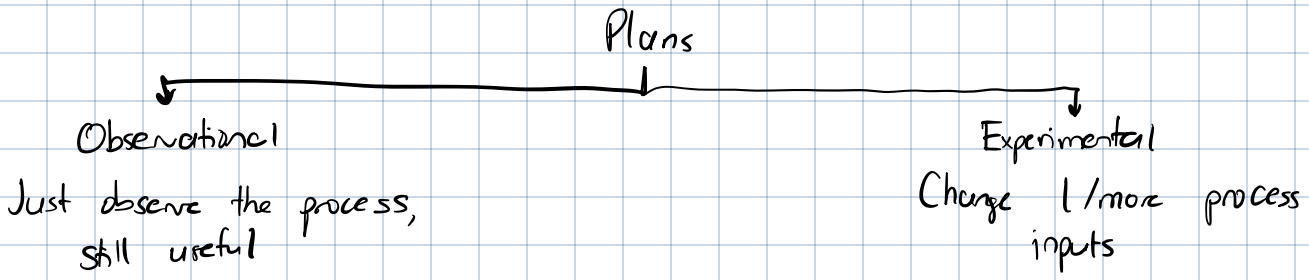
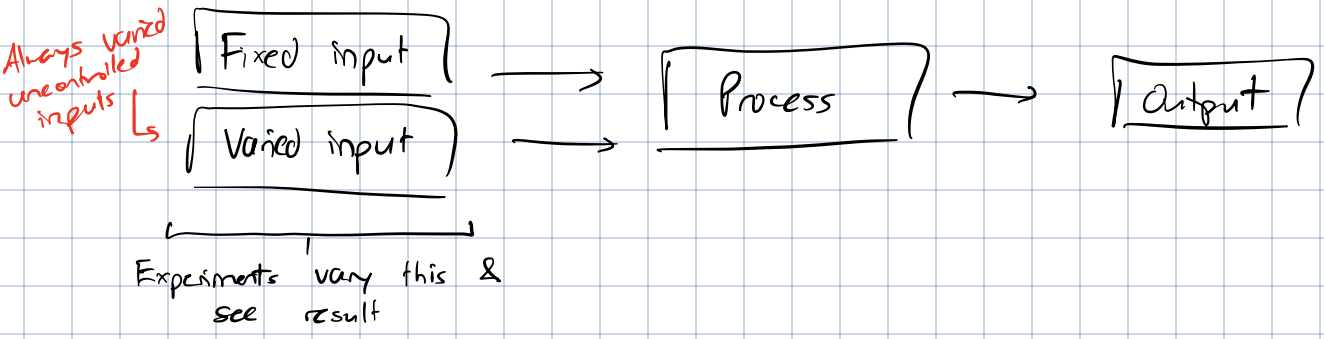
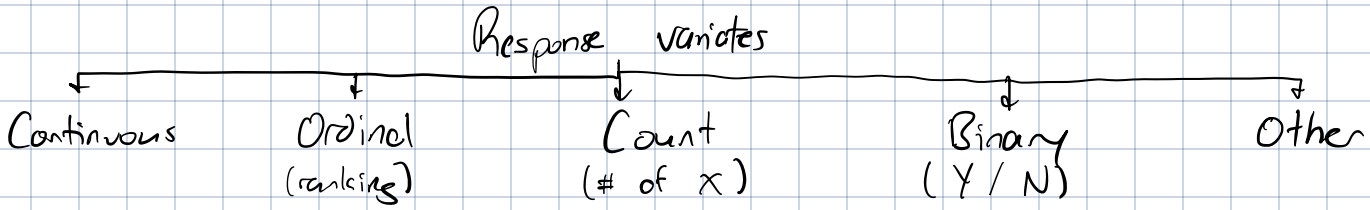
Fundamentals Of Experimental Plans





Factor: single variate that is changed/set on each unit in sample, focus of study
↳ Factor levels: set of values assigned to factor (quantitative/qualitative), categories

Treatment: combo of factor levels that can be applied to a unit



Fundamental Principles of Experimental Plans

Blocking: groups (blocks) of units where 1/more exploratory variates are fixed while diff. treatments applied to units w/in group

↳ Ex:// Sample consists of 300 stores. Stores put in blocks of 4 where all stores of 1 block had similar 3-week sales volume. 4 diff. treatments applied
(expl. variate held fixed.)

↳ Pros:

① Prevents confounding as exploratory variates fixed

② Improves precision of concl. b/c w/in-block variation ↓

Confounding: rel. b/w 2 inputs & vary together ⇒ cannot det. if outcome from 1 or othe

Replication: apply treatment to >1 unit in sample

↳ Ex:// Treatment A applied to 75 stores (once in 75 blocks)

↳ Pros:

① Reduces sample error b/c larger sample possible

② Estimate conclusion precision $\Rightarrow \hat{\sigma}^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$ ✓ $n > 1$!

Random assignment: assign treatments to units randomly

↳ Ex:// Each store in block had an equal chance of getting Treatment A

↳ Pros:

① Reduces confounding risk

② Generates analysis method

INFERENCE OVERVIEW

Gaussian Model

Model: $Y_i = \mu + R_i$, $R_i \overset{iid}{\sim} N(0, \sigma^2)$

↳ 'true' avg. of pop.

Assumption checking:

① Normality $\Rightarrow$ histogram, Q-Q plot

② Constant variance $\Rightarrow$ residual analysis

③ R_i are indep. $\Rightarrow$ can't check, rely on randomization

t-discrepancy measure

To fit model, estimate μ & σ via max. likelihood (MLE) or least squares (LSE)

LSE of μ & σ = MLE of μ & σ for Gaussian model

↳ To estimate μ:

$$\min_{\mu} \sum_{i=1}^{n} (y_i - \mu)^2 \Rightarrow \hat{\mu} = \bar{y}$$

↳ To estimate σ:

$$\hat{\sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Estimator distn. of μ & σ:

$$\tilde{\mu} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

↳ Proof:

① Definition of estimator:

$$\tilde{\mu} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

② Normality:

$Y_i \sim N(\mu, \sigma^2)$. Since $\tilde{\mu}$ is a linear combo. of Y_i (indep.), $\tilde{\mu}$ is also normally distributed.

③ Mean:

$$E\left(\frac{1}{n} \sum_{i=1}^{n} Y_i\right) = \frac{1}{n} \sum_{i=1}^{n} E(Y_i) = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu$$

④ Variance:

$$\text{Var}\left(\frac{1}{n} \sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(Y_i) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$\frac{\tilde{\sigma}^2 (n-1)}{\sigma^2} \sim \chi^2_{n-1}$$

↳ Proof: We know that

$$\sum_{i=1}^{n} (Y_i - \mu)^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + n(\bar{Y} - \mu)^2$$

Now:

$$\underbrace{\sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right)^2}_{U} = \underbrace{\sum_{i=1}^{n} \left(\frac{Y_i - \bar{Y}}{\sigma}\right)^2}_{V} + \underbrace{\left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}\right)^2}_{W}$$

Since $\frac{Y_i - \mu}{\sigma} \sim N(0,1)$, $U \sim \chi^2_n$.

From above, $\bar{Y} \sim N(\mu, \sigma^2/n) \Rightarrow \frac{\bar{Y} - \mu}{\sigma/n} \sim N(0,1) \Rightarrow W \sim \chi^2_1$

Furthermore:

$$\tilde{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Thus:

$$V = \frac{(n-1)\tilde{\sigma}^2}{\sigma^2}$$

Assuming $\bar{Y}$ is indep. of σ^2, V is indep. of W. Thus

$$M_U(t) = M_V(t)\, M_W(t)$$

$$\frac{1}{(1-2t)^{n/2}} = M_V(t) \cdot \frac{1}{(1-2t)^{1/2}}$$

$$M_V(t) = \frac{1}{(1-2t)^{n-1/2}}$$

Thus:

$$V = \frac{(n-1)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$$

Note that:

$$Z \sim N(0,1),\ U \sim \chi^2_K,\ Z \text{ is indep. of } U \Rightarrow T = \frac{Z}{\sqrt{U/K}} \sim t_K$$

Claim: $\dfrac{\tilde{\mu} - \mu}{\sqrt{\tilde{\sigma}^2/n}} \sim t_{n-1}$

Ls Proof:

Let

$$Z = \frac{\tilde{\mu} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1), \quad U = \frac{(n-1)\tilde{\sigma}^2}{\sigma} \sim \chi^2_{n-1}$$

We know that $\tilde{\mu}$ is indep. of σ^2, so Z & U are indep. b/c func. of indep. R.V. are also indep.

Thus:

$$\frac{Z}{\sqrt{U/K}} = \frac{\dfrac{\tilde{\mu} - \mu}{\sqrt{\sigma^2/n}}}{\sqrt{\dfrac{(n-1)\tilde{\sigma}^2}{\sigma^2} \cdot \dfrac{1}{n-1}}} = \frac{\tilde{\mu} - \mu}{\sqrt{\tilde{\sigma}^2/n}} \sim t_{n-1}$$

In a test:

① Find $\hat{\mu}$ & $\hat{\sigma}$

② Discrepancy measure: $H_0: \mu = \mu_0$

$$t^* = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$$

③ p-value:

$$p\text{-val} = 2P(t_{n-1} > t^*) \quad \rightarrow \text{2-sided } t\text{-test}$$

Single Treatment Confidence Interval

Pivotal quantity: f(data, unknown param from known distr.)

When μ is param. of interest:

$$Q: \frac{\tilde{\mu} - \hat{\mu}}{\sigma/\sqrt{n}} \sim N(0,1) \Rightarrow \text{If } \sigma \text{ known}$$

OR

$$Q: \frac{\tilde{\mu} - \hat{\mu}}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1} \rightarrow \sigma \text{ unknown}$$

From these quantities, we can derive $(1-\alpha)\cdot 100$ C.I.

$$C.I. = \hat{\mu} \pm c \cdot s.e.(\hat{\mu}) \rightarrow s.e.(\hat{\mu}) = \frac{\hat{\sigma}}{\sqrt{n}}$$

$$\hookrightarrow P(|t_{n-1}| \leq c) = 1-\alpha$$

Interpretation:

✗ NOT 95% prob. that C.I. contain μ ⇒ it either does or doesn't

✓ One of many intervals that covers the mean μ 95% of time

✓ If you sample data many times, mean μ will lie in C.I. 95% of time

Hypothesis test for mean

① Establish H_0 & H_a

② Discrepancy measure:

$$\frac{\hat{\mu} - \mu_0}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$$

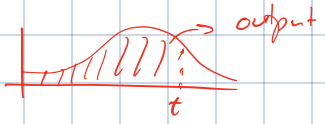
③ Calculate p-value:

a) $H_a: \mu \neq 0 \Rightarrow$ p-val $= 2P(t_{n-1} > t_{obs})$

b) $H_a: \mu > 0 \Rightarrow$ p-val $= P(t_{n-1} > t_{obs})$

c) $H_a: \mu < 0 \Rightarrow$ p-val $= P(t_{n-1} < t_{obs})$

In R: pt(t, df) is CDF of t-distr.



④ Statistical interpretation of p-val

Write in degree of evidence against H_0. Close to zero ⇒ more evidence

⑤ Conclusion in context of study

Note that you cannot compare p-vals! Sample size is huge factor

C.I. will reflect results of hypothesis test. If p-val < 0.05, $H_0 \notin$ 95% C.I.

COMPARATIVE EXPERIMENTAL PLANS

Two Treatments w/out Blocking

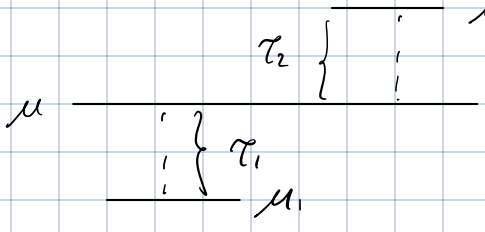
Model: $Y_{ij} = \mu + \tau_i + R_{ij}$, $i = 1, 2$, $j = 1 \dots n_i$, $R_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

R_{ij}: unit j's error in treatment i

Y_{ij}: Response; μ: Overall mean; τ_i: effect of treatment i

Model is subject to identifiability condition: $n_1 \tau_1 + n_2 \tau_2 = 0$

↳ Forces:



↳ If $n_1 = n_2$ ⇒ balanced plan. O.W., unbalanced

↳ Importance:

A: Notation:

$$Y_{i+} = \sum_{j=1}^{n_i} Y_{ij}, \quad \bar{Y}_{i+} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}, \quad Y_{++} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} Y_{ij}, \quad \bar{Y}_{++} = \frac{1}{n_1 + n_2} Y_{++}$$

Y_{i+}: Summing on other variable; $\bar{Y}_{i+}$: mean

B: Estimability of parameters

We could write the following:

$$Y_{ij} = \mu + \tau_i + R_{ij} = (\mu + \alpha) + (\tau_i - \alpha) + R_{ij}$$
$$= \mu^* + \tau_i^* + R_{ij}$$

This makes μ & τ_i hard to estimate. Condition prevents this. b/c:

$$n_1 \tau_1^* + n_2 \tau_2^* = n_1 \tau_1 + n_2 \tau_2 - 2\alpha \neq 0$$

$\therefore$ τ_i is strictly treatment effect

C: Interpretability:

We want $\mathbb{E}[\bar{Y}_{++}] = \mu \Rightarrow$ pop. avg. of response variate.

$$\mathbb{E}[\bar{Y}_{++}] = \frac{1}{n_1 + n_2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} \mathbb{E}[Y_{ij}]$$
$$= \frac{1}{n_1 + n_2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (\mu + \tau_i)$$
$$= \frac{1}{n_1 + n_2} \sum_{i=1}^{2} n_i (\mu + \tau_i)$$
$$= \frac{1}{n_1 + n_2} (n_1 \mu + n_1 \tau_1 + n_2 \mu + n_2 \tau_2)$$
$$= \frac{1}{n_1 + n_2} (n_1 + n_2)\mu + (n_1 \tau_1 + n_2 \tau_2)$$
$$= \mu + (n_1 \tau_1 + n_2 \tau_2)$$

If we want $\mathbb{E}[\bar{Y}_{++}] = \mu \Rightarrow n_1 \tau_1 + n_2 \tau_2 = 0$

We are interested in diff. b/w treatments:

$$\theta = \tau_1 - \tau_2$$

which is:

$$\mathbb{E}[\bar{Y}_{1+} - \bar{Y}_{2+}] = \mathbb{E}[\bar{Y}_{1+}] - \mathbb{E}[\bar{Y}_{2+}]$$
$$= \mu + \tau_1 - \mu - \tau_2$$
$$= \tau_1 - \tau_2$$

Thus $\hat{\theta} = \bar{Y}_{1+} - \bar{Y}_{2+}$ is unbiased estimator of θ

Next goal: estimate μ & τ_i

↳ Proof: We know $Y_{ij} = \mu + \tau_i + R_{ij} \Rightarrow R_{ij} = Y_{ij} - \mu - \tau_i$

Loss function:

$$\min_{\mu, \tau_1, \tau_2} \underbrace{\sum_{i=1}^{2} \sum_{j=1}^{n_i} R_{ij}^2}_{L(\mu, \tau_1, \tau_2)}, \text{ subject to } n_1\tau_1 + n_2\tau_2 = 0$$

We use method of Lagrange multipliers to solve this:

$$\min_{\mu, \tau_1, \tau_2, \lambda} \left[L(\mu, \tau_1, \tau_2) + \lambda(n_1\tau_1 + n_2\tau_2) \right]$$

$$\Rightarrow \min_{\mu, \tau_1, \tau_2, \lambda} \Big[\underbrace{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \mu - \tau_i)^2 + \lambda(n_1\tau_1 + n_2\tau_2)}_{Q} \Big]$$

Thus:

① $$\frac{\partial Q}{\partial \mu} = -2 \sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu} - \hat{\tau}_i) = 0$$

② $$\frac{\partial Q}{\partial \tau_i} = \sum_{j=1}^{n_i} (-2(y_i - \hat{\mu} - \hat{\tau}_i) + n_i\hat{\lambda}) = 0$$

③ $$\frac{\partial Q}{\partial \lambda} = n_1\hat{\tau}_1 + n_2\hat{\tau}_2 = 0$$

Solving ①

$$y_{++} - (n_1 + n_2)\hat{\mu} - \underbrace{n_1\hat{\tau}_1 - n_2\hat{\tau}_2}_{0 \text{ b/c } ③} = 0$$

$$y_{++} - (n_1 + n_2)\hat{\mu} = 0$$

$$\hat{\mu} = \frac{y_{++}}{n_1 + n_2} = \bar{y}_{++}$$

Solving ②:

$$-2(y_{i+} - n_i\hat{\mu} - n_i\hat{\tau}_i) + n_i\hat{\lambda} = 0$$

$$\Rightarrow -2 \sum_{i=1}^{2} (y_{i+} - n_i\hat{\mu} - n_i\hat{\tau}_i) + \sum_{i=1}^{2} n_i\hat{\lambda} = 0$$

Common trick: add treatment effect

$$\Rightarrow y_{1+} - n_1\hat{\mu} - n_1\hat{\tau}_1 + y_{2+} - n_2\hat{\mu} - n_2\hat{\tau}_2 + \frac{1}{2}n_1\hat{\lambda} + \frac{1}{2}n_2\hat{\lambda} = 0$$

$$\Rightarrow y_{1+} + y_{2+} - (n_1 + n_2)\hat{\mu} + \frac{1}{2}\hat{\lambda}(n_1 + n_2) = 0$$

$$\Rightarrow \underbrace{y_{1+} + y_{2+}}_{y_{++}} - y_{++} + \frac{1}{2}\hat{\lambda}(n_1 + n_2) = 0$$

$$\Rightarrow \frac{1}{2}\hat{\lambda}(n_1 + n_2) = 0$$

$$\therefore \hat{\lambda} = 0$$

Redoing ②

$$\sum_{j=1}^{n_i} \left(-2(y_{ij} - \hat{\mu} - \hat{\tau}_i)\right) = 0$$

$$\Rightarrow \quad y_{i+} - n_i \hat{\mu} - n_i \hat{\tau}_i = 0$$

$$\Rightarrow \quad \hat{\tau}_i = \hat{\mu} - \overline{y_{i+}} = \overline{y_{++}} - \overline{y_{i+}}$$

Variance estimation:

① Variance w/in treatment i:

$$\hat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu} - \hat{\tau}_i)^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i+})^2$$

② Pooled / overall variance

$$\hat{\sigma}^2 = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2 = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}$$

Statistical inference

Note that $\overline{Y}_{i+} \sim N(\mu + \tau_i, \sigma^2/n_i)$

↳ Proof:

① Normality:

$\overline{Y}_{i+} = \frac{1}{n_i}(Y_{i0} + \dots Y_{in})$ ⇒ indep. L.C. of normal R.V.

$\therefore$ $\overline{Y}_{i+}$ is normal

② Mean:

$$\mathbb{E}[\overline{Y}_{i+}] = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbb{E}[\mu + \tau_i + R_{ij}] = \mu + \tau_i$$

③ Variance:

$$\mathrm{Var}[\overline{Y}_{i+}] = \frac{1}{n_i^2} \sum_{j=1}^{n_i} \mathrm{Var}[\mu + \tau_i + R_{ij}] = \frac{\sigma_i^2}{n_i}$$

Therefore:

$$\overline{Y}_{1+} - \overline{Y}_{2+} \sim N\left(\tau_1 - \tau_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

We also know that

$$\frac{(n_1+n_2-2)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$$

Same argument as $\frac{\bar{\mu}-\mu}{\sigma/\sqrt{n}} \sim t_{n-1}$ derivation

Thus:

$$\frac{(\bar{Y}_{1+}-\bar{Y}_{2+})-\theta}{\tilde{\sigma}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

pooled std. dev.

Use this to make C.I. & tests

This all relies on β_{ij} iid $N(0,\sigma^2)$

Two Treatments w Blocking

Complete randomized block design: all treatments randomly assigned w/in each block, $\forall$ blocks

Model:

$$Y_{ij} = \mu + \tau_i + \beta_j + \beta_{ij}, \quad \beta_{ij} \overset{iid}{\sim} N(0,\sigma^2), \quad i=1,2, \; j=1,\dots,b$$

(β_j: block effect)

subject to: $\tau_1+\tau_2=0$, $\sum_{j=1}^{b}\beta_j = 0$ } identifiability cond.

Parameter estimation:

obj: $\min_{\mu,\vec{\tau},\vec{\beta}} \sum_{i=1}^{2}\sum_{j=1}^{b}(y_{ij}-\mu-\tau_i-\beta_j)^2$ subj. $\begin{cases}\tau_1+\tau_2=0\\ \sum_{j=1}^{b}\beta_j=0\end{cases}$

Use the same methods (Lagrange):

$$\min_{\mu,\vec{\tau},\vec{\beta},\lambda_1,\lambda_2} \underbrace{\left[\sum_{i=1}^{2}\sum_{j=1}^{b}(y_{ij}-\mu-\tau_i-\beta_j)^2 + \lambda_1(\tau_1+\tau_2) + \lambda_2\sum_{j=1}^{b}\beta_j\right]}_{Q}$$

Set derivatives:

$$\frac{\partial Q}{\partial \mu}=0, \quad \frac{\partial Q}{\partial \tau_i}=0, \quad \frac{\partial Q}{\partial \beta_j}=0, \quad \frac{\partial Q}{\partial \lambda_i}=0$$

Estimates:

$$\hat{\mu} = \bar{y}_{++}, \quad \hat{\tau}_i = \bar{y}_{i+}-\bar{y}_{++}, \quad \hat{\beta}_j = \bar{y}_{+j}-\bar{y}_{++}$$

Estimating variance:

$$\hat{\sigma}^2 = \frac{1}{b-1} \sum_{i=1}^{2} \sum_{j=1}^{b} (y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j)^2$$

$$= \frac{1}{b-1} \sum_{i=1}^{2} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{i+} - \bar{y}_{+j} + \bar{y}_{++})^2$$

$$= \frac{1}{2} \frac{\sum_{j=1}^{b} (d_j - \bar{d})^2}{b-1}$$

$d_j = y_{1j} - y_{2j}$

$\bar{d} = \frac{1}{b} \sum_{j=1}^{b} d_j$

* Prove this *

Why do we have $b-1$ d.f.?

d.f. = sample size − # of estimated params. + # of constraints

$= 2b - (1 + 2 + b) + 2$ (2 people per block in ex.)

(μ, τ_i, β_j; τ, β)

$= b - 1$

$= OR \quad n_1 + n_2 - 2$

Statistical inference

$\theta = \tau_1 - \tau_2 \Rightarrow$ param. of interest. Derive distn. of $\hat{\theta}$

Claim: $\hat{\theta} \sim N(\tau_1 - \tau_2, \frac{2\sigma^2}{b})$

Proof:

① Normality

$$\hat{\theta} = \bar{Y}_{1+} - \bar{Y}_{2+} = \frac{1}{b} \sum_{j=1}^{b} Y_{1j} - \frac{1}{b} \sum_{j=1}^{b} Y_{2j} = \frac{1}{b} \sum_{j=1}^{b} (Y_{1j} - Y_{2j})$$

This is a linear combo of indep. normal R.V. $\Rightarrow \hat{\theta}$ is Gaussian

② Mean

$$E[\hat{\theta}] = \frac{1}{b} \sum_{j=1}^{b} (E[Y_{1j}] - E[Y_{2j}])$$

$$= \frac{1}{b} \sum_{j=1}^{b} (\tau_1 - \tau_2)$$

$$= \tau_1 - \tau_2$$

③ Variance

$Var(a-b) = Var(a) + Var(b) + Cov(a,b)$

$$Var[\hat{\theta}] = \frac{1}{b^2} \sum_{j=1}^{b} (Var[Y_{1j}] + Var[Y_{2j}])$$

$$= \frac{1}{b^2} \sum_{j=1}^{b} (2\sigma^2)$$

$= \frac{2\sigma^2}{b}$

Let pivotal quantity be:

$$\frac{\hat{\theta} - (\tau_1 - \tau_2)}{\sigma\sqrt{2/b}} \sim N(0,1)$$

We know $\frac{(b-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{b-1}$ (indep.)

$$\Rightarrow \frac{\hat{\theta} - (\tau_1 - \tau_2)}{\hat{\sigma}\sqrt{2/b}} \sim t_{b-1}$$

Use this for C.I. & tests.

EXPERIMENTAL PLANS: ≥ 2 TREATMENTS

Introduction

Experimental design

- Completely randomized
  - No blocking
- Randomized block design
  - Each treatment repeated 1x in each block, random assignment

Completely Randomized Design - Balanced

Balanced plan: same # of units receive each treatment

Goal: develop methodology to test $H_0: \mu_A = \mu_B = \mu_C = \ldots = \mu_t$

① MODEL

Assuming balanced plan:

$$Y_{ij} = \mu + \tau_i + R_{ij}, \quad i = 1, \ldots, t, \quad j = 1, \ldots, r, \quad R_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

(t = # of treatments, r = # of repl.)

Constraint: $\sum_{i=1}^{t} \tau_i = 0$ ← exact same arg. as previous constraint

Meaning:

μ: mean response across all treatments

τ_i: effect of treatment i relative to μ

R_{ij}: random error

Constraint leads us to claim $\mathbb{E}[\bar{Y}_{i+}] = \mu$

Proof:

$$\mathbb{E}[\bar{Y}_{i+}] = \frac{1}{r \cdot t} \sum_{i=1}^{t} \sum_{j=1}^{r} \mathbb{E}[Y_{ij}]$$

$$= \frac{1}{r \cdot t} \sum_{i=1}^{t} \sum_{j=1}^{r} (\mu + \tau_i)$$

$$= \frac{1}{r \cdot t} \sum_{i=1}^{t} (\mu + \tau_i) r$$

$$= \frac{t\mu}{t} \cdot \frac{1}{t} \sum_{i=1}^{t} \tau_i$$

$\sum_{i=1}^{t} \tau_i \to 0$ b/c of constraint

$$= \mu$$

We care about treatment effect τ_i:

$$\mathbb{E}[\bar{Y}_{i+}] = \mu + \tau_i \Rightarrow \tau_i = \mathbb{E}[\bar{Y}_{i+}] - \mathbb{E}[\bar{Y}_{++}]$$

Unbiased estimate of τ_i

② Parameter Estimation

Loss function:

$$\sum_{i=1}^{t} \sum_{j=1}^{r} (y_{ij} - \underbrace{(\mu + \tau_i)}_{\hat{y}_{ij}})^2 + \lambda \sum_{i=1}^{t} \tau_i$$

Solving this leads to:

$$\hat{\mu} = \bar{y}_{++} \Rightarrow \tilde{\mu} = \bar{Y}_{++}$$

$$\hat{\tau}_i = \bar{y}_{i+} - \bar{y}_{++} \Rightarrow \tilde{\tau}_i = \bar{Y}_{i+} - \bar{Y}_{++}$$

$$\hat{\tau}_i - \hat{\tau}_j = \bar{y}_{i+} - \bar{y}_{j+} \Rightarrow \tilde{\theta} = \bar{Y}_{i+} - \bar{Y}_{j+}$$

Variance estimator:

$$\hat{\sigma}^2 = \frac{1}{rt - (1+t) + 1} \sum_{i,j} (y_{ij} - \hat{\mu} - \hat{\tau}_i)^2 = \frac{1}{t(r-1)} \sum_{i,j} (y_{ij} - \hat{\mu} - \hat{\tau}_i)^2$$

rt: sample size; $1+t$: # of parameters; $+1$: # of constraints

Note:

$$\hat{\sigma}^2 = \frac{(r-1)\hat{\sigma}_1^2 + \dots + (r-1)\hat{\sigma}_t^2}{t(r-1)}$$

③ DISTRIBUTION

We know $\bar{Y}_{it} \overset{ind.}{\sim} N(\mu + \tau_i, \sigma^2/r)$

$\Rightarrow \hat{\theta} = \hat{\tau}_i - \hat{\tau}_i = \bar{Y}_{it} - \bar{Y}_{jt} \sim N(\tau_i - \tau_j, \frac{2\cdot\sigma^2}{r})$ $\Rightarrow$ CONTRAST

$$\Rightarrow \frac{\hat{\theta} - (\tau_i - \tau_j)}{\sigma\sqrt{2/r}} \sim N(0,1)$$

We also know $U = \frac{t(r-1)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{t(r-1)}$

$$\Rightarrow \frac{\tilde{\theta} - (\tau_i - \tau_j)}{\tilde{\sigma}\sqrt{2/r}} \sim t_{t(r-1)}$$

Completely Randomized Design: Generalizing via Contrasts

Contrast defn: linear combo. of treatment effects s.t. weights sum to 0

- Mathematically: $a_1, \dots, a_t \in \mathbb{R}$ (weights), $\tau_1, \dots, \tau_t \in \mathbb{R}$ (treatment effects)

① $\sum_t a_i \tau_i$ (linear combo)

② $\sum_{i=1}^{t} a_i = 0$

Ex// Identify contrast

1) $\tau_1 - \frac{\tau_2 + \tau_3}{2}$ $\Rightarrow$ Yes! L.C., $\sum a_i = 0$

2) $\tau_i^2 - \tau_j$ $\Rightarrow$ No, not L.C.

3) $\frac{\tau_1 + \tau_2}{2} - \frac{\tau_3 + \tau_4 + \tau_5}{3}$ $\Rightarrow$ Yes, $\sum a_i = 0$

Ex// Write contrasts to compare:

1) Treatment A vs. B, C, D

Trick: if A vs. B

$$\tau_A - \frac{\tau_B + \tau_C + \tau_D}{3}$$

⇒ 1. Make coef. of A & B sum to 1:

$\frac{\sum \text{treatments}}{\#\text{ of treatments}}$

2. A − B

2) A & B vs. C & D

$$\frac{\tau_A + \tau_B}{2} - \frac{\tau_C + \tau_D}{2} \Rightarrow (\tau_A + \tau_B) - (\tau_C + \tau_D)$$

Find the estimator of a generic contrast:

1. Parameter: $\theta = \sum_{i=1}^{t} a_i \tau_i$, $\sum_{i=1}^{t} a_i = 0$

2. Estimate:

$$\hat{\theta} = \sum_{i=1}^{t} a_i \hat{\tau}_i = \sum_{i=1}^{t} a_i (\bar{y}_{i+} - \bar{y}_{++})$$

$$= \sum_{i=1}^{t} a_i \bar{y}_{i+} - \bar{y}_{++} \sum_{i=1}^{t} a_i \quad (\sum a_i = 0)$$

$$= \sum_{i=1}^{t} a_i \bar{y}_{i+}$$

3. Estimator:

$$\tilde{\theta} = \sum_{i=1}^{t} a_i \bar{Y}_{i+}$$

What is distn. of $\tilde{\theta}$? Claim: $\tilde{\theta} \sim G\left(\theta, \sigma \sqrt{\frac{\sum_{i=1}^{t} a_i^2}{r}}\right)$ (STDEV, not var.)

Proof:

① Normality: $\tilde{\theta}$ is sum of indep. normal R.V.

② Expectation:

$$\mathbb{E}[\tilde{\theta}] = \sum_{i=1}^{t} a_i \, \mathbb{E}[\bar{Y}_{i+}] = \sum_{i=1}^{t} a_i (\mu + \tau_i)$$

$$= \mu \sum_{i=1}^{t} a_i + \sum_{i=1}^{t} a_i \tau_i \quad (\sum a_i = 0)$$

$$= \theta$$

③ Variance:

$$\mathrm{Var}[\tilde{\theta}] = \sum_{i=1}^{t} a_i^2 \, \mathrm{Var}(\bar{Y}_{i+}) = \sum_{i=1}^{t} a_i^2 \cdot \frac{\sigma^2}{r}$$

Discrepancy measure for generic contrast:

$$\frac{\hat{\theta} - \theta}{\tilde{\sigma}\sqrt{\sum a_i^2/r}} \sim t_{t(r-1)}$$

We can test diff. b/w generic combos of treatments

Completely Randomized Design: ANOVA

Goal: develop methodology to test $H_0: \mu_a = \mu_b = \mu_c = \ldots = \mu_0$

Idea: Construct:

A: $\tilde{\sigma}_r^2$: estimator for σ^2, unbiased regardless if H_0 true/not

From model, we know:

$$\tilde{\sigma}_r^2 = \frac{1}{t(r-1)} \sum_{i,j} (y_{ij} - \bar{y}_{i+})^2 \Rightarrow \text{unbiased}$$

B: $\tilde{\sigma}_t^2$: " , unbiased $\Leftrightarrow$ H_0 is true

From model: $\bar{y}_{i+} \sim N(\mu + \tau_i, \sigma^2/r)$

If H_0 is true, then $\tau_i = \ldots \tau_t = 0$.

$$\Rightarrow \bar{y}_{i+} \sim N(\mu, \sigma^2/r)$$

Estimate variance:

$$\frac{\hat{\sigma}^2}{r} = \frac{\sum_t (\bar{y}_{i+} - \bar{y}_{++})^2}{t-1} = \frac{1}{t-1} \sum \hat{\tau}_i^2$$

$$\therefore \hat{\sigma}_{tr}^2 = \frac{r}{t-1} \sum \hat{\tau}_i^2 \Rightarrow \tilde{\sigma}_{tr}^2 = \frac{r}{t-1} \sum \tilde{\tau}_i^2$$

$\therefore$ Following estimator:

$$F = \frac{\tilde{\sigma}_t^2}{\tilde{\sigma}_r^2} \Rightarrow$$

($\tilde{\sigma}_t^2$: treatment mean square; $\tilde{\sigma}_r^2$: residual mean square)

If H_0 is true $\Rightarrow$ value ≈ 1

If H_0 is false $\Rightarrow$ B↑ → F↑

Distribution of F:

We know:

① $\frac{t(r-1)\tilde{\sigma}_r^2}{\sigma^2} \sim \chi^2_{r(t-1)}$

② $\frac{(t-1)\sigma_t^2}{\sigma^2} \sim \chi^2_{t-1}$

Also:

$$X \sim \chi^2_{df_X}, \quad Y \sim \chi^2_{df_Y} \Rightarrow F = \frac{X/df_X}{Y/df_Y} \sim F_{df_X, df_Y}$$

Thus:

$$\frac{\tilde{\sigma}_t^2}{\tilde{\sigma}_r^2} \sim F_{t-1,\, r(t-1)}$$

ANOVA & Sum of Squares

Theorem: sum of squared error can be decomposed like so:

$$\sum_{i,j} (y_{ij} - \bar{y}_{++})^2 = \sum_{i,j} (y_{ij} - \bar{y}_{i+})^2 + r \sum_i (\bar{y}_{i+} - \bar{y}_{++})^2$$

Total sum of squares | Residual/error sum of sq. | Treatment sum of squares

Proof:

$$\sum_{i=1}^{t}\sum_{j=1}^{r} (y_{ij} - \bar{y}_{++})^2 = \sum_{i=1}^{t}\sum_{j=1}^{r} (y_{ij} - \bar{y}_{i+} + \bar{y}_{i+} - \bar{y}_{++})^2$$

$$= \sum_{i=1}^{t}\sum_{j=1}^{r} (y_{ij} - \bar{y}_{i+})^2 + \sum_{i=1}^{t}\sum_{j=1}^{r} (\bar{y}_{i+} - \bar{y}_{++})^2 - 2\sum_{i,j} (\bar{y}_{i+} - \bar{y}_{++})(y_{ij} - \bar{y}_{i+})$$

$$= \sum_{i=1}^{t}\sum_{j=1}^{r} (y_{ij} - \bar{y}_{i+})^2 + \sum_{i=1}^{t}\sum_{j=1}^{r} (\bar{y}_{i+} - \bar{y}_{++})^2 \quad \blacksquare$$

$$\sum_i^t (\bar{y}_{i+} - \bar{y}_{++}) = -t\bar{y}_{++} + \sum_{i=1}^t \bar{y}_{i+} = -t\bar{y}_{++} + t\bar{y}_{++} = 0$$

Intuition: if the treatment S.S. >> error S.S. ⇒ treatments are diff.

- SST (total): total variation in response
- SSTr (treatment): variability b/w treatment groups
- SSE (error): variability w/in group. If large → harder to detect diff. b/w groups

ANOVA table:

	Sum of squares	DF	Mean square	F_{obs}
Treatment	$r\sum_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$t-1$	$SSTr/df_{tr}$	$\frac{MS_{tr}}{MS_r}$
	+	+		
Residual	$\sum_{ij} (y_{ij} - \bar{y}_{i\cdot})^2$	$t(r-1)$	SSR/df_r	
	=	=		
Total	$\sum_{ij} (y_{ij} - \bar{y}_{\cdot\cdot})^2$	$tr-1$		

↳ Be able to construct w/ only summary stats

Intuition: large diff. in treatment $\Rightarrow$ large F_{obs} ($MS_{tr} > MS_r$)

ANOVA test:

$\begin{cases} H_0: \tau_1 = \dots = \tau_t = 0 / \mu \\ H_a: \text{not the same} \end{cases} \Rightarrow$

① Calculate F_{obs}

② p-value: $P(F_{t-1, t(r-1)} \geq F_{obs})$

Note: If $t = 2$: $F_{obs} = (t_{obs})^2$ $\Rightarrow$ know proof

After this, perform post-hoc analysis to see which treatments are actually diff.

Completely Randomized Design - Unbalanced

Defn: t treatments assigned to group randomly but # of replications / treatment diff

① Model

$$Y_{ij} = \mu + \tau_i + R_{ij}, \quad i = 1 \dots t, \; j = 1 \dots r_i$$

$$R_{ij} \sim N(0, \sigma^2), \quad \sum_{i=1}^{t} \tau_i r_i =$$

Interpretation:

μ: weighted avg. of treatment effects

$\mu + \tau_i$: average resp. for treatment i

② Parameter Estimation

$$\min_{\mu, \tau_1, \dots, \tau_t, \lambda} \left[\sum_{i=1}^{t} \sum_{j=1}^{r_i} (y_{ij} - \mu - \tau_i)^2 + \lambda \sum_{i=1}^{t} r_i \tau_i \right]$$

$$\hat{\mu} = \bar{y}_{++}, \quad \hat{\tau}_i = \bar{y}_{i+} - \bar{y}_{++}$$

$$\Downarrow \qquad \Downarrow$$

$$\tilde{\mu} = \bar{Y}_{++} \qquad \tilde{\tau}_i = \bar{Y}_{i+} - \bar{Y}_{++}$$

Main change:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{t} \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{i+})^2}{\sum_{i=1}^{t} (r_i - 1)}$$

→ Equiv. to $y_{ij} - \hat{y}_{ij}$

↳ d.f.

Why is $\sum_{i=1}^{t} (r_i - 1)$ our d.f. ?

$$\sum_{i=1}^{t} (r_i - 1) = \sum_{i=1}^{t} r_i - (t + 1) + 1$$

$\sum r_i$: # of obs. $(t+1)$: # of param. $+1$: constraint

① ANOVA

	Sum of squares		DF		Mean square	F_{obs}
Treatment	$\sum_{ij} (\bar{y}_{i+} - \bar{y}_{++})^2$	÷	$t-1$	=	$SSTr/df_{Tr}$	$\frac{MS_{Tr}}{MS_r}$
	+		+			
Residual	$\sum_{ij} (y_{ij} - \bar{y}_{i+})^2$	÷	$\sum_i (r_i - 1)$	=	SSR/df_r	
	=		=			
Total	$\sum_{ij} (y_{ij} - \bar{y}_{++})^2$		$\sum_i r_i - 1$			

→ Equiv. if r_i is same $\forall i$ to par. ANOVA

Ex // Fill in ANOVA for following data:

	T_1	T_2	T_3	Overall
Avg.	3.64	3.40	3.21	3.41
Variance	0.123	0.067	0.105	0.125
n	13	14	14	41

① Estimate params:

$\hat{\mu} = \bar{y}_{1+} = 3.91, \quad \hat{\tau}_1 = \bar{y}_{1+} - \bar{y}_{7+} = 0.23, \quad \hat{\tau}_2 = -0.01. \quad \hat{\tau}_3 = -02$

② Sum of squares:

Total sum of squares = $d.f._+ \times$ overall variance

$= (13 + 14 + 14 - 1) \times 0.125$

$= 5$

$SS_{tr} = \sum_{i=1}^{t} \sum_{j=1}^{r_i} (\underbrace{\bar{y}_{i+} - \bar{y}_{7+}}_{\hat{\tau}_i})^2 = \sum_{i=1}^{t} r_i \hat{\tau}_i^2 = 1.249$

$SSE = SST - SST_r = 3.751$

③ D.f.

$d.f._t = 2, \quad d.f._r = d.f._T - d.f._{tr} = 40 - 2 = 38$

④ Fill in MS & Fobs.

How to do tests for contrasts:

① Estimate

$\hat{\theta} = \sum a_i \hat{\tau}_i = \sum a_i \bar{y}_{i+}$

② Distr.:

$\bar{Y}_{i+} \sim N(\mu + \tau_i, \frac{\sigma^2}{r_i})$

$\Rightarrow \tilde{\theta} = \sum a_i \bar{Y}_{i+} \sim N\left(\sum_{i=1}^{t} a_i \tau_i, \sigma^2 \sum_{i=1}^{t} \frac{a_i^2}{r_i}\right)$

$\therefore \quad \hat{\theta} \pm t^* \hat{\sigma} \sqrt{\sum_{i=1}^{t} \frac{a_i^2}{r_i}}$ is 95% C.I.

Randomized Block Design

Defn: multiple treatments assigned randomly w/in blocks

① MODEL

$Y_{ij} = \mu + \tau_i + \beta_j + B_{ij}, \quad i = 1, \ldots, t, \quad j = 1, \ldots, b$

$B_{ij} \sim N(0, \sigma^2), \quad \sum_{i=1}^{t} \tau_i = \sum_{j=1}^{b} \beta_j = 0$

Interpretations:

μ: overall avg. across treat. & blocks.

τ_i: i^{th} treatment effect

β_j: j^{th} block effect

Additive model: no interaction b/w block & treatment

① Treatment effect does not affect block (τ_i, not τ_{ij})

② Diff. b/w treatments is same $\forall$ blocks (β_j, not β_{ij})

③ Residual variance constant across treat. & blocks

② PARAM. ESTIMATES

$$\min_{\mu, \tau_1, \dots, \tau_t, \beta_1, \dots, \beta_b, \lambda_1, \lambda_2} \left[\sum_{i=1}^{t} \sum_{j=1}^{b} (y_{ij} - \mu - \tau_i - \beta_j)^2 + \lambda_1 \sum_i \tau_i + \lambda_2 \sum_j \beta_j \right]$$

Estimate:

$$\hat{\mu} = \bar{y}_{++} \quad \hat{\tau}_i = \bar{y}_{i+} - \bar{y}_{++}, \quad \hat{\beta}_j = \bar{y}_{+j} - \bar{y}_{++}$$

Variance estimate:

① D.f. = total # − # of params + # of const.

$$= tb - (t + 1 + b) + 2$$

$$= (t-1)(b-1)$$

② Estimate:

$$\hat{\sigma}^2 = \frac{1}{(t-1)(b-1)} \sum_{i,j} (y_{ij} - \bar{y}_{i+} - \bar{y}_{+j} + \bar{y}_{++})^2$$

③ TEST

Goal: test treatment effects simultaneously

Sum of squares decomp:

$$\sum_{i,j}(y_{ij}-\bar{y}_{++})^2 = \sum_{i,j}(y_{ij}-\bar{y}_{i+}-\bar{y}_{+j}+\bar{y}_{++})^2 + b\sum_{i=1}^{t}(\bar{y}_{i+}-\bar{y}_{++})^2 + t\sum_{j=1}^{b}(\bar{y}_{+j}-\bar{y}_{++})^2$$

SS_{total} | SS_{error} | $SS_{treatment}$ | SS_{block}

pulled out of SSE

Biased variance estimate if $H_0: \tau_1 = \dots = \tau_t = 0$

↳ Assume H_0 true $\Rightarrow \bar{Y}_{i+} \sim N(\mu, \sigma^2/b)$

$$\therefore \tilde{\sigma}_t^2 = \frac{b}{t-1}\sum_{i=1}^{t}(\bar{Y}_{i+}-\bar{Y}_{++})^2$$

$$\therefore \frac{SS_{treatment}}{\sigma^2} \sim \chi^2_{t-1}$$

$$F_{obs} = \frac{MS_T}{MS_r} \sim F_{t-1,(t-1)(b-1)}$$

ANOVA table:

	Sum of squares		DF		Mean square	F_{obs}
Treatment	$b\sum_i(\bar{y}_{i+}-\bar{y}_{++})^2$	$\div$	$t-1$	$=$	$SSTr/df_{tr}$	$\frac{MS_{tr}}{MS_r}$
+	+		+			
Blocks	$t\sum_j(\bar{y}_{+j}-\bar{y}_{++})^2$	$\div$	$b-1$	$=$	SSB/df_B	
+	+		+			
Residual	$\sum_{i,j}(y_{ij}-\bar{y}_{i+}-\bar{y}_{+j}+\bar{y}_{++})^2$	$\div$	$(t-1)(b-1)$	$=$	SSR/df_r	
=	=		=			
Total	$\sum_{i,j}(y_{ij}-\bar{y}_{++})^2$		$tb-1$			

↳ Compute via summary stats.

If blocking is important:

$$MSR < \frac{SS_B + SS_{resid}}{df_B + df_{resid}}$$

→ Meaningful amt. of variance explained by blocking

Dealing w/ contrasts:

$$\hat{\theta} = \sum_i a_i \bar{y}_{i+} \sim N\left(\theta, \sigma^2 \frac{\sum a_i^2}{b}\right)$$

$$U = \frac{(t-1)(b-1)\sigma_r^2}{\sigma^2} \sim \chi^2_{(t-1)(b-1)}$$

$$\Rightarrow \frac{\tilde{\theta} - \theta}{\tilde{\sigma}_r \sqrt{\frac{\sum a_i^2}{b}}} \sim t_{(t-1)(b-1)} \Rightarrow \text{Use for tests}$$

Why is blocking useful? Reducing residual error → easier to detect treatment diff if effect exists.

↳ Blocking prevents treatments to be assigned to same type of units

FACTORIAL TREATMENT STRUCTURE & INTERACTION

Blocking acted like a nuisance variable: we don't care about it but it helps identify treatment diff.

Factorial treatment structure: compare combo of factor levels, each combo considered as a treatment.

Diff. b/w factorial vs. blocked design:

① Intent: factorial cares about all factors, blocked only cares about treatment effect

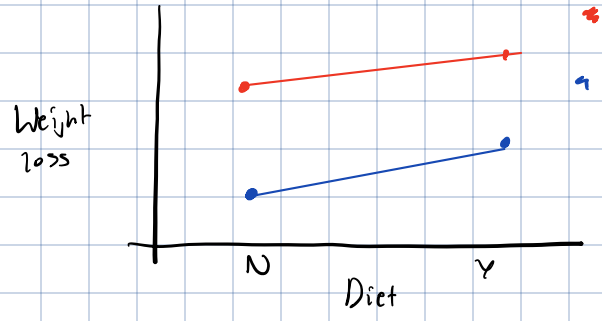
② Randomization: blocked could only randomize treatments in a block, factorial randomizes everything.

Interaction

Defn: effect of 1 factor ← level of another factor

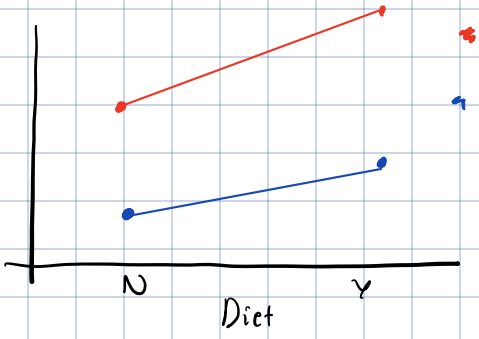
Interaction plot: all treatments in 1 plot

Ex:



⇒ Parallel, so exercise doesn't impact weight loss regardless of diet

However



⇒ Not parallel! If exercising, diet impact is diff.
This is an interaction

+ve interaction plot: effect on response is same direction, but magnitude ↑

-ve " : " differs depending on other factor.
↳ crosses lines in interaction plot.

We can test interaction like a contrast if only 2 factors w/ 2 levels:

$$\theta = (\tau_a - \tau_b) - (\tau_c - \tau_d)$$

Both treatments have same factor level for factor X

Both treatments have same factor level for factor X

⇒ Tests if changing factor X impacts response of changing factor Y

$$\therefore \quad d = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}\sqrt{\frac{1}{r}\sum a_i^2}}$$

↳ Assuming balanced

Interaction Model

Model: $Y_{ijk} = \mu + \alpha_i + \lambda_j + \gamma_{ij} + \beta_{ijk}$, $\beta_{ijk} \sim N(0, \sigma^2)$
$i \in [1, t_a]$, $j \in [1, t_b]$

Intrepretations:

① Y_{ijk}: response of replicate k for factor A at level i, factor B at level j

② μ: mean response across all treatments

additive:

③ α_i: treatment effect for level i for factor A

④ λ_j: " j " B

⑤ γ_{ij}: interaction effect b/w level i for factor A & level j for fact B

Note: $E[Y_{ijk}] = \mu + \alpha_i + \lambda_j + \gamma_{ij} = \mu + \tau$

Num of params = $1 + t_a + t_b + t_a t_b$

Identifiability conditions:

① $\sum_i \alpha_i = 0$

② $\sum_j \lambda_j = 0$

③ $\sum_i \gamma_{ij} = 0 \quad \forall j \in [1, t_b]$

④ $\sum_j \gamma_{ij} = 0 \quad \forall i \in [1, t_a]$

→ $t_a + t_b - 1$ constraints for γ_{ij}

Parameter estimation

① Notation:

$$\bar{Y}_{+++} = \frac{1}{r t_a t_b} \sum_i \sum_j \sum_k Y_{ijk}, \quad \bar{Y}_{++k} = \frac{1}{t_a t_b} \sum_i \sum_j Y_{ijk}$$

avg. of k^{th} replication $\forall$ treatments.

② Estimates:

$$\min_{\mu, \alpha_i, \lambda_j, \gamma_{ij}} \sum_i \sum_j \sum_k (y_{ijk} - \mu - \alpha_i - \lambda_j - \gamma_{ij})^2$$

yields:

① $\hat{\mu} = \bar{y}_{+++}$

② $\hat{\alpha}_i = \bar{y}_{i++} - \bar{y}_{+++}$

③ $\hat{\lambda}_j = \bar{y}_{+j+} - \bar{y}_{+++}$

④ $\hat{\gamma}_{ij} = \bar{y}_{ij+} - \bar{y}_{i++} - \bar{y}_{+j+} + \bar{y}_{+++}$ ⇒ hard to interpret

Also:

$$\hat{\sigma}^2 = \frac{1}{t_a t_b (r-1)} \sum_i \sum_j \sum_k (y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\lambda}_j - \hat{\gamma}_{ij})^2$$

Blocking in interaction model

of replication of each treatment = # of blocks ⇒ 1 treatment per block

$\therefore \quad y_{ijk} = \mu + \alpha_i + \gamma_j + \gamma_{ij} + \beta_k + \varepsilon_{ijk} \quad \Rightarrow \quad i \in [1, t_a] \quad j \in [1, t_b] \quad k \in [1, b]$

Parameter estimation is same, w/ $\hat{\beta}_k = \bar{y}_{++k} - \bar{y}_{+++}$

Decomposition of treatment S.S.

Recall that:

$$SS_{treat} = b \sum_i \sum_j (\bar{y}_{ij+} - \bar{y}_{+++})^2$$

We want:

$$SS_{treat} = SS_A + SS_B + SS_{int}$$

↓ factor S.S. ↓ ↓ interaction

Proof sketch: $\pm \bar{y}_{i++}, \bar{y}_{+j+}, \bar{y}_{+++}$ & $(a+b+c)^2$ identity

$$b \sum_i \sum_j (\bar{y}_{ij+} - \bar{y}_{+++})^2 = bt_b \sum_i^{t_a} (\bar{y}_{i++} - \bar{y}_{+++})^2 \} SS_A$$

$$+ bt_a \sum_j^{t_b} (\bar{y}_{+j+} - \bar{y}_{+++})^2 \} SS_B$$

$$+ b \sum_i \sum_j (\bar{y}_{ij+} - \bar{y}_{i++} - \bar{y}_{+j+} + \bar{y}_{+++})^2 \} SS_{int}$$

* Be able to show this *

D.f. : $df_A + df_B + df_{int} = (t_a - 1) + (t_b - 1) + (t_a - 1)(t_b - 1)$

$= t_a t_b - 1$

↳ SS_{treat} d.f.

ANOVA table:

Source	Sum of squares	DF	Mean square
Treatments	$b\sum_{i,j}(\bar{y}_{ij+} - \bar{y}_{+++})^2$	$t_a t_b - 1$	$\frac{SS_T}{(t_a t_b - 1)}$
- Factor A	$t_b b\sum_{i}(\bar{y}_{i++} - \bar{y}_{+++})^2$	$t_a - 1$	$\frac{SS_A}{(t_a - 1)}$
- Factor B	$t_a b\sum_{j}(\bar{y}_{+j+} - \bar{y}_{+++})^2$	$t_b - 1$	$\frac{SS_B}{(t_b - 1)}$
- Interactions	$b\sum_{i,j}(\bar{y}_{ij+} - \bar{y}_{i++} - \bar{y}_{+j+} + \bar{y}_{+++})^2$	$(t_a - 1)(t_b - 1)$	$\frac{SS_I}{(t_a - 1)(t_b - 1)}$
Blocks	$t_a t_b \sum_{k}(\bar{y}_{++k} - \bar{y}_{+++})^2$	$b - 1$	$\frac{SS_B}{(b - 1)}$
Residual	$\sum_{i,j,k}(y_{ijk} - \bar{y}_{ij+} - \bar{y}_{++k} + \bar{y}_{+++})^2$	$(t_a t_b - 1)(b - 1)$	$\frac{SS_R}{(t_a t_b - 1)(b - 1)}$
Total	$\sum_{i,j,k}(y_{ijk} - \bar{y}_{+++})^2$	$t_a t_b b - 1$	

Testing Interaction Effects

Idea: exact same as testing treatment effect

1. Biased estimator of var. if H_0: no interaction b/w factors is true

$$MS_{int} = \frac{SS_{int}}{(t_a - 1)(t_b - 1)}$$

2. Unbiased estimator for variance

$$MS_{error} = \frac{SS_{error}}{(t_a t_b - 1)(b - 1)}$$

3. F-statistic

$$F = \frac{MS_{int}}{MS_{error}} \sim F_{df_{int}, df_{error}}$$

Important note: can only test treatment if no interaction effect b/c interaction implies that treatment effect changes based on another factor

Testing treatments:

$$\frac{MS_A}{MSE} \sim F_{df_A, df_{error}} \quad \text{OR} \quad \frac{MS_B}{MSE} \sim F_{df_B, df_{error}}$$

$H_0: \alpha_1 = \ldots = \alpha_{t_a} = 0$ $\qquad$ $H_0: \lambda_1 = \ldots = \lambda_{t_b} = 0$

Advantages of factorial structure

Main advantage: enables discussion on interaction effects & more granular levels on factors.

- Having only treatment ANOVA can't let you conclude if factor levels have diff. b/w each other

For post-hoc analysis, we often use regular treatment structure

Overall interaction analysis process

1. If no treatment effect, finish

2. Look for interaction. If there is, interpret & finish

3. Look at factor effects

4. Prove model assumptions are correct

SAMPLE SURVEY ISSUES

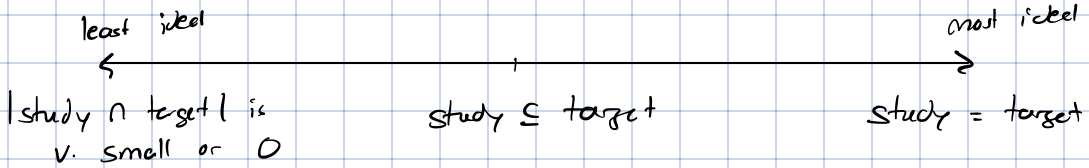
Population is finite → adjust sampling analysis

Terminology

① Sampling unit: individual element in sample which we measure

② Target pop.: pop. we want to study & generalize results

③ Study pop.: pop. that we select samples from

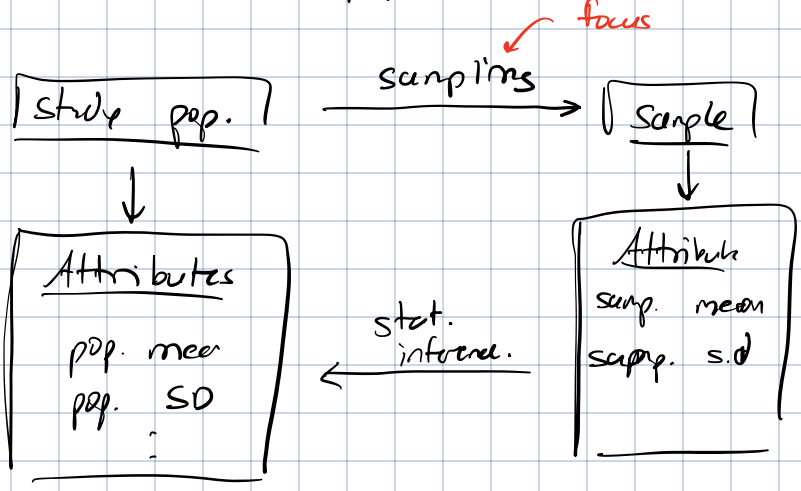


④ Frame: list of units in study pop.

⑤ Sampling: process of gathering data from study pop.

⑥ Attribute: function of pop. that we are interested in (eg. mean, variance, count)

⑦ Census: every unit of pop. is included



Sampling Protocols

Defn: how to select units from study pop.

Sampling protocols

Probabilistic

Assign prob. to ∀ subset of frame → select sample based off distn

Benefit: stat. framework to capture uncertainty due to sampling & stat. inference

Non-probabilistic

Subjective selection of sample

Es:// • Coveniene sampling: take what you get

• Self-selection: units choose to be in sample (eg:// poll)

• Quota: select units if they match target pop. attr.

• Judgment: selected s.t. sample is rep. of target pop. off judgment

No statistical framework to map sample attr. ⟶ pop. attr.

Sampling survey errors

① Study error: | target pop. attr. − study pop. attr |, also called frame error

↳ usually negligible

② Sample error: | study pop. attr − sample attr. |

↳ Es:// MAR (missing at random): missing data indep. of value of var. of interest

↳ No bias issues

MNAR (missing not at random): // depends on //

↳ Bias issues

③ Measurement error: | true sample unit attr. − measured sample unit attr. |

↳ Caused by mistakes, poor technique, lack of precision

Should be able to determine sources of error given scenario.

PROBABILITY SAMPLING

Major advantage: understand error mathematically

Notation:

N: # of units in frame / pop. (in this course)

U: frame

s: sample ($s \subset U$)

n: sample size

y_i: response variable of unit i

Sampling protocol

Defn: how to choose samples. Uniquely det. by $P(s)$ / steps to choose s

↳ Constraint: $\sum_{s \in D} P(s) = 1$

↳ All possible samples

Ex:// $U = \{1, 2, 3\}$. Potential samples $D = \{s_1, s_2, s_3, s_{12}, s_{13}, s_{23}, s_{123}\}$

Protocol A: $P(s_{12}) = P(s_{23}) = P(s_{13}) = 1/3$

All other: $P(s) = 0$

$$P(s) = \begin{cases} 1/3 & \text{if } |s| = 2 \\ 0 & \text{o.w.} \end{cases}$$

Protocol B: $P(s_1) = P(s_{12}) = P(s_{23}) = P(s_{13}) = 1/4$

$P(s)$ ($s \neq$) $= 0$.

Inclusion & joint probabilities

Inclusion prob: $p_i = P(i \in s)$

Joint incl. prob: $p_{ij} = P(i \in s \wedge j \in s)$

} Does not define protocol.

Ex://

s	s_1	s_2	s_3	s_{12}	s_{13}	s_{23}	s_{123}
Sample	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
$P(s)$	1/4	0	0	1/4	1/4	0	1/4

$$p_1 = P(s_1) + P(s_{12}) + P(s_{13}) + P(s_{123}) = 1$$

$$p_2 = P(s_2) + P(s_{12}) + P(s_{23}) + P(s_{123}) = 1/2$$

$$\vdots$$

$$p_{12} = P(s_{12}) + P(s_{123}) = 1/4 + 1/4 = 1/2$$

$$p_{23} = P(s_{23}) + P(s_{123}) = 1/4$$

Common sampling protocols

① Simple random sampling Wout replacement (SRSWOR)

Design: every sample of size n has equal chance

Intuition: 1 unit from pop. → sample until you have n units.

$$P(s) = \begin{cases} \frac{1}{\binom{N}{n}} & |s| = n \\ 0 & \text{o.w.} \end{cases}$$

$$p_i = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

$$p_{ij} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

② Stratified random sampling

Idea: sep. pop. into strata (variation w/in strata ↓, b/w strata var. ↑)
then choose n_h units from strata h, $\sum n_h = n$ ⇒ sample.

$$P(s) = \begin{cases} \frac{1}{\prod_{i=0}^{h} \binom{N_i}{n_i}} & |s| = n \\ 0 & \text{o.w.} \end{cases}$$

$$p_i = \frac{\binom{N_i-1}{n_i-1} \prod_{i=0, j\neq i}^{h} \binom{N_j}{n_j}}{\prod_{i=0}^{h} \binom{N_i}{n_i}}$$

$$\vdots$$

③ Cluster sampling

Idea: sep. pop. into clusters (var. in cluster ↓, var. btw clusters ↓)

Randomly sample clusters → census on each chosen cluster

④ Systematic sampling

Idea: Choose every m^{th} unit in pop., increment m, do again → each time is new sample.

⑤ Two-stage sampling

First stage: cluster sample. 2nd stage: SRSWOR / cluster chosen.

Population parameters

Example params. of interest

$$\mu = \frac{1}{N}\sum Y_i \qquad \text{avg.}$$

$$\tau = \sum Y_i \qquad \text{total.}$$

$$\sigma^2 = \frac{1}{N-1}\sum (y_i - \mu)^2 \qquad \text{variance.}$$

Proportion:

Assume variable of interest is an indicator: $z_i = \begin{cases} 1 & \text{--} \\ 0 & \text{--} \end{cases}$

$$\mu_z = \frac{\sum z_i}{N} = \frac{\tau}{N} = \pi \Rightarrow \text{proportion is avg. of indicator var.}$$

Properties of variance:

① $$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2 = \frac{1}{N-1}\left\{\sum_{i=1}^{N} y_i^2 - N\mu^2\right\}$$

② $\sigma_z^2 \approx \pi(1-\pi)$ as $N \to \infty$

Proof:

$$\sigma_z^2 = \frac{1}{N-1}\sum_{i=1}^{N}(z_i - \mu_z)^2$$

$$= \frac{1}{N-1}\left(\sum_{i=1}^{N} z_i^2 - N\mu_z^2\right)$$

(① ↘, $z_i = \begin{cases} 1 & \\ 0 & \end{cases}$ ↘)

$z_i^2 = \begin{cases} 1 & \cdots \\ 0 & \cdots \end{cases}$

$$= \frac{1}{N-1}\left(\sum_{i=1}^{N} z_i - N\mu_2^2\right)$$

$$= \frac{1}{N-1}\left(N\mu_2 - N\mu_2^2\right)$$

$$= \frac{N(\mu_2 - \mu_2^2)}{N-1}$$

$$= \frac{N}{N-1}\pi(1-\pi)$$

$$\therefore \lim_{N\to\infty} \sigma_2^2 = \pi(1-\pi)$$

Estimates of param.

$$\hat{\mu} = \frac{1}{n}\sum y_i \quad , \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum (y_i - \hat{\mu})^2 \quad , \quad \hat{\tau} = N\hat{\mu}$$

Estimators of params:

$$\tilde{\mu} = \frac{1}{n}\sum_{i\in S} y_i \qquad \tilde{\sigma}^2 = \frac{1}{n-1}\sum_{i\in S}(y_i - \bar{x})^2$$

randomness ↓ det. by protocol.

$$E[R.V.] = \sum_{s\in D} (R.V.\ value)\cdot P(s)$$

Simple random sampling w/out replacement (SRSWOR)

Three results:

① $E[\tilde{\mu}] = \mu$

Proof:

$$E[\tilde{\mu}] = E\left[\frac{1}{n}\sum_{i\in S} y_i\right]$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{N} y_i \cdot I_i\right] \Rightarrow I_i = \begin{cases} 1 & i \in S \\ 0 & \text{o.w.} \end{cases}$$

$$= \frac{1}{n} \sum_{i=1}^{N} y_i \, E[I_i]$$

$$= \frac{1}{n} \sum_{i=1}^{N} y_i \cdot P(i \in s)$$

$$= \frac{1}{n} \cdot \frac{n}{N} \sum_{i=1}^{N} y_i$$

$$= \frac{1}{N} \sum_{i=1}^{N} y_i$$

$$= \mu$$

② $\mathrm{Var}(\tilde{\mu}) = \left(1 - \frac{n}{N}\right) \frac{\sigma^2}{n}$

finite pop. correction → $N \to \infty$, $\mathrm{Var}(\tilde{\mu}) \to \frac{\sigma^2}{n}$

Proof:

$$\mathrm{Cov}(I_i, I_j) = E[I_i I_j] - E[I_i] E[I_j]$$

$$= P_{ij} - P_i P_j$$

$$= \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2$$

$$= -\frac{n}{N} \cdot \frac{1}{N-1} \left(1 - \frac{n}{N}\right)$$

$$\mathrm{Var}(\tilde{\mu}) = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{N} y_i I_i\right)$$

$$= \frac{1}{n^2} \left[\sum_{i=1}^{N} y_i^2 \, \mathrm{Var}(I_i) + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ i \neq j}}^{N} y_i y_j \, \mathrm{Cov}(I_i, I_j) \right]$$

$$= \frac{1}{n^2} \left\{ \sum_{i=1}^{N} y_i^2 \cdot \frac{n}{N} \left(1 - \frac{n}{N}\right) - \sum_{i=1}^{N} \sum_{\substack{j=1 \\ i \neq j}}^{N} y_i y_j \, \frac{n}{N} \cdot \frac{1}{N-1} \left(1 - \frac{n}{N}\right) \right\}$$

$$= \frac{1}{n^2} \cdot \frac{n}{N} \cdot \left(1 - \frac{n}{N}\right) \left[\sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N} \sum_{\substack{j=1 \\ i \neq j}}^{N} \frac{y_i y_j}{N-1} \right]$$

$$= \frac{1}{n^2} \cdot \frac{n}{N} \cdot \left(1 - \frac{n}{N}\right) \cdot N\sigma^2$$

$$= \left(1 - \frac{n}{N}\right) \frac{\sigma^2}{n}$$

① $E[\tilde{\sigma}^2] = \sigma^2$

Do proof

Can derive:

$$\widehat{Var}(\tilde{\mu}) = \left(1 - \frac{n}{N}\right)\frac{\tilde{\sigma}^2}{n}$$

$$\downarrow$$

Standard error: $\sqrt{\widehat{Var}(\tilde{\mu})}$

} Useful for inference about μ

C.I. for μ:

$$\frac{\tilde{\mu} - \mu}{\sqrt{(1-f)\,\tilde{s}^2/n}} \sim N(0,1) \quad \text{if } N, n \text{ \& } (N-n) \text{ are large by CLT}$$

$$\Rightarrow \hat{\mu} \pm c_\alpha \cdot se(\hat{\mu}) = \hat{\mu} \pm c_\alpha \sqrt{(1-f)\frac{\hat{\sigma}^2}{n}} \quad \text{is } 100(1-\alpha)\% \text{ C.I.}$$

Estimation of total:

$$\tilde{\tau} = N \cdot \tilde{\mu}$$

Thus:

$$E[\tilde{\tau}] = N \times E[\tilde{\mu}] = N\mu$$

$$Var(\tilde{\tau})^2 = N^2 \times Var[\tilde{\mu}] \Rightarrow SE(\hat{\tau}) = N \times SE(\hat{\mu})$$

$$\text{C.I. for total} = N\hat{\mu} \pm c \cdot N\sqrt{(1-f)\frac{\hat{\sigma}^2}{n}}$$

$$= N(\text{C.I. } \mu) \Rightarrow \text{multiply endpoints of C.I. } \mu \text{ for C.I. } \tau$$

Confidence Intervals and Coverage Probability

Interpretation of C.I.: if we repeat the sampling process & calculate 95% C.I.s, we expect 95% of C.I. to contain true param.

↳ NOT: $P(\text{param in C.I.}) = 0.95$. Why? Param is fixed! Cannot add prob. statements on fixed #

For CLT, we want $N, n, (N-n) \sim 30$.

↳ If all of these are not large, we can still proceed but look at coverage

Coverage: proportion of C.I. that contain actual param. Should be close to $100(1-\alpha)$

Proportion estimation

If $Y = \{0, 1\}$, then $Var(Y) = \pi(1-\pi)$, π is mean. (if $\frac{N}{N-1} \approx 1$)

$$\therefore \quad \widehat{Var}(Y) = \hat{\sigma}^2 = \hat{\pi}(1-\hat{\pi})$$

$$\therefore \quad S.E. = \sqrt{\left(1-\frac{n}{N}\right)} \cdot \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{\left(1-\frac{n}{N}\right)} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

By CLT (since $E(Y) = \pi$), we have:

$$C.I._{\text{proportion}} = \hat{\pi} \pm c \sqrt{\left(1-\frac{n}{N}\right)} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

If calculating totals of proportions, same trick of multiplying C.I. endpoints exist.

Sample size determination

We often want a C.I. of a certain length → need to determine what is optimal sample size.

Just have to use math: $2 \cdot c \cdot SE(\text{param}) \leq$ desired size.

Ex:// We have $n = 50$ & $N = 362$. How many units should be sampled if 98% C.I. mean should be no longer than 10?

$$\therefore \quad 2 \cdot c \cdot SE(\hat{\mu}) \leq 10$$

$$2 \cdot 2.33 \cdot \sqrt{1-\frac{n}{362}} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \leq 10$$

$\frac{1}{\sqrt{n}} \cdot \sqrt{1-\frac{n}{N}} = \sqrt{\frac{1}{n} - \frac{1}{N}}$

$$2.33 \sqrt{\frac{1}{n} - \frac{1}{362}} \cdot 17.149 \leq 5$$

$$n \geq 54.26$$

Should sample at least 55 units.

Sometimes, we run a pilot study to get data & estimate params ⇓ Sample more to get n.

Ex:// In prev. example, how many units should be sampled if total should be w/in 500 of true value 95% of the.

$$c \cdot SE(\tau) \leq 500$$

$$1.96 \cdot \hat{\sigma} \cdot N \sqrt{\frac{1}{n} - \frac{1}{N}} \leq 500$$

$$n \geq 224.62$$

We need sample total of 225, i.e. additional $225 - 50 = 175$ for req. precision. (50 = pilot)

Ex:// How many units should be sampled if we want to be 95% confident that proportion is within 0.02 of true proportion?

$$C \cdot SE(\hat{\pi}) \leq 0.02$$

$$1.96 \sqrt{\frac{1}{n} - \frac{1}{362}} \cdot \sqrt{0.28(1-0.28)} \leq 0.02$$

$$n \geq 304.97 \Rightarrow \text{Size should be at least 305.}$$

Ex:// $N = 1000$ chips. Want to estimate proportion of defective chips within 5% of the proportion 19 times out of 20. How large should sample be for this precision

$$C.S.E.(\hat{\pi}) \leq 0.05$$

$$1.96 \sqrt{1 - \frac{n}{N}} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \leq 0.05$$

$$n \geq \frac{1}{\frac{1}{1000} + \left[\frac{0.05}{1.96\sqrt{\hat{\pi}(1-\hat{\pi})}}\right]^2}$$

We don't know what $\hat{\pi}$ is, so choose most conservative option. We want n to be as big as possible, so denom should be as small as possible, so $\hat{\pi}(1-\hat{\pi})$ should be as big as possible.

This is when $\hat{\pi}(1-\hat{\pi}) = 0.25$ $\left(\hat{\pi} = 0.5\right)$

$$\therefore \quad n \geq \frac{1}{\frac{1}{1000} + \left[\frac{0.05}{1.96\sqrt{0.25}}\right]^2}$$

$$\geq 277.5$$

278 units should be selected.

Ratio And Regression Estimating w SRS

Estimating a Ratio

Goal: Want to estimate ratios (proportion of avgs. / totals of 2 sub-pops.)

in Waterloo

Ex:// Want to estimate grocery expenses / person across households, took a sample of 10 households

Household # (i)	1	2	3	4	5	6	7	8	9	10
Y Last week's Groceries ($)	200	120	400	80	90	140	210	130	220	60
X # of individuals in the household	3	3	6	2	2	3	4	3	5	1

a) Frame?

List of addresses in Waterloo

b) Parameter:

$$\theta = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \frac{\frac{1}{N}\sum_{i=1}^{N} y_i}{\frac{1}{N}\sum_{i=1}^{N} x_i} = \frac{\mu_Y}{\mu_X}$$

⇒ Both of these are unknown & change b/w samples

c) Estimate:

$$\hat{\theta} = \frac{\sum y_i}{\sum x_i} = \frac{\bar{y}}{\bar{x}} = \frac{200 + \ldots + 60}{3 + \ldots + 1}$$

Note that:

Param:	Estimate	Estimator:
$\theta = \frac{\mu_Y}{\mu_X}$	$\hat{\theta} = \frac{\bar{y}}{\bar{x}}$	$\tilde{\theta} = \frac{\bar{Y}}{\bar{X}}$

To find distn. of $\tilde{\theta}$, we will linearize via Taylor series

- Recall that linear approx. of $f(x,y)$ about (x_0, y_0) is:

$$f(x,y) \approx f(x_0,y_0) + \left.\frac{\partial f(x,y)}{\partial x}\right|_{(x_0,y_0)} (x - x_0) + \left.\frac{\partial f(x,y)}{\partial y}\right|_{(x_0,y_0)} (y - y_0)$$

- Deriving linear approx. for $\tilde{\theta}$:

$$\frac{y}{x} \approx \frac{y_0}{x_0} + \left(\frac{-y_0}{x_0^2}\right)(x - x_0) + \left(\frac{1}{x_0}\right)(y - y_0)$$

$y \to \tilde{\mu}_y$, $x \to \tilde{\mu}_x$

$$\Rightarrow \tilde{\theta} = \frac{\tilde{\mu}_y}{\tilde{\mu}_x} = \frac{\mu_y}{\mu_x} + \frac{1}{\mu_x}(\tilde{\mu}_y - \mu_y) - \frac{\mu_y}{\mu_x^2}(\tilde{\mu}_x - \mu_x)$$

- Mean:

$$E(\tilde{\theta}) \approx \frac{\mu_y}{\mu_x} + \frac{1}{\mu_x}\underbrace{(\mu_y - \mu_y)}_{0} - \frac{\mu_y}{\mu_x^2}\underbrace{(\mu_x - \mu_x)}_{0} = \frac{\mu_y}{\mu_x}$$

$\therefore \tilde{\theta}$ is approx. unbiased

- Variance:

$$\mathrm{Var}(\hat{\theta}) \approx \mathrm{Var}\left(\frac{\mu_y}{\mu_x} + \frac{1}{\mu_x}(\hat{\mu}_y - \mu_y) - \frac{\mu_y}{\mu_x^2}(\hat{\mu}_x - \mu_x)\right)$$

Remove constants

$$= \mathrm{Var}\left(\frac{1}{\mu_x}\hat{\mu}_y - \frac{\mu_y}{\mu_x^2}\hat{\mu}_x\right)$$

$$= \frac{1}{\mu_x^2}\mathrm{Var}\left(\hat{\mu}_y - \frac{\mu_y}{\mu_x}\hat{\mu}_x\right)$$

$$= \frac{1}{\mu_x^2}\mathrm{Var}\left(\hat{\mu}_y - \theta\hat{\mu}_x\right)$$

Note that: $\hat{\mu}_y - \theta\hat{\mu}_x = \frac{\sum_{i\in s} y_i}{n} - \theta\frac{\sum_{i\in s} x_i}{n}$

$$= \frac{1}{n}\sum_{i\in s}(y_i - \theta x_i)$$ (r_i)

$$= \frac{1}{n}\sum_{i\in s} r_i$$

$$= \bar{r}$$

$\theta x_i = \frac{\mu_y}{\mu_x} \cdot x_i \approx \hat{y}_i$

B/c this is SRSWOR:

$$\mathrm{Var}(\bar{r}) = \left(1 - \frac{n}{N}\right)\frac{\sigma_r^2}{n} \quad \Leftarrow \quad \sigma_r^2 = \frac{\sum_{i=1}^{N}(r_i - \mu_r)^2}{N-1}$$

$$\therefore \quad \mathrm{Var}(\hat{\theta}) \approx \frac{1}{\mu_x^2}\left(1 - \frac{n}{N}\right)\frac{\sigma_r^2}{n}$$

Unknown

- Estimate of σ_r^2

$$\hat{\sigma}_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{\theta}x_i)^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(y_i^2 + \hat{\theta}^2 x_i^2 - 2\hat{\theta}x_i y_i)$$

$$= \frac{1}{n-1}\left(\sum_{i\in s} y_i^2 + \hat{\theta}^2\sum_{i\in s} x_i^2 - 2\hat{\theta}\sum_{i\in s} y_i x_i\right)$$

Since $\hat{\theta}$ is approx. linear in $\hat{\mu}_y$ & $\hat{\mu}_x$ $\Rightarrow$ $\hat{\theta} \sim N$

Confidence interval:

$$\frac{\hat{\mu}_y}{\hat{\mu}_x} \pm c \cdot \frac{1}{|\mu_x|}\sqrt{1 - \frac{n}{N}} \cdot \frac{\hat{\sigma}_r}{\sqrt{n}}$$

Ex:// Create 95% C.I. for prev. interval

① Get point estimates:

$$\hat{\theta} = \frac{\bar{y}}{\bar{x}} = \frac{\sum y_i}{\sum x_i} = 51.60 \qquad \hat{\mu}_x = \bar{x} = 3.2$$

$$\hat{\sigma}_r^2 = \frac{1}{n-1}\left(\sum y_i^2 + \hat{\theta}^2 \sum x_i^2 - 2\hat{\theta}\sum x_i y_i\right) = \frac{14517.6992}{10-1}$$

② Combine:

$$\text{C.I.}: \quad \hat{\theta} \pm 1.96 \cdot \frac{1}{|\hat{\mu}_x|}\sqrt{1-\frac{n}{N}} \cdot \frac{\hat{\sigma}_r}{\sqrt{n}}$$

$$\Rightarrow 51.56 \pm 1.96 \cdot \frac{1}{3.2}\sqrt{1-\frac{10}{30000}} \cdot \frac{1}{\sqrt{10}} \cdot \sqrt{\frac{14517.6992}{10-1}}$$

$$\Rightarrow [43.78,\ 59.39]$$

Note that this approach works for looking at ratios of π as well, simply replace.

Ratio estimation of the mean

Idea: We want to estimate μ_y, but if X known & X & Y are corr, can make the estimation of μ_y better?

↳ Only works if $\text{Corr}(X,Y) > 0$

↳ 3 cond. must be met:

① Can measure x in sample

② μ_x (pop.) is known

③ $\text{Corr}(X,Y) > 0$

Mathematically, if we know diff b/w μ_x & $\hat{\mu}_x$, we can correct $\hat{\mu}_y$.

Derivation:

Assume $y_i = \theta x_i + \varepsilon_i \Rightarrow \mu_y = \theta \mu_x \Rightarrow \theta = \frac{\mu_y}{\mu_x}$

Retire: $\hat{\mu}_{ratio} = \hat{\theta}\mu_x = \frac{\hat{\mu}_y}{\hat{\mu}_x}\mu_x = \hat{\mu}_y \cdot \left[\frac{\mu_x}{\hat{\mu}_x}\right]$

Correction!

If X & Y are neg. corr., then $\hat{\mu}_y$ modification would make no sense.

Ex:// Want to estimate brain size (y) w.r.t. height (x).

$$\mu_x = 172.5, \quad \hat{\mu}_x = 151.1, \quad \hat{\mu}_y = 906754.2$$

$$\therefore \hat{\mu}_{ratio}(y) = \frac{\bar{y}}{\bar{x}} \cdot \mu_x = \frac{\mu_x}{\bar{x}} \cdot \bar{y} = \frac{172.5}{151.1} \cdot 906754.2 = 1035166$$

Sample height too low → inflation needed!

Mean of ratio:

$$E[\tilde{\mu}_{ratio}(y)] = E\left[\frac{\tilde{\mu}_y}{\tilde{\mu}_x} \cdot \mu_x\right]$$

$$= \mu_x \cdot E[\tilde{\theta}]$$

$$\approx \mu_x \cdot \theta \quad \text{(Taylor approx.)}$$

$$= \mu_x \cdot \frac{\mu_y}{\mu_x}$$

$$= \mu_y \Rightarrow \text{unbiased}$$

Variance of ratio

$$Var(\tilde{\mu}_{ratio}(y)) = Var\left(\frac{\tilde{\mu}_y}{\tilde{\mu}_x} \cdot \mu_x\right)$$

$$= \mu_x^2 \cdot Var(\tilde{\theta})$$

$$\approx \mu_x^2 \cdot \frac{1}{\mu_x^2} \cdot Var(\tilde{\mu}_y - \theta\tilde{\mu}_x)$$

$$= Var(\tilde{\mu}_y - \theta\tilde{\mu}_x) \in \sigma_r^2$$

$$= \left(1 - \frac{n}{N}\right) \cdot \frac{\sigma_r^2}{n}, \quad \sigma_r^2 = \frac{\sum_{i=1}^{n} (r_i - \mu_r)^2}{n-1}$$

($r_i = y_i - \theta x_i$)

C.I.:

$$C.I. = \hat{\mu}_{ratio} \pm c \cdot \sqrt{\left(1 - \frac{n}{N}\right) \cdot \frac{\sigma_r^2}{n}}$$

When is the ratio estimate more precise than the sample avg.?

$$Var(\tilde{\mu}_{ratio}) < Var(\bar{y})$$

$$\Leftrightarrow \sum_{i \in s} (y_i - \theta x_i)^2 < \sum_{i \in s} (y_i - \bar{y}) \Rightarrow \text{use } \tilde{\mu}_{ratio}$$

- Another way is if the line through the origin of y vs. x explains some of the variation. Why? Model of ratio is $y = \theta x_i + \varepsilon_i$, no intercept!

We could also use LS to estimate θ if $y_i = \theta x_i + \varepsilon_i$

$$\therefore \hat{\theta}_{LS} = \frac{\sum_{i \in s} x_i y_i}{\sum_{i \in s} x_i^2}$$

Regression Estimation of Mean

Ratio estimate better than sample avg. if line of Y vs. X goes through origin

better = less variance

If line doesn't go though origin → regression estimate needed

Regression model is use:

$$Y_i = \alpha + \beta(x_i - \mu_x) + R_i, \quad R_i \sim N(0, \sigma^2)$$

From least squares, we have:

$$\hat{\alpha} = \bar{y}, \quad \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Develop regression estimate:

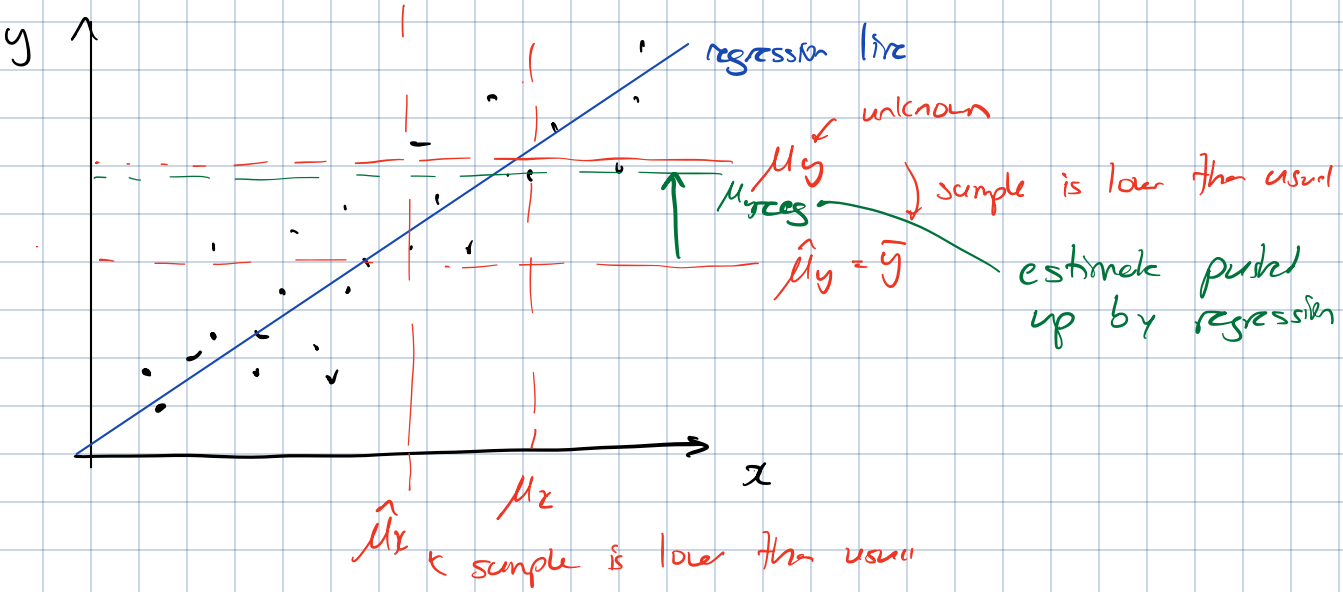
$$\hat{\mu}_{reg} = \hat{\mu}_y + \hat{\beta}(\mu_x - \hat{\mu}_x)$$

$\hat{\mu}_y$: corrected; $\hat{\beta}(\mu_x - \hat{\mu}_x)$: correction; μ_x: pop. avg.; $\hat{\mu}_x$: sample avg.

- $\hat{\beta} > 0$: large values of x → large values of y
  - $\mu_x - \hat{\mu}_x > 0 \Rightarrow \hat{\mu}_{reg}$ adj. up
  - $\mu_x - \hat{\mu}_x < 0 \Rightarrow \hat{\mu}_{reg}$ adj. down
- $\hat{\beta} < 0$: large values of x → small values of y
  - $\mu_x - \hat{\mu}_x > 0 \Rightarrow \hat{\mu}_{reg}$ adj. down
  - $\mu_x - \hat{\mu}_x < 0 \Rightarrow \hat{\mu}_{reg}$ adj. up

Derive from 1st principles of signs.

Ex:



Regression estimator properties

① $E[\tilde{\mu}_{reg}] \approx \mu_y$

Proof: $\tilde{\mu}_{reg} = \bar{y} + \tilde{\beta}(\mu_x - \tilde{\mu}_x)$

$= \bar{y} + (\tilde{\beta} - \beta + \beta)(\mu_x - \tilde{\mu}_x)$

$= \bar{y} + \beta(\mu_x - \tilde{\mu}_x) + (\tilde{\beta} - \beta)(\mu_x - \tilde{\mu}_x)$

$E[\tilde{\mu}_{reg}] = E(\bar{y}) + \beta E[\mu_x - \tilde{\mu}_x] + E((\tilde{\beta} - \beta)(\mu_x - \tilde{\mu}_x))$ ($\mu_x - \tilde{\mu}_x \to 0$)

$= \mu_y + E[(\tilde{\beta} - \beta)(\mu_x - \tilde{\mu}_x)]$

$(\tilde{\beta} - \beta) \to 0$ as $n \to \infty$; $(\mu_x - \tilde{\mu}_x) \to 0$ as $n \to \infty$

$\approx \mu_y$

② $\hat{Var}(\tilde{\mu}_{reg}) = \left(1 - \frac{n}{N}\right) \cdot \frac{1}{n} \cdot \frac{\sum_{i \in s} (y_i - (\bar{y} + \tilde{\beta}(x_i - \bar{x}))^2}{n-1}$ ← SSE

Confidence interval:

$$\tilde{\mu}_{reg} \pm c \times \underbrace{\sqrt{\left(1 - \frac{n}{N}\right) \cdot \frac{1}{n} \cdot \frac{SSE}{n-1}}}_{S.E.(\tilde{\mu}_{reg})}$$

Side note: could we have fit an uncentralized model $Y_i = \lambda + \beta x_i + \varepsilon_i$?

↳ Yup! β is same, $\hat{\lambda} = \bar{y} - \hat{\beta}\bar{x}$

Centralized model in R: $lm(Y \sim I(x - mean(x))$

Uncentralized " : $lm(Y \sim X)$

Comparison

① Sample avg.

② Ratio estimate

③ Regression estimate

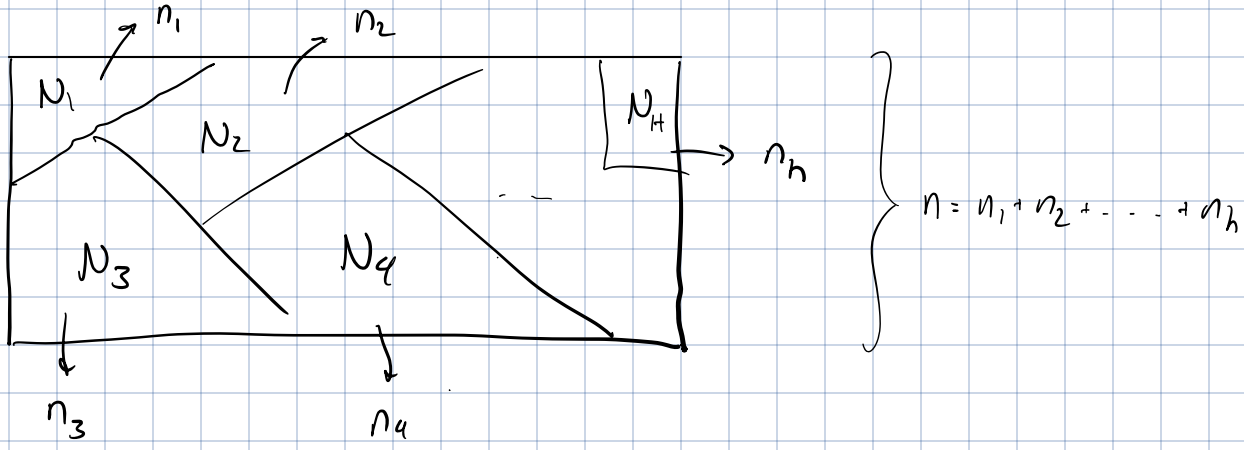
↓ Variance ↓, bias & technicality ↑

STRATIFIED RANDOM SAMPLING

Stratified sampling can yield lower variance if strata chosen carefully

↳ Even though strata is categorical, continuous variable can be bucketed to create strata.

Notation:



μ_h: mean for stratum h, σ_h^2: variance for stratum h

$$= \frac{1}{N_h - 1} \sum_{j=1}^{N_h} (y_{hj} - \mu_h)^2$$

Mean can be considered a weighted avg. of strata:

$$\mu = \frac{N_1}{N}\mu_1 + \dots + \frac{N_h}{N}\mu_h$$

$$= \sum_{h=1}^{H} W_H \mu_H, \quad W_h = \frac{N_h}{N}$$

→ ① $W_h > 0$

② $\sum W_h = 1$

If we are sampling, then:

① $\hat{\mu}_h = \bar{y}_h$

② $\hat{\sigma}_h^2 = \frac{1}{n_h - 1} \sum (y_{hj} - \hat{\mu}_h)^2$

③ $E[\hat{\mu}_h] = \mu_h$ & $Var(\hat{\mu}_h) = \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h}$

} B/c SRSWOR applied per stratum.

This leads to our estimate of the pop. mean:

$$\hat{\mu}_{strata} = \sum_{h=1}^{H} W_h \hat{\mu}_h$$

$$= \sum_{h=1}^{H} \frac{N_h}{N} \cdot \hat{\mu}_h$$

→ NOT $\frac{n_h}{n}$

Properties:

① $E[\hat{\mu}_{strata}] = \mu$

Proof:

$$E[\hat{\mu}_{strata}] = E\left(\sum_{h=1}^{H} W_h \hat{\mu}_h\right)$$

$$= \sum_{h=1}^{H} W_h \cdot E[\hat{\mu}_h]$$

$$= \sum_{h=1}^{H} W_h \cdot \mu_h$$

$$= \mu$$

↓ SRSWOR / stratum, so $\hat{\mu}_h$ is unbiased

② $Var[\hat{\mu}_{strata}] = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \cdot \frac{\sigma_h^2}{n_h}$

Proof:

$$Var[\hat{\mu}_{strata}] = Var\left(\sum_{h=1}^{H} W_h \hat{\mu}_h\right)$$

$$= \sum_{h=1}^{H} W_h^2 \, Var(\hat{\mu}_h) + \sum_{h=1}^{H} \sum_{h^*=1}^{H} Cov(W_h \hat{\mu}_h, W_{h^*} \hat{\mu}_{h^*})$$

0 b/c non-overlapping strata

$Var(\hat{\mu}_h)$ → SRSWOR

$$= \sum_{h=1}^{H} W_h^2 \cdot \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h}$$

Confidence interval

$$\hat{\mu}_{strata} \pm C \cdot \sqrt{\sum_{h=1}^{H} W_h^2 \cdot \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{\sigma}_h^2}{n_h}}$$

Important note: estimates for each strata using diff. techniques can be combined

↳ Remember general formula:

$$\hat{\mu}_{strata} = \sum_{h=1}^{H} W_h \cdot \hat{\mu}_h \quad , \quad Var\left[\hat{\mu}_{strata}\right] = \sum_{h=1}^{H} W_h^2 \cdot Var(\hat{\mu}_h)$$

Could be simple avg., ratio or regression

Estimating proportion via stratified sampling

Since π_h is just an avg. of an indicator variable:

$$\pi_{strata} = \sum_{h=1}^{H} W_h \cdot \pi_h \quad , \quad Var(\hat{\pi}) = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_H}\right) \cdot \frac{\pi_h(1-\pi_h)}{n_h}$$

∴ C.I. is:

$$\hat{\pi}_{strata} \pm C \cdot \sqrt{Var(\hat{\pi})}$$

Stratified vs. SRSWOR

① Variance of mean estimator

$\frac{N_h}{N}$ $\frac{n_h}{n}$

$$Var(\hat{\mu}_{strat}) \approx \left(\frac{1}{n}\right)\left(\frac{W_1^2}{w_1}\sigma_1^2 + \cdots + \frac{W_h^2}{w_h}\sigma_h^2\right)$$

Possible that $Var(\hat{\mu}_{strat}) > Var(\hat{\mu})$, but usually σ_i^2 is very small if strata chosen carefully.

② Pop. structure & strata formation

$$\sigma^2 \approx \underbrace{\sum_{h=1}^{H} W_h \sigma_h^2}_{A} + \underbrace{\sum_{h=1}^{H} W_h (\mu_h - \mu)^2}_{B}$$

A: win group var. B: btw group var.

If strata chosen carefully, A↓, B↑. $SE(\hat{\mu}_{strata})$↓ so C.I. narrower

Allocation

① Proportional

Make samples s.t. $\frac{n_h}{n} = \frac{N_h}{N}$ $\rightarrow$ $n_h = W_h \times n$

Then:

$$Var(\hat{\mu}_{strata}) = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h}$$

$$= \sum_{h=1}^{H} \frac{n_h^2}{n^2} \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h}$$

$$= \frac{\left(1 - \frac{n}{N}\right)}{n} \sum_{h=1}^{H} W_h \cdot \sigma_h^2$$

→ Weighted avg. of strata variances.

↓ smaller than σ^2 if strata are homogenous

② Optimal

Choose $n_1, \ldots, n_h$ s.t. $Var(\hat{\mu}_{str})$ is minimized.

Minimize via:

$$\min_{n_1, \ldots, n_h, \lambda} \left[\sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h} + \lambda \left(\sum_{h=1}^{H} n_h - n \right) \right]$$

Result:

$$n_h = \frac{W_h \sigma_h}{\sum_{h=1}^{H} W_h \sigma_h} \times n$$

$$= \frac{N_h \sigma_h}{\sum_{h=1}^{H} N_h \sigma_h} \times n$$

From this, 2 factors should lead us to sample more from a stratum:

a) Higher stratum variance

b) Bigger stratum

Sample Size Determination

We want to find n s.t.

$$C \times S.E.(\tilde{\mu}_{strata}) \leq \ell$$

We know:

$$Var(\tilde{\mu}_{strata}) = W_1^2\left(1 - \frac{n_1}{N_1}\right)\frac{\sigma_1^2}{n_1} + \ldots + W_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{\sigma_h^2}{n_h}$$

Assuming optimal allocation:

$$n_h = \frac{W_h \hat{\sigma}_h}{\sum W_h \hat{\sigma}_h} \cdot n$$

Assuming $n_i \ll N_i \rightarrow \left(1 - \frac{n_h}{N_h}\right) \approx 1$

Now:

$$\widehat{Var}(\tilde{\mu}_{strata}) \simeq W_1^2 \cdot \frac{\hat{\sigma}_1^2}{n_1} + \ldots + W_h^2 \cdot \frac{\hat{\sigma}_h^2}{n_h}$$

↓ Replacing n_i w/ opt. alloc.

$$= \frac{W_1^2 \hat{\sigma}_1^2}{W_1 \hat{\sigma}_1 n}\sum W_h \sigma_h + \ldots + \frac{W_h^2 \hat{\sigma}_h^2}{W_h \hat{\sigma}_h n} \cdot \sum W_h \sigma_h$$

$$= \left(\frac{\sum W_h \sigma_h}{n}\right)\left(W_1 \hat{\sigma}_1 + \ldots + W_h \hat{\sigma}_h\right)$$

Going back to sample size det:

$$C\sqrt{\left(\frac{\sum W_h \sigma_h}{n}\right)\left(W_1 \hat{\sigma}_1 + \ldots + W_h \hat{\sigma}_h\right)} \leq \ell$$

$$C^2 \cdot \frac{\left(\sum W_h \hat{\sigma}_h\right)^2}{n} \leq \ell^2$$

Not variance, std. dev!

$$n \geq \left(\frac{C \sum W_h \hat{\sigma}_h}{\ell}\right)^2$$

From pilot study, estimate $\hat{\sigma}_h$ & calculate n. Use optimal alloc. scheme to set $n_h \; \forall h$

Ex:// We want to create 95% C.I. of avg. winter hydro usage in Kwh, within 50 Kwh margin of error.

Stratified households by type of heating system:

		Sample size	Mean	Variance
1	Electricity	24	472	202,396
2	Gas	36	463	96,721

$N = 35,000$. $N_1 = 16,450$, $N_2 = 18,550$

a) How much should we sample?

$$n \geq \left(\frac{c \sum W_h \hat{\sigma}_h}{\ell} \right)^2$$

$$\geq \left[\frac{c}{50} \cdot \left(\frac{16450}{35000} \sqrt{202,396} + \frac{18550}{35000} \cdot \sqrt{96721} \right) \right]^2$$

$$\geq 217.6$$

We need at least 218 households sampled.

b) How can we allocate to the different strata?

$$n_1 = \frac{1}{\sum W_h \sigma_h} \cdot \frac{16450}{35000} \cdot \sqrt{202,396} \cdot 218$$

$$= 122.28$$

$$n_2 = 218 - 122 = 96$$

We already sampled 24 & 36 units $\rightarrow$ $n_1^* - n_1 = 122 - 24$

$n_2^* - n_2 = 96 - 36$

Estimating a Controversial / Sensitive Proportion

Sometimes, surveys ask sensitive question which may yield untruthful responses.

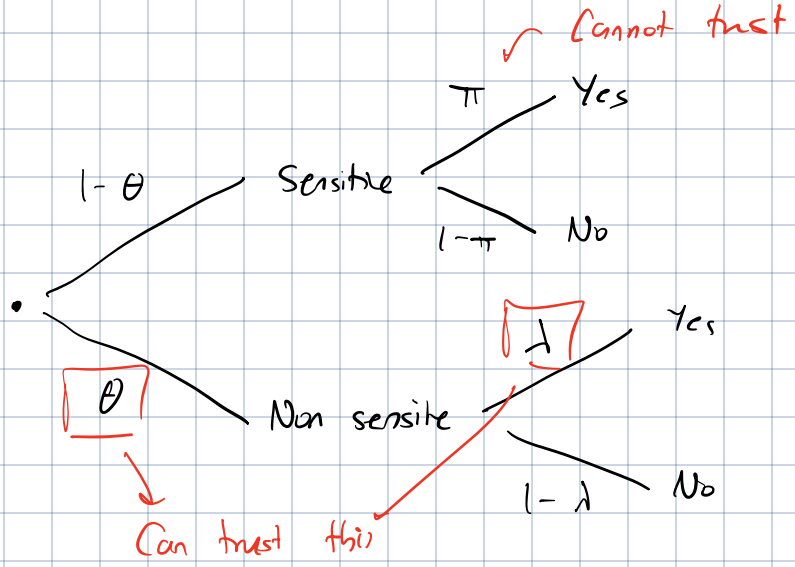
To estimate succesfully, we have 3 potential methods:

① Mix question w/ neutral questions

② Let unit answer question @ random (may / may not be sensitive)

③ Estimate via total probability rule

Methodology



$$P(\text{Yes}) = P(\text{Yes} \mid \text{Non-sensitive}) + P(\text{Yes} \mid \text{Sensitive})$$

$$\Rightarrow \quad p = \theta\lambda + (1-\theta)\pi$$

$$\therefore \quad \pi = \frac{\hat{p} - \theta\lambda}{1-\theta}$$

$$\Rightarrow \quad \hat{\pi} = \frac{\tilde{p} - \theta\lambda}{1-\theta}$$

Deriving C.I. of $\hat{\pi}$:

$$\text{Var}(\hat{\pi}) = \left(\frac{1}{1-\theta}\right)^2 \cdot \text{Var}(\tilde{p} - \theta\lambda)$$

$$= \left(\frac{1}{1-\theta}\right)^2 \cdot \text{Var}(\tilde{p})$$

$$\approx \left(\frac{1}{1-\theta}\right)^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{p(1-p)}{n}$$

Thus:

$$\text{C.I.} \quad \frac{\hat{p} - \theta\lambda}{1-\theta} \pm c_\alpha \cdot \sqrt{\left(\frac{1}{(1-\theta)}\right)^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{\hat{p}(1-\hat{p})}{n}}$$

This method works best if $n \uparrow$, $\theta \downarrow$ (might lead to invalid ans.)

$\hat{\pi}$ may be > 1 or < 0 $\rightarrow$ set to 1 or 0 respectively.