Chapter 4. Probability models

"probability is the only tool we have to quantify uncertainty."

- A **probability model** is an assignment of probabilities to possible outcomes.

- We don't observe these probabilities. We observe a particular dataset.

- If we treat the dataset as an outcome of a probability model, we can answer questions such as,

  "*If there really is no association between unemployment and life expectancy, what is the probability we would see a least squares linear model coefficient as large as the one we actually observed, due to random fluctuations in the data?*"

- Here, we are particularly interested in developing a probability model for the linear model.

- First, we need some basic tools for probability models: random variables, the normal distribution, mean, variance and standard deviation.

Random variables and events

- A **random variable** X is a random number with probabilities assigned to outcomes.

  Example: Let X be a roll of a fair die. A natural probability model is to assign probability of $1/6$ to each of the possible outcomes $1, 2, 3, 4, 5, 6$.

- An **event** is a set of possible outcomes.

  Example: For a die, $E = \{X \geq 4\} = \{4, 5, 6\}$ is the event that the die shows 4 or more.

- We can assign probabilities to events just like to outcomes.

  Example: For a die, $\mathrm{P}(E) = \mathrm{P}(X \geq 4) = 3/6 = 1/2$.

Question 4.1. If an experiment can be repeated many times (like rolling a die) how can you check whether the probability model is correct?

The average of many outcomes should match the probability. This is the "law of large numbers".

For the die, if we roll it long enough we can see if it is fair. If we roll it long enough, we'll detect imperfections.

.

Notation for combining events

- $\{E \text{ or } F\}$ is the event that either E or F or both happens.
- Since E and F are sets, we can write this as a union, $\{E \text{ or } F\} = E \cup F$
- $\{E \text{ and } F\}$ is the event that both E and F happen.
- We can write this as an intersection,

$$\{E \text{ and } F\} = E \cap F$$

- Usually, we prefer "and/or" to "intersection/union".

Question 4.2. When does this formal use of "and" and "or" agree with usual English usage? When does it disagree?

English sometimes uses exclusive or, E or F but not both.
"Would you like a hotdog or a hamburger"
Use of "and" usually matches the technical meaning.

The basic rules of probability

we can use percent, but usually don't for technical work.

1. Probabilities are numbers between 0 (impossible) and 1 (certain).
2. Let $\mathcal{S}$ be the set of all possible outcomes. Then, $\mathrm{P}(\mathcal{S}) = 1$.
   Example: For a die, $\mathrm{P}(X \in \{1, 2, 3, 4, 5, 6\}) = 1$.
3. Events E and F are called **mutually exclusive** if they cannot happen at the same time. In other words, their intersection is the empty set. In this case, the addition rule for probability.

X is in {1, ... 6}

X is an element of {1, ... 6}

$$\mathrm{P}(E \text{ or } F) = \mathrm{P}(E) + \mathrm{P}(F).$$

Question 4.3. You roll a red die and a blue die. Let
$E = \{\text{red die shows } 1\}$, $F = \{\text{blue die shows } 1\}$, $G = \{\text{red die shows } 6\}$.
(a) Are E and F mutually exclusive? (b) How about E and G? (c) How about F and G?

(a) Not mutually exclusive: red & blue die are independent.
It is possible for both dice to show 1.

(b) Yes. E & G cannot happen at the same time.

(c) Not. Same as (a).

Discrete random variables

- X is a **discrete random variable** if we can list all its possible outcomes. Let's call them $x_1, x_2, \ldots$.
- A discrete random variable is specified by probability that the random variable takes each possible outcome,

Upper Case.

lower case

$$p_i = \mathrm{P}[X = x_i], \text{ for } i = 1, 2, 3, \ldots$$

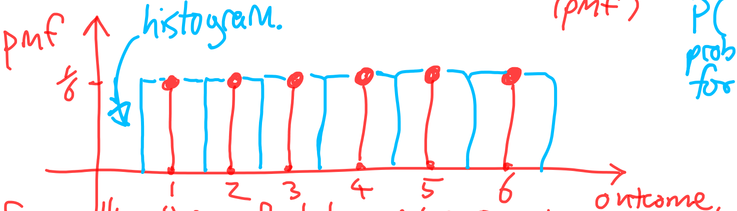
$\{X = x_i\}$ is an event.

- It can be helpful to plot a graph of p_i against x_i.
- This graph is called the **probability mass function**.

Question 4.4. Sketch the probability mass function for a fair die.

(pmf)

P( ) finds probabilities for events.



Formally, the pmf takes value 0 outside possible values.
We sometimes draw a histogram with unit bin widths

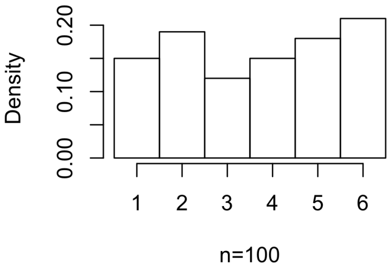
Simulating the law of large numbers

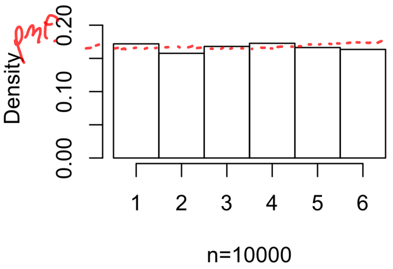
- The "law of large numbers" says that the proportion of each outcome i in a large number of draws of a discrete random variable approaches p_i.
- We can test this by simulation, using the `replicate()` command.

Worked example 4.1. In R, a random draw with replacement from $\{1, 2, 3, 4, 5, 6\}$ can be obtained by `sample(1:6,size=1)` This is equivalent to one roll of a fair die.

?hist

```r
hist(replicate(n=100,sample(1:6,size=1) ),
  main="",prob=TRUE,breaks=0.5:6.5,xlab="n=100",ylim=c(0,0.21))
```





Continuous random variables: the normal distribution

- A **continuous random variable** is one which can take any value in an interval of the real numbers.

  Example: physical quantities such as time and speed are not limited to a discrete set of possible values.

- We will often see the **normal distribution**.

- Let's look at **normal random variables** simulated by R using `rnorm()`.

?rnorm

```
rnorm(n=10,mean=20,sd=5)

##  [1] 14.01141 26.18597 17.18972 21.12226 15.20211 30.34660
##  [7] 19.94277 22.05179 27.73804 25.94906
```

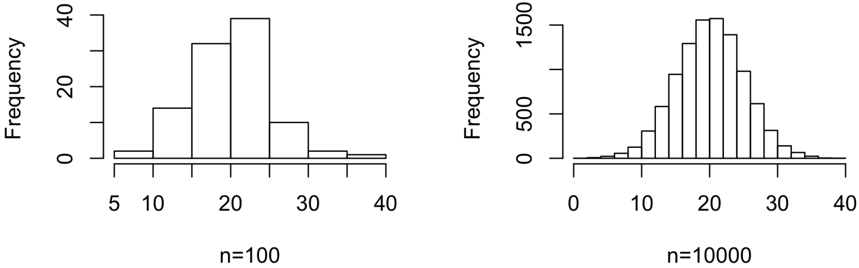
- The arguments `mean=20,sd=5` of `rnorm()` are the **parameters** of the normal distribution.

- A normal random variable can take any numeric value: it is continuous.

- Values are centered on the mean and are usually less than twice the standard deviation (sd) from the mean.

A histogram of normal distribution simulations

```r
hist(rnorm(n=100,mean=20,sd=5),
  main="",xlab="n=100")
```



- Large samples from the normal distribution follow a **bell curve** histogram.
- From smaller samples, this is harder to see.

Finding probabilities for a continuous random variable

- A continuous random variable X has a **probability density function** $f(x)$ which is integrated to find the probability that X falls in any interval: [in this course, we want to know that probabilities can be represented as integrals, but we won't be doing integral calculus].

$$P(a < X < b) = \int_a^b f(x)\,dx$$

- Write $X \sim \text{normal}(\mu, \sigma)$ to mean X is a normal random variable with mean μ and sd σ. The probability density function of X is

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

and so

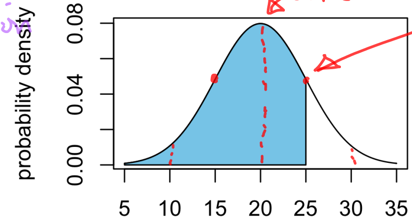
$$P(a < X < b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}\,dx$$

- This integral has no closed form solution.
- Fortunately, R provides `pnorm()` and `qnorm()` that let us work with probabilities for the normal distribution numerically.

Calculating probabilities for the normal distribution

- pnorm() finds the **left tail** of the normal distribution.

f(x) = x^2 is the same function as f(z) = z^2. x & z are dummy variables.



center

inflection point (+/- one SD from the mean)

x is in the limit, so I am using y as the dummy variable

Example: `pnorm(25,mean=20,sd=5)` computes the shaded area above.

- We don't have to do calculus with the normal integral, but we do use the relationship between the curve, area under the curve, and probability. And we must know how to compute these things in R.

- For $X \sim \text{normal}(\mu, \sigma)$,

R code

probability statement

integral

$$\boxed{\texttt{pnorm(x,mu,sigma)} = \mathrm{P}(X \le x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2}\, dy}$$

Finding probabilities that are not a left tail

Question 4.5. Let $X \sim \mathrm{normal}(10, 10)$. Sketch a shaded area under a curve giving $\mathrm{P}(0 \leq X \leq 10)$. Write this probability as an integral and as R code.

Hint: To use `pnorm()`, think of the shaded area as the difference of two left tails.

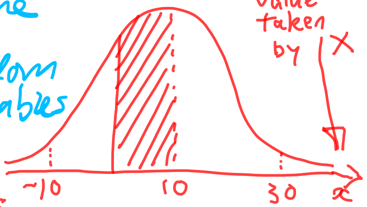
< and ≤ have the same probability for continuous random variables

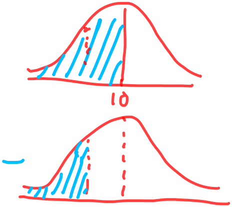
X x

$$P(0 \leq X \leq 10) = P(0 < X < 10)$$

$$= \int_0^{10} \frac{1}{\sqrt{2\pi \cdot 10^2}} e^{-\frac{(x-10)^2}{2 \cdot 100}} dx$$

```
= pnorm(10, mean=10, sd=10)
   - pnorm(0, mean=10, sd=10)
```

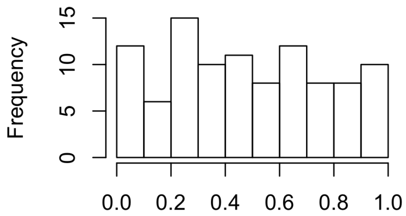




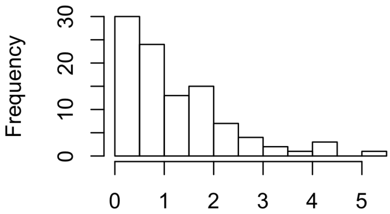
Other continuous distributions

- There are many continuous random variables that are not normally distributed.
- For example, we could explore the **uniform** or **exponential** distributions.

```r
hist(runif(100))
```



```r
hist(rexp(100))
```



- Normal random variables are the only ones we work with in this class.
- Let's investigate why the normal distribution is so important.

Sums and averages follow the normal distribution

- When we sum many random quantities, the sum often follows a normal distribution even if each term in the sum is not normally distributed.
- This property is called **the central limit theorem**.
- It is an empirical fact as well as a mathematical theorem!

A crash would lead to an outlier: a big effect hitting all companies simultaneously.

- Averaging is multiplying the sum by a constant ($1/n$). A bell curve is still a bell curve when we rescale by multiplication.

Question 4.6. Would you expect a histogram of student heights to follow a normal curve? Why? Why not?

"The heights should all fall around their average". Perhaps, we would expect a normal curve to hold better if we break down by gender. Height is the sum of many small genetic & enviromental effects, so might look normal. If a disease/mutation leads to a large effect, this might give outliers

Question 4.7. Would you expect a histogram of the daily change in the Dow Jones stock market index to follow a normal curve? Why? Why not?

- Standard measures of investment risk assume a normal model.
- A normal model is a fair approximation, but tails are too long. A CLT might apply to the sum of many stocks.

Demonstrating the central limit theorem with dice

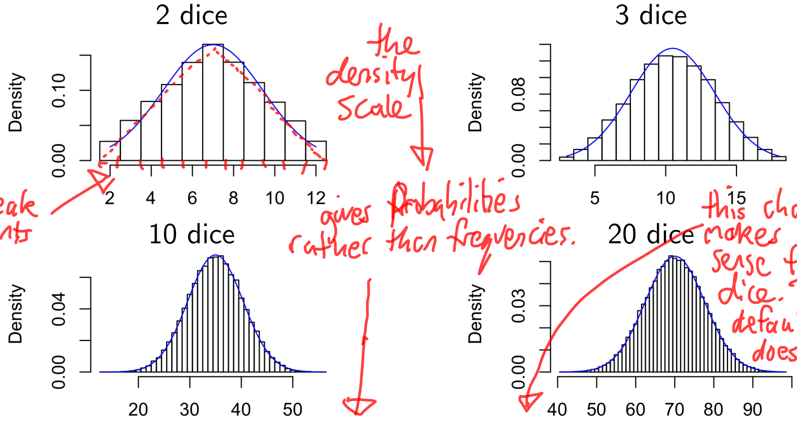
- `sample(1:6,2,replace=TRUE)` simulates the outcome of rolling two dice.
- `sum(sample(1:6,2,replace=TRUE))` simulates the sum of two dice.
- `replicate()` lets us see what happens if we do this many times

```r
dice2 <- replicate(50000,sum(sample(1:6,2,replace=TRUE)))
dice3 <- replicate(50000,sum(sample(1:6,3,replace=TRUE)))
dice10 <- replicate(50000,sum(sample(1:6,10,replace=TRUE)))
dice20 <- replicate(50000,sum(sample(1:6,20,replace=TRUE)))
```

- Guess how many dice you have to add up before the histogram looks normal?

?hist

```
hist(dice2,prob=TRUE,breaks=(min(dice2)-0.5):(max(dice2)+0.5))
normal.x <- seq(from=min(dice2),to=max(dice2),length=100)
normal.y <- dnorm(normal.x,mean=mean(dice2),sd=sd(dice2))
lines(normal.x,normal.y,col="blue")
```



break points

the density scale

gives Probabilities rather than frequencies.

this choice makes sense for dice. The default doesn't.

Question 4.8. Why do we use `prob` and `breaks` arguments to `hist()`?

More normal approximation situations

Question 4.9. Would you expect detrended data on life expectancy at birth to follow a normal distribution? Explain.

Once it is detrended there will still be variation

Mortality is the sum of many chance events happening to a large number of people, so the central limit theorem should be a good approximation.

Large events like wars, or pandemics would be outliers

Maybe there is skew: life expectancy can drop more dramatically than it can rise.

Question 4.10. Consider the mice weight data for HW1 with mice randomized to two treatments: a high fat diet and a usual lab diet. (a) Would you expect the weights of mice in each treatment group to follow a normal distribution? (b) If the experiment were replicated ten times, and an average weight calculated for each of these ten replications, would you expect the ten averages to follow a normal distribution? Are your answers different for (a) and (b)?

(a). Maybe, since weight is a result of many genetic & enviromental factors. (b) Certainly, since sums & averages do follow a central limit theorem, as we saw for dice.

The sample mean and the expectation of a random variable

- The **sample mean** or **average** of $\mathbf{y} = (y_1, \ldots, y_n)$ is

$$\bar{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

similar equations.

- The **expected value** of a random variable X taking possible values $x_1, x_2, \ldots$ with probabilities $p_1, p_2, \ldots$ is

$$\mathrm{E}[X] = \sum_{i=1}^{\infty} x_i p_i$$

- If we have many draws of X, the sample proportion taking value x_i becomes close to p_i and so the sample mean becomes close to the expected value.

- If X is a continuous random variable with density $f(x)$ the sum for expected value becomes an integral,

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx$$

The sample mean and expectation using R

- We've already seen the `mean()` function which computes the sample mean of a numeric vector.
- One way to compute the expected value of a random variable is to take the sample mean of many realizations of the random variable.

```
x <- rnorm(10000,mean=20,sd=5)
mean(x)
## [1] 19.96731
```

expected value, also called population mean

Sample mean

- The same calculation can be done using `replicate()`:

```
y <- replicate(n=10000,rnorm(1,mean=20,sd=5))
mean(y)
## [1] 19.96731
```

apart from some error because $10\,000 \neq \infty$

- We can guess (correctly!) that the expected value of a normal random variable matches its `mean` parameter.
- The expected value of a random variable is sometimes called its mean. We prefer "expected value" to distinguish from the "sample mean."

The expected value is not necessarily a possible value

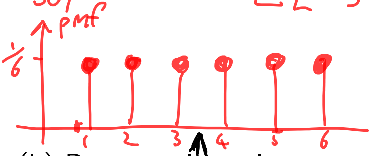
Question 4.11. Compute the expected value of a roll of a fair six-sided die.

(a) By using the definition $\mathrm{E}[X] = \sum_{i=1}^{\infty} x_i p_i$.

For a fair die, possible values are 1, 2, ... 6 each with $p_i = \frac{1}{6}$

So, $E[X] = \sum_{i=1}^{6} i \cdot \frac{1}{6} = \frac{1}{6}[1+2+3+4+5+6] = \frac{21}{6}$

$= 3\frac{1}{2}$.



expected value is the center of mass of a graph of the pmf.
probability mass function.

(b) By averaging a large number of simulated dice using R. Write some R code that is a starting point for testing and debugging.

```
x <- sample(1:6, size = 10000, replace = T)
mean(x)
```

The sample variance and the variance of a random variable

- The **sample variance** of $\mathbf{y} = (y_1, \ldots, y_n)$ is

$$\mathrm{var}(\mathbf{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{\mathbf{y}})^2$$

- The **variance** of a random variable X is defined in terms of the expected value as

$$\mathrm{Var}(X) = \mathrm{E}\Big[\big(\,X - \mathrm{E}[X]\,\big)^2\Big] = E[X^2] - (E[X])^2$$

- If X is a random variable, then so is $Y = \big(\,X - \mathrm{E}[X]\,\big)^2$. Each possible outcome of X (say, $X = x$) matches an outcome $Y = \big(\,x - \underline{\mathrm{E}[X]}\,\big)^2$.

- Collections of numbers have a *sample variance* computed by var (not capitalized). Random variables have a *variance* computed with Var (capitalized).

E[X] is a constant number, not a random variable.

- People do not always make this distinction, but we will try to.

- In R, `var()` calculates the sample variance.

Standard deviation

- The **sample standard deviation** of $\mathbf{y} = (y_1, \ldots, y_n)$ is the square root of the sample variance.

$$\mathrm{sd}(\mathbf{y}) = \sqrt{\mathrm{var}(\mathbf{y})}$$

- The **standard deviation** of a random variable X is the square root of its variance.

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$$

- In R, `sd()` computes the sample standard deviation.
- We can compute $\mathrm{SD}(X)$ as the sample standard deviation of a large number of draws of the random variable X.

set.seed(1284)

```
x <- rnorm(10000,mean=20,sd=5)
sd(x)
```

```
## [1] 5.061782
```

- As we might anticipate, this confirms that the `sd` parameter of the normal distribution matches its standard deviation.

Expectation, variance and standard deviation of $mX + c$

- Let X be a random variable and let $Y = mX + c$.
- Y is also a random variable. If X takes value x, Y takes value $mx + c$.
- Expectation is **linear**, meaning

$$\mathrm{E}[mX + c] = m\mathrm{E}[X] + c$$

- Variance doesn't depend on c. It is a measure of **spread**. Adding a constant shifts the center of a distribution but doesn't change the spread.
- Variance is quadratic in m.

$$\mathrm{Var}(mX + c) = m^2\,\mathrm{Var}(X)$$

- Standard deviation therefore scales with m.

$$\mathrm{SD}(mX + c) = m\,\mathrm{SD}(X)$$

- SD scales nicely. Var can be easier to use for calculations.

The standard normal distribution

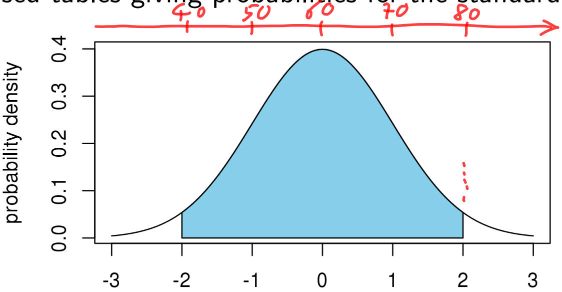
- Let $Z \sim \mathrm{normal}(0,1)$, so $\mathrm{E}[Z] = 0$ and $\mathrm{SD}(Z) = 1$.
- Z is called a **standard normal random variable**.
- Let $X = \mu + \sigma Z$.
- We use linearity of expectation and the scaling property of standard deviation to calculate

$$\underline{\mathrm{E}[X] = \mu + \sigma\,\mathrm{E}[Z] = \mu, \qquad\qquad \mathrm{SD}(X) = \sigma\,\mathrm{SD}(Z) = \sigma}$$

- A bell curve is still a bell curve if you shift or rescale it, so X also follows a normal distribution.
- Therefore, $X \sim \mathrm{normal}(\mu, \sigma)$.
- We can work the other way around: if $X \sim \mathrm{normal}(\mu, \sigma)$ then $Z = (X - \mu)/\sigma$ has $Z \sim \mathrm{normal}(0,1)$.

Standardizing into standard units

- After subtracting the mean and dividing by the standard deviation, any normal random variable follows the standard normal distribution.
- This is called **standardizing**. We say we are working in **standard units**.
- Calculating in standard units was essential before computers: people used tables giving probabilities for the standard normal distribution.



Mean:60,
SD=10

```
pnorm(2)-pnorm(-2)
```

```
## [1] 0.9544997
```

=pnorm(80, Mean=60, sd=10)

- Thinking in standard units remains helpful. For example, as shown by the shaded area above, 95% of normal random variables are within 2 SD units of their mean.

* The general US population is likely not a good representation of UM

Question 4.12. Wikipedia:List_of_average_human_height_worldwide says the average height of an American male aged 20-29 is 176.4 cm ($5'\,9.5''$). Suppose the standard deviation of height is $2.5''$. The average height of an NBA basketball player is about $6'\,7.5''$.

Meaning as tall or taller!

(a) Write $6'\,7.5''$ in standard units for this population.

$6'7.5'' - 5'9.5'' = 10'' = 4 \times 2.5''$

So, $6'7.5''$ is 4 standard units above average.

$$= \int_{6'7.5''}^{\infty} \frac{1}{\sqrt{2\pi \cdot 2.5^2}} e^{-\frac{(x-5'9.5'')^2}{2\times 2.5^2}}\, dx$$

(b) Estimate what percentage of male UM students are as tall as an average NBA player. Explain your assumptions. Sketch the answer as an area under the standard normal curve. Write this probability as an integral and show how to compute it via a call to pnorm().

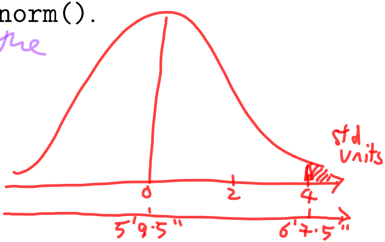
Let X be a random variable modeling the height of a UM student

Suppose $X \sim \text{normal}(5'9.5'', 2.5'')$.

$$P[X > 6'7.5''] = \int_4^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

[3 in a million]

$= 0.0003\,\%$


std units
natural units
5'9.5''
6'7.5''