SAVANNAH math labs

Lab Module 10: Statistical Bias

Learning Objectives

- Define statistical bias with respect to a parameter (θ) and a point estimate ($\hat{\theta}$) derived from a sample.
- Consider different approaches for obtaining a random sample that is representative of the population.
- Quantify the magnitude of statistical bias with respect to a known parameter (θ) and a point estimate ($\hat{\theta}$) for a range of data sources.

Key Concepts

Statistical Bias

Nothing throws a wrench into the best-laid statistical plans like **statistical bias**. Recall that your friend Joanna was running for class president. As Joanna's campaign manager, you conducted a poll to assess her current level of support relative to three other candidates: Eli, Alexa and Jose. Instead of obtaining a simple random sample, you unwittingly polled a substantial proportion of students at the library on a Friday night where Joanna spends much of her time. The results of your poll were as follows:

Candidate	Number of Votes (x)	Relative Frequency ($\frac{x}{N}$)	Election Results
Joanna	255	0.510	0.28
Eli	154	0.308	0.51
Alexa	72	0.144	0.15
Jose	19	0.038	0.06
Total	**500**	**1**	**1**

It appeared from the poll that Joanna was set to win the election by a landslide. However, on Election Day, Eli won with 51% of the total vote, nearly twice the level of support for Joanna! Your poll was statistically biased and the results were very misleading.

Definition of Statistical Bias

The large discrepancies between your poll and the election results cannot be explained by random sampling fluctuations alone. Your poll was statistically biased, but how is statistical bias precisely defined? First, define the parameter theta (θ), the true proportion of votes in the population that Joanna will receive on Election Day. Next, define the point estimate ($\hat{\theta}$) of θ, the relative frequency of votes for Joanna ($\hat{\theta} = \frac{x}{N}$) in a poll of N individuals. Statistical bias is defined as follows:

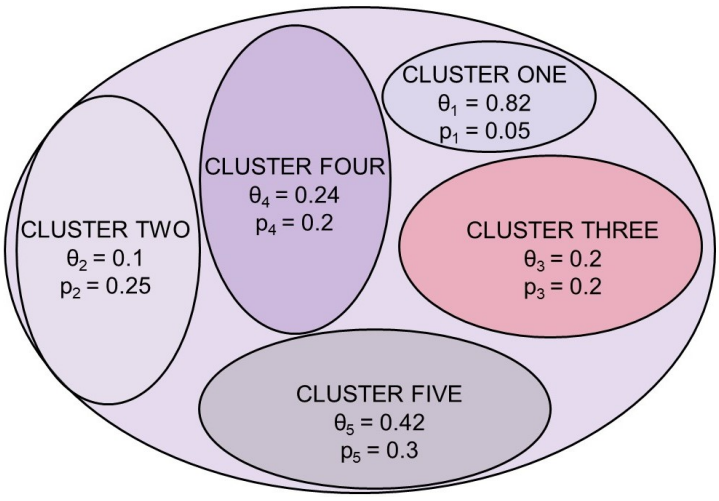
$$\text{Statistical Bias} = E(\hat{\theta}) - \theta$$

Statistical bias is non-zero when $E(\hat{\theta})$ does not equal θ and a data source (i.e. poll) *systematically* overestimates or underestimates the parameter θ among repeated samples. The magnitude of statistical bias with respect to θ equals $\left|E(\hat{\theta}) - \theta\right|$ and may be large or small.

Sources of Statistical Bias in a Poll or Sample

Why does statistical bias occur? Suppose the population of students at Joanna's school is comprised of sub-populations, or **clusters** of individuals in which the level of support for Joanna differs from θ.

Population comprised of Sub-Populations ($\theta \neq \theta_1 \neq \theta_2 \neq \theta_3 \neq \theta_4 \neq \theta_5$)



Cluster One consists of the small sub-population ($p_1 = 0.05$) of Joanna's classmates who routinely spend time in the library on Friday night and overwhelmingly support

Joanna for class president (i.e. $\theta_1 = 0.82$).

Nevertheless, when a polling sample is **representative** of the population, $E(\hat{\theta})$ will equal θ, resulting in zero statistical bias. A representative sample implies that the sample proportion of Cluster One students approximately equals p_1, the sample proportion of Cluster Two students approximately equals p_2, etc. For a representative sample, $E(\hat{\theta})$ can be computed as follows:

$$E(\hat{\theta}) = p_1\theta_1 + p_2\theta_2 + p_3\theta_3 + p_4\theta_4 + p_5\theta_5 = \theta.$$

Joanna received 28% of the votes on Election Day, and therefore θ equals 0.28. A representative poll is expected, on average, to indicate that Joanna will receive 28% of all votes as well.

$$E(\hat{\theta}) = 0.05 \times 0.82 + 0.25 \times 0.1 + 0.2 \times 0.2 + 0.2 \times 0.24 + 0.3 \times 0.42 = 0.28 = \theta.$$

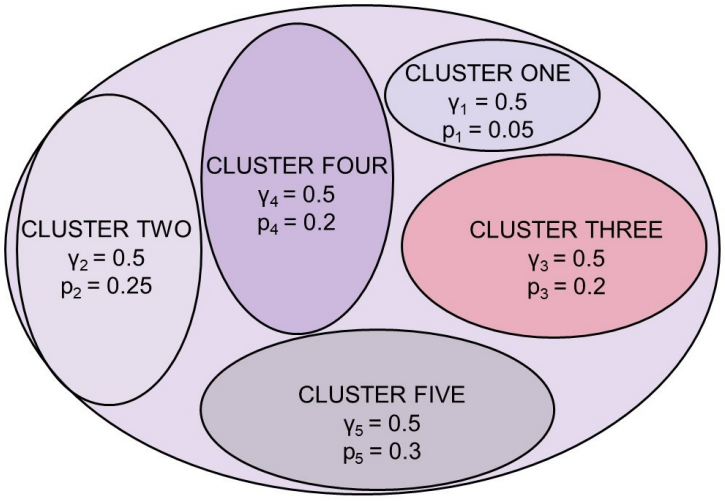
However, the **over-representation** of Joanna's friends (Cluster One) explains why the polling results were erroneously skewed in her favor. Suppose 46% of the polling sample consisted of Joanna's friends in Cluster One, whereas the sample proportion of Clusters Two, Three, Four and Five equaled 0.1, 0.2, 0.1 and 0.14, respectively. $E(\hat{\theta})$ for this non-representative sample can be computed as follows:

$$E(\hat{\theta}) = 0.46 \times 0.82 + 0.1 \times 0.1 + 0.2 \times 0.2 + 0.1 \times 0.24 + 0.14 \times 0.42 = 0.51.$$

The magnitude of statistical bias equals $|0.51 - 0.28|$, or 0.23. When a sub-population or cluster has a dissimilar value of theta (i.e. $\theta_1 \neq \theta$) and is **also** systematically over-represented (or under-represented) in the polling process, statistical bias often results ($E(\hat{\theta}) \neq \theta$).

When a sample is not representative of the population, statistical bias can be present for some variables, but not others. Define a second parameter (γ), the proportion of students at Joanna's school whose favorite ice cream flavor is Vanilla. Suppose the proportion of individuals whose favorite ice cream flavor is Vanilla approximately equals $\gamma = 0.5$ in each sub-population.

Population comprised of Sub-Populations ($\gamma = \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5$)



$E(\hat{\gamma})$ for a "representative" sample can be computed as follows:

$$E(\hat{\gamma}) = 0.05 \times 0.5 + 0.25 \times 0.5 + 0.2 \times 0.5 + 0.2 \times 0.5 + 0.3 \times 0.5 = 0.5 = \gamma.$$

$E(\hat{\gamma})$ for a "non-representative" sample can be computed as follows:

$$E(\hat{\gamma}) = 0.46 \times 0.5 + 0.1 \times 0.5 + 0.2 \times 0.5 + 0.1 \times 0.5 + 0.14 \times 0.5 = 0.5 = \gamma.$$

When a polling sample is representative of the student population, $E(\hat{\gamma})$ equals γ and the magnitude of statistical bias with respect to the parameter γ equals zero. However, even when Joanna's friends (Cluster One) are over-represented in the polling sample, $E(\hat{\gamma})$ still equals γ and the magnitude of statistical bias with respect to the parameter γ also equals zero.

Approaches to Sampling

Census Samples

A **census** is a sample that includes every individual in the population. A census represents the population exactly and statistical bias will equal zero for every defined parameter. Furthermore, statistical inference methods are extraneous for a census because random sampling fluctuations do not occur. Every census sample is identical! However, a census is challenging to obtain and often costly, especially for large populations. As a result, census samples are uncommonly used in statistical studies.

Random Sampling

Random sampling minimizes and may eliminate statistical bias entirely. To obtain a **simple random sample**, a method of random chance identifies individuals, ensuring that every possible sample of size N is equally likely to be chosen. A random number generator, available within most statistical programs, or even a simple table of random numbers can determine a simple random sample. **Systematic random sampling** is less optimal, whereby a random starting point is determined and individuals are subsequently chosen according to a fixed, periodic interval.

Stratified random sampling entails first stratifying the population into a set of homogenous groups (strata). Within each strata, a simple random sample is selected. Simple and stratified random samples are representative subsets of a larger population whereby sample composition will vary among repeated samples. Statistical inference methods can account for the uncertainty associated with these random sampling fluctuations.

Cluster sampling entails first stratifying the population into a set of heterogenous groups or clusters. A simple random sample of clusters is then selected and a census sample within each selected cluster is subsequently obtained.

Sources of Statistical Bias

Selection bias occurred in the polling process for Joanna's campaign, a type of **statistical bias**. Certain members of the population (Joanna's friends) had a higher likelihood of selection than others. Selection bias is synonymous with **undercoverage bias** or **overcoverage bias**, implying that certain groups of individuals will be underrepresented or over-represented in a sample, respectively. In practice, it is difficult to obtain a simple random sample. Instead of polling students in the library on a Friday night, a better choice would be to poll students approaching the cafeteria line at lunch time. Although this polling strategy is an improvement, some sub-populations of students may still be disfavored over others (i.e. students who pack lunch, online students).

When the willingness to respond to a poll is associated with a particular response category, **response bias** or **non-response bias** may result. Suppose you sent out your polling questionnaire in a mailing with a large label reading "Vote for Joanna". Students who favor Joanna may be more likely to respond to the mailing and students who do not favor Joanna may be less likely to respond.

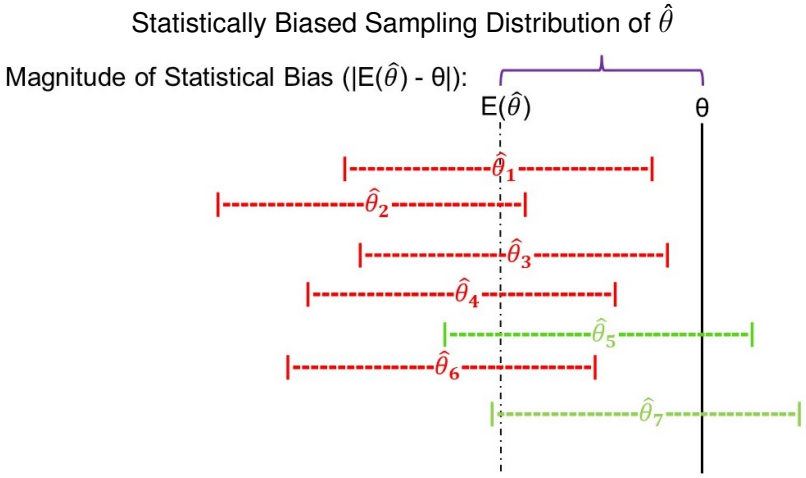
At a physical polling location, curious students may also approach the poll without being previously selected, leading to **voluntary response bias**. Respondents may also answer dishonestly to an in-person poll, resulting in **observer bias**, a type of **measurement error**. If you were to ask leading questions, such as "Don't you agree that Joanna is the best candidate"? students may agree simply to appear agreeable in response to your prompting.

Statistical Inference in the Presence of Statistical Bias

When a point estimate ($\hat{\theta}$) is statistically biased, the stated error rates associated with common statistical inference procedures are no longer reliable. In particular, a 100 $\times$ (1 $-$ α)% confidence interval can fail to circumscribe θ over 100 $\times$ α% of the time in the presence of statistical bias. The actual Type I error probability (α) for a statistical hypothesis test for which H_0 specifies the true value of θ is also likely to be underreported.

100 $\times$ (1 - α)% Confidence Intervals

The construction of a 100 $\times$ (1 $-$ α)% confidence interval for θ implicitly assumes that the point estimate of θ ($\hat{\theta}$) follows a sampling distribution having mean E($\hat{\theta}$) equal to θ (i.e. Statistical Bias = 0). Consider the scenario depicted below in which $\hat{\theta}$ is statistically biased (E($\hat{\theta}$) $<$ θ):



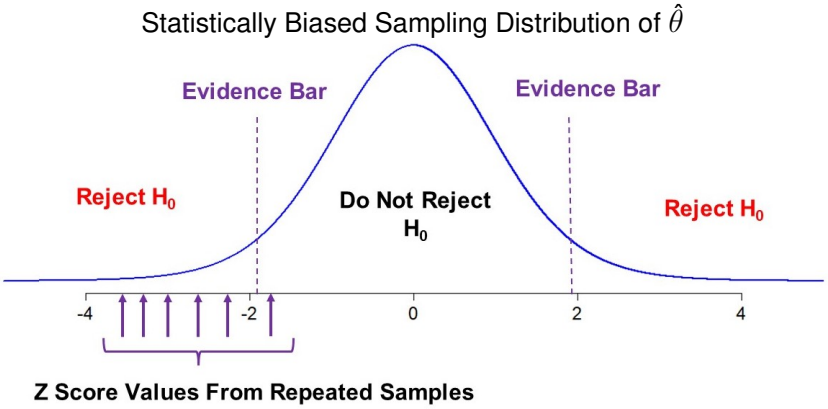
Just two out of seven confidence intervals depicted above circumscribe the true value of θ, whereas all circumscribe E($\hat{\theta}$). The extent to which the error probability (α) for

a 100 $\times$ (1 $-$ α)% confidence interval is under-stated depends on the magnitude of statistical bias.

Statistical Hypothesis Testing

When a researcher rejects the null hypothesis (H_0) of a statistical hypothesis test, even though H_0 is true, a Type I error occurs. The Type I error probability (α) is typically set to be desirably small. This value of α determines the position of the evidence bars, demarcating the rejection region based on a test statistic with a known sampling distribution assuming H_0 is true. The positioning of the evidence bars ***also implicitly assumes*** that the point estimate of θ ($\hat{\theta}$) follows a sampling distribution having mean E($\hat{\theta}$) equal to θ (i.e. Statistical Bias = 0).

Consider the test statistic $Z = \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}}$. When E($\hat{\theta}$) equals θ, the test statistic Z will (usually) follow an approximately standard normal distribution with a mean equal to zero. When E($\hat{\theta}$) $< \theta$, the expected value (E(Z)) of the test statistic (Z) will be less than zero. Consider the scenario depicted below in which $\hat{\theta}$ is statistically biased (E($\hat{\theta}$) $< \theta$):



Five out of the six Z-Score values depicted above lie below the lower evidence bar, resulting in the rejection of H_0. The extent to which the Type I error probability (α) for a statistical hypothesis test is under-stated will also depend on the magnitude of statistical bias.

Examples of Statistical Bias

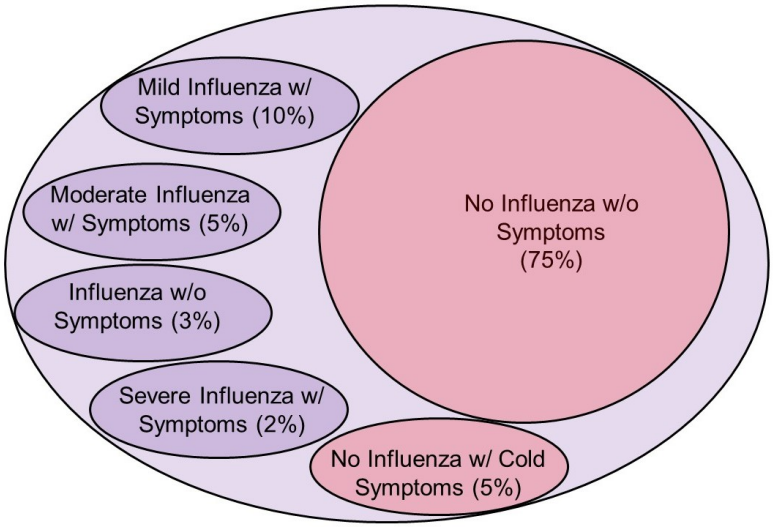
Influenza Pandemic Data

Suppose that a particularly virulent influenza virus has swept the nation into a frenzy! Public health scientists would like to determine the proportion of individuals in the country infected with the virus (θ) (i.e. disease prevalence). They would also like to determine the proportion of infected individuals who become severely ill and subsequently hospitalized (β).

An optimal strategy is to select a random sample from the population and test each individual for the influenza virus. The proportion of sampled individuals ($\hat{\theta}$) who test positive for influenza is an unbiased point estimate of θ. Among those individuals who test positive, the proportion who become severely ill and subsequently hospitalized is an unbiased point estimate of β. However, this ideal plan is not feasible. Many individuals may be unwilling to be tested for influenza, and it is not ethical or legal to force an individual to receive a medical test that they do not want.

Whereas compelled testing for influenza is not practicable, many individuals will nonetheless *voluntarily* be tested for influenza. The voluntary test results are reported to health information organizations and can be analyzed. However, the proportion of volunteers who test positive for influenza is a biased estimate of θ. Consider the composition of the population at the height of the pandemic depicted below:

Population of Individuals During Influenza Pandemic



The disease prevalence of influenza (θ) equals 0.2 and the proportion of infected individuals who become severely ill and subsequently hospitalized (β) equals 0.1. However, individuals with symptoms will tend to be over-represented in a sample of "voluntarily tested" individuals whereas healthy individuals and those with no symptoms will tend to be under-represented. Consider the composition of the "voluntarily tested" sub-population at the height of the pandemic depicted in the table below for a population of one million:

Voluntarily-Tested Sub-population of Individuals During Influenza Pandemic

Status	**Symptoms**	**Population Composition**	**Proportion Tested**	**Total Tested**
No Influenza	No Symptoms	0.75	0.1	75,000
No Influenza	Cold Symptoms	0.05	0.6	30,000
Influenza	No Symptoms	0.03	0.1	3,000
Influenza	Mild Symptoms	0.1	0.6	60,000
Influenza	Moderate Symptoms	0.05	0.85	42,500
Influenza	Severe Symptoms	0.02	1	20,000
All	**Any**	**1**	**1**	**230,500**

Out of one million individuals, 230,500 individuals voluntarily tested for influenza and 125,500 tested positive. Therefore, a point estimate of disease prevalence $\hat{\theta}$ equals $\frac{125{,}500}{230{,}500}$ or 0.544. Since the true disease prevalence (θ) equals 0.2, the magnitude of statistical bias due to voluntary testing approximately equals $\left|0.544 - 0.2\right|$, or 0.344.

Out of the 125,500 voluntarily-tested individuals who tested positive for influenza, 20,000 had severe influenza, leading to hospitalization. Therefore, a point estimate of the proportion of infected individuals who become severely ill and subsequently hospitalized ($\hat{\beta}$) equals $\frac{20{,}000}{125{,}500}$ or 0.159. Since the value of β equals 0.1, the magnitude of statistical bias due to voluntary testing approximately equals $\left|0.159 - 0.1\right|$, or 0.059.

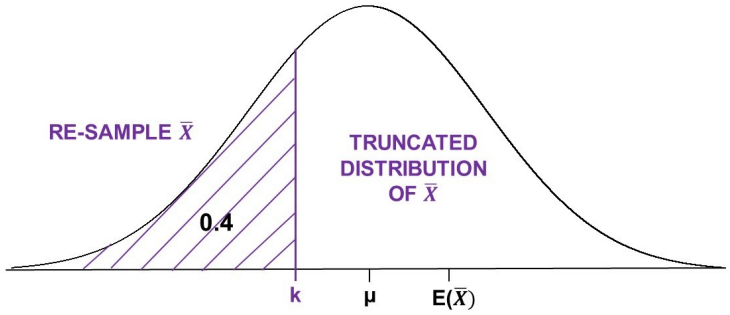
Re-Sampling

You accepted your first job as a marketer for a weight loss company, promoting a new diet plan. Congratulations! Your first task is to compute a 95% lower confidence bound (LCB) for μ, the average monthly weight loss for individuals on the first month of the diet plan. You select a random sample of 100 individuals and find that the average monthly weight loss ($\bar{X}$) for the sample was just 4 pounds! Since this number is a bit low, you repeat the study and find that the average monthly weight loss ($\bar{X}$) in the new sample is now 7.5 pounds. You throw out the first sample and report the results for the new

sample only. Re-sampling data may appear to be harmless, but is a common source of statistical bias.

Consider a sampling distribution of $\bar{X}$ having mean μ (i.e. Statistical Bias = 0). However, any sample that arises having $\bar{X}$ less than a value k will be re-sampled. Let the probability that $\hat{\theta}$ is less than k equal 0.4. Re-sampling introduces statistical bias by truncating the original sampling distribution of $\bar{X}$ and shifting E($\bar{X}$) to the right of μ as depicted below:

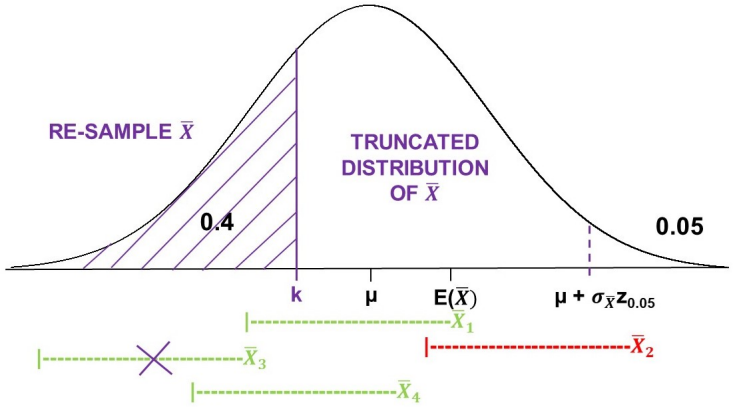
Truncated Distribution of $\bar{X}$ Resulting From Re-Sampling when $\bar{X}$ is less than k.



Since E($\bar{X}$) is now greater than μ, the resulting truncated sampling distribution of $\bar{X}$ is statistically biased.

A 95% lower confidence bound for μ equals $\bar{X} - z_{0.05}\sigma_{\bar{X}}$ (assuming a large sample size N). In the presence of re-sampling, the error rate associated with this 95% lower confidence bound for μ will now exceed 5%. Consider the set of 95% lower confidence bounds for μ depicted below:

95% Lower Confidence Bounds for μ after Re-Sampling when $\bar{X}$ is less than k.



In the absence of re-sampling, three out of the four depicted values of $\bar{X}$ produce a valid lower bound for μ. However, since $\bar{X}_3$ lies below k, it is re-sampled as though it never occurred. Three values of $\bar{X}$ remain, two of which produce a valid lower bound for μ. Over many samples, re-sampling artificially inflates the proportion of $\bar{X}$ values associated with a failed lower bound for μ, increasing the actualized error probability (α).

In the absence of re-sampling, the probability that a 95% lower confidence bound fails (LCB $> \mu$) is as follows:

$$P(\bar{X} - z_{0.05}\sigma_{\bar{X}} > \mu) = P(\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} > z_{0.05}) = 0.05$$

However, if re-sampling occurs whenever $\bar{X}$ is less than k, the probability that a 95% lower confidence bound fails (LCB $> \mu$) is now as follows:

$$P(\bar{X} - z_{0.05}\sigma_{\bar{X}} > \mu | \bar{X} > k) = \frac{P(\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} > z_{0.05})}{P(\bar{X} > k)} = \frac{0.05}{0.6} = 0.083$$

In the presence of re-sampling, a 95% lower confidence bound on μ is expected to fail 8.3% of the time.

Assignment - Calculations

Question 1: Assessing Statistical Bias in an Influenza data-set using Microsoft® Excel®.

The data set labeled ”Influenza” reports simulated data for a medium-sized population (N = 10,000) affected by an outbreak of influenza. The ”Influenza” data set reports the influenza status of all individuals in the population as well as the influenza status for voluntarily-tested individuals. Note that the influenza status of non-tested individuals is a ”hidden” or unknown variable in a research study. The following table lists all twelve variables provided in the ”Influenza” data set.

List of Variables in the ”Influenza” Data Set

Whole Population - Influenza Status	**Tested Individuals - Influenza Status**
No Influenza, No Symptoms	No Influenza, No Symptoms
No Influenza, Cold Symptoms	No Influenza, Cold Symptoms
Influenza, No Symptoms	Influenza, No Symptoms
Influenza, Mild Symptoms	Influenza, Mild Symptoms
Influenza, Moderate Symptoms	Influenza, Moderate Symptoms
Influenza, Severe Symptoms	Influenza, Severe Symptoms

The "Influenza" data set includes n = 2,500 replicates, each having a population size of N = 10,000.

Watch the Lab 10 video tutorial entitled **"Assessing Statistical Bias in a Simulated Influenza Data Set Using Microsoft® Excel®"**.

(a) Estimate the average proportion of individuals in the total population who voluntarily tested for influenza.

(b) For each replicate having a population size of N = 10,000, compute the proportion of each sub-population listed in the table below for the total population and the voluntarily-tested population separately.

Total Population Influenza Status	**Proportion**	**Voluntarily-Tested Influenza Status**	**Proportion**
No Influenza, No Symptoms		No Influenza, No Symptoms	
No Influenza, Cold Symptoms		No Influenza, Cold Symptoms	
Influenza, No Symptoms		Influenza, No Symptoms	
Influenza, Mild Symptoms		Influenza, Mild Symptoms	
Influenza, Moderate Symptoms		Influenza, Moderate Symptoms	
Influenza, Severe Symptoms		Influenza, Severe Symptoms	
Totals	1	—	1

(c) Estimate the average proportion of individuals with no influenza and no symptoms in the total population.

(d) Estimate the average proportion of individuals with no influenza and no symptoms among those who were voluntarily tested.

(e) Are individuals with no influenza and no symptoms over-represented, under-represented, or proportionally represented among those who were voluntarily tested?

(f) Estimate the average proportion of individuals with no influenza and cold symptoms in the total population.

(g) Estimate the average proportion of individuals with no influenza and cold symptoms among those who were voluntarily tested.

(h) Are individuals with no influenza and cold symptoms over-represented, under-represented, or proportionally represented among those who were voluntarily tested?

(i) Estimate the average proportion of individuals with influenza and no symptoms in the total population.

(j) Estimate the average proportion of individuals with influenza and no symptoms among those who were voluntarily tested.

(k) Are individuals with influenza and no symptoms over-represented, under-represented, or proportionally represented among those who were voluntarily tested?

(l) Estimate the average proportion of individuals with influenza and mild symptoms in the total population.

(m) Estimate the average proportion of individuals with influenza and mild symptoms among those who were voluntarily tested.

(n) Are individuals with influenza and mild symptoms over-represented, under-represented, or proportionally represented among those who were voluntarily tested?

(o) Estimate the average proportion of individuals with influenza and moderate symptoms in the total population.

(p) Estimate the average proportion of individuals with influenza and moderate symptoms among those who were voluntarily tested.

(q) Are individuals with influenza and moderate symptoms over-represented, under-represented, or proportionally represented among those who were voluntarily tested?

(r) Estimate the average proportion of individuals with influenza and severe symptoms in the total population.

(s) Estimate the average proportion of individuals with influenza and severe symptoms among those who were voluntarily tested.

(t) Are individuals with influenza and severe symptoms over-represented, under-represented, or proportionally represented among those who were voluntarily tested?

(u) Estimate the average proportion of individuals with influenza in the total population. Set your result equal to θ.

(v) Estimate the average proportion of individuals with influenza among those who were voluntarily tested. Set your result equal to E($\hat{\theta}$).

(w) Compute the magnitude of statistical bias with respect to the variable θ, using your answers in (u) and (v).

(x) Estimate the average proportion of individuals with influenza and severe symptoms among those with influenza in the total population. Set your result equal to β.

(y) Estimate the average proportion of individuals with influenza and severe symptoms among those with influenza who were voluntarily tested. Set your result equal to E($\hat{\beta}$).

(z) Compute the magnitude of statistical bias with respect to the variable β, using your answers in (x) and (y).

(aa) Using Microsoft® Excel®, complete the following steps 1 - 6 for each replicated population of size 10,000.

1. Compute a 99% confidence interval for θ: $\hat{\theta} \pm z_{0.005}\sqrt{Var(\hat{\theta})}$, where $Var(\hat{\theta}) \approx \frac{\hat{\theta}\times(1-\hat{\theta})}{N_{\text{Tested}}}$.
2. Compute a 99% confidence interval for β: $\hat{\beta} \pm z_{0.005}\sqrt{Var(\hat{\beta})}$, where $Var(\hat{\beta}) \approx \frac{\hat{\beta}\times(1-\hat{\beta})}{N_{\text{Tested, Influenza}}}$.
3. Set θ equal to your answer in (u). Compute a Z-Score test statistic: $\frac{\hat{\theta}-\theta}{\sqrt{Var(\hat{\theta})}}$ where $Var(\hat{\theta}) \approx \frac{\hat{\theta}\times(1-\hat{\theta})}{N_{\text{Tested}}}$.
4. Set β equal to your answer in (x). Compute a Z-Score test statistic: $\frac{\hat{\beta}-\beta}{\sqrt{Var(\hat{\beta})}}$ where $Var(\hat{\beta}) \approx \frac{\hat{\beta}\times(1-\hat{\beta})}{N_{\text{Tested, Influenza}}}$.
5. Create an indicator variable to denote whether the magnitude of the Z-Score test statistic ($|Z|$) in (3) exceeds $z_{0.005}$.

6. Create an indicator variable to denote whether the magnitude of the Z-Score test statistic ($|Z|$) in (4) exceeds $z_{0.005}$.

(ab) Compute the proportion of 99% confidence intervals for θ that fail to circumscribe the designated value of θ.

(ac) Compute the proportion of 99% confidence intervals for β that fail to circumscribe the designated value of β.

(ad) Compute the proportion of Z-Score test statistics (Z = $\frac{\hat{\theta}-\theta}{\sqrt{Var(\hat{\theta})}}$) that exceed $z_{0.005}$.

(ae) Estimate the Type I error probability (α) for a statistical hypothesis test about θ based on your result in (ad).

(af) Compute the proportion of Z-Score test statistics (Z = $\frac{\hat{\beta}-\beta}{\sqrt{Var(\hat{\beta})}}$) that exceed $z_{0.005}$.

(ag) Estimate the Type I error probability (α) for a statistical hypothesis test about β based on your result in (af).

Question 2: Assessing Statistical Bias in the Estimation of μ in the Presence of Re-Sampling using Microsoft® Excel®.

Watch the Lab 10 video tutorial entitled **"Assessing Statistical Bias in the Estimation of μ in the Presence of Re-Sampling Using Microsoft® Excel®"**.

An environmental research group hypothesizes that the average frog weight (μ) in a marsh population has increased over time. Ten years prior, a different research group derived a precise estimate of the average frog weight in the marsh equal to μ = 23 grams. In order to produce evidence supporting an increase in the average frog weight (μ), the group plans to randomly sample frogs from the marsh, measure each frog's weight in grams and test the following one-sided statistical hypothesis:

$$H_0: \mu_0 = 23 \text{ grams}$$
$$H_A: \mu_0 > 23 \text{ grams}$$

They also plan to compute a 100 $\times(1-\alpha)$% one-sided lower confidence bound on the parameter μ. With this specific research goal in mind, the research group would not like the point estimate of μ to be "too low" and plan to re-sample if the sample mean ($\bar{X}$) is less than k grams. While re-sampling appears unobjectionable (and untraceable)

enough, it can lead to statistical bias!

The data set labeled "Frog Population C" reports individual frog weights in grams for 2,500 different samples of size N = 10 and size N = 20 frogs. Individual frog weights in Population C are normally distributed with μ equal to 23 grams and σ equal to 2 grams.

(a) Using Microsoft® Excel®, complete the following steps for replicated samples of size N = 10 and N = 20 frogs:

1. Compute the sample mean ($\bar{X}$) for each of the 2,500 replicated samples.
2. Compute the sample standard deviation (s) for each of the 2,500 replicated samples.
3. Create an indicator variable to denote re-sampling when $\bar{X}$ is less than 22.75 grams (Re-sampling Scenario A).
4. Create an indicator variable to denote re-sampling when when $\bar{X}$ is less than 23.5 grams (Re-sampling Scenario B).
5. Compute a T-Score test statistic (T) where $T = \frac{\bar{X}-23}{s_{\bar{X}}}$ where $s_{\bar{X}} = \frac{s}{\sqrt{N}}$.
6. Compute a one-sided 99% lower confidence bound ($\bar{X}$ - d) for the parameter μ where d = $t_{0.01,DF=N-1} \times s_{\bar{X}}$.

(b) Compute the average value of $\bar{X}$ in the absence of re-sampling for samples of N = 10 frogs.

(c) Compute the magnitude of statistical bias in the absence of re-sampling for samples of N = 10 frogs. Use your answer in (b) as an estimate of E($\bar{X}$), setting μ equal to 23 grams.

(d) Compute the average value of $\bar{X}$ under Re-sampling Scenario A for samples of N = 10 frogs.

(e) Compute the magnitude of statistical bias under Re-sampling Scenario A for samples of N = 10 frogs. Use your answer in (d) as an estimate of E($\bar{X}$), setting μ equal to 23 grams.

(f) Compute the average value of $\bar{X}$ under Re-sampling Scenario B for samples of N

= 10 frogs.

(g) Compute the magnitude of statistical bias under Re-sampling Scenario B for samples of N = 10 frogs. Use your answer in (f) as an estimate of E($\bar{X}$), setting μ equal to 23 grams.

(h) Compute the average value of $\bar{X}$ in the absence of re-sampling for samples of N = 20 frogs.

(i) Compute the magnitude of statistical bias in the absence of re-sampling for samples of N = 20 frogs. Use your answer in (h) as an estimate of E($\bar{X}$), setting μ equal to 23 grams.

(j) Compute the average value of $\bar{X}$ under Re-sampling Scenario A for samples of N = 20 frogs.

(k) Compute the magnitude of statistical bias under Re-sampling Scenario A for samples of N = 20 frogs. Use your answer in (j) as an estimate of E($\bar{X}$), setting μ equal to 23 grams.

(l) Compute the average value of $\bar{X}$ under Re-sampling Scenario B for samples of N = 20 frogs.

(m) Compute the magnitude of statistical bias under Re-sampling Scenario B for samples of N = 20 frogs. Use your answer in (l) as an estimate of E($\bar{X}$), setting μ equal to 23 grams.

(n) Compute the proportion of T-Score test statistics (T) that exceed the evidence bar ($t_{\alpha=0.05,DF=9}$) in the absence of re-sampling for samples of N = 10 frogs.

(o) Use your answer in (n) to estimate the Type I Error Probability (α) in the absence of re-sampling for samples of N = 10 frogs.

(p) Compute the proportion of T-Score test statistics (T) that exceed the evidence bar ($t_{\alpha=0.05,DF=9}$) in the presence of re-sampling (Re-sampling Scenario A) for samples of N = 10 frogs.

(q) Use your answer in (p) to estimate the Type I Error Probability (α) in the presence

of re-sampling (Re-sampling Scenario A) for samples of N = 10 frogs.

(r) Compute the proportion of T-Score test statistics (T) that exceed the evidence bar ($t_{\alpha=0.05,DF=9}$) in the presence of re-sampling (Re-sampling Scenario B) for samples of N = 10 frogs.

(s) Use your answer in (r) to estimate the Type I Error Probability (α) in the presence of re-sampling (Re-sampling Scenario B) for samples of N = 10 frogs.

(t) Compute the proportion of T-Score test statistics (T) that exceed the evidence bar ($t_{\alpha=0.05,DF=19}$) in the absence of re-sampling for samples of N = 20 frogs.

(u) Use your answer in (t) to estimate the Type I Error Probability (α) in the absence of re-sampling for samples of N = 20 frogs.

(v) Compute the proportion of T-Score test statistics (T) that exceed the evidence bar ($t_{\alpha=0.05,DF=19}$) in the presence of re-sampling (Re-sampling Scenario A) for samples of N = 20 frogs.

(w) Use your answer in (v) to estimate the Type I Error Probability (α) in the presence of re-sampling (Re-sampling Scenario A) for samples of N = 20 frogs.

(x) Compute the proportion of T-Score test statistics (T) that exceed the evidence bar ($t_{\alpha=0.05,DF=19}$) in the presence of re-sampling (Re-sampling Scenario B) for samples of N = 20 frogs.

(y) Use your answer in (x) to estimate the Type I Error Probability (α) in the presence of re-sampling (Re-sampling Scenario B) for samples of N = 20 frogs.

(z) Compute the proportion of 99% lower confidence bounds that lie above the true mean of 23 grams in the absence of re-sampling for samples of N = 10 frogs.

(aa) Use your answer in (z) to estimate the error probability (α) in the absence of re-sampling for samples of N = 10 frogs.

(ab) Compute the proportion of 99% lower confidence bounds that lie above the true mean of 23 grams in the presence of re-sampling (Re-sampling Scenario A) for samples of N = 10 frogs.

(ac) Use your answer in (ab) to estimate the error probability (α) in the presence of re-sampling (Re-sampling Scenario A) for samples of N = 10 frogs.

(ad) Compute the proportion of 99% lower confidence bounds that lie above the true mean of 23 grams in the presence of re-sampling (Re-sampling Scenario B) for samples of N = 10 frogs.

(ae) Use your answer in (ad) to estimate the error probability (α) in the presence of re-sampling (Re-sampling Scenario B) for samples of N = 10 frogs.

(af) Compute the proportion of 99% lower confidence bounds that lie above the true mean of 23 grams in the absence of re-sampling for samples of N = 20 frogs.

(ag) Use your answer in (af) to estimate the error probability (α) in the absence of re-sampling for samples of N = 20 frogs.

(ah) Compute the proportion of 99% lower confidence bounds that lie above the true mean of 23 grams in the presence of re-sampling (Re-sampling Scenario A) for samples of N = 20 frogs.

(ai) Use your answer in (ah) to estimate the error probability (α) in the presence of re-sampling (Re-sampling Scenario A) for samples of N = 20 frogs.

(aj) Compute the proportion of 99% lower confidence bounds that lie above the true mean of 23 grams in the presence of re-sampling (Re-sampling Scenario B) for samples of N = 20 frogs.

(ak) Use your answer in (aj) to estimate the error probability (α) in the presence of re-sampling (Re-sampling Scenario B) for samples of N = 20 frogs.

Assignment - Short Report

Question 1: Assessing Statistical Bias in an Influenza data-set using Microsoft® Excel®.

(a) Complete the following table based on your answers in (c) - (s).

Representation of Sub-Populations in the "Influenza" Data Set

Total Population Influenza Status	**Proportion**	**Voluntarily-Tested Influenza Status**	**Proportion**
No Influenza, No Symptoms		No Influenza, No Symptoms	
No Influenza, Cold Symptoms		No Influenza, Cold Symptoms	
Influenza, No Symptoms		Influenza, No Symptoms	
Influenza, Mild Symptoms		Influenza, Mild Symptoms	
Influenza, Moderate Symptoms		Influenza, Moderate Symptoms	
Influenza, Severe Symptoms		Influenza, Severe Symptoms	
Totals	1	—	1

(b) Which sub-populations of individuals are over-represented, under-represented and proportionally represented among those individuals voluntarily-tested for influenza?

(c) How might reliance on voluntary testing versus random sampling give rise to statistical bias in the estimation of θ? Did you observe statistical bias in the estimation of θ based on voluntarily-tested individuals?

(d) How might reliance on voluntary testing versus random sampling contribute to statistical bias in the estimation of β? Did you observe statistical bias in the estimation of β based on voluntarily-tested individuals?

(e) Compare your computed error probability (α) associated with the 99% confidence intervals for θ with the pre-specified α equal to 0.01. Did the presence of statistical bias affect the integrity of your 99% confidence intervals of θ?

(f) Compare your computed error probability (α) associated with the 99% confidence intervals for β with the pre-specified α equal to 0.01. Did the presence of statistical bias affect the integrity of your 99% confidence intervals of β?

(g) Compare your computed Type I error probability (α) associated with the statistical hypothesis tests for θ with the pre-specified α equal to 0.01. Did the presence of

statistical bias affect the Type I error probability (α)?

(h) Compare your computed Type I error probability (α) associated with the statistical hypothesis tests for β with the pre-specified α equal to 0.01. Did the presence of statistical bias affect the Type I error probability (α)?

Question 2: Assessing Statistical Bias in the Estimation of μ in the Presence of Re-Sampling using Microsoft® Excel®.

(a) Report your estimates of the magnitude of statistical bias in the table below.

Magnitude of Statistical Bias by Re-Sampling Scenario

	Absence of Re-Sampling	**Re-Sampling Scenario A**	**Re-Sampling Scenario B**
N = 10			
N = 20			

(b) Did the presence of re-sampling impact the magnitude of statistical bias? Was the magnitude of statistical bias in the presence of re-sampling different for Re-sampling Scenarios A and B? Did the sample size (N) impact the magnitude of statistical bias?

(c) Why would re-sampling introduce statistical bias when each sample is a simple random sample?

(d) When reviewing the results of a statistical analysis, how can you tell whether or not a data set had previously been re-sampled?

(e) Report your estimates of the Type I Error Probability (α) in the table below.

Type I Error Probability (α) Associated with a One-Sided Statistical Hypothesis test for μ by Re-Sampling Scenario

	Absence of Re-Sampling	**Re-Sampling Scenario A**	**Re-Sampling Scenario B**
N = 10			
N = 20			

(f) Did re-sampling influence the Type I error probability (α) of your one-sided statistical hypothesis test for μ? Discuss your estimates of the Type I error probability (α) with respect to Re-sampling scenarios A and B and sample size.

(g) Report your estimates of the error probability (α) associated with the 99% lower confidence bound for μ in the table below.

Error Probability (α) Associated with the 99% lower confidence bound for μ by Re-Sampling Scenario

	Absence of Re-Sampling	**Re-Sampling Scenario A**	**Re-Sampling Scenario B**
N = 10			
N = 20			

(h) Did re-sampling influence the attained error probability (α) of your 99% lower confidence bounds for μ? Discuss your estimates of the attained error probability (α) with respect to Re-sampling scenarios A and B and sample size.