# Hypothesis Testing

CS 700

1

## Previously

- Comparing systems
- Using confidence intervals
- Paired, unpaired observations
- Analysis of variance ANOVA

- Next hypothesis testing

2

## Hypothesis Testing

- Now need to make decisions
- Purpose: make inferences about a population parameter by analyzing differences between observed sample statistics and the results one expects to obtain if some underlying assumption is true.
- Null hypothesis: $H_0 : \mu = x$
- Alternative hypothesis: $H_1 : \mu \neq x$
- If the null hypothesis is rejected then the alternative hypothesis is accepted
- Paint drying example (black-board)

3

| | Actual Situation | |
|---|---|---|
| | $H_o$ true | $H_o$ false |
| Accept $H_o$ | Correct decision<br>Confidence=1-α | Type II Error:<br>Pr[Type II]=β |
| Reject $H_o$ | Type I Error<br>P[Type I]=α | Correct Decision<br>Power=1-β |

4

## Risks in Decision Making

- Type I Error occurs if $H_o$ is rejected when it is true.
  - Pr [$H_o$ is rejected | true] = $\alpha$
- Type II Error occurs if $H_o$ is not rejected when it is false.
  - Pr[$H_o$ is not rejected | false] = $\beta$
- Confidence coefficient:
  - Pr [$H_o$ not rejected | true]= 1- $\alpha$
- Power of the test:
  - Pr[$H_o$ is rejected |false]= 1-$\beta$

5

## One-sided and two-sided alternatives

- Traditionally, the null hypothesis is used for a hypothesis set up primarily to see if it can be rejected
  - When the goal of an experiment is to establish an assertion, the negation of the assertion should be taken as the null hypothesis, and the assertion becomes the alternative hypothesis
- Alternative hypotheses usually specify that the population mean (or whatever other parameter is of concern) is not equal to, greater than, or less than the value assumed under the null hypothesis
  - Two-sided alternative $H_1 : \mu \neq x$
  - One-sided alternatives: $H_1 : \mu > x$ or $H_1 : \mu < x$

6

## Critical regions for two-sided and one-sided alternative hypotheses – depends on the decision problem

Null hypothesis: $\mu = \mu_0$

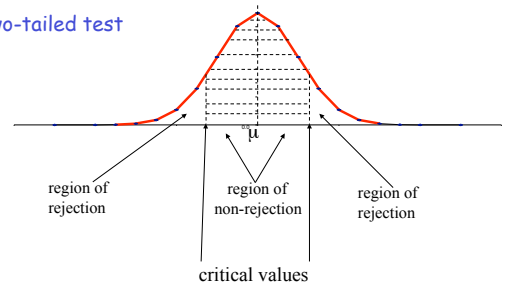
| Alternative hypothesis | Reject null hypothesis if: |
|---|---|
| $\mu < \mu_0$ | $Z < -z_\alpha$ |
| $\mu > \mu_0$ | $Z > z_\alpha$ |
| $\mu \neq \mu_0$ | $Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$ |

Note that the critical region for accepting the null hypothesis can be used to compute the (1-$\alpha$)100% confidence intervals for the population mean $\mu$, i.e.

$$(\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}})$$

7

## Two-tailed test




Test statistic: $Z = \dfrac{\bar{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$

8

## Steps in Hypothesis Testing

1. State the null and alternative hypothesis.
2. Choose the level of significance $\alpha$.
3. Choose the sample size n. Larger samples allow us to detect even small differences between sample statistics and true population parameters. For a given $\alpha$, increasing n decreases $\beta$.
4. Choose the appropriate statistical technique and test statistic to use (Z or t).

9

## Steps in Hypothesis Testing

5. Determine the critical values that divide the regions of acceptance and non-acceptance.
6. Collect the data and compute the sample mean and the appropriate test statistic (e.g., Z).
7. If the test statistic falls in the non-reject region, $H_o$ cannot be rejected. Else $H_o$ is rejected.

10

## Example of Hypothesis Testing

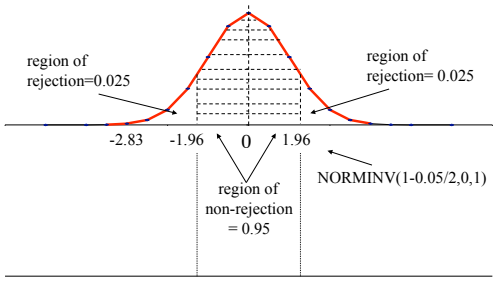
- A sample of 50 files from a file system is selected. The sample mean is 12.3 Kbytes. The standard deviation is known to be 0.5 Kbytes.

  $H_0$: $\mu = 12.5$ Kbytes

  $H_1$: $\mu \neq 12.5$ Kbytes

  Confidence: 0.95

11




$$Z = \frac{12.3 - 12.5}{\frac{0.5}{\sqrt{50}}} = -2.83$$

Reject $H_o$

If Z falls in the interval -1.96 to 1.96 hypothesis cannot be rejected

12

| Z Test of Hypothesis for the Mean | |
|---|---|
| | |
| **Null Hypothesis** μ= | **12.5** |
| **Level of Significance** | **0.05** |
| **Population Standard Deviation** | **0.5** |
| **Sample Size** | **50** |
| **Sample Mean** | **12.3** |
| Standard Error of the Mean | 0.070710678 |
| Z Test Statistic | -2.828427125 |
| | |
| **Two-Tailed Test** | |
| **Lower Critical Value** | **-1.959961082** |
| **Upper Critical Value** | **1.959961082** |
| ***p*-Value** | **0.00467786** |
| **Reject the null hypothesis** | |

The null hypothesis is rejected because p (0.0047) is less than the level of significance (0.05).

13

## Hypothesis Tests with Unknown σ

- We can estimate the variance by the sample variance
- For large samples, we can use the Z statistic
- For small samples, if the population is assumed to be normally distributed the sampling distribution for the mean follows a t distribution with n-1 degrees of freedom
- t statistic for unknown σ:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

sample standard deviation

14

## Example of Hypothesis Testing

- A sample of 5 files from a file system is selected. Assume that file sizes are normally distributed. The sample mean is 12.3 Kbytes. The sample standard deviation is 0.5 Kbytes.

  Ho: $\mu = 12.35$ Kbytes

  $H_1$: $\mu \neq 12.35$ Kbytes

  Confidence: 0.95

15

## Example

t = (12.3 - 12.35)/(0.5/√5) = -0.2236

$\alpha$ = 0.05, degrees of freedom = 4

$t_{\alpha/2}$ = 2.776 for 4 degrees of freedom

In EXCEL, TINV(0.05,4)

The t test statistic (-0.2236) is between the lower and upper critical values (i.e. -2.776 and 2.776)

So the null hypothesis should not be rejected.

16

## Example of One-Tailed Test

- A sample of 50 files from a file system is selected. The sample mean is 12.35 Kbytes. The standard deviation is known to be 0.5 Kbytes.

  Ho: $\mu = 12.3$ Kbytes

  $H_1$: $\mu < 12.3$ Kbytes

  Confidence: 0.95

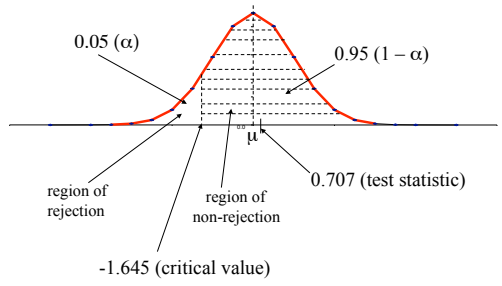
17

## Example of One-Tailed Test

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{12.35 - 12.3}{0.5 / \sqrt{50}} = 0.707$$ Statistic

Critical value = NORMINV(0.05,0,1) = -1.645.
Region of non-rejection: Z ≥ –1.645.
So, do not reject $H_o$. (Z exceeds critical value)

18




Test statistic: $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$

19

## One-tailed Test

| Z Test of Hypothesis for the Mean | | |
|---|---|---|
| | | |
| Null Hypothesis | μ= | 12.3 |
| Level of Significance | | 0.05 |
| Population Standard Deviation | | 0.5 |
| Sample Size | | 50 |
| Sample Mean | | 12.35 |
| Standard Error of the Mean | | 0.070710678 |
| Z Test Statistic | | 0.707106781 |
| | | |
| Lower-Tail Test | | |
| Lower Critical Value | | -1.644853 |
| *p*-Value | | 0.760250013 |
| Do not reject the null hypothesis | | |

20

## Steps in Determining the p-value.

1. State the null and alternative hypothesis.
2. Choose the level of significance $\alpha$.
3. Choose the sample size n. Larger samples allow us to detect even small differences between sample statistics and true population parameters. For a given $\alpha$, increasing n decreases $\beta$.
4. Choose the appropriate statistical technique and test statistic to use (Z or t).

21

## Steps in Determining the p-value.

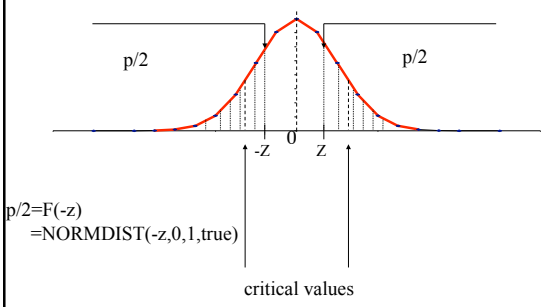
5. Collect the data and compute the sample mean and the appropriate test statistic (e.g., Z)
6. Calculate the p-value based on the test statistic
7. Compare the p-value to $\alpha$
8. If $p \geq \alpha$ then do not reject $H_o$, else reject $H_o$

22

| Z Test of Hypothesis for the Mean | |
|---|---|
| | |
| **Null Hypothesis** μ= | **12.5** |
| **Level of Significance** | **0.05** |
| **Population Standard Deviation** | **0.5** |
| **Sample Size** | **50** |
| **Sample Mean** | **12.3** |
| Standard Error of the Mean | 0.070710678 |
| Z Test Statistic | -2.828427125 |
| | |
| **Two-Tailed Test** | |
| **Lower Critical Value** | **-1.959961082** |
| **Upper Critical Value** | **1.959961082** |
| ***p*-Value** | **0.00467786** |
| **Reject the null hypothesis** | |

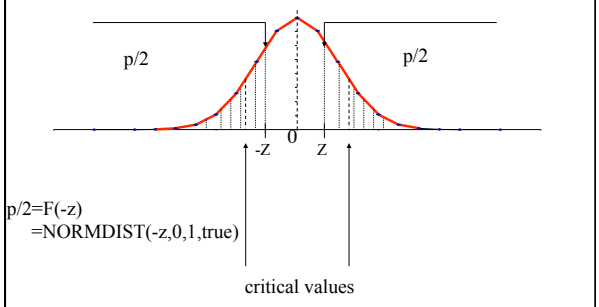
The null hypothesis is rejected because p (0.0047) is less than the level of significance (0.05).

23



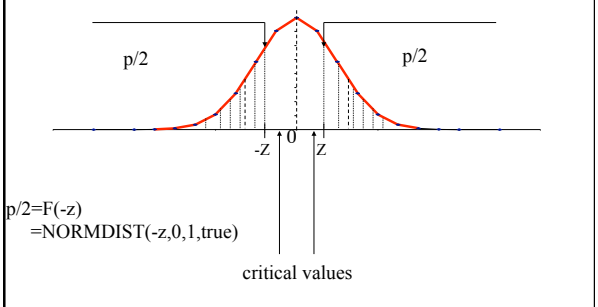

If $p \geq \alpha$ then do not reject $H_o$, else reject $H_o$.

24




Do not reject $H_o$

25




Reject $H_o$

26

## Computing p-values

| Z Test of Hypothesis for the Mean | | |
|---|---|---|
| | | |
| **Null Hypothesis** | **μ=** | **12.5** |
| **Level of Significance** | | **0.05** |
| **Population Standard Deviation** | | **0.5** |
| **Sample Size** | | **50** |
| **Sample Mean** | | **12.3** |
| Standard Error of the Mean | | 0.070710678 |
| Z Test Statistic | | -2.828427125 |
| | | |
| **Two-Tailed Test** | | |
| **Lower Critical Value** | | **-1.959961082** |
| **Upper Critical Value** | | **1.959961082** |
| ***p*-Value** | | **0.00467786** |
| **Reject the null hypothesis** | | |

The null hypothesis is rejected because p (0.0047) is less than the level of significance (0.05).

27

## Hypothesis testing vs estimating confidence intervals

- Textbooks on statistics devote a chapter to hypothesis testing
  - Example: Hypothesis test for a zero mean
  - Hypothesis test has a yes-no answer so either a hypothesis is accepted or rejected
  - Jain argues that confidence intervals provide more information
    - The difference between two systems has a confidence interval of (-100,100) vs a confidence interval of (-1,1)
    - In both cases, the interval includes zero but the width of the interval provides additional information

28

## Design of Experiments

CS 700

29

## Design of Experiments

- Goals
- How to find most about the system with minimal effort
- Terminology
- Full factorial designs
  - *m*-factor ANOVA
- Fractional factorial designs
- Multi-factorial designs

30

## Recall: One-Factor 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)?
- *One-factor experimental design*

31

## 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 $F$ | $F_{[1-\alpha;(k-1),k(n-1)]}$ | | |

32

## 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

33

## 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 (inputs)
    - Continuous (~bytes) or discrete (type of system)

34

## Terminology

- Replication
  - Completely re-run experiment with same input levels
  - Used to determine impact of measurement error
- Interaction
  - *Effect* of one input factor depends on *level* of another input factor

35

## Simplest strategy

- Vary one factor at the time – ignores interactions (e.g. clock time vs cache size)
- Full factorial design with replications Measure all possible input combinations – large number of experiments
- 4 factors, 5 possible level 4^5 experiments + repetition to gather some statistics

36

## One-factor Experiments

- ANOVA before: only compare types of system
- Separate variation due to error, variation due to alternative
- 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

37

## 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 $F$ | $F_{[1-\alpha;(k-1),k(n-1)]}$ | | |

38

## 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

39

## 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 |

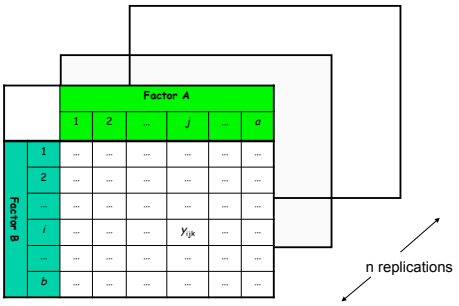
40

## Two-factor ANOVA

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

41

## Two Factors, *n* Replications




k-th measurement

42

## Recall: One-factor ANOVA

- Each individual measurement is composition of
  - Overall mean
  - Effect of alternatives
  - Measurement errors

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

$\bar{y}_{..}$ = overall mean

$\alpha_i$ = effect due to A

$e_{ij}$ = measurement error

43

## Two-factor ANOVA

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

$$y_{ijk} = \bar{y}_{...} + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

$\bar{y}_{...}$ = overall mean

$\alpha_i$ = effect due to A

$\beta_j$ = effect due to B

$\gamma_{ij}$ = effect due to interaction of A and B

$e_{ijk}$ = measurement error

44

## Computation of effects

$$\overline{y_{ij.}} = \overline{y_{...}} + \alpha_j + \beta_i + \gamma_{ij}$$
$$\alpha_j = \overline{y_{.j.}} - \overline{y_{...}}$$
$$\beta_i = \overline{y_{i..}} - \overline{y_{...}}$$
$$\gamma_{ij} = \overline{y_{ij.}} - \overline{y_{i..}} - \overline{y_{.j.}} + \overline{y_{...}}$$

4

## Sum-of-Squares

- As before, use sum-of-squares identity separate total variation

SST = SSA + SSB + SSAB + SSE

- Degrees of freedom
  - *df*(SSA) = *a* – 1
  - *df*(SSB) = *b* – 1
  - *df*(SSAB) = (*a* – 1)(*b* – 1)
  - *df*(SSE) = *ab*(*n* – 1)
  - *df*(SST) = *abn* - 1

46

## Two-Factor ANOVA

- Compute variances – mean squared values
- We can use F test to compare two variances
- If F is statistically significant if it is larger then critical F value

| | A | B | AB | Error |
|---|---|---|---|---|
| Sum of squares | $SSA$ | $SSB$ | $SSAB$ | $SSE$ |
| Deg freedom | $a-1$ | $b-1$ | $(a-1)(b-1)$ | $ab(n-1)$ |
| Mean square | $s_a^2 = SSA/(a-1)$ | $s_b^2 = SSB/(b-1)$ | $s_{ab}^2 = SSAB/[(a-1)(b-1)]$ | $s_e^2 = SSE/[ab(n-1)]$ |
| Computed $F$ | $F_a = s_a^2/s_e^2$ | $F_b = s_b^2/s_e^2$ | $F_{ab} = s_{ab}^2/s_e^2$ | |
| Tabulated $F$ | $F_{[1-\alpha;(a-1),ab(n-1)]}$ | $F_{[1-\alpha;(b-1),ab(n-1)]}$ | $F_{[1-\alpha;(a-1)(b-1),ab(n-1)]}$ | |

47

## Need for Replications

- If n=1
  - Only one measurement of each configuration
- Can then be shown that
  - SSAB = SST – SSA – SSB
- Since
  - SSE = SST – SSA – SSB – SSAB
- We have
  - SSE = 0

48

## Need for Replications

- Thus, when n=1
  - SSE = 0
  - → No information about measurement errors
- Cannot separate effect due to interactions from measurement noise
- Must *replicate* each experiment at least twice

49

## 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 |

50

## Example

| | B (Mbytes) | | |
|---|---|---|---|
| **A** | 32 | 64 | 128 |
| **1** | 0.25 | 0.21 | 0.15 |
| | 0.28 | 0.19 | 0.11 |
| **2** | 0.52 | 0.45 | 0.36 |
| | 0.48 | 0.49 | 0.30 |
| **3** | 0.81 | 0.66 | 0.50 |
| | 0.76 | 0.59 | 0.61 |
| **4** | 1.50 | 1.45 | 0.70 |
| | 1.61 | 1.32 | 0.68 |

51

## Example

| | A | B | AB | Error |
|---|---|---|---|---|
| Sum of squares | 3.3714 | 0.5152 | 0.4317 | 0.0293 |
| Deg freedom | 3 | 2 | 6 | 12 |
| Mean square | 1.1238 | 0.2576 | 0.0720 | 0.0024 |
| Computed $F$ | 460.2 | 105.5 | 29.5 | |
| Tabulated $F$ | $F_{[0.95;3,12]} = 3.49$ | $F_{[0.95;2,12]} = 3.89$ | $F_{[0.95;6,12]} = 3.00$ | |

52

## Conclusions From the 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

53

## Generalized *m*-factor Experiments

$m$ factors $\Rightarrow$

$m$ main effects

$\binom{m}{2}$ two-factor interactions

$\binom{m}{3}$ three-factor interactions

$\vdots$

$\binom{m}{m} = 1$ $m$-factor interactions

$2^m - 1$ total effects

Effects for 3 factors:

A
B
C
AB
AC
BC
ABC

54

## Degrees of Freedom for *m*-factor Experiments

- df(SSA) = (a-1)
- df(SSB) = (b-1)
- df(SSC) = (c-1)
- df(SSAB) = (a-1)(b-1)
- df(SSAC) = (a-1)(c-1)
- …
- df(SSE) = abc(n-1)
- df(SSAB) = abcn-1

55

## Procedure for Generalized *m*-factor Experiments

1. Calculate ($2^m$-1) sum of squares terms (SSx) and SSE
2. Determine degrees of freedom for each SSx
3. Calculate mean squares (variances)
4. Calculate F statistics
5. Find critical F values from table
6. If F(computed) > F(table), (1- $\alpha$ ) confidence that effect is statistically significant

56

## 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!

How to reduce the number of experiments ?

57

## 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

58

## Finding Sum of Squares Terms

| Sum of *n* measurements with (A,B) = (High, Low) | Factor A | Factor B |
|---|---|---|
| $Y_{AB}$ | High | High |
| $Y_{Ab}$ | High | Low |
| $Y_{aB}$ | Low | High |
| $Y_{ab}$ | Low | Low |

59

## $n2^m$ Contrasts

• Difference in systems responses when values are set to high and low for A, for B, and when A,B are set to different values

$$w_A = y_{AB} + y_{Ab} - y_{aB} - y_{ab}$$

$$w_B = y_{AB} - y_{Ab} + y_{aB} - y_{ab}$$

$$w_{AB} = y_{AB} - y_{Ab} - y_{aB} + y_{ab}$$

60

## $n2^m$ Experiments

| | A | B | AB | Error |
|---|---|---|---|---|
| Sum of squares | $SSA$ | $SSB$ | $SSAB$ | $SSE$ |
| Deg freedom | 1 | 1 | 1 | $2^m(n-1)$ |
| Mean square | $s_a^2 = SSA/1$ | $s_b^2 = SSB/1$ | $s_{ab}^2 = SSAB/1$ | $s_e^2 = SSE/[2^m(n-1)]$ |
| Computed $F$ | $F_a = s_a^2/s_e^2$ | $F_b = s_b^2/s_e^2$ | $F_{ab} = s_{ab}^2/s_e^2$ | |
| Tabulated $F$ | $F_{[1-\alpha;1,2^m(n-1)]}$ | $F_{[1-\alpha;1,2^m(n-1)]}$ | $F_{[1-\alpha;1,2^m(n-1)]}$ | |

61

## $n2^m$ Sum of Squares

$$SSA = \frac{w_A^2}{n2^m}$$

Total number on observations at all levels

$$SSB = \frac{w_B^2}{n2^m}$$

$$SSAB = \frac{w_{AB}^2}{n2^m}$$

$$SSE = SST - SSA - SSB - SSAB$$

62

## To Summarize -- $n2^m$ Experiments

| | A | B | AB | Error |
|---|---|---|---|---|
| Sum of squares | $SSA$ | $SSB$ | $SSAB$ | $SSE$ |
| Deg freedom | 1 | 1 | 1 | $2^m(n-1)$ |
| Mean square | $s_a^2 = SSA/1$ | $s_b^2 = SSB/1$ | $s_{ab}^2 = SSAB/1$ | $s_e^2 = SSE/[2^m(n-1)]$ |
| Computed $F$ | $F_a = s_a^2/s_e^2$ | $F_b = s_b^2/s_e^2$ | $F_{ab} = s_{ab}^2/s_e^2$ | |
| Tabulated $F$ | $F_{[1-\alpha;1,2^m(n-1)]}$ | $F_{[1-\alpha;1,2^m(n-1)]}$ | $F_{[1-\alpha;1,2^m(n-1)]}$ | |

63

## Contrasts for $n2^m$ with $m = 2$ factors -- revisited

| Measurements | Contrast | | |
|---|---|---|---|
| | $w_a$ | $w_b$ | $w_{ab}$ |
| $y_{AB}$ | + | + | + |
| $y_{Ab}$ | + | - | - |
| $y_{aB}$ | - | + | - |
| $y_{ab}$ | - | - | + |

$$w_A = y_{AB} + y_{Ab} - y_{aB} - y_{ab}$$

$$w_B = y_{AB} - y_{Ab} + y_{aB} - y_{ab}$$

$$w_{AB} = y_{AB} - y_{Ab} - y_{aB} + y_{ab}$$

Table specifying the signs

64

## Contrasts for $n2^m$ with $m$ = 3 factors

| Meas | Contrast | | | | | | |
|---|---|---|---|---|---|---|---|
| | $w_a$ | $w_b$ | $w_c$ | $w_{ab}$ | $w_{ac}$ | $w_{bc}$ | $w_{abc}$ |
| $y_{abc}$ | - | - | - | + | + | + | - |
| $y_{Abc}$ | + | - | - | - | - | + | + |
| $y_{aBc}$ | - | + | - | - | + | - | + |
| ... | ... | ... | ... | ... | ... | ... | ... |

2^8 combinations must be measured

$$w_{AC} = y_{abc} - y_{Abc} + y_{aBc} - y_{abC} - y_{ABc} + y_{AbC} - y_{aBC} + y_{ABC}$$

65

## $n2^m$ with $m$ = 3 factors

$$SSAC = \frac{w_{AC}^2}{2^3 n}$$

- df(each effect) = 1, since only two levels measured
- SST = SSA + SSB + SSC + SSAB + SSAC + SSBC + SSABC
- df(SSE) = $(n-1)2^3$
- Then perform ANOVA as before
- Easily generalizes to m > 3 factors

66

## Important Points

- Experimental design is used to
  - Isolate the effects of each input variable.
  - Determine the effects of interactions.
  - Determine the magnitude of the error
  - Obtain maximum information for given effort
- Expand 1-factor ANOVA to $m$ factors
- Use $n2^m$ design to reduce the number of experiments needed
  - But loses some information

67

## 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 $O(v^m)$

68

## Plackett and Burman Designs

- PB designs exist only in sizes that are multiples of 4
- Requires *X* experiments for *m* parameters
  - *X* = next multiple of 4 ≥ *m*
- PB design matrix
  - Rows = configurations of low and highs
  - Columns = parameters' values in each config
    - High/low = +1/ -1
  - First row = from P&B paper
  - Subsequent rows = circular right shift of preceding row
  - Last row = all (-1)

69

## PB Design Matrix

| Config | Input Parameters (factors) | | | | | | | Response |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | |
| 1 | +1 | +1 | +1 | -1 | +1 | -1 | -1 | 9 |
| 2 | -1 | +1 | +1 | +1 | -1 | +1 | -1 | |
| 3 | -1 | -1 | +1 | +1 | +1 | -1 | +1 | |
| 4 | +1 | -1 | -1 | +1 | +1 | +1 | -1 | |
| 5 | -1 | +1 | -1 | -1 | +1 | +1 | +1 | |
| 6 | +1 | -1 | +1 | -1 | -1 | +1 | +1 | |
| 7 | +1 | +1 | -1 | +1 | -1 | -1 | +1 | |
| 8 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | |
| Effect | | | | | | | | |

7 factors, 8 experiments

70

## PB Design Matrix

| Config | Input Parameters (factors) | | | | | | | Response |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | |
| 1 | +1 | +1 | +1 | -1 | +1 | -1 | -1 | 9 |
| 2 | -1 | +1 | +1 | +1 | -1 | +1 | -1 | 11 |
| 3 | -1 | -1 | +1 | +1 | +1 | -1 | +1 | |
| 4 | +1 | -1 | -1 | +1 | +1 | +1 | -1 | |
| 5 | -1 | +1 | -1 | -1 | +1 | +1 | +1 | |
| 6 | +1 | -1 | +1 | -1 | -1 | +1 | +1 | |
| 7 | +1 | +1 | -1 | +1 | -1 | -1 | +1 | |
| 8 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | |
| Effect | | | | | | | | |

71

## PB Design Matrix

| Config | Input Parameters (factors) | | | | | | | Response |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | |
| 1 | +1 | +1 | +1 | -1 | +1 | -1 | -1 | 9 |
| 2 | -1 | +1 | +1 | +1 | -1 | +1 | -1 | 11 |
| 3 | -1 | -1 | +1 | +1 | +1 | -1 | +1 | 2 |
| 4 | +1 | -1 | -1 | +1 | +1 | +1 | -1 | 1 |
| 5 | -1 | +1 | -1 | -1 | +1 | +1 | +1 | 9 |
| 6 | +1 | -1 | +1 | -1 | -1 | +1 | +1 | 74 |
| 7 | +1 | +1 | -1 | +1 | -1 | -1 | +1 | 7 |
| 8 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | 4 |
| Effect | | | | | | | | |

72

## PB Design Matrix

| Config | Input Parameters (factors) | | | | | | | Response |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | |
| 1 | +1 | +1 | +1 | -1 | +1 | -1 | -1 | 9 |
| 2 | -1 | +1 | +1 | +1 | -1 | +1 | -1 | 11 |
| 3 | -1 | -1 | +1 | +1 | +1 | -1 | +1 | 2 |
| 4 | +1 | -1 | -1 | +1 | +1 | +1 | -1 | 1 |
| 5 | -1 | +1 | -1 | -1 | +1 | +1 | +1 | 9 |
| 6 | +1 | -1 | +1 | -1 | -1 | +1 | +1 | 74 |
| 7 | +1 | +1 | -1 | +1 | -1 | -1 | +1 | 7 |
| 8 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | 4 |
| Effect | 65 | | | | | | | |

73

## PB Design Matrix

| Config | Input Parameters (factors) | | | | | | | Response |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | |
| 1 | +1 | +1 | +1 | -1 | +1 | -1 | -1 | 9 |
| 2 | -1 | +1 | +1 | +1 | -1 | +1 | -1 | 11 |
| 3 | -1 | -1 | +1 | +1 | +1 | -1 | +1 | 2 |
| 4 | +1 | -1 | -1 | +1 | +1 | +1 | -1 | 1 |
| 5 | -1 | +1 | -1 | -1 | +1 | +1 | +1 | 9 |
| 6 | +1 | -1 | +1 | -1 | -1 | +1 | +1 | 74 |
| 7 | +1 | +1 | -1 | +1 | -1 | -1 | +1 | 7 |
| 8 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | 4 |
| Effect | 65 | -45 | | | | | | |

74

## PB Design Matrix

| Config | Input Parameters (factors) | | | | | | | Response |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | |
| 1 | +1 | +1 | +1 | -1 | +1 | -1 | -1 | 9 |
| 2 | -1 | +1 | +1 | +1 | -1 | +1 | -1 | 11 |
| 3 | -1 | -1 | +1 | +1 | +1 | -1 | +1 | 2 |
| 4 | +1 | -1 | -1 | +1 | +1 | +1 | -1 | 1 |
| 5 | -1 | +1 | -1 | -1 | +1 | +1 | +1 | 9 |
| 6 | +1 | -1 | +1 | -1 | -1 | +1 | +1 | 74 |
| 7 | +1 | +1 | -1 | +1 | -1 | -1 | +1 | 7 |
| 8 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | 4 |
| Effect | 65 | -45 | 75 | -75 | -75 | 73 | 67 | |

75

## PB Design

- Only magnitude of effect is important
  - Sign is meaningless
- In example, most → least important effects:
  - [C, D, E] → F → G → A → B

76

## PB Design Matrix with Foldover

- Add *X* additional rows to matrix
  - Signs of additional rows are opposite original rows
- Provides some additional information about selected interactions

77

## Case Study #1

- Determine the most significant parameters in a processor simulator.
- [Yi, Lilja, & Hawkins, HPCA, 2003.]

79

## Determine the Most Significant Processor Parameters

- Problem
  - So many parameters in a simulator
  - How to choose parameter values?
  - How to decide which parameters are most important?
- Approach
  - Choose reasonable upper/lower bounds.
  - Rank parameters by impact on total execution time.

80

## Simulation Environment

- SimpleScalar simulator
  - sim-outorder 3.0
- Selected SPEC 2000 Benchmarks
  - *gzip, vpr, gcc, mesa, art, mcf, equake, parser, vortex, bzip2, twolf*
- MinneSPEC Reduced Input Sets
- Compiled with gcc (PISA) at O3

81

## Functional Unit Values

| Parameter | Low Value | High Value |
|---|---|---|
| Int ALUs | 1 | 4 |
| Int ALU Latency | 2 Cycles | 1 Cycle |
| Int ALU Throughput | 1 | |
| FP ALUs | 1 | 4 |
| FP ALU Latency | 5 Cycles | 1 Cycle |
| FP ALU Throughputs | 1 | |
| Int Mult/Div Units | 1 | 4 |
| Int Mult Latency | 15 Cycles | 2 Cycles |
| Int Div Latency | 80 Cycles | 10 Cycles |
| Int Mult Throughput | 1 | |
| Int Div Throughput | Equal to Int Div Latency | |
| FP Mult/Div Units | 1 | 4 |
| FP Mult Latency | 5 Cycles | 2 Cycles |
| FP Div Latency | 35 Cycles | 10 Cycles |
| FP Sqrt Latency | 35 Cycles | 15 Cycles |
| FP Mult Throughput | Equal to FP Mult Latency | |
| FP Div Throughput | Equal to FP Div Latency | |
| FP Sqrt Throughput | Equal to FP Sqrt Latency | |

82

## Memory System Values, Part I

| Parameter | Low Value | High Value |
|---|---|---|
| L1 I-Cache Size | 4 KB | 128 KB |
| L1 I-Cache Assoc | 1-Way | 8-Way |
| L1 I-Cache Block Size | 16 Bytes | 64 Bytes |
| L1 I-Cache Repl Policy | Least Recently Used | |
| L1 I-Cache Latency | 4 Cycles | 1 Cycle |
| L1 D-Cache Size | 4 KB | 128 KB |
| L1 D-Cache Assoc | 1-Way | 8-Way |
| L1 D-Cache Block Size | 16 Bytes | 64 Bytes |
| L1 D-Cache Repl Policy | Least Recently Used | |
| L1 D-Cache Latency | 4 Cycles | 1 Cycle |
| L2 Cache Size | 256 KB | 8192 KB |
| L2 Cache Assoc | 1-Way | 8-Way |
| L2 Cache Block Size | 64 Bytes | 256 Bytes |

83

## Memory System Values, Part II

| Parameter | Low Value | High Value |
|---|---|---|
| L2 Cache Repl Policy | Least Recently Used | |
| L2 Cache Latency | 20 Cycles | 5 Cycles |
| Mem Latency, First | 200 Cycles | 50 Cycles |
| Mem Latency, Next | 0.02 * Mem Latency, First | |
| Mem Bandwidth | 4 Bytes | 32 Bytes |
| I-TLB Size | 32 Entries | 256 Entries |
| I-TLB Page Size | 4 KB | 4096 KB |
| I-TLB Assoc | 2-Way | Fully Assoc |
| I-TLB Latency | 80 Cycles | 30 Cycles |
| D-TLB Size | 32 Entries | 256 Entries |
| D-TLB Page Size | Same as I-TLB Page Size | |
| D-TLB Assoc | 2-Way | Fully-Assoc |
| D-TLB Latency | Same as I-TLB Latency | |

84

## Processor Core Values

| Parameter | Low Value | High Value |
|---|---|---|
| Fetch Queue Entries | 4 | 32 |
| Branch Predictor | 2-Level | Perfect |
| Branch MPred Penalty | 10 Cycles | 2 Cycles |
| RAS Entries | 4 | 64 |
| BTB Entries | 16 | 512 |
| BTB Assoc | 2-Way | Fully-Assoc |
| Spec Branch Update | In Commit | In Decode |
| Decode/Issue Width | 4-Way | |
| ROB Entries | 8 | 64 |
| LSQ Entries | 0.25 * ROB | 1.0 * ROB |
| Memory Ports | 1 | 4 |

85

## Determining the Most Significant Parameters

1. Run simulations to find **response**
   - With input parameters at high/low, on/off values

| Config | Input Parameters (factors) | | | | | | | Response |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | |
| 1 | +1 | +1 | +1 | -1 | +1 | -1 | -1 | 9 |
| 2 | -1 | +1 | +1 | +1 | -1 | +1 | -1 | |
| 3 | -1 | -1 | +1 | +1 | +1 | -1 | +1 | |
| … | … | … | … | … | … | … | … | |
| Effect | | | | | | | | |

86

## Determining the Most Significant Parameters

2. Calculate the **effect** of each parameter
   - Across configurations

| Config | Input Parameters (factors) | | | | | | | Response |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | |
| 1 | +1 | +1 | +1 | -1 | +1 | -1 | -1 | 9 |
| 2 | -1 | +1 | +1 | +1 | -1 | +1 | -1 | |
| 3 | -1 | -1 | +1 | +1 | +1 | -1 | +1 | |
| … | … | … | … | … | … | … | … | |
| Effect | 65 | | | | | | | |

87

## Determining the Most Significant Parameters

3. For each benchmark
   **Rank** the parameters in descending order of effect
   (1=most important, …)

| Parameter | Benchmark 1 | Benchmark 2 | Benchmark 3 |
|---|---|---|---|
| A | 3 | 12 | 8 |
| B | 29 | 4 | 22 |
| C | 2 | 6 | 7 |
| … | … | … | … |

88

## Determining the Most Significant Parameters

4. For each parameter
   **Average** the ranks

| Parameter | Benchmark 1 | Benchmark 2 | Benchmark 3 | Average |
|---|---|---|---|---|
| A | 3 | 12 | 8 | 7.67 |
| B | 29 | 4 | 22 | 18.3 |
| C | 2 | 6 | 7 | 5 |
| … | … | … | … | … |

89

## Most Significant Parameters

| Number | Parameter | gcc | gzip | art | Average |
|---|---|---|---|---|---|
| 1 | ROB Entries | 4 | 1 | 2 | 2.77 |
| 2 | L2 Cache Latency | 2 | 4 | 4 | 4.00 |
| 3 | Branch Predictor Accuracy | 5 | 2 | 27 | 7.69 |
| 4 | Number of Integer ALUs | 8 | 3 | 29 | 9.08 |
| 5 | L1 D-Cache Latency | 7 | 7 | 8 | 10.00 |
| 6 | L1 I-Cache Size | 1 | 6 | 12 | 10.23 |
| 7 | L2 Cache Size | 6 | 9 | 1 | 10.62 |
| 8 | L1 I-Cache Block Size | 3 | 16 | 10 | 11.77 |
| 9 | Memory Latency, First | 9 | 36 | 3 | 12.31 |
| 10 | LSQ Entries | 10 | 12 | 39 | 12.62 |
| 11 | Speculative Branch Update | 28 | 8 | 16 | 18.23 |

90

## General Procedure

- Determine upper/lower *bounds* for parameters
- Simulate configurations to find *response*
- Compute *effects* of each parameter for each configuration
- *Rank* the parameters for each benchmark based on effects
- *Average* the ranks across benchmarks
- Focus on *top-ranked* parameters for subsequent analysis

91

## Summary

- Design of experiments
  - Isolate effects of each input variable.
  - Determine effects of interactions.
  - Determine magnitude of experimental error
- *m*-factor ANOVA (*full factorial design*)
  - All effects, interactions, and errors

111

## Summary

- $n2^m$ designs
  - Fractional factorial design
- All effects, interactions, and errors
- But for only 2 input values
  - high/low
  - on/off

112

## Summary

- Plackett and Burman (*multi-factorial design*)
- O(*m*) experiments
- Main effects only
  - *No interactions*
- For only 2 input values (high/low, on/off)
- Examples – rank parameters, group benchmarks, overall impact of an enhancement

113