Notes on STAT-341:
Computational Statistics and Data Analysis

Unversity of Waterloo

David Duan

Last Updated: April 19, 2022 (V1.0)

Contents

1 Samples **1**
1 All Possible Samples . . . 3
2 Selecting Samples . . . 5
3 Inclusion Probabilities . . . 9
4 Estimating Totals . . . 10
5 Sampling Design . . . 13

2 Inductive Inference **14**
1 Sources of Error . . . 16
2 Comparing Sub-Populations . . . 18
3 Interval Estimation . . . 23
4 Resampling . . . 26

3 Prediction **33**
1 Prediction Accuracy . . . 35
2 Prediction over Multiple Samples . . . 37
3 Back to Reality: Predictions with a Single Sample . . . 39

Chapter 1. Samples

1.1. Motivation: Until now, we have been assuming that when calculating attributes, we have access to the entire population. This is often not possible in practice. Thus, we need to consider using *samples* from the population, which can be viewed intuitively as an estimate of the population.

1.2. (Cont'd): Given a sample $\mathcal{S} \subseteq \mathcal{P}$ of $n \ll N = |\mathcal{P}|$ units, the attribute $a(\mathcal{S})$ calculated based on this sample is an estimate of its population counterpart $a(\mathcal{P})$, i.e.,

$$a(\mathcal{S}) = \widehat{a(\mathcal{P})} = a(\widehat{\mathcal{P}}).$$

There are two interpretations to this equality:

- $a(\mathcal{S}) = \widehat{a(\mathcal{P})}$: The sample attribute $a(\mathcal{S})$ is an estimate of the population attribute $\widehat{a(\mathcal{P})}$.
- $a(\mathcal{S}) = a(\widehat{\mathcal{P}})$: The sample $\mathcal{S}$ is an estimate of the population $\mathcal{P}$.

1.3. When using a sample instead of the entire population, we should consider *sample error* and *Fisher consistency*.

1.4. Note: The **sample error** is the difference between the estimate value $a(\mathcal{S})$ and the quantity being estimated $a(\mathcal{P})$, i.e.,

$$\text{sample error} = a(\mathcal{S}) - a(\mathcal{P}).$$

The nature of this error will depend on the sample and the attribute. For obvious reasons, an attribute with lower sampling error is preferable.

1.5. Note: An estimator a is said to be **Fisher consistent** if $\widehat{a(\mathcal{P})} = a(\mathcal{P})$, i.e., if the sample $\mathcal{S}$ is set to be the entire population $\mathcal{P}$, then the sample error should be zero.

Section 1. All Possible Samples

1.6. Note: In a population $\mathcal{P}$ of size N, there are $M := \binom{N}{n}$ different possible samples $\mathcal{S}$ of size n. The **sample error** for a sample $\mathcal{S}$ of size n is

$$\text{sample error} = a(\mathcal{S}) - a(\mathcal{P}) = \frac{1}{n}\sum_{u\in\mathcal{S}} y_u - \frac{1}{N}\sum_{u\in\mathcal{P}} y_u.$$

The **average sample error** over all possible samples of size n is

$$\text{average sample error } = \left(\frac{1}{M}\sum_{i=1}^{M} a\left(\mathcal{S}_i\right)\right) - a(\mathcal{P}).$$

Consistency and the Effect of Sample Size

1.7. Note: The nature of sample error depends largely on the sample size. As the sample size increases, the sample approaches the population, so the sample attribute values will concentrate more around the population attribute value. To quantify the concentration, we look at the **absolute sample error**, defined as the absolute difference of the sample attribute value and the population attribute value

$$|a(\mathcal{S}) - a(\mathcal{P})| = \left|\frac{1}{n}\sum_{u\in\mathcal{S}} y_u - \frac{1}{N}\sum_{u\in\mathcal{P}} y_u\right| < c$$

for some $c > 0$, which in turn helps us calculate the proportion of samples that satisfy this.

1.8. Definition: For $n \leq N$, define the *set of all possible samples of size n*:

$$\mathcal{P}_{\mathcal{S}}(n) = \{\mathcal{S} : \mathcal{S} \subset \mathcal{P} \text{ and } |\mathcal{S}| = n\}.$$

Given $c > 0$, define the *set of samples of size n with absolute sample error bounded by c* as

$$\mathcal{P}_a(c, n) = \{\mathcal{S} : \mathcal{S} \in \mathcal{P}_{\mathcal{S}}(n) \text{ and } |a(\mathcal{S}) - a(\mathcal{P})| < c\}$$

and define the *proportion of samples with absolute sample error bounded above by c*:

$$p_a(c, n) = \frac{|\mathcal{P}_a(c, n)|}{|\mathcal{P}_{\mathcal{S}}(n)|}.$$

For a fixed c, $p_a(c, n)$ increases with n.

1.9. Remark: Be familiar with how sample size affects the concentration of various attributes. For example, as n increases, sample means/trimmed means will concentrate more around the population mean/trimmed mean; the same holds for sample medians; the sample ranges will approach the population range as more samples will include $y_{(1)}$ and $y_{(N)}$; the sample IQRs will become more symmetric and concentrate about the population IQR.

Comparisons Across Attributes

1.10. Motivation: Previously, we defined consistency in terms of *absolute* sample error, which allowed us to evaluate the impact of sample size on concentration. However, if we want to compare different attributes, we need to use the **relative absolute** sample error, which is obtained by normalizing the absolute sample error by $a(\mathcal{P})$.

1.11. Definition: Let $c > 0$. Given a sample $\mathcal{S}$ of population $\mathcal{P}$, the **relative absolute sample error** is given by

$$\frac{|a(\mathcal{S}) - a(\mathcal{P})|}{|a(\mathcal{P})|}.$$

For a fixed $c > 0, n \leq N$, define the *set of samples of size n with relative absolute sample error bounded above by c* as

$$\mathcal{P}_a^\star(c, n) = \left\{ \mathcal{S} : \mathcal{S} \subset \mathcal{P}_\mathcal{S}(n) \text{ and } \frac{|a(\mathcal{S}) - a(\mathcal{P})|}{|a(\mathcal{P})|} < c \right\}$$

and the *proportion of samples of size n with relative absolute sample error bounded above by c* is given by

$$p_a^\star(c, n) = \frac{|\mathcal{P}_a^\star(c, n)|}{|\mathcal{P}_\mathcal{S}(n)|}.$$

1.12. Intuition: Compare $\mathcal{P}_a^*(c, n)$ and $p_a^*(c, n)$ with $\mathcal{P}_a(c, n)$ and $p_a(c, n)$ we defined previously. Intuitively, $p_a^*(c, n)$ measures the consistency of the sample attribute with respect to the *same* population attribute. When comparing between attributes, we are evaluating each attribute on how well its sample values track its population value on the *same scale*.

Section 2. Selecting Samples

1.13. Motivation: When we approximate the population attribute with the sample attribute, i.e., $a(\mathcal{S}) = \widehat{a(\mathcal{P})}$, we must acknowledge the possibility that $a(\mathcal{S}) \neq \widehat{a(\mathcal{P})}$ and the existence of **sample error**

$$a(\mathcal{S}) - a(\mathcal{P}).$$

Thus, we must be careful about how to select the sample and — if possible — do so in such a way to mitigate sample error.

1.14. Note: Given the reality that sample error is inevitable, it would be nice to understand the *magnitude* of error that can be expected. The **sampling distribution** of the attribute $a(\mathcal{S})$ gives insight into this. Properties of this distribution can be determined

- *exactly*, when all possible samples are available.
- *approximately*, when a subset of all possible samples is considered.
- *in expectation*, when a probabilistic sampling mechanism is used to draw a single sample.

We've already studied the first case. We will briefly discuss the second below, then devote the majority of our effort to the third case, which is also the most realistic of the three.

Selecting Samples: Approximation with Subset of Samples

1.15. Motivation: Let us first consider the case where we randomly select a subset of m samples. Consider drawing samples of size n from population $\mathcal{P}$.

1.16. Definition: Define the **population** of samples

$$\mathcal{P}_{\mathcal{S}} = \{\mathcal{S}_1, \ldots, \mathcal{S}_M\}$$

where $M = \binom{N}{n}$. From this, we can derive a **population of attributes**,

$$\mathcal{P}_{a(\mathcal{S})} = \{a(\mathcal{S}_1), \ldots, a(\mathcal{S}_M)\}.$$

1.17. Note: We often don't need all possible samples as a subset could provide a good approximation to the sampling distribution, from which we could compute the sampling error. Thus, we are not required to generate all possible samples to get an idea of the sample error or variability of the sample attribute values.

Quantifying Sample Error

1.18. Definition: Define $p(\mathcal{S}) > 0$ to be the probability of selecting sample $\mathcal{S}$ from the population of samples $\mathcal{P}_{\mathcal{S}}$. Note that $\sum_{\mathcal{S} \in \mathcal{P}_{\mathcal{S}}} p(\mathcal{S}) = 1$.

1.19. Motivation: The *average sample error* is given by

$$\frac{1}{M}\sum_{\mathcal{S}\in\mathcal{P}_\mathcal{S}}(a(\mathcal{S})-a(\mathcal{P})),$$

where $M=|\mathcal{P}_\mathcal{S}|$. We can quantify the concentration of sample errors in expectation using the following quantities. (Note that all expectations below are taken wrt the probabilities $p(\mathcal{S})$ of choosing a sample $\mathcal{S}$ from $\mathcal{P}_\mathcal{S}$.)

1.20. Definition: The **sampling bias** is the expected sample error induced by the repeated random sampling of $\mathcal{S}$ from $\mathcal{P}_\mathcal{S}$:

$$\mathbb{E}[a(\mathcal{S})]-a(\mathcal{P})=\sum_{\mathcal{S}\in\mathcal{P}_\mathcal{S}}a(\mathcal{S})p(\mathcal{S})-a(\mathcal{P})=\sum_{\mathcal{S}\in\mathcal{P}_\mathcal{S}}[a(\mathcal{S})-a(\mathcal{P})]p(\mathcal{S}).$$

- If $p(\mathcal{S})=1/M$, the sampling bias is identical to the **average sample error** of $a(\mathcal{P})$.
- If sampling bias is zero, then $a(\mathcal{S})$ is called an **unbiased** estimator of $a(\mathcal{P})$.

1.21. Definition: The **sampling variance** quantifies the *dispersion* in the sample errors:

$$\mathrm{Var}[a(\mathcal{S})]=\mathbb{E}[(a(\mathcal{S})-\mathbb{E}[a(\mathcal{S})])^2].$$

1.22. Definition: The **means squared error** quantifies the expected squared distance between $a(\mathcal{S})$ and $a(\mathcal{P})$:

$$\mathrm{MSE}[a(\mathcal{S})]=\mathbb{E}[(a(\mathcal{S})-a(\mathcal{P}))^2]=\mathrm{Var}[a(\mathcal{S})]+\text{Sampling Bias}^2.$$

1.23. Intuition: Intuitively, we could like $a(\mathcal{S})$ and $a(\mathcal{P})$ to be as close as possible. From the derivation above, we see that ideally, we could like to choose $p(\mathcal{S})$ and/or $\mathcal{P}_\mathcal{S}$ so that both the square of sampling bias and the sampling variance are as small as possible.

1.24. Note: Thinking of the sampling distribution of an attribute $a(\mathcal{S})$ gives rise to the notion of an attribute as an **estimator** (i.e., as a random variable). Let us introduce a random variable, say A, that takes values a from the distinct values of $a(\mathcal{S})$ for all $\mathcal{S}\in\mathcal{P}_\mathcal{S}$. The induced probability distribution is

$$\Pr(A=a)=\sum_{\mathcal{S}\in\mathcal{P}_\mathcal{S}}p(\mathcal{S})\cdot 1[a(\mathcal{S})=a].$$

- It follows that A is a discrete random variable.
- Probability statements about its values can be made using its distribution, including its expectation, variance, etc.
- Each of the definitions above (sampling bias/variance/MSE) can be defined in terms of this random variable and the corresponding probability distribution.

Sampling Mechanisms

1.25. Motivation: Rather than selecting $\mathcal{S}$ with probability $p(\mathcal{S})$ from the population of samples $\mathcal{P}_{\mathcal{S}}$, we can form $\mathcal{S}$ directly by selecting n units from the population of units $\mathcal{P}$. We select one unit at a time; a *sequence* of the first k units u_i is selected from $\mathcal{P}$ is $s_k = (u_{i_1}, \ldots, u_{i_k})$, where $i_j \in \{1, \ldots, N\}$ is the index of the original unit in $\mathcal{P} = \{u_1, \ldots, u_N\}$.

1.26. Definition: A **sampling mechanism** is defined by the probabilities $\Pr(u)$ and $\Pr(u \mid k, s_{k-1})$, the probability of selecting the first unit as well as the probability of selecting the kth unit given the first $k-1$ units. We can derive that the probability of the sequence of the first k units selected is

$$\Pr(s_k) = \Pr(u_{i_1}) \times \Pr(u_{i_2} \mid 2, s_1) \times \Pr(u_{i_3} \mid 3, s_2) \times \cdots \times \Pr(u_{i_k} \mid k, s_{k-1}).$$

To determine $p(\mathcal{S})$ from a sampling mechanism, *observe that the order in which the units appear does not matter*. Thus, $p(S)$ is the sum of $\Pr(s_n)$ over all permutations s_n:

$$p(S) = \sum_{s_n \text{ is a permutation of } S} \Pr(s_n)$$

1.27. Note (SRSWOR)**:** The **Simple Random Sampling Without Replacement** is defined as

$$\Pr(u) = \frac{1}{N}, \quad \Pr(u \mid k, s_{k-1}) = \frac{1}{N-k+1}.$$

The probability of the sequence s_n is

$$\Pr(s_n) = \frac{1}{N} \times \frac{1}{N-1} \times \frac{1}{N-2} \times \cdots \times \frac{1}{N-n+1},$$

which is the same for all $n!$ permutations. Thus,

$$p(\mathcal{S}) = \frac{n!}{N(N-1)(N-2)\cdots(N-n+1)} = \frac{1}{\binom{N}{n}}.$$

As expected, this probability is the same as the probability for selecting n distinct units from a population of N distinct units. The advantage of this is that we can select a sample *without first enumerating* all $M = \binom{N}{n}$ possible samples in $\mathcal{P}_{\mathcal{S}}$.

1.28. Note (SRSWR)**:** The **Simple Random Sampling With Replacement** is defined as

$$\Pr(u) = \frac{1}{N} = \Pr(u \mid k, s_{k-1})$$

and thus a sample $\mathcal{S}$ can contain one or more replicated units. The probability of the sequence s_n is

$$\Pr(s_n) = \left(\frac{1}{N}\right)^n.$$

Unlike SRSWOR, we typically treat each s_n as an ordered sample and so

$$\Pr(\mathcal{S}) = \Pr(s_n) = \frac{1}{N^n}.$$

The population of all samples $\mathcal{P}_\mathcal{S}$ in this case contains $M = N^n$ ordered samples. If we treat each s_n as an unordered sample similar to SRSWOR, we obtain

$$\Pr(\mathcal{S}) = \Pr(s_n) = \frac{\binom{N+n-1}{n}}{N^n}.$$

1.29. Note (SRSWH)**:** We also have another sampling mechanism proposed by Basu, sometimes known as **SRSWH**. Suppose we perform simple random sampling with replacement except that we remove any duplicate units. The samples produced will have sizes anywhere from 1 to n according to how many distinct units were selected in a sample (sampling with replacement).

1.30. Intuition: Let's see how these mechanisms are connected to the *balls in an urn* problem. Suppose that we have an urn containing N different balls that are either white or black. We would like to estimate the proportion of balls in the urn which are black by drawing n balls at random from the urn.

- SWSWOR: Randomly draw n balls from the urn one after another, without replacing any at any time. The estimation is the proportion of black balls in your sample.
- SRSWR: Randomly draw n balls from the urn one after another, each time replacing the ball after drawing it. The estimation is again the proportion of black balls in your sample.
- SRSWH: Select one ball at a time and record its colour; mark it with an X and return it to the urn. If a ball drawn already has an X marked on it, then it counts as a draw, but is returned to the box without recording its colour. Continue until n draws have been made. The estimate is the proportion of black balls observed with X's.

Section 3. Inclusion Probabilities

1.31. Definition: The **inclusion probability** for unit u is the probability of unit u being included in an arbitrary sample $\mathcal{S}$:

$$\pi_u := \Pr(u \in \mathcal{S}) = \sum_{\mathcal{S} \in \mathcal{P}_\mathcal{S}} p(\mathcal{S}) \cdot \mathbf{1}[u \in \mathcal{S}].$$

1.32. Note: Define the indicator variable D_u where

$$D_u = \begin{cases} 1 & u \in \mathcal{S} \\ 0 & u \notin \mathcal{S}. \end{cases}$$

The expectation and variance of D_u is given by

$$\begin{aligned} \mathbb{E}[D_u] &= 1 \times \Pr(D_u = 1) + 0 \times \Pr(D_u = 0) = \Pr(u \in \mathcal{S}) = \pi_u \\ \operatorname{Var}[D_u] &= \mathbb{E}[D_u^2] - \mathbb{E}[D_u]^2 \\ &= 1^2 \times \Pr(D_u = 1) + 0^2 \times \Pr(D_u = 0) - \pi_u^2 = \pi_u - \pi_u^2 = \pi_u(1 - \pi_u) \end{aligned}$$

1.33. Definition: The **joint inclusion probability** of units u and v is the probability that both u and v are included in a sample $\mathcal{S}$:

$$\pi_{uv} := \Pr(u \in \mathcal{S} \wedge v \in \mathcal{S}).$$

1.34. Note: The covariance of D_u and D_v are given by

$$\begin{aligned} \operatorname{Cov}(D_u, D_v) &= \mathbb{E}[D_u \cdot D_v] - \mathbb{E}[D_u] \cdot \mathbb{E}[D_v] \\ &= \mathbb{E}[D_u \cdot D_v] - \pi_u \cdot \pi_v \\ &= \pi_{uv} - \pi_u \cdot \pi_v \end{aligned}$$

1.35. Note: We now derive the inclusion and joint inclusion probabilities for the three sampling mechanisms discussed in the previous section.

- SRSWOR:

$$\pi_u = \Pr(u \in \mathcal{S}) = \frac{1 \times \binom{N-1}{n-1}}{\binom{N}{n}} = \frac{1 \times \binom{N-1}{n-1}}{\frac{N}{n} \times \binom{N-1}{n-1}} = \frac{n}{N}$$

$$\pi_{uv} = \Pr(u \in \mathcal{S} \wedge v \in \mathcal{S}) = \frac{1 \times 1 \times \binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

- SRSWR and SRSWH:

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

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

Section 4. Estimating Totals

1.36. Motivation: Many attributes are either a *total* of some variate y_u observed on every unit $u \in \mathcal{P}$, i.e.,

$$a(\mathcal{P}) = \sum_{u \in \mathcal{P}} y_u,$$

or a function of such a total, i.e.,

$$a(\mathcal{P}) = f\left(\sum_{u \in \mathcal{P}} y_u\right).$$

1.37. Example:

- The population average:

$$a(\mathcal{P}) = \frac{1}{N} \sum_{u \in \mathcal{P}} y_u.$$

- The population variance (and standard deviation):

$$a(\mathcal{P}) = \sum_{u \in \mathcal{P}} \frac{(y_u - \bar{y})^2}{N}, \quad a(\mathcal{P}) = \sqrt{\sum_{u \in \mathcal{P}} \frac{(y_u - \bar{y})^2}{N}},$$

- The proportion of units satisfying a predicate (represented by an indicator function I):

$$a(\mathcal{P}) = \sum_{u \in \mathcal{P}} \frac{I(y_u)}{N}.$$

- The CDF at a specific value y:

$$F_{\mathcal{P}}(y) = \frac{1}{N} \sum_{u \in \mathcal{P}} I(y_u \leq y).$$

1.38. Definition: The **Horvitz-Thompson estimate** of a population total is

$$\widehat{a(\mathcal{P})} = a_{HT}(\mathcal{S}) = \sum_{u \in \mathcal{S}} \frac{y_u}{\pi_u},$$

where π_u denotes the probability of inclusion of u in $\mathcal{S}$.

1.39. Intuition: Observe that the contribution for each unit u in the sample $\mathcal{S}$ is weighted inversely by π_u, i.e.,

- if the probability of inclusion is small, then the weight will be high;
- if the probability of inclusion is large, then the weight will be low.

Intuitively, we want to counteract the inclusion probability to compensate for the unobserved elements of the population.

1.40. Next, we consider properties of the HT estimator, $\tilde{a}_{HT}(\mathcal{S})$. These properties inform what can be expected under repeated sampling. In particular, the HT estimator is unbiased. 9

1.41. Note: Start by writing the HT estimator in terms of the indicator variable

$$\tilde{a}_{HT}(\mathcal{S}) = \sum_{u\in\mathcal{S}} \frac{y_u}{\pi_u} = \sum_{u\in\mathcal{P}} D_u \frac{y_u}{\pi_u}, \qquad \text{where } D_u = \begin{cases} 1 & u \in \mathcal{S} \\ 0 & u \notin \mathcal{S}. \end{cases}$$

Using what we derived about D_u from the previous section, we have

$$\mathbb{E}\left[\sum_{u\in\mathcal{P}} D_u\right] = \sum_{u\in\mathcal{P}} \mathbb{E}[D_u] = \sum_{u\in\mathcal{P}} \pi_u = \mathbb{E}[n] = n.$$

When the sample size is fixed,

$$\sum_{u\in\mathcal{P}} \pi_{uv} = \sum_{u\in\mathcal{P}} \mathbb{E}[D_u D_v] = \mathbb{E}\left[\sum_{u\in\mathcal{P}} D_u D_v\right] = \mathbb{E}\left[D_v \sum_{u\in\mathcal{P}} D_u\right] = \mathbb{E}[D_v \cdot n] = n\mathbb{E}[D_v] = n\pi_v.$$

In that in both derivations above, we assume that the sample size n is fixed.

1.42. (Cont'd): We now show that the HT estimator is unbiased

$$\begin{aligned}
\mathbb{E}[\tilde{a}_{HT}(\mathcal{S}) - a(P)] &= \mathbb{E}\left[\sum_{u\in\mathcal{P}} D_u \frac{y_u}{\pi_u}\right] - a(\mathcal{P}) \\
&= \sum_{u\in\mathcal{P}} \frac{y_u}{\pi_u} \mathbb{E}[D_u] - a(\mathcal{P}) \\
&= \sum_{u\in\mathcal{P}} \frac{y_u}{\pi_u} \pi_u - a(\mathcal{P}) \\
&= \sum_{u\in\mathcal{P}} y_u - a(\mathcal{P}) = a(\mathcal{P}) - a(\mathcal{P}) = 0.
\end{aligned}$$

Recall that MSE = Var + Bias^2. Since Bias is 0, the MSE is simply equal to the variance.

1.43. (Cont'd): It can be shown that the **variance** of the HT estimator is

$$\text{Var}\left[\tilde{a}_{HT}(\mathcal{S})\right] = \sum_{u\in\mathcal{P}} \sum_{v\in\mathcal{P}} \left(\pi_{uv} - \pi_u \pi_v\right) \frac{y_u}{\pi_u} \frac{y_v}{\pi_v} \equiv \sum_{u\in\mathcal{P}} \sum_{v\in\mathcal{P}} \Delta_{uv} \frac{y_u}{\pi_u} \frac{y_v}{\pi_v}.$$

which can be equivalently written in the **Yates-Grundy** or the **Sen-Yates-Grundy** formulation:

$$\text{Var}\left[\tilde{a}_{HT}(\mathcal{S})\right] = -\frac{1}{2} \sum_{u\in\mathcal{P}} \sum_{v\in\mathcal{P}} \Delta_{uv} \left(\frac{y_u}{\pi_u} - \frac{y_v}{\pi_v}\right)^2$$

The second version provides an intuition for choosing sampling mechanisms that minimize the variance (see next section).

1.44. Note: For the HT estimate, we define the following quantity. Let P_{uv} denote the population of all pairs (u, v) where $u, v \in \mathcal{P}$. For $(u, v) \in P_{uv}$, define

$$q_{uv} = \Delta_{uv} \frac{y_u}{\pi_u} \frac{y_v}{\pi_v}.$$

Then the variance of the HT estimator can be written as

$$\text{Var}\left[\tilde{a}_{HT}(\mathcal{S})\right] = \sum_{u \in \mathcal{P}} \sum_{v \in \mathcal{P}} \Delta_{uv} \frac{y_u}{\pi_u} \frac{y_v}{\pi_v} = \sum_{(u,v) \in \mathcal{P}_{uv}} q_{u,v}.$$

Written in this way, we can see that this variance is also a total, so we can estimate the variance of the HT estimator using a HT estimate, which gives us:

$$\widehat{\text{Var}}\left[\tilde{a}_{HT}(\mathcal{S})\right] = \sum_{(u,v) \in \mathcal{S}_{uv}} \frac{q_{u,v}}{\pi_{uv}} = \sum_{u \in \mathcal{S}} \sum_{v \in \mathcal{S}} \frac{\Delta_{uv}}{\pi_{uv}} \frac{y_u}{\pi_u} \frac{y_v}{\pi_v} = \sum_{u \in \mathcal{S}} \sum_{v \in \mathcal{S}} \left(\frac{\pi_{uv} - \pi_u \pi_v}{\pi_{uv}} \right) \frac{y_u}{\pi_u} \frac{y_v}{\pi_v}.$$

- Note that the sample $\mathcal{S}_{uv}$ is obtained by sampling from the population $\mathcal{P}_{uv}$ of all possible pairs (u, v). The probability that any particular (u, v) is included in the sample is given by the joint inclusion probability $\pi_{uv} > 0$.
- The square root of this variance estimate, or equivalently the HT estimate of the standard deviation, is commonly referred to as the **standard error** of the estimate:

$$SE\left(\tilde{a}_{HT}(\mathcal{S})\right) = \sqrt{\widehat{\text{Var}}\left[\tilde{a}_{HT}(\mathcal{S})\right]}.$$

1.45. Note: Thus, using HT estimation, we are able to construct

- an estimate of the population total, and
- an estimate of the variance of this estimator.

Both estimators are unbiased. Why is this useful?

(1). Many attributes can be written as a total, so the HT framework gives us an effective method of estimation.

(2). Understanding the sampling error requires just one sample.

1.46. Note: The 95% confidence interval estimates of the HT estimate is of the form

$$a_{HT}(\mathcal{S}) \pm 2\hat{SD}\left(\tilde{a}_{HT}(\mathcal{S})\right).$$

Section 5. Sampling Design

1.47. Definition: A **sampling design** refers to a pair $(\mathcal{P}_{\mathcal{S}}, p(\mathcal{S}))$, i.e., a set of samples along with the probability that each sample is getting chosen.

1.48. Example: The SRSWOR, SRSWR, SRSWH frameworks provide examples of different sampling designs.

1.49. Motivation: The sampling design is ours to choose. For example, we may choose $\mathcal{P}_{\mathcal{S}}$ so that the values of $a(\mathcal{S})$ for $\mathcal{S} \in \mathcal{P}_{\mathcal{S}}$ are constrained to be near $a(\mathcal{P})$. Alternatively, we may choose $p(\mathcal{S})$ so that samples $\mathcal{S} \in \mathcal{P}_{\mathcal{S}}$ that have $a(\mathcal{S})$ close to $a(\mathcal{P})$ have higher probability of being selected.

1.50. (Cont'd): Within the HT framework, we know that

$$\text{MSE}[\tilde{a}_{HT}(\mathcal{S})] = \text{Var}[\tilde{a}_{HT}(\mathcal{S})] = -\frac{1}{2}\sum_{u\in\mathcal{P}}\sum_{v\in\mathcal{P}}\Delta_{uv}\left(\frac{y_u}{\pi_u} - \frac{y_v}{\pi_v}\right)^2.$$

This provides insight into how we might best choose a sampling design.

- For example, if we could choose $\pi_u \propto y_u$, then the variance and MSE will be zero!
- If there is a variate x_u that is highly positively correlated with y_u for all $u \in \mathcal{P}$, then choosing $\pi_u \propto x_u$ could reduce MSE.
- If we knew when $y_u \approx y_v$, we could choose $\pi_u \approx y_v$ which might reduce MSE.

Much of the survey sampling is concerned with how best to choose the sampling design $(\mathcal{P}_{\mathcal{S}}, p(\mathcal{S}))$ to reduce the MSE of an estimator (attribute) of interest.

1.51. Note (Stratified Sampling)**:** Split the population $\mathcal{P}$ into H non-overlapping groups called **strata** and we sample without replacement from each stratum.

- Each stratum has N_h units, where $N_1 + \cdots + N_H = N = |\mathcal{P}|$.
- We sample n_h from each stratum, where $n_1 + \cdots + n_h = n = |\mathcal{S}|$.

Chapter 2. Inductive Inference

2.1. Motivation: Probabilistic reasoning can be used to quantify the potential magnitude of sample error and bias, provided that a probabilistic sampling mechanism is used.

The sampling behaviour of any population attribute is examined by repeatedly drawing samples according to the sampling mechanism and calculating the attribute on the samples. The sampling behaviour of a particular attribute can be summarized by its **sampling bias** and **sampling variability**.

Probabilistic sampling allows us to quantify the relative frequency in which any sample attribute value might be realized. This provides an insurance policy in what is learned about the attributes based on the sample and how it compares the attributes based on the population.

Section 1. Sources of Error

2.2. Motivation: Consider the *Path of Inductive Inference* shown below.

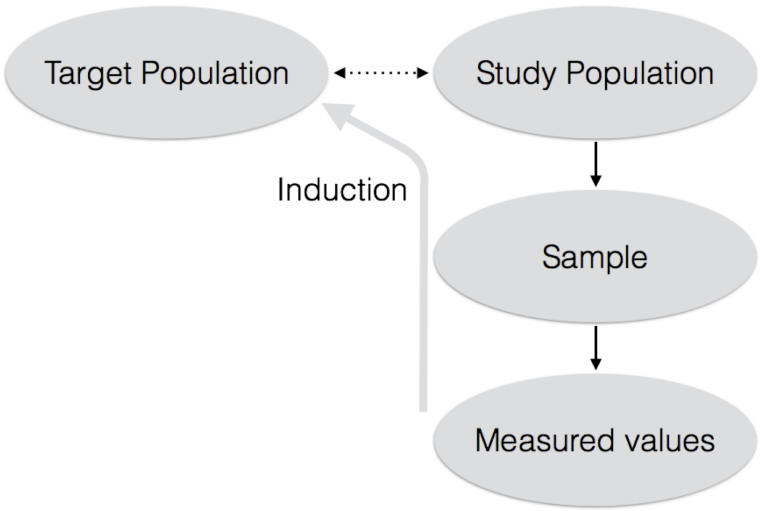

Figure 2.1: Path of Inductive Inference.

Carefully planned sampling designs can provide considerable assurance that the conclusions drawn from a sample will not likely be that different from those which would have been drawn were we able to access the entire population. However, there is almost always another source of error in our inferences that is not resolved by probabilistic sampling. To be more precise, the population which we are able to draw samples from is often not the population about which we would like to draw inferences.

2.3. Definition:

- The **target population** is the population about which we want to draw inferences.
- The **study population** is the population from which samples are taken.
- The difference between the attribute evaluated on them is the **study error**:

$$\text{Study Error} = a(\mathcal{P}_{\text{study}}) - a(\mathcal{P}_{\text{target}}).$$

2.4. Note: If we obtain a sample $\mathcal{S}$ to draw inferences about the target population $\mathcal{P}_{\text{target}}$, the error for a given attribute $a(\cdot)$ is

$$\begin{aligned} a(\mathcal{S}) - a(\mathcal{P}_{\text{target}}) &= [a(\mathcal{S}) - a(\mathcal{P}_{\text{study}})] + [a(\mathcal{P}_{\text{study}}) - a(\mathcal{P}_{\text{target}})] \\ &= \text{Sample Error} + \text{Study Error} \end{aligned}$$

Probabilistic sampling can be used to control the sample error but not the study error. It is challenging to keep study error small, and sometimes, it is not even within the domain of statistics.

2.5. Example: In *medical studies*, interest often lies in the progression of a disease or the efficacy of its treatment in humans. The target population is the set of all humans. However, for ethical and other reasons, the study cannot be conducted on humans but must instead be constructed on some other animals, such as mice, which serve as a model for humans. Thus, the population of mice available from which we can select is the study population. Sampling from this population provides some assurance about the quality and uncertainty of our inferences about study population's attributes, but this assurance does not carry over to inferences about the target population. Mice, after all, might be fundamentally different from humans for these particular attributes, so the quality of the inference depends on how close are the target and study populations.

2.6. Example: In *forecasting problems*, our target population often includes *future realization* of units which are not available at the time of study, e.g., a natural phenomenon such as the monthly tidal patterns and the daily closing price of a stock. In either case, arguing the study error must be small requires that the future should be much like past.

2.7. Note: Measurement errors refer to the errors made in measurement, which can also affect conclusions draw about attributes. Every measuring system has at least three sources of potential errors:

(1). the measuring device, sometimes called the **gauge**;

(2). the person reading or recording the measurement, sometimes called the **operator**;

(3). the method followed to take the measurement (i.e., anything independent of the gauges and operators).

The discipline known as **measurement system analysis** is devoted to the practical and academic study of measurement systems and methods by which their adequacy and comparability is determined.

Section 2. Comparing Sub-Populations

2.8. Motivation: Suppose we have two sub-populations, $\mathcal{P}_1$ and $\mathcal{P}_2$ and interest lies in comparing some attribute across the two sub-populations: $a(\mathcal{P}_1)$ and $a(\mathcal{P}_2)$.

- When the attribute is a measure of location, sub-population comparisons are typically based on the *difference* $a(\mathcal{P}_1) - a(\mathcal{P}_2)$.
- When the attribute is a measure of spread, sub-population comparisons are typically based on the *ratio* $a(\mathcal{P}_1)/a(\mathcal{P}_2)$.

2.9. If the two sub-populations are essentially the same, then the sub-populations observed should not look too different if we were to mix them up with one another. In other words, swapping units would not dramatically change the features of the resulting sub-populations. On the other hand, if the two sub-populations were very different, then shuffling the units could dramatically change the features of the resulting sub-populations.

More precisely, we combine the two sub-populations together into one, $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$, and then randomly draw two new sub-populations $\mathcal{P}_1^*$ and $\mathcal{P}_2^*$ while ensuring that their sizes are kept the same. We then compare the attributes of $\{\mathcal{P}_1, \mathcal{P}_2\}$ with $\{\mathcal{P}_1^*, \mathcal{P}_2^*\}$. If the sub-populations were similar to begin with, there shouldn't be a very large difference between attributes calculated on $\{\mathcal{P}_1, \mathcal{P}_2\}$ versus those calculated on $\{\mathcal{P}_1^*, \mathcal{P}_2^*\}$.

2.10. Next, we would like to quantify, numerically, how unusual the difference between $a(\mathcal{P}_1)$ and $a(\mathcal{P}_2)$ is relative to randomly mixed sub-populations.

- If the two sub-populations are actually similar, we want to provide numerical evidence *in favour of* the notion that the two sub-populations are similar to a randomly mixed sub-population.
- If the two sub-populations are actually different, we want to provide numerical evidence *against* the notation that the two sub-populations are similar to a randomly mixed sub-population.

We use the following steps to gather such evidence:

(1). We suppose the sub-populations were randomly drawn from the same population. This is known as the **null hypothesis**.

(2). We construct a **discrepancy measure** that quantifies how inconsistent our data is with the null hypothesis. A large value indicates evidence against the null hypothesis.

(3). We obtain the **observed discrepancy** by calculating the discrepancy measure on the two observed (i.e., unshuffled) sub-populations.

(4). Finally, we obtain the **observed p-value** by calculating the probability that a randomly shuffled sub-population has a discrepancy measure at least as large as the observed discrepancy. A small value indicates evidence against the null hypothesis.

Let's elaborate on each of these steps.

2.11. Note (Step 1: Null Hypothesis)**:** In our case, each of the following (equivalent) statements constitutes the **null hypothesis** we are testing:

- H_0: The sub-populations $\mathcal{P}_1$ and $\mathcal{P}_2$ were randomly drawn from the same population.
- H_0 $\mathcal{P}_1$ and $\mathcal{P}_2$ were created by randomly assigning units in the same population to one of the two sub-populations.
- H_0: $\mathcal{P}_1$ and $\mathcal{P}_2$ were generated by random mixing.

The **alternative hypothesis** H_A is the complement of H_0. Note that we do not state the null hypothesis in terms of the equivalence of attribute values, i.e., $a(\mathcal{P}_1) = a(\mathcal{P}_2)$. Although such a statement is true if H_0 holds (i.e., implied by H_0), it is weaker than H_0, and thus we avoid using it.

2.12. Note (Step 2: Discrepancy Measure)**:** A **discrepancy measure** (or **test statistics**) $D(\mathcal{P}_1, \mathcal{P}_2)$ quantifies how *inconsistent* our data is with the null hypothesis, and is defined so that large values indicate evidence against the null hypothesis. As a point of interest, the discrepancy measure is technically an attribute for the population, so we could consider properties such as *equivariance* and *invariance*.

The form of $D(\mathcal{P}_1, \mathcal{P}_2)$ depends on how we want to compare $\mathcal{P}_1$ and $\mathcal{P}_2$.

- For measures of location, the discrepancy measure is based on *differences* $a(\mathcal{P}_1) - a(\mathcal{P}_2)$.
- For measures of spread, the discrepancy measure is based on *ratio* $a(\mathcal{P}_1)/a(\mathcal{P}_2)$.

2.13. Example: Consider the following hypotheses and discrepancy measures.

- H_0: The averages from the two sub-populations were the same.
$$D(\mathcal{P}_1, \mathcal{P}_2) = |\overline{y}_1 - \overline{y}_2|.$$
- H_0: The standard deviation from the two sub-populations were the same.
$$D(\mathcal{P}_1, \mathcal{P}_2) = \left| \frac{SD(\mathcal{P}_1)}{SD(\mathcal{P}_2)} - 1 \right|.$$
- H_0: The average from the first population was smaller than the average of the second.
$$D(\mathcal{P}_1, \mathcal{P}_2) = \overline{y}_1 - \overline{y}_2$$
Here, $D(\mathcal{P}_1, \mathcal{P}_2)$ is large if $\overline{y}_1 > \overline{y}_2$ by a lot, which is inconsistent with our hypothesis.
- H_0: The average from the first population was larger than the average of the second.
$$D(\mathcal{P}_1, \mathcal{P}_2) = \overline{y}_2 - \overline{y}_1$$
Here, $D(\mathcal{P}_1, \mathcal{P}_2)$ is large if $\overline{y}_1 < \overline{y}_2$ by a lot, which is inconsistent with our hypothesis.

The last two examples should explain why $D(\mathcal{P}_1, \mathcal{P}_2)$ quantifies how *inconsistent* our data is with the null hypothesis, and how it is defined so that large values indicate evidence against the null hypothesis.

2.14. Note (Step 3: Observed Discrepancy)**:** The **observed discrepancy**, d_{obs}, is the value of the discrepancy measure D calculated on the two observed (i.e., unshuffled) sub-populations:

$$d_{\text{obs}} = D(\mathcal{P}_1, \mathcal{P}_2).$$

It is important to recognize that the discrepancy measure quantifies only one type of discrepancy between the population; all other differences are completely ignored.

2.15. Note (Step 4: Observed p-Value)**:** The **observed p-value** is the probability that a randomly shuffled sub-population has a discrepancy measure at least as large as the observed discrepancy.

$$p\text{-value} = Pr\left(D \geq d_{obs} \,\middle|\, H_0 \text{ is true}\right).$$

If the p-value is very small, then either the null hypothesis is true and we have observed a very unusual value of d_{obs}, or the null hypothesis is false. The smaller the p-value, the greater the evidence against the null hypothesis.

2.16. Example: In our case, to calculate the *exact* p-value, one must consider all

$$\binom{N_1 + N_2}{N_1} = \binom{N_1 + N_2}{N_2}$$

possible permutations of the observed data; the exact p-value is the fraction of $D(\mathcal{P}_1^*, \mathcal{P}_2^*)$ values greater than or equal to d_{obs}. Since this is too expensive, we typically just use M (a large number) of them, i.e., generate M shuffled pairs $(\mathcal{P}_{1,1}^\star, \mathcal{P}_{2,1}^\star), (\mathcal{P}_{1,2}^\star, \mathcal{P}_{2,2}^\star), \ldots, (\mathcal{P}_{1,M}^\star, \mathcal{P}_{2,M}^\star)$ and calculate the *approximated* p-value as

$$\frac{1}{M}\sum_{i=1}^{M} I\left(D(\mathcal{P}_{1,i}^\star, \mathcal{P}_{2,i}^\star) \geq d_{obs}\right) \qquad \text{where} \qquad d_{obs} = D(\mathcal{P}_1, \mathcal{P}_2)$$

2.17. Example: Putting everything together, we have:

- H_0: $\mathcal{P}_1$ and $\mathcal{P}_2$ are drawn from the same population.
- Construct $D = D(\mathcal{P}_1, \mathcal{P}_2)$ where large values indicate evidence against H_0.
- Calculate $d_{\text{obs}} = D(\mathcal{P}_1, \mathcal{P}_2)$.
- Shuffle the sub-populations M times and calculate the observed p-value:

$$p\text{-value} = \Pr\left(D \geq d_{obs} \,\middle|\, H_0 \text{ is true}\right) \approx \frac{1}{M}\sum_{i=1}^{M} I\left(D(\mathcal{P}_{1,i}^\star, \mathcal{P}_{2,i}^\star) \geq d_{obs}\right)$$

2.18. Remark: A test of significance neither accepts nor rejects a null hypothesis; it simply provides a measure of the evidence against it. The decision taken in light of this evidence is the choice of the researcher. Also, the fact that the evidence against the null hypothesis is statistically significant based on some discrepancy measure does not imply that the discrepancy is practically significant.

2.19. Next, we investigate a special discrepancy measure that looks like a *t-test*. When comparing two sub-populations on the basis of a measure of location, one particularly useful discrepancy measure is

$$D(\mathcal{P}_1, \mathcal{P}_2) = \frac{a(\mathcal{P}_1) - a(\mathcal{P}_2)}{\mathrm{SD}\big[a(\mathcal{P}_1) - a(\mathcal{P}_2)\big]} = \frac{\text{difference in attribute values}}{\text{standard deviation of difference}}.$$

This discrepancy measure is "physically dimensionless", i.e., whatever scale the numerator is measured in, the scale of the denominator will match, leaving the ratio free of any measurement scale. This naturally makes this discrepancy measure *scale-invariant*.

2.20. (Cont'd): The numerator is merely the difference in attribute values; the real challenge is determining the denominator of the discrepancy measure. In rare cases, the denominator might be known and then this discrepancy measure is a rescaling of denominator $a(\mathcal{P}_1) - a(\mathcal{P}_2)$ and would not yield different results. However, more commonly, we will need to estimate the denominator using information from $\mathcal{P}_1$ and $\mathcal{P}_2$.

2.21. (Cont'd): Suppose that the population $\mathcal{P}_1$ and $\mathcal{P}_2$ are drawn randomly and independently from the same larger population. Then the denominator should be

$$\mathrm{SD}\left[\tilde{a}(\mathcal{P}_1) - \tilde{a}(\mathcal{P}_2)\right] = \sqrt{\mathrm{Var}[\tilde{a}(\mathcal{P}_1) + \tilde{a}(\mathcal{P}_2)]} = \sqrt{\mathrm{Var}[\tilde{a}(\mathcal{P}_1)] + \mathrm{Var}[\tilde{a}(\mathcal{P}_2)]}$$

However, determining the form of $\mathrm{SD}[\tilde{a}(\mathcal{P}_1) - \tilde{a}(\mathcal{P}_2)]$ can also be difficult, except in the common special case when $a(\mathcal{P})$ is an average. Let us explore this now.

2.22. Note: Suppose we are interested in differences in *averages*, i.e., the attribute of interest is $a(\mathcal{P}_i) = \overline{Y}_i$, with $|\mathcal{P}_i| = N_i$ for $i = 1, 2$. The discrepancy measure becomes

$$D\left(\mathcal{P}_1, \mathcal{P}_2\right) = \frac{\overline{Y_1} - \overline{Y_2}}{\sqrt{\frac{\tilde{\sigma}^2}{N_1} + \frac{\sigma^2}{N_2}}},$$

where $\tilde{\sigma}$ is an estimator of the standard deviation of the Y values in the population $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2\}$. If $\tilde{\sigma}_1$ and $\tilde{\sigma}_2$ denote the estimators of the standard deviations from each of $\mathcal{P}_1$ and $\mathcal{P}_2$ respectively, then the pooled estimator of σ would be

$$\widetilde{\sigma} = \left(\frac{(N_1 - 1)\widetilde{\sigma}_1^2 + (N_2 - 1)\widetilde{\sigma}_2^2}{(N_1 - 1) + (N_2 - 1)}\right)^{\frac{1}{2}}.$$

If it were inappropriate to assume the variability in the two sub-populations was equivalent, we could instead use the denominator

$$\sqrt{\frac{\widetilde{\sigma}_1^2}{N_1} + \frac{\widetilde{\sigma}_2^2}{N_2}}.$$

This is the "two-sample" Student-t statistic used to test the equality of the means of two normal distributions with common but unknown standard deviation σ. If the Y values were in fact normally distributed, the discrepancy measure would follow a Student-t distribution

with $N_1 + N_2 - 2$ degrees of freedom under H_0 : the means were identical.

2.23. Remark: Note, however, in our procedure of randomly mixing the populations, we make no such normality assumption. We simply proceed with this discrepancy measure just as we did with the earlier measures. The only difference is that now we need to first calculate the denominator (the standard error).

Section 3. Interval Estimation

2.24. Motivation: Let us revisit sampling distributions first. Recall that when we look at $a(\mathcal{S})$ for all possible samples $\mathcal{S}$ of some size n from a population $\mathcal{P}$ that the values of $a(\mathcal{S})$ have a distribution, we call this the **sampling distribution** of $\tilde{a}(\mathcal{S})$. Quantifying the spread of this distribution is useful, but to do so exactly we require having observed all possible samples. Alternatively, we could approximate the sampling distribution in the following ways:

- consider a large number of possible samples;
- assume it takes a normal distribution;
- use resampling techniques such as bootstrap.

The normal distribution that best approximate the sampling distribution is the one with mean and standard deviation equal to the mean. To summarize, the normal approximation provides a model for the sampling distribution and can be used as a basis to construct confidence intervals for population averages.

2.25. In the rest of this section, we see how to construct a confidence interval.

2.26. Note: Suppose the attribute of interest is the population average $a(\mathcal{P}) = \bar{y}$. Recall that the estimator $\tilde{a}(\mathcal{P}) = \tilde{\mu} = \overline{Y}$ (which is a random variable) has the following properties:

$$\mathbb{E}[\overline{Y}] = \mu, \quad \text{Var}[\overline{Y}] = \frac{\sigma^2}{n},$$

where $\sigma^2 = \frac{1}{N} \cdot \sum_{\mu \in \mathcal{P}} (y_u - \mu)^2$ is the population variance. If the normality assumption holds (may be appropriate due to CLT), then the estimator $\tilde{a}(\mathcal{P}) = \overline{Y}$ satisfies

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

Standardizing this random variable yields

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

Using Z and a specified $p \in (0, 1)$, we can find a constant $c > 0$ such that

$$1 - p = \Pr(-c \leq Z \leq c).$$

Plugging in and rearranging, we obtain a **random interval** which contains μ with probability $1 - p$:

$$\left[\bar{Y} - c \times \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}, \bar{Y} + c \times \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right]$$

This interval is random because it is defined in terms of random variables and is never actually observed.

2.27. (Cont'd): However, observed intervals calculated by substituting $\overline{Y}$ with $\overline{y}$ are guaranteed to contain μ for $100(1-p)\%$ of the time.

- $1-p$ is therefore called the **coverage probability**.
- μ is *contained in* (or *covered by*) such an interval $100(1-p)\%$ of the time.
- All these intervals have the same width, just at different centers.

2.28. Remark (Determining c)**:** Since the normal distribution is symmetric about its mean μ, we see that

$$1-p = \Pr(-c \leq Z \leq c) \iff 1 - \frac{p}{2} = \Pr(Z \leq c).$$

Therefore, given any $p \in (0,1)$, the value of c can be determined through the quantile function of a standard normal random variable

$$c = Q_Z\left(1 - \frac{p}{2}\right),$$

or in R: `qnorm(1 - p/2)`.

2.29. Note: In practice, we will have only one sample, so there is only one single numerical average $\overline{y}$ and one instance of these randomly generated intervals:

$$\left[\bar{Y} - c \times \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}, \bar{Y} + c \times \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right].$$

In particular, we observe the following one:

$$\left[\bar{y} - c \times \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}, \bar{y} + c \times \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right].$$

Such observed intervals are referred to as **confidence intervals** and you must take care to distinguish them conceptually from a *random interval*:

- In the context of random intervals, probabilistic statements are sensible.
- As observed confidence intervals is not random, it either contains μ or it does not.
- Probability statements are in reference to the method used to generated the intervals, NOT to the particular interval we have observed.

If the normality assumption holds up, then $100(1-p)\%$ of such intervals will contain μ. We thus have some *confidence* that our particular observed interval will contain μ as well, but unfortunately we'll never know if it does. The larger $1-p$, the *more* confident we are that the interval will contain μ.

2.30. Next, we see how to construct a confidence interval using a t-distribution as an approximation to the sampling distribution.

2.31. Note: In the previous section, the CI was calculated assuming $\mathrm{SD}(\overline{Y})$ was known.

This is often unrealistic. However, for many sample attributes $a(\mathcal{S})$ (e.g., HT estimators), we can estimate the standard deviation $\mathrm{SD}(\tilde{a}(\mathcal{S}))$ of the sampling distribution of $\tilde{a}(\mathcal{S})$.

2.32. (Cont'd): The *standard error* is an estimate of the standard deviation of the corresponding estimator:

$$SE[a(\mathcal{S})] = \widehat{SD}[\tilde{a}(\mathcal{S})].$$

Note that the standard error is based on one sample; its corresponding estimator with a sampling distribution is denoted by

$$\widetilde{SE}[a(\mathcal{S})] \quad \text{or} \quad \widetilde{SD}[\tilde{a}(\mathcal{S})].$$

We could use this estimator instead in the test statistic:

$$\frac{a(\mathcal{S}) - a(\mathcal{P})}{SE[a(\mathcal{S})]}.$$

Note that using the estimated SE in place of SD will increase the variability of the random intervals. The corresponding estimator has much more variability, since we have to estimate SD now as well.

2.33. Remark: Under the normality assumption, we have the following distributional result:

$$\frac{\overline{Y} - \mu}{\tilde{\sigma}/\sqrt{n}} \sim t_{n-1}$$

This statistic is known as a **pivotal quantity**, because it is a function of the sample data $Y_u, u \in \mathcal{S}$ and unknown parameter μ and its sampling distribution is completely unknown.

Pivotal quantities are the basis for constructing random intervals. The term *pivot* comes from the fact that with this quantity (which is a function of both $\mathcal{S}$ and $\mathcal{P}$, we are able to pivot and isolate for $a(\mathcal{P})$. This is the general prescription for constructing random intervals.

2.34. Remark (Connection to Hypothesis Testing)**:** Note that there is a 1:1 correspondence between confidence intervals and hypothesis tests. Thus, if one wished to test a hypothesis, e.g., $H_0 : a(\mathcal{P}) = a$, about a population attribute, they could do so with an appropriately defined confidence interval.

Indeed, suppose that there exists a threshold below which a p-value would be sufficiently small so as to disbelieve the null hypothesis. This is often referred to as the **significance level** of the test. If the same pivotal quantity is used as both a discrepancy measure for the test and the basis for a confidence interval, the $100(1-p)\%$ confidence interval $a(\mathcal{P})$ contains all values of a for which the test of $H_0 : a(\mathcal{P}) = a$ would yield a p-value $> p$. Stated more usefully, do not reject $H_0 : a(\mathcal{P}) = a$ at a $100p\%$ significance level iff a is contained in the $100(1-p)\%$ confidence interval for $a(\mathcal{P})$.

Section 4. Resampling

2.35. Motivation: As shown in the previous sections, understanding the sampling behaviour of sample attributes is essential for making inferences about any population attribute. In particular, knowing the sampling distribution of a *discrepancy measure* allows us to do *hypotheses testing* and knowing the sampling distribution of a *pivotal quantity* allows us to *construct confidence intervals*.

2.36. (Cont'd): Let's review what we've learned so far. To perform inference, we draw a sample $\mathcal{S}$ of size n from a study population $\mathcal{P}$ according to some sampling mechanism, then calculate the sample attribute $a(\mathcal{S})$ to estimate its population counterpart $a(\mathcal{P})$. To understand the sampling distribution of the attribute $a(\mathcal{S})$, we draw M samples $S_1, \ldots, S_M$, and use the values $a(\mathcal{S}_1), \ldots, a(\mathcal{S}_M)$ to inform us about the sampling distribution of $a(\mathcal{S})$. Observe that this procedure requires repeated sampling from the population, but we often only have just one sample in practice. To mimic the process, we use *resampling methods*.

Resampling Intuition

2.37. Note (Resampling)**:** The goal is to mimic the repeated sampling process, i.e., obtain samples a set of samples as if they were from the population $\mathcal{P}$. In particular, we draw B samples $\mathcal{S}_1^*, \ldots, \mathcal{S}_B^*$ of size n independently from a population $\mathcal{P}^*$. Ideally, $\mathcal{P}^*$ would be the study population $\mathcal{P}$, but as already mentioned, this would require repeated sampling from the population which is generally impossible.

2.38. (Cont'd): To fix this issue, recall that a sample $\mathcal{S}$ can be viewed as an approximation to the population $\mathcal{P}$, i.e., $\widehat{\mathcal{P}} = \mathcal{S}$, or in usual *bootstrap notation*, $\mathcal{P}^* = \mathcal{S}$. Since the sample population has only n units, using without-replacement sampling mechanisms will immediately exhaust the population. Therefore, we sample *with replacement*.

The Bootstrap Method

2.39. Note (Bootstrap)**:** More formally, an **approximate sampling distribution** is obtained by drawing B samples $\mathcal{S}_1^*, \ldots, \mathcal{S}_B^*$ of size n from $\mathcal{P}^*$ *with replacement* and on each **bootstrap sample** we calculate the attribute value $a(\mathcal{S}_1^*), \ldots, a(\mathcal{S}_B^*)$. This approach is known as the **bootstrap** method; the distribution for any attribute over the bootstrap samples $\mathcal{S}_i^*$ from $\mathcal{P}^*$ is a **bootstrap estimate** of the distribution of the same attribute over all possible samples $\mathcal{S}_i$ from $\mathcal{P}$. With a single sample, we are now able to construct an estimate of the sampling distribution of an attribute that does not rely on assumptions about the form of the attribute.

2.40. Note (Bootstrap Sample Error)**:** The **bootstrap sample error** is given by

$$a(\mathcal{S}^*) - a(\mathcal{S}).$$

2.41. Note (Bootstrap Standard Deviation)**:** For any attribute $a(\mathcal{P})$, the standard deviation of the corresponding sample attribute's estimator can be estimated from the bootstrap distribution with

$$\widehat{SD}_\star\left[\tilde{a}\left(\mathcal{S}^\star\right)\right] = \sqrt{\frac{\sum_{b=1}^{B}\left(a\left(\mathcal{S}_b^\star\right)-\bar{a}^\star\right)^2}{B-1}},$$

where

$$\bar{a}^\star = \frac{1}{B}\sum_{b=1}^{B} a\left(\mathcal{S}_b^\star\right)$$

is the average of the attribute over the bootstrap samples $\mathcal{S}_1^*, \ldots, \mathcal{S}_B^*$. Since B is usually large, it does not make any practical difference whether we use B or $B-1$ in the denominator of the standard deviation. This is an estimate of the standard deviation of the sampling distribution for the attribute $a(\mathcal{S})$ which is called the **standard error**.

2.42. Note: In the special case of the arithmetic average

$$a(\mathcal{S}) = \frac{1}{n}\sum_{u\in\mathcal{S}} y_u,$$

the bootstrap estimate of its standard deviation is

$$\widehat{SD}_\star[\bar{Y}] = \sqrt{\frac{\sum_{b=1}^{B}\left(\bar{y}_b^\star-\bar{y}^\star\right)^2}{B-1}} \quad \text{where} \quad \bar{y}^\star = \frac{1}{B}\sum_{b=1}^{B}\bar{y}_b^\star.$$

But we also know that the standard deviation can be estimated with

$$\widehat{SD}[\bar{Y}] = \frac{\widehat{\sigma}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \quad \text{where} \quad \widehat{\sigma} = \sqrt{\frac{\sum_{u\in\mathcal{S}}\left(y_u-\bar{y}\right)^2}{n}}.$$

With some experiments, we see that the bootstrap estimator of standard deviation has produced, on average, slightly larger values that the standard approach.

2.43. Remark: For bootstrap interval calculations, a divisor of n is preferred (since we are treating the sample as a population) as this version is *replication invariant*. Replication invariant estimates are preferred and often called **plug in estimates** in the bootstrap literature. Anyway, when n is reasonably large, there will be little practical difference between the two.

2.44. Note: Recall that the sampling bias is given by $\mathbb{E}[a(\mathcal{S})] - a(\mathcal{P})$. We can use bootstrap to estimate sampling bias via:

$$\widehat{\text{Sampling Bias}} \;=\; \text{average bootstrap sample error} \;= \bar{a}^\star - a(\mathcal{S}),$$

where $\bar{a}^\star = \frac{1}{B}\sum_{b=1}^{B} a\left(\mathcal{S}_b^\star\right)$ is the average of the attribute over the bootstrap samples $\mathcal{S}_1^\star, \ldots, \mathcal{S}_B^\star$.

2.45. (Cont'd): Whenever an estimator is biased, we could like to "correct" it, i.e., make it unbiased. If, in theory, $a(\mathcal{S})$ was biased and we knew the bias, then we could subtract the correction from our attribute to make a new attribute $a^*(\mathcal{S})$ that is unbiased:

$$a^\star(\mathcal{S}) = a(\mathcal{S}) - \text{ Sampling Bias }.$$

Indeed,

$$\mathbb{E}\left[a^\star(\mathcal{S})\right] = \mathbb{E}[a(\mathcal{S}) - \text{Sampling Bias}] = \mathbb{E}[a(\mathcal{S}) - \mathbb{E}[a(\mathcal{S})] + a(\mathcal{P})] = a(\mathcal{P}).$$

We don't typically know the sampling bias, but we can use the bootstrap estimate of it:

$$a(\mathcal{S}) - \text{Sampling Bias}[a(\mathcal{S})] = a(\mathcal{S}) - [\bar{a}^\star - a(\mathcal{S})] = 2a(\mathcal{S}) - \bar{a}^\star.$$

2.46. Approximate confidence intervals and hypothesis tests can now also be based on the bootstrap estimate of a sampling distribution. We will explore bootstrap-based confidence interval next; bootstrap-based hypothesis testing is not covered in this course.

Bootstrap Confidence Intervals

2.47. The bootstrap distribution provides a proxy for the sampling distribution for any sample attribute $a(\mathcal{S})$. Thus, we can use it to construct (or at least approximate) confidence intervals for the unknown population attribute $a(\mathcal{P})$.

2.48. Note (Normal Bootstrap Interval)**:** Recall that confidence intervals for sample averages tend to have the following structure:

$$[\overline{y} - c \cdot \widehat{\text{SD}}(\overline{Y}), \overline{y} + c \cdot \widehat{\text{SD}}(\overline{Y})].$$

Under the normality assumption, we might pick c such that $\Pr(Z \leq c) = 1 - p/2$ to generate a $100(1-p)\%$ confidence interval.

2.49. (Cont'd): If the bootstrap distribution is approximately normal, rather than estimating $\widehat{\text{SD}}(\overline{Y})$ by $\frac{\widehat{\sigma}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$, we might estimate $\widehat{\text{SD}}(\overline{Y})$ using the standard deviation of the bootstrap distribution of $\overline{Y}$. The attraction of this approach, if it works, is that the same approach could be used for any attribute $a(\mathcal{S})$.

2.50. (Cont'd): To summarize, we can construct a 95% **normal bootstrap interval** for a population attribute $a(\mathcal{P})$ as

$$a(\mathcal{S}) \pm 1.96 \cdot \widehat{\text{SD}_*}[a(\mathcal{S})]$$

where $\widehat{\text{SD}_*}$ is the bootstrap estimate of the standard deviation.

2.51. Next, we use bootstrap to approximate the sampling distribution of a pivotal quantity and use this approximation to construct a CI, which turns out to be similar to using the t-distribution to approximate the sampling distribution of a pivotal quantity.

2.52. Note (Bootstrap-t Confidence Intervals)**:** When $a(\mathcal{S}) = \overline{y}$, we have seen that the quantity

$$Z = \frac{\tilde{a}(\mathcal{S}) - a(\mathcal{P})}{\widetilde{SD}[\tilde{a}(\mathcal{S})]}$$

is approximately pivotal and its sampling distribution (over all possible samples) is well approximated by a t-density.

2.53. (Cont'd): To confirm this approximation, consider the simulation below. Start by generating $\mathcal{S}_1, \ldots, \mathcal{S}_M$, then for each sample calculate

$$Z_i = \frac{a\left(\mathcal{S}_i\right) - a(\mathcal{P})}{\widehat{SE}\left[a\left(\mathcal{S}_i\right)\right]} = \frac{a\left(\mathcal{S}_i\right) - a(\mathcal{P})}{\widehat{SD}\left[\tilde{a}\left(\mathcal{S}_i\right)\right]}.$$

An estimate of t_{n-1} can be constructed with $\{Z_1, \ldots, Z_M\}$. This means that we require an estimate of the standard deviation of the estimator of a standard error.

Because this follows a t_{n-1} density, for $p \in (0, 1)$, we can find t_{lower} and t_{upper} such that

$$\begin{aligned} 1 - p &= \Pr\left(t_{\text{lower}} \ \leq \frac{\tilde{a}(\mathcal{S}) - a(\mathcal{P})}{\widetilde{SD}[\tilde{a}(\mathcal{S})]} \leq t_{\text{upper}} \ \right) \\ &= \Pr\left(\tilde{a}(\mathcal{S}) - t_{\text{upper}} \ \times \widetilde{SD}[\tilde{a}(\mathcal{S})] \leq a(\mathcal{P}) \leq \tilde{a}(\mathcal{S}) - t_{\text{lower}} \ \times \widetilde{SD}[\tilde{a}(\mathcal{S})] \right) \end{aligned}$$

This suggests a confidence interval can be constructed with

$$\left[a(\mathcal{S}) - t_{\text{upper}} \ \times \widehat{SD}[\tilde{a}(\mathcal{S})], a(\mathcal{S}) - t_{\text{lower}} \ \times \widehat{SD}[\tilde{a}(\mathcal{S})] \right].$$

Since the t-distribution is symmetric, we have that

$$t_{\text{upper}} \ = -1 \times t_{\text{lower}} \ = c$$

which gives us the following confidence interval:

$$a(\mathcal{S}) \pm c \times \widehat{SD}[\tilde{a}(\mathcal{S})].$$

2.54. (Cont'd): The t-distribution approximation of a sampling distribution may work for certain attributes, but not all of them, as it requires $\tilde{a}(\mathcal{S})$ to be approximately over all possible samples. For example, if $a(\mathcal{P})$ is the median or a measure of skewness, we would not expect the t-distribution to be a good approximation.

2.55. Note (Bootstrap-F Confidence Intervals)**:** Instead of approximating the sampling distribution with a t-distribution, we suppose that

$$Z = \frac{\tilde{a}(\mathcal{S}) - a(\mathcal{P})}{\widetilde{SD}[\tilde{a}(\mathcal{S})]} \sim F$$

, which motivates the following CI:

$$\left[a(\mathcal{S}) - f_{\text{upper}} \times \widehat{SD}[\tilde{a}(\mathcal{S})], a(\mathcal{S}) - f_{\text{lower}} \times \widehat{SD}[\tilde{a}(\mathcal{S})]\right]$$

2.56. (Cont'd): We can estimate F by the following simulation. Start by generating $\mathcal{S}_1, \ldots, \mathcal{S}_M$, then for each sample, we calculate

$$Z_i = \frac{a\left(\mathcal{S}_i\right) - a(\mathcal{P})}{\widehat{SD}\left[\tilde{a}\left(\mathcal{S}_i\right)\right]}$$

and estimate F using $\{Z_1, \ldots, Z_M\}$. Again, this means that we require an estimate of the standard deviation of the estimator of a standard error.

2.57. Note (Bootstrap)**:** CI] We have seen that the bootstrap may be used to approximate the sampling distribution of $\tilde{a}(\mathcal{S})$. Here, we will use bootstrap to estimate the sampling distribution of

$$Z = \frac{\tilde{a}(\mathcal{S}) - a(\mathcal{P})}{\widehat{SD}[\tilde{a}(\mathcal{S})]}$$

and use it to construct confidence intervals. On the negative side, this requires a lot of computation, but on the positive side, the bootstrap automatically adjusts its shape to the form of $\tilde{a}(\mathcal{S})$.

2.58. (Cont'd): To use bootstrap to approximate the sampling distribution of Z, we estimate the population $\mathcal{P}$ with the estimate $\mathcal{P}^* = \mathcal{S}$. Construct bootstrap sample $\mathcal{S}^*$, then generate $\mathcal{S}_1^*, \ldots, \mathcal{S}_B^*$ and calculate

$$Z_b^\star = \frac{\tilde{a}\left(\mathcal{S}_b^\star\right) - a(\mathcal{S})}{\widehat{SD}\left[\tilde{a}\left(\mathcal{S}_b^\star\right)\right]}$$

and then bootstrap estimate of the sampling distribution is $\{Z_1^*, \ldots, Z_B^*\}$. Then using a $p \in (0, 1)$, the bootstrap estimate we can find Z_{lower}^* and Z_{upper}^* such that

$$1 - p = \Pr\left(Z_{\text{lower}}^\star \leq Z^\star \leq Z_{\text{upper}}^\star\right) \approx \Pr\left(Z_{\text{lower}}^\star \leq Z \leq Z_{\text{upper}}^\star\right).$$

A CI using the bootstrap estimate of the pivotal quantity is

$$\left[a(\mathcal{S}) - Z_{\text{upper}}^\star \times \widehat{SD}[\tilde{a}(\mathcal{S})], a(\mathcal{S}) - Z_{\text{lower}}^\star \times \widehat{SD}[\tilde{a}(\mathcal{S})]\right].$$

Note that Z_{lower}^* and Z_{upper}^* are quantities from $\{Z_1^*, \ldots, Z_B^*\}$.

2.59. Note: We here summarize the general approach. For a given sample $\mathcal{S}$, attribute $a(\mathcal{S})$, and standard error $\widehat{\text{SD}}[\tilde{a}(\mathcal{S})]$. Calculate $a(\mathcal{S})$ and $\widehat{\text{SD}}[\tilde{a}(\mathcal{S})]$ and generate B bootstrap samples $\mathcal{S}_1^*, \ldots, \mathcal{S}_B^*$ from $\mathcal{S}$. For each of the B bootstrap samples from above, calculate $a(\mathcal{S}_b^*)$

and $\widehat{\text{SD}}[\tilde{a}(\mathcal{S}_b^*)]$ and then compute

$$z_b = \frac{a\left(\mathcal{S}_b^\star\right) - a(\mathcal{S})}{\widehat{SD}\left[\tilde{a}\left(\mathcal{S}_b^\star\right)\right]}.$$

From the values $z_1, \ldots, z_B$, find $c_{\text{lower}} = Q_z(p/2)$ and $c_{\text{upper}} = Q_z(1-p/2)$. Then a $100(1-p)\%$ bootstrap-t CI is $\left[a(\mathcal{S}) - c_{\text{upper}} \times \widehat{SD}[\tilde{a}(\mathcal{S})], a(\mathcal{S}) - c_{\text{lower}} \times \widehat{SD}[\tilde{a}(\mathcal{S})]\right]$. Note the signs and positions of the constants c_{lower} and c_{upper} in the interval definition.

2.60. Remark: So far, this method requires an analytic form ot calculate $\widehat{\text{SD}}[\tilde{a}(\mathcal{S})]$ (i.e., the standard deviation of the estimator given a single sample). Another interval can be constructed using the bootstrap estimate of the standard error $\widehat{\text{SD}}_*[\tilde{a}(\mathcal{S})]$. When this quantity is used, this approach is called the *double bootstrap* which we will look at next.

Double Bootstrap

2.61. Motivation: We saw that we can use the bootstrap method to approximate the sampling distribution of a pivotal quantity, which can then be used to construct a confidence interval. However, this requires an estimate of the standard error of an attribute. Here, we explore a procedure where we use the bootstrap to obtain an estimate of the standard error.

2.62. Note: When $a(\mathcal{S}) = \overline{y}$, we have an analytic form for its standard deviation:

$$SD[\bar{Y}] = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}.$$

Replacing σ by $\hat{\sigma}$ gives an estimate $\widehat{SD}[\overline{Y}]$ based on the sample values y_u for $u \in \mathcal{S}$. More generally, when $a(\mathcal{S})$ is an HT estimate, we also have an analytic form for $\widehat{SD}[\tilde{a}(\mathcal{S})]$. However, very often an analytic solution is not available. In this case, an estimate can be obtained by using bootstrap, by generating $\mathcal{S}_1^*, \ldots, \mathcal{S}_B^*$ and calculate

$$\widehat{SD}_\star[\tilde{a}(\mathcal{S})] = \sqrt{\frac{\sum_{b=1}^B \left(a\left(\mathcal{S}_b^\star\right) - \bar{a}^\star\right)^2}{B-1}}$$

where $\bar{a}^\star = \frac{1}{B}\sum_{b=1}^B a\left(\mathcal{S}_b^\star\right)$.

The next problem is that for bootstrap-t, we need an estimate of $SD[\tilde{a}(\mathcal{S}_b^*)]$ for each bootstrap sample $\mathcal{S}_b^*$. To address this, we use the **double bootstrap**, i.e., apply bootstrap to each bootstrap sample $\mathcal{S}_b^*$.

2.63. (Cont'd): To apply a bootstrap within a bootstrap for each bootstrap sample $\mathcal{S}_b^*$, we generate D bootstrap samples, $\mathcal{S}_1^{**}, \ldots, \mathcal{S}_D^{**}$, each with replacement from a population now defined as $\mathcal{P}^{**} = \mathcal{S}_b^*$. The standard deviation of the corresponding values $a(\mathcal{S}_1^{**}), \ldots, a(\mathcal{S}_D^{**})$ will provide the estimate $\widehat{SD}_*[a(\mathcal{S}_b^*)]$. This estimate is then substituted into the general approach for bootstrap-t confidence intervals.

The Percentile Method

2.64. Motivation: The sampling distribution of $\tilde{a}(\mathcal{S})$ can be estimated using a sample $\mathcal{S}$ and bootstrap, so why not simply use quantiles from the bootstrap distribution to directly construct a confidence interval?

2.65. Note: The **percentile method** for bootstrap CI is described below. For a given sample $\mathcal{S}$, generate B bootstrap samples $\mathcal{S}_1^*, \ldots, \mathcal{S}_B^*$ by sampling with replacement from sample $\mathcal{S}$. For the bth bootstrap sample, calculate $a_b = a(\mathcal{S}_b^*)$. From the values $a_1, \ldots, a_B$, find $a_{\text{lower}} = Q_a(p/2)$ and $a_{\text{upper}} = Q_a(1 - p/2)$. Then the $100(1-p)\%$ CI is $[a_{\text{lower}}, a_{\text{upper}}]$.

2.66. Note: This approach is *equivariant* to any 1:1 transformation of the attribute, say $T(a(\mathcal{P}))$. For an increasing function $T(\cdot)$, the corresponding interval for $T(a(\mathcal{P}))$ is simply $[T(a_{\text{lower}}), T(a_{\text{upper}})]$. For a decreasing function $T(\cdot)$, the corresponding interval is $[T(a_{\text{upper}}), T(a_{\text{lower}})]$. Hence, we only need to determine the values a_{lower} and a_{upper} once for $a(\mathcal{P})$ and we have them for any $T(a(\mathcal{P}))$.

2.67. Remark: Simplicity and transformation equivalent are the main attraction of this method. However, the coverage probability is often incorrect if the distribution of the estimator is not nearly symmetric.

Chapter 3. Prediction

3.1. Motivation: Oftentimes, interest lies in *predicting* the value of the **response variate** given the values of one or more **explanatory variates**. We build a **response model** that encodes how that prediction is to be carried out:

$$y = \mu(\mathbf{x}) + \varepsilon.$$

The explanatory variates $\mathbf{x} = (x_1, \ldots, x_p)$ are used to explain or predict the values of the response. We use our observed data to estimate the function $\mu(\mathbf{x})$, yielding the **predictor function** $\widehat{\mu}(\mathbf{x})$. This predictor function $\widehat{\mu}(\mathbf{x})$ is then used to predict y at any given value $\mathbf{x}$.

For example, for an SLR model, we assume the underlying model $\mu(x) = \alpha + \beta x$ and estimate the parameters of the function using least squares to obtain

$$\widehat{\mu}(x) = \widehat{\alpha} + \widehat{\beta} x.$$

But how do we know if our predictions are good? What metrics should we use?

Section 1. Prediction Accuracy

3.2. Motivation: One intuitive way of measuring prediction error is to see how *far* our prediction is from the true value. The **APSE** metric quantifies the *distance* between true and predicted values. In particular, it is the average squared distance between a response observation and its corresponding prediction across the population.

3.3. Definition (APSE)**:** Given observations $\mathbf{x}_u$ and true responses y_u for $u \in \mathcal{P}$, the **average prediction squared error** (APSE) is given by

$$\frac{1}{N}\sum_{u\in\mathcal{P}}(y_u - \widehat{\mu}(\mathbf{x}_u))^2.$$

3.4. Remark: Note that APSE is proportional the the **residual sum of squares**

$$\sum_{i=1}^{N}\hat{r}_i^2 = \sum_{i=1}^{N}\left(y_i - \widehat{\mu}\left(\mathbf{x}_i\right)\right)^2.$$

3.5. Note: So far, we estimate the predictor function and measure its accuracy using the same set of observations. Thus, we can write the measure as

$$\begin{aligned}\text{APSE}\left(\mathcal{P}, \widehat{\mu}_{\mathcal{P}}\right) &= \frac{1}{N}\sum_{u\in\mathcal{P}}\left(y_u - \widehat{\mu}\left(\mathbf{x}_u\right)\right)^2 \\ &= \frac{1}{N}\sum_{u\in\mathcal{P}}\left(y_u - \widehat{\mu}_{\mathcal{P}}\left(\mathbf{x}_u\right)\right)^2.\end{aligned}$$

The notation $\widehat{\mu}_{\mathcal{P}}(\mathbf{x}_u)$ here emphasizes the fact that the predictor function was determined from the entire population. However, this will underestimate the APSE for predictions at values of x not existing in the data, because the training set is exactly the same as the test set, inevitably leading to overfitting.

3.6. (Cont'd): Ideally, to provide a more fair evaluation of prediction accuracy, we would use different data to train and test the model. More precisely, we estimate the predictor function using a sample $\mathcal{S}$ (known as the **training set**) and measure the inaccuracy over the population $\mathcal{P}$, or over the units in the population not included in the sample,

$$\mathcal{T} := \mathcal{P} \setminus \mathcal{S},$$

sometimes called the **test set**. With the notion of train set introduced, we would write

$$\text{APSE}\left(\mathcal{P}, \widehat{\mu}_{\mathcal{S}}\right) = \frac{1}{N}\sum_{u\in\mathcal{P}}\left(y_u - \widehat{\mu}_{\mathcal{S}}\left(\mathbf{x}_u\right)\right)^2.$$

Note that the notation $\widehat{\mu}_{\mathcal{S}}$ emphasizes that the estimate of the predictor function $\widehat{\mu}$ is based on a sample $\mathcal{S}$, and we are evaluating it based on elements $u \in \mathcal{P}$.

3.7. (Cont'd): Since $\mathcal{P} = \mathcal{S} \cup \mathcal{T}$ and $\mathcal{S} \cap \mathcal{T} = \varnothing$, the APSE defined in the way above can be decomposed into a sum of two pieces:

$$\begin{aligned}
\text{APSE}\left(\mathcal{P}, \widehat{\mu}_{\mathcal{S}}\right) &= \frac{1}{N} \sum_{u \in \mathcal{P}} \left(y_u - \widehat{\mu}_{\mathcal{S}}\left(\mathbf{x}_u\right)\right)^2 \\
&= \underbrace{\left(\frac{n}{N}\right) \frac{1}{n} \sum_{u \in \mathcal{S}} \left(y_u - \widehat{\mu}_{\mathcal{S}}\left(\mathbf{x}_u\right)\right)^2}_{\text{APSE from the train set } \mathcal{S}} + \underbrace{\left(\frac{N-n}{N}\right) \frac{1}{N-n} \sum_{u \in \mathcal{T}} \left(y_u - \widehat{\mu}_{\mathcal{S}}\left(\mathbf{x}_u\right)\right)^2}_{\text{APSE from the test set } \mathcal{T}} \\
&= \left(\frac{n}{N}\right) \text{APSE}\left(\mathcal{S}, \widehat{\mu}_{\mathcal{S}}\right) + \left(\frac{N-n}{N}\right) \text{APSE}\left(\mathcal{T}, \widehat{\mu}_{\mathcal{S}}\right)
\end{aligned}$$

Given that interest often lies in the quality of the predictions outside of the sample (i.e., how well the model generalizes to unseen data), we might exclusively calculate the average prediction squared error over $\mathcal{T}$:

$$\text{APSE}\left(\mathcal{T}, \widehat{\mu}_{\mathcal{S}}\right) = \frac{1}{N-n} \sum_{u \in \mathcal{T}} \left(y_u - \widehat{\mu}_{\mathcal{S}}\left(\mathbf{x}_u\right)\right)^2$$

Clearly, if $n \ll N$, the value $\text{APSE}\left(\mathcal{T}, \widehat{\mu}_{\mathcal{S}}\right)$ will not be too different from $\text{APSE}\left(\mathcal{P}, \widehat{\mu}_{\mathcal{S}}\right)$.

3.8. Remark: Since $\widehat{\mu}_{\mathcal{S}}(\mathbf{x})$ is based on a single sample $\mathcal{S}$, the quality of the predictor function depends crucially on the quality of the sample. If the sample is not a good/fair representation of the population, then any predictor function is bound to perform poorly. In practice, we tend to assume our sample is a good representation of the population, but in case that's not true, it is important to choose a predictor function that performs well no matter which sample was used to estimate it. Simpler is often better.

Section 2. Prediction over Multiple Samples

3.9. Motivation: Previously, the performance of the estimated predictor function depends highly on the particular choice of sample and could vary quite a lot from one sample to another. Let us now try to estimate the predictor function using multiple samples.

3.10. Note: Suppose that we have many samples $\mathcal{S}_j$ for $j = 1, \ldots, N_{\mathcal{S}}$. For each sample S_j, we can come up with a model $\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x})$ and hence calculate $\text{APSE}(\mathcal{P}, \widehat{\mu}_{\mathcal{S}_j})$. The average APSE over all $N_{\mathcal{S}}$ samples should be a better measure of the quality of a predictor function:

$$\begin{aligned}\text{APSE}(\mathcal{P}, \tilde{\mu}) &= \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \text{APSE}\left(\mathcal{P}, \widehat{\mu}_{\mathcal{S}_j}\right) \\ &= \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \frac{1}{N} \sum_{u \in \mathcal{P}} \left(y_u - \widehat{\mu}_{\mathcal{S}_j}\left(\mathbf{x}_u\right)\right)^2\end{aligned}$$

Note that in $\text{APSE}(\mathcal{P}, \tilde{\mu})$, the *estimator notation* $\tilde{\mu}$ is used to emphasize that the APSE metric is evaluating $\widehat{\mu}$ over many samples $\mathcal{S}_j$.

3.11. Note: Let us try to decompose the above APSE into separable pieces, much like what we did in the previous section. Starting with

$$\overline{\widehat{\mu}}(\mathbf{x}) = \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}),$$

i.e., the average of the estimated predictor function $\widehat{\mu}$ over all $N_{\mathcal{S}}$ samples, we can write

$$\begin{aligned}&\text{APSE}(\mathcal{P}, \widetilde{\mu}) \\ &= \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \frac{1}{N} \sum_{u \in \mathcal{P}} (y_u - \widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u))^2 \\ &= \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \frac{1}{N} \sum_{u \in \mathcal{P}} (y_u - \mu(\mathbf{x}_u))^2 + \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \frac{1}{N} \sum_{u \in \mathcal{P}} (\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u) - \mu(\mathbf{x}_u))^2 \\ &= \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \frac{1}{N} \sum_{u \in \mathcal{P}} (y_u - \mu(\mathbf{x}_u))^2 + \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \frac{1}{N} \sum_{u \in \mathcal{P}} (\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u) - \overline{\widehat{\mu}}(\mathbf{x}_u))^2 + \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \frac{1}{N} \sum_{u \in \mathcal{P}} (\overline{\widehat{\mu}}(\mathbf{x}_u) - \mu(\mathbf{x}_u))^2 \\ &= \frac{1}{N} \sum_{u \in \mathcal{P}} (y_u - \mu(\mathbf{x}_u))^2 + \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \frac{1}{N} \sum_{u \in \mathcal{P}} (\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u) - \overline{\widehat{\mu}}(\mathbf{x}_u))^2 + \frac{1}{N} \sum_{u \in \mathcal{P}} (\overline{\widehat{\mu}}(\mathbf{x}_u) - \mu(\mathbf{x}_u))^2\end{aligned}$$

The second term reflects the variability of the estimator $\tilde{\mu}$ and the third term reflects the (squared) estimator's bias.

3.12. Note: To interpret the first term, define

$$\mathcal{A}_k = \{u : u \in \mathcal{P}, \mathbf{x}_u = \mathbf{x}_k\}.$$

to be the collection of units who all have the same value $\mathbf{x} = \mathbf{x}_k$. (Indeed, it is common to see many data points in the same dataset to have the same value.) Now rewrite the sum over $u \in \mathcal{P}$ as a sum of u over the *unique* $\mathbf{x}$ values:

$$\frac{1}{N}\sum_{u\in\mathcal{P}}(y_u - \mu(\mathbf{x}_u))^2 = \sum_{k=1}^{K}\frac{n_k}{N}\sum_{u\in\mathcal{A}_k}\frac{1}{n_k}(y_u - \mu(\mathbf{x}_k))^2.$$

Thus, the first term can be interpreted as the *conditional variance of y given $\mathbf{x}$, averaged over all of the unique $\mathbf{x}$ values.*

3.13. (Cont'd): To summarize, we have

$$\begin{aligned}
&\mathrm{APSE}(\mathcal{P}, \widetilde{\mu}) \\
&= \sum_{k=1}^{K}\frac{n_k}{N}\sum_{u\in\mathcal{A}_k}\frac{1}{n_k}(y_u - \mu(\mathbf{x}_k))^2 + \frac{1}{N_{\mathcal{S}}}\sum_{j=1}^{N_{\mathcal{S}}}\frac{1}{N}\sum_{u\in\mathcal{P}}(\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u) - \overline{\widehat{\mu}}(\mathbf{x}_u))^2 + \frac{1}{N}\sum_{u\in\mathcal{P}}(\overline{\widehat{\mu}}(\mathbf{x}_u) - \mu(\mathbf{x}_u))^2 \\
&\equiv \mathrm{Avg}_{\mathbf{x}}\left(\mathrm{Var}[Y|\mathbf{x}]\right) + \mathrm{Var}[\widetilde{\mu}] + \mathrm{Bias}^2[\widetilde{\mu}].
\end{aligned}$$

- The first term is the average of the conditional variance of the response y given $\mathbf{x}$.
- The second term is the variance of the estimator.
- The third term is the squared bias of the estimator.

3.14. Note: We can further decompose $\mathrm{APSE}(\mathcal{P}, \tilde{\mu})$ as

$$\begin{aligned}
\mathrm{APSE}(\mathcal{P}, \widetilde{\mu}) = \left(\frac{n}{N}\right)&\left\{\ \widehat{\mathrm{APSE}}(\mathcal{P}, \widetilde{\mu}) \text{ based on the } \mathbf{same} \text{ samples used by } \widehat{\mu}\ \right\} \\
&+ \left(\frac{N-n}{N}\right)\left\{\ \widehat{\mathrm{APSE}}(\mathcal{P}, \widetilde{\mu}) \text{ based on samples } \mathbf{not} \text{ used by } \widehat{\mu}\ \right\}
\end{aligned}$$

Notice that if $n \ll N$, then the second term dominates the overall APSE. However, regardless of the size of n, we may sometimes want to focus our evaluation only on the second term, since this evaluation is based only on values not used in the actual estimation process. This provides the most fair assessment of the model's out of sample performance.

Section 3. Back to Reality: Predictions with a Single Sample

3.15. Motivation: Predictive accuracy provides insights into the performance of a predictor function and can be used to choose between competing ones. The key to this usefulness, however, is that the predictive accuracy can be measured on population $\mathcal{P}$ about which we want to make inference. Unfortunately, we typically only have $\mathcal{S}$, the training set. What should we do?

3.16. Note: This is the basic problem of inductive inference. Experience says that whenever interest lies in some attribute of the population $a(\mathcal{P})$, we might use $a(\mathcal{S})$ as an estimate of that attribute. More specifically, we cast predictive accuracy as an attribute of population $\mathcal{P}$ and then use the corresponding attribute evaluated on $\mathcal{S}$ as its estimate. In particular, we care about the attribute

$$a_1(\mathcal{P}) = \text{APSE}\left(\mathcal{P}, \widehat{\mu}_{\mathcal{S}}\right) = \frac{1}{N}\sum_{u\in\mathcal{P}}\left(y_u - \widehat{\mu}_{\mathcal{S}}\left(\mathbf{x}_u\right)\right)^2.$$

We call it the **single subset version** of APSE as it only relies on the single sample $\mathcal{S}$.

3.17. (Cont'd): If we have multiple samples, we care about

$$a_2(\mathcal{P}) = APSE(\mathcal{P}, \widetilde{\mu}) = \frac{1}{N_{\mathcal{S}}}\sum_{j=1}^{N_{\mathcal{S}}} APSE\left(\mathcal{P}, \widehat{\mu}_{\mathcal{S}_j}\right),$$

known as the **multiple subset version** of APSE.

3.18. (Cont'd): These are two distinct population attributes, each a slightly different measure of an average prediction squared error. However, we are usually more concerned with how well each predictor function performs on the population that was not used to construct the estimate. With this in mind, we can define these notions as

$$a_3(\mathcal{P}) = \text{APSE}\left(\mathcal{T}, \widehat{\mu}_{\mathcal{S}}\right) = \frac{1}{N-n}\sum_{u\in\mathcal{T}}\left(y_u - \widehat{\mu}_{\mathcal{S}}\left(\mathbf{x}_u\right)\right)^2$$

$$a_4(\mathcal{P}) = \text{APSE}(\mathcal{T}, \widetilde{\mu}) = \frac{1}{N_{\mathcal{S}}}\sum_{j=1}^{N_{\mathcal{S}}} \text{APSE}\left(\mathcal{T}_j, \widehat{\mu}_{\mathcal{S}_j}\right).$$

3.19. Note: Let us first look at the single subset version. Suppose we were interested in estimating

$$\text{APSE}\left(\mathcal{T}, \widehat{\mu}_{\mathcal{S}}\right) = \frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}}\left(y_i - \widehat{\mu}_{\mathcal{S}}\left(\mathbf{x}_i\right)\right)^2.$$

Note that the predictor function is constructed using $\mathcal{S}$ and the prediction error is evaluated on $\mathcal{T} = \mathcal{P} \setminus \mathcal{S}$. If all we observed was the sample $\mathcal{S}$ from $\mathcal{P}$, we might approximate the single subset version of APSE by partitioning $\mathcal{S}$ into $\mathcal{S}_0$ and $\mathcal{T}_0$ and then use these as the training

and test set:

$$\widehat{\text{APSE}}\left(\mathcal{T}, \widehat{\mu}_{\mathcal{S}}\right) = \text{APSE}\left(\widehat{\mathcal{T}}, \widehat{\mu}_{\widehat{\mathcal{S}}}\right) = \text{APSE}\left(\mathcal{T}_0, \widehat{\mu}_{\mathcal{S}_0}\right) = \frac{1}{|\mathcal{T}_0|} \sum_{u \in \mathcal{T}_0} \left(y_u - \widehat{\mu}_{\mathcal{S}_0}\left(\mathbf{x}_u\right)\right)^2$$

$$\widehat{\text{APSE}}\left(\mathcal{P}, \widehat{\mu}_{\mathcal{S}}\right) = \text{APSE}\left(\widehat{\mathcal{P}}, \widehat{\mu}_{\widehat{\mathcal{S}}}\right) = \text{APSE}\left(\mathcal{P}_0, \widehat{\mu}_{\mathcal{S}_0}\right) = \frac{1}{|\mathcal{P}_0|} \sum_{u \in \mathcal{P}_0} \left(y_u - \widehat{\mu}_{\mathcal{S}_0}\left(\mathbf{x}_u\right)\right)^2$$

3.20. (Cont'd): In this setting, $\mathcal{S}_0$ is typically called the **training set**, $\mathcal{T}_0$ is typically called the **validation set** or **hold-out sample**, and performing such a partitioning is referred to as **cross validation**. Of course, the real challenge is how to pick $\mathcal{S}_0$ from $\mathcal{P}_0 = \mathcal{S}$.

3.21. Note: Now suppose we were interested in estimating the average performance over all $N_{\mathcal{S}}$ possible samples:

$$\text{APSE}(\mathcal{T}, \widetilde{\mu}) = \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \text{APSE}\left(\mathcal{T}_j, \widehat{\mu}_{\mathcal{S}_j}\right)$$

Again, we may use an observed sample $\mathcal{S}$ as an estimate of $\mathcal{P}$, then mimic taking many samples and sets from $\mathcal{P}$, i.e., define $(\mathcal{S}_{0,j}, \mathcal{T}_{0,j})$ where j denotes the jth sample from the set of $N_{\mathcal{S}}$ samples. We then estimate $\text{APSE}(\mathcal{T}, \tilde{\mu})$ by

$$\widehat{APSE}(\mathcal{T}, \tilde{\mu}) = \frac{1}{N_S} \sum_{j=1}^{N_S} APSE\left(\mathcal{T}_{0,j}, \widehat{\mu}_{\mathcal{S}_{0,j}}\right).$$

As with the single subset version, the key remains as to how to pick the subsets.

3.22. Note: Indeed, it is not always obvious how one should choose $\mathcal{S}_0$ and $\mathcal{T}_0$ in a given situation. One guide is that the method of selecting $\mathcal{S}_0$ from $\mathcal{P}_0 = \mathcal{S}$ should be as similar as possible to that of selecting the sample $\mathcal{S}$ from the study population $\mathcal{P}$; that is, the same sampling mechanism should be used. We now address the following concerns.

- Should be sampling be done with or without replacement?
- How large should the sample $\mathcal{S}_0$ be?
- Should $\mathcal{T}_0$ be the full complement of $\mathcal{S}_0$ or just a sample from $\mathcal{S} \setminus \mathcal{S}_0$? If the latter, how large should it be?

3.23. Note (Sampling Mechanism)**:** If predictive accuracy is meant to be an out-of-sample assessment, then we should restrict ourselves to sampling without replacement, which results in a clear distinction between the training and test set and reduces the possibility of overestimating the predictor's accuracy. On the other hand, sampling with replacement would require redefining APSE to account for duplicates in the sample, unless APSE was calculated using only out-of-sample units.

3.24. Note (Training Set Size)**:** We can gain insight into how large the training set

should be from the fact that the predicted squared errors are averaged. Recall that $\text{SD}(\overline{Y}) = \sigma/\sqrt{n}$. If the test set $\mathcal{T}_0$ contains $|\mathcal{T}_0|$ units, then the standard deviation of the APSE will decrease proportionately to $1/\sqrt{|\mathcal{T}_0|}$. In other words, the larger $|\mathcal{T}_0|$ is, the better (i.e., less variable) will be our estimate of the APSE. However, the larger $|\mathcal{T}_0|$ is, the smaller $\mathcal{S}_0$ will be, so the estimated predictor function will have lower quality. Thus, choosing a sample size requires some trade-off between the variability and the bias of the estimate predictor function.

3.25. The above questions also apply to the multiple subset version, along with additional consideration:

- How many samples $\mathcal{S}_j$ should we take?
- how do we ensure that every unit the sample is selected in the training and test set?

3.26. Note: A simple way to create a collection of samples $\mathcal{S}_j$ is to partition $\mathcal{P}_0$ into pieces or groups then selected some groups to form $\mathcal{S}_{0,j}$ and the remainder to form $\mathcal{T}_{0,j}$. Typically, $\mathcal{P}_0$ is partitioned into k groups $G_1, \ldots, G_k$ of equal size, called a **k-fold partition** of $\mathcal{P}_0$. The most common defined method of selecting the groups would be to select $k-1$ groups to form $\mathcal{S}_{0,j}$ and the remaining group form $\mathcal{T}_{0,j}$. Calculating

$$\widehat{\text{APSE}}(\mathcal{T}, \widetilde{\mu}) = \frac{1}{N_S} \sum_{j=1}^{N_S} \text{APSE}\left(\mathcal{T}_{0,j}, \widehat{\mu}_{\mathcal{S}_{0,j}}\right)$$

using sampling that selects all $k-1$ groups from a k-fold partition is called **k-fold cross-validation**.