Random Variables and Probability Models

Ryan Miller

Outline

1. Random variables
   - continuous vs. discrete, expected value, variance and standard deviation, linear combinations
2. Normal distributions
   - parameters, probability calculations, Z-scores
3. Binomial distributions
   - parameters, expected value and variance, probability calculations, normal approximation

Introduction

- A **random variable** is a variable used to represent the unknown numerical outcome of a random process
  - A **continuous** random variable can take on *infinitely many* different numerical values
  - A **discrete** random variable can take on *countably many* different numerical values

Introduction

- A **random variable** is a variable used to represent the unknown numerical outcome of a random process
    - A **continuous** random variable can take on *infinitely many* different numerical values
    - A **discrete** random variable can take on *countably many* different numerical values
- For any random variable:
    - **Expected value** describes the average outcome we'd expect if the random process were observed many times
    - **Variance** and **standard deviation** describe the expected variation in outcomes around the expected value

Practice

A student enrolls in a course that might have one of three instructors:

- The first instructor requires a textbook that is free
- The second instructor requires a textbook costing $90
- The third instructor requires a textbook costing $120

The student estimates a 50% chance the course is staffed by the first instructor, a 30% chance its staffed by the second instructor, and a 20% chance its staffed by the third instructor.

1) Let the random variable, X, denote the amount of money the student spends on the textbook for this course. Is X a discrete or continuous random variable?
2) Create a table to represent the probability distribution of X.

Practice (solution)

1) The random variable, X, is discrete as there only 3 distinct outcomes
2) X follows the probability distribution:

Instructor	1	2	3
x_i	0	90	120
$P(X = x_i)$	0.5	0.3	0.2

Expected value (discrete random variables)

For a discrete random variable, **expected value** is the sum of each outcome weighted by that outcome's probability:

$$\begin{aligned} E(X) &= x_1 * P(X = x_1) + x_2 * P(X = x_2) + \dots \\ &= \sum_{i=1}^{k} x_i * P(X = x_i) \end{aligned}$$

Expected value (discrete random variables)

For a discrete random variable, **expected value** is the sum of each outcome weighted by that outcome's probability:

$$\begin{aligned} E(X) &= x_1 * P(X = x_1) + x_2 * P(X = x_2) + \dots \\ &= \sum_{i=1}^{k} x_i * P(X = x_i) \end{aligned}$$

Practice: What is the expected value of X in the textbook cost example? How do you interpret $E(X)$?

Practice (solution)

In the textbook example:

$$E(X) = 0 * 0.5 + 90 * 0.3 + 120 * 0.2 = 51$$

This is the amount the student can *expect* to pay (ie: the long-run average if the random process were observed many times). Notice that it's not particularly close to any of the actual outcomes. . .

Variance (discrete random variables)

For a discrete random variable, **variance** is the sum of squared deviations of each outcome from the expected value weighted by the outcome's probability:

$$
\begin{aligned}
Var(X) &= (x_1 - E(X))^2 * P(X = x_1) + (x_2 - E(X))^2 * P(X = x_2) + \dots \\
&= \sum_{i=1}^{k} (x_i - E(X))^2 * P(X = x_i)
\end{aligned}
$$

Variance (discrete random variables)

For a discrete random variable, **variance** is the sum of squared deviations of each outcome from the expected value weighted by the outcome's probability:

$$\begin{aligned} Var(X) &= (x_1 - E(X))^2 * P(X = x_1) + (x_2 - E(X))^2 * P(X = x_2) + \dots \\ &= \sum_{i=1}^{k} (x_i - E(X))^2 * P(X = x_i) \end{aligned}$$

For the textbook example (recall $E(X) = 51$):

$$Var(X) = (0 - 51)^2 * 0.5 + (90 - 51)^2 * 0.3 + (120 - 51)^2 * 0.2 = 2709$$

Standard deviation (discrete random variables)

Standard deviation is defined as the square-root of a random variable's variance:

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

For the textbook example:

$$Sd(X) = \sqrt{2709} = 52.05$$

Practice: How would interpret the standard deviation of $52.05 in the textbook example?

Practice (solution)

- Standard deviation roughly describes the *expected deviation* of outcomes from the expected value over many repetitions of a random process
- In the textbook example, a standard deviation of \$52.05 suggests the student should plan for a large degree of variability
    - That is, a textbook cost of \$0 (~ 1 SD below the expected value) or a cost $>$ \$100 (~ 1 SD above the expected value) would not be unexpected

Linear combinations of random variables

Let $aX + bY$ denote a *linear combination* of two random variables, X and Y

- $E(aX + bY) = a * E(X) + b * E(Y)$
- $Var(aX + bY) = a^2 * Var(X) + b^2 * Var(Y)$ (for independent random variables only!)

Note: if X and Y are not independent, we'd need to consider the *covariance* between them. In this scenario:
$Var(aX + bY) = a^2 * Var(X) + b^2 * Var(Y) + 2ab * Cov(X, Y)$

Practice

- Suppose an individual investor has \$6000 invested in the SPY and \$2000 in the QQQ
  - For simplicity, we'll assume each fund's returns are independent
- We'll let X to denote the percentage change in price over the next month for SPY, and Y denote the change for QQ
  - Historical data indicates SPY has increased in value by an average of 0.006 each month (0.6% monthly gain) with a standard deviation of 0.04
  - QQQ has increased in value by 0.008 each month (0.8% monthly gain) with a standard deviation of 0.07

Practice

- Suppose an individual investor has \$6000 invested in the SPY and \$2000 in the QQQ
  - For simplicity, we'll assume each fund's returns are independent
- We'll let X to denote the percentage change in price over the next month for SPY, and Y denote the change for QQ
  - Historical data indicates SPY has increased in value by an average of 0.006 each month (0.6% monthly gain) with a standard deviation of 0.04
  - QQQ has increased in value by 0.008 each month (0.8% monthly gain) with a standard deviation of 0.07

1) What is the *expected return* of this portfolio?
2) What is the accompanying *standard deviation*?
3) Why should both of these be important to the investor?

Practice (solution)

1) The expected return is $E(6000 * X + 2000 * Y) = 6000 * E(X) + 2000 * E(Y) = 6000 * 0.006 + 2000 * 0.008 = 52$

2) First, $Var(X) = 0.04^2 = 0.0016$ and $Var(Y) = 0.07^2 = 0.0049$. Then, $Var(6000 * X + 2000 * Y) = 6000^2 * 0.0016 + 2000^2 * 0.0049 = 77200$. So, $SD(6000 * X + 2000 * Y) = \sqrt{Var(6000 * X + 2000 * Y)} = \sqrt{77200} = 277.85$

3) The individual can expect an average monthly return of \$52 on their \$8000 investment. However, since the standard deviation is $\sqrt{277.85}$, they should anticipate a sizable degree of month-to-month fluctuation, with some months resulting in losses and others resulting in gains.

Continuous random variables

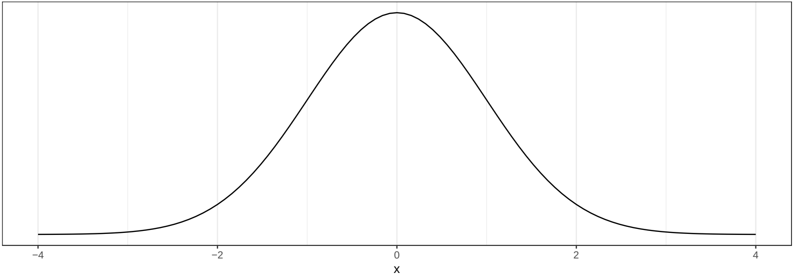
With the help of calculus, the same definitions can be extended to continuous random variables:

- $E(X) = \int_{S_x} x * p(x)dx$
- $Var(X) = \int_{S_x} (x - E(X))^2 * p(x)dx$
- $SD(X) = \sqrt{Var(X)}$

where S_x denotes the sample space of X, and $p(X)$ denotes the *probability density function* of X.

The Normal distribution

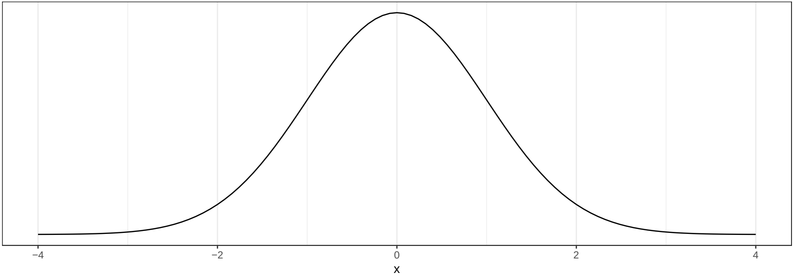
We'll focus heavily on the **normal distribution**:



- The normal curve is a symmetric, bell-shaped distribution defined by its expected value (center), μ, its a standard deviation, σ

The Normal distribution

We'll focus heavily on the **normal distribution**:



- The normal curve is a symmetric, bell-shaped distribution defined by its expected value (center), μ, its a standard deviation, σ
- The **standard normal** distribution is shown above, it's centered at $\mu = 0$ with a standard deviation of $\sigma = 1$
  - We'll use the shorthand: $N(0, 1)$

The Normal distribution

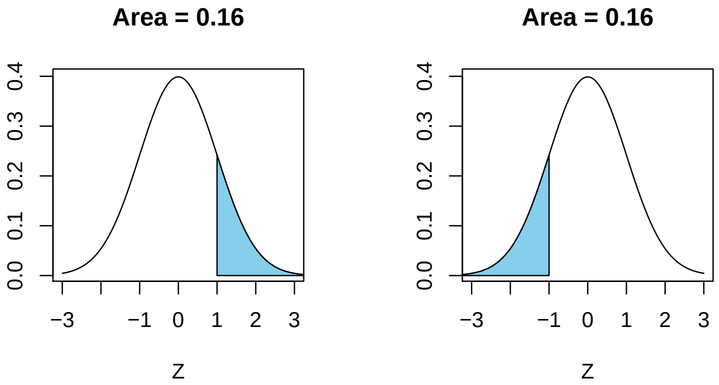
The normal curve's *probability density function* is defined:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Recall that μ and σ are constants defining the center and spread of the curve
- There is no closed-form integral for the normal curve

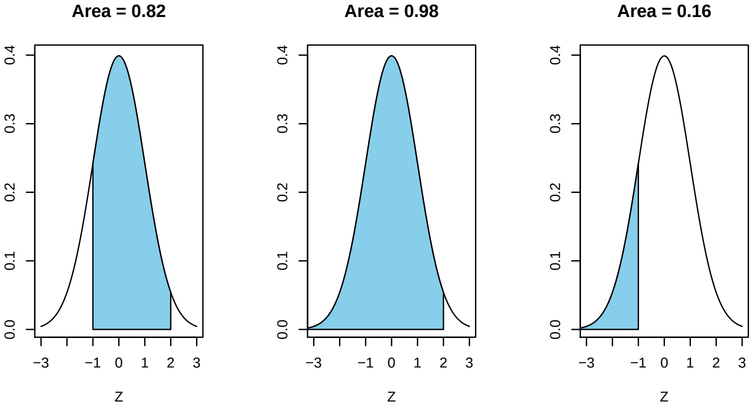
The Normal distribution and probability

For the standard normal distribution: $P(Z \geq t) = P(Z \leq -t)$



The Normal distribution and probability

For the standard normal distribution:
$P(a \leq Z \leq b) = P(Z \leq b) - P(Z \leq a)$



Area = 0.82

Area = 0.98

Area = 0.16

Normal approximations

- While few (if any) random variables will *exactly* follow a Normal distribution, it serves as a *useful model* in a wide variety of applications
  - Shown below are two important R functions for working with Normal models:

```
## pnorm accepts a value (quantile) and returns a probability
pnorm(q = 0.5, mean = 5, sd = 10, lower.tail = TRUE)
```

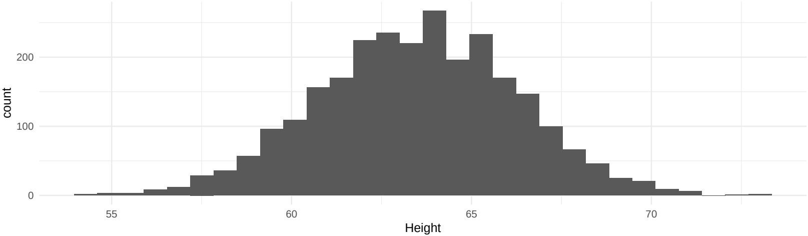
```
## [1] 0.3263552
```

```
## qnorm accepts a probability and returns a value (quantile)
qnorm(p = 0.5, mean = 5, sd = 10)
```

```
## [1] 5
```

Practice

The National Health and Nutrition Examination Survey (NHANES) collected the heights of 2,649 adult women. The data showed a mean of 63.5 inches and a standard deviation of 2.75 inches:



1. Estimate the probability that a randomly selected woman is under 5 ft tall (60 in)
2. Estimate the probability that a randomly selected woman is between 5'3 and 5'6 (63 in and 66 in)
3. At what height would you expect a woman to be taller than 95% of her peers?

Practice (solution)

1. $P(X < 60) = 0.102$ (found using `pnorm`)
2. $P(63 < X < 66) = 0.390$ (found using `pnorm` twice and subtracting)
3. Using a Normal model, at 68.02 inches (approximately 5'8) we'd expect a woman to be taller than 95% of her peers (found using `qnorm`)

Standardization and Z-scores

- There are infinitely many different Normal distributions (one for each combination of μ and σ)
  - Statisticians historically needed to *transform* their data to follow the standard Normal curve, then use a large table to find probabilities or quantities

Standardization and Z-scores

- There are infinitely many different Normal distributions (one for each combination of μ and σ)
  - Statisticians historically needed to *transform* their data to follow the standard Normal curve, then use a large table to find probabilities or quantities
- The Z-transformation is used to produce Z-scores, which are the unit-free measurements on the scale of “standard deviations”:

$$Z_i = \frac{X_i - E(X)}{SD(X)}$$

- Although no longer necessary for probability calculations, Z-scores are popular as method for comparing variables measured on different scales

Example (Z-scores)

- Suppose a blood test reveals the concentration of urea in your blood is 50 mg/dl above average, what do you conclude?

Example (Z-scores)

- Suppose a blood test reveals the concentration of urea in your blood is 50 mg/dl above average, what do you conclude?
  - Probably not very much, to a non-expert these units are meaningless
- Now suppose you're told the concentration is 4 standard deviations above average (a Z-score of $+4$), what do you conclude?

Example (Z-scores)

- Suppose a blood test reveals the concentration of urea in your blood is 50 mg/dl above average, what do you conclude?
  - Probably not very much, to a non-expert these units are meaningless
- Now suppose you're told the concentration is 4 standard deviations above average (a Z-score of $+4$), what do you conclude?
  - You should be very worried! 4 standard deviations above average is extremely high - it's higher than 99.99% of people assuming a Normal model

The Bernoulli distribution

- The Normal distribution is not feasible for random variables with a small number of discrete outcomes

The Bernoulli distribution

- The Normal distribution is not feasible for random variables with a small number of discrete outcomes
- A **Bernoulli random variable** takes a value of 1 with a probability of success defined by p, and a value of 0 otherwise (with probability $1 - p$)
  - The **Bernoulli distribution** models a random process with a *binary outcome*
  - If X follows the Bernoulli distribution, $E(X) = p$ and $SD(X) = \sqrt{p * (1 - p)}$

The binomial distribution

- Many random processes can be viewed as aggregations of multiple Bernoulli trials
  - For example, suppose a couple plans on having 4 children, what is the probability that exactly 2 are boys?

The binomial distribution

- Many random processes can be viewed as aggregations of multiple Bernoulli trials
  - For example, suppose a couple plans on having 4 children, what is the probability that exactly 2 are boys?
- An *incorrect* approach would be to calculate this probability as $P(B) * P(B) * P(G) * P(G) = 0.5^4 = 0.0625$
  - This fails to account for the different *orderings* or *combinations* of boys/girls that are possible

The binomial distribution

- Many random processes can be viewed as aggregations of multiple Bernoulli trials
  - For example, suppose a couple plans on having 4 children, what is the probability that exactly 2 are boys?
- An *incorrect* approach would be to calculate this probability as $P(B) * P(B) * P(G) * P(G) = 0.5^4 = 0.0625$
  - This fails to account for the different *orderings* or *combinations* of boys/girls that are possible
- There are 6 different ways for the couple to have exactly 2 boys (BBGG, BGBG, BGGB, GBBG, GGBB, GBGB)
  - Since each of these combinations is equally likely, we must multiply $6 * 0.5^4$ to correctly calculate the probability of observing exactly 2 boys (which is 0.375)

The binomial distribution

The **binomial distribution** describes the number of successes in a fixed number of independent Bernoulli trials:

$$P(X = x) = \binom{n}{x}(p)^x(1-p)^{n-x}$$

- p is the probability of success in each trial and n is the number trials
- If X follows the binomial distribution, $E(X) = n * p$ and $SD(X) = \sqrt{n * p * (1-p)}$
- The binomial distribution assumes all trials are *independent*, have a *binary outcome*, and have the *same success probability*

The binomial distribution in R

Below are three R functions related to binomial probability models:

```
## P(X >= 8)
pbinom(q = 7, size = 10, prob = 0.7, lower.tail = FALSE)
```

```
## [1] 0.3827828
```

```
## P(X = 7)
dbinom(x = 7, size = 10, prob = 0.7)
```

```
## [1] 0.2668279
```

```
##
qbinom(0.5, size = 10, prob = 0.7)
```

```
## [1] 7
```

- `pbinom` calculates tail-area probabilities
- `dbinom` calculates probabilities for individual values of X
- `qbinom` returns the value of X corresponding to a certain percentile of the distribution

Comparing pbinom and dbinom

Below is an illustration of how `dbinom` and `pbinom` are related:

```r
## Calculating P(X >= 8) using dbinom
dbinom(x = 8, size = 10, prob = 0.7) +
  dbinom(x = 9, size = 10, prob = 0.7) +
  dbinom(x = 10, size = 10, prob = 0.7)
```

```
## [1] 0.3827828
```

```r
## Calculating P(X >= 8) using pbinom
pbinom(q = 7, size = 10, prob = 0.7, lower.tail = FALSE)
```

```
## [1] 0.3827828
```

```r
## Calculating P(X <= 1) using dbinom
dbinom(x = 0, size = 10, prob = 0.7) +
  dbinom(x = 1, size = 10, prob = 0.7)
```

```
## [1] 0.0001436859
```

```r
## Calculating P(X <= 1) using pbinom
pbinom(q = 1, size = 10, prob = 0.7, lower.tail = TRUE)
```

```
## [1] 0.0001436859
```

Conclusion: the threshold value given to `pbinom` is included in the "lower tail" (but not the upper tail)

Practice

Data collected by the Substance Abuse and Mental Health Services Administration (SAMSHA) suggests that 69.8% of 18-20 year olds consume an alcohol beverage in any given year. For the questions below, consider a random sample of $n = 50$ individuals aged 18-20.

1) Explain why the binomial distribution is an appropriate model for the number of individuals in the sample that consume alcohol (in the past year).
2) Using R, find the probability that *exactly* 40 individuals in the sample had consumed alcohol (in the past year).
3) Using R, find the probability that *at least* 40 individuals in the sample had consumed alcohol (in the past year).

Practice (solution)

1) The binomial distribution is appropriate because the sampling of each individual is independent (approximately), and the observed outcome is binary with a fixed probability of success.
2) $P(X = 40) = 0.0368$ (use `dbinom`)
3) $P(X \geq 40) = 0.0745$ (use `pbinom`)

Summary

- Random variables are used to represent unknown numeric outcomes of a random process
  - The *expected value* and *standard deviation* are important aspects of a random variable to consider when making decisions
- The *Normal distribution* is a common probability model for continuous random variables
  - $E(X) = \mu$ and $SD(X) = \sigma$
- The *binomial distribution* is a common probability model for certain discrete random variables (those representing the "successes" across *n* independent Bernoulli trials)
  - $E(X) = n * p$ and $SD(X) = \sqrt{n * p * (1 - p)}$