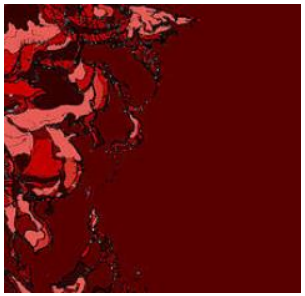
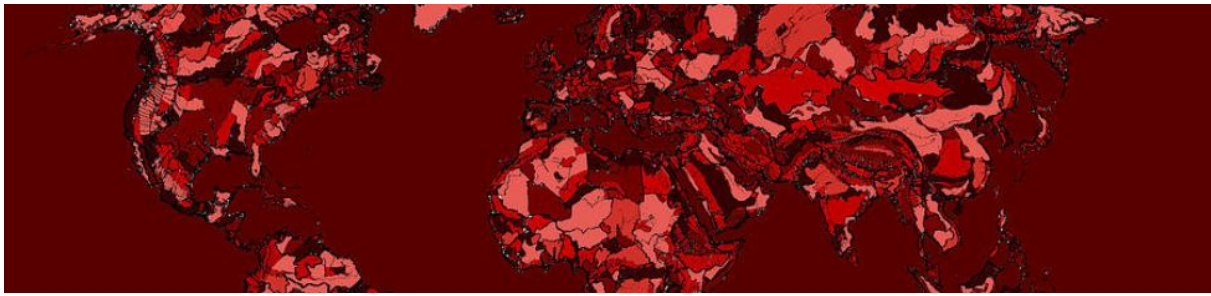
Insert the class / client / audience logo here

Math and Stats for MBA

Lecture note 3

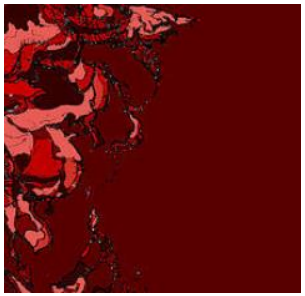
Swiss Institute of
Artificial Intelligence

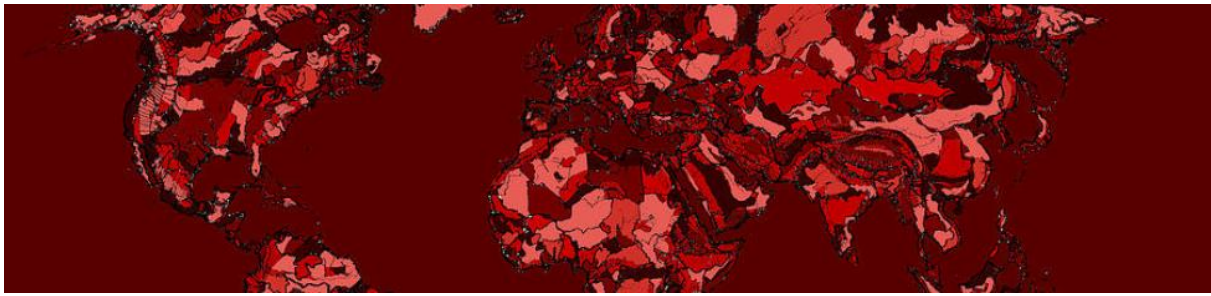




6. Categorical data

Swiss Institute of
Artificial Intelligence





Categorical data

- Categorical data and Simpson's paradox

Definition of categorical data

Categorical data is a collection of information that is divided into a number of groups. It can take on discrete values such as 1 for 'Yes' and 2 for 'No'.

■ **Simple example of categorical data**

	Outcome 1	Outcome 2
Treatment 1	A	B
Treatment 2	C	D

– From the table above, we can calculate "**risk**" of each outcome

– The risk of outcome 1 is $\frac{A}{A+B}$ for treatment 1 and $\frac{C}{C+D}$ for treatment 2. The **relative risk** is $\frac{A/(A+B)}{C/(C+D)}$.

– **Odd ratio** is $\frac{A/B}{C/D}$, comparing how frequently the outcome 1 occurs between both treatments.

■ **Why should we consider categorical data?**

– What are the variance of the categorical data?

– What happens when categorical variable is in the regression model? Are there any issue to calculate $\widehat{\boldsymbol{\beta}}$?

EXAMPLE – SIMPSON'S PARADOX

In the analysis of categorical data, we need to avoid Simpson's paradox caused by confounding factors.

■ **Two hospitals on a risky surgical procedure**

	Lived	Died	Survival rate
Hospital A	80	120	40%
Hospital B	20	80	20%

– Which hospital are you going to choose?

– Is hospital B really worse than hospital A?

■ **Titanic survival rates (survival rate by passenger class)**

	First	Second	Third	Crew	Total
Lived	203	188	178	212	711
Died	122	167	528	696	1513
Survival	62%	41%	25%	23%	32%

– What do you think about "Money can't buy everything"?

– Why do we need to delve into the dataset in a multiple perspectives?

■ **Why Simpson's paradox occurs?**

– The confounding factors due to disproportional sampling

Categorical data

- Categorical data and Simpson's paradox

EXAMPLE – SIMPSON'S PARADOX (cont.)

Suppose that we now have significant factors for both cases.

- **Two hospitals on a risky surgical procedure**

	Good condition			Bad condition		
	Lived	Died	Survival	Lived	Died	Survival
Hospital A	80	100	**44%**	0	20	**0%**
Hospital B	10	10	**50%**	10	70	**13%**

  – Do you still want to choose hospital A?
  – Why the confounding effect of patient condition matters?

- **Titanic survival rates (Survival rate by Class)**

	First	Second	Third	Crew	Total
Children	6	24	79	0	**109 (4%)**
Women	144	93	165	23	**425 (19%)**
Men	175	168	462	885	**1690 (75%)**

  – What does this result imply for?
  – As we know that the lifeboats were generally filled with women and children first. Does this fact affect your interpretation?
  – In terms of regression model, what would happen in this case?

EXAMPLE – SIMPSON'S PARADOX (cont.)

- **Confounding effect**
  – Suppose our one-variable model for Titanic example is like below:

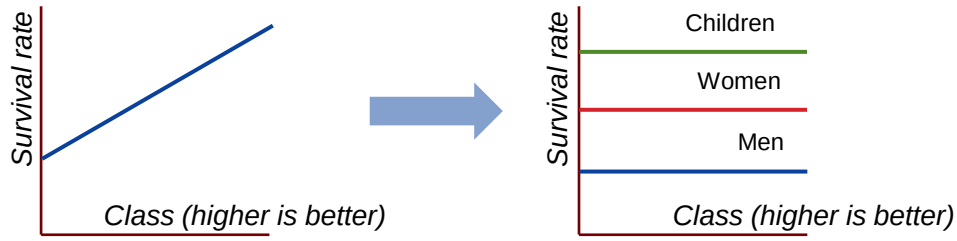
$$y_i = \beta_0 + \beta_1 \times Class_i + e_i$$

  – In this case, we might be able to draw a conclusion that Class significantly has impact on Survival Rate(y_i).

- **Controlling factors**
  – However, we found that there is confounding effect on y_i, the model would become:

$$y_i' = \beta_0 + \beta_1' \times Class_i + \beta_2 \times Gender_i + \beta_3 \times Age_i + \epsilon_i$$

  – Are both β_1 and β_1' would be same?
  – What would be happened in t statistic of β_1'?



Significance testing for categorical data

- The chi-square(χ^2) test

Non-parametric test

Why we need to study non-parametric test?

Z-test, t-test and F-test we covered are assuming that the population follows (asymptotically) normal distribution and homoskedasticity between two groups.

– What if the population does not follow normal distribution?

– What if the data we collect is nominal or ordinal variable?

■ **Test of independence**

– H_0 : each variable is independent each other (H_a : Not H_0)

■ **Test of goodness-of-fit (test of homogeneity)**

– H_0 : each variable is homogeneous (H_a : Not H_0)

# of samples	Parametric test	Non-parametric test	
k	ANOVA	χ^2-test	Kruskal-Wallis H test (Median)
2	t-test (mean diff.) F-test (variance diff.)	χ^2-test	Mann-Whitney test
1	Z-test, t-test	χ^2-test	Kolmogorov-Smirnov test
Scale	*Interval or Ratio*	*Nominal*	*Ordinal*

Chi-square test of independence

■ **χ^2 distribution**

– It consists of k random variables that follow squared standard normal distribution. (*i.i.d*)

$$X = Z_1^2 + Z_2^2 + \cdots Z_n^2 \ \ where\ X \sim \chi^2, Z \sim N(0,1)$$

– What is expected from squared standard normal distribution?

– **Test hypothesis :** Reject null hypothesis when χ^2 statistic is big enough compared to χ^2 critical value with *d.o.f*

– Note that χ^2 statistic assume that the data follows binomial or multinomial distribution. (sampling with replacement)

■ **One-sample test**

– Degrees of freedom : $(k-1)$

$$\chi^2 = \sum_{\#\ of\ groups} \frac{(observed\ - expected)^2}{expected}$$

■ **Two-sample test** *(similar to k-sample test)*

– Degrees of freedom : $(r-1)(c-1)$

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{(observed_i\ - expected_i)^2}{\sum expected_{ij}}$$

Significance testing for categorical data

- The chi-square(χ^2) test

Chi-square test of independence (cont.)

■ **Suppose we have a contingency table regarding the level of wage of data scientists for each specialty they have.**

Wage	Statistics	Engineering	Coding	Total
High	28	35	4	**67**
Middle	18	57	49	**124**
Low	5	13	91	**109**
Total	**51**	**105**	**144**	**300**

– $\boldsymbol{H_0}$: Both level of wage and specialty of data scientists are independent each other, which means that there are no relationship between the level of wage and the specialty. (H_a : Not H_0)

– χ^2 critical value with $(3-1)(3-1)$ degrees of freedom in 5% significance level is $\chi^2_{0.05}(3-1)(3-1) = 9.49$

– χ^2 statistic can be derived as below:

$$\chi^2 = \frac{(28-11.4)^2}{11.4} + \frac{(35-23.5)^2}{23.5} + \cdots + \frac{(91-52.3)^2}{52.3} = 116.2$$

– We have $\chi^2(=116.2) > \chi^2_c(=9.49)$, hence we can conclude that the specialty of data science have an influence on the level of wage by rejecting the null hypothesis.

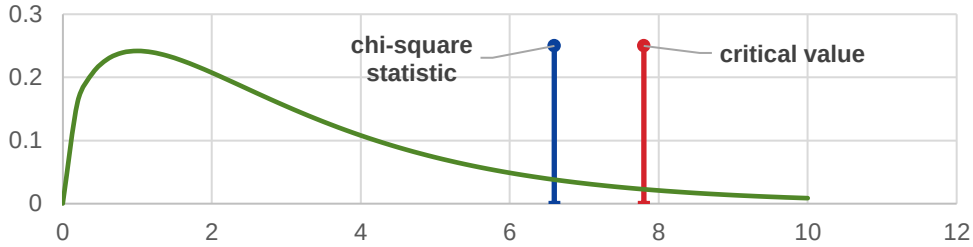
Chi-square test of goodness-of-fit (homogeneity)

■ **Goodness-of-fit (homogeneity) test is comparing each variable in order to check its homogeneity.**

	A	B	O	AB	Total
Male	21	98	555	1,440	2,114
Female	50	123	702	1,768	2,643
Total	71	221	1,257	3,208	4,757

– Suppose the contingency table comparing blood type by gender.

– $\boldsymbol{H_0}$: The distribution of blood type by gender in homogenous.



– According to the test statistic, we fail to reject null hypothesis.

■ **Dose χ^2 work for any case of categorical data analysis?**

– What if the data does not guarantee sampling with replacement?

– What if population is not big enough, or the expected frequency is less than 5? (the data will follow hypergeometric distribution)

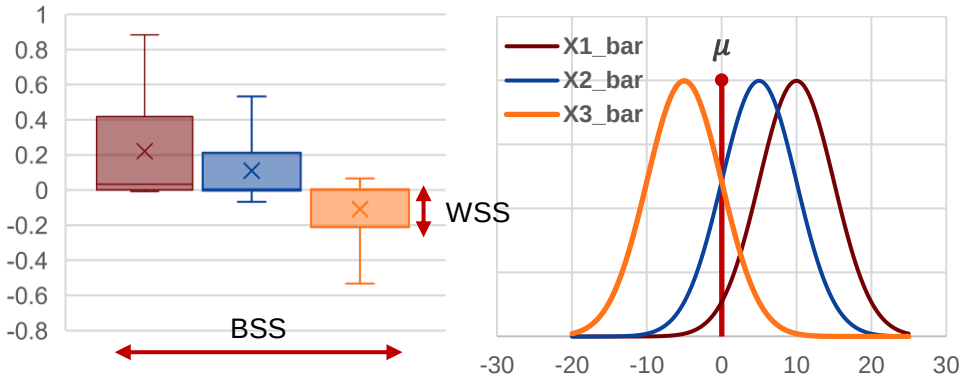
– What are *Fisher's Exact Test* and *Yates correction*?

Categorical inputs with continuous outputs

- ANOVA (Analysis of Variance)

ANOVA

- **Testing whether the continuous outputs depend on the input**
  - ANOVA test compares each mean of k groups, and variances within each group.



  - ANOVA assumes there are k group-specific means, and test whether all the means μ_k are equal. The model would be

$$y_i = \overbrace{\underbrace{\mu}_{global\ mean} + \underbrace{\tau_{x_i}}_{group-specific\ offset}}^{mean\ for\ this\ point's\ group(=\mu_{x_i})} + \underbrace{\epsilon_i}_{random\ noise}$$

  - The test reduces to seeing whether the offset τ are all 0.
- **Assumptions**
  - Identical variance between groups. (homoskedasticity)
    - Why homoskedasticity should be assumed in ANOVA test?
  - Output values for ANOVA are normally distributed.

ANOVA and F-test

- **ANOVA can be represented in a regression model**
  - Suppose we have dummy variables for each student group.

$$\boldsymbol{X} = \begin{pmatrix} \overbrace{Under-graduate}^{} & \overbrace{Master}^{} & \overbrace{Doctor}^{} \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

  - $\boldsymbol{X}$ above can be a perfect input for multiple regression. In this case, we can evaluate how well the model fits using F-test by comparing the variance both *explained* and *unexplained*.

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X})^2}_{TSS\ (Total\ sum\ of\ squares)} = \underbrace{n\sum_{i=1}^{k}(\bar{X}_i-\bar{X})^2}_{BSS\ (Between\ group)} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2}_{WSS\ (Within\ in\ group)}$$

$$\boldsymbol{F}_{(k-1,\,n-k)} = \frac{S_B^2\ (Between\ group\ variance)}{S_W^2\ (Within\ group\ variance)} = \frac{\frac{BSS}{k-1}}{\frac{WSS}{n-k}}$$

  - $\boldsymbol{H_0}: S_B^2 = S_W^2$ same as $\mu_1 = \mu_2 = \cdots = \mu_k$
  - Why F-test? (more t-tests, more errors / variance information)
  - How to test homoskedasticity of between groups? (Levene Test)
  - Post-hoc test can be used to see which group rejects the null
    - A series of independent t-tests comparing each group. Note that significance level must be adjusted for the aggregate test.

Categorical inputs with continuous outputs

- Extensions of ANOVA

Two-way ANOVA

■ **What if we are interested in measuring the effects of two input factors?**

$$y_i = \overbrace{\underbrace{\mu}_{global\ mean} + \underbrace{\tau_{x_i}}_{offset\ for\ factor\ 1} + \underbrace{\eta_{z_i}}_{offset\ for\ factor\ 2} + \underbrace{\gamma_{x_i z_i}}_{offset\ for\ factor\ 1}}^{mean\ for\ this\ point's\ group} + \epsilon_i$$

– We here include second input factor, z_i while the interaction term γ is also appended to the equation.

– We need to test three hypotheses for the factors on Row and Column respectively, as well as the interaction term for the two.

– The data table in two-way ANOVA will be

Column j (Factor B) \ Row i (Factor A)			
$Y_{11,1}$ $Y_{11,2}$ $Y_{11,3}$ ⋮	$Y_{12,1}$ $Y_{12,2}$ $Y_{12,3}$ ⋮	…	$\overline{Y_{1j}}$
$Y_{21,1}$ $Y_{21,2}$ $Y_{21,3}$ ⋮	$Y_{22,1}$ $Y_{22,2}$ $Y_{22,3}$ ⋮	…	$\overline{Y_{2j}}$
⋮	⋮	⋱	$\overline{Y_{+j}}$
$\overline{Y_{i1}}$	$\overline{Y_{i2}}$	$\overline{Y_{i+}}$	$\overline{Y_{++}}$

Two-way ANOVA

■ **The model of two-way ANOVA with interaction term is**

$$TSS = SS_A + SS_B + SS_{AB} + WSS$$

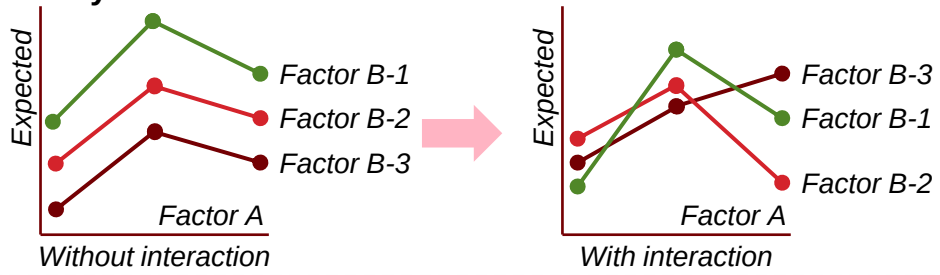
	Sum of Squared	DoF	Mean Sq.	F
SS_A	$cn\sum_i(\overline{Y_{i+}} - \overline{Y_{++}})^2$	$r-1$	$\frac{SS_A}{r-1}$	$F_A = \frac{MS_A}{MSW}$
SS_B	$rn\sum_j(\overline{Y_{+j}} - \overline{Y_{++}})^2$	$c-1$	$\frac{SS_B}{r-1}$	$F_B = \frac{MS_B}{MSW}$
SS_{AB}	$k\sum_i\sum_j(\overline{Y_{ij}} - \overline{Y_{i+}} - \overline{Y_{+j}} + \overline{Y_{++}})^2$	$(r-1)\times(c-1)$	$\frac{SS_{AB}}{(r-1)(c-1)}$	$F_{AB} = \frac{MS_{AB}}{MSW}$
WSS (error)	$\sum_i\sum_j\sum_k(Y_{ij,k} - \overline{Y_{ij}})^2$	$N-rc$	$\frac{WSS}{N-rc}$	
TSS	$\sum_i\sum_j\sum_k(Y_{ij,k} - \overline{Y_{++}})^2$	$N-1$		

– $H_{R,0}: \mu_{A,1} = \mu_{A,2} = \cdots = \mu_{A,r}$ testing with F_1

– $H_{R,0}: \mu_{B,1} = \mu_{B,2} = \cdots = \mu_{B,c}$ testing with F_2

– $H_{R,0}: \mu_{AB,1} = \mu_{AB,2} = \cdots = \mu_{AB,r}$ testing with F_3

■ **Why we should consider interaction between factors?**



Categorical inputs with continuous outputs

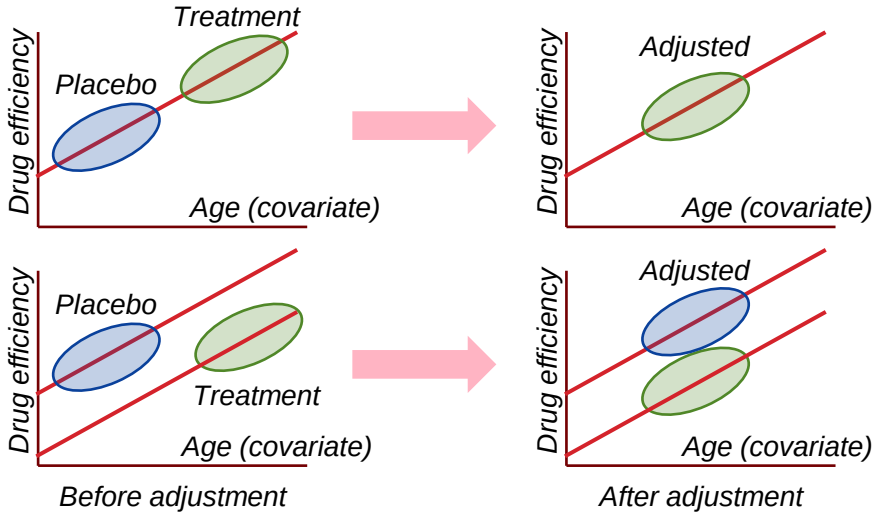
- Extensions of ANOVA

ANCOVA

- **What if we have both categorical and continuous variable?**

$$y_{ij} = \overbrace{\underbrace{\mu}_{global\ mean} + \underbrace{\tau_i}_{\substack{the\ effect\ of\ the\ i^{th}\ level\ of \\ independent\ variable}} + \underbrace{\boldsymbol{\beta}(x_{ij} - \bar{x})}_{\substack{covariate\ for\ j^{th}\ observation \\ under\ the\ i^{th}\ group}}}^{mean\ for\ this\ point's\ group} + \epsilon_{ij}$$

  - **ANCOVA** uses regression model to estimate the size of treatment effects given the covariate information.
  - *e.g)* Efficiency of a drug by two groups(treatment and placebo), controlling for the effect of patient age on outcome.
  - ANCOVA adjusts the dependent variable by covariates.
  - It is based on critical assumptions as like a regression model.



EXAMPLE – GROCERY STORE AWARENESS IN OXFORD

- **Bowlby(1979) investigated the grocery store awareness for households in Oxford.**

Car usage	Age group: The elder	Age group: Youth (Single)	Age group: Youth (Married)
Not at all	59.34 (17.77)	61.00 (13.16)	48.69 (87.24)
Usually	55.89 (75.19)	57.62 (39.23)	55.10 (84.57)
Always	53.98 (75.38)	59.35 (29.70)	59.72 (22.39)

* expected value (variance in parenthesis), some are modified only for this example

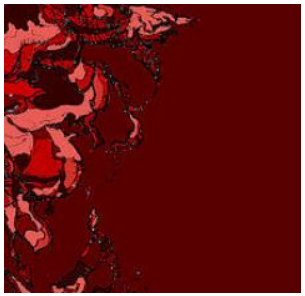
  - We can conduct two-way ANOVA including interaction effect.

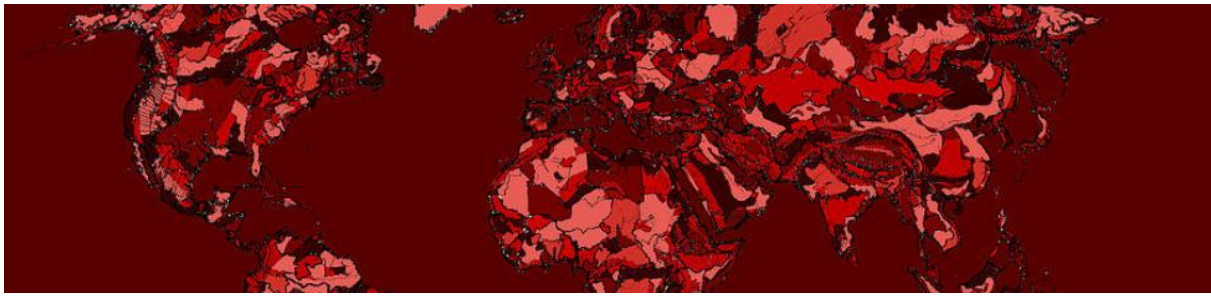
	Sum of Squared	DoF	Mean Sq.	F
SS_A	3.933	2	1.966	$F_A = 7.891 > F_{0.05}(2,180)$
SS_B	0.751	2	0.376	$F_B = 1.507 < F_{0.05}(2,180)$
SS_{AB}	6.664	4	1.666	$F_{AB} = 6.686 > F_{0.05}(2,180)$
WSS	44.853	180	0.249	
TSS	154.17	188		*where* $F_{0.05}(2,180) = 3.04$

  - From the result above, what conclusion we can make?
  - Why the interaction term is significant while the awareness remain unchanged across the car usage?

7. Experimental Design

Swiss Institute of
Artificial Intelligence





Core element of good experimental design

- Comparison, Blocking, Replication, Randomization

Comparison and Blocking

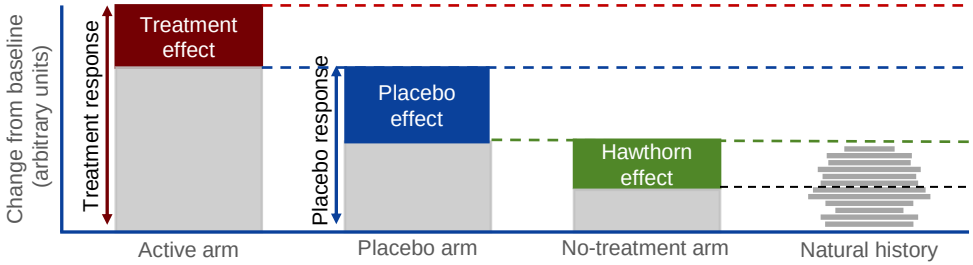
■ **Representativeness**

– Samples should be a representative of the population about which a conclusion is going to be drawn.

– e.g.) A study surveyed 10 million people who subscribe to the literary digest and have their car, found out that Landon would win the Roosevelt. Evaluate this survey.

– Why ***apple-to-apple*** and ***random sampling*** are important?

■ **Comparison/Control/Baseline**



– To know the effectiveness of the experiment, a comparison object is needed.

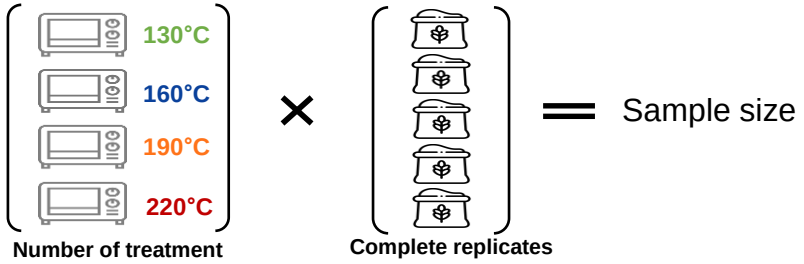
– Placebo and Hawthorne effect

■ **Blocking/controlling for confounds**

– By blocking, one removes the source of variation due to potential confounding factors, and thus improves the efficiency of the inference of treatment effect.

Replication and Randomization

■ **Replication (consistency)**



note : In a study of baking temperature on the volume of quick bread prepared from a package mix, four oven temperatures were tested by randomly assigning each temperature to 5 package mixes.

– When a treatment is repeated under the same experimental conditions, any difference is due to random errors.

– Can we know the evidence about treatment effect?

■ **Randomization**

– Randomization tends to average out between treatments, so that the comparison between treatments measure only the pure treatment effect.

– e.g.) In a study of light effects on plant growth rate, brighter vs. darker. 100 plants are randomly assigned to each.

– What if there is only one growth chamber which can grow 20 plants at one time?

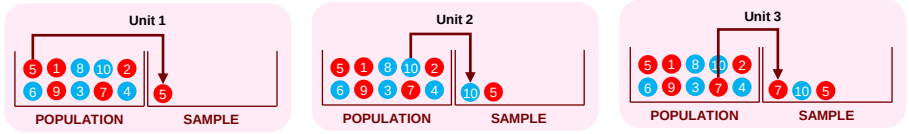
– Also, is the conditions of the growth chamber the same for all time periods?

Probability Sampling methods

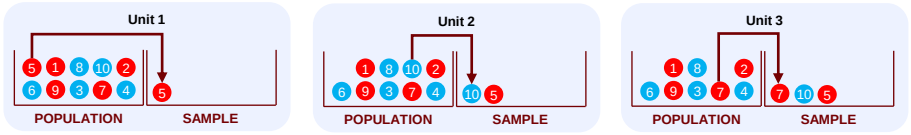
- Simple random, Systematic, Stratified, Cluster sample

Simple Random Sample

- **Simple Random Sample(SRS)**
  - Draw samples of the population uniformly at random without replacement.
  - Is it possible to randomly select samples of students from New York City and students from developing countries?
  - Can research on rare disease patients be conducted with only 1000 samples?
- **Samples independence and correction factor**



SRS *WITH* REPLACEMENT



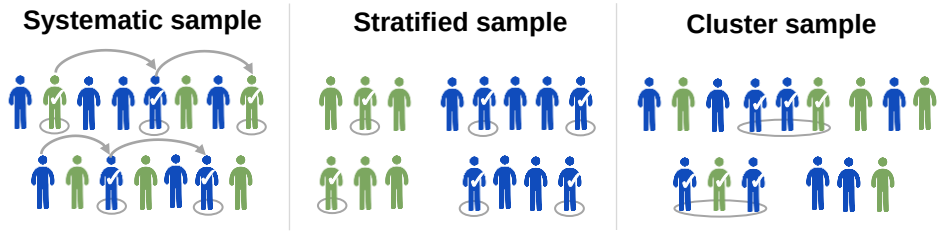
SRS *WITHOUT* REPLACEMENT

$$E[\bar{x}] = \mu,$$

$$Var[\bar{x}] = \underbrace{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}_{correction\ factor}$$

  - If the sample is not independent, the average is adjusted unevenly through the correction factor.

Other sampling methods



Systematic sampling

- It is similar to simple random sampling, but instead of randomly generating numbers, individuals are chosen at regular intervals.

Stratified Random Sample

- Will multiplying the result value of the sample extracted at a ratio of 1:100 by 100 equal the population value?
- The disproportional distribution can be adjusted by control the weight of the factor.

Cluster sampling

- Case of collecting data on 10 companies that have a similar number of employees and roles.
- How to guarantee that the sampled clusters are representative of the whole population?

How to remove sample bias

- It is difficult to obtain a perfect sample without bias

Bias of the sample

- **If sampling techniques are not used**
  - If the reasonable sampling methods mentioned on the previous page are not used, bias occurs. And we can't meet only data without bias. At this time, we have to choose two things.

Fix the sample and use it

Throw away the sample

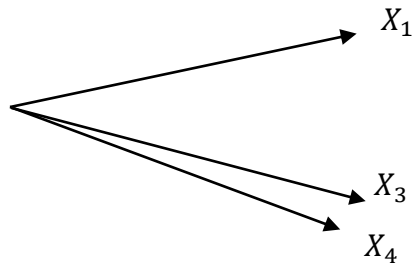
- **How to remove sample bias**
  - If we can't claim new data, we have to know how to control the bias of the sample.
  - Suppose we do multiple regression analysis and we have an error term in X_3.

$$y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots$$

  - How can we redefine X_3?

Simple replacement

- **If there is a vector that has a vector space similar to X_3.**
  - If X_4 is a variable similar to X_3 or has some components of X_3, in other words, if X_4 has a vector space similar to X. We can discard X_3 and use X_4.



- **If X_3 has an error as much as the constant K,**
  - Subtract K from X_3.

$$X_3 = \hat{X}_3 - K_3$$

- **If K cannot be removed by simple subtraction,**
  - Do 2 stage least square method

2 Stage Least Square(2SLS)

- Vector and Correction perspective.

2 Stage least square method and vector perspective

■ **If X_3 has as much bias as K_3 and cannot be removed by simple subtraction, it can be approached in a 2SLS method.**

– First, we perform a regression analysis of K_3 to X_3.

$$X_3 = m + nK_3 + \epsilon$$

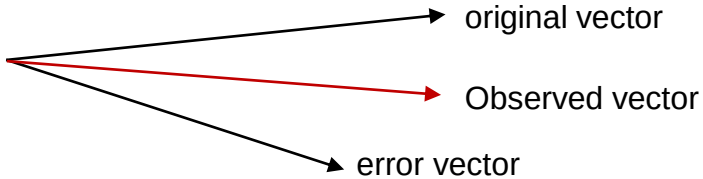
– Then, it is divided into a portion nK_3 described by K_3 and a portion $\mathrm{m} + \epsilon$ not described by K_3. remove nK_3.

$$\hat{X}_3 = m + \epsilon$$

– Regression analysis can be modified by putting this $\hat{X}_3$ in the X_3 position.

$$y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \hat{X}_3 + \cdots$$

■ **2SLS from the perspective of vector space.**



– If the error vector was added to the original vector, the vector would have been observed with a red line. Then, if we remove the error vector from the observed vector, the original vector come out.

Correction perspective

■ **Why did we do correction?**

– In 11p, we multiply the var by the correction factor. Why did we do that?

$$Var[\bar{x}] = \frac{\sigma^2}{n} \underbrace{\left(\frac{N-n}{N-1}\right)}_{correction\ factor}$$

– It was estimated that the variance would have been greater because the X value was not extracted independently. The variance becomes smaller by multiplying a value less than 1.

– It is the same concept as above. Modifying and using a sample even if it's a little less accurate than throwing away all the work you've done so far.

– It may be difficult to measure again. For example, students took an exam at school, and there was an error in one question. Then, do all students have to take the test again? Will the students' scores be the same as before and only that part be corrected? It will never be like that.