Part VII

Study Designs

Chapter 26

Design of Experiments, Randomization, and Sample Size Planning

After a brief general discussion about the design of experiments and randomization, we spend the majority of the time describing various procedures used to determine sample sizes that ensure "good" tests and confidence intervals.

26.1 Design of Experiments

Up to this point, we have looked at *completely randomized designs*. In this last part of the course, we look at other designs. Design of experiments involve

- identifying treatments,
- identifying experimental units,
- devising the rules that assign treatments to experimental units (or vis–versa),
- making measurements made on experimental units after the treatments have been applied.

We will spend most of the rest of the semester looking at the third item: clever ways of assigning treatments to experimental units so that efficient use is made of the experimental units.

26.2 Contributions of Statistics to Experimentation

Statistics has made the following contributions to experimentation,

- factorial experiments,
- replication,
- randomization,
- local control (blocking, stratification).

We will look out for two types of bias,

- *selection bias* which is minimized using the *random* assignment of treatments to experimental units.
- *measurement bias* which is minimized using either *single–blind* or *double–blind* studies.

26.3 Randomization Tests

Randomization can also be used to created so–called *randomization tests*, which require fewer assumptions about the distribution of the error terms in ANOVA models. This idea is not covered in this course.

26.4 Planning of Sample Sizes with Power Approach

A "good" test is one which has large *power*, $1 - \beta$. Consequently, it makes sense to choose a sample size for a test such that the resulting power of the test is large. We look at three cases of using power as a way of deciding what sample size to choose for a test.

Exercise 26.1 (Planning of Sample Sizes with Power Approach)

1. *Reviewing power and hypothesis testing: average student height.*
   We may be interested in collecting a sample of thirty students and using their average height, $\overline{Y}$, to *test* whether the true average height, μ, of *all* students at Purdue University North Central, is either less than or equal to 5.6 feet tall or greater than 5.6 feet tall.

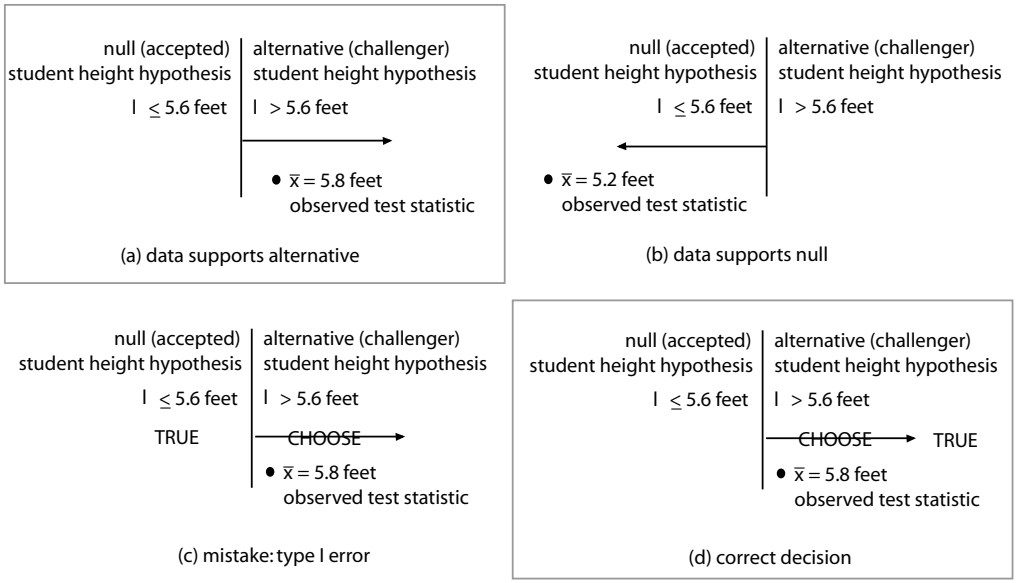

Figure 26.1 (Simple Hypothesis Testing Reviewed)

(a) *Null and alternative hypotheses*
Although you could test other alternatives, you have decided to compare only two alternatives. These alternatives are (circle one)

i. $H_o : \mu \leq 5.6$ and $H_a : \mu \geq 5.6$
ii. $H_o : \mu = 5.6$ and $H_a : \mu > 5.6$
iii. $H_o : \mu \leq 5.6$ and $H_a : \mu > 5.6$
iv. $H_o : \mu < 5.6$ and $H_a : \mu < 5.6$

(b) *Choosing between null and alternative hypotheses using observed samples*
In order to test $H_o : \mu \leq 5.6$ versus $H_a : \mu > 5.6$, you could collect a random sample of 30 students from PU/NC and use their average height, $\overline{Y}$, to decide which alternative to choose. For example, if $\bar{y} = 5.8$, it would seem fairly obvious to choose the alternative hypothesis $H_a : \mu > 5.6$; in other words, if the average height of the 30 students was 5.8 feet tall, this would seem to indicate the average height of all students at PU/NC would be greater than 5.6 feet tall. In a similar way, if $\bar{y} = 5.2$, it seems fairly obvious to choose the (circle one)

i. $H_o : \mu \leq 5.6$
ii. $H_a : \mu > 5.6$

(c) *Mistaken choices in hypothesis testing: α and β*
It is possible to make a mistake when deciding between the null and alternative hypotheses. For instance, it is possible that although the average height of the 30 students chosen, $\bar{y} = 5.8$, say, indicates the alternative hypothesis, $H_a : \mu > 5.6$, would be chosen, when, in fact, the true average

height of all students at PU/NC was less than or equal to 5.6 feet tall. The chance of this type of an error, of *mistakenly rejecting the null*, is called a *type I* error and denoted α. A chance of a type II error, denoted β, on the other hand, occurs when (check none, one or more)

i. it is decided to reject the alternative when, in fact, the alternative is true
ii. it is decided the average height of all students at PU/NC is less than 5.6 feet tall, when, in fact, it is greater than 5.6 feet tall
iii. mistakenly rejecting the alternative
iv. mistakenly accepting the null

(d) *Correct choices in hypothesis testing:* $1-\alpha$ *and* $1-\beta$
It is possible to *correctly* decide between the null and alternative hypotheses. For instance, it is possible that because the average height of the 30 students chosen, $\bar{y} = 5.2$, say, is observed, this leads to correctly choosing the null hypothesis, $H_o : \mu \leq 5.6$. The chance of this type of an error, of *correctly accepting the null*, is denoted $1-\alpha$. A chance of *correctly accepting the alternative*, called *power* and denoted $1-\beta$, on the other hand, occurs when (check none, one or more)

i. it is decided to accept the alternative when, in fact, the alternative is true
ii. it is decided the average height of all students at PU/NC is greater than 5.6 feet tall, when, in fact, it is greater than 5.6 feet tall
iii. correctly rejecting the null
iv. correctly accepting the alternative

(e) *Four probabilities:* α, $1-\alpha$, β, $1-\beta$
In this (or *any*) test, there are four associated probabilities,

actual $\downarrow$ observed $\rightarrow$	$H_o : \mu \leq 5.6$	$H_a : \mu > 5.6$
$H_o : \mu \leq 5.6$	$1-\alpha$ (correct)	α (mistake)
$H_a : \mu > 5.6$	β (mistake)	$1-\beta$ (correct)

A test is a "good" one if the associated power, $1-\beta$, is
(choose one) **small** / **large**.
Consequently, we would like to choose a sample size for the test that makes the power as large as possible.

2. *First procedure: power of test, single factor ANOVA, Table B.11.*
One way to measure how "good" the test is, is to calculate the power, $1-\beta$, of a test (the chance the alternative is correct if, in fact, the alternative is

correct). In the single factor ANOVA case, knowing (or assuming) the following information[1]; in particular, the sample sizes n_i,

$$\phi = \frac{1}{\sigma}\sqrt{\frac{\sum n_i(\mu_i - \mu_\cdot)^2}{r}}$$
$$\mu_\cdot = \frac{\sum n_i \mu_i}{n_T}$$
$$\nu_1 = r - 1,\ \nu_2 = n_T - r,\ \alpha,\ \sigma$$

the power of a test, $1 - \beta$, can be found in Table B.11, page 1356 of the Neter et al. text.

(a) *A first example: patients*
Twelve different patients are subjected to three drugs.

drug 1	5.90	5.92	5.91	5.89	5.88	$\bar{Y}_1 = 5.9$
drug 2	5.50	5.50				$\bar{Y}_2 = 5.5$
drug 3	5.01	5.00	4.99	4.98	5.02	$\bar{Y}_3 = 5.0$

Determine the power of a test which tests if at least two of the three average patient responses to the drug are different at $\alpha = 0.05$, where $\mu_1 = 5.90$, $\mu_2 = 5.50$, $\mu_3 = 5.00$, and $\sigma = 0.5$ and where $n_1 = 5$, $n_2 = 2$ and $n_3 = 5$.

In this case, the sample sizes are $n_1 = 5$, $n_2 = 2$ and $n_3 = 5$
and assume $\mu_1 = 5.90$, $\mu_2 = 5.50$, $\mu_3 = 5.00$,
and so $\mu_\cdot = \frac{\sum n_i \mu_i}{n_T} = \frac{5(5.9)+2(5.5)+5(5)}{12} = \frac{131}{24}$,
and so

$$\begin{aligned}\phi &= \frac{1}{\sigma}\sqrt{\frac{\sum n_i(\mu_i - \mu_\cdot)^2}{r}} \\ &= \frac{1}{0.5}\sqrt{\frac{5(5.9 - 131/24)^2 + 2(5.5 - 131/24)^2 + 5(5 - 131/24)^2}{3}} \\ &\approx 1.645 \approx 1.5\end{aligned}$$

$\nu_1 = r - 1 = 3 - 1 = 2$
$\nu_2 = n_T - r = 12 - 3 = 9$
$\alpha = 0.05$
and so, using Table B.11, page 1356,
$1 - \beta =$ (choose one) **0.49** / **0.74** / **0.91**
So, if the sample sizes are $n_1 = 5$, $n_2 = 2$ and $n_3 = 5$,
then the power[2] of the test is $1 - \beta = 0.49$.

[1]The parameter ϕ, by the way, is the *noncentrality parameter* and is a measure of how unequal are the treatments.

[2]The power means, in this case, there is a 49% chance that the mean differences will be detected by this test, when, in fact, the means are different in the specified way.

(b) *Another example*
Do the same question as before, only let $n_1 = 10$, $n_2 = 7$ and $n_3 = 10$. Notice that the samples sizes are larger in this case, than before.

In this case $\mu_\cdot = \frac{\sum n_i \mu_i}{n_T} = \frac{10(5.9)+7(5.5)+10(5)}{27} = \frac{295}{54}$,
and so

$$\begin{aligned} \phi &= \frac{1}{\sigma}\sqrt{\frac{\sum n_i(\mu_i - \mu_\cdot)^2}{r}} \\ &= \frac{1}{0.5}\sqrt{\frac{10(295/54 - 5.9)^2 + 7(295/54 - 5.5)^2 + 10(295/54 - 5)^2}{3}} \\ &\approx 2.3275 \approx 2.5 \end{aligned}$$

$\nu_1 = r - 1 = 3 - 1 = 2$
$\nu_2 = n_T - r = 27 - 3 = 24$
$\alpha = 0.05$
and so, using Table B.11, page 1356,
$1 - \beta =$ (choose one) **0.49** / **0.74** / **0.96**
So, the power of the test where the
sample sizes are $n_1 = 10$, $n_2 = 7$ and $n_3 = 10$,
is (choose one) **smaller** / **larger**
than the power of the test where the
sample sizes are $n_1 = 5$, $n_2 = 2$ and $n_3 = 5$.
That is, the power increases when the sample size increases.

(c) *And another example*
Do the same question as the first one, only let $\mu_1 = 5$, $\mu_2 = 6$, $\mu_3 = 7$. Notice that the treatment means are closer together (the treatment mean differences are larger) in this case, than before.

In this case, $\mu_\cdot = \frac{\sum n_i \mu_i}{n_T} = \frac{5(5)+2(6)+5(7)}{12} = 6$,

$$\begin{aligned} \phi &= \frac{1}{\sigma}\sqrt{\frac{\sum n_i(\mu_i - \mu_\cdot)^2}{r}} \\ &= \frac{1}{0.5}\sqrt{\frac{5(6-5)^2 + 2(6-6)^2 + 5(6-7)^2}{3}} \\ &\approx 3.65 \approx 3.5 \end{aligned}$$

$\nu_1 = r - 1 = 3 - 1 = 2$
$\nu_2 = n_T - r = 12 - 3 = 9$
$\alpha = 0.05$
and so, using Table B.11, page 1356,
$1 - \beta =$ (choose one) **0.24** / **0.49** / **1.00**

So, the power of the test where the
sample sizes are $\mu_1 = 5.9$, $\mu_2 = 5.5$ and $\mu_3 = 5.0$,
is (choose one) **smaller / larger**
than the power of the test where the
sample sizes are $\mu_1 = 5$, $\mu_2 = 6$ and $\mu_3 = 7$.
The power increases when the treatment mean differences increase.

3. *Second procedure: power of test, two factor ANOVA, Table B.11.*
Consider the effect of air temperature *and* noise on the ROC of deer mice.

	noise →	low	medium	high
temperature	0^o F	10.3, 7.2	9.1, 5.4	6.1, 2.1
	10^o F	1.8, 9.8	12.1, 4.2	5.1, 6.2
	20^o F	1.2, 8.1	6.5, 4.1	1.2, 2.1
	30^o F	12.4, 15.1	16.1, 17.2	18.1, 19.1

(a) *Factor A effects*
Determine the power of a test which tests if at least two of the three average mice ROCs to temperature (A main effects) are different at $\alpha = 0.05$, where $\alpha_1 = 4$, $\alpha_2 = -5$, $\alpha_3 = -7$, $\alpha_4 = 8$ and where $\sigma = 7$ and $n = 2$, $i = 1, 2, 3$.

In this case,

$$\begin{aligned} \phi &= \frac{1}{\sigma}\sqrt{\frac{\sum nb\alpha_i^2}{a}} \\ &= \frac{1}{7}\sqrt{\frac{2(3)[4^2 + (-5)^2 + (-7)^2 + 8^2]}{4}} \\ &\approx 2.17 \approx 2 \end{aligned}$$

$\nu_1 = a - 1 = 4 - 1 = 3$
$\nu_2 = ab(n-1) = 4(3)(2-1) = 12$
$\alpha = 0.05$
and so, using Table B.11, page 1356,
$1 - \beta =$ (choose one) **0.35 / 0.82 / 0.91**
So, if $n = 2$, then the power is $1 - \beta = 0.82$.

(b) *Factor B effects*
Determine the power of a test which tests if at least two of the three average mice ROCs to noise (B main effects) are different at $\alpha = 0.05$, where $\beta_1 = 1$, $\beta_2 = -3$, $\beta_3 = 2$ and where $\sigma = 7$ and $n = 2$, $i = 1, 2, 3$.

In this case,

$$\begin{aligned}\phi &= \frac{1}{\sigma}\sqrt{\frac{\sum na\alpha_i^2}{b}} \\ &= \frac{1}{7}\sqrt{\frac{2(4)[1^2+(-3)^2+(2)^2]}{3}} \\ &\approx 0.87 \approx 1\end{aligned}$$

$\nu_1 = b - 1 = 3 - 1 = 2$
$\nu_2 = ab(n-1) = 4(3)(2-1) = 12$
$\alpha = 0.05$
and so, using Table B.11, page 1356,
$1 - \beta =$ (choose one) **0.26** / **0.82** / **0.91**
So, if $n = 2$, then the power is $1 - \beta = 0.26$.

4. *Third procedure: power and treatment difference, single factor, Table B.12.* The sample size can be calculated to meet a required power $1-\beta$ and *detection*, Δ, of a test[3]. Knowing (or assuming) the following information; in particular, the power $1 - \beta$,

$$\begin{gathered}\Delta = \max(\mu_i) - \min(\mu_i) \\ \text{number of treatments, } r \\ 1 - \beta,\ \alpha,\ \sigma\end{gathered}$$

the sample size, n, can be found in Table B.12, page 1361.

   (a) Determine the sample sizes, if the number of treatments is $r = 10$, $\alpha = 0.20$, $\beta = 0.05$ (and so $1 - \beta = 0.95$), $\sigma = 10$, and the required detectable difference between the treatments means, Δ, is as given in the table below, use Table B.12, page 1361, to fill in the blanks,

Δ	10	15	20	30
$\frac{\Delta}{\sigma}$	1.0	1.5	2.0	3.0
n (Table B.12)	34	16	(a)	(b)

   (a) = (choose one) **3** / **5** / **9**
   (b) = (choose one) **3** / **5** / **9**

   (b) As in (a), only $\alpha = 0.05$,

Δ	10	15	20	30
$\frac{\Delta}{\sigma}$	1.0	1.5	2.0	3.0
n (Table B.12)	(a)	(b)	(c)	(d)

[3] In the first two procedures, where we determined the power of the test for a given sample size; here, we determine the sample size for a given power.

(a) = (choose one) **30** / **48** / **57**
(b) = (choose one) **22** / **35** / **49**
(c) = (choose one) **13** / **15** / **19**
(d) = (choose one) **3** / **5** / **7**

(c) As the required detectable difference between means increases, the required samples size (choose one) **decreases** / **increases**

26.5 Planning of Sample Sizes with Estimation Approach

Consider a study with,
g contrasts, L_i, $i = 1, \ldots, g$,
each with variance $\sigma^2\{\hat{L}\} = \frac{\sigma^2}{n}\sum c_i^2$
and Bonferroni[4] critical value $t(1 - \frac{\alpha}{2g}; n_T - r)$

then the sample size n is chosen so that the *width* of the corresponding confidence interval,

$$\sigma\{\hat{L}\}t(1 - \frac{\alpha}{2g}; n_T - r)$$

equals some (preset) required amount. This is an iterative procedure where n is chosen so that the preset width is approached more and more closely.

Exercise 26.2 (Planning of Sample Sizes with Estimation Approach)

1. *Example: sample size when CI width* $= \pm 1.0$*?*
   Determine the sample size of the confidence for the contrast

   $$L_1 = \mu_3 - \mu_4$$

   if this is one of seven contrasts, $g = 7$, and $\sigma = 0.5$, $\alpha = 0.10$, $n_T = 5 \times 10 = 50$ (initial guess), there are five treatments, $r = 5$, and width of the confidence interval is equal ± 1.0.

   Using Bonferroni, $t(1 - \frac{\alpha}{2g}; n_T - r) = t(1 - \frac{0.10}{2(7)}; 50 - 5) = t(0.9928; 45) \approx 2.54573$
   and since $\sigma = 0.5$,
   and so, since $L_1 = \mu_3 - \mu_4$, and *assuming* $n = 20$,
   $\sigma\{\hat{L}_1\} = \sqrt{\frac{\sigma^2}{n}\sum c_i^2} = \sqrt{\frac{0.5^2}{20}((-1)^2 + (1)^2)} \approx 0.158$
   trying sample sizes $n = 20, 10, 3$,

[4]Scheffe, Tukey or others could be used here.

contrast	$n = 20$	$n = 10$	$n = 3$
$\mu_3 - \mu_4$	0.158	0.224	0.408
width of CI	$\pm(0.158)(2.545) = \pm 0.402$	(a)	(b)

(a) = (choose one) **±0.403** / **±0.570** / **±1.03**
(b) = (choose one) **±0.403** / **±0.570** / **±1.03**
in other words, the width of the CI is closest to ±1.000 when
n = (choose one) **20** / **10** / **3**

2. *Example: sample size when CI width = ±0.4?*
   As above, only the required width of the confidence interval is equal ±0.4.
   In this case, the width of the CI is closest to ±0.400 when
   n = (choose one) **20** / **10** / **3**
   As the sample size increases, the CI width (choose one) **increases** / **decreases**

26.6 Planning of Sample Sizes to Find "Best" Treatment

Consider a single factor ANOVA study with,
a difference, λ, between highest and second–highest treatment means[5],
r treatments,
σ, $1 - \alpha$,

then the sample size n is calculated so that

$$\frac{\lambda\sqrt{n}}{\sigma} = \text{ Table B.13 value}$$

Exercise 26.3 (Planning of Sample Sizes to Find "Best" Treatment)

1. *Sample size when $\lambda = 0.5$, $r = 4$?*
   Determine the sample size in a study with four treatments, $r = 4$, and $\sigma = 2.5$, and $1 - \alpha = 0.99$, and the difference between "best" and "second–best" is required to be $\lambda = 0.5$.

   Using Table B.13,

   $$\frac{\lambda\sqrt{n}}{\sigma} = \frac{(0.5)\sqrt{n}}{2.5} = 3.797$$

   and so

   $$n = \left(\frac{2.5(3.797)}{0.5}\right)^2 =$$

[5]The difference λ could also be between the lowest and second–lowest treatment means, as well.

(choose one) **320** / **340** / **360**

2. *Sample size when* $\lambda = 0.5$*,* $r = 9$*?*
   As above, but let $r = 9$,

   Using Table B.13,

$$\frac{\lambda\sqrt{n}}{\sigma} = \frac{(0.5)\sqrt{n}}{2.5} = 4.1999$$

   and so

$$n = \left(\frac{2.5(4.1999)}{0.5}\right)^2 =$$

   (choose one) **420** / **441** / **460**