

INTRODUCTION TO BASIC STATISTICS AND PROBABILITY

Justin Shows Betsy Enstrom

North Carolina State University
Duke University

May 22nd, 2008

Outline

- 1 Definitions
- 2 Probability Laws
- 3 Descriptive Statistics
- 4 Distribution
- 5 Sampling Distribution of Sample Mean

Preliminary Definitions

- Statistics: Science of collecting, displaying, and analyzing data
- Population: Complete set of objects or people of interest
Examples: All people in NC, cars in the US
- Sample: A subset of the population
- Data: Information

Preliminary Definitions (cont.)

- **Parameter:** Value that describes the population (usually unknown)
Examples: Proportion of Democrats in Wake County, mean height of women at NCSU
- **Statistic:** Value calculated from a sample often used to estimate a parameter
- **Inference:** Drawing conclusions based on data
- **Random Variable:** A variable whose value is a numerical outcome of a random phenomenon
- **Bias:** Difference between an estimator's expectation and the true value of the parameter being estimated

Probability Definitions

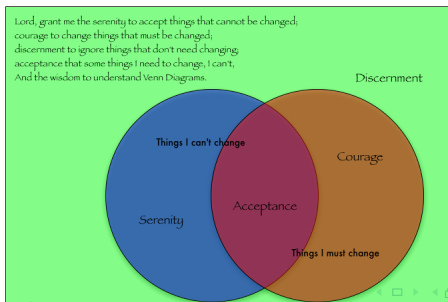
- Sample Space: The set of all possible outcomes of a random variable
Example: Roll a die. Sample space = $\{1,2,3,4,5,6\}$
- Event: Any set of outcome of interest
- Probability of an event: The proportion of times an event occurs over an infinite number of trials

Probability Definitions

- Independent events: Knowledge that one event has occurred has no effect on the probability of the other
- Mutually exclusive events: Events cannot simultaneously occur
Example: Roll a 6 on a die, and roll an odd number

Probability Notation

- $Pr(A)$ = Probability that Event A occurs
- $Pr(A^c)$ = $Pr(A \text{ does not occur}) = 1 - Pr(A)$
- $Pr(A \cup B)$ = $Pr(\text{Either } A, B, \text{ or both occur})$
- $Pr(A \cap B)$ = $Pr(\text{Both } A \text{ and } B \text{ occur})$
- $Pr(A|B)$ = $Pr(A \text{ occurs given } B \text{ occurred})$



Probability Laws

- $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$
- $Pr(A|B) = \frac{Pr(A \text{ and } B)}{Pr(B)}$
- If A and B are independent: $Pr(A \cap B) = Pr(A) \cdot Pr(B)$ and $Pr(A|B) = Pr(A)$
- If A and B are mutually exclusive: $Pr(A \cap B) = 0$ and $Pr(A|B) = 0$

Descriptive Statistics

- Used to describe data
- Can be numerical or graphical
- Helpful in understanding the nature of the data
- Inference often relies heavily on descriptive statistics

Sample Mean and Median

- Both measure the "center" of the data points x_1, x_2, \dots, x_n
- The mean, usually denoted by \bar{x} , is the arithmetic average:
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$
- The median is the middle value of the ordered data

Example: $\mathbf{X} = \{48, 49, 50, 51, 52\}$, so $\bar{\mathbf{X}} = 50$, and median is also 50.

Sample Variance

- Measures how spread out the data points are from the mean
- Usually denoted by s^2
- The formula is: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
- The sample standard deviation is $s = \sqrt{s^2}$

Example: $\mathbf{X}_1 = \{48, 49, 50, 51, 52\}$ and $\mathbf{X}_2 = \{1, 2, 50, 98, 100\}$
 , so $\bar{\mathbf{X}}_1 = \bar{\mathbf{X}}_2 = 50$, and medians are also 50, but $s_{X_1}^2 = 2.5$ and
 $s_{X_2}^2 = 2352.5$

Random Variables

- A random variable is usually denoted by X , Y , or Z
- Data points are realizations (observations) of a random variable
- A discrete random variable has a countable number of possible values. Example: the number of days that it rains yearly
- A continuous random variable has intervals for its set of possible values. Example: amount of prep time for the SAT

Distribution

- The probability distribution for a random variable X gives the possible values (sample space) for X , and the probability for each
- The methods used to specify discrete probability distributions are similar to (but slightly different from) those used to describe continuous probability distributions
- The distribution for a discrete random variable is given by a probability mass function (pmf)
- The distribution for a continuous random variable is given by a probability density function (pdf)

Probability Mass Function

- $f(x)$ is the pmf for a discrete random variable X having possible values x_1, x_2, \dots
- $f(x_i) = Pr(X = x_i)$ is the probability that X takes the value x_i
- $0 \leq f(x_i) \leq 1$ and $\sum_i f(x_i) = 1$
- $f(x)$ can be expressed as a table or a mathematical function

PMF Example

- X is the number of rooms in a randomly chosen house in Anaheim, CA (Moore p. 244)
- The distribution of X is:

x_j	1	2	3	4	5	6	7
$f(x_j)$.083	.071	.076	.139	.210	.224	.197

Expected Value

- Expected value of X or (population) mean:

$$\mu = E(X) = \sum_{i=1}^R x_i f(x_i)$$

where the sum is over R possible values, and R could be finite or infinite

- Analogous to the sample mean \bar{X}
- Represents the "average" value of X
- Parameter

Variance

- Population variance (parameter)

$$\begin{aligned}\sigma^2 &= \text{Var}(X) \\ &= \sum_{i=1}^R (x_i - \mu)^2 f(x_i) \\ &= \sum_{i=1}^R x_i^2 f(x_i) - \mu^2\end{aligned}$$

- Represents the spread, relative to the expected value, of all values with positive probability
- The standard deviation of X is $\sigma = \sqrt{\sigma^2}$

Covariance

The covariance between two random variables X and Y , with expected values $E(X) = \mu$ and $E(Y) = \nu$ is defined as

$$\text{Cov}(X, Y) = E[(X - \mu)(Y - \nu)]$$

- $\text{Cov}(X, Y) = E(XY) - \mu\nu$
- This is a measure of how much the random variable change *together*
- If X and Y are independent, $\text{Cov}(X, Y) = 0$

Room Example

For the room example, find the following:

- $E(X)$
- $Var(X)$
- $\Pr(\text{a house has at least 5 rooms})$

Binomial Distribution

Structure

- Two possible outcomes: Success (S) and Failure (F)
- Repeat the situation (trial) n times
- $Pr(S) = p$ is constant on each trial
- The trials are independent

Binomial Distribution (cont.)

- Let X be the number of S in n independent trials (X can take the values $x = 0, 1, 2, \dots, n$)
- Then X has the binomial distribution with parameter n and p
- The pmf of X is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \dots, n$$

- $\mu = np$
- $\sigma^2 = np(1-p)$

Example

- Each child born to a particular set of parents has a probability of 0.25 of having Type O blood. If these parents have 5 children, what is the probability that exactly 2 of them have Type O blood? (Moore p. 306)
- Let X be the number of children with Type O blood

$$Pr(X = 2) = f(2) = \binom{5}{2} (.25)^2 (.75)^3 = .2637$$

- What are $E(X)$ and $Var(X)$?
- What is the probability of at least 2 children with Type O blood?

$$\begin{aligned}Pr(X \geq 2) &= \sum_{k=2}^5 \binom{5}{2} (.25)^k (.75)^{5-k} \\ &= 1 - \sum_{k=0}^1 \binom{5}{2} (.25)^k (.75)^{5-k} \\ &= .3671875\end{aligned}$$

Continuous Random Variable

- $f(x)$ is the pdf for a continuous random variable X
- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $Pr(a \leq X \leq b) = \int_a^b f(x)dx$
- $Pr(a < X < b) = Pr(a \leq X \leq b)$
- $Pr(X = a) = 0$

Mean and Variance

- Mean (expected value)

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

- Variance

$$\begin{aligned}\sigma^2 &= \text{Var}(X) \\ &= \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx \\ &= \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2\end{aligned}$$

Example

- Let X be the fraction of the population in a city who obtain the flu vaccine.

$$f(x) = \begin{cases} 2x & \text{when } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$



$$\begin{aligned} Pr(1/4 \leq X \leq 1/2) &= \int_{1/4}^{1/2} f(x) dx \\ &= \int_{1/4}^{1/2} 2x dx \\ &= 3/16 \end{aligned}$$

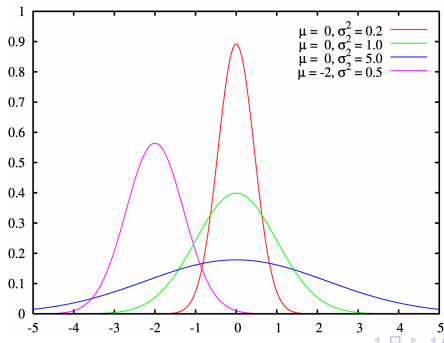
Example (cont.)

- Find $Pr(X \geq 1/2)$
- Find $E(X)$
- Find $Var(X)$

Normal Distribution

Most widely used continuous distribution (also known as Gaussian distribution)

- The pdf is symmetric around μ
- The pdf is bell-shaped



Normal Distribution (cont.)

- The pdf is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right]$$

- $E(X) = \mu$
- $Var(X) = \sigma^2$
- $X \sim N(\mu, \sigma^2)$ means that X is Normally distributed with mean μ and variance σ^2

Standard Normal Distribution

- A Normal distribution with mean 0 and variance 1 is called a standard Normal distribution.
- Cumulative density function: Let $Z \sim N(0, 1)$

$$\begin{aligned}Pr(Z \leq z) &= \int_{-\infty}^z f(z)dz \\ &= \Phi(z)\end{aligned}$$

These values can be found in tables or computers.

- Symmetry property:

$$\Phi(-z) = 1 - \Phi(z)$$

Standardization

- Suppose $X \sim N(\mu, \sigma^2)$ and $Z = \frac{X-\mu}{\sigma}$. Then $Z \sim N(0, 1)$.
- If $X \sim N(\mu, \sigma^2)$, what is $Pr(a < X < b)$?

$$\begin{aligned} Pr(a < X < b) &= Pr\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$

Example

- Suppose that the heights of young women are Normally distributed with $\mu = 64$ and $\sigma^2 = 2.7^2$. What is the probability that a randomly selected young women will be between 60 and 70 inches tall? (Moore, pp. 65-67)

$$\begin{aligned} Pr(60 < X < 70) &= Pr\left(\frac{60-64}{2.7} < Z < \frac{70-64}{2.7}\right) \\ &= Pr(-1.48 < Z < 2.22) \\ &= \Phi(2.22) - \Phi(-1.48) \\ &= .9868 - .0694 \\ &= .9174 \end{aligned}$$

Sampling Distribution of \bar{X}

- A natural estimator for the population mean μ is the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- Consider \bar{x} to be a single realization of a random variable \bar{X} over all possible samples of size n .
- The sampling distribution of \bar{X} is the distribution of the values of \bar{x} over all possible samples of size n that could be selected from the population.

Expected Value of \bar{X}

- The average of sample means when taken over a large number of random samples of size n will approximate μ .
- Let X_1, X_2, \dots, X_n be a random sample from some population with mean μ . Then for the sample mean \bar{X} , $E(\bar{X}) = \mu$.
- \bar{X} is an unbiased estimator of μ .

Standard Deviation of \bar{X}

- Let X_1, X_2, \dots, X_n be a random sample from some population with mean μ and variance σ^2 .
- The variance of the sample mean \bar{X} is given by

$$\text{Var}(\bar{X}) = \sigma^2/n$$

- The standard deviation of the sample mean is given by σ/\sqrt{n} .

Standard Error of \bar{X}

- The standard deviation σ/\sqrt{n} is estimated by the standard error of \bar{X} , which is s/\sqrt{n} .
- The standard error measures the variability of sample means from repeated samples of size n drawn from the same population.
- A larger sample provides a more precise estimate \bar{X} of μ .

Sampling Distribution of \bar{X}

- Let X_1, X_2, \dots, X_n be a random sample from a population that is Normally distributed with mean μ and variance σ^2 .
- Then the sample mean \bar{X} is Normally distributed with mean μ and variance σ^2/n .
- That is,

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

Central Limit Theorem

- Let X_1, X_2, \dots, X_n be a random sample from any population with mean μ and variance σ^2 .
- Then the sample mean \bar{X} is approximately Normally distributed with mean μ and variance σ^2/n when n is large.

References

- Heyward, Shenek. SAMSI UGW Presentation: 2006.
- Moore, David S. "The Basic Practice of Statistics." Third Edition. W.H. Freeman and Company, New York: 2003.
- Weems, Kimberly. SIBS Presentation: 2005.