ONE-FACTOR ANALYSIS MODELS AND ESTIMATORS

Chapter 3-5, 8

LEARNING OBJECTIVES

- Explain conditional distribution
- Write cell-means, effects, and polynomial regression models
- Identify which model is appropriate by identifying type of factor
- Derive expected value and standard error of a linear estimator
- Define estimability

NOTATION
INDICES AND VARIABLES

- Single factor under study that has ***t* unique levels**
  - Index levels by $i = 1, \ldots t$
- Observe $r_i \geq 1$ responses for level ***i***
  - Allow r_i to depend on ***i***, so may not be equal # observations
  - Index responses for given ***i*** by $j = 1, \ldots, r_i$
  - If $r_i = 1$ for all ***i***
- y_{ij} : represents ***j***-th response under factor level ***i***
  - A realization of the random variable, Y_{ij}
- x_{ij} : factor level for y_{ij} $(x_{i1} = x_{i2} = \cdots = x_{ir_i})$

NOTATION
INDICES AND VARIABLES

- Used for both observational studies and designed experiments
- Smoking study design has factor with ***t*=2**
  - x_{1j} = **"Smoking"**
  - x_{2j} = **"Non-Smoking"**
- If equal # of subjects in two groups, $\boldsymbol{r_1 = r_2 = r}$
- Ignore smoking factor and consider Age as a factor?
  - Probably many unique values (large *t*)
  - $r_i = 1$ for many *i* (many 18 year olds but few 77 year olds)

CATEGORICAL AND NUMERIC VARIABLES

- Categorical factor: takes on a finite number of values that may or may not be ordered
  - Ordinal → values have natural ordering but differencing the values doesn't make sense (think rankings)
  - Nominal → no obvious order
- Numeric factor: discrete or continuous but values can be ordered and differences make sense
  - Count data
  - Temperature
  - Age
- Type of factor influences your analysis!

CONDITIONAL DISTRIBUTIONS

- Distribution of Y_{ij} dictates how the y_{ij} are generated
- Analysis goal: does the Y_{ij} distribution change if the factor levels change?
- Asking about the conditional distribution of Y_{ij} given/conditioned on x_{ij}
- If conditional distributions all the same, then no relationship between Y_{ij} and x_{ij}

CONDITIONAL DISTRIBUTIONS EXPECTED VALUE

- Lots of ways the conditional distribution can change
  - Mean (i.e. Expected value)
  - Variance

- Focus solely on changes in expected value
- Represent this dependence mathematically as

$$E(Y_{ij}) = \mu_i \qquad Var(Y_{ij}) = \sigma^2$$

PRACTICE
SOAP EXPERIMENT

- Factor with 3 levels: Regular, Deodorant, Moisturizing
  - Relabel as 1, 2, 3
- Response is weight loss (g)
- 4 cubes per soap type, 1 measurement each
- Draw pictures of distributions assuming normality:
  - $\mu_1 = 0, \mu_2 = 2.5, \mu_3 = 2$
  - $\sigma = 0.25$

CELL-MEANS MODEL
CATEGORICAL FACTORS

- Cell-means model has different mean for each *i*

$$Y_{ij} = \mu_i + E_{ij}$$

- Y_{ij} depends on x_{ij} through μ_i
- Randomness of response comes from error E_{ij} having mean 0 and variance σ^2
  - Assume E_{ij} are independent and normally distributed

- Analysis goal: are the μ_i equal or different?
- If at least one μ_i is different from rest then the conditional distribution changes

CELL-MEANS AND EFFECTS MODEL CATEGORICAL FACTORS

- Cell-means model doesn't clearly state the effect of the treatment, only that means are different
- Rewrite $\mu_i = \mu + \tau_i$
  - μ : overall, constant effect on expected value
  - τ_i : effect specific to x_{ij} (really just *i*)

- Entire effects model is written as

$$Y_{ij} = \mu + \tau_i + E_{ij}$$

$$i = 1, \ldots, t$$
$$j = 1 \ldots, r_i$$
$$E_{ij} \sim^{iid} N(0, \sigma^2)$$

- iid = independent, identically distributed

VISUALIZING MODELS
CELL MEANS

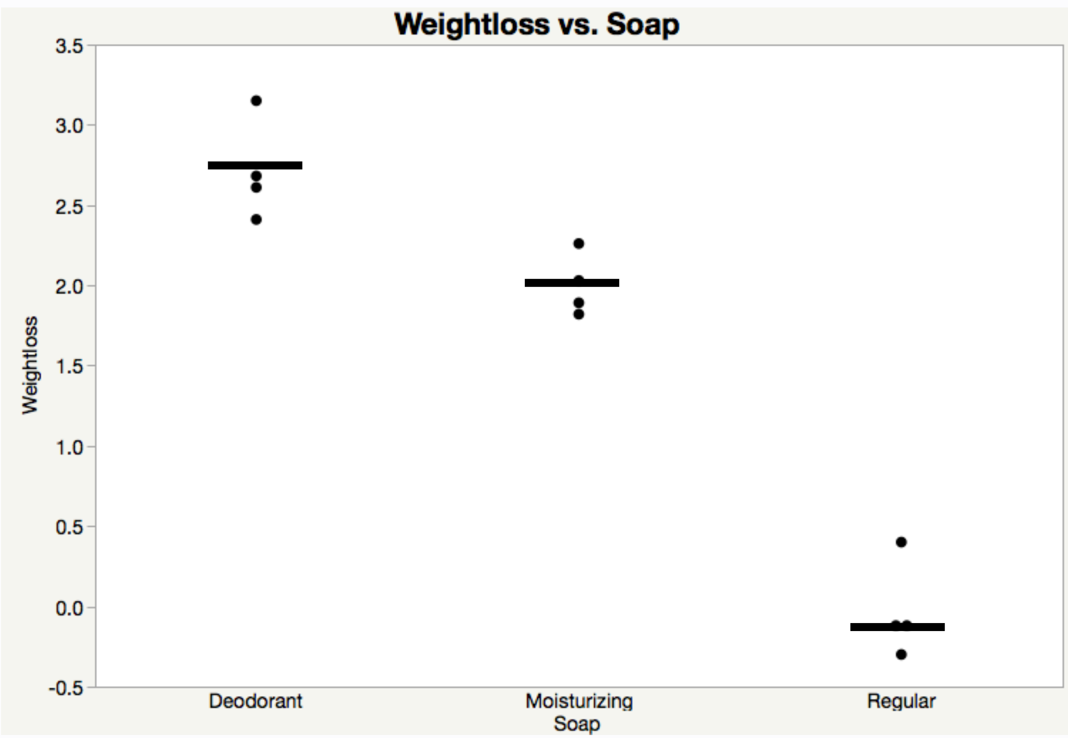
- Recall soap experiment:
  - x_{ij} = "Regular", "Deodorant", "Moisturizing"
  - Y_{ij} = weight loss (in grams)

$$E\left(Y_{ij}\right) = \mu = 1.5?$$

Probably not. We expect points to be fairly symmetric about their expected value

VISUALIZING MODELS CELL MEANS

- Recall soap experiment:
  - x_{ij} = "Regular", "Deodorant", "Moisturizing"
  - Y_{ij} = weight loss (in grams)

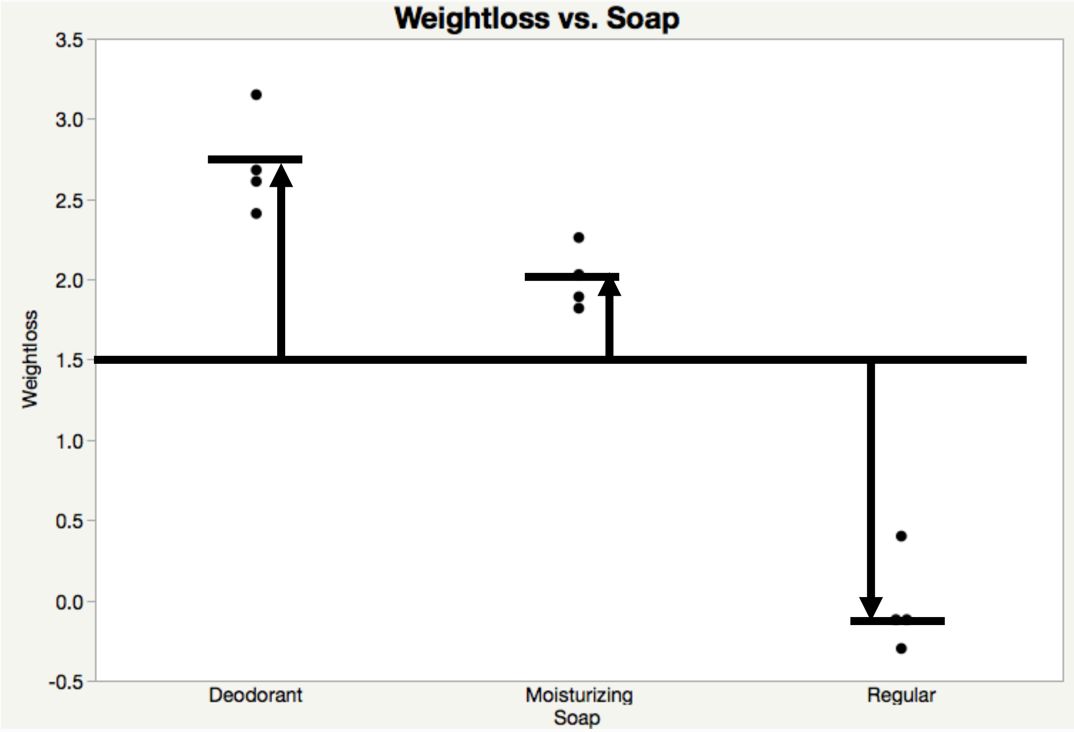


$$\mu_1 = 2.70$$
$$\mu_2 = 1.99$$
$$\mu_3 = -0.04$$

Looks pretty good!

VISUALIZING MODELS
EFFECTS MODEL

- Recall soap experiment:
  - x_{ij} = "Regular", "Deodorant", "Moisturizing"
  - Y_{ij} = weight loss (in grams)



$$\mu = 1.50$$
$$\tau_1 = 1.20$$
$$\tau_2 = 0.49$$
$$\tau_3 = -1.54$$

$$\mu_1 = 2.70$$
$$\mu_2 = 1.99$$
$$\mu_3 = -0.04$$

Same as before!

VISUALIZING MODELS
EFFECTS MODEL

- Recall soap experiment:
  - x_{ij} = "Regular", "Deodorant", "Moisturizing"
  - Y_{ij} = weight loss (in grams)

$$
\begin{aligned}
\mu &= 0.00 \\
\tau_1 &= 2.70 \\
\tau_2 &= 1.99 \\
\tau_3 &= -0.04
\end{aligned}
$$

$$
\begin{aligned}
\mu_1 &= 2.70 \\
\mu_2 &= 1.99 \\
\mu_3 &= -0.04
\end{aligned}
$$

Wait....same as before?

OVERPARAMETERIZED MODELS

- Cell-means model has *t* unique x_{ij}'s and *t* μ_i
- Effects model has *t* unique x_{ij}'s but *t+1* parameters
  - Say it is overparameterized

- To make the model parameters uniquely identifiable, you must impose side conditions such as

$$\mu = 0 \qquad \tau_t = 0 \qquad \sum_i \tau_i = 0 \qquad \sum_i r_i \tau_i = 0$$

- Avoid this and talk about estimability later on

SIMPLE LINEAR REGRESSION MODEL NUMERIC FACTORS

- If x_{ij} is numeric then can use the cell-means or effects model but not recommended
- Reason: *t* is usually large and $r_i = 1$ so there are many parameters that we need to estimate

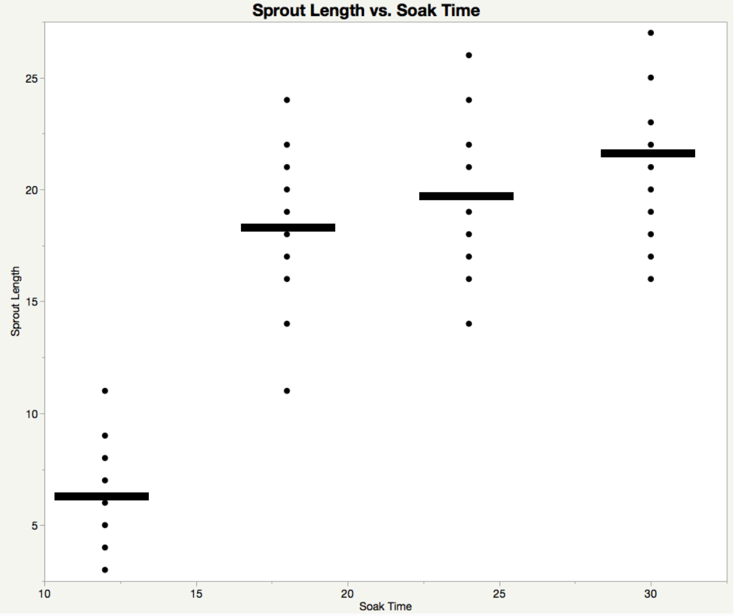
- Simple linear regression proposes a simple relationship using only two parameters

$$\mu_i = \beta_0 + \beta_1 x_{ij}$$

- Again assume $x_{i1} = x_{i2} = \cdots = x_{ir_i}$
- Mean increases/decreases linearly as x_{ij} increases

VISUALIZING MODELS
CELL MEANS FOR NUMERIC

- **Bean-soaking experiment:** packaging says to soak mung bean seed sprouts overnight but no specific time is given
  - x_{ij} = 12, 18, 24, 30 hours
  - Y_{ij} = sprout length (mm) after 48 hours

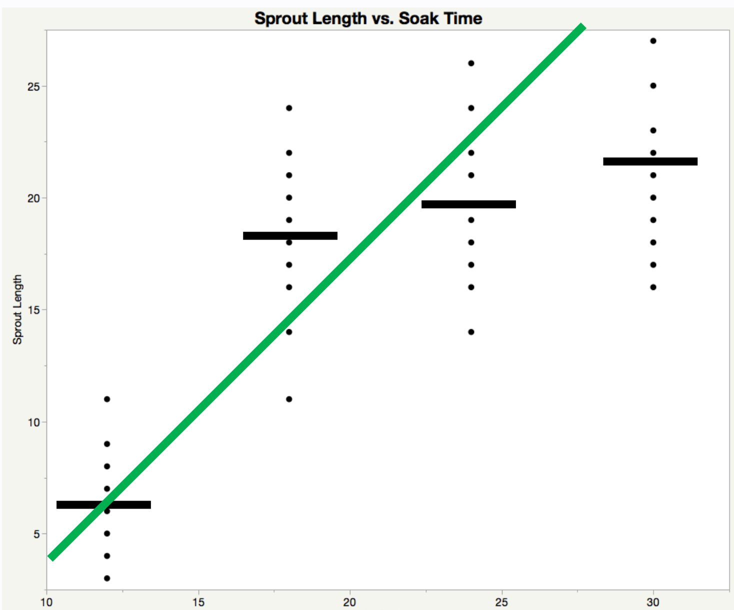


Cell-means model could be

$$\begin{aligned} \mu_1 &= 5.94 \\ \mu_2 &= 18.41 \\ \mu_3 &= 19.53 \\ \mu_4 &= 21.29 \end{aligned}$$

VISUALIZING MODELS
SIMPLE LINEAR REGRESSION

- **Bean-soaking experiment:** packaging says to soak mung bean seed sprouts overnight but no specific time is given
  - x_{ij} = 12, 18, 24, 30 hours
  - Y_{ij} = sprout length (mm) after 48 hours



Regression model could be

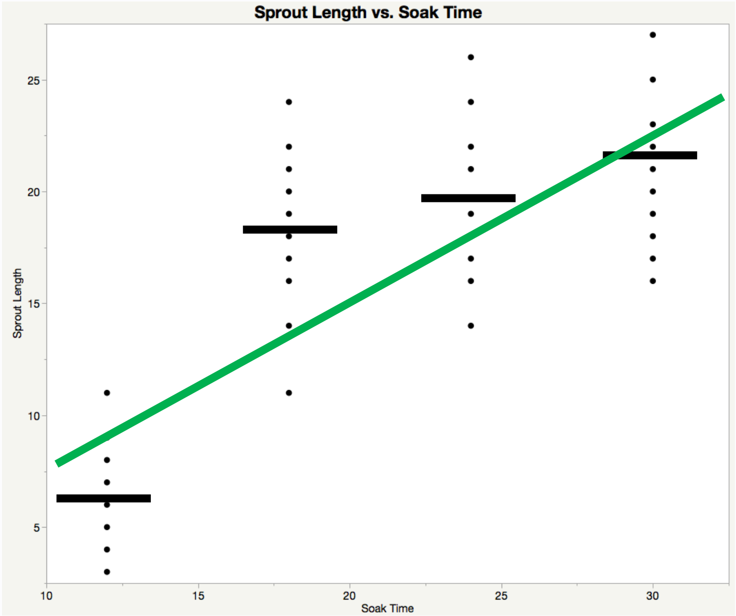
$$\mu_i = \beta_0 + \beta_1 x_{ij}$$

$$\beta_0 = 0 \quad \beta_1 = 1$$

Probably not. Poor mean for $x_i = 30$

VISUALIZING MODELS
SIMPLE LINEAR REGRESSION

- Bean-soaking experiment: packaging says to soak mung bean seed sprouts overnight but no specific time is given
  - x_{ij} = 12, 18, 24, 30 hours
  - Y_{ij} = sprout length (mm) after 48 hours $r_i = 17$



Regression model could be

$$\mu_i = \beta_0 + \beta_1 x_{ij}$$

$\beta_0 = -0.217 \quad \beta_1 = 0.786$

Looks better, but is still poor

POLYNOMIAL REGRESSION MODEL

- A linear relationship may be too simplistic
- The quadratic regression model allows for curvature

$$\mu_i = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2$$

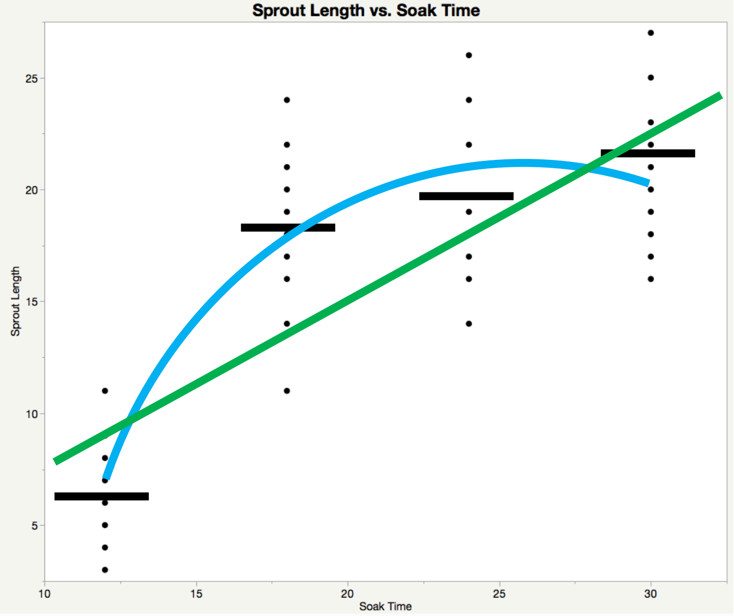
- A polynomial regression model is of the form

$$\mu_i = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \cdots + \beta_p x_{ij}^p$$

- These are still linear models because we never take any nonlinear functions of the parameters

VISUALIZING MODELS
QUADRATIC REGRESSION

- **Bean-soaking experiment:** packaging says to soak mung bean seed sprouts overnight but no specific time is given
  - x_{ij} = 12, 18, 24, 30 hours
  - Y_{ij} = sprout length (mm) after 48 hours $r_i = 17$



Regression model could be

$$\mu_i = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2$$

$$\beta_0 = -29.66 \quad \beta_1 = 3.91$$
$$\beta_2 = -0.07$$

Better than linear!

CELL-MEANS VERSUS POLYNOMIALS NUMERIC FACTORS

- Cell-mean models
  - Capture complicated relationships but require many parameters
  - Can’t predict for unobserved factor values
- Polynomial models
  - Approximate relationships fairly well with fewer parameters
  - Can predict for unobserved factor values
  - Do not extrapolate predictions outside of the observed values!
- What do we do after we decide on a model?

STATISTICAL INFERENCE

- Statistical models involve unknown parameters
- Use observed data to infer what the parameters are
- An estimator of a parameter is a function of the data that informs us about the parameter
- Where do these estimators come from?
- How to compare competing estimators?

MEAN SQUARED ERROR

- Let $\hat{\mu}_i$ denote some estimator for μ_i

- $MSE(\hat{\mu}_i) = E\left[\left(\hat{\mu}_i - \mu_i\right)^2\right]$
  - Want this difference to be as small as possible

- Has the following decomposition

$$MSE(\hat{\mu}_i) = Var(\hat{\mu}_i) + Bias(\hat{\mu}_i)^2 \qquad Bias(\hat{\mu}_i) = E(\hat{\mu}_i - \mu_i)$$

- If $E(\hat{\mu}_i) = \mu_i$ then $Bias(\hat{\mu}_i) = 0$

- Call $\hat{\mu}_i$ an **unbiased estimator**

PARAMETER ESTIMATION USING LEAST-SQUARES

- Least-squares (LS) estimators minimize

$$\sum_i \sum_j \left(Y_{ij} - \hat{\mu}_i\right)^2$$

- Fact: LS estimators can be represented by

$$\sum_i \sum_j h_{ij} Y_{ij}$$

- Estimators of this form are called linear estimators
  - Linear combination of Y_{ij}

STATISTICAL PROPERTIES OF LINEAR ESTIMATORS

- Expected value always distribute over sums

$$E\left(\sum_i \sum_j h_{ij} Y_{ij}\right) = \sum_i \sum_j E(h_{ij} Y_{ij})$$

- Distribute over constants (non-random)

$$\sum_i \sum_j E(h_{ij} Y_{ij}) = \sum_i \sum_j h_{ij} E\left(Y_{ij}\right) = \sum_i \sum_j h_{ij} \mu_i$$

- Since μ_i doesn't have a j subscript we can simplify to

$$\sum_i \sum_j h_{ij} \mu_i = \sum_i h_{i.} \mu_i \qquad\qquad h_{i.} = \sum_j h_{ij}$$

- Result: linear estimator is unbiased for some linear combination of μ_i

STATISTICAL PROPERTIES OF LINEAR ESTIMATORS

- A linear combination, $\sum_i c_i \mu_i$, is **estimable** if there exists a linear, unbiased estimator:

$$E\left(\sum_i \sum_j h_{ij} Y_{ij}\right) = \sum_i c_i \mu_i$$

- From before we must have $h_{i\cdot} = c_i$
- Extract the c_i from given expression
  - μ_1 has $c_1 = 1$ and $c_2 = \cdots = c_t = 0$
  - $\mu_1 - \mu_2 = \mu_1 + (-\mu_2)$ has $c_1 = 1, c_2 = -1, c_3 = \cdots c_t = 0$

- A **contrast** is a $\sum_i c_i \mu_i$ where $\sum_i c_i = 0$

LEAST-SQUARES ESTIMATORS CELL-MEANS MODEL

- For cell-means model we have

$$\hat{\mu}_i = \sum_j \frac{1}{r_i} Y_{ij} = \frac{1}{r_i} \sum_j Y_{ij} = \frac{1}{r_i} Y_{i.} = \bar{Y}_{i.}$$

- Average of the responses for value *i*

- $E(\bar{Y}_{i.}) = \mu_i \qquad Var(\bar{Y}_{i.}) = \sigma^2 / r_i$

- **Design impact:** if you increase r_i you decrease variance of your estimator

LEAST-SQUARES ESTIMATORS SIMPLE LINEAR REGRESSION

- For simple linear regression

$$\hat{\beta}_0 = \bar{Y}_{..} - \hat{\beta}_1 \bar{x}_{..} \qquad \hat{\beta}_1 = \frac{\sum_i \sum_j (x_{ij} - \bar{x}_{..})(Y_{ij} - \bar{Y}_{..})}{\sum_i \sum_j (x_{ij} - \bar{x}_{..})^2}$$

- $E(\hat{\beta}_0) = \beta_0 \qquad Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_{..}^2}{\sum_i \sum_j (x_{ij} - \bar{x}_{..})^2} \right)$
- $E(\hat{\beta}_1) = \beta_1 \qquad Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i \sum_j (x_{ij} - \bar{x}_{..})^2}$
- $n = \sum_i r_i$ is the total number of observations
- **Design impact:** if you increase $\sum_i \sum_j (x_{ij} - \bar{x}_{..})^2$ the variance decreases for BOTH parameters

LEAST-SQUARES ESTIMATORS EFFECTS MODEL

- Remember that identifiability issue? Tells us that the individual parameters may not be estimable
- Every estimable function of the form

$$\sum_i c_i \mu_i = \sum_i c_i(\mu + \tau_i) = \mu \sum_i c_i + \sum_i c_i \tau_i$$

- For μ to be estimable by itself we need to pick c_i so that $\sum_i c_i \tau_i = 0$ for every possible τ_i (does not exist)
  - Can't estimate individual τ_i either

- What functions are estimable?
  - $\mu_i = \mu + \tau_i$
  - Contrasts: $\sum_i c_i \tau_i$ where $\sum_i c_i = 0$ (e.g. $\tau_i - \tau_{i'}$)

LEASTSQUARES ESTIMATORS EFFECTS MODEL

- Even though parameters aren't estimable we still have least-squares estimators for them
  - An infinite number of them and none of them are unbiased
  - Different software give different estimators

- Still use these estimators for estimable functions
  - Say we have estimators $\hat{\mu}$ and $\hat{\tau}_i$
  - Least-squares estimators for estimable functions are then

$$\sum_i \widehat{c_i \tau_i} = \sum_i c_i \hat{\tau}_i \quad \text{and} \quad \widehat{\mu + \tau_i} = \hat{\mu} + \hat{\tau}_i$$

- **Important:** estimable function estimator same regardless of the chosen $\hat{\mu}$ and $\hat{\tau}_i$

LEAST-SQUARES ESTIMATORS EFFECTS MODEL

- Simplifications for this model, but not generally
- $\sum_i c_i \hat{\tau}_i = \sum_i c_i \overline{Y}_{i.}$ $\qquad Var(\sum_i c_i \overline{Y}_{i.}) = \sigma^2 \sum_i \frac{c_i^2}{r_i}$
- $\hat{\mu} + \hat{\tau}_i = \overline{Y}_{i.}$ $\qquad Var(\overline{Y}_{i.}) = \frac{\sigma^2}{r_i}$
- **Design impact:** increasing r_i for all treatments in a given contrast decreases variance of that contrast
- If equally interested in all contrasts then maximize r_i
  - Why equal replication is recommended!

LEARNING OBJECTIVES REVIEW

- Explain a conditional distribution
- Write cell-means, effects, and polynomial regression models
- Identify when which model is appropriate by identifying type of factor
- Derive expected value and standard error of a linear estimator
- Define estimability

APPENDIX: MORE ON NOTATION

- **Subscripts $i = 1, \ldots, t$ and $j = 1, \ldots, r_i$ are necessary tools for framework that applies to many analyses**
- **Linear combinations involve real numbers h_{ij} indexed by *i* and *j* in table**
- **Think of arranging these numbers in a table**
  - **Example:** $i = 1, .., 3$ **with** $r_1 = 5, r_2 = 3, r_4 = 9$

					j					
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	1	h_{11}	h_{12}	h_{13}	h_{14}	h_{15}				
i	**2**	h_{21}	h_{22}	h_{23}						
	3	h_{31}	h_{32}	h_{33}	h_{34}	h_{35}	h_{36}	h_{37}	h_{38}	h_{39}

APPENDIX: MORE ON NOTATION

- The sums for each row are denoted by $\sum_j h_{ij} = h_{i\cdot}$
- Previous example:
  - $h_{1\cdot} = h_{11} + h_{12} + h_{13} + h_{14} + h_{15}$
  - $h_{2\cdot} = h_{21} + h_{22} + h_{23}$

- The sums for each column are denoted by $\sum_i h_{ij} = h_{\cdot j}$
- Previous example:
  - $h_{\cdot 1} = h_{11} + h_{21} + h_{31}$
  - $h_{4\cdot} = h_{14} + h_{34}$ (why is h_{24} missing from here?)

- The overall sum is $\sum_i \sum_j h_{ij} = h_{\cdot\cdot}$

APPENDIX: MORE ON NOTATION

- The overall sum can be expressed in two other ways
  - $h.. = \sum_i h_i.$
  - $h.. = \sum_j h._j$
- Practice notation with the following table

		j								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
i	**1**	1	1	2	2	10				
	2	0.5	−0.5	0						
	3	1	2	3	4	5	6	7	8	9

$h_1. = 16$ $\quad h._2 = 2.5$ $\quad h._9 = 9$ $\quad h.. = 16 + 0 + 45 = 61$