Key Definitions and Properties

Lee Kennedy-Shaffer

1. Probability Review

1.1. Expectations

For any random variables (RVs) X and Y, constants a and b, and events $C_1, \dots, C_n$ that partition the sample space, *expectation/expected value* is defined as:

- $E[X] = P(X = 1)$ for a binary RV X
- $E[X] = \sum_{x \in supp(X)} xP(X = x)$ for a discrete RV X
- $E[X] = \int_{-\infty}^{\infty} xf_X(x)$ for a continuous RV X with PDF $f_X(x)$

and has the following properties:

- *Linearity of expectations:* $E[aX + bY] = aE[X] + bE[Y]$
- *Law of total expectation:* $E[X] = \sum_{i=1}^{n} E[X|C_i]P(C_i)$
- *Law of total probability:* $P[X = 1] = \sum_{i=1}^{n} P[X = 1|C_i]P(C_i)$
- *Law of total expectation/Adam's Law:* $E[X] = E\left[E[X|Y]\right]$

1.2. Variances

For any RVs X and Y and constants a and b, the *variance* is given by

$$Var(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2,$$

the *covariance* is given by:

$$Cov(X, Y) = E\left[(X - E[X])(Y - E[Y])\right] = E[XY] - E[X]E[Y],$$

and the *correlation* is given by:

$$\rho_{XY} = Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

They have the following properties:

- *Range:* $Var(X) \geq 0$ and $-1 \leq \rho_{XY} \leq 1$
- *Independence:* If X, Y are independent, $Cov(X,Y) = 0$ and $Corr(X,Y) = 0$
- *Covariance-Variance Relationship:* $Var(X) = Cov(X,X)$
- *Squared Constants:* $Var(aX) = a^2 Var(X)$
- *Bilinearity:* $Var(aX+bY) = a^2Var(X)+2abCov(X,Y)+b^2Var(Y)$ and $Cov(aX,bY) = abCov(X,Y)$
  - $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
  - $Var(X-Y) = Var(X) + Var(Y) - 2Cov(X,Y)$
  - If X, Y are independent, $Var(X+Y) = Var(X-Y) = Var(X) + Var(Y)$ and $Var(aX+bY) = a^2Var(X) + b^2Var(Y)$
- *Law of Total Variance/Eve's Law:* $Var(X) = E[Var(X|Y)] + Var(E[X|Y])$

1.3. Limiting Theorems

- *i.i.d.:* A set of random variables are *independent and identically distributed (iid)* if they are all mutually independent and follow the same distribution.
- *Weak Law of Large Numbers:* Let $X_1, X_2, ...$ be i.i.d. random variables with $E[X_i] = \mu$ for all i. Then $\lim_{n\to\infty} \bar{X}_n = \mu$, where $\bar{X}_n$ is the average of the first n X_i's: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- *Central Limit Theorem (CLT):* Let $X_1, X_2, ...$ be i.i.d. random variables with mean μ and variance σ^2. Then as $n \to \infty$:

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{\mathcal{D}} N(0,1).$$

- *Approximation Using CLT:* For i.i.d. RVs $X_1, X_2, ...$, for large n, the following distributions hold approximately:

$$\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right) \dot{\sim} N(0,1).$$

$$\bar{X}_n \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right).$$

1.4. Key Distributions

1.4.1. Bernoulli and Binomial Distributions

Definitions: Let Y be a random variable that can take values 0 or 1. If $P(Y = 1) = p$, then we say Y is Bernoulli distributed and write $Y \sim Ber(p)$. Suppose $Y_1, \dots, Y_n$ are i.i.d. with $Y_i \sim Ber(p)$ for all $i \in \{1, \dots, n\}$. Then the sum of the Y_i's given by $S_n = \sum_{i=1}^{n} Y_i$ is Binomial distributed with parameters n and p. We write $S_n \sim Bin(n, p)$.

Properties: Let $Y \sim Ber(p)$ and $S \sim Bin(n, p)$. Then:

$$E[Y] = pE[1-Y] = 1-pVar(Y) = p(1-p)E[S] = npVar(S) = np(1-p)E\left[\frac{S}{n}\right] = pVar\left(\frac{S}{n}\right) = \frac{p(1-p)}{n}$$

1.4.2. Normal (Gaussian) Distribution

Definitions: Let $Z \sim N(0, 1)$. Then Z follows a **standard normal distribution**, with $E[Z] = 0$ and $Var(Z) = 1$. Let $X = \mu + \sigma Z$. Then X follows a **normal distribution**, with $E[X] = \mu$ and $Var(X) = \sigma^2$, and we write $X \sim N(\mu, \sigma^2)$. Note that then $\frac{X-\mu}{\sigma} \sim N(0, 1)$. A normally distributed random variable is **symmetric** about its mean.

Rules of thumb: If $X \sim N(\mu, \sigma^2)$, then:

$$P(|X-\mu| < \sigma) \approx 0.68P(|X-\mu| < 1.96\sigma) \approx 0.95P(|X-\mu| < 2\sigma) \approx 0.954P(|X-\mu| < 3\sigma) \approx 0.997.$$

Location shifting and scaling of normal RVs: Let $X \sim N(\mu, \sigma^2)$ be a normally-distributed RV. For any real number a and positive real number b, $a + bX$ is also normally distributed and has distribution $a + bX \sim N(a + b\mu, b^2\sigma^2)$.

Sums of independent normal RVs: Let $X_1, \dots, X_n$ be independent normal random variables with means $\mu_1, \dots, \mu_n$ and variances $\sigma_1^2, \dots, \sigma_n^2$, respectively. Let $c_1, \dots, c_n$ be constants. Then:

$$\sum_{i=1}^{n} c_i X_i \sim N\left(\sum_{i=1}^{n} c_i \mu_i, \quad \sum_{i=1}^{n} c_i^2 \sigma_i^2\right).$$

1.4.3. Poisson and Related Distributions

Definition: Let Y be a random variable that can take discrete values $\{0, 1, 2, ...\}$. Then $Y \sim Pois(\lambda)$ if it has the PMF $P[Y = y] = f_Y(y) = \frac{\lambda^y e^{-\lambda}}{y!}$.

Properties: Let $Y \sim Pois(\lambda)$. Then:

$$E[Y] = \lambda Var(Y) = \lambda P[Y = 0] = e^{-\lambda}$$

The sum of independent Poisson-distributed random variables follows a Poisson distribution with expectation equal to the sum of the expectations of the random variables.

Related Distributions:

- The negative binomial distribution $Y \sim NB(r, p)$ has the same support and PMF $f_Y(y) = \binom{k+r-1}{k}(1-p)^k p^r$, expectation $\frac{r(1-p)}{p}$, and variance $\frac{r(1-p)}{p^2}$. In the limit as $p \to 1$ with fixed expectation, the negative binomial distribution converges to the Poisson distribution with the same expectation.
- In some cases, the Poisson distribution under-represents the number of zeros in an otherwise Poisson-distributed population. A zero-inflated Poisson distribution can then be used:
$$Z = \begin{cases} 0 & \text{, with probability } p \\ Y & \text{, with probability } 1-p, \end{cases}$$
where $Y \sim Pois(\lambda)$ and p is some positive probability. Then $E[Z] = (1-p)\lambda$ and $P[Z = 0] = p + (1-p)e^{-\lambda}$.
- In other cases, the Poisson distribution where zeroes are removed (truncated) is a better model. We call this the zero-truncated Poisson distribution. If $Z \sim Pois(\lambda)$, then the zero-truncated version of Z, $Z|Z > 0$ has the following properties:

$$P[Z = z|Z > 0] = \frac{\lambda^z}{(e^\lambda - 1)z!} E[Z|Z > 0] = \frac{\lambda e^\lambda}{e^\lambda - 1} Var(Z|Z > 0) = \frac{\lambda + \lambda^2}{1 - e^{-\lambda}} - \frac{\lambda^2}{(1 - e^{-\lambda})^2}$$

1.4.4. Exponential Distribution

Definition: Let Z be a random variable that can take positive values $(0, \infty)$ and has expectation $E[Z] = \frac{1}{\lambda}$. Then $Z \sim Exp(\lambda)$ if it has the PDF $f_Z(z) = \lambda e^{-\lambda z}$ on the support $(0, \infty)$.

Properties: Let $Z \sim Exp(\lambda)$. Then:

$$E[Z] = 1/\lambda Var(Z) = 1/\lambda^2$$

The CDF is given by:

$$F_Z(z) = P[Z \leq z] = 1 - e^{-\lambda z}, \text{ for } z > 0$$

The exponential distribution is also *memoryless*: the probability of the value falling in any fixed-width subsequent interval, conditional on not being below that, is always the same:

$$P[Z \geq s + t | Z \geq t] = P[Z \geq s] \text{ for any } s, t > 0$$

2. Key Definitions

2.1. Foundations of Study Design

- *Estimand:* the quantity we want to estimate in our analysis. It is a population parameter (not dependent on the data we collect). Notation generally used is a Greek letter (θ, β, μ, π etc.)
- *Estimator:* a statistic/function/algorithm of the collected that estimates the estimand. Notation generally used is the estimand with a hat ($\widehat{\theta}$, $\widehat{\beta}$, etc.), and can be expressed as a function of the observations. For example, if the estimator is the mean of the n sample observations, it could be written $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}_n$.
- *Estimate:* the numerical value we get from applying the estimator to the data. Notation is the same as the estimator.
- *Operating Characteristics:* the operating characteristics of an estimator are any statistical quantities or properties of the estimator under assumptions we define for the study and population.
- *Sampling Distribution:* the sampling distribution of an estimator is the distribution of the estimates if we repeatedly performed the full experiment (with new randomization, sampling, etc.). It is an operating characteristic of the estimator.
- *Population:* the set of all possible study units of interest. The group on which the estimand is defined.
- *Study units/sample:* the individuals upon which data will be collected. Each study unit corresponds to one observation in the data set.
- *Sample size:* the number of study units that will be used in the study. Generally denoted by n or N.
- *External validity/generalizability/transportability:* these all refer to the ability of the results (estimates) of one study to apply to another population, treatment, or outcome. In other words, do the results only apply to the narrow sample population or can we assume they apply in broader settings?

2.2. Quantifying Error and Variability

- *Bias:* the difference between the expectation of the estimator and the true value of the estimand. Lower bias corresponds to higher *accuracy* of the estimator.
  - $Bias(\hat{\theta};\theta) = E[\hat{\theta}] - \theta$.
  - An estimator $\hat{\theta}$ is *unbiased* for θ if $Bias(\hat{\theta};\theta) = 0$; that is, if $E[\hat{\theta}] = \theta$.
  - Note the $;\theta$ part of the notation is often dropped, with the estimand of comparison implicit.

- *Sampling Variance/Variability:* the variance of the estimator, an operating characteristic and key part of the sampling distribution. Denoted $Var(\hat{\theta})$. Lower sampling variance corresponds to higher *precision* of the estimator.
- The *standard error* of an estimator is its sampling standard deviation, or the square root of the sampling variance: $SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$.
- A *X% confidence interval (CI)* is the set of values of the parameter that gives us X % confidence that the true parameter is within those values. That is, if we repeated the experiment an infinite number of times, X % of the CIs would contain the true parameter.
  - For a 95% CI where the estimator has a normal sampling distribution, the interval is $\hat{\theta} \pm 1.96SE(\hat{\theta})$.
  - The *margin of error* for an estimator is one-half of the width of the confidence interval, or the amount added to the estimate to get the upper limit of the CI and subtracted from the estimate to get the lower limit of the CI. It is generally calculated based on a 95% CI.

- The *mean squared error (MSE)* of an estimator is a measure of its accuracy and precision and is the expected squared difference between the estimator and estimand.
  - $MSE(\hat{\theta};\theta) = E[(\hat{\theta} - \theta)^2]$.
  - $MSE(\hat{\theta};\theta) = [Bias(\hat{\theta},\theta)]^2 + Var(\hat{\theta})$.

- To compare the *efficiency* of design 1 (with estimator $\hat{\theta}_1$) compared to design 2 (with estimator $\hat{\theta}_2$), with the same sample size, we can use the *design effect (DE)*. A design effect less than 1 indicates that design 1 is more efficent (has a lower variance for the same sample size) than design 2. Usually, design 2 is some simple design (e.g., SRS). It is calculated as follows:
  - $DE = \frac{Var(\hat{\theta}_1)}{Var(\hat{\theta}_2)}$.

2.3. Sampling Designs

2.3.1. General Definitions

- *Nonprobability sampling* refers to any sampling scheme that does not involve randomness. E.g., convenience sampling or a fixed set of study units.
- *Probability sampling* refers to any sampling scheme that does involve randomness.
- A *systematic sample* is a sample of the population chosen by a rule to be balanced or representative along some set of known covariates or characteristics of the population. It is usually a probability sample but with a set repetition.
- *Simple random sampling (SRS)* is a probability sampling technique where each individual in the population is equally likely to be selected into the sample.
- *Nonresponse bias* occurs when study units with a certain value of the characteristic of interest are more likely not to respond to the survey/more likely not to be sampled than others.
- *Measurement error* occurs when the recorded response differs from the truth. This causes *measurement bias* if it is more likely to occur in one direction than another.
- *Stratified sampling* is a probability sampling technique where the population is divided into J groups (called *strata*) and SRS is conducted within each stratum. For each stratum j, n_j study units are chosen at random, and a within-stratum estimator $\hat{\theta}_j$ is calculated. Then the overall estimator is some weighted average of the within-stratum estimators: $\hat{\theta} = \sum_{j=1}^{J} w_j \hat{\theta}_j$. Weights are chosen for desired operating characteristics; to get an unbiased $\hat{\theta}$, the weights should be equal to the proportion of the population that is in that stratum. That is, we should choose $w_j = P[S_i = j]$ where $P[S_i = j]$ is the true probability that a randomly-chosen unit from the population is in stratum j.
- *Self-weighted stratified sampling* refers to a stratified sample where the n_j values (the number of study units sampled in stratum j) are selected to be proportional to the relative size of that stratum in the population. That is, $n_j = P[S_i = j] \cdot n$, where n is the total sample size and $P[S_i = j]$ is the true probability that a randomly-chosen unit from the population is in stratum j. This is generally the lowest-variance unbiased stratified sampling scheme.
- *Poststratification* is the weighting of responses in the analysis phase to match predetermined proportions of key covariates. E.g., the population is divided into J distinct groups and the average response for each group is weighted to match the proportion of the population made up by that group. Note that it gives no control over n_j, the number of study units in each group.

2.3.2. Operating Characteristics of Random Samples

For a random sample of n individuals, whose values are denoted $Y_1, \dots, Y_n$, where each individual has expectation $E[Y] = \mu_Y$ and variance $Var(Y) = \sigma_Y^2$, the mean result $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$ has properties:

$$E[\bar{Y}_n] = \frac{nE[Y_1]}{n} = \mu_Y Var(\bar{Y}_n) = \frac{Var(Y_1)}{n} = \frac{\sigma_Y^2}{n},$$

where the variance result requires independence of individual outcomes.

If n is large, the Central Limit Theorem kicks in and we get the approximate sampling distribution: $\bar{Y}_n \dot{\sim} N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$.

Note that for a proportion (where $Y_i = 1$ if individual i responds Yes and 0 if they respond No), the expectation $E[Y] = P[Y = 1]$ and the variance is $Var(Y) = P[Y = 1] (1 - P[Y = 1])$. If the true value of the proportion in the population (the estimand) is θ, then $E[Y] = \theta$ and $Var(Y) = \theta(1 - \theta)$, and you can plug these into the above equations as μ_Y and σ_Y^2, respectively.

Additional Resource: See Lessons 1 and 2 of (https://online.stat.psu.edu/stat506/lesson/1) for more details on SRS operating characteristics.

2.3.3. Operating Characteristics of Stratified Samples

A within-stratum estimator $\hat{\theta}_s$ is found for each group. Generally, $\hat{\theta}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{j,i}$, where n_j is the sample size within group j (for $j = 1, \ldots, J$) and the individual results for people in group j are labelled $Y_{j,1}, \ldots, Y_{j,n_j}$. n_j can be chosen by the investigator in advance.

Then the overall estimator is constructed as $\hat{\theta} = \sum_{j=1}^{J} w_j \hat{\theta}_j$, where $\sum_{j=1}^{J} w_j = 1$. The most common choice for w_j is $P[S_i = j]$, the proportion of the total population that falls in that stratum (i.e., the probability that a randomly chosen individual from the population is in stratum j). That gives an unbiased estimator for θ, **as long as the weights are correct**. Note that the weights rely on external information and cannot be chosen from the data.

We can think of each stratum as having a true within-stratum estimand θ_s, where the overall estimand $\theta = \sum_{j=1}^{J} \theta_j P[S_i = j]$ by the law of total expectation.

2.3.4. Poststratification Details

The notation is the same as for Stratified Samples. In this case, however, n_j is not specified in advance, but is a result of the random sampling. The construction of the within-stratum and overall estimators follows the same pattern, however.

Additional Resource: See Lesson 6 of (https://online.stat.psu.edu/stat506/lesson/6) for more details on stratified sampling and poststratification.

2.4. Randomized Controlled Trials

2.4.1. General Definitions

- A *randomized controlled trial (RCT)* is a study that compares two or more *treatment arms*, often a new intervention and a standard of care or placebo. Investigators controls factors as much as possible by specifying a population, treatment regimens, and outcomes. Investigators also randomize participants to the trial arms.
- A *clinical trial* is a study, usually a RCT, in which a new medicine or medical project is compared to a standard of care and/or placebo.
- A *placebo* treatment is an inactive treatment (one that should not affect the outcome biologically) that mimics the appearance or process of the new intervention. It is especially used to ensure concealed allocation and masking.
- *Masking*, allocation concealment, or "blinding" is the process of concealing from study participants (single-masked) and providers/assessors (double-masked) which participant is in which treatment arm. It aims to reduce bias from differential loss-to-follow-up or measurement error.
- *Positivity* is needed for causal inference. Study units must have a positive probability of being in each treatment arm in order to enable comparisons across those arms.
- *Consistency* in causal inference refers to the specific following of the treatment arm's protocol. In other words, that the outcomes reflect the treatment arm as designed.
- *Exchangeability* is the property that treatment arm groups would have, on average, the same outcome value in the absence of the new intervention. Randomization is used to ensure exchangeability by eliminating *confounding bias*: bias arising from other factors that are predictive of both the treatment arm and the outcome. This protects *internal validity*: are we estimating what we think we are?
- *Intercurrent events* are events that occur after the randomization/initiation of the intervention. This can include failure to adhere to the assigned treatment regimen, drop-out from the study or loss to follow-up, or emergency "rescue" interventions. They can cause violations of exchangeability because they are not randomized. Two common approaches to handling these are:

  - *Intention-to-treat (ITT) analysis:* analyze all data in the group the participant was assigned regardless of adherence to assignment. High validity, targets policy of treatment assignment.
  - *Per-protocol (PP) analysis:* analyze only the data from those who follow their assignment perfectly. Targets actual effect of intervention itself, but risks bias (targets effect on compliers).

- *Generalizability* is the ability of the study's results to be applicable to other populations, subgroups, or experimental conditions. Also known as **transportability** or **external validity**. There is often a tradeoff between internal validity (can we reasonably assume that our sample represents the population of interest) and external validity (how broad is our population of interest? How well will our results generalize to other populations?).

2.4.2. Hypothesis Testing

- A *Type I Error* of a hypothesis test occurs when the null hypothesis is incorrectly rejected (i.e., when the null is true). A *Type II Error* occurs when the null hypothesis is incorrectly not rejected (i.e., fail to reject when the null is not true).
- The *significance level* is the probability of a Type I Error, generally denoted α. $\alpha = P(\text{reject } H_0|H_0)$. Common α values include 0.05 (most common, corresponds to a 95% confidence interval), 0.01, and 0.1.
- The *p-value* is the result of the hypothesis test that is compared to the significance level. If $p < \alpha$, we reject the null hypothesis; otherwise we fail to reject the null hypothesis.
- The *power* of the test is one minus the probability of a Type II Error, often denoted $1 - \Delta$. $1 - Delta = 1 - P(\text{fail to reject } H_0|H_A)$. Note that it depends on the specific value of the parameter.
- The *minimum detectable effect* or *meaningful effect size* is the minimum value of the estimand, θ, that we want to have our chosen power to detect. So it is the value of the effect size that we assume in order to calculate the power of our test.

3. Randomized Controlled Trials

3.1. Mean Difference Estimand and Estimator for RCTs

If we want to estimate the difference in mean outcomes (whether they are continuous or binary) due to receiving the intervention compared to the placebo, we use the estimand $\theta = \mu_1 - \mu_0 = E[Y_1] - E[Y_0]$. Note that we could also estimate a ratio by dividing or a more complex quantity, like a regression parameter. For the difference estimand, we might use the estimator $\hat{\theta} = \bar{Y}_1 - \bar{Y}_0$, the mean observed outcome among the intervention group minus the mean observed outcome among the control group. Then:

$$
\begin{aligned}
E[\hat{\theta}] &= E\left[\bar{Y}_1 - \bar{Y}_0\right] = E[\bar{Y}_1] - E[\bar{Y}_0] \\
&= \frac{1}{n_1}\sum_{i=1}^{n_1} E[Y_{1,i}] - \frac{1}{n_0}\sum_{j=1}^{n_0} E[Y_{0,j}], \text{ by linearity of expectation} \\
&= E[Y|X=1] - E[Y|X=0], \text{ since we assume our individuals are representative of the population} \\
&= E[Y_1|X=1] - E[Y_0|X=0], \text{ by consistency} \\
&= E[Y_1] - E[Y_0], \text{ by exchangeability} \\
&= \mu_1 - \mu_0 = \theta.
\end{aligned}
$$

So our estimand is unbiased for θ.

3.2. Variance for RCTs

For an RCT (like in the previous section) where the estimand is $\theta = \mu_1 - \mu_0 = E[Y_1] - E[Y_0]$ and the estimator is $\hat{\theta} = \bar{Y}_1 - \bar{Y}_0$, the variance of the estimator is given by:

$$\begin{aligned} Var(\hat{\theta}) &= Var\left(\frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1,j} - \frac{1}{n_0}\sum_{j=1}^{n_0} Y_{0,j}\right) \\ &= Var\left(\frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1,j}\right) + Var\left(\frac{1}{n_0}\sum_{j=1}^{n_0} Y_{0,j}\right) \\ &= \frac{1}{n_1} Var(Y|X=1) + \frac{1}{n_0} Var(Y|X=0) \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}, \end{aligned}$$

where σ_1^2 is the variance of the outcome of a single outcome in the intervention arm and σ_0^2 is the variance of the outcome of a single outcome in the control arm.

From this, we can get a 95% confidence interval for the parameter value:

$$\theta \in \hat{\theta} \pm 1.96\sqrt{Var(\hat{\theta})}$$

3.3. Power and Sample Size Determination for RCTs

Instead of (or in addition to) a confidence interval, we often conduct a hypothesis test of the null hypothesis that there is no difference between the mean outcomes in the two groups:

$$H_0 : \ \theta = 0 H_A : \ \theta \neq 0$$

This can be conducted using a t-test with the test statistic:

$$t_{obs} = \frac{\hat{\theta}}{Var(\hat{\theta})} = \frac{\bar{Y}_1 - \bar{Y}_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}}} \sim t_{n_1+n_0-1}$$

Using a normal approximation to the t-distribution (valid for large enough sample sizes), we can find a formula for the power $(1 - \Delta)$ of the hypothesis test:

$$1 - \Delta = 1 - \Phi\left(\Phi^{-1}\left(1 - \alpha/2\right) - \frac{\theta}{\sqrt{Var(\hat{\theta})}}\right),$$

where Φ is the CDF of the standard normal distribution, Φ^{-1} is the inverse of Φ (note $\Phi^{-1}\left(1 - \alpha/2\right) = 1.96$ if $\alpha = 0.05$), α is the significance level, and θ is the effect size.

If this is a standard (unstratified, unmatched) RCT, then $Var(\hat{\theta}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}$. If it is balanced ($n_1 = n_0 = n$ sample size per arm) and we assume the intervention does not affect the variance ($\sigma_1^2 = \sigma_0^2$), then this formula simplifies to:

$$1 - \Delta = 1 - \Phi\left(\Phi^{-1}\left(1 - \alpha/2\right) - \sqrt{\frac{n}{2}}\frac{\theta}{\sigma_0}\right)$$

In the design phase, if we specify our desired power $1 - \Delta$, significance level α, minimum detectable effect size θ_{MDE}, and assume values for the variance of the outcomes σ_0^2 and σ_1^2, then we can find the needed sample size by:

$$n = \left[\Phi^{-1}\left(1 - \alpha/2\right) - \Phi^{-1}(\Delta)\right]^2 \cdot 2 \cdot \left[\frac{\sqrt{\frac{1}{2}(\sigma_0^2 + \sigma_1^2)}}{\theta_{MDE}}\right]^2$$

For 80% power, $Phi^{-1}(\Delta) = -0.842$; for 90% power, $Phi^{-1}(\Delta) = -1.282$.

For $\alpha = 0.1$, $\Phi^{-1}(1 - \alpha/2) = 1.65$; for $\alpha = 0.05$, $\Phi^{-1}(1 - \alpha/2) = 1.960$; and for $\alpha = 0.01$, $\Phi^{-1}(1 - \alpha/2) = 2.576$.

The value in the second set of brackets is often called $1/d$, where $d = \frac{\theta_{MDE}}{\sigma_p}$ is the standardized effect size (or Cohen's d value), the ratio of the minimum detectable effect size to the pooled (average across arms) standard deviation.

3.3.1. Calculating Power and Sample Size

Online formula sheet: Various power and sample size formulae can be found at this BU SPH site. Be sure to find the right setting and read the definitions of the inputs carefully, as the notation may be different from what we used in class.

Online calculator: Some power and sample size calculations can be done at this UBC site. Click on "Comparing Means for Two Independent Samples" for a t-test (difference in means for continuous RVs) calculation and "Comparing Proportions for Two Independent Samples" for a test of difference of proportions (risk difference for binary RVs) calculation.

Using R: power or sample size for a comparison of two proportions (risk difference) can be conducted in `R` using `power.prop.test()`. Power or sample size for a comparison of two continuous outcome means (mean difference) can be conducted in R using `pwr.t.test()` in the `pwr` library. Note that the values may be slightly different than using the above formula because they use more precise distributions instead of a normal approximation. For both functions, you can input the assumed parameters (probability of the outcome in both arms for the proportions test or Cohen's d value for the means test) and 2 of the 3 of sample size, significance level, and power, and it will give you the one you do not enter. Use `alternative="two.sided"` and, for the means test, `type="two.sample"`.

3.4. Ethics and Feasibility

Key principles of ethics of research in the U.S. context:

1. **Respect for persons:** this includes informed consent.
2. **Equipoise:** a genuine uncertainty about which treatment condition is better, and a valid approach to gaining knowledge from the research.
3. **Beneficence:** do no harm. Maximize potential benefits and minimize potential risks.
4. **Justice:** have reasonable, non-exploitative, well-considered procedures that are administered fairly. Distribute potential benefits and risks as fairly as possible.

External validity/generalizability and **interpretability** of results are also key considerations for trials, often limited by the feasibility of conducting randomization, recruiting a representative sample of people, and measuring the outcomes that we are most interested in.

Addressing ethical or logistical limitations:

- Change the intervention (risk to consistency and generalizability)
- Change the outcome (risk to internal validity and generalizability)
- Change the study population (risk to generalizability)
- Find a "natural experiment" where the intervention is randomized for another reason (hard to identify, often not exactly the intervention of interest, still may have bias, often high loss to follow up)
- Conduct an observational study (loses the exchangeability benefits of randomization)

In particular, it can be hard to randomize or even measure properly exposures/interventions related to large structural issues in society (e.g., race, gender, wealth, income, education, socioeconomic status). So a nuanced approach to research must be taken that appreciates the statistical properties of various approaches without becoming doctrinaire about only accepting a certain type of evidence.

3.5. Stratification and Matching

A stratified trial with S strata, where $\bar{Y}_{j,s}$ indicates the average outcome in treatment arm j in stratum s (similar subscripts for expected outcomes $\mu_{j,s}$, variances $\sigma^2_{j,s}$, and number per group $n_{j,s}$ can be conducted. A difference estimator then is given by:

$$\hat{\theta}_{strat} = \sum_{s=1}^{S} w_s \hat{\theta}_s,$$

where w_s is the weight for stratum s, with the weights adding to 1, and $\hat{\theta}_s$ is the stratum-specific estimator:

$$\hat{\theta}_s = \bar{Y}_{1,s} - \bar{Y}_{0,s}$$

Under this approach:

$$E[\hat{\theta}_s] = \mu_{1,s} - \mu_{0,s} Var(\hat{\theta}_s) = \frac{\sigma^2_{1,s}}{n_{1,s}} + \frac{\sigma^2_{0,s}}{n_{0,s}} E[\hat{\theta}_{strat}] = \sum_{s=1}^{S} w_s(\mu_{1,s} - \mu_{0,s}) Var(\hat{\theta}_{strat}) = \sum_{s=1}^{S} w_s^2 \left(\frac{\sigma^2_{1,s}}{n_{1,s}} + \frac{\sigma^2_{0,s}}{n_{0,s}} \right)$$

Usually, a self-weighted estimator is used, where $w_s = \frac{n_{0,s}+n_{1,s}}{\sum_{s=1}^{S}(n_{0,s}+n_{1,s})}$, we assume that $\sigma^2_{1,s} = \sigma^2_{0,s} = \sigma^2_s$ for all s, and we randomize 1:1 within each stratum so that $n_{1,s} = n_{0,s} = n_s$. In that case:

$$Var(\hat{\theta}_{strat}) = \frac{1}{(\sum_{s=1}^{S} n_s)^2} \sum_{s=1}^{S} 2 n_s \sigma^2_s$$

This variance can be used in power calculations. The average outcome variance is given by $\sigma^2_{strat} = \frac{\sum_{s=1}^{S} n_s \sigma^2_s}{\sum_{s=1}^{S} n_s}$, which can be plugged into formulae to get the Cohen's effect size and used to calculate power/sample size in `R`.

Stratification reduces the required sample size for any given power if $\sigma^2_{strat} < \sigma^2$. That is, if the stratification groups have within-stratum variance that's lower than the overall variance. These variances are connected through the law of total variance.

Matching works similarly to stratification, but generally there is only one study unit in each arm within each matched pair. The relevant variance then is $\sigma^2_D = Var(Y_{1,m} - Y_{0,m})$, where m indexes the matched pair. If this variance is known, it can be used in power and sample size calculations with `type="one.sample"`. We simply compare the average of the within-pair differences to 0 for our hypothesis test.

Blocking is a type of stratification that stratifies on when participants are enrolled, ensuring equal (or the set proportion) of allocation between the two arms within every group (of some pre-determined size) of participants. More details on specific randomization schemes can be found at this site from Penn State.

3.6. Multi-Arm and Factorial Designs

Multi-arm trials compare more than 1 intervention group with a single control group. They are often *unbalanced*: have a higher number of study units in the control group than in either intervention group.

The *Bonferroni Correction* adjusts for multiple comparisons by adjusting the significance level α used for all hypothesis tests, dividing it by the total number of hypothesis tests conducted. For example, for a trial with 2 intervention arms (plus a control), where there will be two tests conducted, and a desired overall risk of Type I Error of $\alpha = 0.05$, $\alpha/2 = 0.025$ would be used as the significance level.

The *Type I Error Rate* of a hypothesis test is the probability of having a significant result given that the null hypothesis is true. This is equal to the significance level α for a single test. The Bonferroni Correction is *conservative*: the overall Type I Error rate ends up strictly less than α.

Factorial trials test many combinations of the possible *levels* of several different *factors*. They are often described as $x \times y \times \ldots$ factorial designs, where each value represents the number of levels for one of the factors. For example, a 3×2 factorial design has two factors: one with 3 levels and the other with 2 levels.

Factorial trials may be:

- *complete:* having every combination of possible levels represented) or incomplete, and either
- *balanced:* (each combination of possible levels represented has the same number of study units or *replicates*) or *unbalanced*.

Factorial trials can be used to assess several different estimands or causal contrasts:

- Main effects* refer to the average effects comparing two or more levels of a single factor. The main effects estimated are the average across all levels of the other factors.
- Simple effects* refer to the specific effects comparing two or more levels of a single factor, within a specific level of another factor (or factors).
- *Interaction effects* refer to the difference in the simple effect based on the level of some other factor, or the difference between the main and simple effects.
- *Higher-order interaction effects* can occur if there are more than 2 factors.
- Comparisons across the individual cells (combinations) can also be made directly.

These effects can often be put into a multiple regression model, and different effects identified as different linear combinations of the regression coefficients.

3.7. Cluster Randomized Trials

Cluster randomized trials randomize entire clusters (or groups) of study units to the treatment or control conditions. This can alleviate consistency issues by reducing the risk that individuals in different arms interact and affect one another's outcomes. It is also useful for identifying effects at the group level or testing interventions that naturally occur at the group level.

3.7.1. Variance Inflation

For individual k in cluster j in treatment arm i, let $\mu_{i,j}$ be the underlying expected value for an individual in cluster j in arm i. let $\sigma_B^2 = Var(E[Y|\mu_{i,j}])$ be the between-cluster variance (the variance of the cluster-level expectations) and $\sigma_W^2 = Var(Y_{i,j,k}|\mu_{i,j})$ be the within-cluster variance (the conditional variance of the outcomes given a cluster's expectation), assumed to be constant across all clusters. Then the variance of the standard mean difference CRT estimator $\hat{\theta} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{m}\sum_{k=1}^{m} Y_{1,j,k} - \frac{1}{N}\sum_{j=1}^{N}\frac{1}{m}\sum_{k=1}^{m} Y_{0,j,k}$ is given by:

$$Var(\hat{\theta}_{CRT}) = \frac{2}{Nm}(\sigma_W^2 + m\sigma_B^2),$$

where m is the (average) number of study units per cluster and N is the number of clusters per treatment arm.

The variance of an individual outcome, by comparison, is:

$$Var(Y_{i,j,k}) = \sigma_W^2 + \sigma_B^2$$

The *design effect (DE)* comparing the CRT to a (hypothetical) individual RCT on the same outcome with the same overall variance and total sample size Nm is:

$$DE = \frac{Var(\hat{\theta}_{CRT})}{Var(\hat{\theta}_{RCT})} = \frac{\frac{2}{N}\left(\frac{\sigma_W^2}{m} + \sigma_B^2\right)}{\frac{2}{Nm}(\sigma_W^2 + \sigma_B^2)} = 1 + (m-1)\frac{\sigma_B^2}{\sigma_W^2 + \sigma_B^2}$$

We define the *intracluster correlation coefficient (ICC)* by $\rho \equiv \frac{\sigma_B^2}{\sigma_W^2 + \sigma_B^2}$, so that:

$$DE = 1 + (m-1)\rho$$

Note that ρ is constrained between 0 and 1 here (some other definitions/settings allow it to be negative). $\rho = 0$ corresponds to $\sigma_B^2 = 0$, meaning that cluster is not predictive of the outcome (there is no correlation between units in the same cluster). $\rho = 1$ corresponds to $\sigma_W^2 = 0$, meaning that cluster is fully predictive of the outcome (all units in a cluster have the same outcome, so adding more individuals within a cluster does not give more information).

The design effect represents the inflation in the variance for the CRT compared to a similar RCT. This also tells us by what factor we need to inflate the sample size (in total study units) compared to the sample size we calculate for a specific power and MDE for the similar RCT.

So the process for determining the sample size for a CRT is:

1. Find the sample size for a corresponding individually-randomized RCT, with outcome variance $\sigma^2 = \sigma_B^2 + \sigma_W^2$ and desired power, MDE, significance level.
2. Find the design effect, using an estimate of ρ (usually between 0.01 and 0.2, but a conservative approach would use 1) and the planned number of units measured within each cluster m.
3. Multiply the results from (a) and (b) to get the total number of individuals required. You can then divide this by m to get the total number of clusters required.

Note that in settings where you have a choice of m, it will affect the sample size needed. There are often cost and feasibility benefits to having a higher m and smaller N, although that will be less statistically efficient than a smaller m and larger N, so trade-offs occur!

4. Observational Studies

4.1. Notation

Let Y be the outcome of interest (can be binary, discrete, or continuous) and X be an indicator of the exposure status (for now, we'll think of a binary exposure, but it can extend to categorical or continuous). Below, we will use $X = 1$ for "exposed" and $X = 0$ for "unexposed".

- $Y(1)$ denotes the potential outcome if an individual had been exposed (status 1). $Y(1) = Y$ among the units with $X = 1$.
- $Y(0)$ denotes the potential outcome if an individual had been unexposed (status 0). $Y(0) = Y$ among the units with $X = 0$.
- $E[Y(1)]$ is the expected (or average) value if everyone were exposed; $E[Y(0)]$ is the expected value if everyone were unexposed.
- $E[Y|X = 1]$ is the expected value of individuals who actually were assigned treatment 1; $E[Y|X = 0]$ is the expected value of individuals who actually were assigned treatment 0.

4.2. Requirements for Causal Inference

- *Consistency/Stable Unit Treatment Value Assumption (SUTVA):* the outcome observed for an individual is their potential outcome corresponding to their actual exposure status. This gives $E[Y|X=1] = E[Y(1)|X=1]$ and $E[Y|X=0] = E[Y(0)|X=1]$. To get consistency, we need individuals to have a well-defined exposure status and for units not to affect one another.

- *Exchangeability:* the potential outcome under either condition would not be different in the two groups except for the effect of the exposure. That is, $E[Y(0)|X=1] = E[Y(0)|X=0]$ and $E[Y(1)|X=0] = E[Y(1)|X=1]$. To get exchangeability in RCTs, we randomize treatment assignment. This eliminates **confounding factors**, factors which affect both the treatment received and the outcome of interest. Note that anything that causes the two groups to differ after randomization (including differential loss-to-follow-up or dropout of the trial, differential misclassification or measurement of the outcomes, or different behaviors) can cause a violation of exchangeability. In observational studies, we rely on restricting the population, stratification/matching, and regression adjustment to handle confounding. Other threats to exchangeability can still occur, however, including selection bias.

- *Positivity:* units have a positive (nonzero) probability of receiving any exposure status. That ensures that we can estimate both $E[Y(1)]$ and $E[Y(0)]$ so that we can get a contrast.