**Analysis of Winter Olympics Participation between 1924 and 2018.**

**Data Set**

This dataset was acquired from <https://www.kaggle.com/datasets>. Compiled by le Minh Nguyen, the dataset ‘disciplie\_details.csv’ includes Olympic game number, year, event discipline, event category, and date of event competition. The data also includes number of overall participants per event, number of competing countries per event, and the gold, silver, and bronze winners along with their home countries.

**Data Preparation**

The data was loaded from .csv format into an R data frame. As part of the analysis centers around a comparison of gender participation, a parse of the ‘category’ column was needed to extract the male/female classification of the event. This information was appended to the data frame as a new column for easier extraction. Additionally, several encoding anomalies existed in the events and dates. These were corrected using string manipulation within the data frame.

Graphical user interface, application, table

Description automatically generated

**Objective**

The goal of this analysis is to measure the growth of the Winter Olympic games between the years 1924 and 2018 across multiple facets:

1. Participant Statistics
   1. Overall participation and event number increase
2. Event Statistics
   1. Distribution of overall participation in top events across all Winter Olympics
   2. Distribution of participant countries in top events across all Winter Olympics
3. Country Statistics
   1. Medal counts of top winning countries across all Winter Olympics
   2. Medal count spread across 1924 (min), 1976(median) and 2018(max) Olympics
4. Population vs. Samples
   1. Central Limit Theorem
   2. Sampling Methods

**Participant Statistics**

**Overall Participation and Event Increase**

The first analysis to perform is to look at overall participation of events across all Winter Olympic years. The raw data needed manipulation to split participation counts by event year. A tibble of the Olympic data frame was used to group the years and sum the yearly participants together. At the same time, the unique event counts were tallied, creating a new tibble column.

**Chart, line chart

Description automatically generated**

The blue bars represent total participation per year. The red line with markers represents the trend of the participation across the same time. To overlay both the number of events and number of yearly participants, I used two y axes in plot\_ly to represent each data range’s max and min values.

Analysis:

Concerning the visual anomalies in the data, the Olympics did not occur during wartime in 1940 and 1944. Also, in 1994 the International Olympic Committee decided to separate the Summer and Winter Olympic years, resulting in the two-year split between the 1992 and 1994 games. Despite these impacts, a very consistent correlation exists between both an increase in participation and an increase in events over time. This demonstrates that the driving factor behind the growth of participation is the creation of new events, and not so much new country participation. Regarding the shape of the plot, the data almost represents an incomplete left-skewed normal distribution with the later years coming close to the median of the bell. There is not any reason to believe that the number of events or the participation overall will drop in subsequent years to fit the normal curve, but the plateau of events and participation may be related to several logistical factors. Increasing stress on the host country to provide slope, arena and building venues balanced against the increased time commitment and event planning will eventually peak the creation of new events and added participants.

**Event Statistics**

**Distribution of Overall Participation Across Top 15 Participated Events Over Time**

Beyond looking at the participation across the entire Winter Olympics per year, the next plot shows the distribution of participants across the same event over all the Olympic years. This data was not present in the raw data, so a tibble was used to first sort the events in order of overall participation. Once the top 15 from this data was found, a second tibble sorted the data by participation per event, with all events other than the top 15 being omitted from the data. For graphical representation, a last sort of the participation per event data tracked the median values to specify the y-axis order.

Analysis:

The boxplot of this distribution shows a team sport, Ice Hockey, at the lead followed by many variations of skiing, which is a long-time staple of the Winter Olympics. All events show a decent spread between min and max participation, the greatest of which is men’s ice hockey. As in other team events, the rules do not commonly change over time to allow more players to participate in an event, unless there are alternates. Also applicable to the single person competitions, the only explanation for a large spread in participation over the years is a change in the number of participant countries. For the same reason, if an outlier existed in an event, it makes sense that it remained relatively close to the distribution.

Graphical user interface, table

Description automatically generated with medium confidence

**Distribution of Country Participation Across Top 15 Participated Events Over Time**

Comparing notes with the distribution of participants above, it is no surprise to see many of the same events on the distribution of country participation per top 15 participated events over the Winter Olympic years. The same steps were taken to manipulate participant data as in the last boxplot, with the exception being a focus on the ‘n\_country\_participation’ variable for means of trimming and sorting.

This plot is dominated by downhill and cross-county skiing events, sorted by maximum country participation. A correlation between the spread of participant countries per event and the spread of overall event participation is now visible.

Chart, box and whisker chart

Description automatically generated

**Country Statistics**

**Medal Counts of Most Winning Countries**

Now will begin the focus on performances between countries overall by measure of cumulative gold, silver, and bronze medals. To achieve this, a table summary was used to capture the frequencies of the gold, silver, and bronze winning country columns after the data was grouped using a tibble. The first chart compares the winning percentages of the top 10 most winning countries, achieved by sorting the tibble by frequency and using only the top 10 rows. The second pie chart uses the same data but adds an 11th wedge which represents all of the remaining medals won by other countries.

Chart, pie chart

Description automatically generated

Among the top 10 most winning countries across Winter Olympic years, there exists a good distribution of percentages. Countries within the arctic regions are well represented here, which makes sense as these Olympic events are all winter activities. Pulling in the results of the second pie chart, the top 10 performing countries contribute to roughly half of overall medals won. The top four performing countries have over a quarter of the total, suggesting that these countries see great Winter Olympics success across time. To see if this trend is representative of all the years of the Winter Olympics, lets now compare medal winnings over set years.

**Medal Counts: 1924 vs 1976 vs 2018**

The three years in this analysis represent the starting, middle, and ending years in this dataset. Comparing the proportions of medal winnings by country across these three years should give insight into winning trends, if any. The multivariate bar plot takes information from the ‘event\_year’, ‘gold\_country’, ‘silver\_country’ and ‘bronze\_country’ data columns. A tibble was used to filter the 1924, 1976 and 2018 event entries out of the Olympics data frame. A table within the tibble gave the unique countries and their frequency in the data, representing their overall medal count for that year. The data was added as three individual traces into a plot\_ly representation.

Chart, bar chart

Description automatically generated

Analysis:

Isolating the 1924 data shows eleven countries receiving medals, with Finland and Norway achieving great success. The 1976 overlay begins to show a spread of medal winnings across more countries, with East Germany and the Soviet Union at the front. The 2018 data shows a wide spread of medals across all participant nations (FRG, DGR, TCH, and URS are no longer countries at this time, and the MIX category only existed in the 1924 plot). The peaks are now forming in the 30 and 40 counts. Combining the statistics from the pie charts, once again the countries with Arctic climates are well represented. In later years, gold medals were won by more countries with greater overall frequency than in earlier years. This trend corroborates the idea that participation and performance in the Winter Olympics has grown with time.

**Population vs. Samples**

**Central Limit Theorem**

Do assumptions made to the whole data population apply to random samples? The Central Limit Theorem states that if a population of data follows a normal distribution (i.e. has a mean and a standard deviation), then random samples of increasing size will be increasingly normally distributed. In other words, the mean of the samples should approach that of the population, and the standard deviation of the samples should approach the value of the equation: SD(sample) = SD(population)/sqrt(sample size).

Revisiting the participants per event over the course of all Winter Olympics can demonstrate this relationship. The data was formed using a tibble to isolate the ‘n\_particpants’ column of the Olympic data frame again. This time, the data is fed into a histogram, with additional line segment traces added to visually represent mean and +/- standard deviations.

Application

Description automatically generated with low confidence

Here is the behavior of four sample sizes: 25, 50, 75 and 100. The mean sampling was taken using normal random selection from a pool of 500 samples of the cumulative participant data, which was extracted using a tibble and isolated down to the ‘n\_participants’ category.

A screenshot of a computer

Description automatically generated with medium confidence

Analysis:

Keeping the x-axis equal to the population data, each of the plots show a noticeable relationship to the original plot. The mean of the original data was 50.99 -> ~51 participants per event, which the sampling means closely match. The standard deviations of the samples follow the expected mathematical calculation shown below. All sample sizes chosen here make very similar plots, with the only difference being the standard deviation growing smaller. This pools more values within one standard deviation of the mean, creating ever narrower and taller distributions.

A picture containing table

Description automatically generated

**Other Sampling Methods**

Several other sampling methods exist beyond random normal sampling. Observe below samples gathered using the following methods:

* Single Random Sampling without Replacement (SRSWOR)

Randomly sampling the entire data set for a given number of results. Each time a sample is taken, the result is omitted from the data set for future samples to eliminate overlap.

* Systematic Sampling with Equal Probability (Systematic Equal)

Dividing the population data into equal segments, the number of which represents the number of samples desired. A random number is selected within the interval of the data, which represents where in the segment the sample will be drawn uniformly across all segments. This method may result in null samples returned if the desired sample size is not a factor of the population size.

* Systematic Sampling with Unequal Probability (Systematic Unequal)

Sampling using the inclusionprobabilities() function to set the fractional probability for each element of the population set according to the desired sample size. Then the UPsystematic() function returns the selected sample indices in binary form to extract from the population data.

* Stratified Sampling with Proportional Sizes (Stratified)

The original dataset is ordered by proportionally by the size of a ranking category. Strata size and probability are determined by the proportion of entries versus the whole data. Based on the weighted probability, sample selections are biased to the strata with larger numbers.

Chart

Description automatically generated

All the sample data created visually resembled the master population plot, demonstrating the viability of using sampling to demonstrate behavior and trends of an entire data set. See below for the means and standard deviations of each sampling method.

Table

Description automatically generated with low confidence

Analysis

The sample size was adjusted from the small samples used in the previous random normal sampling analysis. When stratifying the results per their event discipline for ranking, some disciplines contained hundreds of values, while others contained only one. If a small sample set is chosen, the R strata() function fails with a data frame sizing error. Essentially there cannot be a stratum with 0 selected samples. Additionally, applying a ceiling function to the values ensured that there was always a sample at the lowest strata. The overall sampling size was only slightly bigger than 200 and within an acceptable error to compare to the other sampling results.

The original plot showed overall distribution of participation per event across all the Olympic years. Each of these methods generated a mean and standard deviation extremely close to the original data. The systematic equal sampling performed the best versus the population data, followed by the stratification method. Systematic unequal sampling using inclusion probabilities fared the worst in the comparison. This may be attributed to the resampling loop if all samples were not achieved in the first pass, which biases leading values.