Experimental Design – Part I

Yi-Ju Li, Ph.D.

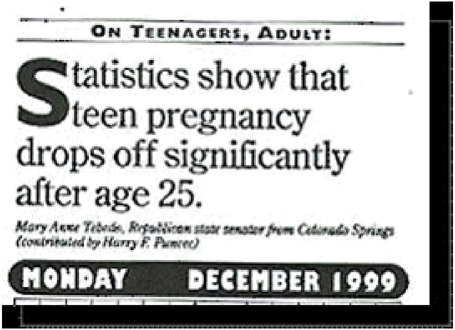
Department of Biostatistics & Bioinformatics
Duke University Medical Center

July 13, 2015

Outline

- 7/13 **(Monday) Part I**
  - Definition of Design of Experiment (DOE)
  - Basic statistics related to DOE
  - Types of designs
- 7/22 **(Tuesday next week) Part II**
  - Designs related to genetic research
  - Designs related to RNA-Seq

Contact Info: yiju.li@duke.edu



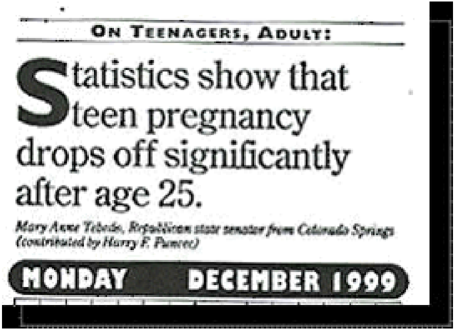
ON TEENAGERS, ADULT:

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

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

MONDAY DECEMBER 1999

What problems 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. Puncec)

MONDAY DECEMBER 1999

What problems do you see?

- Did they form the right hypothesis?
- Did they have the right pool of study subjects?

Section 1

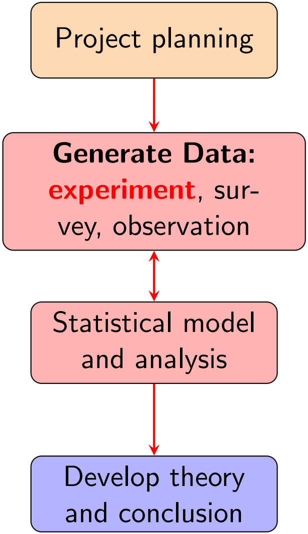
Definition of DOE

Definition of design of experiment (DOE)

- **Experiment:** A process that generates data to achieve specific objective
- **All data are subject to variation.**
- **Experimental design:** A plan for conducting an effective experiment that can
  - eliminate known sources of bias
  - prevent unknown source of bias
  - obtain precise data

  to answer predefined questions (hypothesis, theory, or model).

Stages of a study



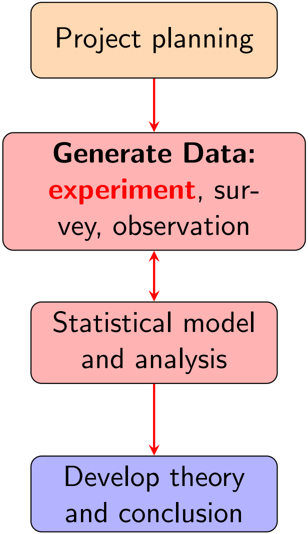
Project planning

Generate Data: **experiment**, survey, observation

Statistical model and analysis

Develop theory and conclusion

Stages of a study



Project planning

Generate Data: experiment, sur-vey, observation

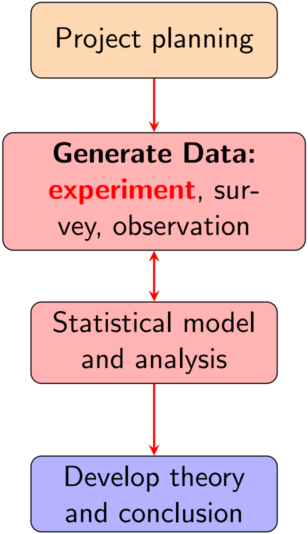
Statistical model and analysis

Develop theory and conclusion

Project planning

Hypothesis; what to be measured; and what influential factors are

Stages of a study


Project planning
Generate Data: **experiment**, survey, observation
Statistical model and analysis
Develop theory and conclusion

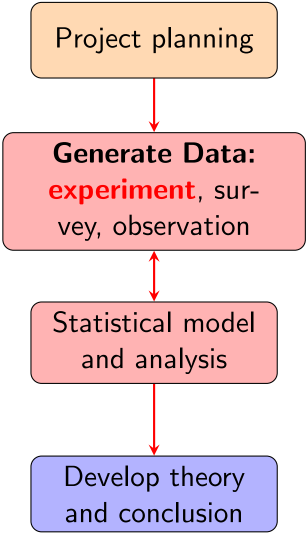
Project planning

Hypothesis; what to be measured; and what influential factors are

Experimental studies

Ability to control the source of variability

Stages of a study



Project planning

Generate Data: **experiment**, survey, observation

Statistical model and analysis

Develop theory and conclusion

Project planning

Hypothesis; what to be measured; and what influential factors are

Experimental studies

Ability to control the source of variability

Observational studies

No controls over the source of variability

Main elements

- Formulate problems and hypothesis
- **Experimental units:** The entities that experimental procedures (*e.g.* treatments) are applied to.
  - **Examples:** Mice, patients, plants, RNA, etc.
  - 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)
  - **Accuracy:** How close the measurement to the true value?
  - **Precision:** How much variation in the data?

More on main elements

- **Factors:** Variables to be investigated to determine its effect to the response variable (*e.g.* different treatment groups.)
  - It should be defined prior to the experiment.
  - It can be controlled by experimenter.
- **Covariate:** May affect the response but cannot be controlled in an experiment.
  - It is not affected by factors.

Formulate hypothesis

- Need to have a scientific question first
- Need to be able to translate the scientific question 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
- Therefore, need to know what statement you would like to make in your scientific question.

Examples: problems in establishing a hypothesis

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

   Concerns:

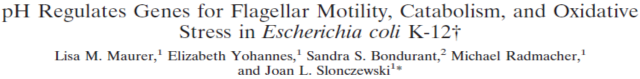
   - Examine = estimate rates? or Examine = compare rates?
   - Compare outcomes in disease X with some a fixed rate or with outcomes in another disease?

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

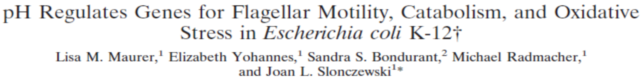


pH Regulates Genes for Flagellar Motility, Catabolism, and Oxidative Stress in *Escherichia coli* K-12†

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.

Example: Maurer et al. 2005



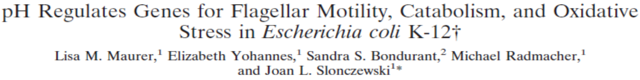
pH Regulates Genes for Flagellar Motility, Catabolism, and Oxidative Stress in *Escherichia coli* K-12†

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.

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

Example: Maurer et al. 2005



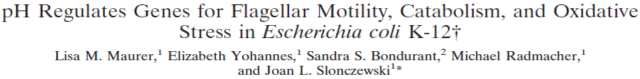
pH Regulates Genes for Flagellar Motility, Catabolism, and Oxidative Stress in *Escherichia coli* K-12†

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.

- **Null Hypothesis:** Expression level of 'a gene' is same between pH conditions.
- **Experimental units:** RNA of E *coli* grown under different pH treatments.

Example: Maurer et al. 2005



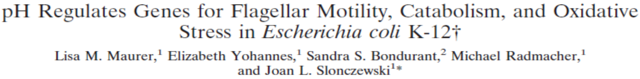
pH Regulates Genes for Flagellar Motility, Catabolism, and Oxidative Stress in *Escherichia coli* K-12†

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.

- **Null Hypothesis:** Expression level of 'a gene' is same between pH conditions.
- **Experimental units:** RNA of E *coli* grown under different pH treatments.
- **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.

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

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 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 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

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

Representative

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

Example:

Representative

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

Example:

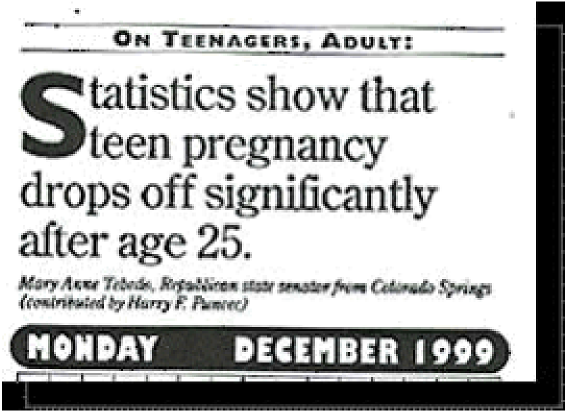
- Study objective: To identify genes with expression changes after treatment A in liver patients.
- Experimental units: Liver tissues were obtained from liver patients over age 50 before and after treatment for RNA-Seq study.

Representative

Can the experimental units used in the study 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 liver patients over age 50 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.

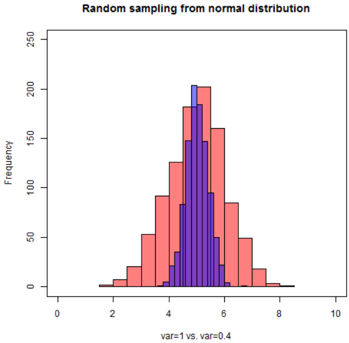
Mary Anne Tebedo, Republican state senator from Colorado Springs (contributed by Harry F. Pamrec)

MONDAY DECEMBER 1999

One possible scenario is that they tested the data from teens, but assuming that they can represent women with age > 25.

Accuracy & Precision

- **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.



Section 2

Basic Statistics

Population and Samples

- **Population:** All possible items or units from an experimental or observational condition.
- **Samples:** A group of observation taken from a 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
- **Mean and variance**
  - Mean, μ: Expected value of Y
  - Variance, σ^2: Expected variation of Y.

For an observed y_i,

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

ϵ is the error term that represents the difference between y_i and μ. $Var(\epsilon) = \sigma^2 = Var(Y)$.

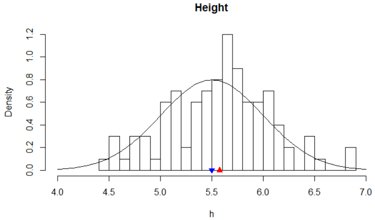
Illustration

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

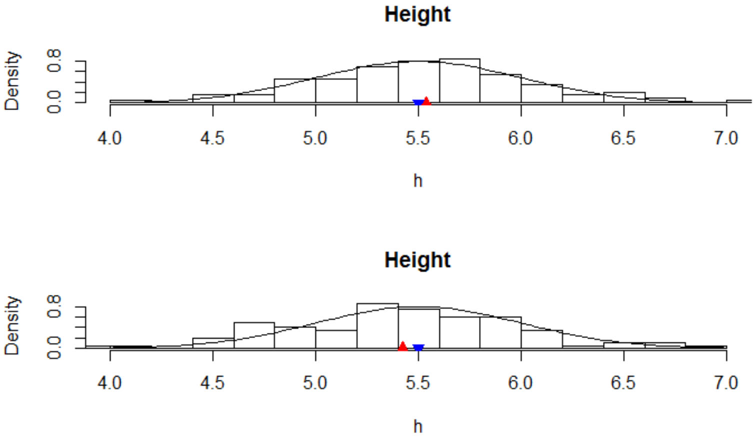
- **Sample mean** $\bar{y} = \frac{\sum_i^n y_i}{n}$
- **Sample variance** $S^2 = \frac{\sum_i^n (y_i - \bar{y})^2}{n-1}$

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 matters for good estimates!

A well-design experimental study is often interested in assessing the relationship between dependent and independent variables.

- **Dependent variable:** The random variable of the outcome measures from the experiment
- **Independent variable:** The variable doesn't change by other variables (*e.g.* age, gender, treatment group)

Example:

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

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

Consideration behind analysis methods

- Types of experimental design
- Types of dependent variable:
  - continuous or discrete data
  - binary or categorical
  - distribution of the data
- Types of independent variable: continuous vs. categorical

Section 3

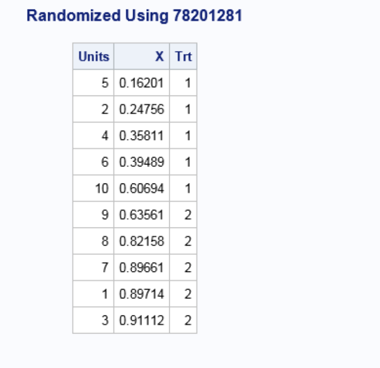
Types of Design

Completely Randomized Design (CRD)

- Assume homogenous experimental units.
- Treatment is assigned randomly to experimental units. That is, each experimental unit has an equal likely chance to be assigned to each treatment group.
- Assume t treatment groups and n experimental units per group, totally nt experimental units. One type of randomization:
    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 10 bacteria samples to grow under two treatment groups before RNA extraction.

- Designate sample ID number 1 to 10.
- Use a seed number, 78201281, to generate 10 random numbers (X) for each sample.
- Sort X from low to high
- Assign the first 5 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 $(t-1, t(n-1))$.

CRD Pros and Cons

- **Pros:**
  - Easy to randomize experimental units
  - Simple statistical analysis: one-way ANOVA
  - 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 exist.
  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 $i = 1, \cdots, b$, $j = 1, \cdots, t$, and $\epsilon_{ij} \approx N(0, \sigma^2)$

Data analysis for RCBD

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:

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

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

What is missing in the model described earlier?

For simplicity, last RCBD model:

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

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

What is missing in the model described earlier?

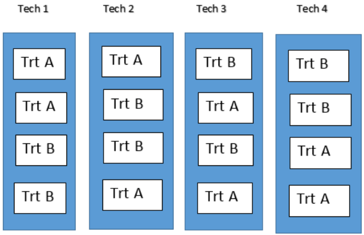
Only one sample per block per treatment. **No replicates**

Remember that RCBD can include replicates.

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).



Tech 1	Tech 2	Tech 3	Tech 4
Trt A	Trt A	Trt B	Trt B
Trt A	Trt B	Trt A	Trt B
Trt B	Trt B	Trt B	Trt A
Trt B	Trt A	Trt A	Trt A

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:**
  - Since it requires homogenous blocks, it is better for a study with a small number of treatments to test.
  - It can only control variability from one nuisance factor.

More experimental designs

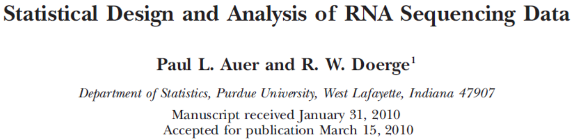
More experimental designs, not covered here;

- **Latin square designs:** When there are are two distinct criteria to assign experimental units into groups.
- **Split-plot design** Consider two different treatments and each have different dosage levels. Use CRD to assign samples to different dosage levels of the first treatment. Then, the dosage levels of 2nd treatment is randomly assign to nested within each plot (dosage level) of the first treatment.
    - Sources to consider: Mean, Trt A, Error from A, Trt B, $A \times B$, Error from B.
- **Factorial design:** Consider a number of factors with the same level (*e.g.* . 2^N factorial for two levels, 3^N for three levels)
- $\cdots$

Summary

- A well-design experiment contribute significantly to the success of the research.
- Establish a testable hypothesis that meets your scientific question(s).
- Be mindful on the data quality, accuracy and precision.
- Follow the four key principles of experimental design

Reading for the next lecture



Statistical Design and Analysis of RNA Sequencing Data

Paul L. Auer and R. W. Doerge[1]

Department of Statistics, Purdue University, West Lafayette, Indiana 47907

Manuscript received January 31, 2010
Accepted for publication March 15, 2010