

Section 1: Introduction and Probability Concepts

Jared S. Murray

Readings: Naked Statistics chapters 5, 5 1/2 and 6

Getting Started

- ▶ Syllabus
- ▶ General Expectations
 1. Read the notes
 2. Read the book
 3. Work on homework assignments

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?

Introduction

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

- ▶ If I am vaccinated, how likely am I to get Covid?
- ▶ 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?
- ▶ Bank runs, chatGPT, etc...

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:

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

- ▶ 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

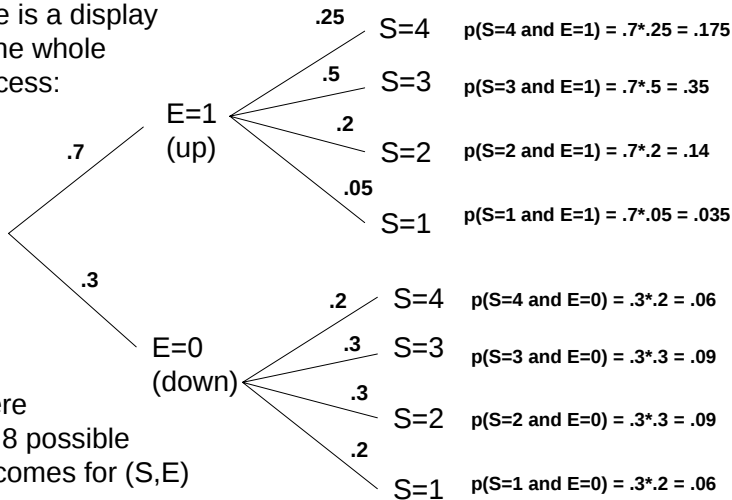
The conditional distributions tell us about about what can happen to S for a given value of E ... but what about S and E jointly?

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

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

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				
		1	2	3	4	
E	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,

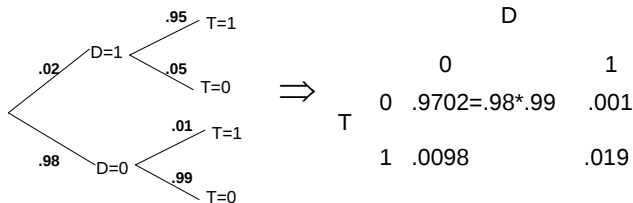
knowing X tells you nothing about Y !

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?

Disease Testing Example

Let $D = 1$ indicate you have a disease

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



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

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

Disease Testing Example

- ▶ 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?

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

- ▶ isn't that the same?!

Prosecutor's Fallacy

- ▶ Assume you are in a jury for a murder trial in a city of roughly 10 million people.
- ▶ In his opening remarks, the prosecutor tells you that the defendant has been arrested on the strength of a single, overwhelming piece of evidence: that his DNA matched a sample of DNA taken from the scene of the crime.
- ▶ To convince you of the strength of this evidence, the prosecutor calls a forensic scientist to the stand, who testifies that the probability that an innocent person's DNA would match the sample found at the crime scene is only one in a million. The experts also testifies that the probability that a guilty person DNA would match is one.
- ▶ Would you vote to convict this man?

Prosecutor's Fallacy

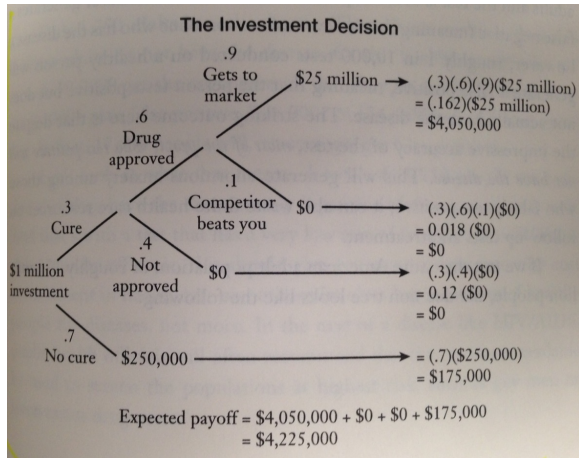
- ▶ Denote G for guilty... what is $P(G)=?$
- ▶ Denote DNA for a positive DNA match... what is $P(DNA|not\ G)?$
- ▶ What is the relevant probability to inform your decision?
 $P(DNA|not\ G)$ or $P(not\ G|DNA)$

Differential Privacy

- ▶ Let's play a little game...
- ▶ You will flip a coin... if heads, you will write down the number 1 if your social security number ends with an even digit, otherwise write down 0.
- ▶ if tails, you will write down the number 1 if you ever got a positive Covid test, otherwise 0.
- ▶ Question: What percentage of students have ever tested positive for Covid?

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...

<i>Revenue</i>	<i>P(Revenue)</i>
\$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

What if you could choose this investment instead?

<i>Revenue</i>	<i>P(Revenue)</i>
\$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

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

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

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

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 $sd(X)$?

Targeted 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 ???

Targeted 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)

Targeted Marketing

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

<i>Profits</i>	<i>P(Profits)</i>
\$-0.8	0.95
\$39.20	0.05

Should you do the promotion?

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

Option Pricing

Suppose Nvidia stock price is \$900 today

After some analysis of the companies fundamentals you predict two scenarios for next month: (i) a 50% chance the stock moves up to \$1000 and (ii) 50% chance the stock moves down to \$800.

Assuming a risk-free rate of return of zero (why am I even bringing this up?), what is the fair price for a Nvidia call option (option to buy) for next month, with a strike price of \$900?

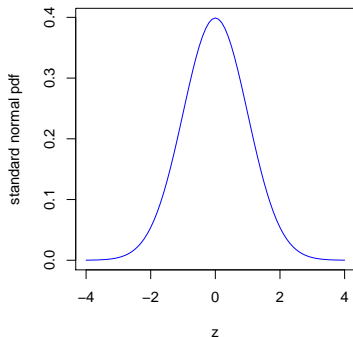
Continuous Random Variables

- ▶ 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 Galton Board

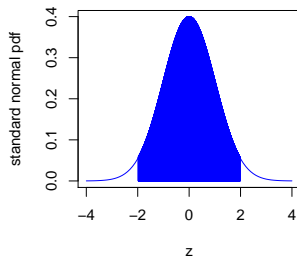
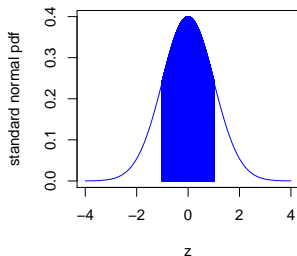
<https://youtu.be/AwEaHCjgeXk?si=ojrZnMAHX1itxtom>

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:

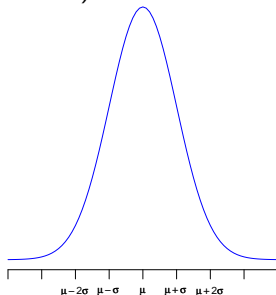
- ▶ What is $Pr(Z < 2)$? How about $Pr(Z \leq 2)$?
- ▶ What is $Pr(Z < 0)$?

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

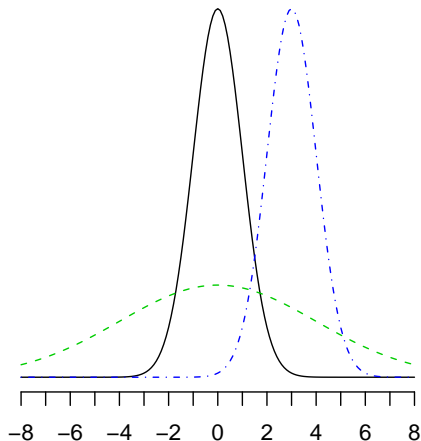


Mean and Variance of a Random Variable

- ▶ 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 an 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

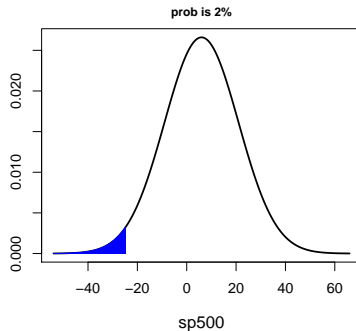
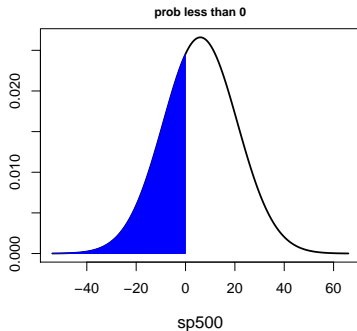
- ▶ **Example:** Below are the pdfs of $X_1 \sim N(0, 1)$, $X_2 \sim N(3, 1)$, and $X_3 \sim N(0, 16)$.
- ▶ Which pdf goes with which X ?



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 zvalue? **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...