Air Quality Index by State

Pranjal Singh

Probability and Applied Statistics

Professor Byron Hoy

November 20th, 2024

Summary

This report provides an analysis of air quality trends across the United States from 1980 to 2022, utilizing a dataset that details the Air Quality Index (AQI) for each state. Our goal is to better understand how air quality changes over time and what this means for public health and the environment. By using various statistical methods, we aim to identify the key factors that influence air quality in different regions.

Gaining insights into these trends is important for policymakers and environmental organizations. The findings from this analysis can contribute to the development of better strategies for monitoring, controlling, and reducing air pollution. This is vital for safeguarding public health and ensuring that communities across the nation have access to cleaner air.

This report seeks to help ongoing efforts to improve air quality management practices. It also offers valuable recommendations for future environmental policies. Our goal is to contribute to cleaner air across the United States, benefiting the well-being of individuals and communities alike.

Chapter 2

Section 2.3 - A Review of Set Notation

From the study of air quality data, we have the following information for the year 2022. A total of 50 states were monitored. 15 states had a maximum AQI (Air Quality Index) above 150, indicating at least one day of unhealthy air quality. 30 states had more than 100 “Good” AQI days (where the AQI was 50 or below). 10 of these states with more than 100 “Good” AQI days also had a maximum AQI above 150.

Find the number of states in 2022 that were:

a. States with more than 100 “Good” AQI days, states with a maximum AQI above 150, or both.

b. States with more than 100 “Good” AQI days but did not have a maximum AQI above 150.

c. States with a maximum AQI of 150 or below.

Answer:

a) $30 + 15 - 10 = 35$ states

b) $30 - 10 = 20$ states

c) $50 - 15 = 35$ states

Section 2.4 - A Probabilistic Model for an Experiment: The Discrete Case

The proportions of different AQI categories – Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous. The proportions of days in each category are approximately 0.30 (Good), 0.50 (Moderate), 0.10 (Unhealthy for Sensitive Groups), 0.07 (Unhealthy), 0.02 (Very Unhealthy), and 0.01 (Hazardous), respectively. A single day is chosen at random from the year 2022 for this state.

a. List the sample space for this experiment.

b. Make use of the information given above to assign probabilities to each of the simple events.

c. What is the probability that the chosen day at random falls under either Good or Unhealthy AQI categories?

Answer:

a) S = {Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, Hazardous}

b) P(Good) = 0.30, P(Moderate) = 0.50, P(Moderate) = 0.50, P(Unhealthy for Sensitive Groups) = 0.10, P(Unhealthy for Sensitive Groups) = 0.10, P(Unhealthy) = 0.07, P(Unhealthy) = 0.07, P(Very Unhealthy)=0.02, P(Very Unhealthy) = 0.02, P(Hazardous)=0.01, P(Hazardous) = 0.01

c) P(Good or Unhealthy) = P(Good) + P(Unhealthy) = 0.30 + 0.07 = 0.37

Section 2.5 - Calculating the Probability of an Event: The Sample-Point Method

Four states are being considered for a special air quality monitoring program in 2022. One of these states is known historically for having higher pollution levels. The program will select two of these states at random for intensive monitoring.

a. List the possible outcomes for this experiment.

b. Assign reasonable probabilities to the sample points, assuming each state has an equal chance of being selected.

c. Find the probability that State X is selected for the monitoring program.

Answer:

a) S = {(A,B),(A,C),(A,X),(B,C),(B,X),(C,X)}

b) 1/6

c) 0.5 = 50%

Section 2.6 - Tools for Counting Sample Points

A research team plans to set up air quality monitoring stations in two phases. In the first phase, they will choose from 6 different urban areas, and in the second phase, they will select from 7 different rural areas. Each urban and rural area will have one station, and the selections for urban and rural areas are made independently. How many different sample points of urban and rural areas can the research team choose for setting up these stations?

Answer:

$7 \times 6 = 42$ arrangements

Section 2.7 – Conditional Probability and Independence of Events

Consider that 1000 air quality measurements are recorded randomly one at a time from different monitoring stations across a region. If the first two measurements recorded both exceed an AQI of 150, what is the probability that the next three measurements will also exceed an AQI of 150?

Answer:

$P = 98/998 \times 97/997 \times 96/996 \approx 0.00092$

Section 2.8 - Two Laws of Probability

Two events A (New Jersey has more than 100 "Unhealthy for Sensitive Groups Days") and B (New Jersey has more than 50 "Hazardous Days") are such that $P(A) = 0.2$, $P(B) = 0.3$, and $P(A \cup B) = 0.4$.

Find the following:

a. $P(A \cap B)$

b. P(A|B)

Answer:

$P(A \cap B) = 0.5 - 0.4 = 0.1$

$P(A|B) = 0.1/0.3 = 1/3 = .333$

Section 2.10 - The Law of Total Probability and Bayes' Rule

A dataset tracks air quality across various states, categorizing them into two types: states with industrial economies and states with service-based economies. Assume 40% of the states have industrial economies and 60% have service-based economies. From historical data, 30% of the states with industrial economies and 70% of the states with service-based economies exceeded air quality safety thresholds last year. A state is chosen at random from those that exceeded safety thresholds. Find the conditional probability that this state has a service-based economy.

Answer:

$P(A|B) = ((.70)(.60))/.54 = 0.42/.54 = 42/54 = 0.7778$

Chapter 3

Section 3.2 - The Probability Distribution for a Discrete Random Variable

In a class designed to raise awareness about air quality, students are given a task to match three states with their respective highest recorded AQI value for the year from a list of three possible values. If the students assign the AQI values at random to the three states, find the probability distribution for Y, the number of correct matches.

Answer:

$p(0) = 2/6$

$p(1) = 3/6$

$p(3) = 1/6$

Section 3.4 - The Binomial Probability Distribution

Consider a series of measures taken to reduce air pollution levels in a state, each with a known probability of success p. These measures are applied independently for five consecutive years. Calculate the probability that:

a. all five measures are successful in reducing pollution levels if p=0.8

b. exactly four of the measures are successful if p=0.6

c. less than two of the measures are successful if p=0.3

Answer:

a) $P(X = 5) = \binom{5}{5}(.8)^5(1 - .8)^0 = 0.32768$

b) $P(X = 4) = \binom{5}{4}(.6)^4(1 - .6)^1 = 0.2592$

c) $P(X < 2) = P(X{=}0) + P(X{=}1) = 0.16807 + 0.36015 = .528$

Section 3.5 - The Geometric Probability Distribution

Suppose that 30% of the states have implemented air quality monitoring systems. States are evaluated randomly to check for these. Find the probability that the first state with an advanced monitoring system is evaluated on the fifth evaluation.

Answer:

$P(X = 5) = (1 - .3)^4(.3) = .07203$

Section 3.6 - The Negative Binomial Probability Distribution

A study indicates that the likelihood of a state exceeding a specific pollution threshold (considered a "significant pollution event") is 0.2.

a. What is the probability that the first significant pollution event is detected on the third state reviewed?

b. Find the mean and variance of the number of states that must be reviewed if the goal is to identify three states with significant pollution events.

Answer:

a) $P(X = 3) = (1 - .2)^2(.2) = .128$

b) $\mu = 15,\ \sigma^2 = 60$

Section 3.7 - The Hypergeometric Probability Distribution

A state's environmental agency oversees ten air quality monitoring stations, four of which are known to malfunction frequently due to hardware issues. An inspector chooses five of these stations at random for an annual performance review, under the assumption that all are functioning correctly. What is the probability that all five of the stations selected are non-defective?

Answer:

$P(X=5) = \frac{\binom{6}{5}\binom{10-6}{5-5}}{\binom{10}{5}} = 1/42 = .02380$

Section 3.8 - The Poisson Probability Distribution

Let Y denote the number of significant pollution events in a region per year that exceed a critical AQI level. Assume Y has a Poisson distribution with a mean λ=2 events per year.

Find the following:

a. Probability of exactly 4 events.

b. Probability of 4 or more events.

c. Probability of fewer than 4 events.

d. Probability of 4 or more events given there are at least 2 events.

Answer:

a) $P(Y = 4) = .090$

b) $P(Y \geq 4) = .143$

c) $P(Y < 4) = .857$

d) $P(Y \geq 4 \mid Y \geq 2) = .241$

Section 3.11 - Tchebysheff's Theorem

Suppose air quality monitoring over three consecutive days in a particular state is conducted to check for "Good" AQI days. Let's assume the probability of a day being classified as "Good" in terms of AQI is 50%, the same as the flip of a balanced coin. Define Y as the number of days observed with "Good" AQI out of three days. Use the formula for the binomial probability distribution to calculate the probabilities associated with $Y = 0$, $Y = 1$, $Y = 2$, and $Y = 3$.

Answer:

$p(0) = \binom{3}{0}(.5)^0(.5)^3 = 1/8 = 0.125$

$p(1) = \binom{3}{1}(.5)^1(.5)^2 = 3/8 = 0.375$

$p(2) = \binom{3}{2}(.5)^2(.5)^1 = 3/8 = 0.375$

$p(3) = \binom{3}{3}(.5)^3(.5)^0 = 1/8 = 0.125$

Chapter 4

Section 4.4 - The Uniform Probability Distribution

An air quality data transmission from a monitoring station is sent at random within a one-hour interval. The station's data transmission system was offline for maintenance for 15 minutes during this one-hour period. What is the probability that the data transmission occurred when the station was not offline?

Answer:

3/4

Section 4.6 - The Gamma Probability Distribution

Historical data suggests that the intervals between days when air quality reaches hazardous levels in a particular city have an approximately exponential distribution. Assume that the mean interval between such hazardous air quality days is 44 days.

a. If a day with hazardous air quality occurred on July 1 of a randomly selected year, what is the probability that another hazardous air quality day will occur within that same month (i.e., within the next 30 days)?

b. What is the variance of the times between hazardous air quality days?

Answer:

a) $P(Y \leq 30) = \frac{1}{44}e^{-30/44} = .5057$

b) $\beta^2 = 44^2 = 1936$

Chapter 5

Section 5.2 - Bivariate and Multivariate Probability Distributions

Funding for two air quality improvement projects is being randomly assigned to one or more of three organizations: A, B, and C. Each organization can receive funding for 0, 1, or 2 projects. Let $Y1$ denote the number of projects funded for Organization A and $Y2$ the number for Organization B.

a. Find the joint probability function for $Y1$ and $Y2$.

b. Find F(1, 0).

Answer:

a)

		y1	
	0	1	2
0	1/9	2/9	1/9
1	2/9	2/9	0
2	1/9	0	0

b) F(1, 0) = 1/9+2/9= ⅓

References

https://www.kaggle.com/datasets/adampq/air-quality-index-by-state-1980-2022