INFERENCE OVERVIEW

Distribution of $\tilde{\mu}$

Context: ① $Y_i = \mu + R_i$, $R_i \sim N(0, \sigma^2)$.

② $\min_{\mu} \sum (y_i - \mu)^2 = \bar{y} = \hat{\mu}$

Claim: $\tilde{\mu} \sim N(\mu, \sigma^2/n)$

Proof: ① Normality:

Since $\tilde{\mu} = \frac{1}{n} \sum Y_i$, it is a linear combination of independent normal variables. Thus, $\tilde{\mu}$ is normally distributed

② Mean:

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

③ Variance

$$Var[\tilde{\mu}] = \frac{1}{n^2} \sum_{i=1}^{n} Var[Y_i] = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

Distribution of discrepancy measure

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

② $Z \sim N(0,1)$, $U \sim \chi^2_k$, Z is indep. of $U \Rightarrow T = \frac{Z}{\sqrt{U/k}} \sim t_k$

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

Proof: We know that $\tilde{\mu} \sim N(\mu, \sigma^2/n)$. Thus:

$$Z = \frac{\tilde{\mu} - \mu}{\sigma\sqrt{1/n}} \sim N(0,1)$$

Then from ②, we have the following:

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

COMPARATIVE EXPERIMENTAL PLANS

Two treatments w/out blocking - parameter estimation

Context: ① $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, $\varepsilon_{ij} \sim N(0, \sigma^2)$, $i \in [1, 2]$, $j \in [1, n_i]$

② Identifiability condition: $n_1\tau_1 + n_2\tau_2 = 0$

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

Proof: ① Define loss function

Lagrange method for constraint optimization

$$\min_{\mu, \tau_1, \tau_2, \lambda} \underbrace{\left[\sum_{i=1}^{2} \sum_{j=1}^{n} (y_{ij} - \overbrace{\mu - \tau_i}^{\hat{y}_{ij}})^2 + \lambda (n_1\tau_1 + n_2\tau_2) \right]}_{Q}$$

② Partial differential system of equations

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

B: $$\frac{\partial Q}{\partial \tau_i} = \sum_{j=1}^{n_i} \left(-2 (y_{ij} - \hat{\mu} - \hat{\tau}_i) + \lambda n_i \right) = 0$$

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

③ Solving by expanding sums & using system

0 from Equation C

A: $$y_{++} - (n_1 + n_2)\hat{\mu} - \cancel{(n_1\hat{\tau}_1 + n_2\hat{\tau}_2)} = 0$$

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

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

Add across all treatments

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

Add & divide by -2

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

0 from C

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

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

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

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

B: $$-2y_{i+} + 2n_i\hat{\mu} + 2n_i\hat{\tau}_i = 0$$

$$\Rightarrow \quad \hat{\tau}_i = \frac{1}{n_i} y_{i+} - \hat{\mu}$$
$$= \bar{y}_{i+} - \bar{y}_{++}$$

Two treatments w/out blocking - parameter distribution

Claim: $\bar{Y}_{1+} - \bar{Y}_{2+} \sim N(\tau_1 - \tau_2, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))$

Proof: ① Prove $\bar{Y}_{i+} \sim N(\mu + \tau_i, \sigma^2/n_i)$

A: Normelility

$\bar{Y}_{i+} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$. Since each Y_{ij} is an independent normal variable, that makes $\bar{Y}_{i+}$ a linear combination of 〃,

so $\bar{Y}_{i+} \sim N$

B: Mean:

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

C: Variance:

$$Var[\bar{Y}_{i+}] = \frac{1}{n_i^2} \sum_{j=1}^{n_i} Var[Y_{ij}] = \frac{1}{n_i^2} \sum_{j=1}^{n_i} \sigma^2 = \frac{\sigma^2}{n_i}$$

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

A: Normality:

Since $\bar{Y}_{i+} \sim N$ and $\bar{Y}_{1+} - \bar{Y}_{2+}$ is a linear combination of indep. normal variables, $\bar{Y}_{1+} - \bar{Y}_{2+} \sim N$

B: Mean:

$$E[\bar{Y}_{1+} - \bar{Y}_{2+}] = \frac{1}{n_1} \sum_{j=1}^{n_1} E[Y_{1j}] - \frac{1}{n_2} \sum_{j=1}^{n_2} E[Y_{2j}]$$
$$= \frac{1}{n_1} \sum_{j=1}^{n_1} (\mu + \tau_1) - \frac{1}{n_2} \sum (\mu + \tau_2)$$
$$= \mu + \tau_1 - \mu - \tau_2$$
$$= \tau_1 - \tau_2$$

C: Variance

$$Var[\bar{Y}_{1+} - \bar{Y}_{2+}] = \frac{1}{n_1^2} \sum Var[Y_{1j}] + \frac{1}{n_2} \sum Var[Y_{2j}]$$
$$= \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$$

Two treatments w/out blocking - discrepancy distribution

Context: ① $U = \frac{(n_1 + n_2 - 1)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{n_1 + n_2 - 1}$

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

Proof: Let:

$$Z = \frac{(\bar{Y}_{1+} - \bar{Y}_{2+}) - \overbrace{(\tau_1 - \tau_2)}^{\theta}}{\sigma \sqrt{1/n_1 + 1/n_2}} \sim N(0,1)$$

We know that if Z & U are indep, $Z \sim N(0,1)$, $U \sim \chi^2_k$, the $\frac{Z}{\sqrt{U/k}} \sim t_k$.

Z & U are indep., so:

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

Two treatments w/ blocking - parameter estimation

Context: ① $Y_{ij} = \mu + \tau_i + \beta_j + R_{ij}$, $R_{ij} \sim N(0, \sigma^2)$, $i \in [1,2]$, $j \in [1, b]$

② Identificability condition 1: $\tau_1 + \tau_2 = 0$

③ " 2: $\sum_{j=1}^{b} \beta_j = 0$

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

Proof: ① Define loss function

$$\min_{\mu, \tau_i, \beta_j, \lambda_1, \lambda_2} \Big[\underbrace{\sum_{i=1}^{2} \sum_{j=1}^{b} (y_{ij} - \mu - \tau_i - \beta_j)^2 + \lambda_1 \sum_{i=1}^{2} \tau_i + \lambda_2 \sum_{j=1}^{b} \beta_j}_{Q} \Big]$$

② Partial derivative system

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

$$B: \frac{\partial Q}{\partial \tau_i} = -2 \sum_{j=1}^{b} (y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j) + \hat{\lambda}_1 = 0$$

$$C: \frac{\partial Q}{\partial \beta_j} = -2 \sum_{i=1}^{2} (y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j) + \hat{\lambda}_2 = 0$$

$$D: \frac{\partial Q}{\partial \lambda_1} = \hat{\tau}_1 + \hat{\tau}_2 = 0$$

$$E: \frac{\partial Q}{\partial \lambda_2} = \sum_{j=1}^{b} \hat{\beta}_j = 0$$

③ Solving system

$$A: y_{++} - 2b\hat{\mu} - b \underbrace{\cancel{\sum_{i=1}^{2} \tau_i}}_{0 \text{ b/c } D} - 2 \underbrace{\cancel{\sum_{j=1}^{b} \hat{\beta}_j}}_{0 \text{ b/c } E} = 0$$

$$\hat{\mu} = \frac{y_{++}}{2b} = \bar{y}_{++}$$

B: $y_{i+} - b\hat{\mu} - b\hat{\tau}_i - \underbrace{\sum_{j=1}^{b} \hat{\beta}_j}_{0 \text{ b/c } E} = \hat{\lambda}_1/2$

$\Rightarrow \quad y_{i+} - \frac{1}{2} y_{++} - b\hat{\tau}_i = \hat{\lambda}_1/2$

$\Rightarrow \quad \sum_{i=1}^{2} \left(y_{i+} - \frac{1}{2} y_{++} - b\hat{\tau}_i \right) = \sum_{i=1}^{2} \hat{\lambda}_1/2$

$\Rightarrow \quad y_{++} - y_{++} - b\underbrace{\cancel{(\hat{\tau}_1 + \hat{\tau}_2)}}_{0 \text{ b/c } D} = \hat{\lambda}_1$

$\Rightarrow \quad \hat{\lambda}_1 = 0$

B: $y_{i+} - \frac{1}{2} y_{++} - b\hat{\tau}_i = 0$

$\Rightarrow \quad b\hat{\tau}_i = y_{i+} - \frac{1}{2} y_{++}$

$\Rightarrow \quad \hat{\tau}_i = \frac{1}{b} y_{i+} - \frac{1}{2b} y_{++}$

$= \bar{y}_{i+} - \bar{y}_{++}$

C: $y_{+j} - 2\hat{\mu} - \underbrace{\cancel{\sum_{i=1}^{2} \hat{\tau}_i}}_{0 \text{ b/c } D} - 2\hat{\beta}_j = \hat{\lambda}_2/2$

$\Rightarrow \quad \sum_{j=1}^{b} \left(y_{+j} - 2\hat{\mu} - 2\hat{\beta}_j \right) = \sum_{j=1}^{b} \hat{\lambda}_2/2$

$\Rightarrow \quad y_{++} - 2b\hat{\mu} - 2\underbrace{\cancel{\sum_{j=1}^{b} \hat{\beta}_j}}_{0 \text{ b/c } E} = \frac{b}{2} \hat{\lambda}_2$

$\Rightarrow \quad y_{++} - y_{++} = \frac{b}{2} \hat{\lambda}_2$

$\Rightarrow \quad \hat{\lambda}_2 = 0$

C: $y_{+j} - 2\hat{\mu} - 2\hat{\beta}_j = 0$

$\Rightarrow \quad \hat{\beta}_j = \frac{y_{+j}}{2} - \hat{\mu}$

$\Rightarrow \quad \hat{\beta}_j = \bar{y}_{+j} - \bar{y}_{++}$

Prove: $\hat{\sigma}^2 = \frac{1}{2} \cdot \frac{1}{b-1} \sum_{j=1}^{b} (d_j - \bar{d})^2$

if $d_j = y_{1j} - y_{2j}$ & $\bar{d} = \frac{1}{b} \sum_j d_j$

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

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

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

$= \frac{1}{b-1}$

Two treatments w/ blocking - parameter distribution

Claim: Let $\theta = \tilde{\tau}_1 - \tilde{\tau}_2$. $\theta \sim N(\tau_1 - \tau_2, 2\sigma^2/b)$

Proof: ① Normality:

$\tilde{\theta} = \bar{Y}_{1+} - \bar{Y}_{2+} = \frac{1}{b} \sum_{j=1}^{b} Y_{1j} - \frac{1}{b} \sum_{j=1}^{b} Y_{2j}$. This is a linear combination of independent normally distributed variables, so $\tilde{\theta} \sim N$

② Mean:

$E[\tilde{\theta}] = \frac{1}{b} \sum_{j=1}^{b} \left\{ E[Y_{1j}] - E[Y_{2j}] \right\} = \frac{1}{b} \sum_{j=1}^{b} (\mu + \tau_1 + \beta_j - \mu - \tau_2 - \beta_j)$

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

$= \tau_1 - \tau_2$

⑤ Variance:

$$\mathrm{Var}[\tilde{\theta}] = \frac{1}{b^2}\sum_{j=1}^{b}\left[\mathrm{Var}[Y_{1j}] + \mathrm{Var}[Y_{2j}]\right] = \frac{1}{b^2}\sum_{j=1}^{b}\left[\sigma^2+\sigma^2\right] = \frac{2\sigma^2}{b}$$

Two treatments w/ blocking - discrepancy distribution

Context: ① $U = \dfrac{(b-1)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{b-1}$

② U & θ are independent

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

Proof: Let Z be:

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

Recall that if $Z \sim N(0,1)$, $U \sim \chi^2_k$ and Z & U are indep., then $\dfrac{Z}{\sqrt{U/k}} \sim t_k$

Since Z is a function of θ and U & θ are indep., then Z & U are indep.
Thus:

$$\frac{\tilde{\theta} - (\tau_1 - \tau_2)}{\sigma\sqrt{2/b}} \times \frac{\sqrt{b-1}\;\sigma}{\sqrt{b-1}\;\tilde{\sigma}} = \frac{\tilde{\theta} - (\tau_1 - \tau_2)}{\tilde{\sigma}\sqrt{2/b}} \sim t_{b-1}$$

EXPERIMENTAL PLANS: ≥ 2 TREATMENTS

Balanced completely randomized design - parameter estimation

Context: ① $Y_{ij} = \mu + \tau_i + R_{ij}$, $i \in [1, t]$, $j \in [1, r]$, $R_{ij} \sim N(0, \sigma^2)$

② Identifiability condition: $\sum_{i=1}^{t} \tau_i = 0$

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

Proof: ① Loss function definition

$$\min_{\mu, \tau_i, \lambda} \Big[\underbrace{\sum_{i=1}^{t}\sum_{j=1}^{r} (y_{ij} - \mu - \tau_i)^2 + \lambda \sum_{i=1}^{t} \tau_i}_{Q} \Big]$$

② Partial derivative system of equations

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

B: $\frac{\partial Q}{\partial \tau_i} = -2\sum_{j=1}^{r}(y_{ij} - \hat{\mu} - \hat{\tau}_i) + \hat{\lambda} = 0$

C: $\frac{\partial Q}{\partial \lambda} = \sum_{i=1}^{t} \hat{\tau}_i = 0$

③ Solving system

A. $y_{++} - tr\hat{\mu} - r\underbrace{\cancel{\sum_{i=1}^{t}\hat{\tau}_i}}_{0 \text{ from } C} = 0$

$$\Rightarrow \hat{\mu} = \frac{y_{++}}{tr} = \bar{y}_{++}$$

B: $y_{i+} - r\hat{\mu} - r\hat{\tau}_i = \hat{\lambda}/2$

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

$\Rightarrow y_{++} - rt\hat{\mu} - r\underbrace{\cancel{\sum_{i=1}^{t}\hat{\tau}_i}}_{0 \text{ from } C} = \frac{t}{2}\hat{\lambda}$

$\Rightarrow y_{++} - y_{++} = t/2\,\hat{\lambda}$

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

B: $y_{i+} - r\hat{\mu} - r\hat{\tau}_i = 0$

$\Rightarrow y_{i+} - t y_{++} - r\hat{\tau}_i = 0$

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

Balanced completely randomized design - parameter distribution

Claim: $\tilde{\theta} := \tilde{\tau}_i - \tilde{\tau}_j \sim N(\tau_i - \tau_j, \frac{2\sigma^2}{r})$

Proof: ① Normality:

$\tilde{\theta} = (\bar{Y}_{i+} - \bar{Y}_{++}) - (\bar{Y}_{j+} - \bar{Y}_{++}) = \bar{Y}_{i+} - \bar{Y}_{j+}$

This is a linear combo. of independent normally distributed variables. Thus, $\tilde{\theta} \sim N$.

② Mean:

$E[\tilde{\theta}] = \frac{1}{r}\sum_{a=1}^{r}(E[Y_{ia}] - E[Y_{ja}]) = \frac{1}{r}\sum_{a=1}^{r}(\mu + \tau_i - \mu - \tau_j) = \tau_i - \tau_j$

③ Variance:

$Var[\tilde{\theta}] = \frac{1}{r^2}\sum_{a=1}^{r}(Var[Y_{ia}] + Var[Y_{ja}]) = \frac{1}{r^2}\sum_{a=1}^{r}(2\sigma^2) = \frac{2\sigma^2}{r}$

Balanced completely randomized design - discrepancy distribution

Context: $U = \dfrac{t(r-1)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{t(r-1)}$ & indep. of $\tilde{\theta}$

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

Proof: From above:

$$Z := \frac{\tilde{\theta} - (\tau_i - \tau_j)}{\sigma\sqrt{2/r}} \sim N(0,1)$$

Recall that if Z & U are indep., $Z \sim N(0,1)$ & $U \sim \chi^2_k$, then $\dfrac{Z}{\sqrt{U/k}} \sim t_k$.

Since U is indep. of θ and $Z = f(\theta)$, Z is indep. of U. Thus:

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

Contrast distribution

Context: ① $\tilde{\theta} = \sum_{i=1}^{t} a_i \bar{Y}_{i+}$, $\sum_{i=1}^{t} a_i = 0$

Claim: $\tilde{\theta} \sim G\left(\theta, \sigma\sqrt{\dfrac{\sum_{i=1}^{t} a_i^2}{r}}\right)$

Proof: ① Normality:

We know $\bar{Y}_{i+}$ is a L.C of indep. normally distributed random variables.
Since a_i are constants, $\tilde{\theta}$ is also a L.C. of " , making it normal.

② Mean:

$$\mathbb{E}[\tilde{\theta}] = \sum_{i=1}^{t} a_i \mathbb{E}[\bar{Y}_{i+}] = \sum_{i=1}^{t} a_i(\mu + \tau_i) = \mu \underbrace{\cancel{\sum_{i=1}^{t} a_i}}_{0} + \sum_{i=1}^{t} a_i\tau_i = \sum_{i=1}^{t} a_i\tau_i = \theta$$

③ Variance:

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

ANOVA Sum of Squares Decomposition

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

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

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

Note that:

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

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

$$= 1/r \; y_{++} - t \cdot \frac{1}{t \cdot r} y_{++}$$

$$= 1/r \; y_{++} - 1/r \; y_{++}$$

$$= 0$$

Thus:

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

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

ANOVA treatment variance

Claim: $\tilde{\sigma}_{tr}^2 = \frac{r}{t-1} \sum_{i=1}^{t} \tilde{\tau}_i^2$

Proof: Assume $H_0: \tau_0 = \dots = \tau_t = 0$ is true.

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

$$\therefore \frac{\hat{\sigma}^2}{r} = \frac{1}{t-1} \sum_{i=1}^{t} (\bar{y}_{i+} - \bar{y}_{++})^2 \Rightarrow \hat{\sigma}_{tr}^2 = \frac{r}{t-1} \sum_{i=1}^{t} \hat{\tau}_i^2$$

$$\therefore \tilde{\sigma}_{tr}^2 = \frac{r}{t-1} \sum_{i=1}^{t} \tilde{\tau}_i^2$$

Randomized block design - parameter estimates

Context: ① $Y_{ij} = \mu + \tau_i + \beta_j$, $i \in [1, t]$, $j \in [1, b]$, $\beta_{ij} \sim N(0, \sigma^2)$

② Identifiability cond. 1: $\sum_i \tau_i = 0$

③ " 2: $\sum_j \beta_j = 0$

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

Proof: ① Loss function:

$$\min_{\mu, \tau_i, \beta_j, \lambda_1, \lambda_2} \underbrace{\left[\sum_{i=1}^{t} \sum_{j=1}^{b} (y_{ij} - \mu - \tau_i - \beta_j)^2 + \lambda_1 \sum_{i=1}^{t} \tau_i + \lambda_2 \sum_{j=1}^{b} \beta_j \right]}_{\mathcal{L}}$$

② Partial derivative system of equations

$$A: \quad \frac{\partial Q}{\partial \mu} = -2 \sum_{i=1}^{t} \sum_{j=1}^{b} (y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j) = 0$$

$$B: \quad \frac{\partial Q}{\partial \tau_i} = -2 \sum_{j=1}^{b} (y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j) + \lambda_1 = 0$$

$$C: \quad \frac{\partial Q}{\partial \beta_j} = -2 \sum_{i=1}^{t} (y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j) + \lambda_2 = 0$$

$$D: \quad \frac{\partial Q}{\partial \lambda_1} = \sum_{i=1}^{t} \hat{\tau}_i = 0$$

$$E: \quad \frac{\partial Q}{\partial \lambda_2} = \sum_{j=1}^{b} \hat{\beta}_j = 0$$

③ Solving:

$$A: \quad y_{++} - tb\hat{\mu} - b \cancelto{0 \text{ from D}}{\sum_{i=1}^{t} \hat{\tau}_i} - t \cancelto{0 \text{ from E}}{\sum_{j=1}^{b} \hat{\beta}_j} = 0$$

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

$$B: \quad y_{i+} - b\hat{\mu} - b\hat{\tau}_i - \cancelto{0 \text{ from E}}{\sum_{j=1}^{b} \hat{\beta}_j} = \hat{\lambda}/2$$

$$\sum_{i=1}^{t} (y_{i+} - b\hat{\mu} \cdot b\hat{\tau}_i) = \sum_{i=1}^{t} \hat{\lambda}_1/2$$

$$y_{++} - tb\hat{\mu} \cdot b \cancelto{0 \text{ from D}}{\sum_{i=1}^{t} \hat{\tau}_i} = \frac{t}{2} \hat{\lambda}_1$$

$$\hat{\lambda}_1 = 0$$

$$B: \quad y_{i+} - b\hat{\mu} - b\hat{\tau}_i = 0$$

$$\hat{\tau}_i = \frac{1}{b} y_{i+} - \hat{\mu} = \bar{y}_{i+} - \bar{y}_{++}$$

$$C: \quad y_{+j} - t\hat{\mu} - \cancelto{0}{\sum_{i=1}^{t} \hat{\tau}_i} - t\hat{\beta}_j = \hat{\lambda}_2/2$$

$$\sum_{j=1}^{b} (y_{+j} - t\hat{\mu} - t\hat{\beta}_j) = \sum_{j=1}^{b} \hat{\lambda}_2/2$$

$$y_{++} - \cancelto{y_{++}}{tb\hat{\mu}} - t \cancelto{0}{\sum_{j=1}^{b} \hat{\beta}_j} = \frac{b}{2} \hat{\lambda}_2$$

$$\hat{\lambda}_2 = 0$$

$$C: \quad y_{+j} - t\hat{\mu} - t\hat{\beta}_j = 0$$

$$\beta_j = \frac{1}{t} y_{+j} - \hat{\mu} = \bar{y}_{+j} - \bar{y}_{++}$$

ANOVA equivalence to t^2

Claim: $F_{obs} = (t_{obs})^2$ if $t = 2$.

Proof: If $t = 2$, then

$$H_0: \tau_1 = \tau_2 = 0$$
$$H_a: \tau_1 \neq \tau_2$$

Our ANOVA discrepancy measure would be:

$$F_{obs} = \frac{MS_{Tr}}{MSE} = \frac{\sum_{i=1}^{2} (\bar{y}_{i+} - \bar{y}_{++})^2}{2-1} \div \frac{\sum_{i=1}^{2}\sum_{j=1}^{r} (y_{ij} - \bar{y}_{i+})^2}{N-2}$$

t_{obs}^2 is:

$$t_{obs}^2 = \frac{(\bar{y}_{1+} - \bar{y}_{2+})^2}{\hat{\sigma}_p^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$\hat{\sigma}_p^2 = MSE \Rightarrow$ already know this.

Now:

$$MST_r = \sum_{i=1}^{2} n_i (\bar{y}_{i+} - \bar{y}_{++})^2 = n_1 (\bar{y}_{1+} - \bar{y}_{++})^2 + n_2 (\bar{y}_{2+} - \bar{y}_{++})^2$$
$$= n_1 \left(\bar{y}_{1+} - \left(\frac{n_1 \bar{y}_{1+} + n_2 \bar{y}_{2+}}{N}\right)\right)^2 + n_2 \left(\bar{y}_{2+} - \left(\frac{n_1 \bar{y}_{1+} + n_2 \bar{y}_{2+}}{N}\right)\right)^2$$

Rewriting: $n_1 \left(\bar{y}_{1+} - \left(\frac{n_1 \bar{y}_{1+} + n_2 \bar{y}_{2+}}{N}\right)\right)^2$

$$\Rightarrow n_1 \left(\frac{N\bar{y}_{1+}}{N} - \left(\frac{n_1 \bar{y}_{1+} + n_2 \bar{y}_{2+}}{N}\right)\right)^2$$

$$\Rightarrow n_1 \left(\frac{N\bar{y}_{1+} - n_1 \bar{y}_{1+} - n_2 \bar{y}_{2+}}{N}\right)^2$$

$$\Rightarrow n_1 \left(\frac{n_2 \bar{y}_{1+} - n_2 \bar{y}_{2+}}{N}\right)^2$$

$$\Rightarrow \frac{n_1 n_2^2}{N^2} (\bar{y}_{1+} - \bar{y}_{2+})^2$$

$\Rightarrow$ Similarly:

$$n_2 \left(\bar{y}_{2+} - \left(\frac{n_1 \bar{y}_{1+} + n_2 \bar{y}_{2+}}{N}\right)\right)^2$$
$$= \frac{n_2 n_1^2}{N^2} (\bar{y}_{2+} - \bar{y}_{1+})^2$$

Thus:

$$MST_r = \frac{n_1 n_2^2}{N^2} (\bar{y}_{1+} - \bar{y}_{2+})^2 + \frac{n_2 n_1^2}{N^2} (\bar{y}_{1+} - \bar{y}_{2+})^2$$
$$= \left(\frac{n_1 n_2^2}{N^2} + \frac{n_2 n_1^2}{N^2}\right) (\bar{y}_{1+} - \bar{y}_{2+})^2$$

Now:

$$\frac{n_1 n_2^2}{N^2} + \frac{n_2 n_1^2}{N^2} = \frac{n_1 n_2 (n_1 + n_2)}{N^2} = \frac{n_1 n_2 N}{N^2} = \frac{n_1 n_2}{N}$$

$$\Rightarrow \frac{1}{\frac{N}{n_1 n_2}} = \frac{1}{\frac{n_1+n_2}{n_1 n_2}} = \frac{1}{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$\therefore \quad MSTr = \frac{(\bar{y}_{1+} - \bar{y}_{2+})^2}{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\therefore \quad \frac{MSTr}{MSE} = \frac{(\bar{y}_{1+} - \bar{y}_{2+})^2}{\hat{\sigma}_p^{\,2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \quad \boxtimes$$

ANOVA Decomposition - blocked.

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

Proof:

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

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

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

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

Note:

$$\left.\begin{aligned}
\sum_{i=1}^{t}\sum_{j=1}^{b}(\bar{y}_{i+} - \bar{y}_{++}) &= b\sum_{i=1}^{t}\bar{y}_{i+} - tb\,\bar{y}_{++} \\
&= tb\,\bar{y}_{++} - tb\,\bar{y}_{++} \\
&= 0 \\
\sum_{i=1}^{t}\sum_{j=1}^{b}(\bar{y}_{+j} - \bar{y}_{++}) &= t\sum_{j=1}^{b}\bar{y}_{+j} - tb\,\bar{y}_{++} \\
&= tb\,\bar{y}_{++} - tb\,\bar{y}_{++} \\
&= 0
\end{aligned}\right\} \quad \therefore \; SST = SSTr + SSB + SSE$$

Factorial design - decomposition of treatment sum of squares

Claim: $SS_{\text{treatment}} = SS_A + SS_B + SS_{int}$

Proof: ① Inject SS_A, SS_B, SS_{int}

$$SS_{treatment} = b \sum_{i=0}^{t_a} \sum_{j=0}^{t_b} \left(\overline{y_{ij+}} - \overline{y_{+++}} \right)^2$$

$$= b \sum_{i=0}^{t_a} \sum_{j=0}^{t_b} \Big[(\overline{y_{i++}} \underset{SS_A}{-} \overline{y_{+++}}) + (\overline{y_{+j+}} \underset{SS_B}{-} \overline{y_{+++}}) + (\overline{y_{ij+}} \underset{SS_{int}}{-} \overline{y_{i++}} - \overline{y_{+j+}} + \overline{y_{+++}}) \Big]^2$$

$$= b t_b \sum_{i=0}^{t_a} (\overline{y_{i++}} - \overline{y_{+++}})^2 + b t_a \sum_{j=0}^{t_b} (\overline{y_{+j+}} - \overline{y_{+++}})^2 + \sum_{i,j} (\overline{y_{ij+}} - \overline{y_{i++}} - \overline{y_{+j+}} + \overline{y_{+++}})^2$$

$$\left.\begin{aligned} &+ 2b \sum_{i,j} (\overline{y_{i++}} - \overline{y_{+++}})(\overline{y_{+j+}} - \overline{y_{+++}}) \\ &+ 2b \sum_{i,j} (\overline{y_{+j+}} - \overline{y_{+++}})(\overline{y_{ij+}} - \overline{y_{i++}} - \overline{y_{+j+}} + \overline{y_{+++}}) \\ &+ 2b \sum_{i,j} (\overline{y_{i++}} - \overline{y_{+++}})(\overline{y_{ij+}} - \overline{y_{i++}} - \overline{y_{+j+}} + \overline{y_{+++}}) \end{aligned}\right\} \text{Prove } 0$$

② Prove cross products $= 0$

$$\sum_{i=0}^{t_a} (\overline{y_{i++}} - \overline{y_{+++}}) = \sum_{i=0}^{t_a} \overline{y_{i++}} - t_a \overline{y_{+++}} \qquad \overline{y_{+++}} = \frac{\sum_i \overline{y_{i++}}}{t_a}$$

$$= t_a \overline{y_{+++}} - t_a \overline{y_{+++}}$$

$$= 0$$

$$\sum_{j=0}^{t_b} (\overline{y_{+j+}} - \overline{y_{+++}}) = \sum_{j=0}^{t_b} \overline{y_{+j+}} - t_b \overline{y_{+++}} \qquad \overline{y_{+++}} = \frac{\sum_j \overline{y_{+j+}}}{t_b}$$

$$= t_b \overline{y_{+++}} - t_b \overline{y_{+++}}$$

$$= 0$$

Probability Sampling

Variance decomposition

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

Proof: $\sum_{i=1}^{N} (y_i - \mu)^2 = \sum_{i=1} (y_i^2 - 2 y_i \mu + \mu^2)$

$$= \sum_{i=1}^{N} y_i^2 - 2\mu \sum_{i=1}^{N} y_i + N\mu^2$$

$$= \sum_{i=1}^{N} y_i^2 - 2\mu \cdot N\mu + N\mu^2$$

$$= \sum_{i=1}^{N} y_i^2 - 2N\mu^2 + N\mu^2$$

$$= \sum_{i=1}^{N} y_i^2 - N\mu^2$$

Variance of proportion

Context: $Z = \begin{cases} 1 & \text{if cond. T} \\ 0 & \text{if cond. F} \end{cases}$ $\cdot$ $\mu_Z = \pi = P(\text{cond.} == T)$

Claim: $\sigma_Z^2 = \pi(1-\pi)$ if $N \to \infty$

Proof:

$$\sigma_Z^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Z_i - \mu_Z)^2$$

$$= \frac{1}{N-1} \left[\sum_{i=1}^{N} Z_i^2 - N\mu_Z^2 \right] \qquad Z_i = \begin{cases} 1 & \dots \\ 0 & \dots \end{cases}$$

$$= \frac{1}{N-1} \left[\sum_{i=1}^{N} Z_i - N\pi^2 \right] \qquad Z_i^2 = \begin{cases} 1 & \dots \\ 0 & \dots \end{cases}$$

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

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

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

$\lim_{N \to \infty} \frac{N}{N-1} = 1$, so $\lim_{N \to \infty} \sigma_Z^2 = \pi(1-\pi)$

SRSWOR - unbiasedness of mean estimator

Claim: $E[\hat{\mu}] = \mu$

Proof:

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

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

$$= \frac{1}{n}\sum_{i\in D} y_i \cdot E[I_i]$$

$$= \frac{1}{n}\sum_{i\in D} y_i \cdot p_i$$

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

$$= \frac{1}{N}\sum_{i\in D} y_i$$

$$= \mu$$

SRSWOR - variance of mean estimator

Claim: $Var(\hat{\mu}) = \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}$

Proof: Let $I_i = \begin{cases} 1 & i \in S \\ 0 & \text{o.w.} \end{cases}$

$$Var(I_i) = E[I_i^2] - E[I_i]^2$$

$$= E[I_i] - E[I_i]^2$$

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

Now: $Cov(I_i, I_j) = E[I_i I_j] - E[I_i]E[I_j]$

$$= p_{ij} - p_i p_j$$

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

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

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

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

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

Na:

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

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

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

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

We know that:

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

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

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

$$= \frac{N}{N-1}\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j$$

$$= \sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j$$

$\frac{N}{N-1} \to 1$ as $N \to \infty$

Thus:

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

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

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

SRSWOR – Unbiasedness of variance estimator

Claim: $E[\tilde{\sigma}^2] = \sigma^2$

Proof:

$$(n-1)\tilde{\sigma}^2 = \sum_{i\in U} y_i^2 I_i - n\tilde{\mu}^2$$

$$(n-1)E[\tilde{\sigma}^2] = \sum_{i\in U} y_i^2 E[I_i] - nE[\tilde{\mu}^2]$$

$$E[\tilde{\mu}^2] = Var[\tilde{\mu}] + E[\tilde{\mu}]^2$$
$$= (1-f)\sigma^2/n + \mu^2$$

$$E[\tilde{\sigma}^2] = \frac{1}{n-1}\left[\sum_{i\in U} y_i^2 E[I_i] - n\left((1-f)\frac{\sigma^2}{n} + \mu^2\right)\right]$$

$$= \frac{1}{n-1}\left[\sum_{i\in U} y_i^2 \frac{n}{N} - n\left(\left(1-\frac{n}{N}\right)\frac{\sigma^2}{n} + \mu^2\right)\right]$$

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

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

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

$$= \frac{1}{n-1}\left[n\sigma^2 - \frac{n}{N}\sigma^2 - \sigma^2 + \frac{n}{N}\sigma^2\right]$$

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

$$= \sigma^2$$

Estimating a Ratio – Properties

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

Claim 1: $E[\tilde{\theta}] = \frac{\mu_y}{\mu_x}$

↳ Proof:

$$E[\tilde{\theta}] = E\left[\frac{\mu_y}{\mu_x} + \frac{1}{\mu_x}(\tilde{\mu}_y - \mu_y) - \frac{\mu_y}{\mu_x^2}(\tilde{\mu}_x - \mu_x)\right]$$

$$= \frac{\mu_y}{\mu_x} + \frac{1}{\mu_x}(E[\tilde{\mu}_y] - \mu_y) - \frac{\mu_y}{\mu_x^2}(E[\tilde{\mu}_x] - \mu_x)$$

$$= \frac{\mu_y}{\mu_x} + \frac{1}{\mu_x}(\mu_y - \mu_y) - \frac{\mu_y}{\mu_x^2}(\mu_x - \mu_x)$$

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

Claim 2: $Var[\tilde{\theta}] = \frac{1}{\mu_x^2}\left(1 - \frac{n}{N}\right)\frac{\sigma_r^2}{n}$

↳ Proof:

$$Var[\tilde{\theta}] = Var\left(\underbrace{\cancel{\frac{\mu_y}{\mu_x}}}_{0 \text{ b/c constant}} + \frac{1}{\mu_x}(\tilde{\mu}_y - \mu_y) - \frac{\mu_y}{\mu_x^2}(\hat{\mu}_x - \mu_x)\right)$$

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

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

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

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

Now: $\tilde{\mu}_y - \theta\tilde{\mu}_x = \frac{1}{n}\sum_{i \in s} y_i - \frac{\theta}{n}\sum_{i \in s} x_i$

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

$$= \bar{r}$$

Therefore: $Var(\tilde{\mu}_y - \theta\hat{\mu}_x) = Var(\bar{r})$ ↓ From SRSWOR

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

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

Thus: $Var(\tilde{\theta}) = \frac{1}{\mu_x^2}\left(1 - \frac{n}{N}\right) \cdot \frac{\sigma_r^2}{n}$

Ratio Estimate of Mean – Properties

Context: $\hat{\mu}_{ratio} = \theta\mu_x = \frac{\tilde{\mu}_y}{\tilde{\mu}_x}\mu_x$

Claim 1: $E[\tilde{\mu}_{ratio}] = \mu_y$

↳ Proof: $E[\tilde{\mu}_{ratio}] = E\left[\frac{\hat{\mu}_y}{\hat{\mu}_x}\mu_x\right]$

$= \frac{\mu_y}{\mu_x}\mu_x$ we know $\tilde{\theta}$ is unbiased for θ

$= \mu_y$

Claim 2: $Var[\tilde{\mu}_{ratio}] \approx \left(1-\frac{n}{N}\right)\frac{\sigma_r^2}{n}$

↳ Proof: $Var[\tilde{\mu}_{ratio}] = Var\left[\frac{\hat{\mu}_y}{\hat{\mu}_x}\mu_x\right]$

$= \mu_x^2 \cdot Var[\tilde{\theta}]$

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

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

Regression Estimate of Mean - Properties

Context: $\hat{\mu}_{res} = \bar{y} + \hat{\beta}(\mu_x - \hat{\mu}_x)$

Claim 1: $E[\tilde{\mu}_{res}] = \mu_y$

↳ Proof: $\tilde{\mu}_{res} = \bar{Y} + \tilde{\beta}(\mu_x - \tilde{\mu}_x)$

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

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

$E[\tilde{\mu}_{res}] = E[\bar{Y}] + \beta(\underbrace{\mu_x - E[\tilde{\mu}_x]}_{0}) + (\underbrace{E[\tilde{\beta}] - \beta}_{0})(\underbrace{\mu_x - E[\tilde{\mu}_x]}_{0})$

$= \mu_y$

Claim 2: $Var[\tilde{\mu}_{res}] = \left(1-\frac{n}{N}\right)\cdot\frac{1}{n}\cdot\frac{SSE}{n-1}$

↳ Proof: $Var[\tilde{\mu}_{res}] = \left(1-\frac{n}{N}\right)\cdot\frac{1}{n}\sigma_r^2$

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

$$\sum_{i=1}^{N}(r_i-\mu_r)^2 = \sum_{i=1}^{N}\left[y_i-\hat{\beta}(x_i-\mu_x)-\bar{y}+\hat{\beta}(\bar{x}-\mu_x)\right]^2$$

$$= \sum_{i=1}^{N}\left(y_i-\hat{\beta}x_i+\hat{\beta}\mu_x-\bar{y}+\hat{\beta}\bar{x}-\hat{\beta}\mu_x\right)^2$$

$$= \sum_{i=1}^{N}\left(y_i-\bar{y}-\hat{\beta}(x_i-\bar{x})\right)^2$$

$$= \sum_{i=1}^{N}\left[\left(y_i-(\bar{y}+\hat{\beta}(x_i-\bar{x})\right)\right]^2 \Rightarrow$$ This is sum of squares of residuals $\to$ SSE!

$$\therefore \text{Var}[\tilde{\mu}_{res}] = \left(1-\frac{n}{N}\right)\cdot\frac{1}{n}\cdot\frac{SSE}{n-1}$$

Stratified Sampling - Mean Estimator Properties

Context: $\hat{\mu}_{stra} = \sum_h W_h \mu_h$

Claim 1: $E[\tilde{\mu}_{stra}] = \mu$

$\hookrightarrow$ Proof:

$$E[\tilde{\mu}_{stra}] = \sum_{h=1}^{H} W_h E[\tilde{\mu}_h]$$

$\tilde{\mu}_h$ is taken via SRSWOR

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

$$= \mu_h$$

Claim 2: $\text{Var}[\tilde{\mu}_{stra}] = \sum_{h=1}^{H} W_h^2 \left(1-\frac{n_h}{N_H}\right)\frac{\sigma_h^2}{n_h}$

$\hookrightarrow$ Proof: $\text{Var}[\tilde{\mu}_{stra}] = \sum_{h=1}^{H}\text{Var}[W_h\tilde{\mu}_h] + \cancel{\sum_i\sum_j \text{Cov}(W_i\hat{\mu}_i, W_j\hat{\mu}_j)}$

0 b/c strata do not overlap

$$= \sum_{h=1}^{H} W_h^2\cdot\text{Var}[\tilde{\mu}_h]$$

from SRSWOR / stratum

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

Stratified random sampling decomposition

Claim. $\sigma^2 \approx \sum_{h=1}^{H} W_h \sigma^2 + \sum_{h=1}^{H} W_h(\mu_h - \mu)^2$

Proof:

$$\sigma^2 = \frac{1}{N-1} \sum_{h=1}^{H} \sum_{j=1}^{N_h} (y_{hj} - \mu)^2$$

$$\Rightarrow (N-1)\sigma^2 = \sum_h \sum_j (y_{hj} - \mu_h + \mu_h - \mu)^2$$

$$= \sum_h \sum_j (y_{hj} - \mu_h)^2 + \sum_h \sum_j (\mu_h - \mu)^2 + 2 \sum_h \sum_j (\mu_h - \mu)(y_{hj} - \mu_h)$$

Note:

$$\sum_h \sum_j (\mu_h - \mu)(y_{hj} - \mu_h) = \sum_h (\mu_h - \mu) \sum_j (y_{hj} - \mu_h)$$

$$\sum_h (\mu_h - \mu) = \mu H - H\mu = 0$$

$$\Rightarrow \sum_h \sum_j (\mu_h - \mu)(y_{hj} - \mu_h) = 0$$

$$\Rightarrow (N-1)\sigma^2 = \sum_h \sum_j (y_{hj} - \mu_h)^2 + \sum_h \sum_{j}^{N_h} (\mu_h - \mu)^2$$

$$= \underbrace{\sum_h (N_h - 1)\sigma_h^2}_{\text{Within strata var.}} + \underbrace{\sum_h N_h (\mu_h - \mu)}_{\text{B/w strata var.}}$$

Proportional allocation - variance

Claim: $Var(\tilde{\mu}_{str}) = (1 - \frac{n}{N}) \cdot \frac{1}{n} \sum W_h \sigma_h^2$

Proof:

$$Var(\hat{\mu}_{str}) = \sum_h W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h}$$

Under prop. alloc.: $\frac{n_h}{n} = \frac{N_h}{N}$

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

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