6.780 Semiconductor Manufacturing
Spring, 2003

DESIGN and ANALYSIS of EXPERIMENTS

- The goals of the next portion of the course are to develop the methods to
  (1) determine if proposed process/equipment modifications improve or impact the results of concern

  (2) devise experiments to aid in modeling or optimization of the process

- Comparison of Treatments
- Blocking and Randomization
- Reference Distributions
- ANOVA
- MANOVA
- Factorial Designs
- Two-Level Factorials
- Fractional Factorials
- Regression Analysis
- Robust (Taguchi) Design

↑ Analysis

↓ Modeling

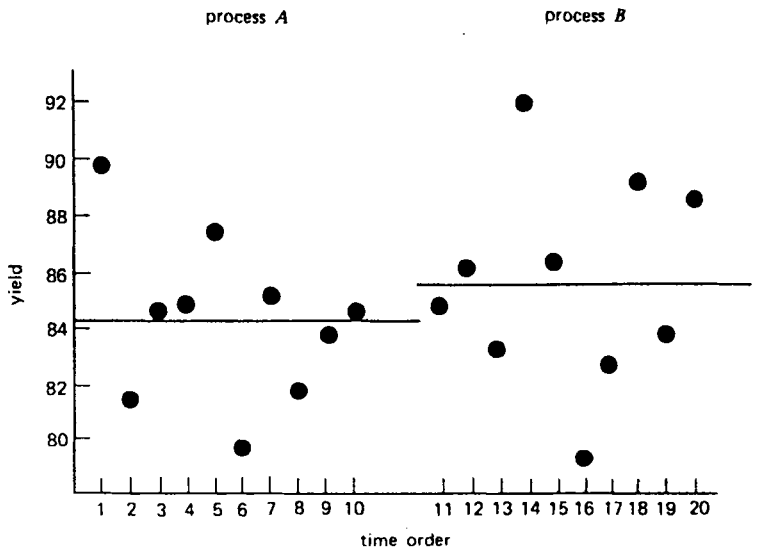
A SIMPLE EXPERIMENT: COMPARISON OF TREATMENTS

- A new process, process B, is to be compared against the process of record – process A. Process 10 lots w. Process A and 10 with B, and measure the yield for each lot:

time order	method	yield
1	*A*	89.7
2	*A*	81.4
3	*A*	84.5
4	*A*	84.8
5	*A*	87.3
6	*A*	79.7
7	*A*	85.1
8	*A*	81.7
9	*A*	83.7
10	*A*	84.5
11	*B*	84.7
12	*B*	86.1
13	*B*	83.2
14	*B*	91.9
15	*B*	86.3
16	*B*	79.3
17	*B*	82.6
18	*B*	89.1
19	*B*	83.7
20	*B*	88.5

$\bar{y}_A = 84.24, \quad \bar{y}_B = 85.54$

$\bar{y}_B - \bar{y}_A = 1.30$



QUESTION: Is Process B really better than Process A?

$H_0: \mu_A = \mu_B$ vs. $H_1: \mu_A < \mu_B$

APPROACHES:

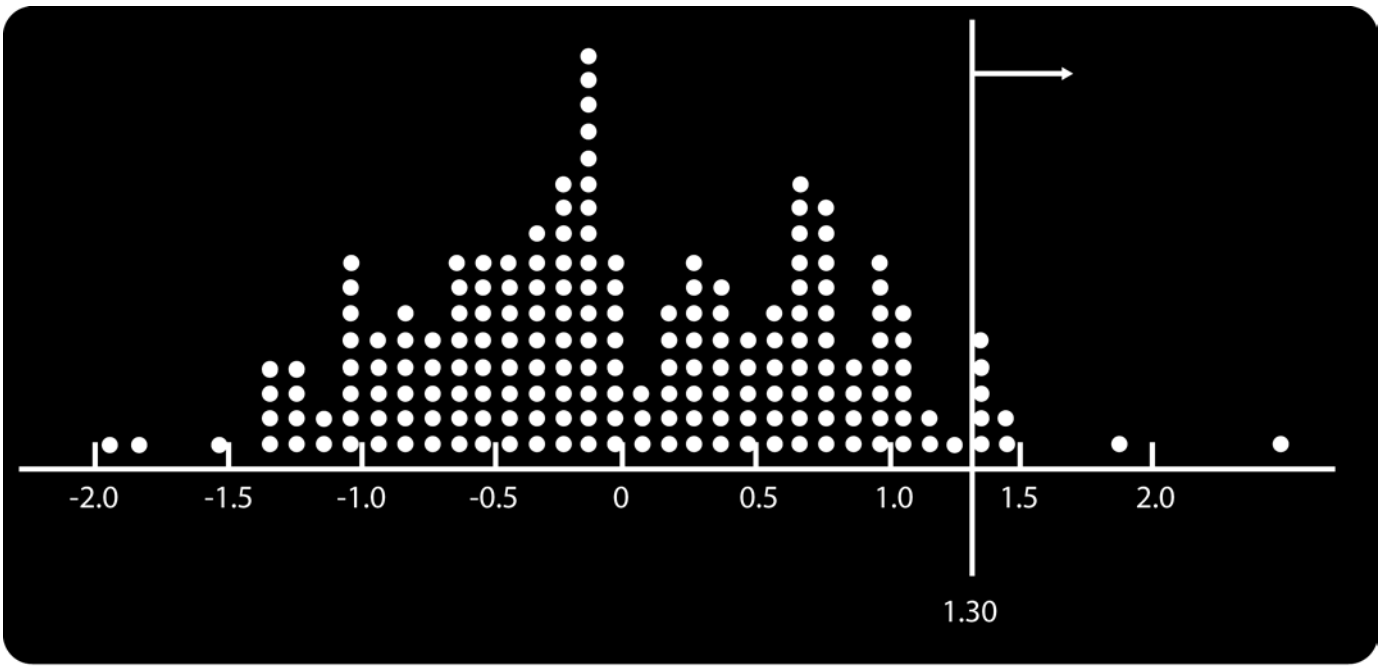
1. External historical data – empirical distribution
2. External t-distribution
3. External normal distribution with random sampling assumption
4. Internal estimate of standard deviation – assume random sampling from normal populations

Approach 1: External Reference Distribution

- Suppose we have historical data for process A:

- Since our experiment is comparing $\bar{y}_A$ to $\bar{y}_B$, where each constructed by 10 consecutive runs, we can build a reference distribution of all possible differences between adjacent runs of 10 from historical data

"Randomization Distribution"



Reference distribution of 191 differences between averages of adjacent sets of 10 observations.

$\Rightarrow$ EMPIRICAL level of significance / confidence:

$\frac{4}{191}$ runs gave a difference $\geq 1.30 = \bar{y}_B - \bar{y}_A$

$\therefore$ difference is statistically significant at $\alpha = 0.047$ level. //

NOTE: No assumption of independence!

Approach 2: External - t distribution

- Can build up a set of 10 differences in averages between non-overlapping historical data... with small gaps in between.

  What we gain: (1) 10 differences ~ Normally distributed - CLT
  (2) approximate independence of differences in averages ... even if some auto-correlation in individual runs

- Compute a SAMPLE std. dev of mean differences (assuming population mean of 0):

Ten nearly independent differences

observed results	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_2 - \bar{y}_1$
from past records	83.94	83.51	−0.43
	83.99	84.42	0.43
	84.18	84.01	−0.17
	85.18	84.28	−0.90
	83.58	84.38	0.80
	84.42	83.99	−0.43
	84.72	84.21	−0.51
	84.78	83.96	−0.82
	84.09	84.58	0.49
	83.62	84.26	0.64
	$\bar{y}_A$	$\bar{y}_B$	$\bar{y}_B - \bar{y}_A$
from plant trial	84.24	85.54	1.30
variance of differences		$s^2 = 0.36$	
standard deviation of differences		$s = 0.60$	

- Now we base our test on difference between means with an estimated population variance:

$$t_0 = \frac{(\bar{y}_B - \bar{y}_A) - (\mu_B - \mu_A)}{s} \quad ; \text{ under } H_0,\ \mu_B = \mu_A$$

$$= \frac{1.30}{0.60} = 2.17 \Rightarrow \alpha = 0.028 \text{ level of confidence} \ //$$

t with 10 d.o.f.

Approach 3: Assume random sampling from normal distrib. with external value for σ

- Now we're close to one of our more standard statistical tests $\Rightarrow$ inferences on sampling distributions

$n_A = 10$, $n_B = 10$, $\quad Var(\bar{y}_A) = \frac{\sigma^2}{n_A}$ & $Var(\bar{y}_B) = \frac{\sigma^2}{n_B}$

- How about the sampling distribution for the difference in means?

$$Var(\bar{y}_B - \bar{y}_A) = \frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B} = \sigma^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)$$

$$s_{B-A} = \sigma\sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \quad \text{.. standard error}$$

- Even if original process moderately nonnormal, distribution of $\bar{y}_B - \bar{y}_A$ ≈ normal by CLT

$$z = \frac{(\bar{y}_B - \bar{y}_A) - (\mu_B - \mu_A)}{\sigma\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

Using $s = \hat{\sigma}$ from historical (individual) data

$$= \sigma\sqrt{\frac{1}{10} + \frac{1}{10}} = \frac{2.88}{\sqrt{5}} = 1.29$$

Under H_0, $\mu_B = \mu_A$, so

$$z_0 = \frac{1.30}{1.29} = 1.01 \Rightarrow Pr(z > z_0) = 0.156$$

- Technically, we're estimating σ; really, we OUGHT to use

$t_0 = 1.01$ with $N = 210$ or 209 degrees of freedom

- Disadvantage: Still requires external historical data

Approach 4: Random Sampling with Internal σ Estimate

- Now suppose we do not have historical data – only $n_A = 10$ & $n_B = 10$ runs.

- Task 1: Estimate σ from samples

  Individual variances – $\nu_A = n_A - 1$ & $\nu_B = n_B - 1$

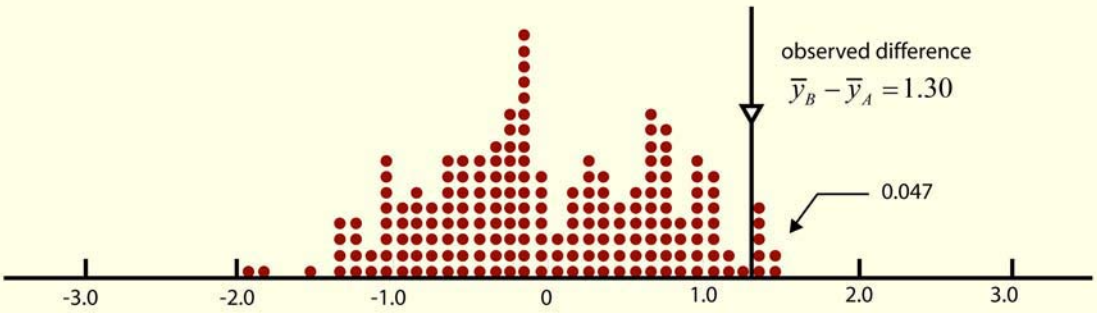
$$s_A^2 = \frac{1}{\nu_A} \sum (y_A - \bar{y}_A)^2 \quad \& \quad s_B^2 = \frac{1}{\nu_B} \sum (y_B - \bar{y}_B)^2$$

  Pooled variance –

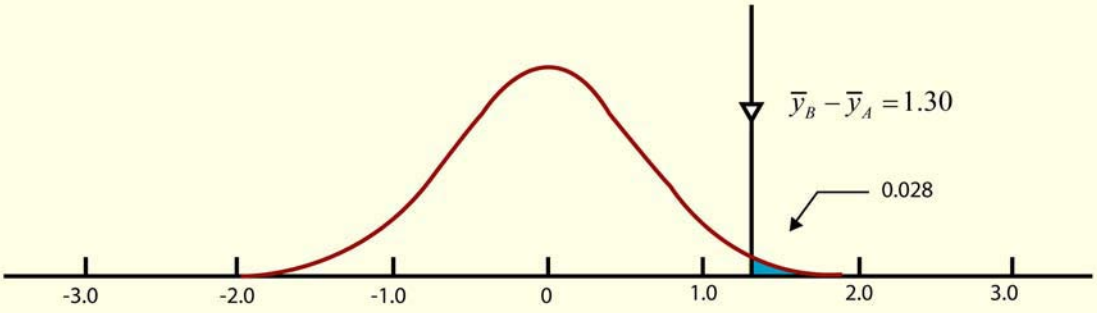
$$s^2 = \frac{\sum (y_A - \bar{y}_A)^2 + \sum (y_B - \bar{y}_B)^2}{n_A + n_B - 2} = \frac{\nu_A s_A^2 + \nu_B s_B^2}{\nu_A + \nu_B}$$

- Task 2: Form t test using pooled estimate for variance

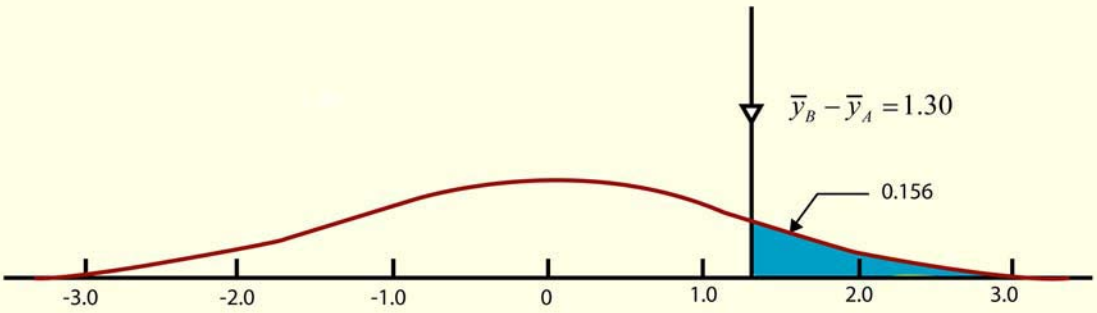
Comparison of Approaches: Testing Experiment Mean Difference



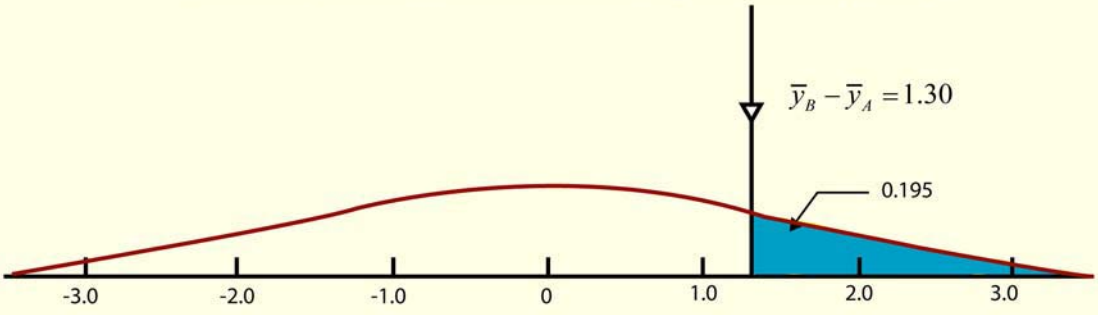
(A) External reference distribution



(B) External *t* reference distribution



(C) Reference distribution based on random sampling model, external value for σ .



(D) Reference distribution based on random sampling model, internal estimate of σ .

Four reference distributions for industrial experiment, with significance levels shown at right.

Observations:
- The more rough our estimate of distribution, the less confidence we have in mean shift
- Random sampling – IID – assumption is crucial to use of internal data in inferences!

Effect of Autocorrelation on Variance Estimates

- Recall our definitions & properties

$$X = a_1 y_1 + a_2 y_2 + a_3 y_3 = \sum_{i=1}^{3} a_i y_i$$

$$E(x) = \mu_x = a_1 \mu_1 + a_2 \mu_2 + a_3 \mu_3$$

$$\text{Var}(x) = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + a_3^2 \sigma_3^2 + 2a_1 a_2 \sigma_1 \sigma_2 \rho_{12} + 2a_1 a_3 \sigma_1 \sigma_3 \rho_{13} + 2a_2 a_3 \sigma_2 \sigma_3 \rho_{23}$$

If y_1, y_2, y_3 are NOT independent

$$= \sum_{i=1}^{n} a_i^2 \sigma_i^2 + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} a_i a_j \underbrace{\sigma_i \sigma_j \rho_{ij}}_{\text{Cov}(y_i, y_j)}$$

- Suppose y_i have $\mu_i = \mu$, $\sigma_i = \sigma$, and a lag 1 autocorrelation $\rho_1 = \rho_{i-1,i} \neq 0$

$$\text{Var}(x) = \text{Var}(\Sigma y) = \text{Var}(n\bar{y})$$

Since $a_i = 1$, $\sigma_i = \sigma$, $\rho_1 \neq 0$ use above to find

$$\text{Var}(n\bar{y}) = n\sigma^2 + 2\sigma^2(n-1)\rho_1 = \sigma^2\left[n + 2(n-1)\rho_1\right]$$

- Since $-1 \leq \rho \leq 1$, the sampling variance can be either INFLATED or DEFLATED quite largely.

$$\text{Var}(\bar{y}) = \frac{\sigma^2}{n}\left[1 + \frac{2(n-1)}{n}\rho_1\right] = \frac{\sigma^2}{n} C$$

E.g. for $-1/2 \leq \rho_1 \leq 1/2$, $n = 10 \Rightarrow 0.1 \leq C \leq 1.9$

RANDOMIZATION & BLOCKING

- Randomization: to ensure validity in face of unknown disturbances
- Blocking: to eliminate unwanted sources of variability

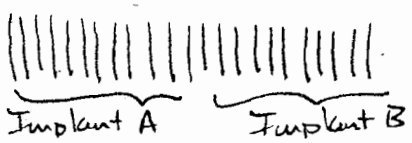
(1) Randomization

- we depend on an experimental realization being representative of the "randomization distribution" formed by the internal pool of data ⇒ allow use of t

- Important to ensure that treatments are applied randomly in time and/or space

  E.g. • $n_A, n_B = 10$ process experiments should be randomly intermixed to avoid correlation, trend issues

  • Suppose we've subjected half of 24 wafers each to 2 different implants. In the subsequent furnace anneal (batch), how should we order the wafers?



→ how tell if any difference is due to implant, or due to tube end?

- Especially important for dealing with time or space trends

(2) Blocking

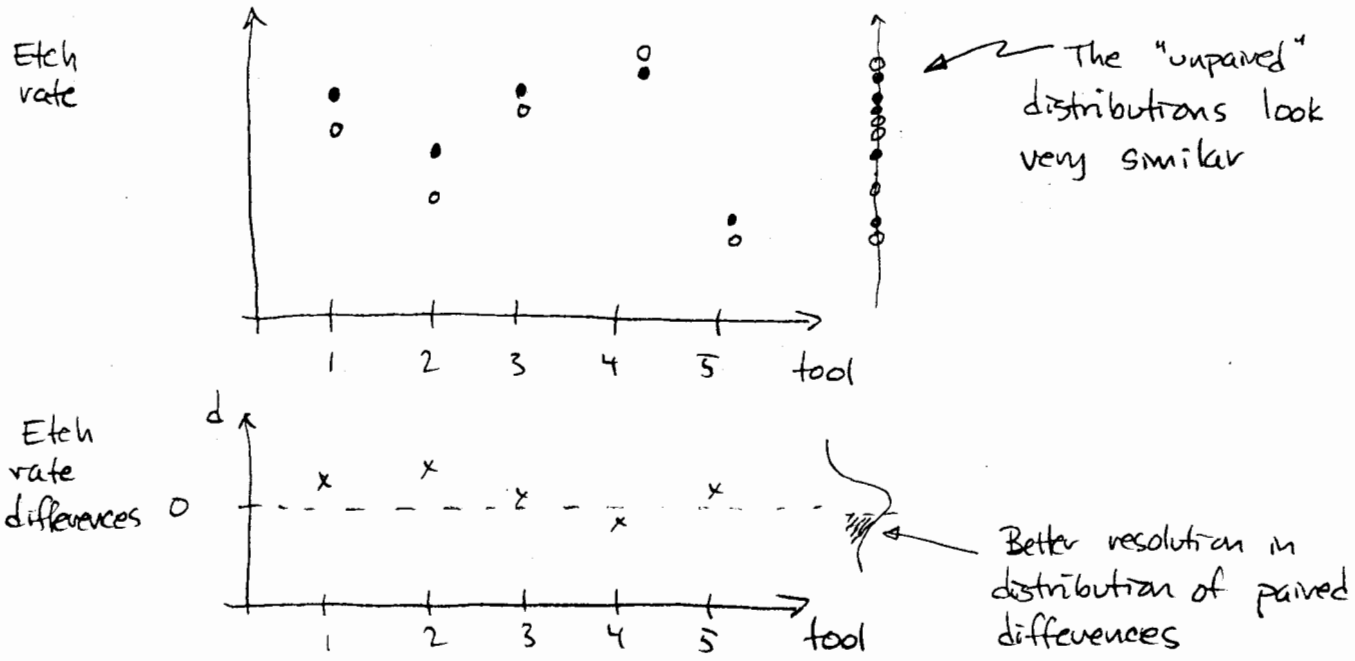
- Experimental precision can often be greatly increased if we can make comparisons within matched pairs of experimental items.

- Example: Compare recipes A & B, across five machines

AB	BA	BA	AB	BA

⇒ Run each recipe on each machine
⇒ Randomize order of runs within machines

- Advantage: Can now examine differences (B - A) across the tools:

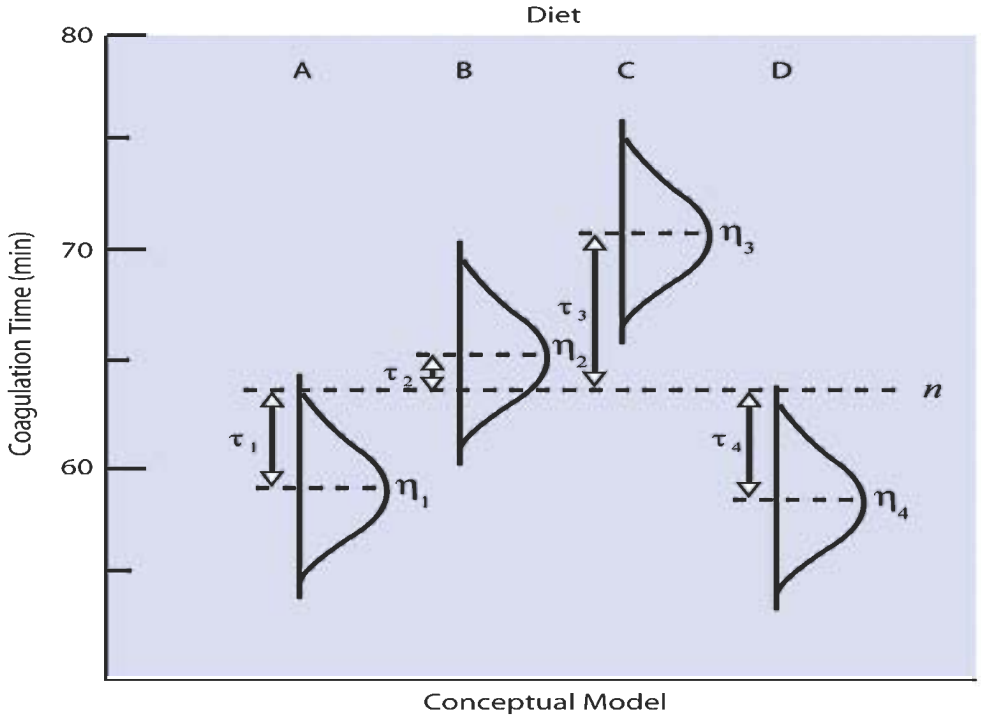


$\bar{d} = \frac{d_1 + d_2 + d_3 + d_4 + d_5}{5}$; can compare to a t-distribution approximation of $\bar{d}$ created by randomization $\frac{\pm d_1 \pm d_2 \pm d_3 \pm d_4 \pm d_5}{5}$

So $\frac{\bar{d} - \delta}{s/\sqrt{n}} \sim t_{n-1}$ to test for probability of observing deviation given H_0: $\delta = 0$

Analysis Of Variance (ANOVA)
-Comparing Several Treatment Means

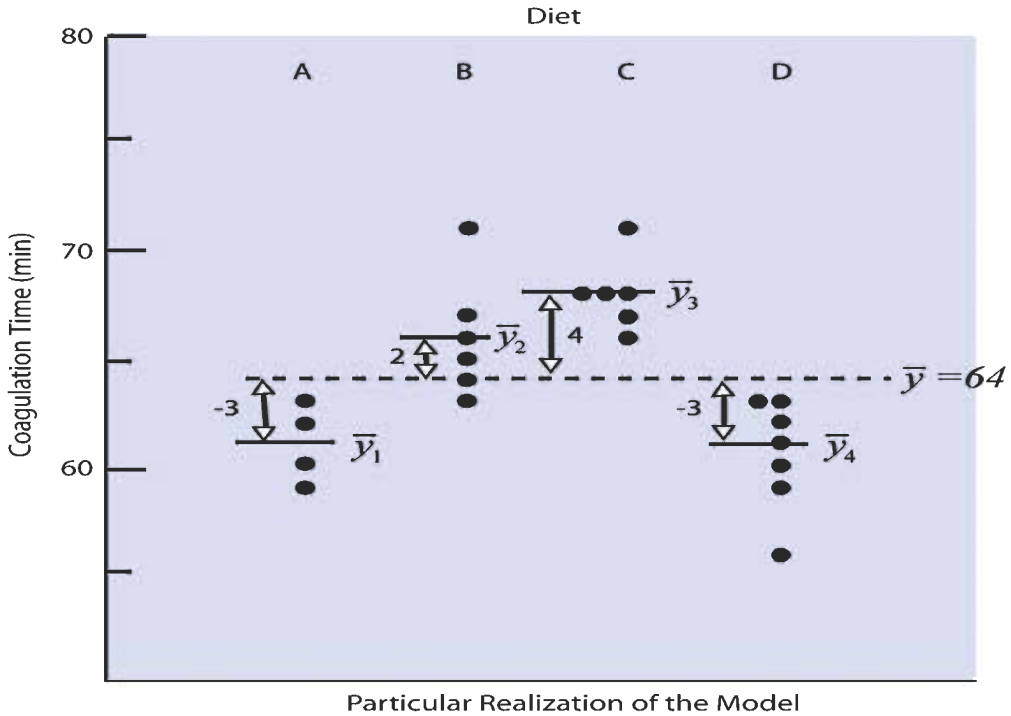
- Suppose we want to compare 4 different process options - A, B, C, & D. How do we tell if these "treatments" has any effect?



Conceptual Model

POPULATION

- True difference in means?



Particular Realization of the Model

EXPERIMENT SAMPLES

- estimates of population parameters

BETWEEN vs. WITHIN GROUP Variation!

DATA for diet/coagulation Example

– Run order in time shown in parentheses

Coagulation time (seconds) for blood drawn from 24 animals randomly allocated to four different diets

	diet (treatment)			
	A	*B*	*C*	*D*
	62(20)	63(12)	68(16)	56(23)
	60(2)	67(9)	66(7)	62(3)
	63(11)	71(15)	71(1)	60(6)
	59(10)	64(14)	67(17)	61(18)
		65(4)	68(13)	63(22)
		66(8)	68(21)	64(19)
				63(5)
				59(24)
treatment average	61	66	68	61
grand average	64			

(1) Within Group Variation

- Assume that each group is normally distributed and share a common σ^2

$$S_t \triangleq \text{Sum of squares, within } t^{th} \text{ group} = \sum_{j=1}^{n_t} (y_{tj} - \bar{y}_t)^2$$

where n_t = # of samples for t^{th} treatment

$$s_t^2 = S_t / \nu_t = \frac{S_t}{n_{t-1}} \quad \text{since } \nu_t = n_t - 1$$

POOLING these to get estimate of common, w/in group σ^2:

$$s_R^2 = \frac{\nu_1 s_1^2 + \nu_2 s_2^2 + \dots + \nu_K s_K^2}{\nu_1 + \nu_2 + \dots \nu_K} = \frac{S_R}{N-k} = \frac{S_R}{\nu_R}$$

$\triangleq$ WITH-IN TREATMENT MEAN SQUARE

(2) Between Group Variation

$\Rightarrow$ We will be testing hypothesis $\mu_1 = \mu_2 = \dots = \mu_K$;

In this case, a 2nd estimate of σ^2 would be

$$s_T^2 = \frac{\sum_{t=1}^{k} n_t (\bar{y}_t - \bar{y})^2}{k-1} = \frac{S_I}{\nu_T}$$

$\triangleq$ BETWEEN TREATMENT MEAN SQUARE

(3) Key Question: What if treatments <u>are</u> different?

Then s_T^2 estimates $\sigma^2 + \left[\sum_{t=1}^{k} n_t \tau_t^2 / (k-1) \right]$

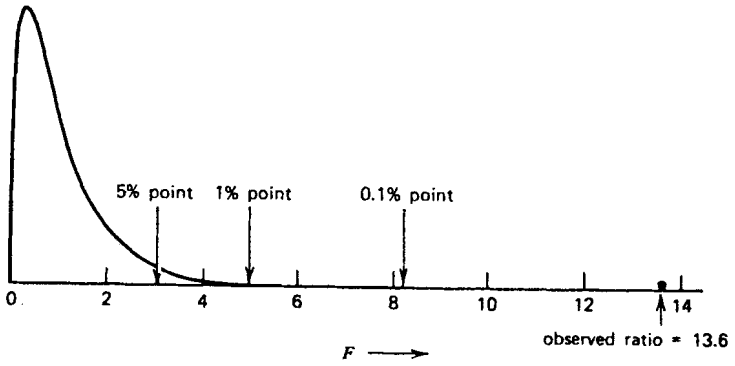
where $\tau_t = \mu_t - \mu$

$\Rightarrow$ That is, s_T^2 is <u>inflated</u> by some factor related to the difference between treatments!

(4) Formal Test for Treatment Significance

Reject H_0 if $\frac{S_T^2}{S_R^2}$ is significantly > 1

$\Rightarrow$ use F distribution, since $\frac{S_T^2}{S_R^2} \sim F_{k-1, N-k}$



$\leftarrow F_{3,20}$

(5) Total Variation

$$S_D' = \sum_{t=1}^{k} \sum_{i=1}^{n_t} (y_{ti} - \bar{y})^2$$

$$s_D^2 = \frac{S_D'}{\nu_D} = \frac{S_D'}{N-1}$$ Total sample variance

(6) Table:

An analysis of variance table

source of variation	sum of squares	degrees of freedom	mean square
between treatments	$S_T = 228$	$\nu_T = 3$	$s_T^2 = 76.0$
within treatments	$S_R = 112$	$\nu_R = 20$	$s_R^2 = 5.6$
total about the grand average	$S_D = 340$	$\nu_D = 23$	$s_D^2 = 14.8$

ANOVA AS DECOMPOSITION of VARIATION

- Algebraic identity:

$$\sum_{t=1}^{k}\sum_{i=1}^{n_t}(y_{ti}-\bar{y})^2 = \sum_{t=1}^{k} n_t(\bar{y}_t-\bar{y})^2 + \sum_{t=1}^{k}\sum_{i=1}^{n_t}(y_{ti}-\bar{y}_t)^2$$

$$S_D = S_T + S_R$$

Total Sum of Squares of deviations from grand ave = between-treatment sum of squares + within-treatment sum of squares

- One can further decompose the total sum of squares:

$$S' = S'_A + S'_D$$

$$\sum_{t=1}^{k}\sum_{i=1}^{n_t} y_{ti}^2 = N\bar{y}^2 + \sum_{t=1}^{k}\sum_{i=1}^{n_t}(y_{ti}-\bar{y})^2$$

Total Sum of Squares (about zero origin) = Sum of squares due to the average + Total sum of squares of deviations from average

- So $S' = S'_A + S_T + S_R$

d.o.f: $N = 1 + k-1 + N-k$

Full analysis of variance table

source of variation	sum of squares	degrees of freedom	mean square
average	98,304	1	98,304
between treatments	228	3	76.0
within treatments	112	20	5.6
total	98,644	24	

ANOVA: CHECKING THE MODEL & RESIDUAL ANALYSIS

- A key assumption in resolving differences between variation components is that the residuals are "random" - IID & $N(0, \sigma^2)$

- Assumed mathematical models:

$$y_{ti} = \mu_t + \epsilon_{ti} \qquad ; \text{ plot } \; y_{ti} - \hat{y}_{ti}$$

(μ_t: treatment mean; ϵ_{ti}: residuals)

- CHECKS:
  1. Plot residuals against time order
     - attempt to catch any time trends. While it is possible to randomize against such trends, we lose resolving power if such a trend is large.
  2. Examine distribution of residuals
     - check IID $N(0, \sigma^2)$ assumption; look for gross non-normality
     $\Rightarrow$ examine residuals for each treatment group
  3. Plot residuals vs. estimates
     - be especially alert to dependencies on size of estimate (e.g. proportional vs, absolute errors)
  4. Plot residuals vs. other variables of interest
     - consider other variables of possible relevance (e.g. environmental factors)

Two-Way Analysis of Variance: MANOVA

- Suppose we carefully structure our experiment to randomize our treatments, but also to block against some undesired source of variation (not of interest)
  ⇒ How do we analyze our results?

  More generally: Our treatments (e.g. process A, B) are one factor
  Our blocks (e.g. tools 1-5) are another factor
  ⇒ How analyze experiments with 2 or more factors?

- Assumed Model:

  AVG BLOCK TREATMENT RESIDUAL

$$y_{ti} = \mu + \beta_i + \tau_t + \varepsilon_{ti}$$

With decomposition

$$y_{ti} = \bar{y} + (\bar{y}_i - \bar{y}) + (\bar{y}_t - \bar{y}) + (y_{ti} - \bar{y}_i - \bar{y}_t + \bar{y})$$

$$\vec{Y} = \vec{A} + \vec{B} + \vec{T} + \vec{R}$$ each with $N = nk$ elements

with corresponding sum of squares

$$S = S_A + S_B + S_T + S_R$$

and d.o.f. $$nk = 1 + (n-1) + (k-1) + (n-1)(k-1)$$

Results from randomized block design, general case

block \ treatment	1	2	---	t	---	k	block average
1	y_{11}	y_{21}	---	y_{t1}	---	y_{k1}	
2	y_{12}	y_{22}	---	y_{t2}	---	y_{k2}	
⋮	⋮	⋮		⋮		⋮	⋮
i	y_{1i}	y_{2i}	---	y_{ti}	---	y_{ki}	$\bar{y}_i$
⋮	⋮	⋮		⋮		⋮	⋮
n	y_{1n}	y_{2n}	---	y_{tn}	---	y_{kn}	
treatment average			---	$\bar{y}_t$	---		$\bar{y}$ = grand average

The corresponding ANOVA table is constructed:

Algebraic decomposition of sums of squares for the randomized block design, general formulas

source of variation	sum of squares	degrees of freedom
average (correction factor)	$S_A = nk\bar{y}^2$	1
between blocks	$S_B = k\sum_{i}^{n}(\bar{y}_i - \bar{y})^2$	$n-1$
between treatments	$S_T = n\sum_{i}^{k}(\bar{y}_i - \bar{y})^2$	$k-1$
residuals	$S_R = \sum_{t}^{k}\sum_{i}^{n}(y_{ti} - \bar{y}_i - \bar{y}_t + \bar{y})^2$	$(n-1)(k-1)$
total	$S = \sum_{t}^{k}\sum_{i}^{n} y_{ti}^2$	$N = nk$

With expected values $S_B^2 = \sigma^2 + k\sum_{i=1}^{k}\beta_i^2/(k-1)$

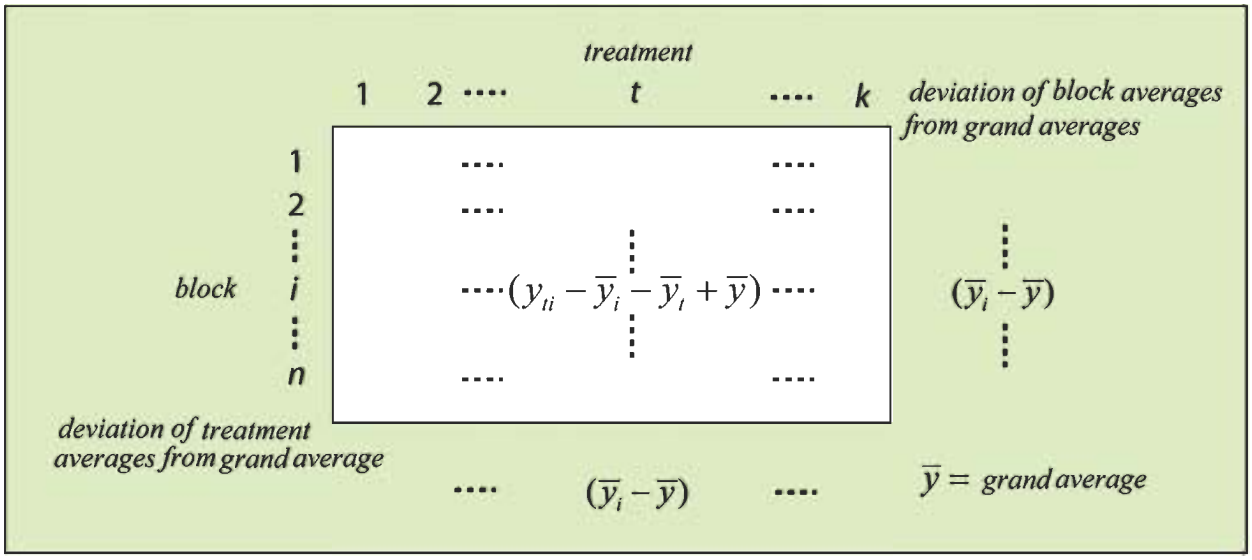
$S_T^2 = \sigma^2 + n\sum_{t=1}^{n}\tau_t^2/(n-1)$

So again, we can test the significance of observed "inflation" in block or treatment std. dev.

Assumptions to keep in mind! (1) IID, Normal residuals
(2) additivity of effects

- Deviation: or "corrected total SS" ANOVA: $S' = S'_A + S'_D$

Summary table: residuals and treatment and block deviations for a general randomized block design



Example: Particle Contamination

Two LPCVD tubes, three gas suppliers. Does supplier matter in average particle counts on wafers?

Experiment: 3 lots on each tube, for each gas; report average # particles added

Block (Tube)	A	B	C	
1	7	36	2	*15*
2	13	44	18	*25*
	10	*40*	*10*	*20*

Treatment (Gas)

Decompose according to the equation:

$$y_{ti} = \mu + \beta_i + \tau_t + \varepsilon_{ti}$$

$$y_{ti} = \bar{y} + (\bar{y}_i - \bar{y}) + (\bar{y}_t - \bar{y}) + (y_{ti} - \bar{y}_t - \bar{y}_i + \bar{y})$$

7	36	2		20	20	20		-5	-5	-5		-10	20	-10		2	1	-3
13	44	18		20	20	20		5	5	5		-10	20	-10		-2	-1	3

$S = S_A + S_B + S_T + S_R$

MANOVA for our Example

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	1350.0000	450.000	32.1429
Error	2	28.0000	14.000	Prob > F
C Total	5	1378.0000		0.0303

Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Block	1	1	150.0000	10.7143	0.0820
Treatment	2	2	1200.0000	42.8571	0.0228