Sums and Averages of Independent Random Variables

Probability and Statistics for Data Science

Carlos Fernandez-Granda

NYU COURANT INSTITUTE OF MATHEMATICAL SCIENCES

NYU DATA SCIENCE

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

Plan

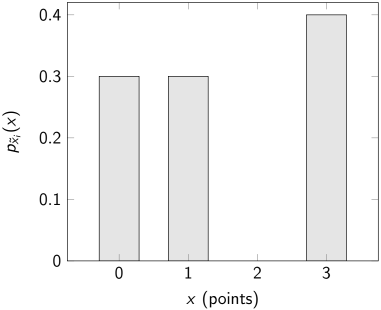
Study the distribution of sums and averages of independent quantities

Soccer league

Points $\tilde{x}_i$ in game i

Games are independent

Points won over n games? $\tilde{s}_n := \sum_{i=1}^{n} \tilde{x}_i$



$$p_{\tilde{x}_i}(0) = 0.3 \qquad p_{\tilde{x}_i}(1) = 0.3 \qquad p_{\tilde{x}_i}(3) = 0.4$$

Two games

$$p_{\tilde{x}_i}(0) = 0.3 \qquad p_{\tilde{x}_i}(1) = 0.3 \qquad p_{\tilde{x}_i}(3) = 0.4$$

$$\begin{aligned}
p_{\tilde{s}_2}(0) &= \mathrm{P}\left(\tilde{x}_1 + \tilde{x}_2 = 0\right) \\
&= \mathrm{P}\left(\tilde{x}_1 = 0, \tilde{x}_2 = 0\right) \\
&= \mathrm{P}\left(\tilde{x}_1 = 0\right)\mathrm{P}\left(\tilde{x}_2 = 0\right) \\
&= 0.09
\end{aligned}$$

$$\begin{aligned}
p_{\tilde{s}_2}(1) &= \mathrm{P}\left(\tilde{x}_1 + \tilde{x}_2 = 1\right) \\
&= \mathrm{P}\left(\tilde{x}_1 = 0, \tilde{x}_2 = 1\right) + \mathrm{P}\left(\tilde{x}_1 = 1, \tilde{x}_2 = 0\right) \\
&= \mathrm{P}\left(\tilde{x}_1 = 0\right)\mathrm{P}\left(\tilde{x}_2 = 1\right) + \mathrm{P}\left(\tilde{x}_1 = 1\right)\mathrm{P}\left(\tilde{x}_2 = 0\right) \\
&= 0.18
\end{aligned}$$

Two independent discrete random variables

Independent discrete random variables $\tilde{a}$ and $\tilde{b}$ (range A and B)

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

$$\begin{aligned} p_{\tilde{s}}(s) &= \mathrm{P}\left(\tilde{a} + \tilde{b} = s\right) \\ &= \sum_{a \in A} \mathrm{P}\left(\tilde{a} = a, \tilde{b} = s - a\right) \\ &= \sum_{a \in A} \mathrm{P}\left(\tilde{a} = a\right) \mathrm{P}\left(\tilde{b} = s - a\right) \\ &= \sum_{a \in A} p_{\tilde{a}}(a)\, p_{\tilde{b}}(s - a) \end{aligned}$$

If A and B are subsets of the integers

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

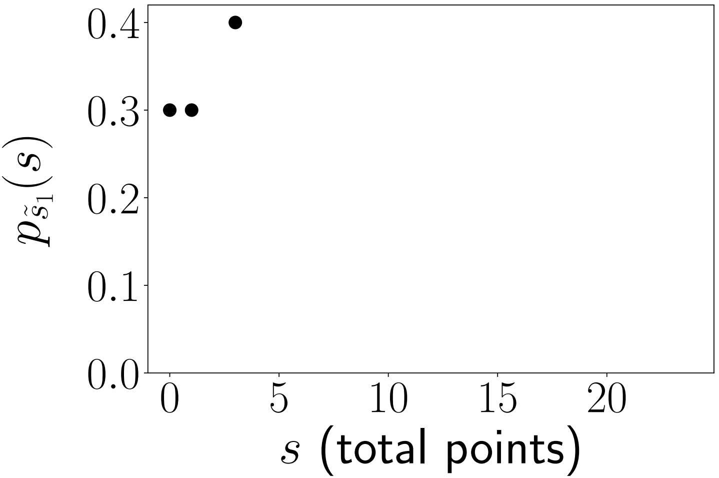
Sum of n independent discrete random variables

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

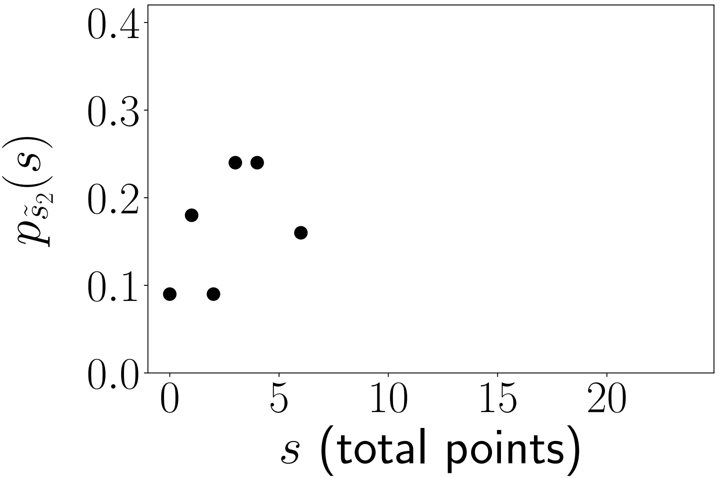
The pmf of $\tilde{s}_n = \sum_{i=1}^{n} \tilde{a}_i$ is

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

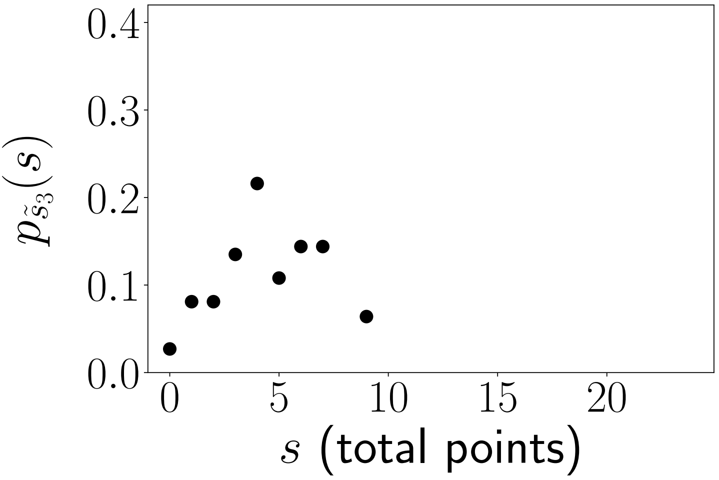
Soccer league: 1 game



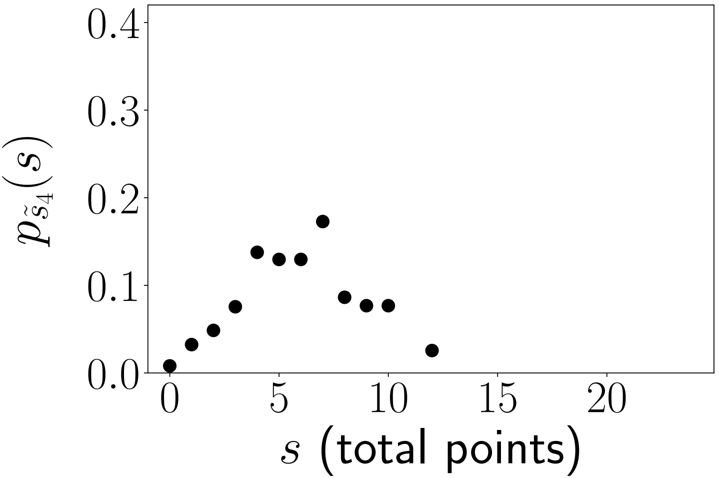
Soccer league: 2 games



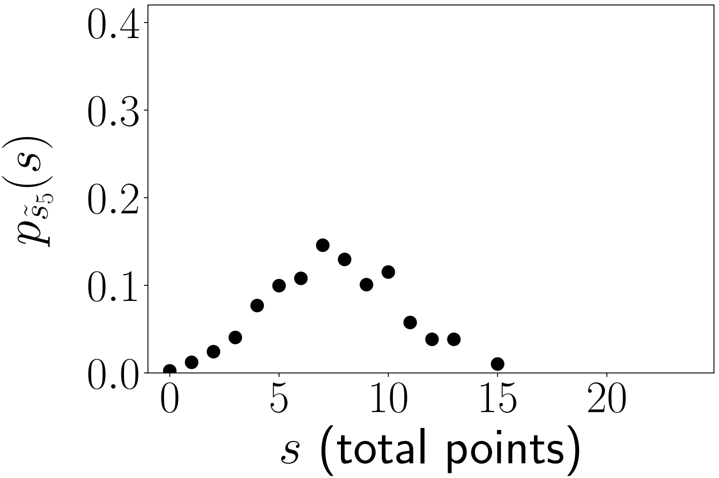
Soccer league: 3 games



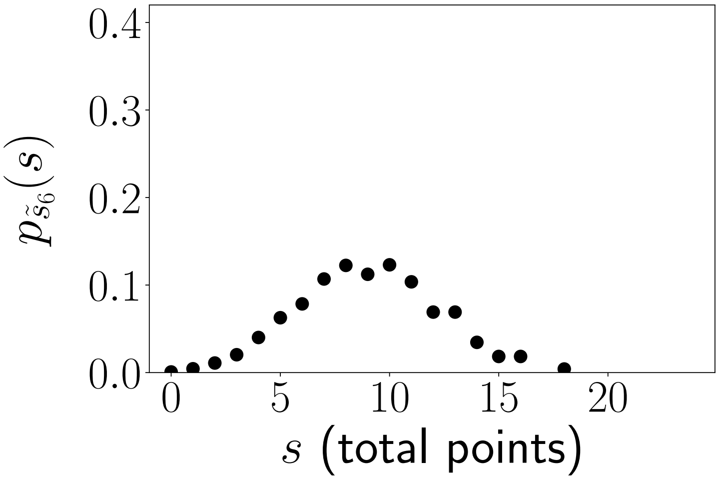
Soccer league: 4 games



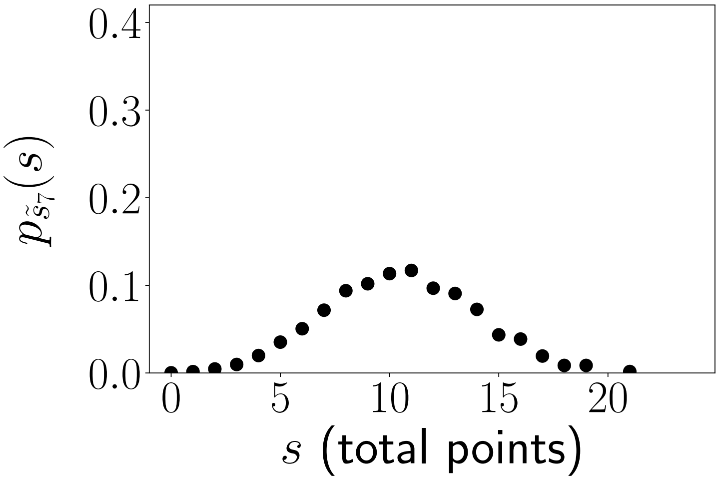
Soccer league: 5 games



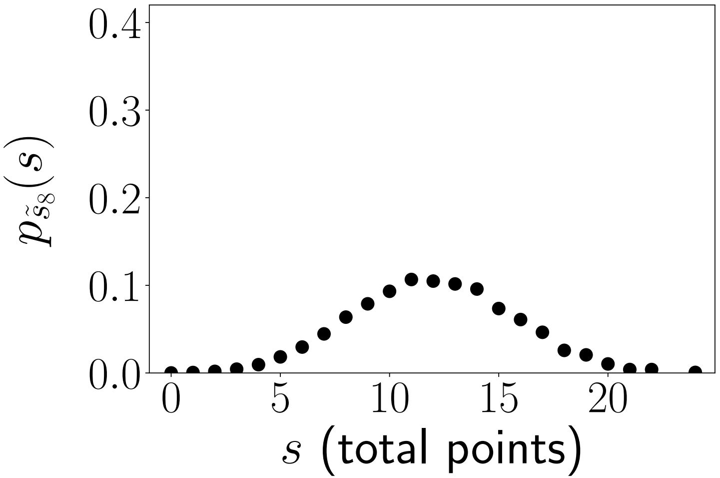
Soccer league: 6 games



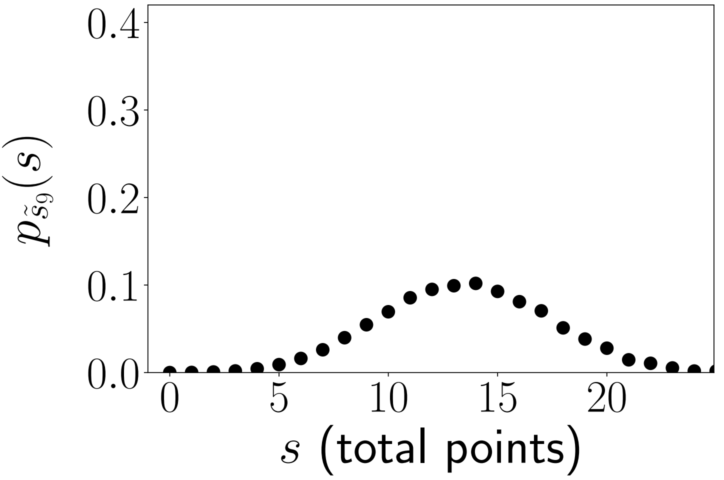
Soccer league: 7 games



Soccer league: 8 games



Soccer league: 9 games

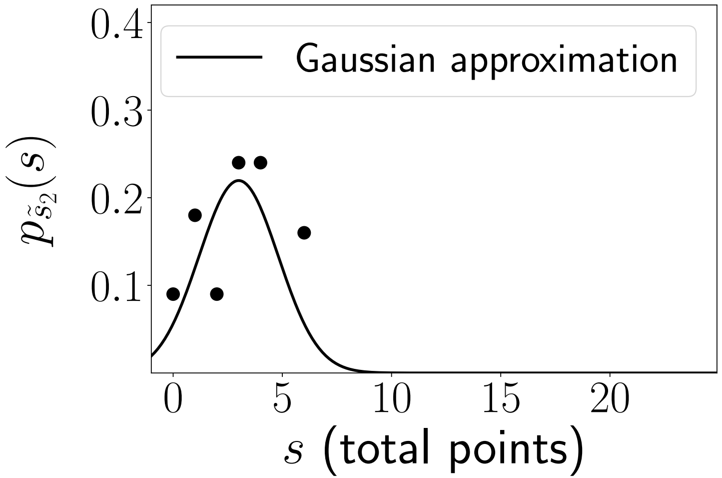


Gaussian approximation

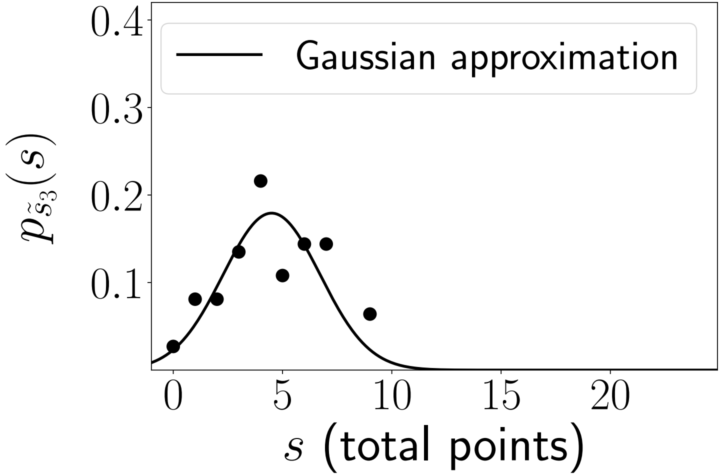
$$\mathrm{E}\left[\tilde{s}_n\right] = \sum_{i=1}^{n} \mathrm{E}\left[\tilde{x}_i\right] = 1.5n$$

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

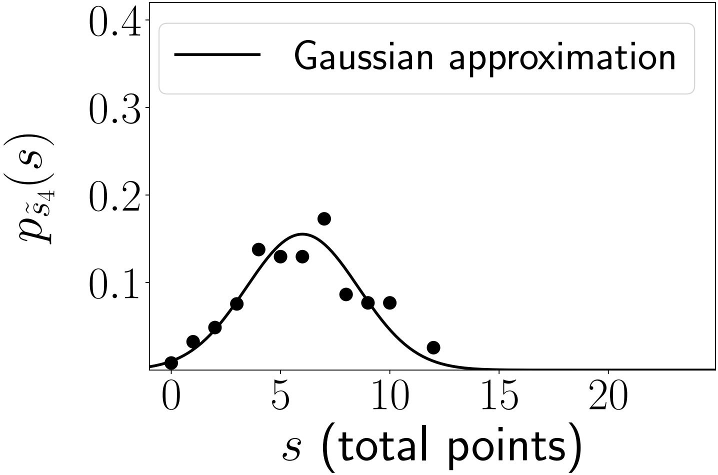
Soccer league: 2 games



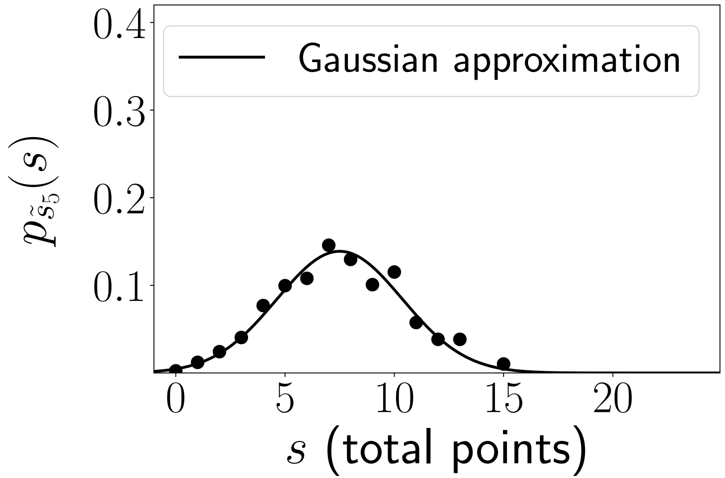
Soccer league: 3 games



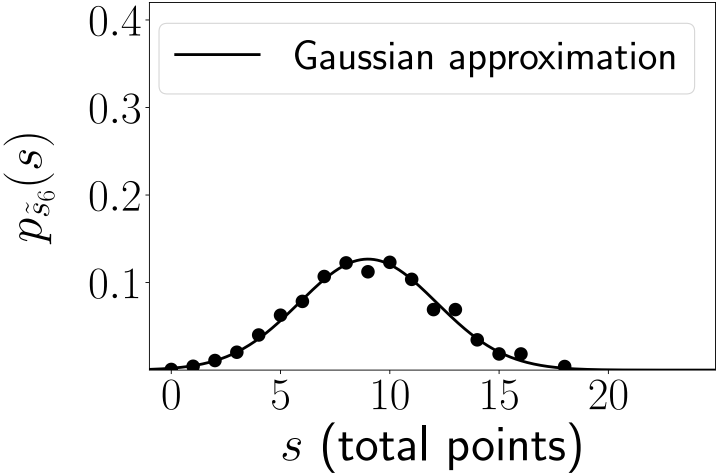
Soccer league: 4 games



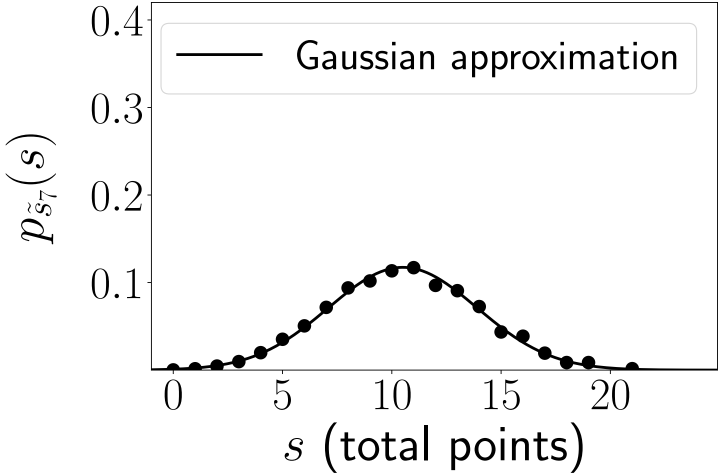
Soccer league: 5 games



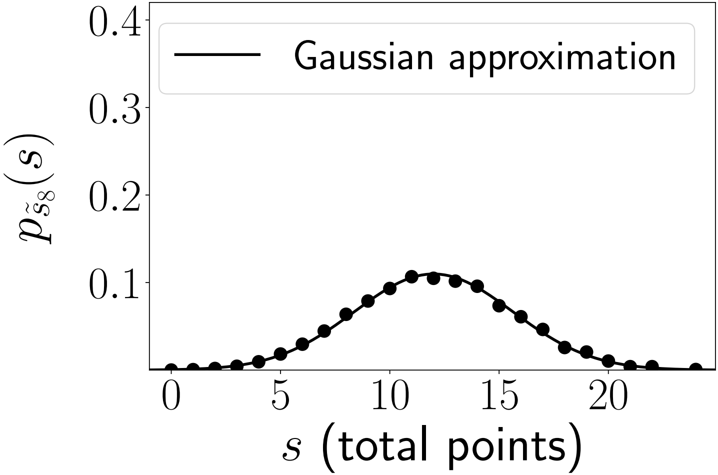
Soccer league: 6 games



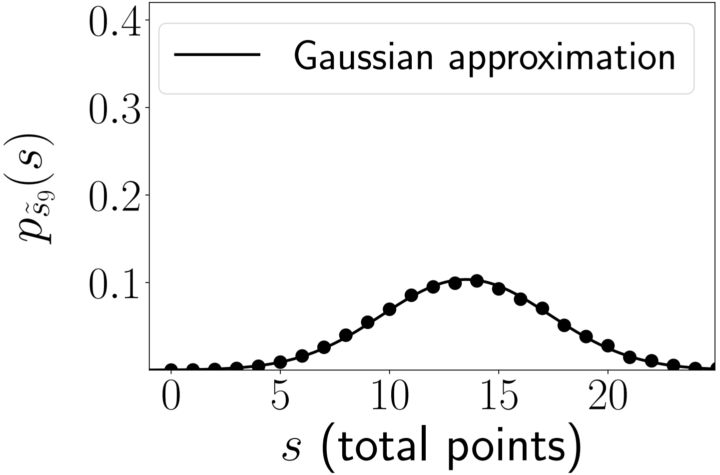
Soccer league: 7 games



Soccer league: 8 games



Soccer league: 9 games



Coffee supply

n independent suppliers

Coffee from ith supplier: uniform between 0 and 1 ton

Total available coffee

$$\tilde{s}_n := \sum_{i=1}^{n} \tilde{c}_i$$

Purchased coffee

$$\widetilde{m}_n := \frac{\tilde{s}_n}{n}$$

Two suppliers

$$
\begin{aligned}
F_{\tilde{s}_2}(s) &= \mathrm{P}\left(\tilde{c}_1 + \tilde{c}_2 \leq s\right) \\
&= \int_{a=-\infty}^{\infty} \int_{b=-\infty}^{s-a} f_{\tilde{c}_1}(a) f_{\tilde{c}_2}(b)\, \mathrm{d}a\, \mathrm{d}b \\
&= \int_{a=-\infty}^{\infty} f_{\tilde{c}_1}(a)\, F_{\tilde{c}_2}(s-a)\, \mathrm{d}a
\end{aligned}
$$

$$
\begin{aligned}
f_{\tilde{s}_2}(s) &= \frac{\mathrm{d}}{\mathrm{d}s} \lim_{t\to\infty} \int_{a=-t}^{t} f_{\tilde{c}_1}(a)\, F_{\tilde{c}_2}(s-a)\, \mathrm{d}a \\
&= \lim_{t\to\infty} \frac{\mathrm{d}}{\mathrm{d}s} \int_{a=-t}^{t} f_{\tilde{c}_1}(a)\, F_{\tilde{c}_2}(s-a)\, \mathrm{d}a \\
&= \lim_{t\to\infty} \int_{a=-t}^{t} \frac{\mathrm{d}}{\mathrm{d}s} f_{\tilde{c}_1}(a)\, F_{\tilde{c}_2}(s-a)\, \mathrm{d}a \\
&= \int_{a=-\infty}^{\infty} f_{\tilde{c}_1}(a)\, f_{\tilde{c}_2}(s-a)\, \mathrm{d}a
\end{aligned}
$$

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 $\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_n$

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

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

Two suppliers: Total supply

$$f_{\tilde{s}_2}(s) = \int_{a=-\infty}^{\infty} f_{\tilde{c}_1}(a)\, f_{\tilde{c}_2}(s-a)\, \mathrm{d}a$$

$$\begin{aligned} f_{\tilde{c}_1}(a) &= 1 \qquad \text{if } 0 \leq a \leq 1 \\ f_{\tilde{c}_2}(s-a) &= 1 \qquad \text{if } 0 \leq s-a \leq 1 \implies s-1 \leq a \leq s \end{aligned}$$

If $0 \leq s \leq 1$

$$f_{\tilde{s}_2}(s) = \int_{a=0}^{s} \mathrm{d}a = s$$

If $1 \leq s \leq 2$

$$f_{\tilde{s}_2}(s) = \int_{a=s-1}^{1} \mathrm{d}a = 2 - s$$

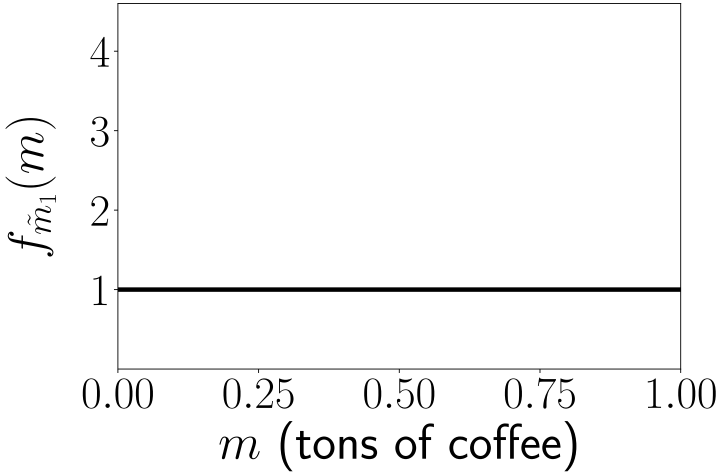
If $s < 0$ or $s > 2$ $\qquad f_{\tilde{s}_2}(s) = 0$

Two suppliers: Purchased coffee

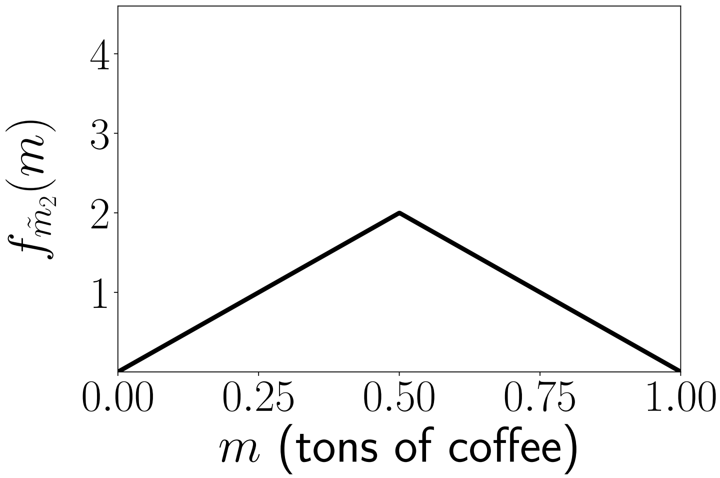
$$
\begin{aligned}
F_{\tilde{m}_2}(m) &= \mathrm{P}\left(\tilde{m}_2 \leq m\right) \\
&= \mathrm{P}\left(\frac{\tilde{s}_2}{2} \leq m\right) \\
&= F_{\tilde{s}_2}\left(2m\right)
\end{aligned}
$$

$$
f_{\tilde{m}_2}(m) = 2 f_{\tilde{s}_2}\left(2m\right) = \begin{cases} 4m & \text{for } 0 \leq s \leq \frac{1}{2} \\ 4(1-m) & \text{for } \frac{1}{2} \leq s \leq 1 \\ 0 & \text{otherwise} \end{cases}
$$

Purchased coffee: 1 supplier



Purchased coffee: 2 suppliers

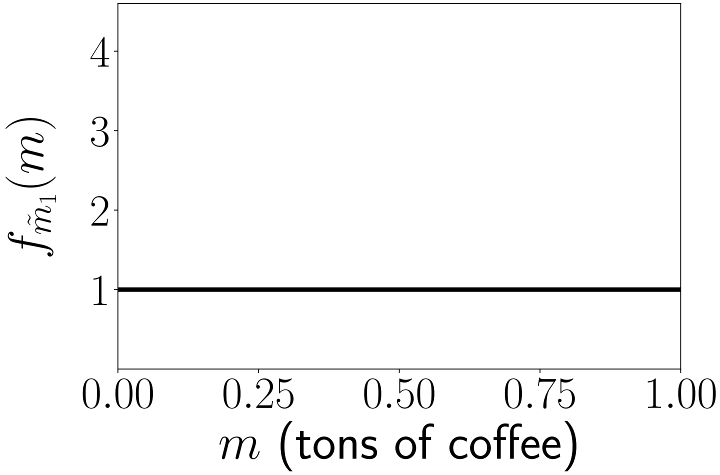


n suppliers

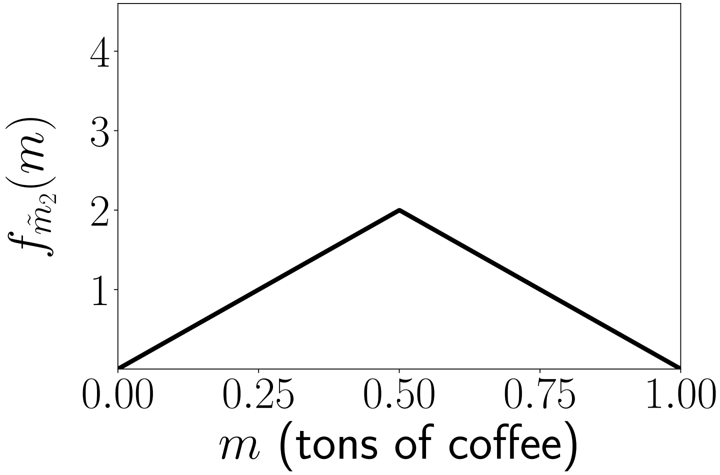
$$f_{\tilde{s}_n}(s) = f_{\tilde{c}_1} * f_{\tilde{c}_2} * \cdots * f_{\tilde{c}_n}(s)$$

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

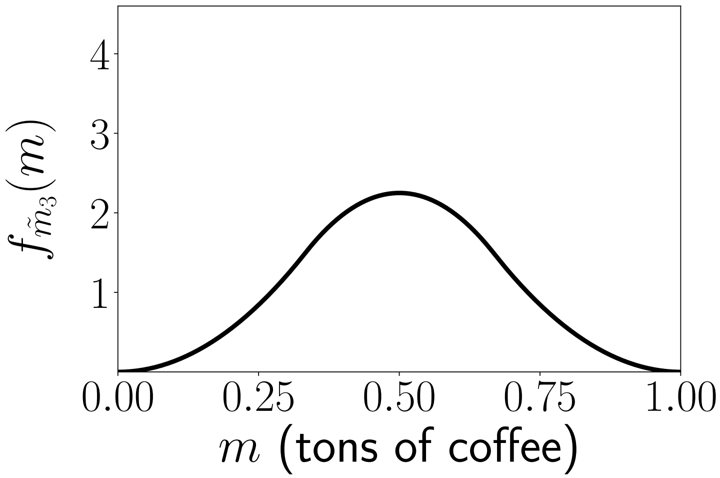
Purchased coffee: 1 supplier



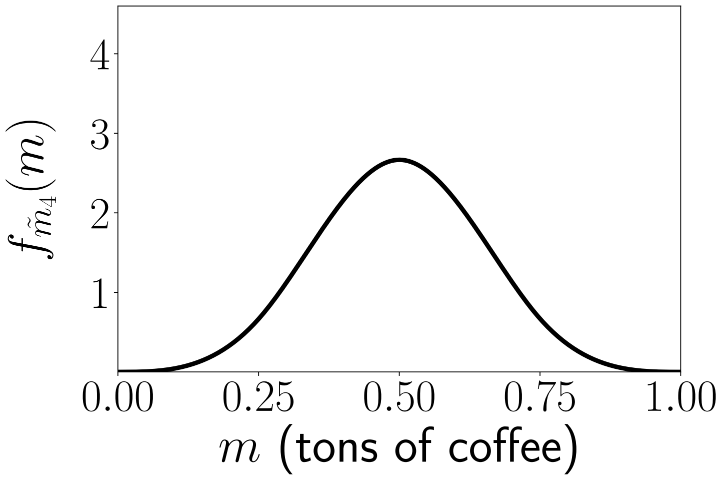
Purchased coffee: 2 suppliers



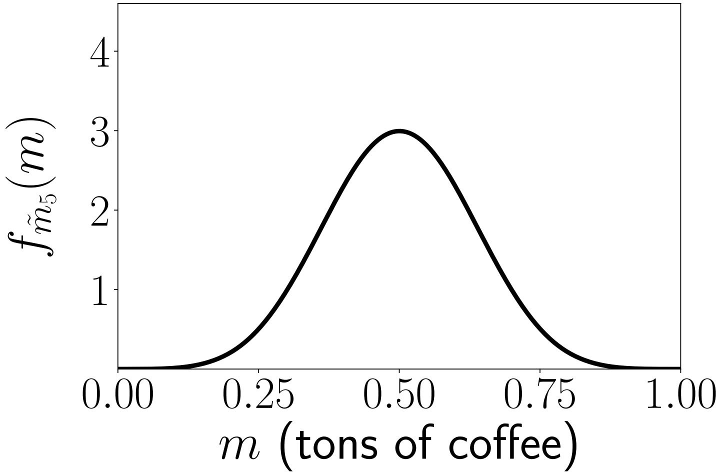
Purchased coffee: 3 suppliers



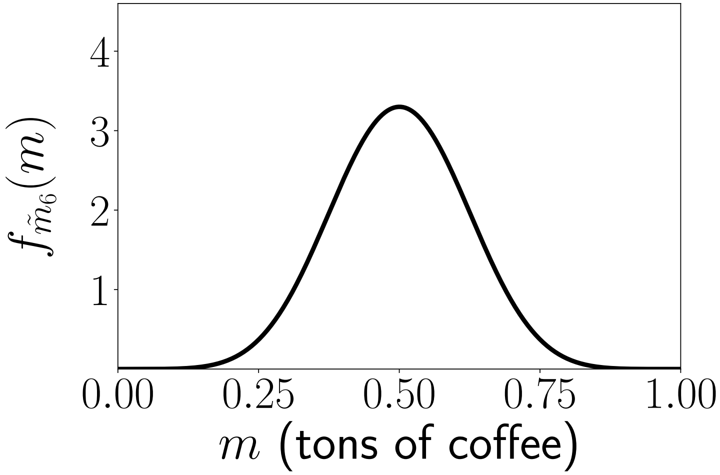
Purchased coffee: 4 suppliers



Purchased coffee: 5 suppliers



Purchased coffee: 6 suppliers

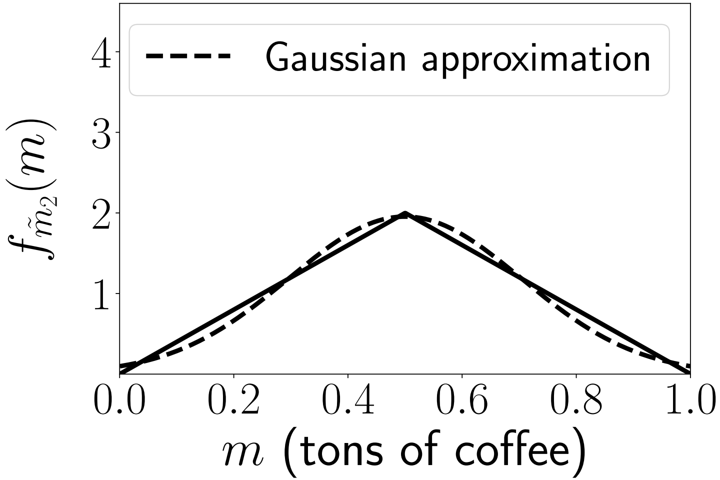


Gaussian approximation

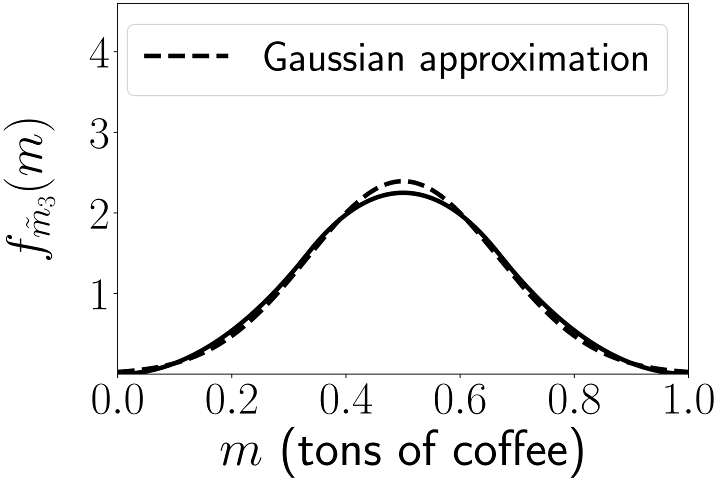
$$\mathrm{E}\left[\widetilde{m}_n\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{c}_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\tilde{c}_i\right] = 0.5$$

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

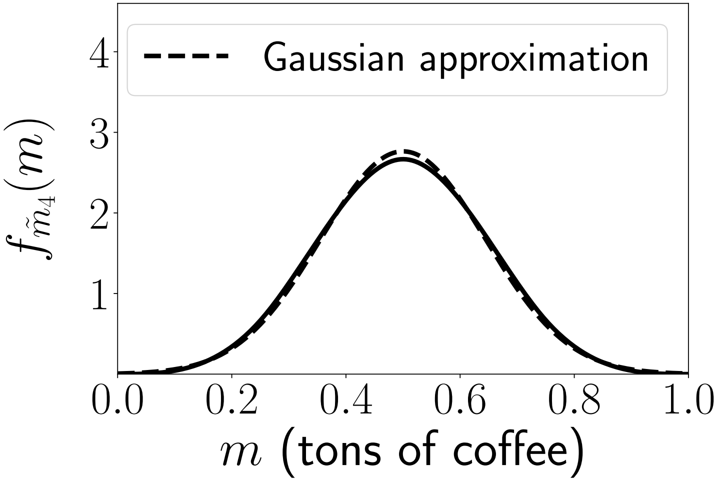
Purchased coffee: 2 suppliers



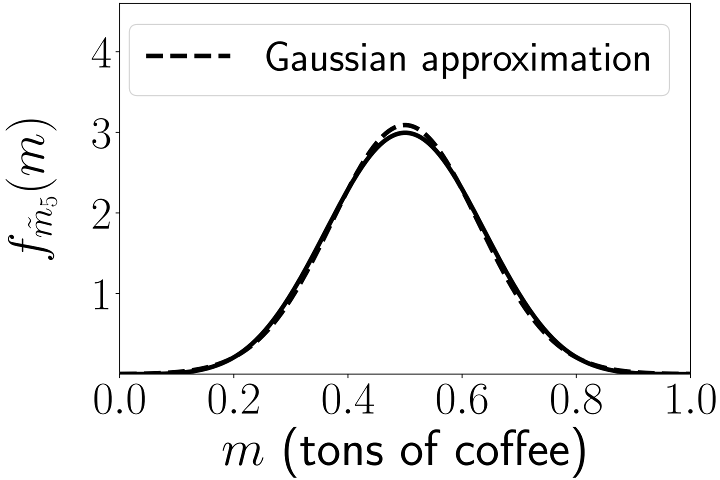
Purchased coffee: 3 suppliers



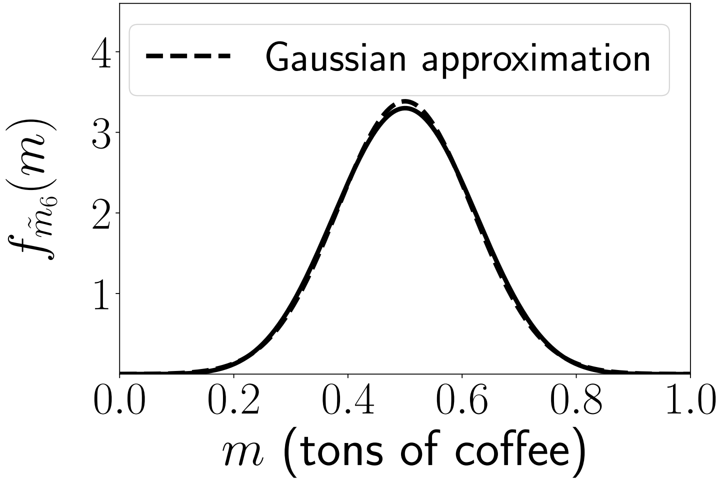
Purchased coffee: 4 suppliers



Purchased coffee: 5 suppliers



Purchased coffee: 6 suppliers



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

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

$$
\begin{aligned}
f_{\tilde{s}}(s) = & \int_{a=-\infty}^{\infty} f_{\tilde{a}}(a)\, f_{\tilde{b}}(s-a)\, \mathrm{d}a \\
& \int_{a=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{a^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(s-a)^2}{2}\right) \mathrm{d}a \\
& = \int_{a=-\infty}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{1}{2}\left(a^2 + (s-a)^2\right)\right) \mathrm{d}a \\
& = \int_{a=-\infty}^{\infty} \frac{1}{2\pi} \exp\left(-a^2 - as + \frac{s^2}{2}\right) \mathrm{d}a \\
& = \int_{a=-\infty}^{\infty} \frac{1}{2\pi} \exp\left(-\left(a - \frac{s}{2}\right)^2 - \frac{s^2}{4}\right) \mathrm{d}a \\
& = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{s^2}{2\sigma^2}\right) \int_{a=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma^{-1}} \exp\left(-\frac{\left(a - \frac{s}{2}\right)^2}{2\sigma^{-2}}\right) \mathrm{d}a \\
& = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{s^2}{2\sigma^2}\right) \qquad \sigma^2 := 2
\end{aligned}
$$

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

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

What have we learned

Distribution of sums and averages of independent random variables

Distribution tends to look Gaussian-like

Sum of independent Gaussians is Gaussian