Week 4: Estimators and their distributions

STA238: Probability, Statistics, and Data Analysis II

Luis Sierra Muntané

2026-01-29

Addendum: Law of total Variance

Proposition. $\mathbf{Var}(Y) = \mathbf{Var}(\mathbb{E}[Y|X]) + \mathbb{E}[\mathbf{Var}(Y|X)]$.

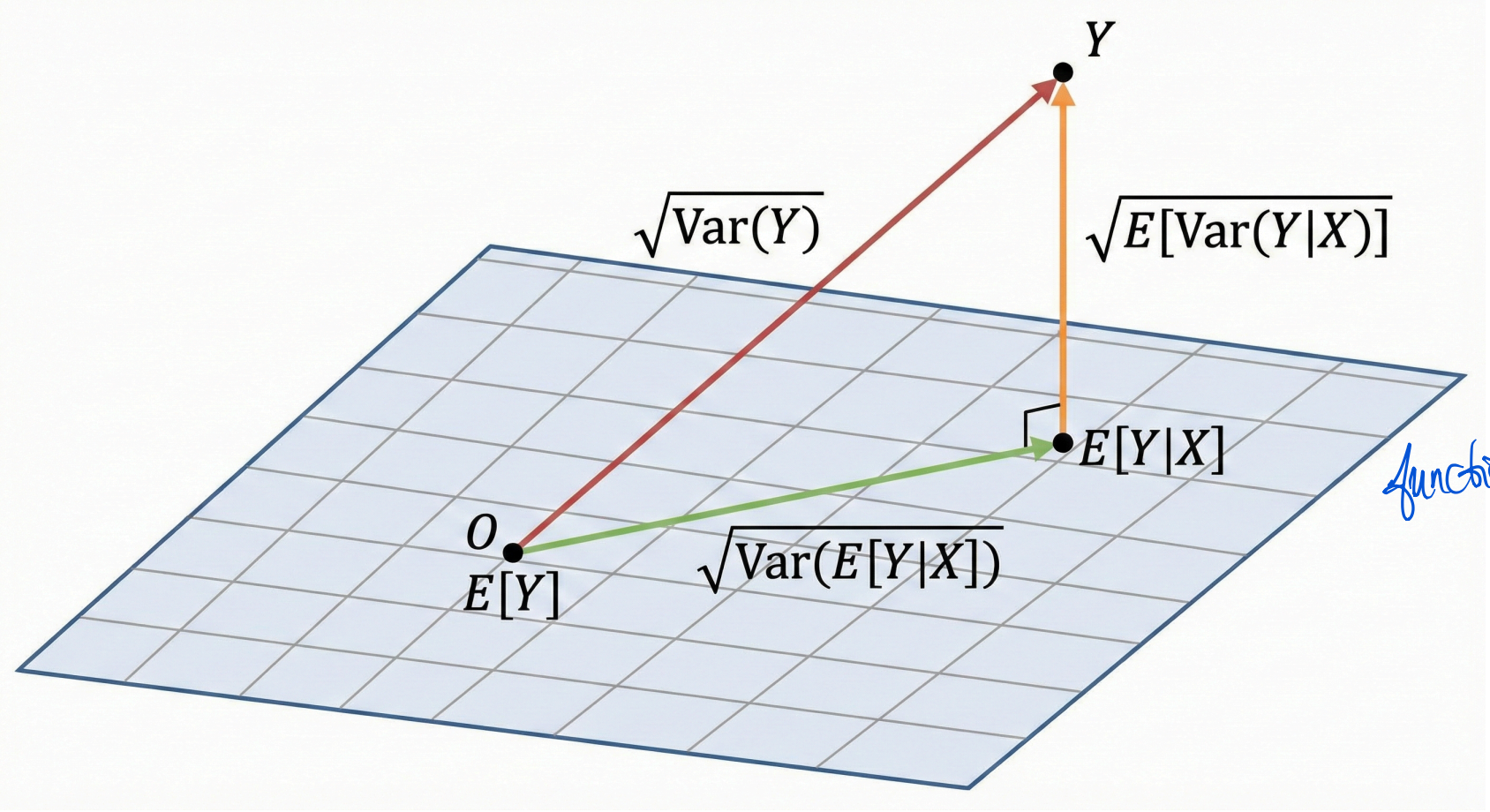
Proof. Separate the deviation $Y - \mathbb{E}[Y]$ as

$$Y - \mathbb{E}[Y] = \underbrace{(\mathbb{E}[Y|X] - \mathbb{E}[Y])}_{\text{Explained by } X} + \underbrace{(Y - \mathbb{E}[Y|X])}_{\text{Residual (Unexplained)}}$$

Since the two components are orthogonal, the squared norm of the sum is the sum of the squared norms:

$$\|Y - \mathbb{E}[Y]\|^2 = \|\mathbb{E}[Y|X] - \mathbb{E}[Y]\|^2 + \|Y - \mathbb{E}[Y|X]\|^2$$

Addendum: Law of total Variance

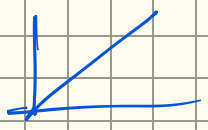


Exercises

Suppose $X_1, \ldots, X_n \sim F_X$ is a random sample. For each of the following estimators $\hat{\theta}$, compute the MSE:

1. F_X is $\mathbf{Unif}(0, \theta)$ and $\hat{\theta} = X_{(n)}$.
2. F_X is $\mathbf{Binom}(10, \theta)$ and $\hat{\theta}$ is $\bar{X}_n/10$.
3. F_X is $\mathcal{N}(\theta, 1)$ and $\bar{X}_n$.

1. Estimator $X_{(n)}$ for θ, $\text{Unif}(0,\theta)$



$$MSE(\hat\theta,\theta) = \text{Bias}(\hat\theta)^2 + \text{Var}(\hat\theta), \quad X_{(n)}: \text{CDF } F_{X_{(n)}}(x) = \left(\frac{x}{\theta}\right)^n, \quad \text{PDF } f_{X_{(n)}}(x) = \frac{nx^{n-1}}{\theta^n}$$

$$\text{Bias: } E[X_{(n)}] = \int_0^\theta x \cdot f_{X_{(n)}}(x)\,dx$$

$$= \int_0^\theta n x^n/\theta^n\,dx = \frac{n}{\theta^n}\int_0^\theta x^n dx = \frac{n}{\theta^n}\left[\frac{x^{n+1}}{n+1}\right]_0^\theta$$

$$= \frac{n}{\theta^n}\cdot\frac{\theta^{n+1}}{n+1} = \frac{n}{n+1}\theta \longrightarrow |\text{Bias}\,\hat\theta| = \left|\frac{n}{n+1}\theta - \theta\right|$$

$$= \frac{1}{n+1}\theta$$

$$\text{Var}\,\hat\theta = E\hat\theta^2 - [E\hat\theta]^2$$

$$E[\hat\theta^2] = \int_0^\theta x^2 f_{X_{(n)}}(x)\,dx = \int_0^\theta x^2 \cdot \frac{nx^{n-1}}{\theta^n}\,dx = \frac{n}{\theta^n}\int_0^\theta x^{n+1}\,dx$$

$$= \frac{n}{\theta^n}\left[\frac{x^{n+2}}{n+2}\right]_0^\theta = \frac{n}{n+2}\theta^2$$

$$\text{Var}\,\hat\theta = \frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2 = \theta^2\left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right)$$

$$= \theta^2\left(\frac{n(n+1)^2 - n^2(n+2)}{(n+2)(n+1)^2}\right)$$

$$= \theta^2\left(\frac{n}{(n+2)(n+1)^2}\right)$$

$$MSE(\hat\theta,\theta) = \theta^2\left(\frac{n}{(n+2)(n+1)^2}\right) + \left(\frac{1}{n+1}\theta\right)^2$$

$$= \theta^2\left(\frac{2}{(n+2)(n+1)^2}\right). \xrightarrow{n\to\infty} 0$$

$\therefore \hat\theta$ is consistent.

2. $Bin(10,\theta)$, $\hat{\theta} = \bar{X}_n/10$ $\quad n\cdot p = 10\theta$

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}\left[\frac{\bar{X}_n}{10}\right] = \frac{1}{10}\sum_{i=1}^{n}\frac{1}{n}\mathbb{E}[X_i] = \frac{1}{10}\cdot\frac{n}{n}\cdot 10\theta = \theta \Longrightarrow \text{unbiased}$$

$$MSE(\hat{\theta},\theta) = Var(\hat{\theta}) \qquad n\cdot p\cdot(1-p) = 10\theta(1-\theta)$$

$$= Var\left(\frac{1}{10}\bar{X}_n\right) = \frac{1}{10^2}\cdot\sum_{i=1}^{n} Var\left(\frac{X_i}{n}\right) = \frac{1}{10^2}\cdot\frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$$

$$= \frac{1}{10^2}\cdot\frac{n}{n^2}\cdot 10\theta(1-\theta)$$

$$= \frac{1}{10n}\theta(1-\theta).$$

3. $\hat{\theta} = \bar{X}_n$, $N(\theta, 1)$,

$$Bias: \quad \mathbb{E}[\hat{\theta}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\cdot n\cdot\theta = \theta.$$

$$Var(\hat{\theta}) = Var(\bar{X}_n) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$$

$$= \frac{n}{n^2}\cdot\sigma^2 = \frac{\sigma^2}{n}.$$

$$MSE(\hat{\theta},\theta) = \cancel{Bias^2(\hat{\theta})}^{\,0} + Var(\hat{\theta}) = \frac{\sigma^2}{n}.$$

Central Limit Theorem

Let $X_1, \ldots, X_n$ be a random sample from a distribution with finite variance $\sigma^2 > 0$. Let $\mu = \mathbb{E}[X]$.

Then, the random variable $Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ satisfies

$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

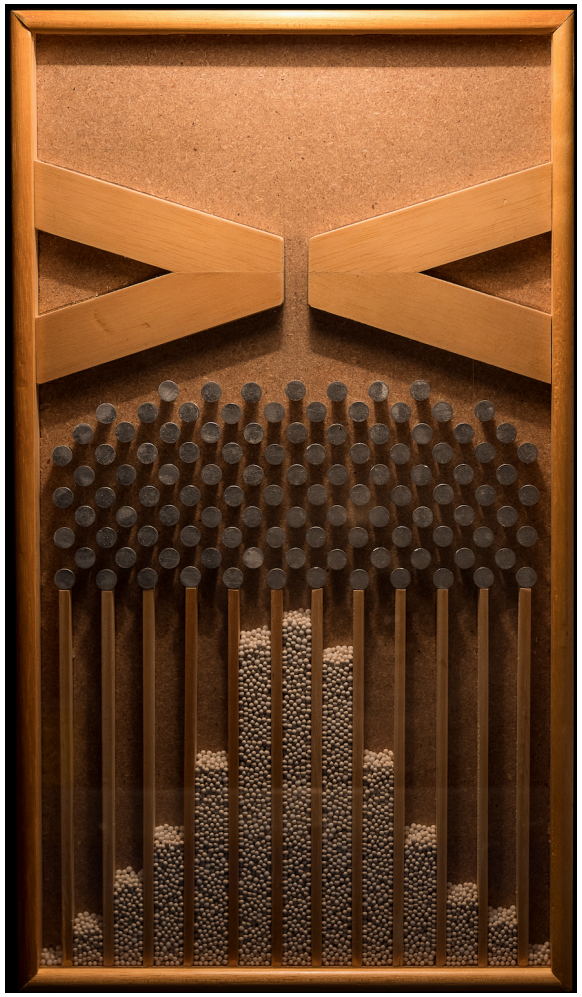
$$Z_n \xrightarrow{d} \mathcal{N}(0, 1).$$

In particular, we can approximate $\mathbb{P}(Z_n \leq z) \approx \Phi(z)$, where $\Phi(z)$ is the CDF of a standard normal distribution.

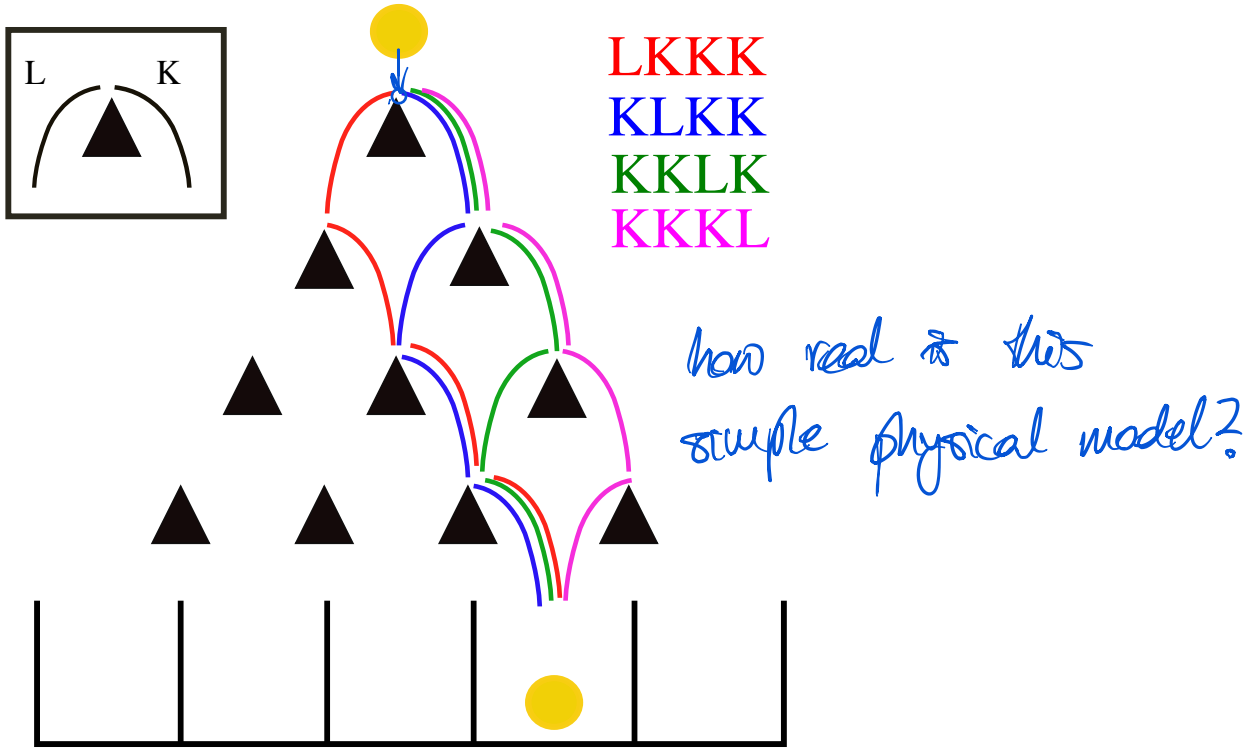
> The **sampling distribution** of the sample mean will *approximately* follow a normal distribution when the sample size is sufficiently large.

> Informally: with enough samples, the sample mean is *approximately* normally distributed.

Central Limit Theorem

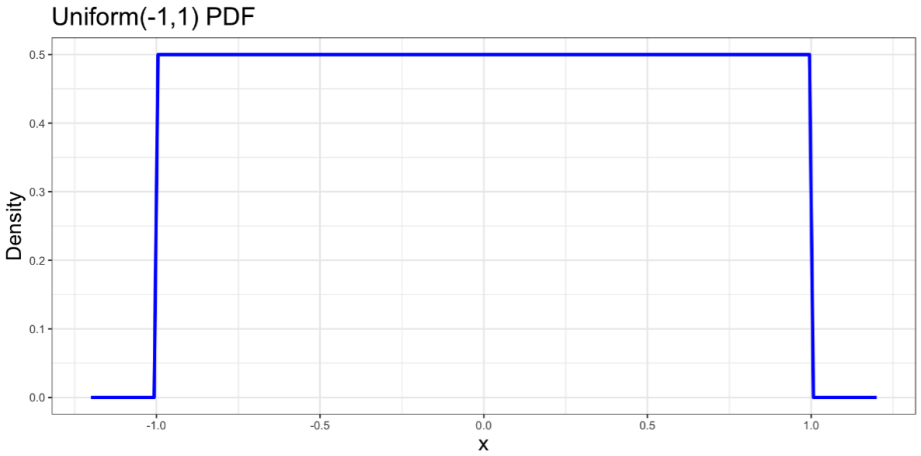


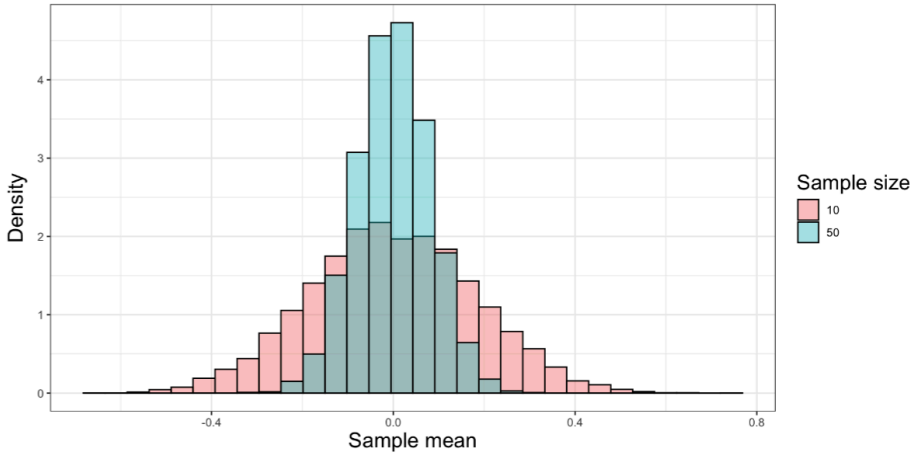
How necessary are the assumptions of the CLT?

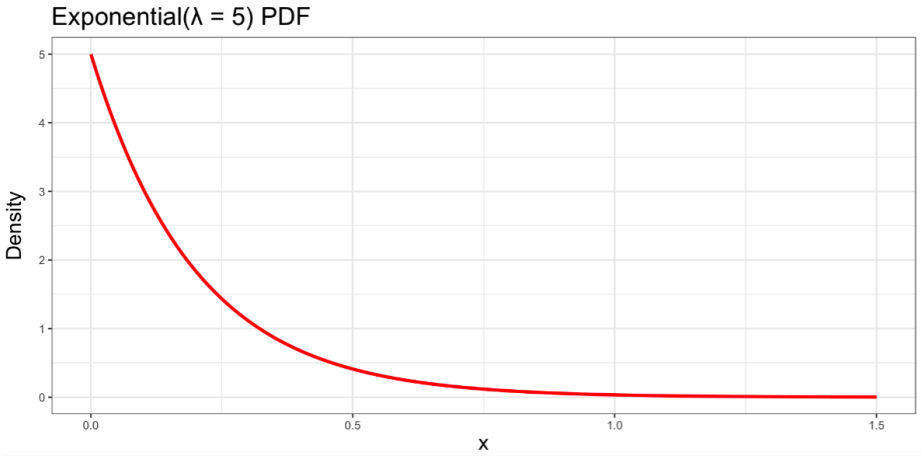


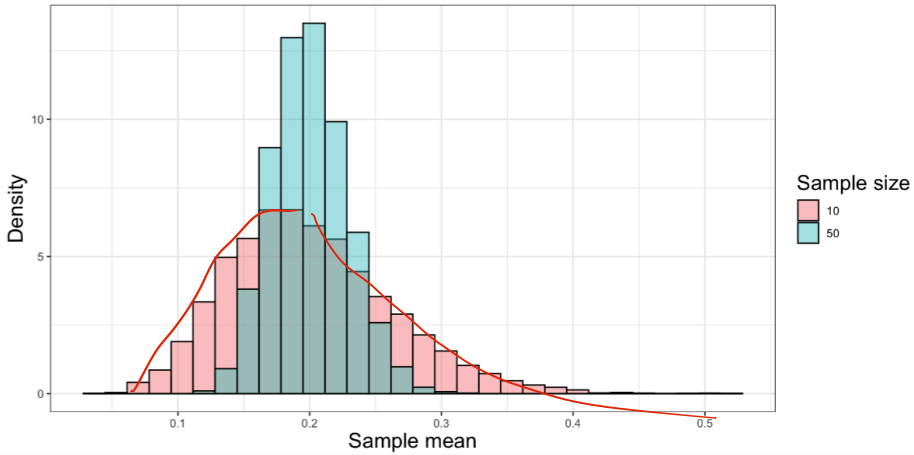
Example: CLT

$\frac{\sigma^2}{n}$









Example: CLT for Binomial approximation

Suppose we have a binomial distributed random variable $X \sim \mathbf{Bin}(n, p)$ and we want to calculate $\mathbb{P}(X \leq a)$:

$$\mathbb{P}(X \leq a) = \sum_{k=0}^{\lfloor a \rfloor} \binom{n}{k} p^k (1-p)^{n-k}$$

expensive to compute

Instead, we can use the much simpler *normal approximation*:

$\mathbb{E}[X] = np$, $\mathrm{Var}(X) = np(1-p)$, X is a sum of n iid Bernoulli r.v.s

$X = \sum_{i=1}^{n} Y_i$, $Y_i \sim \mathrm{Bern}(p) \longrightarrow X \rightsquigarrow N(np, np(1-p))$ CDF $\longrightarrow \Phi\left(\frac{a-np}{\sqrt{np(1-p)}}\right)$

$\frac{X-\mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1) \Longrightarrow X \dot{\sim} N(\mu, \frac{\sigma^2}{n})$, well approximated for n large enough ($\begin{matrix} np \geq 10 \\ n(1-p) \geq 10 \end{matrix}$).

Example: Efficiency of the median

The variance of the sample median of a distribution $f(x)$ is $\frac{1}{4nf(m)^2}$, asymptotically.[1]

Applying this to the normal pdf:

$$f(m) = f(\mu) = \frac{1}{\sqrt{2\pi}\sigma_N} e^{-\frac{1}{2}\left(\frac{\mu-\mu}{\sigma_N}\right)^2} = \frac{1}{\sqrt{2\pi}\sigma_N},$$

so we have that the asymptotic variance of the sample median is

$$\mathbf{Var}(m) \to \frac{\sigma^2}{n} \cdot \frac{\pi}{2}.$$

Example: Efficiency of the median

Using a stronger version of the CLT, one can show that the asymptotic distribution of the sample median is normal. Under our previous assumptions:

$$m \sim \mathcal{N}\left(\mu = m, \sigma^2 = \frac{1}{4nf(m)^2}\right) = \mathcal{N}\left(\mu = m, \sigma^2 = \frac{\sigma_N^2}{n} \cdot \frac{\pi}{2}\right)$$

Recall that $\mathbf{Var}(\bar{X}) = \frac{\sigma^2}{n}$, so the relative efficiency between the sample median and mean is

$$\frac{\mathbf{Var}(m)}{\mathbf{Var}(\bar{X})} \longrightarrow \frac{\sigma^2/n \cdot \frac{\pi}{2}}{\sigma^2/n} = \frac{\pi}{2} > 1$$

> Intuitively, we can say that the median trades efficiency for robustness.

Example: CLT

Suppose we have iid. random variables distributed following $\mathbf{Gamma}(\alpha = 2, \beta = 1)$. Recall that if $X \sim \mathbf{Gamma}(\alpha, \beta)$ then $\mathbf{E}[X] = \alpha\beta$ and $V(X) = \alpha\beta^2$. In an experiment, after 400 realizations of the experiment, $\bar{x}_{400} = 2.0009$ and in 500 realizations, $\bar{x}_{500} = 2.0369$.

- The observed sample means are moving away from 2! Does this contradict the WLLN? No, the WLLN is a probabilistic result, so it can fluctuate.
- What is the asymptotic distribution of $\bar{X}_n$?

CLT: $\bar{X}_n$ approximately $\sim N\left(\alpha\beta, \frac{\alpha\beta^2}{n}\right)$

mean $\quad$ $\frac{1}{n}$ variance

We can estimate error between $\bar{x}_n$ and $\alpha\beta$: $\sqrt{n}\,\frac{\bar{X}_n - \alpha\beta}{\sqrt{\alpha\beta^2}} \dot{\sim} N(0,1)$

A note on sample sizes required for CLT

How large should n be before the CLT provides a "good" approximation? How fast is convergence to a normal distribution?

- There is no universal sample size that works for all distributions.
- The required sample size will depend on the distribution of X
- For symmetric distributions, close values to the normal tend to occur with smaller samples; even $n \approx 5 - 10$ can be sufficient.
- For asymmetric or skewed distributions, convergence is slower.
  - Larger samples are needed to offset the effect of skewness.
  - Common rules of thumb (e.g., $n = 30$ or $n = 50$) are often used, but can still fail in highly skewed settings.
- In short, skewness slows down the convergence towards normality.
- Common rule of thumb for binary variables: n should be large enough so that $np \geq 10$ and $n(1-p) \geq 10$.

29th January ends here, we'll continue with the rest next Tuesday 2nd Feb.

Likelihood function

We commonly use the notation $f(x;\theta)$ to denote the density (or probability mass) at the point x when the underlying population parameters are θ.

If we have a random realization $x_1, \ldots, x_n$ and assume that it is drawn from a distribution with density (or PMF) f, we can think about how likely it would have been to observe that value: $f(x_i;\theta)$, where aggregating for the entire realization we would get

$$\mathcal{L}(x;\theta) := \prod_{i=1}^{n} f(x_i;\theta),$$

where we know think of the realizations as fixed, but θ as an unknown variable!

We call this the *Likelihood function*.

Likelihood function

Products are hard to work with, so we often prefer to analyze the *log*-likelihood

$$\ell(x;\theta) = \log \mathcal{L}(x;\theta) = \log \prod_{i=1}^{n} f(x_i;\theta) = \sum_{i=1}^{n} \log f(x_i;\theta).$$

Example: Coin Probability I

Suppose I have two unfair coins, one falls heads with probability $\theta_1 = 0.8$ and the other falls heads with probability $\theta_2 = 0.4$.

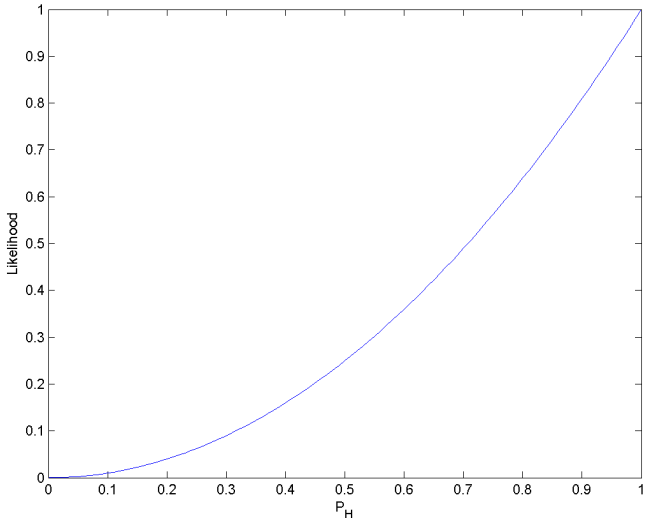
If I randomly pick a coin and it lands TT, which is more *likely* to be the one I picked?

To compute the likelihood, note that $\mathbb{P}(\mathrm{T}) = (1-p)$, so by independence $\mathbb{P}(\mathrm{TT}) = (1-p)^2$.

Example: Coin Probability II

Consider throwing an unfair coin two times and obtaining the sequence $\mathbf{HH}$.

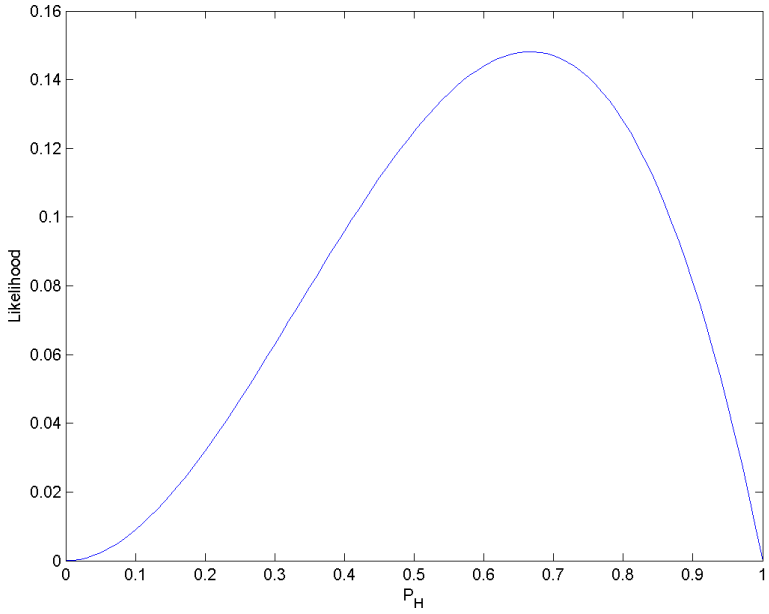
To compute the likelihood, note that $\mathbb{P}(\mathbf{T}) = (1-p)$, so by independence $\mathbb{P}(\mathbf{TT}) = (1-p)^2$.



Example: Coin Probability II

Consider throwing an unfair coin three times and obtaining the sequence **HHT**.

What is the likelihood function in this case?



Example: Normal distribution

For a normal distribution, the likelihood of a random sample $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ can be found from the density of the normal distribution

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Assumptions:

In order to write the likelihood function as we have

$$\ell(x;\theta) = \sum_{i=1}^{n} \log f(x_i;\theta),$$

we need to make a few assumptions about our model.

- $X_1, \ldots, X_n$ is a random sample (iid.)
- $f(x_i; \theta)$ is not equal to zero (we don't observe any "impossible" points in our sample).
- The support of the distribution (the set of x values where $f(x; \theta) > 0$) does not depend on the parameter θ.
- Distinct parameter values ($\theta_1 \neq \theta_2$) must produce distinct probability distributions ($f(x; \theta_1) \neq f(x; \theta_2)$)
- The probability f is differentiable (or easy to maximize) $\longrightarrow$ More on this on Tuesday!

Summary

- We have seen how the central limit theorem can be used in statistics

Next week:

- How to construct a class of efficient and unbiased* estimators
- What is the best possible efficiency we can reach?