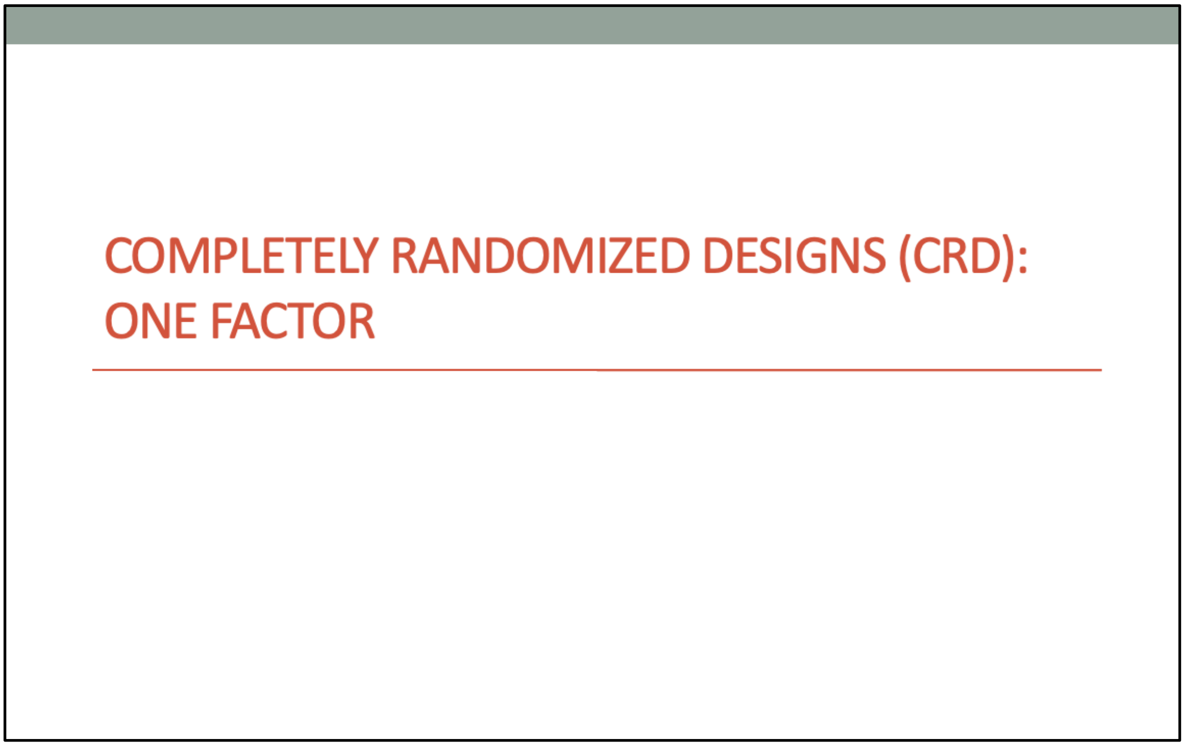
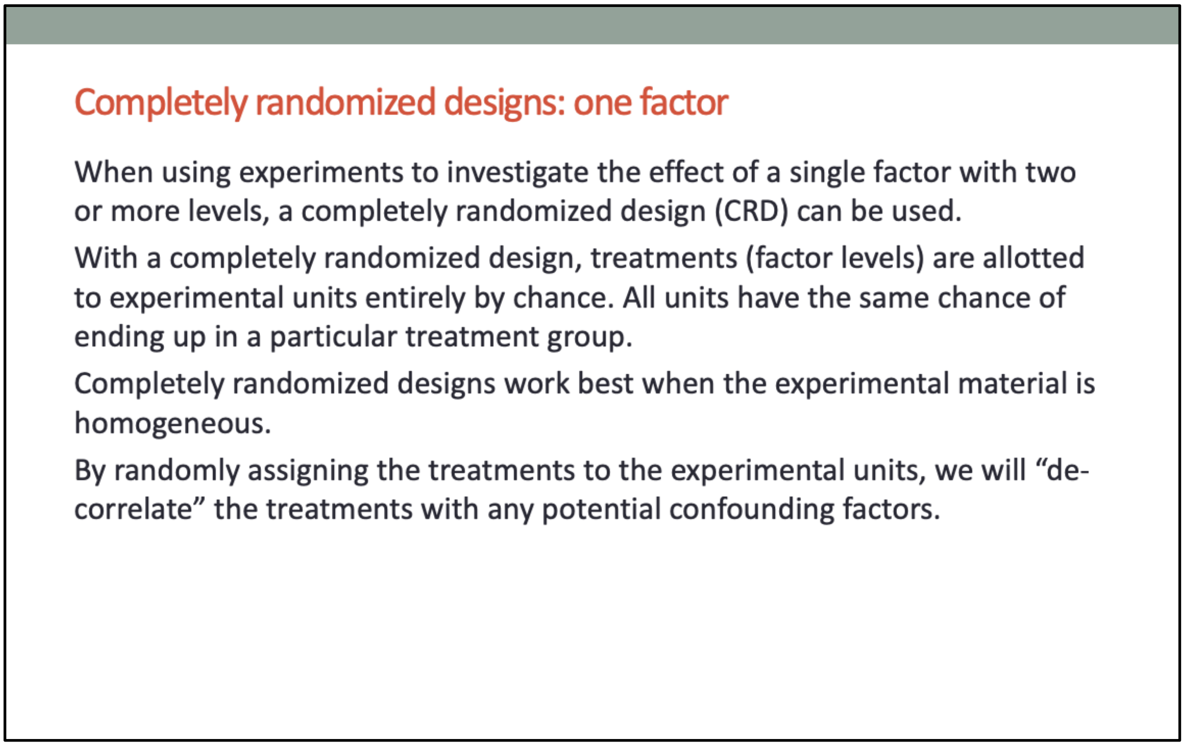


COMPLETELY RANDOMIZED DESIGNS (CRD): ONE FACTOR

THE DESIGN



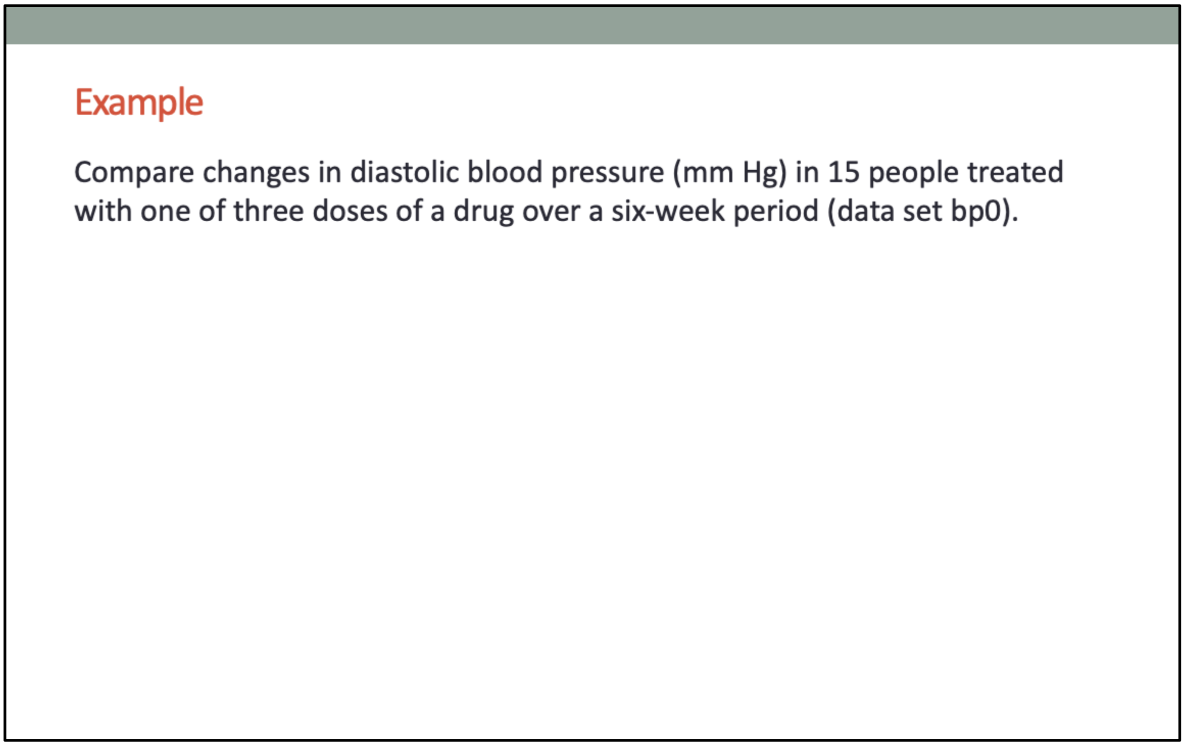
Completely randomized designs: one factor

When using experiments to investigate the effect of a single factor with two or more levels, a completely randomized design (CRD) can be used.

With a completely randomized design, treatments (factor levels) are allotted to experimental units entirely by chance. All units have the same chance of ending up in a particular treatment group.

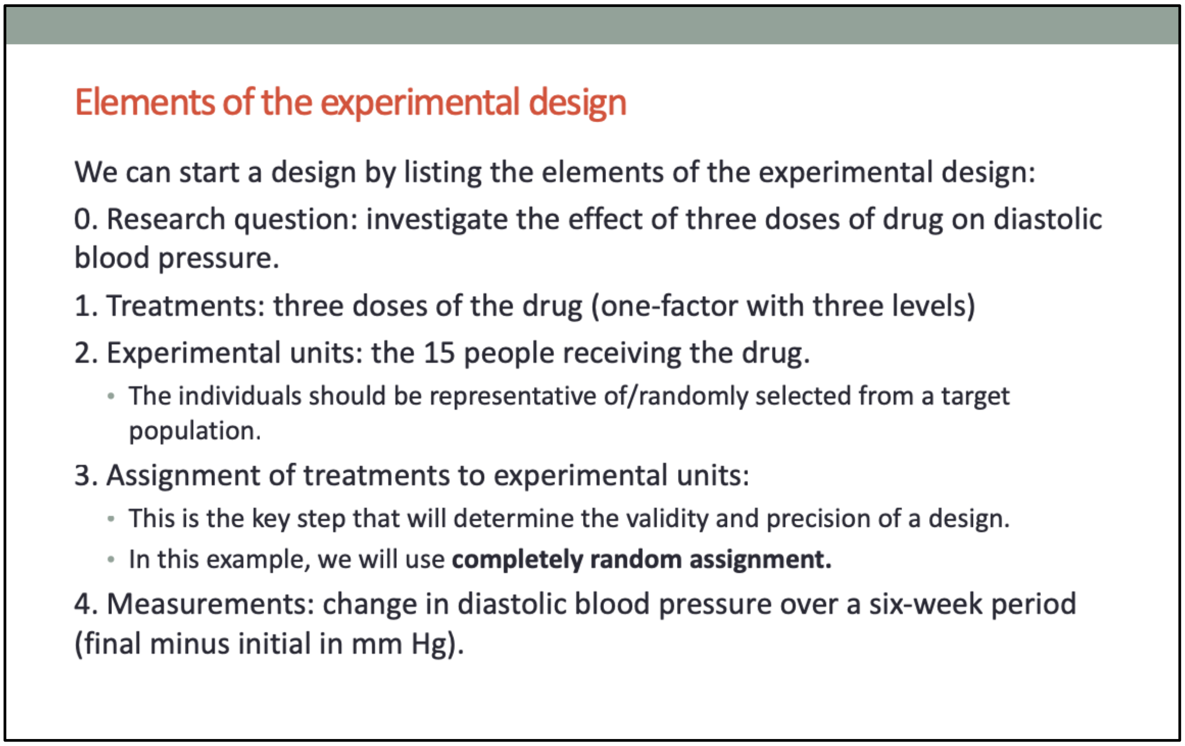
Completely randomized designs work best when the experimental material is homogeneous.

By randomly assigning the treatments to the experimental units, we will “de-correlate” the treatments with any potential confounding factors.



Example

Compare changes in diastolic blood pressure (mm Hg) in 15 people treated with one of three doses of a drug over a six-week period (data set bp0).



Elements of the experimental design

We can start a design by listing the elements of the experimental design:

0. Research question: investigate the effect of three doses of drug on diastolic blood pressure.

1. Treatments: three doses of the drug (one-factor with three levels)

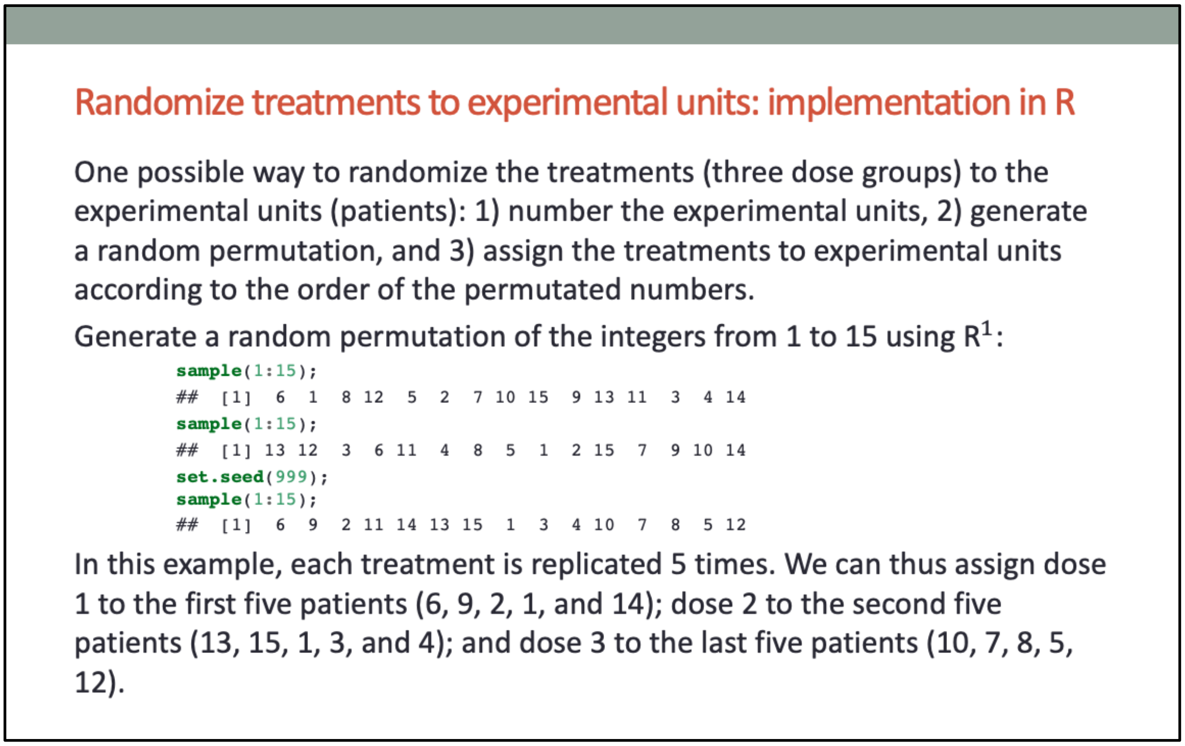
2. Experimental units: the 15 people receiving the drug.

   - The individuals should be representative of/randomly selected from a target population.

3. Assignment of treatments to experimental units:

   - This is the key step that will determine the validity and precision of a design.
   - In this example, we will use **completely random assignment.**

4. Measurements: change in diastolic blood pressure over a six-week period (final minus initial in mm Hg).



Randomize treatments to experimental units: implementation in R

One possible way to randomize the treatments (three dose groups) to the experimental units (patients): 1) number the experimental units, 2) generate a random permutation, and 3) assign the treatments to experimental units according to the order of the permutated numbers.

Generate a random permutation of the integers from 1 to 15 using R[1]:

```
sample(1:15);
##  [1]  6  1  8 12  5  2  7 10 15  9 13 11  3  4 14
sample(1:15);
##  [1] 13 12  3  6 11  4  8  5  1  2 15  7  9 10 14
set.seed(999);
sample(1:15);
##  [1]  6  9  2 11 14 13 15  1  3  4 10  7  8  5 12
```

In this example, each treatment is replicated 5 times. We can thus assign dose 1 to the first five patients (6, 9, 2, 1, and 14); dose 2 to the second five patients (13, 15, 1, 3, and 4); and dose 3 to the last five patients (10, 7, 8, 5, 12).

1. The R command

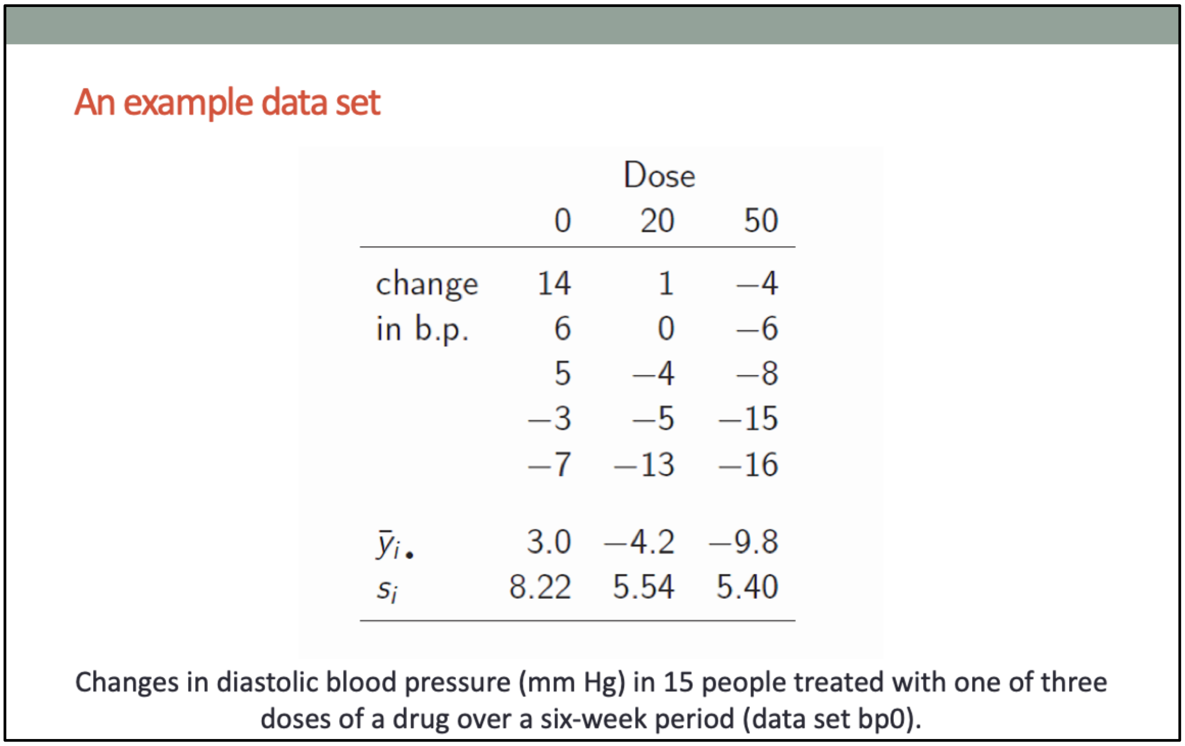
```
sample(1:15)
```

will give a random permutation of the numbers 1, 2, …, 15. Each time we run the command, a different permutation will be generated and all possible permutations are equally likely to happen.

All commands in R that produce random outputs depend on an random number seed to work. If we want our results to be reproducible, we can fix the random number seed using the command “set.seed”:

```
set.seed(999)
```

In the code chunk, lines starting with “##” indicating R output, you will not see them when you run the R code.

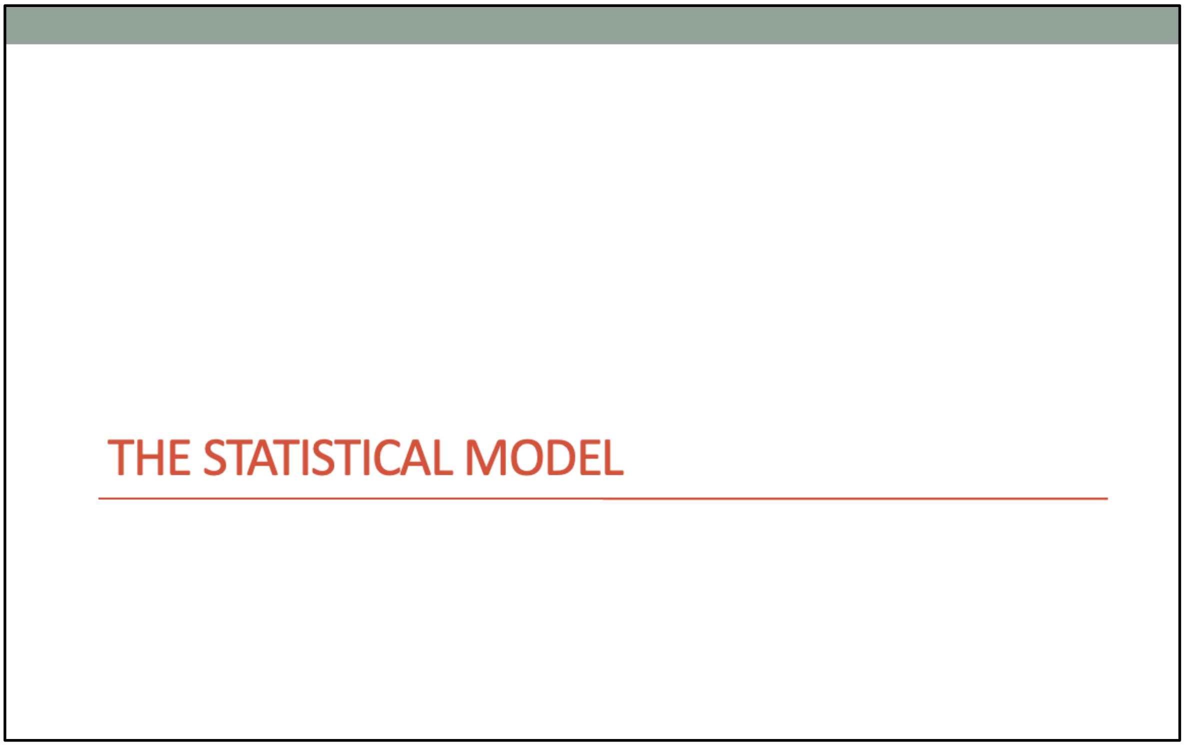


An example data set

	Dose		
	0	20	50
change in b.p.	14	1	−4
	6	0	−6
	5	−4	−8
	−3	−5	−15
	−7	−13	−16
$\bar{y}_{i\bullet}$	3.0	−4.2	−9.8
s_i	8.22	5.54	5.40

Changes in diastolic blood pressure (mm Hg) in 15 people treated with one of three doses of a drug over a six-week period (data set bp0).

The data we use here is a subset of the full data. The actual clinical trial actually has a much bigger sample size.

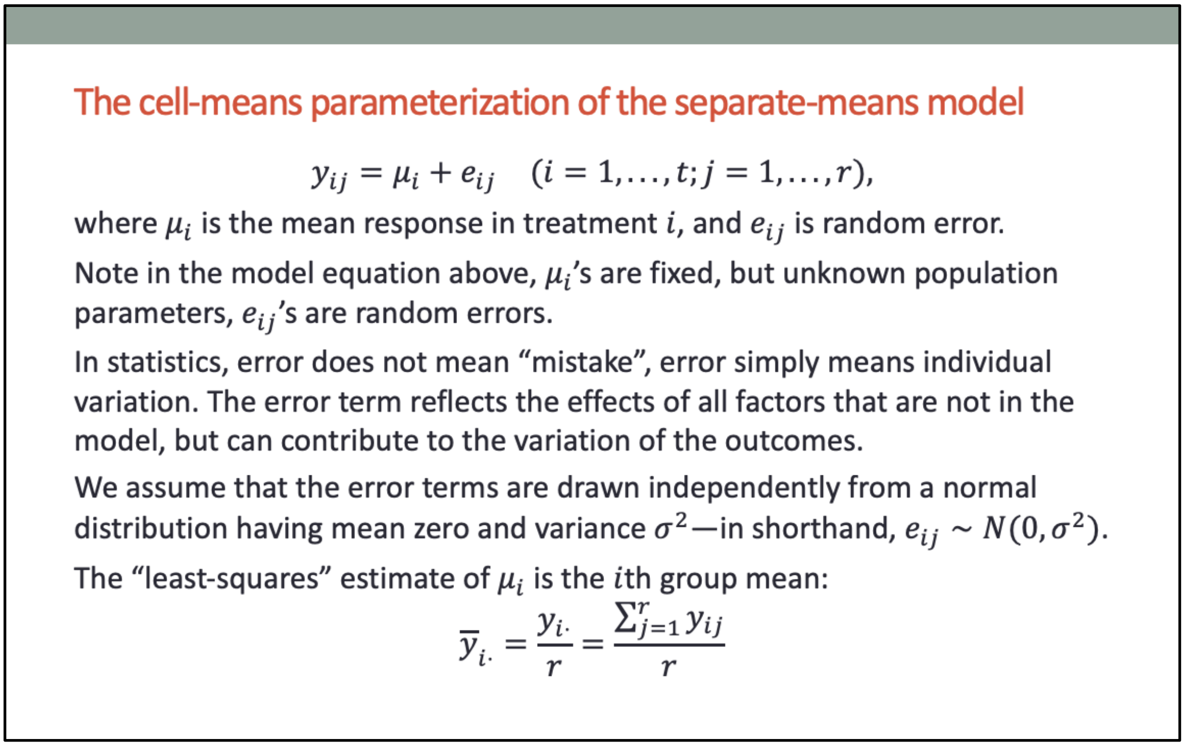


THE STATISTICAL MODEL

The separate-means model

Let y_{ij} be the jth observation in the ith group. In the blood-pressure example, $i = 1{,}2{,}3$ and $j = 1{,}2, \ldots, 5$.

There are two ways to 'parameterize' the single-factor analysis of variance (ANOVA) model, also called the separate-means model.



The cell-means parameterization of the separate-means model

$$y_{ij} = \mu_i + e_{ij} \quad (i = 1, \ldots, t; j = 1, \ldots, r),$$

where μ_i is the mean response in treatment i, and e_{ij} is random error.

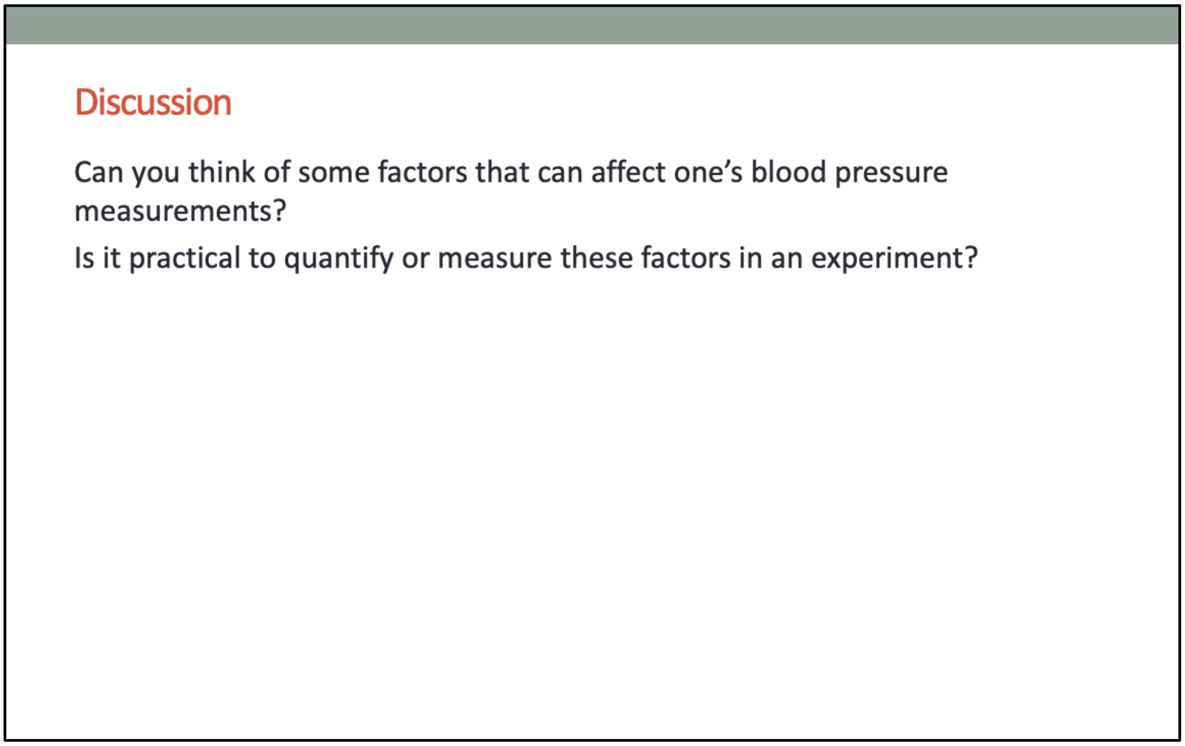
Note in the model equation above, μ_i's are fixed, but unknown population parameters, e_{ij}'s are random errors.

In statistics, error does not mean “mistake”, error simply means individual variation. The error term reflects the effects of all factors that are not in the model, but can contribute to the variation of the outcomes.

We assume that the error terms are drawn independently from a normal distribution having mean zero and variance σ^2—in shorthand, $e_{ij} \sim N(0, \sigma^2)$.

The “least-squares” estimate of μ_i is the ith group mean:

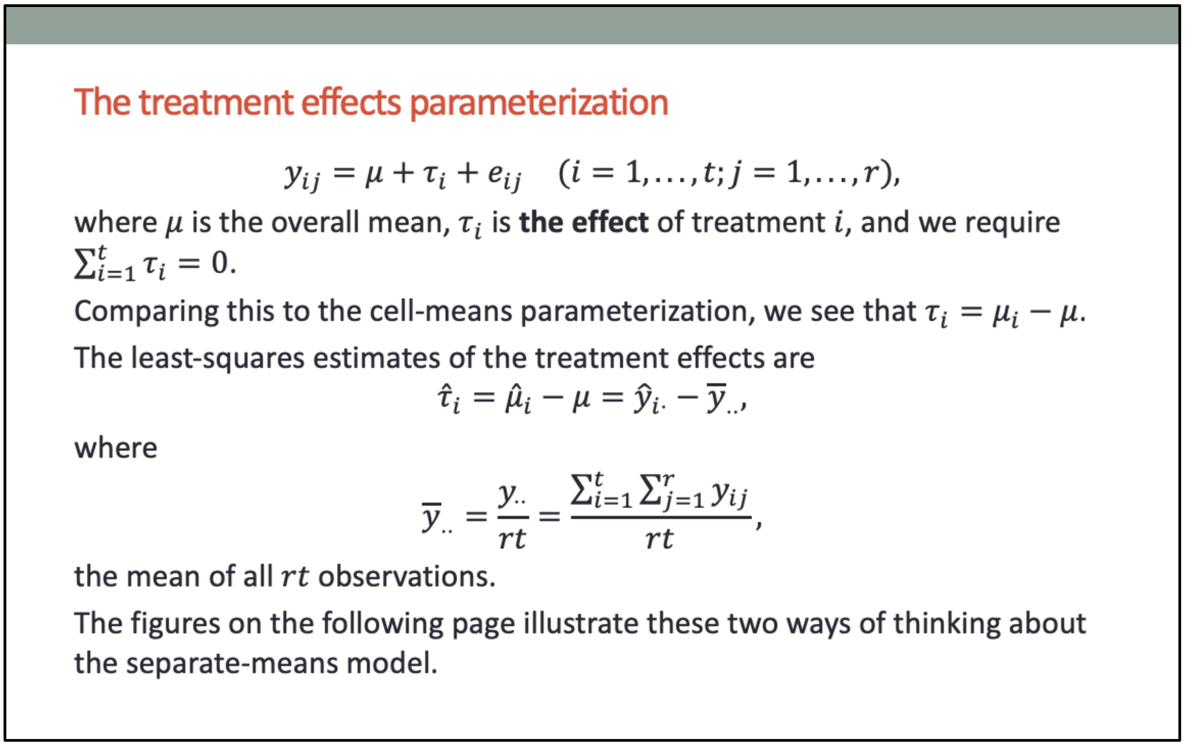
$$\overline{y}_{i\cdot} = \frac{y_{i\cdot}}{r} = \frac{\sum_{j=1}^{r} y_{ij}}{r}$$



Discussion

Can you think of some factors that can affect one’s blood pressure measurements?

Is it practical to quantify or measure these factors in an experiment?



The treatment effects parameterization

$$y_{ij} = \mu + \tau_i + e_{ij} \quad (i = 1, \dots, t; j = 1, \dots, r),$$

where μ is the overall mean, τ_i is **the effect** of treatment i, and we require $\sum_{i=1}^{t} \tau_i = 0$.

Comparing this to the cell-means parameterization, we see that $\tau_i = \mu_i - \mu$.

The least-squares estimates of the treatment effects are

$$\hat{\tau}_i = \hat{\mu}_i - \mu = \hat{y}_{i\cdot} - \overline{y}_{\cdot\cdot},$$

where

$$\overline{y}_{\cdot\cdot} = \frac{y_{\cdot\cdot}}{rt} = \frac{\sum_{i=1}^{t} \sum_{j=1}^{r} y_{ij}}{rt},$$

the mean of all rt observations.

The figures on the following page illustrate these two ways of thinking about the separate-means model.

Treatment effects and cell means

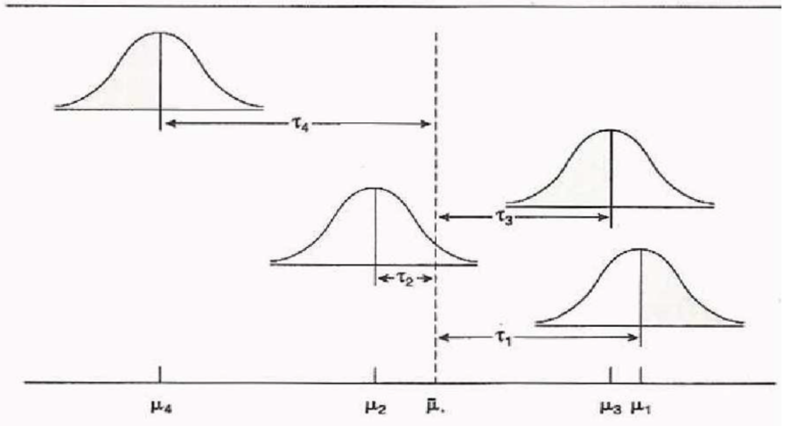
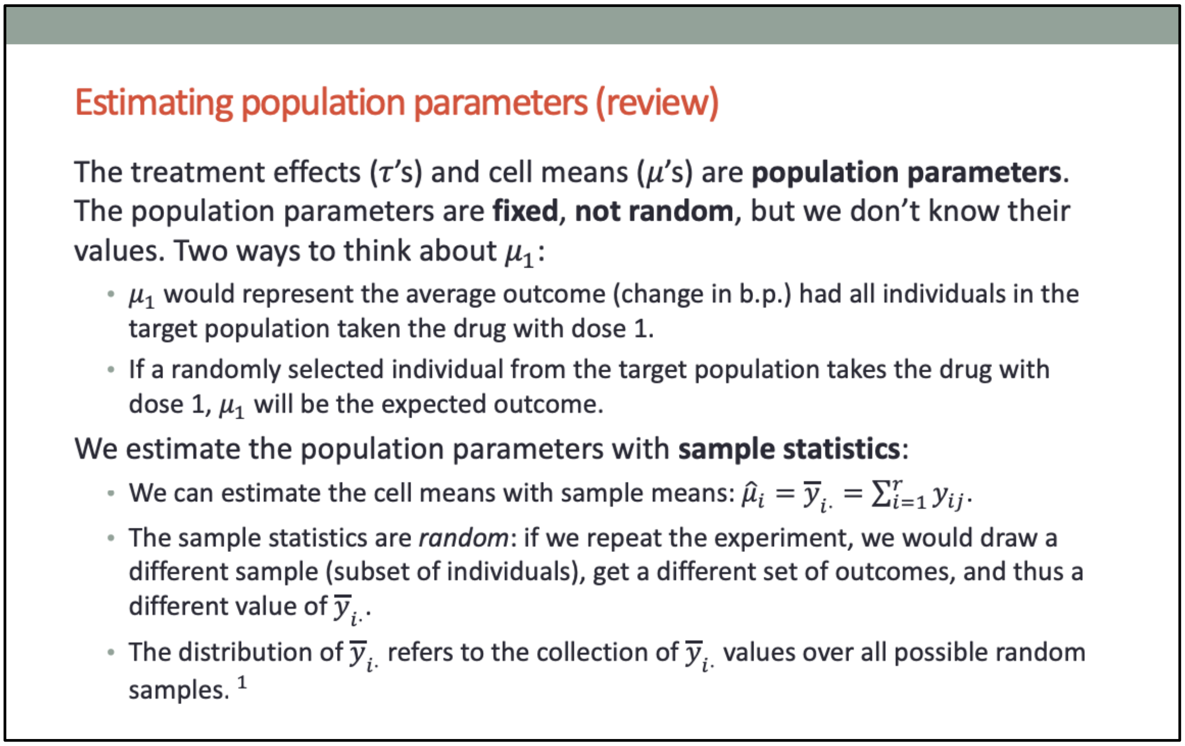


Figure 2.3 A graphical representation of treatment effects

Figure 2.3 from Kuehl.



Estimating population parameters (review)

The treatment effects (τ's) and cell means (μ's) are **population parameters**. The population parameters are **fixed**, **not random**, but we don't know their values. Two ways to think about μ_1:

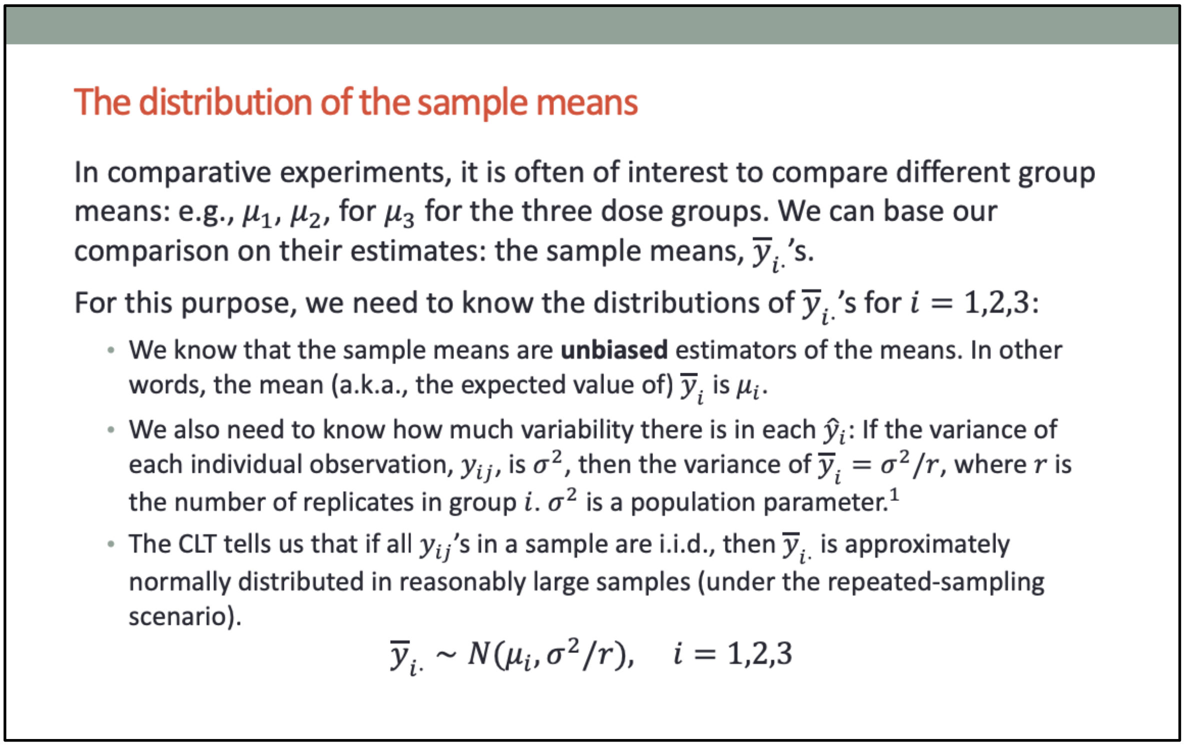
- μ_1 would represent the average outcome (change in b.p.) had all individuals in the target population taken the drug with dose 1.
- If a randomly selected individual from the target population takes the drug with dose 1, μ_1 will be the expected outcome.

We estimate the population parameters with **sample statistics**:

- We can estimate the cell means with sample means: $\hat{\mu}_i = \overline{y}_{i\cdot} = \sum_{i=1}^{r} y_{ij}$.
- The sample statistics are *random*: if we repeat the experiment, we would draw a different sample (subset of individuals), get a different set of outcomes, and thus a different value of $\overline{y}_{i\cdot}$.
- The distribution of $\overline{y}_{i\cdot}$ refers to the collection of $\overline{y}_{i\cdot}$ values over all possible random samples. [1]

If we want to see the distribution of a statistic, we can simulate a large number of random samples (of the same size), and compute the sample statistic value from each of the sample, and plot the histogram of the result values.

R is really powerful tool for performing such simulations.



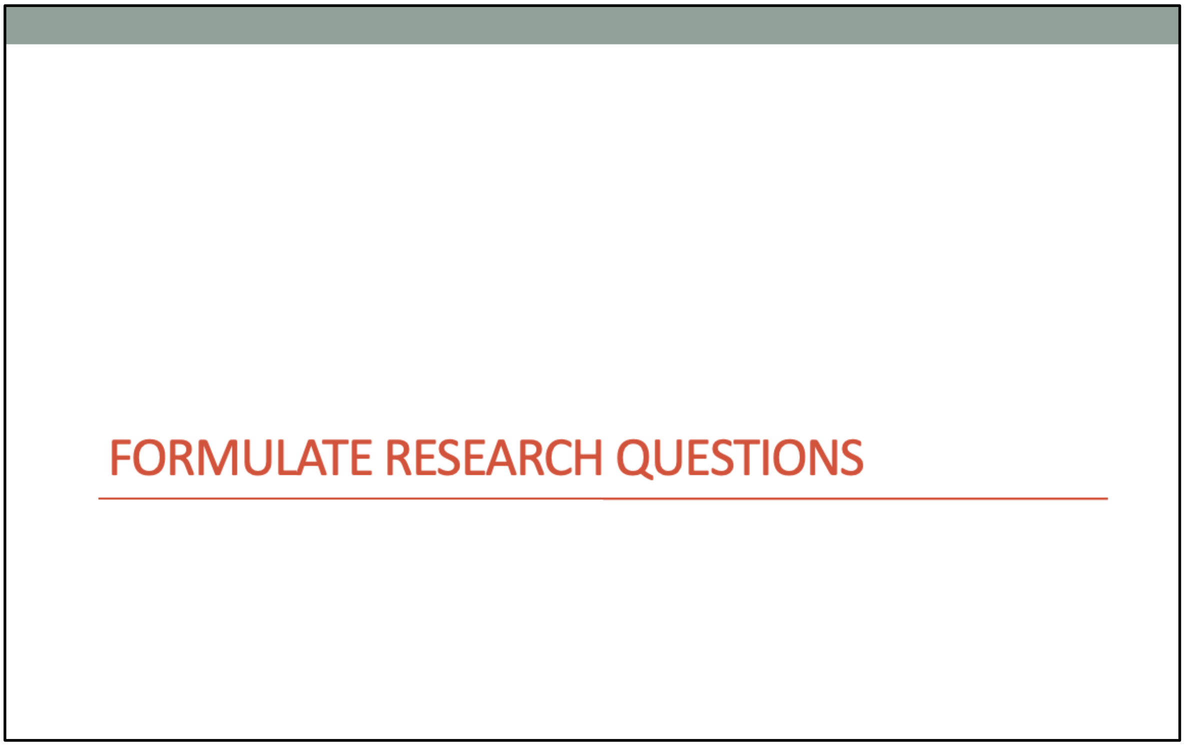
The distribution of the sample means

In comparative experiments, it is often of interest to compare different group means: e.g., μ_1, μ_2, for μ_3 for the three dose groups. We can base our comparison on their estimates: the sample means, $\overline{y}_{i.}$'s.

For this purpose, we need to know the distributions of $\overline{y}_{i.}$'s for $i = 1,2,3$:

- We know that the sample means are **unbiased** estimators of the means. In other words, the mean (a.k.a., the expected value of) $\overline{y}_i$ is μ_i.
- We also need to know how much variability there is in each $\hat{y}_i$: If the variance of each individual observation, y_{ij}, is σ^2, then the variance of $\overline{y}_i = \sigma^2/r$, where r is the number of replicates in group i. σ^2 is a population parameter.[1]
- The CLT tells us that if all y_{ij}'s in a sample are i.i.d., then $\overline{y}_{i.}$ is approximately normally distributed in reasonably large samples (under the repeated-sampling scenario).

$$\overline{y}_{i.} \sim N(\mu_i, \sigma^2/r), \quad i = 1,2,3$$



FORMULATE RESEARCH QUESTIONS

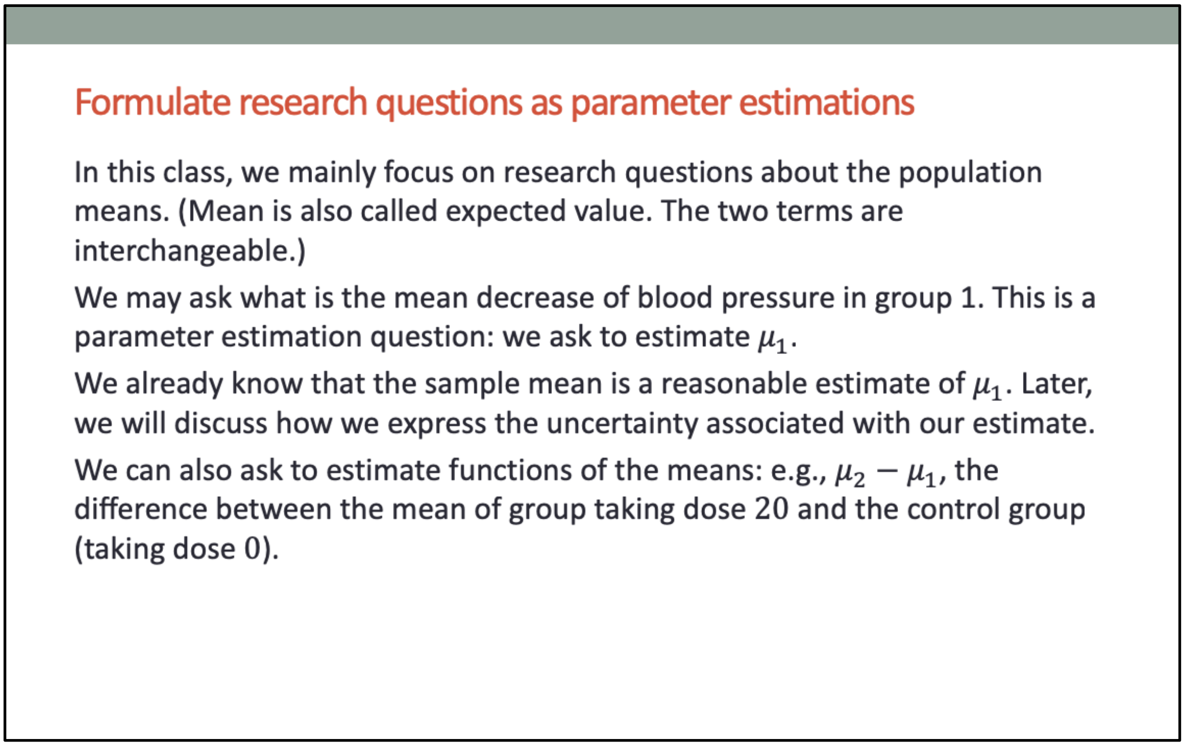
Research questions

In statistics, research questions are often formulated either as parameter estimation or hypothesis testing.

Note that in the context of experimental design, we should clearly state the research questions before we design and carry out the experiment.

If one looks at the data first and then formulate research questions, he or she runs the risk of “data snooping”.

Basically, that will increase the chance of getting false signals: finding patterns in the data that is not really there.



Formulate research questions as parameter estimations

In this class, we mainly focus on research questions about the population means. (Mean is also called expected value. The two terms are interchangeable.)

We may ask what is the mean decrease of blood pressure in group 1. This is a parameter estimation question: we ask to estimate μ_1.

We already know that the sample mean is a reasonable estimate of μ_1. Later, we will discuss how we express the uncertainty associated with our estimate.

We can also ask to estimate functions of the means: e.g., $\mu_2 - \mu_1$, the difference between the mean of group taking dose 20 and the control group (taking dose 0).

Formulate research questions as statistical hypotheses

We can formulate our research questions as hypotheses about the population parameters.

For example, we may ask: Do the patients from all three dose groups have the same mean outcomes? Or we can rephrase the sentence as testing true or false of statement “Patients from all three dose groups have the same mean outcomes.”

We can write the quoted statement using a mathematical expression:

$$H_0 : \mu_1 = \mu_2 = \mu_3.$$

We call such a statement a null hypothesis.

To test this H_0 means we want to see whether our data are consistent with the statement. We will reject H_0 if our data are not consistent with H_0 and conclude that there is evidence from data that the three dose groups do not have the same means.

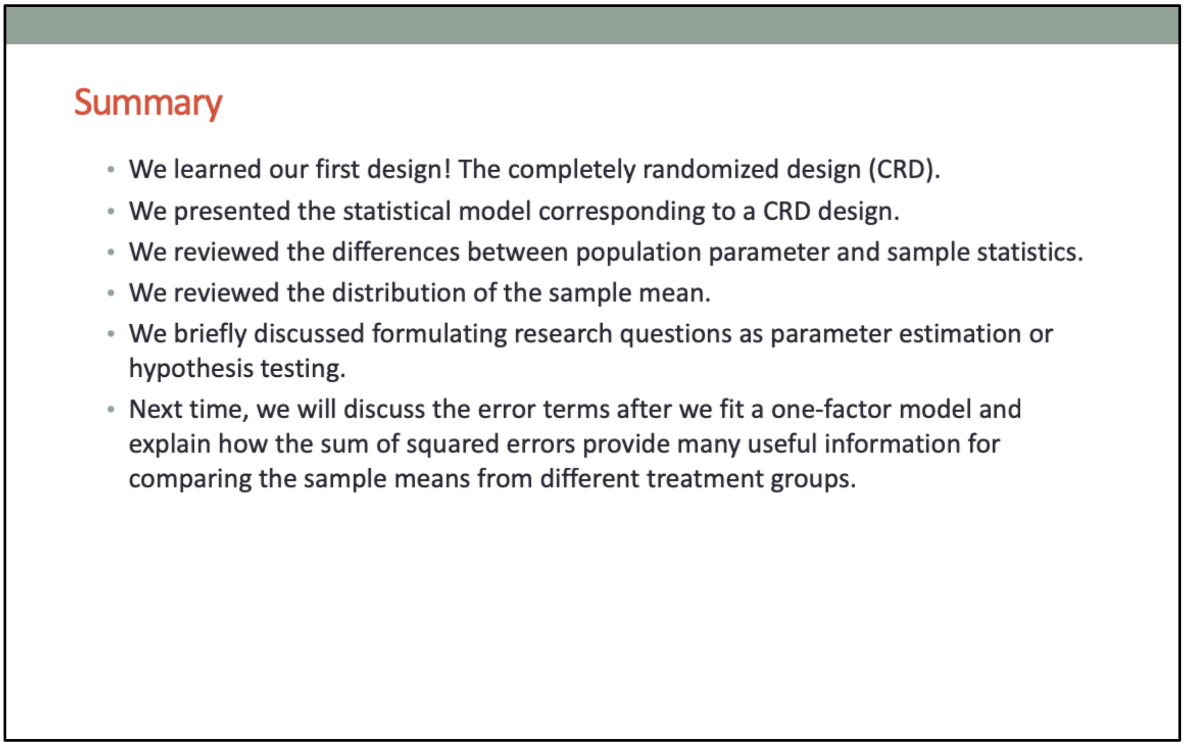
More examples of statistical hypotheses

We can also ask to test $H_0: \mu_1 < 10$. This null hypothesis says "the mean drop in blood pressure measurement in group 1 is less than 10?"

This is a one-side hypothesis. We will reject H_0 if the statistical evidence from the observed data indicates it is unlikely that the our data are from a ditribution with $\mu_1 \leq 10$.

The null hypothesis $H_0: (\mu_2 + \mu_3)/2 \leq \mu_1$, says "the average mean decrease in blood pressure in group 2 and 3 is no greater than the mean decrease in blood pressure in group 1." Again, this is a one-sided hypothesis.

Often times, the null hypotheses correspond to the "boring" cases/situations, researchers often hope that the data provide evidence to reject the null hypotheses.



Summary

- We learned our first design! The completely randomized design (CRD).
- We presented the statistical model corresponding to a CRD design.
- We reviewed the differences between population parameter and sample statistics.
- We reviewed the distribution of the sample mean.
- We briefly discussed formulating research questions as parameter estimation or hypothesis testing.
- Next time, we will discuss the error terms after we fit a one-factor model and explain how the sum of squared errors provide many useful information for comparing the sample means from different treatment groups.