STA 371G: Statistics and Modeling

Review of Basic Probability and Statistics: Random Variables and Probability Distributions

Mingyuan Zhou
McCombs School of Business
The University of Texas at Austin

http://mingyuanzhou.github.io/STA371G

Getting Started

- Course website
  `http://mingyuanzhou.github.io/STA371G`
- Syllabus
- Outline (day by day)
- Slides
- General Expectations
  1. Be on schedule
  2. Read the assigned materials
  3. Turn in homework on time
  4. Participate in discussions
  5. Get familiar with data analysis using Excel and/or R

Course Overview

- Review of Basic Probability and Statistics
- Linear Regression
- Time Series
- Decision Making
- Simulation with Excel and R

Course Overview

- Review of Basic Probability and Statistics
  - Random Variables and Probability Distributions
  - Normal and Binomial Distributions
  - Estimation and Sampling Distributions

Course Overview

- Simple Linear Regression
  - Least Squares Estimation
  - Covariance and Correlation, Goodness of Fit
  - Model Assumptions
  - Sampling Distributions and Confidence Intervals for Regression Parameters
- Multiple Linear Regression
  - Multiple Linear Regression
  - Dummy Variables and Interactions
  - Diagnostics and Transformations
  - Logistic regression

Course Overview

- Time Series
  - Fitting a Trend
  - Autoregressive Models
  - Modeling Seasonality

Course Overview

- Decision Making
  - Probability, Betting Odds and Bayes' Theorem
  - Decision Criteria and Utility Functions
  - Decision Trees
  - The Value of Information
- Introduction to Monte Carlo Simulation

Cases to be Studied

- Amore Frozen Foods
- Waite First Securities
- Milk and Money
- Orion Bus Industries: Contract Bidding Strategy
- Oakland A's – A
- Oakland A's – B
- Northern Napa Valley Winery, Inc.
- Freemark Abbey

Introduction to Data

Statistics is the study of how to collect, analyze, and draw conclusions from data

- Identify the problem
- Collect relevant data
- Analyze the data
- Make conclusions and/or decisions

Data Basics

- Types of variables
  - Continuous
  - Discrete
  - Categorical
- Relationships between variables
  - Association
  - Correlation
  - Causation
  - Independence

Populations and Samples

- Population
- Sample
- Sampling from a population
- Explanatory and response variables
- Inference

Review of Basic Concepts

Probability and statistics help us deal with uncertainty.

- if I only ask 1,000 voters out of 10 million, how sure can I be about how they all will vote? What is the *true* proportion of "yes voters".
- if I am trying to choose my portfolio, how sure am I about returns on the assets next period?
- if I am trying to predict sales next quarter, how sure am I?
- if I want to do target marketing, which customers are more "likely" to respond to a promotion?

All of these involve inferring or predicting unknown quantities!!

Random Variables

- *A Random Variable* represents each possible random outcome with a numerical value; each numerical value is associated with a probability; the probabilities of all possible outcomes sum to one.
- Example 1: Bernoulli random variable $X \sim \text{Bernoulli}(p)$. The two possible outcomes $X = 1$ and $X = 0$ happen with probabilities p and $1 - p$, respectively.
- Example 2: Using a random variable X to describe the total number of heads observed by flipping a coin twice.
- Example 3: Using a random variable X to describe the average price per square foot of a house at Austin.

Probability

Probability assigns an event a number between 0 and 1 that measures how likely the event is to occur. The closer the probability is to 1, the more likely the event is to happen.

1. If an event A is certain to occur, it has probability 1, denoted $P(A) = 1$.
2. If two events A and B are mutually exclusive (both cannot occur simultaneously), then $P(A \text{ or } B) = P(A) + P(B)$.
3. $P(\text{not-}A) = 1 - P(A)$.

$\mathrm{P}(\text{Salary} \leq 120\text{K}) = 0.70$, $\mathrm{P}(\text{Salary} \geq 200\text{K}) = 0.05$

$\mathrm{P}(\text{Salary} \leq 120\text{K or Salary} \geq 200\text{K}) = ?$

$\mathrm{P}(120\text{K} < \text{Salary} < 200\text{K}) = ?$

Probability

- Probability is always a positive number
- It can take values between 0 and 1.
- The probabilities of all possible values of a random variable sum to 1.
- If events A and B are not mutually exclusive, then

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

Probability Distribution

- We describe the probabilities of all possible values of a random variable with a Probability Distribution.
- Example: If X is the random variable denoting the number of heads in two *independent* coin tosses, we can describe its random values through the following probability distribution:

$$X = \begin{cases} 0 & \text{with prob.} \quad 0.25 \\ 1 & \text{with prob.} \quad 0.50 \\ 2 & \text{with prob.} \quad 0.25 \end{cases}$$

- A *Discrete Random Variable* has a finite (or countably infinite) number of possible discrete values. Question: What is $P(X = 0)$? How about $P(X \geq 1)$?
- Future topic: Continuous Random Variables.

Mean and Variance of Random Variables

We can describe a discrete random variable X by listing all its possible values x_i and their probabilities $P(x_i)$, which shall satisfy

$$\sum_{i=1}^{k} P(X = x_i) = 1, \quad P(X = x_i) \geq 0.$$

It is usually convenient to summarize X with its mean and variance.
The Mean or Expected Value is defined as (for a discrete X):

$$\mu = E(X) = \sum_{i=1}^{k} x_i P(X = x_i)$$

We weight all possible values by how likely they are... this provides us with a measure of centrality of the distribution... a “good” prediction for X!

Mean and Variance of Random Variables

Suppose $X \sim \text{Bernoulli}(p)$

$$
\begin{aligned}
E(X) &= \sum_i x_i P(X = x_i) \\
&= 0 \times (1 - p) + 1 \times p \\
E(X) &= p
\end{aligned}
$$

Mean and Variance of Random Variables

The Variance is defined as (for a discrete X):

$$\sigma^2 = Var(X) = \sum_{i=1}^{k} [x_i - E(X)]^2 P(X = x_i)$$
$$= E(X^2) - [E(X)]^2$$

Weighted average of squared prediction errors... This is a measure of spread of a distribution.

Mean and Variance of Random Variables

Suppose $X \sim Bernoulli(p)$

$$
\begin{aligned}
Var(X) &= \sum_i [x_i - E(X)]^2 \, P(x_i) \\
&= (0-p)^2 \times (1-p) + (1-p)^2 \times p \\
&= p(1-p) \times [(1-p) + p] \\
Var(X) &= p(1-p)
\end{aligned}
$$

Question: For which value of p is the variance the largest?

Standard Deviation

- A more intuitive way to understand the spread of a distribution is to look at the standard deviation:

$$\sigma = sd(X) = \sqrt{Var(X)}$$

Population/Sample Mean and Variance

How can we find out the mean of the starting salaries of all US college students graduated in 2013?

- Mean and variance of random variable X:

$$\mu = E(X),\ \sigma^2 = Var(X) = E[(X - E(X))^2]$$

- Population mean and variance:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N},\ \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

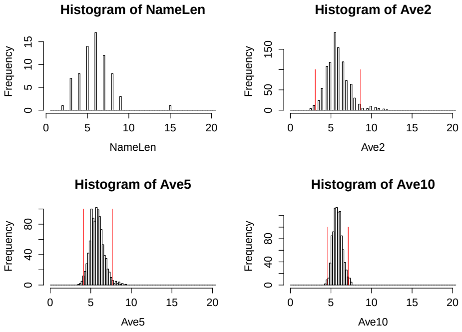
- Sample mean and variance:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n},\ s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- How accurate is the sample mean $\bar{x}$ as an estimate of the population mean? We will come back to this later.

Example: STA 371G Student Last Name Length

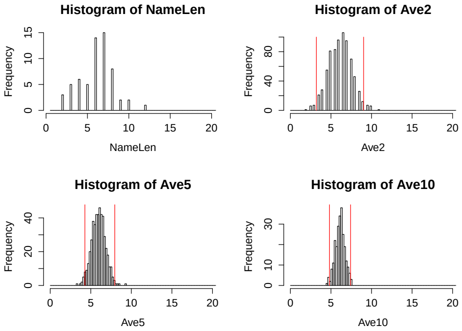
- 2014 Session 04635, 12:30-2:00 PM, Enrollment = 71
- Repeat 1000 times and plot the histogram of the averages
  - Randomly sample *n* students from the class
  - Record the average length of the students' last names.



- What can you observe from the figure?
- Run the R code: Jan14_NameLength.R

Example: STA 371G Student Last Name Length

- 2014 Session 04640, 2:00-3:30 PM, Enrollment = 61
- Repeat 1000 times and plot the histogram of the averages
  - Randomly sample *n* students from the class
  - Record the average length of the students' last names.



- What can you observe from the figure?
- Run the R code: Jan14_NameLength.R

Adding or Multiplying a Constant

Both a and b are constants:

- If $Y = a + X$, then

$$E(Y) = a + E(X), \quad Var(Y) = Var(X), \quad sd(Y) = sd(X).$$

- If $Y = bX$, then

$$E(Y) = bE(X), \quad Var(Y) = b^2 Var(X), \quad sd(Y) = |b| \times sd(X).$$

- If $Y = a + bX$, then

$$E(Y) = a+bE(X), \quad Var(Y) = b^2 Var(X), \quad sd(Y) = |b| \times sd(X).$$

Conditional, Joint and Marginal Distributions

In general, we want to use probability to address problems involving more than one variable.

We need to be able to describe what we think will happen to one variable relative to another... we want to answer questions like: How are my sales impacted by the overall economy?

Conditional, Joint and Marginal Distributions

Let E denote the performance of the economy next quarter... for simplicity, say $E = 1$ if the economy is expanding and $E = 0$ if the economy is contracting (what kind of random variable is this?) Let's assume $P(E = 1) = 0.7$

Let S denote my sales next quarter... and let's suppose the following probability statements:

S	$P(S\|E = 1)$	S	$P(S\|E = 0)$
1	0.05	1	0.20
2	0.20	2	0.30
3	0.50	3	0.30
4	0.25	4	0.20

These are called *Conditional Distributions*

Conditional, Joint and Marginal Distributions

S	$P(S\|E=1)$	S	$P(S\|E=0)$
1	0.05	1	0.20
2	0.20	2	0.30
3	0.50	3	0.30
4	0.25	4	0.20

- In blue is the conditional distribution of S given $E = 1$
- In red is the conditional distribution of S given $E = 0$
- We read: *the probability of Sales of 4 ($S = 4$)* **given(or conditional on)** *the economy is growing ($E = 1$) is 0.25*

Conditional, Joint and Marginal Distributions

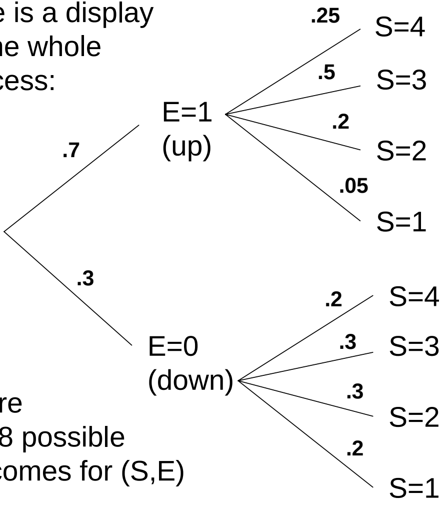
The conditional distributions tell us about about what can happen to S for a given value of E... but what about S and E jointly?

$$
\begin{aligned}
P(S = 4 \quad \text{and} \quad E = 1) &= P(E = 1) \times P(S = 4|E = 1) \\
&= 0.70 \times 0.25 = 0.175
\end{aligned}
$$

In english, 70% of the times the economy grows and 1/4 of those times sales equals 4... 25% of 70% is 17.5%

Conditional, Joint and Marginal Distributions

here is a display of the whole process:



p(S=4 and E=1) = .7*.25 = .175

p(S=3 and E=1) = .7*.5 = .35

p(S=2 and E=1) = .7*.2 = .14

p(S=1 and E=1) = .7*.05 = .035

p(S=4 and E=0) = .3*.2 = .06

p(S=3 and E=0) = .3*.3 = .09

p(S=2 and E=0) = .3*.3 = .09

p(S=1 and E=0) = .3*.2 = .06

There are 8 possible outcomes for (S,E)

Conditional, Joint and Marginal Distributions

We call the probabilities of E and S together the joint distribution of E and S.

In general the notation is...

- $P(Y = y, X = x)$ is the joint probability of the random variable Y equal y **AND** the random variable X equal x.
- $P(Y = y|X = x)$ is the conditional probability of the random variable Y takes the value y **GIVEN** that X equals x.
- $P(Y = y)$ and $P(X = x)$ are the marginal probabilities of $Y = y$ and $X = x$

Important relationships

Relationship between the joint and conditional...

$$P(y|x) = \frac{P(x,y)}{P(x)}$$

$$\begin{aligned} P(y,x) &= P(x) \times P(y|x) \\ &= P(y) \times P(x|y) \end{aligned}$$

Relationship between joint and marginal...

$$\begin{aligned} P(x) &= \sum_y P(x,y) \\ P(y) &= \sum_x P(x,y) \end{aligned}$$

Conditional, Joint and Marginal Distributions

Why we call marginals marginals... the table represents the joint and at the margins, we get the marginals.

		S				
		1	2	3	4	
E	0	.06	.09	.09	.06	.3
	1	.035	.14	.35	.175	.7
		.095	.23	.44	.235	1

Conditional, Joint and Marginal Distributions

Example... Given $E = 1$ what is the probability of $S = 4$?

		S				
		1	2	3	4	
E	0	.06	.09	.09	.06	.3
	1	.035	.14	.35	.175	.7
		.095	.23	.44	.235	1

$$P(S = 4|E = 1) = \frac{P(S = 4, E = 1)}{P(E = 1)} = \frac{0.175}{0.7} = 0.25$$

Conditional, Joint and Marginal Distributions

Example... Given $S = 4$ what is the probability of $E = 1$?

		S				
		1	2	3	4	
E	0	.06	.09	.09	.06	.3
	1	.035	.14	.35	.175	.7
		.095	.23	.44	.235	1

$$P(E = 1 | S = 4) = \frac{P(S = 4, E = 1)}{P(S = 4)} = \frac{0.175}{0.235} = 0.745$$

Independence

Two random variables X and Y are *independent* if

$$P(Y = y|X = x) = P(Y = y)$$

for all possible x and y.

In other words,

knowing X tells you nothing about Y!

e.g.,tossing a coin 2 times... what is the probability of getting H in the second toss given we saw a T in the first one?

Independence

Joint distribution of independent random variables:

$$P(X = x \text{ and } Y = y) = P(X = x)P(Y = y)$$

Flipping a coin three times, what's the probability of seeing three heads ($X_1 = X_2 = X_3 = 1$)?

Sum of Independent Random Variables

- if $Y = aX$, then $E(Y) = aE(X)$ and $Var(Y) = a^2 Var(X)$
- If $Y = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$, then

$$E(Y) = a_0 + a_1 E(X_1) + a_2 E(X_2) + \cdots + a_n E(X_n)$$

- If X_i and X_j are independent for $i \neq j$, then we further have

$$Var(Y) = a_1^2 Var(X_1) + a_2^2 Var(X_2) + \cdots + a_n^2 Var(X_n)$$

- Future topic: Correlated random variables.
- Uncorrelated $\neq$ independent!