

22S:30/105
Statistical Methods and Computing

Probability Distributions

Lecture 10
 Feb. 23, 2009

Kate Cowles
 374 SH, 335-0727
 kcowles@stat.uiowa.edu

- **discrete random variable:** a random variable for which there exists a discrete set of possible values
 - Example: Let X be a random variable that represents the number of episodes of otitis media in the first 2 years of a child's life
 Then X is a discrete random variable which can take on values 0, 1, 2, ...

Random variables

- **variable:** a characteristic that can be measured or categorized
 - takes on different values for different members of a population or of a sample
- **random variable:** a numeric quantity that takes different values depending on chance
 - i.e., the value of a random variable is a numeric outcome of a random phenomenon
 - usually denoted by uppercase letters near end of alphabet

- **continuous random variable:** a random variable that is not discrete, i.e., whose values naturally from a continuum
 - Example: Let X be the random variable that measures cumulative lifetime exposure of shipyard workers to radiation. In one study, values of this random variable varied from 0.000 rem to 91.414 rem, which may be regarded as taken on an essentially infinite number of values.

Probability distributions

- A **probability distribution** describes the behavior of a random variable.
- The probability distribution of a discrete random variable specifies all the possible values of the variable, along with the probability of each.

- Example: number of middle ear infections experienced by children between the ages of birth and 2 years

| x | P(X = x) |
|----|----------|
| 0 | 0.129 |
| 1 | 0.264 |
| 2 | 0.271 |
| 3 | 0.185 |
| 4 | 0.095 |
| 5 | 0.039 |
| 6+ | 0.017 |

Note that the probabilities for all the possible values must sum to 1.

Assigning probabilities to intervals of outcomes

For continuous random variable, possible values are on a continuum.

- We can not assign an individual probability to each possible value because there are infinitely many possible values.
- solution: use a density curve to assign probabilities to *intervals* of values
- Example: Suppose we draw a birth record at random from a nationwide medical database. What is the probability that the birth-weight of the infant was between 80 and 96 ounces (or between 5 and 6 pounds)?

Using density curves to describe the distribution of values of a quantitative variable

- density curve – a curve that describes the overall pattern of a distribution
- total area under a probability density curve is 1.0
- the curve never drops below the horizontal axis
- normal curve is only one example

Measures of center and spread can be used to describe density curves.

– To distinguish between these measures in the idealized curve vs. in actual sample data, we use different symbols:

- * μ for the *mean* of a density curve
- * σ for the *standard deviation* of a density curve

Example

For women in the US between 18 and 74 years of age, diastolic blood pressure follows a normal distribution with mean is $\mu = 77$ mm Hg and standard deviation $\sigma = 11.6$ mm Hg.

We want to know the proportion of US women in this age group who have dbp between 60 and 100.

1. Call the variable representing a woman's dbp X , and call the specific value for an individual woman x . X has a normal distribution with $\mu = 77$ and $\sigma = 11.6$. We want to compute to compute the proportion of women such that

$$60 \leq X \leq 100$$

2. Standardize x to produce z , a draw from a standard normal distribution.

$$60 \leq X \leq 100$$

$$\frac{60 - 77}{11.6} \leq \frac{X - 77}{11.6} \leq \frac{100 - 77}{11.6}$$

$$-1.47 \leq Z \leq 1.98$$

3. Use Table A to find
 - the proportion of Z values ≤ -1.47 , which = .0708
 - and the proportion of Z values ≤ 1.98 , which = .9761.
4. So the percent of women with diastolic blood pressure between 60 and 100 is about $97.61\% - 7.08\% = 90.5\%$.

Normal calculations going the other direction

What is the value of dbp such that 10% of women have values greater than or equal to it?

1. Use Table A to find the z-score such that 10% of a standard normal population would have values greater than or equal to it. This is the same value such that 90% of values are less than or equal to it, namely 1.28.
2. Convert $z = 1.28$ into x .

$$\frac{x - \mu}{\sigma} = z$$

$$\frac{x - 77}{11.6} = 1.28$$

$$x = 77 + (11.6)(1.28)$$

$$x = 91.85$$

we have been using normal distributions to describe populations

- we can use z-scores to compare values from 2 different populations that follow normal distributions
- example:
 - * Former NBA superstar Michael Jordan is 78 in. tall.
 - * WNBA basketball player Rebecca Lobo is 76 in. tall.
 - * Data from the National Health Survey indicate that men's heights are roughly normally distributed with a mean of 69.0 in and standard deviation of 2.8 in, while women's heights are approximately normally distributed with mean 63.6 in and standard deviation 2.5 in.

General formula for *un*standardizing a z-score:

$$x = \mu + z\sigma$$

- * Is Michael Jordan's height among men more or less extreme than Rebecca Lobo's height among women? Show any calculations that justify your answer.

Another density curve for continuous outcomes

The *uniform* density curve distributes probability evenly over the interval from 0 to 1.

- Where do probability distributions come from?
 - Informally, we may think of a probability distribution as a model based on an infinitely large sample.
 - In some cases there is previous data on the same type of random variable in a large enough number of observations that we can use it to compute a probability distribution
 - In other cases, we try to use a well-known theoretical probability distribution and see how well it fits with some sample data.

Review of terminology

A **population** is the entire set of items that we would like to investigate or draw conclusions about.

A **population parameter** is a numerical quantity that describes a characteristic of a population.

- The exact value of a parameter can be obtained only if the values of a variable are known for every single item in a population.
- Population parameters are usually designated by Greek letters.

The location and spread of a random variable

- **population mean** or **expected value** of a random variable: the average value assumed by a random variable
 - analogous to the arithmetic mean \bar{x} in a sample
- **population variance** and **population standard deviation** are measures of the dispersion of values of the random variable around the population mean

A **sample** is a subset of items that is selected from a population.

- The sample is of a manageable size, so we can actually measure the values of the variable of interest for all members of the sample.

A **simple random sample** is a sample drawn in such a way that every item in the population has an equal chance of being selected.

Statistical inference is the process of drawing conclusions and making decisions about a population based on information contained in a sample drawn from that population.

The methods of statistical inference that we will study in this class assume that the sample being used is a simple random sample.

A **sample statistic** is a number that can be computed from sample data without our having to know any unknown parameters. We often use a statistic to *estimate* the value of an unknown parameter.

Example 1:

The Current Population Survey reported the mean income of the sample of households they interviewed to be $\bar{x} = \$49,692$.

- The number \$49,692 is a *statistic* because it describes the particular sample of households included in the CPS.
- The *population* that the poll wants to draw conclusions about is all 103 million U.S. households.
- The *parameter* of interest is the mean income of all these 103 million households. We do not know the value of this parameter.

Example 2:

We wish to study body fat levels in Chinese adult males. The particular variable that we are interested in is the upper arm skinfold thickness in mm.

- The population of interest is all Chinese males aged 18-24. There are approximately 300,000,000 of them.
- The parameter of interest is μ , the population mean of upper arm skinfold thickness. We will never know the exact value of this parameter because we cannot measure all members of the population.
- We will take a random sample of Chinese males and determine the sample mean \bar{x} of upper arm skinfold thickness.
- We will use \bar{x} to *estimate* the unknown μ .