STAT 390 A
Statistical Methods in Engineering and Science
Week 6 Lectures - Part 1 & 2 – Spring 2023
Asymptotic Results

Alexander Giessing

Department of Statistics

University of Washington

April 28, 2023

Outline

Averages vary less

- For many experiments concerning natural phenomena one finds that performing the pocedure twice under (what seem) identical conditions results in two different outcomes.
  - measuring the speed of light.
  - your performance in an exam.
  - ...
- Why? Because uncontrollable factors cause "random" variation.
- Solution: Repeat the experiment a number of times and average the result in some way.
- We now discuss why this works so well in practice!

Simple Random Sample (Mathematical Definition)

Simple Random Sample (Heuristic Definition)

A simple random sample is a sample which is constructed in such a way that each member of the population has equal chance of being part of the sample.

Simple Random Sample (Rigorous Definition)

A sequence of $X_1, \ldots, X_n$ random variables is called a simple random sample if

- the X_i's are independent, and
- the X_i's have the same probability distribution.

We say that the X_i's are independent and identically distributed (i.i.d.).

- Think of a simple random sample as the realizations of n repetitions of a particular measurement or experiment.

Expectation and Variance of an Average

Expectation and variance of an average

Let $X_1, \ldots, X_n$ be a simple random sample random variables with mean μ and variance σ^2. Let $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ be the average. Then,

$$\mathrm{E}[\bar{X}_n] = \mu \quad \text{and} \quad \mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

- Observe that the standard deviation $S.D.(\bar{X}_n)$ is less than that of a single X_i by a factor of $\sqrt{n}$.
- Thus, the "typical distance" of $\bar{X}_n$ from μ is smaller than the "typical distance" of X_i from μ.

Expectation and Variance of an Average (Cont.)

Derivation:

$$\begin{aligned}
\mathrm{E}[\bar{X}] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=}^{n} X_i\right] \\
&\overset{(a)}{=} \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[X_i] \\
&\overset{(b)}{=} \frac{1}{n}\sum_{i=1}^{n} \mu \\
&= \mu,
\end{aligned}$$

where (a) holds by linearity of expectation and (b) because the X_i's are identical (and thus $\mathrm{E}[X_1] = \ldots = \mathrm{E}[X_n] = \mu$).

Expectation and Variance of an Average (Cont.)

$$
\begin{aligned}
\mathrm{Var}(\bar{X}_n) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&\overset{(a)}{=} \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu, \frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \\
&\overset{(b)}{=} \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}\left(X_i - \mu, X_j - \mu\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Cov}(X_i - \mu, X_i - \mu) + \frac{1}{n^2}\sum_{i\neq j} \mathrm{Cov}(X_i - \mu, X_j - \mu) \\
&\overset{(c)}{=} \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) + 0 \\
&\overset{(d)}{=} \frac{\sigma^2}{n},
\end{aligned}
$$

where (a) follows from the definition of variance and covariance, (b) holds by linearity of the covariance, (c) holds because the X_i's are independent, and (d) hold because the X_i's are identical (hence, $\mathrm{Var}(X_1) = \ldots = \mathrm{Var}(X_n) = \sigma^2$).

Example: Average of Random Variables

- Let $X_1, \ldots X_n$ be a simpel random sample from the $Gamma(2,1)$ distribution with density

$$f_{X_i}(x) = xe^{-x} \quad \text{for } x \geq 0.$$

- One can show that the average $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ has density

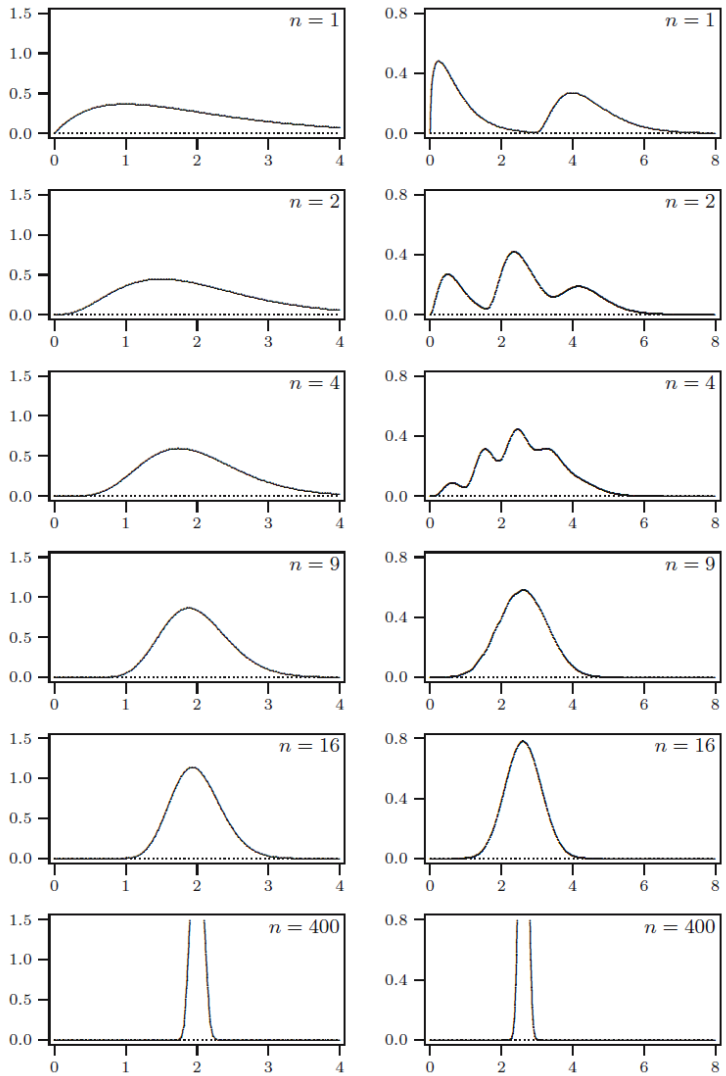
$$f_{\bar{X}_n}(x) = \frac{n(nx)^{2n-1}e^{-nx}}{(2n-1)!} \quad \text{for } x \geq 0.$$

  This is the density of the $Gamma(2n, n)$ distribution.

- Having determined explicitely the distribution of $\bar{X}_n$ we can investigate what happens as n increases (see next slid, left column).

Example: Average of Random Variables (Cont.)

Densities of $\bar{X}_n$ for different sample sizes n.



(left column: Gamma density; right column: a bimodal density.)

Example: Average of Random Variables (Cont.)

- Previous slide:
  - left column: Gamma distribution.
  - right column: a bimodal distribution.
- The graphs on the previous slide show that, as n increases, there is a "contraction" of the probability mass near the expected value μ:
  - μ of the gamma distribution $= 2$
  - μ of the bimodal distirbution $= 2.625$
- This contraction happens for (almost) any distribution, symmetric, asymmetric, uni- or multi-modal.
- Why? How can this be explained?

Chebyshev's inequality

Chebyshev's inequality

For an arbitrary random variable Y and any $a > 0$,

$$P(|Y - \mathrm{E}[Y]| \geq a) \leq \frac{1}{a^2}\mathrm{Var}(Y).$$

- The probability that Y deviates from its mean $\mathrm{E}[X]$ by more than a decreases quadratically in a.

Derivation (for continuous random variable only):

$$\begin{aligned}
\mathrm{Var}(Y) &= \int_{\infty}^{\infty} (y-\mu)^2 f_Y(y)dy \\
&\geq \int_{|y-\mu|\geq a} (y-\mu)^2 f_Y(y)dy \\
&\geq \int_{|y-\mu|\geq a} a^2 f_Y(y)dy \\
&= a^2 P(|Y-\mu| \geq a).
\end{aligned}$$

Interpretation: Chebyshev's inequality

- Setting $a = k\sigma$ in Chebyshev's inequality (and switching to the complement), we find

$$P(|Y - \mu| < k\sigma) \geq 1 - \frac{\text{Var}(Y)}{k^2\sigma^2} = 1 - \frac{1}{k^2}.$$

- For $k = 2, 3, 4$ the right-hand side is $3/4, 8/9$ and $15/16$, respectively. In plain English, "the probability that Y is within 2 (3, or 4) standard deviations from μ is at least $3/4$ ($8/9$, or $15/16$)".

- This phenomenon is also known as the following rule:

THE "$\mu \pm \sigma$" RULE

Most of the probability mass of a random variable is within a few standard deviations σ from its expectation μ.

The Weak Law of Large Numbers (WLLN)

The "$\mu \pm \sigma$" rule can be significantly sharpened.

The Weak Law of Large Numbers (WLLN)

Let $X_1, \ldots, X_n$ be a simple random sample of random variables with mean μ and finite variance σ^2. Let $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ be the average. Then, for any $\varepsilon > 0$,

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0.$$

- The probability that $\bar{X}_n$ deviates from μ by more than ϵ goes to zero as the sample size increases.
- This is a quantitative version of the "contraction" observed in the plots of the density of $\bar{X}_n$.

Derivation:

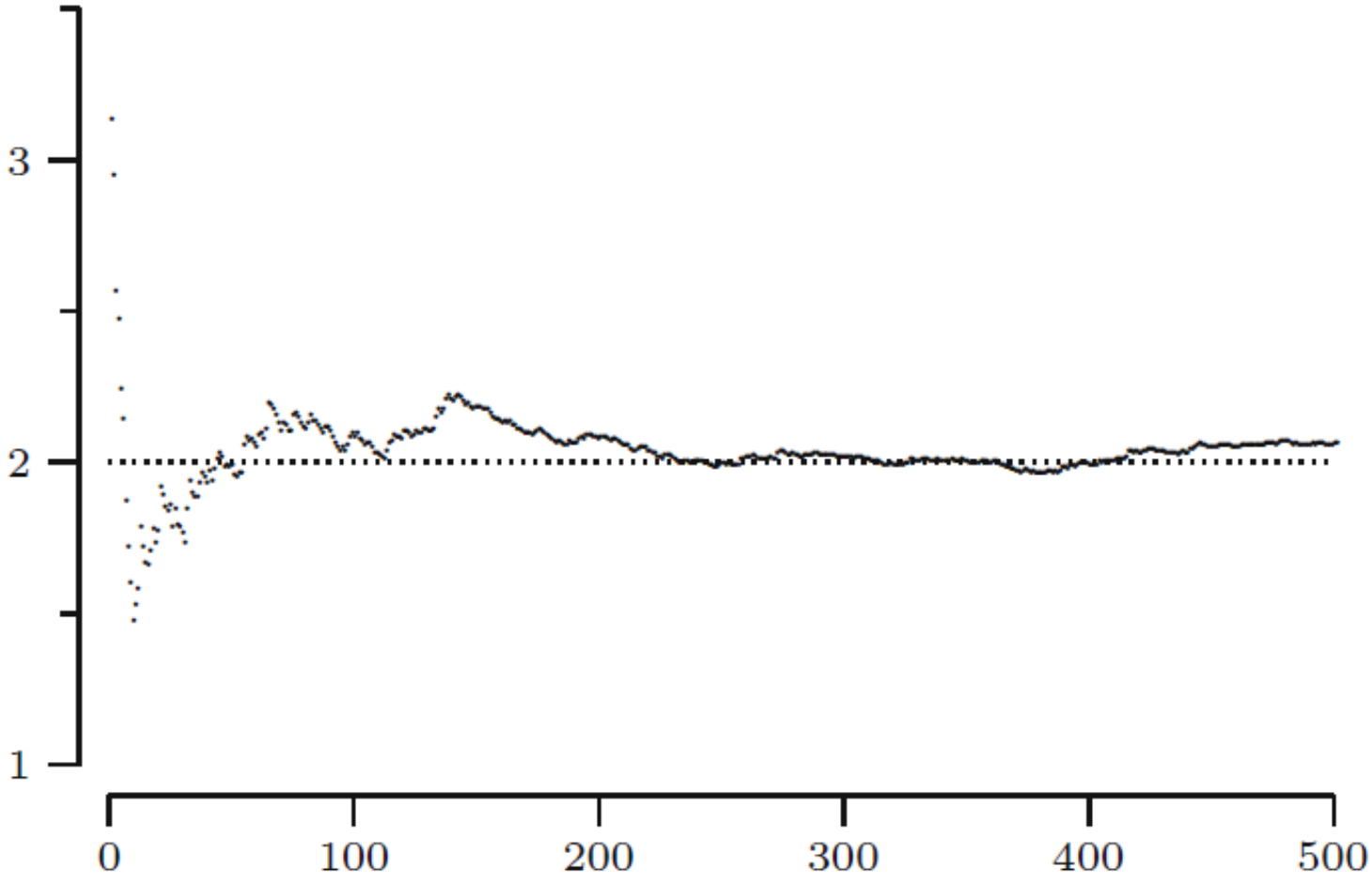
$$P(|\bar{X}_n - \mu| > \varepsilon) = P(|\bar{X}_n - \mathrm{E}[\bar{X}_n]| > \varepsilon) \leq \frac{1}{\varepsilon^2} \mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{\varepsilon^2 n} \to 0 \text{ as } n \to \infty.$$

WLLN and connection to experimental work

- Imagine conducting a series of n identical experiments
  - measuring the speed of light,
  - duration of a chemical reaction,
  - ...
- Since the experimental setup is complicated, your measurements vary quite a bit around the "true" value you are after.
- You know about the WLLN and therefore decide to average over all measurements and report the average as your estimate of the "true" value.
- We can simulate this approach to get a feeling for what the statement "$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0$" means in practice.

WLLN and connection to experimental work (Cont.)

Average of realizations of a sequence of Gamma(2,1) distributed random variables with mean $\mu = 2$ (x-axis: sample size, y-axis: average). The average converges to $\mu = 2$ as the sample size increases.



Example: Recovering the probability of an event

Suppose we want to know the probability $p = P(X \in C)$ *where* $C = (a, b]$.

- Idea: Use the relative frequency of many i.i.d. X_i's hitting the set $C = (a, b]$, i.e. define for $i = 1, \ldots n$,

$$Y_i = \begin{cases} 1 & \text{if } X_i \in C, \\ 0 & \text{if } X_i \notin C. \end{cases}$$

- Note that

$$\begin{aligned} \mathrm{E}[Y_i] &= 1 \times P(X_i \in C) + 0 \times P(X_i \notin C) = P(X_i \in C) = p, \\ \mathrm{Var}(Y_i) &= P(X_i \in C) - P(X_i \in C)^2 = p(1-p). \end{aligned}$$

- Define the relative frequency with which the X_i's hit C as

$$\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i = \frac{\#\{i : X_i \in C,\ i \leq n\}}{n}.$$

- The WLLN applied to $\bar{Y}_n$ yields,

$$\lim_{n \to \infty} P(|\bar{Y}_n - p| > \varepsilon) = 0.$$

Thus, $\bar{Y}_n \approx p$ if the sample size is large.

Example: Recovering the probability density function

Suppose X is a continuous random variable with distribution F and corresponding density f.

- Let $C = (a-h, a+h]$ for some (small) $h > 0$ and for a simple random sample $X_1, \ldots, X_n$ define

$$Y_i = \begin{cases} 1 & \text{if } X_i \in C, \\ 0 & \text{if } X_i \notin C. \end{cases}$$

- By the result on the previous slide, for large sample sizes n,

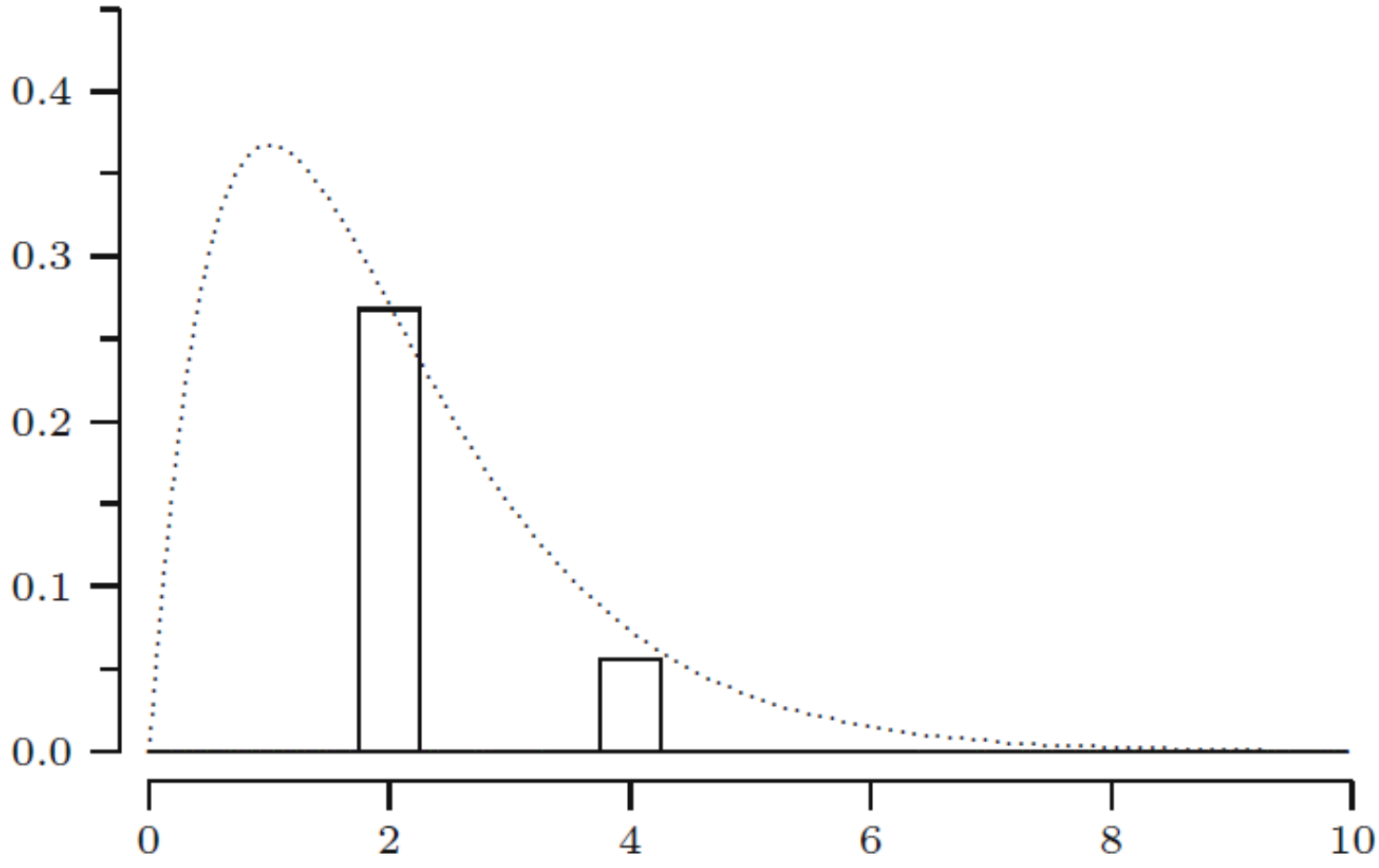
$$\bar{Y}_n \approx p = P(X \in C) = \int_{a-h}^{a+h} f(x)dx \approx 2hf(a).$$

- This approximate identity suggest to estimate the density f evaluated at a as follows:

$$f(a) \approx \frac{\bar{Y}_n}{2h} = \frac{\#\{i : X_i \in C,\ i \leq n\}}{n \times \text{length of } C}.$$

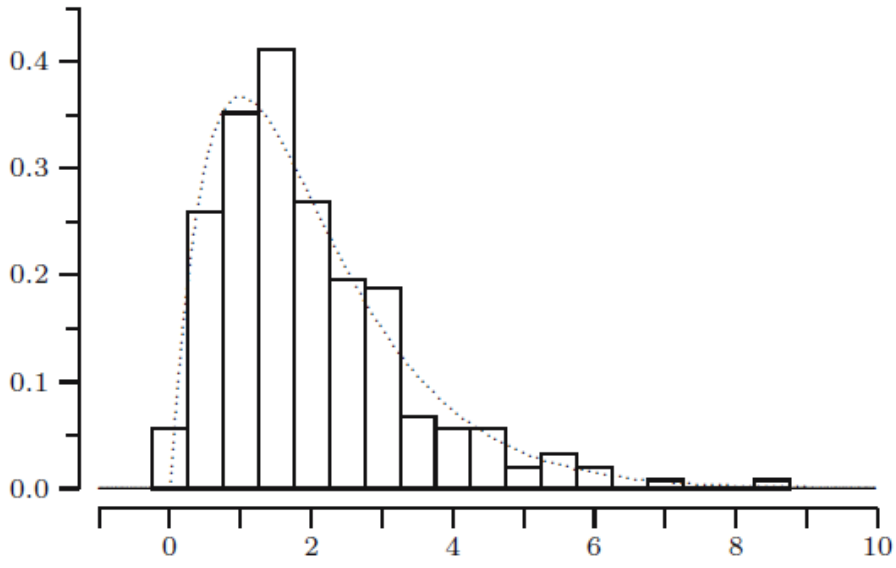
Example: Recovering the pdf (Cont.)

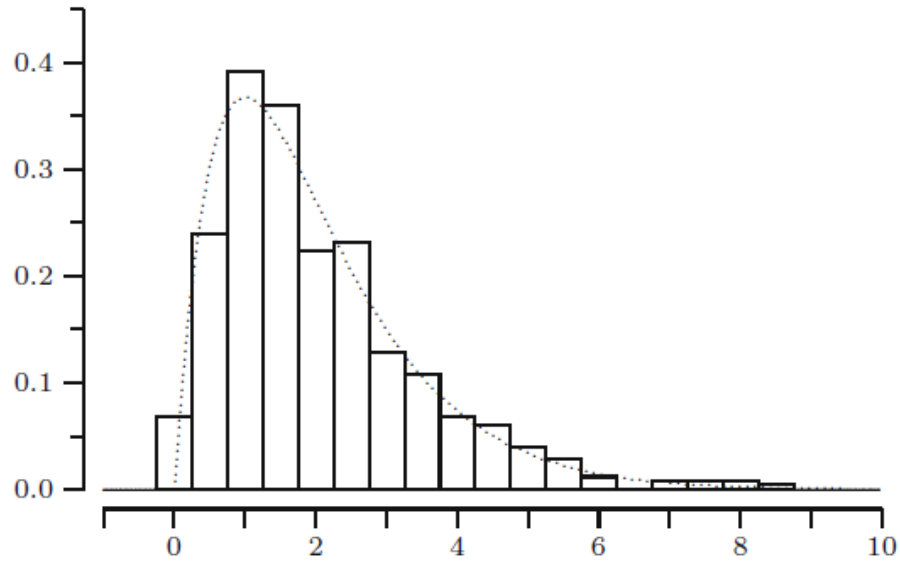
The following plot is based on a simple random sample of $n = 500$ *Gamma(2,1) distributed random variables. The hight of the bars are estimates of* f *at* $a = 2$ *and* $a = 4$ *for* $h = 0.25$*, the width of the bars is* $2h$*. The broken line is the density of the Gamma(2,1) distribution.*



Example: Recovering the pdf (Cont.)

We are usually not interested of estimates of the density in just a few points. The following two plots show histograms with bin size $2h$ and $h = 0.25$ based on two different simple random samples of size $n = 500$ of Gamma(2,1) distributed random variables.





Both graphs match the general shape of the density, with some bumps and valleys tht are particular for the two different simple random samples.

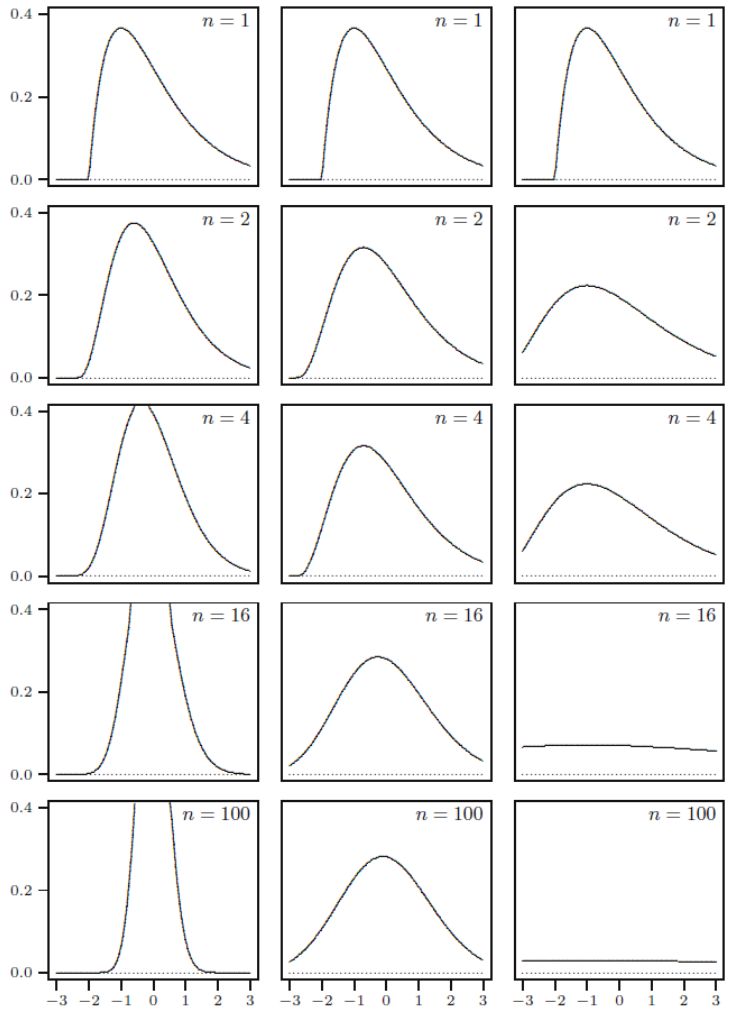
Outline

Standardizing averages

- In the first part of the letcture we have seen that averages vary less:
  - The density of an average $\bar{X}_n$ of i.i.d. random variables $X_1, \ldots X_n$ with mean μ and finite variance σ^2 concentrates around μ as n grows.
  - The WLLN provides the mathematical precise formulation of this phenomenon.
- Reconsider the plots on Slide 6:
  - As n increases, the density of $\bar{X}_n$ becomes more symmetrical around and bell shaped around μ.
  - Eventually, the density of $\bar{X}_n$ collapses into a single spike at μ.
- Is it possible to transform (aka *standardize*) the average $\bar{X}_n$ in some way so that its density "settles down"?
  - Note: $\mathrm{E}[\bar{X}_n] = \mu$ is independent of the sample size n.
  - But: $\mathrm{Var}(\bar{X}_n) = \mathrm{Var}\left(n^{-1}\sum_{i=1}^n X_i\right) = n^{-2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{\sigma^2}{n} \to 0$ as $n \to \infty$!

Finding the correct (stabilizing) standardization

The plots show $\alpha(\bar{X}_n - \mu)$ of n Gamma(2,1) distributed random variables for different values of α and n.



(left column: $\alpha = n^{1/4}$; middle column: $\alpha = \sqrt{n}$; right column: $\alpha = n$.)

Finding the correct (stabilizing) standardization (Cont.)

- Previous slide: $\alpha(\bar{X}_n - \mu)$ of n $Gamma(2,1)$distributed random variables
  - left column: $\alpha = n^{1/4}$, density collapses into a single spike.
  - middle column: $\alpha = \sqrt{n}$, density appears to "settles down".
  - right column: $\alpha = n$, density spreads out/ vanishes.
- The graphs the previous slide show that, as n increases, $\sqrt{n}$ seems to be the standardization that stabilizes the density. Intuitively, this is sensible since

$$\mathrm{Var}\big(\alpha(\bar{X}_n - \mu)\big) = \frac{\alpha^2\sigma^2}{n} \to \begin{cases} 0 & \text{if } \alpha < \sqrt{n} \\ \sigma^2 & \text{if } \alpha = \sqrt{n} \\ \infty & \text{if } \alpha > \sqrt{n} \end{cases} \qquad \text{as} \quad n \to \infty.$$

- Since the density of $\bar{X}_n - \mu$ seems to stabilize with this standardization, what can we say about the (limiting) density as n grows?

Central Limit Theorem (CLT)

The Central Limit Theorem (CLT)

Let $X_1, \ldots, X_n$ be a simple random sample of random variables with mean μ and finite variance σ^2. For $n \geq 1$, define $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ and

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

Then, for any $a \in \mathbb{R}$,

$$\lim_{n \to \infty} F_{Z_n}(a) = \Phi(a),$$

where Φ is the cdf of the standard normal distribution $N(0, 1)$.

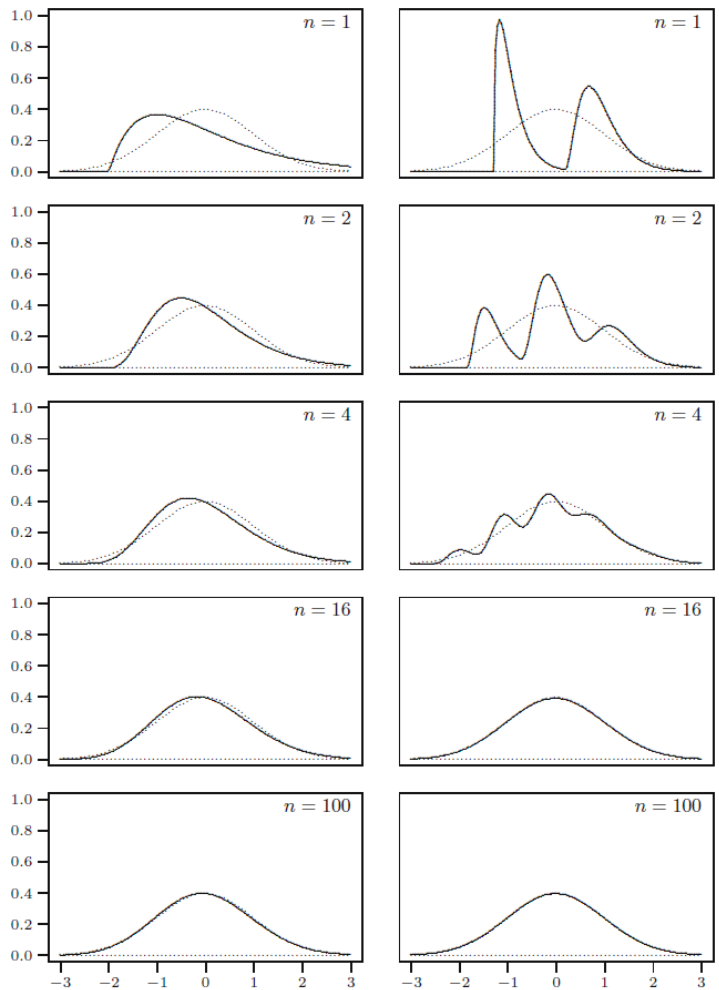
- We say that the distribution function of Z_n converges to the distribution function Φ of the standard normal distribution.
- Note that

$$Z_n = \frac{\bar{X}_n - \mathrm{E}[\bar{X}_n]}{\sqrt{\mathrm{Var}(\bar{X}_n)}},$$

  which is a more direct way to see that Z_n is the standardized average $\bar{X}_n$.

Central Limit Theorem (Cont.)

Densities of standardized averages Z_n for different sample sizes n. The dotted line denotes the density of $N(0,1)$.



(left column: Gamma density; right column: bimodal density.)

Central Limit Theorem (Cont.)

- Previous slide:
  - left column: density of standardize Gamma random variables.
  - right column: density of standardized bimodal distribution.
- The graphs on the previous slide illustrate the CLT:
  - As n increases, the densities of the standardized averages converge to the density of the $N(0, 1)$ distribution.
  - This phenomenon occurs for any distribution with finite mean μ and finite variance σ^2.
- This is the rigorous justification for the "Normal approximation of data histograms" (Lecture Week 4, Part 1, Slide 23).
- When does the CLT (and the WLLN) fail?
  - See the example in homework 5.

Example: Approximating exceedance probabilities

Consider a simple random sample of $n = 500$ *Gamma(2,1) distributed random variables. What is the probability that the sample average* $\bar{X}_n$ *exceeds the mean* $\mu = 2$ *by* 0.06*?*

- Recall that $\mu = \mathrm{E}[X_i] = 2$ and $\sigma^2 = \mathrm{Var}(X_i) = 2$. Therefore,

$$\begin{aligned} P(\bar{X}_n \geq 2.06) &= P(\bar{X}_n - \mu \geq 2.06 - \mu) \\ &= P\left(\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \geq \sqrt{n}\frac{2.06 - \mu}{\sigma}\right) \\ &= P\left(Z_n \geq \frac{0.06\sqrt{n}}{\sqrt{2}}\right) \\ &= P\left(Z_{500} \geq 0.95\right) \\ &= 1 - P\left(Z_{500} < 0.95\right). \end{aligned}$$

- By the CLT we can approxiamte above probability using the density/ distribution of the standard normal $N(0,1)$ distribution,

$$P(\bar{X}_n \geq 2.06) \approx 1 - \Phi(0.95) = 17.11\%.$$

Example: Rounding to the nearest integer

Suppose that an accounted wants to simplify his bookkeeping by rounding amounts to the nearest integer (or, think of him as pocketing some extra money). What is the probability that the cumulative rounding error of 100 transactions exceeds \$10 (which would trigger an external audit)?

- For simplicity we assume that the rounding errors $X_1, \ldots X_{100}$ can be modeled as independent $Unif(-0.5, 0.5)$ distributed random variables.
- Then, we want to comnpute the probability that

$$P(|X_1 + \ldots + X_{100}| > 10) = P(X_1 + \ldots + X_{100} \leq -10) + P(X_1 + \ldots + X_{100} > 10).$$

- Next, recall that $\mu = \mathrm{E}[X_i] = 0$ and $\sigma^2 = \mathrm{Var}(X_i) = 1/12$ and consider

$$\begin{aligned} P(X_1 + \ldots + X_{100} > 10) &= P(X_1 + \ldots + X_{100} - n\mu > 10 - n\mu) \\ &= P\left(\frac{X_1 + \ldots + X_{100} - n\mu}{\sigma\sqrt{n}} > \frac{10 - n\mu}{\sigma\sqrt{n}}\right) \\ &= P\left(Z_{100} > \frac{10 - 100 \times 0}{\sqrt{1/12}\sqrt{100}}\right) \\ &= P(Z_{100} > 3.46). \end{aligned}$$

Example: Rounding to the nearest integer (Cont.)

- By the CLT we have

$$P(Z_{100} > 3.46) \approx 1 - \Phi(3.46) = 0.0003 = 0.03\%.$$

- Similarly, we can compute (verify at home!)

$$\begin{aligned} P(X_1 + \ldots + X_{100} < -10) &= P(Z_{100} < -3.46) \\ &\approx \Phi(-3.46) = 0.0003 = 0.03\%. \end{aligned}$$

- Thus, we find that the probability of the accountant being caught of pocketing extra money is 0.06%.

- Note that the expected amount of money the accountant pockets is $\mathrm{E}[\bar{X}_n] = 0$. How do the results change if he only pockets positive amounts of money (i.e. $X_i \sim Unif(0, 0.5)$, which is a more reasonable assumption to model a fraudster)?

Normal approximation to the Binomial distr.

- Recall that $X \sim Bin(n, p)$ can be written as

$$X = Y_1 + \ldots + Y_n,$$

where the Y_i's are i.i.d. $Ber(p)$ random variables.

- By the CLT we conclude that for large n, the distribution of

$$\frac{\sqrt{n}(\bar{Y} - p)}{\sqrt{p(1-p)}} = \frac{\sqrt{n}(X/n - p)}{\sqrt{p(1-p)}} = \frac{X - np}{\sqrt{np(1-p)}}$$

can be approximated by the standard normal distribution $N(0, 1)$.

- By Week 4 Lecture, Part 1, Slide 20, this is equivalent to saying that for large n, the distribution of $X \sim Bin(n, p)$ can be approximated by $N\big(np, np(1-p)\big)$.

- **Caveat:** This approximation is often quite poor. Any idea why?

Normal approximation to the Binomial distr. (Cont.)

Since the Binomial distribution is discrete, the normal approximation is more accurate with a so-called "continuity correction".

Normal approximation to Binomial with continuity correction

Let $X \sim Bin(n, p)$. If $n \min(p, 1-p) \geq 10$, then for integers $a \leq b$,

$$\begin{aligned} P(a \leq X \leq b) &= P(a - 0.5 \leq X \leq b + 0.5) \\ &\approx \Phi\left(\frac{b + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - 0.5 - np}{\sqrt{np(1-p)}}\right). \end{aligned}$$

- **Continuity correction:** modification of lower and upper bounds by ± 0.5.
- **Careful!** Continuity correction does not always yields more accurate approximation (. . . but we don't worry about these subtleties).

Example: Normal approximation of many coin tosses

Let X be the number of heads in 100 coin tosses. Thus, $X \sim Bin(100, 0.5)$.

- Note that $\mathrm{E}[X] = 50$ and $SD(X) = \sqrt{100 \times 0.5 \times 0.5} = 5$. Then, for $Z \sim N(0, 1)$,

$$\begin{aligned} P(X = 50) &= P(49.5 \leq X \leq 50.5) \\ &\approx P\left(\frac{49.5 - 50}{5} \leq Z \leq \frac{50.5 - 50}{5}\right) \\ &= 0.5398 - 0.4602 \\ &= 7.96\%. \end{aligned}$$

$$P(X \geq 60) = P(X \geq 59.5) \approx 1 - P(Z \leq 9.5/5) = 2.87\%.$$

$$P(45 \leq X \leq 60) = P(44.5 \leq X \leq 60.5) \approx \Phi(2.1) - \Phi(-1.1) = 83.64\%.$$

Example: WLLN and Normal approx. of lotteries

In a large Statistics class of size 100, students are asked to simulate drawing 400 tickets with replacement from a box with values $0, 2, 5, 8, 10$.

1. What percentage of students will get a sum of 2100 or more?

- By the WLLN, the percentage is approximately the probability of getting a sum larger than 2100.
- Let $X_1, X_2 \ldots, X_{400}$ be the random variables corresponding to the 1st, 2nd, ..., 400th draw, respectively. Then, $X_1, \ldots X_{400}$ are i.i.d. with

$$P(X_1 = 0) = P(X_1 = 2) = P(X_1 = 5) = P(X_1 = 8) = P(X_1 = 10) = \frac{1}{5},$$

and we want to find

$$p = P\left(\sum_{i=1}^{400} X_i \geq 2100\right).$$

Example: WLLN & Normal approx. of lotteries (Cont.)

- To use the CLT, we need to compute the population mean and SD, i.e.

$$\mu = \mathrm{E}[X_i] = \frac{1}{5} \times 0 + \frac{1}{5} \times 2 + \ldots + \frac{1}{5} \times 10 = 5.$$

$$\mathrm{E}[X_i^2] = \frac{1}{5} \times 0^2 + \frac{1}{5} \times 2^2 + \ldots + \frac{1}{5} \times 10^2 = 38.6.$$

Hence,

$$\sigma^2 = \mathrm{Var}(X_i) = \mathrm{E}[X_i^2] - \mu^2 = 38.6 - 5^2 = 13.6 \quad \text{and} \quad \sigma = 3.7.$$

Consequently,

$$p \approx 1 - \Phi\left(\frac{2100 - 400 \times 5}{\sqrt{400} \times 3.7}\right) = 1 - \Phi(1.35) = 9\%.$$

Example: WLLN & Normal approx. of lotteries (Cont.)

2. What is the likely size of the error of the estimate?

- Let $Y_i = 1$ if the ith student gets a sum larger than 2100 and 0 otherwise. Then, $Y_1, \ldots Y_{100}$ are i.i.d. $Ber(p)$.
- Let $\hat{p} := \bar{Y} = \sum_{i=1}^{100} Y_i/100$ be the proportion of students who get a sum larger than 2100. It follows that

$$\mathrm{E}[\hat{p}] = \mathrm{E}[\bar{Y}] = p = 9\%$$

and

$$SD(\hat{p}) = SD(\bar{Y}) = \frac{\sqrt{p(1-p)}}{\sqrt{100}} = 2.86\%,$$

i.e. the estimation (chance) error is about 2.86%.

Example: Classification and Counting

In a city, the average family income is \$40K with an SD of \$30K. Among those, 20% of families have income larger than \$80K. A survey organization takes a random sample of 900 families.

1. What is the chance that the sample average falls between \$38K and \$42K?

- Let $\bar{X}$ be the sample average. Then,

$$\mathrm{E}[\bar{X}] = 40, \qquad SD(\bar{X}) = \frac{30}{\sqrt{900}} = 1.$$

Hence, by the CLT

$$P(38 \le \bar{X} \le 42) \approx \Phi\left(\frac{42-40}{1}\right) - \Phi\left(\frac{38-40}{1}\right) = \Phi(2) - \Phi(-2) \approx 95\%.$$

Example: Classification and Counting

2. What is the chance that between 18% and 22% of the selected families have an income larger than \$80K (i.e. poll error 2%)?

- Let $Y_i = 1$ if the ith draw has income larger than \$80K. Then, $\hat{p} = \bar{Y}$ is the proportion of selected families having income larger than \$80K.
- We want to compute the probability

$$p = P(0.18 \leq \hat{p} \leq 0.22) = P\left(162 \leq \sum_{i=1}^{900} Y_i \leq 198\right).$$

Since $\mathrm{E}[Y_i] = 0.2$ and $SD(Y_i) = \sqrt{0.2 \times 0.8} = 0.4$, it follows that

$$\mathrm{E}\left[\sum_{i=1}^{900} Y_i\right] = 900 \times 0.2 = 180, \quad SD\left(\sum_{i=1}^{900} Y_i\right) = \sqrt{900 \times 0.2 \times 0.8} = 12.$$

By the CLT,

$$p \approx \Phi\left(\frac{198.5 - 180}{12}\right) - \Phi\left(\frac{161.5 - 180}{12}\right) = 87.68\%.$$

Conclusion

- Chance errors or poll errors depend on sample size n, not the population size N (see lottery example $n = 100$, $N = 400$)
- Random sampling has advantages of avoiding biases and allows us to calculate chance errors (variances) (see all examples above).
- A simple random sample is NOT an "arbitrary" sample. "Arbitrary sampling" (without ensuring that the sampled units are independent and identically distributed) creates unintentional biases and leads to unknown chance errors (variances).