# STAT 3119 ANOVA

*08/29/2019 @GWU*

## Outline Week 1 (Chapter 15, 16.1)

**Class websites: https://github.com/npmldabook/Stat3119**

**Email: xtian@gwu.edu**

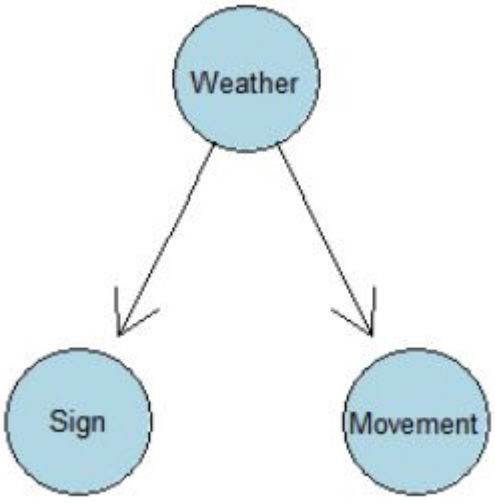
***Blackboard and TA will be updated soon in the coming weeks.**

***No homework to submit**; No **Quiz next week.**

- A1. Introduction of DOE, basic concepts, and Software setup
- B1. Design of Observational Studies
- B2. Overview of Basic Statistical Concepts
- B3. Software demo of basic analysis

## Observational Studies (chapter 15.4)

- In contrast to the comparative experiment, in an observational studies: factors are **not randomized** assigned and can only establish association.
- One potential danger is the existence of **confounding** variables (short: confounders). A confounder is a common cause for two variables.
  - Is the seatbelt sign on an airplane causing a plane to shake? If we could switch it on ourselves, would the plane start shaking?




  - Turbulent weather at the same time makes the pilot switch on the seatbelt sign and the plane shake. What we observe is an association between the appearance of the seatbelt sign and a shaking plane. The seatbelt sign is not a cause of the shaking plane.

- There are 3 common types of observational studies. $\Longrightarrow$

## Cross-sectional observational study

- A cross-sectional observational study involves measurements taken from one or more populations at a single point in time or a single time interval
  - E.g., a cross-sectional study of household incomes by geographic location
  - National surveys. E.g., National Health and Nutrition Examination Survey (*NHANES*)) by National Center for Health Statistics (NCHS): interview includes demographic, socioeconomic, dietary, and health-related questions. The survey examines a nationally representative sample of ~ 5,000 persons/ year.

## Prospective observational study

- In a prospective observational study (cohort study), one or more groups are formed in a nonrandom manner according to the levels of a hypothesized causal factor, and then these groups are observed over time with respect to an outcome variable of interest. e.g. Small or large cohort studies. It answers the question: "What is going to happen?"

- Framingham Heart Study(FHS). Launched in 1948, these studies involve studying the health of various populations to uncover patterns, trends, and outcomes to identify common factors or characteristics that contribute to cardiovascular disease.

- Factors: high vs. low blood pressures, high vs. low lipids, diabetes or not.
- Outcome: cardiovascular disease, heart attack, cancer …

## Retrospective observational study

- Groups are defined on the basis of an observed outcome, and the differences among the groups at an earlier point in time are identified as potential causal effects. It answers the question: "What has happened?"
  - Also known as the **case-control** studies: save study time when when an outcome of interest occurs infrequently.
  - A retrospective of cancer study would identify persons who have a certain cancer (cases) and a matching persons who do not have this cancer (controls) and look back in time to assess differences in their risk exposure variable. For example, finding the lung cancer cases and matching non-disease controls, then collect their smoking history and environmental factors.

## Review: Fundamental Statistical Concepts

- Prerequisite for this course: STAT 2118 (Regression Analysis), Other basic stat courses
- Appendix A of the textbook covers *Some Basic Results in Probability and Statistics*
- You are also expected to know the *basic statistical concepts*, and have knowledge in calculus and matrix algebra.
- **Population**: Any large collection of objects or individuals, such as Americans, students, or certain disease population about which information is desired
- **Parameter**: any summary number, like an average or percentage, that describes the entire population.

- **Sample**: a representative group drawn from the population. Most of statistical theory assumes random samples are used, where every observation in the sample has an equal probability of being selected
- **Statistic**: any summary number that describes the sample. It is any function of the observations in a sample that does not contain unknown parameters, such as sample mean and sample variance.

## Type of variables and sampling distributions (Appendix A)

- Type of variables:

```
- Continuous: (takes any value in a range or interval)

- Discrete: categorical/factor (commonly takes finite/countable values)

- summary: mean, variance, SD, median, range, percentage
```

- Common distribution functions for variables:

```
- Normal (Gaussian) distribution
- t-distribution (df= degrees of freedom)
- chi-square distribution (df)
- F-distribution (df1, df2)
```

## Estimation and Testing

- **Estimation** of parameter and **confidence Intervals**: for model parameters, regression coefficients.
- **Hypothesis testing** ($H_0$ vs. $H_a$): A statistical hypothesis is a statement either about the parameters of a probability distribution or the parameters of a model.
- Null vs. Alternative hypotheses (e.g. making initial assumptions of no difference in treatments vs. not)
- Test statistic and decision rule (rejection rule): To test a hypothesis, we devise a procedure for taking a random sample, computing an appropriate test statistic, and then rejecting or failing to reject the null hypothesis H0 based on the computed value of the test statistic.

| | | **Truth** | |
|---|---|---|---|
| | | Null Hypothesis | Alternative Hypothesis |
| **Decision** | Do not Reject Null | OK | Type II Error |
| | Reject Null | Type I Error | OK |

- Type I ($\alpha$) and type II error ($\beta$):
  - If the null hypothesis is rejected when it is true, a type I error has occurred (false positive).
  - If the null hypothesis is not rejected when it is false, a type II error has been made.
  - Power =1- $\beta$= probability of rejecting null when the alternative is true.

## Hypotheses Testing

- **P-value** : is defined as the smallest level of significance that would lead to rejection of the null hypothesis $H_0$. Once the P-value has been determined, we can make decision if the test result is significant by comparing P-value with the given significant level $\alpha$ (such as 0.05, 0.01).
  - If P-value $\leq \alpha$ => reject H0 at level $\alpha$.
  - If P-value $> \alpha$ => do not reject H0 at level $\alpha$.
- Inference about one sample mean, proportion, or variance.
- The statistical tests for comparison of two sample means, proportions, or variances.
- At design stage: Sample size and power analysis: a procedure that researchers can use to determine if the test contains enough power to make a reasonable conclusion for a study. Also, power analysis can also be used to calculate the number of samples required to achieve a specified level of power. For clinical trials, we set power 80% ~ 90% to calculate the sample size.

## Correlation and Regression;

- X (explanatory/independent variables, predictors, covariates, input)
- Y (outcome/response/dependent variable, output)
- Correlation coefficient between a pair of variables
- Simple linear regression (Y ~ X) : $Y_i = \beta_0 + \beta_1 X_i + e_i$
- Multiple linear regression ($Y \sim X_1, X_2, ..., X_k$) :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + e_i$$

## SINGLE-FACTOR Experimental studies (chapter 16.1 )

### Introduction

In a single-factor experimental study, the treatments correspond to the levels of the factor, and randomization is used to assign the treatments to the experimental units.

Example 1.

- A hospital research staff wished to determine the best dosage level for a standard type of drug therapy to treat a medical condition.
- In order to compare the effectiveness of three dosage levels, 30 patients with the medical problem were recruited to participate in a pilot study. Each patient was randomly assigned to one of the three drug dosage levels.
- Randomization was performed in such a way that an equal number of patients ended up being evaluated for each drug dosage level, i.e., with exactly 10 patients studied in each drug dosage level group.
- This is an example of **completely randomized design**, based on **a single, three-level factor**. This particular design is said to be **balanced**, because each treatment is replicated the same number of times.

  - **In a single factor observation study, the factor levels are not randomized. We can still analyze the data the same way with ANOVA or regression model, but it leads to the 'association' conclusions instead of 'causation'.**

## R: How to randomizely assign treatments (or assign patients into different treatment levels)

First, we generate the treatment labels, each with 10 replications (balanced design).

```
treat.order <- rep(c("Dose-1", "Dose-2", "Dose-3"), each = 10)
treat.order
```

```
##  [1] "Dose-1" "Dose-1" "Dose-1" "Dose-1" "Dose-1" "Dose-1" "Dose-1"
##  [8] "Dose-1" "Dose-1" "Dose-1" "Dose-2" "Dose-2" "Dose-2" "Dose-2"
## [15] "Dose-2" "Dose-2" "Dose-2" "Dose-2" "Dose-2" "Dose-2" "Dose-3"
## [22] "Dose-3" "Dose-3" "Dose-3" "Dose-3" "Dose-3" "Dose-3" "Dose-3"
## [29] "Dose-3" "Dose-3"
```

In the old days (15+ years ago), people use random number tables or generate a random number 0~1 from a uniform distribution, etc. With current statistical software, it is easy to randomize the treatment by getting a random permutation of the treatment label. Then we use this for the treatment assignment to the 30 patients.

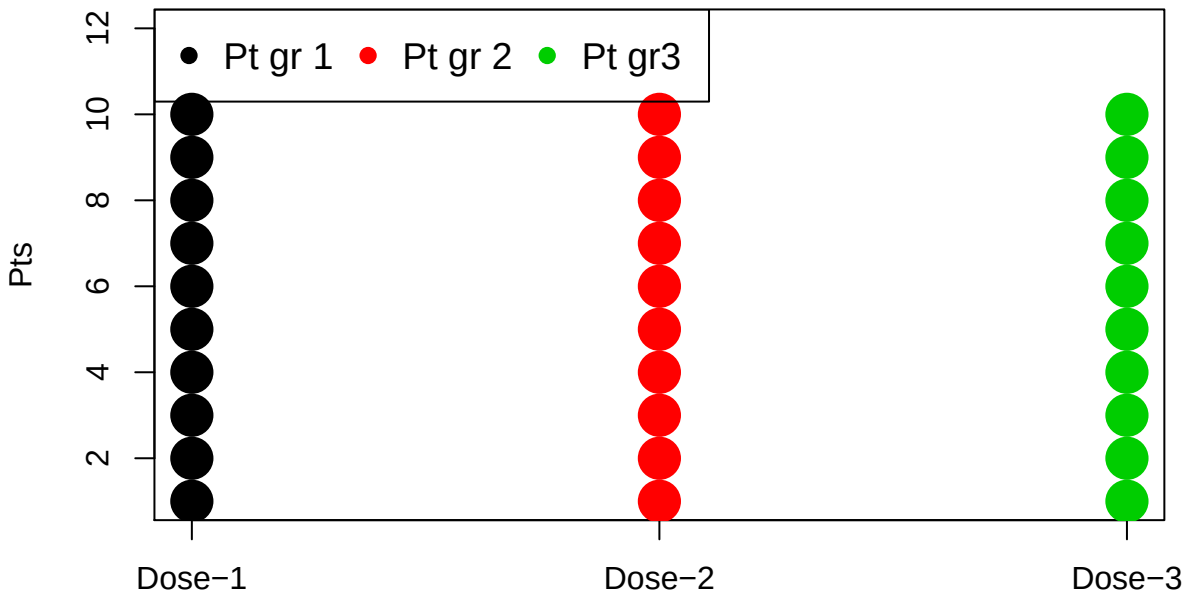
```
set.seed(111)
# set the seed of random number generator  which is useful for creating
# simulations or random objects that can be reproduced.
# Without setting the seed, the random permutation will be different each time
(NewOrder<-sample(treat.order))
```

```
##  [1] "Dose-2" "Dose-3" "Dose-2" "Dose-2" "Dose-2" "Dose-3" "Dose-1"
##  [8] "Dose-1" "Dose-1" "Dose-2" "Dose-2" "Dose-1" "Dose-1" "Dose-3"
## [15] "Dose-3" "Dose-3" "Dose-3" "Dose-1" "Dose-2" "Dose-2" "Dose-1"
## [22] "Dose-1" "Dose-3" "Dose-3" "Dose-3" "Dose-1" "Dose-3" "Dose-2"
## [29] "Dose-2" "Dose-1"
```

The randomized treatment sequence is necessary to prevent the effects of unknown nuisance variables (confounding), e.g., in a nonrandomized setting to study a lung disease such as asthma, each dose cohort is given during 3 separate months (10 patients per month), there might be a seasonable effect related to a disease incidence in these months that is cofounded with the drug dose effect.

Another possibility is that patients may come in different types or groups (sicker patients are easier to be recruited at the beginning of a clinical trial). Let me illustrate:

```
# Treat patients in order ##
Pts<- rep(1:10, 3)
stripchart(Pts ~ treat.order, vertical = TRUE,  pch=16, col=1:3, cex=3, ylim=c(1, 12))
legend('topleft', legend=c("Pt gr 1", "Pt gr 2", "Pt gr3"), col=1:3, cex=1.2, horiz = T, pch=16 )
```

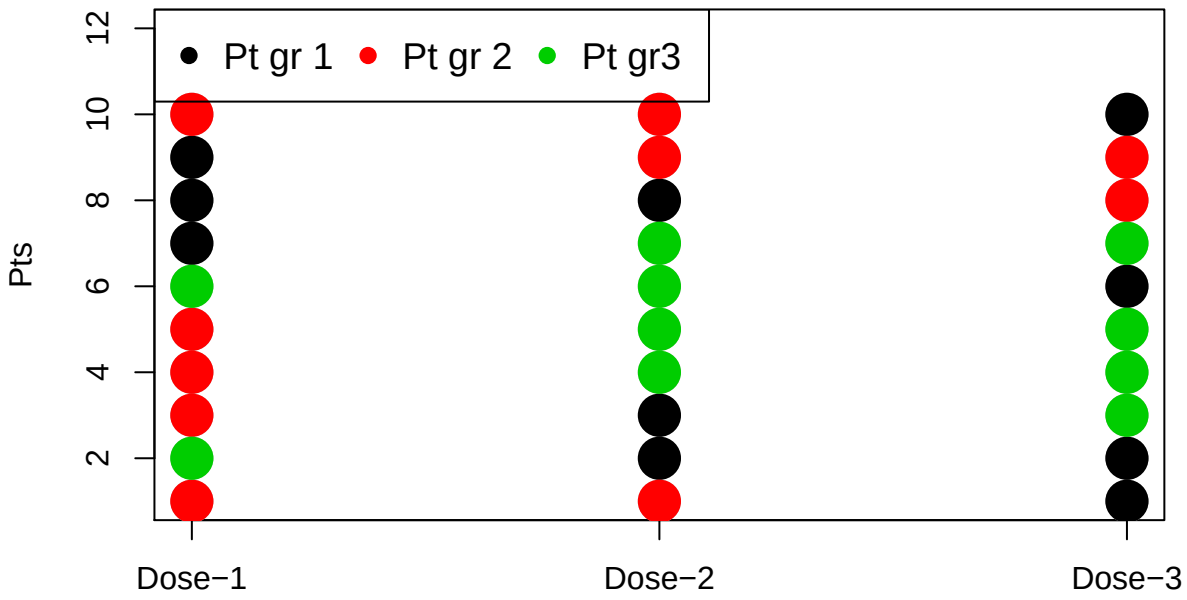



In this cases, the patient group and the treatment groups are completely confounded ! We can't separate the difference due to the treatment dose level or patient group difference.

```
# Assign patients randomly into different treatment levels  ##
Pts<- rep(1:10, 3)
stripchart(Pts ~ treat.order, vertical = TRUE,  pch=16, col='gray', cex=3, ylim=c(1, 12))

 PtGr<- sort(rep(1:3, 10))
 points(PtGr[NewOrder=='Dose-1'], Pts[NewOrder=='Dose-1'], col=1, cex=3, pch=16  )
 points(PtGr[NewOrder=='Dose-2'], Pts[NewOrder=='Dose-2'], col=2, cex=3, pch=16  )
 points(PtGr[NewOrder=='Dose-3'], Pts[NewOrder=='Dose-3'], col=3, cex=3, pch=16  )

legend('topleft', legend=c("Pt gr 1", "Pt gr 2", "Pt gr3"), col=1:3, cex=1.2, horiz = T, pch=16 )
```




Now, if we have 3 patient groups coming together as cohorts, we can assign them into different treatment dose levels by a randomized order. For each treatment level, there are patients from different patient groups. Now we can separate the effects due to treatment levels and patient groups.

Note, this is just an example to show how we can randomize a one-factor study with 3 levels (n=10 each level). For a real clinical trial, we need to have a target for the treatment effect (the difference between treatment levels) that we want to establish, and we may need to use a large sample size to achieve the adqequate power for the study.

## A walkthrough for the example using R in Chapter 15.5

Example (a matched-pair design; this is also a simplest case for a randomized complete block design with block size=2 ): The objective of a product-improvement project at a major pharmaceutical company was to reduce the sensitivity of skin to the injection of an allergen. A new experimental allergen was developed and dermatologists were interested in comparing the new formulation to the existing product. Reactions to allergen injections vary greatly from person to person, and it was decided that all comparisons of the newtreatment and standard control treatment should be conducted on a within-subject basis.

- Thus a randomized complete block experiment was utilized, where blocks correspond to subjects, and each subject was injected with both the experimental and control allergens, once in each arm. (Blocking or matched pair design is to reduce the cofounding.)

- Here, the experimental units are the subjects' arms, and each block consists of two experimental units. Randomization is accomplished by randomly assigning the treatments to the right or left arms for each subject. (In other words: if the two treatments are labeled with

(A, B). Some patients will get A/B for their left/right arms, and some patients get B/A for their left/right arms. )

- Twenty subjects were randomly chosen from a pool of available subjects for testing. The experimental layout, randomization, and results of the 40 tests are shown in Table 15.1 (textbook). The response is skin sensitivity (diameter of the red area surrounding the injection in centimeters).

## Example (continued)

**TABLE 15.1** Data and Descriptive Statistics—Skin Sensitivity Experiment.

| Subject | Control Treatment | Experimental Treatment | Within-Subject Difference |
|---|---|---|---|
| 1 | 0.59 | 0.43 | −0.16 |
| 2 | 0.69 | 0.53 | −0.16 |
| 3 | 0.82 | 0.58 | −0.24 |
| ... | ... | ... | ... |
| 18 | 0.85 | 0.60 | −0.25 |
| 19 | 0.85 | 0.65 | −0.20 |
| 20 | 0.74 | 0.58 | −0.16 |
| Sample Mean: | .7315 | .5400 | −.1915 |
| Sample Std Dev: | .0758 | .0807 | .0501 |

From (15.15) a linear statistical model for the experiment is:

$$Y_{ij} = \beta_0 + \beta_1 X_{i1} + \sum_{j=2}^{20} \beta_j X_{ij} + \varepsilon_{ij} \qquad i = 1, 2 \qquad \textbf{(15.18)}$$

where:

$$X_{i1} = \begin{cases} 1 & \text{if experimental treatment} \\ 0 & \text{if control treatment} \end{cases}$$

$$X_{ij} = \begin{cases} 1 & \text{if response is from subject } j-1, \text{ for } j = 2, \ldots, 20 \\ 0 & \text{otherwise} \end{cases}$$

In this regression model, here the subscript $i$ indicates the arms, and $j$ indicates the subject. So, for each subject, we have two measurements of outcome $Y_{ij}$. As you have learned from the regression course, we can set up this model:

1. A indicator variable for the treatment,
2. 19 indicator variables to indicate 19 different subjects (subject 2 to 20),
3. The intercept term $\beta_0$ would be the subject effect for the 1st subject when all $X_{ij} = 0, j = 2, ..19$.

**Now we illustrate how we analyze this data in R**

- **Step 1**. We read the data into R. You can go to **stat3119/Week1 folder** and click the data file "CH15TA01.txt" to see its content. Then you click the "raw" Tab to open the data file in a browser

such as Chrome. From that window, you can right-click the mouse to save it to your local computer. If you have the data in your local folder, say, "c:/stat3119/CH15TA01.txt", then you can read it from you local directory into R in 2 ways: (1) use **read.table(file.choose())**, then you choose where the file is (user input). (2) use **read.table("c:/stat3119/CH15TA01.txt")**. Note, you have to replace the 'c:/stat3119/CH15TA01.txt' with the actual file path in your computer. Or, in the third way, you can read it directly from the class website.

```
#Ex15 <- read.table(file.choose(), header=F) # find the file location from file browser
#Ex15 <- read.table("c:/stat3119/CH15TA01.txt")   # specify where the file is in your local computer

Ex15 <- read.table(url("https://raw.githubusercontent.com/npmldabook/Stat3119/master/Week1/CH15TA01.txt"))

# check the data. This data has no headers or variable name.
dim(Ex15)
```

```
## [1] 20  4
```

```
head(Ex15)
```

```
##   V1   V2   V3    V4
## 1  1 0.59 0.43 -0.16
## 2  2 0.69 0.53 -0.16
## 3  3 0.82 0.58 -0.24
## 4  4 0.80 0.65 -0.15
## 5  5 0.66 0.38 -0.28
## 6  6 0.67 0.48 -0.19
```

- **Step 2**. The raw data file has no variable names. We rename the variables and check data again. We will csee the same data as in the Table 15.1.

```
names(Ex15) <- c("subject", "Control","Exp", "Dif")
head(Ex15)
```

```
##   subject Control  Exp   Dif
## 1       1    0.59 0.43 -0.16
## 2       2    0.69 0.53 -0.16
## 3       3    0.82 0.58 -0.24
## 4       4    0.80 0.65 -0.15
## 5       5    0.66 0.38 -0.28
## 6       6    0.67 0.48 -0.19
```

- **Step 3**. Run the statistical tests for paried comparison. Here, we run paired comparison for control and experimental groups with a paired t-test. The results match the textbook (p. 671).

The dermatologists were primarily interested in determining whether the experimental allergen formulation led to difference in skin sensitivity, i.e., testing $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.

We don't have to use MINITAB to run the analysis. This is simply a paired t-test that can be done in R with a line of code.

```r
t.test(Ex15$Exp, Ex15$Control, paired=T )
```

```
##
##  Paired t-test
##
## data:  Ex15$Exp and Ex15$Control
## t = -17.1, df = 19, p-value = 5.377e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2149389 -0.1680611
## sample estimates:
## mean of the differences
##                 -0.1915
```

Since t= 17.10 and p-value is very small, we conclude Ha, that $1 = 0$. Since the mean estimate b1 was =-0.1915 <0, the dermatologists concluded that the new formulation significantly reduces skin irritation.

- **Step 4**. In the paired t-test, we are only interested in the treatment effect, not the subject difference. However, if the investigators were not primarily interested in determining whether or not subject (block) effects were present. Here, blocking was used here to increase the precision of the comparisons between the experimental and control treatments and it was fully expected that significant subject-to-subject differences would be present. We can use ANOVA table to differentiate the source of variability, difference due to the treatment or the subjects (study participants). (*Don't worry if this is not clear here, we will go over the model and details in later chapters*).

Basically, we need to transform this data to specify the factors (treatment and subject) and outcome $Y$, then we run the ANOVA to get the test reesults.

```r
# Data transformation
Ex15v2 <- data.frame( Treatment=c(rep(0,20), rep(1,20)), Subject=factor(c(1:20, 1:20)),
                      Y= c( Ex15$Control ,Ex15$Exp))
head(Ex15v2)
```

```
##   Treatment Subject    Y
## 1         0       1 0.59
## 2         0       2 0.69
## 3         0       3 0.82
## 4         0       4 0.80
## 5         0       5 0.66
## 6         0       6 0.67
```

```r
summary(aov(Y~  Treatment+Subject, data=Ex15v2 ))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Treatment    1 0.3667  0.3667 292.424 5.38e-13 ***
## Subject     19 0.2120  0.0112   8.898 7.33e-06 ***
## Residuals   19 0.0238  0.0013
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

And the results show that both the treatment effect and the effect due to subject difference are significant. Therefore, blocking is important to use here to reduce the error variance. Without blocking, you may need more subjects to test the difference between two treatments.

## Homework

- **Reading**: Chapter 15; Appendix A.
- Rerun the example R code in the lecture note to see if you can the same results. You get better quickly by practicing the examples.
- **Homework**: Problem 15.9, 15.14 (finish at home, no need to turn-in this time, we discuss in the next class)

*15.9. An economist compiled data on productivity improvements last year for a sample of firms producing electronic computing equipment. The firms were classified according to the level of their average expenditures for research and development in the past three years (low, moderate, high).

a. Is this study experimental, observational, or mixed experimental and observational? Why?
b. Identify all factors, factor levels, and factor-level combinations.
c. What type of study design is being implemented here?
d. What is the basic unit of study?

*15.14. A research laboratory was developing a new compound for the relief of severe cases of hay fever. The amounts of two active ingredients (low, medium, high) in the compound were varied at three levels each using 18 volunteers. Randomization was used in assigning volunteers to each of the treatment combinations. Data were collected on hours of relief.

a. Is this study experimental, observational, or mixed? Why?
b. Identify all factors, factor levels, and factor-level combinations.
c. Describe how randomization would be performed in this study.
d. What type of study design is being implemented here?
e. What is the basic unit of study?