Section 1: Introduction, Probability Concepts and Decisions

Tengyuan Liang, Chicago Booth

https://tyliang.github.io/BUS41000/

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

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

Section 6: Additional Topics (time permitting)

Course Schedule

Week 1 Section 1

Week 2 Section 1

Week 3 Section 2

Week 4 Section 2

Week 5 Section 3 and Review

Week 6 Midterm

Week 7 Section 3

Week 8 Section 4

Week 9 Section 5

Week 10 Section 5 and Review

Finals Week Final

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 2019?

Another question... (Target Marketing)

Suppose you are deciding whether or not to target a customer with a promotion(or an ad)...

It will cost you \$.80 (eighty cents) to run the promotion and a customer spends \$40 if they respond to the promotion.

Should you do it ???

Introduction

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

- How likely is Andrew Yang to win the election?
- 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 aggregating 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 $Pr(X \geq 1)$?

Pete Rose Hitting Streak

Pete Rose of the Cincinnati Reds set a National League record of hitting safely in 44 consecutive games. . .

- Rose was a 300 hitter.
- Assume he comes to bat 4 times each game.
- Each at bat is assumed to be independent, i.e., the current at bat doesn't affect the outcome of the next.

What probability might reasonably be associated with a hitting streak of that length?

Pete Rose Hitting Streak

Let A_i denote the event that "Rose hits safely in the i^{th} game"

Then $P(\text{Rose Hits Safely in 44 consecutive games}) =$
$P(A_1 \text{ and } A_2 \ldots \text{ and } A_{44}) = P(A_1)P(A_2)...P(A_{44})$

We now need to find $P(A_i)\ldots$ It is easier to think of the complement of A_i, i.e., $P(A_i) = 1 - P(\text{not } A_i)$

$$\begin{aligned} P(A_i) &= 1 - P(\text{Rose makes 4 outs}) \\ &= 1 - (0.7 \times 0.7 \times 0.7 \times 0.7) \\ &= 1 - (0.7)^4 = 0.76 \end{aligned}$$

So, for the winning streak we have $(0.76)^{44} = 0.0000057$!!! (Why?)
(btw, Joe DiMaggio's record is 56!!!!)

New England Patriots and Coin Tosses

For the past 25 games the Patriots won 19 coin tosses!

What is the probability of that happening?

Let T be a random variable taking the value 1 when the Patriots win the toss or 0 otherwise.
It's reasonable to assume $Pr(T = 1) = 0.5$, right??

Now what? It turns out that there are 177,100 different sequences of 25 games where the Patriots win 19... it turns out each potential sequence has probability 0.5^{25} (why?)

Therefore the probability for the Patriots to win 19 out 25 tosses is $177,100 \times 0.5^{25} = 0.005$

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

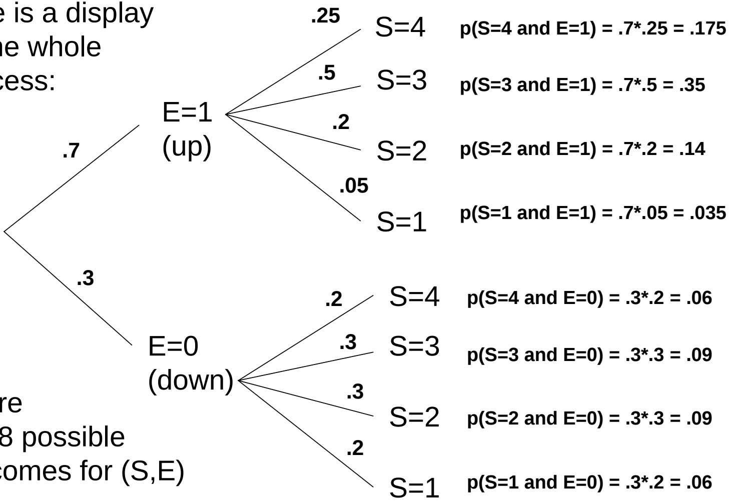
The conditional distributions tell us about what can happen to S for a given value of E... but what about S and E jointly?

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

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 $S = 4$?

S

E		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?

Republicans' victory

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

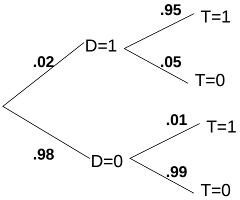
I am simplifying things a bit, but starting the day, Republicans had to win 5 states to get the presidency: Florida, North Carolina, Pennsylvania, Michigan and Wisconsin. One could also say that each of these states had a 50-50 change for Republicans and Democrats.

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

[FiveThirtyEight article on 2016 election, further reading]

Disease Testing Example

Let $D = 1$ indicate you have a disease
Let $T = 1$ indicate that you test positive for it



$\Longrightarrow$

		D	
		0	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?

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

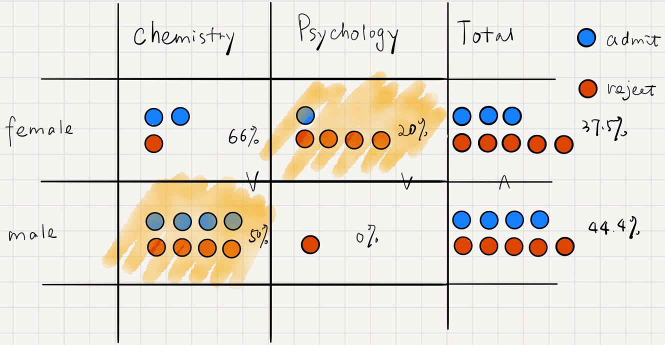
- isn't that the same?!

Prosecutor's fallacy

The following demonstrates the fallacy in the context of a prosecutor questioning an expert witness:

- "the odds of finding this evidence on an innocent man are so small that the jury can safely disregard the possibility that this defendant is innocent."
- What's wrong? Explain with conditional probability

The Berkeley gender bias case



The Berkeley gender bias case

Let $A =$"admitted to Berkeley". In 1973 it was noted that $P(A \mid \text{male}) = 0.44$ while $P(A \mid \text{female}) = 0.35$. Meanwhile, individual departments showed no signs of discrimination. Consider the chemistry department and the psychology department.

	Chemistry	Psychology
$P(A \mid \text{female})$	0.6	0.3
$P(A \mid \text{male})$	0.5	0.25

What is going on?

The Berkeley gender bias case

For Females:

$$P(A) = P(A \mid \text{chem})P(\text{chem}) + P(A \mid \text{psych})P(\text{psych})$$
$$0.35 = 0.6P(\text{chem}) + 0.3(1 - P(\text{chem}))$$
$$\text{(hence)} \quad P(\text{chem}) = 0.167$$

For Males:

$$P(A) = P(A \mid \text{chem})P(\text{chem}) + P(A \mid \text{psych})P(\text{psych})$$
$$0.44 = 0.5P(\text{chem}) + 0.25(1 - P(\text{chem}))$$
$$\text{(hence)} \quad P(\text{chem}) = 0.76$$

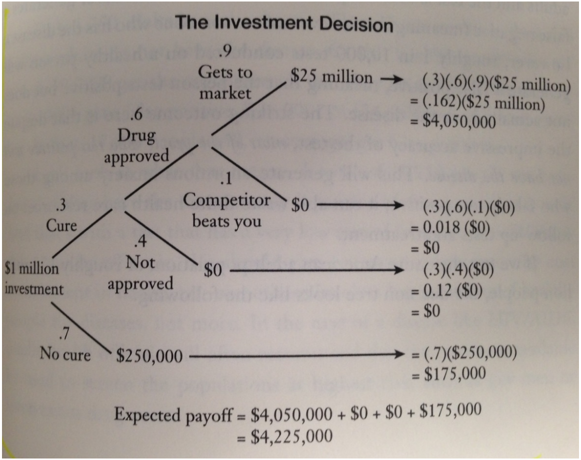
The Berkeley gender bias case

The explanation for the apparent overall bias was that women have a higher probability of applying to Psychology than to Chemistry (assuming for simplicity that these are the only two options) and overall Psychology has a lower admissions rate!

This is a cautionary tale! Before we can act on a apparent association between two variables (for example, sue Berkeley) we need to account for potential lurking variables that are the real cause of the relationship. We will talk a lot more about this... but keep in mind, association is NOT causation!

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?

Back to Target Marketing

Should we send the promotion ???

Well, it depends on how likely it is that the customer will respond!!

If they respond, you get 40-0.8=$39.20.

If they do not respond, you lose $0.80.

Let's assume your "predictive analytics" team has studied the conditional probability of customer responses given customer characteristics... (say, previous purchase behavior, demographics, etc)

Back to Target Marketing

Suppose that for a particular customer, the probability of a response is 0.05.

Revenue	*P*(*Revenue*)
$-0.8	0.95
$39.20	0.05

Should you do the promotion?

Homework question: How low can the probability of a response be so that it is still a good idea to send out the promotion?

Probability and Decisions

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

What if you could choose this investment instead?

Revenue	$P(\mathit{Revenue})$
\$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 $sd(X)$?

Covariance

- A measure of *dependence* between two random variables...
- It tells us how two unknown quantities tend to move together

The Covariance is defined as (for discrete X and Y):

$$Cov(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} Pr(x_i, y_j) \times [x_i - E(X)] \times [y_j - E(Y)]$$

- What are the units of $Cov(X,Y)$?
- What is the $Cov(Sales, Economy)$ in our example above?

Ford vs. Tesla

- Assume a very simple joint distribution of monthly returns for Ford (F) and Tesla (T):

	t=-7%	t=0%	t=7%	Pr(F=f)
f=-4%	0.06	0.07	0.02	**0.15**
f=0%	0.03	0.62	0.02	**0.67**
f=4%	0.00	0.11	0.07	**0.18**
Pr(T=t)	**0.09**	**0.80**	**0.11**	**1**

Let's summarize this table with some numbers...

Ford vs. Tesla

	t=-7%	t=0%	t=7%	Pr(F=f)
f=-4%	0.06	0.07	0.02	**0.15**
f=0%	0.03	0.62	0.02	**0.67**
f=4%	0.00	0.11	0.07	**0.18**
Pr(T=t)	**0.09**	**0.80**	**0.11**	**1**

- $E(F) = 0.12$, $E(T) = 0.14$
- $Var(F) = 5.25$, $sd(F) = 2.29$, $Var(T) = 9.76$, $sd(T) = 3.12$
- What is the better stock?

Ford vs. Tesla

	t=-7%	t=0%	t=7%	Pr(F=f)
f=-4%	0.06	0.07	0.02	**0.15**
f=0%	0.03	0.62	0.02	**0.67**
f=4%	0.00	0.11	0.07	**0.18**
Pr(T=t)	**0.09**	**0.80**	**0.11**	**1**

$$
\begin{aligned}
Cov(F,T) =& (-7-0.14)(-4-0.12)0.06 + (-7-0.14)(0-0.12)0.03 + \\
& (-7-0.14)(4-0.12)0.00 + (0-0.14)(-4-0.12)0.07 + \\
& (0-0.14)(0-0.12)0.62 + (0-0.14)(4-0.12)0.11 + \\
& (7-0.14)(-4-0.12)0.02 + (7-0.14)(0-0.12)0.02 + \\
& (7-0.14)(4-0.12)0.07 = 3.063
\end{aligned}
$$

Okay, the covariance in positive... makes sense, but can we get a more intuitive number?

Correlation

$$Corr(X, Y) = \frac{Cov(X, Y)}{sd(X)sd(Y)}$$

- What are the units of $Corr(X, Y)$? It doesn't depend on the units of X or Y!
- $-1 \leq Corr(X, Y) \leq 1$

In our Ford vs. Tesla example:

$$Corr(F, T) = \frac{3.063}{2.29 \times 3.12} = 0.428 \text{ (not too strong!)}$$

Linear Combination of Random Variables

Is it better to hold Ford or Tesla? How about half and half?

To answer this question we need to understand the behavior of the weighted sum (linear combinations) of two random variables...

Let X and Y be two random variables:

- $E(aX + bY) = aE(X) + bE(Y)$
- $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2ab \times Cov(X, Y)$

Linear Combination of Random Variables

Applying this to the Ford vs. Tesla example...

- $E(0.5F + 0.5T) = 0.5E(F) + 0.5E(T) = 0.5 \times 0.12 + 0.5 \times 0.14 = 0.13$
- $Var(0.5F + 0.5T) = (0.5)^2 Var(F) + (0.5)^2 Var(T) + 2(0.5)(0.5) \times Cov(F, T) = (0.5)^2(5.25) + (0.5)^2(9.76) + 2(0.5)(0.5) \times 3.063 = 5.28$
- $sd(0.5F + 0.5T) = 2.297$

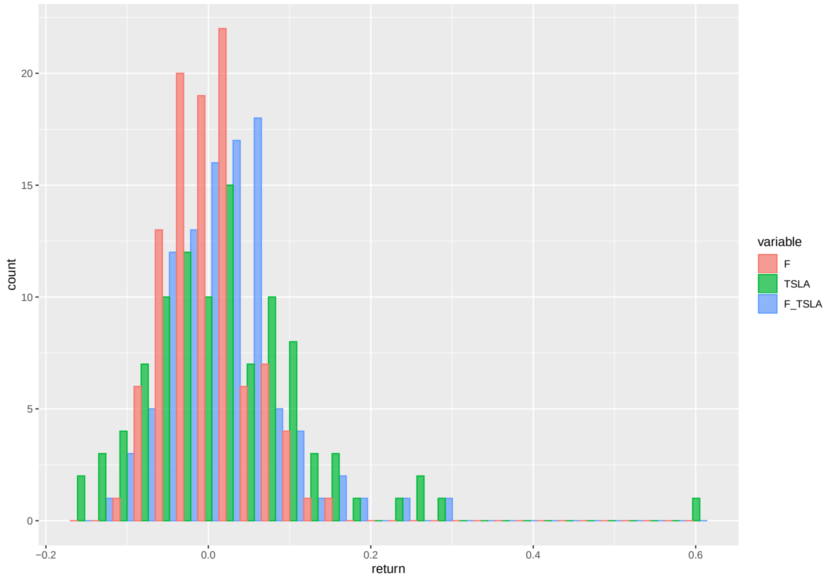
so, what is better? Holding Ford, Tesla or the combination?

Linear Combination of Random Variables

More generally...

- $E(w_1 X_1 + w_2 X_2 + ... w_p X_p) = w_1 E(X_1) + w_2 E(X_2) + ... + w_p E(X_p) = \sum_{i=1}^{p} w_i E(X_i)$
- $Var(w_1 X_1 + w_2 X_2 + ... w_p X_p) = w_1^2 Var(X_1) + w_2^2 Var(X_2) + ... + w_p^2 Var(X_p) + 2 w_1 w_2 \times Cov(X_1, X_2) + 2 w_1 w_3 Cov(X_1, X_3) + ... = \sum_{i=1}^{p} w_i^2 Var(X_i) + \sum_{i=1}^{p} \sum_{j \neq i} w_i w_j Cov(X_i, X_j)$

Real Stock Data: Ford and Tesla



Real Stock Data: Ford and Tesla

Portfolio vs. Single Project (from Misbehaving)

In a meeting with 23 executives plus the CEO of a major company economist Richard Thaler poses the following question:

Suppose you were offered an investment opportunity for your division (each executive headed a separate/independent division) that will yield one of two payoffs. After the investment is made, there is a 50% chance it will make a profit of \$2 million, and a 50% chance it will lose \$1 million. Thaler then asked by a show of hands who of the executives would take on this project. Of the twenty-three executives, only three said they would do it.

Anything wrong with that?

Portfolio vs. Single Project (from Misbehaving)

Then Thaler asked the CEO a question. If these projects were independent, that is, the success of one was unrelated to the success of another, how many of the projects would he want to undertake? His answer: all of them! By taking on twenty three projects, the firm expects to make \$11.5 million (since each of them is worth an expected half million), and a bit of mathematics reveals that the chance of losing any money overall is less than 10%. (Homework: why? trick one... dont worry too much about it...)

Homework: compare the "sharpe ratio" of a single project vs. taking on all the projects

Portfolio vs. Single Project (from Misbehaving)

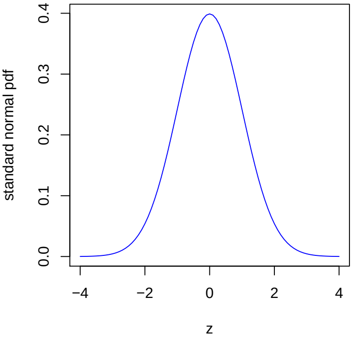
Companies, CEO's, managers have to be careful in setting incentives that avoid what psychologist and behavior economists call "narrow framing"... otherwise, what can be perceived to be bad for one manager may be very good for the entire company!

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)

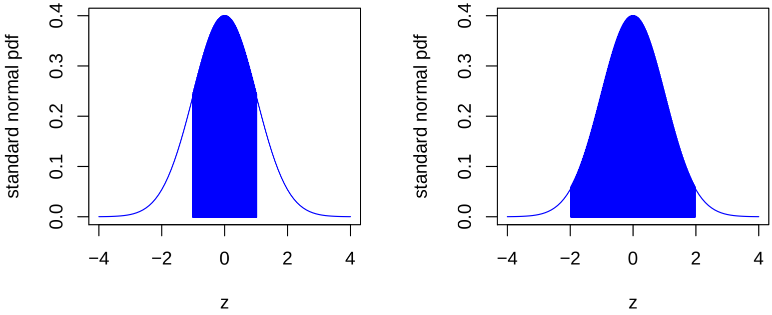


The Normal Distribution

- 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:

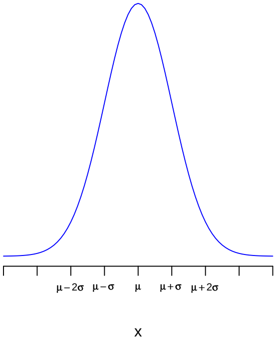
- What is $Pr(Z < 2)$? How about $Pr(Z \leq 2)$?
- What is $Pr(Z < 0)$?

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%. $Pr(\mu - 2\,\sigma < X < \mu + 2\,\sigma) \approx 0.95$

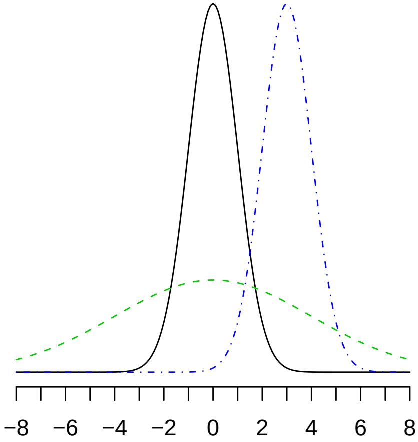


Mean and Variance of a Random Variable

- 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

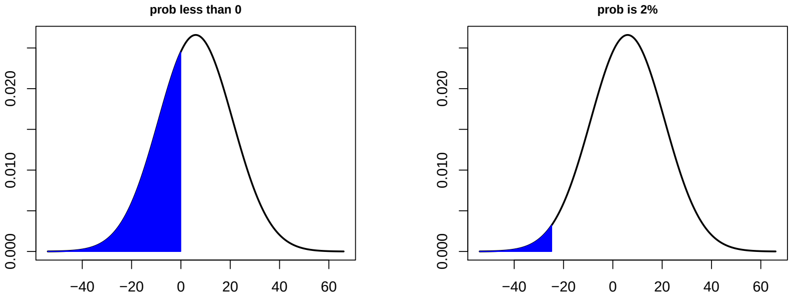
- Example: Below are the pdfs of $X_1 \sim N(0, 1)$, $X_2 \sim N(3, 1)$, and $X_3 \sim N(0, 16)$.
- Which pdf goes with which X?



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




```r
pnorm(0, mean = 6, sd = 15)
```

```
## [1] 0.3445783
```

```r
qnorm(0.02, mean = 6, sd = 15)
```

```
## [1] -24.80623
```

- (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!!

The Normal Distribution – Approximating repeated trials

Let's revisit the Patriots coin toss example... the $E(wins) = 0.5 \times 25$ and the $Var(wins) = 0.25 \times 25$, right?

We can now approximate the number of wins via a $N(12.5, 6.25)$... therefore the $Pr(wins = 19) \approx Pr(wins \in [18.5, 19.5]) \approx 0.00543$ (why?, and compare to $\binom{25}{19} 0.5^{25} = 0.00528$)

```
dnorm(19, mean=12.5, sd=2.5)
```

```
## [1] 0.005433188
```

```
dbinom(19, size = 25, prob = 0.5)
```

```
## [1] 0.005277991
```

The Normal Distribution – Approximating repeated trials

We can do the same for the "Narrow Framing" example (Portfolios of projects vs. single project)... There, if we are taking on all 23 projects, $E(Profits) = 0.5 \times 23$ and the $Var(Profits) = 2.25 \times 23$... (correct?)

We can now approximate the distribution of Profits via a $N(11.5, 7.19^2)$... therefore the $Pr(Profits > 0) \approx 0.94$ (why?)

```
pnorm(0, mean=11.5, sd=sqrt(23*2.25), lower.tail=FALSE)
```

```
## [1] 0.9450464
```

In summary, in many situations, if you can figure out the **mean and variance** of the random variable of interest, you can use a **normal distribution** to **approximate** the calculation of probabilities.

Portfolios, once again...

- As before, let's assume that the annual returns on the SP500 are normally distributed with mean 6% and standard deviation of 15%, i.e., $SP500 \sim N(6, 15^2)$
- Let's also assume that annual returns on bonds are normally distributed with mean 2% and standard deviation 5%, i.e., $Bonds \sim N(2, 5^2)$
- What is the best investment?
- What else do I need to know if I want to consider a portfolio of SP500 and bonds?

Portfolios once again...

- Additionally, let's assume the correlation between the returns on SP500 and the returns on bonds is -0.2.
- How does this information impact our evaluation of the best available investment?

Recall that for two random variables X and Y:

- $E(aX + bY) = aE(X) + bE(Y)$
- $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2ab \times Cov(X, Y)$

- One more very useful property... sum of normal random variables is a new normal random variable!

Portfolios once again...

- What is the behavior of the returns of a portfolio with 70% in the SP500 and 30% in Bonds?
- $E(0.7SP500 + 0.3Bonds) = 0.7E(SP500) + 0.3E(Bonds) = 0.7 \times 6 + 0.3 \times 2 = 4.8$
- $Var(0.7SP500 + 0.3Bonds) = (0.7)^2 Var(SP500) + (0.3)^2 Var(Bonds) + 2(0.7)(0.3) \times Corr(SP500, Bonds) \times sd(SP500) \times sd(Bonds) = (0.7)^2(15^2) + (0.3)^2(5^2) + 2(0.7)(0.3) \times -0.2 \times 15 \times 5 = 106.2$
- $Portfolio \sim N(4.8, 10.3^2)$
- Homework: good or bad? What now? Is there a better combination?

Median, Skewness

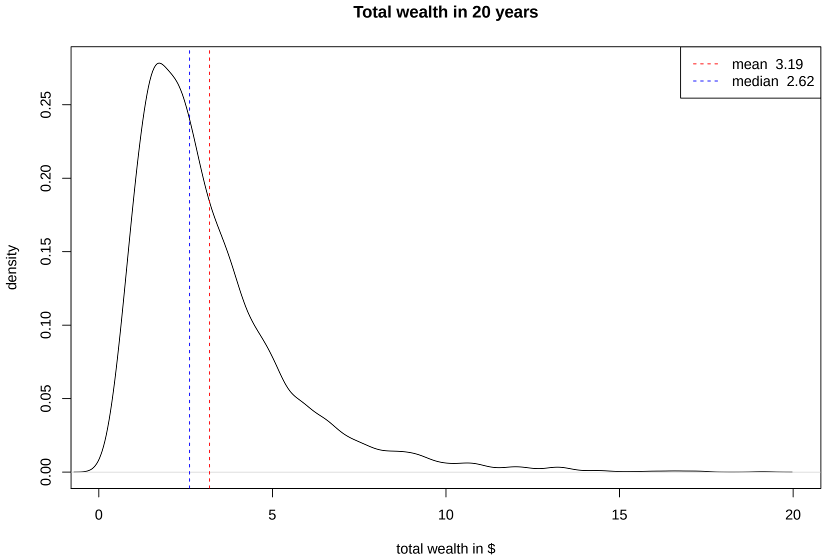
- The median of a random variable X is the point such that there is 50% chance X is above it, and hence a 50% chance X is below it.
- For symmetric distributions, the expected value (mean) and the median are always the same... look at all of our normal distribution examples.
- But sometimes, distributions are skewed, i.e., not symmetric. In those cases the median becomes another helpful summary!

Median, Skewness

- Let's think of an example... imagine you invest \$1 in the SP500 today and want to know how much money you are going to have in 20 years. We can assume, once again, that the returns on the SP500 on a given year follow $N(6, 15^2)$
- Let's also assume returns are independent year after year...
- Are my total returns just the sum of returns over 20 years? Not quite... compounding gets in the way.

Let's simulate potential "futures"

Median, Skewness



Median, Skewness: R Code

```
# Generate 5000 worlds, each simulate 20 years
returns = matrix(rnorm(n = 5000*20, mean = 6, sd = 15),
                 nrow = 5000, ncol=20)/100
total_wealth = apply(1+returns, 1, prod)
```

```
# Plotting
d = density(total_wealth)
plot(d, xlab="total wealth in $", ylab = "density",
     main = "Total wealth in 20 years", xlim = c(0,20))
abline(v = mean(total_wealth), col = 'red', lty=2)
abline(v = median(total_wealth), col = 'blue', lty=2)
legend("topright",
   legend = c(paste("mean ", round(mean(total_wealth),2)),
         paste("median ", round(median(total_wealth),2))),
        col = c('red', 'blue'), lty = c(2,2))
```