Statistics

Lecture 08: |Quartiles - Percentiles

Prof. Dr. Mostafa Elhosseini

https://youtube.com/drmelhosseini

Mode and the Range

- The **mode** and the **range** are summary statistics that are of limited use but are occasionally seen.
- The sample **mode** is the **most frequently occurring value** in a sample. If **several values occur** with equal frequency, **each one is a mode**.
- The **range** is the **difference between the largest and smallest values** in a sample. It is a **measure of spread**, but it is **rarely used**, because it **depends only on the two extreme values** and provides no information about the rest of the sample.

Mode and the Range

In the article "Evaluation of Low-Temperature Properties of HMA Mixtures" (P. Sebaaly, A. Lake, and J. Epps, *Journal of Transportation Engineering*, 2002: 578–583), the following values of fracture stress (in megapascals) were measured for a sample of 24 mixtures of hot-mixed asphalt (HMA).

30	75	79	80	80	105	126	138	149	179	179	191
223	232	232	236	240	242	245	247	254	274	384	470

- Find the modes and the range for the sample
- There are **three modes: 80, 179, and 232**. Each of these values appears twice, and no other value appears more than once. The range is **470 – 30 = 440**

Quartiles

- The median divides the sample in half. Quartiles divide it as nearly as possible into quarters.
- A sample has three quartiles. There are several different ways to compute quartiles, but all of them give approximately the same result.
- The simplest method when computing by hand is as follows: Let n represent the sample size. Order the sample values from smallest to largest. To find the first quartile, compute the value $0.25(n + 1)$. If this is an integer, then the sample value in that position is the first quartile. If not, then take the average of the sample values on either side of this value.

Quartiles

- The third quartile is computed in the same way, except that the value 0.75(n+1) is used. The second quartile uses the value 0.5(n + 1).
- The second quartile is identical to the median.

Quartiles

In the article "Evaluation of Low-Temperature Properties of HMA Mixtures" (P. Sebaaly, A. Lake, and J. Epps, *Journal of Transportation Engineering*, 2002: 578–583), the following values of fracture stress (in megapascals) were measured for a sample of 24 mixtures of hot-mixed asphalt (HMA).

30	75	79	80	80	105	126	138	149	179	179	191
223	232	232	236	240	242	245	247	254	274	384	470

- Find the first and third quartiles of the asphalt data

Quartiles

- The sample size is n = 24. To find the first quartile, compute (0.25)(25) = 6.25.
- The first quartile is therefore found by averaging the 6th and 7th data points, when the sample is arranged in increasing order.
- This yields (105 + 126)/2 = 115.5.
- To find the third quartile, compute (0.75)(25) = 18.75.
- We average the 18th and 19th data points to obtain (242 + 245)/2 = 243.5.

Percentiles

- The pth percentile of a sample, for a number p between 0 and 100, divides the sample so that as nearly as possible p% of the sample values are less than the pth percentile, and (100 − p)% are greater
- Order the sample values from smallest to largest, and then compute the quantity (p/100)(n + 1), where n is the sample size.
- If this quantity is an integer, the sample value in this position is the pth percentile. Otherwise average the two sample values on either side.

Percentiles

- Note that the first quartile is the 25th percentile, the median is the 50th percentile, and the third quartile is the 75th percentile
- **Percentiles** are often used to **interpret scores on standardized tests**. For example, if a student is informed that her score on a college entrance exam is on the 64th percentile, this means that 64% of the students who took the exam got lower scores

Percentiles

In the article "Evaluation of Low-Temperature Properties of HMA Mixtures" (P. Sebaaly, A. Lake, and J. Epps, *Journal of Transportation Engineering*, 2002: 578–583), the following values of fracture stress (in megapascals) were measured for a sample of 24 mixtures of hot-mixed asphalt (HMA).

30	75	79	80	80	105	126	138	149	179	179	191
223	232	232	236	240	242	245	247	254	274	384	470

- Find the 65th percentile of the asphalt data

Percentiles

- The sample size is n = 24. To find the 65th percentile, compute (0.65)(25) = 16.25.
- The 65th percentile is therefore found by averaging the 16th and 17th data points,
- when the sample is arranged in increasing order. This yields (236 + 240)/2 = 238

Note that…

- The **summary statistics** are sometimes called **descriptive statistics** because they describe the data
- The standard error of the mean (SE Mean) is equal to the standard deviation divided by the square root of the sample size.
- This is a quantity that is not used much as a descriptive statistic, although it is important for applications such as constructing confidence intervals and hypothesis tests

Summary Statistics for Categorical Data

- With **categorical** data, each sample **item is assigned a category** rather than a quantitative value.
- But to work with categorical data, **numerical summaries** are needed.
- The two most commonly used ones are the **frequencies and the sample proportions** (sometimes called **relative frequencies**).
- The frequency for a given category is simply the number of sample items that fall into that category.
- The sample proportion is the frequency divided by the sample size

Summary Statistics for Categorical Data

- A process manufactures crankshaft journal bearings for an internal combustion engine. Bearings whose thicknesses are between 1.486 and 1.490 mm are classified as conforming, which means that they meet the specification. Bearings thicker than this are reground, and bearings thinner than this are scrapped. In a sample of **1000 bearings**, **910 were conforming**, **53 were regro**und, and **37 were scrapped**. Find the frequencies and sample proportions.
- The frequencies are 910, 53, and 37.
- The sample proportions are 910/1000 = 0.910, 53/1000 = 0.053, and 37/1000 = 0.037

Sample Statistics and Population Parameters

- any **numerical summary** used for a **sample** can be used for a **finite population**, just by **applying the methods** of calculation to the population values rather than the sample values.
  - One **small exception** occurs for the population **variance**, where we **divide by n rather than n – 1**.

Sample Statistics and Population Parameters

- There is a **difference in terminology** for numerical summaries of populations as opposed to samples
- **Numerical summaries of a sample** are called **statistics**, while **numerical summaries of a population** are called **parameters**.
- Of course, in practice, the **entire population is never observed**, so the **population parameters cannot be calculated directly**.
- Instead, the **sample statistics are used to estimate the values** of the population parameters.

To conclude

- A numerical summary of a sample is called a **statistic**.
- A numerical summary of a population is called a **parameter**.
- **Statistics are often used to estimate paramete**rs.