STAT 135, Concepts of Statistics

Helmut Pitters

Simple random sampling

Department of Statistics
University of California, Berkeley

January 25, 2017

Simple random sampling

Simple random sampling

Simple random sampling: motivation

Usually, we <u>do not know</u>

$$N : \text{population size}$$

$$x_1, \ldots, x_N : \text{individual characteristics}$$

Can sample some of the individuals and record their characteristics. However, sampling all individuals is usually not feasible, e.g. it might

- be too costly, time-consuming (e.g. polling opinions of all US citizens)
- require to destroy products (e.g. canned food)
- be impossible, since we have no means to observe whole population (e.g. population of atlantic cod)

Simple random sampling

9-year old **Antonio Martinez** of San Lorenzo caught a 12 lb., 7 oz., 27" trout at Don Castro using power bait on 4/6/2008!!

Want to estimate average weight of trouts in Lake Don Castro.
How?
Catch a trout, record its weight w_1, and release the fish again.
Maybe this fish was comparatively small/big.
Let's iterate this procedure, until we recorded the weights

$$w_1, w_2, \ldots, w_n$$

of a "sufficiently large" number n of fishes.
With this knowledge, how do we estimate the weight of a trout?

Simple random sampling: setup

Mostly, one is interested in *population parameters*

$$\mu := \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{population mean}$$

$$\sigma^2 := \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2 \qquad \text{population variance}$$

$$\tau := \sum_{i=1}^{N} x_i \qquad \text{population total}$$

which are unknown.

Goal: Want to estimate (or "learn") population parameters by studying a sample of n individuals drawn radomly from the population.

Simple random sampling: setup

Sample n individuals *randomly with replacement* and record their characteristics

$$X_1, X_2, \dots, X_n.$$

In other words: X_1 is chosen at random from $x_1, \dots, x_N$, and so is $X_2, \dots, X_n$, and the X_is are independent.

Here, sampling introduces the randomness.

Later, we'll also be interested in *sampling without replacement* (referred to as simple random sampling).

Remark

Always consider carefully whether sampling is conducted with or without replacement.

Simple random sampling

Intuitively, we expect

$$\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \text{sample mean}$$

to be a good estimator for population mean μ.

With some work we can show rigorously that this intuition is good!

Simple random sampling

Notation: suppose there are m different values in $x_1, x_2, \ldots, x_N$; denote them by

$$\zeta_1, \zeta_2, \ldots, \zeta_m.$$

Let's say the value ζ_i appears n_i times in the population, i.e.

$$n_i := \#\{j \colon x_j = \zeta_i\}.^1$$

By construction of X_1,

$$\mathbb{P}\{X_1 = \zeta_i\} = \frac{n_i}{n}.$$

We can now study the distribution of X_1 (and $\bar{X}$).

[1] $\#A$ denotes the number of elements in the set A.

Simple random sampling

For the mean of X_1 we find

$$\mathbb{E}[X_1] = \sum_{i=1}^{m} \zeta_i \mathbb{P}\left\{X_1 = \zeta_i\right\} = \sum_{i=1}^{m} \zeta_i \frac{n_i}{N} = \frac{1}{N} \sum_{j=1}^{N} x_j = \mu.$$

It is now straightforward to compute the mean of $\bar{X}$:

$$\mathbb{E}[\bar{X}] = \mathbb{E}[\frac{1}{n} \sum_{i=1}^{n} X_i] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \mu,$$

since $X_1, \ldots, X_n$ all have the same distribution.

This makes our intuition precise: if we were to study the population by drawing SRSs many times the average of the sample means would be μ.

E.g. think of a large number of scientists independently studying the same population by drawing SRSs.

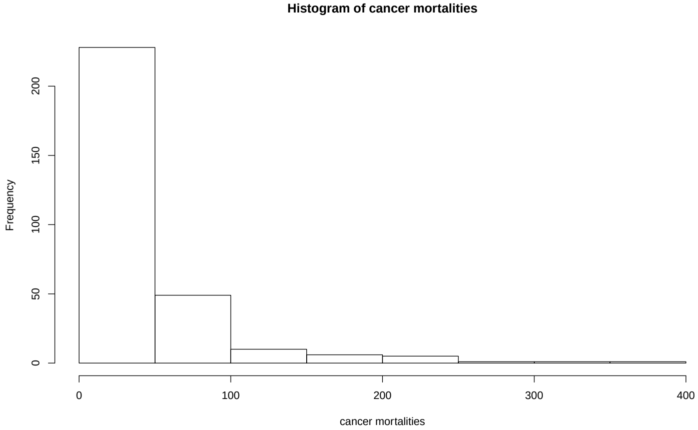
Simple random sampling

The results and calculations derived so far are still correct if our sampling is *without* replacement.

Why?

Simple random sampling

data: Values for breast cancer mortality 1950–1960 for 301 counties in North Carolina, South Carolina and Georgia.[2]



[2]You can find these data in `data/cancer.txt`, on bCourses.

Simple random sampling

data: Values for breast cancer mortality 1950–1960 for 301 counties in North Carolina, South Carolina and Georgia.[3]

population mean $\mu = 39.86$

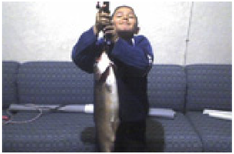
The following are five means computed from five independent SRSs from the population, each of size 20:

$$64.7, 22.2, 29.3, 30.4, 42.2.$$

[3] You can find these data in `data/cancer.txt`, on bCourses.

Aside: Estimating population size

Usually, we do not even know the population size N.



9-year old **Antonio Martinez** of San Lorenzo caught a 12 lb., 7 oz., 27" trout at Don Castro using power bait on 4/6/2008!!

Lake Don Castro contains a population of trout, the number of which (call it N) is unknown. Come up with a simple method to estimate N.
You are allowed to catch (some) fish, and mark them with a color.

Simple random sampling

On average, sample mean agrees with population mean:

$$\mathbb{E}\bar{X} = \mu.$$

How accurate is $\bar{X}$ as an estimator for μ?

A reasonable measure for the accuracy of $\bar{X}$ is its standard deviation or *standard error*

$$\sigma_{\bar{X}} := \sqrt{\mathrm{Var}(\bar{X})}.$$

Since

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n}\mathrm{Var}(X_1),$$

let's compute $\mathrm{Var}(X_1)$.

Simple random sampling

We find

$$\operatorname{Var}(X_1) = \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \sum_{i=1}^{m} \zeta_i^2 \frac{n_i}{N} - \mu^2$$
$$= \frac{1}{N} \sum_{i=1}^{n} x_i^2 - \mu^2 = \sigma^2,$$

and therefore

$$\operatorname{Var}(\bar{X}) = \frac{1}{n} \operatorname{Var}(X_1) = \frac{\sigma^2}{n},$$

so $\bar{X}$ has standard error

$$\boxed{\sigma_{\bar{X}} = \sqrt{\operatorname{Var}(\bar{X})} = \frac{\sigma}{\sqrt{n}}.}$$

What does this formula tell us about how accurate we can estimate ("guess") μ from $\bar{X}$?

Review: Covariance.

Recall from STAT 134: *Covariance* of two random variables X and Y is defined by

$$\mathrm{Cov}(X, Y) \coloneqq \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right].$$

The covariance can be interpreted as a measure for joint variability or degree of linear association.

If $\mathrm{Cov}(X, Y) = 0$, we call X and Y *uncorrelated*.

While independence of X, Y implies $\mathrm{Cov}(X, Y) = 0$, the converse is not true.

Review: Covariance.

Recall from STAT 134 some useful formulas:

$$\begin{aligned}
\mathrm{Cov}(X, X) &= \mathrm{Var}(X) \\
\mathrm{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\
\mathrm{Var}(X + Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y) \\
\mathrm{Cov}(X, Y) &= \mathrm{Cov}(Y, X) \qquad \text{(symmetry)} \\
\mathrm{Cov}(aX + b, Y) &= a\,\mathrm{Cov}(X, Y) \qquad \text{(multilinearity)}
\end{aligned}$$

Simple random sampling

What if instead we sample without replacement?
Now

$$\begin{aligned}\mathrm{Var}(\bar{X}) &= \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2}\,\mathrm{Cov}\Big(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\Big)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\mathrm{Cov}(X_i, X_j),\end{aligned}$$

and we need to find $\mathrm{Cov}(X_i, X_j)$ for $i \neq j$.

Simple random sampling

Lemma

For $i \neq j$ we have

$$\mathrm{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}. \tag{1}$$

Proof.

By definition of covariance and using $\mathbb{E}X_i = \mu$,

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}X_iX_j - \mathbb{E}X_i\mathbb{E}X_j = \mathbb{E}X_iX_j - \mu^2$$

$$\mathbb{E}X_iX_j = \sum_{k,l=1}^{m} \zeta_k\zeta_l\mathbb{P}\left\{X_i = \zeta_k, X_j = \zeta_l\right\}.$$

$\square$

Proof.

$$\mathbb{P}\{X_i = \zeta_k, X_j = \zeta_l\} = \mathbb{P}\{X_j = \zeta_l\}\,\mathbb{P}\{X_i = \zeta_k | X_j = \zeta_l\}$$
$$= \begin{cases} \frac{n_l n_k}{N(N-1)} & k \neq l \\ \frac{n_l(n_k-1)}{N(N-1)} & k = l. \end{cases}$$

Write $\frac{n_l(n_k-1)}{N(N-1)} = \frac{n_l n_k}{N(N-1)} - \frac{n_l}{N(N-1)}$ to obtain

$$\mathbb{E}X_iX_j = \frac{1}{N(N-1)} \sum_{l=1}^{m} \zeta_l n_l \sum_{k=1}^{m} \zeta_k n_k - \frac{1}{N(N-1)} \sum_{l=1}^{m} \zeta_l^2 n_l$$
$$= \frac{N}{N-1}\mu^2 - \frac{1}{N-1}(\sigma^2 + \mu^2) = \mu^2 - \frac{\sigma^2}{N-1}.$$

The claim follows. □

Theorem

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) = \frac{\sigma^2}{n}\frac{N-n}{N-1}. \tag{2}$$

Remark

- $1 - \frac{n-1}{N-1}$ is the *finite population correction*
- if *sampling fraction*

$$\frac{n}{N}$$

  is small, the standard error is close to the one for sampling with replacement:

$$\sigma_{\bar{X}} \approx \frac{\sigma}{\sqrt{n}},$$

  and hardly depends on N.

Proof.

Recall

$$\mathrm{Var}(X_i) = \sigma^2, \qquad \mathrm{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}.$$

Hence

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\mathrm{Cov}(X_i, X_j) \\
&= \frac{\sigma^2}{n} - \frac{n(n-1)}{n^2}\frac{\sigma^2}{N-1} \\
&= \frac{\sigma^2}{n}(1 - \frac{n-1}{N-1}).
\end{aligned}$$

$\square$

Simple random sampling

We now turn to an example where we sample without replacement.

Example (Hospital discharges[4])

For instructional purposes we study an example where we have access to the entire population—this will note be the case in real studies.

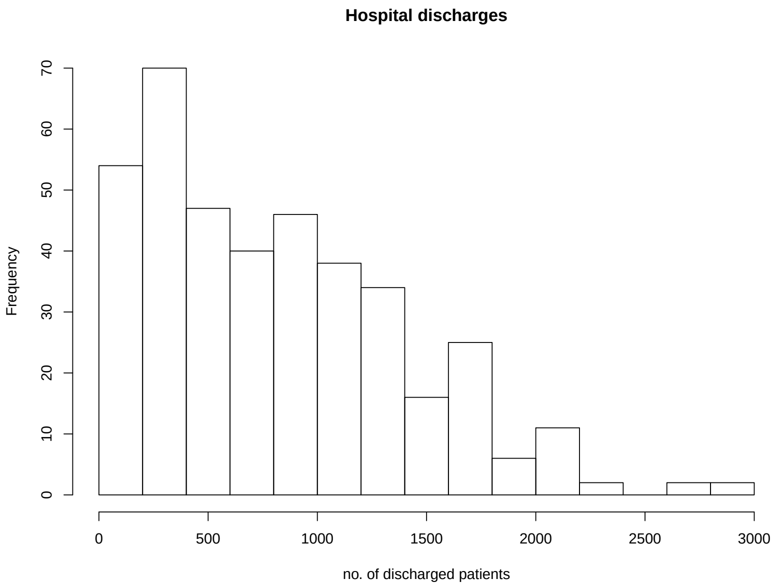
population: $N = 393$ short stay hospitals

$x_i :=$ # patients discharged from ith hospital during January 1968

$$\mu = 814.6$$

[4] Find the data in `data/hospitals.txt`.

Simple random sampling



Example: hospital discharges

Population parameters

$$N = 393, \qquad \mu = 814.6, \qquad \sigma = 589.7$$

Drawing sample of size $n = 50$ yielded $\bar{X} = 818.0$.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{589.7}{\sqrt{50}}\sqrt{\frac{343}{392}} = 83.4 \times 0.94 \approx 78.0.$$

If we were to take samples (of size $n = 50$) repeatedly and independently, most of the sample means would be contained in

$$(\bar{X} - 2\sigma_{\bar{X}}, \bar{X} + 2\sigma_{\bar{X}}) \approx (662.0, 974.0).$$

(Cf. histogram of simulated sample means)

How likely is it that this interval contains what we are actually interested in: the population mean μ?

Review. Binomial distribution.

Example (Lake Don Castro)

Suppose a proportion $p = 0.3$ of the fishes in Lake Don Castro are trouts. If we catch $n = 20$ fishes, releasing each fish after the catch (=sampling with replacement), the number

$$T = T(n, p)$$

of trouts in Lake Don Castro is random. It has the so-called *binomial distribution* with probability mass function

$$\mathbb{P}\{T = k\} = \binom{n}{k} p^k (1-p)^{n-k} \quad (0 \le k \le n).$$

Histogram: `https://www.geogebra.org/m/CmHJuJxs`

Review. Normal distribution.

Definition (Normal density)

The map

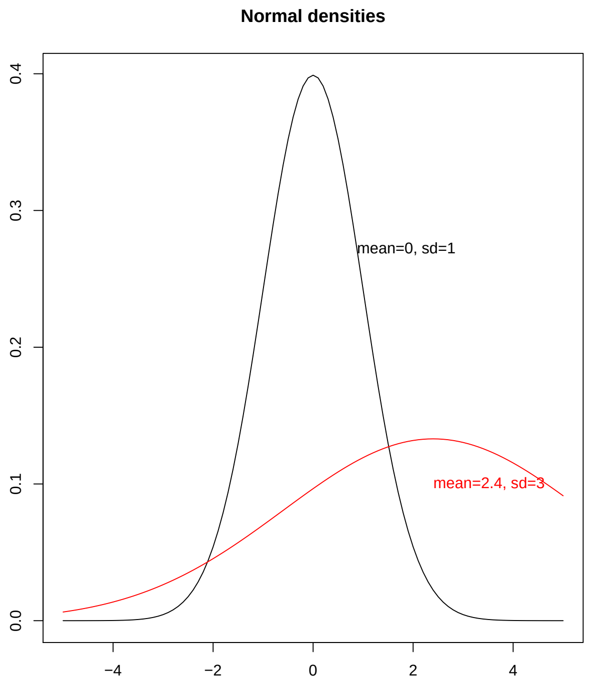
$$\phi_{\mu,\sigma}(x) := \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (-\infty < x < \infty)$$

is the *density of the Normal distribution with mean μ and standard deviation $\sigma > 0$.*

For $\mu = 0$, $\sigma = 1$ we obtain the *density of the standard Normal distribution*[5]

$$\phi(x) := \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (-\infty < x < \infty).$$

[5] We also refer to this density as the "normal curve."


Normal densities
mean=0, sd=1
mean=2.4, sd=3
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4

Review. Normal distribution.

Facts

1. *Total area under the curve* $\phi_{\mu,\sigma}$

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}dx} = 1.$$

2. $\phi_{\mu,\sigma}$ *is symmetric about* $x = \mu$, *i.e.*

$$\phi_{\mu,\sigma}(\mu + x) = \phi_{\mu,\sigma}(\mu - x).$$

3. *The points of inflection of* $\phi_{\mu,\sigma}$ *are*

$$\left(\mu \pm \sigma, \frac{1}{\sqrt{2\pi e}\sigma}\right).$$

Review. Normal distribution. Area under the curve.

Definition (Cumulative distribution function of Normal)

The area under the density $\phi(x)$ up to $x = z$ is denoted

$$\Phi(z) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}} dx.$$

Φ is the so-called *cumulative distribution function of the standard Normal distribution*.[6]

In particular, we have $\Phi(-z) = 1 - \Phi(z)$ and $\Phi(0) = \frac{1}{2}$ from the symmetry of ϕ.

[6] There is no simple formula for the indefinite integral $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}} dx$. Values for Φ are tabulated, see Table 2 in Appendix B.

Review. Normal distribution: empirical rule.

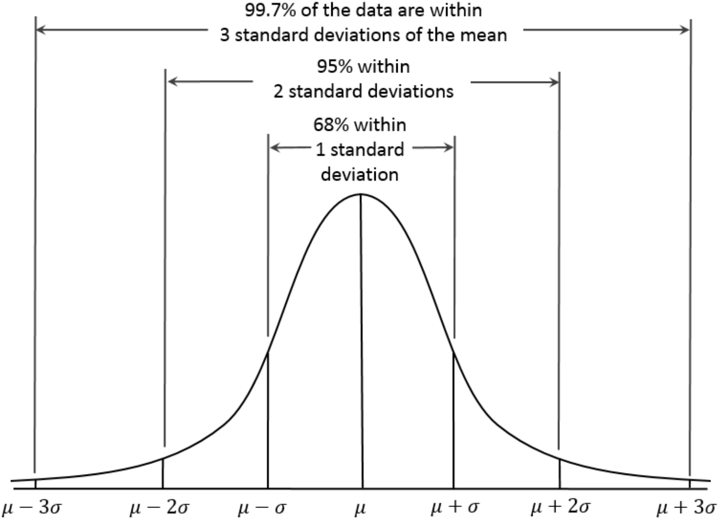


Figure: Empirical rule.

Review. Binomial distribution: Normal approximation.

For large n and p not too close to 0 or 1 the histogram of $\mathrm{binomial}(n, p)$ can be well approximated by $\phi_{\mu,\sigma}$, the normal density with mean $\mu = np$ and standard deviation $\sigma = \sqrt{np(1-p)}$.

[show approximation on `https://www.geogebra.org/m/CmHJuJxs`]

Review. Binomial distribution: Normal approximation.

Fact (Normal approximation to binomial distribution)

For n independent trials with success parameter p

$$\mathbb{P}\left\{a \text{ to } b \text{ successes}\right\} \approx \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right),$$

provided n is large enough, where $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

Review. Binomial distribution: Normal approximation.

Remark (Rule of thumb)

The normal approximation cannot always be accurate, of course.
As a rule of thumb, the Normal approximation works better the larger σ is and the closer p is to $\frac{1}{2}$.
A frequently used rule is that the approximation is reasonable, if

$$np > 5 \text{ and } n(1-p) > 5.$$

Review. Binomial distribution: Fluctuation in number of successes.

Applying the Normal approximation and using the empirical rule, we have

$$\mathbb{P}\{\mu - \sigma \text{ to } \mu + \sigma \text{ successes in } n \text{ trials}\} \approx 68\%$$
$$\mathbb{P}\{\mu - 2\sigma \text{ to } \mu + 2\sigma \text{ successes in } n \text{ trials}\} \approx 95\%$$
$$\mathbb{P}\{\mu - 3\sigma \text{ to } \mu + 3\sigma \text{ successes in } n \text{ trials}\} \approx 99.7\%$$
$$\mathbb{P}\{\mu - 4\sigma \text{ to } \mu + 4\sigma \text{ successes in } n \text{ trials}\} \approx 99.99\%.$$

Typical size of fluctuation in the number of successes is

$$\sigma = \sqrt{np(1-p)}.$$

Typical size of fluctuation in the proportion of successes is

$$\frac{\sigma}{n} = \sqrt{\frac{p(1-p)}{n}}.$$

So: As n grows variability in #successes increases while variability in proportion of successes decreases.

Review. Binomial distribution: Fluctuation in number of successes.

Consider a large number n of independent trials with success probability p on each.

Fact (Square root law)

- *With high probability, the number of successes will lie in an interval centered at mean np with width a moderate multiple of $\sqrt{n}$.*
- *With high probability, proportion of successes will lie in an interval centered at p with width a moderate multiple of $1/\sqrt{n}$.*

In particular, as n increases, proportion of successes tends to p with high probability.

Review. Binomial distribution: Law of large numbers.

Consider n independent trials with success probability p on each. Then for each $\varepsilon > 0$

$$\mathbb{P}\left\{\left|\frac{\text{\#successes in } n \text{ trials}}{n} - p\right| \le \epsilon\right\} \to 1$$

as $n \to \infty$.

In words: as n increases, the proportion of successes in n independent trials will be very close to p with high probability.

Review. Square root law.

We are now more generally interested in distribution of

$$S_n := X_1 + X_2 + \cdots + X_n,$$

where the $X_1, X_2, \ldots$ are independent with some common distribution with $\mu = \mathbb{E}[X_1]$, $\sigma = \mathrm{SD}(X_1)$.

$$\bar{X}_n := \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{S_n}{n}.$$

We find

$$\mathbb{E}[S_n] = n\mu \quad \mathbb{E}[\bar{X}_n] = \mu$$

$$\mathrm{Var}(S_n) = n\sigma^2 \quad \boxed{\mathrm{SD}(S_n) = \sqrt{n}\sigma}$$

$$\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n} \quad \boxed{\mathrm{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}}$$

This is an immediate generalization of the <u>square root law</u> that we saw for the binomial distribution.

Review. Law of large numbers.

Let's take another careful look at

$$\bar{X} \coloneqq \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{S_n}{n}.$$

$$\mathbb{E}[\bar{X}_n] = \mu \quad \mathrm{SD}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

As n grows large, the average

$$\bar{X}_n$$

of a sequence of independent draws

has mean μ, and its spread $\dfrac{\sigma}{\sqrt{n}}$ decreases.

In other words, $\bar{X}_n$ is more and more concentrated around μ and is eventually constant.

Review. Law of large numbers.

Fact ((Weak) Law of large numbers)

Let $X_1, X_2, \dots$ be independent r.v.s with common distribution and mean $\mu = \mathbb{E}[X_1]$. As n grows, the average

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

tends to μ with probability approaching 1.

Review. Normal approximation.

Fact (Normal approximation)

Let $X_1, X_2, \ldots$ *be independent random variables with common distribution with mean* $\mu := \mathbb{E}[X_1]$ *and finite standard deviation* $\sigma := \mathrm{SD}(X_1) > 0$.

Then, for large n *the distribution of the sum*

$$S_n := X_1 + X_2 + \cdots + X_n$$

is approximately normal with mean $n\mu$ *and standard deviation* $\sigma\sqrt{n}$. *Put differently, after standardizing* S_n

$$\mathbb{P}\left\{a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right\} \approx \Phi(b) - \Phi(a) \qquad (a \leq b).$$

As $n \to \infty$ *the error in this approximation tends to* 0.

Review. Normal approximation. Simulations.

How can we statistically confirm the Normal approximation (also known as Central Limit Theorem (CLT))?

Idea: Draw large number of independent samples from quantity of interest,

$$S_n = X_1 + X_2 + \cdots + X_n,$$

and study their histogram!

1. Draw n (=5000) independent samples $X_1, X_2, \ldots, X_n$ from some distribution (here: geometric 0.1).
2. Compute $S_n^1 := X_1 + X_2 + \cdots + X_n$.
3. Repeat 1. and 2. r times (pick a large r, here: $r = 10000$) to obtain sums
$$S_n^1, S_n^2, \ldots, S_n^r.$$
4. Draw the histogram of $S_n^1, S_n^2, \ldots, S_n^r$.

[R script]

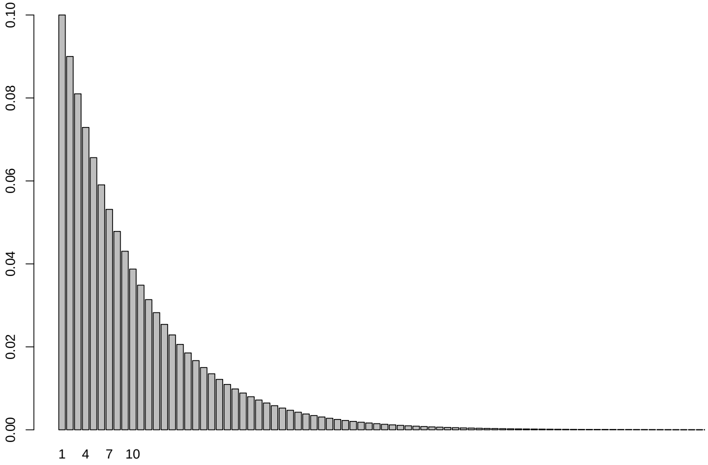


Figure: Probability mass function of geometric distribution.

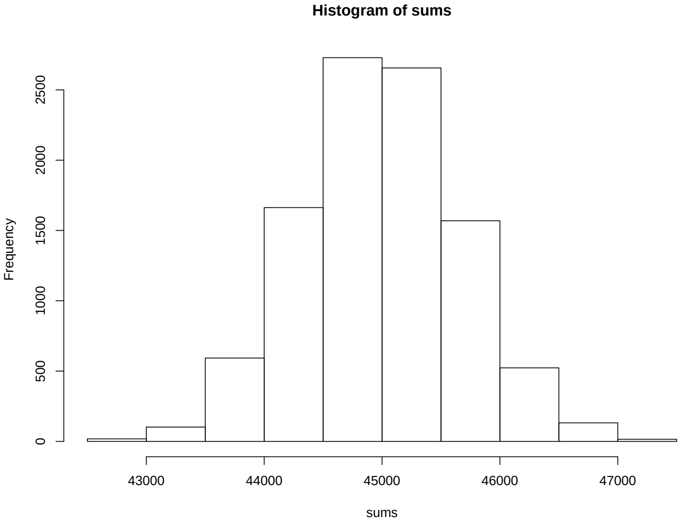


Figure: Histogram of sums $S^1_{5000}, \ldots, S^{10000}_{5000}$.

[Sanity check?!]

Simple Random Sampling. Aside: drawing samples.

Example (Quality control)

Large manufacturer produces cars. Number N of cars produced in one day differs considerably from day to day and is not known beforehand. Each day a random sample of $n = 200$ cars is to be drawn (without replacement) to be tested for failures.

Cars arrive from the assembly line one by one. For each car test drivers have to decide immediately whether the car is shipped or tested. They can park n cars that are to be tested.

Study a nice algorithm that solves this problem in HW 7.7.27.

Example (Hospital discharges)

Would like to quantify how likely it is that the interval

$$(\bar{X} - 2\sigma_{\bar{X}}, \bar{X} + 2\sigma_{\bar{X}}) \approx (662.0, 974.0)$$

contains μ.
To compute this probability, need distribution of $\bar{X}$.

In principle, we know the distribution of $\bar{X}$, as we can work it out in terms of the $\zeta_1, \ldots, \zeta_m$. However, studying this distribution analytically is unfeasible.

[Chebyshev's inequality gives a crude upper bound on this probability.]

Recurring theme: as the setting seems to be too complicated to answer our question, let's make (reasonable) simplifying assumptions.

Example (Hospital discharges)

Simplifying assumption: suppose sample size n is large.

Idea: approximate $\bar{X}$ by Normal distribution (provided n large enough).[7]
That is, approximately $\bar{X} \sim \mathcal{N}(\mu, \sigma_{\bar{X}}^2)$, so

$$\mathbb{P}\left\{a \leq \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \leq b\right\} \approx \Phi(b) - \Phi(a).$$

Recall

$$\mathbb{E}\bar{X} = \mu = 814.6, \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \approx 78.0.$$

[7] For sampling with replacement this approximation is justified by the CLT. If samples are drawn without replacement one has to work harder to show that for large n, but still small compared to N, this is still a good approximation.

Example (Hospital discharges)

Now

$$\mu \in (\bar{X} - 2\sigma_{\bar{X}}, \bar{X} + 2\sigma_{\bar{X}}) \Leftrightarrow -2 \leq \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \leq 2,$$

hence

$$\mathbb{P}\left\{\mu \in (\bar{X} - 2\sigma_{\bar{X}}, \bar{X} + \sigma_{\bar{X}})\right\} = \mathbb{P}\left\{-2 \leq \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \leq 2\right\}$$
$$= \Phi(2) - \Phi(-2) = 2\Phi(2) - 1 = 0.954.$$

Thus, the probability that μ differs from $\bar{X} = 818.0$ by more than $2\sigma_{\bar{X}} = 78.0$ is about 0.046 or 4.6%.
For this reason, $(\bar{X} - 2\sigma_{\bar{X}}, \bar{X} + 2\sigma_{\bar{X}})$ is called a 95.4% *confidence interval* for the population mean μ.

Simple random sampling

Estimating the population variance

Simple random sampling

Saw that standard error

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n-1}{N-1}}$$

of sample mean $\bar{X}$ (and other estimators) depends on

$$n \text{ and } \sigma \text{ (and } N\text{).}$$

However, population standard deviation σ (and N) is usually not known.

Idea: Estimate σ^2 from the data.

Natural candidate: sample variance $\boxed{\hat{\sigma}^2 := \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2.}$

Theorem

$$\mathbb{E}\hat{\sigma}^2 = \sigma^2 \frac{n-1}{n} \frac{N}{N-1},$$

in particular, $\hat{\sigma}^2$ is a biased estimator for σ^2.

Simple random sampling

Remark

For any population parameter that we might be interested in, call it θ, we could come up with an *estimator*

$$\hat{\Theta} = \hat{\Theta}(x_1, \ldots, x_n)$$

for θ.
Morally speaking, $\hat{\Theta}$ is our "best guess" for θ, after seeing the data $x_1, \ldots, x_n$.
Most of the time, we assume $x_1, \ldots, x_n$ to be samples from some r.v.s $X_1, \ldots, X_n$. If

$$\mathbb{E}[\hat{\Theta}(X_1, \ldots, X_n)] = \theta$$

then $\hat{\Theta}$ is called an *unbiased estimator* for θ.

Simple random sampling

Example

1. Since $\mathbb{E}[\bar{X}] = \mu$, $\bar{X}$ is an unbiased estimator for μ.
2. Just saw that $\mathbb{E}[\hat{\sigma}] \neq \sigma$, so $\hat{\sigma}$ is a biased estimator for σ.

Simple random sampling

Consequently, $\hat{\sigma}^2 \frac{n}{n-1} \frac{N-1}{N}$ is an unbiased estimator for σ^2.
Here however, we are interested in an unbiased estimator for $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \frac{N-n}{N-1}$.

Corollary

$$s_{\bar{X}}^2 := \frac{\hat{\sigma}^2}{n} \frac{N-n}{N-1} = \frac{\hat{\sigma}^2}{n} \frac{n}{n-1} \frac{N-1}{N} \frac{N-n}{N-1}$$
$$= \frac{s^2}{n}(1 - \frac{n}{N})$$

is an unbiased estimator for $\sigma_{\bar{X}}^2$, where

$$s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

The estimator $s_{\bar{X}}^2$ is called the *estimated standard error*.

Simple random sampling

Example (Hospital discharges)

Population parameters

$$N = 393, \qquad \mu = 814.6, \qquad \sigma = 589.7$$

Taking a sample of size $n = 50$ yields

$$\bar{X} = 815.92, \qquad s \approx 565.85,$$

hence an estimated standard error of

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}} \approx 74.76.$$

Recall that the true value for the standard error was

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \approx 78.0.$$

Confidence intervals.

Example: Opinion polls. Consider town of $N = 25,000$ eligible voters. Taking a simple random sample $X_1, \ldots, X_{1600}$ of size $n = 1,600 = 40^2$ we find that 917 support Democrats. Suppose we try to estimate percentage

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = p$$

of people who support Democrats, where

$$x_i = \begin{cases} 1 & \text{if } i\text{th person votes for Democrats} \\ 0 & \text{otherwise.} \end{cases}$$

Notice that

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \frac{2}{N}\mu\sum_{i=1}^{N} x_i + \mu^2 = p(1-p).$$

Confidence intervals.

Example: Opinion polls, cont. Idea: use

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{917}{1600} \approx 0.57$$

as estimator for p. We know (simple random sampling):

$$\mathbb{E}\hat{p} = p, \qquad \sigma_{\hat{p}} = \sqrt{\mathrm{Var}(\hat{p})} = \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n-1}{N-1}} \approx \frac{\sigma}{40},$$

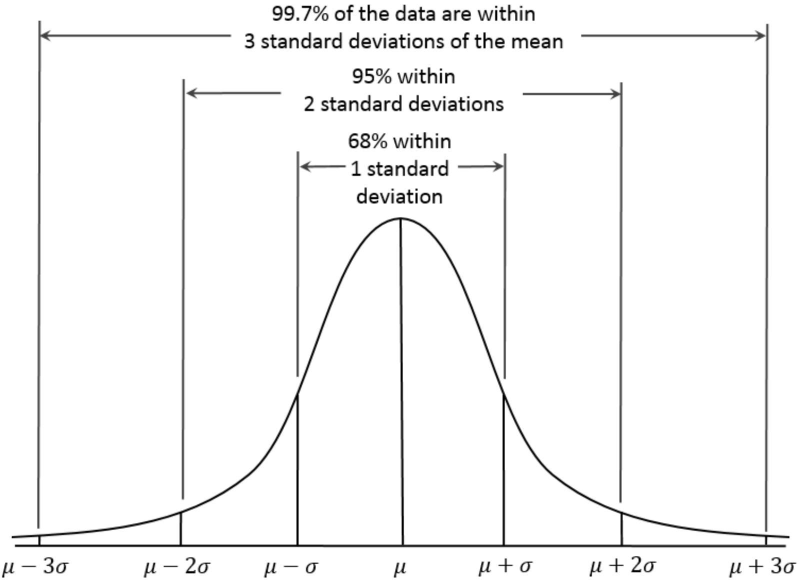
since sampling fraction $n/N = 1,600/25,000 = 0.064$ is small. Moreover, since $\sigma = \sqrt{p(1-p)}$ not known, estimate it by

$$\hat{\sigma} = \sqrt{\hat{p}(1-\hat{p})} \approx 0.5,$$

i.e. estimate standard error $\sigma_{\hat{p}}$ by $0.0125 = 1.25\%$. Put differently, the percentage $\hat{p} \approx 0.57$ of Democrats in the sample is likely to be off the percentage of Democrats among all $25,000$ eligible voters by 1.25 percentage points or so.

Confidence intervals.

Recall the empirical rule for the standard Normal distribution.



Confidence intervals.

Example: Opinion polls, cont. Approximate $\hat{p}$ ($= \bar{X}$) by $\mathcal{N}(\hat{p}, \hat{p}(1-\hat{p})/n) = \mathcal{N}(0.57, 0.25/1600)$ (due to CLT). Hence

Table: Confidence intervals

$\hat{p} \pm 1.25\% = [0.55, 0.58]$	68.3%	confidence interval for p
$\hat{p} \pm 2 \times 1.25\% = [0.54, 0.6]$	95.5%	"
$\hat{p} \pm 3 \times 1.25\% = [0.53, 0.61]$	99.7%	"