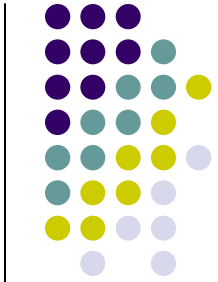


Comparing alternatives

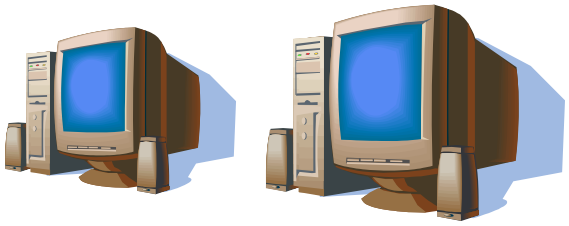
Prof Anja Feldmann
based on slides by David J. Lilja

Comparing alternatives

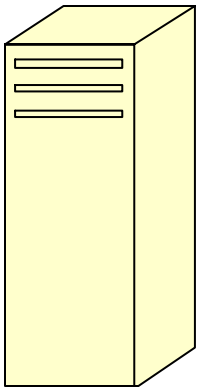
- ANOVA
  - Analysis of Variance
- Partitions total variation in a set of measurements into
  - Variation due to real differences in alternatives
  - Variation due to errors

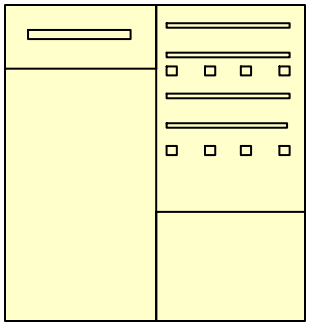
Comparing two alternatives

1. *Before-and-after*
   Did a change to the system have a statistically significant impact on performance?

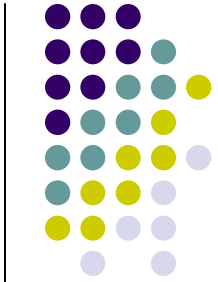


2. *Non-corresponding measurements*
   Is there a statistically significant difference between two different systems?





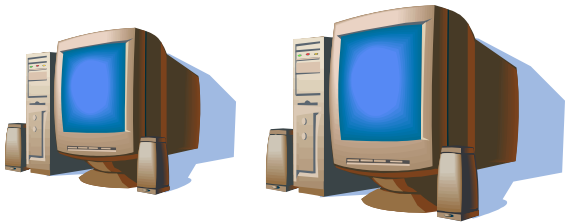
Before-and-after comparison



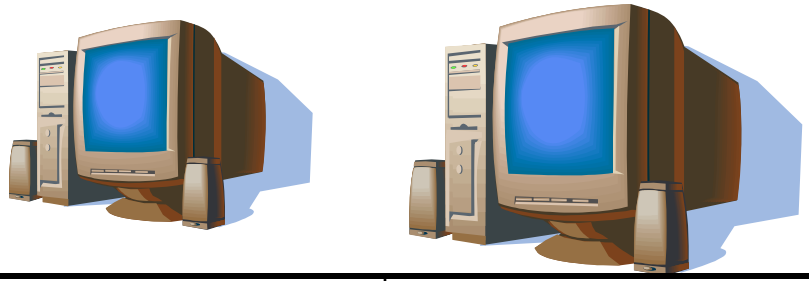
- Assumptions
  - Before-and-after measurements are not independent
  - Variances in two sets of measurements may not be equal

→ Measurements are related

- Use *mean of differences*

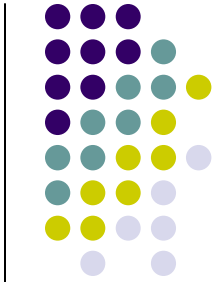


Before-and-after comparison



Measurement (i)	Before (b_i)	After (a_i)	Difference ($d_i = b_i - a_i$)
1	85	86	-1
2	83	88	-5
3	94	90	4
4	90	95	-5
5	88	91	-3
6	87	83	4

Before-and-after comparison



$$\text{Mean of differences} = \bar{d} = -1$$

$$\text{Standard deviation} = s_d = 4.15$$

- From mean of differences, appears that change reduced performance.
- However, standard deviation is large.

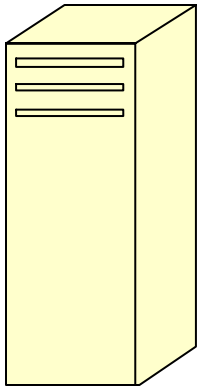
95% Confidence interval for mean of differences

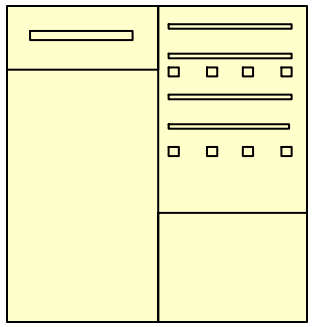
- $c_{1,2}$ = [-5.36, 3.36]
- Interval includes 0

→ With 95% confidence, there is *no statistically significant difference* between the two systems.

Noncorresponding measurements

- No direct correspondence between pairs of measurements
- *Unpaired* observations
- n_1 measurements of system 1
- n_2 measurements of system 2





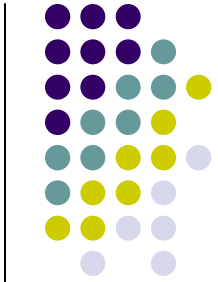
Confidence interval for difference of means

1. Compute means
2. Compute difference of means
3. Compute standard deviation of difference of means
4. Find confidence interval for this difference
5. No statistically significant difference between systems if interval includes 0

OS example

- Initial operating system
  - n1 = 1,300,203 interrupts (3.5 hours)
  - m1 = 142,892 interrupts occurred in OS code
  - p1 = 0.1099, or 11% of time executing in OS
- Upgrade OS
  - n2 = 999,382
  - m2 = 84,876
  - p2 = 0.0849, or 8.5% of time executing in OS
- Statistically significant improvement?

OS example (2.)



- p = p1 – p2 = 0.0250
- s_p = 0.0003911
- 90% confidence interval
  - (0.0242, 0.0257)
- Statistically significant difference?

Important points

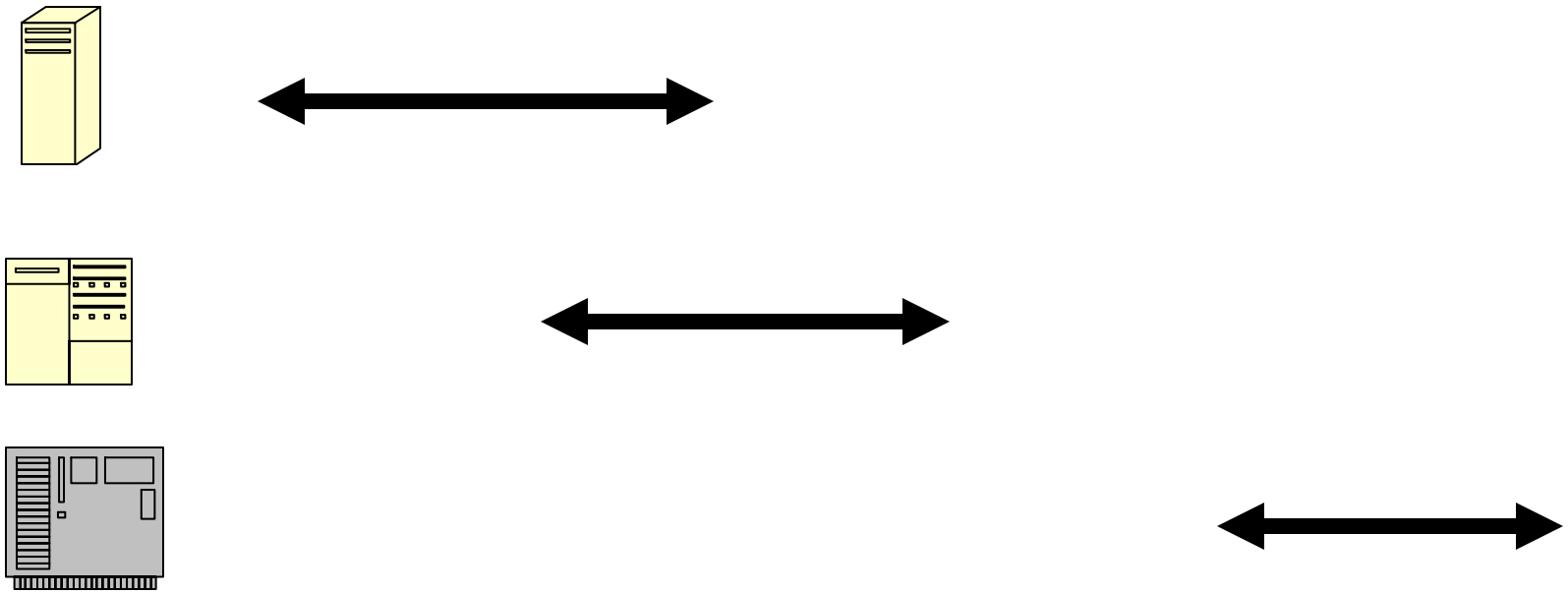
- Use confidence intervals to determine if there are statistically significant differences
  - Before-and-after comparisons
    - Find interval for *mean of differences*
  - Noncorresponding measurements
    - Find interval for *difference of means*

- If interval includes zero

  → No statistically significant difference

Comparing > two alternatives

- Naïve approach
  - Compare confidence intervals



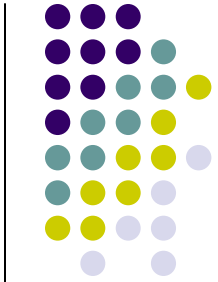
One-factor Analysis of Variance (ANOVA)

- Very general technique
  - Look at total *variation* in a set of measurements
  - Divide into meaningful components
- Also called
  - One-way classification
  - One-factor experimental design

One-factor Analysis of Variance (ANOVA)

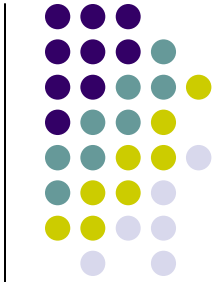
- Separates total variation observed in a set of measurements into:
  1. Variation within one system
     - Due to random measurement errors
  2. Variation between systems
     - Due to real differences + random error
- Is variation(2) statistically > variation(1)?

ANOVA



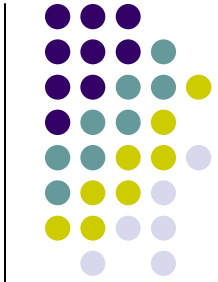
- Make *n* measurements of *k* alternatives
- y_{ij} = *i*th measurment on *j*th alternative
- Assumes errors are:
  - Independent
  - Gaussian (normal)

Measurements for all alternatives



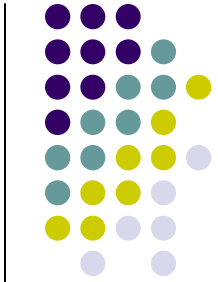
	Alternatives					
Measurements	1	2	…	*j*	…	*k*
1	y_{11}	y_{12}	…	y_{1j}	…	y_{k1}
2	y_{21}	y_{22}	…	y_{2j}	…	y_{2k}
…	…	…	…	…	…	…
i	y_{i1}	y_{i2}	…	y_{ij}	…	y_{ik}
…	…	…	…	…	…	…
n	y_{n1}	y_{n2}	…	y_{nj}	…	y_{nk}
Col mean	$y_{.1}$	$y_{.2}$	…	$y_{.j}$	…	$y_{.k}$
Effect	α_1	α_2	…	α_j	…	α_k

Column means:
Average performance of one alternative



	Alternatives					
Measurements	1	2	…	*j*	…	*k*
1	y_{11}	y_{12}	…	y_{1j}	…	y_{k1}
2	y_{21}	y_{22}	…	y_{2j}	…	y_{2k}
…	…	…	…	…	…	…
i	y_{i1}	y_{i2}	…	y_{ij}	…	y_{ik}
…	…	…	…	…	…	…
n	y_{n1}	y_{n2}	…	y_{nj}	…	y_{nk}
Col mean	$y_{.1}$	$y_{.2}$	…	$y_{.j}$	…	$y_{.k}$
Effect	α_1	α_2	…	α_j	…	α_k

Error:
Deviation from column mean

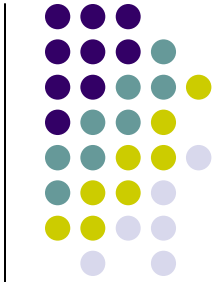


	Alternatives					
Measurements	1	2	…	*j*	…	*k*
1	y_{11}	y_{12}	…	y_{1j}	…	y_{k1}
2	y_{21}	y_{22}	…	y_{2j}	…	y_{2k}
…	…	…	…	…	…	…
i	y_{i1}	y_{i2}	…	y_{ij}	…	y_{ik}
…	…	…	…	…	…	…
n	y_{n1}	y_{n2}	…	y_{nj}	…	y_{nk}
Col mean	$y_{.1}$	$y_{.2}$	…	$y_{.j}$	…	$y_{.k}$
Effect	α_1	α_2	…	α_j	…	α_k

Overall mean:
Average performance of all alternatives

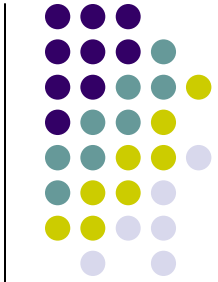
	Alternatives					
Measurements	1	2	…	*j*	…	*k*
1	y_{11}	y_{12}	…	y_{1j}	…	y_{k1}
2	y_{21}	y_{22}	…	y_{2j}	…	y_{2k}
…	…	…	…	…	…	…
i	y_{i1}	y_{i2}	…	y_{ij}	…	y_{ik}
…	…	…	…	…	…	…
n	y_{n1}	y_{n2}	…	y_{nj}	…	y_{nk}
Col mean	$y_{.1}$	$y_{.2}$	…	$y_{.j}$	…	$y_{.k}$
Effect	α_1	α_2	…	α_j	…	α_k

Effect:
Deviation from overall mean



	Alternatives					
Measurements	1	2	…	*j*	…	*k*
1	y_{11}	y_{12}	…	y_{1j}	…	y_{k1}
2	y_{21}	y_{22}	…	y_{2j}	…	y_{2k}
…	…	…	…	…	…	…
i	y_{i1}	y_{i2}	…	y_{ij}	…	y_{ik}
…	…	…	…	…	…	…
n	y_{n1}	y_{n2}	…	y_{nj}	…	y_{nk}
Col mean	$y_{.1}$	$y_{.2}$	…	$y_{.j}$	…	$y_{.k}$
Effect	α_1	α_2	…	α_j	…	α_k

Effects and errors



- *Effect* is distance from overall mean
  - Horizontally across alternatives
- *Error* is distance from column mean
  - Vertically within one alternative
  - Error across alternatives, too
- Individual measurements are then:

$$y_{ij} = \bar{y}_{..} + \alpha_j + e_{ij}$$

Sum of squares of differences

- ***SST*** = differences between each measurement and overall mean
- ***SSA*** = variation due to effects of alternatives
- ***SSE*** = variation due to errors in measurments

$$SST = SSA + SSE$$

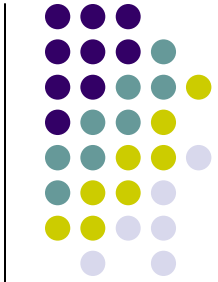
Sum of squares of differences

$$SSA = n\sum_{j=1}^{k}\left(\bar{y}_{.j} - \bar{y}_{..}\right)^2$$

$$SSE = \sum_{j=1}^{k}\sum_{i=1}^{n}\left(y_{ij} - \bar{y}_{.j}\right)^2$$

$$SST = \sum_{j=1}^{k}\sum_{i=1}^{n}\left(y_{ij} - \bar{y}_{..}\right)^2$$

ANOVA – Fundamental idea



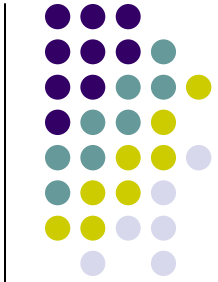
Separates variation in measured values into:

1. Variation due to effects of alternatives
   - SSA – variation across columns
2. Variation due to errors
   - SSE – variation within a single column

If differences among alternatives are due to real differences,

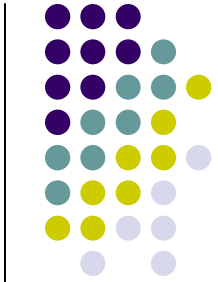
- SSA should be statistically > SSE

Comparing SSE and SSA



- Simple approach
  - *SSA* / *SST* = fraction of total variation explained by differences among alternatives
  - *SSE* / *SST* = fraction of total variation due to experimental error
- But is it statistically significant?

Comparing variances

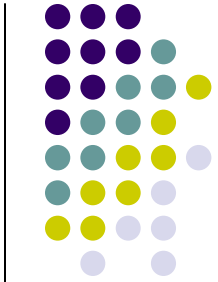


- Use F-test (statistics) to compare ratio of variances
- If $F_{computed} > F_{table}$
  $\rightarrow$ We have $(1 - \alpha) * 100\%$ confidence that variation due to actual differences in alternatives, SSA, is statistically greater than variation due to errors, SSE.

ANOVA example

	Alternatives			
Measurements	**1**	**2**	**3**	**Overall mean**
1	0.0972	0.1382	0.7966	
2	0.0971	0.1432	0.5300	
3	0.0969	0.1382	0.5152	
4	0.1954	0.1730	0.6675	
5	0.0974	0.1383	0.5298	
Column mean	0.1168	0.1462	0.6078	0.2903
Effects	-0.1735	-0.1441	0.3175	

Conclusions from example



- SSA/SST = 0.7585/0.8270 = 0.917

  → 91.7% of total variation in measurements is due to differences among alternatives

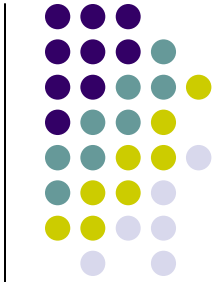
- SSE/SST = 0.0685/0.8270 = 0.083

  → 8.3% of total variation in measurements is due to noise in measurements

- Computed *F* statistic > tabulated *F* statistic

  → 95% confidence that differences among alternatives are statistically significant.

Contrasts



- ANOVA tells us that there is a statistically significant difference among alternatives
- But it does *not* tell us *where* difference is
- Use method of contrasts to compare subsets of alternatives
  - A vs B
  - {A, B} vs {C}
  - Etc.
- Contrast = linear combination of *effects* of alternatives

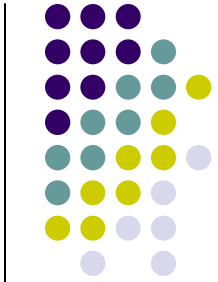
Important Points

- Use one-factor ANOVA to separate total variation into:
  - Variation within one system
    - Due to random errors
  - Variation between systems
    - Due to real differences (+ random error)
- Is the variation due to real differences *statistically* greater than the variation due to errors?

Generalized *design of experiments*

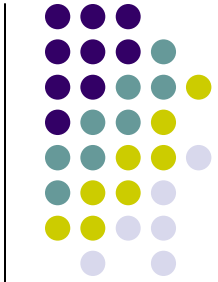
- Goals
  - Isolate effects of each input variable.
  - Determine effects of interactions.
  - Determine magnitude of experimental error
  - Obtain maximum information for given effort
- Basic idea
  - Expand 1-factor ANOVA to *m* factors

Terminology



- Response variable
  - Measured output value: e.g., total execution time
- Factors
  - Input variables that can be changed
    - E.g.: cache size, clock rate, bytes transmitted
- Levels
  - Specific values of factors:
    - Continuous (~bytes) or discrete (type of system)
- Replication
  - Completely re-run experiment with same input levels
- Interaction
  - *Effect* of input factor A depends on *level* of input factor B

Two-factor experiments



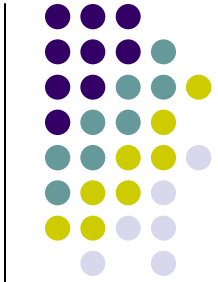
- Two factors (inputs)
  - A, B
- Separate total variation in output values into:
  - Effect due to A
  - Effect due to B
  - Effect due to interaction of A and B (AB)
  - Experimental error

Example – User response time

- A = degree of multiprogramming
- B = memory size
- AB = interaction of memory size and degree of multiprogramming

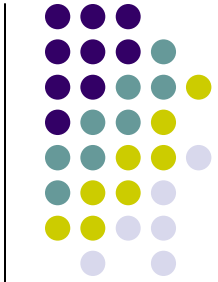
A	B (Mbytes)		
	32	64	128
1	0.25	0.21	0.15
2	0.52	0.45	0.36
3	0.81	0.66	0.50
4	1.50	1.45	0.70

Two-factor ANOVA



- Factor A – *a* input levels
- Factor B – *b* input levels
- *n* measurements for each input combination
- *abn* total measurements

Two-factor ANOVA



- Each individual measurement is composition of
  - Overall mean
  - Effects
  - Interactions
  - Measurement errors

Example

- Output = user response time (seconds)
- Want to separate effects due to
  - A = degree of multiprogramming
  - B = memory size
  - AB = interaction
  - Error
- Need replications to separate error

A	B (Mbytes) 32	64	128
1	0.25	0.21	0.15
2	0.52	0.45	0.36
3	0.81	0.66	0.50
4	1.50	1.45	0.70

Conclusions from example

- 77.6% (SSA/SST) of all variation in response time due to degree of multiprogramming
- 11.8% (SSB/SST) due to memory size
- 9.9% (SSAB/SST) due to interaction
- 0.7% due to measurement error
- 95% confident that all effects and interactions are statistically significant

A problem

- *Full factorial design with replication*
  - Measure system response with all possible input combinations
  - Replicate each measurement *n* times to determine effect of measurement error
- *m* factors, *v* levels, *n* replications

  $\rightarrow n\, v^m$ experiments
- *m* = 5 input factors, *v* = 4 levels, *n* = 3
  - $\rightarrow 3(4^5) =$ 3,072 experiments!

Fractional factorial designs: $n2^m$ experiments

- Special case of generalized *m*-factor experiments
- Restrict each factor to two possible values
  - High, low
  - On, off
- Find factors that have largest impact
- Full factorial design with only those factors

Still too many experiments with $n2^m$!

- Plackett and Burman designs (1946)
  - Multifactorial designs
- Effects of main factors only
  - Logically minimal number of experiments to estimate effects of *m* input parameters (factors)
  - Ignores interactions
- Requires O(*m*) experiments
  - Instead of O(2^m) or O(v^m)