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

- Syllabus, Outline and Slides

  `http://mingyuanzhou.github.io/STA371G`

- General Expectations

  1. Read the assigned materials
  2. Turn in homework on time
  3. Be on schedule

Course Overview

- Review of Basic Probability and Statistics
- Simple Linear Regression
- Multiple Regression
- Dummy Variables and Interactions
- Regression Diagnostics and Transformations
- Time Series
- Decision Making
- Simulation with Excel and R

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

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 predict sales next quarter, how sure am I?
- if I am trying to choose my portfolio, how sure am I about returns on the assets next period?
- 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!!

Example: STA 371G Spring 2014 Enrollment

- I teach two sessions of STA 371G, as of January 13
  - Session 04635, 12:30-2:00 PM, Enrollment = 71
  - Session 04640, 2:00-3:30 PM, Enrollment = 61
- Null Hypothesis: students do not prefer the early afternoon session to the late afternoon one
  - I repeat the following experiment 10,000 times
    - Flip a coin 71+61 times, record the difference between the number of heads and the number of tails
  - I find that more than 2000 times that the difference is larger or equal to 10
- Do I have enough evidence to reject the Null Hypothesis?
- What if I observe the same difference for the next 7 years?
- Run the R code: Jan14_ClassEnrollment.R

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. $P(\text{not-}A) = 1 - P(A)$.

P(Salary$\leq$120K) = 0.70, P(Salary$\geq$200K) = 0.05

P(Salary$\leq$120K or Salary$\geq$200K) = ?

P(120K$<$Salary$<$200K)=?

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 sums 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 $Pr(X = 0)$? How about $Pr(X \geq 1)$?
- 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 each possible value 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?

- Population mean and variance:

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

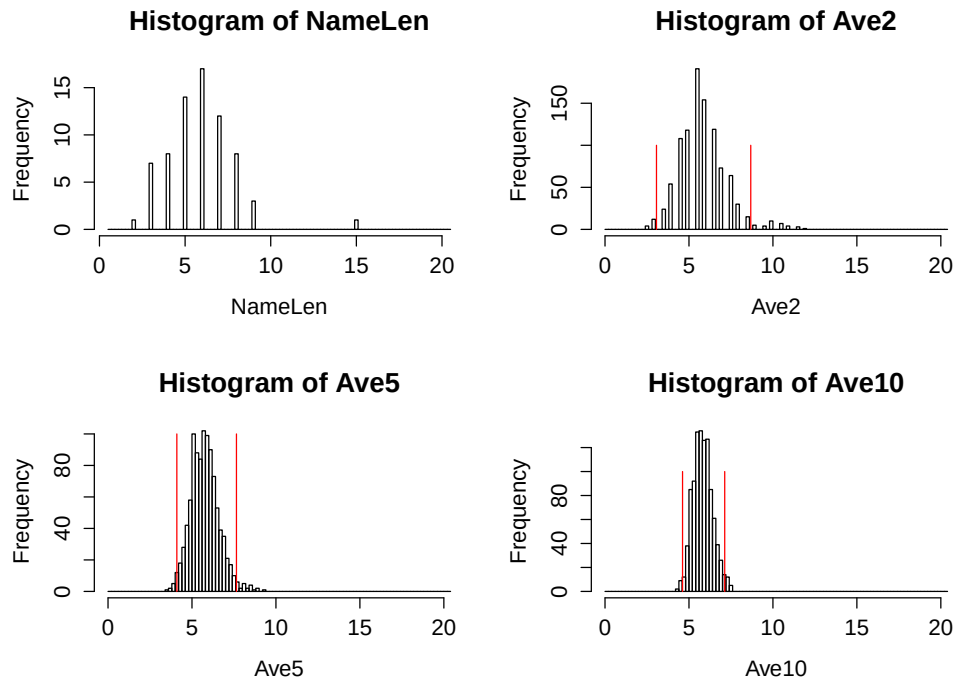
- Sample mean and variance:

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

- 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

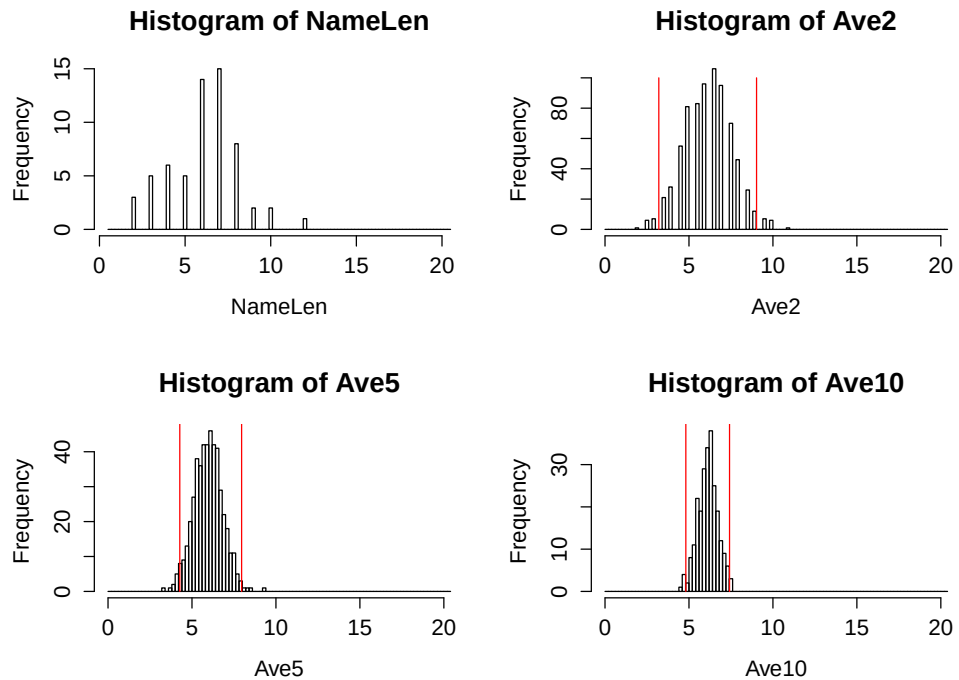
- Session 04635, 12:30-2:00 PM, Enrollment = 71
- Repeat 1000 times and plot the histogram of the averages
  - Randomly sample n students from the class
  - Record the average length of the students' last names.



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

Example: STA 371G Student Last Name Length

- 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 at the time

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 $pr(E = 1) = 0.7$

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

S	$pr(S\|E = 1)$	S	$pr(S\|E = 0)$
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	$pr(S\|E=1)$	S	$pr(S\|E=0)$
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

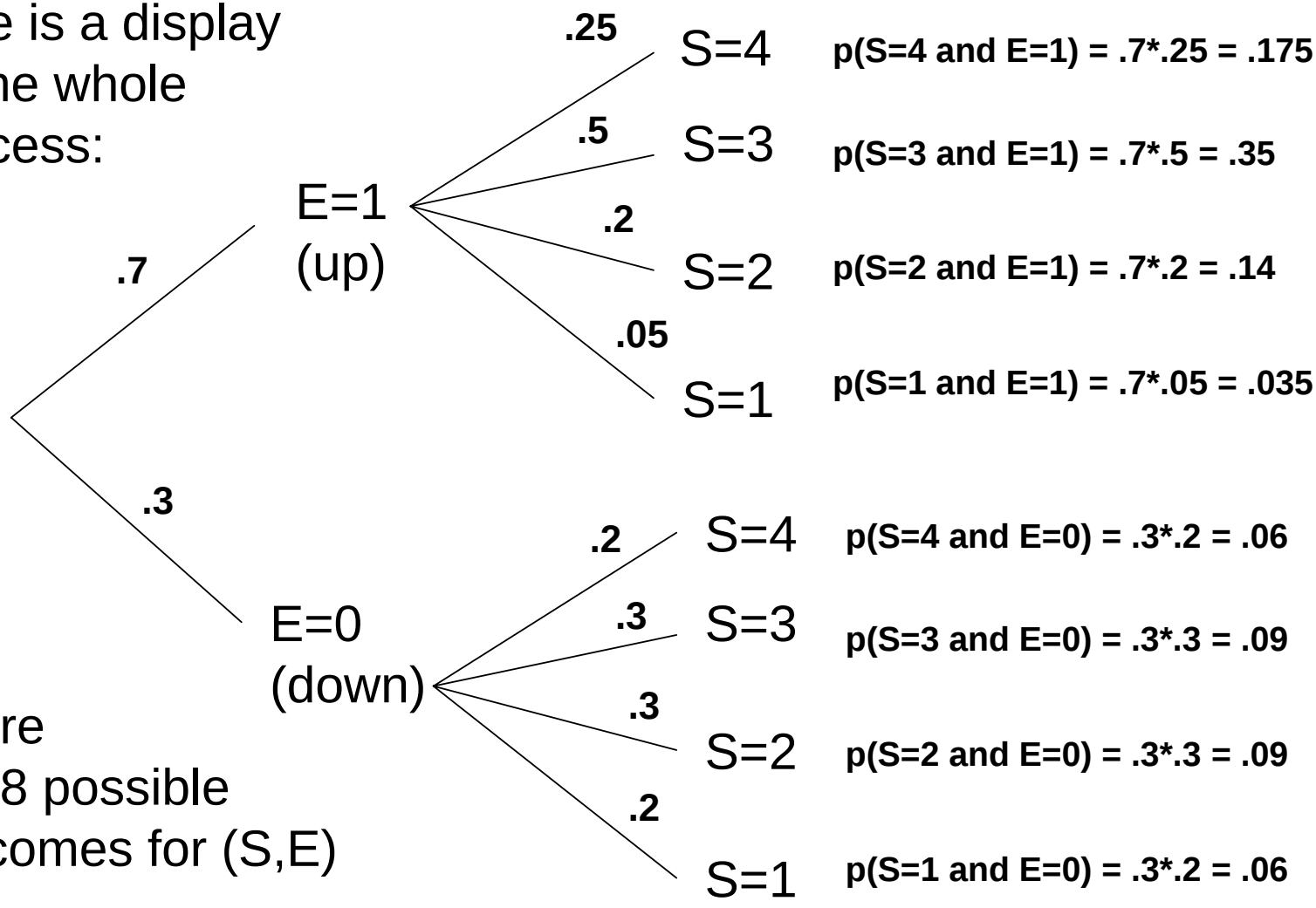
The conditional distributions tell us about about what can happen to S for a given value of E... but what about S and E jointly?

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

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:



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...

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

Important relationships

Relationship between the joint and conditional...

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

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

Relationship between joint and marginal...

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

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 $S = 4$?

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

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

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

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

Independence

Two random variables X and Y are *independent* if

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

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!