

GOVT 6019: Homework 5 — Solutions

Fall 2026 · Estimation and Large Samples

Melissa Sands (Instructor)

Izza Malik (Teaching Assistant)

```
library(tidyverse)
source("../..//scripts/gov6019_ggplot.R") # theme_gov() and course colors
set.seed(6019)
```

Question 1: Arrival times at a polling place

(a) $f(x) = 1/13$ for $0 < x < 13$ and 0 otherwise. The area under it is a rectangle of width 13 and height $1/13$, so it equals 1.

(b) $\mathbb{P}(X = 5) = 0$. X is continuous. The probability of an interval is the area above it, and a single point has width zero.

(c) The last two hours are $11 < X < 13$. By hand, $2/13 \approx .154$.

```
punif(13, 0, 13) - punif(11, 0, 13)
```

```
[1] 0.1538462
```

(d) $F(x) = 0$ for $x \leq 0$, $F(x) = x/13$ for $0 < x < 13$, and $F(x) = 1$ for $x \geq 13$. Setting $x/13 = .5$ gives a median of 6.5 hours.

```
qunif(0.5, 0, 13)
```

```
[1] 6.5
```

(e)

```
arrive <- runif(10000, 0, 13)
mean(arrive > 11)
```

```
[1] 0.1595
```

The simulated share is 0.160, against $2/13 = .154$.

(f) No. A probability cannot exceed 1. The height of a density is a probability per unit of x . A short interval of width h near x has probability close to $f(x)h$. For $\mathcal{U}(0, 0.5)$ the density is 2 on an interval of width 0.5, so the total area is still 1.

Question 2: County turnout

(a) $z = (75 - 60)/10 = 1.5$.

```
1 - pnorm(1.5)
```

```
[1] 0.0668072
```

```
pnorm(75, mean = 60, sd = 10, lower.tail = FALSE)
```

```
[1] 0.0668072
```

Both give 0.0668.

(b)

```
pnorm(80, 60, 10) - pnorm(40, 60, 10)
```

```
[1] 0.9544997
```

40 and 80 are two standard deviations either side of the mean. This is the “95” of the rule. The exact share is .9545.

(c) 50 to 70 is one standard deviation either side of 60: about 68 percent. 90 is three standard deviations above the mean. About 99.7 percent lie within three standard deviations, so about .3 percent lie outside, and by symmetry about .15 percent lie above 90.

(d)

```
qnorm(0.10, 60, 10)
```

```
[1] 47.18448
```

```
qnorm(c(0.025, 0.975), 60, 10)
```

```
[1] 40.40036 79.59964
```

The lowest 10 percent of counties have turnout below 47.2. The middle 95 percent lie between 40.4 and 79.6, which is $60 \pm 1.96 \times 10$.

(e) Write $Z = -\mu/\sigma + (1/\sigma)X$, which has the form $a + bX$ with $a = -\mu/\sigma$ and $b = 1/\sigma$. Then $\mathbb{E}[Z] = -\mu/\sigma + \mathbb{E}[X]/\sigma = -\mu/\sigma + \mu/\sigma = 0$. Adding a constant does not change a variance, so $\mathbb{V}[Z] = \mathbb{V}[X/\sigma] = \mathbb{V}[X]/\sigma^2 = \sigma^2/\sigma^2 = 1$.

Question 3: Interest and participation

(a) Each conditional expectation is a sum over the row for that level of interest:

$$\begin{aligned}\mathbb{E}[Y \mid X = \text{low}] &= 0(.70) + 1(.20) + 2(.10) = .40, \\ \mathbb{E}[Y \mid X = \text{medium}] &= 0(.40) + 1(.40) + 2(.20) = .80, \\ \mathbb{E}[Y \mid X = \text{high}] &= 0(.10) + 1(.40) + 2(.50) = 1.40.\end{aligned}$$

(b) $\mathbb{E}[Y \mid X]$ takes the value $\mathbb{E}[Y \mid X = x]$ whenever $X = x$. So it equals .40 with probability .40, .80 with probability .45, and 1.40 with probability .15. It equals 1.4 exactly when X is high, so $\mathbb{P}(\mathbb{E}[Y \mid X] = 1.4) = \mathbb{P}(X = \text{high}) = .15$.

(c) The outer expectation is taken over the distribution of X :

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid X]] = .40(.40) + .80(.45) + 1.40(.15) = .16 + .36 + .21 = .73.$$

(d) Each probability is a sum over the levels of interest: $\mathbb{P}(Y = y) = \sum_x \mathbb{P}(Y = y \mid X = x) \mathbb{P}(X = x)$.

$$\mathbb{P}(Y = 0) = .70(.40) + .40(.45) + .10(.15) = .280 + .180 + .015 = .475,$$

$$\mathbb{P}(Y = 1) = .20(.40) + .40(.45) + .40(.15) = .080 + .180 + .060 = .320,$$

$$\mathbb{P}(Y = 2) = .10(.40) + .20(.45) + .50(.15) = .040 + .090 + .075 = .205.$$

The three probabilities sum to 1. Then $\mathbb{E}[Y] = 0(.475) + 1(.320) + 2(.205) = .73$, which matches part (c).

```
p_x <- c(low = .40, medium = .45, high = .15)
p_y_given_x <- rbind(low = c(.70, .20, .10),
                      medium = c(.40, .40, .20),
                      high = c(.10, .40, .50))
cond_mean <- p_y_given_x %*% 0:2          # part (a)
sum(cond_mean * p_x)                    # part (c)
```

```
[1] 0.73
```

```
p_y <- colSums(p_y_given_x * p_x)        # part (d)
p_y
```

```
[1] 0.475 0.320 0.205
```

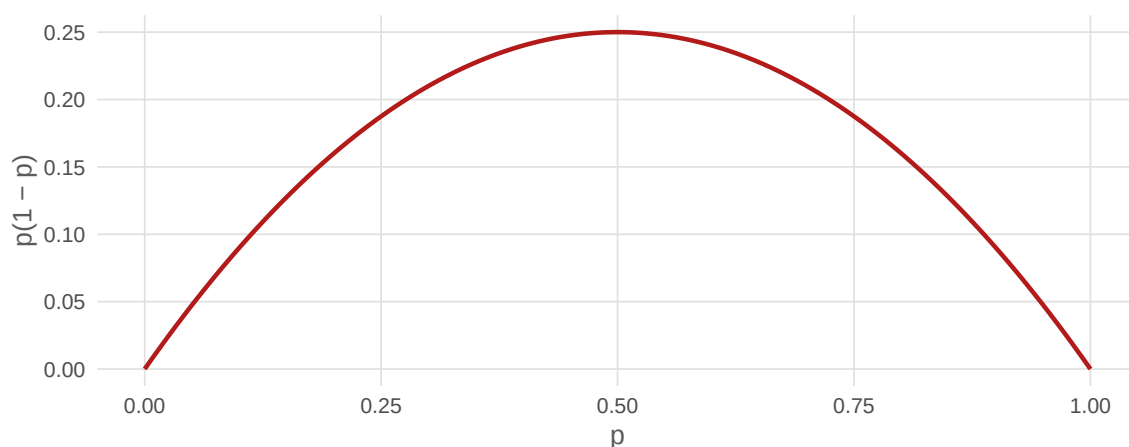
```
sum(p_y * 0:2)
```

```
[1] 0.73
```

Question 4: The variance of a sample proportion

(a)

```
tibble(p = seq(0, 1, length.out = 400), v = p * (1 - p)) |>
  ggplot(aes(p, v)) +
  geom_line(color = gov_colors[["carnelian"]], linewidth = 0.8) +
  labs(x = "p", y = "p(1 - p)") +
  theme_gov(base_size = 10)
```



The curve is largest at $p = .5$, where $p(1 - p) = .25$. It equals zero at $p = 0$ and at $p = 1$.

(b) Expand the right-hand side:

$$.25 - (p - .5)^2 = .25 - (p^2 - p + .25) = p - p^2 = p(1 - p).$$

The square $(p - .5)^2$ is at least zero for every p . So $p(1 - p) = .25 - (p - .5)^2 \leq .25$, with equality only at $p = .5$.

(c) $SD[\bar{X}] = \sigma/\sqrt{n} = \sqrt{p(1-p)}/\sqrt{n}$. By part (b), $\sqrt{p(1-p)} \leq \sqrt{.25} = .5$. So $SD[\bar{X}] \leq .5/\sqrt{n}$ at every p .

(d)

```
n_poll <- 1000
bound <- 0.5 / sqrt(n_poll)
true_sd <- sqrt(c(0.38, 0.05) * (1 - c(0.38, 0.05))) / n_poll)
c(bound = bound, p38 = true_sd[1], p05 = true_sd[2])
```

```
      bound      p38      p05
0.015811388 0.015349267 0.006892024
```

```
bound - true_sd
```

```
[1] 0.0004621212 0.0089193639
```

```
bound / true_sd
```

```
[1] 1.030107 2.294157
```

The bound is 0.0158. At $p = .38$ the true value is 0.0153, so the bound exceeds it by 0.0005 and is 1.03 times as large. At $p = .05$ the true value is 0.0069. There the bound exceeds it by 0.0089 and is 2.3 times as large. The bound is close near $p = .5$ and loose near 0 or 1.

(e)

```
props <- function(p) replicate(5000, mean(rbinom(n_poll, 1, p)))
sim <- c(p50 = sd(props(0.5)), p05 = sd(props(0.05)))
sim
```

```
      p50      p05
0.015823740 0.006893381
```

At $p = .5$ the simulated standard deviation is 0.0158, against the bound of 0.0158. At $p = .5$ the bound is the true value. At $p = .05$ it is 0.0069, against the true value of 0.0069 from part (d).

(f) Set $.5/\sqrt{n} = .01$. Then $\sqrt{n} = 50$ and $n = 2,500$.

Question 5: Rolling dice in R

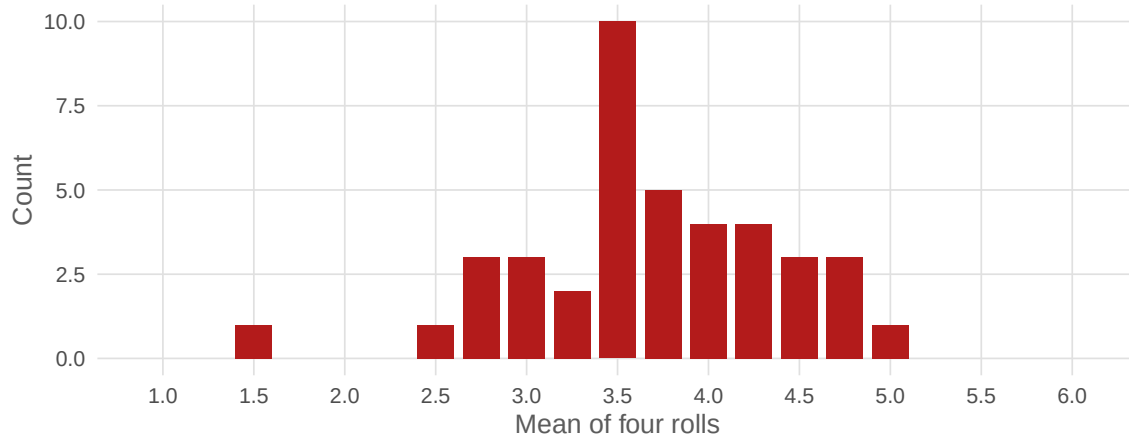
(a)

```
trial <- function() mean(sample(1:6, 4, replace = TRUE))
```

(b)

```
xbar40 <- replicate(40, trial())

tibble(xbar = xbar40) |>
  ggplot(aes(xbar)) +
  geom_bar(fill = gov_colors[["carnelian"]], width = 0.2) +
  scale_x_continuous(breaks = seq(1, 6, by = 0.5), limits = c(0.9, 6.1)) +
  labs(x = "Mean of four rolls", y = "Count") +
  theme_gov(base_size = 10)
```



```
in_band <- function(x) x >= 2.75 & x <= 4.25
sum(in_band(xbar40))
```

```
[1] 31
```

In this run, 31 of the 40 means fall in the band. Your count will differ unless you used the same seed.

(c) $\bar{x}$ moves in steps of 0.25. The band $3.5 \pm 0.854 = [2.646, 4.354]$ therefore contains exactly the values 2.75, 3.00, ..., 4.25.

```
outcomes <- expand_grid(r1 = 1:6, r2 = 1:6, r3 = 1:6, r4 = 1:6) |>
  mutate(xbar = (r1 + r2 + r3 + r4) / 4)
nrow(outcomes)
```

```
[1] 1296
```

```
p_band <- mean(in_band(outcomes$xbar))
p_band
```

```
[1] 0.6820988
```

```
pnorm(1) - pnorm(-1)
```

```
[1] 0.6826895
```

884 of the 1,296 outcomes fall in the band, so the exact probability is $884/1296 = 0.682$. A normal distribution places 0.683 within one standard deviation of its mean. The two agree to two decimal places, but the close match depends on where the band's edges fall. $\bar{X}$ takes values in steps of 0.25.

Any band $3.5 \pm w$ with $0.75 \leq w < 1$ contains the same values, 2.75 to 4.25, so the exact probability is .682 for every such w . Over the same range of w , the share a normal distribution places within w of its mean runs from .62 to .76. The band of one standard deviation, $w = 0.854$, falls at the point where the two happen to agree.

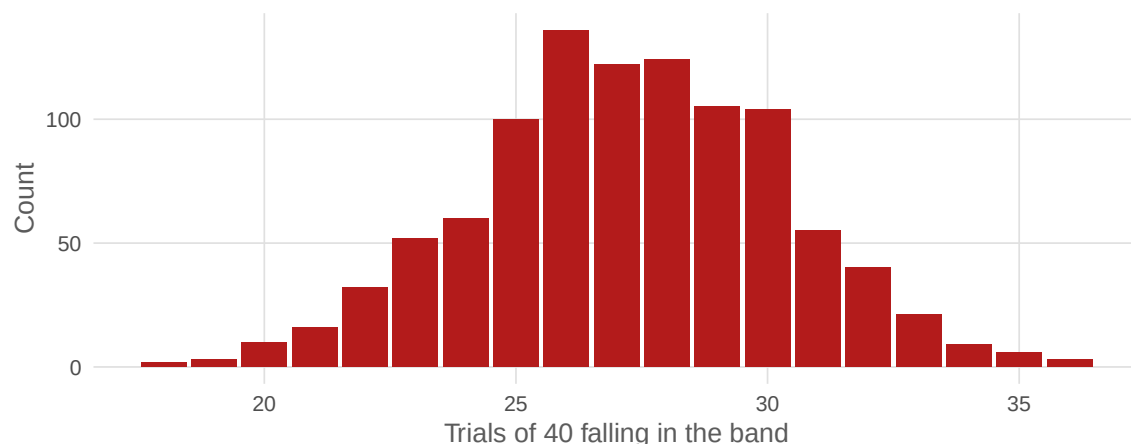
(d) Each trial falls in the band or it does not. The trials are independent, and each falls in the band with the same probability $p = 884/1296$. So $K \sim \text{Binomial}(40, 884/1296)$, and

$$\mathbb{E}[K] = np = 40 \times .682 = 27.3, \quad \text{SD}[K] = \sqrt{np(1-p)} = 2.95.$$

(e)

```
K <- replicate(1000, sum(in_band(replicate(40, trial()))))
```

```
tibble(K) |>
  ggplot(aes(K)) +
  geom_bar(fill = gov_colors[["carnelian"]]) +
  labs(x = "Trials of 40 falling in the band", y = "Count") +
  theme_gov(base_size = 10)
```



```
c(mean = mean(K), sd = sd(K))
```

```
      mean      sd
27.238000  3.043784
```

```
range(K)
```

```
[1] 18 36
```

```
mean(K <= 24)
```

```
[1] 0.175
```

```
pbinom(24, 40, 884 / 1296)
```

```
[1] 0.1716119
```

The simulated mean and standard deviation, 27.24 and 3.04, match part (d). A pair with a fair die

counts 24 or fewer about 18 percent of the time. The exact binomial probability is 0.172. The 1,000 counts run from 18 to 36, so two pairs with fair dice can hold up sheets with very different counts. The probability .682 describes the distribution of $\bar{X}$. Any one sheet of 40 trials is a sample from that distribution, and its count is a random variable.

(f)

```
xbar10k <- replicate(10000, trial())
mean(in_band(xbar10k))
```

```
[1] 0.6769
```

The share is 0.677, close to .682. The share of trials in the band is the sample mean of 10,000 iid indicators, each with expected value .682. By the law of large numbers it converges in probability to .682 as the number of trials grows.

Question 6: Small donations

```
mu    <- exp(4 + 1 / 2)
sigma <- sqrt((exp(1) - 1) * exp(2 * 4 + 1))
c(mu = mu, sigma = sigma)
```

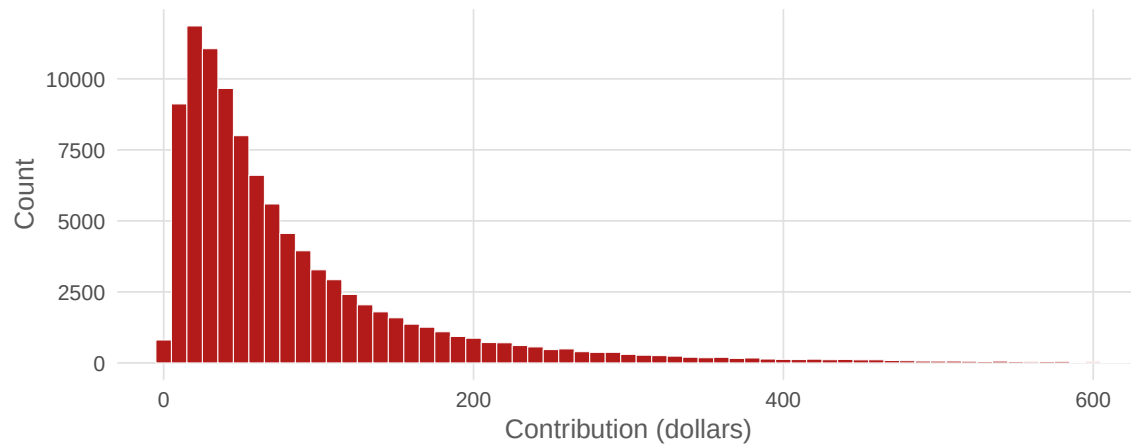
```
      mu      sigma
90.01713 117.99738
```

(a)

```
donations <- rlnorm(1e5, meanlog = 4, sdlog = 1)
c(mean = mean(donations), median = median(donations))
```

```
      mean    median
89.69500  54.32307
```

```
tibble(x = donations) |>
  ggplot(aes(x)) +
  geom_histogram(binwidth = 10, fill = gov_colors[["carnelian"]],
                 color = "white", linewidth = 0.1) +
  coord_cartesian(xlim = c(0, 600)) +
  labs(x = "Contribution (dollars)", y = "Count") +
  theme_gov(base_size = 10)
```



The distribution is right-skewed. Most contributions are small. A long right tail of large contributions runs well past the plotted range. The median, about \$54, is the point with half the contributions on each side. The mean, about \$90, weights each contribution by its size. The large contributions in the right tail raise the mean and leave the median where it is.

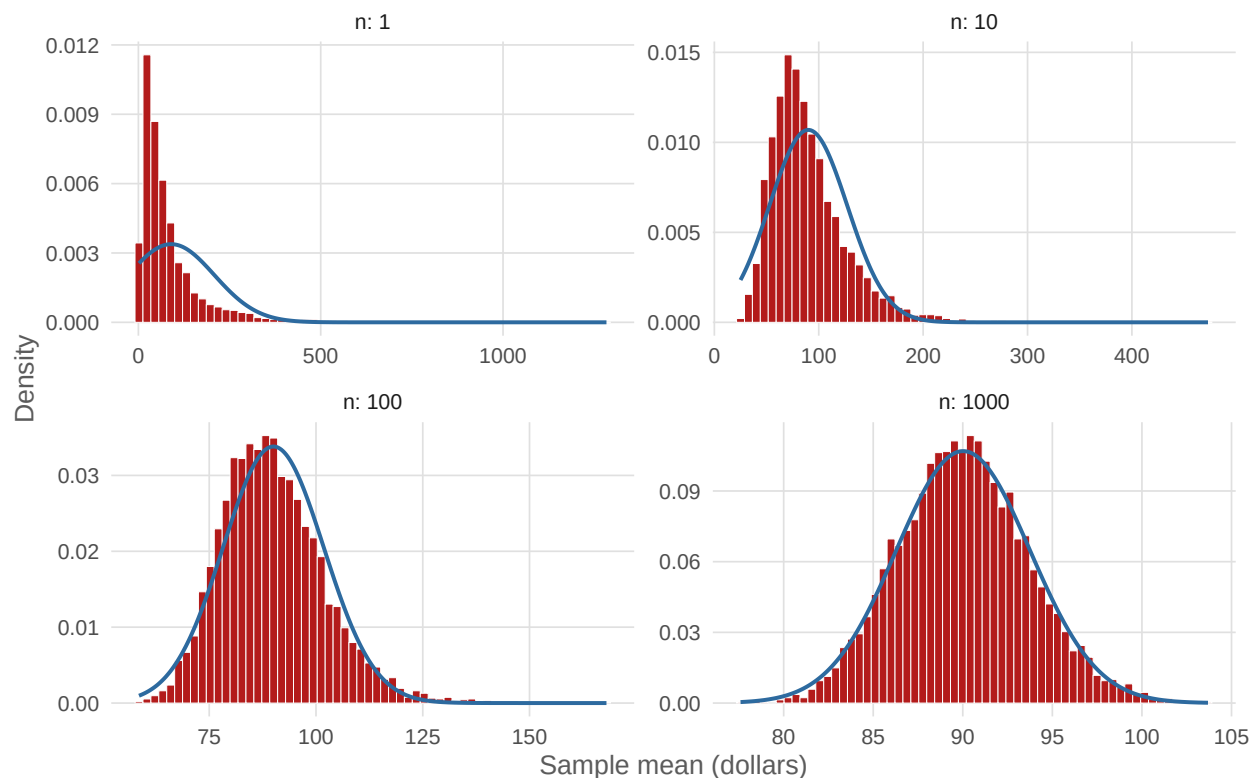
(b)

```
ns <- c(1, 10, 100, 1000)

sims <- map_dfr(ns, \(n) tibble(n = n,
  xbar = replicate(5000, mean(rlnorm(n, meanlog = 4, sdlog = 1))))))

curves <- map_dfr(ns, \(n) {
  s <- sims$xbar[sims$n == n]
  grid <- seq(min(s), max(s), length.out = 300)
  tibble(n = n, x = grid, dens = dnorm(grid, mu, sigma / sqrt(n)))
})

ggplot(sims, aes(xbar)) +
  geom_histogram(aes(y = after_stat(density)), bins = 60,
    fill = gov_colors[["carnelian"]], color = "white",
    linewidth = 0.1) +
  geom_line(data = curves, aes(x, dens), color = gov_colors[["blue"]],
    linewidth = 0.7) +
  facet_wrap(~ n, scales = "free", labeller = label_both) +
  labs(x = "Sample mean (dollars)", y = "Density") +
  theme_gov(base_size = 10)
```



At $n = 1$ the histogram is the population itself, and the normal curve misses it badly. By $n = 1000$ the histogram and the curve coincide.

(c)

```
summ <- sims |>
  group_by(n) |>
  summarize(mean_xbar = mean(xbar), sd_xbar = sd(xbar),
            below_mu = mean(xbar < mu),
            within_1 = mean(abs((xbar - mu) / (sigma / sqrt(n))) <= 1),
            z_025 = quantile((xbar - mu) / (sigma / sqrt(n)), .025),
            z_975 = quantile((xbar - mu) / (sigma / sqrt(n)), .975)) |>
  mutate(sigma_root_n = sigma / sqrt(n))
summ |> select(n, mean_xbar, sd_xbar, sigma_root_n)
```

A tibble: 4 x 4

	n	mean_xbar	sd_xbar	sigma_root_n
	<dbl>	<dbl>	<dbl>	<dbl>
1	1	86.5	105.	118.
2	10	90.4	37.3	37.3
3	100	89.9	12.1	11.8
4	1000	90.0	3.64	3.73

At every n the mean of the sample means is close to $\mu = 90.02$, and their standard deviation is close to $\sigma/\sqrt{n}$. Both hold at $n = 1$, where the distribution is far from normal. The $n = 1$ row shows the largest gaps. There each “sample mean” is a single contribution, and the long right tail

makes the mean and SD of 5,000 single draws noisy. The gaps are simulation error. They shrink as n grows. They are the two exact results derived in lecture: $\mathbb{E}[\bar{X}] = \mu$ and $\mathbb{V}[\bar{X}] = \sigma^2/n$. Neither result depends on n being large, and neither says anything about shape.

(d)

```
summ |> select(n, below_mu)
```

```
# A tibble: 4 x 2
      n below_mu
  <dbl>   <dbl>
1     1     0.702
2    10     0.589
3   100     0.544
4  1000     0.502
```

A normal distribution is symmetric about its mean, so it would put .5 below μ . At small n the sampling distribution inherits the population's skew: the typical sample contains no large contribution and so has a mean below μ . A few samples contain one or more very large contributions and have means far above μ . Those few pull the balance point up to μ and leave most of the mass below it. As n grows the share falls toward .5: here it runs 0.70, 0.59, 0.54, 0.50 at $n = 1, 10, 100, 1000$.

(e)

```
summ |> select(n, within_1, z_025, z_975)
```

```
# A tibble: 4 x 4
      n within_1 z_025 z_975
  <dbl>   <dbl> <dbl> <dbl>
1     1     0.911 -0.702  2.37
2    10     0.753 -1.32   2.44
3   100     0.693 -1.75   2.24
4  1000     0.695 -1.83   1.99
```

```
pnorm(1) - pnorm(-1)
```

```
[1] 0.6826895
```

```
qnorm(c(.025, .975))
```

```
[1] -1.959964  1.959964
```

The share with $|Z| \leq 1$ falls from about 0.91 at $n = 1$ to 0.69 at $n = 1000$, against the normal's .683. The percentiles start lopsided, near -0.70 and 2.37, and approach ± 1.96 . The lower percentile converges more slowly than the upper. The right skew of the population compresses the lower tail of the sampling distribution.

(f) Skewness. The distribution of one roll of a fair die is symmetric about 3.5. The distribution of the mean of n rolls is then symmetric about 3.5 at every n , and at $n = 4$ it is already close to normal. The contributions are strongly right-skewed. The sample mean inherits that skew and loses it only gradually as n grows. The central limit theorem says the sampling distribution approaches a normal whatever the population's shape. The shape determines how large n must be before the approximation is good.

(g)

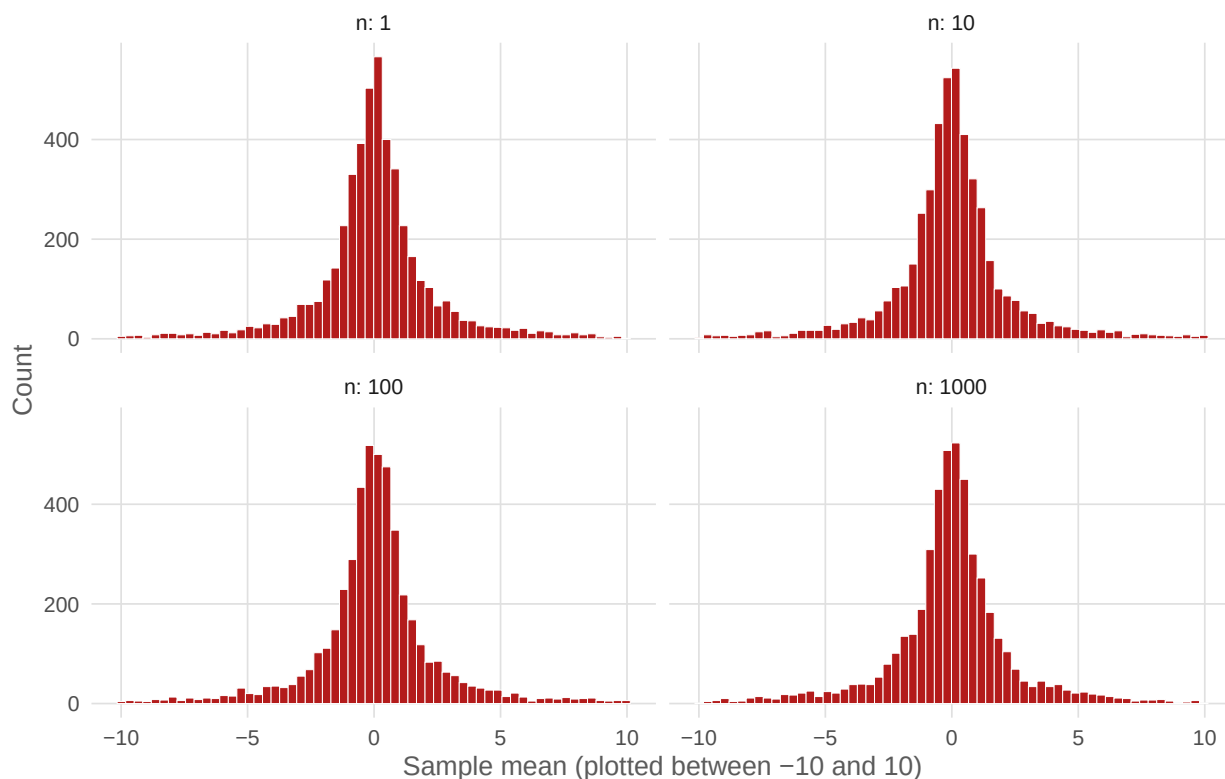
```
cauchy <- map_dfr(ns, \(n) tibble(n = n,  
  xbar = replicate(5000, mean(rcauchy(n)))))
```

```
cauchy |>  
  group_by(n) |>  
  summarize(q25 = quantile(xbar, .25), median = median(xbar),  
    q75 = quantile(xbar, .75))
```

A tibble: 4 x 4

	n	q25	median	q75
	<dbl>	<dbl>	<dbl>	<dbl>
1	1	-0.974	0.0244	1.01
2	10	-1.02	-0.0102	0.975
3	100	-0.964	0.00979	0.982
4	1000	-0.967	0.0150	1.03

```
cauchy |>  
  filter(abs(xbar) < 10) |>  
  ggplot(aes(xbar)) +  
  geom_histogram(bins = 60, fill = gov_colors[["carnelian"]],  
    color = "white", linewidth = 0.1) +  
  facet_wrap(~ n, labeller = label_both) +  
  labs(x = "Sample mean (plotted between -10 and 10)", y = "Count") +  
  theme_gov(base_size = 10)
```



The distribution of the sample mean does not narrow as n grows. Its quartiles stay near ± 1 at every n . The mean of n independent standard Cauchy draws is itself standard Cauchy, whatever n is. The central limit theorem requires iid draws with a finite variance σ^2 . The Cauchy has no finite variance and no finite mean, so there is no $\sigma/\sqrt{n}$ to shrink and no μ for the law of large numbers to converge to.

Question 7: India's missing girls

(a)

```
p <- 0.487
n_births <- 15e6
sd_xbar <- sqrt(p * (1 - p) / n_births)
c(mean = p, sd = sd_xbar, bound = 0.5 / sqrt(n_births))
```

```
      mean      sd      bound
0.4870000000 0.0001290558 0.0001290994
```

$\mathbb{E}[\bar{X}] = p = .487$. $\text{SD}[\bar{X}] = \sqrt{p(1-p)/n} \approx 0.000129$. The bound $.5/\sqrt{n}$ from Question 4(c) is 0.000129. At $p = .487$ the two agree to three significant figures, because p is close to .5.

(b) By the central limit theorem, $\bar{X} \approx \mathcal{N}(.487, .487 \times .513/15,000,000)$.

```
z <- (0.47 - p) / sd_xbar
z
```

```
[1] -131.726
```

The observed share lies about 132 standard deviations below $\mathbb{E}[\bar{X}]$.

(c)

```
pnorm(0.47, mean = p, sd = sd_xbar)
```

```
[1] 0
```

```
pnorm(z, log.p = TRUE) / log(10) # the probability's power of 10
```

```
[1] -3770.399
```

The probability is positive: the normal density is positive everywhere. It is smaller than the smallest positive number R can store, so R rounds it to 0. That number is about 10^{-308} (about 5×10^{-324} counting R's least precise numbers). R's answer says the probability is too small to represent. It does not say the event is impossible. Asking `pnorm()` for the logarithm of the probability shows its size: about 10^{-3770} .

(d)

```
sims_births <- rbinom(10000, size = n_births, prob = p) / n_births
range(sims_births)
```

```
[1] 0.4864489 0.4875606
```

The 10,000 simulated shares run from 0.4864 to 0.4876. All of them lie within about .001 of .487. The observed .47 is about 30 times farther from .487 than the farthest simulated value.

(e)

```
ns_births <- c(1000, 4000, 10000)
pnorm(0.47, mean = p, sd = sqrt(p * (1 - p) / ns_births))
```

```
[1] 0.1410669648 0.0157351009 0.0003355152
```

```
n2 <- 4 * p * (1 - p) / 0.017^2
n2
```

```
[1] 3457.869
```

At $n = 1,000$ the probability is about .14: one sample in seven would show 47 percent girls or fewer. At $n = 4,000$ it is about .016, and at $n = 10,000$ about .0003. Setting $.017 = 2\sqrt{p(1-p)/n}$ gives $n = 4p(1-p)/.017^2 \approx 3,457.9$. The smallest whole n at which .47 lies at least 2 standard deviations below .487 is 3,458. The gap of .017 is the same at every n . $SD[\bar{X}]$ shrinks as n grows, so the same gap counts as more standard deviations.

(f) For a sum, $V[\sum_i X_i] = \sum_i V[X_i] + 2\sum_i \sum_{j>i} \text{Cov}(X_i, X_j)$. Identical twins have the same sex, so for a twin pair $\text{Cov}(X_i, X_j) > 0$. Every other pair has covariance 0. The covariance sum is positive, so twins make $V[\bar{X}]$ larger than $p(1-p)/n$.

```
factor_sd <- (0.017 / 2) / sd_xbar
c(sd_factor = factor_sd, var_factor = factor_sd^2)
```

```
sd_factor var_factor
65.86298 4337.93244
```

For .47 to lie 2 standard deviations below .487, $SD[\bar{X}]$ would have to be .0085, about 66 times its value under independence. The variance would have to grow by a factor of about 4,300. Twins cannot do this. In the extreme case every birth is one of an identical pair. Then the 15 million births are 7.5 million independent pairs, each counted twice, and the variance doubles. The SD grows by a factor of $\sqrt{2} \approx 1.4$. Identical twins are well under 1 percent of births, so the real effect is far smaller still.

(g)

```
girls <- 0.47 * n_births
boys <- 0.53 * n_births
miss_total <- p * n_births - girls # (i): total births fixed
miss_boys <- boys * p / (1 - p) - girls # (ii): boys fixed
c(i = miss_total, ii = miss_boys)
```

```
      i      ii
255000 497076
```

Count (i) is 255,000. It holds total births at 15 million: each missing girl was replaced by a boy, as when a pregnancy is ended and the next one produces a son. Count (ii) is about 497,000. It holds the number of boys fixed and asks how many girls a process with $p = .487$ would have produced alongside them. Each missing girl is then simply absent from the count, and nothing replaced her. Under (ii), total births would have been about 15.5 million.

(h) Any two of the following, or others that work as well:

- Sex-selective abortion: girls are conceived at the usual rate, but fewer are born.
- Under-registration: girls are born at the usual rate, but fewer of their births are recorded. The recorded share then measures the registration system, not the births.
- Female infanticide or neglect in the first days of life, before a birth is registered.
- A different biological p . Oster (2003) argued that hepatitis B raises the probability of a male birth. Later work found that hepatitis B accounts for little of the gap.

The CLT calculation rules out $p = .487$ with independent births. It cannot choose among these processes. Each of them produces a recorded share below .487.

Question 8: Five estimators of a share

(a) Each X_i has $\mathbb{E}[X_i] = \theta$. A share among any fixed set of m respondents is a sample mean of m iid draws, so its expected value is θ , whatever m is.

Estimator	Expected value	Bias
$\hat{\theta}_A = \bar{X}_n$	θ	0
$\hat{\theta}_B$, first 4	θ	0
$\hat{\theta}_C$, first 10	θ	0
$\hat{\theta}_D = \bar{X}_n + 1/n$	$\theta + 1/n$	$1/n$
$\hat{\theta}_E = \bar{X}_n + .05$	$\theta + .05$	.05

For $\hat{\theta}_D$ and $\hat{\theta}_E$, $\mathbb{E}[a + bX] = a + b\mathbb{E}[X]$ with $b = 1$. $\hat{\theta}_A$, $\hat{\theta}_B$ and $\hat{\theta}_C$ are unbiased. $\hat{\theta}_D$ and $\hat{\theta}_E$ are biased upward.

(b) $\mathbb{V}[X_i] = \theta(1 - \theta)$. The variance of a mean of m iid draws is $\theta(1 - \theta)/m$. Adding a constant does not change a variance.

Estimator	Variance	At $\theta = .50$
$\hat{\theta}_A$	$\theta(1 - \theta)/n$	$.25/n$
$\hat{\theta}_B$	$\theta(1 - \theta)/4$	.0625
$\hat{\theta}_C$	$\theta(1 - \theta)/10$	.025
$\hat{\theta}_D$	$\theta(1 - \theta)/n$	$.25/n$
$\hat{\theta}_E$	$\theta(1 - \theta)/n$	$.25/n$

The variances of $\hat{\theta}_B$ and $\hat{\theta}_C$ do not depend on n . Neither estimator uses the respondents after the first 4 or the first 10.

(c) For $n \geq 10$, $1/n \leq 1/10 < 1/4$, so $\mathbb{V}[\hat{\theta}_A] \leq \mathbb{V}[\hat{\theta}_C] < \mathbb{V}[\hat{\theta}_B]$. $\hat{\theta}_A$ is the most efficient of the three and $\hat{\theta}_B$ the least. At $n = 10$, $\hat{\theta}_A$ and $\hat{\theta}_C$ are the same estimator, so they tie. For every $n > 10$, $\hat{\theta}_A$ is strictly more efficient than $\hat{\theta}_C$.

(d) $\text{MSE} = \text{Bias}^2 + \text{Variance}$, at $n = 100$:

```
theta <- 0.5
n_mse <- 100
v <- theta * (1 - theta)
```

```
tibble(estimator = c("A", "B", "C", "D", "E"),
       bias = c(0, 0, 0, 1 / n_mse, 0.05),
       variance = c(v / n_mse, v / 4, v / 10, v / n_mse, v / n_mse)) |>
  mutate(mse = bias^2 + variance)
```

```
# A tibble: 5 x 4
  estimator bias variance    mse
  <chr>      <dbl>    <dbl> <dbl>
1 A          0      0.0025 0.0025
2 B          0      0.0625 0.0625
3 C          0      0.025  0.025
4 D        0.01      0.0025 0.0026
5 E        0.05      0.0025 0.005
```

The MSEs are .0025, .0625, .025, .0026 and .005. $\hat{\theta}_D$ and $\hat{\theta}_E$ are biased, but both have a smaller MSE than the unbiased $\hat{\theta}_B$ and $\hat{\theta}_C$. For example, $\hat{\theta}_E$ has MSE .005, one fifth of $\hat{\theta}_C$'s .025. $\hat{\theta}_E$'s bias costs $.05^2 = .0025$, and it uses all 100 respondents, which saves far more than that in variance.

(e) The error is $\hat{\theta}_A - \theta = .54 - .50 = .04$.

No. Bias is $\mathbb{E}[\hat{\theta}_A] - \theta$, the average error across repeated polls. One poll gives one error, which includes the chance variation of that sample. $\hat{\theta}_A$ is unbiased (part a). With $n = 100$ its standard deviation is $\sqrt{.25/100} = .05$, so an error of .04 is less than one standard deviation and is unremarkable.

A pollster cannot compute the error of a real poll, because the error requires θ . If the pollster knew θ , there would be no reason to run the poll.

(f) For each n , the code draws the first 10 respondents of each poll individually, so that $\hat{\theta}_B$ and $\hat{\theta}_C$ can be computed. It draws the number of supporters among the remaining $n - 10$ in one step.

```
one_n <- function(n, polls = 5000) {
  first10 <- matrix(rbinom(polls * 10, 1, theta), nrow = polls)
  rest <- rbinom(polls, n - 10, theta)
  xbar <- (rowSums(first10) + rest) / n
  tibble(n = n,
        A = xbar,
        B = rowMeans(first10[, 1:4]),
        C = rowMeans(first10),
        D = xbar + 1 / n,
        E = xbar + 0.05)
}
ests <- map_dfr(c(10, 100, 1000, 10000), one_n)
```

```
# round() guards against floating-point error: in R, 0.52 - 0.5 is slightly
# larger than 0.02, so an estimate of exactly .52 would otherwise be excluded.
```

```
ests |>
  pivot_longer(A:E, names_to = "estimator", values_to = "est") |>
  group_by(n, estimator) |>
  summarize(within = mean(round(abs(est - theta), 10) <= 0.02),
            .groups = "drop") |>
```

```
pivot_wider(names_from = estimator, values_from = within)
```

```
# A tibble: 4 x 6
      n      A      B      C      D      E
<dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1    10 0.254 0.381 0.254 0.210 0
2   100 0.387 0.381 0.245 0.375 0.236
3  1000 0.808 0.377 0.243 0.810 0.0282
4 10000 1     0.382 0.242 1     0
```

- $\hat{\theta}_A$ and $\hat{\theta}_D$: their shares go to 1 as n grows. Both are consistent. $\hat{\theta}_D$'s bias, $1/n$, and its variance both shrink to zero, so its distribution piles up at θ .
- $\hat{\theta}_B$ and $\hat{\theta}_C$: their shares stay flat at every n . $\hat{\theta}_B$ is within .02 of .50 only when exactly 2 of the first 4 respondents support the measure, with probability $\binom{4}{2}(.5)^4 = .375$. $\hat{\theta}_C$ is within .02 only when exactly 5 of the first 10 do, with probability $\binom{10}{5}(.5)^{10} \approx .246$. More respondents do not change either estimator, so both are unbiased but inconsistent.
- $\hat{\theta}_E$: its share is 0 at $n = 10$, about .24 at $n = 100$, and then falls to 0. At $n = 10$ it can only take values $\bar{X} + .05$, and the two closest to .50, .45 and .55, are both .05 away. At larger n its distribution narrows around .55, which is .05 from θ , more than the .02 window. It is inconsistent.

The consistent estimators are $\hat{\theta}_A$ and $\hat{\theta}_D$. The share within .02 is $\mathbb{P}(|\hat{\theta}_n - \theta| \leq .02)$. Consistency requires it to go to 1 for every window, however small, not just .02.

Question 9: Dividing by n or by $n - 1$

(a) Expand the square and use $\sum_i X_i = n\bar{X}$:

$$\sum_i (X_i - \bar{X})^2 = \sum_i X_i^2 - 2\bar{X} \sum_i X_i + n\bar{X}^2 = \sum_i X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 = \sum_i X_i^2 - n\bar{X}^2.$$

(b) With $W = X_i$: $\mathbb{E}[X_i^2] = \sigma^2 + \mu^2$. With $W = \bar{X}$, and $\mathbb{E}[\bar{X}] = \mu$, $\mathbb{V}[\bar{X}] = \sigma^2/n$ from lecture: $\mathbb{E}[\bar{X}^2] = \sigma^2/n + \mu^2$.

(c) By linearity and part (b),

$$\mathbb{E}[\text{SSD}] = \sum_i \mathbb{E}[X_i^2] - n \mathbb{E}[\bar{X}^2] = n(\sigma^2 + \mu^2) - n \left(\frac{\sigma^2}{n} + \mu^2 \right) = (n - 1) \sigma^2.$$

(d) $\mathbb{E}[s^2] = \mathbb{E}[\text{SSD}]/(n - 1) = \sigma^2$, so s^2 is unbiased. $\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}[\text{SSD}]/n = \frac{n-1}{n} \sigma^2$, so $\text{Bias}(\hat{\sigma}^2) = -\sigma^2/n$. It underestimates σ^2 on average. The reason: $\bar{X}$ is the value of c that makes $\sum_i (X_i - c)^2$ as small as possible. So $\text{SSD} \leq \sum_i (X_i - \mu)^2$ in every sample, and the right-hand side has expectation $n\sigma^2$. Measuring the deviations from $\bar{X}$ instead of μ makes them too small on average.

(e)

```
ssd <- replicate(1e5, {x <- rnorm(10); sum((x - mean(x))^2)})
q9_est <- tibble(estimator = c("s2 (n - 1)", "sigma-hat2 (n)"),
  est = list(ssd / 9, ssd / 10)) |>
```

```
mutate(bias = map_dbl(est, \(e) mean(e) - 1),
       variance = map_dbl(est, var),
       mse = map_dbl(est, \(e) mean((e - 1)^2))) |>
select(-est)
q9_est
```

A tibble: 2 x 4

	estimator	bias	variance	mse
	<chr>	<dbl>	<dbl>	<dbl>
1	s2 (n - 1)	-0.00272	0.221	0.221
2	sigma-hat2 (n)	-0.102	0.179	0.190

s^2 has bias near zero and MSE near 0.221. $\hat{\sigma}^2$ has bias near -0.1 , which is $-\sigma^2/n$, but its variance is smaller by a factor of $(9/10)^2$. Its MSE is about 0.190, the smaller of the two. The bias costs $.1^2 = .01$ in MSE. The lower variance saves more than that.

(f) s^2 is right on average: across repeated samples it centers on σ^2 . $\hat{\sigma}^2$ has the smaller mean squared error: its average squared distance from σ^2 is smaller. The two differ by the factor $(n-1)/n$, which goes to 1. As n grows the bias $-\sigma^2/n$ and the difference between the estimators both go to zero.