Analysis of Average Winter Temperatures in College Park, MD

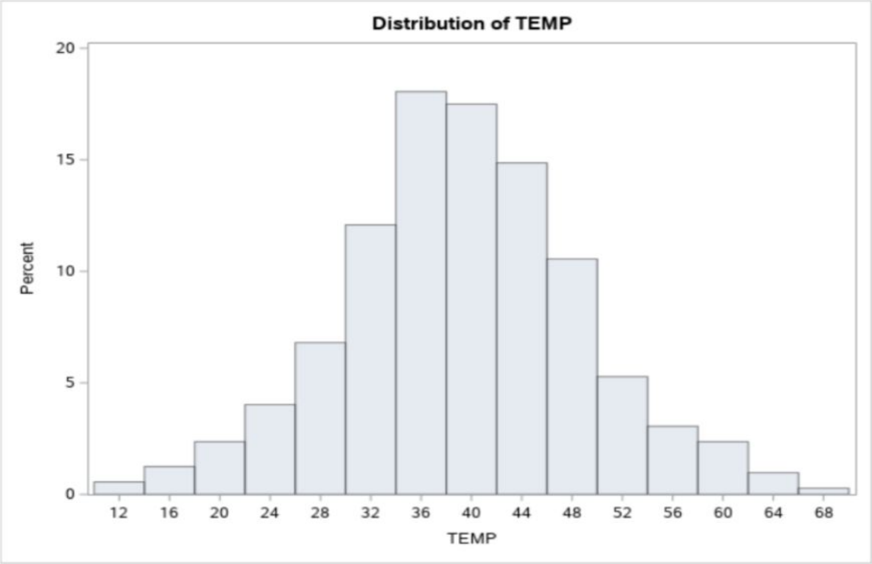
Sophia Hu

Introduction

- Our Question
  - *What is the average temperature (in Fahrenheit) in College Park, Maryland during the winter season (December, January, February months)?*
- Dataset
  - Winter weather in College Park, Maryland
  - Specifically, the entries are from December, January, and February months from 12/1/14 to 2/28/22 (N = 720 entries total)
- Dataset Source: Visual Crossing Weather
  - https://www.visualcrossing.com/resources/documentation/weather-data/weather-data-documentation/
- Software
  - Mostly used R to conduct analyses
  - Used some Python for Report 3
- Variable of Interest
  - Average Winter Temperature (˚F)
- Other Variables Used
  - Dew, Humidity, Date

Population

Mean: $$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{28207.60}{720} \approx 39.177\ ^{\circ}\mathrm{F}$$



- Symmetrical
- Bell-shaped
- Approximately normally distributed
- Median (37.100) ≈ Mean (39.177)
- Standard deviation ≈ 9.6
- **N = 720**
- From Report 2, we based sample sizes on **n = 84** as optimal sample size and most convenient

Real Life Application

- Interesting to know the average temperature for local College Park town during the winter season
- Discover trends over past few years
- Winter temperature tells us the amount of snow that falls, timing of snowmelt runoff, loss of soil moisture, frozen duration of rivers, and more!
    - This information can be crucial for business and personal matters and have great impacts on our lifestyle
- Tells us more information about climate change
    - Increase in extreme weather activity, rise in sea levels, melting of glaciers and sea ice, wildfires, droughts, loss of wildlife species

Stratified Sampling Overview

- Obtained best results using Stratified Sampling with Optimum and then Proportional Allocation
- Tried using Dew and Humidity as variable for stratification, but the confidence intervals using Humidity did not include the temperature's true mean
    - Makes sense because Temperature is much more correlated with Dew than Humidity

$$d = (N-1)\sigma^2 - \sum_{h=1}^{L}(N_h - 1)\sigma_h^2 \quad \text{where}$$

$$\sigma_h^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(y_{hi} - \mu_h)^2,\ h = 1, \ldots, L$$

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

- Tried three different ways to create stratas and chose method with largest d
    1. Used winter months to create 3 strata to divide population into data from December, January, and February -> d = 4719.88.
    2. Used dew to stratify every 10˚F -> d = 51024.52
    3. Used dew to stratify every 20˚F -> d = 41389.77
- Performed Stratified Sampling using Equal, Proportional, and Optimum Allocation

Estimator #1: Stratified Sample with Unbiased Estimator with Optimum Allocation

- Stratified using Dew variable every 10 ˚F
- L = 7 strata
- Overall, Optimum Allocation had smallest estimated variance, and Proportional allocation had second smallest estimated variance

Threshold	N	Optimum n	Proportional n
Dew < 0	N1 = 17	n1 = 1	n1 = 2
0 <= Dew < 10	N2 = 49	n2 = 6	n2 = 6
10 <= Dew < 20	N3 = 133	n3 = 16	n3 = 16
20 <= Dew < 30	N4 = 236	n4 = 27	n4 = 28
30 <= Dew < 40	N5 = 167	n5 = 20	n5 = 19
40 <= Dew < 50	N6 = 92	n6 = 11	n6 = 11
Dew > 50	N7 = 26	n7 = 3	n7 = 3
Total	N = 720	n = 84	n = 84

Formulas:

Optimum Allocation: $n_h = \frac{n N_h \sigma_h}{\sum_{h=1}^{L} N_h \sigma_h}$ Proportional Allocation: $n_h = \frac{n N_h}{N}$

$$\bar{y}_{st} = \frac{1}{N} \sum_{h=1}^{L} (N_h * \bar{y}_h)$$

$$\widehat{var}(\bar{y}_{st}) = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \left(\frac{N_h - n_h}{N_h}\right) \left(\frac{s_h^2}{n_h}\right)$$

$$95\% \text{ CI} = \bar{y}_{st} \pm t\sqrt{\widehat{var}(\bar{y}_{st})}$$ with d degrees of freedom where

$$d = \left(\sum_{h=1}^{L} a_h s_h^2\right)^2 / \left(\sum_{h=1}^{L} (a_h s_h^2)^2 / (n_h - 1)\right) \text{ where } a_h = N_h(N_h - n_h)/n_h$$

Optimum Allocation

Estimated Values:

$\bar{y}_{st} = 39.0705$

$\widehat{var}(\bar{y}_{st}) = 0.012$

95% CI = (38.852, 39.289)

Proportional Allocation

Estimated Values:

$\bar{y}_{st} = 39.308$

$\widehat{var}(\bar{y}_{st}) = 0.016$

95% CI = (39.055, 39.560)

Estimator #2:
Regression Estimator using Dew as Auxiliary Variable

- Also tried using Humidity as auxiliary variable but those results yielded larger estimated variances
  - Makes sense because Temperature is more correlated with Dew than Humidity
- n = 84 SRS
- Performed diagnostic analysis
  - Slope p-value less than 0.05 -> Linear relationship between Dew and Temperature
  - Intercept p-value less than 0.05 -> Does not cross through origin
- Makes sense Regression Estimator performed better than Ratio Estimator from regression analysis

Formulas and Results

$$\widehat{\mu_L} = a + b * \mu_x \text{, with } a = \bar{y} - b\bar{x} \text{, and } b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad = 39.513$$

$$\widehat{var(\mu_L)} = \left(\frac{N-n}{Nn(n-2)}\right)\sum_{i=1}^{n}(y_i - a - bx_i)^2 \quad = 0.217$$

$$CI_{95} = \widehat{\mu_L} \pm t_{n-2,\alpha/2}\sqrt{\widehat{var(\mu_L)}} \quad = (38.586, 40.439)$$

Systematic Sampling Overview

- Also obtained pretty good results using Systematic Sampling based on estimated variances
  - Though estimated variances were small, the confidence intervals were fairly large compared to other estimators due to low degrees of freedom
- Other estimators had narrower confidence intervals
- Performed four Systematic Samples
  - 2-in-16 Systematic Sampling on unsorted Temperature data
  - 3-in-24 Systematic Sampling on unsorted Temperature data
  - 2-in-16 Systematic Sampling on sorted Temperature data
  - 3-in-24 Systematic Sampling on sorted Temperature data
- Using sorted data yielded lower estimated variances than the respective unsorted data

Estimator #3:
Sorted Systematic 3-in-24 Sample with Unbiased Estimator

- Also has fairly big confidence interval despite small estimated variance but is smaller than 2-in-16 unsorted (degrees of freedom)
- 3 groups of 30 samples each = 90 total
- Degrees of freedom = 2
- Smaller estimated variance than 3-in-24 unsorted

Formulas:

$$\hat{\mu} = \hat{\tau}/M = \frac{N}{n^*M}\sum_{i=1}^{n} y_i = \frac{N\bar{y}}{M}$$

$$\widehat{var}(\hat{\mu}) = \frac{\widehat{var}(\hat{\tau})}{M^2} = N(N-n)\frac{s_u^2}{nM^2}$$

where $s_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

$$95\%\ CI = \hat{\mu} \pm t_{\alpha/2,\, n-1}\sqrt{\widehat{var}(\hat{\mu})}$$

N = # of primary units in population so N = 24

n = # of primary units in sample so n = 3

M = # of secondary units so M = 720

Estimated Values:

$\hat{\mu} = 39.37$

$\widehat{var}(\hat{\mu}) = 0.1458333$

95% CI = (37.7269, 41.0131)

Estimator #4:
Sorted Systematic 2-in-16 Sample with Unbiased Estimator

- Fairly large confidence interval despite small estimated variance
- 2 groups of 45 samples each = 90 total
- Degrees of freedom = 1
- Smaller estimated variance than 2-in-16 unsorted

Formulas:

$$\hat{\mu} = \hat{\tau}/M = \frac{N}{n^*M}\sum_{i=1}^{n} y_i = \frac{N\bar{y}}{M}$$

$$\widehat{var}(\hat{\mu}) = \frac{\widehat{var}(\hat{\tau})}{M^2} = N(N-n)\frac{s_u^2}{nM^2}$$

where $s_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

$95\%\ CI = \hat{\mu} \pm t_{\alpha/2,\,n-1}\sqrt{\widehat{var}(\hat{\mu})}$

N = # of primary units in population so N = 16

n = # of primary units in sample so n = 2

M = # of secondary units so M = 720

Estimated Values:

$\hat{\mu} = 39.32$

$\widehat{var}(\hat{\mu}) = 0.06118951$

95% CI = (36.17693, 42.46307)

Summary Of Top Estimators

Rank (Using Estimated Variance)	Estimator	Estimated Mean	Estimated Variance	Estimated Standard Deviation	95% Confidence Interval
1	Stratified Sample with Unbiased Estimator with Optimum Allocation (using Dew)	39.07	0.012	0.110	(38.852, 39.289)
2	Stratified Sample with Unbiased Estimator with Proportional Allocation (using Dew)	39.31	0.016	0.126	(39.055, 39.560)
3	Sorted Systematic 2-in-16 Sample with Unbiased Estimator	39.32	0.061	0.247	(36.177, 42.463)
4	Sorted Systematic 3-in-24 Sample with Unbiased Estimator	39.37	0.146	0.382	(37.727, 41.013)
5	Regression Estimator (using Dew)	39.51	0.217	0.487	(38.586, 40.439)

Best Estimator

- Based on smallest estimated variance, the best estimator is using the Stratified Sample with Unbiased Estimator with Optimum Allocation when we stratified using Dew as an auxiliary variable every 10 ˚F using seven strata

Estimated Values:

$\bar{y}_{st} = 39.0705$

$\widehat{var}(\bar{y}_{st}) = 0.012$

95% CI = (38.852, 39.289)

- Makes sense using Dew as auxiliary variable yielded good results because Dew and Temperature are highly correlated
  - Correlation coefficient between Dew and Temperature: r = 0.9025
  - Correlation coefficient between Humidity and Temperature: r = 0.4340

Interpretation and Conclusion

- Some data was missing or were mostly 0s so unable to use some variables like precipitation, snow, rain
  - Also had entry with Dew of 0 which caused issues in Report 3 when calculating true variances, so we omitted that entry for Report 3
- Out of all estimators calculated this semester, we obtained best results in terms of smallest estimated variance using Stratified Sampling and Systematic Sampling
- Can further examine weather in College Park
  - Can look at another season like summer, spring, or fall
  - Analyze weather in the future or prior to 2014 to see how results compare holistically
- Also can conduct similar analyses in other locations and make connections to big weather events occurring