ST 517 Note Outline 3: Distributions Part 1 (Foundations)

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

- **Probability** is a mathematical model to quantify and help us make sense of uncertainty
- 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

- Ways to Interpret Probability
    1. Basket Model

    2. Long term relative frequency

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

a. What is the probability we roll a 4? Not a 4?

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

c. What is the probability we roll a 4 *or* an odd number? 4 *and* an odd number?

d. 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
    - How often possible values occur = how likely they are to occur
    - This is formally measured by calculating a probability

- Recall: 3 major elements of a distribution
    - Shape: Skew vs. symmetry, number of modes, outliers [if any]
    - 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.

Probability Distributions for Discrete RVs (continued)

- 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

- **<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 and Variance

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
    - 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

Example: Heights for a certain population of females have an expected value of 65 inches. Write a sentence interpreting this value in context.

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 ~ Bernoulli(α)

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

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

Variance

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

- Recall: the **<u>standard deviation</u>** is the square root of the variance

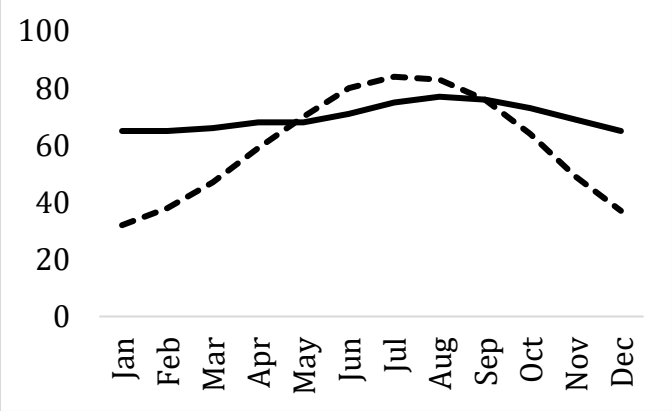
Understanding Variance and Standard Deviation

- Recall: Both are measures of the spread of a distribution.
- Recall: variance has squared units; standard deviation has the same units as the rv, so standard deviation is typically reported in practice
    - But variance has some nice mathematical properties we will reply on
- Recall: (Variance and) Standard deviation represent the average distance between the points and their mean (roughly speaking)
- Variance and standard deviation give a sense of how reliable mean is as "typical" value

- Variance and standard deviations are useful for comparing distributions
    - Ex: Standard deviation in average monthly high temperature (degrees Fahrenheit)

San Diego, CA

Chicago, IL



Example: Heights for a certain population of females have an expected value of 65 inches and a standard deviation of 2.8 inches. Write a sentence interpreting the standard deviation in context.

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 3.4: Determining Which Distribution to Use

- Every probability distribution is a mathematical function used to model a variable
    - Quantifies how likely is each value of the variable is to occur, what value is expected, and how spread out are other values from what is expected
    - Many "famous" examples, e.g. Bernoulli, Binomial, Normal, Student's t
- How do you know which distribution is the best model for a given variable?
    - Discrete random variables: Many discrete distributions follow certain rules for counting the number of "successes"

    - Continuous random variables: Continuous distributions are often identified by their shape
        - Several continuous distributions have the same general shape
        - Specifics determined by the parameter values for that distribution
        - How to distinguish between distributions and/or parameter values?
            - Parametric method: Probability plots
            - Non-parametric method: Kernel density estimation

Parametric vs. Non-parametric

- Sometimes a statistical method is referred to as *parametric* or *non-parametric*
- **Parametric** means that
    - All of the methods for conducting statistical inference we will learn this semester are parametric because they assume the population data follows a specific distribution, such as the normal distribution
- **Non-parametric** means that you
    - Non-parametric equivalents of the methods for conducting statistical inference are beyond the scope of this course
    - Methods to summarize data are typically non-parametric
        - One exception is the probability plot

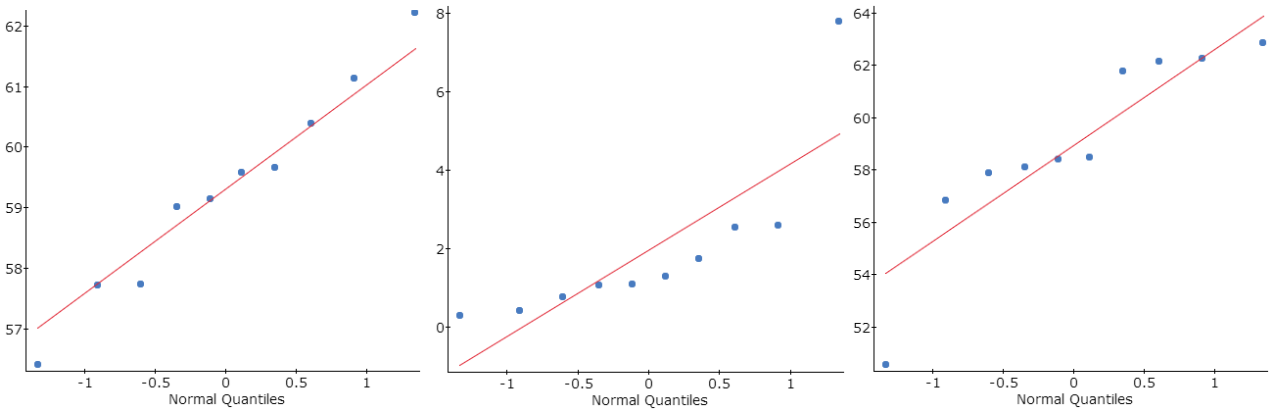
Probability Plots

- **Probability plot** can be used to determine if data follows a particular parameterized distribution
- Basic idea:

Probability Plots (continued)

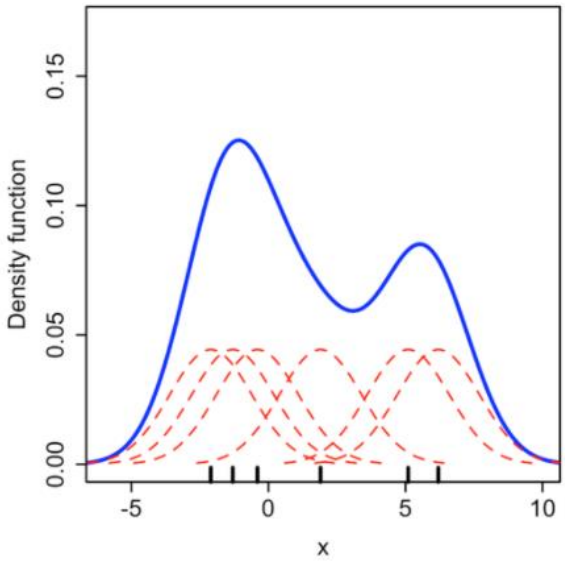
- Different software follows different conventions in creating the plots, but one axis will represent the percentiles of the sample data and the other axis will represent the corresponding percentiles of the pdf
    - The pdf fits the data well if the points fall along the line y=x
    - In practice, quantities may not be plotted on the same scale (e.g. values may be standardized), so the points may not fall exactly on this line
- If the distribution is a good fit for the data, the points to fall along a straight, upward sloping (approximately 45°) line
- Any strong deviation from this indicates the pdf is not a good fit for the data

Example: Below are three probability plots, each checking if a different dataset is well modeled by a Normal distribution. Which plot indicates a good fit?



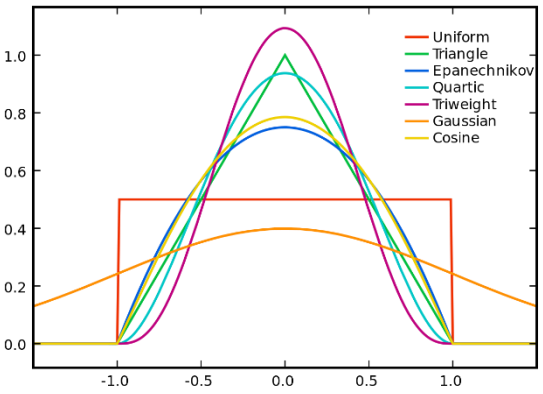
Kernel Density Estimation

- **Kernel density estimation** (KDE) is a non-parametric method used to summarize the probability distribution of a random variable
    - Similar to a histogram, but imagine if we could smooth the histogram so that the 'boxes' blurred together; or you could think of it as drawing a smooth line over the top of the boxes
- Main idea: KDE weights the distances for each data point based on a kernel function of a specified bandwidth (red dashed curves in plot at right), then builds those up to get an estimate of the probability density function (solid blue curve)

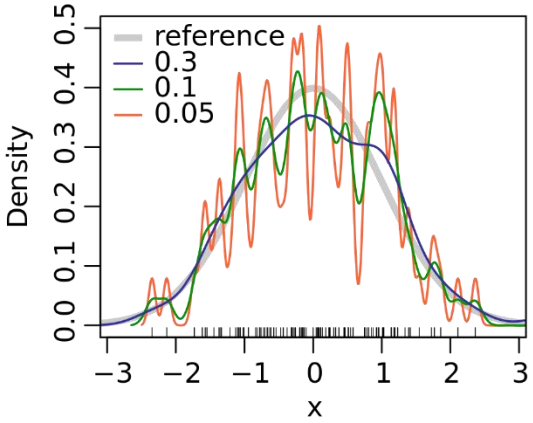


- **Kernel:** Weighting function
    - Common examples include: Gaussian, Uniform, Triangular, Epanechnikov (see picture below, left)
- **Bandwidth** (also called **smoothing parameter** and often denoted h or λ): Controls the width of the kernel
    - Choice of the value for the bandwidth has an important impact on the resulting estimate (see picture below, right)
        - Larger values of λ result in smoother curves
        - Smaller values of λ result in 'spikier' curves

Kernel Options



Impact of Bandwidth



* All pictures on this page from Wikipedia, see full articles here and here (accessed on August 17, 2020)

KDE (continued):

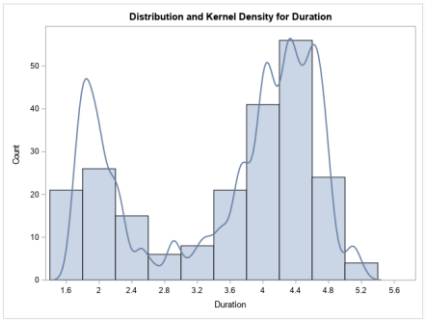
- Want to find value of λ that minimizes the **<u>mean square error</u>** (MSE)
    - In general, MSE is the expected value of the squared distance between an estimator and the parameter you are trying to estimate
    - In this context, MSE is the expected value of the squared distance between the kernel density estimate and the true probability distribution for the data:

$$E[(f_{\text{kde}}(x) - f(x))^2]$$

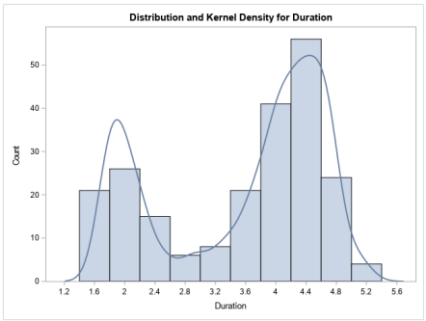
    - Algorithms can be used to find the optimal value of λ for the data

Example: Below are kernel density estimates of the probability density function for the Old Faithful geyser eruption data using three different bandwidths. Which does best at representing the underlying histogram?

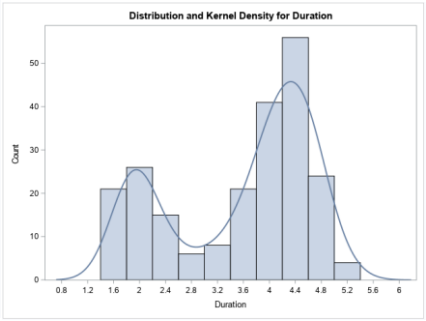
$\lambda = 0.08$



$\lambda = 0.16$



$\lambda = 0.32$



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 measures of center but not variability
    - Multiplicative transformations affect both measures of center and variability
- 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 = average monthly high temperature in °Fahrenheit for Chicago; Y = corresponding temperature in °Celsius $= \frac{5}{9}(X - 32)$. Find the mean and variance of Y.

Example: Let X be a rv with mean μ and standard deviation σ. Find the mean and standard deviation of $Z = \frac{X-\mu}{\sigma}$.

Lecture 3.6: Distributions for Linear Combinations of Variables

Recall: Transformations

- If a and b are constants, then $Y = a + bX$ is a linear transformation of the rv X
- Expectation: $E(a + bX) = a + b\,E(X) = a + b\mu$
- Variance: $V(a + bX) = b^2V(X) = b^2\sigma^2$
- We derived these properties in the case of a single random variable
    - They can be extended to multiple random variables

Propositions:

- Let $X_1,\ldots,X_n$ be rvs with $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$.
- Let Y be a **<u>linear combination</u>** of the X_i's: $Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n$.

1. Expectation of a linear combination, for any $X_1,\ldots,X_n$:

2. Variance of a linear combination, for <u>independent</u> $X_1,\ldots,X_n$:

3. If $X_1,\ldots, X_n$ are normally distributed rvs (possibly with different means and/or variances), then any linear combination of the X_i's also has a normal distribution.

- These will all be important when we discuss sampling distributions—an special type of distribution that is key for statistical inference