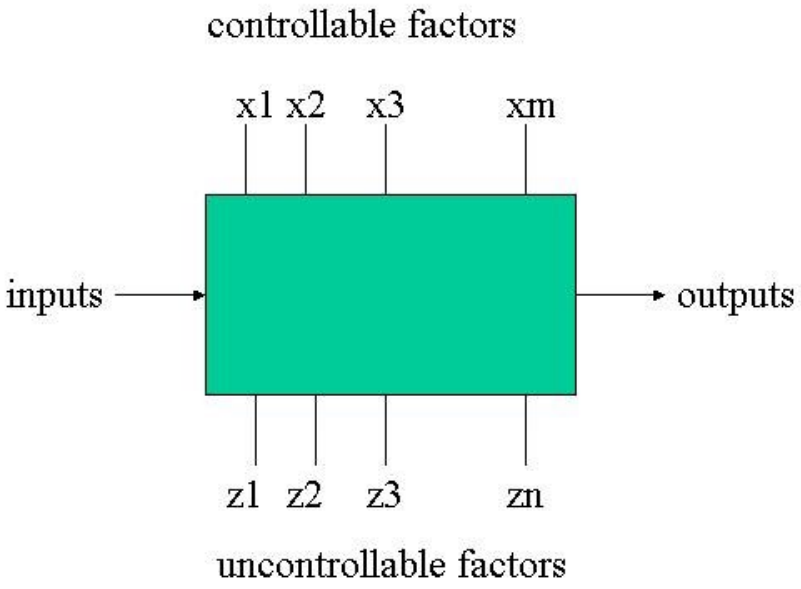
Purdue-NCKU program

Lecture 4

Design of Experiment
Single Factor Analysis

Dr. Qifan Song

Experimental study



"Controllable factor" means: you can direct tune its value (e.g., pressure in a manufacturing process) or you can select subjects such that the factor value can be tuned (e.g., sex in a medical study)

Observational Study vs Experimental Study

- Causality and Association
  - Response Y, controllable factors X, other factors Z
  - Observational studies observe $E(Y|X=1) - E(Y|X=0)$
  - Causality $\int_{\mathcal{Z}}[E(Y|X=1,Z=z) - E(Y|X=0,Z=z)]f_Z(z)dz$
- Goal of Experiment: Design the assignment of X to control the joint distribution of (X,Z)

Terminology

- Experimental factor (or variable): Controlled aspect of the experiment. One may choose not to control all controllable factors.

- Factor level: Specific value of factor.

- Treatment: A single factor level or combinations of two or more factors.

- Unit: "the smallest division of experimental material such that any two units may receive different treatments in the actual experimen" (Cox,1992)

- Experimental run (trial): *One* experiment which applies one treatment to one unit.

- Experimental error: Variation between repeated runs
  Source of experimental error: variation among units due to uncontrolled variables and background noise such as measurement error. (A better design can reduce experimental error)

Machine Tool Life Experiment

An engineer is interested in the effects of **cutting speed** (A), **tool geometry** (B) and **cutting angle** (C) on the lifespan (in hours) of a machine tool. Two levels of each factor are chosen (hence 8 possible treatments) and three replicates of for each treatment are run. The results:

Factor			Replicate		
A	B	C	I	II	III
−	−	−	22	31	25
+	−	−	32	43	29
−	+	−	35	34	50
+	+	−	55	47	46
−	−	+	44	45	38
+	−	+	40	37	36
−	+	+	60	50	54
+	+	+	39	41	47

Unit: The batch of raw material, which may produce multiple tools

Results: average lifespan for a batch of products

Fundamental Principles: Randomization

- The selection of unites, allocation of treatments to units, run order and (if possible) measurement order need to be randomized.
- Protect against all observable and non-observable latent variables
- Ensure the independence between experimental factors and other variables. Therefore, sub-populations corresponding to different treatments are almost identical.
- Ensure the indepdence between runs
- Ensure the validity of experimental error estimation.
- Ensure the validity of statistical inferences.

Permutation Test

Complete randomization makes possible to derive and perform *Two Sample Permutation Test*

An experiment was conducted by an amateur gardener whose object was to discover whether a change in the fertilizer mixture applied to his tomato plants would result in an improved yield. He had 11 plants set out in a single row; 5 were given the standard fertilizer mixture A, and the remaining 6 were fed a supposedly improved mixture B. The A's and B's were randomly applied to the positions in the row.

Pos	1	2	3	4	5	6	7	8	9	10	11
Trt	A	A	B	B	A	B	B	B	A	A	B
Yds	29.9	11.4	26.6	23.7	25.3	28.5	14.2	17.9	16.5	21.1	24.3

Mean difference (modified minus standard)= $\bar{y_B} - \bar{y_A} = 1.69$

Hypothesis Test on the effect

H_0: the modified fertilizer does not improve the (mean) yield.
H_a: the modified fertilizer improves the (mean) yield.

Under the null hypothesis, A and B are mere labels and should not affect the yield. For example, the first plant would yield 29.9 pounds of tomatoes no matter it had been labeled as A or B (or fed A or B).

There are $\frac{11!}{5!6!} = 462$ ways of allocating 5 A's and 6 B's to the 11 plants, any one of which could equally be chosen. The used design is just one of 462 equally likely possibilities. (why?)

For example:

Pos	1	2	3	4	5	6	7	8	9	10	11
Yds	29.9	11.4	26.6	23.7	25.3	28.5	14.2	17.9	16.5	21.1	24.3
LL1	A	A	A	A	A	B	B	B	B	B	B
LL2	A	A	A	A	B	A	B	B	B	B	B
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
LL462	B	B	B	B	B	B	A	A	A	A	A

LL1, LL2, etc are equally likely.

LL1: mean difference between B and A is -2.96

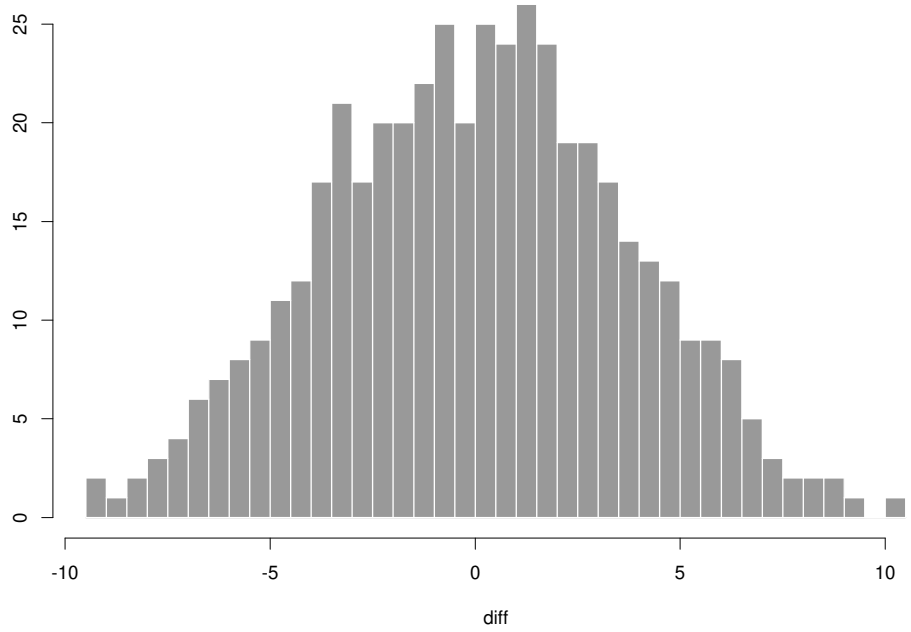
LL2: mean difference between B and A is -4.14

⋮

LL462: mean difference between B and A is 5.43

Under the null hypothesis, these differences are equally likely.

Randomization Distribution (Histogram) of the Mean Differences



Observed Diff $= 1.69$

P-value $= Pr(\text{Diff} \geq 1.69 \mid \text{randomization}) = \frac{155}{462} = .335$

Because P-value $\geq \alpha$, fail to H_0.

Principles: Replication, Blocking, Double Blind

- Replication
  - Each treatment is applied to a number of units representative of the population (of units)
  - Enable the estimation of experimental error. This estimation will be used to assess the accuracy of inferences.
- Blocking
  - Balanced treatment assignment w.r.t controllable nuisance factor (block factor); reduce experimental error
  - Restricted randomization: total randomization within block and the block factor may not independent to uncontrolled factors
- Double Blind
  - Remove subject biased and placebo effect

Single Factor Analysis

- Interested in comparing several treatments, i.e., one factor with several levels
- Could do numerous two-sample t-tests; we want to test equality of all treatments simultaneously.
- Statistical Model:

$$y_{ij} = \mu_i + \epsilon_{ij} = \mu + \tau_i + \epsilon_{ij} \quad \begin{cases} i = 1, 2 \ldots a \\ j = 1, 2, \ldots n_i \end{cases}$$

  μ - grand mean; τ_i - ith treatment effect; μ_i - ith treatment mean; ϵ_{ij} - error term

  Constraint: $\sum_{i=1}^{a} \tau_i = 0$.

- Completely Randomized Design ensures that except receiving different treatements, all units are statistically equivalent and indepdendent, i.e., ϵ_{ij} are i.i.d. We further assume that $\epsilon_{ij} \sim N(0, \sigma^2)$.

Partitioning the Sum of Squares

- Basic Hypotheses:
  $H_0 : \tau_1 = \tau_2 = \ldots = \tau_a = 0$ vs $H_1 : \tau_i \neq 0$ for at least one i
- Notation
  - $y_{i.} = \sum_{j=1}^{n_i} y_{ij} \rightarrow \overline{y}_{i.} = y_{i.}/n_i$ (treatment sample mean, or row mean)
  - $y_{..} = \sum\sum y_{ij} \rightarrow \overline{y}_{..} = y_{..}/N$ (grand sample mean)
  - $\widehat{\epsilon}_{ij} = y_{ij} - \overline{y}_{i.}$
- Decomposition of y_{ij}: $y_{ij} = \overline{y}_{..} + (\overline{y}_{i.} - \overline{y}_{..}) + (y_{ij} - \overline{y}_{i.})$
- Can show

$$\begin{aligned} \textstyle\sum_i \sum_j (y_{ij} - \overline{y}_{..})^2 &= \textstyle\sum_i n_i (\overline{y}_{i.} - \overline{y}_{..})^2 &&+ \textstyle\sum_i \sum_j (y_{ij} - \overline{y}_{i.})^2 \\ \text{Total SS} &= \text{Treatment SS} &&+ \text{Error SS} \\ \text{Total Variation} &= \text{Variation between} &&+ \text{Variation within} \end{aligned}$$

$\text{SS}_\text{T} = \text{SS}_\text{Treatments} + \text{SS}_\text{E}$

Test Statistic

$$F = \frac{\text{SS}_{\text{Treatments}}/(a-1)}{\text{SS}_{\text{E}}/(N-a)} = \frac{\text{MS}_{\text{Treatments}}}{\text{MS}_{\text{E}}}$$

- Under null hypothesis, both $\overline{y}_{i.}$ and $\overline{y}_{..}$ are consistent for μ, i.e., $\text{SS}_{\text{Treatments}}$ will be small.
- F_0 tends to small under null, and tends to be large under alternative.
- $\mathcal{C} = (F_{\text{critical}}, \infty)$
- What is the distribution F_0 under null?
- Under null, SS_{T}, $\text{SS}_{\text{Treatments}}$ and $\text{SS}_{\text{Treatments}}$ follows $\sigma^2\chi^2$ distribution with d.f. $N-1$, $a-1$ and $N-a$ respectively. $\text{SS}_{\text{Treatments}}$ and $\text{SS}_{\text{Treatments}}$ are independent. Then, F follows $F_{a-1,N-a}$ distribution.

Analysis of Variance (ANOVA) Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Between	$SS_{Treatment}$	$a-1$	$MS_{Treatment}$	F_0
Within	SS_E	$N-a$	MS_E	
Total	SS_T	$N-1$		

Alternative way of computing

$$SS_T = \sum\sum y_{ij}^2 - y_{..}^2/N; \quad SS_{Treatment} = \sum \frac{y_{i.}^2}{n_i} - y_{..}^2/N$$

SS_E=SS_T - $SS_{Treatment}$

- **Decision Rule**: If $F_0 > F_{a-1,N-a,1-\alpha}$ or p-value $= Pr(F_{a-1,N-a} > F_0) < \alpha$, then reject H_0
- When $a = 2$, F-test is equivalent to 2-sample t test.

Model Diagnose

- Model Assumptions

  1 Assumption on means is correct

  2 Independent observations

  3 Errors normally distributed

  4 Constant variance

$$
\begin{array}{rcccc}
y_{ij} & = & \mu_i & + & \epsilon_{ij} \\
y_{ij} & = & \overline{y}_{i.} & + & \widehat{\epsilon}_{ij} \\
\text{observed} & = & \text{predicted} & + & \text{residual}
\end{array}
$$

- View residuals as observable surrogate of ϵ's
- Diagnostics use predicted responses and residuals.

Diagnostic Method

- Normality
  - Histogram of residuals
  - Normal probability plot / QQ plot of residuals
  - Formal Tests: e.g., Shapiro-Wilk Test
  - Minor deviation from normality is acceptable
- Constant Variance
  - Plot $\widehat{\epsilon}_{ij}$ vs $\widehat{y}_{ij}$ (residual plot)
  - Hartley Test
  - Non-constant variance leads to incorrect MSE
- Independence
  - Plot $\widehat{\epsilon}_{ij}$ vs time/space
  - Plot $\widehat{\epsilon}_{ij}$ vs variable of interest

Normal Probability Plot

$Y_1, Y_2, \ldots, Y_n$ is a random sample from a population with mean μ and variance σ^2.

Order Statistics: $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ where $Y_{(i)}$ is the ith smallest value.

if the population is normal, i.e., $N(\mu, \sigma^2)$, then

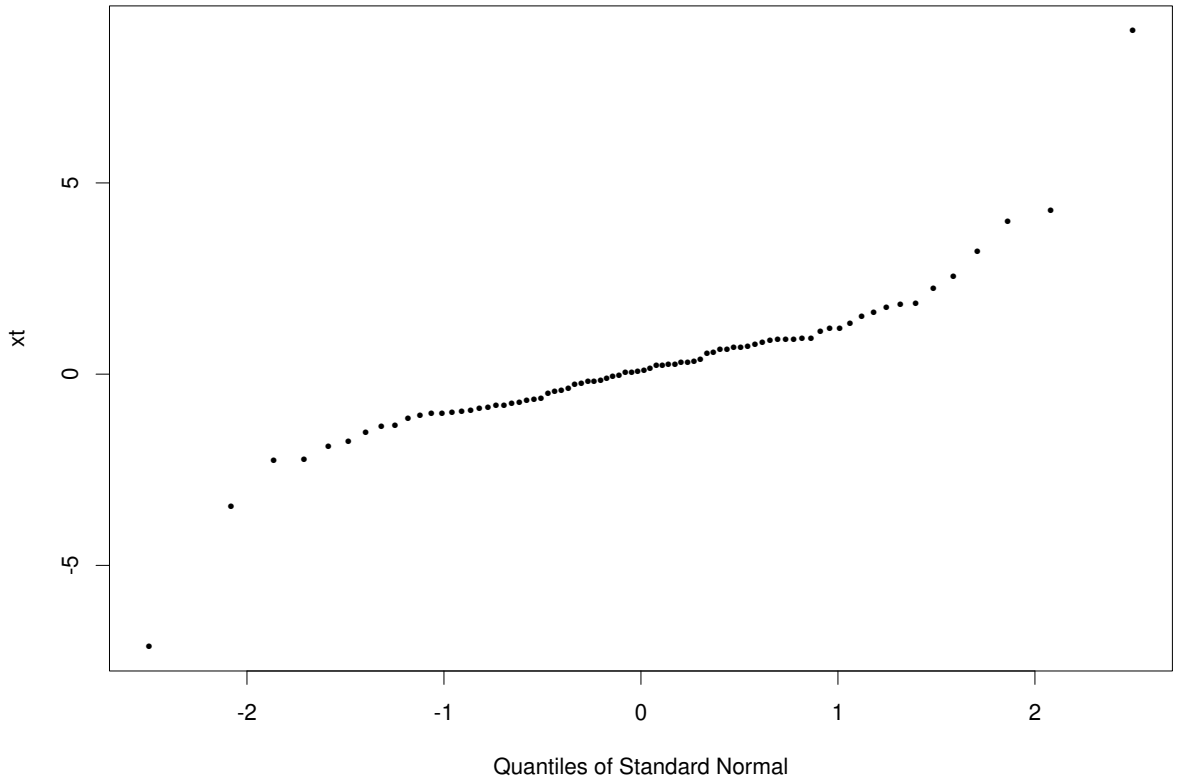
$E(Y_{(i)}) \approx \mu + \sigma Z_{\alpha_i}$ with $\alpha_i = \frac{i-3/8}{n+1/4}$ for $1 \leq i \leq n$.

Given a sample $y_1, y_2, \ldots, y_n$, the plot of $(Z_{\alpha_i}, y_{(i)})$ is called the normal probability plot or QQ plot.

the points falling around a straight line indicate normality of the population; Deviation from a straight line pattern indicates non-normality (the pen rule)

QQ plot

Don't just focus on the middle portion of the QQ plot Below is a *bad* QQ plot



This QQ plot corresponds to a heavier tailed distribution than Normal distribution (e.g. t distribution)

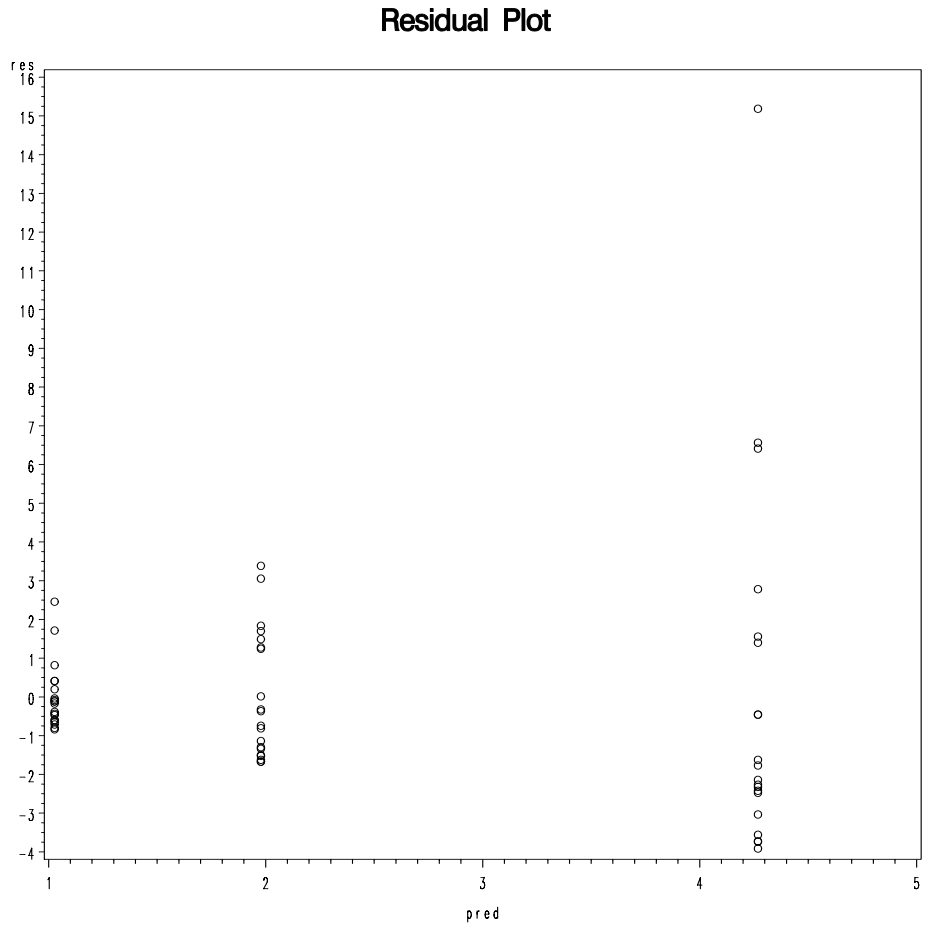
Formal Test

- Plots are usually enough for identifying gross violations of assumptions (since inferences are quite robust)
- Shapiro-Wilk test: a normality test based on the correlation between the residuals and their expected value.

Plots vs Tests: Test results are very dependent on n. With a large enough sample size, a good formal test are likely to reject null and claim violation, even if the deviation is slight. This is unnecessary, since most of the inferences are robust.

- Plots for large-sample data set
- Tests for small-sample data set, where it is difficult to make a judgment call

Residual Plot



- Similar information as $\widehat{\epsilon}_{ij}$ vs i
- Never plot $\widehat{\epsilon}_{ij}$ vs y_{ij}

Homogeneity of Variance: Hartley Test

- Hartley statistic,

$$H = \frac{\max(s_i^2)}{\min(s_i^2)}$$

- Under H_0, H's distribution only depends on a and n_i's. The rejection region can be determined by simulations
- For balanced case (i.e., $n_i = n$), statisticians have created the table of critical values.

Hartley Test is sensitive to normality, hence more robust tests are developed as well.

Non-constant Variance: Impact and Remedy

Usually, when non-constant variance occurs, the variances (σ_i^2's) depend on treatment means (μ_i's), i.e. $\sigma_i^2 = g(\mu_i)$.

- Does not affect F-test dramatically when experiment is balanced
- Why concern?
- Lead to unreliable confidence intervals (to be discuss later).
- Variance-Stabilizing Transformations
  - Transform data y_{ij} to $f(y_{ij})$, e.g. y_{ij} to $\sqrt{y_{ij}}$, with the hope that the transformed data $f(y_{ij})$ do not violate the constant variance assumption.
  - f is called a variance-stabilizing transformation; $\sqrt{y}$, $\log(y)$, $1/y$, $\arcsin(\sqrt{y})$, and $1/\sqrt{y}$ are some commonly used transformations.
  - Transformations are also used as remedies for nonnormality

Ideas for Finding Proper Transformations

- Denote $\tilde{Y} = f(Y)$; What is the mean and variance of $\tilde{Y}$?
- Approximate $f(Y)$ by a linear function (Delta Method):

$$f(Y) \approx f(\mu) + (Y - \mu) f'(\mu)$$

Mean $\tilde{\mu} = \mathrm{E}(\tilde{Y}) = \mathrm{E}(f(Y)) \approx \mathrm{E}(f(\mu)) + \mathrm{E}((Y - \mu) f'(\mu)) = f(\mu)$

Variance $\tilde{\sigma}^2 = \mathrm{Var}(\tilde{Y}) \approx [f'(\mu)]^2 \mathrm{Var}(Y) = [f'(\mu)]^2 \sigma^2 = [f'(\mu)]^2 g(\mu)$

- Need to choose f such that $[f'(\mu)]^2 g(\mu) =$ constant
- When $g(\mu)$ is known, f can be derived explicitly.

Examples (c is some unknown constant)

$g(\mu) = c\mu$	(Poisson)	$f(Y) = \int \frac{1}{\sqrt{\mu}} d\mu \to f(Y) = \sqrt{Y}$
$g(\mu) = c\mu(1-\mu)$	(Binomial)	$f(Y) = \int \frac{1}{\sqrt{\mu(1-\mu)}} d\mu \to f(Y) = \arcsin(\sqrt{Y})$
$g(\mu) = c\mu^{2\beta} (\beta \neq 1)$	(Box-Cox)	$f(Y) = \int \mu^{-\beta} d\mu \to f(Y) = Y^{1-\beta}$
$g(\mu) = c\mu^2$	(Box-Cox)	$f(Y) = \int \frac{1}{\mu} d\mu \to f(Y) = \log Y$

Box-Cox Transformations

- Assume $\sigma^2 = c\mu^{2\beta}$, then the variance-stabilizing transform should be

$$f(Y) = \begin{cases} Y^{1-\beta} & \beta \neq 1; \\ \log\ Y & \beta = 1 \end{cases}$$

  These transformations are referred to as Box-Cox transformations.

  Clearly it is crucial to know what β is.

  As a matter of fact, β can be regarded as a parameter, and it can be estimated (identified) from data.

Approximate Box-Cox Transformations:

- From the assumption $\sigma^2 = c\mu^{2\beta}$, we have

$$\sigma_i^2 = c\mu_i^{2\beta} \text{ for treatments } i = 1, 2, \ldots, a.$$

  Take logarithm of both sides,

$$\log\sigma_i = \frac{1}{2}\log c + \beta\log\mu_i$$

- Let s_i and $\bar{y}_{i.}$ be the sample standard deviations and means. Because $\widehat{\sigma}_i = s_i$ and $\widehat{\mu}_i = \bar{y}_{i.}$, **approximately**,

$$\log s_i = \text{ constant } + \beta\log\bar{y}_{i.},$$

  where $i = 1, \ldots, a$.

- We can plot $\log s_i$ against $\log\bar{y}_{i.}$, fit a straight line and use the slope to estimate β (i.e., simple linear regression, Lecture 7).

Multiple Comparison

A significant F test (i.e., rejection of null) only asserts the existence of difference between μ_i's, but doesn't tell us where is the difference. We still need to perform individual tests for possible differences

- A contrast is defined as $\Gamma = \sum_{i=1}^{a} c_i \mu_i$ is a contrast if $\sum_{i=1}^{a} c_i = 0$.

  Equivalently, $\Gamma = \sum_{i=1}^{a} c_i \tau_i$.

- Examples

  1. $\Gamma_1 = \mu_1 - \mu_2 = \mu_1 - \mu_2 + 0\mu_3 + 0\mu_4$,
     $c_1 = 1, c_2 = -1, c_3 = 0, c_4 = 0$
     Comparing μ_1 and μ_2.

  2. $\Gamma_2 = \mu_1 - 0.5\mu_2 - 0.5\mu_3 = \mu_1 - 0.5\mu_2 - 0.5\mu_3 + 0\mu_4$
     $c_1 = 1, c_2 = -0.5, c_3 = -0.5, c_4 = 0$
     Comparing μ_1 and the average of μ_2 and μ_3.

Inference for a single contrast

- Estimation of Γ: $C = \sum_{i=1}^{a} c_i \bar{y}_{i.}$
- Test $H_0 : \Gamma = 0$

$$t = \frac{C}{\sqrt{\text{MSE} \sum \frac{c_i^2}{n_i}}} \sim t(N-a), \text{or}$$

$$F = t^2 = \frac{(\sum c_i \bar{y}_{i.})^2}{\text{MSE} \sum \frac{c_i^2}{n_i}} = \frac{(\sum c_i \bar{y}_{i.})^2 / \sum c_i^2 / n_i}{\text{MSE}}$$

Under H_0, $t \sim t_{N-a}, F \sim F_{1,N-a}$, due to the independence between $\bar{y}_{i.}$'s and MSE.

- Rejection region: $|t| > t_{N-a,1-\alpha/2}$ or $F > F_{1,N-a,1-\alpha}$
- C.I., based on the same pivotal quantity

$$\sum c_i \bar{y}_{i.} \pm t_{N-a,1-\alpha/2} \sqrt{MS_E \sum \frac{c_i^2}{n_i}}$$

Inference for multiple contrasts

- If each inference task has α probability of failure (type I error for test, or C.I. fail to cover true parameter), then this failure probability can accumulate when performing multiple inference tasks
- Simultaneously inferences requires: the probability of failing at least one inference task is no more than α
- $Pr(\{\text{fail at least one task}\}) = Pr(\cup_{i=1}^{m}\{\text{fail the } i\text{th task}\}) \leq \sum_i Pr(\{\text{fail the } i\text{th task}\})$
- Bonferroni Correction: for each individual task, we set $\alpha' = \alpha/m$. This ensure the overall error rate is not bigger than α
- Bonferroni Contrast C.I.: $\sum c_i \bar{y}_{i\cdot} \pm t_{N-a,1-\alpha/(2m)}\sqrt{MS_E \sum \frac{c_i^2}{n_i}}$ for m contrasts.
- *universal* but conservative correction

Scheffe's Method for Testing All Contrasts

- Consider all possible contrasts: $\Gamma = \sum c_i \mu_i$

  Estimate: $C = \sum c_i \bar{y}_{i.}$, St. Error: S.E.$_C = \sqrt{\text{MS}_E \sum \frac{c_i^2}{n_i}}$

- Critical value: $\sqrt{(a-1)F_{\alpha,a-1,N-a}}$

- Scheffe's simultaneous CI: $C \pm \sqrt{(a-1)F_{\alpha,a-1,N-a}}$ S.E.$_C$

- Scheffe's simultaneous Test: C/S.E.$_C$ vs $\sqrt{(a-1)F_{\alpha,a-1,N-a}}$

- Overall error rate for all possible (infinity many) contrasts

$$P(\text{at least one type I error}) \leq \alpha$$

- Comparison between Scheffe and Bonferroni

Chapter Review

- DoE
- Principle of DoE
- One factor analysis and ANOVA
- Contrast testing