An Introduction to Basic Statistics and Probability

Shenek Heyward

NCSU

Outline

- Basic probability concepts
- Conditional probability
- Discrete Random Variables and Probability Distributions
- Continuous Random Variables and Probability Distributions
- Sampling Distribution of the Sample Mean
- Central Limit Theorem

Idea of Probability

- Chance behavior is unpredictable in the short run, but has a regular and predictable pattern in the long run.
- The probability of any outcome of a random phenomenom is the proportion of times the outcome would occur in a very long series of repetitions.

Terminology

- Sample Space - the set of all possible outcomes of a random phenomenon
- Event - any set of outcomes of interest
- Probability of an event - the relative frequency of this set of outcomes over an infinite number of trials
- Pr(A) is the probability of event A

Example

- Suppose we roll two die and take their sum
- $S = \{2, 3, 4, 5, .., 11, 12\}$
- $\Pr(sum = 5) = \frac{4}{36}$
- Because we get the sum of two die to be 5 if we roll a (1,4),(2,3),(3,2) or (4,1).

Notation

- Let A and B denote two events.
  - $A \cup B$ is the event that either A or B or both occur.
  - $A \cap B$ is the event that both A and B occur simultaneously.
  - The complement of A is denoted by $\overline{A}$.
    - $\overline{A}$ is the event that A does not occur.
    - Note that $\Pr(\overline{A}) = 1 - \Pr(A)$.

Definitions

- A and B are mutually exclusive if both cannot occur at the same time.
- A and B are independent events if and only if

$$\Pr(A \cap B) = \Pr(A)\Pr(B).$$

Laws of Probability

- Multiplication Law: If $A_1, \cdots, A_k$ are independent events, then

$$\Pr(A_1 \cap A_2 \cap \cdots \cap A_k) = \Pr(A_1)\Pr(A_2)\cdots\Pr(A_k).$$

- Addition Law: If A and B are any events, then

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

Note: This law can be extended to more than 2 events.

Conditional Probability

- The conditional probability of B given A

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

- A and B are independent events if and only if

$$\Pr(B|A) = \Pr(B) = \Pr(B|\overline{A})$$

Random Variable

- A random variable is a variable whose value is a numerical outcome of a random phenomenon
- Usually denoted by X, Y or Z.
- Can be
  - Discrete - a random variable that has finite or countable infinite possible values
    - Example: the number of days that it rains yearly
  - Continuous - a random variable that has an (continuous) interval for its set of possible values
    - Example: amount of preparation time for the SAT

Probability Distributions

- The probability distribution for a random variable X gives
  - the possible values for X, and
  - the probabilities associated with each possible value (i.e., the likelihood that the values will occur)
- The methods used to specify discrete prob. distributions are similar to (but slightly different from) those used to specify continuous prob. distributions.

Probability Mass Function

- $f(x)$ - Probability mass function for a discrete random variable X having possible values $x_1, x_2, \cdots$
- $f(x_i) = \Pr(X = x_i)$ is the probability that X has the value x_i
- Properties
  - $0 \leq f(x_i) \leq 1$
  - $\sum_i f(x_i) = f(x_1) + f(x_2) + \cdots = 1$
- $f(x_i)$ can be displayed as a table or as a mathematical function

Probability Mass Function

- Example: (Moore p. 244)Suppose the random variable X is the number of rooms in a randomly chosen owner-occupied housing unit in Anaheim, California.
- The distribution of X is:

Rooms X	1	2	3	4	5	6	7
Probability	.083	.071	.076	.139	.210	.224	.197

Parameters vs. Statistics

- A parameter is a number that describes the population. Usually its value is unknown.
- A statistic is a number that can be computed from the sample data without making use of any unknown parameters.
- In practice, we often use a statistic to estimate an unknown parameter.

Parameter vs. Statistic Example

- For example, we denote the population mean by μ, and we can use the sample mean $\bar{x}$ to estimate μ.
- Suppose we wanted to know the average income of households in NC.
- To estimate this population mean income μ, we may randomly take a sample of 1000 households and compute their average income $\bar{x}$ and use this as an estimate for μ.

Expected Value

- Expected Value of X or (population) mean

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

  where the sum is over R possible values. R may be finite or infinite.

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

Variance

- (Population) variance

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

- Represents the spread, relative to the expected value, of all values with positive probability
- The standard deviation of X, denoted by σ, is the square root of its variance.

Room Example

For the Room example, find the following

- $E(X)$
- $Var(X)$
- $\Pr$ [a unit has at least 5 rooms]

Binomial Distribution

- Structure
  - Two possible outcomes: Success (S) and Failure (F).
  - Repeat the situation n times (i.e., there are n trials).
  - The "probability of success," p, is constant on each trial.
  - The trials are independent.

Binomial Distribution

- Let X = the number of S's in n independent trials. (X has values $x = 0, 1, 2, \cdots, n$)
- Then X has a binomial distribution with parameters n and p.
- The binomial probability mass function is

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \cdots, n$$

- Expected Value: $\mu = E(X) = np$
- Variance: $\sigma^2 = Var(X) = np(1-p)$

Example

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

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

Example

- What is the expected number of children with type O blood?
  $\mu = 5(.25) = 1.25$
- What is the probability of at least 2 children with type O blood?

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

Continuous Random Variable

- $f(x)$ - Probability density function for a continuous random variable X
- Properties
  - $f(x) \geq 0$
  - $\int_{-\infty}^{\infty} f(x)dx = 1$
  - $P[a \leq X \leq b] = \int_a^b f(x)dx$
- Important Notes
  - $P[a \leq X \leq a] = \int_a^a f(x)dx = 0$
    This implies that $P[X = a] = 0$
  - $P[a \leq X \leq b] = P[a < X < b]$

Summarizations

Summarizations for continuous prob. distributions

- Mean or Expected Value of X

$$\mu = EX = \int_{-\infty}^{\infty} x f(x) dx$$

- Variance

$$\begin{aligned}\sigma^2 &= VarX \\ &= \int_{-\infty}^{\infty} (x - EX)^2 f(x) dx \\ &= \int_{-\infty}^{\infty} x^2 f(x) dx - (EX)^2\end{aligned}$$

Example

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

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

- Find P$(1/4 \leq X \leq 1/2)$
-

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

Example

- $$f(x) = \begin{cases} 2x & \text{when } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$
- Find P($X \geq 1/2$)
- Find EX
- Find $VarX$

Normal Distribution

- Most widely used continuous distribution
- Also known as the Gaussian distribution
- Symmetric

Normal Distribution

- Probability density function

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

- $EX = \mu$
- $VarX = \sigma^2$
- Notation: $X \sim N(\mu, \sigma^2)$
  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.
- Standard normal probability density function

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

- Standard normal cumulative probability function
  Let $Z \sim N(0,1)$

$$\Phi(z) = P(Z \leq z)$$

- Symmetry property

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

Standardization

Standardization of a Normal Random Variable

- Suppose $X \sim N(\mu, \sigma^2)$ and let $Z = \frac{X-\mu}{\sigma}$. Then $Z \sim N(0, 1)$.
- If $X \sim N(\mu, \sigma^2)$, what is $P(a < X < b)$?
  - Form equivalent probability in terms of Z :

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$

  - Use standard normal tables to compute latter probability.

Standardization

- Example. (Moore pp.65-67) Heights of Women
- Suppose the distribution of 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 woman will have a height between 60 and 70 inches?

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

Sampling Distribution of $\overline{X}$

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

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

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

Expected Value of $\overline{X}$

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

Standard Error of $\overline{X}$

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

$$Var(\overline{X}) = \sigma^2/n.$$

- The standard deviation of the sample mean is given by $\sigma/\sqrt{n}$. This quantity is called the standard error (of the mean).

Standard Error of $\overline{X}$

- The standard error $\sigma/\sqrt{n}$ is estimated by $s/\sqrt{n}$.
- 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 $\overline{X}$ of μ

Sampling Distribution of $\overline{X}$

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

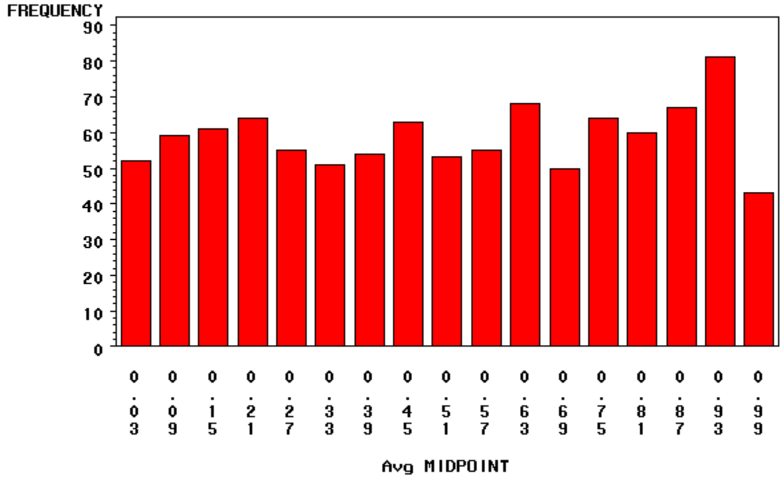
$$\overline{X} \sim N(\mu, \sigma^2/n).$$

Central Limit Theorem

- Let $X_1, \cdots, X_n$ be a random sample from any population with mean μ and variance σ^2.
- Then the sample mean $\overline{X}$ is approximately normally distributed with mean μ and variance σ^2/n.

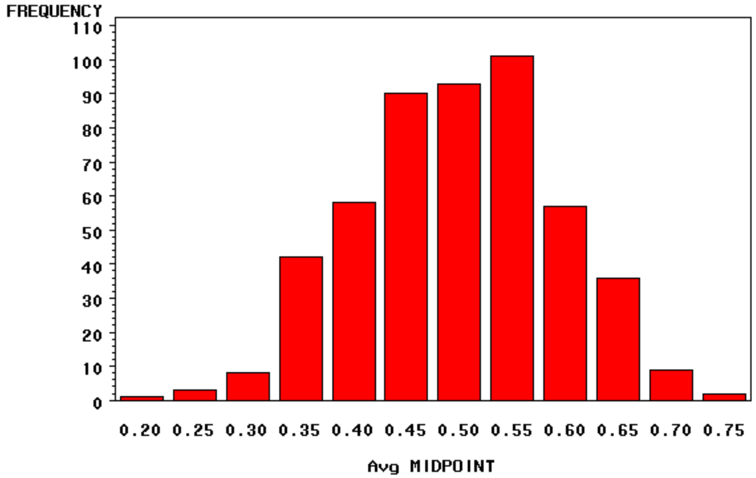
Data Sampled from Uniform Distribution

The following is a distribution of $\overline{X}$ when we take samples of size 1.



Example

The following is a distribution of $\overline{X}$ when we take samples of size 10.



References

- Moore, David S., "The Basic Practice of Statistics." Third edition. W.H. Freeman and Company. New York. 2003
- Weems, Kimberly. SIBS Presentation, 2005.