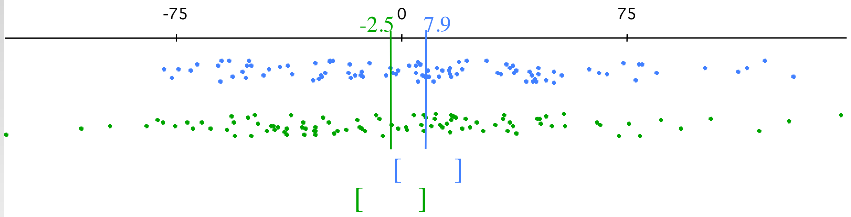


μ	σ
+10	10
-10	10

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

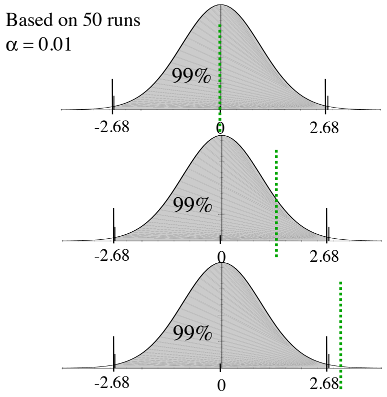
n	X	s_X	$1.96\frac{s_X}{\sqrt{n}}$	Lower	Uppe
100	7.9	47.1	9.2	-1.3	17.1
100	-2.5	52.1	10.2	-12.7	7.7

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

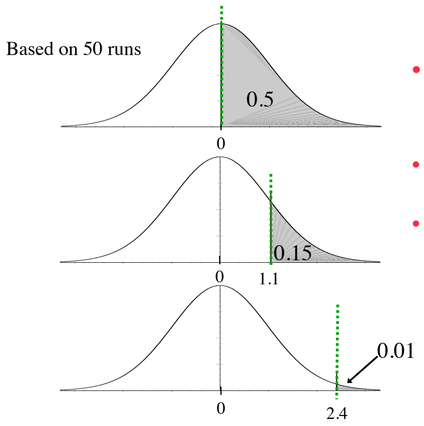


- Normalized difference called the *t* value

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

- 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
  - $1 - \Pr(T < t \text{ score})$
  - 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: $n_1 + n_2 - 2 = 2n - 2$
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
   - *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 $n_1 = n_2 = n$

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

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

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

In Excel: =ttest(A1:A50, B1:B50, 2, 2)

Variance Assumptions and the T Test

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

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_{X_1}^2}{n_1} + \frac{s_{X_2}^2}{n_2}}} \longleftarrow \text{Approximate variance not pooled}$$

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

called the Welch's T test

In Excel: =ttest(A1:A50, B1:B50, 2, 3)

t.test(): Welch's vs Student's

n = 80 for both OEA and NEA

```
> 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        -0.3143734
 average of OEA             average of NEA
5.119665                    7.657929
```

slightly modified for legibility

t.test(): Welch's vs Student's

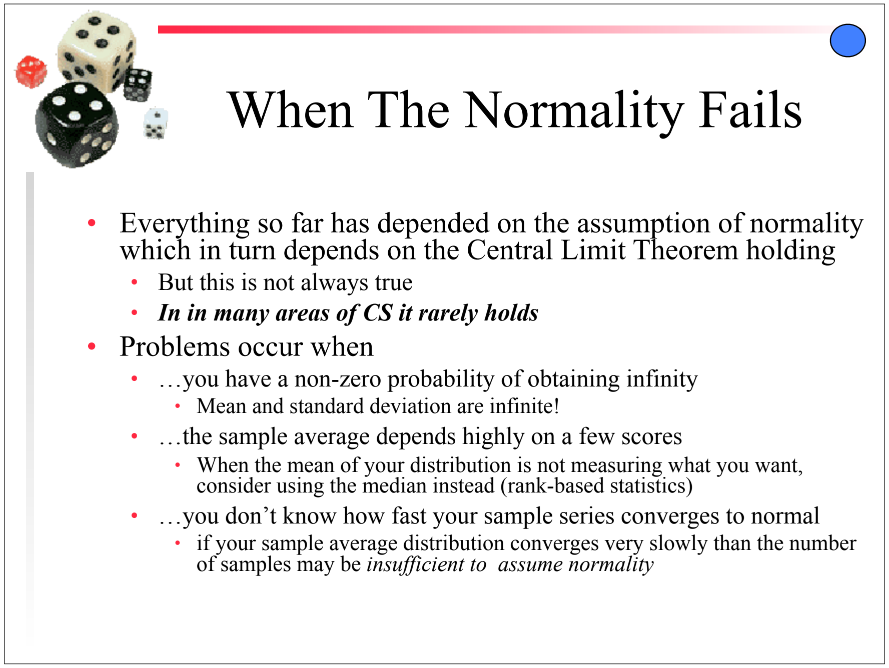
n = 80 for both OEA and NEA

```
> 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:
 -4.7615555        -0.3149714
 average of OEA             average of NEA
5.119665                    7.657929
```

slightly modified for legibility

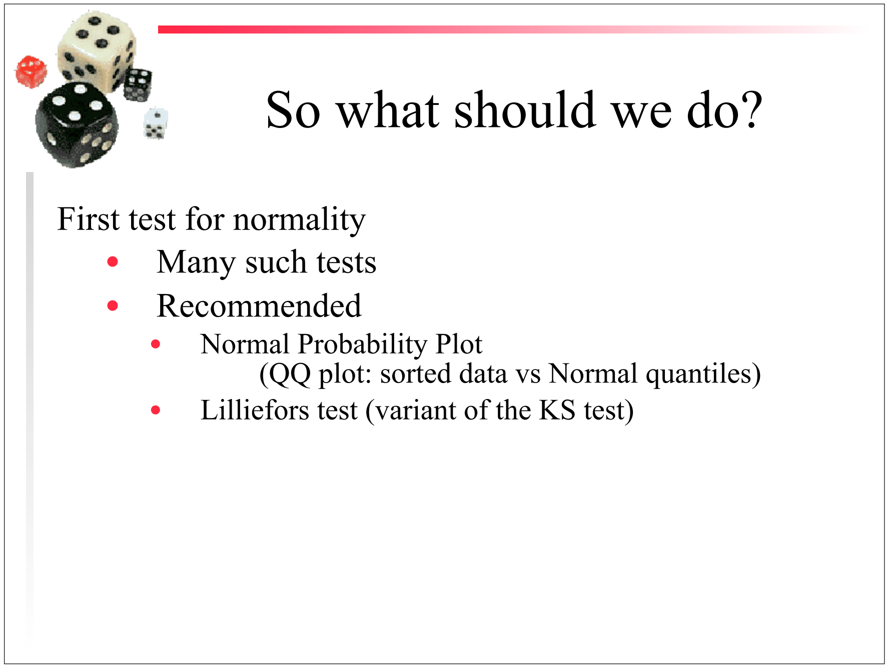
Tests on Non-Normally Distributed Random Variables

Non-Parametric Statistics



When The Normality Fails

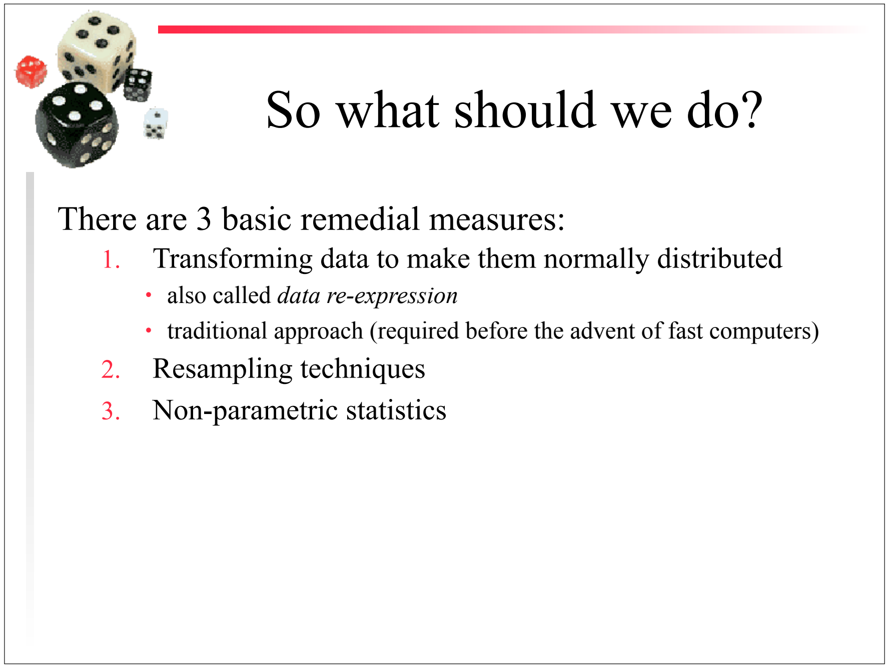
- Everything so far has depended on the assumption of normality which in turn depends on the Central Limit Theorem holding
  - But this is not always true
  - ***In in many areas of CS it rarely holds***
- Problems occur when
  - …you have a non-zero probability of obtaining infinity
    - Mean and standard 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

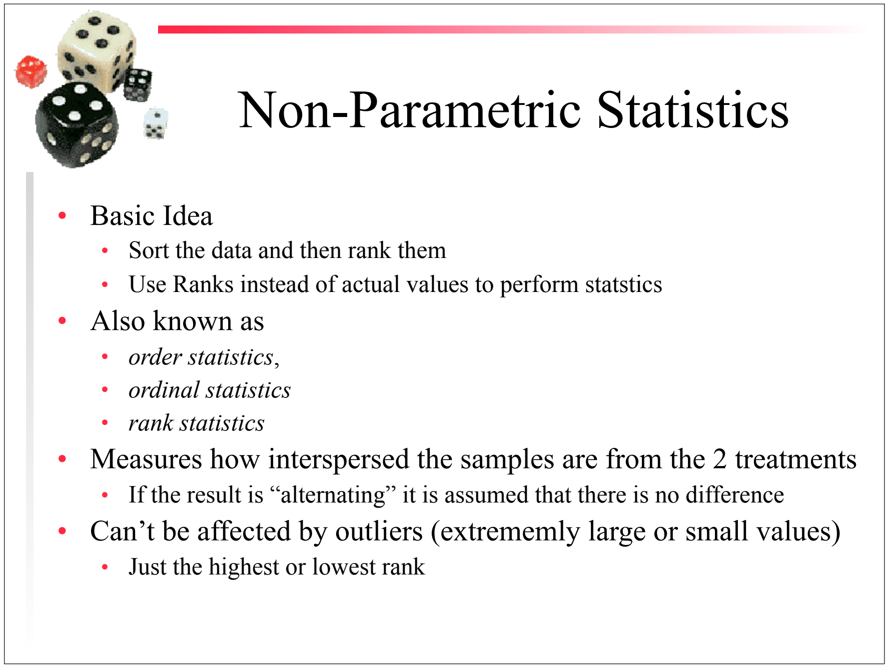
- Many such tests
- Recommended
  - Normal Probability Plot
    (QQ plot: sorted data vs Normal quantiles)
  - Lilliefors test (variant of the KS test)



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

Ranked Example

Two data sets combined 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

		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

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	

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
  - *b* v
- The l
- The u

However: Thompson-Savur is not common

Usually a Box-Plot is used to show where the "mass" of the data points are (based on interquartile range)

Box-Plot has the advantage of finding potential outliers

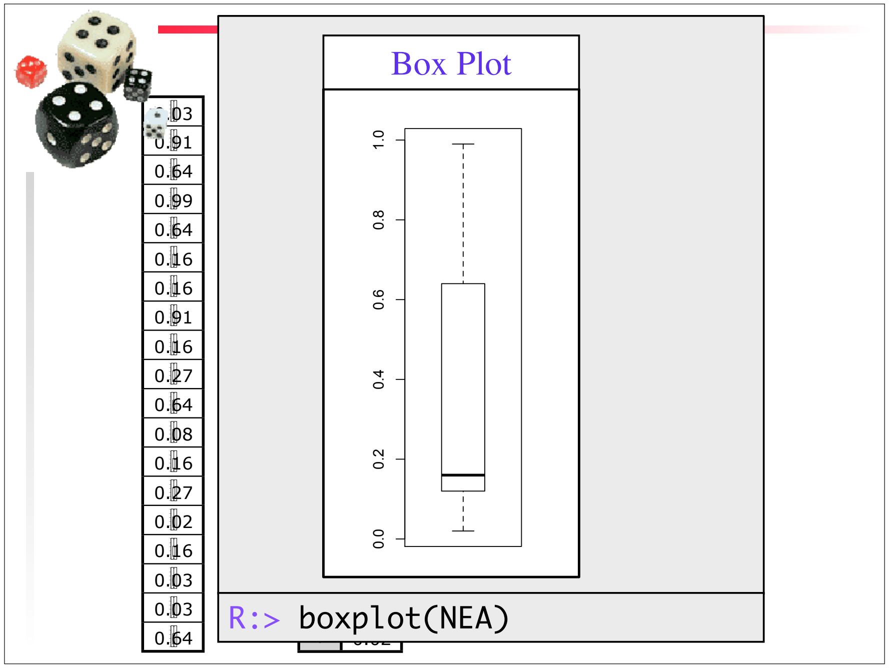
- Confidence Interval is $[value_l, value_u]$
  - i.e. $value_l \leq median \leq value_u$
  - With a confidence level of $1-\alpha$

Box Plot: Example

0.03
0.91
0.64
0.99
0.64
0.16
0.16
0.91
0.16
0.27
0.64
0.08
0.16
0.27
0.02
0.16
0.03
0.03
0.64

Sort Data

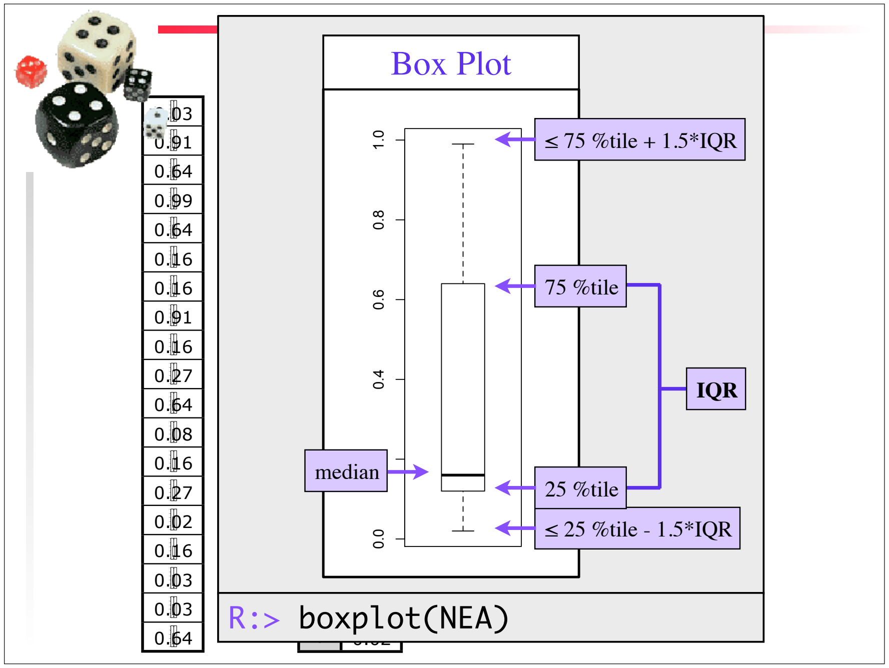
18	0.99
17	0.91
16	0.91
15	0.64
14	0.64
13	0.64
12	0.64
11	0.27
10	0.27
9	0.16
8	0.16
7	0.16
6	0.16
5	0.16
4	0.08
3	0.03
2	0.03
1	0.03
0	0.02



Box Plot

0.03
0.91
0.64
0.99
0.64
0.16
0.16
0.91
0.16
0.27
0.64
0.08
0.16
0.27
0.02
0.16
0.03
0.03
0.64

```
R:> boxplot(NEA)
```



Box Plot

0.03
0.91
0.64
0.99
0.64
0.16
0.16
0.91
0.16
0.27
0.64
0.08
0.16
0.27
0.02
0.16
0.03
0.03
0.64

≤ 75 %tile + 1.5*IQR

75 %tile

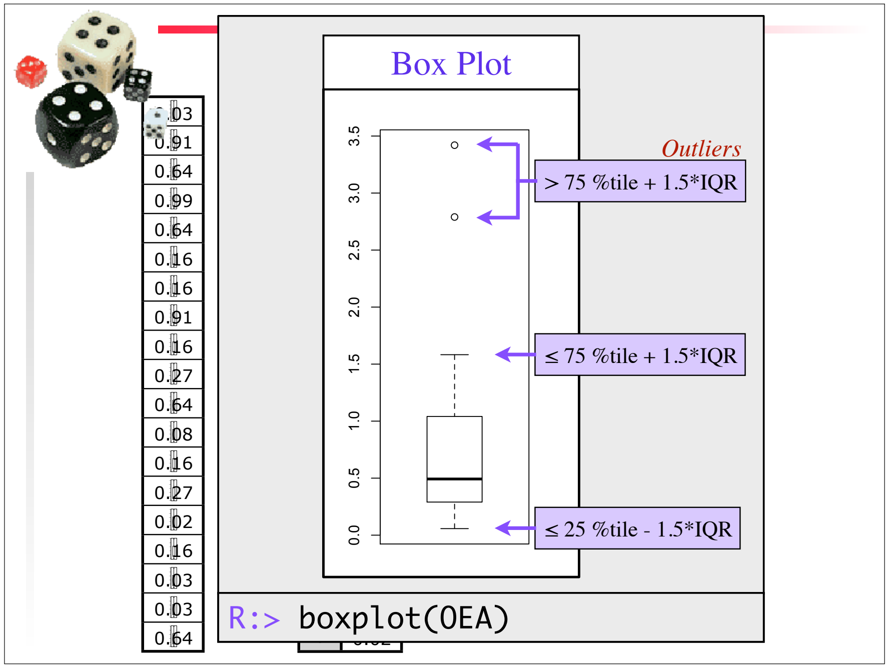
IQR

median

25 %tile

≤ 25 %tile - 1.5*IQR

```
R:> boxplot(NEA)
```



Box Plot

0.03
0.91
0.64
0.99
0.64
0.16
0.16
0.91
0.16
0.27
0.64
0.08
0.16
0.27
0.02
0.16
0.03
0.03
0.64

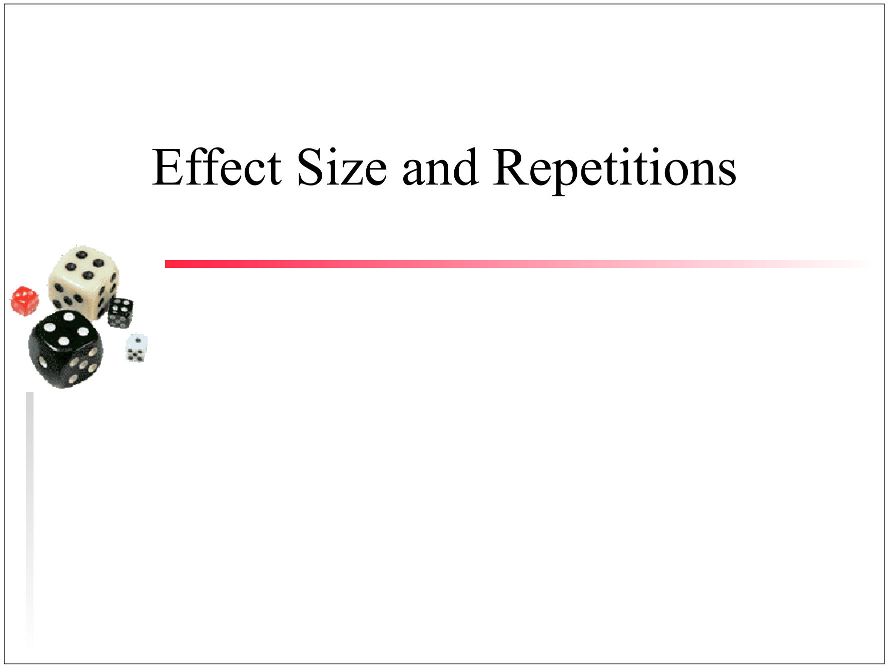
Outliers

> 75 %tile + 1.5*IQR

≤ 75 %tile + 1.5*IQR

≤ 25 %tile - 1.5*IQR

```
R:> boxplot(OEA)
```



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 2.0x10^{-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

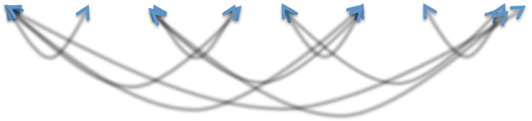
Analysis of Variance (ANOVA)

Fitness Values

no xover	*xover* = 1pt	*xover* = 2pt	*xover* = 3pt	*xover* = 4pt
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
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Question:
Do crossover settings make a difference at all?

avg fitness	4.02	8.13	5.09	7.02	5.76
std dev	0.451	0.313	0.424	0.478	0.471



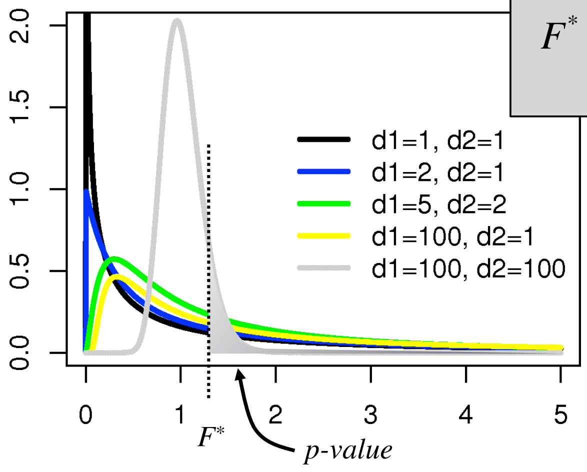
all pairwise T test

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

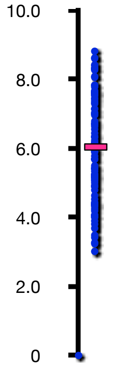
$$F^* = \frac{s^2_{X_1}}{s^2_{X_2}}$$



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

ANOVA: Discrete Levels

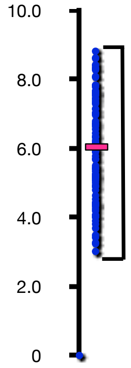
Average of Y (no model)



ANOVA: Discrete Levels

Variance of Y (no model)

represented as a std deviation



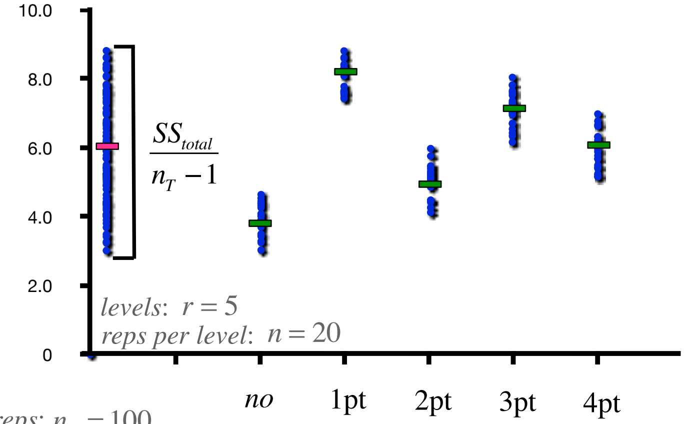
$$s^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2$$

total reps: $n_T = 100$

ANOVA: Discrete Levels

Add average for each level

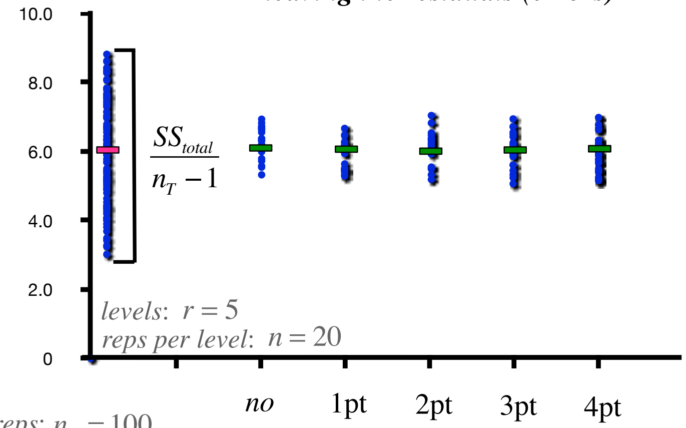
a model of the behavior of the system



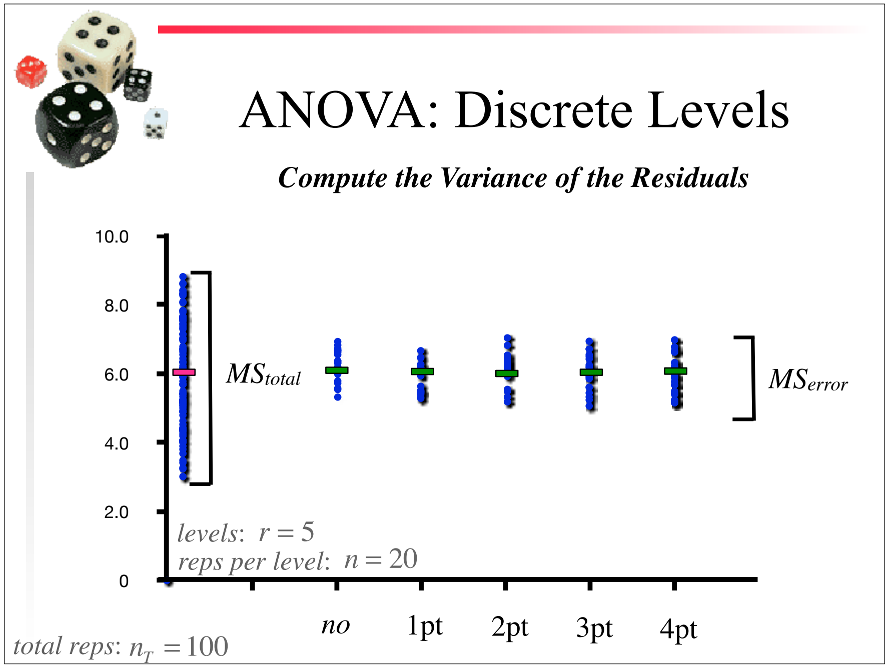
total reps: $n_T = 100$

ANOVA: Discrete Levels

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



total reps: $n_T = 100$



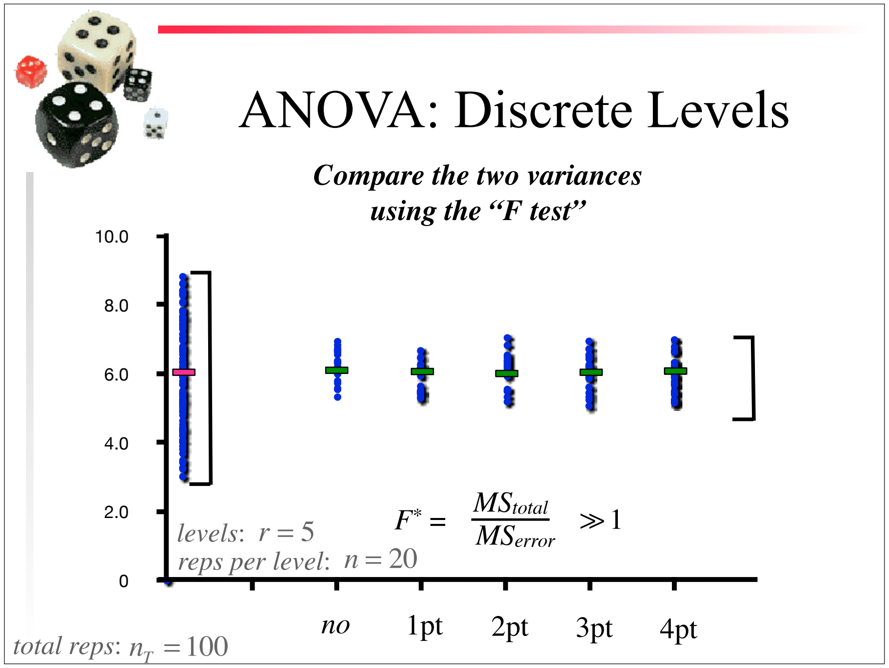
ANOVA: Discrete Levels

Compute the Variance of the Residuals

levels: $r = 5$

reps per level: $n = 20$

total reps: $n_T = 100$



ANOVA: Discrete Levels

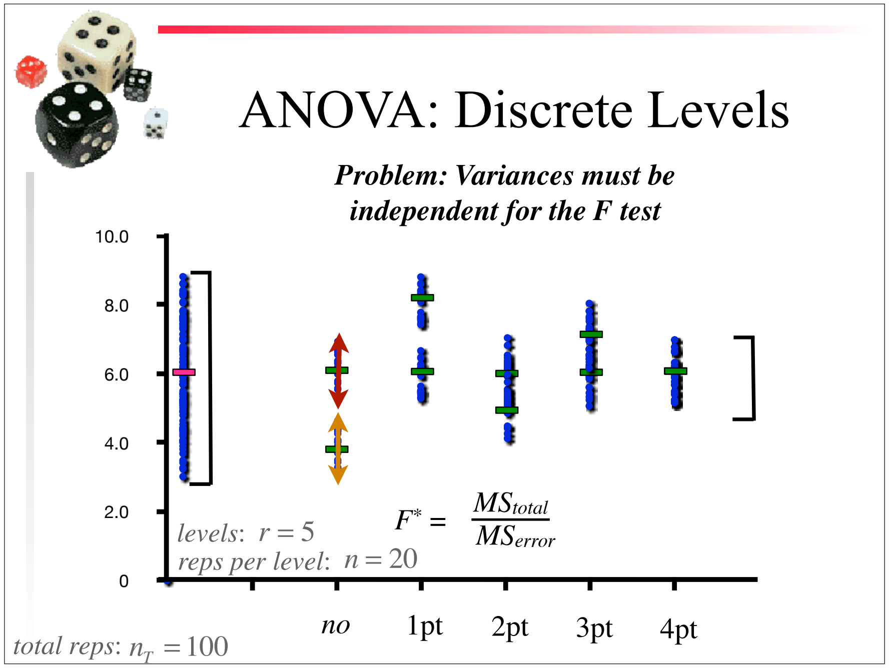
Compare the two variances using the "F test"

$$F^* = \frac{MS_{total}}{MS_{error}} \gg 1$$

levels: $r = 5$

reps per level: $n = 20$

total reps: $n_T = 100$



ANOVA: Discrete Levels

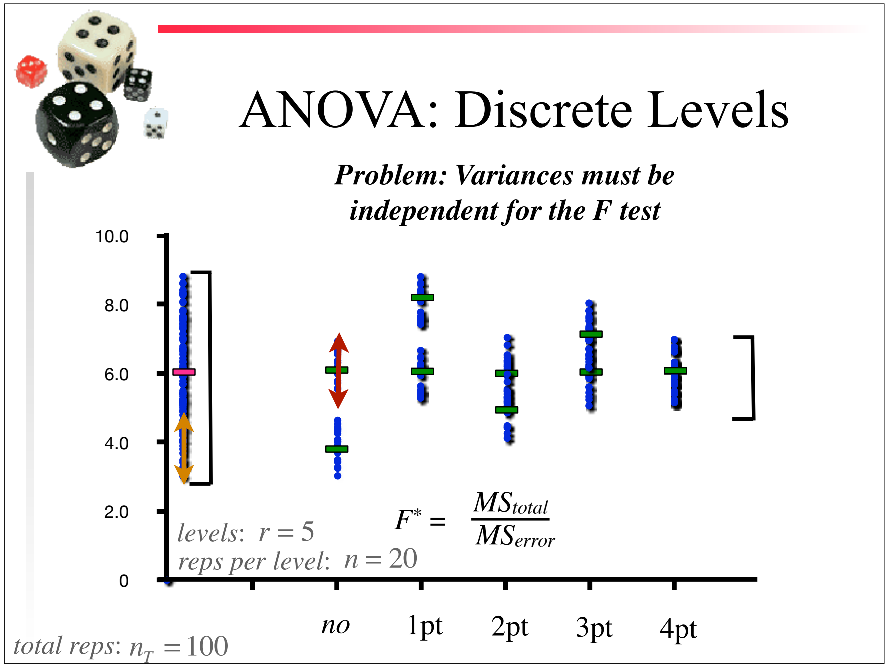
Problem: Variances must be independent for the F test

$$F^* = \frac{MS_{total}}{MS_{error}}$$

levels: $r = 5$

reps per level: $n = 20$

total reps: $n_T = 100$



ANOVA: Discrete Levels

Problem: Variances must be independent for the F test

$$F^* = \frac{MS_{total}}{MS_{error}}$$

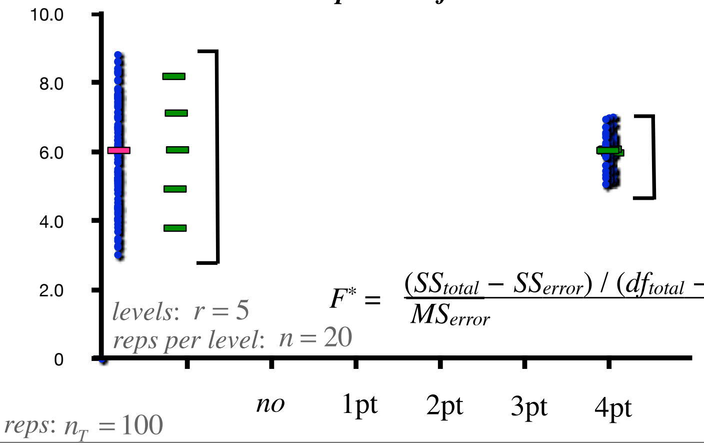
levels: $r = 5$

reps per level: $n = 20$

total reps: $n_T = 100$

ANOVA: Discrete Levels

Problem: Variances must be independent for the F test



levels: $r = 5$

reps per level: $n = 20$

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

total reps: $n_T = 100$

ANOVA: Discrete Levels

Assumption:

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

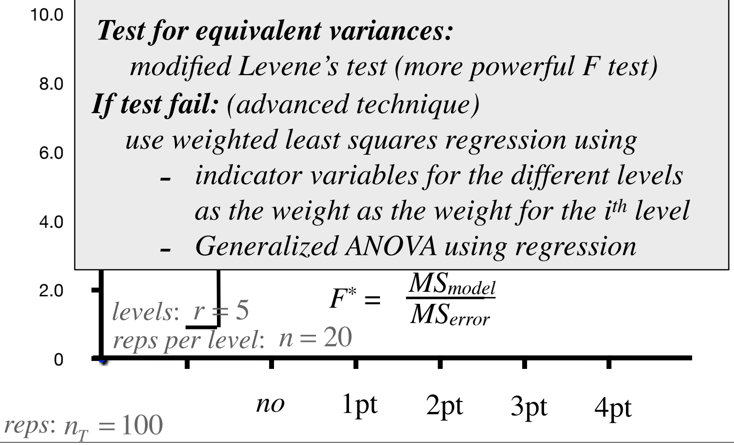
Test for equivalent variances:

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*



levels: $r = 5$

reps per level: $n = 20$

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

total reps: $n_T = 100$

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) = 8.19E-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 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*

original T test comparison

comparing level *i* with level *j* across the ANOVA model

$$t\ value = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_{X_1}^2}{n_1} + \frac{s_{X_2}^2}{n_2}}} \longrightarrow t\ value = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}}$$

Pairwise Comparisons between Factor-Level Means

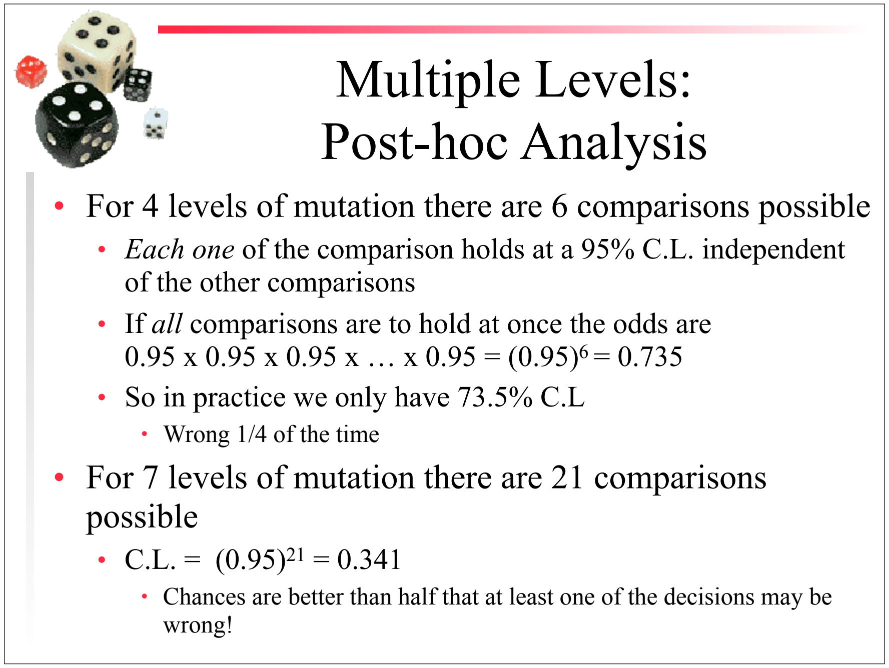
- This is determined by a series of pair-wise T tests
- However, commonly uses pooled information from the

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

original T test comparison

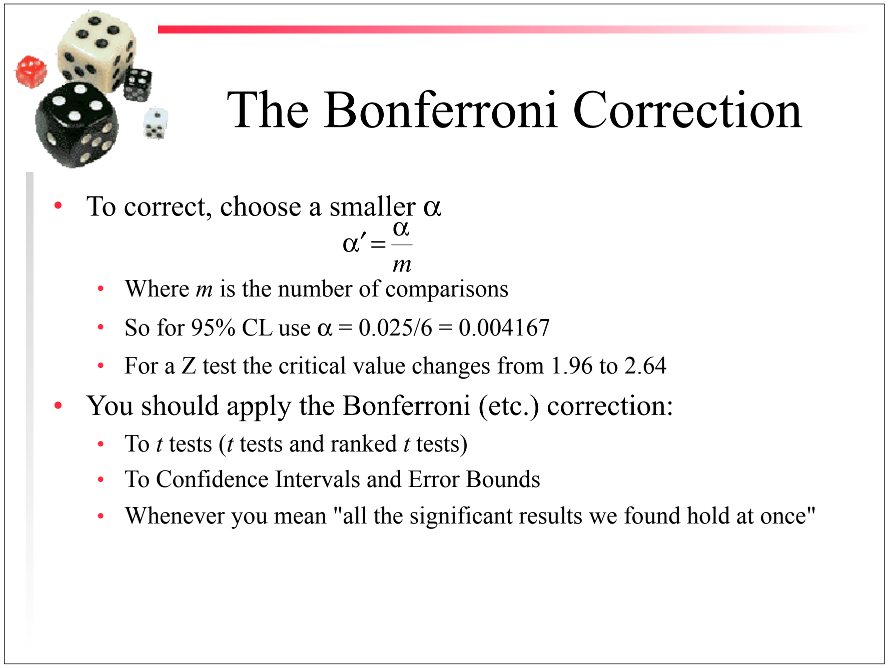
comparing level *i* with level *j* across the ANOVA model

$$t\ value = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_{X_1}^2}{n_1} + \frac{s_{X_2}^2}{n_2}}} \longrightarrow t\ value = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}}$$



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!

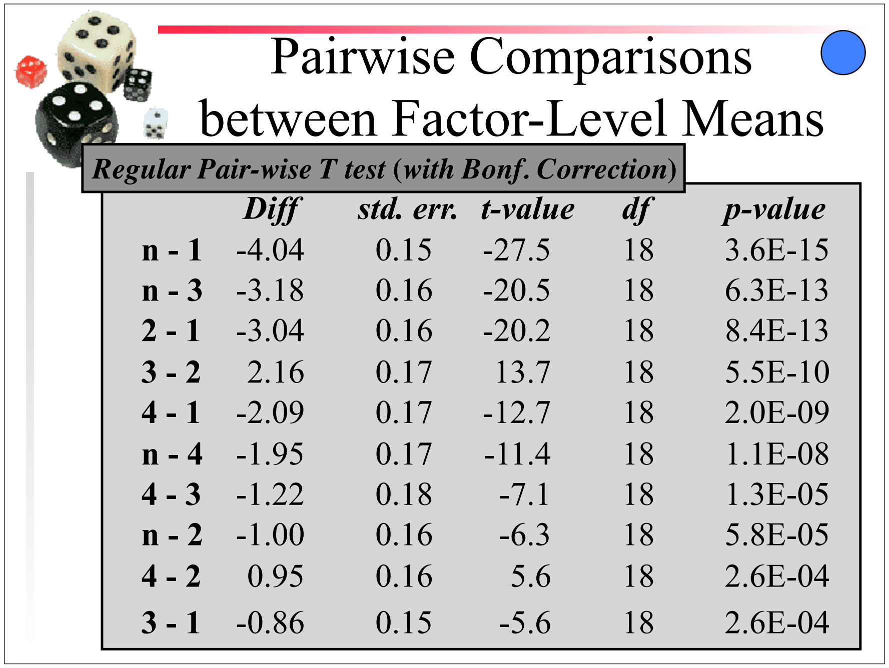


The Bonferroni Correction

- To correct, choose a smaller α

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

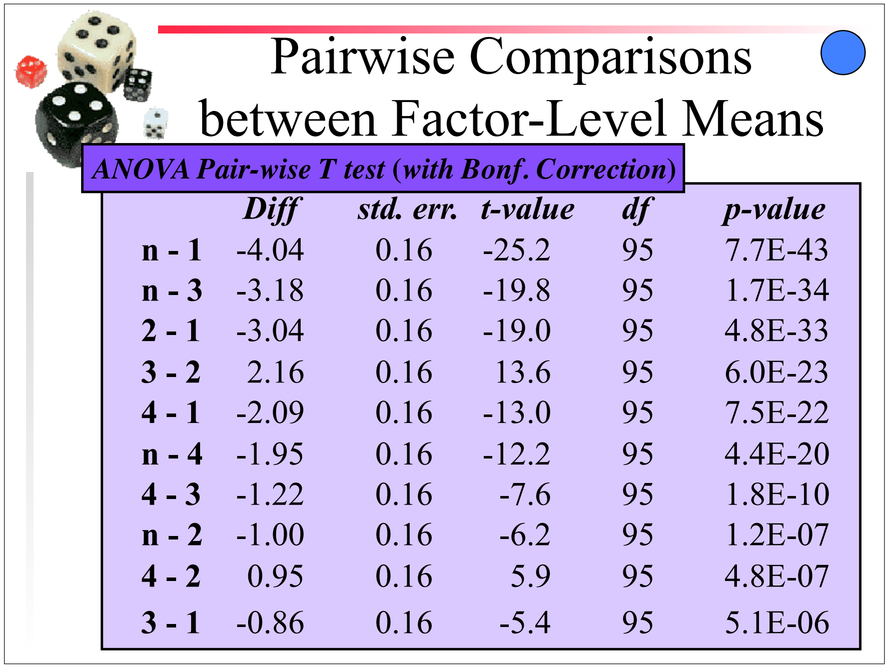
  - Where *m* is the number of comparisons
  - So for 95% CL use α = 0.025/6 = 0.004167
  - For a Z test the critical value changes from 1.96 to 2.64
- You should apply the Bonferroni (etc.) correction:
  - To *t* tests (*t* tests and ranked *t* tests)
  - To Confidence Intervals and Error Bounds
  - Whenever 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

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

	Diff	*std. err.*	*t-value*	*df*	*p-value*
[illegible]	4.04	0.16	25.2	95	7.7E-43

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

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

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

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

Student-T with Bonf. Correction

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

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

Other Post-Hoc Corrections

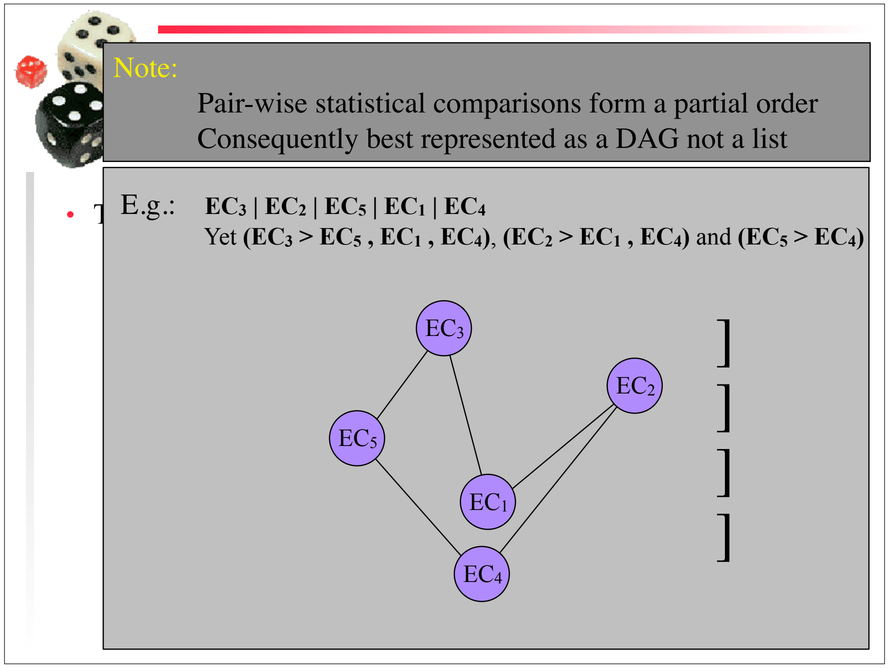
- Holm -Sidak (really Bonferroni done “right”)
  - Order the p-values from smallest to largest
  - Compare the smallest p-value to α/k (regular Bonferroni)
  - If that p-value is less than α/k, then accept that alternative hypothesis
  - Now look at the next smallest p-value at α / (k − 1)
  - Continue 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
      - actually there are 5C2 = 10 implicit comparisons
        - because we did not know how many comparisons there would be apriori

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
      - actually there are 5C2 = 10 implicit comparisons
        - because we did not know how many comparisons there would be apriori



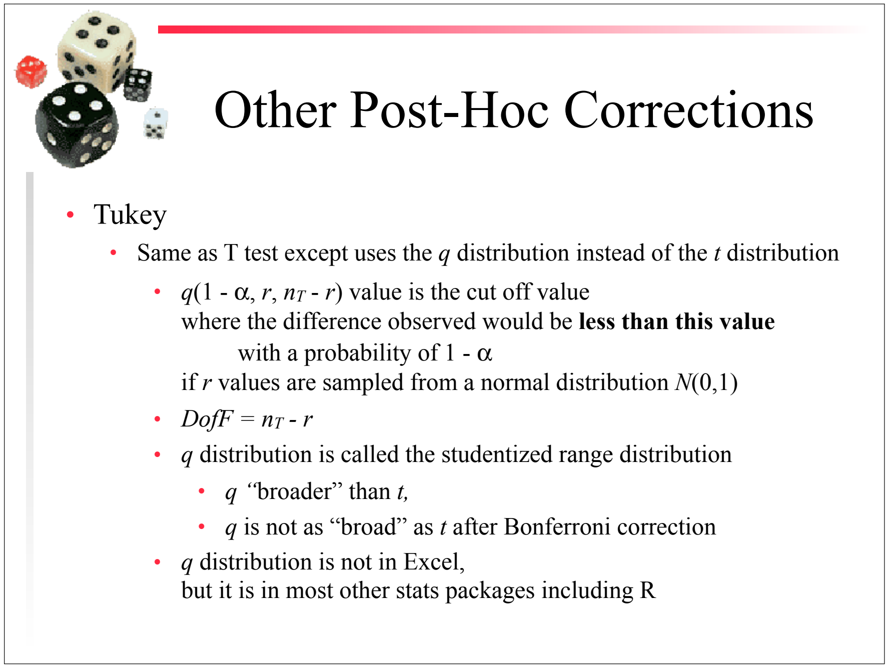
Note:

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

- T

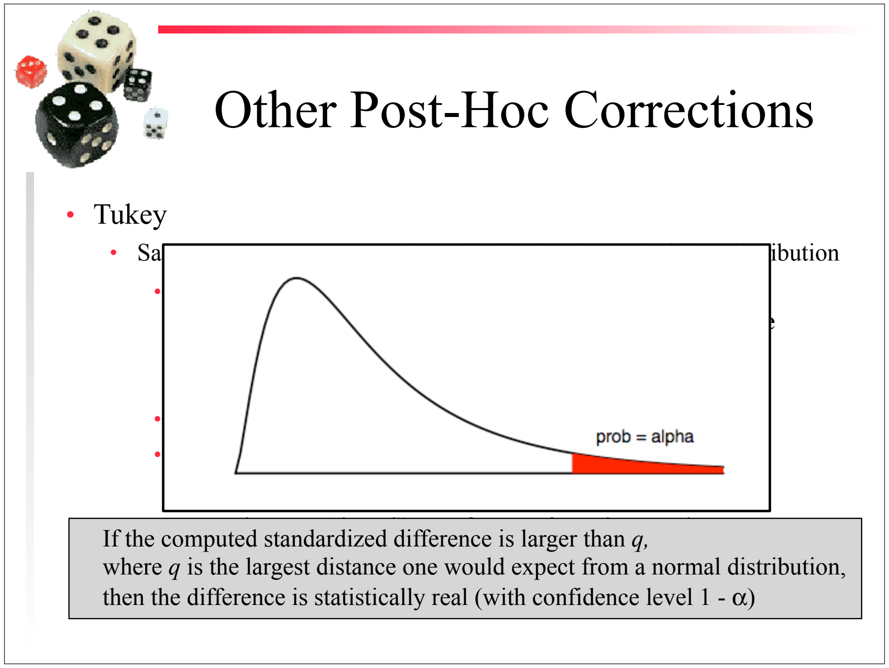
E.g.: **EC_3 | EC_2 | EC_5 | EC_1 | EC_4**
Yet **($EC_3 > EC_5$, EC_1 , EC_4)**, **($EC_2 > EC_1$, EC_4)** and **($EC_5 > EC_4$)**

EC3
EC2
EC5
EC1
EC4
]
]
]
]



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

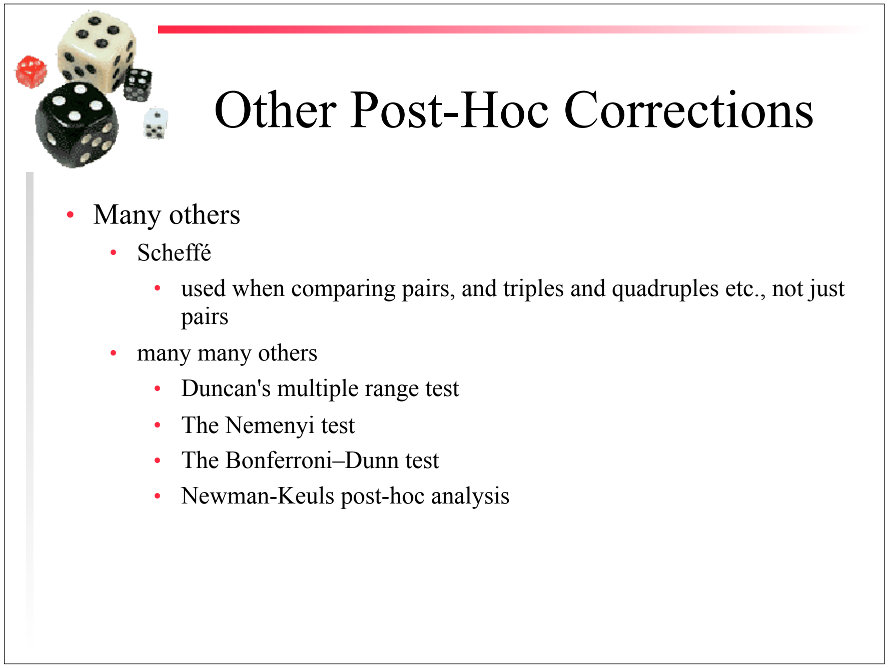


Other Post-Hoc Corrections

- Tukey
  - Sa ibution

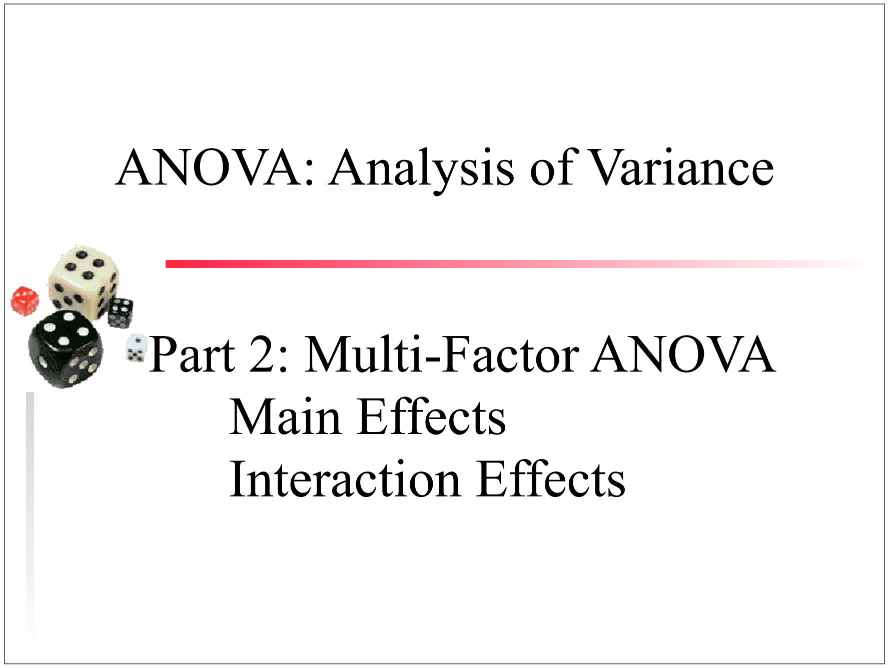
e

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 $1 - \alpha$)



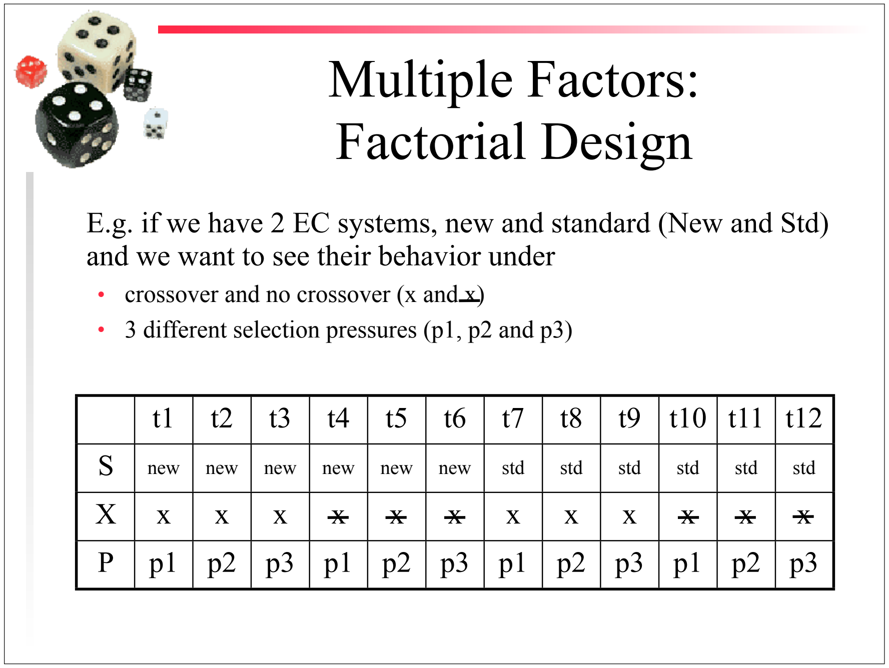
Other Post-Hoc Corrections

- 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



ANOVA: Analysis of Variance

Part 2: Multi-Factor ANOVA
Main Effects
Interaction Effects

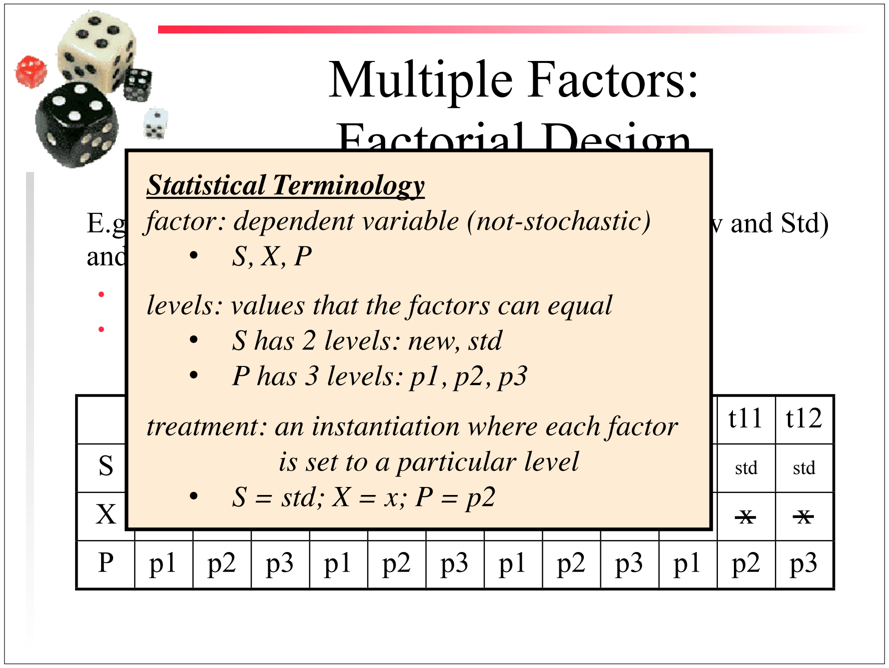


Multiple Factors: Factorial Design

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)

	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

E.g ... v and Std)
and

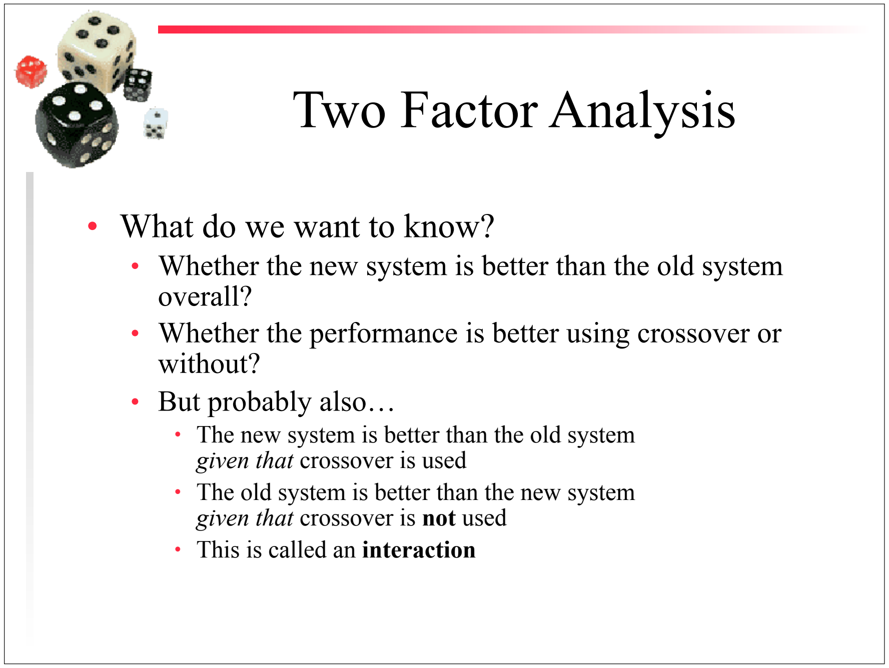
Statistical Terminology

factor: dependent variable (not-stochastic)
- *S, X, P*

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*

											t11	t12
S											std	std
X											~~x~~	~~x~~
P	p1	p2	p3	p1	p2	p3	p1	p2	p3	p1	p2	p3



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 + ε

error term

$n_T = 180$
$a = 2$
$b = 2$
$c = 3$
$n = 15$

Multi-Factor ANOVA: Results Report

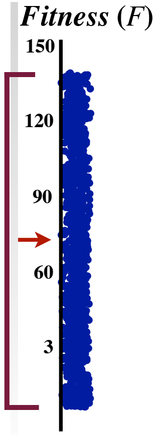
Source	df	SS	MS	F-ratio	p-value
Const	1	16970	16970	12930	≤ 0.0001
S	1	113	113	86.5	≤ 0.0001
X	1	775	775	591.0	≤ 0.0001
P	2	939	469.5	357.7	≤ 0.0001
S*X	1	4.05	4.05	3.1	0.0809
S*P	2	307	153.5	116.8	≤ 0.0001
X*P	2	0.570	0.285	0.217	0.8049
S*X*P	2	0.308	0.154	0.117	0.8892
Error	168	220.5	1.312		
Total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

Part 3

Regression
by means of Least Squares

Linear Regression

Factor in Population Size

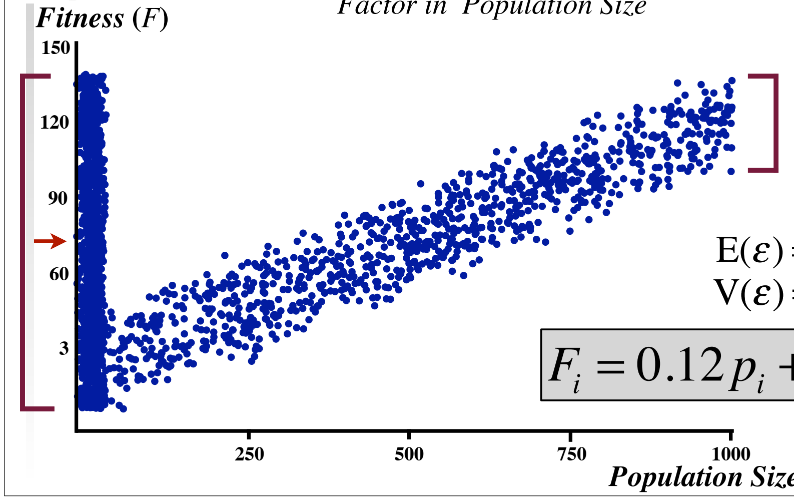


$$E(\varepsilon_F) = 0$$
$$V(\varepsilon_F) = \sigma_F^2$$

$$\boxed{F_i = 72 + \varepsilon_F}$$

Linear Regression

Factor in Population Size

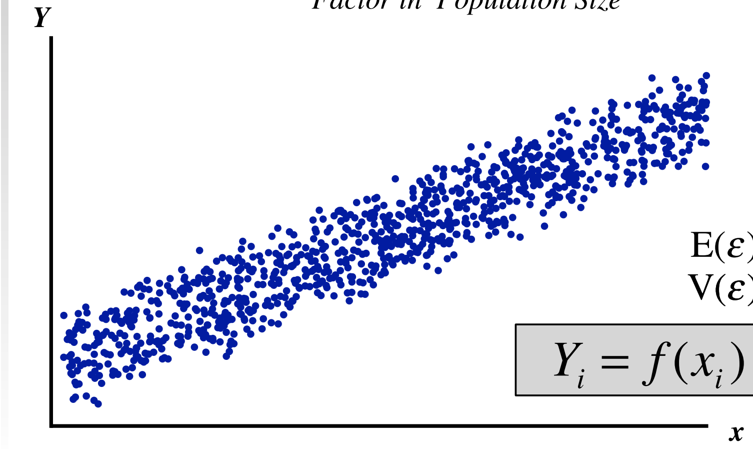


$$E(\varepsilon) = 0$$
$$V(\varepsilon) = \sigma^2$$

$$\boxed{F_i = 0.12 p_i + \varepsilon}$$

Linear Regression

Factor in Population Size



$$E(\varepsilon) = 0$$
$$V(\varepsilon) = \sigma^2$$

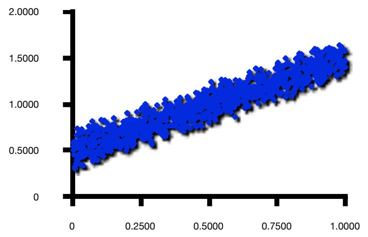
$$\boxed{Y_i = f(x_i) + \varepsilon}$$

Modeling Response Behavior: Treating X as a factor

- Simplest model - linear relationship

$Y_i = f(x_i) + \varepsilon$ with $f(x_i) = \beta_0 + \beta_1 x_i$

$$\Longrightarrow \quad Y_i = \beta_0 + \beta_1 x_i + \varepsilon$$



Two parameters β_0 and β_1 define the function

18

Linear Regression by Means of Least Squares

- Idea:
  - From sample pairs $\{(Y_1, x_1), (Y_2, x_2), \dots, (Y_n, x_n)\}$ determine b_0, b_1
    - Estimates of the two unknowns β_0, β_1

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon \quad \Longrightarrow \quad \hat{Y}_i = b_0 + b_1 x_i$$

- 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 samp …, $(Y_n, x_n)\}$
    determine l
    - Estimates

$$Y_i = \beta_0 \qquad b_0 + b_1 x_i$$

    - chosen su rs is minimized
    - i.e. find th
      t l error

Error

$$e_i = Y_i - \hat{Y}$$

Error

$$e_i = Y_i - b_0 - b_1 x_i$$

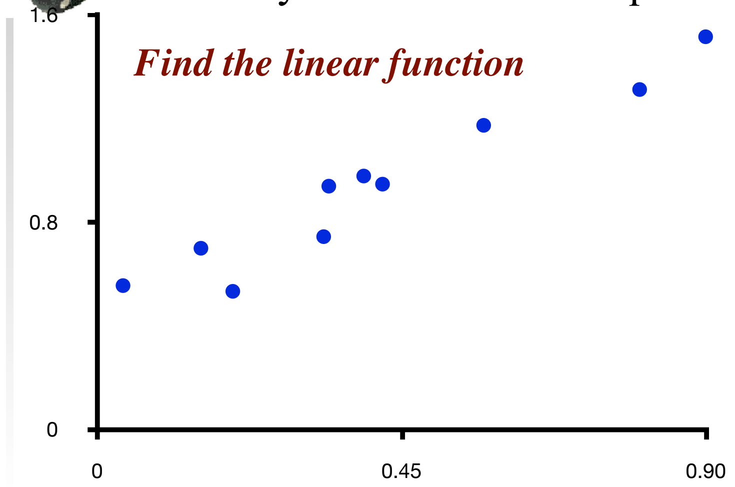
Squared Error

$$e_i^2 = (Y_i - b_0 - b_1 x_i)^2$$

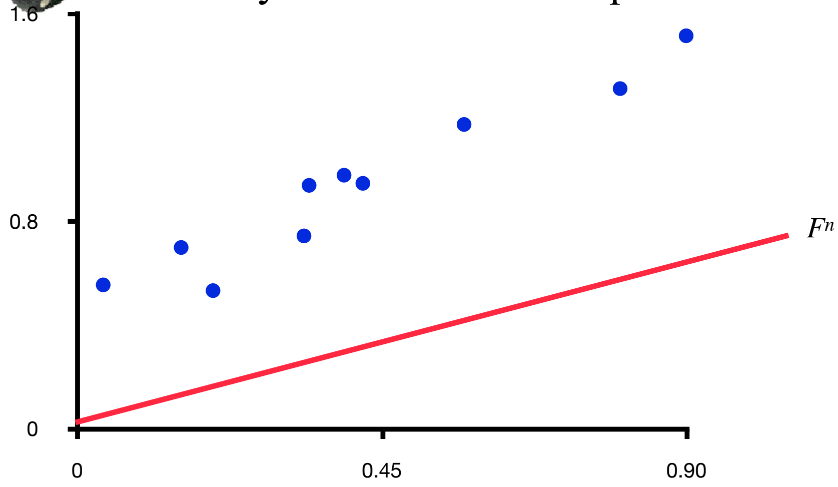
Sum of Squared Errors

$$SSE = \sum e_i^2$$

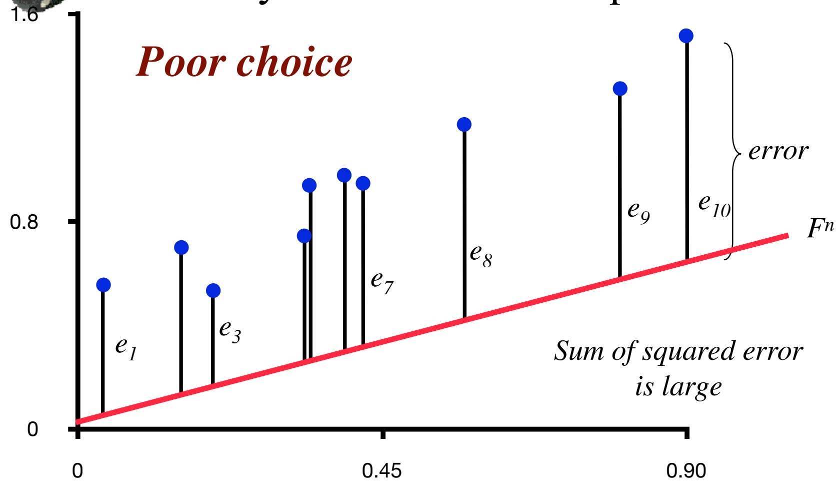
Linear Regression by Means of Least Squares

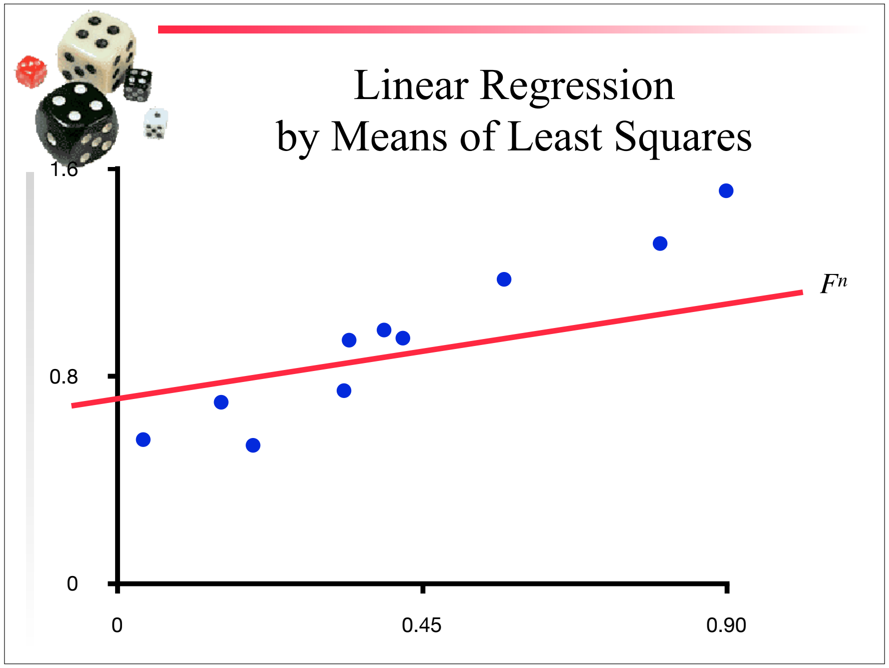


Linear Regression by Means of Least Squares

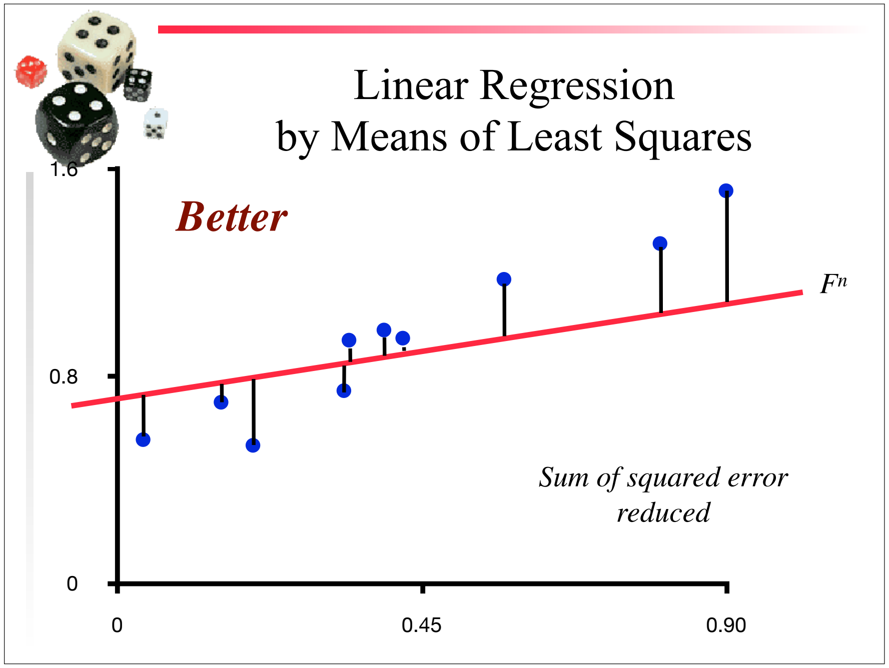


Linear Regression by Means of Least Squares





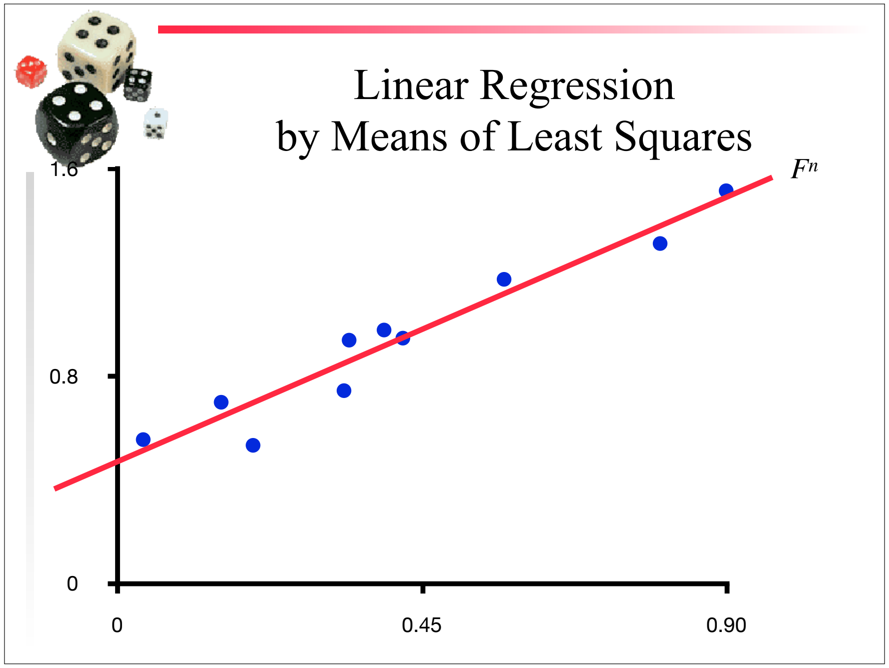
Linear Regression by Means of Least Squares



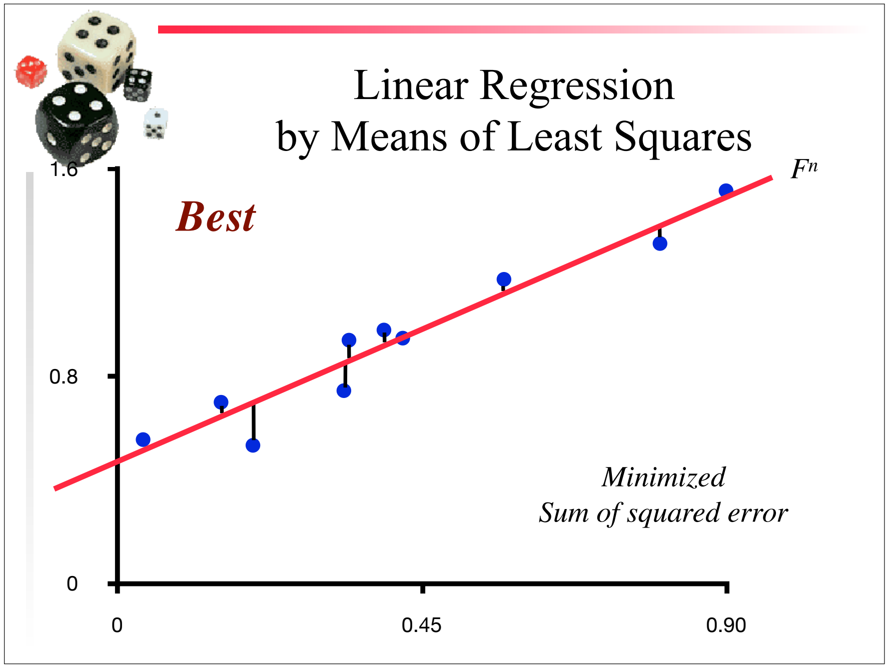
Linear Regression by Means of Least Squares

Better

Sum of squared error reduced



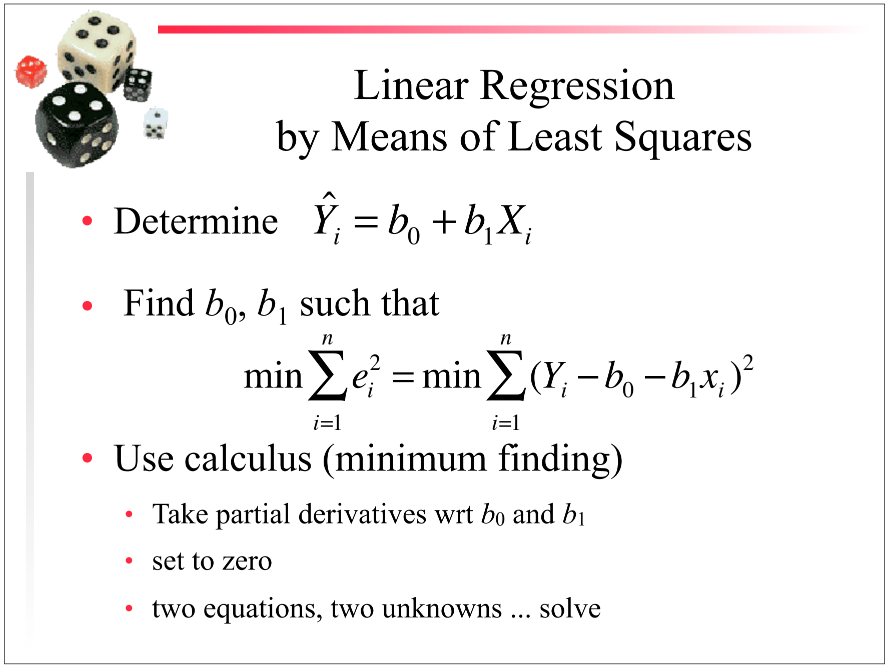
Linear Regression by Means of Least Squares



Linear Regression by Means of Least Squares

Best

Minimized Sum of squared error

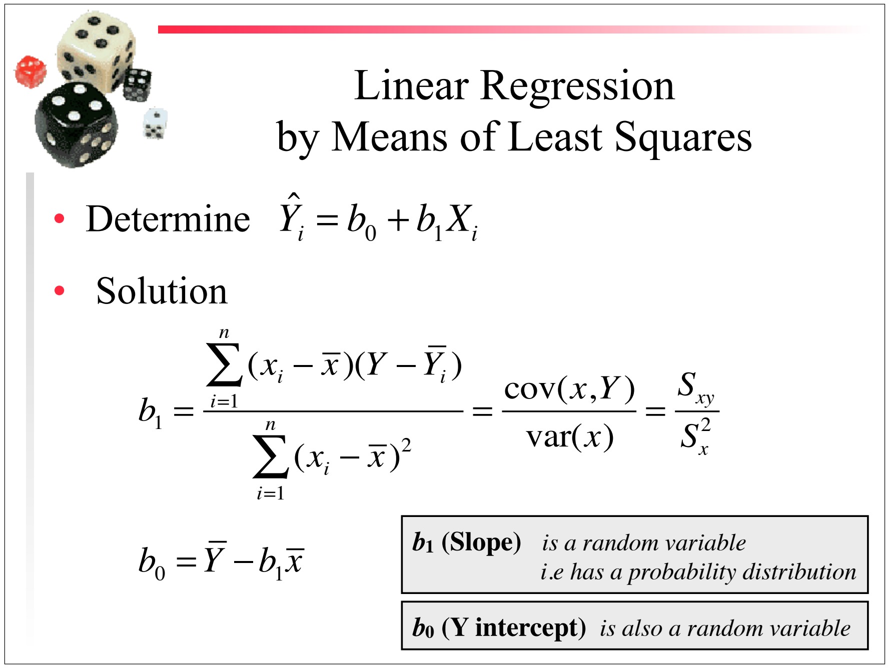


Linear Regression by Means of Least Squares

- 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



Linear Regression by Means of Least Squares

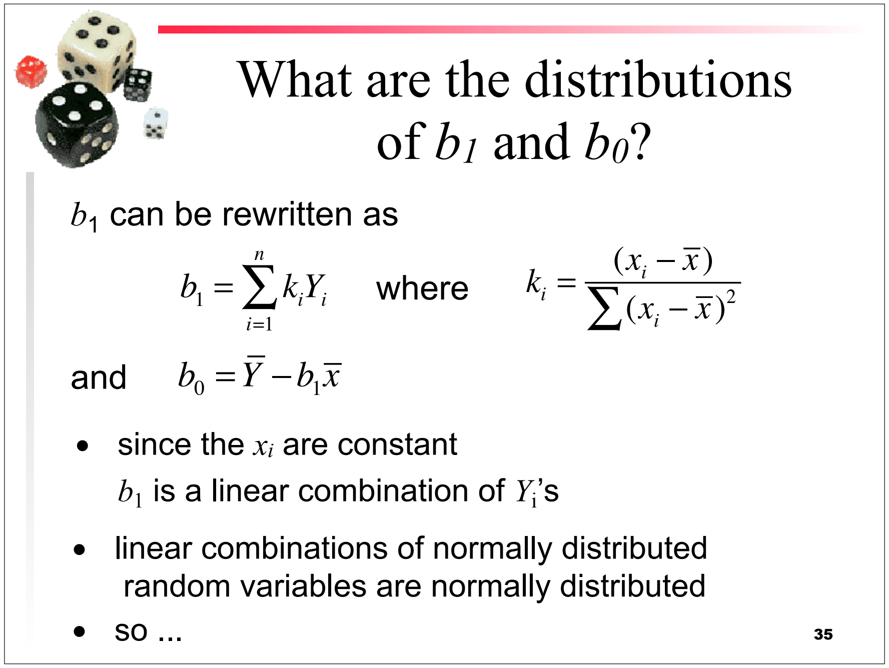
- Determine $\hat{Y}_i = b_0 + b_1 X_i$
- Solution

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

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

b_1 (Slope) *is a random variable i.e has a probability distribution*

b_0 (Y intercept) *is also a random variable*



What are the distributions of b_1 and b_0?

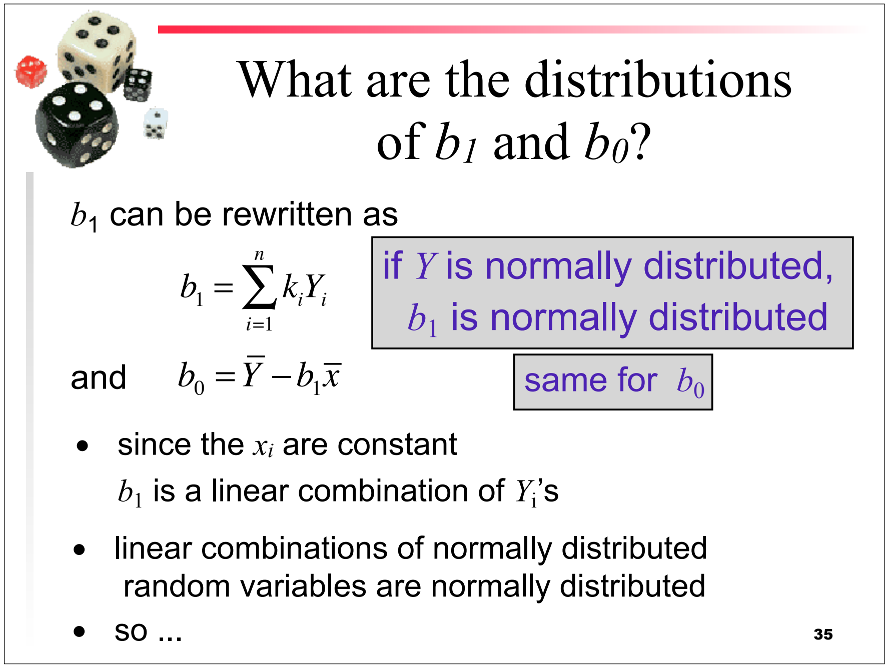
b_1 can be rewritten as

$$b_1 = \sum_{i=1}^{n} k_i Y_i \quad \text{where} \quad k_i = \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

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 ...

35



What are the distributions of b_1 and b_0?

b_1 can be rewritten as

$$b_1 = \sum_{i=1}^{n} k_i Y_i$$

if Y is normally distributed, b_1 is normally distributed

and $b_0 = \bar{Y} - b_1 \bar{x}$

same for b_0

- 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 ...

35

Expectation and Variance of b_1 *and* b_0

b_1 and b_0 can be thought of as sample means

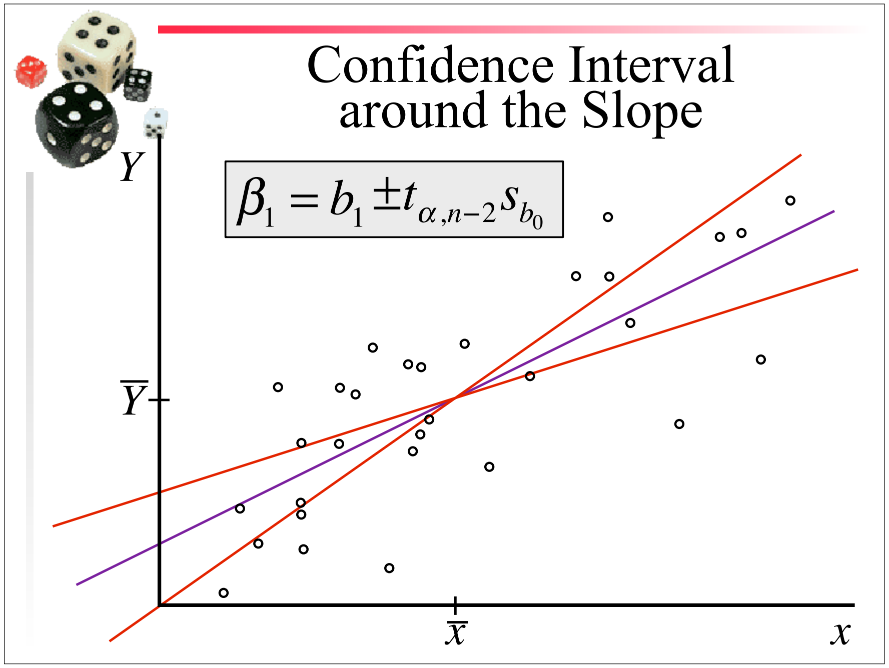
$$E(b_0)=\beta_0 \qquad E(b_1)=\beta_1$$

and they have associated variances

$$V(b_1)=\frac{\sigma_Y^2}{nS_x^2} \quad \Rightarrow \quad s_{b_1}^2=\frac{MS_{error}}{nS_x^2}$$

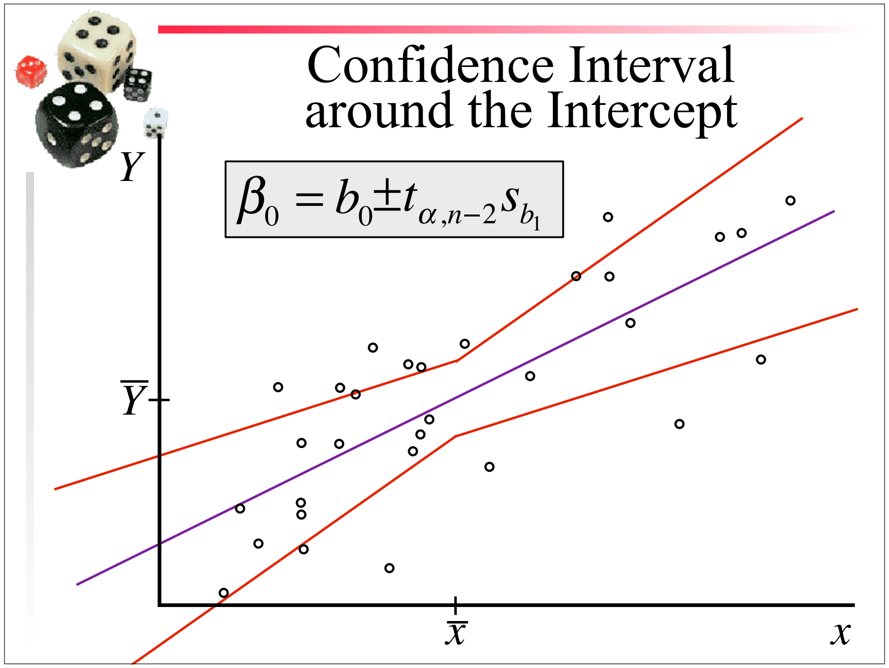
$$V(b_0)=\left(1+\frac{\overline{x}^2}{S_x^2}\right)\frac{\sigma_Y^2}{n} \quad \Rightarrow \quad s_{b_0}^2=\left(1+\frac{\overline{x}^2}{S_x^2}\right)\frac{MS_{error}}{n}$$

37



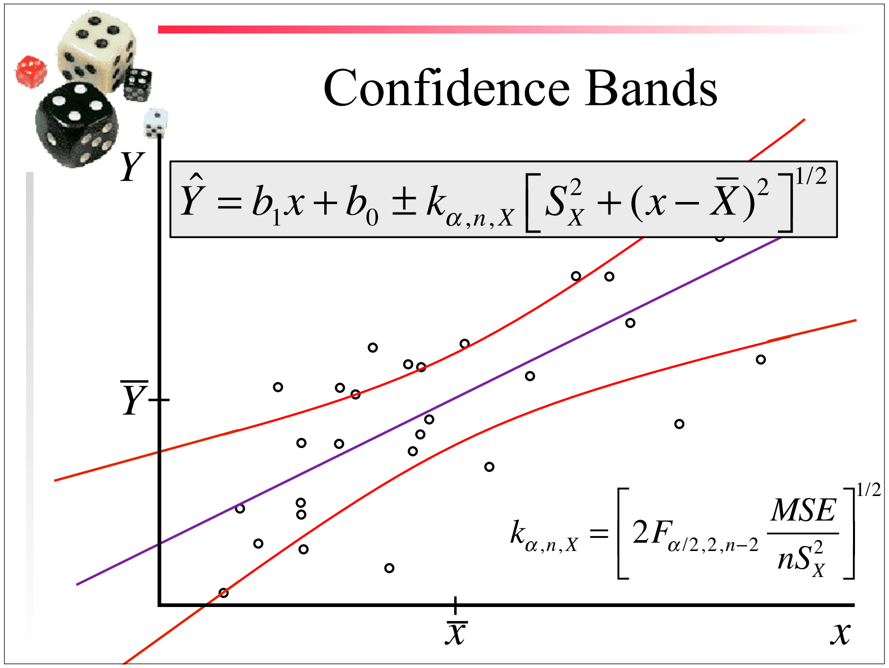
Confidence Interval around the Slope

$$\beta_1=b_1\pm t_{\alpha,n-2}s_{b_0}$$



Confidence Interval around the Intercept

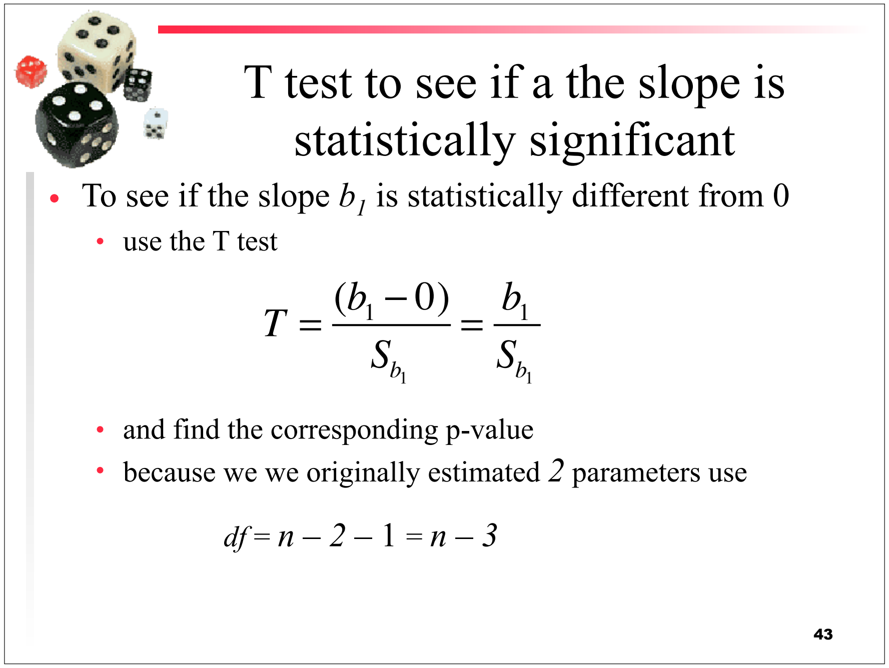
$$\beta_0=b_0\pm t_{\alpha,n-2}s_{b_1}$$



Confidence Bands

$$\hat{Y}=b_1x+b_0\pm k_{\alpha,n,X}\left[S_X^2+(x-\overline{X})^2\right]^{1/2}$$

$$k_{\alpha,n,X}=\left[2F_{\alpha/2,2,n-2}\frac{MSE}{nS_X^2}\right]^{1/2}$$



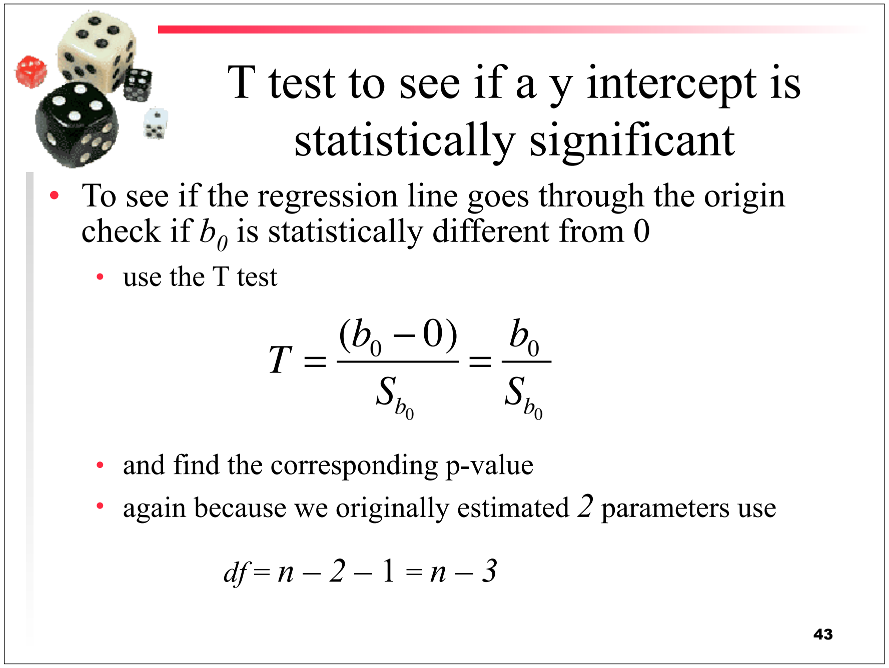
T test to see if a the slope is statistically significant

- 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 we originally estimated *2* parameters use

$df = n - 2 - 1 = n - 3$



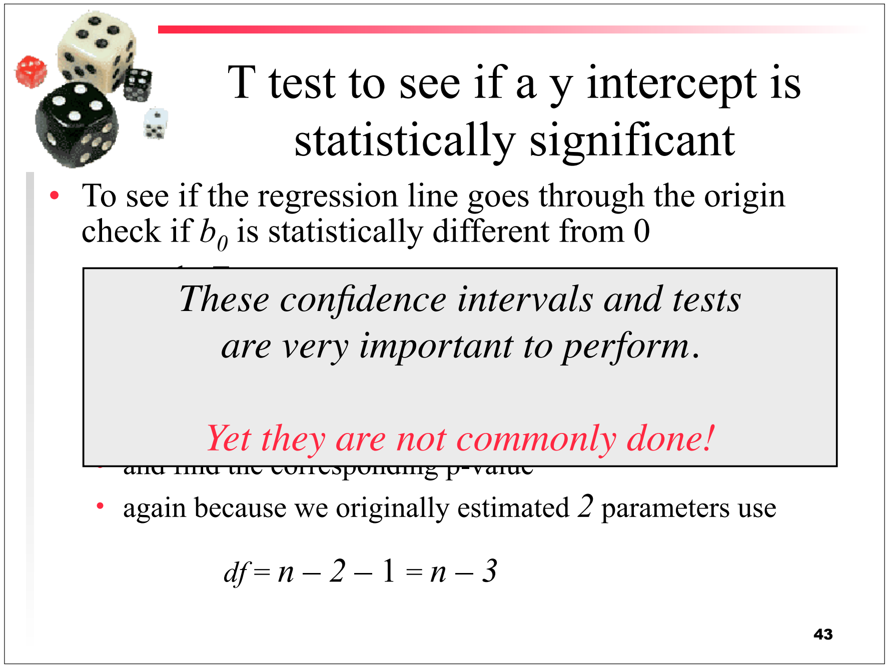
T test to see if a y intercept is statistically significant

- 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

$df = n - 2 - 1 = n - 3$



T test to see if a y intercept is statistically significant

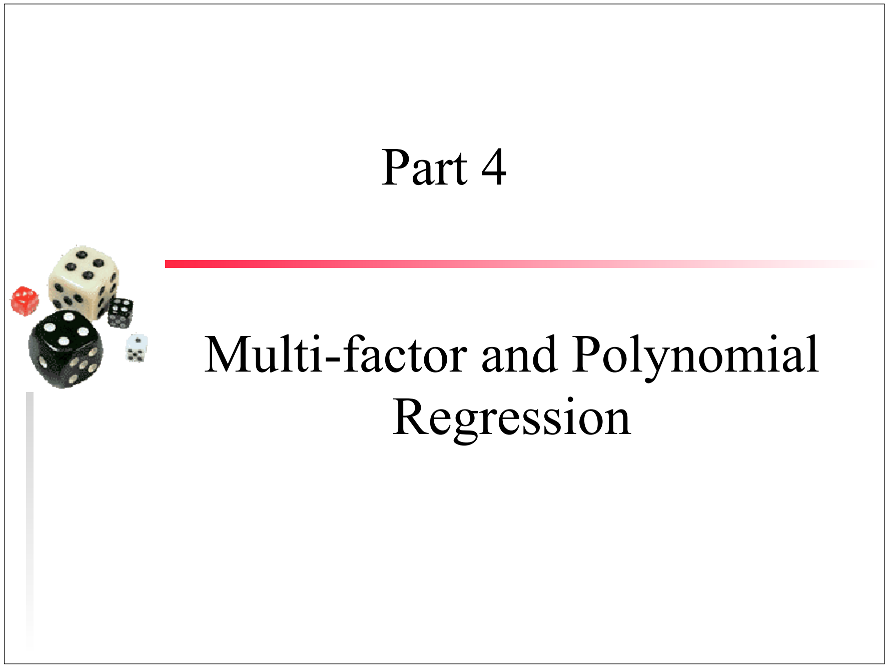
- To see if the regression line goes through the origin check if b_0 is statistically different from 0

These confidence intervals and tests are very important to perform.

Yet they are not commonly done!

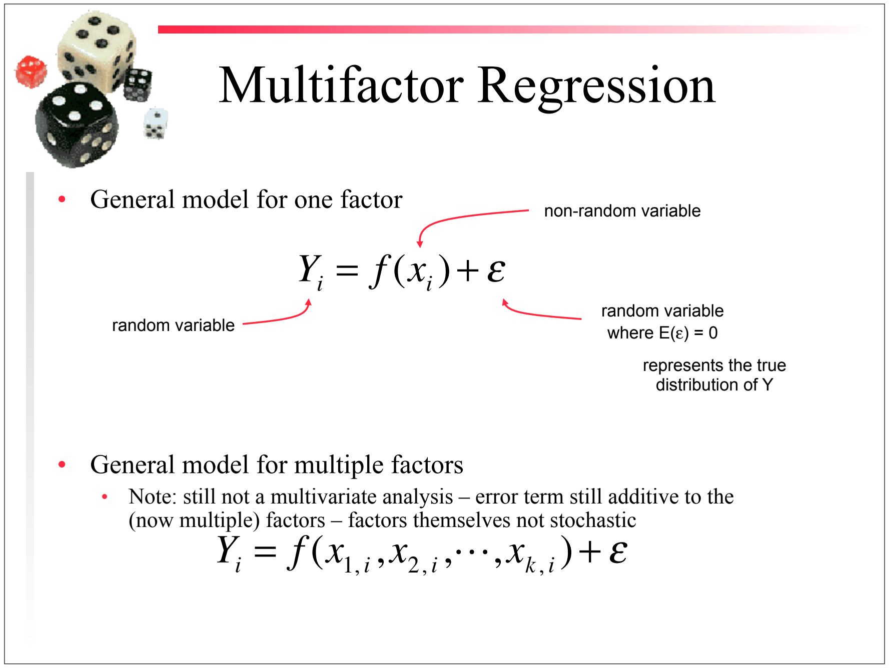
  - and find the corresponding p-value
  - again because we originally estimated *2* parameters use

$df = n - 2 - 1 = n - 3$



Part 4

Multi-factor and Polynomial Regression



Multifactor Regression

- General model for one factor

$$Y_i = f(x_i) + \varepsilon$$

non-random variable

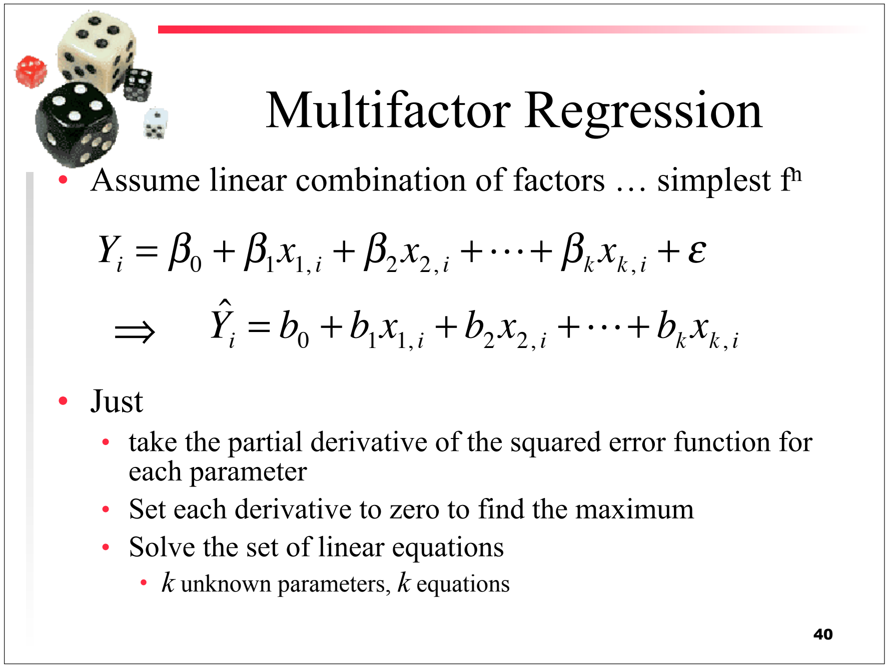
random variable

random variable where E(ε) = 0

represents the true distribution of Y

- General model for multiple factors
  - Note: still not a multivariate analysis – error term still additive to the (now multiple) factors – factors themselves not stochastic

$$Y_i = f(x_{1,i}, x_{2,i}, \cdots, x_{k,i}) + \varepsilon$$



Multifactor Regression

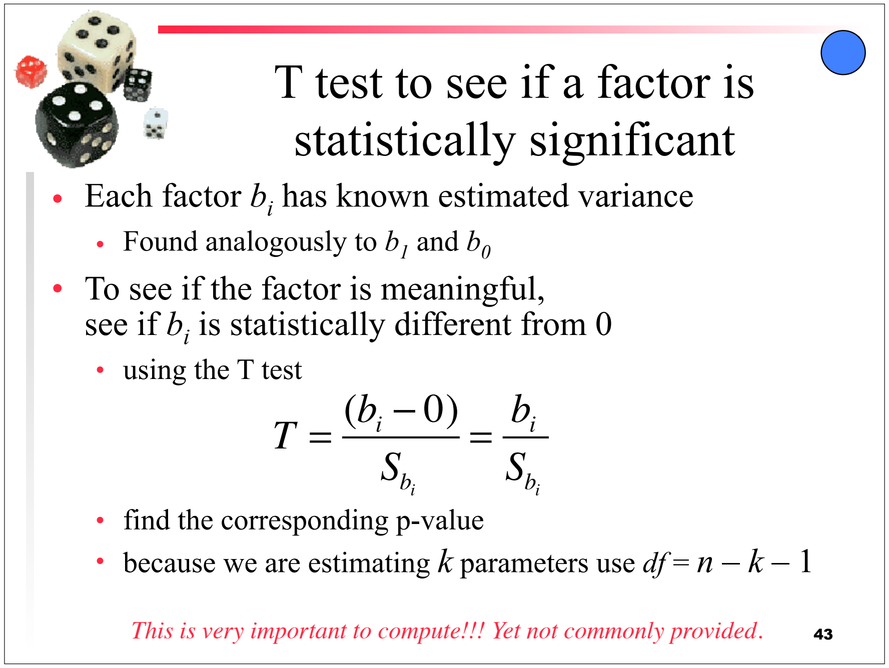
- Assume linear combination of factors … simplest f^{n}

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + \varepsilon$$

$$\Rightarrow \quad \hat{Y}_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + \cdots + b_k x_{k,i}$$

- 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

40



T test to see if a factor is statistically significant

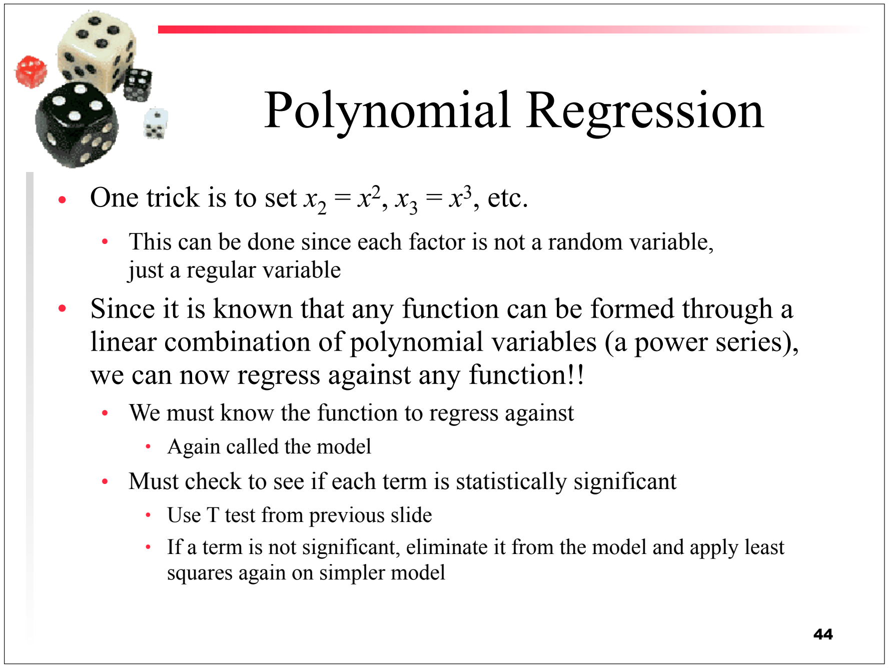
- Each factor b_i has known estimated variance
  - Found analogously to b_1 and b_0
- To see if the factor is meaningful, see if b_i is statistically different from 0
  - using the T test

$$T = \frac{(b_i - 0)}{S_{b_i}} = \frac{b_i}{S_{b_i}}$$

  - 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.

43

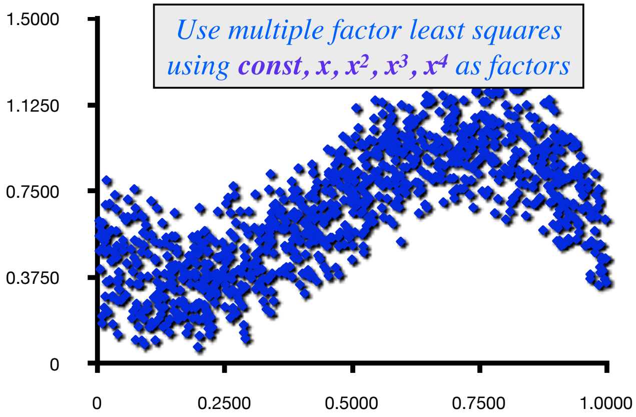


Polynomial Regression

- 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

44

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.

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

X^4 is not statistically significant
…reduce the number of terms by one

Source				*F-ratio*
Regression				*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) = 70.2%
s = 0.1465 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*

Polynomial Regression E.g.

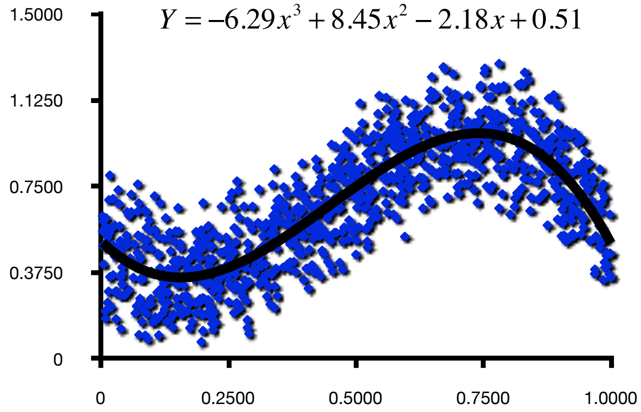
R squared = 70.2%   R squared (adjusted) = 70.2%
s = 0.1465 with 1000 - 4 = 996 degrees of freedom

Source … *atio*
Regress
Residu

All factors statistically significant
… regression function is a cubic polynomial

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: $Y = -6x^3 + 8x^2 - 2x + 0.5 + \varepsilon$

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/