Section 1: Introduction, Probability Concepts and Decisions

Carlos M. Carvalho
The University of Texas at Austin
McCombs School of Business

http://faculty.mccombs.utexas.edu/carlos.carvalho/teaching/

Suggested Reading:
Naked Statistics, Chapters 1, 2, 3, 5, 5.5 and 6

Getting Started

- Syllabus
- General Expectations
  1. Read the notes
  2. Work on homework assignments
  3. Be on schedule

Course Overview

Section 1: Introduction, Probability Concepts and Decisions

Section 2: Learning from Data: Estimation, Confidence Intervals and Testing Hypothesis

Section 3: Simple Linear Regression

Section 4: Multiple Linear Regression

Section 5: More on MLR, Dummy Variables, Interactions

Let's start with a question...

My entire portfolio is in U.S. equities. How would you describe the potential outcomes for my returns in 2017?

Introduction

Probability and statistics let us talk efficiently about things we are unsure about.

- How likely is Trump to finish a four year term?
- How much will Amazon sell next quarter?
- What will the return of my retirement portfolio be next year?
- How often will users click on a particular Facebook ad?

All of these involve inferring or predicting unknown quantities!!

Random Variables

- *Random Variables* are numbers that we are NOT sure about but we might have some idea of how to describe its potential outcomes.
- Example: Suppose we are about to toss two coins.
  Let X denote the number of heads.

  We say that X, is the random variable that stands for the number we are not sure about.

Probability

Probability is a language designed to help us talk and think about aggregate properties of random variables. The key idea is that to each event we will assign a number between 0 and 1 which reflects how likely that event is to occur. For such an immensely useful language, it has only a few basic rules.

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

Probability Distribution

- We describe the behavior of random variables 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 behavior through the following probability distribution:

$$X = \begin{cases} 0 & \text{with prob.} \quad 0.25 \\ 1 & \text{with prob.} \quad 0.5 \\ 2 & \text{with prob.} \quad 0.25 \end{cases}$$

- X is called a *Discrete Random Variable* as we are able to list all the possible outcomes
- Question: What is $Pr(X = 0)$? How about $Pr(X \geq 1)$?

Conditional, Joint and Marginal Distributions

In general we want to use probability to address problems involving more than one variable at the time

Think back to our first question on the returns of my portfolio... if we know that the economy will be growing next year, does that change the assessment about the behavior of my returns?

We need to be able to describe what we think will happen to one variable relative to another...

Conditional, Joint and Marginal Distributions

Here's an example: we want to answer questions like: How are my sales impacted by the overall economy?

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 $pr(E = 1) = 0.7$

Conditional, Joint and Marginal Distributions

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

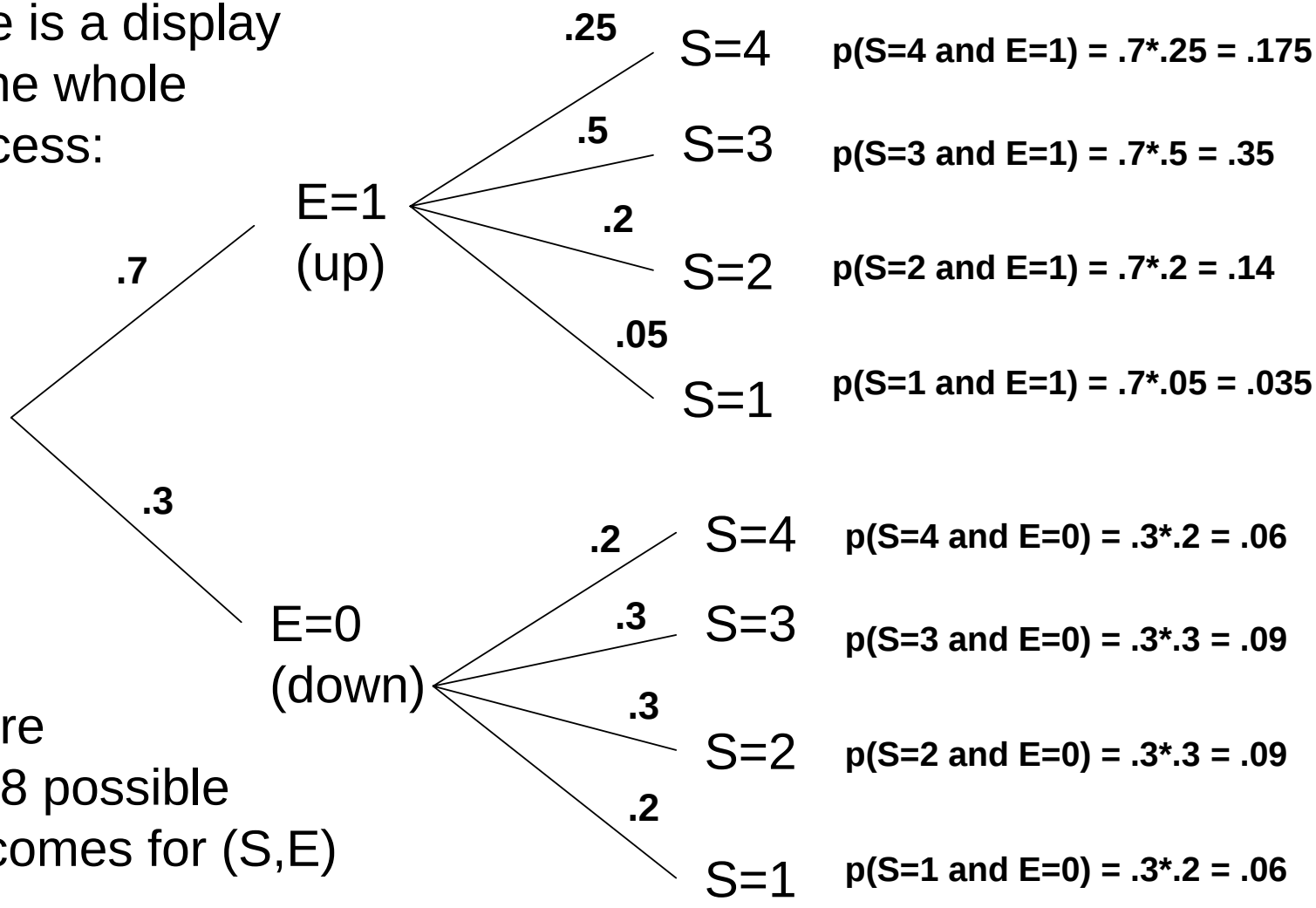
The conditional distributions tell us about about what can happen to S for a given value of E... but what about S and E jointly?

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

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 $X = x$

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 variable 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?

Trump's victory

Let's try to figure out why were people so confused on November 8th 2016...

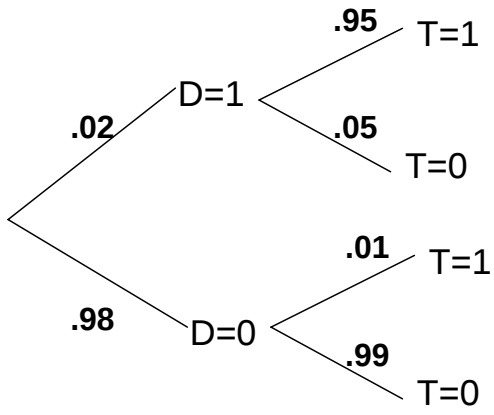
I am simplifying things a bit, but starting the day, Trump had to win 5 states to get the presidency: Florida, North Carolina, Pennsylvania, Michigan and Wisconsin. One could also say that each of these states had a 50-50 change for Trump and Hillary.

So, based on this information, what was the probability of a Trump victory? (Homework: make sure to revisit this at home.)

Disease Testing Example

Let $D = 1$ indicate you have a disease

Let $T = 1$ indicate that you test positive for it



$\Longrightarrow$

		D = 0	D = 1
T	0	.9702=.98*.99	.001
	1	.0098	.019

If you take the test and the result is positive, you are really interested in the question: Given that you tested positive, what is the chance you have the disease?

Disease Testing Example

		D	
		0	1
T	0	.9702	.001
	1	.0098	.019

$$pr(D = 1|T = 1) = \frac{0.019}{(0.019 + 0.0098)} = 0.66$$

Bayes Theorem (aside)

The computation of $pr(x|y)$ from $pr(x)$ and $pr(y|x)$ is called Bayes theorem...

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

In the disease testing example:

$$p(D=1|T=1) = \frac{p(T=1|D=1)p(D=1)}{p(T=1|D=1)p(D=1)+p(T=1|D=0)p(D=0)}$$

$$pr(D=1|T=1) = \frac{0.019}{(0.019+0.0098)} = 0.66$$

Bayes Theorem (aside)

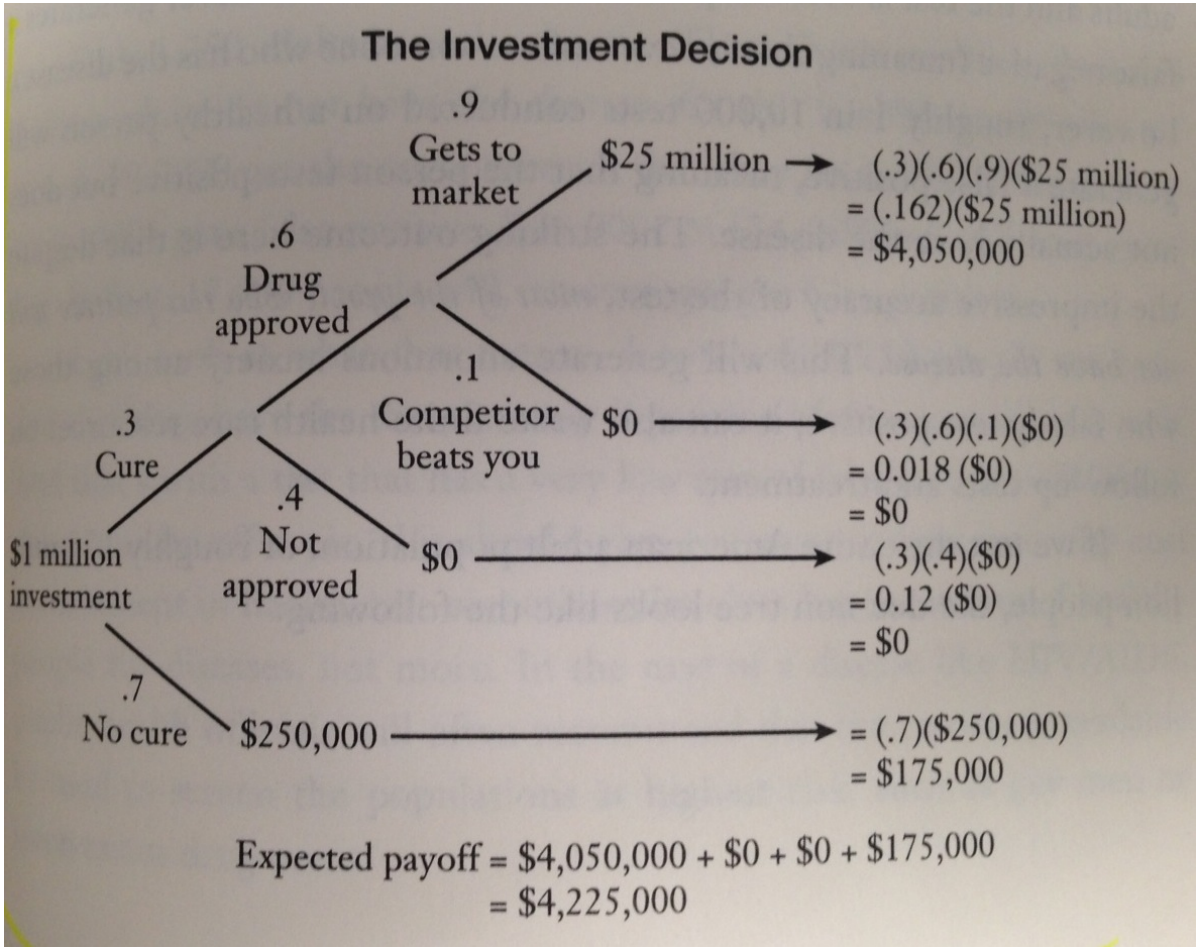
- Try to think about this intuitively... imagine you are about to test 100,000 people.
- we assume that about 2,000 of those have the disease.
- we also expect 1% of the disease-free people to test positive, ie, 980, and 95% of the sick people to test positive, ie 1,900. So, we expect a total of 2,880 positive tests.
- Choose one of the 2,880 people at random... what is the probability that he/she has the disease?

$$p(D = 1|T = 1) = 1,900/2,880 = 0.66$$

- isn't that the same?!

Probability and Decisions

Suppose you are presented with an investment opportunity in the development of a drug... probabilities are a vehicle to help us build scenarios and make decisions.



Probability and Decisions

We basically have a new random variable, i.e, our revenue, with the following probabilities...

Revenue	$P(Revenue)$
\$250,000	0.7
\$0	0.138
\$25,000,000	0.162

The expected revenue is then \$4,225,000...

So, should we invest or not?

Probability and Decisions

Let's get back to the drug investment example...

What if you could choose this investment instead?

Revenue	$P(Revenue)$
\$3,721,428	0.7
\$0	0.138
\$10,000,000	0.162

The expected revenue is still \$4,225,000...

What is the difference?

Mean and Variance of a Random Variable

The Mean or Expected Value is defined as (for a discrete X):

$$E(X) = \sum_{i=1}^{n} Pr(x_i) \times x_i$$

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 a Random Variable

Suppose

$$X = \begin{cases} 1 & \text{with prob.} \quad p \\ 0 & \text{with prob.} \quad 1-p \end{cases}$$

$$\begin{aligned} E(X) &= \sum_{i=1}^{n} Pr(x_i) \times x_i \\ &= 0 \times (1-p) + 1 \times p \\ E(X) &= p \end{aligned}$$

What is the $E(Sales)$ in our example above?

Didn't we see this in the drug investment problem?

Mean and Variance of a Random Variable

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

$$Var(X) = \sum_{i=1}^{n} Pr(x_i) \times [x_i - E(X)]^2$$

Weighted average of squared prediction errors... This is a measure of spread of a distribution. More risky distributions have larger variance.

Mean and Variance of a Random Variable

Suppose

$$X = \begin{cases} 1 & \text{with prob.} \quad p \\ 0 & \text{with prob.} \quad 1-p \end{cases}$$

$$\begin{aligned} Var(X) &= \sum_{i=1}^{n} Pr(x_i) \times [x_i - E(X)]^2 \\ &= (0-p)^2 \times (1-p) + (1-p)^2 \times p \\ &= p(1-p) \times [(1-p) + p] \\ Var(X) &= p(1-p) \end{aligned}$$

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

What is the $Var(Sales)$ in our example above?

How about the drug problem?

The Standard Deviation

- What are the units of $E(X)$? What are the units of $Var(X)$?
- A more intuitive way to understand the spread of a distribution is to look at the standard deviation:

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

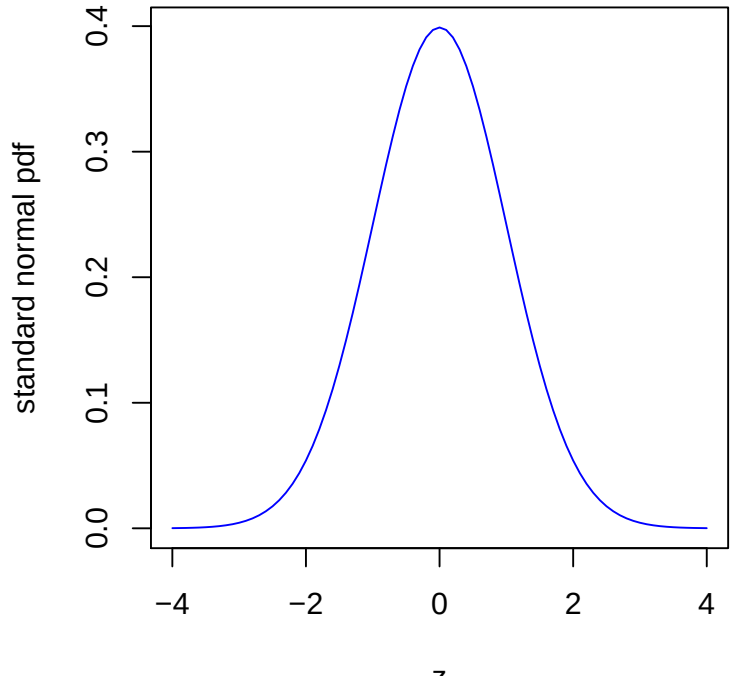
- What are the units of $sd(X)$?

Continuous Random Variables

- Suppose we are trying to predict tomorrow's return on the S&P500...
- Question: What is the random variable of interest?
- Question: How can we describe our uncertainty about tomorrow's outcome?
- Listing all possible values seems like a crazy task... we'll work with intervals instead.
- These are call continuous random variables.
- The probability of an interval is defined by the area under the probability density function.

The Normal Distribution

- A random variable is a number we are NOT sure about but we might have some idea of how to describe its potential outcomes. The Normal distribution is the most used probability distribution to describe a random variable
- The probability the number ends up in an interval is given by the area under the curve (pdf)

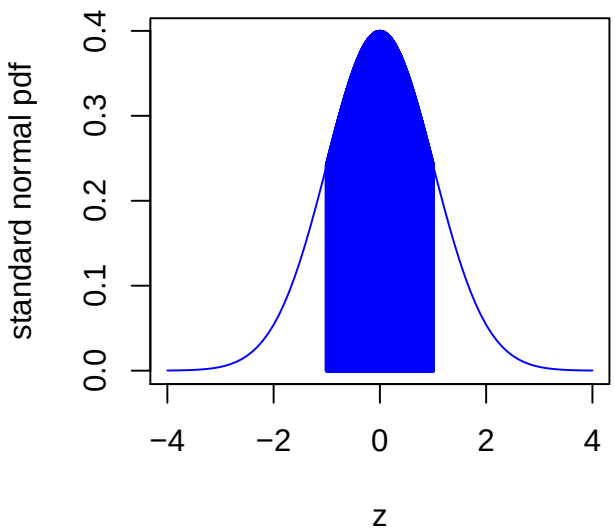


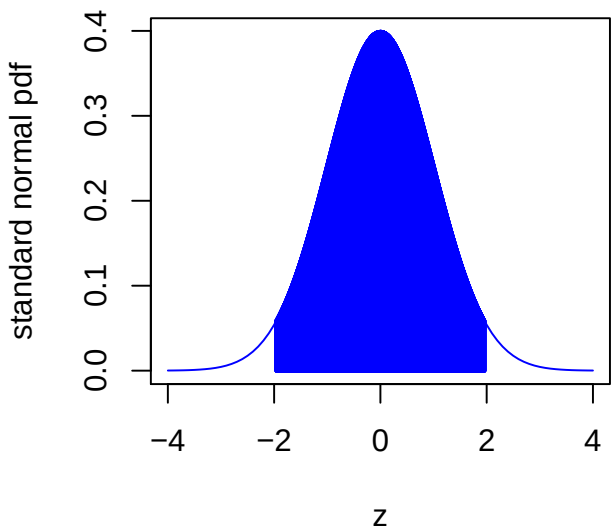
The Normal Distribution

- The standard Normal distribution has mean 0 and has variance 1.
- Notation: If $Z \sim N(0,1)$ (Z is the random variable)

$$Pr(-1 < Z < 1) = 0.68$$

$$Pr(-1.96 < Z < 1.96) = 0.95$$





The Normal Distribution

Note:

For simplicity we will often use $P(-2 < Z < 2) \approx 0.95$

Questions:

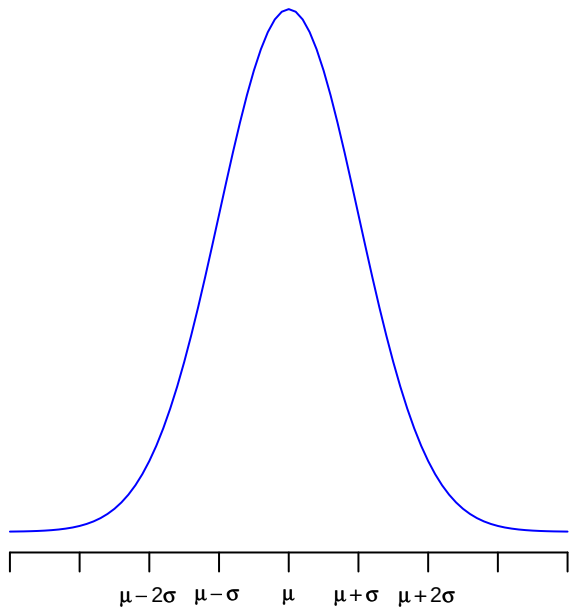
- What is $Pr(Z < 2)$? How about $Pr(Z \leq 2)$?
- What is $Pr(Z < 0)$?

The Normal Distribution

- The standard normal is not that useful by itself. When we say "the normal distribution", we really mean a family of distributions.
- We obtain pdfs in the normal family by shifting the bell curve around and spreading it out (or tightening it up).

The Normal Distribution

- We write $X \sim N(\mu, \sigma^2)$. "Normal distribution with mean μ and variance σ^2.
- The parameter μ determines where the curve is. The center of the curve is μ.
- The parameter σ determines how spread out the curve is. The area under the curve in the interval $(\mu - 2\sigma, \mu + 2\sigma)$ is 95%.
  $Pr(\mu - 2\,\sigma < X < \mu + 2\,\sigma) \approx 0.95$

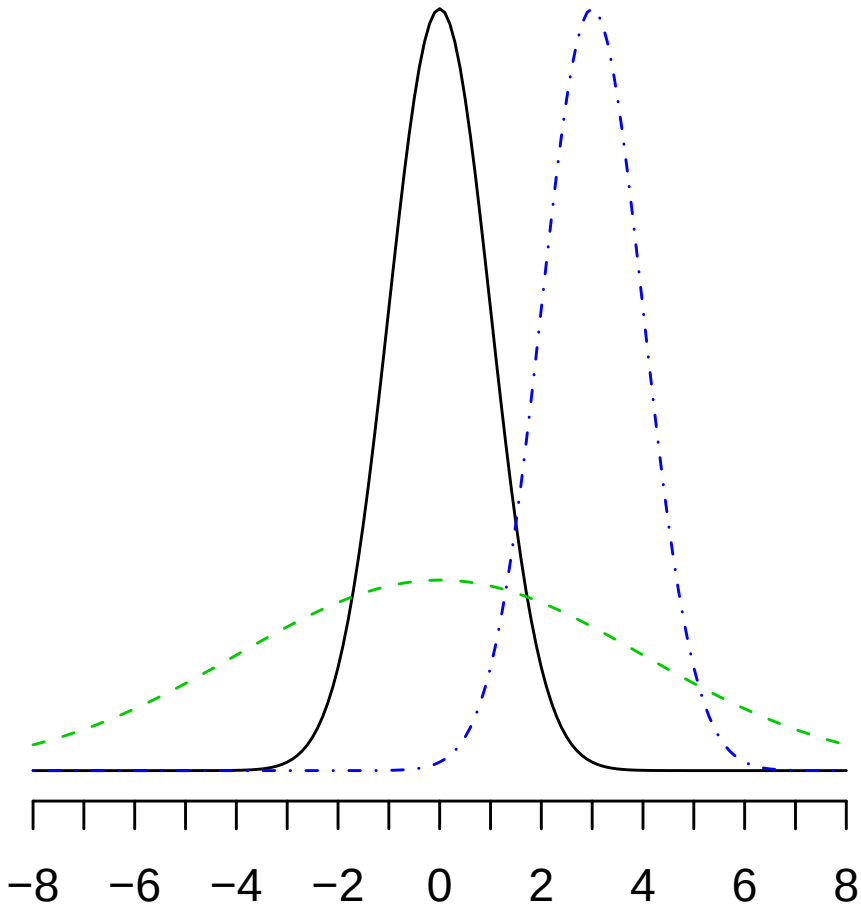


Mean and Variance of a Random Variable

- For the normal family of distributions we can see that the parameter μ talks about "where" the distribution is *located* or *centered*.
- We often use μ as our best guess for a *prediction*.
- The parameter σ talks about how *spread out* the distribution is. This gives us and indication about how *uncertain* or how *risky* our prediction is.
- If X is any random variable, the mean will be a measure of the location of the distribution and the variance will be a measure of how spread out it is.

The Normal Distribution

- Example: Below are the pdfs of $X_1 \sim N(0,1)$, $X_2 \sim N(3,1)$, and $X_3 \sim N(0,16)$.
- Which pdf goes with which X?

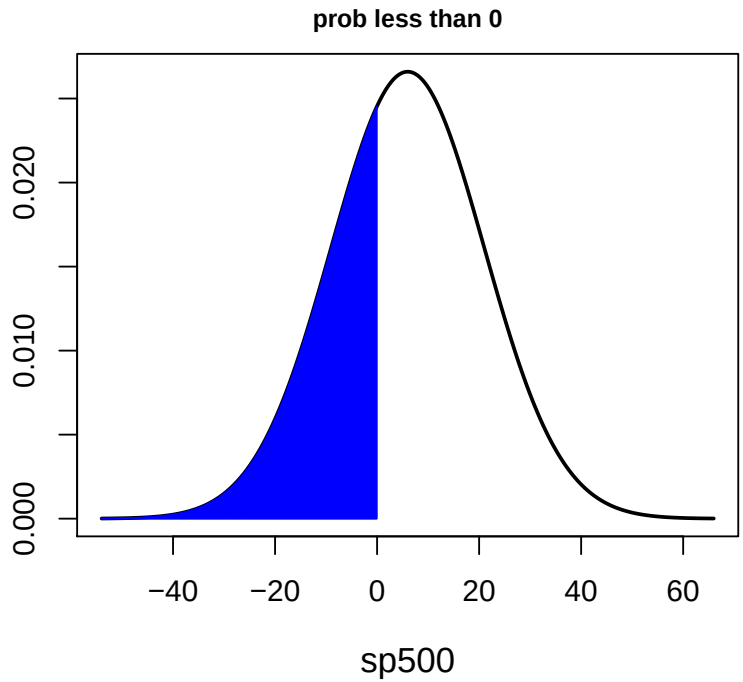


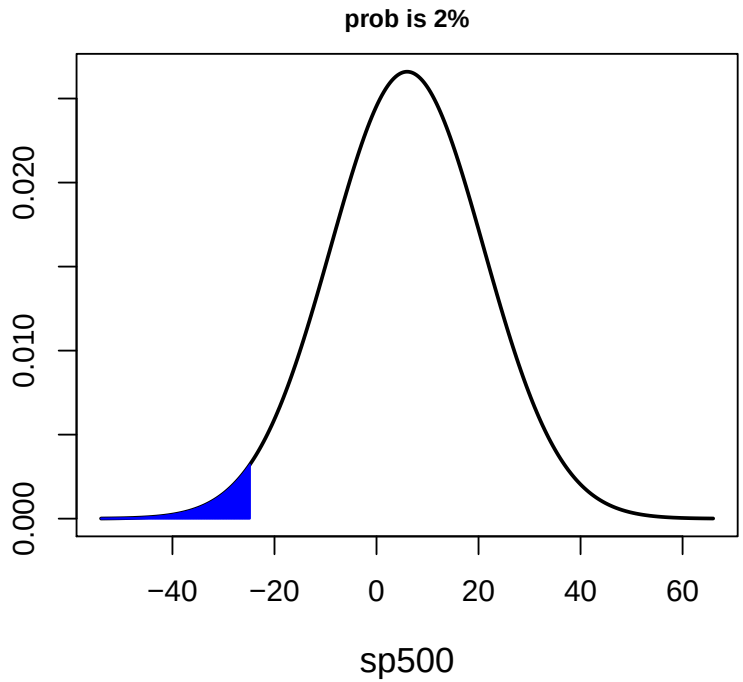
The Normal Distribution – Example

- Assume the annual returns on the SP500 are normally distributed with mean 6% and standard deviation 15%. SP500 $\sim N(6, 225)$. (Notice: $15^2 = 225$).
- Two questions: (i) What is the chance of losing money on a given year? (ii) What is the value that there's only a 2% chance of losing that or more?
- Lloyd Blankfein: *"I spend 98% of my time thinking about 2% probability events!"*

- (i) $Pr(SP500 < 0)$ and (ii) $Pr(SP500 <?) = 0.02$

The Normal Distribution – Example





- (i) $Pr(SP500 < 0) = 0.35$ and (ii) $Pr(SP500 < -25) = 0.02$
- In Excel: NORMDIST and NORMINV (homework!)

The Normal Distribution

1. Note: In

$$X \sim N(\mu, \sigma^2)$$

μ is the mean and σ^2 is the variance.

2. Standardization: if $X \sim N(\mu, \sigma^2)$ then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

3. Summary:

$X \sim N(\mu, \sigma^2)$:

μ: where the curve is

σ: how spread out the curve is

95% chance $X \in \mu \pm 2\sigma$.

The Normal Distribution – Another Example

Prior to the 1987 crash, monthly S&P500 returns (r) followed (approximately) a normal with mean 0.012 and standard deviation equal to 0.043. How extreme was the crash of -0.2176? The standardization helps us interpret these numbers...

$$r \sim N(0.012, 0.043^2)$$

$$z = \frac{r - 0.012}{0.043} \sim N(0, 1)$$

For the crash,

$$z = \frac{-0.2176 - 0.012}{0.043} = -5.27$$

How extreme is this $z$value? 5 standard deviations away!!

Regression to the Mean

- Imagine your performance on a task follows a standard normal distribution, i.e., $N(0, 1)$... Say you perform that task today and score 2.
- If you perform the same task tomorrow, what is the probability you are going to do worse? 97.5%, right?
- This is called regression to the mean!!

Make sure to read the article on this topic available in the class website...