# **Dataset Description**

The Indian monsoon affects crop and water supplies across India and surrounding countries, thus impacting the lives of over a billion people [1]. The system brings large amounts of rainfall during the months of June through September with extended dry periods in November through February.  Meteorologists have shown that the timing of the heaviest rains and driest days depend on subtropical westerly jet stream location within the continent.  Monthly precipitation totals in the 36 meteorological subdivisions of India [2] from 1901 to 2017 compose the data[3]. Suppose we introduce the India precipitation dataset with the goal of describing total precipitation during the monsoon months of June-September.

Diagram

Description automatically generated

# **Data Exploration and Visualization**

**Prompt I**: Familiarize yourself with the provided data. Then, use the data to calculate the following quantities

1. Median total precipitation during Monsoon months
2. Mean total precipitation during Monsoon months
3. Standard deviation of total precipitation during Monsoon months
4. Using the calculated answers above, do you expect that the total precipitation in India during Monsoon months that was observed from 1901-2017 to have a symmetric, left-skewed or right-skewed distribution? Briefly explain why.

**Answer I**:

1. 880.6 mm
2. 1063.9 mm
3. 706.7 mm
4. Since the mean is larger than the median in the data, we may expect to see a right-skewed distribution.

**Prompt II**:

Create a histogram which displays the total precipitation during monsoon months in India. Then, explain why the histogram is not the most informative display of total monsoon precipitation, given the information in this dataset. Finally, suggest a way to improve how the information is displayed.

**Answer II**:

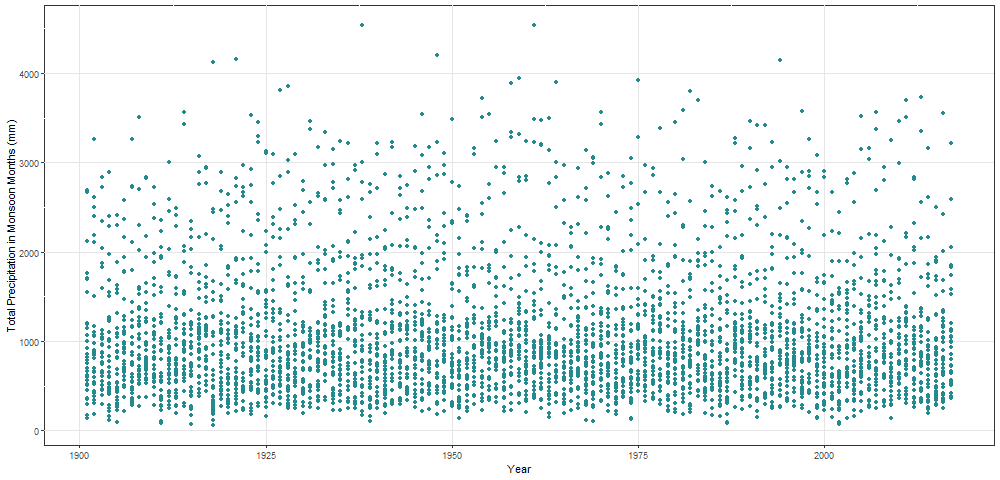
Chart, histogram

Description automatically generated

There are several other variables within the dataset and the information that they carry is neglected within this display. For example, the dataset contains information about the year of data collection, the subdivision, and other monthly precipitation totals. One display which would potentially be more informative is a scatterplot showing total monsoon precipitation by year.

**Prompt III**:

Consider the scatterplot of total precipitation during monsoon months vs. year provided below.



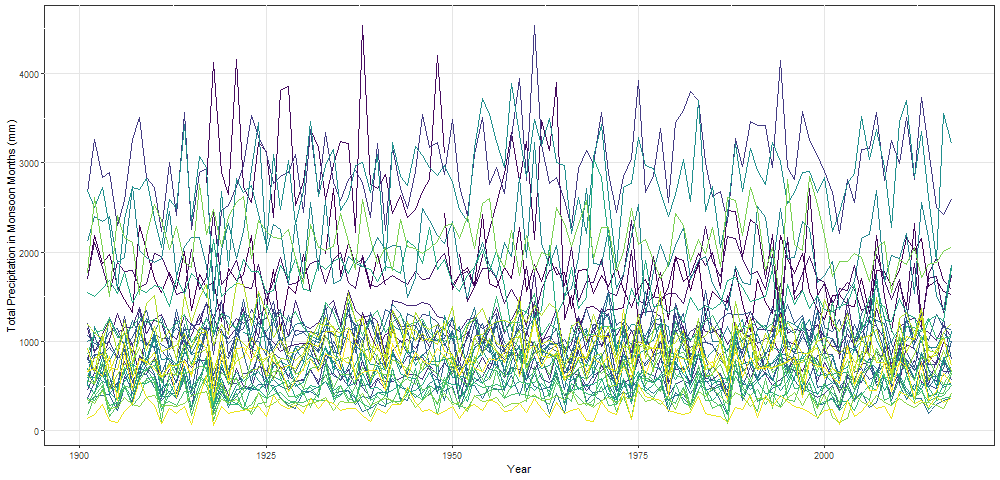
May we use this scatterplot as evidence that it seems reasonable to assume all of the observations/cases are independent? Why or why not?

**Answer III**:

This scatterplot alone cannot be used to support the broad statement that the individual cases of total precipitation are independent. Although we do not see any clear relationship between all of the cases and time, we cannot view information related to subdivision. For example, we might guess that subdivisions that are proximate may receive similar precipitation totals each year and the total precipitation within any subdivision might be related year-to-year.

**Prompt IV**:

Notice that these data actually contain yearly sequences of monsoon season precipitation values for each location:  36 time series. Time series are better displayed with the points connected by lines. The graph below displays total precipitation observed during the months of June-September vs. year for all India subdivisions. The timeseries for each subdivision is denoted by a different line color.



Unfortunately, this timeseries plot also does not adequately address all dependencies within the data. What types of dependence are not shown? How could these dependencies be further explored?

**Answer IV**:

Although the timeseries plot displays relationships year-to-year within each location, it does not display potential dependencies among locations. Moreover, it does not adequately display the types of dependence that could be among cases within any given location. For example, the plot does not directly show if a previous year’s rainfall is related to the current year’s rainfall within any location.

# **Cautioning Univariate Statistical Inference**

**Prompt V:**

Consider the total monsoon precipitation observed in India subdivisions from 1901 – 2017. Explain why it is unreasonable to use any traditional univariate statistics method (i.e. normal distribution, t-distribution, bootstrapping, randomization test) to create a confidence interval for the population mean.

**Answer V:**

All of the univariate methods we covered require that the cases within the data are independent. This is not necessarily reasonable. For example, there could be dependence among the observations across time, among locations, and within time and location.

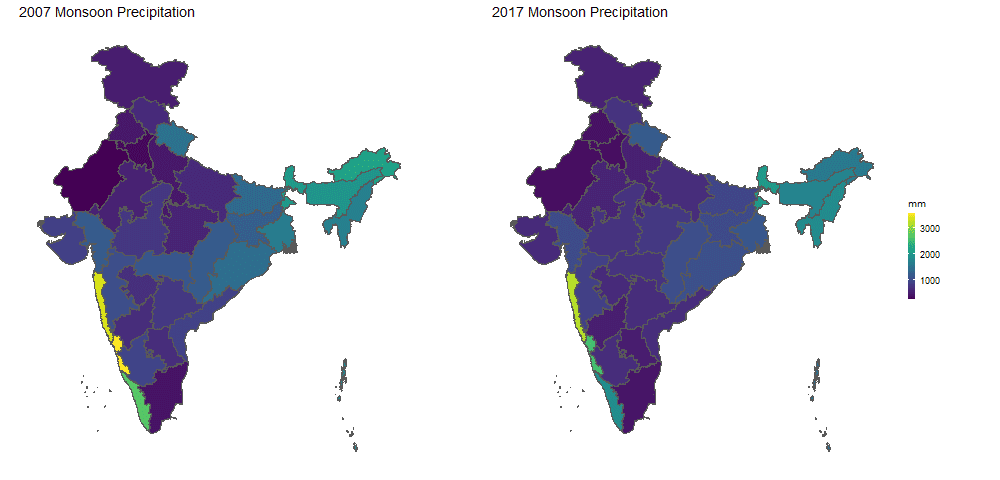
**Further Discussion of Prompt V:**

Here we can easily demonstrate the importance of the independence assumption and implications of ignoring dependence in analysis.  Suppose that we attempt to remove spatial dependence by just focusing on the n=61 Arunachal Pradesh precipitation values from 1957-2017.  The CLT method to build a confidence interval for the population mean has a standard error of . When fitting a time series model to account for dependence while also obtaining an estimate for the intercept (the population mean), the standard error of the estimate is 204.4.  The temporal dependence resulted in an estimated standard error using CLT to be about 2.5 times smaller than the appropriate standard error from the time series model.  Thus, using the CLT approach results in much more confidence in the estimate than reasonable.

This example demonstrates that the variance calculation depends on whether observations are independent.  While it is true that the statistical estimator from the appropriate time series model here, AR(2) or autoregressive of order 2, is not exactly the sample mean since the model typically uses maximum likelihood estimation, we aren’t comparing apples to oranges.  There is no need to present details of the AR(2) time series model to make this point.

# **Inference for Two Population Means**

The graphs below contain maps of the total monsoon precipitation for the years 2007 and 2017.



**Prompt VI**:

Suppose that we would like to use the subset of the data displayed above to attempt to answer the question, “Is there evidence that the average total monsoon precipitation across India was less in 2017 than ten years earlier in 2007?” Explain why we need to use a statistical hypothesis test.

**Answer VI**:

This is not a question that can be answered by just directly comparing the observed averages from 2007 to 2017. These data are summaries of the recorded precipitation that occurred across large land areas, not measurements of the total monsoon precipitation across every square centimeter of land within India during those years. Thus, it is necessary to assume that data are representative and use a hypothesis test for population means

**Prompt VII**:

Suppose that we use the subset of the data for years 2007 and 2017 (displayed in the maps above) to test the following set of hypotheses

H0: µ­2007 ≤ µ2017 vs. H1: µ­2007 > µ2017

Explain why a two independent sample test for population means is inadequate for these data.

**Answer VII**:

Observations of total precipitation are paired or grouped by subdivision within India. Some subdivisions may tend to receive more or less monsoon precipitation, as we can see in the provided maps. Thus, observations of precipitation across the years are not necessarily independent. The two independent sample test requires that observations (or errors) must be independent, which would not be reasonable with the paired structure.

**Prompt VIII**:

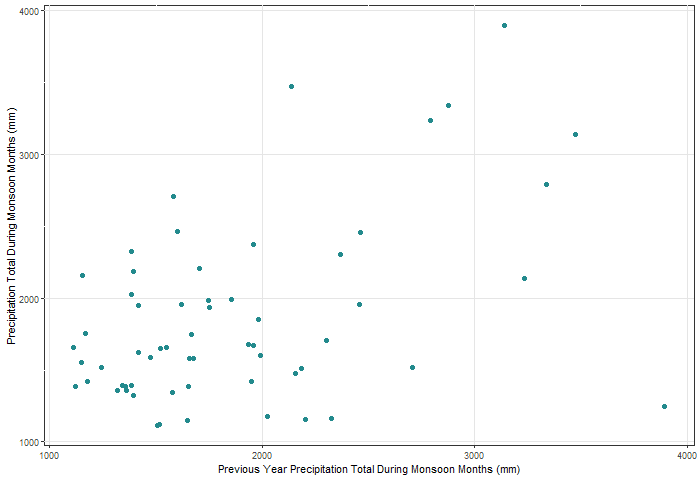
One of the assumptions of the paired test for population means is that the remaining error is independent. Explain why this condition is likely unreasonable within the Indian Monsoon data displayed in the provided maps.

**Answer VIII**:

In order to make the assumption that the remaining error is independent within the paired test, we must be able to assume that the cases/rows associated with the differences are independent. However, the rows are still associated with location and some locations are more proximate. Rainfall in proximate locations may be strongly related.

# **Inference with Simple Linear Regression**

Now, let’s consider total precipitation data from only the Arunachal Pradesh subdivision during Monsoon months. To avoid missing observations within the time series, the data here are limited to cases from 1957-2017. If there was no dependence within this subset of data, one aspect we would expect to observe is no relationship between any observation and the previous observation of total precipitation.



The scatterplot does depict a weak positive linear relationship between precipitation total and what was observed in the previous year within Arunachal Pradesh.  Thus, we have some qualitative evidence of dependence among cases.

**Prompt IX:**

Consider the data composing the scatterplot above. Explain why there isn’t an issue with spatial dependence in these data.

**Answer IX:**

Creating this subset does remove spatial dependence, since we have limited ourselves to one location. Thus, using such a dataset will limit the scope of statistical inference to this specific location (Arunachal Pradesh) instead of the larger area (India). With the other subdivision’s information removed, we don’t have any information related to other locations to use or worry about with respect to dependence.

The Arunachal Pradesh subset clearly contains temporal measurements of precipitation.  Thus, incorporating an aspect related to time into the model would be prudent when attempting to describe dependence so that independent errors may be assumed.  One approach would be to consider the scatterplot above.  The display motivates utilizing a simple linear regression model to describe the monsoon precipitation total observed during any given year by using the previous year’s monsoon total.  That is, if *Y(t)* is the total precipitation at time *t*, then estimate the simple linear model

*Y(t) = β0 + β1 Y(t-1) + ϵ(t)*

But, the result of the simple linear regression model fit, using ordinary least squares, is provided below.

Coefficients:

              Estimate Std. Error t value Pr(>|t|)

(Intercept)  1017.9319   226.8034   4.488 3.46e-05 \*\*\*

PreviousYear    0.4453     0.1146   3.884 0.000266 \*\*\*

Residual standard error: 562.7 on 58 degrees of freedom

Multiple R-squared:  0.2064, Adjusted R-squared:  0.1928

F-statistic: 15.09 on 1 and 58 DF,  p-value: 0.0002657

Prompt X:

Even though we have attempted to directly model temporal dependence within the Arunachal Pradesh data through the simple linear regression, explain why the condition that errors are independent is questionable at best.

Answer X:

The implemented model uses previous observations of total monsoon precipitation to predict the upcoming total monsoon precipitation, specifically the observation from the previous year. However, without further exploration it is not guaranteed that the total monsoon precipitation only depends on one previous year. That is, exploring larger time lag relationships should also be considered.

**Further discussion of Prompt X:**

Note that this simple linear model actually translates to the AR(1), autoregressive model of order 1, time series model.  The prompt V further discussion actually modeled these data with an AR(2) model, so we know that this introductory analysis is not appropriate.

This prompt is extremely challenging for students.  A conceptual discussion of this prompt is not unreasonable for undergraduates in a first statistics course.  Attempting to quantify the remaining dependence within the data could be achieved within the introductory statistics course by considering the remaining correlations among residuals.  At lag 2, there is a slightly significant positive correlation value of 0.3.  Perhaps a similar scatterplot could be used to help display this information to students, if discussing in-depth.

Note that climate patterns with longer temporal dependence such as El-Niño Southern Oscillation (ENSO) are well-studied and also have been examined specifically with respect to the Indian Monsoon in several sources [4, 5, 6, 7, 8].  These types of studies may be interesting to supplement and support discussion of Prompt X.

# **References**

[1] Krishnamurti, T. (2015).Indian monsoon. <https://www.britannica.com/science/Indian-monsoon>. Last Access: 05-25-21

[2] Rao, K., Rakesh, V., Prasad, K., and Mohapatra, G. (2020). Identification of vegetation responses to soil moisture, rainfall, and lulc over different meteorological subdivisions in India using remote sensing data. *Theoretical and Applied Climatology*,**142**.

[3] Open Government Data Platform India (2017). Rainfall in India. https://data.gov.in/resources/sub-divisional-monthly-rainfall-1901-2017. Last Access: 05-25-21

[4] Schulte, J., Policelli, F., and Zaitchik, B. (2021). A continuum approach to understanding changes in the ENSO–Indian monsoon relationship. *Journal of Climate*, **34(4)**, 1549 – 1561.

[5] Feba, F., Ashok, K., and Ravichandran, M. (2019). Role of changed Indo-pacific atmospheric circulation in the recent disconnect between the Indian summer monsoon and ENSO. *Climate Dynamics*, **52**, 1461–1470.

[6] Kumar, K. K., Rajagopalan, B., and Cane, M. A. (1999). On the weakening relationship between the Indian monsoon and ENSO. *Science*, **284(5423)**, 2156–2159.

[7] Krishnamurthy, V. and Goswami, B. N. (2000). Indian monsoon–ENSO relationship on interdecadal timescale. *Journal of Climate*,**13(3)**, 579– 595.

[8] Kripalani, R. H. and Kulkarni, A. (1997). Rainfall variability over south–east Asia—connections with Indian monsoon and ENSO extremes: new perspectives. *International Journal of Climatology*,**17(11)**, 1155–1168