Random Variables and Probability Distributions

STAT 250

Lecture 4

Outline

- Discrete and Continous Random Variables
  - Probability density functions, density curves
  - Mean and Standard Deviation for Random Variables
- Normal Random Variables
  - Empirical Rule, Z-scores
  - Areas under the curve
  - Percentiles
- Binomial Random Variables
  - Computing binomial probabilities
  - Normal approximation to binomial

Random Variables

A **random variable** assigns a number to each outcome of a random circumstance, or, equivalently, to each unit in a population.

- Use letters near the end of the alphabet, such as x, to symbolize variables.
- Use a capital letter, such as X, to refer to the random variable itself.
- Use a small letter, such as x, to refer to a particular value of the variable.

Types of Random Variables

- A **discrete random variable** has possible values that are isolated points on the number line. It can take one of a countable list of distinct values. For discrete random variables, we can find probabilities for exact outcomes.
- A random variable is **continuous** if its possible values are all points in some interval. For continuous random variables we cannot find probabilities for exact outcomes. We are limited to finding probabilities for __________________ of values.

Which Type?

Determine whether each of the following random variables is discrete or continuous

- Number of students who earn 100% on next homework assignment.
- Time to commute home
- Yearly rain at airport
- Pairs of shoes in your closet
- Number of survey participants who answer yes
- Units of classes enrolled for the semester
- Attempts at legal bar exam before passing

Discrete Random Variables

X = the random variable.

k = a number the discrete random variable could assume.

P(*X = k*) is the probability that *X* equals *k.*

- The **probability distribution function (pdf)** *X* is a table or rule (formula) that assigns probabilities to possible values of *X.* The pdf must meet the following two conditions:
  - The sum of the probabilities over all possible values of a discrete random variable must equal 1. Stated mathematically,
  - The probability of any specific outcome for a discrete random variable must be between 0 and 1. Stated mathematically,

Discrete PDF example

Here is the pdf for the random variable number of people dated in the last month in a population of students

k	P(X=k)
0	.475
1	.45
2	.025
3	.025
4	

- What probability should be assigned to P(X=4)?

Using Sample Space to Compute PDF

Sometimes the sample space can be used to find probabilities for discrete random variables. This is typically the case when the random variable is a count.

- **Step 1:** List all simple events in sample space.
- **Step 2:** Identify the value of the random variable X for each simple event.
- **Step 3:** Find the probability for each simple event (often equally likely).
- **Step 4:** Find $P(X = k)$ as the sum of the probabilities for all simple events where $X = k$.

- Example: A student guesses on 3 True False questions on a quiz. Let X be the number of correct answers.
  - Find the pdf of X (create a table of possible values and probabilities) using the steps above. Also find P(X>0).

CDF

- A cumulative distribution function is a table that gives _________________________ for any real number k. Give the cumulative distribution function for X, the number of correct True/False responses.

Continuous Random Variables

Recall that for a continuous random variable the outcome can be any value in an interval or collection of intervals.

- Assigning probabilities when your sample space is not a finite set of numbers, but rather a continuous interval is a bit different.
- We cannot assign individual probabilities to all possible outcomes; so instead, we use the area under a **density curve** to represent the probability of the random variable taking values in that interval.

Density Curves

A **probability distribution for a continuous random variable x** is specified by a curve called a **density curve.** The function that defines this curve is denoted by f(x) and it is called the density function

- $f(x) \geq$ ____ (so that the curve cannot dip below the horizontal axis)
- The total area under the density curve is equal to _____.
- The probability that x falls in any particular interval is the area under the density curve in that interval.

Density Curve Illustration

$P(x<a)$

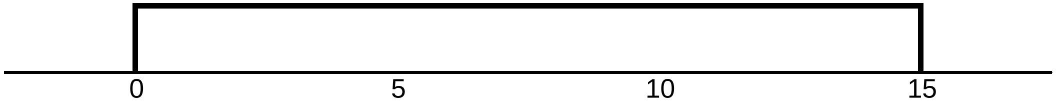
$P(a<x<b)$

$P(x>b)$

$P(x=b)$

Trolley Example

Time spent waiting for a trolley has a probability distribution given by the density curve below (we call this a **uniform** distribution).



- What is the height of the curve?

- What is the probability you will wait more than 10 minutes?

- What is the probability you will wait exactly 10 minutes?

Mean and Standard Deviation for Random Variables

- The mean value of a random variable x, denoted by _______ , describes where the probability distribution of x is centered. This is also called the ____ of the random variable.

- The standard deviation of a random variable x, denoted by _______, describes the variability in the probability distribution.

Expected Value - Discrete

- For a discrete random variable, we calculate expected value as: Expected value = Sum of "value × probability"
- If X is a random variable with possible values $x_1, x_2, x_3, \ldots ,$ occurring with probabilities $p_1, p_2, p_3, \ldots ,$ then the **expected value** of X is calculated as

Expected Value Example

A fundraiser is planning to rent game equipment for an upcoming fair. The non-refundable cost of the rental is \$1000. If the event is rained out (25% chance), her organization will be out \$1000, but if the weather is good, revenue is expected to be \$2500. Let X be the profit(loss) for the rental. Give the pdf of X and compute the expected value.

Standard Deviation

- The **standard deviation** of a random variable is roughly the average distance the random variable falls from its mean, or expected value, over the long run.
- If *X* is a discrete random variable with possible values $x_1, x_2, x_3, \ldots,$ occurring with probabilities $p_1, p_2, p_3, \ldots,$ and expected value $E(X) = \mu$, then
  - $V(X) = \sigma^2$
  - $StDev(X) = \sigma =$
- Note that if we have a population of measurements, the standard deviation can be computed using either the formula above or the formula for σ given in earlier (where we divided by n for the population – we divided by n-1 for a sample). Both will result in the same standard deviation.

Standard Deviation Example

- The fundraiser has the opportunity to rent a different set of equipment for a lower price, but it will not generate as much revenue. This equipment costs $250 to rent, but can only expect to bring in $1500 in revenue. Find the mean and standard deviation of profit with this equipment if the chance of a rain out is still 25%. Compare to the mean and standard deviation of the $1000 rental.

Expected Value and S.D. Continuous

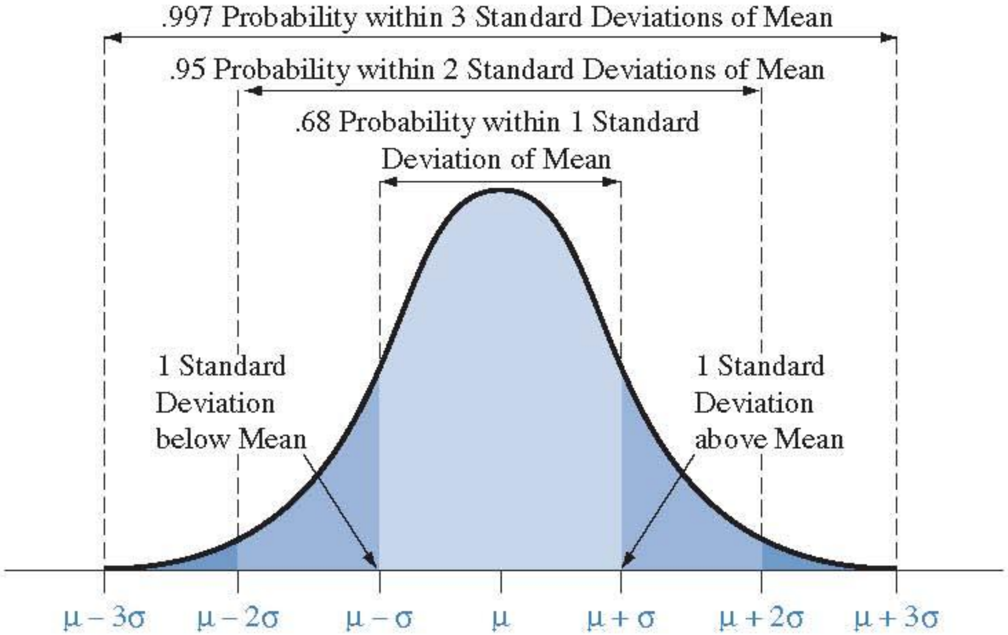
Mean and standard deviation of continuous random variables are found using calculus. We won’t cover them in this course, but their interpretation is the same (measure of center/balance point) and measure of spread (“average” distance from a value in the distribution to the mean)

Normal Random Variables

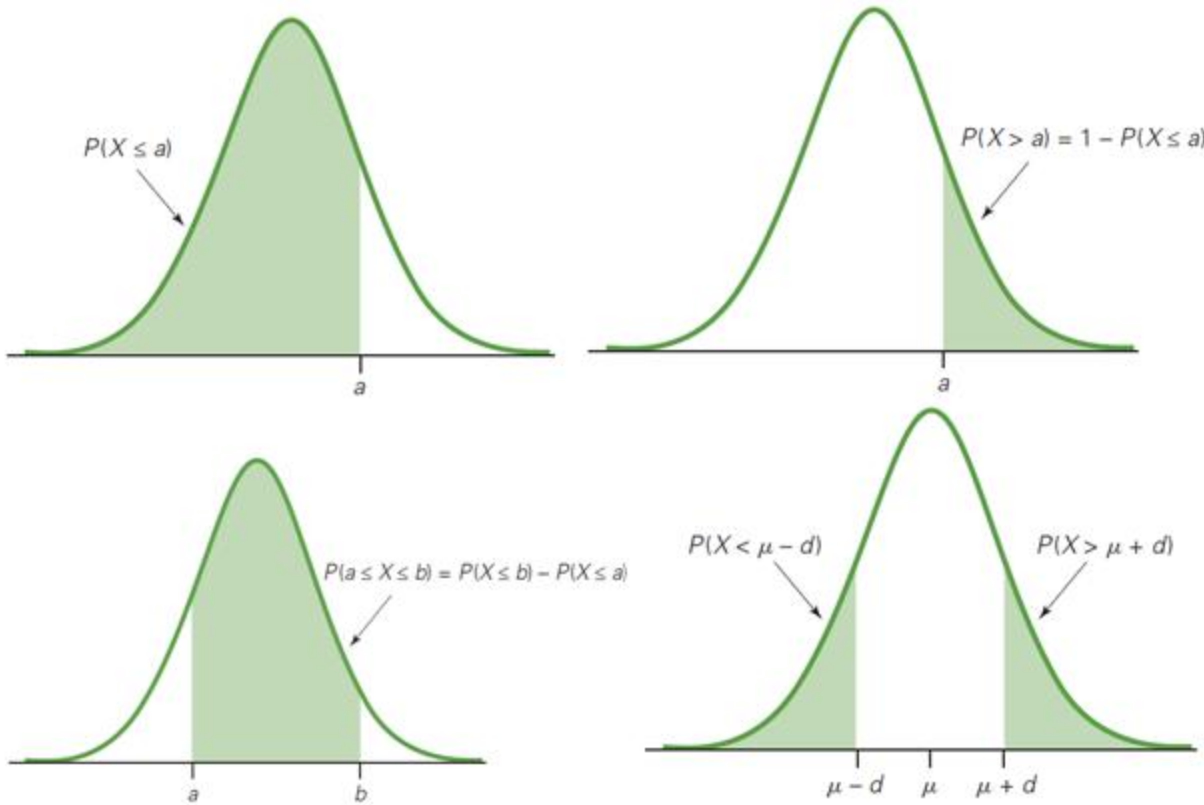
- Most commonly encountered type of continuous random variable is the **normal random variable,** which has a specific form of a bell-shaped probability density curve called a **normal curve.** A **normal random variable** is also said to have a **normal distribution**
- The normal distribution is __________________________, ___________________ and characterized by its ________________________________ and _____________________________.
- Because the curve is symmetric, $P(X \leq \mu) = P(X \geq \mu) =$ _______.

Empirical Rule

The probability of falling within any particular number of standard deviations of μ is the same for all normal distributions.



Useful Probability Relationships



Useful Probability Relationships

- We will use pictures as a tool to help us solve normal distribution problems.
- The rules represented above are
  - Rule 1: $P(X>a)=$
  - Rule 2: $P(a<X<b)=$
  - Rule 3: $P(X> \mu+d)=$

Empirical Rule Example

Suppose the monthly minutes used by customers of a cell phone provider is normally distributed with a mean of 344 minutes and a standard deviation of 22 minutes.

- What is the probability a randomly selected customer talked less than 344 minutes?

- What is the probability a randomly selected customer talked between 322 and 366 minutes?

- What is the probability a randomly selected customer talked more than 388 minutes?

- What is the probability a randomly selected customer talked less than 350 minutes?

Beyond the Empirical Rule

When we want to assess a probability for an x value that is not 1, 2, or 3 standard deviations away from the mean, we need to use a table or software/calculator to find the area under the curve.

Standardizing

- Recall: The _______________ for an observation is the number of standard deviations that it falls from the mean.

  – The formula is

- When X has a normal distribution with mean ___ and standard deviation ____, Z has a _____________ distribution with mean ____ and standard deviation _____. This is called the ____ distribution.

Normal Curve Example

The speeds of cars traveling on a highway are normally distributed with a mean of 73 miles per hour and a standard deviation of 3.5 miles per hour.

- What proportion of the cars traveling on this highway have a speed less than 65 mph? What is the z-score for 65 mph?

- What proportion of the cars have a speed between 69 and 79 mph? Draw a picture of a normal curve and shade the area corresponding to this problem.

Finding Percentiles

- We have shown how to find an area under a normal curve when given an X value from the distribution. Next, we look at how to figure out which X value from the distribution has a given left or right area. This type of computation is sometimes referred to as an _____________ normal computation.

- Example: Scores on a test are normally distributed with a mean of 500 and a standard deviation of 100. How high do you need to score to be in the top 10%?

- This value is referred to as the ______ percentile of the distribution.

More Normal Examples

The distribution of heights of college women is normal, with mean 65 inches and standard deviation 2.7 inches.

- Find the height such that 10% of college women are shorter than that height. Draw a picture of a normal curve and indicate this height on the curve.

- Determine the two heights that make up the middle 90% of the distribution.

More Normal Examples

An athletic association wants to sponsor a footrace. The time it takes to run the course is normally distributed with a mean of 58.6 minutes, and a standard deviation of 3.9 minutes

- The association decides to have a tryout run, and eliminate the slowest 30% of the racers. What should the cutoff time be in the tryout run for elimination?

- What is the value of the first quartile for this distribution?

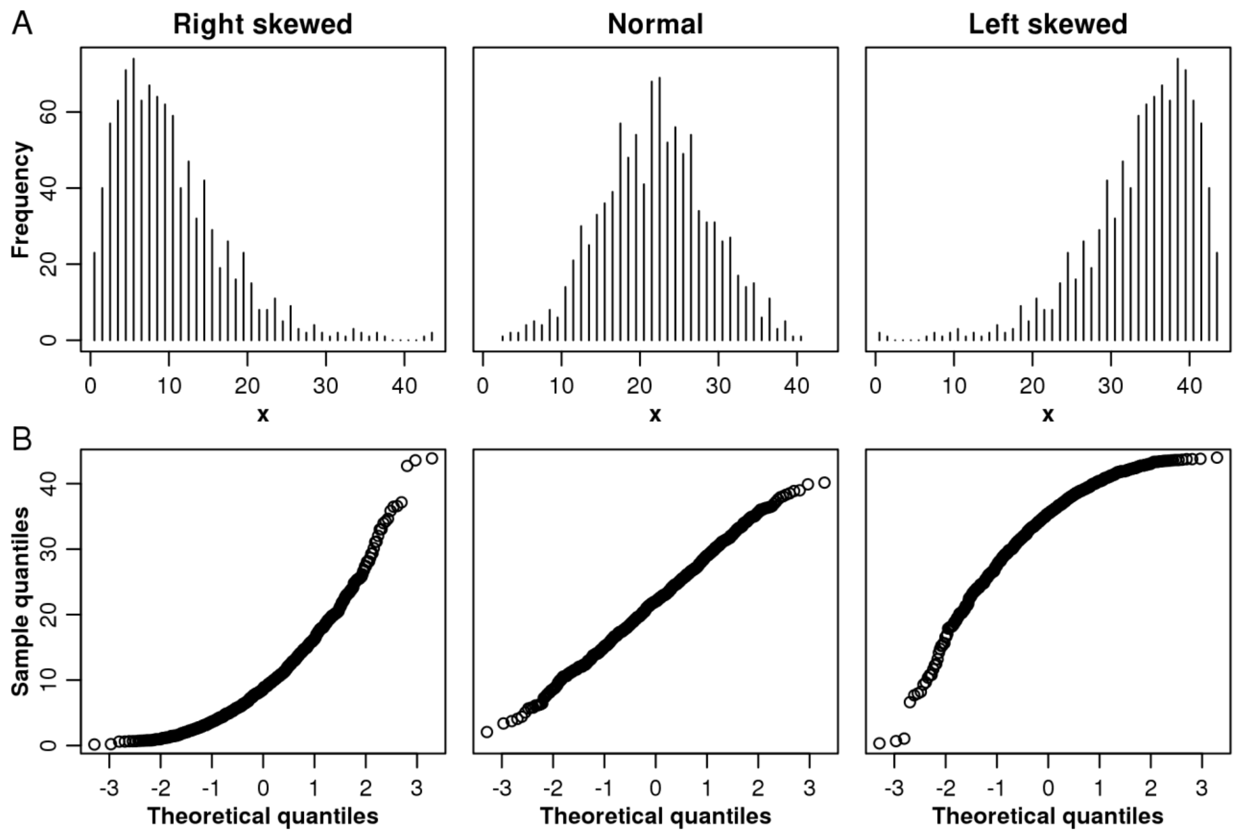
Assessing Normality

Many of the procedures we will be using in the remainder of this course are only valid for data that is normally distributed. If we have a sample of data and want to assess whether it is normal, we can

- Plot a ______________________
- See if the empirical rule fits the data
- Construct a **normal probability plot** using software. This is the best choice for smaller data sets.

Normal Probabilty Plot

A normal probability plot is a scatterplot of normal scores versus observed values. Normal scores are based on the sample size and tell us what the ordered set of z-scores we would expect to see in normally distributed data. When data are normally distributed, a normal probability plot follows a linear pattern.



Binomial Random Variables

The Binomial Distribution

- Each observation is binary:

- Examples:

Conditions for a binomial experiment

1. There are ***n* "trials"** where n is determined in advance and is not a random value.
2. **Two possible outcomes** on each trial, called "success" and "failure" and denoted S and F.
3. **Outcomes are independent** from one trial to the next.
4. **Probability of a "success"**, denoted by p, remains **same** from one trial to the next. Probability of "failure" is $1 - p$.

A **binomial random variable** is defined as X=number of successes in the n trials of a binomial experiment.

Notes on Binomial Experiments

- Variables that are not binary to begin with can often be categorized to be binary.

- Surveys produce binomial random variables when we count how many gave a particular response.
  - When the population is large compared to the sample we consider the observations to be approximately independent even though we are sampling without replacement because removing one person from a large population leaves the population almost the same.
  - Sampling without replacement from a small population does not give a binomial random variable. Outcomes cannot be considered independent.
- Any random circumstance can be thought of as a ___________ random variable, the result of a binomial experiment with $n=1$ and p= probability of a specific outcome.

Finding Binomial Probabilities

Guessing on a quiz: Jane is completely unprepared for her multiple choice quiz and guesses on all of the 3 questions. Each question has 5 possible response options. What is the probability that Jane will make a correct guess on exactly two of the three questions?

- The three ways Jane could make two correct guesses in three trials are:

- Each of these has probability:

- The total probability of two correct guesses is

Formula for Binary Probabilities

For k = 0, 1 , 2, …n

$$P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

Quiz Example Revisited

- The probability of exactly 2 correct guesses is the binomial probability with $n = 3$ trials, $x = 2$ correct guesses and $p = 0.2$ probability of a correct guess.

- What if the quiz had 10 questions and Jane guessed on all 10? The probability of guessing exactly 2 correctly is

Mean and standard deviation of a binomial distribution

If X has a binomial distribution B(n,p), then the mean and standard deviation are:

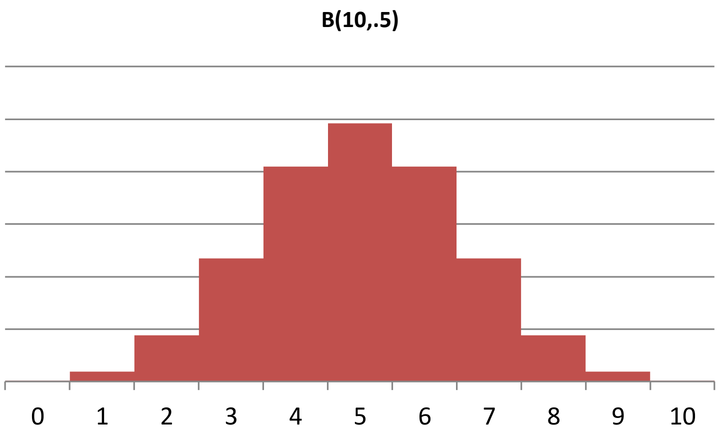
Jury Example

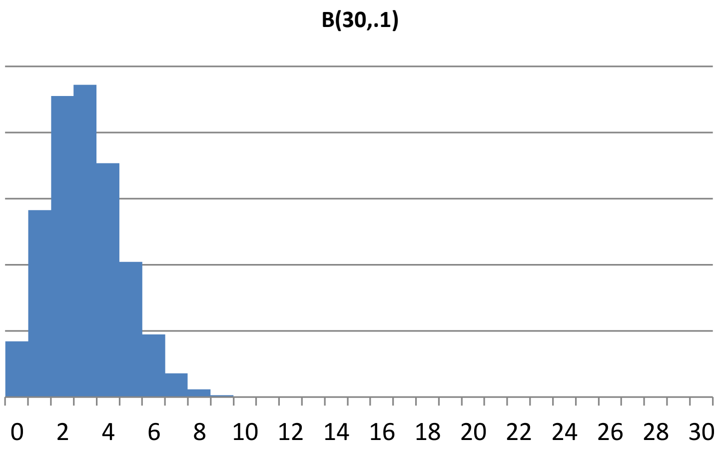
- Civil rights groups claim that a jury is not representative of the community.
  - Of eligible jurors in the community, 40% are non-white.
  - None of the 10 jurors chosen for the trial of a Hispanic defendant were non-white.
- How can we investigate statistically the civil rights groups' claim of bias?
  - If the employees are selected randomly, we expect about ______ of the 10 to be non-white. Due to ordinary sampling variation, it need not happen that exactly ____ % of those selected are non-white.
  - If jurors were actually selected at random for the trial, what are the chances that none of the 10 jurors selected were non-white?
  - Is this very likely to happen by chance?
- Give the full Binomial Probability Distribution for n=10 and p=.4

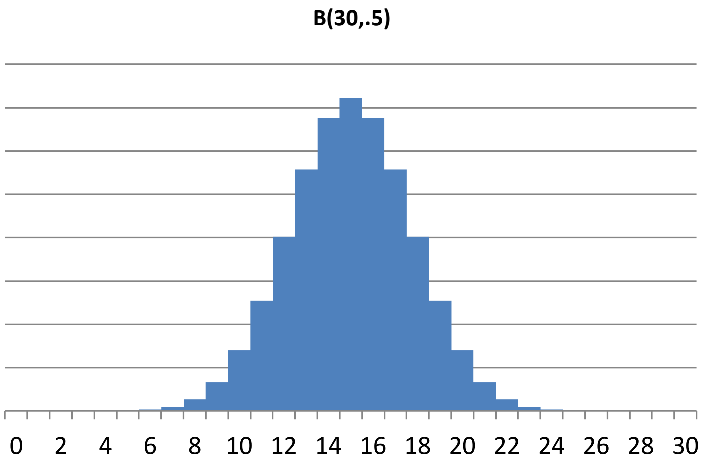
Approximating Binomial Distribution Probabilities

- When X has a binomial distribution with a large number of trials, the binomial probability formula is difficult to use because the factorial expressions in the formula become very large. The work required to find binomial probabilities when n is large is enormous. The normal distribution can be used to approximate these binomial probabilities.
-
- Let X be a binomial random variable based on n trials and probability of success p:
-
- If a SRS of size (n) is drawn from a large population and (n) is sufficiently large, the sampling distribution of X is approximately normal:
- Rule of thumb:

Does the normal approximation apply?







Jury Example 2

Over a period of several years, 456 of 1000 jurors in a particular court have belonged to a certain ethnic group. Census records show that 71% of eligible jurors in the region belong to that group.

- Let X be the number of jurors belonging to the ethnic group chosen using random selection. What is the distribution of X?

- What is the probability of 456 or fewer jurors belong to this group if jurors are selected randomly from the pool?

- What is the probability that between 500 and 650 jurors are belong to this group if selection is fair?