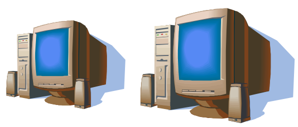
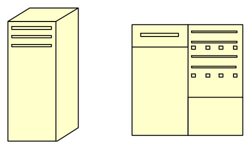
Comparing Alternatives

Samuel Kounev





1

References

- „Measuring Computer Performance – A Practitioner's Guide“ by David J. Lilja, Cambridge University Press, New York, NY, 2000, ISBN 0-521-64105-5
- The supplemental teaching materials provided at http://www.arctic.umn.edu/perf-book/ by David J. Lilja

2

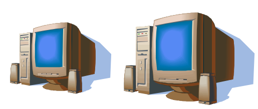
Roadmap

- Comparing two alternatives
  - Before-and-after comparisons
  - Non-corresponding measurements
- Comparing proportions
- Comparing more than two alternatives
  - One-Factor analysis of variance (ANOVA)

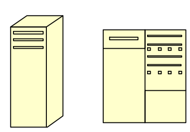
3

Comparing Two Alternatives

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



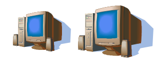
- Is there a statistically significant difference between two different systems?



1. ***Before-and-after comparisons***
   (paired observations)
2. ***Non-corresponding measurements***
   (unpaired observations)

4

Before-and-After Comparisons



- Before-and-after measurements are not independent
- They form corresponding pairs
- Measurements within the 2 sets (before and after) need not be identically distributed

Procedure

- Compute mean and standard deviation of the differences

  $d_i = b_i - a_i \quad b_i = \text{before measurement} \quad a_i = \text{after measurement}$

  $\bar{d} = \text{mean value of } d_i \quad s_d = \text{std. deviation of } d_i$

- Find confidence interval for the mean of differences

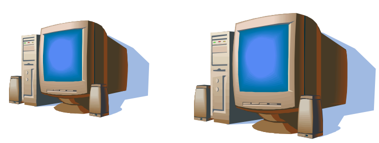
  $\text{If } n \geq 30, \text{a CI for d is given by } (c_1, c_2) = \bar{d} \mp z_{1-\alpha/2} \frac{s_d}{\sqrt{n}}$

  $\text{If } n < 30, \text{a CI for d is given by } (c_1, c_2) = \bar{d} \mp t_{1-\alpha/2;n-1} \frac{s_d}{\sqrt{n}}$

- No statistically significant difference between systems if interval includes 0

5

Ex: Before-and-After Comparisons



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

6

Ex: Before-and-After Comparisons (2)

$$\text{Mean of differences} = \bar{d} = -1 \qquad \text{Std. deviation} = s_d = 4.15$$

- From mean of differences, appears that change reduced performance. However, standard deviation is large!
- 95% Confidence Interval for the mean of differences:

$$t_{1-\alpha/2;n-1} = t_{0.975;5} = 2.571$$

$$c_{1,2} = \bar{d} \mp t_{1-\alpha/2;n-1} \frac{s_d}{\sqrt{n}} = -1 \mp 2.571 \left(\frac{4.15}{\sqrt{6}} \right) = [-5.36, 3.36]$$

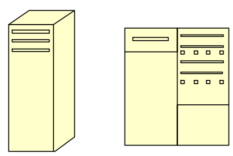
- Interval includes 0

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

7

Non-Corresponding Measurements

- No direct correspondence between pairs of measurements → *Unpaired* observations
- n_1 measurements of system 1
- n_2 measurements of system 2
- Measurements within each set are IID random variables



Procedure

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

This procedure is known as *t*-test

8

Conf. Interval for Difference of Means

$\bar{x} = \bar{x}_1 - \bar{x}_2$ Note: If Y_i are indep. R.V. then $\sigma^2\left[\sum_{i=1}^{n} a_i Y_i\right] = \sum_{i=1}^{n} a_i^2 \sigma^2[Y_i]$

$$\sigma^2(\bar{x}) = \sigma^2(\bar{x}_1 - \bar{x}_2) = 1^2\sigma^2\left(\frac{\sum_{i=1}^{n_1} x_{1,i}}{n_1}\right) + (-1)^2\sigma^2\left(\frac{\sum_{i=1}^{n_2} x_{2,i}}{n_2}\right) =$$

$$= \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sigma^2(x_{1,i}) + \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sigma^2(x_{2,i}) = \frac{n_1\sigma^2(x_1)}{n_1^2} + \frac{n_2\sigma^2(x_2)}{n_2^2} =$$

$$= \frac{\sigma^2(x_1)}{n_1} + \frac{\sigma^2(x_2)}{n_2}$$

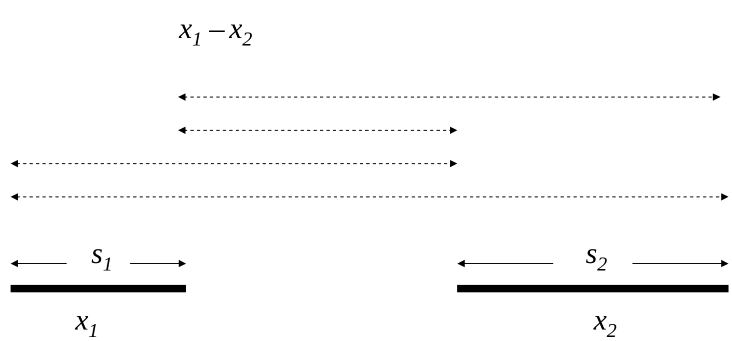
Therefore, we can estimate the standard deviation of $\bar{x}$ using $s_{\bar{x}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Assuming that n_1 and n_2 are both >= 30.

9

Why Add Standard Deviations?

$$s_{\bar{x}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$



10

Conf. Interval for Difference of Means

$\overline{x}_1$ and $\overline{x}_2$ are approx. normally distributed (CLT)

$\Rightarrow \overline{x}$ is also normally distributed, i.e. $\overline{x} \in N(\mu, \sigma_{\overline{x}})$

$\left(\frac{\overline{x} - \mu}{\sigma_{\overline{x}}}\right) \in N(0,1)$ (standard normal distribution)

If $n_1 \geq 30$ and $n_2 \geq 30$, we can approximate $\sigma_{\overline{x}}$ with $s_{\overline{x}}$

$$c_1 = \overline{x} - z_{1-\alpha/2} s_{\overline{x}} \quad c_2 = \overline{x} + z_{1-\alpha/2} s_{\overline{x}}$$

Else we can use the Student-t distribution

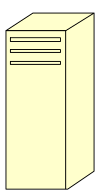
$$c_1 = \overline{x} - t_{1-\alpha/2; n_{df}} s_{\overline{x}} \quad c_2 = \overline{x} + t_{1-\alpha/2; n_{df}} s_{\overline{x}}$$

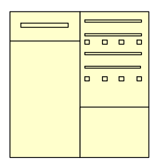
However, $n_{df} \neq n_1 + n_2 - 2$ instead

$$n_{df} \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$$

11

Example





$n_1 = 12$ measurements

$\overline{x}_1 = 1243$ s

$s_1 = 38.5$

$n_2 = 7$ measurements

$\overline{x}_2 = 1085$ s

$s_2 = 54.0$

12

Example (cont.)

$$\bar{x} = \bar{x}_1 - \bar{x}_2 = 1243 - 1085 = 158$$

$$s_{\bar{x}} = \sqrt{\frac{38.5^2}{12} + \frac{54^2}{7}} = 23.24$$

$$n_{df} = \frac{\left(\frac{38.5^2}{12} + \frac{54^2}{7}\right)^2}{\frac{(38.5^2/12)^2}{12-1} + \frac{(54^2/7)^2}{7-1}} = 9.62 \rightarrow 10$$

$$c_{1,2} = \bar{x} \mp t_{1-\alpha/2;n_{df}} s_{\bar{x}} \qquad t_{1-\alpha/2;n_{df}} = t_{0.95;10} = 1.813$$

$$c_{1,2} = 158 \mp 1.813(23.24) = [116,200]$$

13

Special Case

- If only a few measurements available (i.e. $n_1 < 30$ or $n_2 < 30$), but it is known that
  - Errors are normally distributed **and** ($\sigma_1 = \sigma_2$ **or** $n_1 = n_2$)
- Then special case applies…

$$(c_1, c_2) = \bar{x} \mp t_{1-\alpha/2;n_{df}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$n_{df} = n_1 + n_2 - 2$$

$$s_p = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

- Typically produces tighter confidence interval
- Sometimes useful after obtaining additional measurements to tease out small differences

14

Comparing Proportions

$m_1 = \#\text{events of interest in system } 1$

$n_1 = \text{total } \#\text{events in system } 1$

$m_2 = \#\text{events of interest in system } 2$

$n_2 = \text{total } \#\text{events in system } 2$

$\overline{p}_i = \frac{m_i}{n_i}$ is the proportion of the events of interest measured in system i

- The number of events of interest m_i follows a binomial distribution with parameters p_i and n_i

$$\mathrm{E}(m_i) = p_i n_i \qquad \sigma^2[m_i] = p_i(1-p_i)n_i$$

- If $m_i >= 10$ we can approximate the binomial distributions using normal distributions:

$$m_i \approx N(p_i n_i, p_i(1-p_i)n_i) \Rightarrow m_i \approx N(p_i n_i, \overline{p}_i(1-\overline{p}_i)n_i)$$

15

Comparing Proportions (2)

Let $\overline{p} = \overline{p}_1 - \overline{p}_2 = \frac{m_1}{n_1} - \frac{m_2}{n_2}$ Need a CI for the mean of $\overline{\mathrm{p}}$

$$E(\overline{p}) = E\left(\frac{m_1}{n_1} - \frac{m_2}{n_2}\right) = \frac{E(m_1)}{n_1} - \frac{E(m_2)}{n_2} = p_1\text{-}p_2$$

$$m_i \approx N(p_i n_i, \overline{p}_i(1-\overline{p}_i)n_i) \Rightarrow \frac{m_i}{n_i} \approx N\left(p_i, \frac{\overline{p}_i(1-\overline{p}_i)}{n_i}\right)$$

$$\overline{p} = \frac{m_1}{n_1} - \frac{m_2}{n_2} \approx N\left(p_1 - p_2, \frac{\overline{p}_1(1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_2}\right)$$

16

Comparing Proportions (3)

$$\Pr\left(-z_{1-\alpha/2} \le \frac{\bar{p}-(p_1-p_2)}{\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1}+\frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}} \le z_{1-\alpha/2}\right) = 1-\alpha$$

$$\Pr\left(\bar{p}-z_{1-\alpha/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1}+\frac{\bar{p}_2(1-\bar{p}_2)}{n_2}} \le p_1-p_2 \le \bar{p}+z_{1-\alpha/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1}+\frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}\right) = 1-\alpha$$

$$\Rightarrow (c_1, c_2) = \bar{p} \mp z_{1-\alpha/2} s_p \quad \text{where} \quad s_p = \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1}+\frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$$

17

Example

- Initial operating system (OS)
  - 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
- Upgraded OS
  - n2 = 999,382
  - m2 = 84,876
  - p2 = 0.0849, or 8.5% of time executing in OS
- Statistically significant improvement?
  - $\bar{p} = \bar{p}_1 - \bar{p}_2 = 0.0250$
  - $s_p = 0.0003911$
  - 90% confidence interval
    - (0.0242, 0.0257)
  - Statistically significant difference?

18

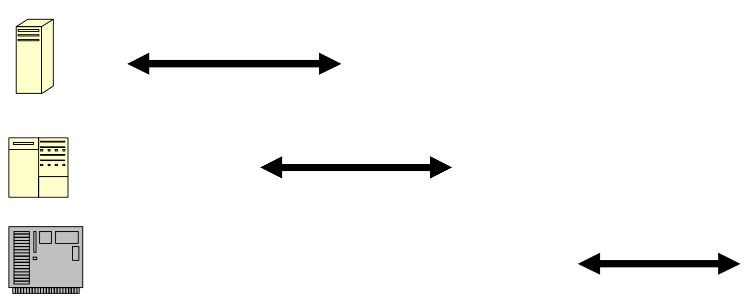
Roadmap

- Comparing two alternatives
  - Before-and-after comparisons
  - Non-corresponding measurements
- Comparing proportions
- Comparing more than two alternatives
  - One-Factor analysis of variance (ANOVA)

19

Comparing More Than Two Alternatives

- Want to find out if there is statistically significant difference between the alternatives
- Naïve approach
  - Compare systems two-by-two using previous techniques



20

One-Factor Analysis of Variance

- Very general technique, also called
  - One-factor ANOVA
  - One-factor experimental design
  - One-way classification
- Separates total variation observed in a set of measurements into:
  1. Variation within individual systems
     - Due to random measurement errors
  2. Variation between systems
     - Due to real differences + random errors
- Aim is to determine if (2) is statistically greater than (1)?

21

One-Factor Analysis of Variance (2)

- Make n measurements of k alternatives
- y_{ij} = i^{th} measurement on j^{th} alternative
- Assumes measurements (errors) for the different alternatives are:
  - Independent
  - Gaussian (normally) distributed

22

Measurements for All Alternatives

	Alternatives					
Measurements	1	2	…	*j*	…	*k*
1	y_{11}	y_{12}	…	y_{1j}	…	y_{1k}
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

23

Column Means

	Alternatives					
Measurements	1	2	…	*j*	…	*k*
1	y_{11}	y_{12}	…	y_{1j}	…	y_{1k}
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

24

Deviation from Column Means

- Column means are average values of all measurements within a single alternative
  - Average performance of one alternative

$$\bar{y}_{.j} = \frac{\sum_{i=1}^{n} y_{ij}}{n}$$

- Each measurements can be represented as

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

$$e_{ij} = \text{deviation of } y_{ij} \text{ from column mean}$$

$$= \text{error in measurements}$$

25

Overall Mean

	Alternatives					
Measurements	1	2	…	*j*	…	*k*
1	y_{11}	y_{12}	…	y_{1j}	…	y_{1k}
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

26

Deviation From Overall Mean

- Average of all measurements made of all alternatives

$$\bar{y}_{..} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n} y_{ij}}{kn}$$

- Column means can be represented as

$$\bar{y}_{.j} = \bar{y}_{..} + \alpha_j$$

$$\alpha_j = \text{deviation of column mean from overall mean}$$

$$= \text{effect of alternative } j$$

$$\text{It can be shown that } \sum_{j=1}^{k} \alpha_j = 0$$

27

Effect = Deviation From Overall Mean

	Alternatives					
Measurements	1	2	…	*j*	…	*k*
1	y_{11}	y_{12}	…	y_{1j}	…	y_{1k}
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

28

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

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

29

Sum of Squares of Differences: SSE

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

$$e_{ij} = y_{ij} - \bar{y}_{.j}$$

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

SSE characterizes the *variation due to errors*

30

Sum of Squares of Differences: SSA

$$\bar{y}_{.j} = \bar{y}_{..} + \alpha_j$$

$$\alpha_j = \bar{y}_{.j} - \bar{y}_{..}$$

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

SSA characterizes the *variation due to the effects*

31

Sum of Squares of Differences: SST

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

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

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

SST characterizes the *total variation*

32

Sum of Squares of Differences

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

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

- Sum of squares of differences
  - ***SSA*** = variation due to effects of alternatives (+ errors)
  - ***SSE*** = variation due to errors in measurements
  - ***SST*** = total variation
- Expanding SST and using $\sum_{j=1}^{k}\alpha_j = 0$ we can show

$$SST = SSA + SSE$$

33

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

34

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?

35

Statistically Comparing SSE and SSA

$$\text{Variance} = \text{mean square value} = \frac{\text{total variation}}{\text{degrees of freedom}}$$

$$s_x^2 = \frac{SSx}{df}$$

df(*SSA*) = *k*–1, since *k* alternatives

df(*SSE*) = *k(n*–1), since *k* alternatives, each with (*n*–1) *df*

df(*SST*) = *kn*–1 = *df*(*SSA*) + *df*(*SSE*)

$$\Rightarrow s_a^2 = \frac{SSA}{k-1} \qquad s_e^2 = \frac{SSE}{k(n-1)} \qquad s_t^2 = \frac{SST}{kn-1}$$

36

Degrees of Freedom for SSA

	Alternatives					
Measure ments	1	2	…	*j*	…	*k*
1	y_{11}	y_{12}	…	y_{1j}	…	y_{1k}
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

Degrees of Freedom for SSE

	Alternatives					
Measure ments	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

38

Degrees of Freedom for SST

	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

39

Variances from Sum of Squares (Mean Square Values)

- Use F-test to compare ratio of variances

$$F = \frac{s_a^2}{s_e^2}$$

$$F_{[1-\alpha;df(num),df(denom)]} = \text{tabulated critical values}$$

- If $F_{computed} > F_{table}$
  we have $(1 - \alpha)$ * 100% confidence that the variation due to actual differences in alternatives, SSA, is statistically greater than the variation due to errors, SSE.

40

One-Factor ANOVA Summary

Variation	Alternatives	Error	Total
Sum of squares	SSA	SSE	SST
Deg freedom	$k-1$	$k(n-1)$	$kn-1$
Mean square	$s_a^2 = SSA/(k-1)$	$s_e^2 = SSE/[k(n-1)]$	
Computed F	s_a^2/s_e^2		
Tabulated F	$F_{[1-\alpha;(k-1),k(n-1)]}$		

41

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	

42

ANOVA Example (cont.)

Variation	Alternatives	Error	Total
Sum of squares	$SSA = 0.7585$	$SSE = 0.0685$	$SST = 0.8270$
Deg freedom	$k-1=2$	$k(n-1)=12$	$kn-1=14$
Mean square	$s_a^2 = 0.3793$	$s_e^2 = 0.0057$	
Computed F	$0.3793/0.0057 = 66.4$		
Tabulated F	$F_{[0.95;2,12]} = 3.89$		

43

ANOVA Example (cont.)

- 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.

44

The Method of Contrasts

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

45

Contrasts

- Contrast = linear combination of *effects* of alternatives

$$c = \sum_{j=1}^{k} w_j \alpha_j \qquad \sum_{j=1}^{k} w_j = 0$$

- Contrasts are used to compare effects of a subset of the alternatives
- E.g. Compare effect of system 1 to effect of system 2

$$w_1 = 1 \qquad w_2 = -1 \qquad w_3 = 0$$
$$c = (1)\alpha_1 + (-1)\alpha_2 + (0)\alpha_3 = \alpha_1 - \alpha_2$$

46

Confidence Interval For Contrasts

- Need
  - Estimate of variance of contrast
  - Appropriate value from Student *t* table
- Compute confidence interval as before
- If interval includes 0
  - Then no statistically significant difference exists between the alternatives included in the contrast

47

Confidence Interval For Contrasts (2)

- Assuming variation due to errors is equally distributed among *kn* total measurements

$$\begin{aligned}\text{Var}[c] &= \text{Var}[\textstyle\sum_{j=1}^{k}(w_j\alpha_j)] \\ &= \textstyle\sum_{j=1}^{k}\text{Var}[w_j\alpha_j] \\ &= \textstyle\sum_{j=1}^{k} w_j^2\,\text{Var}[\alpha_j]\end{aligned}$$

$$s_c^2 = \frac{\sum_{j=1}^{k}(w_j^2 s_e^2)}{kn}$$

$$s_e^2 = \frac{SSE}{k(n-1)}$$

$$df(s_c^2) = k(n-1)$$

$$(c_1, c_2) = c \mp t_{1-\alpha/2;k(n-1)} s_c$$

$$s_c = \sqrt{\frac{\sum_{j=1}^{k}(w_j^2 s_e^2)}{kn}}$$

48

Example

- 90% confidence interval for contrast of [Sys1-Sys2]

$$\alpha_1 = -0.1735$$
$$\alpha_2 = -0.1441$$
$$\alpha_3 = 0.3175$$
$$c_{[1-2]} = -0.1735 - (-0.1441) = -0.0294$$
$$s_c = s_e \sqrt{\frac{1^2 + (-1)^2 + 0^2}{3(5)}} = 0.0275$$
$$90\% : (c_1, c_2) = (-0.0784, 0.0196)$$

- With 90% confidence, the difference between system 1 and system 2 is statistically insignificant

Summary

- **Comparing two alternatives**
  - 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
- **Comparing proportions**

Summary (cont.)

- **Comparing more than two alternatives**
  - Use one-factor ANOVA to separate total variation into:
    - Variation within individual systems
      - 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?
  - Use contrasts to compare effects of subsets of alternatives

51

Further Reading

- *"The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling"* by R. K. Jain, Wiley (April 1991), ISBN: 0471503363, 1991
- *"Performance Evaluation and Benchmarking", e*dited by Lizy Kurian John, Lieven Eeckhout, CRC Press Inc., ISBN: 0849336228, 2005
- *"Probability and Statistics for Engineers and Scientists (7th Edition)"* by Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying Ye, Keying Yee, Prentice Hall, 7 edition (January 2002, ISBN-10: 0130415294, ISBN-13: 978-0130415295

52

Exercises

- Using the “before-and-after” comparison technique with both a 90% and a 99% confidence level, determine whether turning a specific compiler optimization on makes a statistically significant difference. Repeat your analysis using ANOVA test with k=2 alternatives. Explain your results.

- Use the ANOVA test to compare the performances of three different, but roughly comparable, computer systems measured in terms of execution time of an appropriate benchmark program. The ANOVA test shows only whether there is a statistically significant difference among the systems, not how large the difference really is. Use appropriate contrasts to compare the differences between all possible pairs of the systems. Explain and interpret your results.