I. More normal probabilities

A machine produces nails whose lengths are normally distributed with mean $\mu = 2$ inches and standard deviation $\sigma = 0.015$ inches.

- What proportion of nails produced by the machine are larger than 2.03 inches?
- Let $X$ stand for the length of a randomly selected nail.

\[
P(X > 2.03) = P\left(Z > \frac{2.03 - 2}{0.015}\right)
\]

where $Z$ is standard normal.

- Look up

\[
\frac{2.03 - 2}{0.015} = 2
\]

in Table B.2 to find $P(Z \leq 2) = .9772$.

So the answer is $1 - .9772 = .0228$. 

What proportion of nails are within 3 standard deviations of the mean?

We want

\[ P(2 - 0.045 \leq X \leq 2 + 0.045). \]

This is the same as

\[ P(1.955 \leq X \leq 2.045). \]

Standardize by subtracting 2 and dividing by 0.015 to convert this into

\[ P(-3 \leq Z \leq 3). \]

Use Table B.2 to find the answer, 0.9987 \(- 0.0013 = 0.9974. \)
• Fill in the blank: Exactly 20% of the nails greater than \underline{\underline{8}} inches in length.
• We need the 80th percentile.
• Find the 80th percentile of the standard normal: 0.84.
• The answer is

\[
(0.84)(0.015) + 2 = 2.0126.
\]
II. Sampling Distributions (Chapter 5)

- In 1898 Simon Newcomb performed an experiment to measure the speed of light.
  - Light source at his lab.
  - Mirror on Washington monument (about 2 miles away)
  - Detector at his lab.
  - By measuring the time it took for light to travel to the mirror and back, he could compute the speed of light.

- Newcomb took 66 such measurements.

**Question:** Why not just 1 measurement?

**Answer:** He knew there was potential error (variability) in the measurements. (Can you think of some sources of error?)
• The mean of the 66 measurements, 298072100 meters per second, is a reasonable estimate of the true speed of light.

• **Important question:** How “accurate” is this estimate? Can we be sure that it’s within 1000 m/s of the true speed?
• Want to compare Tylenol and Advil for relieving headache pain.

• Randomly assign each of 25 headache sufferers to take either Advil or Tylenol for headaches for 6 weeks.

• Ask each whether the pain relief was adequate or not.

• A total of 15 took Tylenol. Of these 6 said yes.

• A total of 10 took Advil. Of these, 6 said yes.

• Proportion saying yes:
  – For Tylenol, 6/15 = 0.4.
  – For Advil, 6/10 = 0.6.

• **Important Questions:** Is this enough evidence to conclude that Advil is better than Tylenol for headache sufferers? How much faith should we have in the 20% difference in effectiveness?
• In both examples, there is variability in the data.
  – In the speed of light data, because of various uncontrollable sources of error.
  – In the headache data, because different people will respond differently to the medications.
• To answer the “Important” questions, we will use probability models to model the variability. This is what the book means by “sampling distributions.”
• In the first case we’re using a sample mean.
• In the second case we’re using sample proportions.
**Intuition:**

- More data is better.
  - We’d get more information if we had more than 25 people in the headache study.
  - We’d get a better estimate of the speed of light if we collected more than 66 measurements.

- Less variability is better.
  - We’d get a better estimate of the speed of light if we could refine the experiment to produce measurements with less error.

- Much of what we’ll learn just refines this intuition!
More about variability

- Often variability arises because we choose a sample \textit{at random} from a population.
- This variability is good and bad:
- Bad because it reduces the “accuracy” of our estimators, as we’ve seen.
- Good because without it, we’d be in worse shape:
  - Want to estimate the mean height of STT 201 students.
  - Method 1: Choose 10 students at random and use their mean height.
  - Method 2: Choose the 10 students in the front row and use their mean height.
  - Method 2 eliminates variability (as long as students don’t move) since we’ll always get the same sample mean.
  - But the randomization in Method 1 is designed to get a sample that is representative of the population.
Notation and structure

- Let $y_1, \ldots, y_n$ stand for the individual observations.
  For example, $y_1, \ldots, y_{66}$ could stand for Newcomb’s 66 measurements.
- Let $\bar{y}$ stand for the mean of the data:
  $$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}.$$
- Before we collect the data, the observations and hence $\bar{y}$ are random variables.
- Our goal is to understand the probability distribution of $\bar{y}$. 
Throughout we’ll assume that the observations $y_1, \ldots, y_n$ are independent and all have the same probability distribution.

- In the usual “sampling from a population” model this is justified if we’re taking a random sample from the population.
- In an example like Newcomb’s measurements, this is justified if the experiment is not changing over time.
- Potential problem in sampling model: Population changing while sampling proceeds.
- Potential problem in Newcomb example: Experimenter gets better at measurements over time.
III. The distribution of \( \bar{y} \) (section 5.2)

*First important fact:* If the population distribution is normal with mean \( \mu \) and standard deviation \( \sigma \), then the distribution of \( \bar{y} \) is normal with mean \( \mu \) and standard deviation \( \sigma / \sqrt{n} \).
Example:

- Model the weights of 5 year olds by a normal density with mean $\mu = 20$ kg and standard deviation $\sigma = 3$ kg.
- **Question:** Choose one child at random. What’s the probability that his weight is within 3 of the mean weight 20?
- **Answer:** Let $X$ represent the child’s weight.

  - Then $X$ has a normal distribution with $\mu = 20$ and $\sigma = 3$.
  - We want $P(17 < X < 23)$. Standardize:
    \[
P \left( \frac{17 - 20}{3} < \frac{X - 20}{3} < \frac{23 - 20}{3} \right)
    \]
    to convert this into
    \[
P(-1 < Z < 1)
    \]
    where $Z$ is standard normal.
  - Use Table B.2 to compute this probability, namely, 0.6826.
• **Question:** Choose \( n = 9 \) children at random. What is the probability that the mean of their weights is within 3 of the population mean weight 20?

• **Answer:** Let \( \bar{X} \) stand for the mean of the 9 weights.

• Then \( \bar{X} \) has a normal distribution with mean 20 and standard deviation \( \sigma / \sqrt{n} = 3 / \sqrt{9} = 1. \)

• We want \( P(17 < \bar{X} < 23) \). Standardize:

\[
P \left( \frac{17 - 20}{1} < \frac{\bar{X} - 20}{1} < \frac{23 - 20}{1} \right)
\]

to convert this into

\[P(-3 < Z < 3)\]

where \( Z \) is standard normal.

• Use Table B.2 to compute this probability, namely, 0.9974.
Example: (Somewhat harder!) Newcomb’s experiment

- Assume Newcomb’s measurements are normally distributed with mean \( \mu \) equal to the (unknown!) speed of light and \( \sigma = 150000 \text{ m/s} \).

- Question: What’s the probability that the result of one observation will be within 10000 of the true speed of light?

- Answer: Let \( X \) be the observation. We want

\[
P(|X - \mu| < 10000)
\]

which is equal to

\[
P(-10000 < X - \mu < 10000).
\]

- Divide by \( \sigma = 150000 \):

\[
P \left( \frac{-10000}{150000} < \frac{X - \mu}{150000} < \frac{10000}{150000} \right)
\]
• We’ve standardized $X$! So this is the same as

$$P(-0.07 < Z < 0.07)$$

• Find the answer 0.0558 using Table B.2.
• **Question:** What’s the probability that his estimate based on \( n = 66 \) observations is within 10000 of the true speed of light?

• **Answer:** Let \( \overline{X} \) be the estimate. We want

\[
P(|\overline{X} - \mu| < 10000)
\]

which is equal to

\[
P(-10000 < \overline{X} - \mu < 10000).
\]

• Divide by the standard deviation of \( \overline{X} \), which is \( \frac{150000}{\sqrt{66}} \approx 18464 \).

\[
P \left( \frac{-10000}{18464} < \frac{\overline{X} - \mu}{18464} < \frac{10000}{18464} \right)
\]

• We’ve standardized \( \overline{X} \)! So this is the same as

\[
P(-0.54 < Z < 0.54)
\]

• Find the answer 0.411 using Table B.2.
Non-normal populations

- So far we only know the distribution of $\bar{y}$ when the population is normally distributed.
- But many populations are not:
  - Skewed populations
  - Discrete populations
  - Etc.
**Second important fact:** If the population has mean $\mu$ and standard deviation $\sigma$, and the sample size $n$ is large, then the distribution of $\bar{y}$ is approximately normal with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$.

- **Differences:**
  - Normal population: $\bar{y}$ is exactly normally distributed no matter what the sample size.
  - Non-normal population: Need the sample size $n$ to be large for $\bar{y}$ to be approximately normally distributed.

- **Similarities:**
  - $\bar{y}$ always has mean $\mu$
  - $\bar{y}$ always has standard deviation $\sigma/\sqrt{n}$.

- **How large must $n$ be?**
  - No universal answer, but $n$ around 30 is often sufficient.
Example:

- In 1798, Henry Cavendish performed an experiment to estimate the density of the earth (in multiples of the density of water). For this problem we’ll assume that his measurements have standard deviation $\sigma = 0.35$.
- He obtained $n = 29$ measurements.
- The mean $\bar{y}$ of his measurements is his estimate of the true density $\mu$. 
• **Question:** What’s the probability that the estimate is within 0.1 of the true density?

• **Answer:** We know that the distribution of the sample mean $\bar{X}$ is approximately normal with mean $\mu$ and standard deviation $0.35/\sqrt{29}$, which is about 0.065.

• We want to know $P(-0.1 < \bar{X} - \mu < 0.1)$.

• Divide by the standard deviation: We want

$$P \left( \frac{-0.1}{0.065} < \frac{\bar{X} - \mu}{0.065} < \frac{-0.1}{0.065} \right)$$

• This is $P(-1.54 < Z < 1.54)$ which from Table B.2 is about 0.8764.

• **Important point:** If the sample size were very small (e.g. $n = 4$ measurements) then the above method would not work!
Summary

Take a random sample of size $n$ from a population with mean $\mu$ and standard deviation $\sigma$. Then

- The sample mean $\bar{y}$ has mean $\mu$.
- The sample mean $\bar{y}$ has standard deviation $\sigma/\sqrt{n}$.
- If the population distribution is normal, the distribution of $\bar{y}$ is normal.
- If the population distribution is not normal, then
  - If $n$ is small, we don’t know how to proceed.
  - If $n$ is large, we can use the normal density to approximate the distribution of $\bar{y}$.