

ANOVA

Comparing Alternatives

Prof. Anja Feldmann

&

Balakrishnan Chandrasekaran

(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

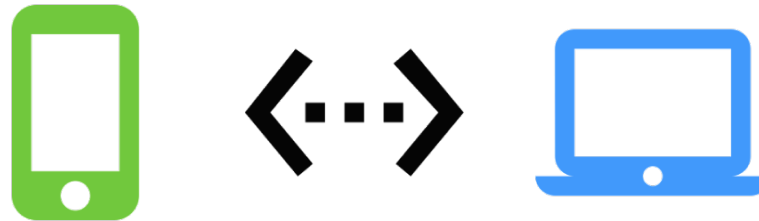
- *Before-and-after*



Did a change to the system have a statistically significant impact on performance?

Comparing *two* alternatives

- *Non-corresponding measurements*



Is there a statistically significant difference between two different systems?

Before-and-after comparison

- Assumptions

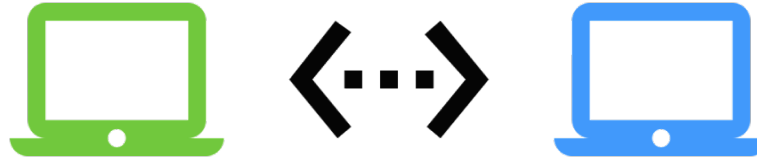
- Before-and-after measurements are *not independent*
- Variances in two sets of measurements *may not be equal*

→ *Measurements are related*

- Use *mean of differences*

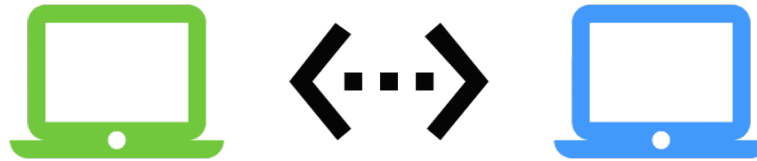


Before-and-after comparison



<i>Measurement</i> <i>i</i>	<i>Before</i> <i>b_i</i>	<i>After</i> <i>a_i</i>
1	85	86
2	83	88
3	94	90
4	90	95
5	88	91
6	87	83

Before-and-after comparison



<i>Measurement</i> <i>i</i>	<i>Before</i> b_i	<i>After</i> a_i	<i>Difference</i> $d_i = b_i - a_i$
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

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

Standard deviation: $S_d = 4.15$

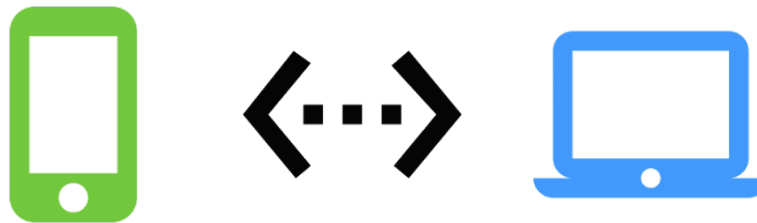
- From mean of differences, it appears that change reduced performance
- Standard deviation, however, is large!

Before-and-after comparison

- **95% confidence interval** for mean of differences
 - $C_{1,2} = [-5.36, 3.36]$
 - Interval includes **zero!**
- With 95% confidence, there is ***no statistically significant difference between the two systems***

Non-corresponding measurements

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



Non-corresponding measurements

- Use 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. If interval includes zero
→ *no statistically significant difference between systems*

Toy Example

- *Initial operating system (OS)*

- $n_1 = 1,300,203$ interrupts (3.5 hours)
- $m_1 = 142,892$ interrupts occurred in OS code
- $p_1 = 0.1099$, or 11% of time executing in OS

- *Upgrade OS*

- $n_2 = 999,382$
- $m_2 = 84,876$
- $p_2 = 0.0849$, or 8.5% of time executing in OS

Statistically significant improvement?

Toy Example contd.

- $p = p_1 - p_2 = 0.0250$
- $s_p = 0.0003911$
- 90% confidence interval
 - $[0.0242, 0.0257]$

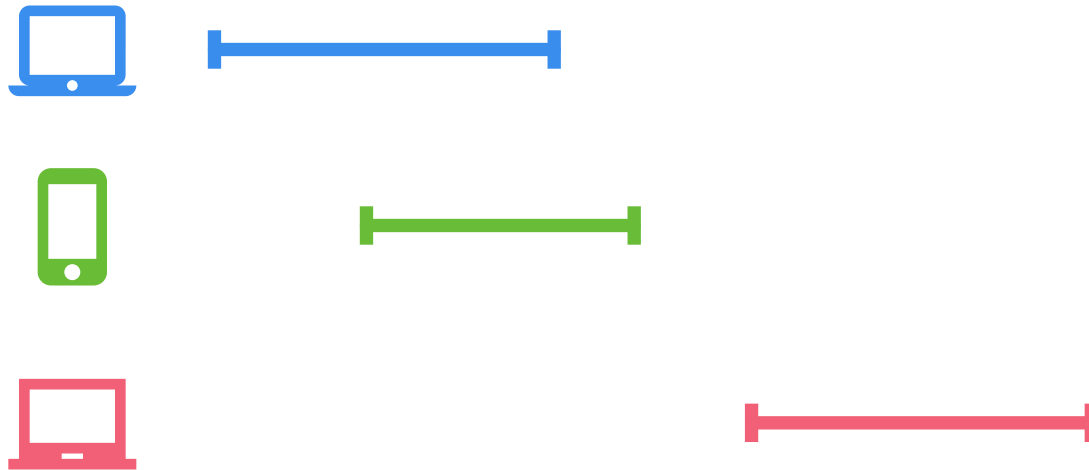
Statistically significant difference?

Key takeaways

- Use confidence intervals to determine if there are statistically significant differences
 - Before-and-after comparisons
 - Find interval for *mean of differences*
 - Non-corresponding measurements
 - Find interval for *difference of means*
- If interval includes zero
 - *No statistically significant difference*

More than two alternatives ...

- Naïve approach: compare confidence intervals



One-factor 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 ANOVA

- Separates *total variation* observed in a set of measurements into:
 1. Variation *within one system*
 - *Due to random measurement error*
 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} measurement on j^{th} alternative

- Assumes *errors* are:
 - *Independent*
 - *Gaussian (normal)*

Measurements for all alternatives

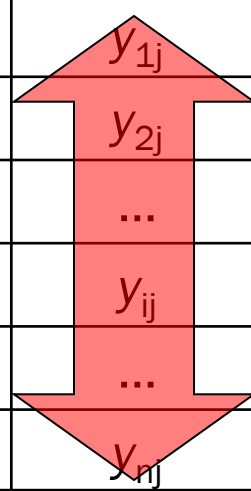
	Alternatives					
Measurements	1	2	...	j	...	k
1	y_{11}	y_{12}	...	y_{1j}	...	y_{k1}
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2k}
...
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{ik}
...
n	y_{n1}	y_{n2}	...	y_{nj}	...	y_{nk}
Col. Mean	$y_{.1}$	$y_{.2}$...	$y_{.j}$...	$y_{.k}$
Effect	α_1	α_2	...	α_j	...	α_k

Col. means: Avg. 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 col. 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: Avg. 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 measurements

$$SST = SSA + SSE$$

Sum of squares of differences

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

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

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

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 $F_{computed} > F_{table}$
 - 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
- $F_{computed} > F_{table}$
→ **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*

Key takeaways

- 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

	B (Mbytes)		
A	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

	B (Mbytes)		
A	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
 - $n v^m$ experiments
- $m = 5$ input factors, $v = 4$ levels, $n = 3$
 - → $3(4^5) = 3,072$ 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 $O(v^m)$