Final

2024-08-11

Contents

Contents

Classical Probability

Theorems

De Morgan

$$\overline{A \vee B} = \overline{A} \wedge \overline{B}$$
$$\overline{A \wedge B} = \overline{A} \vee \overline{B}$$

Counting

Counting rule for k sets, each with n_i possible outcomes. If pick one outcome from each set, these will be $n_1 \times n_2 \times n_k$ ways to do so.

With replacement

n^k

without replacement

$n \times (n-1) \times (n-2) \times \cdots \times (n-k+1)$

Permutation

For n elements, the permutation of taking k of them.

$P_k^n = \frac{n!}{(n-k)!} = n \times (n-1) \times (n-2) \times \cdots \times (n-k+1)$

Combination

$\binom{n}{k} = \frac{P_k^n}{k!} = \frac{n!}{k!(n-k)!}$

Conditional Probabilities

$P(A|B) = \frac{P(AB)}{P(B)}$

$P(A|B) = \frac{P(AB)}{P(B)} \implies P(AB) = P(A|B)P(B)$

$P(A_1 A_2 ... A_n) = P(A_1)P(A_2|A_1)...P(A_n|A_1...A_{n-1})$

Example 1

Question

A shady used car dealer has 30 cars, and 10 of them are "lemons" (that is, mechanically faulty used cars), but you don't know which cars they are. If you buy 3 cars, what is the probability that you will get at least one lemon?

Solution

Event A is defined as 'you will get at least one lemon among the 3 cars you purchase.' It is easier to consider the converse: getting no lemons among the 3 cars you purchase. To calculate this, you can imagine choosing 3 cars from 20 normal cars, while the probability space consists of choosing 3 cars from all 30 cars.

$$P(A) = 1 - P(\overline{A}) = 1 - \frac{\binom{20}{3}}{\binom{30}{3}} \approx 0.719$$

Conclusion: The probability that you will get at least one lemon when you buy 3 cars is approximately 71.9%.

Example 2

Question

We throw two dice (each with the usual 6 sides, numbered 1-6). What is the probability that the sum of the two numbers is odd? What is the probability that the sum of the two numbers is less than 7?

Solution

It is clear that:

odd + even = odd
odd + odd = even
even + even = even

so the numbers of a 1-6 dice could be split to 2 sets.

$$Set_{odd} = \{1, 3, 5\}$$

$$Set_{even} = \{2, 4, 6\}$$

Event A is 'the sum of the two numbers is odd.' This event can be split into two steps. Step 1 is to choose one odd number from the set of odd numbers, and Step 2 is to choose one even number from the set of even numbers. Interestingly, we could choose from the odd set first or the even set first, and it does not affect the sum. Meanwhile, the probability space is clearly the random selection of 2 numbers from 1 to 6.

$$P(A) = \frac{\binom{3}{1}\binom{3}{1}2!}{\binom{6}{1} \cdot \binom{6}{1}} = \frac{1}{2}$$

Conclusion: The probability that the sum of the two numbers is odd is 50%.

Question

What is the probability that the sum of the two numbers is less than 7, given that it is odd?

Solution

The event 'the sum of the two numbers is less than 7' can be divided into subevents, as follows:

$$P(x + y < 7) = P(y < 6, x = 1) + P(y < 5, x = 2) + P(y < 4, x = 3) + P(y < 3, x = 4) + P(y < 2, x = 5) + P(y < 1, x = 6)$$

Conclusion: Through counting the case, the probability that the sum of the two numbers is less than 7 is $\frac{15}{36} = \frac{5}{12}$.

Question

Are these two events independent?

Solution

Event A is "the sum of the two numbers is less than 7".

Event B is "the sum of the two numbers is odd".

So $P(A|B) = \frac{P(AB)}{P(B)}$.

Given that there are 6 cases where the sum of the two numbers is less than 7 and odd, and there are 18 cases where the sum of the two numbers is odd, the probability that the sum of the two numbers is less than 7, given that it is odd, is $\frac{6}{18} = \frac{1}{3}$.

Because $P(A|\overline{B}) = \frac{P(A\overline{B})}{P(\overline{B})} = \frac{9}{18} = \frac{1}{2}$.

Clearly, $P(A|B) \neq P(A|\overline{B}) \neq P(A)$.

Conclusion: The events 'the sum of the two numbers is less than 7' and 'the sum of the two numbers is odd' are **not independent.**

Example 3

Question

Draw a card at random. What is P(card is Ace or Spade)?

Solution

- Key fact: "ace" and "spade" are not mutually exclusive outcomes.
- Therefore can't just add P(ace) + P(spade)! We'll double-count the ace of spades.
- Of the 52 cards:

4 are aces: P(Ace) = 4/52

13 are spades: P(Spade) = 13/52

1 is both an ace and a spade: P(Ace, Spade) = 1/52

P(Ace or Spade) = 4/52 + 13/52 - 1/52 = 16/52

Example 4

Question

Suppose that, among all UT freshmen:

45% take a class with a QR flag in their first semester.

25% take a class with a Writing flag in their first semester.

15% take both a QR class and a Writing class in their first semester.

- Self-test! What is the probability that a randomly selected UT freshman will have taken either a QR or a writing class in their first semester?

Solution

P(QR or W) = P(QR) + P(W) - P(QR, W) (Addition rule)

P(QR or W) = 0.45 + 0.25 - 0.15 = 0.55

Question

What is the probability that a randomly selected UT freshman will have taken neither a QR nor a writing class in their first semester?

Solution

P(no QR, no W) = 1 - P(QR or W) (Negation rule)

P(QR or W) = P(QR) + P(W) - P(QR, W) (Addition rule)

P(QR or W) = 0.45 + 0.25 - 0.15 = 0.55

So P(no QR, no W) = 1 - 0.55 = 0.45

Example 5

Question

- Someone deals you a five-card poker hand from a 52-card deck.
- What is the probability of a flush (all five cards the same suit)?
- Note: this is a very historically accurate illustration of probability, given its origins among bored French aristocrats!

Solution

- Our sample space has 2,598,960 possible hands, each one equally likely.
- How many possible flushes are there? Let's start with hearts:

There are 13 hearts. To make a flush with hearts, you need any 5 of these 13 cards. Thus there are $\binom{13}{5} = 1287$ possible flushes with hearts.

- The same argument works for all four suits, so there are $4 \times 1287 = 5148$ flushes.

Thus

$P(flush) = \frac{|A|}{|\Omega|} = \frac{5148}{2598960} = 0.00198079$

Total Probability and Bayes' Rule

Theorems

Total Probability

$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$

Bayes' Rule

$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}$

Independent

$P(AB) = P(A)P(B)$

Others

$P(A \vee B) = P(A) + P(B) - P(AB)$

$P(A \vee B \vee C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$

$P(A - B) = P(A) - P(AB)$

$P(A\overline{B}) = P(A) - P(AB)$

$P(A) = P(AB) + P(A\overline{B})$

Example 1

There are 100 coins. 99 of them are common, and 1 of them is exotic: head and head.

A: exotic coin B: all 5 flips are head

Question

What is the probability of B?

Solution

$P(A) = \frac{1}{100}$

$P(B|A) = 1^5$

Total Probability

$P(B) = P(B|A) \times P(A) + P(B|\overline{A}) \times P(\overline{A}) = 1^5 \times \frac{1}{100} + (\frac{1}{2})^5 \times \frac{99}{100}$

Question

What is probability of the coin is exotic if B happened?

Solution

Bayes' theorem

$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$

Example 2

Question

There is a fair coin: head and tail, and toss 1000 times. What is the probability that the number of head is even?

Solution

P(even H for 1000)
$= P$(even H for 999) $\times P$(T at 1000) $+ P$(odd H for 999) $\times P$(H at 1000)
$= \frac{1}{2} \times (P$(even H for 999) $+ P$(odd H for 999)$) = \frac{1}{2}$

Example 3

Question

Visitors to your website are asked to answer a single survey question before they get access to the content on the page. Among all of the users, there are two categories: Random Clicker (RC), and Truthful Clicker (TC). There are two possible answers to the survey: yes and no. Random clickers would click either one with equal probability. You are also giving the information that the expected fraction of random clickers is 0.3. After a trial period, you get the following survey results: 65% said Yes and 35% said No. What fraction of people who are truthful clickers answered yes? Hint: use the rule of total probability.

Solution

Let's denote the event "Random clicker" as RC, the event "Truth clicker" as TC, the event "answer yes" as Y, the event "answer no" as N.

$P(Y) = P(Y|RC) \cdot P(RC) + P(Y|TC) \cdot P(TC)$

Given the expected fraction of random clickers is 0.3, it means that $P(RC) = 0.3$, so $P(TC) = 1 - P(\overline{TC}) = 1 - P(RC) = 1 - 0.3 = 0.7$.

Besides, because random clickers would click either one with equal probability, which means $P(Y|RC) = P(N|RC) = 0.5$ and the following survey results: 65% said Yes and 35% said No.

Therefore, $P(Y) = P(Y|RC) \cdot P(RC) + P(Y|TC) \cdot P(TC) = 0.5 \cdot 0.3 + P(Y|TC) \cdot 0.7 = 0.65$, it is clear that $P(Y|TC) = \frac{P(Y) - P(Y|RC) \cdot P(RC)}{P(TC)} = \frac{0.65 - 0.5 \cdot 0.3}{0.7} \approx 0.714$.

Conclusion: The fraction of people who are truthful clickers and answered "Yes" is approximately 71.4%.

Example 4

Question

Imagine a medical test for a disease with the following two attributes:

- The sensitivity is about 0.993. That is, if someone has the disease, there is a probability of 0.993 that they will test positive.

- The specificity is about 0.9999. This means that if someone doesn't have the disease, there is probability of 0.9999 that they will test negative.

- In the general population, incidence of the disease is reasonably rare: about 0.0025% of all people have it (or 0.000025 as a decimal probability).

Suppose someone tests positive. What is the probability that they have the disease?

Solution

Let's denote the event "someone has the disease" as D, the event "test positive" as TP.

Because someone has the disease, there is a probability of 0.993 that they will test positive, it is clear that $P(TP|D) = 0.993$.

Additionally, if someone does not have the disease, there is a 0.9999 probability that they will test negative, which means $P(\overline{TP}|\overline{D}) = 0.9999$ and $P(TP|\overline{D}) = 1 - P(\overline{TP}|\overline{D}) = 0.0001$.

In the general population, incidence of the disease is reasonably rare: about 0.0025% of all people have it, which means $P(D) = 0.000025$ and $P(\overline{D}) = 1 - P(\overline{D}) = 0.999975$.

$P(TP) = P(TP|D) \cdot P(D) + P(TP|\overline{D}) \cdot P(\overline{D}) = 0.993 \cdot 0.000025 + 0.0001 \cdot 0.999975 = 0.0001248225$

According to Bayes' theorem, $P(D|TP) = \frac{P(TP|D) \cdot P(D)}{P(TP)} = \frac{0.993 \cdot 0.000025}{0.0001248225} \approx 0.1989$.

Conclusion: The probability that someone has the disease given that they tested positive is approximately 19.89%.

Example 5

Question

If an aircraft is present in a certain area, a radar correctly registers its presence with probability 0.99. If it is not present, the radar falsely registers an aircraft presence with probability 0.10. Suppose that on average across all days, an aircraft is present with probability 0.05. Let the events A and R be defined as follows: A = an aircraft is present, R = the radar registers an aircraft presence. What is P (A | R), the conditional probability that an aircraft is present, given that the radar registers an aircraft presence?

Solution

Let's denote the event "an aircraft is present in a certain area" as A, the event "a radar correctly registers its presence" as R.

$P(R|A) = 0.99$
$P(R|\overline{A}) = 0.10$
$P(A) = 0.05$ and $P(\overline{A}) = 1 - P(A) = 0.95$

Because $P(R) = P(R|A) \cdot P(A) + P(R|\overline{A}) \cdot P(\overline{A}) = 0.99 \cdot 0.05 + 0.10 \cdot 0.95 = 0.1445$.

According to Bayes' theorem, $P(A|R) = \frac{P(R|A) \cdot P(A)}{P(R)} = \frac{0.99 \cdot 0.05}{0.1445} \approx 0.3426$

Conclusion: The conditional probability that an aircraft is present given that the radar registers an aircraft presence is approximately 34.26%.

Example 6

Question

A 45-year old woman goes to the doctor for a routine screening mammogram. (No family history or clinical symptoms.)

- The mammogram comes back positive.
- What is P(cancer | positive test)?

Some Facts

- For every 1,000 45-year-old women who participate in a routine screening mammogram, about 10 of them actually have breast cancer, and 990 don't.
- Out of 10 cancer cases, we would expect a screening mammogram to correctly detect about 8 of them, on average.
- If a woman does not have breast cancer, the mammogram has a small chance of resulting in a positive test anyway. Out of 100 such cases, it will wrongly flag about 10 of them, on average.

Solution

- For every 1,000 45-year-old women who participate in a routine screening mammogram, about 10 of them actually have breast cancer, and 990 don't.

P(C) = 10/1000 = 0.01 and P(not C) = 990/1000 = 0.99

- Out of 10 cancer cases, we would expect a screening mammogram to correctly detect about 8 of them, on average.

P(T | C) = 0.8

- If a woman does not have breast cancer, the mammogram has a small chance of resulting in a positive test anyway. Out of 100 such cases, it will wrongly flag about 10 of them, on average.

P(T | not C) = 0.1

So,

$P(C|T) = \frac{P(C)P(T|C)}{P(T|C)P(C)+P(T|\overline{C})P(\overline{C})} \approx 0.075$

Distribution

Theorems

Bionomial Distribution

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

$E(X) = np$

$var(X) = np(1 - p)$

Poisson Distribution

$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$

$E(X) = \lambda$

$var(X) = \lambda$

Normal Distribution

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

$\Phi(x) = \int_{-\infty}^{x} \phi(x) dx$

$E(X) = \mu$

$var(X) = \sigma^2$

$P(X \leq x) = P(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}) = P(Y \leq \frac{x-\mu}{\sigma}) = \Phi(\frac{x-\mu}{\sigma})$

Exponential Distribution

For $x > 0$,

$f(x) = \lambda e^{-\lambda x}$

$F(x) = 1 - e^{-\lambda x}$

$E(X) = \lambda^{-1}$

$var(X) = \frac{1}{\lambda^2}$

Uniform Distribution

$f(x) = \frac{1}{b-a}, x \in [a, b]$

$F(x) = \begin{cases} 0, x \leq a \\ \frac{x-a}{b-a}, a < x \leq b \\ 1, x > b \end{cases}$

$E(X) = \frac{a+b}{2}$

$var(X) = \frac{(b-a)^2}{12}$

Example 1

10 students are walking on the speedway. Each student has a probability of 0.8 to receive a flyer.

Question

What is the probability that 3 of them receive flyer?

Solution

$\binom{10}{3} \times 0.8^3 \times (1 - 0.8)^{10-3}$

Question

What is the probability that at least one of them receive a flyer?

Solution

$P(\text{at least on receive}) = 1 - P(\text{none receive}) = 1 - (1 - 0.8)^{10}$

Example 2

Question

How to prove for $X \sim \text{Poisson}(\lambda)$, the sum of PMF $= f(k) = e^{-\lambda}\frac{\lambda^k}{k!}$, for $k = 0, 1, 2, ...$ is 1?

Solution

$\sum f(k) = \sum e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum \frac{\lambda^k}{k!}$

Given Taylor's Formula:

$\because e^x = \frac{x^0}{0!} + \frac{x^1}{1!} + \frac{x^2}{2!} + \cdots$

$\therefore e^{\lambda} = \frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!} + \cdots$

$\therefore \sum f(k) = e^{-\lambda} \times e^{\lambda} = 1$

PMF, PDF and CDF

Theorems

PMF

Conditional PMFs $p_{X|A}(x) = P(X = x|A)$

For joint:

$p_{X|Y}(x, y) = p_Y(y)p_{X|Y}(x|y)$

For marginal:

$p_X(x) = \sum_y p_Y(y)p_{X|Y}(x|y)$

CDF

For discrete:

$F_X(x) = P(X \leq x)$

For continuous:

$F(x) = \int_{-\infty}^{x} f(x)dx$

PDF

For discrete:

$p_X(k) = P(X \leq k) - P(X < k)$

For continuous:

$f(x) = F'(x)$

Example 1

Monotone Transformation: $Y = g(X)$ and X has a density function $f_X(x)$, then the density function of Y is:

$f_Y(y) = f_X(g^{-1}(y))J(y),$

where $J(y) = |\frac{dg^{-1}(y)}{dy}|$.

Non-monotone Transformation: First, calculate the CDF of Y, that is,

$F_Y(y) = P(g(X) \leq y)$

then derive the density from CDF by differentiating,

$f_Y(y) = \frac{dF_Y(y)}{dy}$

Question

What is the CDF and PDF of $Y = g(X)$?

Solution

$F_Y(y)$
$= P(Y \leq y)$
$= P(g(X) \leq y)$
$= P(X \leq g^{-1}(y))$
$= F_X(g^{-1}(y))$

$f_Y(y)$
$= \frac{dF_Y(y)}{dy}$
$= \frac{dF_X(g^{-1}(y))}{dy}$
$= \frac{F_X(g^{-1}(y))}{dg^{-1}(y)} \times \frac{dg^{-1}(y)}{dy}$, based on chain rule.
$= f_X(g^{-1}(y)) \times \frac{dg^{-1}(y)}{dy}$

Example 2

Question

Log-Normal Distribution If $\log(X) \sim N(\mu, \sigma^2)$, then calculate the PDF and expectation of X.

Solution

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

Let $Z = \log(X)$, then $X = e^Z$

$\therefore f_X(x) = f_Z[\log(x)] \times |\frac{d\log(x)}{dx}|$
$= \frac{1}{\sqrt{2\pi}\sigma} \times \exp\{-\frac{(\log(x)-\mu)^2}{2\sigma^2}\} \times \frac{1}{x}$
for $x > 0$.

Example 3

Question

Consider the random variable X with PDF $f_X(x)$ defined as

$f_X(x) = \begin{cases} \frac{2}{3\theta^2}x, & \text{for } x \in (\theta, 2\theta) \\ 0, & \text{otherwise} \end{cases}$

where $\theta > 0$.

Let $Y = X^2$. What is the PDF for Y?

Solution

$F_Y(y) = Pr(X^2 \leq y) = Pr(-\sqrt{y} \leq X \leq \sqrt{y})$

for $y \geq 0$.

For $y \leq \theta^2$, $F_Y(y) = 0$ and for $y \geq (2\theta)^2$, $F_Y(y) = 1$.

For $\theta^2 < y < 4\theta^2$,

$F_Y(y) = \int_{-\sqrt{y}}^{\sqrt{y}} f_X(x)_{\{\theta \le x \le 2\theta\}} dx = \int_0^{\sqrt{y}} \frac{2}{3\theta^2} x\, dx = \frac{y}{3\theta^2} - \frac{2}{6}$

Thus,

$$f_Y(y) = \begin{cases} \frac{1}{3\theta^2}, & \text{for } y \in (\theta^2, 4\theta^2) \\ 0, & \text{otherwise} \end{cases}$$

Example 4

Question

Define a discrete random variable X as follows:

- X = 2 with probability 1/10
- X = 3 with probability 1/10
- X = 5 with probability 8/10

What are $P(2 < X \le 4.5)$ and $P(2 \le X < 4.5)$?

Solution

Given the definition in the Part A, the CDF of X is:

$$\text{CDF}_X = P(X \le x) = \begin{cases} 0, & \text{for } x < 2 \\ 0.1, & \text{for } 2 \le x < 3 \\ 0.2, & \text{for } 3 \le x < 5 \\ 1, & \text{for } 5 \le x \end{cases}$$

therefore, it is clear that:

$P(2 < X \le 4.5) = P(X \le 4.5) - P(X \le 2) = 0.2 - 0.1 = 0.1$

$P(2 \le X < 4.5) = P(X < 4.5) - P(X < 2) = 0.2 - 0 = 0.2$

Example 5

Question

Let X be a (continuous) uniform random variable on [0,1].

i. Compute $P(X2 \le 0.25)$.

ii. For any number a, compute $P(X2 \le a)$.

iii. From (ii), find the PDF of the random variable $Y = X^2$.

iv. Compute $E(Y)$ and $var(Y)$ directly from the PDF.

Solution

Given the definition in the Part B ($X\ U(0,1)$), the CDF of X is:

$$\text{CDF}_X = P(X \leq x) = \begin{cases} 0, & \text{for } x < 0 \\ x, & \text{for } 0 \leq x \leq 1 \\ 1, & \text{for } x > 1 \end{cases}$$

i. Compute $P(X^2 \leq 0.25)$

$$P(X^2 \leq 0.25) = P(X \leq 0.5) = 0.5$$

ii. For any number a, compute $P(X^2 \leq a)$

$$P(X^2 \leq a) = P(X \leq \sqrt{a}) = \begin{cases} \sqrt{a}, & \text{for } 0 \leq a \leq 1 \\ 1, & \text{for } a > 1 \end{cases}$$

iii. From (ii), find the PDF of the random variable $Y = X^2$

From (ii), we have the CDF of $Y = X^2$:

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(X \leq \sqrt{y}) = \begin{cases} \sqrt{y}, & \text{for } 0 \leq y \leq 1 \\ 1, & \text{for } y > 1 \end{cases}$$

therefore, it is clear that:

$$\text{PDF}_Y = f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{1}{2} y^{\frac{-1}{2}}, \text{ for } 0 < y \leq 1$$

iv. Compute $E(Y)$ and $var(Y)$ directly from the PDF

Given (iii), it is clear that:

$$E(Y) = \int_{-\infty}^{+\infty} y \cdot f_Y(y) dy = \int_0^1 \frac{1}{2} y^{\frac{1}{2}} dy = \frac{1}{3}$$

For $var(Y)$, given the formula $var(Y) = E(Y^2) - (E(Y))^2$, it is necessary to calculate $E(Y^2)$ firstly:

$$E(Y^2) = \int_{-\infty}^{+\infty} y^2 \cdot f_Y(y) dy = \int_0^1 \frac{1}{2} y^{\frac{3}{2}} dy = \frac{1}{5}$$

Therefore, $var(Y) = \frac{1}{5} - (\frac{1}{3})^2 = \frac{4}{45}$

Example 6

Question

Suppose that U is a continuous random variable with a uniform distribution on $[0, 1]$. Now suppose that f is the PDF of some continuous random variable of interest, that F is the corresponding CDF, and assume that F is invertible, so that the function $F^{-1}(u)$ exists and assigns a unique value to each $u \in (0, 1)$. Show that the random variable $X = F^{-1}(U)$ has PDF $f(x)$—that is, that X has the desired PDF. Hint: use results on transformations of random variables.

Solution

$\because U \sim U(0, 1)$

$\therefore f_U(u) = \begin{cases} 1, & \text{for } 0 \le u \le 1 \\ 0, & \text{otherwise} \end{cases}$

$\therefore F_U(u) = \begin{cases} 0, & \text{for } u < 0, \\ u, & \text{for } 0 \le u \le 1 \\ 1, & \text{for } u > 1 \end{cases}$

$\because X = F^{-1}(U)$

$\therefore F_X(x) = P(F^{-1}(U) \le x)$

$\therefore F_X(x) = P(U \le F(x))$

$\therefore F_X(x) = F(x)$

$\therefore f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx} F(x) = f(x)$

Example 7

Question

A non-negative continuous random variable X is said to follow an exponential distribution with parameter $\lambda > 0$ if its PDF is given by $f(x) = \lambda e^{-\lambda x}$ for $x \ge 0$, and is zero otherwise. The parameter λ is usually referred to as the rate. Suppose that $X_1, ..., X_N$ are a set of N independent samples from an exponential distribution with rate λ. Let $Y_N = \max X_1, ..., X_N$ be the maximum value in your sample. Derive the PDF of Y_N for fixed N.

Here are some hints:

- For any y, we know that $\max X_1, ..., X_N \le y$ if and only if $Xi \le y$ for all $i = 1, ...N$.

- The following integral will help:

$\int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}$

Solution

For each standard normal random variable X_i picked from $X_1, \cdots, X_N$ is said to follow $X_i \sim Exp(\lambda)$ for $x \ge 0$, and $f(x) = 0$ otherwise.

So the cumulative distribution function of X is $F_X(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}$.

Given $Y_N = \max\{X_1, \cdots, X_N\}$, it is clear that:

$F_{Y_N}(y) = P(Y \le y) = P(X_1 \le y, \cdots, X_N \le y)$

Because $X_1, \cdots, X_N$ are a set of N independent samples, it implies that:

$F_{Y_N}(y) = P(X_1 \le y) \times \cdots \times P(X_N \le y)$

Then,

$F_{Y_N}(y) = \underbrace{(1 - e^{-\lambda y}) \times \cdots \times (1 - e^{-\lambda x})}_{N} = (1 - e^{-\lambda y})^N$

Therefore, the PDF of Y_N for fixed N is:

$f_{Y_N}(y) = \frac{d}{dy}(1 - e^{-\lambda y})^N = \lambda N(1 - e^{-\lambda y})^{N-1} e^{-\lambda y}$

Statistic

Theorems

Expectation

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

For continuous: $E(X) = \int_{-\infty}^{+\infty} x f_X(x) dx$ and $E[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx$

$E(X_1 + X_2 + ... + X_n) = E(X_1) + E(X_2) + ... + E(X_n)$

$E(aX) = aE(X)$

If all independent, then: $E(X_1 X_2 ... X_n) = E(X_1) E(X_2) ... E(X_n)$

$E[g(x)] = \sum |g(a_i)| p_i$

$E[g(x)] = \int_{-\infty}^{+\infty} g(x) f(x) dx$

Conditional Expectation $\quad E[g(X)|A] = \sum_x g(x) p_{X|A}(x|A)$

$E[X] = \sum_y p_Y(y) E[X|Y = y]$ (total expectation theorem)

$E(X|Y = y) = \int_{-\infty}^{+\infty} x f(x|y) dx$

$E[X|Y = y]$ is a number, whose value depends on y.

$E[X|Y]$ is a function of the random variable Y, hence a random variable. Its experimental value is $E[X|Y = y]$ whenever the experimental value of Y is y.

$E(X) = E[E(X|Y)]$ (law of iterated expectations)

Variance

$var(X) = E[(X - E[X])^2]$

For discrete: $var(X) = \sum_{i=1}^{n} P(X = x_i)(x_i - \mu)^2$

For continuous: $var(X) = \int_{-\infty}^{+\infty} f_X(x)[x - E(X)]^2 dx$

$var(X) = E(X^2) - [E(X)]^2 \geq 0$

$var(X + a) = var(X)$

$var(aX) = a^2 var(X)$

If all independent, then:

$var(X_1 + X_2 + ... + X_n) = var(X_1) + ... + var(X_n)$

$var(X - Y) = var[X + (-1)Y] = var(X) + (-1)^2 var(Y) = var(X) + var(Y)$

Standard Deviation

$\sigma = \sqrt{var(X)}$

Covariance

$cov(X,Y) = E[(X - m_1)](Y - m_2)]$

$cov(X,Y) = E(XY) - E(X)E(Y)$

$cov(aX + b, cY + d) = ac \times cov(X,Y)$

If all independent, then:

$cov(X,Y) = 0$

Correlation Coefficient

$corr(X,Y) = \frac{cov(X,Y)}{\sigma_1 \sigma_2}$

If all independent, then:

$corr(X,Y) = 0$

Example 1

Question

For $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ (all independent), why $var(X - Y) = \sigma_X^2 + \sigma_Y^2$?

Solution

$var(X - Y) = var(X + (-1) \times Y) = var(X) + (-1)^2 var(Y) = var(X) + var(Y) = \sigma_X^2 + \sigma_Y^2$

Example 2

Question

Cauchy Inequality: Let $a_i, b_i (i = 1, 2, 3, ..., n)$ be sequences of positive numbers.

Suppose that

$a_1 + a_2 + \cdots + a_n = 1$

Prove that

$\frac{b_1^2}{a_1} + \frac{b_2^2}{a_2} + \cdots + \frac{b_n^2}{a_n} \geq (b_1 + b_2 + \cdots + b_n)^2$

Hint: Consider a discrete random variable X with probability mass function

$f_X(\frac{b_k}{a_k}) = P(X = \frac{b_k}{a_k}) = a_k, k = 1, 2, 3, ..., n$

Solution

Consider a discrete random variable X with probability mass function:

$f_X(\frac{b_k}{a_k}) = P(X = \frac{b_k}{a_k}) = a_k, k = 1, 2, 3, ..., n$

$\because f_X(\frac{b_k}{a_k}) = a_k$

$\therefore E(X) = \sum_{k=1}^{n} \frac{b_k}{a_k} \times f_X(\frac{b_k}{a_k}) = \sum_{k=1}^{n} b_k$

$$\therefore E(X^2) = \sum_{k=1}^{n} \left(\frac{b_k}{a_k}\right)^2 \times f_X\left(\frac{b_k}{a_k}\right) = \sum_{k=1}^{n} \frac{b_k^2}{a_k}$$

$$\because var(X) = E(X^2) - E(X)^2 = \sum \frac{b_k^2}{a_k} - \left(\sum b_k\right)^2 \geq 0$$

$$\therefore var(X) = E\{[X - E(X)]^2\} \geq 0$$

$$\therefore E(X^2) = \frac{b_1^2}{a_1} + \frac{b_2^2}{a_2} + \cdots + \frac{b_n^2}{a_n} \geq [E(X)]^2 = (b_1 + b_2 + \cdots + b_n)^2$$

Example 3

Question

Please prove the Law of Iterated Expectation: $E[E(X|Y)] = E(X)$.

Solution

$E[E(X|P)]$
$= \int_{-\infty}^{+\infty} E(X|Y = y) \times f_Y(y) dy$
$= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x f_{X|Y}(x|y) \times f_Y(y) dx dy$
$= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x f_{X,Y}(x, y) dx dy$
$= E(X)$

Example 4

Suppose that we have d independent standard normal random variables $Z_1, \ldots, Z_d$, where each $Z_i \sim N(0, 1)$. We say that a continuous random variable X follows a chi-squared (χ^2) distribution with d degrees of freedom if $X \overset{D}{=} Z_1^2 + \cdots + Z_d^2$. Remember that $\overset{D}{=}$ means "equal in distribution," i.e. the left and right-hand side of the equation have the same CDF. Our shorthand for the chi-squared distribution is $X \sim \chi_d^2$. The chi-squared distribution plays an important role in statistical inference. For now, your job is simply to compute $E(X)$ using the relationship between the normal and the chi-squared. (That is, do not look up the PDF of the chi-squared distribution and try to compute the mean and variance directly, by integration.)

Solution

For each standard normal random variable Z_i picked from $Z_1, \ldots, Z_d$ is said to follow $Z_i \sim N(0, 1)$.

$f(Z_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z_i^2}{2}}$

$E(Z_i^2) = \int_{-\infty}^{+\infty} Z_i^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{Z_i^2}{2}} dZ_i$

Because of

$\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{Z_i^2}{2}}\right)' = -Z_i\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{Z_i^2}{2}}\right),$

and

$\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{Z_i^2}{2}}\right)'' = Z_i^2\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{Z_i^2}{2}}\right) - \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{Z_i^2}{2}}\right)$

therefore,

$E(Z_i^2) = \int_{-\infty}^{+\infty} \left[\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{Z_i^2}{2}}\right)'' + \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{Z_i^2}{2}}\right)\right] dZ_i$

which is equal to:

$E(Z_i^2) = \int_{-\infty}^{+\infty} (\frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}})'' dZ_i + \int_{-\infty}^{+\infty} (\frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}}) dZ_i$

so,

$E(Z_i^2) = 0 + 1 = 1$

It also is clear that $E(Z_i) = 0$ and $var(Z_i) = 1$.

Because the formula $var(Z_i) = E(Z_i^2) - (E(Z_i))^2$, we also can find:

$E(Z_i^2) = var(Z_i) + (E(Z_i))^2 = 1 + 0^2 = 1$

For $X \overset{D}{=} Z_1^2 + \cdots + Z_d^2$, we can compute $E(X)$ as:

$E(X) = E(Z_1^2) + \cdots + E(Z_d^2) = \underbrace{1 + \cdots + 1}_{d} = d$

Question

To compute var(X) using the same basic strategy.

Solution

Let's denote the PDF of standard normal distribution as $\phi(z)$.

Because of

$\phi'(z) = -z\phi(z)$

and

$\phi''(z) = z^2\phi - \phi$

and

$\phi^{(3)}(z) = 2z\phi(z) - z^3\phi + z\phi$

and

$\phi^{(4)}(z) = z^4\phi - 6z^2\phi + 3\phi$

therefore,

$E(Z_i^4) = \int_{-\infty}^{+\infty} Z_i^4 \phi(Z_i) dZ_i = \int_{-\infty}^{+\infty} [\phi^{(4)}(Z_i) + 6Z_i^2\phi(Z_i) - 3\phi(Z_i)] dZ_i$

which is equal to:

$E(Z_i^4) = \int_{-\infty}^{+\infty} \phi^{(4)}(Z_i) dZ_i + 6 \int_{-\infty}^{+\infty} Z_i^2 \phi(Z_i) dZ_i - 3 \int_{-\infty}^{+\infty} \phi(Z_i) dZ_i = 0 + 6 - 3 = 3$

So $var(Z_i^2) = E(Z_i^4) - (E(Z_i^2))^2 = 3 - 1^2 = 2$

Since X is the sum of the squares of d independent standard normal random variables, the variance of X is the sum of the variances of Z_i^2:

$var(X) = \underbrace{var(Z_1^2) + \cdots + var(Z_d^2)}_{d} = 2d$

Example 5

Question

Suppose that $X_N \sim Binomial(N, P)$ be the (random) number of successes in a sequence of N binary trials. Let $\hat{p}N = X_N/N$ denote the proportion of observed successes. Calculate $E(\hat{p}N)$ and $sd(\hat{p}N)$.

Solution

$\because X_n \sim \text{Binominal}(N, P)$

$\therefore E(X_N) = NP$ and $var(X_N) = NP(1 - P)$

Let $\hat{p}N = X_N/N$ denote the proportion of observed successes.

Based on the formulas:

$E(aX) = aE(X)$ and $var(aX) = a^2 var(X)$

it it clear that:

$E(\hat{p}_N) = E(X_N/N) = \frac{1}{N}E(X_N) = \frac{1}{N}NP = P$

$var(\hat{p}_N) = var(X_N/N) = (\frac{1}{N})^2 var(X_N) = \frac{1}{N^2}NP(1 - P) = \frac{P(1-P)}{N}$

$\therefore sd(\hat{p}_N) = \sqrt{\frac{P(1-P)}{N}}$

Example 6

Question

Suppose that $X_1, ..., X_N \sim Bernoulli(p)$ and that $Y_1, ..., Y_M \sim Bernoulli(q)$ (all independent). We will consider $\hat{p} = X_N$ and $\hat{q} = Y_M$ as estimators of p and q, respectively.

    i. Show that $E(\hat{p} - \hat{q}) = p - q$, the true difference in success probabilities.

    ii. Use what you know of probability to compute the standard error of $\hat{p}$, i.e. the standard deviation of the sampling distribution of $\hat{p}$.

    iii. Compute the standard error of $\hat{\Delta} = \hat{p} - \hat{q}$ as an estimator of the true difference $\Delta = p - q$.

Solution

First of all, there are some assumptions:

$\because X_1, ..., X_N \sim \text{Bernoulli}(p)$ and $Y_1, ..., Y_M \sim \text{Bernoulli}(q)$ (all independent)

$\because \hat{p} = \overline{X}_N$ and $\hat{q} = \overline{Y}_M$

Question i $\because E(\hat{p}) = E(\frac{\sum_{i=1}^{N} X_i}{N})$ and $E(\hat{q}) = E(\frac{\sum_{j=1}^{M} Y_j}{M})$

$\therefore E(\hat{p} - \hat{q}) = E(\hat{p}) - E(\hat{q}) = \frac{\sum_{i=1}^{N} E(X_i)}{N} - \frac{\sum_{j=1}^{M} E(Y_j)}{M} = p - q$

Question ii $\because var(\hat{p}) = var(\frac{\sum_{i=1}^{N} X_i}{N}) = \frac{1}{N^2}var(\sum_{i=1}^{N} X_i) = \frac{1}{N^2}\sum_{i=1}^{N} var(X_i)$

$\because var(X_i) = p(1 - p)$

$\therefore var(\hat{p}) = \frac{1}{N}p(1 - p)$

$\therefore \text{SE}(\hat{p}) = \sqrt{\frac{p(1-p)}{N}}$

Question iii $\because var(\hat{\Delta}) = var(\hat{p} - \hat{q}) = var(\frac{\sum_{i=1}^{N} X_i}{N} + (-1) \times \frac{\sum_{j=1}^{M} Y_j}{M}) = \frac{1}{N^2} \sum_{i=1}^{N} var(X_i) + \frac{1}{M^2} \sum_{j=1}^{M} var(Y_j)$

$\because var(X_i) = p(1-p)$ and $var(Y_j) = q(1-q)$

$\therefore var(\hat{\Delta}) = \frac{1}{N} p(1-p) + \frac{1}{M} q(1-q)$

$\therefore SE(\hat{\Delta}) = \sqrt{\frac{p(1-p)}{N} + \frac{q(1-q)}{M}}$

Question

Suppose we have data on some numerical attribute from two groups: $X_1, ..., X_N$ from group 1, and $Y_1, ..., Y_M$ from group 2. (Notice the unequal sample sizes N and M.) Suppose that:

- $E(X_i) = \mu_X$ and $var(X_i) = \sigma_X^2$, both unknown

- $E(Y_i) = \mu_Y$ and $var(Y_i) = \sigma_Y^2$, again both unknown

Suppose we're interested in the population-level difference in means between the two groups: $\Delta = \mu_X - \mu_Y$.

We use the difference in sample means to estimate this quantity: $\hat{\Delta} = \hat{X}_N - \hat{Y}_M$

Use your knowledge of probability theory to calculate the expected value and standard error of $\hat{\Delta}$. These expressions should involve some combination of the true unknown population parameters and the sample sizes.

Solution

First of all,

$\overline{X}_N = \frac{1}{N} \sum_{i=1}^{N} X_i$

$\overline{Y}_M = \frac{1}{M} \sum_{j=1}^{M} Y_j$

$\Delta = \mu_X - \mu_Y$

$\hat{\Delta} = \overline{X}_N - \overline{Y}_M$

Assume $X \perp Y$

Expected value

$\because E(\hat{\Delta}) = E(\overline{X}_N - \overline{Y}_M) = E(\frac{1}{N} \sum_{i=1}^{N} X_i - \frac{1}{M} \sum_{j=1}^{M} Y_j)$

$\therefore E(\hat{\Delta}) = \frac{1}{N} \sum_{i=1}^{N} E(X_i) - \frac{1}{M} \sum_{j=1}^{M} E(Y_j) = E(X_i) - E(Y_j) = \mu_X - \mu_Y = \Delta$

Standard error

$\because var(\hat{\Delta}) = var(\overline{X}_N - \overline{Y}_M) = var(\frac{\sum_{i=1}^{N} X_i}{N} + (-1) \times \frac{\sum_{j=1}^{M} Y_j}{M}) = \frac{1}{N^2} \sum_{i=1}^{N} var(X_i) + \frac{1}{M^2} \sum_{j=1}^{M} var(Y_j)$

$\therefore SE(\hat{\Delta}) = \sqrt{\frac{\sigma_X^2}{N} + \frac{\sigma_Y^2}{M}}$

Example 7

Question

Here's our basic model equation:

$Y_t = Y_0 \cdot e^{rt}$

- What happens if we work with this model on the log scale, i.e. by taking the log of both sides?

Solution

$logY_t = log(Y_0 \cdot e^{rt}) = log(Y_0) + r \cdot t$

You can prove this using the rules for logs, top right.

- This equation says that:

The log of Y_t is a linear function of the time variable (t).

Exponential = linear on a log scale. To diagnose whether an exponential model makes sense, plot log(y) vs. t and check whether the relationship looks linear.

Example 8

Question

- There are 140 million housing units in the US, and roughly 388,000 had a fire last year. So on average, $P(fire) = 388,000/140,000,000 \approx 0.00063$

- Suppose you own a house worth $320,000. Would you pay $30 to insure it against fire?

Solution

- There are 140 million housing units in the US, and roughly 388,000 had a fire last year. So on average, $P(fire) = 388,000/140,000,000 \approx 0.00063$

- Suppose you own a house worth $320,000. Would you pay $30 to insure it against fire?

- Expected value if uninsured:

Let's make the simplifying assumption that the fire destroys the entire structure.

Your "expected" fire loss is: $EV = (1 ̌0.000063) \times 0 + 0.000063 \times 320,000 \approx \$20$.

- Expected value if insured:

You lose $30, guaranteed—but no longer have the risk of a $320,000 loss.

The difference ($10) is the premium you'd be paying in order to convert the $20 expected loss into a $30 guaranteed loss that won't bankrupt you.

Question

- Suppose each house is worth $320,000, and you have 100,000 clients that all buy insurance.

Again, suppose each fire destroys the whole house.

We "expect" that $0.00063 \times 100,000 = 6.3$ of your clients will suffer a fire and make a $320,000 claim.

Then the total payout is $320,000 \times 6.3 \times \approx \$2$million, spread across 100,000 clients.

- That's an average of $20 per client:

The average fire loss across all the households is the same as each individual household's "expected" fire loss.

That's the LLN in action: the "expected value" for an individual household is the average of the actual outcomes experienced across many different households.

Chebyshev's inequality

Markov's inequality

Let $Y_t = \begin{cases} 0, & \text{if } x < t \\ t, & \text{if } x \geq t \end{cases}$

Therefore,

$E(Y_t) \leq E(X)$

That's because:

$0 \times P(X < t) + t \times P(X \geq t) \leq E(X)$

So,

$t \times P(X \geq t) \leq E(X)$

$\therefore P(X \geq t) \leq \frac{E(X)}{t}$

Chebyshev's inequality

For $|x - \mu| \geq t$

$P(|x - \mu| \geq t) = P([x - \mu]^2 \geq t^2)$

Given Markov's inequality, then:

$P([x - \mu]^2 \geq t^2) \leq \frac{E([x-\mu]^2)}{t^2}$

$P(|x - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$

The Law of Large Numbers (LLN)

Convergence in probability.

It says that the mean of a large sample is close to the expected value of the distribution from which the sample is drawn:

$$\overline{X}_N \xrightarrow{P} \mu$$

Suppose $X_1, X_2, ... X_N$ are all random variables with expected value μ.

Let $\overline{X}_N$ be the mean of these N numbers.

The Law of Large Numbers says that as N gets large, $\overline{X}_N$ becomes very close to μ.

Central Limit Theorem

Suppose that $X_1, ..., X_N$ are IID random variables with mean μ and variance σ^2.

Let $\overline{X}_N$ be the sample mean.

Now define Z_N as the z-score of the sample mean:

$$Z_n = \frac{\overline{X}_n - \mu}{sd(\overline{X}_n)} = \frac{\overline{X}_n - \mu}{\sqrt{var(\overline{X}_n)}} = \frac{\overline{X}_n - \mu}{\sqrt{\frac{\sigma^2}{N}}} = \frac{\sqrt{N}(\overline{X}_n - \mu)}{\sigma}$$

So, for all $n \in [1, N]$, there is:

$$Z_n = \frac{\sqrt{N}(\overline{X}_n - \mu)}{\sigma} \sim N(0, 1)$$

Code 1

```
# Generating Random Variables

# Uniform distribution
runif(n = 10, min = 0, max = 1)
# Normal distribution
rnorm(n = 10, mean = 0, sd = 1)
# Bernoulli distribution (using size = 1)
rbinom(n = 10, size = 1, prob = 0.5)
# Binomial distribution
rbinom(n = 10, size = 1, prob = 0.5)
# Poisson distribution
rpois(n = 10, lambda = 2)
# Exponential distribution
rexp(n = 10, rate = 1)
```

Code 2

```
# PDF/PMF Function

# PDF of Uniform distribution
dunif(x = 1, min = 0, max = 1)
# PDF of Normal distribution
dnorm(x = 1, mean = 0, sd = 1)
# PMF of Bernoulli distribution (using size = 1)
dbinom(x = 10, size = 1, prob = 0.5)
# PMF of Binomial distribution
# this is the probability mass function of the binomial, i.e. P(X = k)
dbinom(x = 10, size = 1, prob = 0.5)
# PMF of Poisson distribution, chance of 2 goals
dpois(x = 10, lambda = 2)
# PDF of Exponential distribution
dexp(x = 10, rate = 1)
```

Code 3

```
# CDF Function

# CDF of Uniform distribution
punif(q = 1, min = 0, max = 1)
# CDF of Normal distribution
pnorm(q = 0, mean = 0, sd = 1)
# CDF of Bernoulli distribution (using size = 1), use pbinom to calculate P(X <= 11).
pbinom(q = 11, size = 1, prob = 0.5)
# CDF of Binomial distribution
pbinom(q = 2, size = 1, prob = 0.5)
# CDF of Poisson distribution, P(X <= 2)
ppois(q = 2, lambda = 2)
# CDF of Exponential distribution
pexp(q = 2, rate = 1)
```

Code 4

```
# Summary and Wrangling

# to calculate summary statistics
titanic %>%
  summarize(avg_age = mean(age))
# used to split the rows of a data set into groups.
titanic %>%
  group_by(sex)
# keep rows that satisfy your conditions, ignore everthing else.
titanic %>%
  filter(sex == "female")
# used to select specific columns (variables) in your data frame
titanic %>%
```

```
  select(sex, age)
# add a column defined in terms of existing columns
titanic %>%
  mutate(children = ifelse(age < 16, "yes", "no"))
# used to sort according to a variable or set of variables
titanic %>%
  arrange(sex, age)
```

Code 5

```
# Linear Model, Example: Heart rate vs. age
# Type of change: Additive
lm_heartrate = lm(hrmax ~ age, data=heartrate)
coef(lm_heartrate)

# For Exponential Model, Example: Population vs. time
# Type of change: Multiplicative
# linear model for log(cases) versus time
lm_ebola = lm(log(totalSus) ~ Day, data=ebola)
coef(lm_ebola)

# Power Law Model, Example: Demand vs. price
# Type of change: Relative proportional
# Let's fit the power law by fitting a linear model on a log-log scale
lm1 = lm(log(sales) ~ log(price), data=milk)
coef(lm1)
```

Code 6

```
# Estimating a mean
do(10000)*mean(~age, data=resample(titanic))
# Estimating a proportion
do(10000)*prop(~age, data=resample(titanic))
# Comparing two means
do(10000)*mean(age~sex, data=resample(titanic))
# Comparing two proportions
do(10000)*prop(age~sex, data=resample(titanic))
# Regression model
do(10000)*lm(age~sex, data=resample(titanic))
```

Code 7

```
# Simulate lots of games between Arsenal and Man City
# under a very simple Poisson model.
# the rates are chosen from average goal-scoring rates last season.
```

```r
# On your homework, you will modify this model to account for:
#   - home and away differences
#   - offensive strength and defensive weakness for each team

NMC = 100000
arsenal = rpois(NMC, 1.45)
ManCity = rpois(NMC, 2.18)

# Compile the results
xtabs(~arsenal + ManCity)

# Monte Carlo estimates of probabilities
sum(arsenal > ManCity)/NMC
sum(arsenal == ManCity)/NMC
sum(arsenal < ManCity)/NMC

# Compare with a calculation from the PMF assuming independence
dpois(1, 1.45) * dpois(1, 2.18) # Arsenal 1 - 1 Man City
dpois(0, 1.45) * dpois(2, 2.18) # Arsenal 0 - 2 Man City
```

Code 8

```r
# Data on annual returns of stocks, bonds, and savings account
annual_returns = read.csv("../data/raw data/annual_returns_since1928.csv")

# first few lines
head(annual_returns)

# Shape of distribution of returns, net of inflation?
hist(annual_returns$SP500 - annual_returns$Inflation)
```

Histogram of annual_returns$SP500 – annual_returns$Inflation

annual_returns$SP500 – annual_returns$Inflation
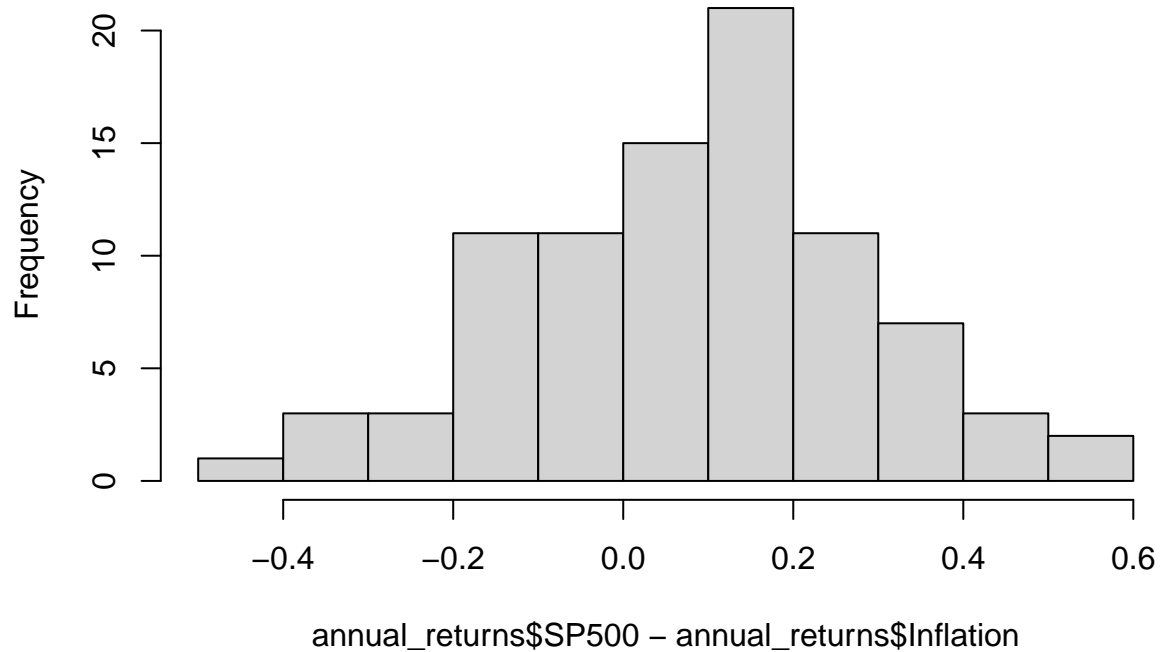
```r
# Mean and standard deviation of net returns for stocks
mean(annual_returns$SP500 - annual_returns$Inflation)
sd(annual_returns$SP500 - annual_returns$Inflation)

# Modeling a risky asset with a positive expected return
mu = mean(annual_returns$SP500 - annual_returns$Inflation)
sigma = sd(annual_returns$SP500 - annual_returns$Inflation)
Horizon = 40

Wealth = 10000  # Initial wealth
WealthTracker = rep(0, Horizon)  # Placeholder
# Sweep through each year and update the value of wealth
for(i in 1:Horizon) {
  ThisReturn = rnorm(1, mu, sigma)  # simulate a random return
  Wealth = Wealth * (1 + ThisReturn) # update wealth recursively
  WealthTracker[i] = Wealth # save the result
}
Wealth  # final value of wealth
plot(WealthTracker, type='l')  # wealth over time
```
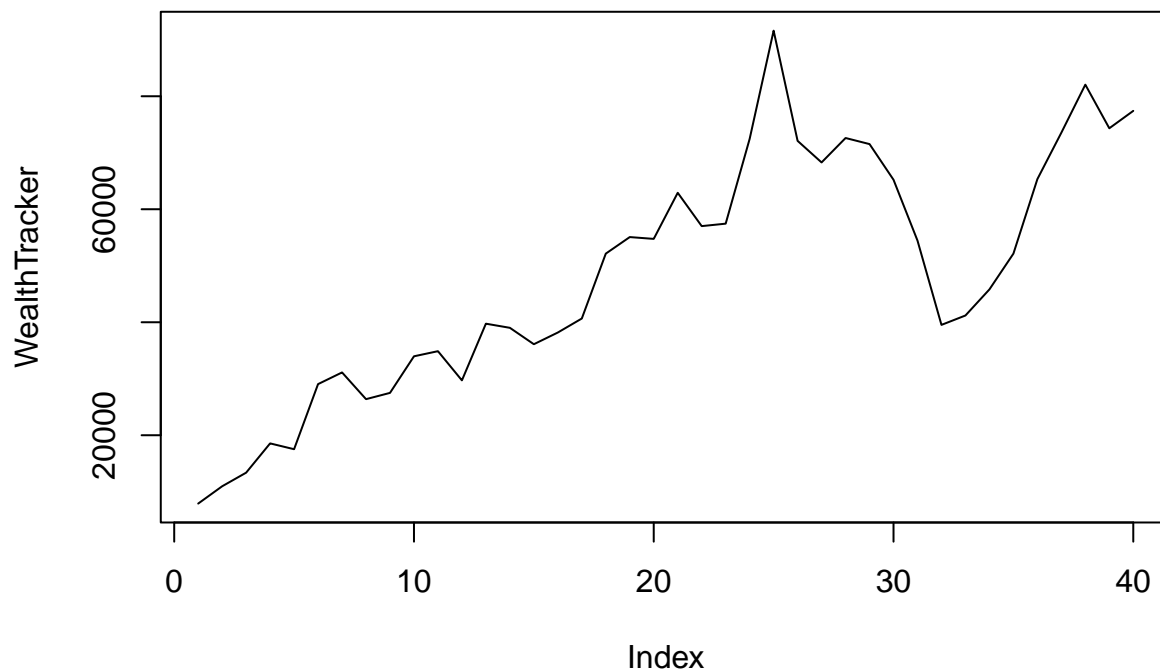
```
# Now a Monte Carlo simulation
InitialWealth = 10000
sim1 = do(1000)*{
  Wealth = InitialWealth  # Reset initial wealth
  WealthTracker = rep(0, Horizon)  # Placeholder
  # Sweep through each year and update the value of wealth
  for(i in 1:Horizon) {
    ThisReturn = rnorm(1, mu, sigma)
    Wealth = Wealth * (1 + ThisReturn)
    WealthTracker[i] = Wealth
  }
  WealthTracker
}
head(sim1)

# calculate some summary statistics
summary(sim1$V40)



# Plot a few simulated scenarios
plot(1:Horizon, sim1[1,], type='l')
lines(1:Horizon, sim1[2,], type='l')
lines(1:Horizon, sim1[3,], type='l')
```
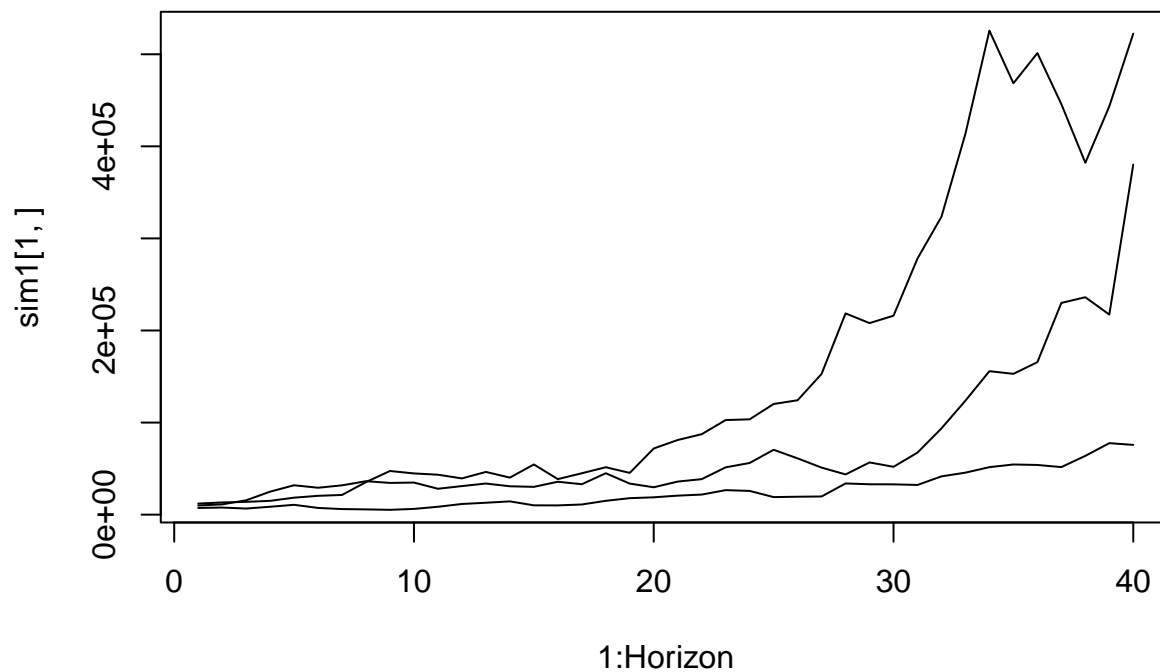
```
# A cool (more advanced) plot to show variability over time

plot(1:Horizon, colMeans(sim1), type='l', col='red',
    xlab="Years into future",
    ylab="Value of portfolio",
    main="Simulated growth of a stock portfolio over 40 years",
    las=1, cex.axis = 0.85,
    ylim=c(10000, 2*max(colMeans(sim1))))

for(sim in 1:50) {
  lines(1:Horizon, sim1[sim,], type='l', col=rgb(0.5,0.5,0.5,0.25))
}
lines(1:Horizon, colMeans(sim1), lwd=3, col='red')
# legend('topleft',
#   legend=c("Average trajectory", "All simulations", "Example trajectories"),
#   lwd=1, col=c("red", "grey", "blue"))
lines(1:Horizon, sim1[1,], col='blue')
lines(1:Horizon, sim1[2,], col='blue')
```

Simulated growth of a stock portfolio over 40 years
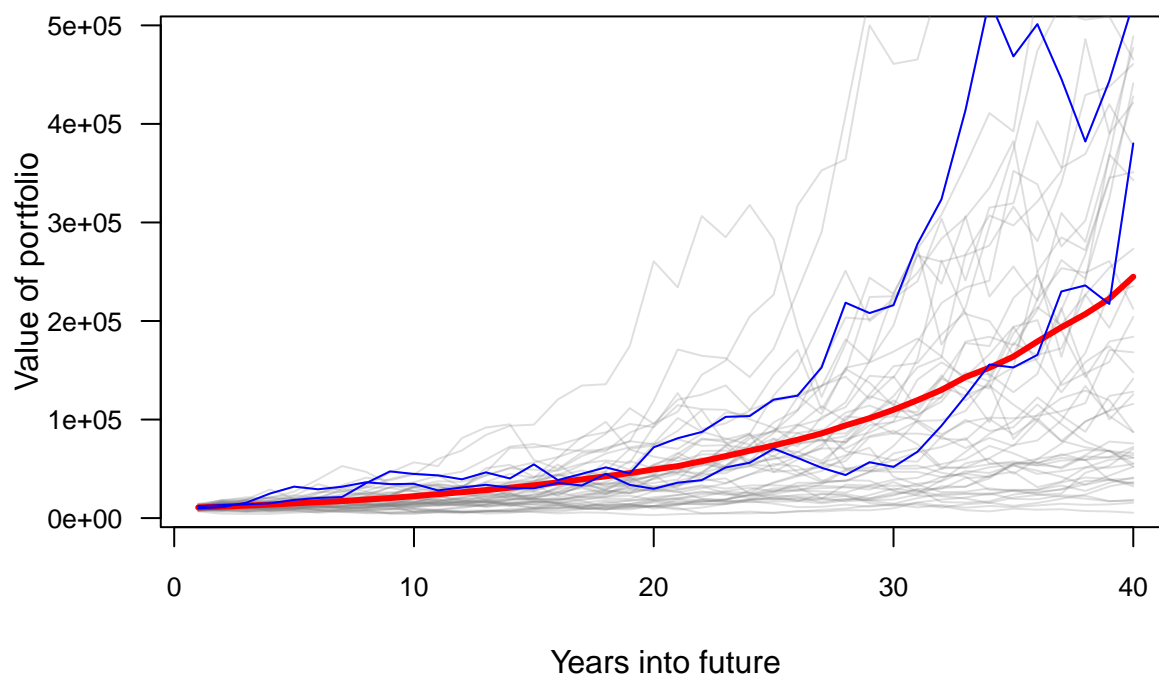
Code 9

```r
# simple computations with the normal distribution
# rnorm: simulating random normal variables
# pnorm: calculating tail areas under the normal distribution
# there's also a dnorm...
# but it's kind of hard to interpret without calculus!
# we won't use dnorm (the normal density function) in this course.

# Since 1918, the US stock market has averaged a 6% annual return
# net of inflation, with a yearly standard deviation of 20%.
ReturnAvg = 0.06
ReturnSD = 0.20

# if yearly stock returns are normally distributed,
# what do they tend to look like?
# rnorm wants to know three things:
#   - how many random draws (here, 10,000)
#   - the mean
#   - the standard deviation
sim_returns = tibble(return = rnorm(10000, ReturnAvg, ReturnSD))
ggplot(sim_returns) +
  geom_histogram(aes(x=return))
```
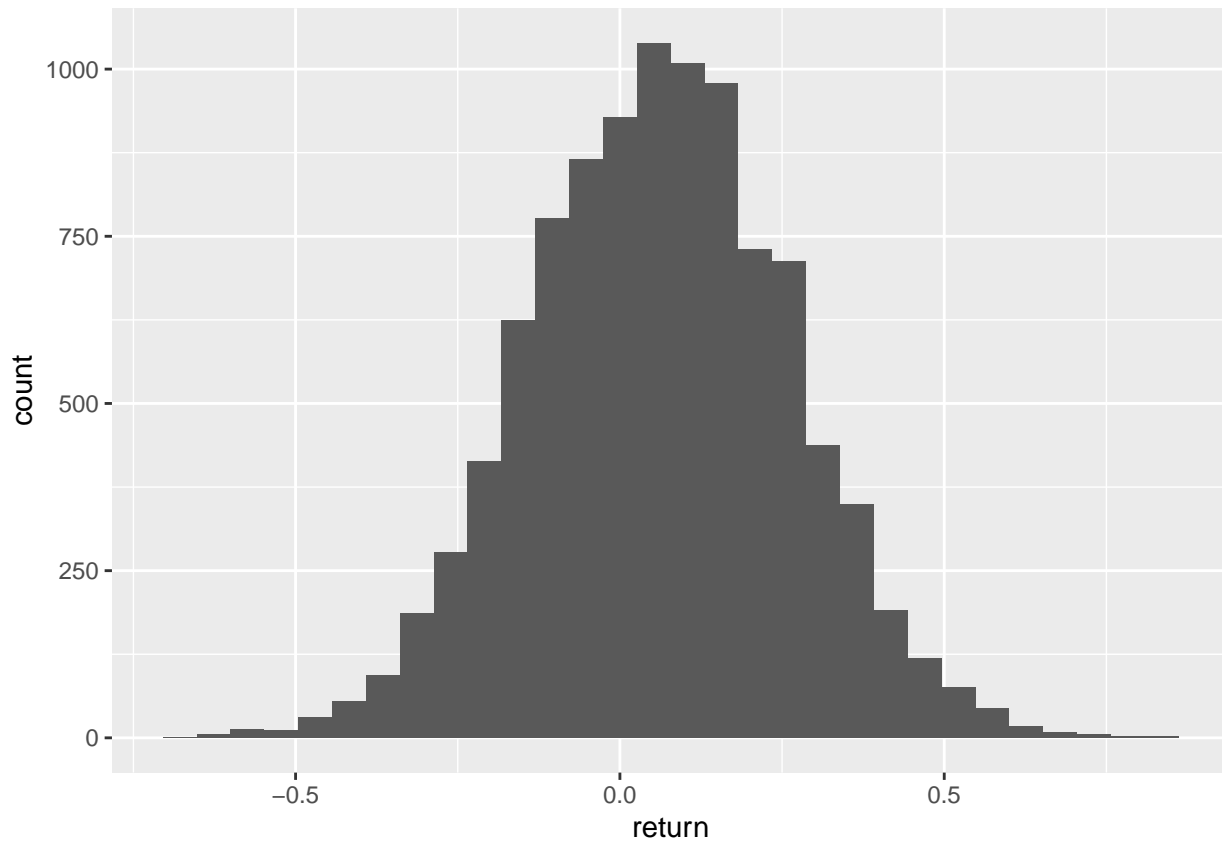
```
# What is the probability of seeing a really bad year,
# with -25% return or worse?
# pnorm(x, mu, sd) gives you P(X <= x) if X ~ N(mu, sd)
pnorm(-0.25, ReturnAvg, ReturnSD)

# What about about a really good year, with >30% growth?

# pnorm gives us the chance of 30% return or less...
pnorm(0.3, ReturnAvg, ReturnSD)

# so this is the chance of >30% return:
1 - pnorm(0.3, ReturnAvg, ReturnSD)
```

Code 10

```
# Let's simulate three years of growth in a portfolio
# that earns 6% interest with zero risk.
# We start with some initial wealth
Wealth = 10000

# Now we update/overwrite the value of Wealth three times.
# this is like compounding the 6.3% interest rate once after each year.
Wealth = Wealth * (1 + 0.06) # year 1
Wealth = Wealth * (1 + 0.06) # year 2
```

```
Wealth = Wealth * (1 + 0.06) # year 3

# where are we after three years?
Wealth

# Let's accomplish the same thing with a "for loop"
# the key idea here is recursion.
# a recursive formula defines the next step in a sequence of numbers
# in terms of the previous number in the sequence.
Wealth = 10000  # reset initial wealth to its initials tate
for(year in 1:3) {
   # We update/overwrite the Wealth variable each pass through the for loop.
   # this update is recursive: previous wealth on the right,
   # new wealth on the left.
   Wealth = Wealth * (1 + 0.06)
}
Wealth


# The nice thing about the for loop:
# you use the same code for any length of your investment horizon
# without copy/pasting a bajillion times
Wealth = 10000
Horizon = 40
for(year in 1:Horizon) {
   Wealth = Wealth * (1 + 0.06)
}
Wealth


# But this is a riskless asset with a 6% real return!
# Good luck finding that.
# So let's make things more realistic:
# we'll invest in a risky asset with a 6% average return: stocks!
# Since 1918, the US stock market has averaged a 6% annual return
# net of inflation, with a yearly standard deviation of 20%.
ReturnAvg = 0.06
ReturnSD = 0.20
Horizon = 40

# Simulate 40 years of investment in the stock market
# where each year's return is a normal random variable
# Try the following block of code several times

# Initial wealth
Wealth = 10000
# Sweep through each year.  In each year we:
#   - Simulate a random interest rate.
#   - Update the value of wealth.
# Last year's final wealth becomes this year's initial wealth.
for(year in 1:Horizon) {
    # Generate a random return
    ThisReturn = rnorm(1, ReturnAvg, ReturnSD)
```

```r
    # Update wealth using simple interest formula
    Wealth = Wealth * (1 + ThisReturn)
}

# Final value of wealth
Wealth


####
# Now a Monte Carlo simulation with
# 1000 simulated trajectories
####

# Outer loop: repeat the basic sim 1000 times
NMC = 1000
sim1 = do(NMC)*{
  Wealth = 10000  # Reset initial wealth
  # Inner loop: sweep through each year and update the value of wealth
  for(year in 1:Horizon) {
    ThisReturn = rnorm(1, ReturnAvg, ReturnSD)
    Wealth = Wealth * (1 + ThisReturn)
  }
  Wealth
}
head(sim1)

# The probability distribution over final wealth
ggplot(sim1) +
  geom_histogram(aes(x=result))
```
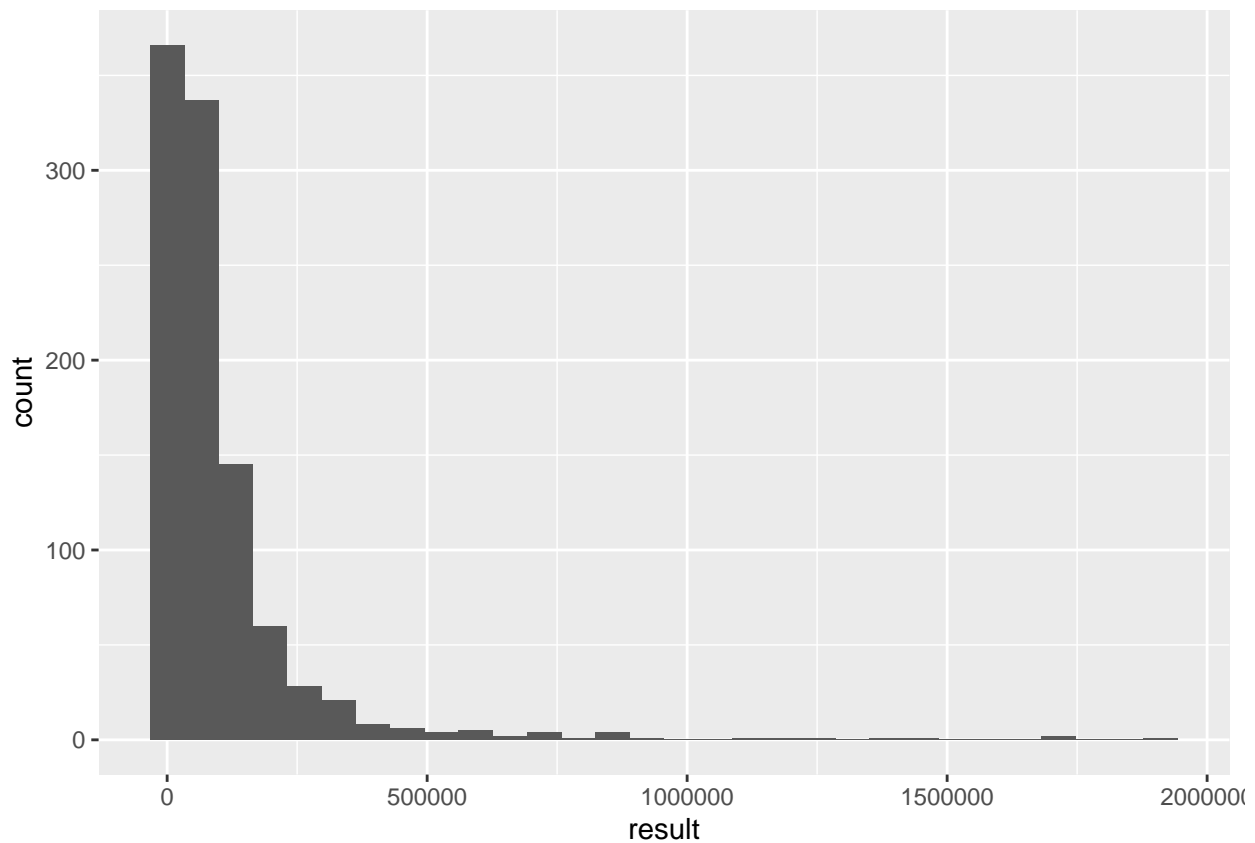
```r
# What fraction of the simulated futures lost money
# at the end of the investment horizon?
sim1 %>%
  count(result < 10000)

# How many made more than $1 million?
sim1 %>%
  count(result > 1e6)

# summary statistics
favstats(~result, data=sim1) %>% round(0)
```

Code 11

```r
# simulate from the normal random walk model:
# y_t = y_{t-1} + e_t, where e_t is N(0, sigma)

# writing our own function simulations a random walk
normal_random_walk = function(y0, sigma, n_steps) {

  # 1) set up a vector that tracks our random walk
  y_history = rep(y0, n_steps)

  # 2) initialize the value of the random walk
```

```
  y = y0

  # 3) Loop n_steps of the random walk
  for(i in 1:n_steps) {
    shock = rnorm(1, mean=0, sd=sigma)
    y = y + shock # update value of y
    y_history[i] = y # save y
  }

  # return the simulated y_history
  y_history
}

# start with an initial price
normal_random_walk(y0 = 50, sigma=1, n=390)

# let's save one in a tibble and plot it
sim1 = tibble(t = 1:390,
  price_history = normal_random_walk(50, sigma=0.05, n=390))

ggplot(sim1) +
  geom_line(aes(x=t, y=price_history))
```
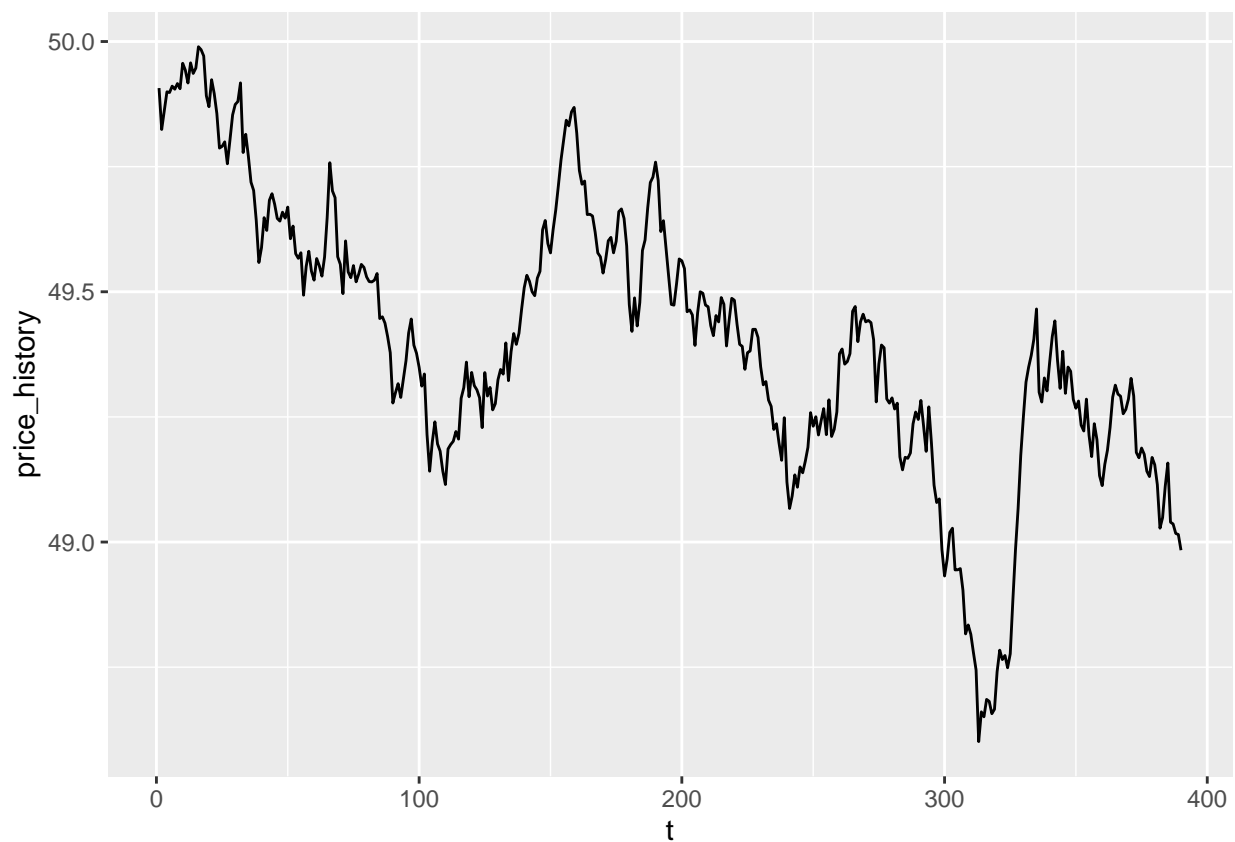


```
####
# now for some data and the real fun!
####
```

```
# read in the AAPL data frame
AAPL = read.csv('../data/raw data/AAPL.csv')
head(AAPL)

# tell R that time_stamp is a date
AAPL = AAPL %>%
  mutate(time_stamp = ymd_hms(time_stamp))

# the intra-day price at the end of each minute
ggplot(AAPL) +
  geom_line(aes(x=time_stamp, y=price))
```
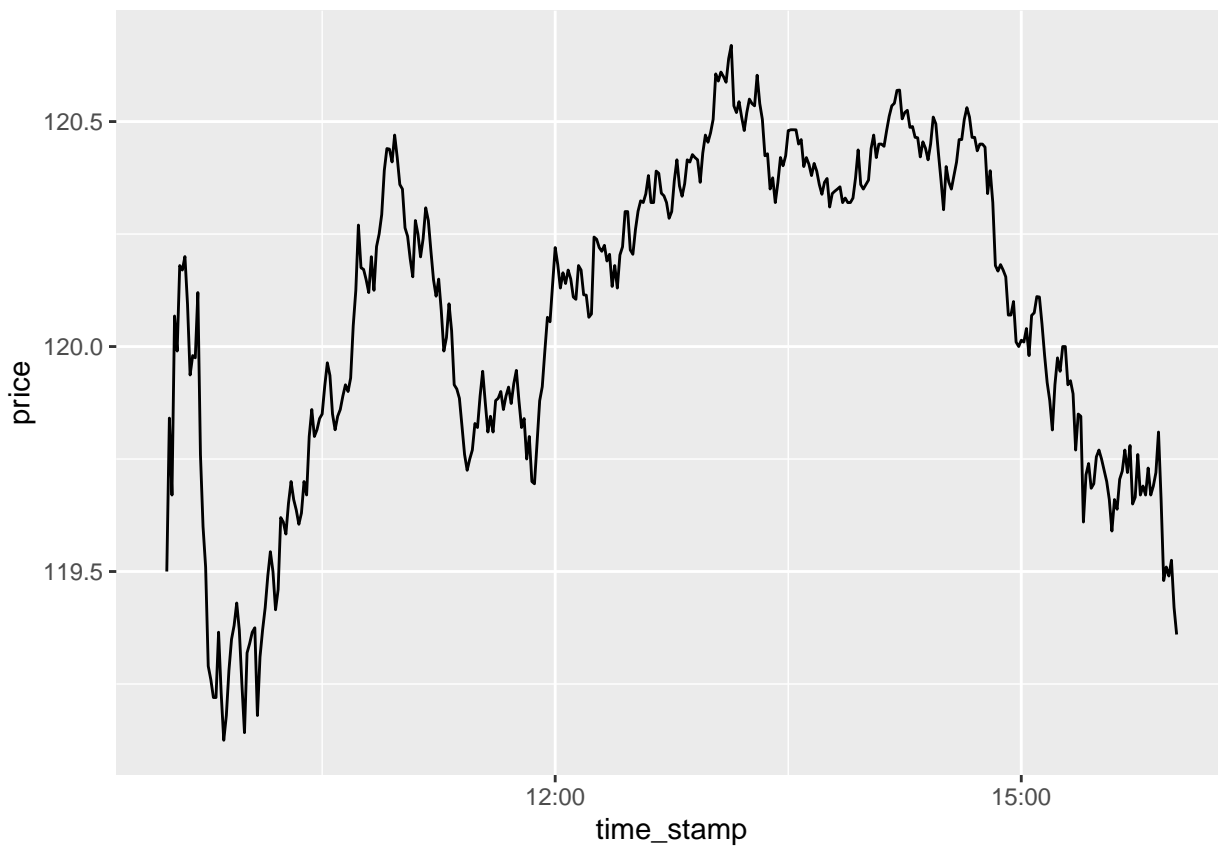


```
# Let's calculate the minute-to-minute differences in price.
# this is easy to do using the "lag" function,
# which takes the price on the previous row
AAPL = AAPL %>%
  mutate(shock = price - lag(price, n=1))


# first several lines
head(AAPL)

# std dev of of the minute-to-minute shocks
favstats(~shock, data= AAPL)

# histogram of these shocks:
```

```
# normal not perfect but not terrible
ggplot(AAPL) +
  geom_histogram(aes(x=shock))
```
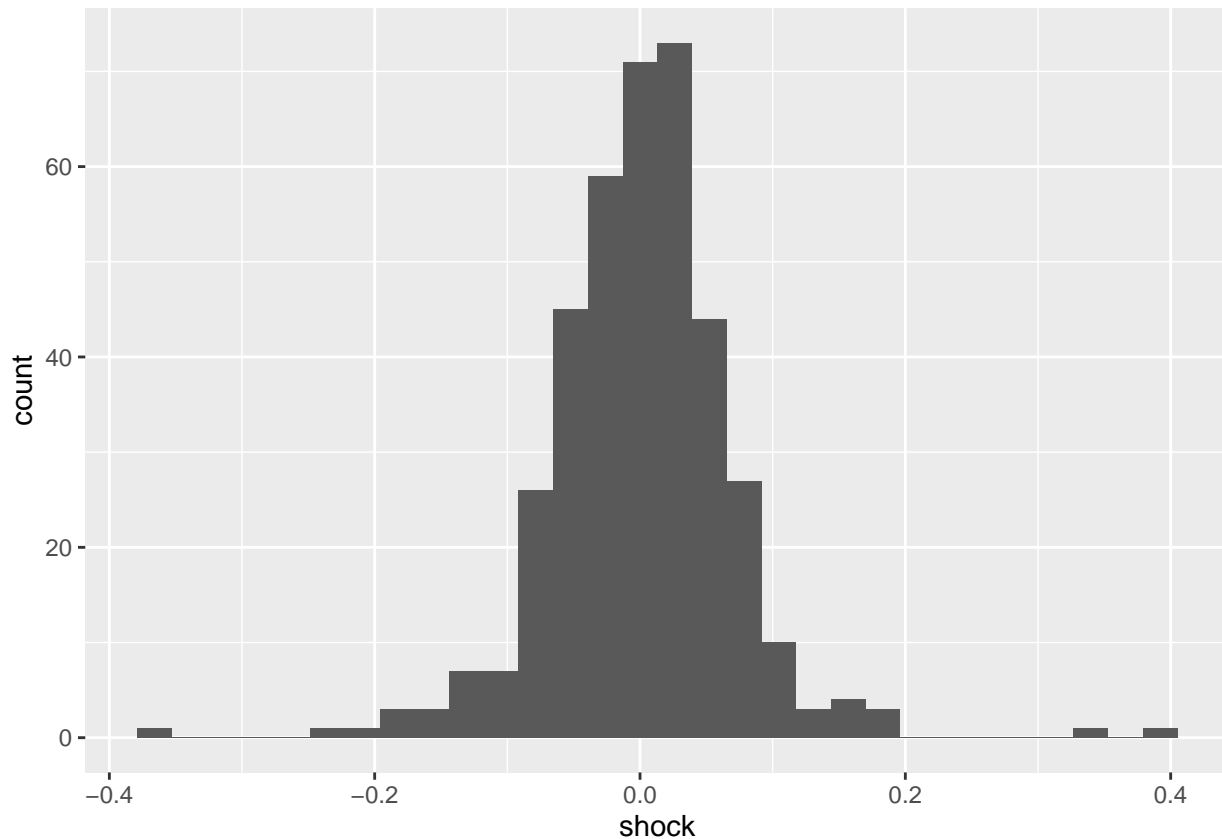


```
# Let's use this to simulate what AAPL stock might look like tomorrow

# what was the previous day's closing price?
tail(AAPL)

# looks like 119.36
# we'll use that as a starting price to simulate the next day
initial_price = 119.36
sigma = 0.0694   # our minute-to-minute standard deviation

# let's simulate 10000 intra-day trajectories for AAPL
# the inner loop involves our normal_random_walk function
AAPL_intraday = do(10000)*{
  # simulate the random walk
  price_history = normal_random_walk(initial_price, sigma, 390)

  # collect the max and min intra-day values
  c(min_intraday=min(price_history), max_intraday=max(price_history))
}


# What can we do with this?
```

```
# Application 1:
# calculate the probability that a stop-loss order will trigger.
# Suppose we put a 2.5% stop-loss on AAPL
# That is, we exit our position and sell AAPL if the price
# hits a level 2.5% below the open.
# This will happen if the price hits our stop-loss threshold
# at any point in the day.  So look at the min intra-day price

stop_loss = 0.975*initial_price
stop_loss

ggplot(AAPL_intraday) +
  geom_histogram(aes(x=min_intraday))
```
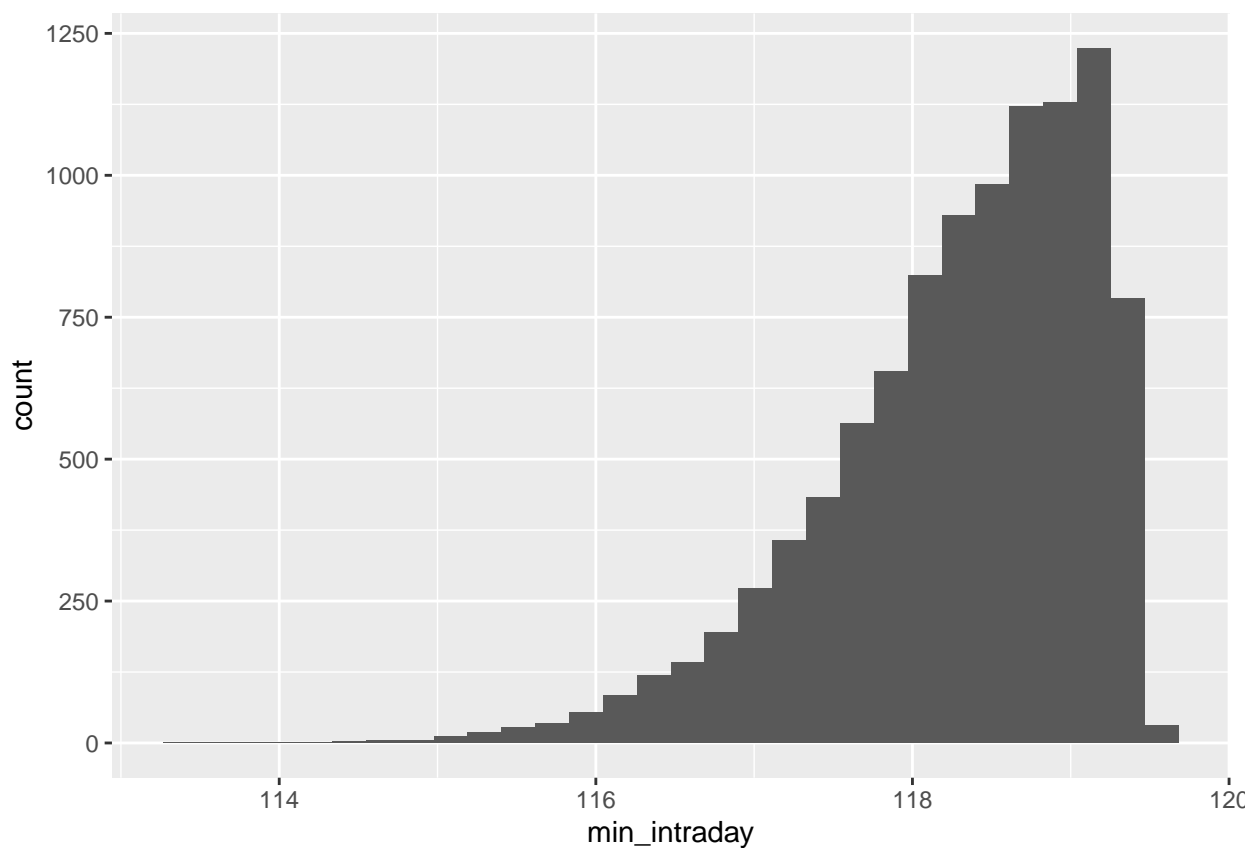


```
# let's calculate how many simulated trajectories
# had a minimum that went below our "stop-loss" level
AAPL_intraday %>%
  summarize(n_stoploss = count(min_intraday < stop_loss))


# Application 2: options pricing.
# Suppose we've bought a "call" option on our AAPL stock
# with a strike price of $121.50.  If the price hits $121.50,
# our call option will be "in the money."
# If that happens, we win, because we can buy AAPL stock
# for $121.50, rather than the current (higher) market price.
```

```
ggplot(AAPL_intraday) +
  geom_histogram(aes(x=max_intraday))
```
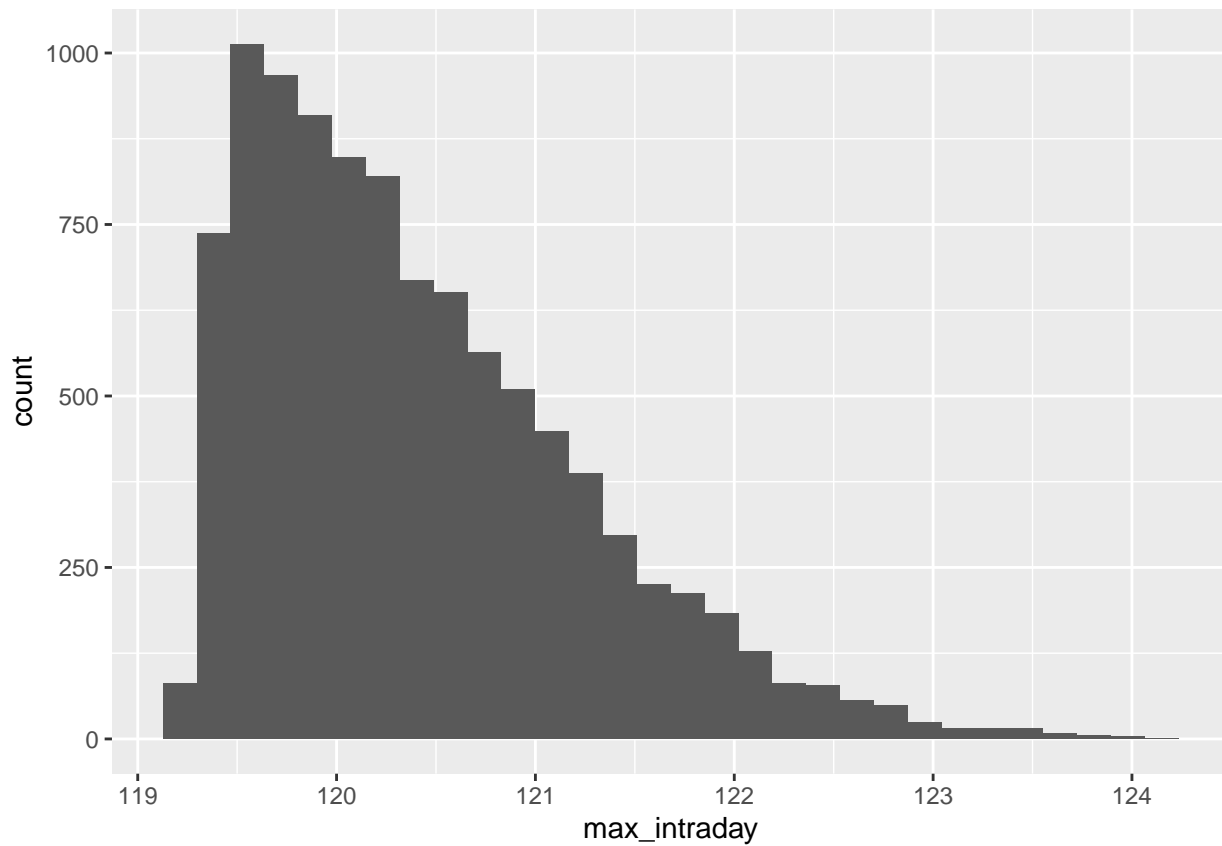


```
# how likely is AAPL stock to hit our strike price?
# look at the max price
AAPL_intraday %>%
  summarize(n_ITM = sum(max_intraday > 121.50))

# roughly a 10% chance our option reaches "in the money" today
# this random walk model is the heart of what's called the Black-Scholes
# model for pricing options in finance
```

Code 12

```
# load NHANES_sleep.csv data
# Make sure to check for a header row!

# how well are Americans sleeping, on average?
# the data distribution of sleeping hours per night in the sample
ggplot(NHANES_sleep) +
  geom_histogram(aes(x = SleepHrsNight), binwidth=1)
```
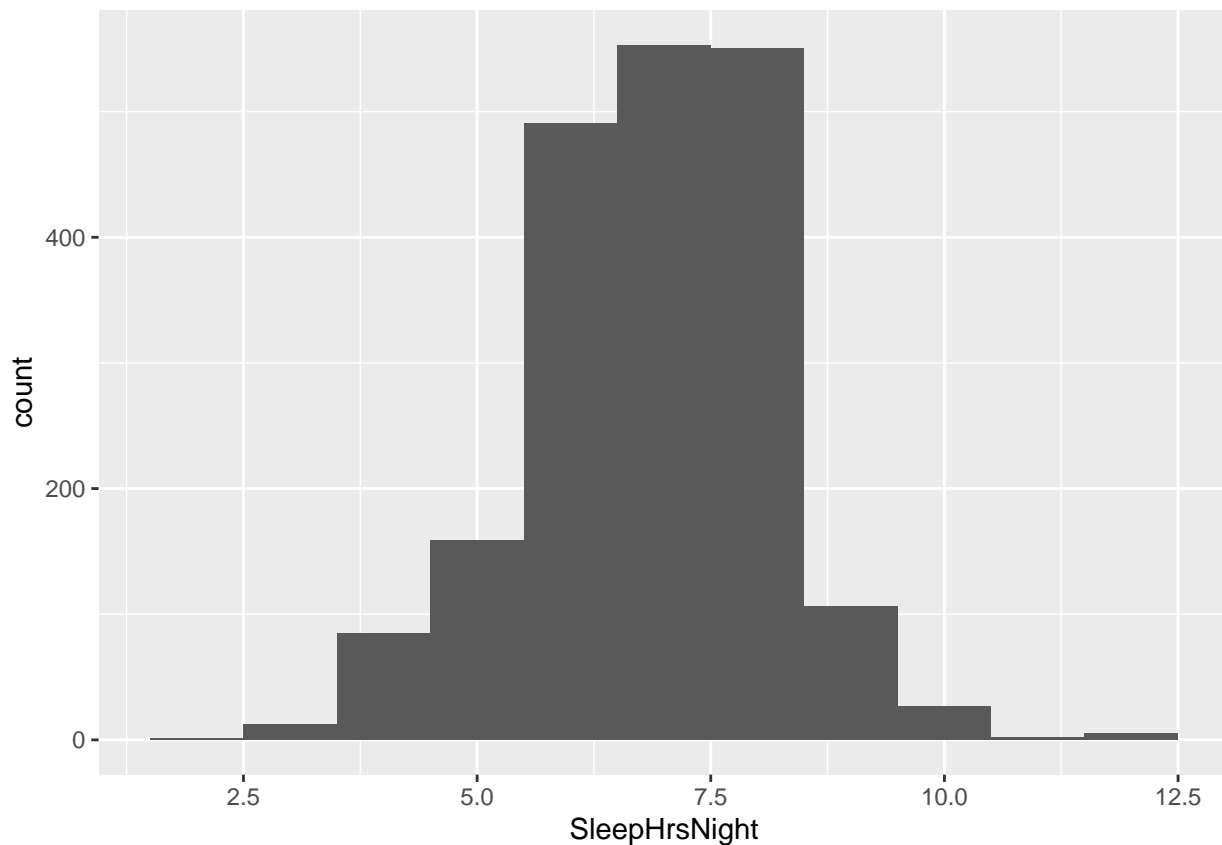
```r
# the sample mean should be a decent estimate for the population mean...
mean(~SleepHrsNight, data=NHANES_sleep)

# but how close is it to the right answer for the whole population?
# remember, this is just a survey...
# a very well-designed survey, but a survey nonetheless!

# let's try a few bootstrap samples
# by repeating these lines below.
# Looks like our bootstrap estimate is pretty close.
# Doesn't seem like a huge amount of sampling variability...
NHANES_sleep_bootstrap = resample(NHANES_sleep)
mean(~SleepHrsNight, data=NHANES_sleep_bootstrap)

# Let's be more systematic, with 10,000 bootstrap samples
boot_sleep = do(10000)*mean(~SleepHrsNight, data=resample(NHANES_sleep))

# What does this boot_sleep object look like?
# Let's examine the first several lines: one column called mean
head(boot_sleep)

# bootstrap sampling distribution
ggplot(boot_sleep) +
  geom_histogram(aes(x=mean))
```
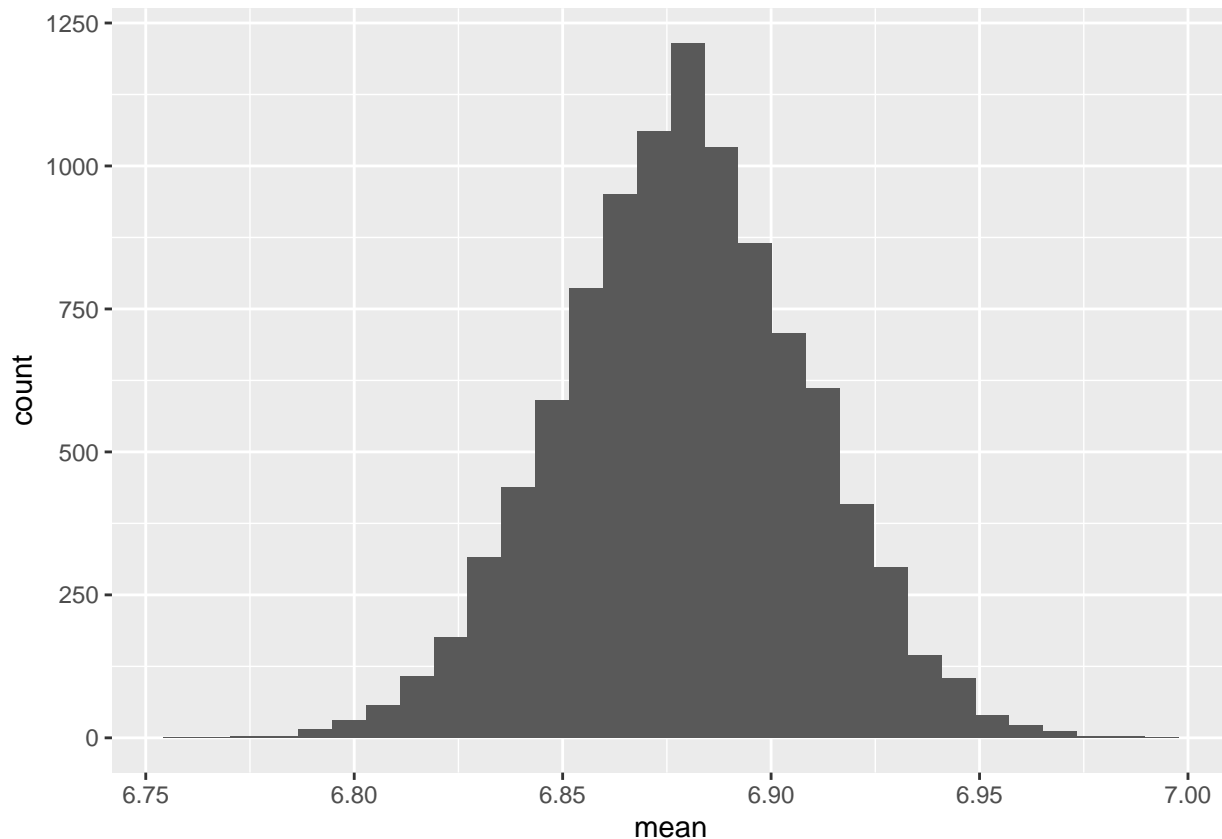
```
# how spread out is the sampling distribution?
sd(~mean, data=boot_sleep)

# So it looks like a typical sampling error is about 0.03 hours,
# or roughly 2 minutes.

# a confidence interval: a range of plausible values,
# based on the sampling distribution
confint(boot_sleep, level=0.95)

# can we eyeball these lower and upper endpoints of the
# confidence interval from the histogram?


###
# Example 2: depression
###

# quick exploratory analysis
NHANES_sleep %>%
  group_by(Depressed) %>%
  summarize(count = n())

# let's make a DepressedAny variable
NHANES_sleep = NHANES_sleep %>%
  mutate(DepressedAny = ifelse(Depressed != "None", yes=TRUE, no=FALSE))
```

```
prop(~DepressedAny, data=NHANES_sleep)

# How precisely does this survey result characterize
# the frequency of depression among all Americans?
# Let's bootstrap to understand the likely magnitude of
# estimation error due to sampling variability

boot_depression = do(10000)*prop(~DepressedAny, data=resample(NHANES_sleep))

# sanity check: single column called prop_TRUE
head(boot_depression)

# histogram of the 10,000 different estimates
# for prop_TRUE
ggplot(boot_depression) +
  geom_histogram(aes(x=prop_TRUE))
```
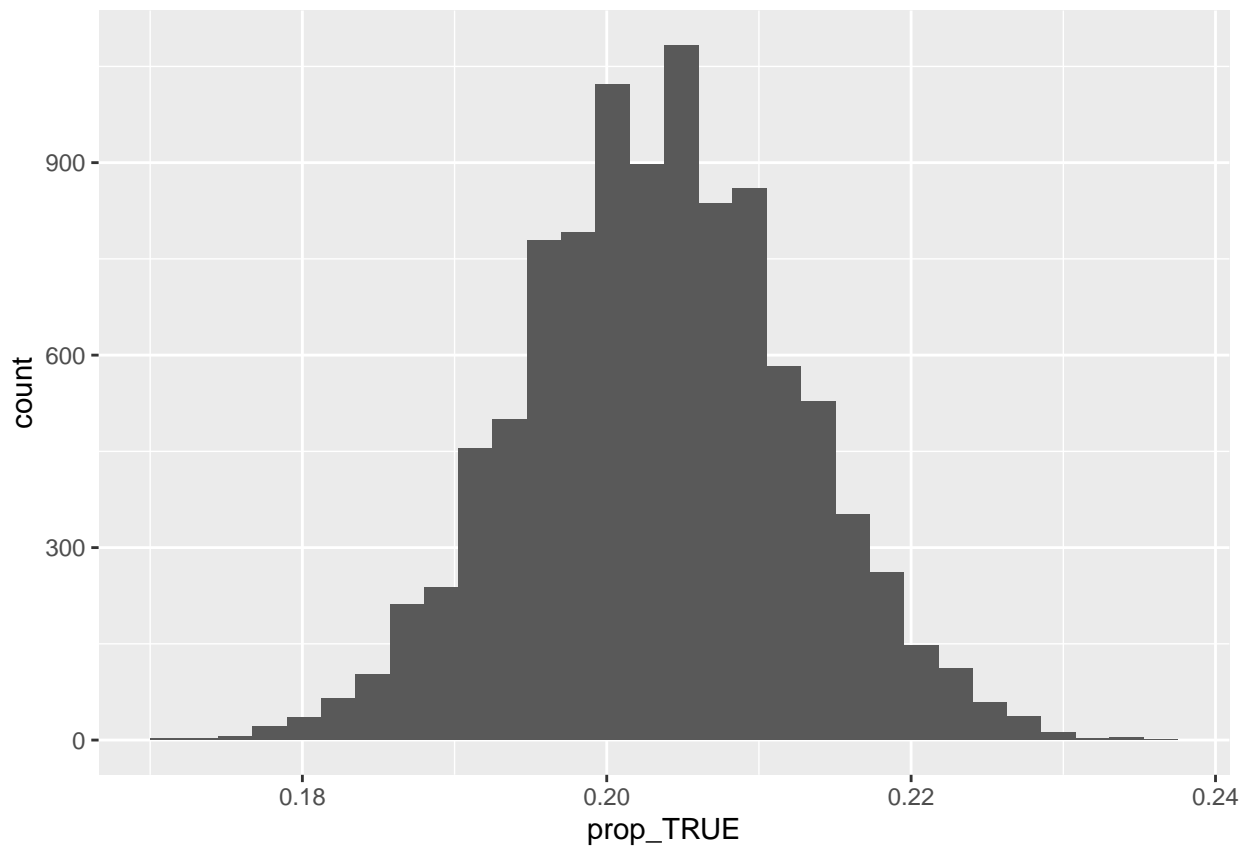


```
# a 95% confidence interval for the population proportion
confint(boot_depression, level = 0.95)

# again, can we eyeball this from the histogram?



###
# Smoke100
```

```
###
```

```
prop(~Smoke100, data=NHANES_sleep)
boot_smoking = do(10000)*prop(~Smoke100, data=resample(NHANES_sleep))
confint(boot_smoking, level=0.95)
```

Code 13

```
# In a 25-game stretch of the 2014-15 NFL seasons,
# the New England Patriots won 19 coin tosses, for a
# winning percentage of 76%. The Patriots have a
# reputation for ethical lapses. Could they have cheated
# at the coin toss, too?

# What does a random set of 25 coin flips look like?
# Answer: like a binomial distribution!

# This simulates the number of heads we'd get
# in a single set of 25 fair coin flips
rbinom(1, size=25, prob=0.5)

# repeat the simulation 50 times
rbinom(50, size=25, prob=0.5)

# repeat 10000 times and store in a vector
sim_flips = rbinom(10000, size=25, prob=0.5)

# Visualize the distribution
hist(sim_flips)
```
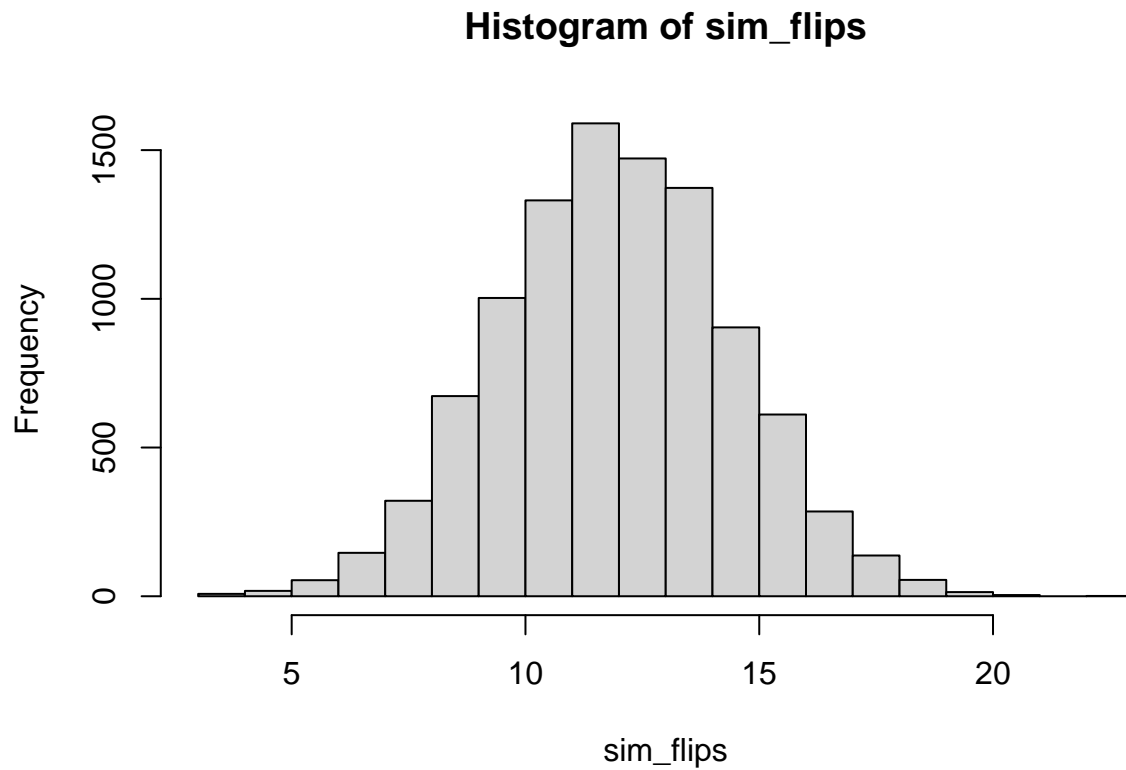
Histogram of sim_flips


```r
# How many are 19 or larger?
sum(sim_flips >= 19)
sum(sim_flips >= 19)/10000

# What is the actual probability under the binomial distribution?
# P(X >= 19) = 1 - P(X <= 18)
pbinom(18, 25, 0.5)
1 - pbinom(18, 25, 0.5)

# or equivalently:
pbinom(18, 25, 0.5, lower.tail=FALSE)
sum(dbinom(19:25, 25, 0.5))
```
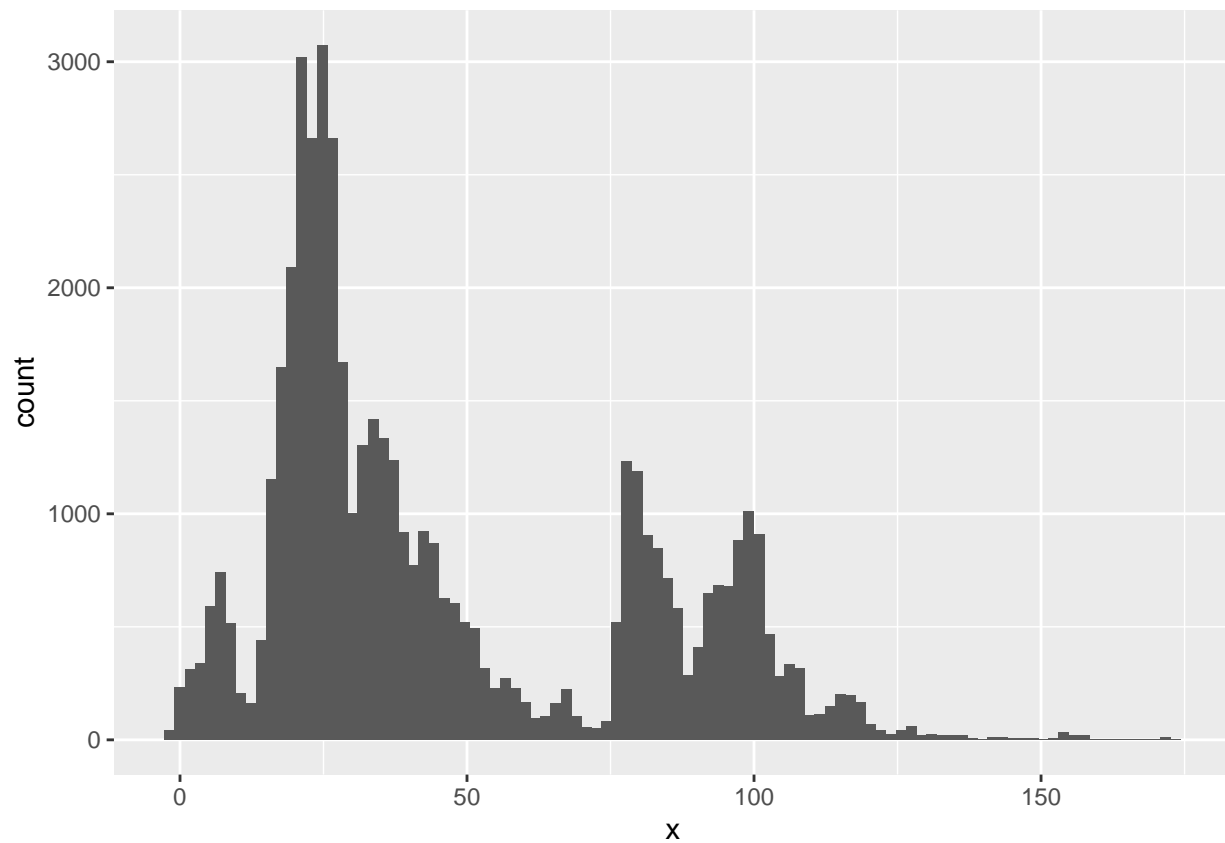
Code 14

```r
###
# Part 1: what happens when N gets bigger?
###

# read in a crazy population in crazypop2.csv
crazypop2 = read.csv('../data/raw data/crazypop2.csv')

# these individual data points are from some weird distribution
ggplot(crazypop2) +
  geom_histogram(aes(x=x), bins=100)
```
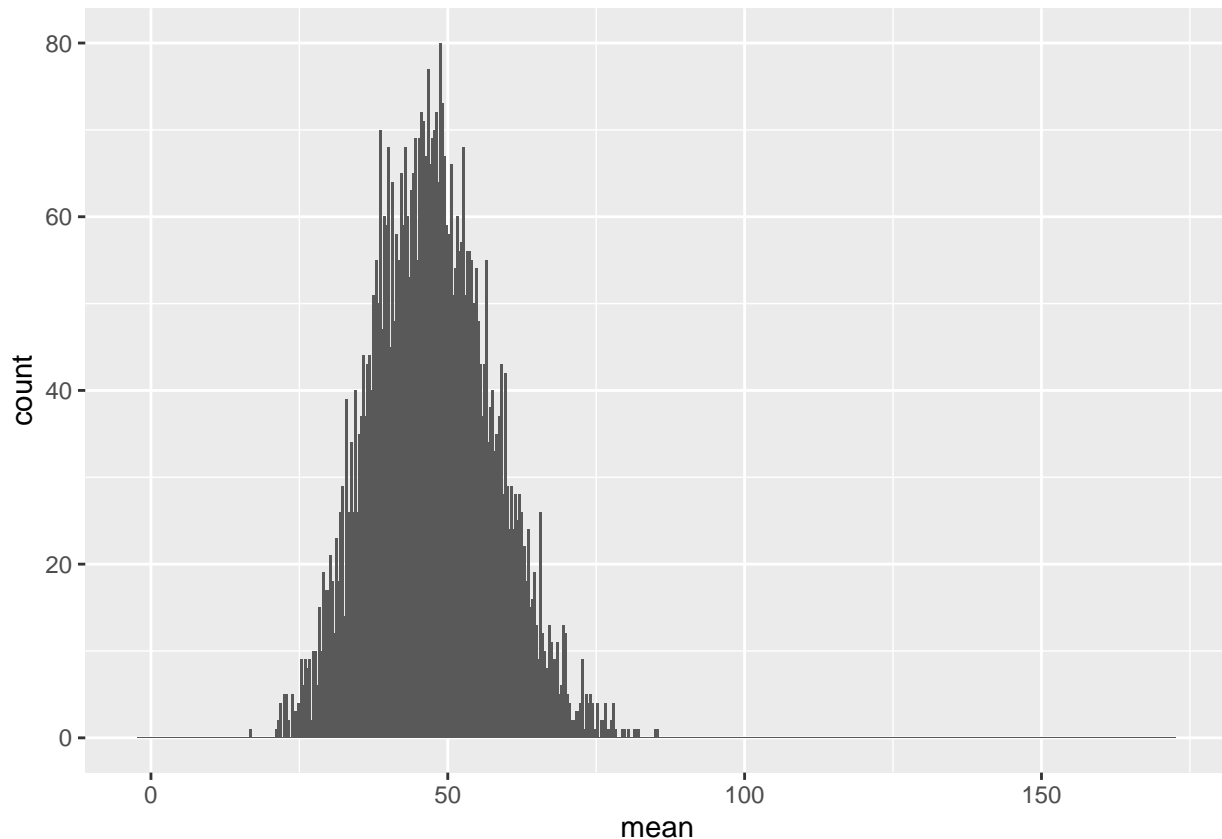
```
# let's see what happens when we repeatedly sample from crazypop2
N = 10 # size of each sample (we'll try 2, 5, 25, and 100)
sim2 = do(5000)*mean(~x, data=sample(crazypop2, size=N))

ggplot(sim2) +
  geom_histogram(aes(x=mean), bins=500) +
  xlim(range(crazypop2))
```

```
# What happens as N grows?
# 1) the distribution of the means gets narrower
# 2) the sampling distribution looks more and more like a normal distribution


###
# part 2: de Moivre's equation
###

# Let's compare the theoretical std error
# with the actual std error from our Monte Carlo sim

# possible values of n
n_grid = c(2, 5, 10, 20, 30, 50, 75, 100, 150, 200, 500, 1000)


# how many cores to use?
detectCores()
registerDoParallel(6)

# outer loop
stderr_grid = foreach(N = n_grid, .combine='c') %dopar% {
  # inner loop: simulate the sampling distribution of the mean
  sim_this_n = do(10000, parallel=FALSE)*mean(~x, data=sample(crazypop2, size=N))
  stderr_xbar = sd(~mean, data=sim_this_n)
  stderr_xbar
}
```
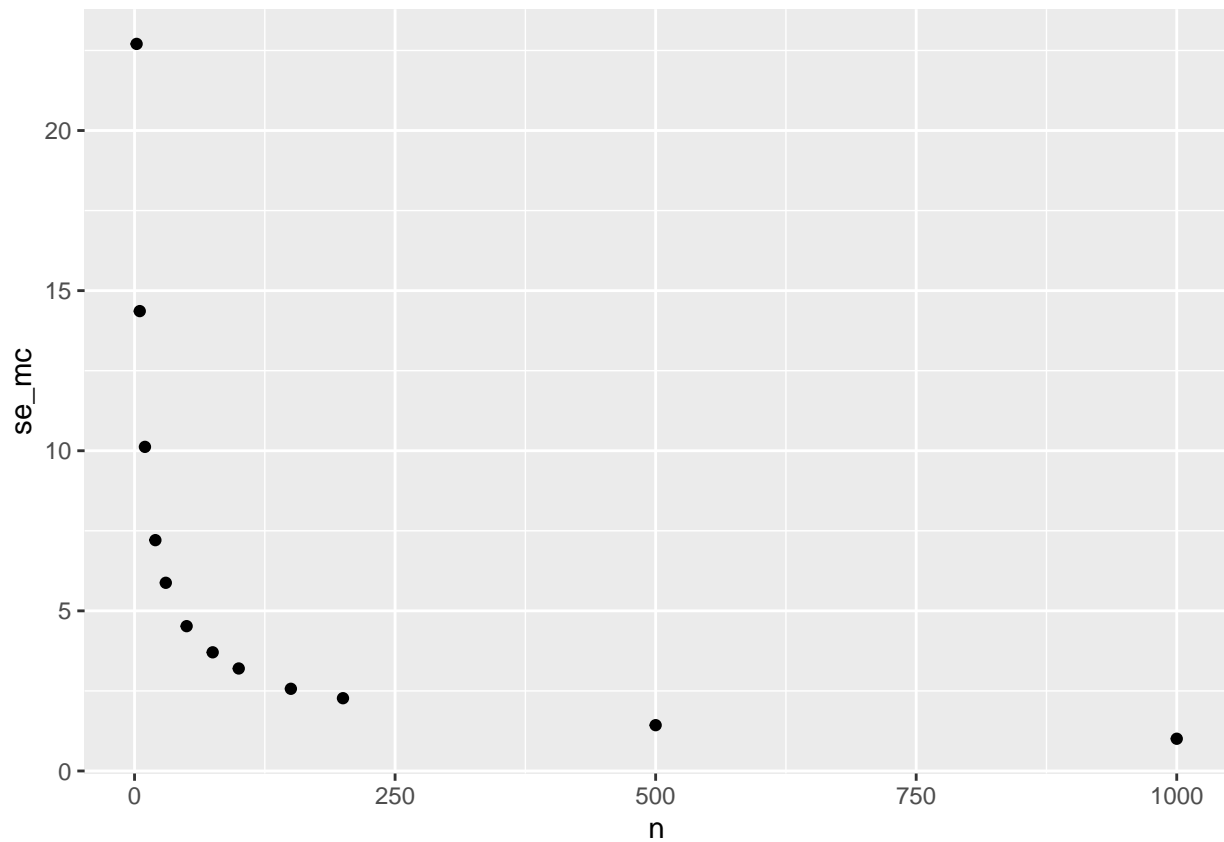
```
# What would de Moivre predict?
sigma = sd(~x, data=crazypop2)
sigma/sqrt(n_grid)

demoivre_grid = tibble(n = n_grid,
                       se_mc = stderr_grid,
                       se_theory = sigma/sqrt(n_grid))

# standard errors from simulation
p0 = ggplot(demoivre_grid) +
  geom_point(aes(x=n, y=se_mc))
p0
```
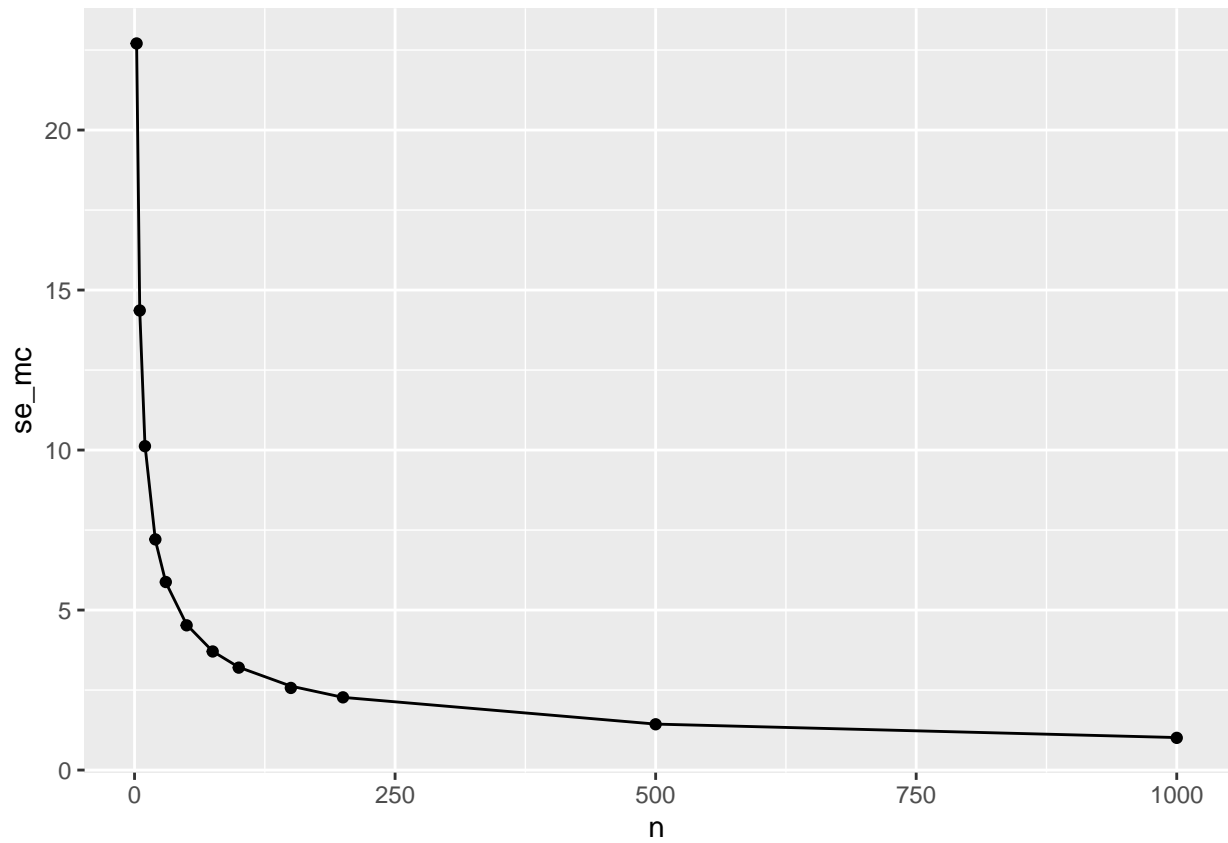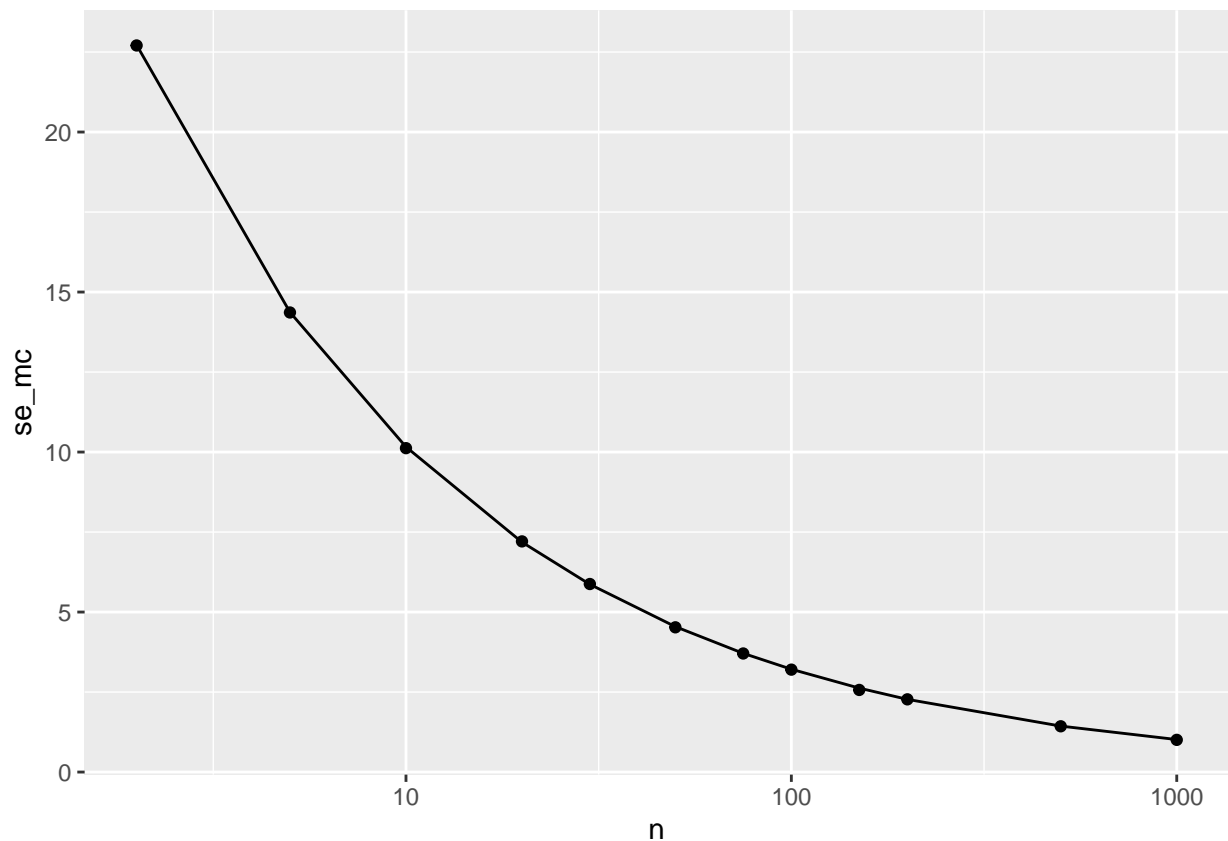


```
# add the theoretical std errs from de Moivre:
p0 + geom_line(aes(x=n, y=se_theory))
```

```
p0 + geom_line(aes(x=n, y=se_theory)) + scale_x_log10()
```

```
###
# Part 3: CLT with de Moivre's equation
###

# population parameters
mu = mean(~x, data=crazypop2)
mu

sigma = sd(~x, data=crazypop2)
sigma

# let's use de Moivre's equation + CLT to predict what the
# sampling distribution of the mean should look like.
N = 5 # size of each sample (we'll try 5, 25, and 100)
sim2 = do(5000)*mean(~x, data=sample(crazypop2, size=N))

# de Moivre's equation:
stderr_thisN = sigma/sqrt(N)

# actual sampling distribution
p1 = ggplot(sim2) +
  geom_histogram(aes(x=mean, y=..density..), bins=30)
p1
```
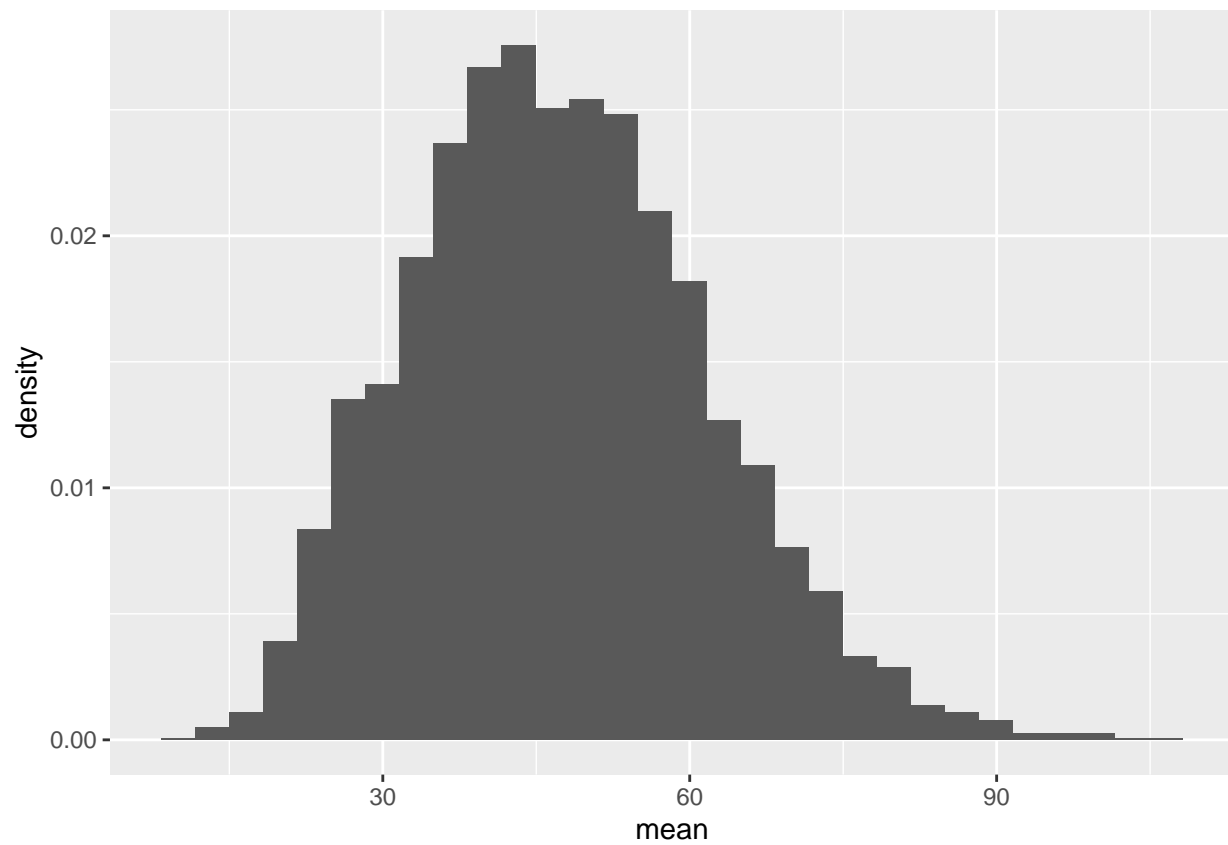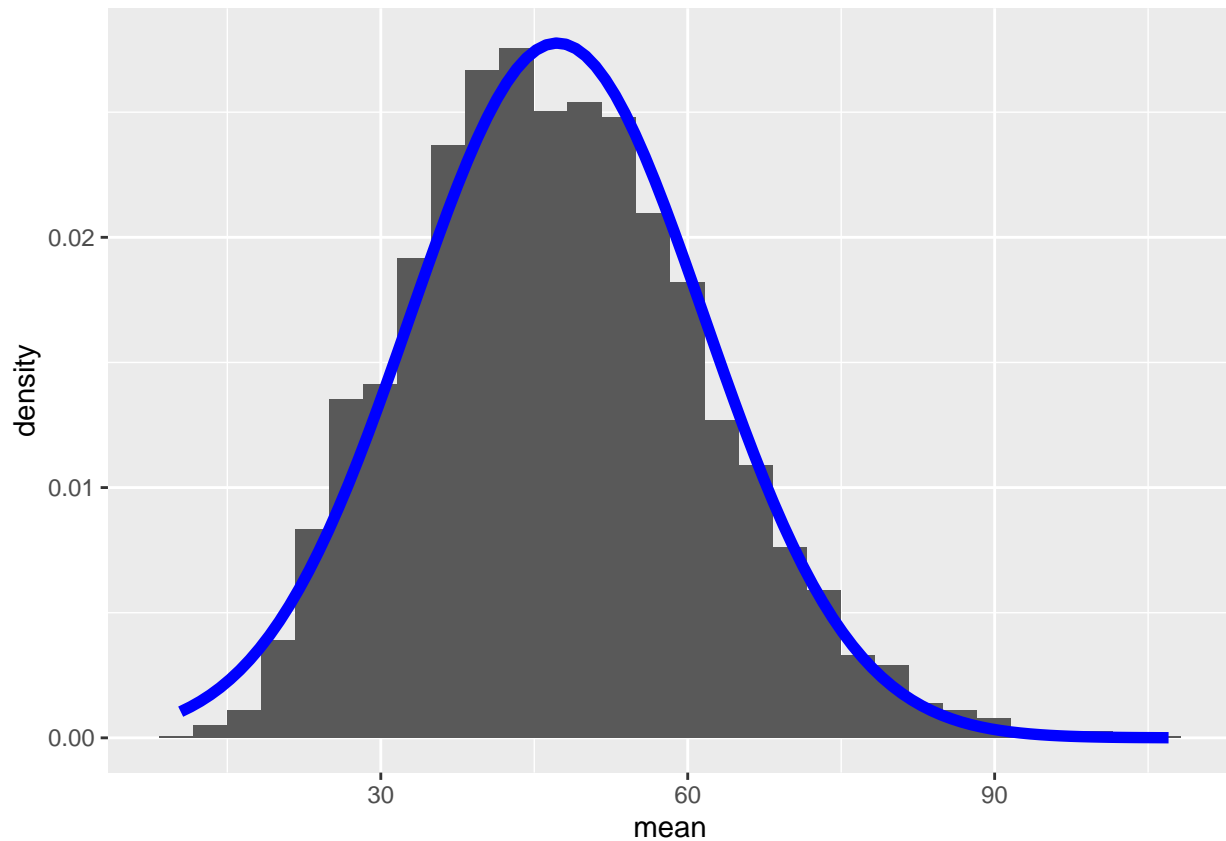
```r
# this is the normal density predicted by de Moivre + CLT
mydens = function(x) {
  dnorm(x, mu, stderr_thisN)
}

# sampling distribution + prediction
p1 + stat_function(fun = mydens, col='blue', size=2)
```

Code 15

```r
mean(~SleepHrsNight, data=NHANES_sleep)

# de Moivre's equation
nrow(NHANES_sleep)
sd(~SleepHrsNight, data=NHANES_sleep)

# standard error of the mean
1.32/sqrt(1991)

# compare with bootstrapping
boot_sleep = do(5000)*mean(~SleepHrsNight, data=resample(NHANES_sleep))

# calculate bootstrapped standard error: really close
sd(~mean, data=boot_sleep)

# confidence interval?

# go out 2 standard errors to either side of our estimate from the sample
6.88 - 2 * 1.32/sqrt(1991)
6.88 + 2 * 1.32/sqrt(1991)

# compare with bootstrapping
```

```r
confint(boot_sleep, level = 0.95)

# Take-home lesson: you get a confidence interval that's
# basically identical to what you get when you bootstrap,
# except using math rather than computational muscle.
```