The Central Limit Theorem

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Law of large numbers

If $\tilde{x}_1$, $\tilde{x}_2$, $\tilde{x}_3$, ... are independent random variables with mean μ and variance σ^2

$$\tilde{m}_n := \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i$$

$$\mathrm{E}\left[\tilde{m}_n\right] = \mu$$

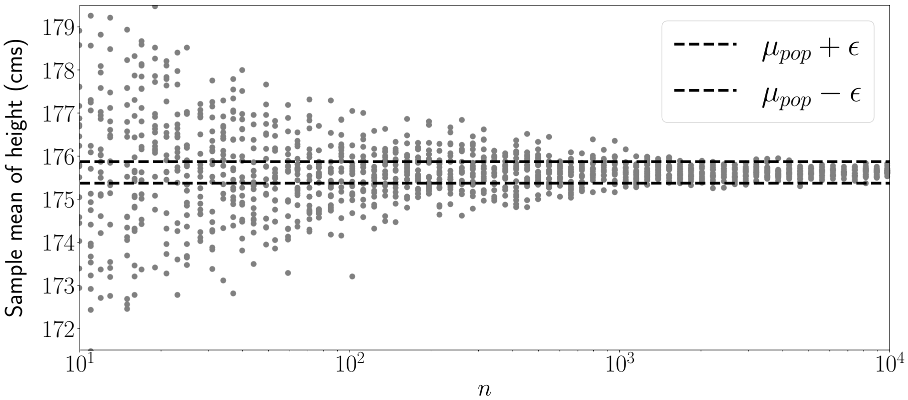
$$\mathrm{Var}\left[\tilde{m}_n\right] = \frac{\sigma^2}{n}$$

$$\mathrm{P}\left(|\tilde{m}_n - \mu| > \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$$

Converges to zero for any ϵ!

Consistency of sample mean

Distribution for fixed n?

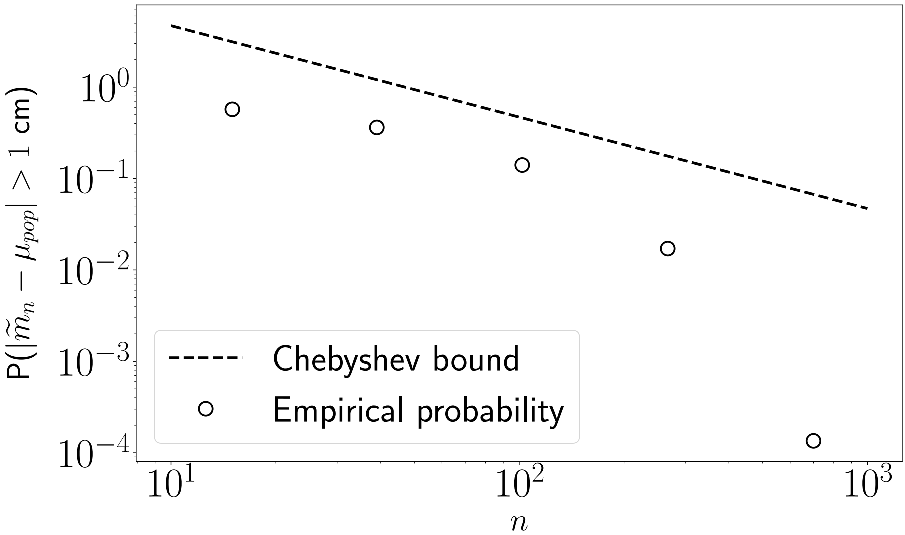


Chebyshev bound

$$\mathrm{P}\left(|\tilde{m}_n - \mu_{\mathsf{pop}}| > \epsilon\right) \leq \frac{\sigma_{\mathsf{pop}}^2}{n\epsilon^2}$$

Is this a good approximation?

No!


Chebyshev bound
Empirical probability
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^1
10^2
10^3
n
P($|\widetilde{m}_n - \mu_{pop}| > 1$ cm)

Goal

Approximate the distribution of the sample mean

$$\tilde{m}_n := \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i$$

Sum of independent discrete random variables

Independent discrete random variables $\tilde{a}$ and $\tilde{b}$ with integer values

The pmf of $\tilde{s} = \tilde{a} + \tilde{b}$ is

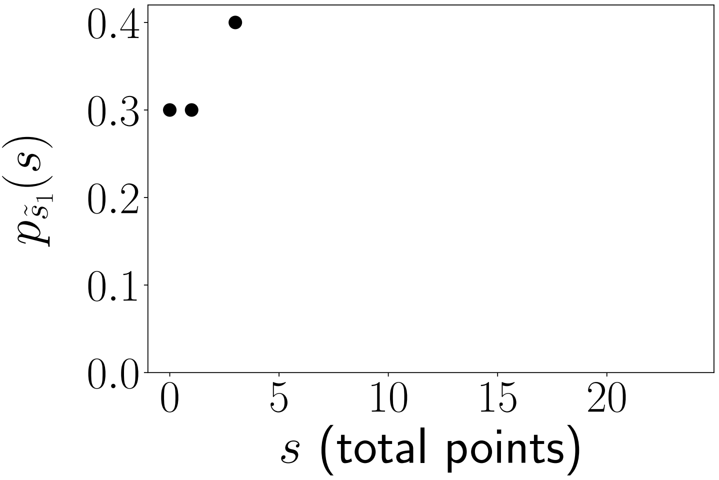
$$p_{\tilde{s}}(s) = \sum_{a=-\infty}^{\infty} p_{\tilde{a}}(a)\, p_{\tilde{b}}(s-a) = p_{\tilde{a}} * p_{\tilde{b}}(s)$$

Independent discrete random variables $\tilde{a}_1$, $\tilde{a}_2$, ..., $\tilde{a}_n$ with integer values

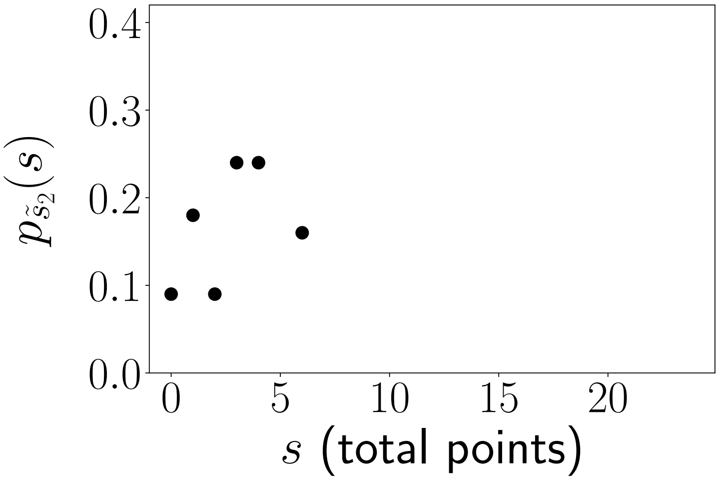
The pmf of $\tilde{s}_n = \sum_{i=1}^{n} \tilde{a}_i$ is

$$p_{\tilde{s}_n}(s) = p_{\tilde{a}_1} * p_{\tilde{a}_2} * \cdots * p_{\tilde{a}_n}(s)$$

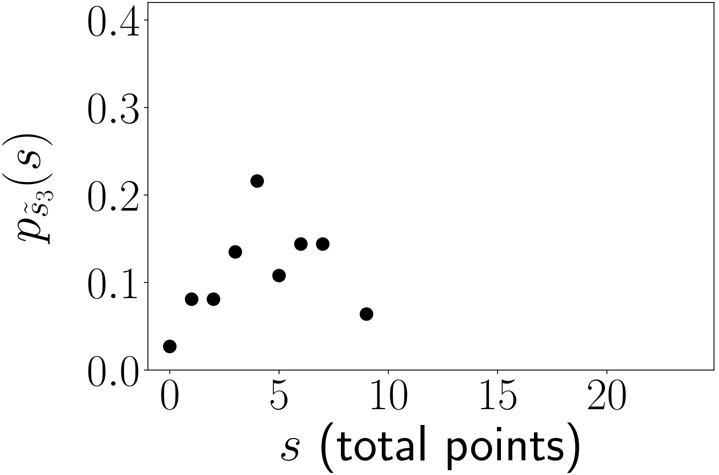
Soccer league: 1 game



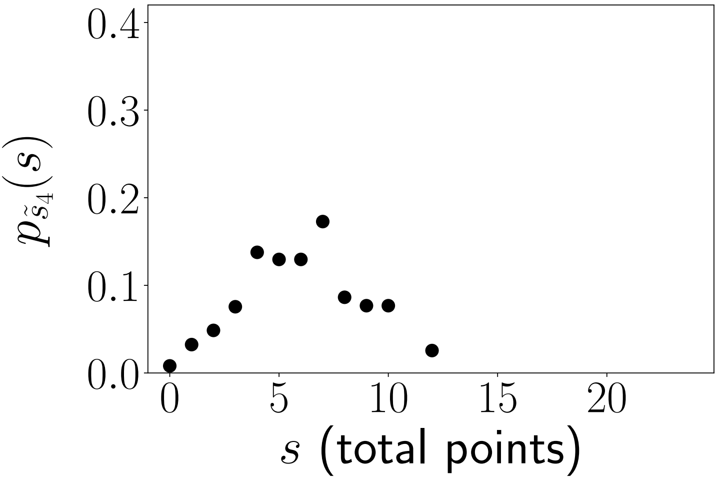
Soccer league: 2 games



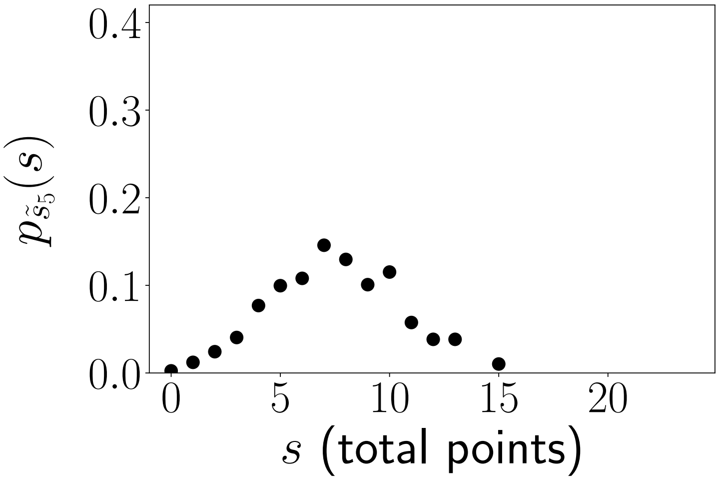
Soccer league: 3 games



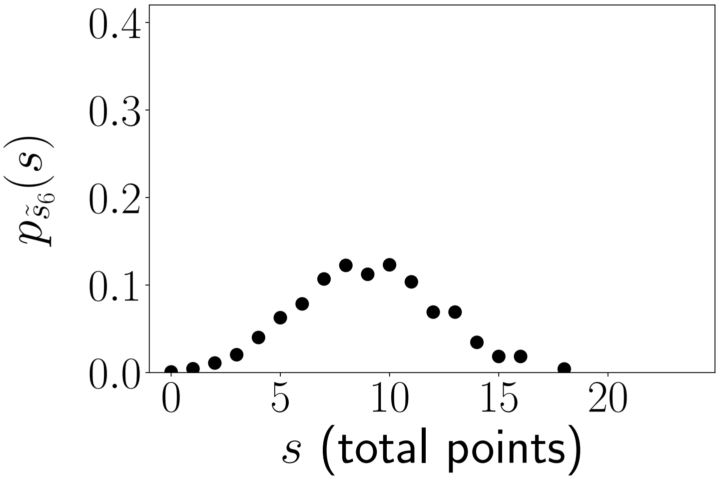
Soccer league: 4 games



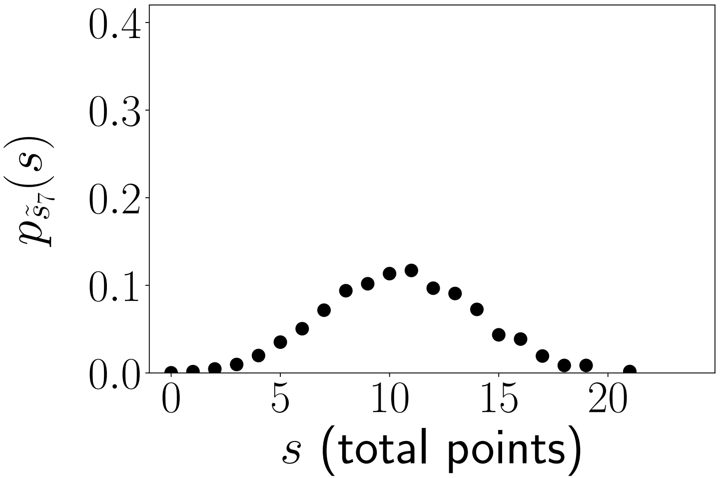
Soccer league: 5 games



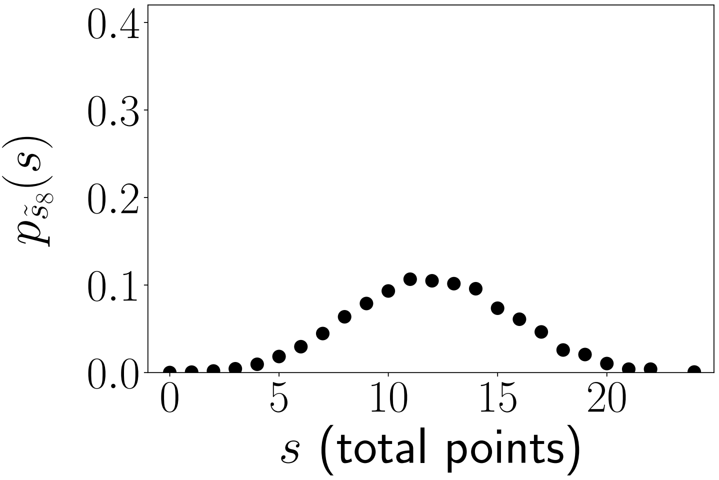
Soccer league: 6 games



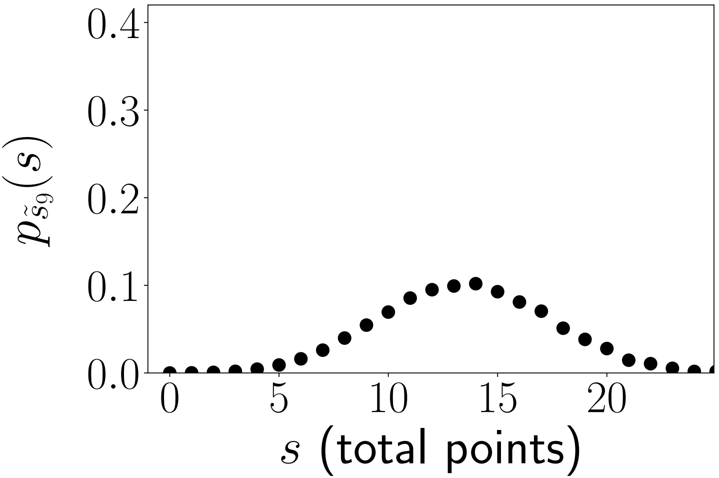
Soccer league: 7 games



Soccer league: 8 games



Soccer league: 9 games

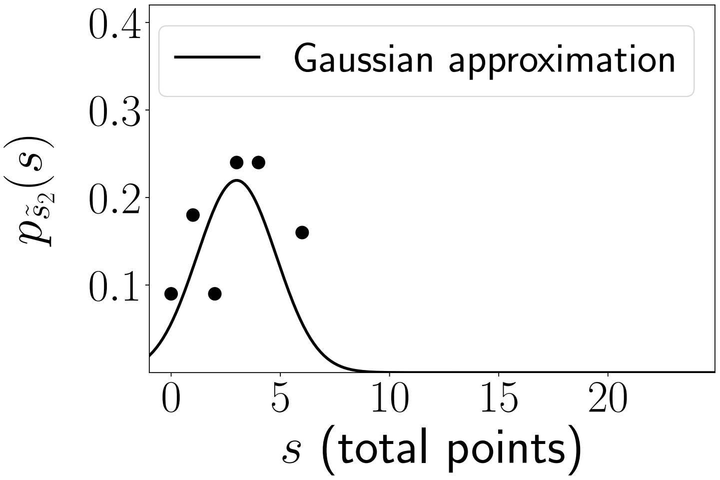


Gaussian approximation

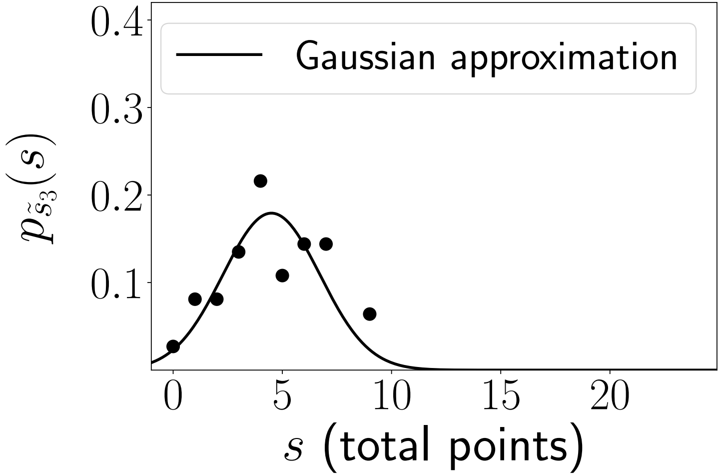
$$\mathrm{E}\left[\tilde{s}_n\right] = \sum_{i=1}^{n} \mathrm{E}\left[\tilde{x}_i\right] = 1.5n$$

$$\mathrm{Var}\left[\tilde{s}_n\right] = \sum_{i=1}^{n} \mathrm{Var}\left[\tilde{x}_i\right] = 1.65n$$

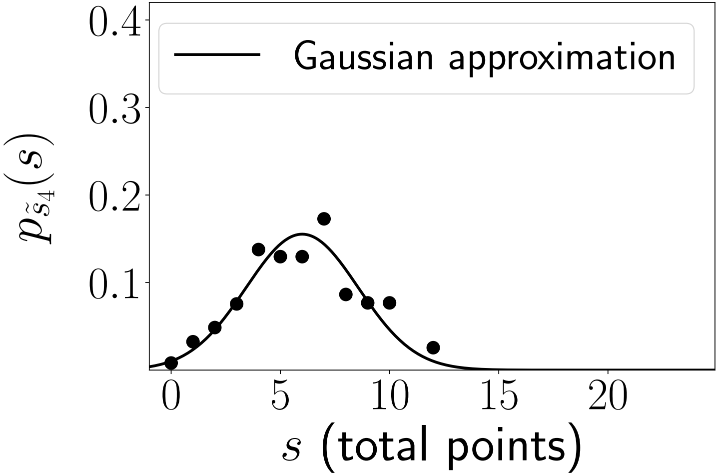
Soccer league: 2 games



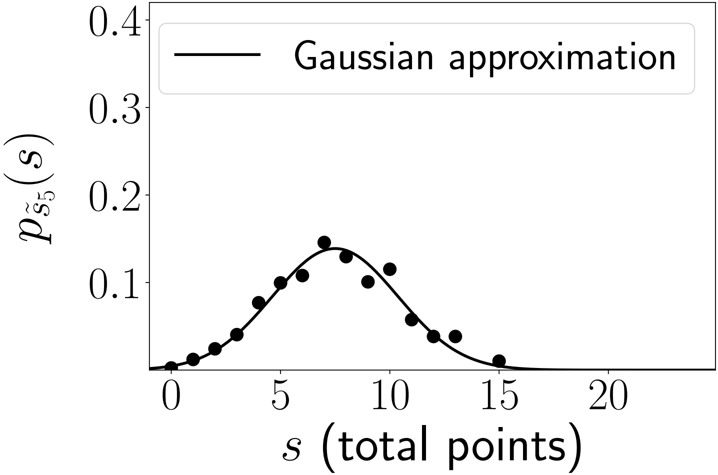
Soccer league: 3 games



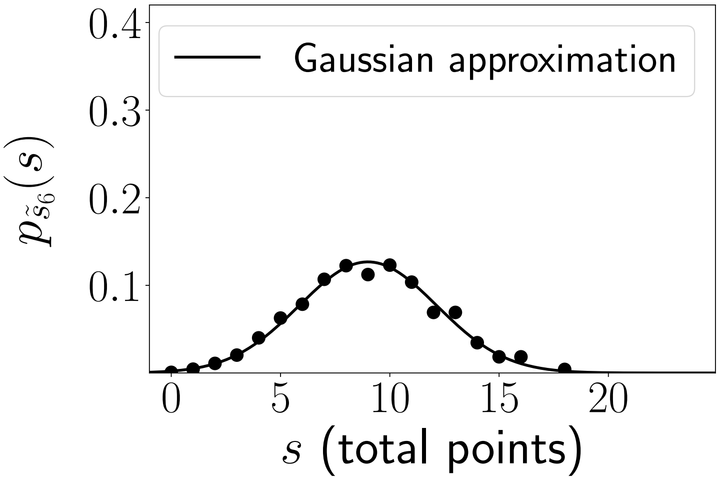
Soccer league: 4 games



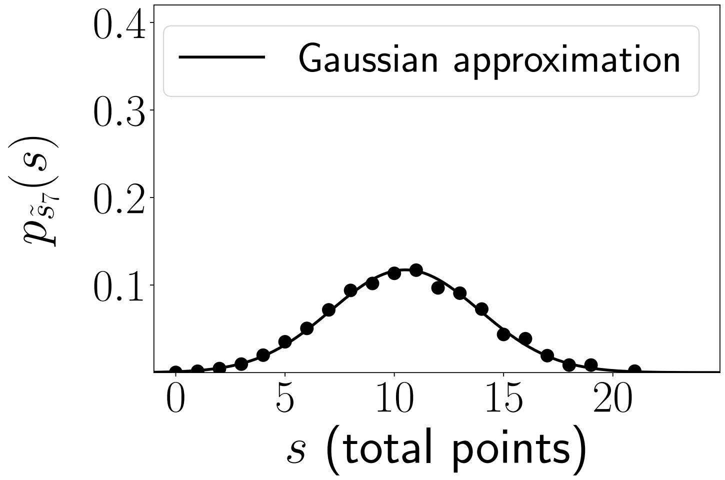
Soccer league: 5 games



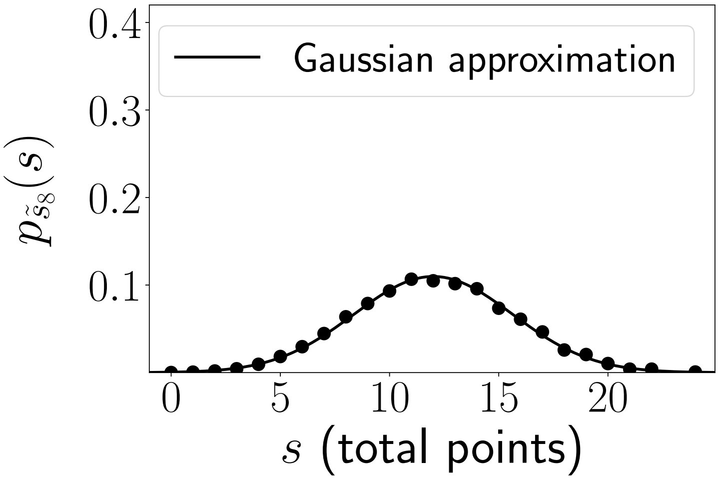
Soccer league: 6 games



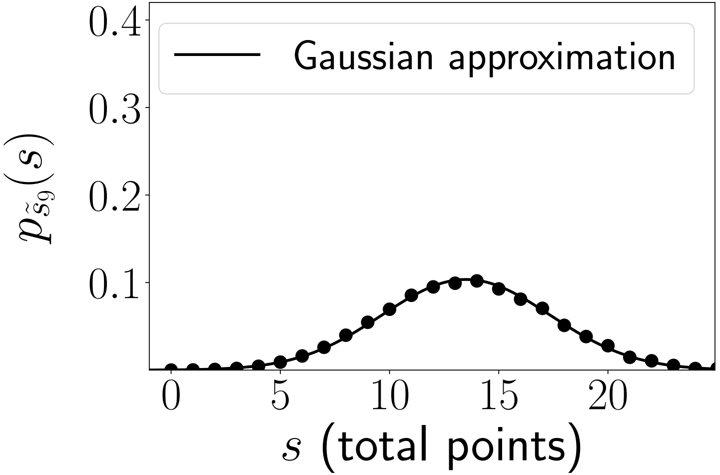
Soccer league: 7 games



Soccer league: 8 games



Soccer league: 9 games



Sum of independent continuous random variables

Independent continuous random variables $\tilde{a}$ and $\tilde{b}$

The pdf of $\tilde{s} = \tilde{a} + \tilde{b}$ is

$$\begin{aligned} f_{\tilde{s}}(s) &= \int_{a=-\infty}^{\infty} f_{\tilde{a}}(a)\, f_{\tilde{b}}(s-a)\, \mathrm{d}a \\ &= f_{\tilde{a}} * f_{\tilde{b}}(s) \end{aligned}$$

Independent continuous random variables $\tilde{a}_1$, $\tilde{a}_2$, ..., $\tilde{a}_n$

The pdf of $\tilde{s}_n = \sum_{i=1}^{n} \tilde{a}_i$ is

$$f_{\tilde{s}_n}(s) = f_{\tilde{a}_1} * f_{\tilde{a}_2} * \cdots * f_{\tilde{a}_n}(s)$$

Sample mean

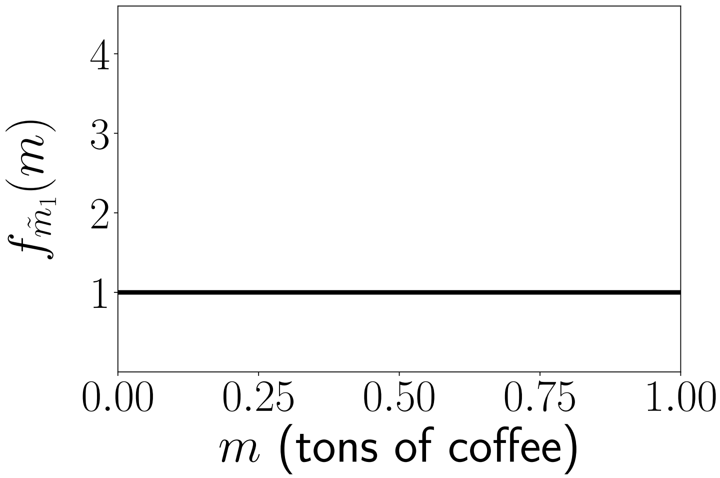
Independent continuous random variables $\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_n$

$\tilde{m}_n := \frac{1}{n}\tilde{s}_n = \frac{1}{n}\sum_{i=1}^{n}\tilde{a}_i$

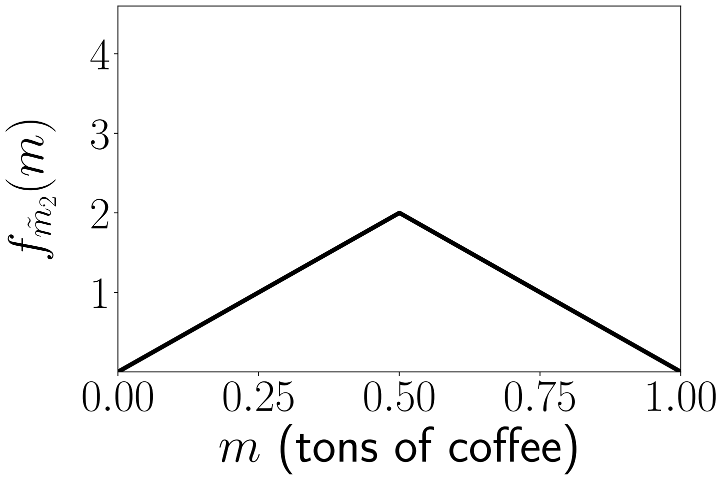
$$f_{\tilde{s}_n}(s) = f_{\tilde{a}_1} * f_{\tilde{a}_2} * \cdots * f_{\tilde{a}_n}(s)$$

$$\begin{aligned} f_{\widetilde{m}_n}(m) &= n f_{\tilde{s}_n}(nm) \\ &= n\left(f_{\tilde{a}_1} * f_{\tilde{a}_2} * \cdots * f_{\tilde{a}_n}\right)(nm) \end{aligned}$$

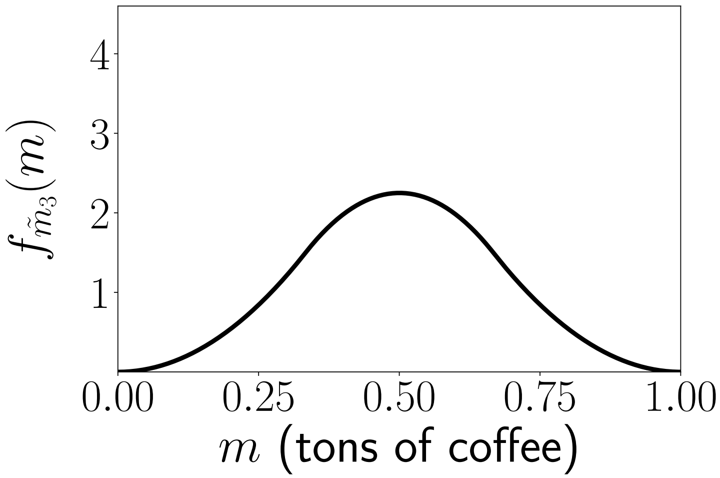
Purchased coffee: 1 supplier



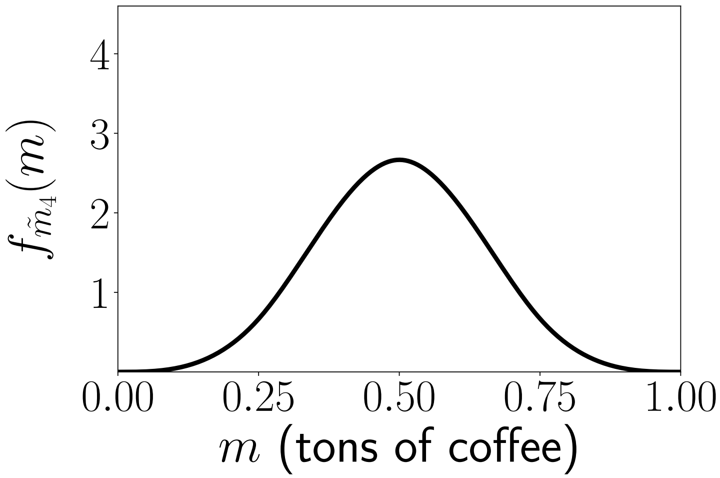
Purchased coffee: 2 suppliers



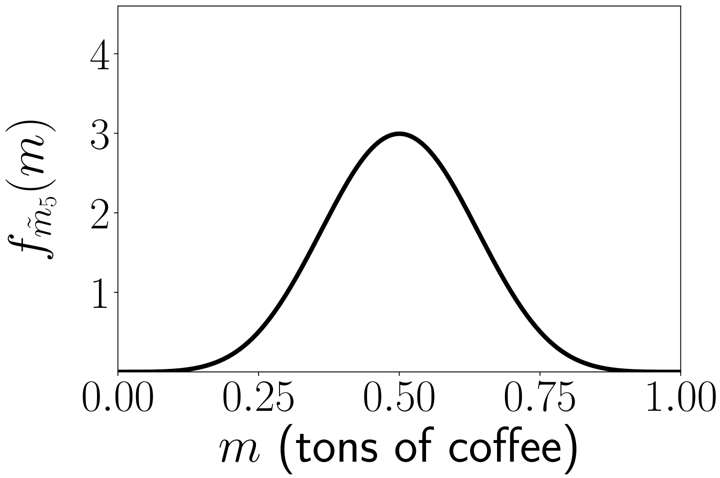
Purchased coffee: 3 suppliers



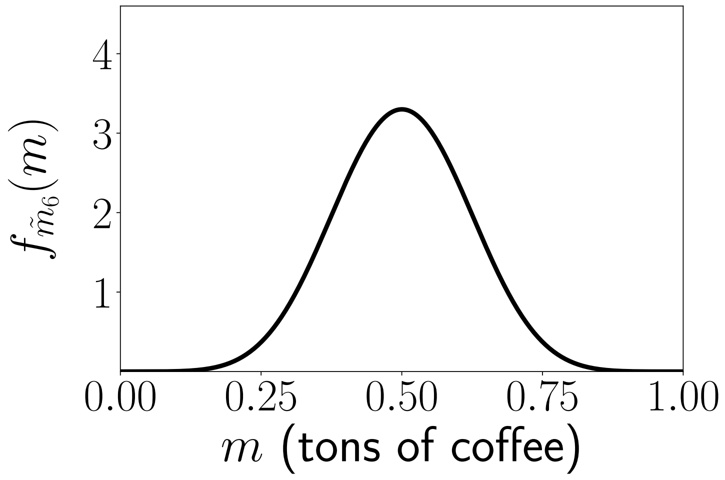
Purchased coffee: 4 suppliers



Purchased coffee: 5 suppliers



Purchased coffee: 6 suppliers

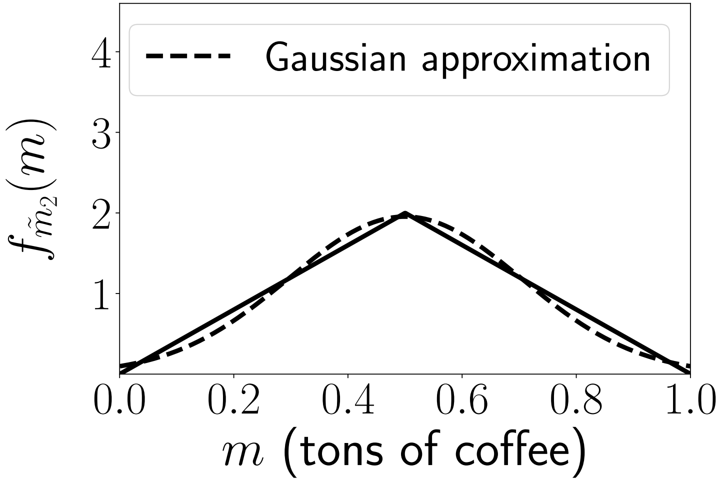


Gaussian approximation

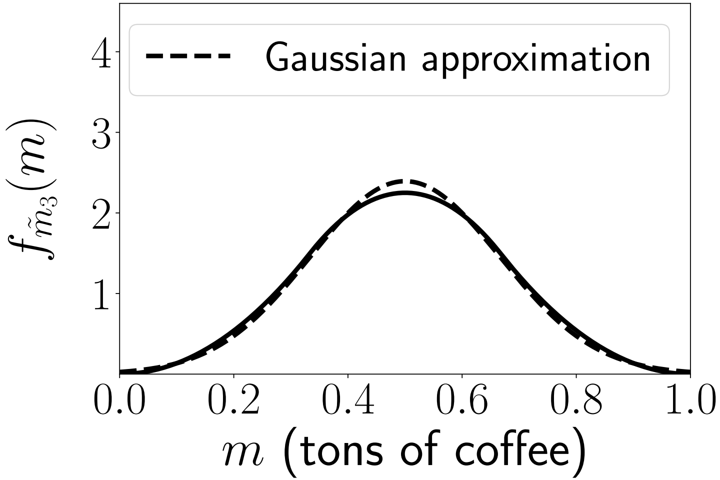
$$\mathrm{E}\left[\widetilde{m}_n\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\tilde{c}_i\right] = 0.5$$

$$\mathrm{Var}\left[\widetilde{m}_n\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[\tilde{c}_i\right] = \frac{1}{12n}$$

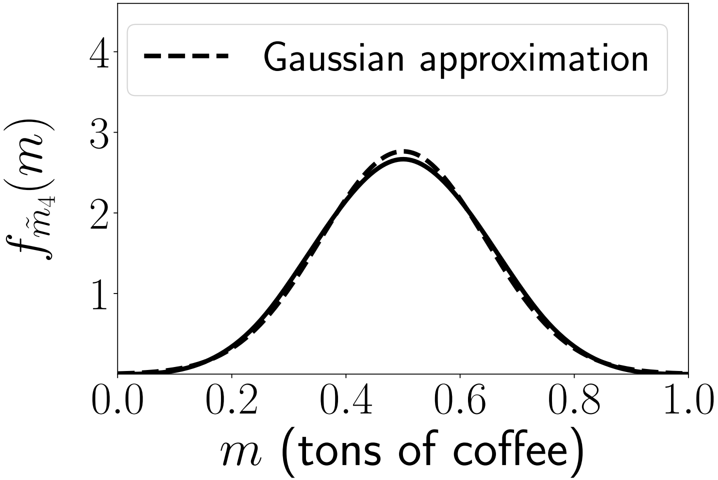
Purchased coffee: 2 suppliers



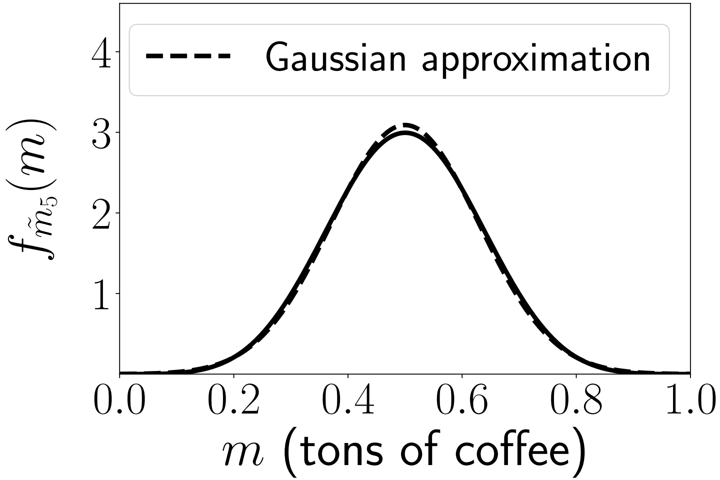
Purchased coffee: 3 suppliers



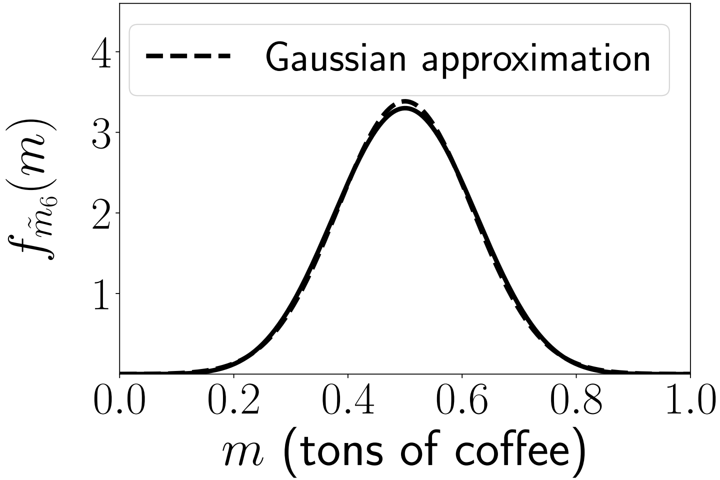
Purchased coffee: 4 suppliers



Purchased coffee: 5 suppliers



Purchased coffee: 6 suppliers



Central limit theorem

If $\tilde{x}_1, \tilde{x}_2, \ldots$ are independent random variables with mean μ and variance σ^2

$$\tilde{m}_n := \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i$$

$$\mathrm{E}\left[\tilde{m}_n\right] = \mu$$

$$\mathrm{Var}\left[\tilde{m}_n\right] = \frac{\sigma^2}{n}$$

As $n \to \infty$ $\tilde{m}_n$ converges in distribution to a Gaussian with mean μ and variance $\frac{\sigma^2}{n}$

Reminder

If $\tilde{a}$ is a Gaussian random variable with mean μ and variance σ^2

$$\tilde{b} := \alpha \tilde{a} + \beta$$

is Gaussian with mean $\alpha\mu + \beta$ and variance $\alpha^2\sigma^2$

More formally

The cdf $F_{s(\widetilde{m}_n)}$ of the standardized sample mean

$$s(\widetilde{m}_n) := \frac{\widetilde{m}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

converges to the cdf of a standard Gaussian with mean zero and unit variance as $n \to \infty$

Binomial distribution

The pmf of a binomial random variable $\tilde{a}$ with parameters n and θ is

$$p_{\tilde{a}}(a) = \binom{n}{a} \theta^a (1-\theta)^{(n-a)} \qquad a = 0, 1, \ldots, n$$

Can be represented as sum of n independent random variables

$$\tilde{a} = \sum_{i=1}^{n} \tilde{b}_i$$

Approximation for $\tilde{a}/n$:

Gaussian with mean θ and variance $\theta(1-\theta)/n$

Approximation for $\tilde{a}$:

Gaussian with mean $n\theta$ and variance $n\theta(1-\theta)$

Basketball strategy

Goal: Compare two strategies

Strategy 2p: only taking 2-point shots

Strategy 3p: only taking 3-point shots

100 shots modeled as i.i.d. Bernoulli random variables with parameter $\theta_2 := 0.5$ and $\theta_3 := 0.35$

Basketball strategy

Shots made: $\tilde{x}_{2p}$ and $\tilde{x}_{3p}$

Binomial with parameters $n := 100$ and $\theta_2 := 0.5$ / $\theta_3 := 0.35$

Score of Strategy 2p: $\tilde{y}_{2p} := 2\tilde{x}_{2p}$

Score of Strategy 3p: $\tilde{y}_{3p} := 3\tilde{x}_{3p}$

Score difference: $\tilde{d} := \tilde{y}_{3p} - \tilde{y}_{2p}$

Gaussian approximation

$\tilde{x}_{2p}$: mean $100\,\theta_2$ and variance $100\,\theta_2(1-\theta_2)$

$\tilde{y}_{2p}$: mean $200\,\theta_2 = 100$ and variance $400\,\theta_2(1-\theta_2) = 100$

$\tilde{x}_{3p}$: mean $100\,\theta_3$ and variance $100\,\theta_3(1-\theta_3)$

$\tilde{y}_{3p}$: mean $300\,\theta_3 = 105$ and variance $900\,\theta_3(1-\theta_3) = 204.75$

$\tilde{d}$?

Independent standard Gaussians $\tilde{a}$ and $\tilde{b}$

If $\tilde{a}_1$ and $\tilde{a}_2$ are Gaussian with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2

$\tilde{s} = \tilde{a}_1 + \tilde{a}_2$ is Gaussian with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$

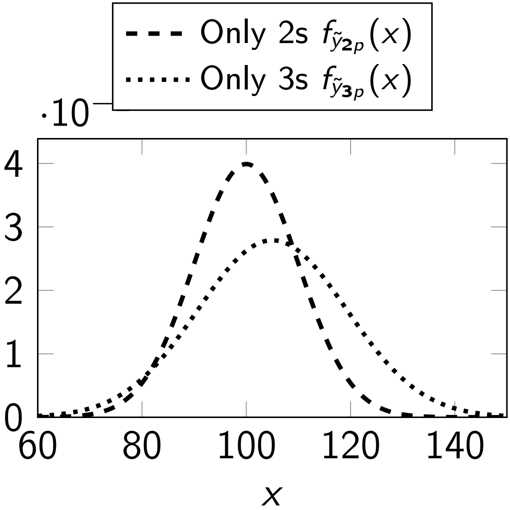
Gaussian approximation

$\tilde{y}_{2p}$: mean 100 and variance 100

$\tilde{y}_{3p}$: mean 105 and variance 204.75

$\tilde{d}$: mean 5 and variance 304.75

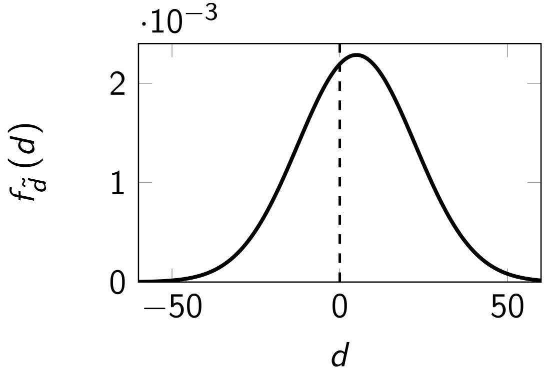
Strategy 2p vs Strategy 3p



Score difference

$\mathrm{P}(\text{Strategy 3p wins}) \approx 61\%$

Monte Carlo simulation: 60%



Distribution of the sample mean

Population mean: μ_{pop} Population variance: σ^2_{pop}

Random samples selected independently and uniformly at random with replacement: $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$

$$\tilde{m}_n := \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i$$

$$\mathrm{E}[\tilde{m}_n] = \mu_{\text{pop}}$$

$$\text{se}[\widetilde{m}_n] = \frac{\sigma_{\text{pop}}}{\sqrt{n}}$$

As $n \to \infty$ $\tilde{m}_n$ converges in distribution to a Gaussian with mean μ_{pop} and standard deviation $\text{se}[\widetilde{m}_n]$

More formally

The cdf $F_{s(\widetilde{m}_n)}$ of the standardized sample mean

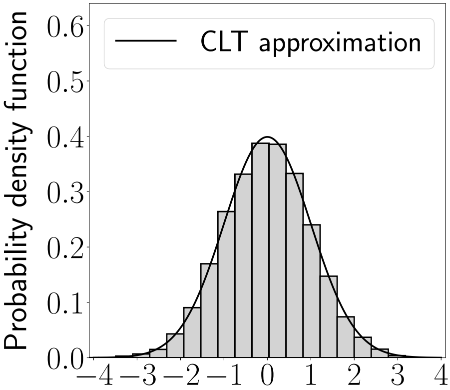
$$s(\widetilde{m}_n) := \frac{\widetilde{m}_n - \mu_{\text{pop}}}{\text{se}\left[\widetilde{m}_n\right]}$$

converges to the cdf of a standard Gaussian with mean zero and unit variance as $n \to \infty$

Height data: $n = 20$

$\mu_{\text{pop}} := 175.6$ cm, $\sigma_{\text{pop}} = 6.85$ cm

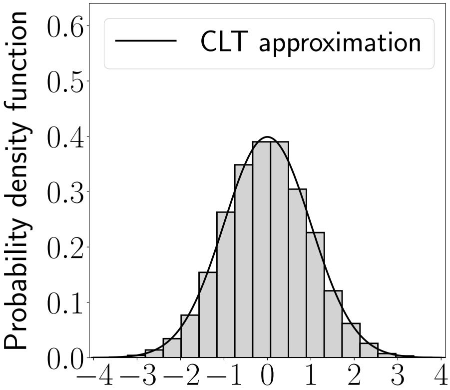
Total population $N := 4{,}082$



Height data: $n = 100$

$\mu_{\text{pop}} := 175.6$ cm, $\sigma_{\text{pop}} = 6.85$ cm

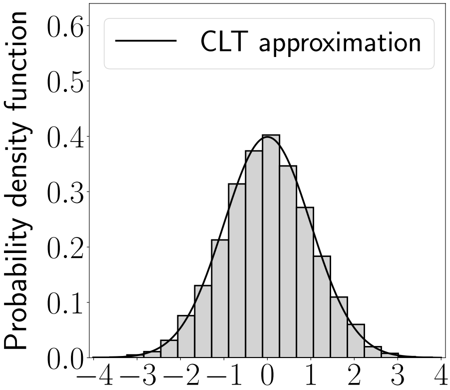
Total population $N := 4{,}082$



Height data: n =1,000

$\mu_{\text{pop}} := 175.6$ cm, $\sigma_{\text{pop}} = 6.85$ cm

Total population $N := 4{,}082$



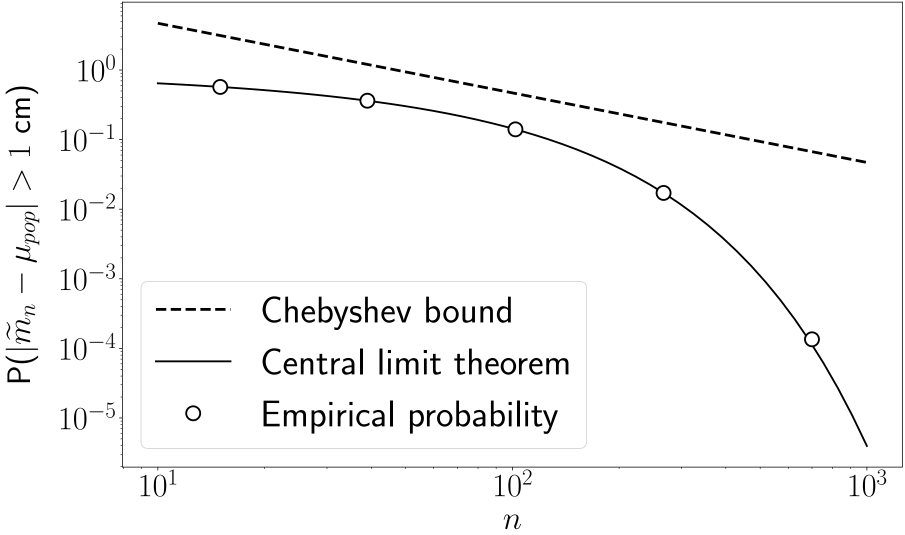
Chebyshev bound

$$\mathrm{P}\left(|\tilde{m}_n - \mu_{\text{pop}}| > \epsilon\right) \leq \frac{\sigma_{\text{pop}}^2}{n\epsilon^2}$$

Terrible approximation...

Do we get a better approximation from the central limit theorem?

Much better



10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
$\mathrm{P}(|\widetilde{m}_n - \mu_{pop}| > 1$ cm$)$
Chebyshev bound
Central limit theorem
Empirical probability
10^1
10^2
10^3
n

What have we learned

Sample mean of independent random variables with finite mean and variance converges in distribution to a Gaussian

Gaussian approximation is often very accurate for finite data in practice