**Air Quality Index by State 1980 - 2022**

Rocco Vulpis

Probability and Applied Statistics

Professor Byron Hoy

April 25, 2024

Contents

[Summary 2](#_Toc165233696)

[Chapter 2 3](#_Toc165233697)

[Section 2.3 - A Review of Set Notation 3](#_Toc165233698)

[Section 2.4 - A Probabilistic Model for an Experiment: The Discrete Case 4](#_Toc165233699)

[Section 2.5 - Calculating the Probability of an Event: The Sample-Point Method 5](#_Toc165233700)

[Section 2.6 - Tools for Counting Sample Points 5](#_Toc165233701)

[Section 2.7 – Conditional Probability and Independence of Events 5](#_Toc165233702)

[Section 2.8 - Two Laws of Probability 6](#_Toc165233703)

[Section 2.10 - The Law of Total Probability and Bayes’ Rule 6](#_Toc165233704)

[Chapter 3 7](#_Toc165233705)

[Section 3.2 - The Probability Distribution for a Discrete Random Variable 7](#_Toc165233706)

[Section 3.4 - The Binomial Probability Distribution 7](#_Toc165233707)

[Section 3.5 - The Geometric Probability Distribution 7](#_Toc165233708)

[Section 3.6 - The Negative Binomial Probability Distribution 8](#_Toc165233709)

[Section 3.7 - The Hypergeometric Probability Distribution 8](#_Toc165233710)

[Section 3.8 - The Poisson Probability Distribution 8](#_Toc165233711)

[Section 3.11 - Tchebysheff’s Theorem 9](#_Toc165233712)

[Chapter 4 10](#_Toc165233713)

[Section 4.2 - The Probability Distribution for a Continuous Random Variable 10](#_Toc165233714)

[Section 4.3 - Expected Values for Continuous Random Variables 10](#_Toc165233715)

[Section 4.4 - The Uniform Probability Distribution 11](#_Toc165233716)

[Section 4.6 - The Gamma Probability Distribution 11](#_Toc165233717)

[Chapter 5 12](#_Toc165233718)

[Section 5.2 - Bivariate and Multivariate Probability Distributions 12](#_Toc165233719)

[Section 5.3 - Marginal and Conditional Probability Distributions 12](#_Toc165233720)

# Summary

This report explores air quality trends across the United States from 1980 to 2022 by analyzing a dataset on 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 different statistical methods, we aim to identify key trends and factors that affect air quality.

Understanding these patterns helps policymakers and environmental agencies make informed decisions about how to manage air quality effectively. The insights gained from our analysis can help improve strategies for monitoring and controlling air pollution. This is crucial for protecting public health and ensuring that communities have clean air to breathe.

By providing clear and actionable information, this report supports efforts to enhance air quality management and offers valuable guidance for future environmental policies. Ultimately, we want to help improve the air quality in different regions of the country, benefiting everyone who lives there.

# Chapter 2

## Section 2.3 - A Review of Set Notation

From a study of air quality data across different states in the United States, 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 Air Quality Index (AQI) categories – Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous – for a particular state in the United States were recorded for the year 2022. 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 arrangements of urban and rural areas can the research team choose for setting up these stations?

**Answer**

7 x 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 = × × ≈ 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 ∪ B) = 0.4.

Find the following:

a. *P*(*A*∩*B*)

b. *P*(*A*∪*B*)

c. *P*(*A|B*)

**Answer**

*P(A∩B) =* 0.5 − 0.4 = 0.1

𝑃(𝐴∪𝐵) = *P*(*A*∪*B*) = 0.4

𝑃(𝐴∣𝐵) = = = .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) = = = = 0.7778

# Chapter 3

## Section 3.2 - The Probability Distribution for a Discrete Random Variable

In an educational program designed to raise awareness about air quality, students are given a task to match three states with their respective highest recorded AQI (Air Quality Index) 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) =

p(1) =

p(3) =

## 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 in 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) = = 0.32768

b) *P*(*X* = 4) = = 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 advanced air quality monitoring systems. States are evaluated sequentially for a special environmental compliance check, and the evaluations are conducted randomly from the pool of states. Find the probability that the first state with an advanced monitoring system is evaluated on the fifth evaluation.

**Answer**

*P*(*X* = 5) = (.30) = .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. What is the probability that the third significant pollution event is detected on the seventh state reviewed?

c. 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) = (.2) = .128

b) *P*(*X* = 7) = = 15 \* .008 \* .4096 = .049

c) μ = 15, σ2 = 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**

= .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**

*P*(*Y* = *k*) = e-*λ*

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

b) *P*(*Y* ≥ 4) = .143

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

d) *P*(*Y ≥* 4 | *Y* ≥ 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) = = = 0.125

p(1) = = = 0.375

p(2) = = = 0.375

p(3) = = = 0.125

# Chapter 4

### Section 4.2 - The Probability Distribution for a Continuous Random Variable

An environmental agency has a protocol for deploying emergency air quality interventions that are triggered based on the severity of pollution levels recorded each week. The agency has resources to handle up to 150 intervention events per week. Historical data show that the weekly demand for these interventions increases steadily as pollution worsens until it reaches 100 events, after which the frequency of demand levels off, even as pollution may continue to increase, until reaching the capacity at 150 events. If *Y* denotes weekly demand in hundreds of interventions, the relative frequency of demand can be modeled by the function:

*f*(*y*) =

a Find *F*(*y*).

b Find *P*(0 ≤ *Y* ≤ .5).

c Find *P*(.5 ≤ *Y* ≤ 1.2).

**Answer**

*F*(*y*) =

## Section 4.3 - Expected Values for Continuous Random Variables

The temperature 𝑌 at which a sensor in an air quality monitoring station activates is controlled by a thermostat. The probability density function (PDF) for this activation temperature is given by:

*f*(*y*) =

Find *E*(*Y* ) and *V*(*Y* ).

**Answer**

= 60

*V*(*Y* ) =

## 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**

## Section 4.6 - The Gamma Probability Distribution

**Question 4.93**

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 ≤ 30) = e-30/44 = .5057

b) *β2* = 442 = 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 |  |  | | | 0 | |
| 2 |  | 0 | 0 | | |

y1

b) *F*(1, 0) =

## Section 5.3 - Marginal and Conditional Probability Distributions

Consider the scenario where 𝑌1​ represents the proportion of the total air quality monitoring capacity that is deployed at the beginning of the week, and 𝑌2 represents the proportion of that capacity that actually recorded significant air quality events during the week. The joint density function of 𝑌1​ and 𝑌2​ is given by:

*f*(*y1, y2*) =

a. Find the marginal density function for 𝑌2.

b. For what values of y2 is the conditional density *f* (*y1| y2*) defined?

**Answer**

a) *f*2**(***y*2) = = - , 0 1

b) Defined over 1 if 1