Limit theorems

Last modified: 2026-11-07 21:34:35 (PST)

The Central Limit Theorem

> Remark

Remark. The sample mean of many independent random variables, none of which dominates the others, has a distribution that is approximately bell-shaped, whatever the distributions of the individual variables. Theorem 0.1 makes this precise.

Theorem 0.1 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be IID random variables with population mean μ and finite variance $\sigma^2 > 0$, and let $\bar{X}_n \stackrel{def}{=} \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for every real number z [1]:*

$$\lim_{n\to\infty} \Pr\left(\frac{\sqrt{n}\,(\bar{X}_n - \mu)}{\sigma} \le z\right) = \Phi(z)$$

where $\Phi(z) \stackrel{def}{=} \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\,du$ is the CDF of the standard normal distribution $\mathrm{N}(0,1)$.

> Remark

Remark. This version is the Lindeberg–Lévy CLT; its proof is beyond these notes' scope [2], Theorem 27.1. Other versions relax the IID assumption, which is why the informal statement asks only that no summand dominate.

Corollary 0.1 (Central Limit Theorem for a sum)**.** *Let $X_1, X_2, \ldots$ be IID random variables with population mean μ and finite variance $\sigma^2 > 0$, and let $S_n \stackrel{def}{=} \sum_{i=1}^{n} X_i$. Then for every real number z:*

$$\lim_{n\to\infty} \Pr\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le z\right) = \Phi(z)$$

Proof

Proof. Since $S_n = n\bar{X}_n$,

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{n(\bar{X}_n - \mu)}{\sigma\sqrt{n}} = \frac{\sqrt{n}\,(\bar{X}_n - \mu)}{\sigma},$$

so the event in this corollary is the event in Theorem 0.1, and the limits agree. □

Theorem 0.2 (Mean and variance of a sum of IID random variables)**.** *Let $X_1, \ldots, X_n$ be IID random variables, each discrete or continuous, with population mean μ and finite variance σ^2, and let $S_n \stackrel{def}{=} \sum_{i=1}^n X_i$. Then, by linearity of expectation for finite sums and the variance of a linear combination:*

$$\mathbb{E}[S_n] = n\mu$$
$$\mathrm{Var}(S_n) = n\sigma^2$$

Proof

Proof. For the mean, apply linearity of expectation for finite sums with every $a_i = 1$:

$$\begin{aligned}
\mathbb{E}[S_n] &= \mathbb{E}\left[\sum_{i=1}^n 1 \cdot X_i\right] && (1 \cdot v = v) \\
&= \sum_{i=1}^n 1 \cdot \mathbb{E}[X_i] && \text{(linearity of expectation for finite sums, all } a_i = 1) \\
&= \sum_{i=1}^n \mathbb{E}[X_i] && (1 \cdot v = v) \\
&= n\mu && \text{(each } X_i \text{ has mean } \mu)
\end{aligned}$$

For the variance, apply the variance of a linear combination with every $a_i = 1$. For $i \neq j$, X_i and X_j are independent (take $A_k = \mathbb{R}$ for every other k in the definition of independence), so $\mathrm{Cov}(X_i, X_j) = 0$ for independent summands; and $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i)$ (covariance of a variable with itself):

$$
\begin{aligned}
\operatorname{Var}(S_n) &= \sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{Cov}(X_i, X_j) && \text{(variance of a linear combination, all } a_i = 1\text{)} \\
&= \sum_{i=1}^{n}\operatorname{Cov}(X_i, X_i) + \sum_{\substack{i,j=1 \\ i\neq j}}^{n}\operatorname{Cov}(X_i, X_j) && \text{(split off the terms with } i = j\text{)} \\
&= \sum_{i=1}^{n}\operatorname{Var}(X_i) + 0 && \text{(covariance with itself; independent summands)} \\
&= n\sigma^2 && \text{(each } X_i \text{ has variance } \sigma^2\text{)}
\end{aligned}
$$

□

Corollary 0.2 (Mean and variance of a sample mean of IID random variables)**.** *Let $X_1, \ldots, X_n$ be IID random variables, each discrete or continuous, with population mean μ and finite variance σ^2, and let $\bar{X}_n \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} X_i$. Then:*

$$\mathbb{E}\left[\bar{X}_n\right] = \mu$$
$$\operatorname{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

Proof

Proof. Let $S_n \overset{\text{def}}{=} \sum_{i=1}^{n} X_i$, so $\bar{X}_n = S_n/n$. By Theorem 0.2, $\mathbb{E}[S_n] = n\mu$ and $\operatorname{Var}(S_n) = n\sigma^2$:

$$
\begin{aligned}
\mathbb{E}\left[\bar{X}_n\right] &= \mathbb{E}\left[\frac{S_n}{n}\right] && \text{(definition of sample mean)} \\
&= \frac{1}{n}\mathbb{E}[S_n] && \text{(linearity of expectation)} \\
&= \frac{1}{n}\cdot n\mu && \text{(mean of a sum of IID variables)} \\
&= \mu && \text{(cancel } n\text{)} \\
\operatorname{Var}(\bar{X}_n) &= \operatorname{Var}\left(\frac{S_n}{n}\right) && \text{(definition of sample mean)} \\
&= \frac{1}{n^2}\operatorname{Var}(S_n) && \text{(variance of a constant multiple)} \\
&= \frac{1}{n^2}\cdot n\sigma^2 && \text{(variance of a sum of IID variables)} \\
&= \frac{\sigma^2}{n} && \text{(cancel one factor of } n\text{)}
\end{aligned}
$$

□

Remark

Remark. In practice, Theorem 0.1 justifies approximating $\bar{X}_n$ by a normal distribution with mean μ and variance σ^2/n, and Corollary 0.1 justifies approximating S_n by a normal distribution with mean $n\mu$ and variance $n\sigma^2$. By Theorem 0.2, these are exactly the means and variances of $\bar{X}_n = S_n/n$ and S_n.

Example 0.1 (The sum of five dice)**.** A single fair die roll has the discrete uniform distribution on $\{1, \dots, 6\}$, which is flat, not bell-shaped (Figure 1). Its mean is $\mu = 3.5$, and its variance is

$$\begin{aligned}\sigma^2 &= \sum_{x=1}^{6} (x - 3.5)^2/6 \\ &= 35/12.\end{aligned}$$

```
dice_sum_pmf <- function(n_dice) {
  totals <- rowSums(expand.grid(rep(list(1:6), n_dice)))
  probs <- prop.table(table(totals))
  tibble::tibble(total = as.numeric(names(probs)), p = as.vector(probs))
}

dice_plot <-
  ggplot2::ggplot(mapping = ggplot2::aes(x = total, y = p)) +
  ggplot2::xlab("sum of dice (x)") +
  ggplot2::ylab("Probability of outcome, Pr(X=x)") +
  ggplot2::expand_limits(y = 0)

dice_plot + ggplot2::geom_col(data = dice_sum_pmf(1))
```

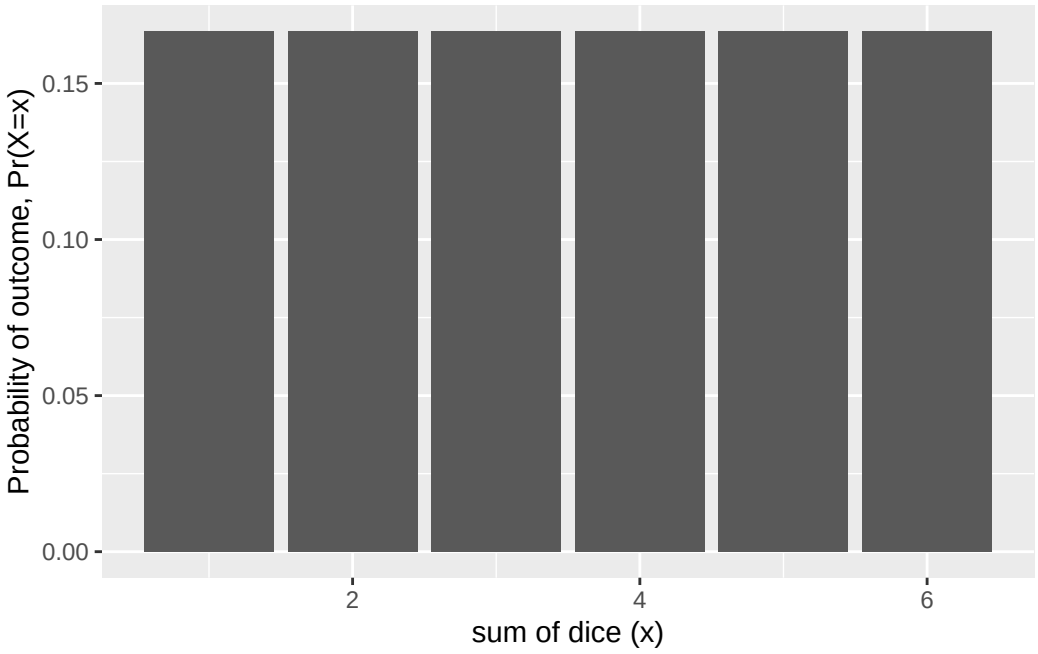


Figure 1: Distribution of the outcome of one die

The sum of five independent rolls, S_5, is already close to bell-shaped (Figure 2).

```
pmf_five_dice <- dice_sum_pmf(5)
dice_plot + ggplot2::geom_col(data = pmf_five_dice)
```

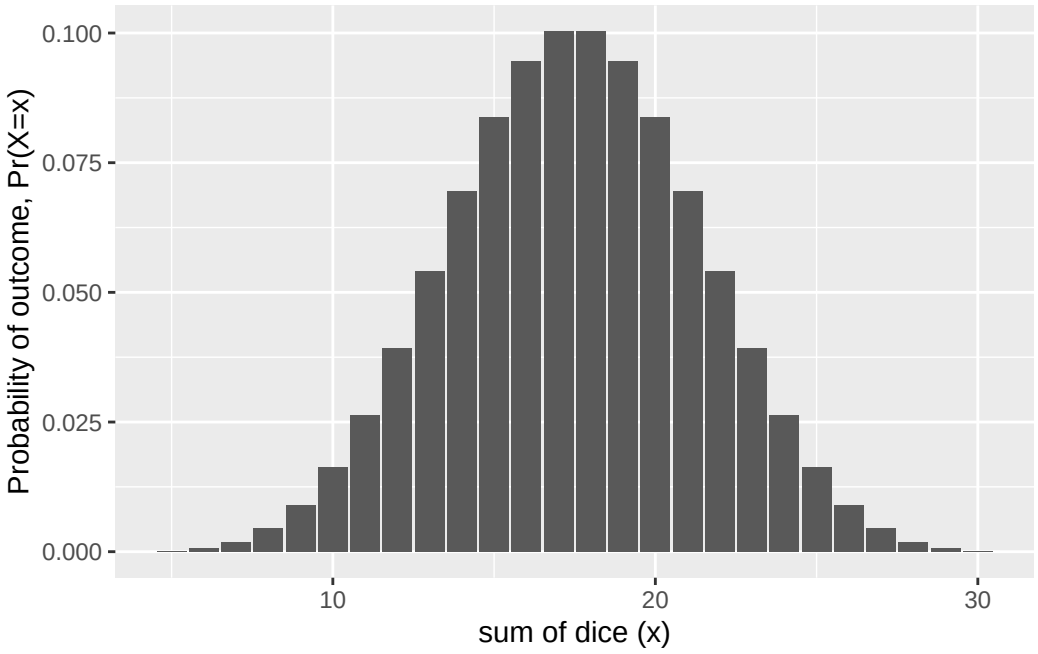


Figure 2: Distribution of the sum of five dice

For example, the exact probability that five dice total at most 15 is $\Pr(S_5 \le 15) = 0.3052$. The normal approximation that Corollary 0.1 suggests, with mean $5 \cdot 3.5 = 17.5$ and variance $5 \cdot 35/12 \approx 14.58$, evaluated at 15.5 to account for S_5 taking only integer values, gives $\Phi\big((15.5 - 17.5)/\sqrt{14.58}\big) \approx 0.3002$.

Example 0.2 (Sample means of car horsepower)**.** The Auto data record the horsepower and fuel economy of 392 car models sold between 1970 and 1982. They come from the StatLib library at Carnegie Mellon University and ship with the book *An Introduction to Statistical Learning*, in the Python package `ISLP` (`ISLP.load_data("Auto")`) and the R package `ISLR2` (`ISLR2::Auto`). This site has a copy at `data/auto.csv`, and both code tabs below read that file.

Horsepower is not bell-shaped: a few powerful cars give it a long right tail. To apply Theorem 0.1, treat each draw of one car at random as one IID observation X_i, so that μ and σ^2 are the population mean and variance of the 392 horsepower values. The sample mean $\bar{X}_n = S_n/n$, where $S_n = \sum_{i=1}^{n} X_i$, has mean μ and standard deviation $\sigma/\sqrt{n}$, because S_n has mean $n\mu$ and variance $n\sigma^2$ by Theorem 0.2. The code below draws 10,000 random samples of each size $n \in \{1, 5, 30\}$, with replacement, and computes each sample's sample mean.

Python

```
import numpy as np
import pandas as pd

hp = pd.read_csv("data/auto.csv")["horsepower"].to_numpy(dtype=float)
mu, sigma = hp.mean(), hp.std(ddof=0)

rng = np.random.default_rng(571)
sizes = [1, 5, 30]
sample_means = {n: rng.choice(hp, size=(10_000, n)).mean(axis=1) for n in sizes}

def skewness(x):
    return np.mean(((x - x.mean()) / x.std()) ** 3)

pd.DataFrame({
    "n": sizes,
    "sd of sample means": [sample_means[n].std(ddof=1) for n in sizes],
    "sigma / sqrt(n)": [sigma / np.sqrt(n) for n in sizes],
    "skewness of sample means": [skewness(sample_means[n]) for n in sizes],
})
#>     n  sd of sample means  sigma / sqrt(n)  skewness of sample means
#> 0   1           38.691521        38.442033                  1.080767
#> 1   5           17.265666        17.191800                  0.522243
#> 2  30            6.971269         7.018523                  0.164940
```

R

```r
hp <- read.csv("data/auto.csv")$horsepower
mu <- mean(hp)
sigma <- sqrt(mean((hp - mu)^2))

set.seed(571)
sizes <- c(1, 5, 30)
sample_means <- lapply(sizes, function(n) {
  replicate(10000, mean(sample(hp, n, replace = TRUE)))
})

skewness <- function(x) mean(((x - mean(x)) / sqrt(mean((x - mean(x))^2)))^3)

data.frame(
  n = sizes,
  sd_of_sample_means = sapply(sample_means, sd),
  sigma_over_root_n = sigma / sqrt(sizes),
  skewness_of_sample_means = sapply(sample_means, skewness)
)
#>    n sd_of_sample_means sigma_over_root_n skewness_of_sample_means
#> 1  1          38.611630         38.442033                1.0696654
#> 2  5          16.987684         17.191800                0.4804001
#> 3 30           7.057655          7.018523                0.1991387
```

Figure 3 shows a histogram of the 10,000 sample means for each n, with the normal density $\mathrm{N}(\mu, \sigma^2/n)$ from Theorem 0.1 drawn over it.

```r
n_labels <- paste("n =", sizes)
means_long <- tibble::tibble(
  n = factor(
    rep(n_labels, each = length(sample_means[[1]])),
    levels = n_labels
  ),
  sample_mean = unlist(sample_means)
)
normal_curves <- seq_along(sizes) |>
  lapply(function(k) {
    se <- sigma / sqrt(sizes[k])
    grid <- seq(mu - 4 * se, mu + 4 * se, length.out = 200)
    tibble::tibble(
      n = factor(n_labels[k], levels = n_labels),
      sample_mean = grid,
      density = dnorm(grid, mu, se)
    )
  }) |>
  dplyr::bind_rows()

ggplot2::ggplot(means_long, ggplot2::aes(x = sample_mean)) +
  ggplot2::geom_histogram(
    ggplot2::aes(y = ggplot2::after_stat(density)),
    bins = 40,
    fill = "grey70",
    color = "white"
  ) +
  ggplot2::geom_line(
    ggplot2::aes(y = density),
    data = normal_curves,
    linewidth = 1
  ) +
  ggplot2::facet_wrap(ggplot2::vars(n), scales = "free") +
  ggplot2::labs(x = "sample mean of horsepower", y = "density") +
  ggplot2::theme_minimal()
```

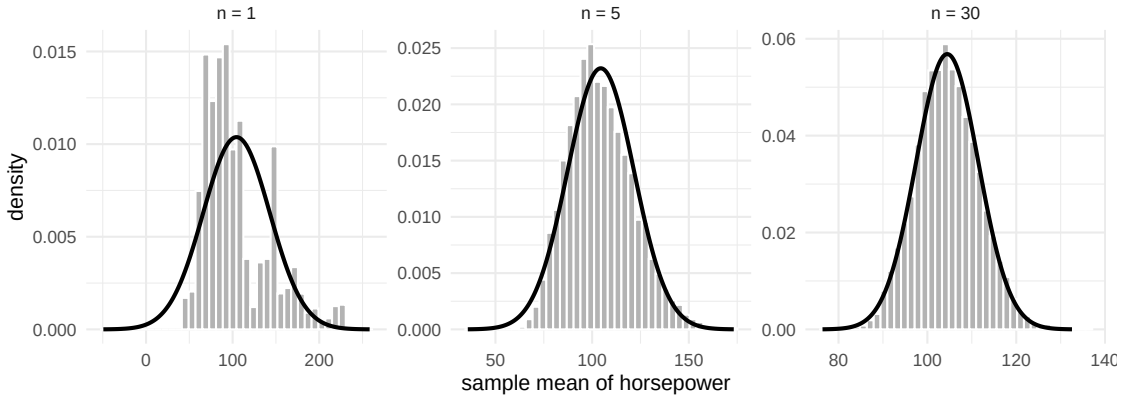


Figure 3: Histograms of 10,000 sample means of horsepower, for samples of n = 1, 5 and 30 cars drawn with replacement from the Auto data. The curve is the normal density with mean mu and standard deviation sigma over the square root of n.

The sample means for $n = 1$ are just single cars, so they are skewed. By $n = 30$ the histogram is close to the normal curve. The standard deviation of the sample means is within 1.2% of $\sigma/\sqrt{n}$ at every n, as Theorem 0.2 predicts. The skewness of the sample means, which is zero for a normal distribution, falls from 1.07 at $n = 1$ to 0.48 at $n = 5$ and 0.2 at $n = 30$.

i Watch

5MinuteAI explains limit theorems and gives a proof of the CLT, following the University of Washington's CSE 312 [3].

Tail bounds

Theorem 0.1 (Markov's inequality)**.** *Let X be a random variable with $X \geq 0$ and $\mathbb{E}[X] < \infty$. Then for every $a > 0$ [4]:*

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Proof

Proof. Because $X \geq 0$, every outcome satisfies $X \geq a \cdot \mathbb{1}_{\{X \geq a\}}$: when $X \geq a$ the right side is $a \leq X$, and otherwise it is $0 \leq X$. The expectation of a nonnegative random variable is nonnegative, so by linearity of expectation, and because the expectation of an indicator is the probability of its event,

$$\mathbb{E}[X] - a \cdot \Pr(X \geq a) = \mathbb{E}\left[X - a \cdot \mathbb{1}_{\{X \geq a\}}\right] \geq 0$$

and dividing by $a > 0$ gives the claim. □

Remark

Remark. The inequality carries Markov's name by convention. Chebyshev's inequality, which follows from it, is older [5], [6]; see [7] for the history of the attribution. A modern treatment is in [8].

Theorem 0.2 (Chebyshev's inequality)**.** *Let X be a random variable with mean μ and finite variance $\sigma^2 > 0$. Then for every $k > 0$ [9]:*

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Proof

Proof. The random variable $(X - \mu)^2$ is nonnegative, and its expectation is σ^2. Because $|X - \mu| \geq k\sigma$ exactly when $(X - \mu)^2 \geq k^2\sigma^2$, Theorem 0.1 gives

$$\Pr(|X - \mu| \geq k\sigma) = \Pr((X - \mu)^2 \geq k^2\sigma^2) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

□

Remark

Remark. Theorem 0.2 needs only a mean and a variance, not a distribution, so it holds for every distribution that has them. That generality makes it loose: for a normal distribution, the chance of landing more than two standard deviations from the mean is about 5%, but the bound at $k = 2$ is 25%.

Remark

Remark. Bienaymé derived this bound in 1853 [5], and Chebyshev proved a more general form in 1867 [6]. A modern treatment is in [8].

Example 0.1 (How loose is Chebyshev's bound for car horsepower?)**.** Take the 392 horsepower values in the Auto data described in Example 0.2 as the whole population, and draw one car at random. For each k, the code below counts the share of cars whose horsepower is at least k standard deviations from the mean, and sets it beside the bound $1/k^2$ from Theorem 0.2.

```
hp <- read.csv("data/auto.csv")$horsepower
mu <- mean(hp)
sigma <- sqrt(mean((hp - mu)^2))

k <- c(1.5, 2, 3)
share_beyond <- sapply(k, function(kk) mean(abs(hp - mu) >= kk * sigma))
data.frame(k = k, share_of_cars = share_beyond, chebyshev_bound = 1 / k^2)
#>     k share_of_cars chebyshev_bound
#> 1 1.5    0.09948980       0.4444444
#> 2 2.0    0.04336735       0.2500000
#> 3 3.0    0.01275510       0.1111111
```

The bound holds at every k, as Theorem 0.2 guarantees. At $k = 2$ the share of cars is 4.3%, against a bound of 25%, and at $k = 3$ it is 1.3% against 11.1%. The bound is loose because it must hold for every distribution with this mean and variance.

Exercise 0.1 (Chebyshev's bound for the mean fuel economy)**.** Let $X_1, \ldots, X_n$ be IID draws of the fuel economy (`mpg`) of a car chosen at random from the Auto data, with population mean μ and variance σ^2, and let $\bar{X}_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} X_i$.
Use Theorem 0.2 and Theorem 0.2 to show that for every $\epsilon > 0$,

$$\Pr\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$$

Solution

Solution 0.1. By Theorem 0.2, $\mathrm{Var}(S_n) = n\sigma^2$ for $S_n \stackrel{\text{def}}{=} \sum_{i=1}^{n} X_i$, so $\bar{X}_n = S_n/n$ has mean μ and variance σ^2/n (a constant factor c scales a variance by c^2). Apply Theorem 0.2 to $\bar{X}_n$ with $k\sigma/\sqrt{n} = \epsilon$, that is, $k = \epsilon\sqrt{n}/\sigma$, so $1/k^2 = \sigma^2/(n\epsilon^2)$.

Exercise 0.2 (Population variance of the fuel economy)**.** Compute σ^2 for the `mpg` column of `data/auto.csv`, treating the 392 cars as the whole population.

Solution

Solution 0.2.

```
mpg <- read.csv("data/auto.csv")$mpg
mu <- mean(mpg)
sigma2 <- mean((mpg - mu)^2)
sigma2
#> [1] 60.76274
```

The population variance of `mpg` is 60.8.

Exercise 0.3 (Smallest sample size from Chebyshev's bound)**.** With σ^2 from Exercise 0.2, find the smallest n for which the bound in Exercise 0.1 guarantees $\Pr\left(\left|\bar{X}_n - \mu\right| \geq 1\right) \leq 0.05$.

Solution

Solution 0.3. Setting $\sigma^2/(n \cdot 1^2) \leq 0.05$ gives $n \geq 20\sigma^2$.

```
n_min <- ceiling(20 * sigma2)
n_min
#> [1] 1216
```

With σ^2 from Solution 0.2, $n = 1216$ cars suffice.

Exercise 0.4 (Simulated misses of the sample mean)**.** Draw 10,000 samples of the size from Exercise 0.3 with replacement, and report how many of the sample means $\bar{X}_n$ miss μ by 1 or more.

Solution

Solution 0.4.

```r
# Uses mpg and mu from sol-chebyshev-variance,
# and n_min from sol-chebyshev-min-n.
set.seed(571)
xbar <- replicate(10000, mean(sample(mpg, n_min, replace = TRUE)))
misses <- sum(abs(xbar - mu) >= 1)
misses
#> [1] 0
```

Out of 10,000 samples of that size, 0 sample means missed μ by 1 or more, against the 5% (500 samples) the bound allows.

Exercise 0.5 (Why Chebyshev's bound is loose here)**.** Explain why the count in Exercise 0.4 is far below the 5% the bound allows.

Solution

Solution 0.5.

```r
normal_approx <- 2 * pnorm(-1 / sqrt(sigma2 / n_min))
normal_approx
#> [1] 7.694705e-06
```

In Solution 0.4, the count is far below the 5% (500 samples) the bound allows. The bound is loose because the sample mean of this many cars is nearly normal (Theorem 0.1), and the normal approximation puts the chance of a miss at about 7.7×10^{-6}. Chebyshev's inequality must hold for every distribution, as in Example 0.1.

Theorem 0.3 (Chebyshev's inequality for a sample mean)**.** *Let $X_1, \ldots, X_n$ be IID random variables with population mean μ and finite variance $\sigma^2 > 0$, and let $\bar{X}_n \stackrel{def}{=} \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for every $\epsilon > 0$:*

$$\Pr\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$$

Proof

Proof. This is the solution to Exercise 0.1, worked in Solution 0.1. It applies Theorem 0.2 to $\bar{X}_n$, whose variance is σ^2/n by Corollary 0.2. □

Lemma 0.1 (Hoeffding's lemma)**.** *Let X be a random variable with $0 \le X \le 1$ and mean μ. Then for every real number s,*

$$\mathbb{E}\left[e^{s(X-\mu)}\right] \le e^{s^2/8}.$$

Proof

Proof. First bound e^{sX} by a line. The function $x \mapsto e^x$ is convex, and for $0 \le X \le 1$ the number sX is the weighted average $(1-X)\cdot 0 + X \cdot s$, so

$$\begin{aligned} e^{sX} &\le (1-X)e^0 + Xe^s && \text{(convexity of the exponential)} \\ &= 1 - X + Xe^s && (e^0 = 1) \end{aligned}$$

Take the expectation of both sides, which preserves the inequality:

$$\begin{aligned} \mathbb{E}\left[e^{sX}\right] &\le 1 - \mathbb{E}[X] + \mathbb{E}[X]e^s && \text{(linearity of expectation)} \\ &= 1 - \mu + \mu e^s && \text{(definition of } \mu) \end{aligned}$$

The number $1 - \mu + \mu e^s$ is a weighted average of 1 and e^s with weights $1-\mu$ and μ, so it is positive. Multiply by the constant $e^{-s\mu}$ and write the result as an exponential:

$$\begin{aligned} \mathbb{E}\left[e^{s(X-\mu)}\right] &= e^{-s\mu}\mathbb{E}\left[e^{sX}\right] && (e^{-s\mu} \text{ is a constant}) \\ &\le e^{-s\mu}(1 - \mu + \mu e^s) && \text{(the bound above)} \\ &= e^{L(s)} && (\text{define } L(s) \stackrel{\text{def}}{=} -s\mu + \log(1 - \mu + \mu e^s)) \end{aligned}$$

It remains to show $L(s) \le s^2/8$. Write

$$p(s) \stackrel{\text{def}}{=} \frac{\mu e^s}{1 - \mu + \mu e^s},$$

which lies in $[0, 1]$. The first derivative of L is

$$\begin{aligned} L'(s) &= -\mu + \frac{\mu e^s}{1 - \mu + \mu e^s} && \text{(chain rule for the logarithm)} \\ &= -\mu + p(s) && \text{(definition of } p) \end{aligned}$$

The derivative of p is

$$\begin{aligned} p'(s) &= \frac{\mu e^s(1 - \mu + \mu e^s) - \mu e^s \cdot \mu e^s}{(1 - \mu + \mu e^s)^2} && \text{(quotient rule)} \\ &= \frac{\mu e^s(1-\mu)}{(1 - \mu + \mu e^s)^2} && (\text{the } \mu^2 e^{2s} \text{ terms cancel}) \\ &= p(s)(1 - p(s)) && \left(1 - p(s) = \tfrac{1-\mu}{1-\mu+\mu e^s}\right) \end{aligned}$$

so $L''(s) = p'(s) = p(s)(1-p(s))$. Completing the square gives

$$\begin{aligned} p(s)(1-p(s)) &= \frac{1}{4} - \left(p(s) - \frac{1}{2}\right)^2 && \text{(complete the square)} \\ &\leq \frac{1}{4} && \text{(a square is not negative)} \end{aligned}$$

At $s = 0$, $L(0) = 0 + \log 1 = 0$ and $L'(0) = -\mu + \mu = 0$. By Taylor's theorem with the Lagrange remainder, there is a number ξ between 0 and s with

$$\begin{aligned} L(s) &= L(0) + L'(0)\, s + \frac{L''(\xi)}{2} s^2 && \text{(Taylor's theorem)} \\ &= \frac{L''(\xi)}{2} s^2 && (L(0) = L'(0) = 0) \\ &\leq \frac{s^2}{8} && (L''(\xi) \leq \tfrac{1}{4}) \end{aligned}$$

Therefore $\mathbb{E}\left[e^{s(X-\mu)}\right] \leq e^{L(s)} \leq e^{s^2/8}$. □

Remark

Remark. Hoeffding's 1963 paper [10] controls the moment generating function $\mathbb{E}\left[e^{s(X-\mu)}\right]$ in a step that the literature calls Hoeffding's lemma.

Theorem 0.4 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be IID random variables with $0 \leq X_i \leq 1$ and population mean μ, and let $\bar{X}_n \stackrel{def}{=} \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for every $\varepsilon > 0$ [11],*

$$\Pr\left(\bar{X}_n - \mu \geq \varepsilon\right) \leq e^{-2n\varepsilon^2}. \tag{1}$$

Proof

Proof. Let the free parameter be $s > 0$. The event $\left\{\bar{X}_n - \mu \geq \varepsilon\right\}$ is the event $\left\{e^{sn(\bar{X}_n - \mu)} \geq e^{sn\varepsilon}\right\}$, because the exponential is increasing and $sn > 0$. The random variable $e^{sn(\bar{X}_n - \mu)}$ is positive, so Markov's inequality applies:

$$\begin{aligned} \Pr\left(\bar{X}_n - \mu \geq \varepsilon\right) &= \Pr\left(e^{sn(\bar{X}_n - \mu)} \geq e^{sn\varepsilon}\right) && \text{(the exponential is increasing)} \\ &\leq e^{-sn\varepsilon}\, \mathbb{E}\left[e^{sn(\bar{X}_n - \mu)}\right] && \text{(Markov's inequality)} \end{aligned}$$

Rewrite the exponent as a sum, so that the exponential becomes a product:

$$
\begin{aligned}
e^{sn(\bar{X}_n-\mu)} &= e^{s\sum_{i=1}^{n}(X_i-\mu)} && \left(n\bar{X}_n = \sum_{i=1}^{n} X_i\right) \\
&= \prod_{i=1}^{n} e^{s(X_i-\mu)} && \left(e^{a+b} = e^a e^b\right)
\end{aligned}
$$

The factors are functions of independent random variables, so they are independent, and the expectation of a product of independent random variables is the product of their expectations (the expectation of a product of independent variables, applied one factor at a time). Then apply Lemma 0.1 to each factor:

$$
\begin{aligned}
\mathbb{E}\left[e^{sn(\bar{X}_n-\mu)}\right] &= \prod_{i=1}^{n} \mathbb{E}\left[e^{s(X_i-\mu)}\right] && \text{(independence)} \\
&\le \prod_{i=1}^{n} e^{s^2/8} && \text{(Hoeffding's lemma)} \\
&= e^{ns^2/8} && (n \text{ equal factors})
\end{aligned}
$$

Combining the two displays, for every $s > 0$,

$$
\Pr\left(\bar{X}_n - \mu \ge \varepsilon\right) \le e^{-sn\varepsilon + ns^2/8}.
$$

The right side is smallest where the exponent $-sn\varepsilon + ns^2/8$ is smallest. The exponent is a quadratic in s with a positive leading coefficient, so it has one minimum, where its derivative is zero. Its derivative in s is $-n\varepsilon + ns/4$, which is zero at $s = 4\varepsilon$, and $4\varepsilon > 0$ is allowed. There the exponent is

$$
\begin{aligned}
-4\varepsilon \cdot n\varepsilon + \frac{n(4\varepsilon)^2}{8} &= -4n\varepsilon^2 + 2n\varepsilon^2 && \text{(substitute } s = 4\varepsilon) \\
&= -2n\varepsilon^2 && \text{(combine like terms)}
\end{aligned}
$$

which is the claim. □

Corollary 0.1 (Hoeffding's inequality, two-sided)**.** *Under the assumptions of Theorem 0.4, for every $\varepsilon > 0$,*

$$
\Pr\left(\left|\bar{X}_n - \mu\right| \ge \varepsilon\right) \le 2e^{-2n\varepsilon^2}. \tag{2}
$$

Proof

Proof. The variables $1 - X_1, \ldots, 1 - X_n$ are IID, lie in $[0, 1]$, and have mean $1 - \mu$. Their sample mean is $1 - \bar{X}_n$, so Theorem 0.4 applied to them gives

$$\begin{aligned}
\Pr\big(\mu - \bar{X}_n \geq \varepsilon\big) &= \Pr\big((1 - \bar{X}_n) - (1 - \mu) \geq \varepsilon\big) && \text{(rearrange the event)} \\
&\leq e^{-2n\varepsilon^2} && \text{(Hoeffding's inequality for } 1 - X_i\text{)}
\end{aligned}$$

The event $\big\{\big|\bar{X}_n - \mu\big| \geq \varepsilon\big\}$ is the union of $\big\{\bar{X}_n - \mu \geq \varepsilon\big\}$ and $\big\{\mu - \bar{X}_n \geq \varepsilon\big\}$. The probability of a union is at most the sum of the probabilities, so

$$\begin{aligned}
\Pr\big(\big|\bar{X}_n - \mu\big| \geq \varepsilon\big) &\leq \Pr\big(\bar{X}_n - \mu \geq \varepsilon\big) + \Pr\big(\mu - \bar{X}_n \geq \varepsilon\big) && \text{(union bound)} \\
&\leq e^{-2n\varepsilon^2} + e^{-2n\varepsilon^2} && \text{(Hoeffding's inequality, twice)} \\
&= 2e^{-2n\varepsilon^2} && \text{(combine like terms)}
\end{aligned}$$

□

Remark

Remark. Hoeffding stated the inequality for sums of independent bounded random variables, with different bounds for each term [10]. This page states the case of the sample mean of IID variables in $[0, 1]$, which is the case that bounds the gap between a sample proportion and the probability it estimates. For an indicator variable, μ is the probability of the event and $\bar{X}_n$ is the share of the n draws in which it occurs.

Example 0.2 (How loose is Hoeffding's bound for the share of efficient cars?)**.** Take the 392 cars in the Auto data described in Example 0.2 as the whole population, and call a car efficient if its fuel economy (`mpg`) is above 25. Let $X_i = 1$ if the ith car drawn at random is efficient and $X_i = 0$ otherwise, so μ is the share of efficient cars and $\bar{X}_n$ is the share in a sample of n draws. Each X_i lies in $[0, 1]$, and the draws are made with replacement, so the X_i are IID and Corollary 0.1 applies. The code below draws 10,000 samples of each size, computes the proportion of samples whose share misses μ by at least $\varepsilon = 0.1$, and sets that proportion beside the bound $2e^{-2n\varepsilon^2}$ of Equation 2 and the bound $\sigma^2/(n\varepsilon^2)$ that Chebyshev's inequality gives (Theorem 0.3).

```r
mpg <- read.csv("data/auto.csv")$mpg
efficient <- as.numeric(mpg > 25)
mu <- mean(efficient)
sigma2 <- mean((efficient - mu)^2)
tol <- 0.1

n <- c(50, 100, 200, 400)
set.seed(571)
share_missed <- sapply(n, function(nn) {
  xbar <- replicate(10000, mean(sample(efficient, nn, replace = TRUE)))
  mean(abs(xbar - mu) >= tol)
})
hoeffding_bound <- pmin(1, 2 * exp(-2 * n * tol^2))
chebyshev_bound <- pmin(1, sigma2 / (n * tol^2))
data.frame(
  n = n,
  share_missed = share_missed,
  hoeffding_bound = hoeffding_bound,
  chebyshev_bound = chebyshev_bound
)
#>     n share_missed hoeffding_bound chebyshev_bound
#> 1  50       0.1481    0.7357588823      0.47917534
#> 2 100       0.0408    0.2706705665      0.23958767
#> 3 200       0.0042    0.0366312778      0.11979384
#> 4 400       0.0000    0.0006709253      0.05989692
```

The simulated proportions lie below both bounds at every sample size, as they should: the bounds apply to the true probabilities these proportions estimate. The share of efficient cars is 0.398. For $n = 50$ and $n = 100$ Chebyshev's bound is the smaller of the two, but the Hoeffding bound falls exponentially in n while the Chebyshev bound falls only like $1/n$, so at $n = 200$ and $n = 400$ Hoeffding's is the smaller. At $n = 400$ the Hoeffding bound is 0.00067, against 0.06 for Chebyshev, and 0 of the 10,000 samples missed. Both bounds are loose because they must hold for every distribution of this kind, not only for this one.

The law of large numbers

Definition 0.1 (Convergence in probability)**.** A sequence of random variables $Y_1, Y_2, \ldots$ **converges in probability** to a constant c [12] if, for every distance $\varepsilon > 0$, the probability that Y_n lies at least ε away from c goes to zero as n grows:

$$\lim_{n\to\infty} \Pr(|Y_n - c| \geq \varepsilon) = 0$$

We write $Y_n \xrightarrow{p} c$.

Source

Casella and Berger [13], Def. 5.5.1.

Theorem 0.1 (Weak law of large numbers)**.** *Let $X_1, X_2, \ldots$ be IID random variables with population mean μ and finite variance $\sigma^2 > 0$, and let $\bar{X}_n \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean. Then the sample mean converges in probability to μ [14]:*

$$\bar{X}_n \xrightarrow{p} \mu$$

Source

Casella and Berger [13], Thm. 5.5.2.

Proof

Proof. By Corollary 0.2, $\mathbb{E}[\bar{X}_n] = \mu$ and $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$.
Let the tolerance be $\varepsilon > 0$ and let the multiplier be $k \overset{\text{def}}{=} \varepsilon\sqrt{n}/\sigma$, so that $k\sigma/\sqrt{n} = \varepsilon$. Chebyshev's inequality for $\bar{X}_n$, whose standard deviation is $\sigma/\sqrt{n}$, gives

$$\begin{aligned} 0 \leq \Pr(|\bar{X}_n - \mu| \geq \varepsilon) &= \Pr\left(|\bar{X}_n - \mu| \geq k\frac{\sigma}{\sqrt{n}}\right) && \text{(definition of } k\text{)} \\ &\leq \frac{1}{k^2} && \text{(Chebyshev's inequality)} \\ &= \frac{\sigma^2}{n\varepsilon^2} && \text{(substitute } k\text{)} \end{aligned}$$

The bound goes to zero as $n \to \infty$, so the probability, squeezed between zero and the bound, goes to zero. This holds for every $\varepsilon > 0$. □

References

[1] Wikipedia contributors, "Central limit theorem — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Central_limit_theorem

[2] P. Billingsley, *Probability and measure*, 3rd ed. in Wiley series in probability and mathematical statistics. New York: Wiley, 1995.

[3] 5MinuteAI, "Probability & statistics." Video playlist on YouTube. Accessed: Oct. 12, 2026. [Online]. Available: https://www.youtube.com/playlist?list=PLeB45KifGiuHesi4PALNZSYZFhViVGQJK

[4] Wikipedia contributors, "Markov's inequality — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Markov%27s_inequality

[5] I.-J. Bienaymé, "Considérations à l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrés," *Comptes Rendus de l'Académie des Sciences*, vol. 37, pp. 309–324, 1853, Available: https://www.numdam.org/item/JMPA_1867_2_12___158_0/

[6] P. L. Tchébychef, "Des valeurs moyennes," *Journal de mathématiques pures et appliquées*, vol. 12, pp. 177–184, 1867, Available: https://www.numdam.org/item/JMPA_1867_2_12___177_0/

[7] E. Seneta, "A tricentenary history of the law of large numbers," *Bernoulli*, vol. 19, no. 4, pp. 1088–1121, 2013, doi: 10.3150/12-BEJSP12.

[8] J. K. Blitzstein and J. Hwang, *Introduction to probability*, 2nd ed. Chapman; Hall/CRC, 2019. doi: 10.1201/9780429428357.

[9] Wikipedia contributors, "Chebyshev's inequality — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Chebyshev%27s_inequality

[10] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963, doi: 10.1080/01621459.1963.10500830.

[11] Wikipedia contributors, "Hoeffding's inequality — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Hoeffding%27s_inequality

[12] Wikipedia contributors, "Convergence of random variables — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Convergence_of_random_variables

[13] G. Casella and R. Berger, *Statistical inference*, 2nd ed. Cengage Learning, 2002. Available: https://www.cengage.com/c/statistical-inference-2e-casella-berger/9780534243128/

[14] Wikipedia contributors, "Law of large numbers — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Law_of_large_numbers