Discrete Random Variables, Mean and Variance

Rob McCulloch

1. Why Probability?
2. Discrete Random Variables, Functions of RV's
3. Mean and Variance
4. Mean and Variance of a Linear Function

1. Why Probability?

Probability plays a fundamental role in statistics.

Probability allows us to deal with uncertainty and this comes up in several ways.

We like to look for relationships between variables and often they are not deterministic.

For example when we say "smoking causes cancer" we do not mean if you smoke you will get cancer.

We mean, if you smoke you are *more likely* to get cancer.

Perhaps most fundamentally, we have to make decisions in the presence of uncertainty.

If the probability you get cancer *given* you smoke is .95, you might decide not to smoke.

If the probability you get cancer *given* you smoke is .05, you might decide to smoke!

We do this all the time.
The probability the plane will crash is low enough that I am willing to fly.

How do we get the probabilities?

Some probabilities we just agree on.

If I toss a coin, the probability of a head is about .5.

But for most interesting probabilities we have to go out into the world, collect data, and *estimate* the probabilities.

What is the probability the hurricane will hit a city?
What is the probability the return on my ETF will be less than 2%?
What is the probability the email is spam?

You can see, there are many, many, interesting and important examples !!!

But, when we go out into the world and collect data, there is usually not enough information to be sure about the probabilities.

We have to quantify the amount of information we actually get from data.

The classic example is the $\pm$ attached to poll results.

We have already used regression.

After we have studied the basics of probability and how it is used in statistics, we will return to regression and study it a *a model* with a "random" component.

This will enable us to answer basic questions like:

- I think the slope is 70, but I only have 100 houses in my sample, could I be wrong?
- I can predict using regression, but what is my $\pm$, how sure am I about my prediction?

2. Discrete Random Variables, Functions of RV's

A random variable is *a number we're not sure about*.

Its *distribution* describes what we think it might turn out to be.

For a discrete random variable, we specify the distribution by:

- Listing all the possible numbers it can turn out to be.
- Assigning a probability to each possible outcome.
- Each probability is between 0 and 1.
- The probabilities add up to 1.

Note: "discrete" refers to the situation where can make the list. Later we will look at continuous random variable where such a list is not practical.

Example:

Suppose we are about to toss two coins.

Let X denote the number of heads.

Then the distribution of X might be given by

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

Notation: $P(X = x)$ is "the probability X turns out to be x".
You will see other notations !!

Example:

Let S denote sales next period (thousands of units).

Then the distribution of S might be given by

s	$P(S = s)$
1	.095
2	.23
3	.44
4	.235

What is $P(S > 2)$?
What is $P(S \geq 2)$?

Example:

Let A denote the annual return on a stock.

Then the distribution of A might given by

a	$P(A = a)$
-.02	.2
.06	.5
.14	.3

That is, if ".06" happens then your money goes up by a factor of (1+.06) or 6%.

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:

You are about to toss a coin.

Let X be 1 if it comes up Heads and 0 if tails.

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

Example:

You are about to target a customer.

Let R be 1 if the respond (buy) and 0 otherwise.

For a particular customer we might have:

$$R \sim \text{Bernoulli}(.05)$$

Functions of Random Variables

Functions are our basic mathematical language for specifying relationships.

We use functions with random variables as well.

In our target marketing example, R is 0 or 1 depending on whether the customer responds.

Suppose it costs .8 dollars to target the customer and a response will get you 40 dollars.

If M is our monetary payoff then we can use

$$M = -.8 + 40\,R$$

We don't know what R will turn out to be and we don't know what M will turn out to be, but however they turn out they will be related this way.

From the distribution of R we can figure out the distribution of M.

R:

r	$P(R = r)$
0	.95
1	.05

M:

m	$P(M = m)$
-.8 + 40(0) = -.8	.95
-.8 + 40(1) = 39.2	.05

Example: Portfolio Return

Suppose we want to invest in the stock with return A, but we are worried that it looks a bit risky.

Rather than putting all our money into the stock, let's put 75% into the stock and 25% into a riskless bond that pays 2% for sure.

Let P denote the return on this portfolio.

What is the distribution of P??

Quick Review: Return on a Portfolio

Suppose you put w_1 of your the total investment amount into asset 1, with return r_1 and you put w_2 into asset 2, with return r_2.

For example, $w_1 = w_2 = .5$ means half of your money goes into each of the two assets.

You are just investing in these two assets so $w_1 + w_2 = 1$.

The w's are called the portfolio weights.

What is the return on the portfolio?

Suppose you invest M (amount of money) in the portfolio.

$$M \Rightarrow (M\,w_1)(1+r_1)+(M\,w_2)(1+r_2) = M(w_1+w_2+(w_1\,r_1+w_2\,r_2))$$

$$= M(1+(w_1\,r_1+w_2\,r_2))$$

So, the return is $w_1\,r_1 + w_2\,r_2$.

Portfolio return:

$$w_1 r_1 + w_2 r_2$$

Example:

Suppose the return on a riskless asset is .02.

The return on the risky asset turns out to be .06.

Portfolio weights .25 (riskless) and .75 (risky).

That is: $r_1 = .02$, $r_2 = .06$, $w_1 = .25$, and $w_2 = .75$.

Then the portfolio return is

$$.25*.02 + .75*.06 = .05$$

or 5%.

In general, with a riskless return of 2%, risky return A, and 75% into the risky asset we have:

$$\begin{aligned} P &= .25(.02) + .75(A) \\ &= .005 + .75\,A \end{aligned}$$

P and A are linearly related !!

$$P = .005 + .75\,A$$

A:

a	$P(A = a)$
-.02	.2
.06	.5
.14	.3

P:

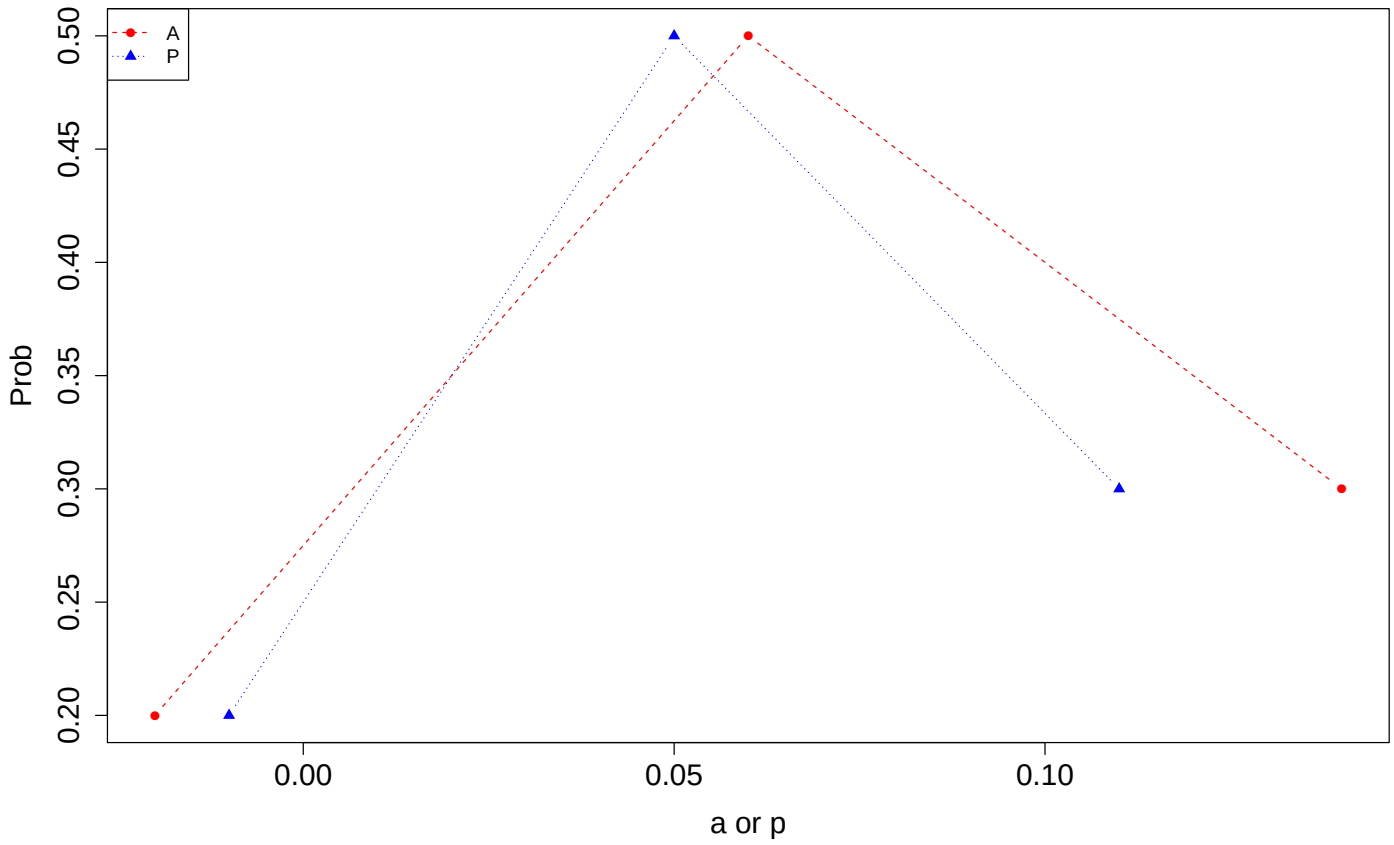
p	$P(P = p)$
.005 + .75*(-.02) = -0.01	.2
.005 + .75*(.06) = .05	.5
.005 + .75*(.14) = 0.11	.3

$$P = .005 + .75\,A$$

a	$P(A = a)$
-.02	.2
.06	.5
.14	.3

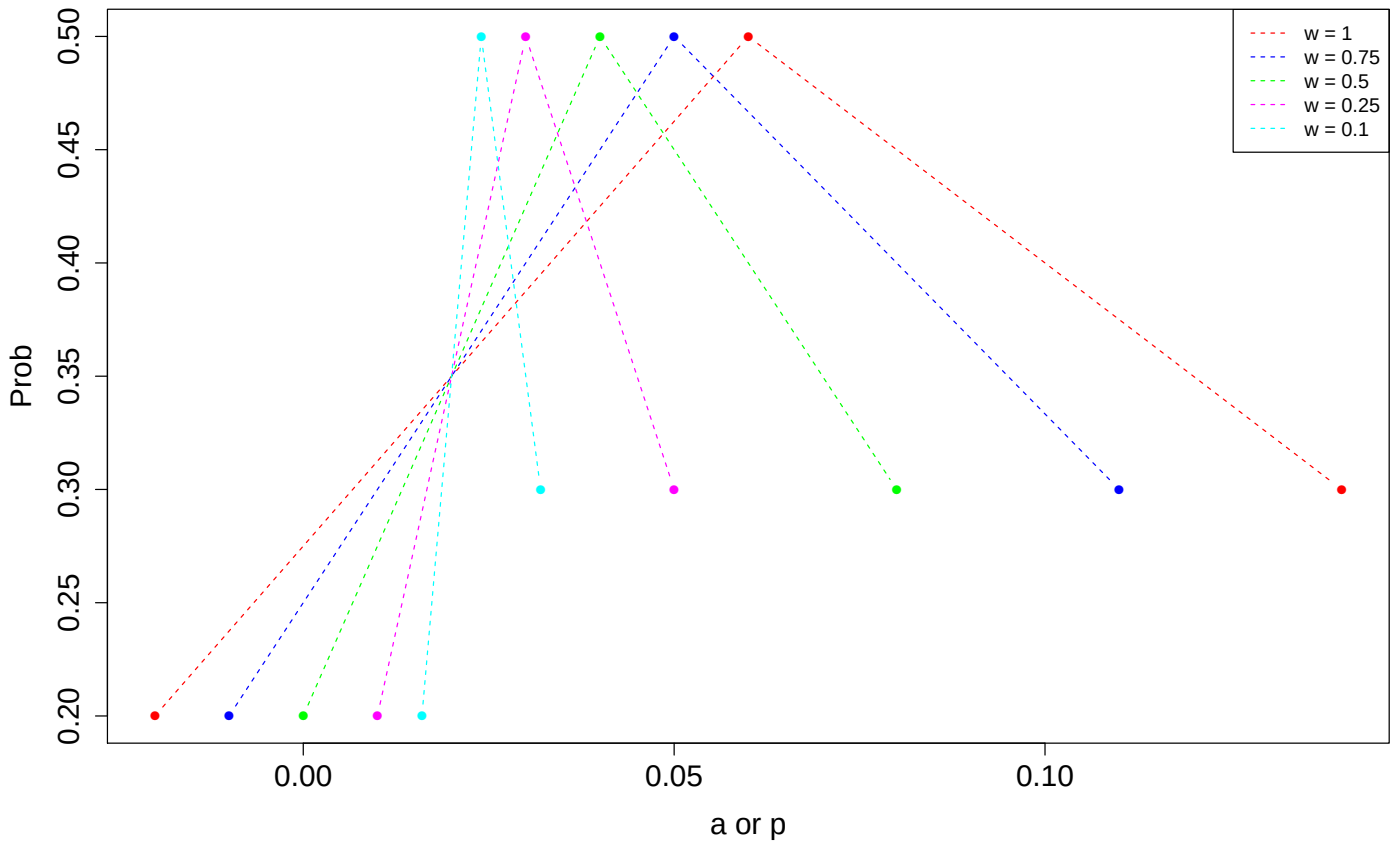
p	P(P=p)
-0.01	.2
.05	.5
0.11	.3

We can visualize the two distributions by plotting the values against the probabilities.



Which would you rather own, A or P ??

Return distributions you could get by varying the weight w on the risky asset.



3. Mean and Variance

The Expected Value

Recall our sales example.

Suppose you "believe" this distribution.

s	$P(S = s)$
1	.095
2	.23
3	.44
4	.235

Suppose people want you to come up with one number which is your prediction for sales.

What number would you give?

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 it is.

Gives a one number summary of what kind of number you get.

$E(S)$

$$E(S) = \sum_{i=1}^{4} P(s_i) \times s_i$$

s	$P(S = s)$	$P(S = s) \times s$
1	.095	.095
2	.23	.46
3	.44	1.32
4	.235	.94

$$E(S) = .095 + .46 + 1.32 + .94 = 2.815.$$

$E(A)$, $E(P)$

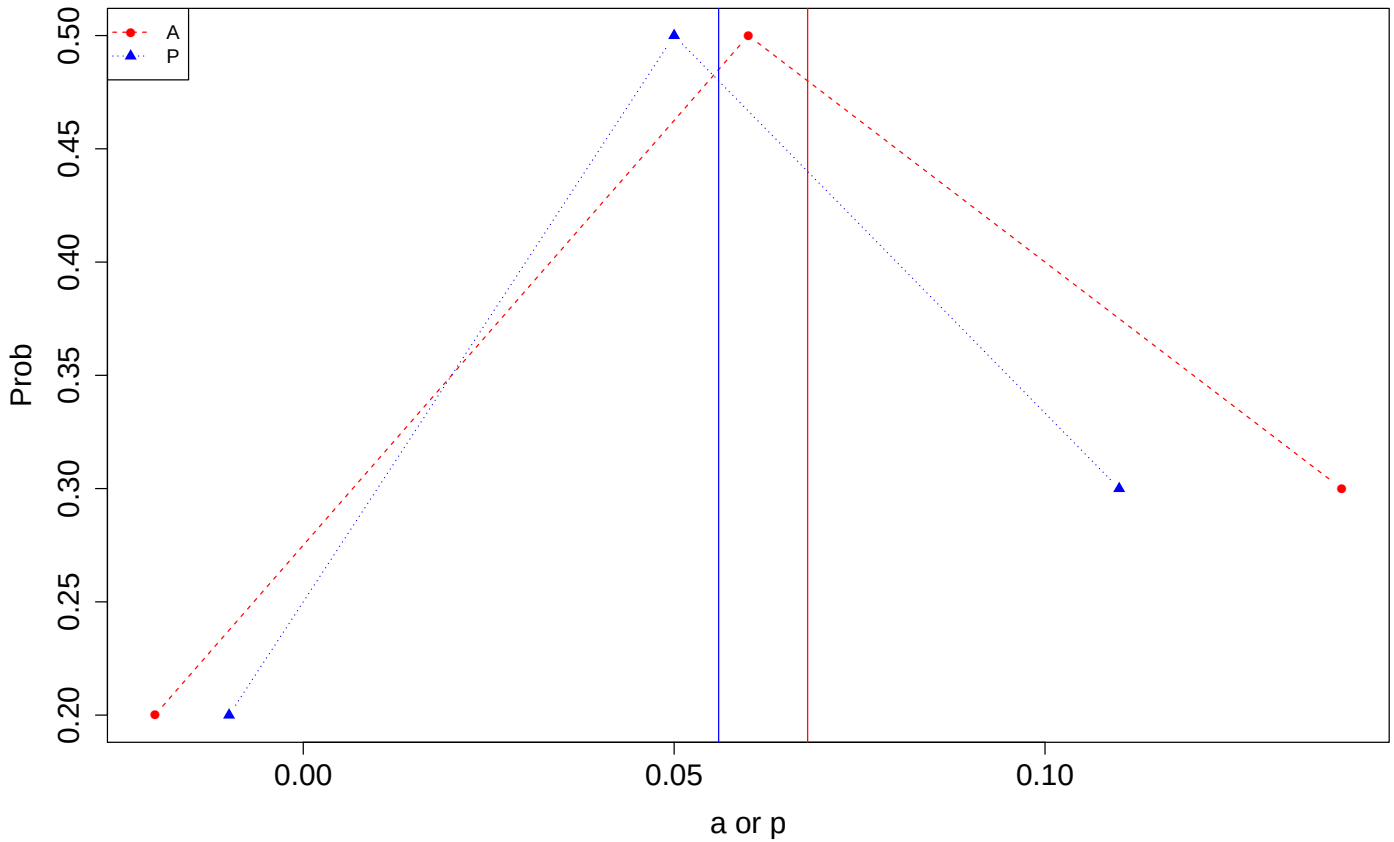
a	$P(A = a)$
-.02	.2
.06	.5
.14	.3

p	P(P=p)
-0.01	.2
.05	.5
0.11	.3

$$E(A) = .2 * (-.02) + .5 * .06 + .3 * .14 = 0.068.$$

$$E(P) = .2 * (-.01) + .5 * .05 + .3 * .11 = .056.$$

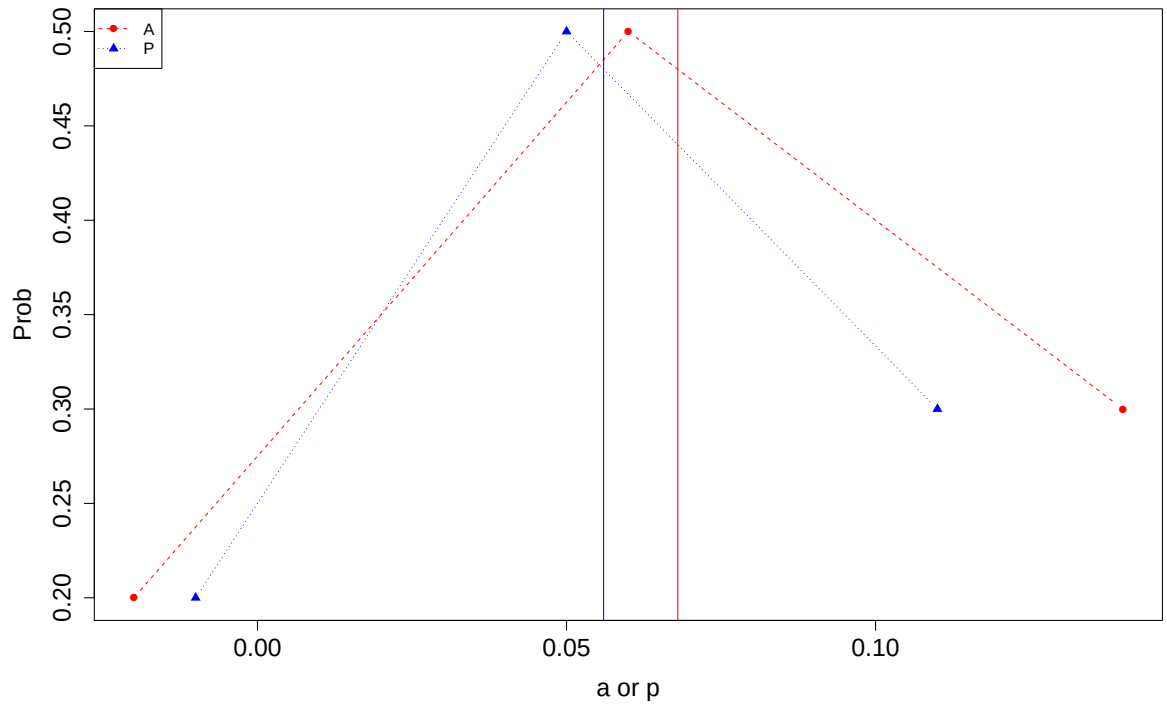
Solid vertical lines at the means.



The Variance and Standard Deviation

The A distribution has a higher mean, but that does not mean we should choose it.

The A distribution is also more spread out or "riskier".



The variance and standard deviation are often used to measure how spread out a distribution is.

Variance of a Random Variable

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.

Example, the Variance of S

s	$P(S = s)$
1	.095
2	.23
3	.44
4	.235

Recall, $E(S) = 2.815$.

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

s	$P(S = s)$	$P(S = s) * (s - E(S))^2$
1	.095	$.095 * (1 - 2.815)^2 = 0.313$
2	.23	$.23 * (2 - 2.815)^2 = 0.153$
3	.44	$.44 * (3 - 2.815)^2 = 0.015$
4	.235	$.235 * (4 - 2.815)^2 = 0.33$

$Var(S) = .313 + .153 + .015 + .33 = 0.811$

Example, the Variance of A

a	$P(A = a)$
-.02	.2
.06	.5
.14	.3

Recall, $E(A) = 0.068$

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

a	$P(A = a)$	$P(A = a) * (a - E(a))^2$
-.02	.2	$.2 * (-.02 - .068)^2 = 0.0015488$
.06	.5	$.5 * (.06 - .068)^2 = .000032$
.14	.3	$.3 * (.14 - .068)^2 = 0.0015552$

$Var(A) = 0.0015488 + .000032 + 0.0015552 = 0.003136$

Example, Variance of P and A

$Var(P) = .2 * (-.01 - .056)^2 + .5 * (.05 - .056)^2 + .3 * (.11 - .056)^2$
$= 0.001764.$

As we expect:

$$Var(P) < Var(A)$$

But the little numbers are hard to interpret!!

What are the units of the variance?

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 $sd(X)$?

sd of P and A

$$sd(A) = \sqrt{0.003136} = .056.$$

$$sd(P) = \sqrt{0.001764} = .042.$$

Compare:

A: mean $= .068$, sd $= .056$
P: mean $= .056$, sd$= .042$.

Expected Value and Variance of a Bernoulli

$X \sim$ Bernoulli(p) means:

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

$E(X) = (1 - p) \times 0 + p \times 1 = p.$

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

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

Don't get confused !!!

If Y is a random variable, we have talked about the "mean of Y", or $E(Y)$.

If y_i is a bunch of numbers we have talked about "the mean of y", or $\bar{y}$.

Later on we will talk about connections between these two different meanings of the mean, but for now, notice they are different animals.

4. Mean and Variance of a Linear Function

We saw that with portfolio weights .25 for the bond and .75 for the stock we got

$$P = .005 + .75\,A, \quad E(P) = .056, \quad sd(P) = .042.$$

In general, if we put w into the stock and $(1 - w)$ into the bond and let P_w denote the consequent portfolio return then

$$P_w = .02\,(1 - w) + w\,A.$$

As w increases the mean should go up but so should the variance.

We'd like to compute the mean and variance (or standard deviation) for various w and choose the one we like. For each w, P_w is a linear function of A.

I turns out that when one random variable is a linear function of another the means and variances are related by a simple formula.

For, $Y = a + b\,X$,

$$E(Y) = a + b\,E(X)$$

$$Var(Y) = b^2\,Var(X)$$

$$sd(Y) = |b|\,sd(X)$$

Example

In our target marketing example, let R be 1 if the customer responds and 0 otherwise.

In our initial example we had $p = .05$.

Then,

$$R \sim \text{Bernoulli}(.05)$$

The monetary payoff is

$$M = -.8 + 40\,R$$

So,

$$E(M) = -.8 + 40 * E(R) = -.8 + 40 * .05 = 1.2.$$

$$R \sim \text{Bernoulli}(.05)$$

$$M = -.8 + 40\,R$$

We did not use the variance and sd of M but, as an example of the formulas,

$$Var(M) = 40^2\, Var(R) = 1600 * (.05 * (1 - .05)) = 76.$$

$$sd(M) = 40\, sd(R) = 40 * \sqrt{.05 * (1 - .05)} = 8.72.$$

Of course $\sqrt{76} = 8.717798$.

Example

We have

$$E(A) = .068, \quad Var(A) = .003136, \quad sd(A) = 0.056$$

$$P = .005 + .75\,A$$

So,

$E(P) = .005 + .75 * .068 = .056.$

$Var(P) = .75^2 * .003136 = 0.001764.$

$sd(P) = .75 * 0.056 = .042.$

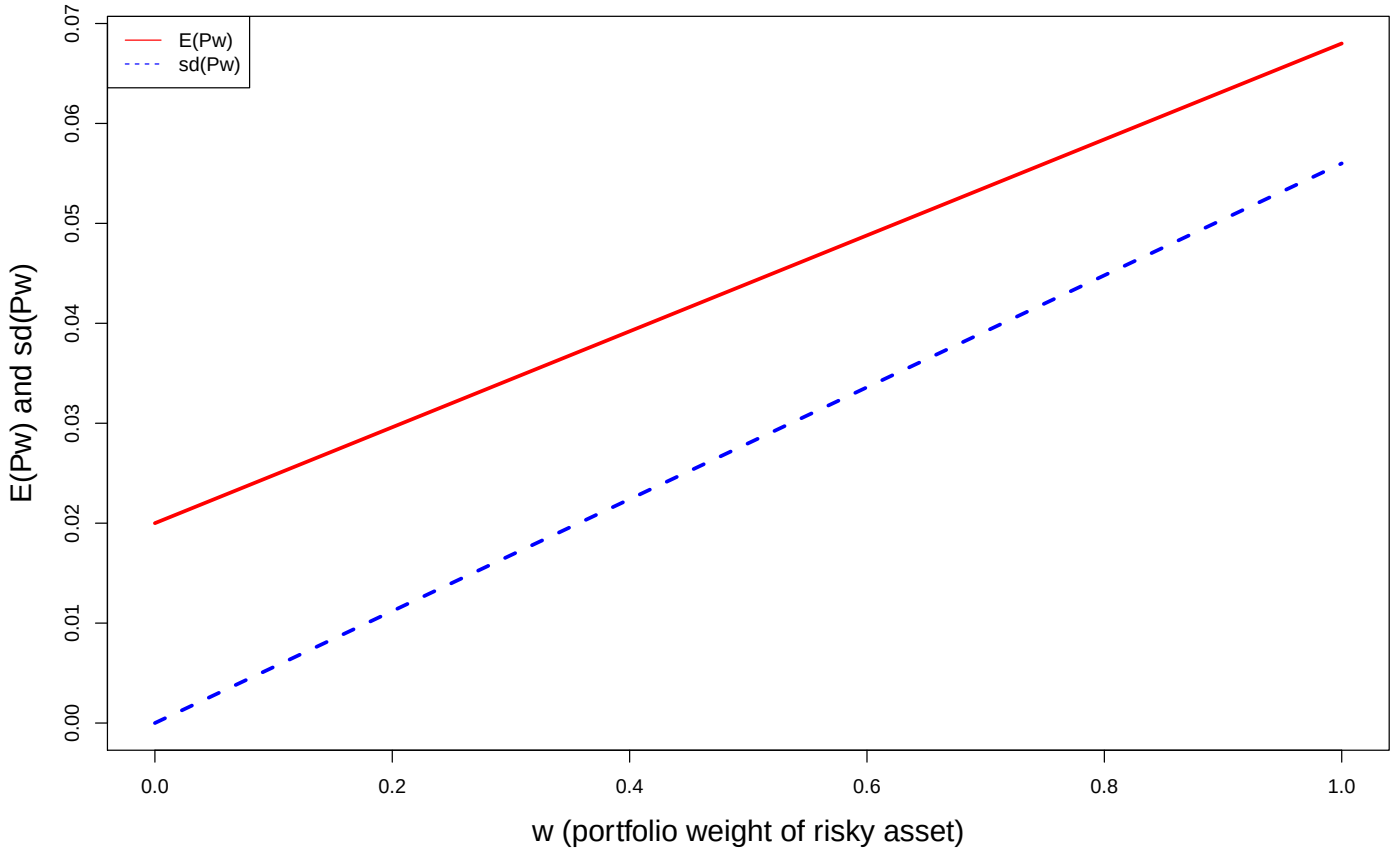
For general w we have,

$$P_w = .02\,(1 - w) + w\,A.$$

$E(P_w) = .02(1 - w) + w * .068 = .02 + .048\,w.$

$sd(P_w) = .056\,w.$

Plot $E(P_w)$ and $sd(P_w)$ against w.



- As w increases the mean increases, *but so does the sd !*.
- At $w = 1$ you get the mean and sd of A.
- At $w = 0$ you get the mean and sd=0 of putting all your money in the riskless asset.

Example

$Z = 2\,X$

$E(Z) = 2\,E(X)$

$Var(Z) = 4\,Var(X)$

$sd(Z) = 2\,sd(X)$

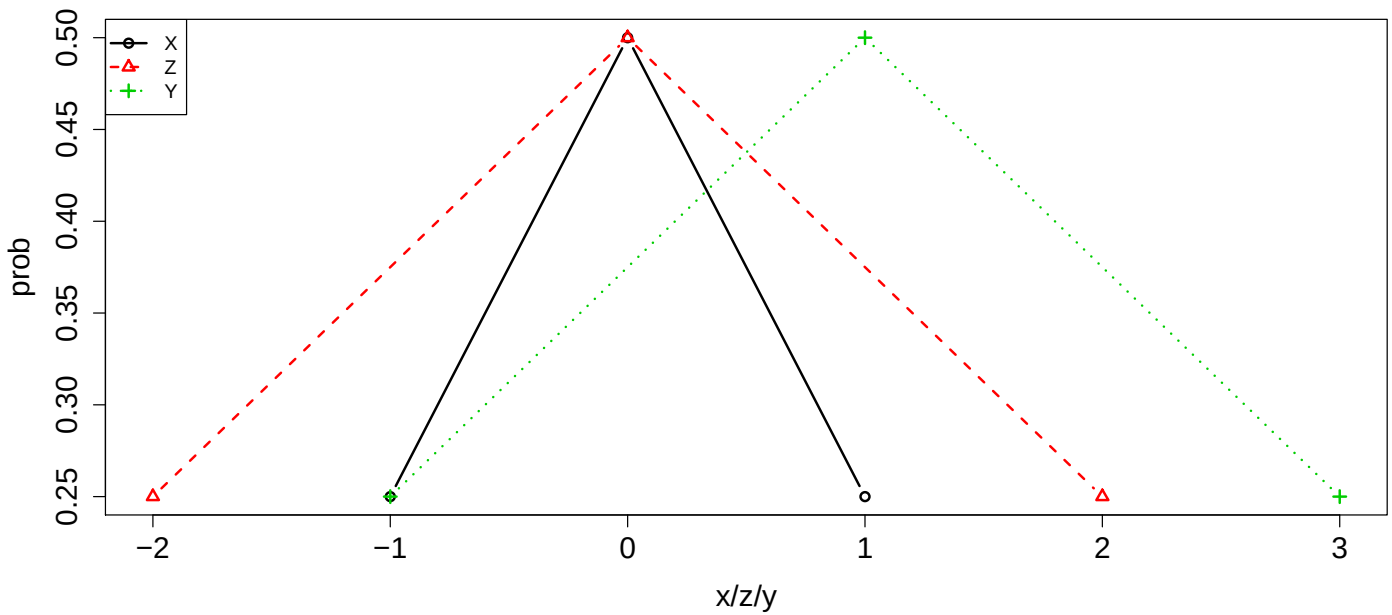
Example

$Y = 1 + Z$

$E(Y) = 1 + E(Z)$

$Var(Y) = Var(Z)$

$sd(Y) = sd(Z)$



Example:

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

$f(x) = x^2$.

What are the mean and variance of $Y = f(X)$?