November 10, 2013

Contents

1 Introduction to Statistics

This chapter contains two separate but related themes, both to do with the understanding of data. The rst idea is to nd graphical representations for the data, which allow one to easily see its most important characteristics. The second idea is to nd simple numbers, like mean or inter-quartile range, which will summarise those characteristics.

Lecture 1 Part A About descriptive statistics

Lecture 1 Part B About inferential statistics

Lecture 1 Part C Variables

Lecture 2 Part A Parameters and Statistics

Lecture 2 Part B Categorical Data

Lecture 2 Part C Summation notation

Lecture 3 Part A Measurement scales

Lecture 3 Part B Computing the sample mean

Lecture 3 Part C Computing the sample median

2 Graphical Methods

Lecture 4 Part A Pie-Charts (?)

Lecture 4 Part B Relative Frequency and Cumulative Frequency

Lecture 4 Part C Histograms

Lecture 5 Part A Ogives

Lecture 5 Part B Stem and Leaf Plots

Lecture 5 Part C Boxplots

3 Descriptive Statistics

Lecture 6 Part A Interquartile Range

Lecture 6 Part B

Lecture 6 Part C

Lecture 7 Part A Types of Data

Lecture 7 Part B Range

Lecture 7 Part C Mean

Lecture 8 Part A Median and Mode

Lecture 8 Part B Outliers

Lecture 8 Part C Trimean and Trimmed Mean (Outliers)

Lecture 9 Part A Skew and Kurtosis

Lecture 9 Part B Summary

Lecture 9 Part C Spread

Lecture 10 Part A

Lecture 10 Part B Semi-Interquartile Range

Lecture 10 Part C Variance

4 Grouped Data

Lecture 10 Part A

Lecture 10 Part B

Lecture 10 Part C

5 Probability

Lecture 11 Part A Introduction to Probability

Lecture 11 Part B Counting and Permutations

Lecture 11 Part C Permuations without Repetition

Lecture 12 Part A

Lecture 12 Part B Axioms of Probability

Lecture 12 Part C Conditional Probability and Independent Events

Random experiment

- **Sample Space, S**. For a given experiment the sample space, S, is the set of all possible outcomes.
- **Event, E**. This is a subset of S. If an event E occurs, the outcome of the experiment is contained in E.

6 Probability Distributions

7 The Normal Distribution

8 Relational operators

- $>$ means 'is greater than
- $\geq$ means 'is greater than or equal to
- $<$ means 'is less than
- $\leq$ means 'is less than or equal to
- $\neq$ means 'is not equal to
- $\approx$ or $\simeq$ means 'is approximately equal to

9 A simple data set

Suppose that we have a data set with n observations. For each observation, a measure is recorded. Conventionally the measures are denoted x unless a more suitable notation is available. A subscript can be used to indicate which observation the measure is for. Hence we would write a data set as follows; $(x_1, x_2, x_3, x_1 \ldots x_n)$ (i.e. the first, second, third ... nth observation).

10 Summation

The summation sign $\sum$ is commonly used in most areas of statistics. Given $x_1 = 3, x_2 = 1, x_3 = 4, x_4 = 6, x_5 = 8$ find:

$$(i) \sum_{i=1}^{i=n} x_i \qquad\qquad (ii) \sum_{i=3}^{i=4} x_i^2$$

$$\begin{aligned} (i) \sum_{i=1}^{i=n} x_i &= x_1 + x_2 + x_3 + x_4 + x_5 \\ &= 3 + 1 + 4 + 6 + 8 \\ &= \mathbf{22} \end{aligned}$$

$$(ii) \sum_{i=1}^{i=n} x_i^2 = x_3^2 + x_4^2 = 9 + 16 = \mathbf{25}$$

When all elements of a data set are used, a simple version of the summation notation can be used. $\sum_{i=1}^{i=n} x_i$ can simply be written as $\sum x$

Example

Given that $p_1 = 1/4, p_2 = 1/8, p_3 = 1/8, p_4 = 1/3, p_5 = 1/6$ find:

- $\sum_{i=1}^{i=n} p_i \times x_i$
- $\sum_{i=1}^{i=n} p_1 \times x_i^2$

11 Arithmetic mean

One of the basic quantities is the arithmetic mean (it is sometimes called the 'average but there are in fact other measures of average apart from the mean). The arithmetic mean is calculated by adding the measures of the number of observations in which you are interested and dividing by the number of observations.

$$\bar{x} = \frac{\sum x}{n}$$

For our data set $\bar{x} = \frac{22}{5} = \mathbf{4.4}$.

12 Medians and modes

The median ($\tilde{x}$) is the value that separates a sample into two groups; 50% of observations are greater than the median and 50% are less than it.

The set of n numbers is arranged in ascending order, say as $x_(1), x_(2), x_(3), \ldots x_(n)$, where $x_(1)$ is the smallest of the observations and $x_(n)$ is the largest.

Computation of the median differs for samples that have an odd number size, and samples with an even number size. If sample size n is odd

$$\tilde{x} = x_{(\frac{n+1}{2})},$$

or if n is even

$$\tilde{x} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n+1}{2})}}{2}.$$

A visual inspection of the ordered data set will be useful for a quick determination of the median. For example, the median of the numbers 1, 3, 4, 5 and 7 is 4 (check this by rearranging the numbers in order!) and for 1, 3, 4 and 5, the median is (3+4)/2, that is 3.5.

The mode is the most frequently occurring value. There is not necessarily only one such value. For example, the figures 1, 2, 2, 3, 5, 9, 9, 11 have two modes: the numbers 2 and 9.

13 Observational studies and experiments

In industrial and agricultural applications of statistics it is possible to control the levels of the important factors that affect the results. Bias from the factors that cannot be controlled is dealt with by randomization. These investigations are designed experiments.
In medical, economic and other social science applications of statistics one usually just observes a sample of the population available, without control of any of the factors that may influence the measures observed. These studies are observational studies.

14 Measures of dispersion

It is unlikely we will only be interested in the average value of our data, we will want to know how large the spread or dispersion of values is about it.
The simplest measure of dispersion is the range. The range is simply the difference of the lowest and highest values. The range is another easy-to-understand measure, but it will clearly be very affected by a few extreme values.

Aside from the range, the most common measures of dispersion are:

- Variance
- Standard deviation
- Mean Absolute Deviation (MAD)
- Inter-quartile range.

The first three are related to the use of the arithmetic mean, and are computed using deviations of each observation from the mean.

The mean absolute deviation (MAD) uses the absolute values of the deviations from the mean and perhaps gives us a more intuitively understandable measure of deviation than variance and standard deviation.

15 Statistics for grouped data

grouped data refers to the arrangement of raw data with a wide range of values into groups. This process makes the data more manageable. Graphs and frequency diagrams can then be drawn showing the class intervals chosen instead of individual values.
An estimate, $\bar{x}$, of the mean of the population from which the data are drawn can be calculated from the grouped data as:

$$\bar{x} = \frac{\sum fx}{\sum f}$$

In this formula, x refers to the mid-point of the class intervals, and f is the class frequency. Note that the result of this will be different from the sample mean of the ungrouped data.

Class limits	Class midpoint	frequency
$240 - 259.99	$250	7
$260 - 279.99	$270	20
$280 - 299.99	$290	33
$300 - 319.99	$310	25
$320 - 339.99	$330	11
$340 - 359.99	$350	4
		Total = 100

16 Cumulative frequency

The graph of a cumulative frequency distribution is called an ogive (pronounced "o-jive"). For the less-than type of cumulative distribution, this graph indicates the cumulative frequency below each exact class limit of the frequency distribution. When such a line graph is smoothed, it is called an ogive curve.

17 Relative frequency

A relative frequency distribution is one in which the number of observations associated with each class has been converted into a relative frequency by dividing by the total number of observations in the entire distribution. Each relative frequency is thus a proportion, and can be converted into a percentage by multiplying by 100.

One of the advantages associated with preparing a relative frequency distribution is that the cumulative distribution and the ogive for such a distribution indicate the cumulative proportion (or percentage) of observations up to the various possible values of the variable. A percentile value is the cumulative percentage of observations up to a designated value of a variable.

18 Coefficient of variation

The coefficient of variation, CV, indicates the relative magnitude of the standard deviation as compared with the mean of the distribution of measure-

ments, as a percentage. Thus, the formulas are

$$\text{Population}: CV = \frac{\sigma}{\mu} \times 100$$
$$\text{Sample}: CV = \frac{s}{\bar{x}} \times 100$$

The coefficient of variation is useful when we wish to compare the variability of two data sets relative to the general level of values (and thus relative to the mean) in each set.

19 Basic definitions of probability

The symbol P is used to designate the probability of an event. Thus P(A) denotes the probability that event A will occur in a single observation or experiment.

The smallest value that a probability statement can have is 0 (indicating the event is impossible) and the largest value it can have is 1 (indicating the event is certain to occur). Thus, in general: $0 \leq P(A) \leq 1$

In a given observation or experiment, an event must either occur or not occur. Therefore, the sum of the probability of occurrence plus the probability of nonoccurrence always equals 1. Thus, where A' indicates the nonoccurrence of event A, we have $P(A) + P(A') = 1$

Mutually exclusive events

Two or more events are mutually exclusive, or disjoint, if they cannot occur together. That is, the occurrence of one event automatically precludes the occurrence of the other event (or events). For instance, suppose we consider the two possible events “ace” and “king” with respect to a card being drawn from a deck of playing cards. These two events are mutually exclusive, because any given card cannot be both an ace and a king. Two or more events are nonexclusive when it is possible for them to occur together.
Note that this definition does not indicate that such events must necessarily always occur jointly. For instance, suppose we consider the two possible events “ace” and “spade”. These events are not mutually exclusive, because

a given card can be both an ace and a spade; however, it does not follow that every ace is a spade or every spade is an ace.

General rule of addition

For events that are not mutually exclusive, the probability of the joint occurrence of the two events is subtracted from the sum of the simple probabilities of the two events. We can represent the probability of joint occurrence by P(A and B). In the language of set theory this is called the intersection of A and B and the probability is designated by P(A and B). Thus, the rule of addition for events that are not mutually exclusive is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Example

When drawing a card from a deck of playing cards, the events "ace" and "spade" are not mutually exclusive. The probability of drawing an ace (A) or spade (S) (or both) in a single draw is

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(S) - P(A \text{ and } B) \\ &= 4/52 + 13/52 - 1/52 \\ &= 16/52 \\ &= \mathbf{4/13} \end{aligned}$$

Independent events

Two events are independent when the occurrence or nonoccurrence of one event has no effect on the probability of occurrence of the other event. Two events are dependent when the occurrence or nonoccurrence of one event does affect the probability of occurrence of the other event.

Conceptual approaches

Historically, three different conceptual approaches have been developed for defining probability and for determining probability values: the classical,

relative frequency, and subjective approaches.
If N(A) possible elementary outcomes are favorable to event A, N(S) possible outcomes are included in the sample space, and all the elementary outcomes are equally likely and mutually exclusive, then the probability that event A will occur is

$$P(A) = \frac{N(A)}{N(S)}$$

Examples

When a fair dice is thrown, what are the possible outcomes? There are 6 possible outcomes. The dice can role any number between one and six. Each outcome is equally likely. The probability of each outcome is 1/6.

In a well-shuffled deck of cards which contains 4 aces and 48 other cards, the probability of an ace (A) being obtained on a single draw is;

$$P(A) = N(A)/N(S) = 4/52 = 1/13$$

20 Bayes theorem

In its simplest algebraic form, Bayes theorem is concerned with determining the conditional probability of event A given that event B has occurred. The general form of Bayes theorem is

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

21 Joint probability tables

A joint probability table is a table in which all possible events (or outcomes) for one variable are listed as row headings, all possible events for a second variable are listed as column headings, and the value entered in each cell of the table is the probability of each joint occurrence.

Often, the probabilities in such a table are based on observed frequencies of occurrence for the various joint events. The table of joint-occurrence frequencies which can serve as the basis for constructing a joint probability table is called a contingency table.

22 Permutations

The number of permutations of n objects is the number of ways in which the objects can be arranged in terms of order:
Permutations of n objects $= n! = (n) \times (n-1) \times (n-2) dots \times (2) \times (1)$

The symbol n! is read "n factorial". In permutations and combinations problems, n is always positive. Also, note that by definition $0! = 1$ in mathematics.

23 Combinations

In the case of permutations, the order in which the objects are arranged is important. In the case of combinations, we are concerned with the number of different groupings of objects that can occur without regard to their order. Therefore, an interest in combinations always concerns the number of different subgroups that can be taken from n objects. The number of combinations of n objects taken r at a time is

24 random variables

A random variable is defined as a numerical event whose value is determined by a chance process. When probability values are assigned to all possible numerical values of a random variable X, either by a listing or by a mathematical function, the result is a probability distribution.

The sum of the probabilities for all the possible numerical outcomes must equal 1.0. Individual probability values may be denoted by the symbol f (x), which indicates that a mathematical function is involved, by P(x = X), which recognizes that the random variable can have various specific values, or simply by P(X).

25 Discrete random variables

For a discrete random variable observed values can occur only at isolated points along a scale of values. For a six sided dice, the only possible observed

values are 1, 2, 3, 4, 5 and 6. It is not possible to observe values such as 5.732.

Therefore, it is possible that all numerical values for the variable can be listed in a table with accompanying probabilities. There are several standard probability distributions that can serve as models for a wide variety of discrete random variables involved in business applications. The standard models described in this course are the binomial, hypergeometric, and Poisson probability distributions.

For a continuous random variable all possible fractional values of the variable cannot be listed, and therefore the probabilities that are determined by a mathematical function are portrayed graphically by a probability density function, or probability curve.

26 solutions 1

1. Assume that the number of weekly study hours for students at a certain university is approximately normally distributed with a mean of 22 and a standard deviation of 6.

   (a) Find the probability that a randomly chosen student studies less than 12 hours.

   (b) Estimate the percentage of students that study more than 37 hours.

   $X \sim (22, 6^2)$
   $P(X \leq 12)$
   $P(X \geq 37)$
   $Z_1 = \frac{12-22}{6} = \frac{-10}{6} = -1.66$
   $Z_2 = \frac{37-22}{6} = \frac{15}{6} = 2.5$

27 Normal - example

In an examination the scores of students who attend schools of type A are normally distributed about a mean of 55 with a standard deviation of 6. The scores of students who attend type B schools are normally distributed about a mean of 60 with a standard deviation of 5.

Which type of school would have a higher proportion of students with marks above 70?

- $\mu_A = 55$
- $\sigma_A = 8$
- $\mu_B = 60$
- $\sigma_B = 5$

We have to fins $P(X_A \geq 70)$ and $P(X_B \geq 70)$.
using the standardisation formula $Z_A = \frac{70-55}{6} = \frac{15}{6} = 2.5$
$Z_B = \frac{70-60}{5} = \frac{10}{5} = 2$

Confidence intervals: example

A random sample of 25 female students is chosen from students at higher education establishments in a particular area of a country, and it is found that their mean height is 165 centimeters with sample variance of 81.

Assuming that the distribution of the heights of the students may be regarded as normally distributed, calculate a 95% confidence interval for the mean height of female students.

You are asked to obtain a 98% confidence interval for the mean height of width 3 centimeters. What sample size would be needed in order to achieve that degree of accuracy?

Solution 1

- Confidence interval formula: $\bar{X} \pm t_{(\alpha/2,df)} \frac{s}{\sqrt{n}}$.
- Small sample, therefore degrees of freedom = 24 (i.e. n-1).
- 95% confidence, therefore $\alpha = 5\%(i.e.0.05)$
- Correct t value from tables: $t_{(\alpha/2,n-1)} = t_{0.025,24}$
- Interval computed: $165 \pm 1.96 \frac{9}{\sqrt{25}} = (160:09; 169:91)$.

Solution 2

- Confidence interval width is 3, so half-width is 1.5
- Seek n such that $1.96 \times \frac{9}{\sqrt{n}} = 1.5$
- Divide both sides by 1.96×9

$$\frac{1}{\sqrt{n}} = \frac{1.5}{1.96 \times 9} =$$

- invert and square boths sides.