4 Modeling Data as Random Variables and Populations as Probability Distributions

MTH 3240 Environmental Statistics

Spring 2020

Notes

Objectives

Objectives:

- Use discrete and continuous probability distributions to obtain probabilities involving random variables.
- Interpret the mean and standard deviation of a probability distribution.
- Recognize binomial and Poisson random variables.

Notes

Random Variables

- Any numerical variable whose value is determined by chance is called a ***random variable***.

  Examples:

  - The E. coli level in a **randomly selected** water specimen from a lake is a **random variable**.
  - The number of occupants in a **randomly selected** automobile is a **random variable**.

- Random variables can be ***discrete*** or ***continuous*** depending on whether the possible values for the variable are isolated numbers (e.g. integers) or a continuum.

Notes

- Random variables are said to be ***discrete*** if they can only take *integer* values, and ***continuous*** if they can take values on a *continuum*.

Notes

Introduction to Probability Distributions

- The set of values a random variable might take and their probabilities form the ***probability distribution*** of the random variable.
- Probability distributions are used to represent **populations** from which individuals are **randomly** selected.

Notes

- We'll use the following notation:
  - Upper case letters such as X, Y, and Z denote **random variables** (whose values *have yet to be determined*).
  - **Probabilities** involving a random variable X will be denoted $P(X = 3)$, $P(X \le 6.5)$, and so on.

Notes

Discrete Probability Distributions

Example

Here are vehicle occupancy rates on urban arterials and freeways in Miami-Dade County, Florida.

Number of Occupants	Percentage of Vehicles
1	82 %
2	12 %
3	4 %
4	2 %

We can interpret each percentage as the **probability** that a **randomly selected** vehicle will have 1, 2, 3, and 4 occupants, respectively.

Notes

Letting X be the **number of occupants** in a **randomly selected vehicle**, X is a **discrete random variable** with possible values 1, 2, 3, and 4.

The **probability distribution** of X is below.

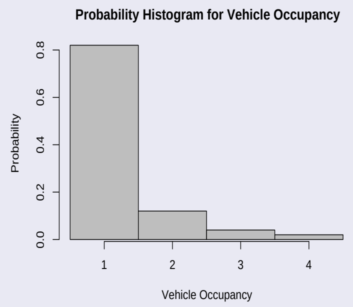
x	1	2	3	4
$P(X = x)$	0.82	0.12	0.04	0.02

For example, that the **probability** of a **randomly selected** vehicle having only **one** occupant is

$$P(X = 1) = 0.82.$$

The probability distribution is on the next slide as a ***probability histogram***.

Notes



Notes

This **probability distribution** represents the **population** of vehicles in Miami-Dade County.

Notes

Mean of a Discrete Probability Distribution

- We measure the **center** of a probability distribution by its ***mean***, denoted μ.
- If the bars in a probability histogram were weights, μ is the point along the x-axis at which they'd balance.
- The value of μ represents the value that the random variable takes **on average** .

Notes

- μ can be thought of as the **population mean** if the probability distribution represents a **population**.

Notes

Standard Deviation of a Discrete Probability Distribution

- We measure the **spread** in a probability distribution by its ***standard deviation***, denoted σ.
- A larger value of σ corresponds to a more spread-out distribution.
- The value of σ represents a **typical deviation** of a the randomly variable away from μ.
- The **square** of the standard deviation is called the ***variance***, denoted σ^2.

Notes

- σ can be thought of as the **population standard deviation** if the probability distribution represents a **population**.

Notes

Theoretical Probability Distributions

- In the vehicle occupancy example, the **probability distribution** was based on *accurate information* about the **population** of vehicles.
- In the absence of such accurate information, we have to choose from a set of stock **theoretical distributions** the one that we *think* describes the population.
- The first step is to identify whether the variable is **discrete** or **continuous**.

Notes

- When the random variable is a **count**, it's **discrete**.

  Two commonly used **theoretical distributions** for **counts** are:

  1. The **binomial** distribution.
  2. The **Poisson** distribution.

Notes

- When the random variable is a **numerical measurement**, it's **continuous**.

  Two commonly used **continuous theoretical distributions** are:

  1. The **normal** distribution.
  2. The **lognormal** distribution.

- We'll look at these four theoretical probability distributions one at a time.

Notes

The Binomial Distribution

- **Examples of Binomial Random Variables**:
  - The number of animal traps, among the 10 traps set, that catch animals.
  - The number of fish, among eight tested, that test positive for a certain disease.
  - The number of sites, among 12 sites visited, at which a certain bird species is present.

Notes

The Binomial Distribution

Conditions for a Binomial Random Variable:

1. There are a certain number of **trials** n.
2. Each trial has **two possible outcomes**, *success* and *failure*, say.
3. The trials all have the same **probability** p of resulting in a **success**. Thus the **probability** of a **failure** is $1 - p$.
4. The trials are **independent**, meaning their outcomes don't affect each other.

Notes

- Under these conditions, the random variable

  X = The **number of successes** among the n trials

  is a ***binomial*** random variable.

- We refer to n and p as the ***parameters*** of the binomial distribution.

Notes

- The two ***parameters*** of the **binomial** distribution, n and p, control the distribution's shape, center, *and* spread.

Notes

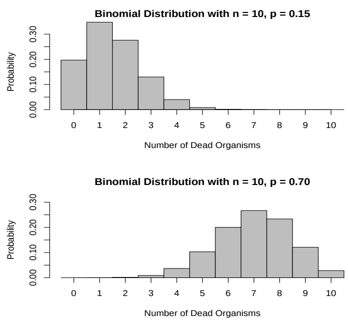


Figure: Probability histograms for two binomial distributions.

Notes

Example

The World Health Organization suggests that fish with mercury (Hg) concentrations greater than 0.5 mg/kg are **unsafe** for human consumption.

In the U.S., much of the fish consumed comes in the form of canned tuna, which is sometimes sold in packages of **four cans** (*trials*).

Each can is either **unsafe** or **safe** (*success* or *failure*).

If **four** randomly selected cans are tested, the **number of cans** that are **unsafe** is a **binomial** random variable.

Notes

- Binomial distribution probabilities $P(X = x)$ can be obtained using any of the following:
  - A table
  - Statistical software
  - A formula (the ***binomial probability function***).

Notes

The Poisson Distribution

- The *Poisson* distribution is used to model **counts** that are are either:
  - **Counts of events** occurring in a certain **period of time**, where the events occur at random in time points, or
  - **Counts of events** occurring in a given **spatial area**, where the events occur at random spatial points.

Notes

- **Examples**:
  - The number of flash floods during a 100-year period.
  - The number of beetles in a 1 m^2 quadrat.
  - The number of shooting stars in the night sky during a one hour period.
  - The number of trees of a certain species on a 100 m^2 plot of land.
  - The number of patients admitted for respiratory problems at a hospital during a month.

Notes

Conditions for a Poisson Random Variable:

1. Events occur at random time points or at random spatial points. The (temporal) rate or (spatial) density of their occurrence is approximately constant (doesn't change over time or across space).
2. The events occur independently of each other in time or space, e.g. they don't occur in "clusters" ("clumps") in time or space.

Notes

- Under these conditions, the random variable

  X = The **number of events that occur** in a specified period of time (or spatial area)

  is a ***Poisson*** random variable.

Notes

- The shape, center, and spread are controlled by the (one) ***parameter*** of the distribution, μ (which is the mean of the distribution).

Notes

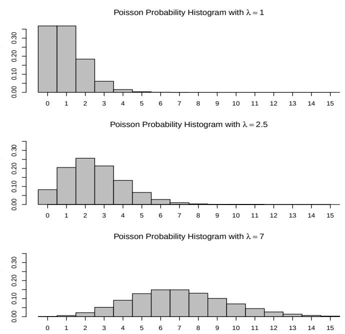


Figure: Probability histograms for three Poisson distributions.

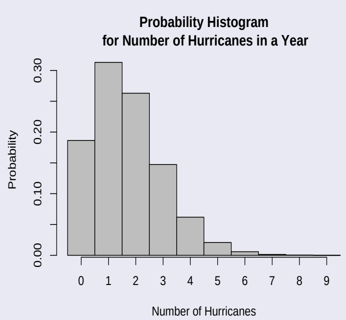
Notes

Example

In any given year, hurricanes that make landfall on the continental U.S. is a random variable that could be modeled by a **Poisson** distribution with parameter $\mu = 1.68$ (based on historical records).

This distribution is depicted below.

Notes



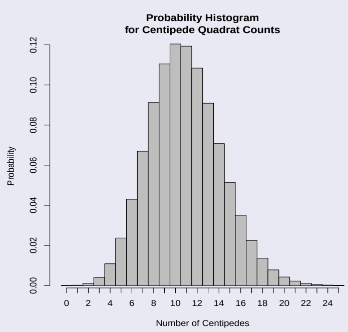
Notes

Example

In a study of the spatial dispersion of the centipede *Lithobius muticus*, the number of centipedes in a randomly selected $1\,\text{m}^2$ quadrat could be modeled by a **Poisson** distribution with parameter $\mu = 10.5$ (based on prior studies).

This distribution is shown below.

Notes



Notes

- Poisson distribution probabilities $P(X = x)$ can be obtained using any of the following:
  - A table
  - Statistical software
  - A formula (the ***Poisson probability function***).

Notes

Exercise

Each random variable below is a **count**. Identify whether it would follow a **binomial** or a **Poisson probability distribution**.

a) At a certain vehicle emissions testing center, let X be the number of cars that **pass** the test out of the next **10** cars that are tested.

b) Let X be the number of meteorites larger than one ft in diameter that strike the Earth **in a given month**.

Notes

c) Let X be the number of *Philonthus fuscipennis* beetles X in a **1 m^2 area**.

d) The **six** public drinking fountains in a town are tested for a hazardous contaminant. Let X be the number of fountains that are found to be **safe**.

Notes

Notes

Notes

Notes