ST 517 Note Outline 3: Probability Distributions

Notes for Lecture 3.1: Random Variables & Probability

Recall:

- A **variable** is any characteristic of individuals in the population we want to learn about
- In general, there are two types of variables (two types of data):
  - **Categorical**: Places a unit into one of several groups or categories
  - **Quantitative**: Takes numeric values

Random Variables

- In simplest terms, a **random variable** is

- A key thing to note about random variables is that they are always quantitative

- Notation:

Basic Probability Facts

- In general the probability that some event A occurs is:

$$P(A) = \frac{\text{number of ways } A \text{ can occur}}{\text{total number of possible outcomes}}$$

- Probabilities must always be between 0 and 1 (inclusive)

- Total probability (over all possible outcomes) must be exactly 1
- Addition Rule for Disjoint Events: If two events A and B are **disjoint** (also called mutually exclusive), meaning they do not overlap or occur at the same time, their probabilities add

- Multiplication Rule for Independent Events: If two events A and B are **independent**, meaning the outcome of one event does not affect the outcome of the other, their probabilities multiply

- Complement Rule: Define $\bar{A}$ or A^c as the **complement** of event A, meaning that is contains all outcomes that are not a part of A. Since A and $\bar{A}$ are disjoint events that together represent all possible outcomes, their probabilities must add to 1

Example: We roll a standard six-sided playing die. The possible outcomes are:

a. What is the probability we roll an odd number?

b. What is the probability we roll a 4 or an odd number? we *don't* roll a 4 or an odd number?

c. We roll the die a second time. What is the probability both rolls are a 4?

Types of Random Variables

- Discrete random variable – Number of possible values is finite or countable
  - Examples:

- Continuous random variable – Can assume any value in an interval
  - Examples:

Notes for Lecture 3.2: Distributions

Distribution

- Overall pattern of how often possible values of a random variable occur
- Recall: 3 major elements of a distribution
  - Shape:
    - Skewed right, skewed left, symmetric
    - Number of modes
    - Outliers—Unusual values that do not fit with the rest of the pattern
  - Center (main chunk of data): Measured by mean, median
  - Variability (inconsistency in data values): Measured by variance, standard deviation, range, IQR
- Last outline: focused on observed distribution for sample data
- This outline: use mathematical model to describe population distribution
  - Nearly all of these are unimodal
  - We don't need to worry about outliers
  - Tend to focus on mean and variance/standard deviation as measures of center and variability

Probability Distributions for Discrete Random Variables

- Discrete random variables often represent counts, e.g. the number of individuals who meet a specified criteria or fall into a specified category

- **Probability mass function (pmf)**: Describes how likely each possible value of a discrete random variable is to occur

  - Properties:

    1.

    2.

      - Anything that satisfies these two criteria is a valid probability distribution.

- Some pmfs are indexed by _____ -- quantities that can take any one of several possible values

Example: $P(X = x) = \begin{cases} \alpha & \text{if } X = 1 \\ (1 - \alpha) & \text{if } X = 0 \end{cases}$ for $0 \leq \alpha \leq 1$

Probability Distributions for Continuous Random Variables

- Recall: a rv *X* is continuous if it can take any value in an interval

- Continuous random variables often represent measurements like heights, weights, speeds, differences or ratios of measurements, etc.
- **<u>Probability density function (pdf):</u>** The pdf, *f(x)*, for a continuous rv *X* has the property that for any two constants a and b ($a \leq b$):

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

  - Notes:

  - Properties:

    1. *f(x)* is a _____ function:

    2. Entire area under *f(x)* is 1:

    - Anything that meets these criteria is a valid pdf

Example: $f(x) = \begin{cases} 3x^2 \text{ if } 0 \leq X \leq 1 \\ 0 \quad \text{otherwise} \end{cases}$

Is this a valid pdf?

What is $P(0.2 < X < 0.5)$?

What is $P(X < 0.5)$?

Notes for Lecture 3.3: Expected Value

Expected Value

- If X is discrete, the expected value of X is:

- If X is continuous, the expected value of X is:

Example: $X \sim$ Bernoulli(α) $\quad P(X = x) = \begin{cases} \alpha & \text{if } X = 1 \\ (1-\alpha) & \text{if } X = 0 \end{cases}$ for $0 \leq \alpha \leq 1$

Example: $f(x) = 3x^2$ for $0 \leq x \leq 1$

Understanding Expected Value

- "Expected value," "mean," and "average" are all synonyms
- Single number summary that gives (some) information about an entire distribution
- Some common interpretations are:
    - A "typical" value for the rv
    - The balancing point
    - Long-run signal for a "noisy process"
- Expected values are useful for comparing distributions
    - Are men taller than women? We know that each individual male is not taller than each individual female, but
- You get more information about a distribution when you consider the expected value in connection with some other summary of the data
    - Comparing mean, median gives information about shape of the distribution
    - Considering mean, standard deviation together gives a much better idea of "typical" values for the rv

Expected Value of a Function

- Let *h(x)* be some function of *X*
- If *X* is discrete, the expected value of *h(X)* is:

- If *X* is continuous, the expected value of *h(X)* is:

Example: For each of the distributions we have considered, calculate $E(1+2X)$.

$X \sim$ Bernoulli(α)

$f(x) = 3x^2$ for $0 \leq x \leq 1$

Notes for Lecture 3.4: Variance and Standard Deviation

Variance

- Special case of an expected value of a function
- For both discrete and continuous random variables, the variance of *X* is:

- The **standard deviation** is the square root of the variance

Understanding Variance

- Both variance and standard deviation are measures of the spread of a distribution.
- Variance has squared units; standard deviation has the same units as the rv

- Standard deviation represent the "expected squared deviation from the mean" or, roughly speaking…

- Standard deviations are useful for comparing distributions
    - Ex: Weights of NFL players more varied than weights of NBA players

- The standard deviation , together with the mean, give a better understanding of the data values

Example: $X \sim$ Bernoulli(α) $\quad P(X = x) = \begin{cases} \alpha & \text{if } X = 1 \\ (1-\alpha) & \text{if } X = 0 \end{cases}$ for $0 \leq \alpha \leq 1$

Example: $f(x) = 3x^2$ for $0 \leq x \leq 1$

Notes for Lecture 3.5: Transformations Revisited

Transformations

- Recall: Data is often transformed (adjusted, rescaled, standardized) to better represent the values or compare variables
    - Additive transformations affect the mean but not the variance/standard deviation
    - Multiplicative transformations affect both the mean and the variance/standard deviation

- Formally: If a and b are constants, then $Y = a + bX$ is a transformation of the rv X

    - Expectation:

    - Variance:

Proofs:

- Expectation:

- Variance:

Example: Let X = temperature in °Fahrenheit and Y = temperature in °Celsius $= \frac{5}{9}(X - 32)$

Notes for Lecture 3.6: Famous Discrete Distribution—Binomial

Example: A multiple choice test has 20 questions. Suppose that the chance a particular student will answer each question correctly is *p*, and that the answers to each question are independent of each other.

a. What is the probability that the student answers all 20 questions correctly?

b. What is the probability that the student answers 0 questions correctly?

c. What is the probability that the student answers 18 out of 20 questions correctly?

- How many ways are there to answer 18 out of 20 questions correctly?
  - This question can be answered by considering the number of **<u>combinations</u>** (unordered groups) of size *r* that can be formed from *n* individuals:

  - So, the number of ways to answer 18 out of 20 questions correctly is:

- Bringing this all together: There are _____ terms that each have probability _____; thus the probability of getting 18 out 20 questions correct is:

- Example illustrates first famous distribution: the Binomial Distribution

The Binomial Distribution:

- PMF: $\binom{n}{x} p^x (1-p)^{n-x}$
- X can take integer values
- Depends on two parameters:
- Conditions for using Binomial Distribution:
- Mean of a Binomial random variable:
- Variance of a Binomial random variable:

Example: Using the Binomial distribution to make decisions

a. What is the probability of getting 70 “heads” out of 100 flips of a fair coin?

b. Now imagine you watch someone flip a coin 100 times and get 70 “heads;” would you believe that the coin was biased?

c. What is the probability of getting 7 “heads” out of 10 flips of a fair coin?

d. Imagine you watch someone flip a coin 10 times and get 7 “heads;” would you believe that the coin was biased?

Example: Suppose that about 10% of Americans are left-handed.
Let X = the number of left-handed Americans in a random sample of 12 Americans.

a. What are the mean and standard deviation of the number of left-handed Americans in the sample? Interpret each of these values.

b. What is the probability that the sample contains at most 2 left-handed Americans?

Notes for Lecture 3.7: Famous Continuous Distribution—Normal

The Normal Distribution

- Also known as the Gaussian distribution
- Most common continuous distribution since many random variables are well modeled by normal distributions

- PDF: $f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$
    - Notation:
- *X* can take any values
- Depends on two parameters:

- Bell-shaped, symmetric:

- To sketch a normal distribution:
    - Put μ at the _____; the curve is _____ here
    - 99.7% of the distribution is between
        - So the smallest value of *X* is approximately
        - And the largest value of *X* is approximately
    - A picture provides a useful frame of reference when problem solving

Example: $X \sim \mathrm{N}(7,2)$

Normal distribution (continued):

- Changing μ shifts entire distribution

- Changing σ controls how tall/flat the distribution is

Example: A professor teaches two sections of the same class, grading each on a curve so that a student's letter grade is based on their performance relative to that of their classmates. Grades on an exam in the first section have a mean of 85 with a standard deviation of 5. Grades on the same exam in the second section also have a mean of 85 but with a standard deviation of 3. The distributions of scores in both sections can be well modeled by a normal distribution. Suppose you are in this class and score 90 points on this exam. Which section would you rather be in (i.e. for which section would your letter grade for this exam be higher)?

Standard score: $z = \frac{x - \mu}{\sigma}$

- Tells you how many standard deviations a particular observation is from the mean

- Follows a **<u>standard normal distribution</u>**

- Most values will be between -3 and +3

Example: Using the Normal distribution to make decisions
One way to help classify dinosaur remains is by taking measurements of bones found. For example, a paleontologist could measure the width of a skull from the tip of the snout to a point at the back of the skull (in millimeters [mm]). Suppose a paleontologist is excavating in an area where two species of dinosaurs are known to have roamed. For the first species, skull widths can be well-modeled by a normal distribution with a mean of 619 mm with a standard deviation of 3.4 mm. For the second species, skull widths can be well-modeled by a normal distribution with a mean of 641 mm with a standard deviation of 10 mm. The paleontologist finds a skull with a width of 629 mm. Which species do you think it belongs to?

Notes for Lecture 3.8: Probabilities & Percentiles for a Normal Distribution

- Recall that we can find probabilities by integrating the pdf *f(x)*:

$$P(a < X < b) = \int_a^b f(x)dx$$

- Recall that for the normal distribution with mean μ and standard deviation σ:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

- Integrals using the normal distribution must be approximated, which was difficult in the days before computers.
    - Not to mention that there are an infinite number of possible normal distributions one could encounter.
    - To deal with this people would "standardize" each normal distribution using the standard score. There is a table which then gives probabilities based on the standard normal distribution:

$$P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$

    - However, tables are old-fashioned and only used in educational settings.
    - Fortunately for us, modern computing has improved to the point where we can use technology calculate probabilities for us.

The Normal Distribution: Finding Probabilities

- Focus on conceptual understand of what probability represents
- Probability represents area under the curve; the percent of the distribution that is covered by the event of interest
- Exams: Communicate understanding through writing the integral or providing a well-labeled picture
    - Need to include mean, standard deviation, and value of interest
- Quizzes, lecture examples: Use technology to calculate probabilities

Using Technology to Find Probabilities under the Normal Distribution

- Graphing calculator (e.g. TI-83 or 84)
  - Function: `normalcdf(`
  - Syntax: `normalcdf(LB, UB, mean, std_dev)`
- Software, e.g.
  - SAS:
```
DATA temp; prob_x = cdf('normal',UB,mean,std_dev);
    PROC PRINT; var prob_x; run;
```
  - Excel: `norm.dist(UB, mean, std_dev, TRUE)`
- Online calculators
  - E.g. stattrek.com/online-calculator/normal.aspx
  - Fill in: Value of x (or z) you are interested in, Mean, & Standard deviation
  - Click "Calculate" and computer will provide the Cumulative probability (e.g. area below the entered value of x)

Example: Scores on a standardized math test follow a normal distribution with a mean of 430 and a standard deviation 40. Janice scored 480; what percent of students scored below her?

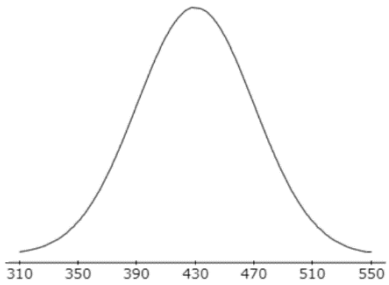
Writing the integral for this example:

$$P(X < 480) = \int_{-\infty}^{480} \frac{1}{\sqrt{2\pi(40^2)}} e^{-(x-430)^2/2(40^2)}\, dx$$

Note: $Z = \frac{X-\mu}{\sigma} = \frac{480-430}{40} = 1.25$, so this is equivalent to:

$$P(Z < 1.25) = \int_{-\infty}^{1.25} \frac{1}{\sqrt{2\pi}} e^{-(1.25^2)/2}\, dz$$

Picture for this example:



Using the technology for this example:

- Graphing calc: `normalcdf(-1000,480,430,40)`
- SAS:
```
DATA temp; prob_x = cdf('normal', 480, 430, 40);
    PROC PRINT; var prob_x; run;
```
- Excel: `norm.dist(480,430,40,TRUE)`
- Online calculator:

Standard score (z)	480			Normal random variable (x)	480
Cumulative probability P(Z ≤ z)		⇒	Calculate ⇒	Cumulative probability: P(X ≤ 480)	0.894
Mean	430			Mean	430
Standard deviation	40			Standard deviation	40

- From each of these:

The Normal Distribution: Finding Percentiles

- **Percentile** = value of variable that divides the distribution so that a specified percentage is below that value
  - Ex: 75th percentile is value of X such that 75% of area is less than x

- To calculate a percentile, you need to work backwards from the provided probability to solve for x:
  - Ex: 75th percentile

$$0.75 = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

- Finding percentiles under a normal distribution: Focus on conceptual understanding
- As before:
  - Exams: Communicate understanding through writing the integral or providing a well-labeled picture
  - Quizzes, lecture examples: Use technology to calculate percentiles

Using Technology to Calculate Percentiles under a Normal Distribution

- Graphing calculator (e.g. TI-83 or 84)
  - Function: `invNorm(`
  - Syntax: `invNorm(proportion to left, mean, std_dev)`
- Software, e.g.
  - SAS uses $N(0,1)$ so it returns a z-score! Need to solve z-score formula for x:

```
DATA temp; z=probit(proportion); x = (z*std_dev) + mean;
PROC PRINT; var x; run;
```

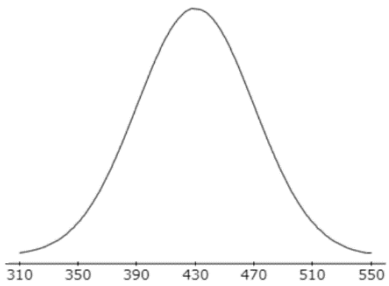
  - Excel: `norm.inv(proportion, mean, std_dev)`
- Online calculators
  - E.g. stattrek.com/online-calculator/normal.aspx
  - Fill in Cumulative probability, Mean, & Standard deviation
  - Click "Calculate" and computer will provide the value of x that is the appropriate percentile

Example: The principal of a high school wants give an award to students who score in the top 10% of the standardized mathematics test [recall: scores $\sim N(430,40)$]. What raw score has the top 10% above it?

Writing the integral for this example:

$$0.90 = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi(40^2)}} e^{-(y-430)^2/2(40^2)}\, dy$$

Picture for this example:



Using the technology for this example:

- Graphing calc: `invNorm(0.9,430,40)`
- SAS:
```
DATA temp; z = probit(proportion); x = (z*std_dev) + mean;
PROC PRINT; var x; run;
```
- Excel: `norm.inv(0.9,430,40)`
- Online calculator:

Standard score (z)				Normal random variable (x)	481.262
Cumulative probability P(Z ≤ z)	0.9	⇒	Calculate	⇒ Cumulative probability: P(X ≤ 481.262)	0.9
Mean	430			Mean	430
Standard deviation	40			Standard deviation	40

- From each of these:

Example: A proposal before the Board of Education has specified certain designations for elementary schools depending on how their students do on the Evaluation of Scholastic Scoring. This test is given to all students in all public schools in grades 2 through 4. Schools that score in the top 20% are labeled excellent. Schools in the bottom 25% are labeled "in danger" and schools in the bottom 5% are designated as failing. Previous data suggests that scores on this test are approximately normal with a mean of 75 and a standard deviation of 5.

a. What is the probability that a randomly selected school will score below 70?

b. What is the probability that a randomly selected school will score above 83?

c. What is the probability that a randomly selected school will score between 80 and 85?

d. What is the score cut-off required for schools to be labeled excellent?