Section 1: Introduction, Probability Concepts and Decisions

Carlos M. Carvalho and Robert McCulloch

Readings:
OpenIntro Statistics, Chapters 2 and 3

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: Topics in Regression

Section 6: Naive Bayes

Let's start with a question...

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

Another question... (Target Marketing)

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

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 about things we are unsure about.

- 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?
- If I give a patient a certain drug, how long are they likely to live?
- Will the Leafs win the Stanley Cup any time soon?

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 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 coin tosses, we can describe its behavior through the following probability distribution:

x	$P(X = x)$
0	.25
1	.5
2	.25

x: a possible outcome

$P(X = x)$: the probability X turns out to be x.

Probability Distribution

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

The Bernoulli Distribution

A very common situation is that we are wondering whether something will happen or not.

Heads or tails, respond or don't respond, ...

It turns out to be very convenient to code up one possibility as a 0, and the other possibility as a 1.

The gives us the *Bernoulli distribution*.

$X \sim$ Bernoulli(p) means:

x	$P(X = x)$
0	1-p
1	p

Example:

I am about to toss a single coin. X is the random variable which is 1 if the coin turns out to be heads and 0, if it is tails.

$$X \sim \text{Bernoulli}(.5)$$

Example:

L is the random variable which is 1 if the Leafs win the Stanley Cup and 0, if not.

$$L \sim \text{Bernoulli}(?????)$$

Conditional, Joint and Marginal Distributions

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

Let's suppose you are thinking about your sales next quarter.

Let S denote your sales (in thousands of units sold).

S is a number you are not sure about !!

In thinking about what S, you find you are thinking about what will happen next quarter for the overall economy.

We need to think about two things we are uncertain about, the economy and sales!

Let E denote the performance of the economy next quarter.

Let $E = 1$ 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*, they describe our beliefs about S conditional on knowing what happens for E.

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

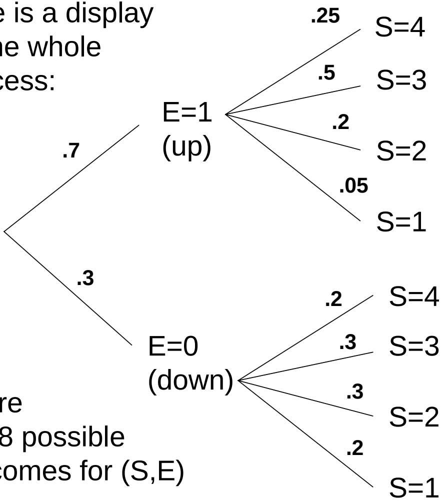
- In blue is the conditional distribution of S given $E = 1$
- In red is the conditional distribution of S given $E = 0$
- We read: *the probability of Sales of 4 ($S = 4$)* **given(or conditional on)** *the economy is growing ($E = 1$) is 0.25*

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%

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)

We can specify the distribution of the pair of random variables (S, E) by listing all possible pairs and the corresponding probability.

(s, e)	$p(S = s, E = e)$
(1,1)	.035
(2,1)	.14
(3,1)	.35
(4,1)	.175
(1,0)	.06
(2,0)	.09
(3,0)	.09
(4,0)	.06

Question: What is $P(S = 1)$?

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 $X = x$

Warning:

The notation can get tricky.

Sometimes rather than writing

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

someone might write just,

$$p(x, y)$$

for the same thing!!

Usually, *but not always*, capitals are used for random variables and small case is used for possible values.

Conditional, Joint and Marginal Distributions and Two-way Tables

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

S

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

Conditionals from Joints

We derived the joint distribution of (E, S) from the marginal for E and the conditional $S \mid E$.

You can also calculate the conditional from the joint by doing it the other way

$$P(Y = y, X = x) = P(X = x)\, P(Y = y \mid X = x)$$

$\Rightarrow$

$$P(Y = y \mid X = x) = \frac{P(Y = y, X = x)}{P(X = x)}$$

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

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

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 variable 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?

Example:

You are about to toss two coins.
Let X_1 be 1 if the first coin is a head and 0 if tails.
Let X_2 be 1 if the second coin is a head and 0 if tails.

$$X_1 \sim \text{Bernoulli}(.5), \quad X_2 \sim \text{Bernoulli}(.5)$$

What is the probability of getting two heads in a row?

$$\begin{aligned} P(X_1 = 1, X_2 = 1) &= P(X_1 = 1)\, P(X_2 = 1 \mid X_1 = 1) \\ &= P(X_1 = 1)\, P(X_2 = 1) \\ &= (.5) \times (.5) \\ &= .25 \end{aligned}$$

IID

Our two coins X_1 and X_2 both have the same distribution and they are independent.

We say they are IID:

- I: independent
- ID: identically distributed

This terminology gets used a lot in statistics.

Example:

We say the two coins are IID Bernoulli with $p = .5$.

Suppose I am about to toss two dice.

Y_1 is the number on the face of the first die.

Y_2 is the number on the face of the second die.

Are Y_1, Y_2 IID?

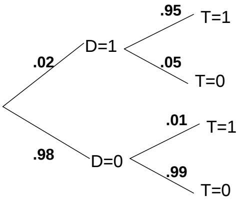
Are Y_1, Y_2 IID Bernoulli?

Bayes Theorem

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?

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

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

Note:

In this example the *sensitivity* is .95.

The probability of a true postitive.

In this example the *specificity* is .99.

The probability of a true negative.

Bayes Theorem:

In the disease testing problem we formulated our understand of the variable T and D using

$$p(t, d) = p(d)p(t|d).$$

Then we use probability theory to compute the quantity we really want which is

$$p(d \mid t).$$

This process of getting the probability "the other way" from how the modeling describes things is called *Bayes Theorem*.

We can develop of more formal statement of Bayes Theorem by writing things out using our basic properties of probability.

Suppose we have $p(y)$ and $p(x \mid y)$.

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

For binary y (y is 0 or 1, as in our Disease testing problem), we have:

$$p(Y=1|x) = \frac{p(Y=1)\, p(x|Y=1)}{p(Y=0)\, p(x|Y=0) + p(Y=1)\, p(x|Y=1)}$$

$$p(Y=1|x) = \frac{p(Y=1)\,p(x|Y=1)}{p(Y=0)\,p(x|Y=0) + p(Y=1)\,p(x|Y=1)}$$

In the disease testing example Y is D and X is T:

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

$$p(D=1|T=1) = \frac{.95*.02}{.95*.02+.01*.98} = \frac{0.019}{(0.019+0.0098)} = 0.66$$

More Than Two Random Variables

Of course, we may want to think about more than two uncertain quantities at a time!!

Our ideas extend nicely to any number of variables.

For example with three random variables X_1, X_2, and X_3 we might want to think about:

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

The probability that X_1 turns out to be x_1 and X_2 turns out to be x_2 and X_3 turns out to be x_3.

We can immediately extend our basic inutitive ideas:

$$P(X_1 = x1, X_2 = x_2, X_3 = x_3) =$$

$$P(X_1 = x_1)\, P(X_2 = x_2 \mid X_1 = x_1)\, P(X_3 = x_3 \mid X_1 = x_1, X_2 = x_2).$$

Example:

Suppose we have 10 voters.

4 are republican and 6 are democratic.

We "randomly" choose 3.

Let Y_i be 1 if the i^{th} voter is a democrat and 0 otherwise, $i = 1, 2, 3$.

What is

$$P(Y_1 = 1, Y_2 = 1, Y_3 = 1)$$

What is the probability of getting three democrats in a row ??

$$P(Y_1 = 1, Y_2 = 1, Y_3 = 1) =$$

$$P(Y_1 = 1)\, p(Y_2 = 1 \mid Y_1 = 1)\, P(Y_3 = 1 \mid Y_1 = 1, Y_2 = 1)$$

$$= (6/10)(5/9)(4/8)$$

$$= (1/6) = .167.$$

When we randomly pick a person from a population of people, and then randomly pick a second from the ones left, and so on, we call it *sampling without replacement*.

If we put the person back each time and randomly choose from the whole group each time, then we call it *sampling with replacement*.

Random Sampling in R

```
> set.seed(99)
> sample(1:10,5)
[1]  6  2 10  7  4
> sample(1:10,5)
[1] 10  7  3  8  2
> set.seed(99)
> sample(1:10,5)
[1]  6  2 10  7  4
> sample(1:10000,20)
 [1] 9667 6714 2946 3583 1753 5486 5052 1938 6364 6872 6396 3575 1025  977 1827
[16] 2276  804 8203 5901 7720
> sample(1:10,5,replace=TRUE)
[1] 4 1 9 1 3
```

Example:

Suppose we are tossing 100 coins.

Let X_i be 1 if the ith coin is a head and 0 otherwise.

What is the probability of 100 heads in a row?

$P(X_1 = 1, X_2 = 1, \ldots, X_{100} = 1) =$

$$P(X_1 = 1)\, P(X_2 = 1 \mid X_1 = 1) \ldots P(X_j = 1 \mid X_1 = 1, X_2 = 1, \ldots, X_{j-1} = 1)$$
$$\ldots P(X_{100} = 1 \mid X_1 = 1, \ldots, X_{99} = 1)$$

$= .5^{100} = 7.888609e - 31.$

The 100 X_i are IID Bernoulli, with $p = .5$.

Question:

Suppose I get 100 heads in a row.

What is the probability the next one is a head?

Example:

Suppose I toss 100 dice.

Let Y_i be number on the ith die.

Are the Y_i IID?

Are the Y_i IID Bernoulli?

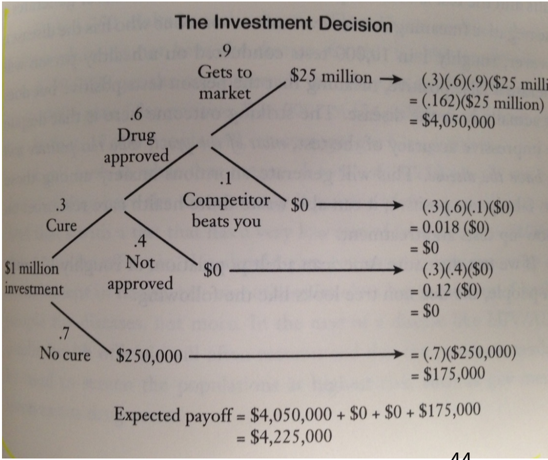
Probability and Decisions

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.

You make a 1 million investment to develop the drug.

If "no cure" (drug does not work) you get 250,000 back you don't spend.

If you do find a cure you have to worry about whether it is approved and whether a competitor beats you out.



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

Note:

$.3*.6*.9 = 0.162$

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)

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

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

Should you do the promotion?

.95*(-.8) + .05*39.20 = 1.2.

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

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

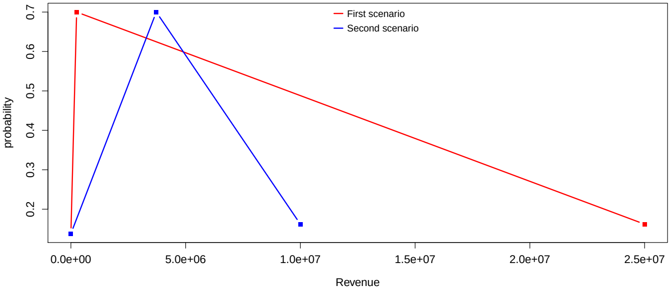
What if you could choose this investment instead?

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?

Here is a plot of the two distribution for the two drug discovery scenarios.



Mean and Variance of a Random Variable

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

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

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!

Suppose

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

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

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

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

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

Suppose

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

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

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

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 $sd(X)$?

Mean and Variance for Drug Development Example

Previously we computed the expected value for the drug development examples.

Let's review those calculations and compute the variances as well.

Same mean,

Different standard deviations!!

```
> pv = c(.7,.138,.162)
>
> d1 = c(250000,0,25000000)
> d2 = c(3721428,0, 10000000)
>
> cat("E1:",sum(pv*d1),"\n")
E1: 4225000
> cat("E2:",sum(pv*d2),"\n")
E2: 4225000
>
> M = sum(pv*d1)
>
> v1 = sum(pv*(d1-M)^2)
> v2 = sum(pv*(d2-M)^2)
> s1 = sqrt(v1)
> s2 = sqrt(v2)
>
> cat("s1,s2: ",s1,", ",s2,"\n")
s1,s2:  9134721 ,  2836141
```

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):

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

- What are the units of $Cov(X,Y)$?

Example:

$\mu_X = .1, \quad \mu_Y = .1.$

$\sigma_X = .05, \;\; \sigma_Y = .05.$

		X	
		.05	.15
Y	.05	.4	.1
	.15	.1	.4

x	y	prob	x-E(X)	y-E(Y)	prod
0.05	0.05	0.4	-0.05	-0.05	0.0025
0.15	0.05	0.1	0.05	-0.05	-0.0025
0.05	0.15	0.1	-0.05	0.05	-0.0025
0.15	0.15	0.4	0.05	0.05	0.0025

$$\text{Cov}(X, Y) = \sigma_{XY}$$
$$= .4 * .0025 + .1 * (-.0025) + .1 * (-.0025) + .4 * .0025 = .0015.$$

Intuition: There is an 80% chance X and Y move *in the same direction*.

Example:

$\mu_X = .1, \quad \mu_Y = .1.$

$\sigma_X = .05, \;\; \sigma_Y = .05.$

		X	
		.05	.15
Y	.05	.1	.4
	.15	.4	.1

x	y	prob	x-E(X)	y-E(Y)	prod
0.05	0.05	0.1	-0.05	-0.05	0.0025
0.15	0.05	0.4	0.05	-0.05	-0.0025
0.05	0.15	0.4	-0.05	0.05	-0.0025
0.15	0.15	0.1	0.05	0.05	0.0025

$\text{Cov}(X, Y) = \sigma_{XY}$

$= .1 * .0025 + .4 * (-.0025) + .4 * (-.0025) + .1 * .0025 = -.0015.$

Intuition: There is an 80% chance X and Y move *in opposite directions*.

Ford vs. Tesla

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

	t=-7%	t=0%	t=7%	P(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**
P(T=t)	**0.09**	**0.80**	**0.11**	**1**

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

	t=-7%	t=0%	t=7%	P(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**
P(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?

	t=-7%	t=0%	t=7%	P(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**
P(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 first example:

Corr(X,Y) = .0015/(.05*.05) = 0.6

In our second example:

Corr(X,Y) = -.0015/(.05*.05) = -0.6

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?

What do we mean by "half and half"?

We mean the *portfolio* where we put half of our money into Ford and half into Tesla.

Return On a Portfolio:

Suppose we form a portfolio in which we put fraction w_1 of our wealth into an asset with return R_1 and fraction w_2 of our wealth into an asset with return R_2. Let P be the return on the portfolio.

$$P = w_1\,R_1 + w_2\,R_2$$

In our Ford/Tesla use use $R_1 =$ Ford, $R_2 =$ Tesla, and $w_1 = w_2 = .5$.

Since the return on Ford and Tesla are random variables, so of course is the return on the portfolio!

Here is the joint distribution of $(R_1, R_2) = (F, T)$ *and* $P = .5\,F + .5\,T$.

	Ford	Tesla	P	prob
1	-4	-7	-5.5	0.06
2	0	-7	-3.5	0.03
3	4	-7	-1.5	0.00
4	-4	0	-2.0	0.07
5	0	0	0.0	0.62
6	4	0	2.0	0.11
7	-4	7	1.5	0.02
8	0	7	3.5	0.02
9	4	7	5.5	0.07

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

We can compare the random returns based on the means and variances:
big mean: good, big variance: bad.

We could could compute the mean and variance of P directly from its distribution, but there are some very handy formulas for the mean and variance of a *linear combination* of random variables.

Let X and Y be two random variables, a, b, and c are **known** constants:

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

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

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

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

```
asset: Ford, Tesla, Portfolio
mean: .12, .14, .13
sd:  2.29, 3.12, 2.297
```

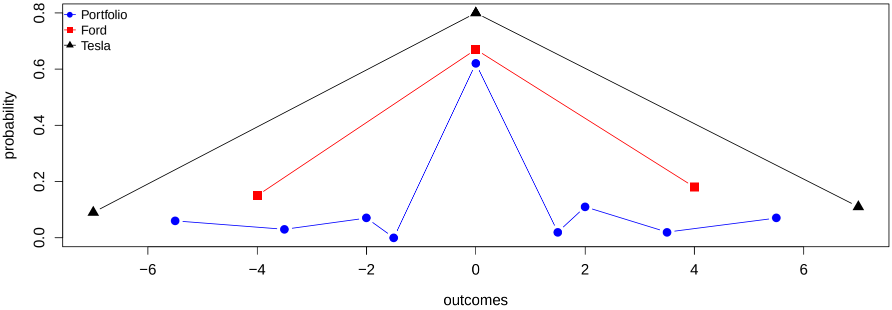
Let's check the mean and variance of P from our basic formulas by computing them in R.

```
> head(ddf)
  Ford Tesla    P prob
1   -4    -7 -5.5 0.06
2    0    -7 -3.5 0.03
3    4    -7 -1.5 0.00
4   -4     0 -2.0 0.07
5    0     0  0.0 0.62
6    4     0  2.0 0.11
> EP = sum(ddf$prob * ddf$P)
> cat("Expected value of port: ",EP,"\n")
Expected value of port:  0.13
> VP = sum(ddf$prob * (ddf$P-EP)^2)
> cat("Variance of port: ",VP,"\n")
Variance of port:  5.2931
```

Same numbers!!

Let's see what is going on graphically!

- plot the three distributions for Ford, Tesla, and the portfolio
- possible values on the x axis, probabilities on the y axis
- easy to see that Tesla has a higher variance than Ford.
- not so easy to see the difference in the means, this is realistic
- you can see you the diversification killed the tails



Note:

If X and Y are independent, then $Cov(X, Y) = 0$.

Covariance measures linear dependence.

If they have nothing to do with each other (independence), then they certainly have nothing to do with each other linearly.

More generally...

- $E(w_0 + w_1X_1 + w_2X_2 + ...w_pX_p) = w_0 + w_1E(X_1) + w_2E(X_2) + ... + w_pE(X_p) = w_0 + \sum_{i=1}^{p} w_iE(X_i)$

- $Var(w_0 + w_1X_1 + w_2X_2 + ...w_pX_p) = w_1^2Var(X_1) + w_2^2Var(X_2) + ... + w_p^2Var(X_p) + 2w_1w_2 \times Cov(X_1, X_2) + 2w_1w_3Cov(X_1, X_3) + ... = \sum_{i=1}^{p} w_i^2Var(X_i) + \sum_{i=1}^{p}\sum_{j\neq i} w_iw_jCov(X_i, X_j)$

Example:

Ford, Tesla, GM.

$$P = w_1 F + w_2 T + w_3 G$$

$$E(P) = w_1 E(F) + w_2 E(T) + w_3 E(G)$$

$$\begin{aligned} Var(P) = {} & w_1^2 Var(F) + w_2^2 Var(T) + w_3^2 Var(G) \\ & + 2w_1 w_2 Cov(F, T) + 2w_1 w_3 Cov(F, G) + 2w_2 w_3 Cov(T, G). \end{aligned}$$

With lots of assests this gets complicated!! There many possible pairs of assests and corresponding covariance pairs representing the high dimensional dependence of the many input assets.

In practice you have to estimate all the covariances from data, another good reason to index!!

Example, Sum and Mean of IID

Suppose you play a game n times and the winning from the ith play is represented by the random variable X_i, $i = 1, 2, \ldots, n$.

We assume the each play of the game is indepedent of the others and it is the same game each time.

So, the X_i are IID.

What are the mean and variance of the *total* winnings?

$$T = X_1 + X_2 + X_3 + \ldots + X_n.$$

Let $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$.

$$T = X_1 + X_2 + X_3 + \ldots + X_n.$$

The Mean:

$$\begin{aligned} E(T) &= E(X_1) + E(X_2) + \ldots + E(X_n) \\ &= \mu + \mu + \ldots + \mu \\ &= n\,\mu \end{aligned}$$

The variance:

For the variance note that *Because all the X_i are independent, the covariances are all 0 !!!!!*

$$\begin{aligned} Var(T) &= Var(X_1) + Var(X_2) + \ldots + Var(X_n) \\ &= \sigma^2 + \sigma^2 + \ldots + \sigma^2 \\ &= n\,\sigma^2 \end{aligned}$$

And the average:

$$\bar{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \frac{1}{n}X_3 + \ldots + \frac{1}{n}X_n.$$

The Mean:

$$\begin{aligned} E(\bar{X}) &= \frac{1}{n}E(X_1) + \frac{1}{n}E(X_2) + \ldots + \frac{1}{n}E(X_n) \\ &= \frac{1}{n}\mu + \frac{1}{n}\mu + \ldots + \frac{1}{n}\mu \\ &= n\left(\frac{1}{n}\right)\mu = \mu \end{aligned}$$

$$\bar{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \frac{1}{n}X_3 + \ldots + \frac{1}{n}X_n.$$

The variance:

$$\begin{aligned} Var(\bar{X}) &= \frac{1}{n^2}Var(X_1) + \frac{1}{n^2}Var(X_2) + \ldots + \frac{1}{n^2}Var(X_n) \\ &= \frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \ldots + \frac{1}{n^2}\sigma^2 \\ &= n\,(\frac{1}{n^2})\sigma^2 \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

Note:

We have, for X_i, IID, $E(X_i) = \mu$, $Var(X_i) = \sigma^2$,

$$E(\bar{X}) = \mu, \quad Var(\bar{X}) = \frac{\sigma^2}{n}.$$

Intuitively this says the average of a lot of IID draws tends to be closer to the mean μ than an individual draw.

We do a lot of averaging in statistics !!

This will turn out to be important !!

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.

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!

What ?????!!!!

How can we understand this?

for an individual executive:

$X_i, i = 1, 2, \ldots, 23.$

x	$P(X = x)$
-1	.5
2	.5

μ: $.5 * (-1) + .5 * 2 = 0.5$

σ^2: $.5 * (-1 - .5)^2 + .5 * (2 - .5)^2 = 2.25$

σ: 1.5

μ/σ: $.5/1.5 = 0.3333333$

for CEO:

$$T = X_1 + X_2 + X_3 + \ldots + X_n.$$

$E(T)$: 23*.5 $=$ 11.5

$Var(T)$: 23*2.25 $=$ 51.75

$sd(T)$: sqrt(51.75) $=$ 7.193747

$E(T)/sd(T)$: 11.5/7.193747 $=$ 1.598611

For the CEO, the mean is much bigger relative to the standard deviation that is it for the individual managers !!!

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!

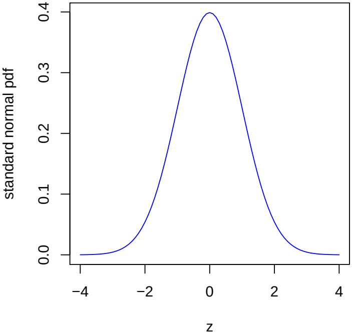
Continuous Random Variables

- 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 called continuous random variables.
- The probability of an interval is defined by the area under the *probability density function, the "pdf"*.

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 probability the number ends up in an interval is given by the area under the curve (pdf)

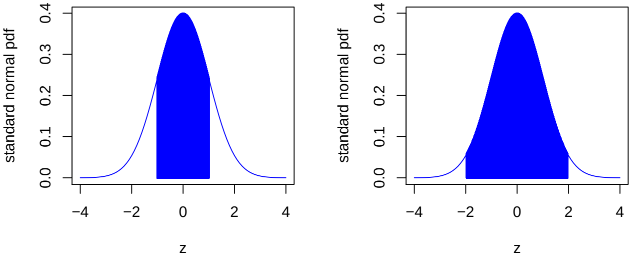
This is the pdf for the *standard normal* distribution.



Notation: We often use Z, to denote a standard normal random variable.

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

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



Note:

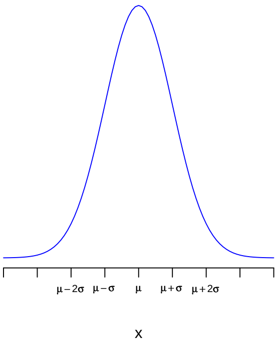
For simplicity we will often use $P(-2 < Z < 2) \approx 0.95$

Questions:

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

- 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).

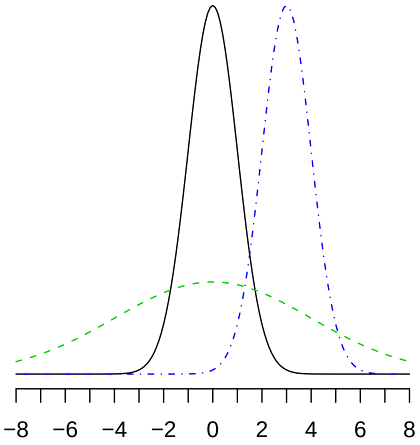
- We write $X \sim N(\mu, \sigma^2)$.
- The parameter μ determines where the curve is. The center of the curve is μ.
- The parameter σ determines how spread out the curve is. The area under the curve in the interval $(\mu - 2\sigma, \mu + 2\sigma)$ is 95%. $P(\mu - 2\,\sigma < X < \mu + 2\,\sigma) \approx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



- 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.
- $Z \sim N(0, 1)$.

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?



Mean and Variance of a Continuous Random Variable

Continuous random variables have expected values and variances analogous to what we have defined for discrete random variables.

But, the definition requires calculus and we don't want to have to remember all that stuff.

Fortunately, our intuition is the same!!!

- The expected value of a random variable is the probability weighted average value.
- The variance of a random variable is the probability weighted average squared distance to the expected value.

Mean and Variance for a Normal

For

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

$$E(X) = \mu, \quad Var(X) = \sigma^2$$

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

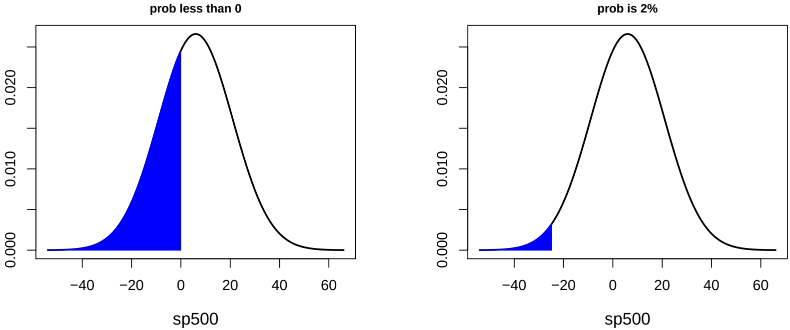
Note:

$$E(Z) = 0, \quad Var(Z) = 1.$$

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) $P(SP500 < 0)$ and (ii) $P(SP500 <?) = 0.02$

The Normal Distribution – Example



$(i)\ P(SP500 < 0) = 0.35$ and $(ii)\ P(SP500 < -25) = 0.02$

In R:

For $X \sim N(\mu, \sigma^2)$, `pnorm(c,...)` gives $P(X < c)$, which is the CDF (cumulative distribution function) evaluated at c.

`qnorm(q,...)` gives the value c such that $P(X < c) = q$.

```
> 1-2*pnorm(-1.96,mean=0,sd=1)
[1] 0.9500042
> 1-2*pnorm(-1.00,mean=0,sd=1)
[1] 0.6826895

> pnorm(0,mean=6,sd=15)
[1] 0.3445783
> pnorm(-25,mean=6,sd=15)
[1] 0.01938279
>
> qnorm(.02,mean=6,sd=15)
[1] -24.80623
```

In Excel see: NORMDIST and NORMINV

The Probability of an Interval:

What is $P(0 < SP500 < 20)$?

That is, what is the probability that the return value ends up being in the interval $(0, 20)$?

```
> pnorm(0,6,15)
[1] 0.3445783
> pnorm(20,6,15)
[1] 0.8246761
> 0.8246761 - 0.3445783
[1] 0.4800978
```

The probability of the interval (a, b) is $CDF(b) - CDF(a)$.

Standardization

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

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

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

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

$\frac{X-\mu}{\sigma}$ should look like a Z !!

The number of standard deviations, X is away from the mean.

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...

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

The month of the crash, R turned out to be $r = -0.2176$. Correspondingly, for the crash,

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

which is pretty wild for a standard normal!!

5 standard deviations away from the mean!!

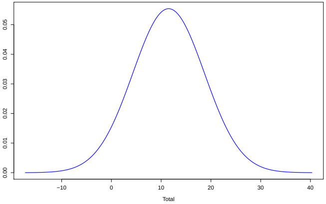
The Normal Distribution – Approximating Combinations of RVs

Recall the Thaler example (Portfolios of projects vs. single project).

A linear combination of independent random variables is approximately normal (the CLT: Central Limit Theorm), so

$$T \sim N(11.5, 7.2^2) \quad \text{approximately}$$

```
> .5 * 23
[1] 11.5
> 2.25*23
[1] 51.75
> sqrt(51.75)
[1] 7.193747
> 1 - pnorm(0,11.5,7.2)
[1] 0.9448919
```



much more compelling than the simple Sharpe ratio we looked at before !!

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?

- 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?

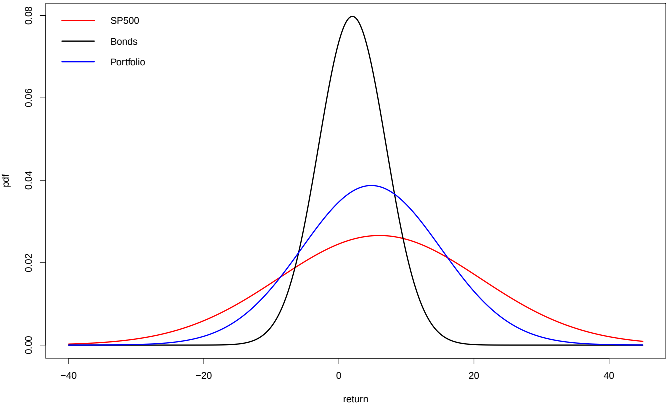
Recall that for two random variables X and Y:

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

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

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

Here are the normal pdfs for our three assets, SP500, Bonds, and the portfolio.



The Uniform Distribution

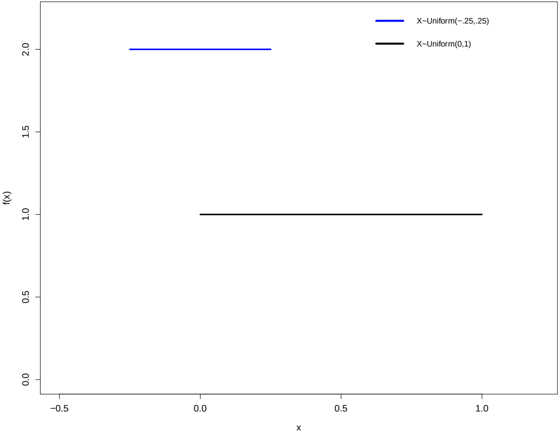
Suppose we think a random variable X can turn out to be any number between -.25 and .25 and the numbers in (-.25,25) are equally likely?

How would we describe this??

Suppose we think a random variable X can turn out to be any number between 0 and 1 and the numbers are equally likely?

How would we describe this?

If X can be any number in (a, b) and the numbers are equally likely, then we say $X \sim Uniform(a, b)$



The density is $\frac{1}{b-a}$ inside (a, b) and 0 elswhere.