1 Introduction

This is mostly review from Stat 231.

1.1 PPDAC

Problem: Define the problem, discuss the target population (T.P.), response (the variable that represents the answer given by the target pop to the problem), and attribute. For example, we might ask "What is the average grade of students in Stat 332?". The target population is Stat 332 students, response is the grade of a Stat 332 student, and attribute is average grade.

Plan: How will we answer the problem? Define the study population (S.P.). Recall the study population is not necessarily a subset of the target population; eg for drug testing, study pop is monkeys, target pop is humans. Define the sample, which is always a subset of the study pop.

Data: Collect the data, according to the plan and record any unusual circumstances.

Analysis: Analyze the data.

Conclusion: Refer back to the problem, note any errors. Three main types of errors: study error (qualitative, hard to quantify since you don't know the true attribute of the T.P.), sample error (for example $\mu - \bar{x}$ but since μ is usually unknown, this error is also qualitatively described), and measurement error.

1.2 Independent and dependent groups

There are three ways to create dependent groups:

1. Matching: randomly select one group and find a match group, having the same explanatory variates, such that each unit of the first group is matched with a unit of the second group; thus the two groups are dependent/related
2. Twins: twins already come in pairs, assign one member of each pair into one group and the other into another
3. Reusing units: starting with a treatment on one group, stop the treatment and start another treatment using the same group. This naturally creates twins (each unit twinned with itself). For example, a psychology experiment which starts people off driving a car simulator sober then again inebriated.

Essentially in each method, the members of both groups have the same explanatory variates thus are related in some way.

In contrast, independent groups are any two groups with no dependency. One way to interpret this (although it is not technically correct) is that units are selected at random from mutually exclusive groups. For example, choosing broken parts and not broken parts of a machine.

1.3 Response models

Response models relate a parameter to a response.

Model 1: $Y_j = \mu + R_j$ where $R_j \sim N(0, \sigma^2)$. Here, Y_j is the response of unit j and is random, μ is the population mean (not random but is unknown), and R_j is the error term which captures the distribution of responses around μ. In this course, always assume R_j is independent of each other.

By Gauss's theorem, any linear combination of Normal rvs is itself Normal. Using this fact, it can be shown that $Y_j \sim N(\mu, \sigma^2)$, since $E[Y_j] = E[\mu + R_j] = \mu + E[R_j] = \mu$ and $\text{Var}(Y_j) = \text{Var}(\mu + R_j) = \text{Var}(R_j) = \sigma^2$.

For example, consider the average grade of Stat 332 students. Model 1 is good to use since the response Y_j (individual grade) is related to the average grade μ plus some error.

Model 2A: $Y_{ij} = \mu_i + R_{ij}$ where $R_{ij} \sim N(0, \sigma^2)$ and the groups are independent but with the same std dev σ. Here, Y_{ij} is the response of the jth unit in the ith group, μ_i is the mean for group i (not random, but typically unknown), and R_{ij} is the error term which captures the distribution of responses around μ_i.

For example, this model can be used to describe heart rates between teenagers and seniors who have different mean heart rates but probably similar variance. Note the two groups do not have to be the same size (typically group 1 has size denoted n_1 and group 2 has n_2).

See pg 234 of the Stat 231 book (Chapter 6.3) for more details.

Model 2B: $Y_{ij} = \mu_i + R_{ij}$ where $R_{ij} \sim N(0, \sigma_i^2)$. The groups are independent but std devs between groups are not the same.

Model 3: Models differences in responses between two dependent groups $Y_{1j} = \mu_1 + R_{1j}, Y_{2j} = \mu_2 + R_{2j}$. Then, $Y_{1j} - Y_{2j} = \mu_1 - \mu_2 + R_{1j} - R_{2j}$. Let $Y_{1j} - Y_{2j} = Y_{dj}$, let $\mu_1 - \mu_2 = \mu_d$, and let $R_{1j} - R_{2j} = R_{dj}$. Thus, Model 3 is: $Y_{dj} = \mu_d + R_{dj}$ where $R_{dj} \sim N(0, \sigma_d^2)$. Note R_{dj} are still independent and that σ_d is the std dev of the differences and has no relation to the std dev within each group.

For example, this model can be used for modelling difference in heart rate before and after exercise. Note that this model is the same as Model 1 but with a sample of differences specifically rather than a sample of any generic data values. So, Model 3 is a "sub-model" of Model 1.

Model 4: Recall $Y \sim \text{BIN}(n, p)$ has n outcomes, each which are each a binary response (true or false, etc) and $E[Y] = np, \text{Var}(Y) = np(1-p)$. By CLT, Y is approximately $N(np, np(1-p))$ and the proportion Y/n is $N(p, p(1-p)/n)$. p is often denoted π in this course.

1.4 Maximum Likelihood Estimate (MLE)

The MLE connects the population parameter (generally denoted θ, eg might be the p in a $BIN(n, p)$ model) to the sample statistic (denoted $\hat{\theta}$). The MLE is the most probable value of θ given our data. Process:

1. Define the likelihood function $L = f(Y_1 = y_1, \ldots, Y_n = y_n)$; assume Y's are independent (recall the definition of a random sample from Stat 231). Thus, $L = f(Y_1 = y_1) \cdots f(Y_n = y_n)$. Note to not drop constants at this step even though it doesn't affect maximum. Rather, when taking the log then derivative, these constants will disappear to 0.
2. Define the log likelihood function which is more convenient when taking derivatives
3. Find $\frac{dl}{d\theta}$ and set it to 0. Possibly need partial derivatives for multi-parameter distributions such as Normal. Put hats on the θ: we are trying to solve for $\hat{\theta}$ which is the MLE in the equation $\frac{dl}{d\hat{\theta}} = 0$

See Lecture 7 MLE example for one involving Model 2A (which is essentially pg 234 of the Stat 231 notes). The overall result is that the MLEs are: $\hat{\mu}_1 = \bar{y}_{1+}, \hat{\mu}_2 = \bar{y}_{2+}$ where $\bar{y}_{1+}$ means the sample average of the Y_{1j}'s (+ in the second index indicates summing over the second index). The MLE for σ can be found by taking the derivative w.r.t σ ($\frac{dl}{d\sigma}$).

For the normal distribution, the MLE for σ^2 is biased although the least squares result is unbiased if the error term is Normal. Instead, use the unbiased sample variance (pooled estimate of variance, see pg 234 of Stat 231).

An estimator for θ ($\tilde{\theta}$) is unbiased iff $E[\tilde{\theta}] = \theta$ where $\tilde{\theta}$.

1.5 Least Squares (LS)

For the Models seen so far (1, 2A, 2B, 3, 4), the general form is always:

$$\text{response} = \text{deterministic part} + \text{random part}$$

or alternatively, $Y = f(\theta) + R$. The realizations of the rv Y are $y_1, y_2, \ldots, y_n$ and the prediction is $\hat{y}_i = f(\hat{\theta})$, eg $\hat{\mu}$, where $f(\hat{\theta})$ is simply $f(\theta)$ with θ replaced by $\hat{\theta}$. The residual is $r_i = y_i - f(\hat{\theta}) = y_i - \hat{y}_i$ (see pg 230 of Stat 231 notes).

LS is another technique in addition to MLE for finding $\hat{\theta}$ which minimizes the residuals. The general procedure for solving is:

1. Define W function which is a function of the non-sigma parameters in the model. W is the sum of the residuals squared
2. Calculate $\frac{\partial W}{\partial \theta}$ for all θ which are non-sigma parameters
3. Solve for $\hat{\theta}$

See Lecture 9 for an example based on Model 2A. In this example, $W = \sum_{ij} r_{ij}^2$ and we calculate $\frac{\partial W}{\partial \hat{\mu}_1}, \frac{\partial W}{\partial \hat{\mu}_2}$ since μ_1, μ_2 are the only non-sigma parameters. The end results are $\hat{\mu}_1 = \bar{y}_{1+}, \hat{\mu}_2 = \bar{y}_{2+}$ which are exactly the MLE estimates. Note that $\hat{\sigma}^2$ is always of the form $\hat{\sigma}^2 = \frac{W}{n-q+c}$ where n is the sample size, q is the number of non-sigma parameters, and c is the number of constraints. In this example, $n = n_1 + n_2, c = 0, q - 2$ and $\frac{W}{n-2}$ is exactly the pooled estimate.

In comparison with MLE, LS is older (from the 1860s) and is unbiased provided the errors R_j are Normally distributed. MLE is a recent technique and is more flexible since it does not require R_j to be Normal.

Can assume the result is a minimum when using LS (don't need to prove using the second derivative).

See chapter 6.4 of Stat 231 notes for more details.

1.6 Estimators

$y_1, y_2, \ldots, y_n$ is sample data which is not random and is a realization of the rvs $Y_1, \ldots, Y_n$. A statistic, denoted $\hat{\theta}$, is a function of the sample data (eg mean, mode) and is also not random but changes when the sample data changes. Thus, $\hat{\theta}$ can be thought of as the realization of a rv $\tilde{\theta}$ called the estimator. To move from $\hat{\theta}$ to $\tilde{\theta}$, just replace the data points with rvs (capitalize y's to Y's). We care about the distribution of the estimator.

How do you know if a rv is Normal? Recall Gauss's theorem: any linear combination of Normal rvs is also Normal. Let $X \sim N(\mu_x, \sigma_x^2), Y \sim N(\mu_y, \sigma_y^2)$ and X, Y be independent and a, b, c be constants where a or b is not 0. Let $L = aX + bY + c$. Then, $L \sim N(E[L], \text{Var}(L))$.

Central Limit Theorem (CLT): Let $Y_1, \ldots, Y_n$ be a sequence of rvs, let $E[Y_i] = \mu$ for all i, $\text{Var}(Y_i) = \sigma^2 < \infty$ for all i, and Y_i be independent of each other. Then, the average of these Y_i, $\bar{Y}$ is approximately $N(\mu, \sigma^2/n)$ and the approximation is better as $n \to \infty$. This theorem tells us that the initial distributions of Y_i doesn't matter, the average is always Normal.

Standard error: Standard error is the standard deviation of the estimator. In many cases, the estimator is for the mean (biased) so the standard error would be the standard deviation divided by $\sqrt{n}$.

Example (see Lec 11.00): For Model 2A, what is the distribution of $\tilde{\mu}_1$, the estimator for $\hat{\mu}_1$ (the unbiased estimate of μ_1)? Using LS or MLE, we obtained $\hat{\mu}_1 = \bar{y}_{1+}$ so the estimator is $\tilde{\mu}_1 = \bar{Y}_{1+}$. Since $\bar{Y}_{1+}$ is a linear combination of Y_{ij}, it is Normal by Gauss's theorem and can be shown that the mean of this Normal distribution is μ_1 since $E[R_{ij}] = 0$ and the variance is σ^2/n_1 since Cov is always 0 by independence. Therefore, this is an unbiased estimator; $E[\tilde{\theta}] = \theta$ (which is μ_1 in this case). Similarly, it can be shown that $\tilde{\mu}_2 \sim N(\mu_2, \sigma^2/n_2)$.

1.7 Sigma

Various theorems first:

1. Theorem 1: Let $Z \sim N(0,1)$. Then $Z^2 \sim \chi^2(1)$. This is Theorem 30 from Pg 146 of Stat 231.
2. Theorem 2: Let $X \sim \chi^2(m), Y \sim \chi^2(n)$ and X, Y are independent. Then $X + Y \sim chi^2(m+n)$.
3. Theorem 3: Let $Z \sim N(0,1), X \sim \chi^2(m)$. Then $\frac{Z}{\sqrt{X/m}} \sim t(m)$. This is Theorem 32 from Pg 148 of Stat 231.
4. Theorem 4: Let $Y = \frac{(n-q+c)\tilde{\sigma}^2}{\sigma^2}$. Then $Y \sim \chi^2(n-q+c)$.

Example (Lecture 13.00): Consider Model 1: $Y_j = \mu + R_j$ where $R_j \sim N(0, \sigma^2)$. What is the distribution of $\frac{\tilde{\mu}-\mu}{\tilde{\sigma}/\sqrt{n}}$? We know by LS or MLE, $\hat{\mu} = \bar{y}_+$ so $\tilde{\mu} = \bar{Y}_+$. By CLT, $\tilde{\mu}$ is approximately $N(\mu, \sigma^2/n)$. Standardizing, we get that $Z = \frac{\tilde{\mu}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. By Theorem 4, $X = \frac{(n-1)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2(n-1)$. By Theorem 3, $\frac{Z}{\sqrt{X/(n-1)}} = \frac{\tilde{\mu}-\mu}{\tilde{\sigma}/\sqrt{n}}$ is $t(n-1)$. Therefore, by replacing σ by $\tilde{\sigma}$ in Z, we get a t distribution instead of a Normal.

1.8 Confidence Intervals (CI)

For this course, we assume our estimator is unbiased and,

$$\tilde{\theta} \sim N(\theta, \text{Var}(\tilde{\theta}))$$

Thus, the CI for θ is:

$$\theta : \text{estimate} \pm c \cdot (\text{standard error})$$

where the standard error is $\sigma = \sqrt{\text{Var}(\tilde{\theta})}$. If σ is unknown, then it is replaced in the CI by $\hat{\sigma}$ (the estimate of σ).

Model 1 example: By LS, we know that $\hat{\mu} = \bar{y}_+$ and the distribution of the estimator $\tilde{\mu}$ is $N(\mu, \sigma^2/n)$. Thus, the CI for μ when σ is known (see pg 139 of Stat 231 book for more details) is:

$$\mu : \text{estimate} \pm c \cdot \sigma/\sqrt{n}$$

where the estimate is $\hat{\mu} = \bar{y}_+$ using the notation for the estimate $\hat{\mu}$ and $C \sim N(0,1)$, where C is the rv associated with c. For example, to find a 95% CI, we compute c such that $P(-c < C < c) = 0.95$. When σ is unknown (see pg 153 of Stat 231):

$$\mu : \bar{y}_+ \pm c \cdot s/\sqrt{n}$$

where s is the estimate (sample variance) from the unbiased estimator for variance, $s = \frac{\sum_i (y_i - \bar{y})^2}{n-1}$ and $C \sim t(n-1)$.

Model 2A example: By LS (MLE gives same result), $\hat{\mu}_1 = \bar{y}_{1+}$ and $\hat{\mu}_2 = \bar{y}_{2+}$ and the estimators are $\tilde{\mu}_1 = \bar{Y}_{1+}$ and $\tilde{\mu}_2 = \bar{Y}_{2+}$. The distributions by CLT are $\tilde{\mu}_1 \sim N(\mu_1, \sigma^2/n_1)$ and $\tilde{\mu}_2 \sim N(\mu_1, \sigma^2/n_2)$. Comparing the means of the two groups (see pg 234 of Stat 231), $\tilde{\mu}_1 - \tilde{\mu}_2 \sim N(\mu_1 - \mu_2, \sigma^2/n_1 + \sigma^2/n_2)$ and the CI if σ is known is:

$$\mu_1 - \mu_2 : \hat{\mu}_1 - \hat{\mu}_2 \pm c\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $C \sim N(0,1)$. If σ is unknown (see pg 235 of Stat 231), we use the pooled estimate of variance s_p,

$$\mu_1 - \mu_2 : \hat{\mu}_1 - \hat{\mu}_2 \pm cs_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $C \sim t(n_1 + n_2 - 2)$. For example, if we want a 95% CI, we solve for c in $P(-c < C < c) = 0.95$.

Model 2B example: Recall in Model 2B, the variance is not the same between the two populations. So, $\tilde{\mu}_1 - \tilde{\mu}_2 \sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$ (see pg 238 of Stat 231) and the CI if σ_1, σ_2 are known is

$$\mu_1 - \mu_2 : \hat{\mu}_1 - \hat{\mu}_2 \pm c\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $C \sim N(0,1)$. Supposing σ_1, σ_2 are unknown, the CI is,

$$\mu_1 - \mu_2 : \hat{\mu}_1 - \hat{\mu}_2 \pm c\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where s_1, s_2 are the sample std devs and C is approximately $t(n_1 + n_2 - 2)$.

Model 3 example: Recall Model 3 is the same as Model 1 but with differences rather than general values. So the CI when σ_d is known is,

$$\mu_d : \bar{y}_{d+} \pm c\sigma_d/\sqrt{n_d}$$

where $C \sim N(0,1)$ and when σ_d is unknown,

$$\mu_d : \bar{y}_{d+} \pm cs_d/\sqrt{n_d}$$

where $C \sim t(n_d - 1)$.

Model 4 example: The CI for p (the proportion) from CLT is,

$$p : \hat{p} \pm c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $C \sim N(0,1)$.

Note that "σ is known" can also mean that σ came from a previous study, even if it was estimated. The important point is that the estimate came from a different one than the one currently being conducted and analyzed. In summary, if σ is the true population std dev or came from a prior study, then use Normal and otherwise, if you calculate σ using sample data, use t.

1.9 Hypothesis Testing (HT)

Confidence intervals are exploratory and builds an interval of plausible values for your parameter. Hypothesis tests are confirmatory and confirms whether or not a certain value is plausible. The four steps for HT are:

1. Define the hypothesis H_0 (null hypothesis $\theta = \theta_0$ or $\theta \geq \theta_0$ or $\theta \leq \theta_0$) and H_A (alternative hypothesis $\theta \neq \theta_0$ or $\theta < \theta_0$ or $\theta > \theta_0$). Note that for the null hypothesis, all three options are equivalent (eg you can have $H_0 : \theta = \theta_0$ and $H_A : \theta > \theta_0$).

2. Calculate discrepancy

$$d = \frac{\text{estimate} - H_0\text{value}}{\text{std error}} = \frac{\hat{\theta} - \theta_0}{\sqrt{\text{Var}(\tilde{\theta})}}$$

   given $\tilde{\theta} \sim N(\theta, \text{Var}(\tilde{\theta}))$. As well, D (the rv associated with d) is $N(0,1)$ when the std error σ is known, or $t(n - q + c)$ when σ is unknown and replaced by sample std dev.

3. Calculate p value. If H_A is $\theta \neq \theta_0$ (two-sided tests), then the p value is $2P(D > |d|)$ (the two tails beyond $-|d|$ and $|d|$). If H_A is $\theta > \theta_0$ (one-sided test), then the p value is $P(D > d)$ and if H_A is $\theta < \theta_0$, then the p value is $P(D < d)$. These various probabilities represent the probability (calculated assuming H_0 is true) of observing a value of the test statistic which is "more extreme" than what the sample indicates. For two sided, "more extreme" means larger on the positive side, smaller on the negative side, etc.

4. Make a conclusion based on the p value. The simple way is to choose a significance value α and reject H_0 if p value is $< \alpha$. The more thorough approach is to use a gradient (see Table 5.1 on pg 184 of Stat 231).

A p value does not provide any actual evidence supporting H_A. If H_0 is rejected in favor of H_A, this just means we have no choice but to accept H_A. Note that for each Model, there is a corresponding discrepancy and distribution.

Model 1: for $D \sim t(n-1)$ (when σ is unknown and s is used instead) or for $D \sim N(0,1)$ (when σ is known and replaces s in the following equation; the discrepancy using true σ is $N(0,1)$ for all models),

$$d = \frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Model 2A: for $D \sim t(n_1 + n_2 - 2)$, using the pooled estimate s_p if σ is unknown,

$$d = \frac{\hat{\mu}_1 - \hat{\mu}_2 - \mu_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Model 2B: for $D \sim t(n_1 + n_2 - 2)$,

$$d = \frac{\hat{\mu}_1 - \hat{\mu}_2 - \mu_0}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

Model 3: for $D \sim t(n_d - 1)$,

$$d = \frac{\hat{\mu}_d - \mu_0}{\frac{\hat{\sigma}_d}{\sqrt{n_d}}}$$

Model 4: for $D \sim N(0,1)$, f

$$d = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \text{ or using } \pi \text{ notation, } d = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

2 Balanced CRD

Model 5: Completely randomized design (CRD), $Y_{ij} = \mu + \tau_i + R_{ij}$ where $R_{ij} \sim N(0, \sigma^2)$ for $i = 1, 2, \ldots, t$ where t is the number of treatments (groups), $j = 1, 2, \ldots, r$ where r is the number of replicates per treatment. So, there are tr units in this model. This model is "balanced" because the number of replicates is the same in every treatment.

μ is the overall (study population) mean, $\mu + \tau_i$ is the group mean, and τ_i is the treatment effect of group i. R_{ij} is the distribution of values about the deterministic part of the model μ. This model also has a constraint: $\tau_1 + \tau_2 + \cdots + \tau_t = 0$, which will be useful for minimizing the least squares regression line. Y_{ij} independent and groups are also independent.

Example: Suppose there are 2 groups with values $60, 65, 70$ in group 1 and $70, 75, 80$ in group 2. The overall average $\hat{\mu}$ is 70, $\hat{\mu} + \hat{\tau}_1 = 65$ (average of group 1), and $\hat{\mu} + \hat{\tau}_2 = 75$. Thus, the treatment effects are $\hat{\tau}_1 = -5$ and $\hat{\tau}_2 = 5$. We can think of treatment effects as the average difference between the elements in a certain group with the overall.

Least Squares for Model 5: The W function is

$$W = \sum_{ij} r_{ij}^2 + \lambda(\tau_1 + \cdots + \tau_t) = \sum_{ij} (y_{ij} - \mu - \tau_i)^2 + \lambda(\tau_1 + \cdots + \tau_t)$$

The λ part is how we incorporate the constraint that all treatment effects sum to 0 into the W function (essentially the Lagrange multipliers method). Next, we find $\frac{\partial W}{\partial \mu}, \frac{\partial W}{\partial \tau_1}, \ldots, \frac{\partial W}{\partial \tau_t}, \frac{\partial W}{\partial \lambda}$ and set them to zero and solve. The resulting estimates are:

$$\hat{\mu} = \bar{y}_{++} \qquad \hat{\tau}_i = \bar{y}_{i+} - \bar{y}_{++} \qquad \sigma^2 = \frac{W}{n - q + c} = \frac{W}{tr - (1 + t) + 1}$$

where $\bar{y}_{++}$ is the average of all values, $\bar{y}_{i+}$ is the average of the values in group i, and with $1 + t$ non-sigma parameters and 1 constraint.

Estimators for Model 5: Consider the example where $t = 2$ and r replicates per treatment. This generalizes for any t but we take $t = 2$ for simplicity. Consider the estimator $\tilde{\mu}$ is $\bar{Y}_{++}$ and its expected value,

$$E[\bar{Y}_{++}] = E\left[\frac{\sum_{i=1}^{2}\sum_{j=1}^{r} Y_{ij}}{2r}\right] = E\left[\frac{\sum_{i=1}^{2}\sum_{j=1}^{r}(\mu + \tau_i + R_{ij})}{2r}\right] = \frac{2r\mu + \sum_{j=1}^{r}(\tau_1 + \tau_2)}{2r} = \mu$$

since $E[R_{ij}] = 0, \tau_1 + \tau_2 = 0$, and by linearity of expectation. So, $E[\tilde{\mu}] = \mu$; this is an unbiased estimator. Next, consider the variance,

$$\begin{aligned}
\text{Var}(\bar{Y}_{++}) &= \text{Var}\left(\frac{\sum_{i=1}^{2}\sum_{j=1}^{r} Y_{ij}}{2r}\right) \\
&= \frac{\sum_{i=1}^{2}\sum_{j=1}^{r} \text{Var}(Y_{ij})}{(2r)^2} \qquad \text{; independence} \\
&= \frac{\sum_{i=1}^{2}\sum_{j=1}^{r} \sigma^2}{(2r)^2} \\
&= \frac{\sigma^2}{2r}
\end{aligned}$$

This is the same result as would be given by CLT. Recall any linear combination of Normal rvs is Normal. Thus, since $\bar{Y}_{++}$ is a linear combination of Normal rvs, $\bar{Y}_{++} \sim N(\mu, \sigma^2/2r)$.

Next, recall $\hat{\tau}_1 = \bar{y}_{1+} - \bar{y}_{++}$, so the estimator is $\tilde{\tau}_1 = \bar{Y}_{1+} - \bar{Y}_{++}$. This is also a linear combination of Normal rvs so will itself be Normal. To find its distribution, we want to find its expectation and variance,

$$E[\tilde{\tau}_1] = E[\bar{Y}_{1+} + \bar{Y}_{++}] = E[\bar{Y}_{1+}] - \mu = E\left[\frac{\sum_{j=1}^{r} Y_{1j}}{r}\right] - \mu = E\left[\frac{\sum_{j=1}^{r}(\mu + \tau_1 + R_{1j})}{r}\right] - \mu = \tau_1$$

So, $\tilde{\tau}_1$ is an unbiased estimator for τ_1. For the variance, note that since the elements which make up $\bar{Y}_{1+}$ are also in $\bar{Y}_{++}$, then they are not independent rvs. Rather than using covariance,

$$\text{Var}(\tilde{\tau}_1) = \text{Var}(\bar{Y}_{1+} - \bar{Y}_{++}) = \text{Var}\left(\bar{Y}_{1+} - \frac{\bar{Y}_{1+} + \bar{Y}_{2+}}{2}\right) = \text{Var}\left(\frac{\bar{Y}_{1+}}{2} - \frac{\bar{Y}_{2+}}{2}\right) = \frac{1}{4}\text{Var}(\bar{Y}_{1+}) + \frac{1}{4}\text{Var}(\bar{Y}_{2+})$$

Taking the variance of averages, $\text{Var}(\bar{Y}_{1+}) = \text{Var}(\bar{Y}_{2+}) = \sigma^2/r$, so $\text{Var}(\tilde{\tau}_1) = \sigma^2/2r$. Now, we construct a CI for τ_1,

$$\tau_1 : \hat{\tau}_1 \pm c\sqrt{\frac{\sigma^2}{2r}}$$

where $C \sim N(0,1)$ for known σ^2 and the discrepancy for hypothesis testing on the constant τ_0 is,

$$d = \frac{\hat{\tau}_1 - \tau_0}{\sqrt{\frac{\sigma^2}{2r}}}$$

where $D \sim N(0,1)$. If σ^2 is unknown, we instead use $\hat{\sigma}^2$ for both the CI and discrepancy and for the CI, use $C \sim t(n-q+c)$ and for the discrepancy, use $D \sim t(n-q+c)$. This is interesting because the denominator for $\hat{\sigma}^2$ will also be $n-q+c$.

A CI for μ is,

$$\mu : \hat{\mu} \pm c\sqrt{\frac{\sigma^2}{2r}}$$

if σ^2 is known then $C \sim N(0,1)$ and if unknown, replace σ^2 above with $\hat{\sigma}^2$ and use $C \sim t(n-q+c)$.

Example 1: See "Example – 2 Trt – With R" document (Lecture 21).

Now suppose there are two groups and we want to find the distribution of the difference between the treatment effects of the two groups, $\theta = \tau_1 - \tau_2$. Note that this is equal to the difference in averages between the two groups as well: $\mu + \tau_1 - (\mu + \tau_2)$. The corresponding estimator is $\tilde{\theta} = \tilde{\tau}_1 - \tilde{\tau}_2$. As shown above, since $\tilde{\tau}_1, \tilde{\tau}_2$ are a linear combination of Normals, then by Gauss, $\tilde{\theta}$ is also Normal. The expectation is,

$$E[\tilde{\theta}] = E[\tilde{\tau}_1 - \tilde{\tau}_2] = E[\tilde{\tau}_1] - E[\tilde{\tau}_2] = \tau_1 - \tau_2$$

since $\tilde{\tau}_1, \tilde{\tau}_2$ are unbiased estimators. For the variance,

$$\text{Var}(\tilde{\theta}) = \text{Var}(\tilde{\tau}_1 - \tilde{\tau}_2) = \text{Var}(\bar{Y}_{1+} - \bar{Y}_{++} - (\bar{Y}_{2+} - \bar{Y}_{++})) = \text{Var}(\bar{Y}_{1+} - \bar{Y}_{2+}) = \text{Var}(\bar{Y}_{1+}) + \text{Var}(\bar{Y}_{2+}) = \frac{\sigma^2}{r} + \frac{\sigma^2}{r} = \frac{2\sigma^2}{r}$$

Now we can build a CI for θ,

$$\theta : \hat{\theta} \pm c \cdot (\text{Standard error}) \rightarrow \theta : \hat{\tau}_1 - \hat{\tau}_2 \pm c\sqrt{\frac{2\hat{\sigma}^2}{r}}$$

where we use $\hat{\sigma}$ since σ is unknown and $C \sim t(2r-q+c)$ (for the more general case with more than two groups, n instead of $2r$). If 0 is in this interval, then we could argue there does not appear to be a difference in the two groups. Alternatively, we can do a hypothesis test,

$$H_0 : \tau_1 = \tau_2 \qquad H_A : \tau_1 \neq \tau_2$$

The discrepancy is,

$$d = \frac{\text{estimate} - H_0 \text{ value}}{\text{standard error}} = \frac{\hat{\tau}_1 - \hat{\tau}_2 - \tau_0}{\sqrt{\frac{2\hat{\sigma}^2}{r}}}$$

where $\tau_0 = 0$ in this case, and the pvalue is $p = 2P(D > |d|)$ where $D \sim t(2r - q + c)$ (for the more general case, n instead of $2r$).

Example 2: See "Example – 2 Trt – With R – residuals" document (Lecture 22). We want to confirm that the residuals follow the assumptions in the model: independent of each other, are Normal, have the same variance and mean of 0. To confirm that they are Normally distributed, we can use a QQ plot (`qqnorm(model$residuals)`) and expect to see a line. To confirm independence, we can plot the residuals (`plot(model$residuals)`) and expect to not see any patterns. We can also plot the fitted values against the residuals (`plot(model$fitted.values, model$residuals)`) and expect to see the residuals be bounded above and below by some constants; do not want to see a funnel effect where as the fitted values increase, the range of the residuals also increase, or the opposite funnel effect. This can confirm that the residuals have the same variance. To check that the mean of the residuals is 0, we can plot the residuals as above and inspect visually.

Example 3: See "Example – 4 Trt – With R" document (Lecture 23). This is an application of Model 5 with 4 treatment groups. Suppose we want to determine whether there is a difference between the treatment effect of group 1 and 2; $\tilde{\theta} = \tilde{\tau}_1 - \tilde{\tau}_2$. By Gauss, this is a Normally distributed estimator. By the same reasoning as in Example 1,

$$E[\tilde{\theta}] = \tau_1 - \tau_2 \qquad \text{Var}(\tilde{\theta}) = \text{Var}(\bar{Y}_{1+}) + \text{Var}(\bar{Y}_{2+}) = \frac{2\sigma^2}{r}$$

The CI is built in the same way as in Example 1, with $C \sim t(n - q + c)$ (specifically, $n = 4r$ now since there are 4 groups),

$$\theta : \hat{\tau}_1 - \hat{\tau}_2 \pm c\frac{\sqrt{2}\hat{\sigma}}{\sqrt{r}}$$

If 0 is in this interval, then there does not appear to be any difference between the two groups.

Now suppose the 4 groups are 4 classes of a course where groups $1, 4$ are taught by instructor 1 and $2, 3$ are taught by instructor 2. How can we use HT to determine whether there is a difference between the average treatment effect of instructor 1 and instructor 2? We want,

$$\tilde{\theta} = \frac{\tilde{\tau}_1 + \tilde{\tau}_4}{2} - \frac{\tilde{\tau}_2 + \tilde{\tau}_3}{2}$$

The expectation (by linearity) and since $\tilde{\tau}_i$ are unbiased estimators for $i = 1, 2, 3, 4$ is,

$$E[\tilde{\theta}] = \frac{\tau_1 + \tau_4}{2} - \frac{\tau_2 + \tau_3}{2}$$

And the variance,

$$\begin{aligned}
\text{Var}(\tilde{\theta}) &= \text{Var}\left(\frac{\tilde{\tau}_1 + \tilde{\tau}_4}{2} - \frac{\tilde{\tau}_2 + \tilde{\tau}_3}{2}\right) \\
&= \text{Var}\left(\frac{\bar{Y}_{1+} + \bar{Y}_{4+}}{2} - \frac{\bar{Y}_{2+} + \bar{Y}_{3+}}{2}\right) \\
&= \frac{\text{Var}(\bar{Y}_{1+}) + \text{Var}(\bar{Y}_{2+}) + \text{Var}(\bar{Y}_{3+}) + \text{Var}(\bar{Y}_{4+})}{4} \\
&= \frac{4\sigma^2}{4r} \\
&= \frac{\sigma^2}{r}
\end{aligned}$$

Now for $H_0 : \theta = 0$ and $H_A : \theta \neq 0$, the discrepancy is,

$$d = \frac{\text{estimate} - H_0 \text{ value}}{\text{standard error}} = \frac{\hat{\theta}}{\frac{\hat{\sigma}}{\sqrt{r}}}$$

where $D \sim t(n - q + c)$ and pvalue is $p = 2P(D > |d|)$.

Contrast: The θ above is an example of a contrast. A contrast is a linear combination of parameters or statistics of the form,

$$a_1\tau_1 + a_2\tau_2 + \cdots + a_n\tau_n$$

where $\sum_{i=1}^{n} a_i = 0$. This form is convenient for comparing different treatments because all the variables which are not independent have components which cancel each other out to give dependent variables that can turn into a sum when computing variance.

3 ANOVA and the F-Test

ANOVA (Analysis of Variance): ANOVA is a methodology used to test hypotheses that involve two or more restrictions on the parameters. It can help answer questions such as whether all group averages are the same (eg all treatment effects are 0). Applied to Model 5 at first, recall,

$$W = \sum_{ij} r_{ij}^2 = \sum_{ij} (y_{ij} - \hat{\mu} - \hat{\tau}_i)^2 = \sum_{ij} (y_{ij} - \hat{\mu})^2 + (-r) \sum_{i} (\bar{y}_{i+} - \bar{y}_{++})^2$$

Rearranging,

$$\sum_{ij} (y_{ij} - \bar{y}_{++})^2 = r \sum_{i} (y_{i+} - \bar{y}_{++})^2 + \sum_{ij} (y_{ij} - \hat{\mu} - \hat{\tau}_i)^2$$

We will call the left hand side $\sum_{ij}(y_{ij} - \bar{y}_{++})^2$ as SS_{tot} (sum of squares total) and $r\sum_i (y_{i+} - \bar{y}_{++})^2$ as SS_{trt} and $\sum_{ij}(y_{ij} - \hat{\mu} - \hat{\tau}_i)^2$ as SS_{res} where trt and res are treatment and residual respectively. So,

$$\boxed{SS_{tot} = SS_{trt} + SS_{res}}$$

SS_{tot} represents a measure of total variability in the data and note that,

$$s^2 = \frac{SS_{tot}}{n-1} = MS_{tot}$$

where MS_{tot} is the mean squared (total). The degrees of freedom here is $n-1$. You get this by fitting Model 1 ($Y_i = \mu + R_i$ where $R_i \sim N(0, \sigma^2)$) and recall $\hat{\sigma} = s$ in Model 1.

SS_{res} is the variability left over after you fit the model (the unexplained variability). This is synonymous with,

$$\hat{\sigma}^2 = \frac{W}{n-q+c} = \frac{SS_{res}}{df_{res}} = MS_{res}$$

where the degrees of freedom is $n - q + c$. Every Model has a SS_{tot} and SS_{res}.

Lastly, SS_{trt} (does not appear in every model) is the variability that is explained by the model and comes from components such as τ in models such as Model 5. And,

$$\frac{SS_{trt}}{df_{trt}} = MS_{trt}$$

where the degrees of freedom for treatments is $t-1$ where there are t treatments. The total degrees of freedom is,

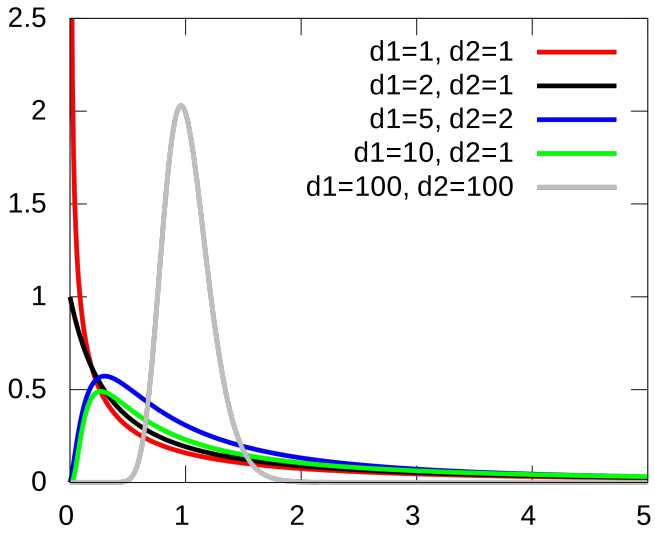
$$\boxed{df_{tot} = df_{trt} + df_{res}}$$

Ideally, we want $SS_{trt} >> SS_{res}$ (variability explained by the model is ideally much bigger than variability that is not explained by the model). To make this comparison, we compare MS_{trt} to MS_{res} using the ratio,

$$\boxed{F = \frac{MS_{trt}}{MS_{res}}}$$

and ideally want a big F which implies our model is "good".

F-distribution: Parameterized by two numbers, $F(n, d)$ where n is the degrees of freedom for the numerator and d is the degrees of freedom for the denominotor, can look an exponential or a chi squared or a Normal depending on the parameters. Cdf illustrated below,



Some useful theorems,

1. Let $X \sim \chi^2(m), Y \sim \chi^2(n)$, then $\frac{X/m}{Y/n} \sim F(m,n)$.
2. Let $X \sim F(m,n), Y = \frac{1}{X}$, then $Y \sim F(n,m)$. This can be proven using the first theorem.

An F table is essentially 3-dimensional. So instead, it is often split up by probability: $\alpha = 0.1, \alpha = 0.05, \alpha = 0.01$ tables (see document "FTableDetails"). If your probability of interest is 0.1, then use the $\alpha = 0.1$ table, etc. For example, to find $P(X > 4)$ where $X \sim F(4, 20)$, we find the critical values for $\alpha = 0.1$ (which is 3.84), $\alpha = 0.05$ (which is 5.80), and $\alpha = 0.01$ (which is 14.0) and note that since $3.84 < 4 < 5.80$, then the probability is between 0.05 and 0.1.

Next, consider the estimator,

$$\tilde{F} = \frac{\widetilde{MS}_{trt}}{\widetilde{MS}_{res}}$$

where $\widetilde{MS}_{res} = \tilde{\sigma}^2$ and we know that,

$$\frac{\tilde{\sigma}^2(df_{res})}{\sigma^2} \sim \chi^2(df_{res})$$

and similarly for $\widetilde{MS}_{trt}$,

$$\frac{\widetilde{MS}_{trt}(df_{trt})}{\sigma^2} \sim \chi^2(df_{trt})$$

Thus,

$$\frac{df_{trt}\widetilde{MS}_{trt}/\sigma^2 df_{trt}}{df_{res}\widetilde{MS}_{res}/\sigma^2 df_{res}} = \frac{\widetilde{MS}_{trt}}{\widetilde{MS}_{res}} \sim F(df_{trt}, df_{res})$$

When is $\tilde{F}$ "large"? The expectation is,

$$E[\tilde{F}] = E\left[\frac{\widetilde{MS}_{trt}}{\widetilde{MS}_{res}}\right] \approx \frac{E\widetilde{MS}_{trt}]}{E[\widetilde{MS}_{res}]} = \frac{\sigma^2 + r\frac{\sum \tau_i^2}{t-1}}{\sigma^2} = 1 + \frac{r}{\sigma^2} \cdot \frac{\sum \tau_i^2}{t-1}$$

Note all these values are positive. If $\tau_1 = \tau_2 = \cdots = \tau_t = 0$, then $E[\tilde{F}] = 1$ and if a single τ_i is non-zero, then $E[\tilde{F}] > 1$ (is "large").

F-test: Consider $H_0 : \tau_1 = \tau_2 = \cdots = \tau_t = 0$ and H_A is that at least one of the τ_i is non-zero. The discrepancy is,

$$d = \frac{MS_{trt}}{MS_{res}}$$

where $D \sim F(df_{trt}, df_{res})$. The pvalue is $P(D > d)$ and this is a one-sided test because the F distribution is always greater than 0. The conclusion for this test based on the pvalue is the same as in regular HT.

In general, F-tests are for determining whether the standard deviations from two sets of data are statistically different. They can be used to determine whether blocking reduced variability, etc.

F-tests are for testing the significance of all parameters (eg $H_0 : \tau_1 = \tau_2 = \cdots = \tau_t = 0$ whereas t-test is for testing the significance of a single parameter (eg $H_0 : \tau_1 = 0$).

Example 1: See "Example – F Test" document (Lecture 26). The R command `anova(model)` gives the degrees of freedom, sum of squares, mean squares, F value, and pvalue for the corresponding F-test. Since there are four groups in this example, $H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$ and H_A is that at least one τ_i is not 0. The discrepancy is,

$$d = \frac{MS_{trt}}{MS_{res}} = \frac{SS_{trt}/df_{trt}}{SS_{res}/df_{res}} = \frac{7315.5/3}{3928.3/20} = \frac{2438.5}{196.42} = 12.415$$

And the pvalue is, where $D \sim F(3, 20)$,

$$P(D > 12.415) = 8.281e - 05$$

This means there is significant evidence to reject H_0. All this information matches the R output.

4 Unbalanced CRD and RBD

Model 6: This is an unbalanced, completely random design (CRD), meaning that not all treatments have the same number of replicates. In this model,

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

where $R_{ij} \sim N(0, \sigma^2)$ and independent of each other and $i = 1, 2, \ldots, t$ and $j = 1, 2, \ldots, r_i$. Treatment i has r_i replicates. There is also a constraint: $\sum_{i=1}^{t} r_i \tau_i = 0$.

Least Squares: Consider the W function,

$$W = \sum_{ij} r_{ij}^2 + \lambda \sum_{i=1}^{t} r_i \tau_i$$

Taking partial derivatives wrt each non-sigma parameter gives,

$$\hat{\mu} = \bar{y}_{++} \qquad \hat{\tau}_i = \bar{y}_{i+} - \bar{y}_{++} \qquad \hat{\sigma}^2 = \frac{W}{n - q + c} = \frac{W}{r_1 + r_2 + \cdots + r_t - t - 1 + 1} = \frac{W}{r_1 + r_2 + \cdots + r_t - t}$$

Example: See "Example – 2 Trt – Unbalanced". Note that grp1 had 4 students, grp2 had 3 students. In R, use `tapply` to find group averages, `summary(lm(Y x))$sigma` to get $\hat{\sigma}$, and `anova(lm(Y x))` to get the analysis of variance table. The treatment effect for being inebriated is $\hat{\tau}_1 = -2.18$. Next, consider the difference between the treatment effects of group 1 and 2, $\theta = \tau_1 - \tau_2$,

$$\tilde{\theta} = \tilde{\tau}_1 - \tilde{\tau}_2 \implies E[\tilde{\theta}] = E[\tilde{\tau}_1] - E[\tilde{\tau}_2] = \tau_1 - \tau_2$$

since $\tilde{\tau}_1, \tilde{\tau}_2$ are unbiased estimators, implying that $\tilde{\theta}$ is an unbiased estimator for θ (and also Normal by Gauss). The variance is,

$$\text{Var}(\tilde{\theta}) = \text{Var}(\tilde{\tau}_1 - \tilde{\tau}_2) = \text{Var}(\bar{Y}_{1+} - \bar{Y}_{2+}) = \frac{\sigma^2}{r_1} + \frac{\sigma^2}{r_2} = \frac{7\sigma^2}{12} \qquad \text{(in this example)}$$

Thus, a 95% CI for θ is (where SE is standard error),

$$\text{Estimate} \pm c \cdot \text{SE} = \hat{\tau}_1 - \hat{\tau}_2 \pm c\sqrt{\frac{7\hat{\sigma}^2}{12}}$$

We can then plug in the values for $\hat{\tau}_1, \hat{\tau}_2, \hat{\sigma}^2$ given by the R output. If 0 is in the CI, then there does not appear to be a difference between the two groups.

Model 7: This is a randomized block design (RBD). Recall blocking in Model 3 with two treatments (creating pairs). This model is used over CRD when there is blocking. In this model, we can assume there are ≥ 2 treatments. The model is,

$$Y_{ij} = \mu + \tau_i + \beta_j + R_{ij}$$

where $R_{ij} \sim N(0, \sigma^2)$ and independent of each other. The terms are all the same as before (τ_i are treatment effects, μ is overall average, Y_{ij} is the response), with the additional β_j being the jth block effect (also called "blk effect"); the effect of being in block j. Note that $i = 1, 2, \ldots, t$ and $\sum_{i=1}^{t} \tau_i = 0$ and $j = 1, 2, \ldots, r$ and $\sum_{j=1}^{r} \beta_j = 0$.

Least Squares: Consider the W function,

$$W = \sum_{ij} r_{ij}^2 + \lambda_1 \sum_{i=1}^{t} \tau_1 + \lambda_2 \sum_{j=1}^{r} \beta_j$$

Solving for the partial derivatives set to 0 (partial derivatives also with respect to λ_1, λ_2), we have,

$$\hat{\mu} = \bar{y}_{++} \qquad \hat{\tau}_i = \bar{y}_{i+} - \bar{y}_{++} \qquad \hat{\beta}_j = \bar{y}_{+j} - \bar{y}_{++} \qquad \hat{\sigma}^2 = \frac{W}{n-q+c} = \frac{W}{rt-t-r-1+2}$$

Note that the $\hat{\tau}_i$ sum across rows (replicates within a group) whereas the $\hat{\beta}_j$ sum across columns (blocks). Also, there are 2 constraints and $t + r + 1$ non-sigma parameters.

Example: See "Example – Blocked". Contains R code about reading a table from a csv file and setting options. Consider,

$$\tilde{\theta} = \tilde{\tau}_1 - \tilde{\tau}_2$$

By Gauss, $\tilde{\theta}$ is Normal and $E[\tilde{\theta}] = \tau_1 - \tau_2$ since these are unbiased estimators. For the variance,

$$\text{Var}(\tilde{\theta}) = \text{Var}(\tilde{\tau}_1 - \tilde{\tau}_2) = \text{Var}(\bar{Y}_{1+} - \bar{Y}_{2+}) = \text{Var}(\bar{Y}_{1+}) - \text{Var}(\bar{Y}_{2+}) = \frac{\sigma^2}{r} + \frac{\sigma^2}{r}$$

where r is the number of blocks. CI is constructed similarly as before, Estimate $\pm\ c \cdot \text{SE}$. Using CRD compared to RBD on a data set increases the residual standard error and degrees of freedom since you are no longer accounting for the blocking and variability in blocks and they instead are in the residuals. The resulting CI are also much wider.

For HT (F test), to test H_0 that all the block effects are 0 and H_A that at least one is not zero, use the discrepancy $d = \frac{MS_{blk}}{MS_{res}}$ and the pvalue is $P(D > d)$ where $D \sim F(df_{blk}, df_{res})$.

5 Factorial Designs

A treatment is essentially a combination of the levels of different factors (explanatory variates). For example, there are two main factors/methods for combatting cancer: chemotherapy (high or low) and radition (high or low). So, a treatment might be high chemo and low radiation. In factorial design, we are interested in the factors/factorials individually as well as how they interact. Interaction means that the effects of the factors alone differ from when the factors are used together, and is one of the main reasons you would use a factorial design. For example, suppose radiation and chemo each kill 1/4 of cancer cells but when used together, kill 5/6 of cancer cells.

Model 8: This is a factorial design and can be considered a balanced CRD. In this model,

$$Y_{ijk} = \mu + \tau_{ij} + R_{ijk}$$

where $R_{ijk} \sim N(0, \sigma^2)$ and independent, where $i = 1, 2, \ldots, l_1$ (l_1 stands for level 1, the number of levels of factor 1), and $j = 1, 2, \ldots, l_2$ (number of levels of factor 2), and $k = 1, 2, \ldots, r$ (number of replicates). μ is the overall average, τ_{ij} is the treatment effect when factor 1 is at level i and factor 2 is at level j, R_{ijk} explains the distribution of values about the deterministic mark, and Y_{ijk} is the response for replicate k when factor 1 is at level i and factor 2 is at level j. There is one constraint: $\sum_{ij} \tau_{ij} = 0$.

Least squares: Let the W function be,

$$W = \sum r_{ijk}^2 + \lambda \sum_{ij} \tau_{ij}$$

and setting the non-sigma partial derivatives to 0 and solving gives,

$$\hat{\mu} = \bar{y}_{+++} \qquad \hat{\tau}_{ij} = \bar{y}_{ij+} - \bar{y}_{+++} \qquad \hat{\sigma}^2 = \frac{W}{n - q + c} = \frac{W}{rl_1l_2 - l_1l_2 - 1 + 1} = \frac{W}{rl_1l_2 - l_1l_2}$$

Model 9: This is a factorial design but RBD (randomized block design). In this model,

$$Y_{ijk} = \mu + \tau_{ij} + \beta_k + R_{ijk}$$

where $R_{ijk} \sim N(0, \sigma^2)$ and independent, where everything is the same as in Model 8 but with β_k which is the block effect, and two constraints: $\sum_{ij} \tau_{ij} = 0$ and $\sum_k \beta_k = 0$.

Least squares: Let the W function be,

$$W = \sum r_{ijk}^2 + \lambda_1 \sum_{ij} \tau_{ij} + \lambda_2 \sum_k \beta_k$$

and setting the non-sigma partial derivatives to 0 and solving gives the same $\hat{\mu}, \hat{\tau}_{ij}$ as in Model 8 and additionally,

$$\hat{\beta}_k = \bar{y}_{++k} - \bar{y}_{+++} \qquad \hat{\sigma}^2 = \frac{W}{n - q + c} = \frac{W}{rl_1l_2 - l_1l_2 - r - 1 + 2}$$

Determining interaction: Suppose there are two factors A, B, each with levels $0, 1$ and the averages of the four treatments are: $\bar{y}_{00+} = 10$ (level 0 of A and level 0 of B), $\bar{y}_{01+} = 20, \bar{y}_{10+} = 20, \bar{y}_{11+} = 30$. There are three ways of determining interaction,

1. Draw an interaction plot from the treatment averages, where the x axis is the levels for A, the y axis is numerical, and the levels for B are plotted as different kinds of lines. See minute 16 of Lec 30. Parallel lines imply that there is no interaction (levels of A move the same way regardless of which level of B we are in). Non-parallel lines imply that the level of B influences how the levels of A move. This works for any number of factors and levels.

2. Create a contrast for interaction and do a hypothesis test. This only works if there are two factors with two levels each. In the interaction plot, the two lines will be parallel iff $\Delta_1 = \bar{y}_{11+} - \bar{y}_{01+}$ equals $\Delta_2 = \bar{y}_{10+} - \bar{y}_{00+}$. Consider the contrast,

$$\tilde{\theta} = \tilde{\Delta}_1 - \tilde{\Delta}_2 = (\tilde{\tau}_{11} - \tilde{\tau}_{01}) - (\tilde{\tau}_{10} - \tilde{\tau}_{00})$$

By Gauss, $\tilde{\theta}$ is Normal. As well, since the $\tilde{\tau}_{ij}$ are unbiased estimators, $E[\tilde{\theta}] = \tau_{11} - \tau_{01} - \tau_{10} + \tau_{00}$. And the variance (since each of the $\bar{Y}_{ij}$ are independent),

$$\text{Var}(\tilde{\theta}) = \text{Var}(\bar{Y}_{11+} - \bar{Y}_{01+} - \bar{Y}_{10+} + \bar{Y}_{00+}) = \text{Var}(\bar{Y}_{11+}) + \text{Var}(\bar{Y}_{01+}) + \text{Var}(\bar{Y}_{10+}) + \text{Var}(\bar{Y}_{00+}) = 4 \cdot \frac{\sigma^2}{r}$$

Now for the hypothesis test, the null hypothesis is that there is no interaction, $H_0 : \theta = 0, H_A : \theta \neq 0$. The discrepancy is,

$$\frac{\text{estimate} - H_0 \text{ value}}{\text{standard error}} = \frac{\hat{\tau}_{11} - \hat{\tau}_{01} - \hat{\tau}_{10} + \hat{\tau}_{00}}{\sqrt{\frac{4\hat{\sigma}^2}{r}}}$$

where $D \sim t(n - q + c) = t(rl_1l_2 - l_1l_2 - 1 + 1)$. In this specific case, $l_1l_2 = 4$. Then calculate the pvalue $2P(D > d)$. If there is no evidence to reject H_0, then it looks like there is no interaction.

3. ANOVA then F-test. See "Example – Factorial" for R code. We are interested in the difference in the heart rate after stepping (HR minus RestHR). In this code, `Freq * Height` is the interaction (*int*) term and note that,

$$df_{freq} + df_{height} + df_{freq:height} = df_{trt} \qquad SS_{freq} + SS_{height} + SS_{freq:height} = SS_{trt}$$

In this particular example, $df_{trt} = 3$ and $SS_{trt} = 1143.0$. The hypothesis test for interaction is then, H_0 : no interactions, H_A : there is interaction. From the R output, the discrepancy is,

$$d = \frac{MS_{int}}{MS_{res}} = \frac{45}{102.37} = 0.4396$$

and the pvalue is 0.51677 where $D \sim F(df_{int}, df_{res})$, which is $F(1, 16)$. Based on this pvalue, there is no evidence to reject H_0 so it appears there is no interaction.

6 Sampling

Probability sampling: Let U be the frame (aka study population), $U = \{1, 2, \ldots, N\}$, let s be our sample with size $n \leq N$ and $s \subseteq U$. Let the sampling protocol refer to the probability of selecting any particular sample. Let π_{ij} be the inclusion probability for unit i and j (note π_{ii} is equivalent to π_i).

Simple Random Sampling Without Replacement (SRSWOR): Consider a frame $\{1, 2, 3, 4\}$ and we select a sample of size $n = 2$. The possible samples without replacement are $s_1 = \{1, 2\}, s_2 = \{1, 3\}, s_3 = \{1, 4\}, s_4 = \{2, 3\}, s_5 = \{2, 4\}, s_6 = \{3, 4\}$. The probability we select s_1 is $1/6$ and the probability that unit 1 (ie π_1) is in the sample is $1/2$. In general, for a frame $\{1, 2, \ldots, N\}$ and we select SRSWOR a sample of size n. Then the probability we select a particular sample (eg sample 1) is $P(s_1) = 1/\binom{N}{n}$. And the probability that unit 1 is selected is $\pi_1 = \binom{N-1}{n-1}/\binom{N}{n} = \frac{n}{N}$.

Model 1 (sampling perspective): Recall Model 1 is,

$$Y_j = \mu + R_j$$

where $R_j \sim N(0, \sigma^2)$ and are independent. The parameters to this model are μ, σ^2 (assumed to be unknown, so need estimates). The LS estimates are $\hat{\mu} = \bar{y}_+, \hat{\sigma}^2 = s^2$ and the estimators are $\tilde{\mu} = \bar{Y}_+, \tilde{\sigma}^2 = S^2$. As well, $\tilde{\mu} \sim N(\mu, \sigma^2/n)$ by Gauss and CLT. The CI for μ is: $\hat{\mu} \pm c\hat{\sigma}/\sqrt{n}$ where $C \sim t(n-1)$ and c is such that $P(-c < C < c) = p$ for a $100p\%$ CI.

From the point of view of a sample using SRSWOR, the parameters are,

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

The best estimates, which make intuitive sense, are,

$$\hat{\mu} = \frac{1}{n}\sum_{i\in s} y_i \qquad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \hat{\mu})^2$$

From a sampling perspective, the estimators start with the basic concept that y_i is not a realization of some random variable Y_i. Instead, y_i is a constant. What is random is whether y_i is selected for the sample. Let I_i be an indicator variable which is 0 if y_i is not in the sample and 1 if it is in the sample. Then,

$$\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{N} I_i y_i \qquad \tilde{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{N} I_i(y_i - \hat{\mu})^2$$

And, $\tilde{\mu} \sim N(\mu, (1 - n/N)\sigma^2/n)$ where n/N is the sampling fraction, $1 - n/N$ is called the finite population correction. The proof is:

$$P(I_i = 1) = \pi_i = \frac{n}{N} \qquad E[I_i] = 0 \cdot P(I_i = 0) + 1 \cdot P(I_i = 1) = \frac{n}{N}$$

And for the variance,

$$E[I_i^2] = P(I_i = 1) = \frac{n}{N} \qquad \text{Var}(I_i) = E[I_i^2] - E[I_i]^2 = \frac{n}{N}\Big(1 - \frac{n}{N}\Big)$$

And,

$$E[I_i I_j] = P(I_i = 1, I_j = 1) = \pi_{ij} = \frac{\binom{N-1}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

Next, by Schlotzsky's theorem, $\tilde{\mu}$ is Normal and since y_i are constant,

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

Thus, $\tilde{\mu}$ is an unbiased estimator. For variance, since I_i is not independent of I_j, we need to consider covariances,

$$\text{Var}(\tilde{\mu}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{N} I_i y_i\right) = \frac{1}{n^2}\sum_{i=1}^{N} y_i^2 \text{Var}(I_i) + \frac{1}{n^2}\sum_{ij} y_i y_j \text{Cov}(I_i, I_j) = \cdots = \frac{(1-n/N)\sigma^2}{n}$$

See course notes for more detail.

With the distribution of $\tilde{\mu}$ in mind, the CI from the SRSWOR perspective is then,

$$\hat{\mu} \pm c\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$$

where $C \sim N(0,1)$. The reason $N(0,1)$ is used rather than $t(n-1)$ is because we are not using regression/MLE to estimate the parameters, we are using sampling. When using sampling, an extension of the CLT lets us assume everything is normal.

The finite population correction causes the interval to be narrower than the one from Model 1 which does not have this term. As well, if the entire population is in the sample, then the width is 0, whereas for Model 1, the width would be non-zero.

Sample size calculations: Recall the CI for Model 1 has a margin of error $E = c\sigma/\sqrt{n}$. Setting $E = c\sigma/\sqrt{n}$ then solving for $n = c^2\sigma^2/E^2$ gives the sample size required to have a CI width. The process is to take a small sample, use it to estimate σ, then solve for n using the formula and perform a large study with n units. For SRSWOR, the error in a CI for the mean, is,

$$E = \frac{c\hat{\sigma}}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$$

and solving for n (possibly rounding up),

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

As $n \to \infty$, the result is the same as in Model 1. This makes sense because in Model 1, the population size is assumed to be infinite.

Example: See "Example – SRSWOR".

SRS perspective of a proportion: Recall Model 4 (π for p),

$$\frac{Y_i}{n} \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

and the CI for π is: $\hat{\pi} \pm c\sqrt{\hat{\pi}(1-\hat{\pi})/n}$. From the SRS perspective, the study population parameter is,

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

where y_i is 0 or 1 (since the Y_i can from a binomial model of true/false). The statistic (estimate) is,

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

And since $y_i^2 = y_i$ (y_i is either 0 or 1) and $\bar{y} = \hat{\pi}$,

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

Assuming $n \to \infty$, then $\hat{\sigma}^2 = \hat{\pi}(1-\hat{\pi})$. Using the same I_i indicator rvs as before, the estimators are,

$$\tilde{\pi} = \frac{1}{n}\sum_{i=1}^{N} y_i I_i$$

This is the same as $\tilde{\mu}$ for Model 1. It can be shown that $\tilde{\pi}$ is an unbiased estimator for π and that $\tilde{\pi}$ is Normal. The variance,

$$\text{Var}(\tilde{\pi}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{N} y_i I_i\right) = \cdots = \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)$$

and we replace σ^2 by $\hat{\sigma}^2 = \hat{\pi}(1-\hat{\pi})$. The CI for π from the SRS perspective is then (with a finite population correction),

$$\pi : \hat{\pi} \pm c\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}\left(1 - \frac{n}{N}\right)}$$

where again $C \sim N(0,1)$. As well, similar as with the SRS view of Model 1, this SRS CI is narrower than the Model 4.

Sample size calculation: The formula is still,

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

but $\hat{\sigma}^2 = \hat{\pi}(1-\hat{\pi})$ and to get the worst case scenario, we calculate using $\hat{\pi} = \frac{1}{2}$.

Example 2: See "Example – SRSWOR" Example 2. The population size is unknown (so assume it is infinite), the CI is,

$$\hat{\pi} \pm c\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

where $\hat{\pi} = 0.42, n = 1053$, and $c = 1.96$ (since $C \sim N(0,1)$, giving $[0.39, 0.45]$. Since 0.45 is in the interval, then we can say there is no significant difference between last year and this year's results.

Example 3: See "Example – SRSWOR" Example 3. This is a sample size calculation with $N = 150000, E = 1$, and $c = 1.96$ (since $C \sim N(0,1)$). Since no proportion is given, we assume $\hat{\pi} = 1/2$. Then use the formula above and round up to get 9027.

Example 4: See "Example – SRSWOR" Example 4. Using a $N(0,1)$ at 90% confidence, then $c = 1.645$. First we create a CI for μ then convert that to the total amount of money. $\hat{\mu} = 143.95$ and $\hat{\sigma}^2 = 81.09$ and $N = 200$ and $n = 15$. Building the CI based on model 1, the final CI is $[140.27, 147.63]$. To convert this CI to a CI for the total, we multiply by the total number of accounts (200) to get $[28054, 29526]$. Since 25000 is out of this interval, then we can conclude the company is not being compliant.

7 Regression Sampling

We want $\mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i$ (population average) and we use $\hat{\mu}_y = \frac{1}{n}\sum_{i\in S} y_i = \bar{y}$ (sample average) to calculate it. Suppose Y_i is linearly related to a continuous explanatory variate x_i. Then, $\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$ with sample mean $\hat{\mu}_x = \frac{1}{n}\sum_{i\in S} x_i = \bar{x}$. Suppose the relationship is (where $R_i \sim N(0, \sigma^2)$),

$$Y_i = \alpha + \beta(x_i - \bar{x}) + R_i$$

Overall, use regression sampling when there is a linear relationship between Y and x, SRS otherwise.

Least squares: Let the W function be,

$$W = \sum_i r_i^2 = \sum_i (y_i - \alpha - \beta(x_i - \bar{x}))^2$$

Taking partial derivatives wrt $\alpha, \beta, \mu_x, \mu_y, \ldots$ gives

$$\hat{\alpha} = \bar{y} \qquad \hat{\beta} = \frac{\sum_i y_i(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} = \frac{s_{xy}}{s_x^2}$$

where $S_{xy} = \sum_i y_i(x_i - \bar{x}) = \sum_i (y_i - \bar{y})(x_i - \bar{x})$ and $s_{xy} = \frac{S_{xy}}{n-1}$ and $S_{xx} = \sum_i (x_i - \bar{x})^2$ and $s_x^2 = \frac{S_{xx}}{n-1}$.

Regression line: We had $Y_i = \alpha + \beta(x_i - \bar{x}) + R_i$ and obtained LS estimates of α, β, and,

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}(x_i - \bar{x})$$

this is the regression line and $\hat{y}_i$ is the prediction. If $x_i = \bar{x}$, then $\hat{y}_i = \hat{\alpha} = \bar{y}$. If $x_i = \mu_x$, then $\hat{y}_i = \hat{\alpha} + \hat{\beta}(\mu_x - \bar{x})$; this particular prediction $\hat{y}_i$ is called $\hat{\mu}_{reg}$ (regression prediction),

$$\hat{\mu}_{reg} = \hat{\alpha} + \hat{\beta}(\mu_x - \bar{x})$$

Estimators: The estimators corresponding to parameters $\alpha, \beta, \mu_x, \mu_y$ are all unbiased. What about the estimator for $\hat{\mu}_{reg}$? To begin with, note that,

$$\begin{aligned}
\hat{\mu}_{reg} &= \hat{\alpha} + \hat{\beta}(\mu_x - \bar{x}) \\
&= \hat{\alpha} - \hat{\beta}(\bar{x} - \mu_x) \\
&= \bar{y} - \hat{\beta}(\bar{x} - \mu_x) \\
&= \frac{1}{n}\sum_{i\in S} y_i - \hat{\beta}\left(\frac{1}{n}\sum_{i\in S} x_i - \frac{n\mu_x}{n}\right) \\
&= \frac{1}{n}\sum_{i\in S} (y_i - \hat{\beta}(x_i - \mu_x)) \\
&= \frac{1}{n}\sum_{i\in S} r_i \qquad \text{; denote with } r_i \text{ (residual)}
\end{aligned}$$

The estimator is then,

$$\tilde{\mu}_{reg} = \frac{1}{n}\sum_{i=1}^{N} I_i r_i$$

where I_i is an indicator variable which is 0 if y_i is not in the sample and 1 if it is. This estimator is Normally distributed and,

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

Although $E[\tilde{\beta}(\tilde{\mu}_x - \mu_x)]$ is small, it is non-zero, so this $\tilde{\mu}_{reg}$ is a biased estimator for μ_y. For the variance,

$$\text{Var}(\tilde{\mu}_{reg}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{N} I_i r_i\right) = \left(1 - \frac{n}{N}\right)\frac{\sigma_r^2}{n}$$

since it is the same form as SRS (see above). We estimate $\hat{\sigma}_r^2$ by,

$$\hat{\sigma}_r^2 = \frac{1}{n-1}\sum_{i \in S}(r_i - \bar{r})^2 = \cdots = \frac{W}{n-1}$$

The CI is then,

$$\text{estimate} \pm c \cdot SE = \hat{\mu}_{reg} \pm c\frac{\hat{\sigma}_r}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}$$

where $C \sim N(0,1)$.

Example: See "Example – Regression – Intro". First we try SRS then find that $\mu_y - \hat{\mu}_y = 1.6$. Using regression sampling, $\hat{\alpha}, \hat{\beta}$ can be taken from `summary(lm(sample_heights sample_weights))`. However, the residual standard error $\hat{\sigma}$ given by R is divided by $n-2$ but we want to divide by $n-1$. So, $\hat{\sigma}_r^2 = \hat{\sigma}^2(n-2)/(n-1)$ As well, note $\hat{\sigma}_{reg}$ is used interchangeably with $\hat{\sigma}_r$. Regression intervals are narrower than SRS intervals, but due to the bias in the regression interval, sometimes may not include the true population mean.

Example 2: See "Example– Regression 1". Note that $N = 94, n = 9$, and $\mu_x = 5.1$ (population mean). And,

$$\hat{\beta} = \frac{s_{xy}}{s_x^2} = \frac{24.75}{8.11} \approx 3 \qquad \hat{\alpha} = \bar{y} = \hat{\mu}_y = 74 \qquad \hat{\mu}_{reg} = \hat{\alpha} + \hat{\beta}(\mu_x - \bar{x}) = 74 + 3(5.1 - 4.89) = 74.63$$

The SRS 95% CI is,

$$\hat{\mu}_y \pm \frac{c\hat{\sigma}_y}{\sqrt{n}}\sqrt{1 - \frac{n}{N}} = 74 \pm 1.96\sqrt{\frac{89.25}{9}}\sqrt{1 - \frac{9}{94}} \implies [68.1, 79.9]$$

The width is roughly 11. The regression sampling 95% CI is,

$$\hat{\mu}_{reg} \pm \frac{c\hat{\sigma}_r}{\sqrt{n}}\sqrt{1 - \frac{n}{N}} = 74.63 \pm 1.96\sqrt{\frac{15.7}{9}}\sqrt{1 - \frac{9}{94}} \implies [72.2, 77.1]$$

The width is much narrower than for the SRS interval. Overall, note that all CI used for sampling (both SRS and regression) use $C \sim N(0,1)$.

8 Ratio Estimation

Ratio estimation of an average: SRS and regression sampling are both useful for building CIs for an average. Similar to in regression sampling, suppose Y_i is linearly related to a continuous explanatory variate x_i. We would like to do ratio estimation of an average. Consider the model,

$$Y_i = \beta x_i + R_i, \text{ where } R_i \sim N(0, x_i\sigma^2) \text{ and are independent}$$

Graphically, the $Y_i = \beta x_i$ part can be drawn as a line in 2D going through the origin and the variance grows with x_i, creating a funnel effect. So, this model is especially useful when the funnel effect is present in the underlying model. To fix the funnel effect in this model, we divide by $\sqrt{x_i}$ to give,

$$\frac{Y_i}{\sqrt{x_i}} = \beta\sqrt{x_i} + \frac{R_i}{\sqrt{x_i}}, \text{ where } \frac{R_i}{\sqrt{x_i}} \text{ is } N(0, \sigma^2) \text{ and are independent}$$

By Gauss, $\frac{R_i}{\sqrt{x_i}}$ is Normal and it is easy to see that the expectation is 0, since x_i is a constant. For the variance,

$$\text{Var}\left(\frac{R_i}{\sqrt{x_i}}\right) = \frac{\text{Var}(R_i)}{x_i} = \sigma^2$$

Let $Y_i' = \frac{Y_i}{\sqrt{x_i}}, x_i' = \sqrt{x_i}, R_i' = \frac{R_i}{\sqrt{x_i}}$. Our model is then,

$$Y_i' = \beta x_i' + R_i', \text{ where } R_i' \sim N(0, \sigma^2) \text{ and are independent}$$

Least squares: Using LS to estimate our parameters, we find that,

$$\hat{\beta} = \frac{\bar{y}}{\bar{x}} \qquad \hat{\sigma}^2_{ratio} = \frac{W}{n-1}$$

where *ratio* stands for ratio estimation.

Prediction (line of best fit): The prediction is,

$$\hat{y}_i = \hat{\beta}x_i' = \frac{\bar{y}}{\bar{x}}x_i'$$

Thus, if $x_i' = \bar{x}$, then the prediction is the sample average $\bar{y}$. And if $x_i' = \mu_x$, then we have $\frac{\bar{y}}{\bar{x}}\mu_x$. We believe this will be μ_y and denote it $\hat{\mu}_{ratio}$ (this is called the ratio estimator).

Next, we want to build a CI for μ_y, using the same method as for regression sampling,

$$\mu_y : \hat{\mu}_{ratio} \pm \frac{c \cdot \hat{\sigma}_{ratio}}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}$$

where $C \sim N(0, 1)$.

Example: See "Example – Ratio – Intro" for how to use R to compute $\hat{\mu}_{ratio}$ of 66.4. Use -1 to tell R to remove the intercept term, regardless of what the actual intercept is. The corresponding CI for μ_y uses this value as well as $c = 1.96, \hat{\sigma}_{ratio} = 0.1572, n = 5, N = 15$ to get $[66.3, 66.5]$. The width of this CI is narrower than if we used SRS. As well, the ratio estimator is biased since the interval does not contain 65 which is the actual μ_y, similar to a regression CI.

Requirements: Requirements for this model are: highly correlated Y_i and x_i (linear relationship and x_i are continuous, same requirement as for regression sampling), intercept of zero. Suitable for data with a funnel effect.

In summary, for SRS, regression, and ratio, if we care about the mean,

1. SRS: estimate is $\hat{\mu}_y$, CI is $\hat{\mu}_y \pm \frac{c\cdot\hat{\sigma}_y}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$
2. Regression sampling: estimate is $\hat{\mu}_{reg} = \bar{y} + \hat{\beta}(\mu_x - \bar{x})$, CI is $\hat{\mu}_{reg} \pm \frac{c\cdot\hat{\sigma}_{reg}}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$
3. Ratio estimate: estimate is $\hat{\mu}_{ratio} = \frac{\bar{y}}{\bar{x}}\mu_x$, CI is $\hat{\mu}_{ratio} \pm \frac{c\cdot\hat{\sigma}_{ratio}}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$

Ratio estimations: Ratio estimation of an average (REA) and ratio estimation (RE) are both used to estimate an average. The difference is that in REA, we take an SRS first and then find the average of the entire sample (i.e. both males and females). In RE we take an SRS but then find the average of a portion of the dataset (i.e. just the males). For example, suppose we take an SRS of a class and get the data,

Gender	M	F	M	F	M	F
Grade	70	70	85	85	90	90

Suppose we want to determine how well the males did on average, $\bar{x}_m = \frac{70+85+90}{3}$. Generalizing, let y_i be a grade and z_i be 1 if the ith person is male, 0 otherwise. Our estimate is,

$$\hat{\theta} = \frac{\sum_{i\in S} y_i z_i}{\sum_{i\in S} z_i}$$

where the parameter is,

$$\theta = \frac{\frac{1}{N}\sum_{i=1}^{N} y_i z_i}{\frac{1}{N}\sum_{i=1}^{N} z_i} = \frac{\mu}{\pi}$$

where π is the proportion of males and μ is the average grade of the males. So, $\hat{\theta} = \frac{\hat{\mu}}{\hat{\pi}}$. The estimator is,

$$\tilde{\theta} = \frac{\frac{1}{n}\sum_{i=1}^{N} I_i y_i z_i}{\frac{1}{n}\sum_{i=1}^{N} I_i z_i} = \frac{\tilde{\mu}}{\tilde{\pi}}$$

where I_i is the indicator rv which is 1 if the ith person is in the sample and 0 otherwise. Computing a ratio of two rvs is difficult, so using the first order Taylor's approximation for multivariate functions,

$$\frac{\tilde{\mu}}{\tilde{\pi}} \approx \frac{\mu}{\pi} + \frac{1}{\pi}(\tilde{\mu}-\mu) - \frac{\mu}{\pi^2}(\tilde{\pi}-\pi)$$

where $\tilde{\mu}, \tilde{\pi}$ are both obtained by SRS. It can be shown that this estimator $\tilde{\theta}$ is approximately Normal. For the expectation,

$$E\left[\frac{\tilde{\mu}}{\tilde{\pi}}\right] \approx E\left[\frac{\mu}{\pi} + \frac{1}{\pi}(\tilde{\mu}-\mu) - \frac{\mu}{\pi^2}(\tilde{\pi}-\pi)\right] = \frac{\mu}{\pi} + \frac{1}{\pi}(E[\tilde{\mu}]-\mu) - \frac{\mu}{\pi^2}(E[\tilde{\pi}]-\pi) = \frac{\mu}{\pi}$$

where by SRS (unbiased), we know that $E[\tilde{\mu}] = \mu, E[\tilde{\pi}] = \pi$. Thus, this estimator is approximately unbiased. For the variance,

$$\mathrm{Var}\left(\frac{\tilde{\mu}}{\tilde{\pi}}\right) \approx \mathrm{Var}\left(\frac{\mu}{\pi} + \frac{1}{\pi}(\tilde{\mu}-\mu) - \frac{\mu}{\pi^2}(\tilde{\pi}-\pi)\right) = \frac{1}{\pi^2}\mathrm{Var}\left(\tilde{\mu} - \mu - \frac{\mu\tilde{\pi}}{\pi} + \mu\right) = \frac{1}{\pi^2}\mathrm{Var}\left(\tilde{\mu} - \frac{\mu\tilde{\pi}}{\pi}\right)$$

This is an average, so just like with SRS, the above simplifies to,

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

The CI is then,

$$\text{estimate} \pm c\cdot SE \implies \hat{\theta} \pm c\frac{1}{\hat{\pi}}\sqrt{1-\frac{n}{N}}\frac{\hat{\sigma}_{ratio}}{\sqrt{n}}$$

where $C \sim N(0,1)$ and,

$$\hat{\sigma}^2_{ratio} = \frac{1}{n-1}\sum_{i \in S}(y_i - \hat{\theta} z_i)^2$$

This CI tends to be very large because there is a lot of variability since you did not perform a SRS on one group, but instead are dividing up a larger group (eg a group containing both males and females but we only care about the males).

Example: See "Example – Ratio Example 1". We have $\hat{\pi} = \frac{42}{80}, \hat{\theta} = \frac{\hat{\mu}}{\hat{\pi}} = 67$. The CI is,

$$\hat{\theta} \pm \frac{c}{\hat{\pi}} \frac{\hat{\sigma}_{ratio}}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

where $c = 1.96, \hat{\sigma}_{ratio} = \sqrt{5.42}, n = 80, N = 89422$.

9 Stratified Sampling

A stratified sample is suitable if you are interested in the population as a whole and at the same time, interested in some subpopulation of the population. We divide the study population into subpopulations (called strata) then sample independently using SRS from each stratum. Then combine the estimates of each stratum average to get an estimate of the population average.

Suppose we have a population frame U and we divide it into subframes $U_1, U_2, \ldots, U_H$ where:

1. $U_1 \cup U_2 \cup \cdots \cup U_H = U$
2. For any $i \neq j$, $U_i \cap U_j = \emptyset$. No unit is in both subframes.
3. Let the size of each subframe be $|U_i| = N_i$ for $i = 1, 2, \ldots, H$ and let $|U| = N$.
4. $\sum_{i=1}^{H} N_i = N$

Stratified sampling for an average: The parameter of interest is the population average,

$$\mu = \frac{N_1\mu_1 + N_2\mu_2 + \cdots + N_H\mu_H}{N} = w_1\mu_1 + w_2\mu_2 + \cdots + w_H\mu_h$$

where μ_i is the stratum average and with weights $w_i = \frac{N_i}{N}$. The estimate is,

$$\hat{\mu} = \sum_{i=1}^{H} w_i\hat{\mu}_i$$

where $\hat{\mu}_i$ is obtained by SRS. The estimator is,

$$\tilde{\mu} = \sum_{i=1}^{H} w_i\tilde{\mu}_i$$

where $\tilde{\mu}_i$ is the Normally distributed rv from SRS. Thus, $\tilde{\mu}$ is Normally distributed as well. As well, since SRS is unbiased, the expectation is,

$$E[\tilde{\mu}] = \sum_{i=1}^{H} w_i E[\tilde{\mu}_i] = \sum_{i=1}^{H} w_i\mu_i = \mu$$

So, the stratified estimator $\tilde{\mu}$ is also unbiased. Since μ_i, μ_j are independent for $i \neq j$ (because no unit is in two stratas), then the variance is,

$$\mathrm{Var}(\tilde{\mu}) = \mathrm{Var}\left(\sum_{i=1}^{H} w_i\tilde{\mu}_i\right) = \sum_{i=1}^{H} w_i^2\mathrm{Var}(\tilde{\mu}_i) = \sum_{i=1}^{H} w_i^2\frac{\sigma_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)$$

Thus, a CI for μ is,

$$\mu : \hat{\mu} \pm c\sqrt{\sum_{i=1}^{H} w_i^2\frac{\sigma_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)}$$

where $C \sim N(0, 1)$.

Stratified sampling for a proportion: For example, rather than wanting to know the average grade of students in a subpopulation, we want to know the proportion which enjoy a certain course. The population proportion parameter is,

$$\pi = \sum_{i=1}^{H} w_i\pi_i$$

The estimate is,

$$\hat{\pi} = \sum_{i=1}^{H} w_i \hat{\pi}_i$$

where $\hat{\pi}_i$ is estimated using SRS. The estimator is,

$$\tilde{\pi} = \sum_{i=1}^{H} w_i \tilde{\pi}_i$$

Using a similar approach as above for stratified sampling for an average, the CI for π is,

$$\pi : \hat{\pi} \pm c\sqrt{\sum_{i=1}^{H} w_i^2 \frac{\sigma_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)}$$

where $C \sim N(0,1)$. As well, recall that $\sigma_i^2 = \pi_i(1-\pi_i)$.

Allocation: Suppose you have a sample of 100 units and four strata. How should you spend those 100 units (eg three quarters go to one strata and the remaining quarter is splilt between the last three strata)? We use allocation to make this decision. There are two types of allocation,

1. Proportional allocation: We sample based on the size of the strata. The bigger the strata size, the bigger the sample should be. Using the weights $w_i = \frac{N_i}{N}$ above, the size for strata i is,

   $$n_i = w_i n$$

   where n is the number of units available. For example, consider four provinces of Canada with populations (in millions): ON (10), QC (5), BC (3), AB (2) for a total of 20 million. Thus, if there are $n = 100$ units, then ON should get $w_{ON} n = (1/2)100 = 50$.

2. Neyman allocation (aka optimal allocation): We select the sample size that minimize stratified variance. Recall the stratified variance is,

   $$\text{Var}(\tilde{\mu}) = \sum_{i=1}^{H} w_i^2 \frac{\sigma_i^2}{n_i}\left(1 - \frac{n_i}{N}\right)$$

   This is done subject to the constraint $n = n_1 + n_2 + \cdots + n_H$ where n_i is the number of units which go to strata i. This is a Lagrange multiplication problem. So, we minimize,

   $$W(\tilde{\mu}) = \sum_{i=1}^{H} w_i^2 \frac{\sigma_i^2}{n_i}\left(1 - \frac{n_i}{N}\right) + \lambda(n - n_1 - n_2 - \cdots - n_H)$$

   Finding $\frac{\partial W}{\partial \lambda}$, $\frac{\partial W}{\partial n_i}$ and set to 0 gives the allocation,

   $$n_i = \frac{n\sigma_i w_i}{\sum_{j=1}^{H} \sigma_j w_j}$$

   Note $n_i \propto \sigma_i$ (proportional to; since a larger sample size reduces variance). As well, note that $n_i \propto w_i$ (larger strata means more units allocated to it). Moreover, if $\sigma = \sigma_1 = \sigma_2 = \cdots = \sigma_H$, then $n_i = \frac{n w_i}{\sum_{i=1}^{H} w_i} = n w_i$, which is the same as in proportional allocation.

   Typically, we take a small sample size to estimate the σ_i then use those estimates to determine how to allocate the larger sample size to the actual strata of interest.

Example 1: See "Example – Stratified 11". The μ, σ should actually be $\hat{\mu}, \hat{\sigma}$ (coming from a SRS sample). The weights W come from dividing the strata size by the total size. Since the samples come from SRS, using techniques from the SRS section, a 95% CI for the mean tuition in math is,

$$\mu_m : \hat{\mu}_m \pm c\frac{\hat{\sigma}_m}{\sqrt{n_m}}\sqrt{1-\frac{n_m}{N_m}} \implies [4298, 4702]$$

where $\hat{\mu}_m = 4500, c = 1.96, \hat{\sigma}_m = 400, n_m = 15, N_m = 6600$. Next, we want a 95% CI for the mean tuition at UW. Firstly,

$$\hat{\mu} = w_m\hat{\mu}_m + w_a\hat{\mu}_a + \cdots + w_{evs}\hat{\mu}_{evs} = 4435$$

Next, the estimate of the variance is 1024.8 using the equation,

$$\text{Var}(\tilde{\mu}) = \sum_{i=1}^{H} w_i^2\frac{\hat{\sigma}_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)$$

Lastly, using these values, a 95% CI for the mean tuition at UW is,

$$\mu : \hat{\mu} \pm c\sqrt{\sum_{i=1}^{H} w_i^2\frac{\sigma_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)} \implies [4372, 4498]$$

Example 2: See "Example – Stratified 21". Again, μ, σ should be $\hat{\mu}, \hat{\sigma}$ and each sample is obtained from the corresponding stratum using SRS. We use the weights W to compute a proportional allocation. For example, the number of students to allocate to math is $0.22n = 0.22(120) = 26.4$ which we round to 26. For science, it is $0.18(120) = 21.6$ which we round to 22. For optimal allocation,

$$n_i = \frac{n\hat{\sigma}_i w_i}{\sum_{j=1}^{6}\hat{\sigma}_j w_j}$$

As a result, we allocate 30 to math, 29 to arts, 17 to science, 27 to engineering, 8 to ahs, and (solving for the last one), $120 - 30 - 29 - 17 - \cdots = 10$ for evs.

Poststratification: Up until now, we have performed SRS in each stratum to get samples, which are then combined to compute population values. In poststratification, we perform SRS on the population then stratify afterwards. Mathematically, the sample sizes end up being random which affects the derivations but the results are still the same,

$$\hat{\mu}_{post} = w_1\mu_1 + \cdots + w_H\mu_H$$

with estimated variance,

$$\text{Var}(\tilde{\mu}_{post}) = \sum_{i=1}^{H} w_i^2\frac{\hat{\sigma}_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)$$

and the CI is again,

$$\mu_{post} : \hat{\mu}_{post} \pm c\sqrt{\sum_{i=1}^{H} w_i^2\frac{\hat{\sigma}_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)}$$

where $C \sim N(0, 1)$.

It is not typically a choice between regular stratification and post-stratification and you would use each in different circumstances.

Poststratification is only done if you do not know the values for the attributes you are basing your strata on until you take your sample (eg stratify based on income level, but a person's income is unknown until they

complete the survey). On the other hand, in regular stratified sampling you already know (eg stratify by province and can plan to collect a different number of samples from each province).

In regular stratification, we can choose $n_1, n_2, \ldots, n_H$ whereas in poststratification, you are stuck with whatever you observed in your sample. If you compare the variances of the two estimators, they are similar. But you can choose $n1, \ldots, nH$ in such a way that you obtain a smaller variance only with regular stratification.

Non-response: Non-response means that someone did not respond to our survey. Almost all surveys with human respondents have non-response. Non-response causes bias and is a form of error that can skew our results. The response rate (proportion of people who respond) is hard to define. To correct non-response, we do two-phase sampling. The first phase is a typical SRS with sample size n and the second phase involves subsampling the non-responders from the first phase. This creates a stratified design with responders and non-responders as the strata. The population estimate is,

$$\hat{\mu} = \frac{n_R}{n}\hat{\mu}_R + \frac{n_M}{n}\hat{\mu}_M$$

where n_R is the number of responders, n_M is the number of non-responders. For the proportion,

$$\hat{\pi} = \frac{n_R}{n}\hat{\pi}_R + \frac{n_M}{n}\hat{\pi}_M$$

Stratified sampling with cost: Suppose $C = c_0 + \sum_{i=1}^{H} c_i n_i$ where C is the total cost, c_0 is the fixed cost, and c_i is the cost per response in stratum i. The optimal allocation to minimize the stratified variance is,

$$n_i = \frac{n w_i \sigma_i / \sqrt{c_i}}{\sum_{j=1}^{H} w_j \sigma_j / \sqrt{c_j}}$$

See A9 Q2 for more details and a derivation.

10 Other

Summary of some important concepts,

1. Suppose θ is the parameter of interest and the corresponding estimator is $\tilde{\theta}$. In every CI, we want the standard error, which is the square root of $\text{Var}(\tilde{\theta})$.
2. In general, the degrees of freedoms (df) means the number of values in the model which can vary independently. This is typically the sample size minus the number of parameters plus the number of constraints. Df is a little different for ANOVA. Consider the ANOVA table of a balanced CRD model (although this easily generalizes to other models): for treatment, the sum of squares is calculated across the t different treatment groups, so we have $t-1$ df. For residual, the sum of squares is callculated within each treatment group (containing r replicates), so we have $(r-1)t$ df; t groups each with $r-1$ df. This results in the total df being the sum of the two above, $rt-1$. This intuitively makes sense since it is the total sum of squares after we combine all groups together.

Other things,

1. The parameter to the Exponential distribution is the mean rather than the reciprocal of the mean (same as in Stat 333 notes)
2. If $X \sim \text{EXP}(\theta)$, then $2X\theta \sim \chi^2(2)$
3. For any integer n, $\Gamma(n, 1/2)$ is $\chi^2(2n)$
4. $\chi^2(n)$ is the sum of n Z^2 rvs where $Z \sim N(0,1)$
5. The coefficient of variation is $\frac{100\sigma}{\mu}$ where σ, μ are pop std dev, mean
6. In this course, if X, Y are independent rvs, the notation is $X \perp Y$