

6.29 Descriptive Statistics

Intro

How Do We Summarize Data?

Outline of What's Coming Up Next

Introduction to Statistics

Statistics forms a rational basis for decision-making using observed or experimental **data**. We make these decisions in the face of uncertainty.

Statistics helps us answer questions concerning:

- * The analysis of one population (or system)
- * The comparison of many populations.

Examples:

(1) Election polling.

(2) Coke vs. Pepsi.

(3) The effect of cigarette smoking on the probability of getting cancer.

(4) The effect of a new drug on the probability of contracting hepatitis.

(5) What drugs are most effective against AIDS?

(6) What's the most popular TV show during a certain time period?

(7) The effect of different heat-treating methods on tensile strength of steel.

(8) Which fertilizers improve crop yield?

(9) King of Siam — etc., etc., etc.

Idea (Election polling example): We can't poll every single voter. Thus, we take a **sample** of data from the **population** of voters, and try to make a reasonable conclusion based on that sample.

Statistics tells us how to conduct the sampling (i.e., how many observations to take, how to take them, etc.), and then how to draw conclusions from the sampled data.

Types of Data:

Continuous variables: Can take on any real value in a certain interval. For example, the lifetime of a light-bulb or the weight of a newborn child.

Discrete variables: Can only take on specific values. For example, the number of accidents this week at a factory or the possible rolls of a pair of dice.

How Do We Summarize Data?

It's nice to have lots of data. But sometimes it's too much of a good thing! Need to summarize.

Example: Grades on a test (i.e., raw data):

23	62	91	83	82	64	73	94	94	52
67	11	87	99	37	62	40	33	80	83
99	90	18	73	68	75	75	90	36	55

Stem-and-Leaf Diagram of grades. Easy way to write down all of the data. Saves some space, and looks like a sideways histogram.

9		9944100
8		73320
7		5533
6		87422
5		52
4		0
3		763
2		3
1		81

Grouped Data

Range	Freq.	Cumul. Freq.	Prop'n of obs'ns so far
0–20	2	2	2/30
21–40	5	7	7/30
41–60	2	9	9/30
61–80	10	19	19/30
81–100	11	30	1

Summary Statistics:

$n = 30$ observations

If X_i is the i th score, then the **sample mean** is

$$\bar{X} \equiv \sum_{i=1}^n X_i/n = 66.5.$$

The **sample variance** is

$$S^2 \equiv \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = 630.6.$$

In general, suppose that we sample i.i.d. data X_1, \dots, X_n from the population of interest.

Example: X_i is the lifespan of the i th lightbulb we observe.

We're most interested in measuring the “center” and “spread” of the underlying distribution of the data.

Measures of Central Tendency:

Sample Mean: $\bar{X} = \sum_{i=1}^n X_i/n$

Sample Median: The “middle” observation when the X_i 's are arranged numerically.

Example: 16, 7, 83 gives a median of 16.

Example: 16, 7, 83, 20 gives a “reasonable” median of $(16 + 20)/2 = 18$.

Sample median is less susceptible to “outlier” data than the sample mean. One bad number can spoil the sample mean’s entire day.

Example: 7, 7, 7, 672, 7 gives a sample mean of 100 and a sample mdn of 7.

Sample Mode: “Most common” value. Not the most useful measure sometimes.

Example: 16, 7, 20, 83, 7 gives a mode of 7.

Measures of Variation (dispersion, spread):

Sample Variance:

$$S^2 \equiv \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right),$$

the latter expression being easier to compute.

Sample Standard Deviation: $S = +\sqrt{S^2}$.

Sample Range: $\max_i X_i - \min_i X_i$.

Remark: Suppose the data takes p different values X_1, \dots, X_p , with frequencies f_1, \dots, f_p , respectively.

How to calculate \bar{X} and S^2 quickly?

$$\bar{X} = \sum_{j=1}^p f_j X_j / n$$

$$S^2 = \frac{\sum_{j=1}^p f_j X_j^2 - n\bar{X}^2}{n - 1}.$$

Example: Suppose we roll a die 10 times.

X_j	1	2	3	4	5	6
f_j	2	1	1	3	0	3

Then $\bar{X} = (2 \cdot 1 + 1 \cdot 2 + \dots + 3 \cdot 6)/10 = 3.7$

If the individual observations can't be determined in frequency distributions, you might just break the observations up into c intervals.

x_j interval	m_j	f_j
100–150	125	3
150–200	175	6
⋮	⋮	⋮

$$\bar{X} \approx \frac{\sum_{j=1}^c f_j m_j}{n}, \quad S^2 \approx \frac{\sum_{j=1}^c f_j m_j^2 - n \bar{X}^2}{n - 1}$$

What's Up Next, Doc?

Point Estimation

Confidence Interval Estimation

Hypothesis Testing