Design of Experiment – Part I

High-throughput Sequencing Course

Department of Biostatistics & Bioinformatics
Duke University Medical Center

July 17, 2017

Outline

- **Part I**
  - Definition of Design of Experiment (DOE)
  - Basic principles of DOE
  - Basic statistics related to DOE
  - Experimental design for basic science
- **Part II**
  - Study design in clinical research
  - Study design in genetic association studies
  - Designs related to RNA-Seq

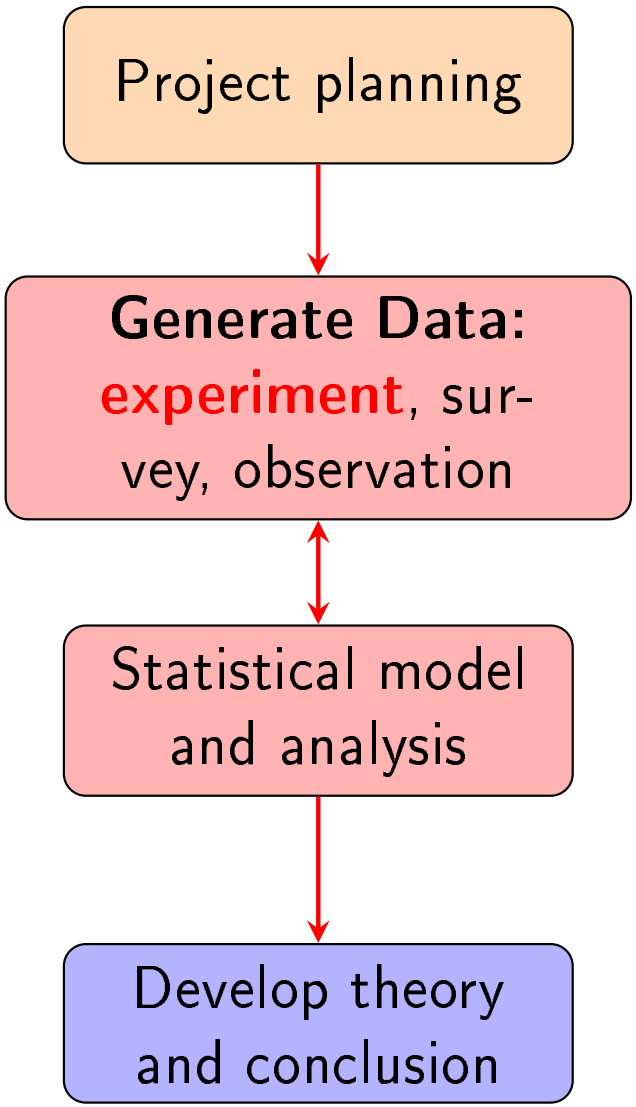
Definition of DOE

Basic definition of design of experiment (DOE)

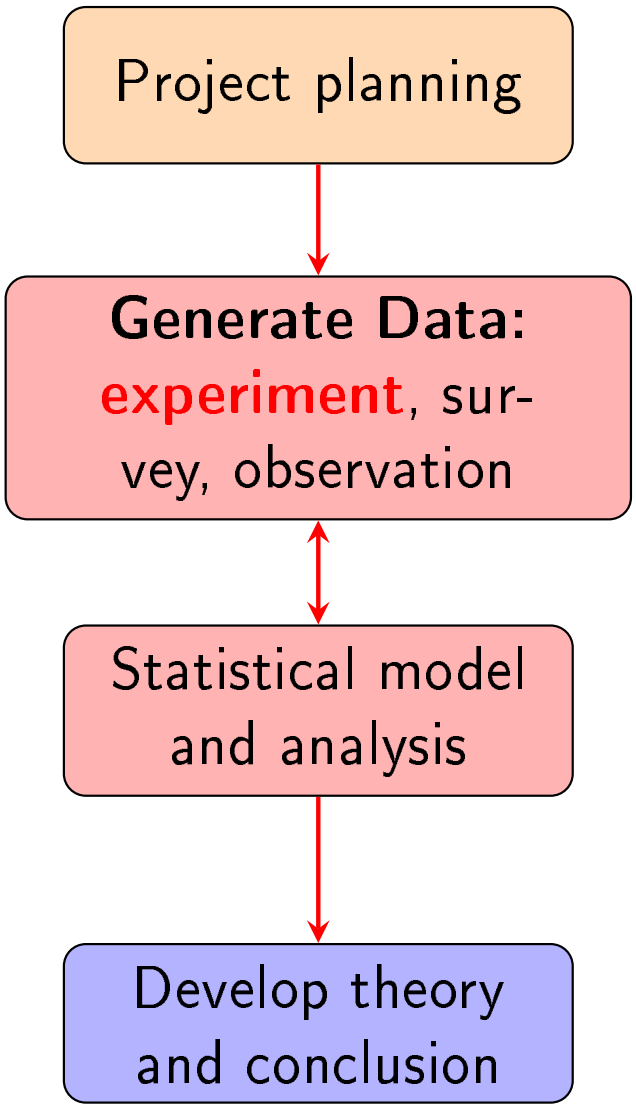
- **Experiment:** A process that generates data to achieve specific objective
- **All data are subject to variation.**
- **DOE:** A systematic method to determine the effect of a factor(s) to the outputs (responses) of the experiment. An effective experiment can
  - eliminate known sources of bias
  - prevent unknown source of bias
  - obtain data with high accuracy and precision.

  So, it can effectively evaluate the effects of individual and/or joint factors to the responses, and to answer predefined questions (*e.g.* hypothesis, theory, model).
- R.A. Fisher pioneered the field of statistical principals of experimental design.

Stages of a study



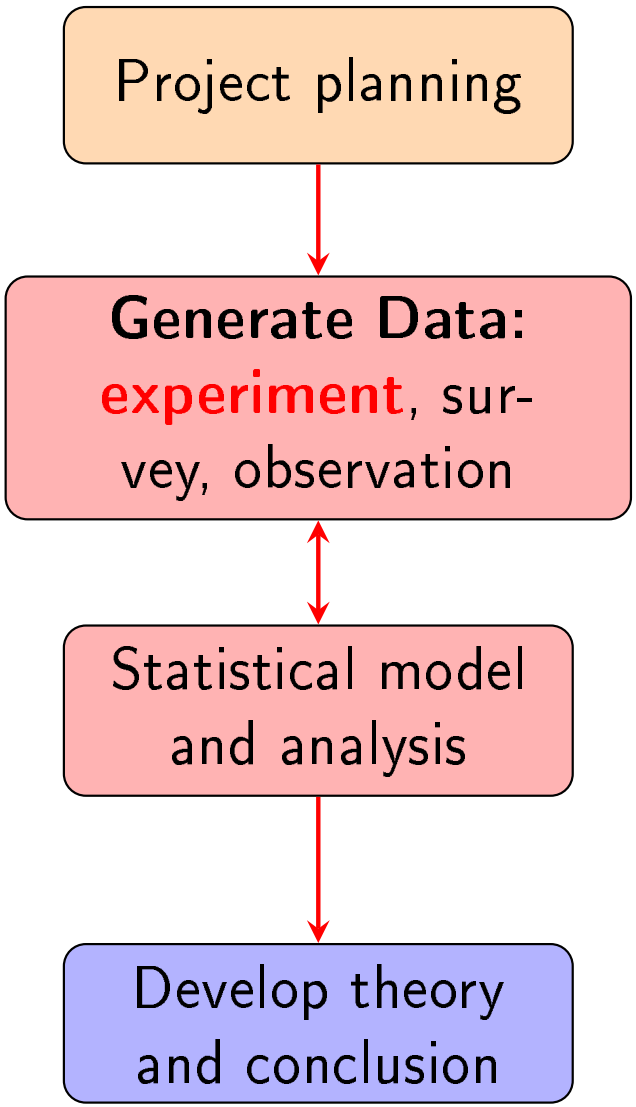
Stages of a study



Project planning

Hypothesis; what to be measured; influential factors

Stages of a study



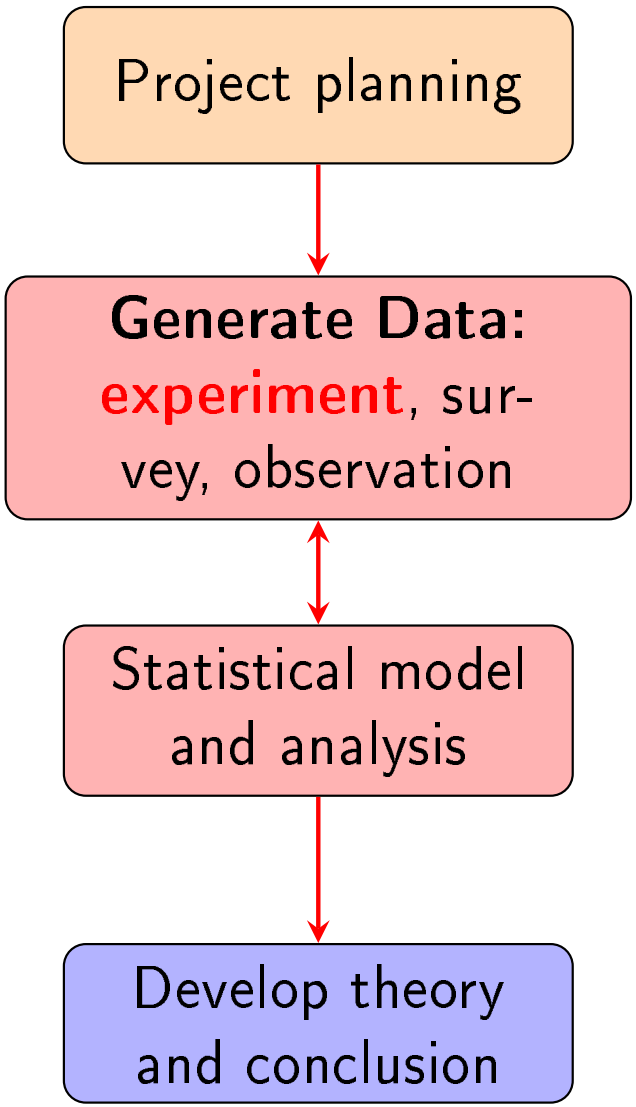
Project planning

Hypothesis; what to be measured; influential factors

Experimental studies

Ability to control the source of variability

Stages of a study



Project planning

Hypothesis; what to be measured; influential factors

Experimental studies

Ability to control the source of variability

Observational studies

No controls over the source of variability

Main elements in EOD

- **Formulate research questions and hypothesis.**

Main elements in EOD

- **Formulate research questions and hypothesis.**
- **Experimental units:** The entities that experimental procedures (*e.g.* treatments) are applied to.
  - Examples: Mice, plants, patients, etc. What is the experimental units in your experiement?
  - Need to be representative for the inference to be made.

Main elements in EOD

- **Formulate research questions and hypothesis.**
- **Experimental units:** The entities that experimental procedures (*e.g.* treatments) are applied to.
  - Examples: Mice, plants, patients, etc. What is the experimental units in your experiement?
  - Need to be representative for the inference to be made.
- **Observation units or response variables:** Any outcomes or results of the experiment (*e.g.* . gene expression of the RNA-Seq study)
  - Quantitative measures: continuous variable, *e.g.* yield of corns, height, gene expression
  - Qualitative measures: binary or categorical
  - responses are only comparable if they are measured from homogeneous experimental units.

More on main elements

- **Factors:** Variables to be investigated to determine its effect to the response variable (*e.g.* treatment effect)
  - It should be defined prior to the experiment.
  - It can be controlled by experimenter.
- **Effect**: changes in the average response between levels of a factor, or between two experimental conditions.
- **Covariate:** May affect the response but cannot be controlled in an experiment.
  - It is not affected by factors.

Formulate hypothesis

- Study objective is established on a scientific question.
- Need to be able to translate the scientific question (study objective) to a hypothesis that can be tested.
  - **Null hypothesis:** hypothesis of no change or no experimental effect
  - **Alternative hypothesis:** hypothesis of change or experimental effect.
    - Mostly is the goal you want to acheive in your study objective.
- Therefore, need to know what statement you would like to make in your scientific question.

Examples: are these correct study objectives or testable hypotheses?

1. **Study objective:** 'To examine the complications, mortality, cost and discharge status of patients with disease X'

Examples: are these correct study objectives or testable hypotheses?

1. **Study objective:** 'To examine the complications, mortality, cost and discharge status of patients with disease X'

   Concerns:

   - Examine = estimate rates? or Examine = compare rates?
   - Compare outcomes in disease X with a fixed rate or with outcomes in another disease?

Examples: are these correct study objectives or testable hypotheses?

1. **Study objective:** 'To examine the complications, mortality, cost and discharge status of patients with disease X'

   Concerns:

   - Examine = estimate rates? or Examine = compare rates?
   - Compare outcomes in disease X with a fixed rate or with outcomes in another disease?

2. **Study objective:** 'This study will identify and characterize patients who had operation X'

Examples: are these correct study objectives or testable hypotheses?

1. **Study objective:** 'To examine the complications, mortality, cost and discharge status of patients with disease X'

   Concerns:

   - Examine = estimate rates? or Examine = compare rates?
   - Compare outcomes in disease X with a fixed rate or with outcomes in another disease?

2. **Study objective:** 'This study will identify and characterize patients who had operation X'

   Concerns:

   - This is NOT a testable hypothesis as no comparable group.
   - Do you want to simply describe these patients?

Example: Maurer et al. 2005

pH Regulates Genes for Flagellar Motility, Catabolism, and Oxidative Stress in *Escherichia coli* K-12†

Lisa M. Maurer,[1] Elizabeth Yohannes,[1] Sandra S. Bondurant,[2] Michael Radmacher,[1] and Joan L. Slonczewski[1]*

Rationale: E *coli* and related enteric bacteria respond to a wide range of pH stress by regulating gene expression and protein profiles. How gene expression regulated by different levels of pH stress was not well studied. Therefore, they wanted to investigate the gene expression pattern of E *coli* under both acid and base condition at low, neutral, and high external pH stress. **What are the hypothesis and main elements?**

Example: Maurer et al. 2005

pH Regulates Genes for Flagellar Motility, Catabolism, and Oxidative Stress in *Escherichia coli* K-12†

Lisa M. Maurer,[1] Elizabeth Yohannes,[1] Sandra S. Bondurant,[2] Michael Radmacher,[1] and Joan L. Slonczewski[1]*

Rationale: E *coli* and related enteric bacteria respond to a wide range of pH stress by regulating gene expression and protein profiles. How gene expression regulated by different levels of pH stress was not well studied. Therefore, they wanted to investigate the gene expression pattern of E *coli* under both acid and base condition at low, neutral, and high external pH stress. **What are the hypothesis and main elements?**

- **Null hypothesis:** Expression level of 'a gene' is same between pH conditions.

Example: Maurer et al. 2005

pH Regulates Genes for Flagellar Motility, Catabolism, and Oxidative Stress in *Escherichia coli* K-12†

Lisa M. Maurer,[1] Elizabeth Yohannes,[1] Sandra S. Bondurant,[2] Michael Radmacher,[1] and Joan L. Slonczewski[1]*

Rationale: E *coli* and related enteric bacteria respond to a wide range of pH stress by regulating gene expression and protein profiles. How gene expression regulated by different levels of pH stress was not well studied. Therefore, they wanted to investigate the gene expression pattern of E *coli* under both acid and base condition at low, neutral, and high external pH stress. **What are the hypothesis and main elements?**

- **Null hypothesis:** Expression level of 'a gene' is same between pH conditions.
- **Alternative hypothesis:** Expression level of 'a gene' is different between pH conditions.

Example: Maurer et al. 2005

pH Regulates Genes for Flagellar Motility, Catabolism, and Oxidative Stress in *Escherichia coli* K-12†

Lisa M. Maurer,[1] Elizabeth Yohannes,[1] Sandra S. Bondurant,[2] Michael Radmacher,[1] and Joan L. Slonczewski[1]*

Rationale: E *coli* and related enteric bacteria respond to a wide range of pH stress by regulating gene expression and protein profiles. How gene expression regulated by different levels of pH stress was not well studied. Therefore, they wanted to investigate the gene expression pattern of E *coli* under both acid and base condition at low, neutral, and high external pH stress. **What are the hypothesis and main elements?**

- **Null hypothesis:** Expression level of 'a gene' is same between pH conditions.
- **Alternative hypothesis:** Expression level of 'a gene' is different between pH conditions.
- **Experimental units:** RNA of E *coli*

Example: Maurer et al. 2005

pH Regulates Genes for Flagellar Motility, Catabolism, and Oxidative Stress in *Escherichia coli* K-12†

Lisa M. Maurer,[1] Elizabeth Yohannes,[1] Sandra S. Bondurant,[2] Michael Radmacher,[1] and Joan L. Slonczewski[1]*

Rationale: E *coli* and related enteric bacteria respond to a wide range of pH stress by regulating gene expression and protein profiles. How gene expression regulated by different levels of pH stress was not well studied. Therefore, they wanted to investigate the gene expression pattern of E *coli* under both acid and base condition at low, neutral, and high external pH stress. **What are the hypothesis and main elements?**

- **Null hypothesis:** Expression level of 'a gene' is same between pH conditions.
- **Alternative hypothesis:** Expression level of 'a gene' is different between pH conditions.
- **Experimental units:** RNA of E *coli*
- **Observation units (Response):** Gene expression level

Example: Maurer et al. 2005

pH Regulates Genes for Flagellar Motility, Catabolism, and Oxidative Stress in *Escherichia coli* K-12†

Lisa M. Maurer,[1] Elizabeth Yohannes,[1] Sandra S. Bondurant,[2] Michael Radmacher,[1] and Joan L. Slonczewski[1]*

Rationale: E *coli* and related enteric bacteria respond to a wide range of pH stress by regulating gene expression and protein profiles. How gene expression regulated by different levels of pH stress was not well studied. Therefore, they wanted to investigate the gene expression pattern of E *coli* under both acid and base condition at low, neutral, and high external pH stress. **What are the hypothesis and main elements?**

- **Null hypothesis:** Expression level of 'a gene' is same between pH conditions.
- **Alternative hypothesis:** Expression level of 'a gene' is different between pH conditions.
- **Experimental units:** RNA of E *coli*
- **Observation units (Response):** Gene expression level
- **Factors**: pH treatments (Low, neutral, high) and condition (base, acid)

Practice: using the course experiment

- Describe the experiement
- Null and Alternative hypothesis?
- Experimental units?
- Observation units?
- Factors?

Common Design Problems

- Experimental variation may mask the factor effects.
  - For data with larger variation, it is more difficult to detect mean differences between two levels of a factor.
- Uncontrolled factors compromised the conclusion such as confound with the factor that you plan to test.
  - *RNA samples with treatment A was run in one batch (or time 1), and RNA samples with treatment B was run in another batch (or time 2).*
- When multiple factors are involved and tested, one-factor at a time design will not work.

Principles of DOE

Principles of Design of Experiments

Four commonly considered principles in the design of experiment (Fisher1935).

- **Representativeness:** Are the experimental units used in the experiment sufficient to represent the conclusion to be made?
- **Randomization:** Help to avoid unknown or systemic bias.
- **Replication:** Increase the precision of the data.
- **Error control or blocking:** Help to reduce known bias (e.g. batch effect).

Principles of Design of Experiments

Four commonly considered principles in the design of experiment (Fisher1935).

- **Representativeness:** Are the experimental units used in the experiment sufficient to represent the conclusion to be made?
- **Randomization:** Help to avoid unknown or systemic bias.
- **Replication:** Increase the precision of the data.
- **Error control or blocking:** Help to reduce known bias (e.g. batch effect).

Experiment needs to be comparative.

Principles of Design of Experiments

Four commonly considered principles in the design of experiment (Fisher1935).

- **Representativeness:** Are the experimental units used in the experiment sufficient to represent the conclusion to be made?
- **Randomization:** Help to avoid unknown or systemic bias.
- **Replication:** Increase the precision of the data.
- **Error control or blocking:** Help to reduce known bias (e.g. batch effect).

Experiment needs to be comparative.

Fisher R.A., 1935 The Design of Experiment, Ed. 2nd Oliver & Boyd, Edingburgh

Representative

"There's a flaw in your experimental design. All the mice are scorpios."

Representative

Can the experimental units allow you to draw the right inference for the hypothesis?

Representative

Can the experimental units allow you to draw the right inference for the hypothesis?

Example:

Representative

Can the experimental units allow you to draw the right inference for the hypothesis?

Example:

- Study objective: To identify genes with expression changes after treatment A in liver patients.

Representative

Can the experimental units allow you to draw the right inference for the hypothesis?

Example:

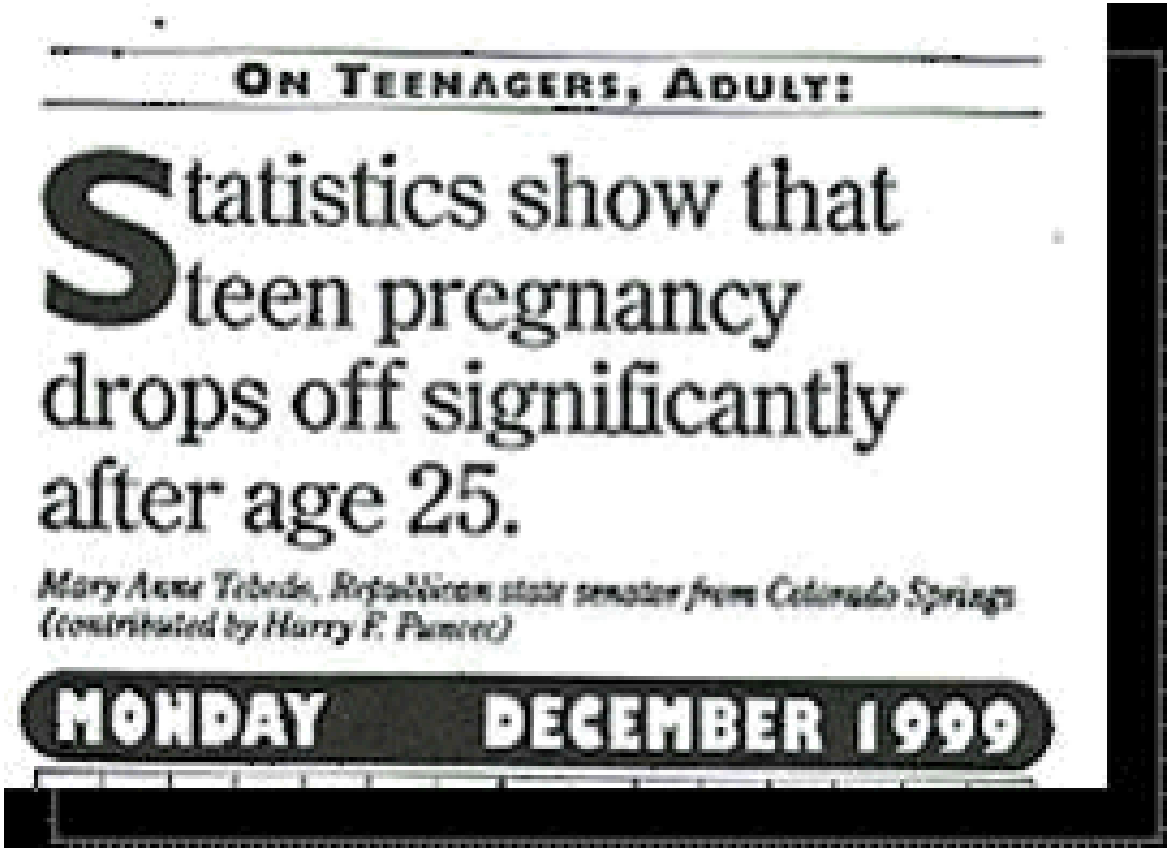
- Study objective: To identify genes with expression changes after treatment A in liver patients.
- Experimental units: Liver tissues were obtained from male liver patients before and after treatment for RNA-Seq study.

Representative

Can the experimental units allow you to draw the right inference for the hypothesis?

Example:

- Study objective: To identify genes with expression changes after treatment A in liver patients.
- Experimental units: Liver tissues were obtained from male liver patients before and after treatment for RNA-Seq study.
- Problem: The inference derived from this experiment cannot be applied to all liver patients. The experimental units are not representative to all liver patients.

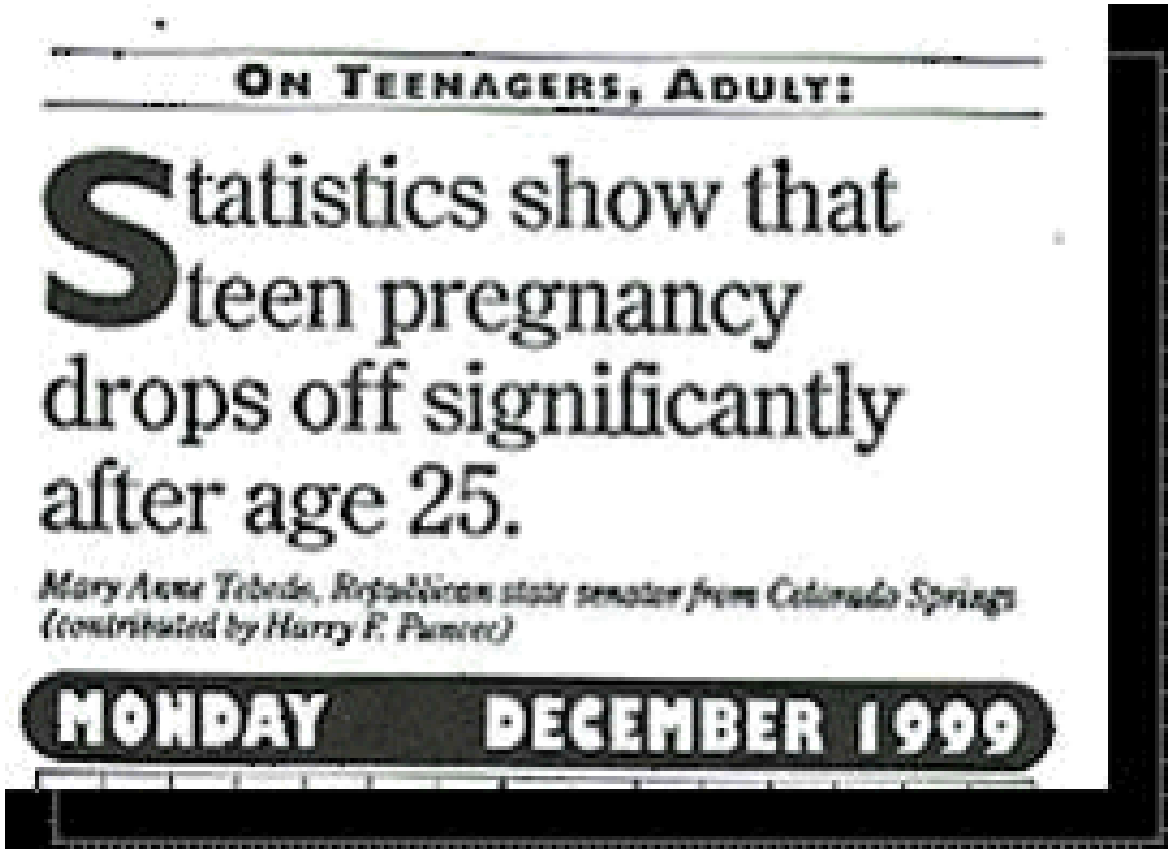


ON TEENAGERS, ADULT:

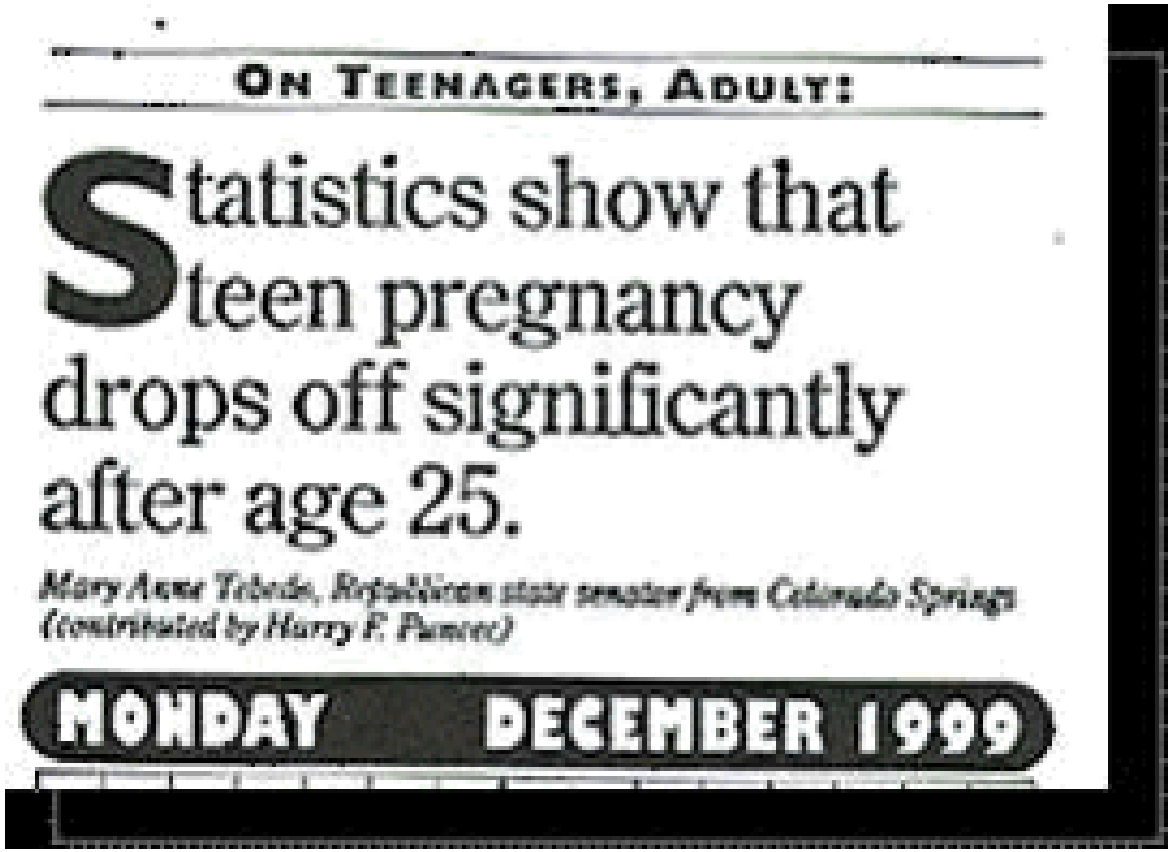
Statistics show that teen pregnancy drops off significantly after age 25.

Mary Anne Tebedo, Republican state senator from Colorado Springs (contributed by Harry F. Pamer)

MONDAY DECEMBER 1999



What problem do you see?


ON TEENAGERS, ADULT:

Statistics show that teen pregnancy drops off significantly after age 25.

Mary Anne Tebedo, Republican state senator from Colorado Springs (contributed by Harry F. Pumeo)

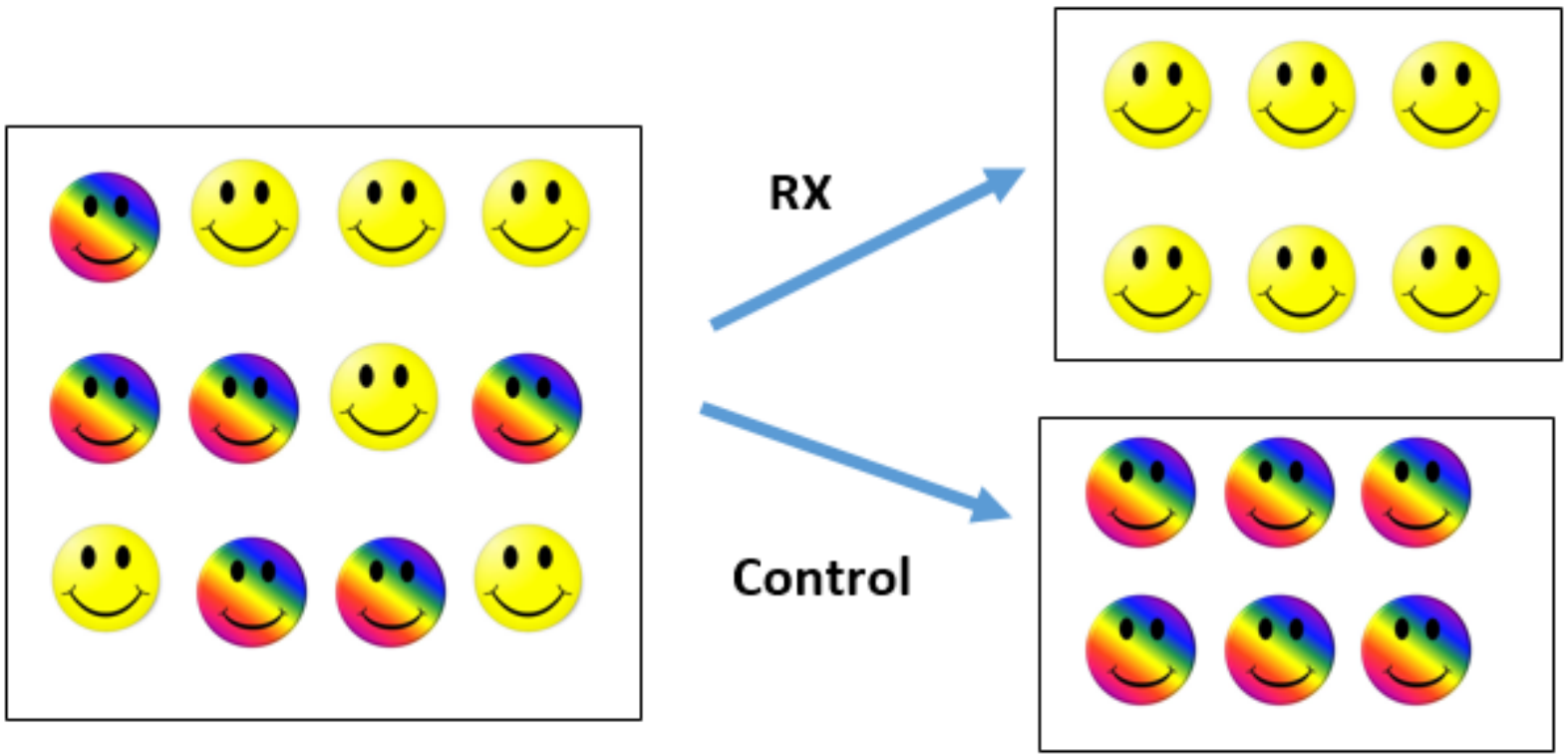
MONDAY DECEMBER 1999

What problem do you see?

The study subjects may not be teens.

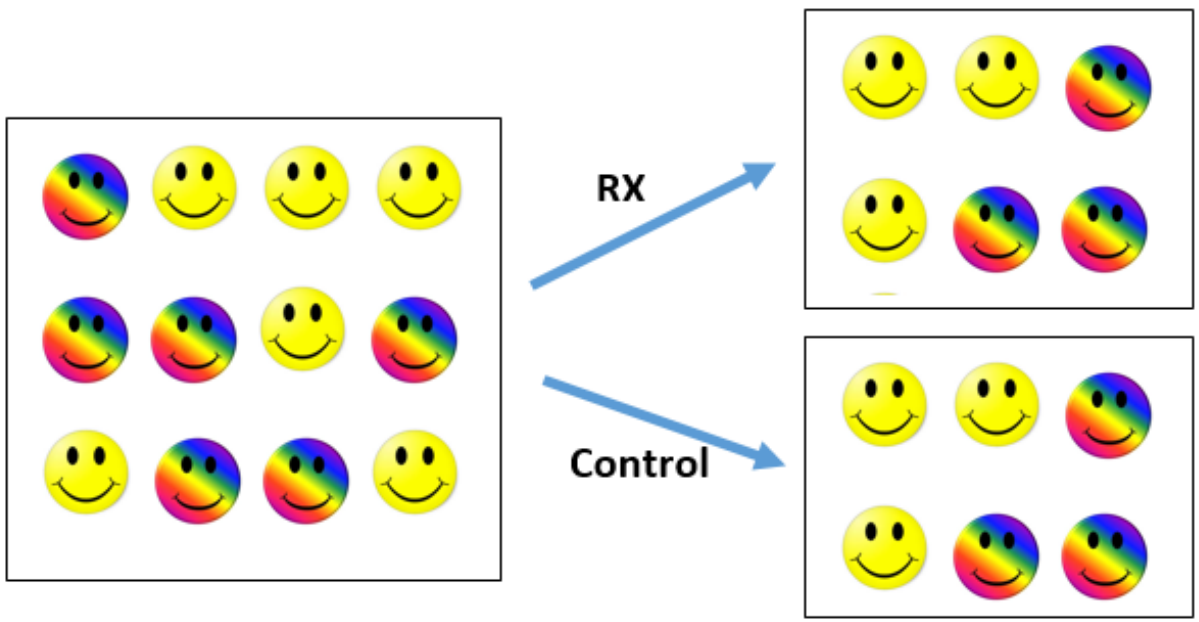
Randomization

Can the following design detect the drug effect?



Randomization

- Each experimental unit should have an equal chance to be assigned to a treatment group or block
- Prevent the introduction of systematic bias into the response of the experiment.
- Allow estimating experimental error.



Replications and Blocking

- **Replications:** Essential for controlling data variation. Why?
  - Observed data: $(Y_1, Y_2, \cdots, Y_n) \sim N(\mu, \sigma^2)$.
  - μ and σ^2 are unknown population parameters.
  - Estimates: $\hat{\mu} = \bar{Y}$ (sample mean) and $\hat{\sigma^2} = S^2$ (sample variance)
  - Standard error of the mean $= \sqrt{S^2/n}$, which determines the confidence interval (CI) of $\hat{\mu}$.
  - larger n (more replications) $\rightarrow$ narrower CI $\rightarrow$ more precision in mean estimate.

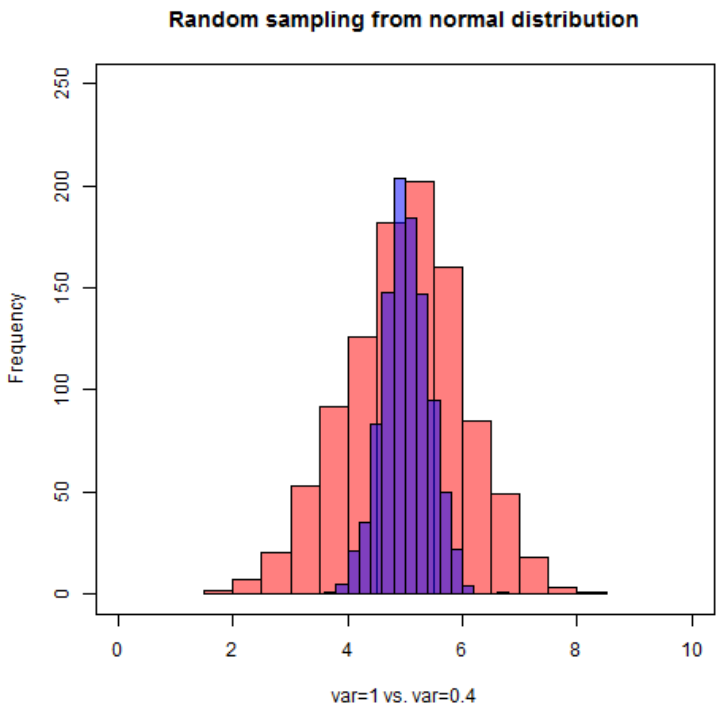
Replications and Blocking

- **Replications:** Essential for controlling data variation. Why?
  - Observed data: $(Y_1, Y_2, \cdots, Y_n) \sim N(\mu, \sigma^2)$.
  - μ and σ^2 are unknown population parameters.
  - Estimates: $\hat{\mu} = \bar{Y}$ (sample mean) and $\hat{\sigma^2} = S^2$ (sample variance)
  - Standard error of the mean $= \sqrt{S^2/n}$, which determines the confidence interval (CI) of $\hat{\mu}$.
  - larger n (more replications) $\rightarrow$ narrower CI $\rightarrow$ more precision in mean estimate.
- **Blocking:**
  - Include other factors that contribute to the unwanted variation in the design.
  - By blocking, we can reduce the source of variation.
  - Reduced standard error $\rightarrow$ narrower CI $\rightarrow$ more precision in mean estimate.

Accuracy & Precision

A well design experiment should generate high quality data.

- **Accuracy:**
  - Focus on if a method or technique produces measurements that are close to the true values.
  - Minimise measurement bias.
  - Microarry vs. RNA-Seq
- **Precision:**
  - Emphasize on smaller variation of the data
  - Lower variation, higher precision because measurements are closer to the mean.



Basic Statistics for DOE

Population and Samples

- **Population:** All possible items or units from an experimental or observational condition.
- **Samples:** A group of observation taken from a population.
- **We are interested in making inference in the population.**

Example:

- All cancer patients in the US vs. cancer patients in Duke hospital
- Tumor vs. tumor cells extracted for an experiment

Random variable

- **Random variable (Y):** A variable whose possible values are subject to variation, such as the responses obtained in an experiement
  - Quantitative: continuous measures
  - Qualitative: Binary, categorical, counts
- For observed data $y_i, i = 1, \cdots, n$

$$y_i = \mu + \epsilon_i, i = 1, \cdots, n$$

  - μ: unknown population parameter of interest.
  - ϵ: random and unobserved variable; $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$ are independent and follow a normal distribution $N(0, \sigma^2)$.
  - $Var(\epsilon) = \sigma^2 = Var(Y)$, an unknown population parameter

Illustration

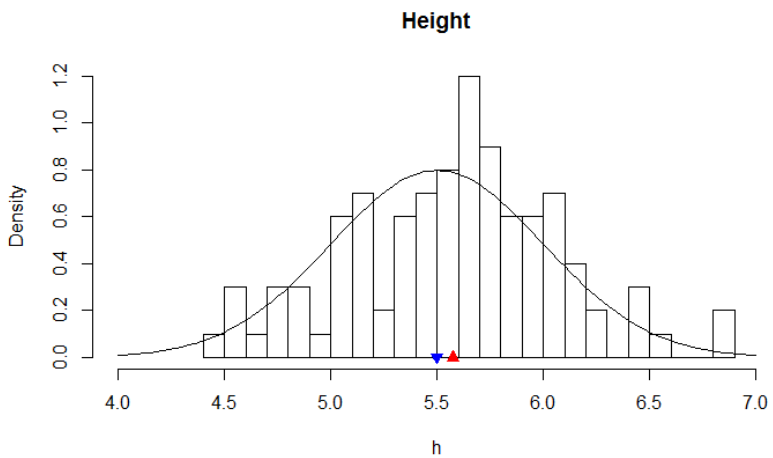
For a random variable y, y_i is the i^{th} observed value, $i = 1, \cdots, n$

- **Sample mean** $\bar{y} = \frac{\sum_i^n y_i}{n}$
- **Sample variance** $S^2 = \frac{\sum_i^n (y_i - \bar{y})^2}{n-1}$

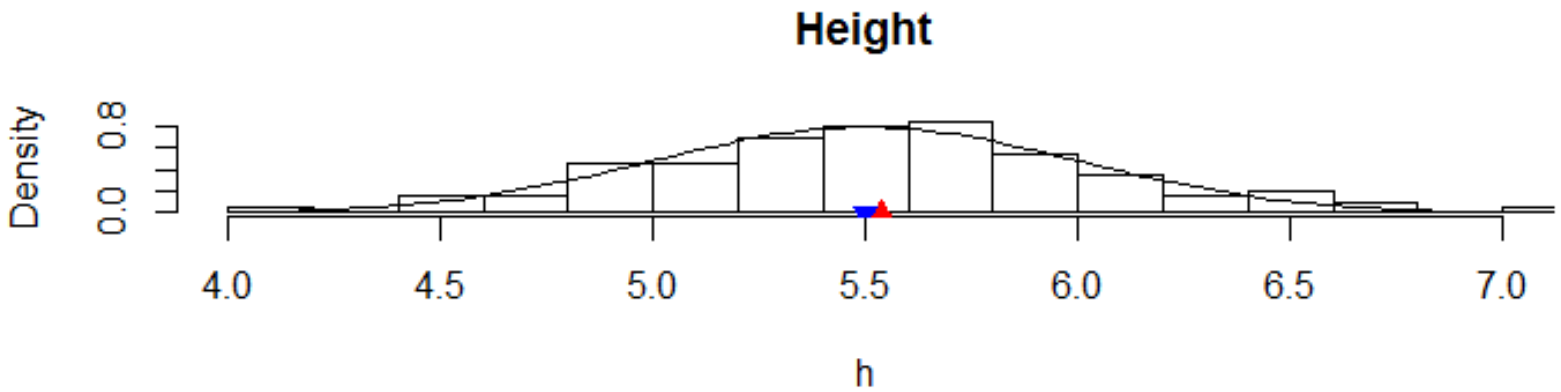
Example: Assume the true distribution of the height of high school Seniors is a normal distribution $N(\mu = 5.5, \sigma^2 = 0.25)$. We randomly survey 100 students for their height.

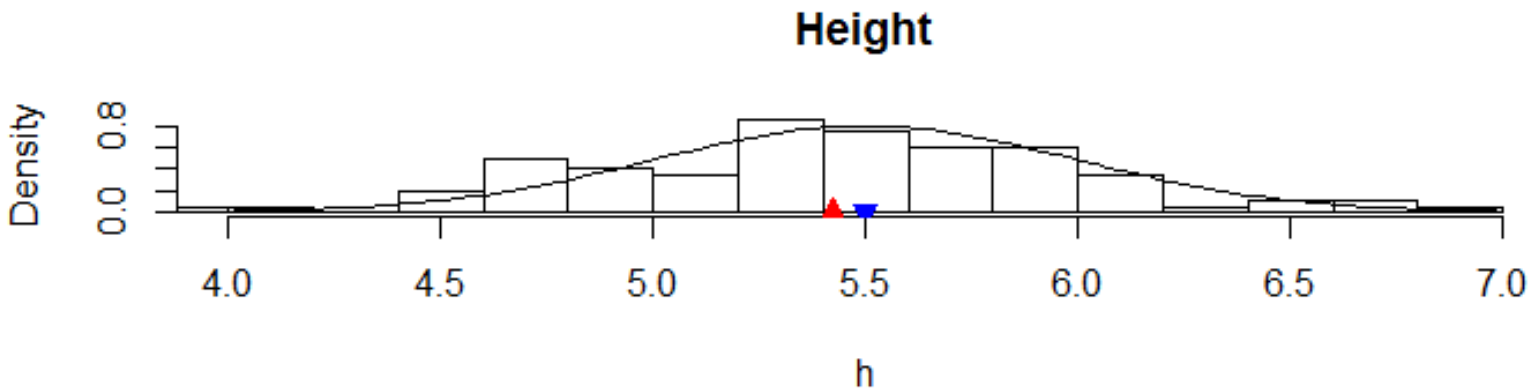
Average height, $\bar{y} = 5.57$

Sample variance, $S^2 = 0.2495$



Sample mean and variance changed by different set of sampling.





Example: height of the high school Seniors

If we survey 20, 100, and 500 students, can we make a good inference for the student height?

- Assume 10,000 random samples from $N(5.5, 0.25)$ as the 'population' of the high school students.
- Randomly draw 20, 100, and 500 values from the population (10,000 data points).

Sample size, n	20	100	500
Sample Mean	5.458	5.509	**5.493**
Sample Variance	0.297	0.191	**0.241**

Example: height of the high school Seniors

If we survey 20, 100, and 500 students, can we make a good inference for the student height?

- Assume 10,000 random samples from $N(5.5, 0.25)$ as the 'population' of the high school students.
- Randomly draw 20, 100, and 500 values from the population (10,000 data points).

Sample size, n	20	100	500
Sample Mean	5.458	5.509	**5.493**
Sample Variance	0.297	0.191	**0.241**

Sample size , sample size!

- Critical for precision of estimates
- Critical for statistical power in hypothesis testing

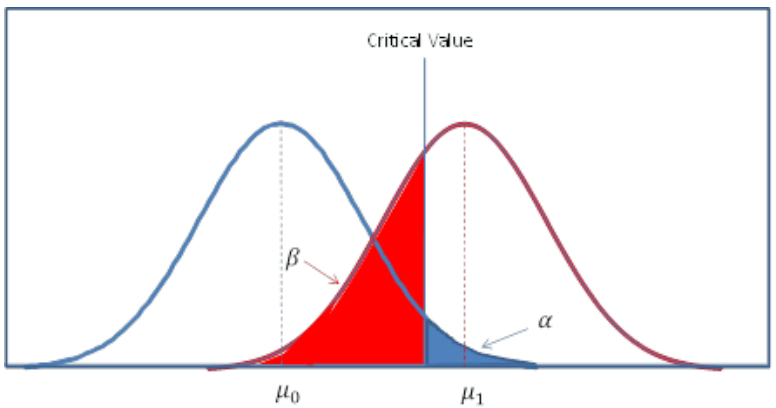
Review statistical power

		Null Hypothesis (H_0)	
		True	**False**
Test Decision	**Reject** (***Significant p***)	Type I error (α) False Positive (FP)	Correct inference True Positive (TP)
	Fail to reject (***Not significant p***)	Correct inference True Negative (TN)	Type II error (β) False Negative (FN)

$$\text{Power} = 1 - \beta$$

Power and Sample Size

A well-designed study should have sufficient statistical power.



- Determine what test statistics to be used for the hypothesis testing.
- Assume a two-sample t-test, the effect size is

$$\Delta = \frac{|\mu_0 - \mu_1|}{\sigma}$$

- The sample size is $n = 2\frac{(Z_{1-\alpha}+Z_{1-\beta})^2}{\Delta^2}$
- The larger the effect size, the smaller n.

Key elements for power calculation: (1) study design; (2) test statistics; (3) some ideas of target 'effect size' to be detected.

Consideration behind analysis methods

- Types of experimental design (or study design).
- Types of dependent variable:
  - continuous or discrete data
  - binary or categorical
  - distribution of the data
- Types of independent variable: continuous vs. categorical
- Covariates to be considered for adjustment.

Example:

A RNA-Seq experiment was performed to investigate the gene expression profile of E *E Coli* under different levels of pH stress.

- Dependent variable: Gene expression
- Independent variable: pH condition (multiple categories)

Types of Designs

Completely Randomized Design (CRD)

- Assume homogenous experimental units.
- Factor considered is 'categorical'. It can be two or multiple levels/groups.
  Example: Placebo vs. treatment group; multiple level of drug doses, different PH levels.
- **Randomization:** Each experimental unit has an equal likely chance to be assigned to a treatment group. Assume t treatment groups and n experimental units per group, totally nt experimental units.
  1. Label experimental units 1 to nt.
  2. Generate a random number for each experimental unit (keep the label and random number paired).
  3. Rank the random number, and the first n units go to treatment 1, 2nd set of n units go to treatment 2, etc.

Example: Plan to randomly assign two different treatments to 10 bacteria samples before RNA extraction.

- Designate sample ID number 1 to 10.
- Use seed number, 78201281, to generate 12 random numbers (x) between 0 and 1 for each sample.
- Sort x from low to high
- Assign the first 6 to treatment 1.

Randomized Using 78201281

Units	X	Trt
5	0.16201	1
2	0.24756	1
4	0.35811	1
6	0.39489	1
10	0.60694	1
9	0.63561	2
8	0.82158	2
7	0.89661	2
1	0.89714	2
3	0.91112	2

Measurements of variation

1. n samples obtained from one group:
   Within group variation: $S^2 = \frac{\sum_i^n (y_i - \bar{y})^2}{n-1}$
2. t treatment groups, n samples per group:
   Between treatment variation:

$$MST = \frac{n \sum_i^t (\bar{y_{i.}} - \bar{y})^2}{t-1}$$

Within treatment variation:

$$MSE = \frac{\sum_i^t \sum_j^n (y_{ij} - \bar{y_{i.}})^2}{t(n-1)}$$

Data analysis for CRD

Dependent variable: Gene expression level (y_{ij})
Independent variable: Treatment group (β_i)

Model: $y_{ij} = \mu + \beta_i + \epsilon_{ij}$, $i = 1, \cdots, t$ and $j = 1, \cdots, n$

Analysis of variance (ANOVA)Table:

Source	df	Mean SS (MS)	F
Treatment	$t-1$	MST	$\frac{MST}{MSE}$
Error	t(n-1)	MSE	

$$F = \frac{\text{Variation between treatments}}{\text{Variation within treatment}},$$

following an F distribution with d.f. of $(t-1, t(n-1))$.

one-way ANOVA example: PlantGrowth

```
plant <- PlantGrowth
plant
```

```
##    weight group
## 1    4.17  ctrl
## 2    5.58  ctrl
## 3    5.18  ctrl
## 4    6.11  ctrl
## 5    4.50  ctrl
## 6    4.61  ctrl
## 7    5.17  ctrl
## 8    4.53  ctrl
## 9    5.33  ctrl
## 10   5.14  ctrl
## 11   4.81  trt1
## 12   4.17  trt1
## 13   4.41  trt1
## 14   3.59  trt1
## 15   5.87  trt1
## 16   3.83  trt1
## 17   6.03  trt1
## 18   4.89  trt1
## 19   4.32  trt1
## 20   4.69  trt1
## 21   6.31  trt2
## 22   5.12  trt2
## 23   5.54  trt2
## 24   5.50  trt2
## 25   5.37  trt2
## 26   5.29  trt2
## 27   4.92  trt2
```

PlantGrowth dataset in R for plant yield (dried weight of plants) of 30 plants, which were randomized to three treatment groups (control, treatment 1, treatment 2).

```
##              Df    Sum.Sq   Mean.Sq  F.value      Pr..F.
## plant$group   2   3.76634 1.8831700 4.846088 0.01590996
## Residuals    27 10.49209 0.3885959       NA          NA
```

CRD Pros and Cons

- **Pros:**
  - Easy to randomize experimental units
  - Simple statistical analysis: two sample t-test for two treatment groups or one-way ANOVA for multiple treatment groups.
  - Flexible in terms of number of experimental units per groups (equal or unequal number per group).
- **Cons:** Can't control the differences between experimental units prior to the randomization.
  Example: If there are more females than males in the study,
  - CRD cannot control the gender effect.
  - Provide incorrect representation of the results, such as drawing the conclusion for both males and female.
- For CRD, it is better to have homogenous experimental units or large sample size.

Randomized Completed Block Design(RCBD)

When experimental units are not uniform $\cdots$

- Probably most frequently used design
- **Goal**: Minimize the effect of nuisance factors to the observation units.
- **Types of nuisance factors**: males, females, different technicians, different days(time) of experiment, etc.
- Restrict randomization to homogenous blocks.
- Block is usually treated as a random effect.

How the RCBD works?

- Identify the nuisance factor to be controlled – block.
- Sort experimental untis into homogeneous batches (blocks). The experimental units within each batch is as uniform as possible.
- Proceed with CRD within each block: randomly assign treatments to experiments units within each block.
- **Model:** Factors to considered: blocks (β_i), treatments (τ_j). ANOVA model:

$$y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij},$$

where $i = 1, \cdots, b$ for blocks, $j = 1, \cdots, t$ for treatments, and $\epsilon_{ij} \sim N(0, \sigma^2)$

Analysis for RCBD

Assume b blocks and samples are randomized to t treatments within each block.

Model: $y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}$, where $i = 1, \cdots, b$ (b blocks), and $j = 1, \cdots, t$ (t treatments).

ANOVA Table:

Source	df	MS	F
Block	$b-1$	MSB	$\frac{MSB}{MSE}$
Treatment	$t-1$	MST	$\frac{MST}{MSE}$
Error	$(b-1)(t-1)$	MSE	

MSB : variation between blocks

MST : variation between treatments

MSE : variation within the same block and same treatment

- **For simplicity, last RCBD model was written as**

$$y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}$$

, where $i = 1, \cdots, b$ (b blocks), and $j = 1, \cdots, t$ (t treatments).

- **For simplicity, last RCBD model was written as**

$$y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}$$

, where $i = 1, \cdots, b$ (b blocks), and $j = 1, \cdots, t$ (t treatments).

- **What is missing in the model above?**

- **For simplicity, last RCBD model was written as**

$$y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}$$

, where $i = 1, \cdots, b$ (b blocks), and $j = 1, \cdots, t$ (t treatments).

- **What is missing in the model above?**

Only one sample per block per treatment. **No repeat tests** in each block-treatment combination

- **For simplicity, last RCBD model was written as**

$$y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}$$

, where $i = 1, \cdots, b$ (b blocks), and $j = 1, \cdots, t$ (t treatments).

- **What is missing in the model above?**

  Only one sample per block per treatment. **No repeat tests** in each block-treatment combination

- Repeats per block-treatment combination allow us to estimate experiment-error variability. The model can be re-written as:

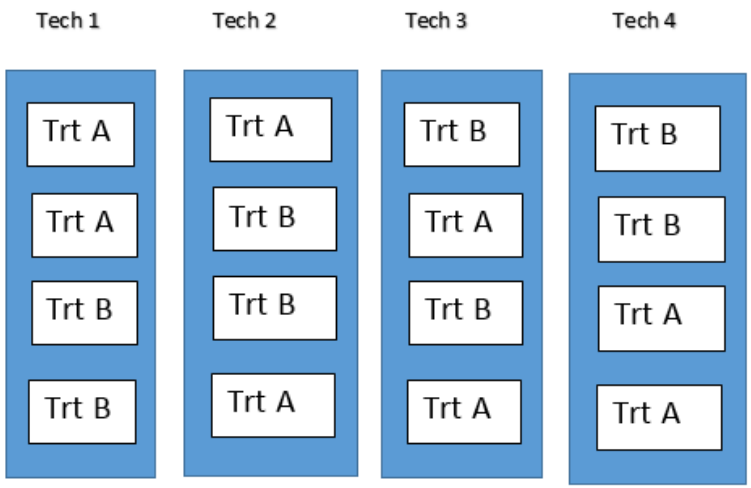
$$y_{ijk} = \mu + \beta_i + \tau_j + \epsilon_{ijk},$$

where $k = 1, \cdots, k$ repeats in each block-treatment combination.

Illustration

Assume 4 technicians working on a sequencing study. To control for the variation among technicians, we can consider each technician as a homogenous block.

- Randomly assign 4 samples to each technician for RNA extraction (*i.e.* 4 samples per block).
- Randomly assign two treatments to samples handled by each technician (within each block).
- Two repeats are for each block-treatment combination.



Two-way ANOVA example: Stress redcution example

```
stress <- read.csv(file = "./data/stress.csv")
stress <- data.frame(stress)
stress
```

```
##    Treatment   Age StressReduction
## 1     mental young               10
## 2     mental young                9
## 3     mental young                8
## 4     mental   mid                7
## 5     mental   mid                6
## 6     mental   mid                5
## 7     mental   old                4
## 8     mental   old                3
## 9     mental   old                2
## 10 physical young                9
## 11 physical young                8
## 12 physical young                7
## 13 physical   mid                6
## 14 physical   mid                5
## 15 physical   mid                4
## 16 physical   old                3
## 17 physical   old                2
## 18 physical   old                1
## 19  medical young                8
## 20  medical young                7
## 21  medical young                6
## 22  medical   mid                5
## 23  medical   mid                4
## 24  medical   mid                3
## 25  medical   old                2
```

27 subjects from three age groups (young, mid, and old ages) were studied for stress reduction by three types of stress redcution treatments (mental, physical, and medical).

```
res <- anova(lm(StressReduction ~ Treatment + Age, data = stress))
res <- data.frame(res)
res
```

```
##            Df Sum.Sq    Mean.Sq F.value       Pr..F.
## Treatment   2     18  9.0000000      11 4.882812e-04
## Age         2    162 81.0000000      99 1.000000e-11
## Residuals  22     18  0.8181818      NA           NA
```

In this example, $b = 3$ for age groups, $t = 3$ for treatment groups, and $k = 3$ for repeats within each block-treatment combination.

RCBD Pros and Cons

- **Pros:**
  - Good for comparing treatment effect when there is one nuisance factor to worry about.
  - Easy to construct the experiment
  - Simple statistical analysis – ANOVA
  - Flexible for any numbers of treatments and blocks.
- **Cons:**
  - It can only control variability from one nuisance factor.
  - Since it requires homogenous blocks, it is better for a study with a small number of treatments (factor levels) to test.
  - It requires the number of experimental units $\geq$ the number of factor-level combinations of interest.

Latin-Square Design

- RCBD can only control one nuisance factor (*i.e.* blocking factor). Latin-square design can control for two sources of variation (two blocking factors).
- Primarily use to test one factor of interest.
- If more than one factors are considered, factor-level combinations are used.
- Assume no interaction between the factor(s) and two blocking factors.

How to construct Latin-Square Design?

Assume one factor with k levels.

1. Set up a $k \times k$ table.
2. Assign $A, B, \cdots, K$ to the cells in the first row.
3. For the 2nd to kth rows, place the first letter of the previous row to the last position, and shift others forward one position.
4. Randomly assign one block factor to rows and one block factor to columns.
5. Randomly assign the levels of factor and blocks to the letters, row positions, and column positions.

Example:

Assume a factors with 4 levels (A, B, C, D), and first block (T1) with 2 levels $(1, 2)$ and 2nd block (T2) with 4 levels.

Block T2 \ Block T1	1	2	2	1
1	A	B	C	D
4	B	C	D	A
2	C	D	A	B
3	D	A	B	C

Analysis model for Latin-Square Design

Let α for the 1st block factor ($B1$), β for the 2nd blocking factor ($B2$), and γ for the factor of interest (treatment).
Model: $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk}$, where $i = 1, \cdots, a$, $j = 1, \cdots, b$, and $k = 1, \cdots, t$.

ANOVA Table:

Source	df	MS	F
$B1$	$a-1$	MSB1	$MSB1/MSE$
$B2$	$b-1$	MSB2	$MSB2/MSE$
Treatment	$t-1$	MST	MST/MSE
Error	$(a-1)(b-1)-(t-1)$	MSE	

$MSB1$: variation between block $B1$
$MSB2$: variation between block $B2$
MST : variation between treatments
MSE : Residual variation

Example: OrchardSprays data in R

An experiment to assess the potency of various constituents of orchard sprays in repelling honeybees, using a Latin square design.

```
spray <- OrchardSprays
spray
```

```
##    decrease rowpos colpos treatment
## 1        57      1      1         D
## 2        95      2      1         E
## 3         8      3      1         B
## 4        69      4      1         H
## 5        92      5      1         G
## 6        90      6      1         F
## 7        15      7      1         C
## 8         2      8      1         A
## 9        84      1      2         C
## 10        6      2      2         B
## 11      127      3      2         H
## 12       36      4      2         D
## 13       51      5      2         E
## 14        2      6      2         A
## 15       69      7      2         F
## 16       71      8      2         G
## 17       87      1      3         F
## 18       72      2      3         H
## 19        5      3      3         A
## 20       39      4      3         E
## 21       22      5      3         D
## 22       16      6      3         C
## 23       72      7      3         G
```

Example: OrchardSprays data in R

- Treatment: 8 levels (no lime sulphur, 7 different concentrations of lime sulphur) to form a 8×8 latin-square design.
- Rowpos and colpos: two block factors.
- decrease: the response variable.

```
spray$block1 <- factor(spray$rowpos)
spray$block2 <- factor(spray$colpos)
res <- anova(lm(decrease ~ block1 + block2 + treatment, data = spray))
res <- data.frame(res)
res
```

```
##           Df    Sum.Sq   Mean.Sq   F.value       Pr..F.
## block1     7  4767.484  681.0692  1.788376 1.151081e-01
## block2     7  2807.234  401.0335  1.053048 4.100372e-01
## treatment  7 56159.984 8022.8549 21.066701 7.454922e-12
## Residuals 42 15994.906  380.8311        NA           NA
```

Other basic experimental designs

More experimental designs, not covered here;

- **Split-plot design:** Consider two treatments. Use CRD to assign the first treatment. Then, the 2nd treatment is randomly assign to each plot of the first treatment.
  - Sources to consider: Mean, Trt A, Error from A, Trt B, $A \times B$, Error from B.
- **Nested design:** A factor is nested within a level of the other factor.
  - Assume factors A and B have two levels. Factor B is nested within each level of factor A.
  - The levels of factor B do not need to be identifical between different levels of factor A.
  - No interaction term
- **Factorial design:** Consider a number of factors with the same level (*e.g.* . 2^N factorial for two levels, 3^N for three levels)
- ...

Summary

- Outline a testable hypothesis.
- Be mindful on the data quality, accuracy and precision.
- A well-design experiment contributes significantly to the success of the research.
- Identify factor(s) of interest and nuisance factors to be controled to determine the type of experimental design to use.
- Follow the four key principles of experimental design.
- Statistical model should reflect to the experiemental design.

Reference Book: Planning, Contruction, and Statistical Analysis of Comparative Experiments, Francis G. Giesbrecht and Marcia L. Gumpertz (Wiley)