Statistics for Astronomers: Lecture 9, 2020.10.28

Prof. Sundar Srinivasan

IRyA/UNAM

Review

The Empirical Rule for normal distributions, the z-score.

Student's t-distribution, the t-score.

Interval estimates: the confidence interval.

Confidence interval: frequentist interpretation

The true parameter value θ is fixed. Repeated observations generate a distribution $p_{\hat{\Theta}}(\hat{\theta})$ of point estimates $\hat{\Theta}$ for θ.

Use this distribution to constrain the true value – interval estimate. Most common interval estimate: confidence interval (CI).

Confidence interval: frequentist interpretation

The true parameter value θ is fixed. Repeated observations generate a distribution $p_{\hat{\Theta}}(\hat{\theta})$ of point estimates $\hat{\Theta}$ for θ.

Use this distribution to constrain the true value – interval estimate. Most common interval estimate: confidence interval (CI).

"The $100(1-\alpha)\%$ CI (for θ) is $[a, b]$" $\Longrightarrow P(a \leq \hat{\Theta} \leq b) = \int_a^b p_{\hat{\Theta}}(\hat{\theta})\ d\hat{\theta} = 1 - \alpha$

$$\Longrightarrow P(\hat{\Theta} < a) + P(\hat{\Theta} > b) = \alpha$$

(Note: definitions don't contain θ, only its estimates $\hat{\theta}$!)

Confidence interval: frequentist interpretation

The true parameter value θ is fixed. Repeated observations generate a distribution $p_{\hat{\Theta}}(\hat{\theta})$ of point estimates $\hat{\Theta}$ for θ.

Use this distribution to constrain the true value – interval estimate. Most common interval estimate: confidence interval (CI).

$$\text{“The } 100(1-\alpha)\% \text{ CI (for } \theta) \text{ is } [a, b]\text{”} \implies P(a \leq \hat{\Theta} \leq b) = \int_a^b p_{\hat{\Theta}}(\hat{\theta})\ d\hat{\theta} = 1 - \alpha$$

$$\implies P(\hat{\Theta} < a) + P(\hat{\Theta} > b) = \alpha$$

(Note: definitions don't contain θ, only its estimates $\hat{\theta}$!)

If number of parameters $p > 1$, confidence set or confidence region.

Confidence interval: frequentist interpretation

The true parameter value θ is fixed. Repeated observations generate a distribution $p_{\hat{\Theta}}(\hat{\theta})$ of point estimates $\hat{\Theta}$ for θ.

Use this distribution to constrain the true value – interval estimate. Most common interval estimate: confidence interval (CI).

$$\text{“The } 100(1-\alpha)\% \text{ CI (for } \theta) \text{ is } [a, b]\text{”} \implies P(a \leq \hat{\Theta} \leq b) = \int_a^b p_{\hat{\Theta}}(\hat{\theta})\, d\hat{\theta} = 1 - \alpha$$

$$\implies P(\hat{\Theta} < a) + P(\hat{\Theta} > b) = \alpha$$

(Note: definitions don't contain θ, only its estimates $\hat{\theta}$!)

If number of parameters $p > 1$, confidence set or confidence region.

Frequentist interpretation of CI convoluted! Bayesian “credible interval” more straightforward.

“95% CI of $[-1.3, 1.3]$” $=$ fixed (unknown) θ such that we observe $\hat{\Theta}$ outside $[-1.3, 1.3] \leq 5\%$ of the time.

Equivalently, if CI computed $N \gg 1$ times using the same procedure, 95% of CIs will contain true value.

Since θ fixed, a single CI will either trap it (probability $= 1$) or it won't (probability $= 0$).

Confidence interval: frequentist interpretation

The true parameter value θ is fixed. Repeated observations generate a distribution $p_{\hat{\Theta}}(\hat{\theta})$ of point estimates $\hat{\Theta}$ for θ.

Use this distribution to constrain the true value – interval estimate. Most common interval estimate: confidence interval (CI).

$$\text{“The } 100(1-\alpha)\% \text{ CI (for } \theta) \text{ is } [a, b]\text{”} \implies P(a \leq \hat{\Theta} \leq b) = \int_a^b p_{\hat{\Theta}}(\hat{\theta})\, d\hat{\theta} = 1 - \alpha$$

$$\implies P(\hat{\Theta} < a) + P(\hat{\Theta} > b) = \alpha$$

(Note: definitions don't contain θ, only its estimates $\hat{\theta}$!)

If number of parameters $p > 1$, confidence set or confidence region.

Frequentist interpretation of CI convoluted! Bayesian “credible interval” more straightforward.

“95% CI of $[-1.3, 1.3]$” = fixed (unknown) θ such that we observe $\hat{\Theta}$ outside $[-1.3, 1.3] \leq 5\%$ of the time.

Equivalently, if CI computed $N \gg 1$ times using the same procedure, 95% of CIs will contain true value.

Since θ fixed, a single CI will either trap it (probability $= 1$) or it won't (probability $= 0$).

“A 95% CI (for θ)” = “fraction of CIs generated in same fashion that trap true value θ is 0.95”.

$\neq$ “the probability that a single CI traps the true value is 0.95”.

Note: “<u>A</u> 95% CI” and not “<u>the</u> 95% CI” – for symmetric distributions, less ambiguous.

Confidence interval: frequentist interpretation

The true parameter value θ is fixed. Repeated observations generate a distribution $p_{\hat{\Theta}}(\hat{\theta})$ of point estimates $\hat{\Theta}$ for θ.

Use this distribution to constrain the true value – interval estimate. Most common interval estimate: confidence interval (CI).

$$\text{“The } 100(1-\alpha)\% \text{ CI (for } \theta) \text{ is } [a, b]\text{”} \Longrightarrow P(a \leq \hat{\Theta} \leq b) = \int_a^b p_{\hat{\Theta}}(\hat{\theta})\, d\hat{\theta} = 1 - \alpha$$

$$\Longrightarrow P(\hat{\Theta} < a) + P(\hat{\Theta} > b) = \alpha$$

(Note: definitions don't contain θ, only its estimates $\hat{\theta}$!)

If number of parameters $p > 1$, confidence set or confidence region.

Frequentist interpretation of CI convoluted! Bayesian “credible interval” more straightforward.

“95% CI of $[-1.3, 1.3]$” $=$ fixed (unknown) θ such that we observe $\hat{\Theta}$ outside $[-1.3, 1.3] \leq 5\%$ of the time.

Equivalently, if CI computed $N \gg 1$ times using the same procedure, 95% of CIs will contain true value.

Since θ fixed, a single CI will either trap it (probability $= 1$) or it won't (probability $= 0$).

“A 95% CI (for θ)” $=$ “fraction of CIs generated in same fashion that trap true value θ is 0.95”.
$\neq$ “the probability that a single CI traps the true value is 0.95”.

Note: “A 95% CI” and not “the 95% CI” – for symmetric distributions, less ambiguous.

Perform an experiment each day, trap a parameter θ_j in a 95% CI on the j^{th} day. As long as you use the same procedure to construct the CI, it doesn't even have to be the same experiment!!. In the long run, 95% of the intervals you constructed would have trapped the true value of whatever parameter you were exploring.

BUT P(parameter trapped in today's CI) $\in \{0, 1\}$.

General procedure to compute confidence intervals

A CI makes a probabilistic statement about the likely range of point estimates for a parameter.
To construct CIs, we need the probability distribution of the point estimates.

General procedure to compute confidence intervals

A CI makes a probabilistic statement about the likely range of point estimates for a parameter. To construct CIs, we need the probability distribution of the point estimates.

In practice, depending on the problem, it might be difficult or impossible to obtain the full distribution. In such a case, we can estimate the variance and then use, *e.g.*, the Chebyshev Inequality to compute an approximate CI.

General procedure to compute confidence intervals

A CI makes a probabilistic statement about the likely range of point estimates for a parameter. To construct CIs, we need the probability distribution of the point estimates.

In practice, depending on the problem, it might be difficult or impossible to obtain the full distribution. In such a case, we can estimate the variance and then use, *e.g.*, the Chebyshev Inequality to compute an approximate CI.

Example: If $\hat{\theta}_0$ is the mean of estimates for θ, and $\sigma^2(\hat{\theta}_0)$ the variance around this mean,

$$\text{a } 100(1-\alpha)\% \text{ CI is such that, for some } \ell_{\alpha/2},\ P\left(\left|\frac{\theta - \hat{\theta}_0}{\sigma(\hat{\theta}_0)}\right| \geq \ell_{\alpha/2}\right) \leq \alpha.$$

General procedure to compute confidence intervals

A CI makes a probabilistic statement about the likely range of point estimates for a parameter. To construct CIs, we need the probability distribution of the point estimates.

In practice, depending on the problem, it might be difficult or impossible to obtain the full distribution. In such a case, we can estimate the variance and then use, *e.g.*, the Chebyshev Inequality to compute an approximate CI.

Example: If $\hat{\theta}_0$ is the mean of estimates for θ, and $\sigma^2(\hat{\theta}_0)$ the variance around this mean,

a $100(1-\alpha)\%$ CI is such that, for some $\ell_{\alpha/2}$, $P\left(\left|\dfrac{\theta - \hat{\theta}_0}{\sigma(\hat{\theta}_0)}\right| \geq \ell_{\alpha/2}\right) \leq \alpha.$

Comparing this to the Chebyshev Inequality: $P\left(\left|\dfrac{\theta - \hat{\theta}_0}{\sigma(\hat{\theta}_0)}\right| \geq k\right) \leq \dfrac{1}{k^2},$

General procedure to compute confidence intervals

A CI makes a probabilistic statement about the likely range of point estimates for a parameter. To construct CIs, we need the probability distribution of the point estimates.

In practice, depending on the problem, it might be difficult or impossible to obtain the full distribution. In such a case, we can estimate the variance and then use, *e.g.*, the Chebyshev Inequality to compute an approximate CI.

Example: If $\hat{\theta}_0$ is the mean of estimates for θ, and $\sigma^2(\hat{\theta}_0)$ the variance around this mean,

a $100(1-\alpha)\%$ CI is such that, for some $\ell_{\alpha/2}$, $P\left(\left|\frac{\theta - \hat{\theta}_0}{\sigma(\hat{\theta}_0)}\right| \geq \ell_{\alpha/2}\right) \leq \alpha$.

Comparing this to the Chebyshev Inequality: $P\left(\left|\frac{\theta - \hat{\theta}_0}{\sigma(\hat{\theta}_0)}\right| \geq k\right) \leq \frac{1}{k^2}$,

we get an approximate $100(1-\alpha)\%$ CI if we choose $\ell_{\alpha/2} = 1/\sqrt{\alpha}$.

The actual probability enclosed by this CI will be smaller than α since the inequality provides an upper bound.

General procedure to compute confidence intervals

A CI makes a probabilistic statement about the likely range of point estimates for a parameter. To construct CIs, we need the probability distribution of the point estimates.

In practice, depending on the problem, it might be difficult or impossible to obtain the full distribution. In such a case, we can estimate the variance and then use, *e.g.*, the Chebyshev Inequality to compute an approximate CI.

Example: If $\hat{\theta}_0$ is the mean of estimates for θ, and $\sigma^2(\hat{\theta}_0)$ the variance around this mean,

a $100(1-\alpha)\%$ CI is such that, for some $\ell_{\alpha/2}$, $P\left(\left|\dfrac{\theta-\hat{\theta}_0}{\sigma(\hat{\theta}_0)}\right| \geq \ell_{\alpha/2}\right) \leq \alpha$.

Comparing this to the Chebyshev Inequality: $P\left(\left|\dfrac{\theta-\hat{\theta}_0}{\sigma(\hat{\theta}_0)}\right| \geq k\right) \leq \dfrac{1}{k^2}$,

we get an approximate $100(1-\alpha)\%$ CI if we choose $\ell_{\alpha/2} = 1/\sqrt{\alpha}$.

The actual probability enclosed by this CI will be smaller than α since the inequality provides an upper bound.

Example of a point estimate: MLE. Find $\hat{\theta}_{\mathrm{MLE}}$ such that $\mathscr{L}(\theta)$ is maximum. If $\mathscr{L}(\theta)$ known for all values allowed for θ, CI computation straightforward. If not, use the CRLB to at least find lower bound on variance. Let's look at some examples for CIs using MLE.

Terminology

Consider X drawn from an unknown distribution with mean μ and variance σ^2.
Perform a trial N times, obtain values $\{x_i\}$ $(i = 1, \cdots, N)$.

Terminology

Consider X drawn from an unknown distribution with mean μ and variance σ^2.
Perform a trial N times, obtain values $\{x_i\}$ $(i = 1, \cdots, N)$.

Uncertainty associated with

each data point x_i: σ ("1σ uncertainty on a single observation").

sample mean $\overline{x}$, from Central Limit Theorem: $\sigma/\sqrt{N}$ ("1σ uncertainty on sample mean").

Terminology

Consider X drawn from an unknown distribution with mean μ and variance σ^2.
Perform a trial N times, obtain values $\{x_i\}$ $(i = 1, \cdots, N)$.

Uncertainty associated with
each data point x_i: σ ("1σ uncertainty on a single observation").
sample mean $\overline{x}$, from Central Limit Theorem: $\sigma/\sqrt{N}$ ("1σ uncertainty on sample mean").

In this context, "1σ" is short for "one standard deviation", not to the literal value σ.

In this particular example, the "1σ uncertainty" *happens* to have the value σ for a single observation and the value $\sigma/\sqrt{N}$ for the sample mean.

These results are true for any distribution. If σ unknown, estimate from data.

Example: CI for mean of normal with known variance

Case 1: $X \sim \mathcal{N}(\mu, \sigma^2)$. Single observation x. MLE for μ:

Case 2: $X \sim \mathcal{N}(\mu, \sigma^2)$. N observations $\{x_i\}$ $(i = 1, \cdots, N)$. MLE for μ:

Example: CI for mean of normal with known variance

Case 1: $X \sim \mathcal{N}(\mu, \sigma^2)$. Single observation x. MLE for μ: $\hat{\mu}_{\mathrm{MLE}} = x$.

Case 2: $X \sim \mathcal{N}(\mu, \sigma^2)$. N observations $\{x_i\}$ $(i = 1, \cdots, N)$. MLE for μ:

Example: CI for mean of normal with known variance

Case 1: $X \sim \mathscr{N}(\mu, \sigma^2)$. Single observation x. MLE for μ: $\hat{\mu}_{\mathrm{MLE}} = x$.

$$\mathscr{L}(\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{\mu-x}{\sigma}\right)^2\right]$$

Case 2: $X \sim \mathscr{N}(\mu, \sigma^2)$. N observations $\{x_i\}$ $(i = 1, \cdots, N)$. MLE for μ:

Example: CI for mean of normal with known variance

Case 1: $X \sim \mathscr{N}(\mu, \sigma^2)$. Single observation x. MLE for μ: $\hat{\mu}_{\mathrm{MLE}} = x$.

$$\mathscr{L}(\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{\mu-x}{\sigma}\right)^2\right]$$

$\implies \mathscr{L}(\mu) \propto \mathscr{N}(\hat{\mu}_{\mathrm{MLE}}, \sigma^2(\hat{\mu}_{\mathrm{MLE}}))$, with $\hat{\mu}_{\mathrm{MLE}} = x$ and $\sigma(\hat{\mu}_{\mathrm{MLE}}) = \sigma$.

Case 2: $X \sim \mathscr{N}(\mu, \sigma^2)$. N observations $\{x_i\}$ $(i = 1, \cdots, N)$. MLE for μ:

Example: CI for mean of normal with known variance

Case 1: $X \sim \mathcal{N}(\mu, \sigma^2)$. Single observation x. MLE for μ: $\hat{\mu}_{\rm MLE} = x$.

$$\mathscr{L}(\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{\mu-x}{\sigma}\right)^2\right]$$

$\implies \mathscr{L}(\mu) \propto \mathcal{N}(\hat{\mu}_{\rm MLE}, \sigma^2(\hat{\mu}_{\rm MLE}))$, with $\hat{\mu}_{\rm MLE} = x$ and $\sigma(\hat{\mu}_{\rm MLE}) = \sigma$.

Case 2: $X \sim \mathcal{N}(\mu, \sigma^2)$. N observations $\{x_i\}$ $(i = 1, \cdots, N)$. MLE for μ: $\hat{\mu}_{\rm MLE} = \overline{x}$.

Example: CI for mean of normal with known variance

Case 1: $X \sim \mathcal{N}(\mu, \sigma^2)$. Single observation x. MLE for μ: $\hat{\mu}_{\mathrm{MLE}} = x$.

$$\mathscr{L}(\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{\mu-x}{\sigma}\right)^2\right]$$

$\Longrightarrow \mathscr{L}(\mu) \propto \mathcal{N}(\hat{\mu}_{\mathrm{MLE}}, \sigma^2(\hat{\mu}_{\mathrm{MLE}}))$, with $\hat{\mu}_{\mathrm{MLE}} = x$ and $\sigma(\hat{\mu}_{\mathrm{MLE}}) = \sigma$.

Case 2: $X \sim \mathcal{N}(\mu, \sigma^2)$. N observations $\{x_i\}$ $(i = 1, \cdots, N)$. MLE for μ: $\hat{\mu}_{\mathrm{MLE}} = \overline{x}$.

$$\mathscr{L}(\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right] = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^2\right].$$

Example: CI for mean of normal with known variance

Case 1: $X \sim \mathcal{N}(\mu, \sigma^2)$. Single observation x. MLE for μ: $\hat{\mu}_{\mathrm{MLE}} = x$.

$$\mathscr{L}(\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{\mu-x}{\sigma}\right)^2\right]$$

$\implies \mathscr{L}(\mu) \propto \mathcal{N}(\hat{\mu}_{\mathrm{MLE}}, \sigma^2(\hat{\mu}_{\mathrm{MLE}}))$, with $\hat{\mu}_{\mathrm{MLE}} = x$ and $\sigma(\hat{\mu}_{\mathrm{MLE}}) = \sigma$.

Case 2: $X \sim \mathcal{N}(\mu, \sigma^2)$. N observations $\{x_i\}$ $(i = 1, \cdots, N)$. MLE for μ: $\hat{\mu}_{\mathrm{MLE}} = \overline{x}$.

$$\mathscr{L}(\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right] = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^2\right].$$

Noting $\displaystyle\sum_{i=1}^{N}(x_i-\mu)^2 = \sum_{i=1}^{N}(x_i-\overline{x})^2 + \sum_{i=1}^{N}(\overline{x}-\mu)^2 = N\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 + (\overline{x}-\mu)^2\right),$

Example: CI for mean of normal with known variance

Case 1: $X \sim \mathscr{N}(\mu, \sigma^2)$. Single observation x. MLE for μ: $\hat{\mu}_{\mathrm{MLE}} = x$.

$$\mathscr{L}(\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{\mu-x}{\sigma}\right)^2\right]$$

$\Longrightarrow \mathscr{L}(\mu) \propto \mathscr{N}(\hat{\mu}_{\mathrm{MLE}}, \sigma^2(\hat{\mu}_{\mathrm{MLE}}))$, with $\hat{\mu}_{\mathrm{MLE}} = x$ and $\sigma(\hat{\mu}_{\mathrm{MLE}}) = \sigma$.

Case 2: $X \sim \mathscr{N}(\mu, \sigma^2)$. N observations $\{x_i\}$ $(i = 1, \cdots, N)$. MLE for μ: $\hat{\mu}_{\mathrm{MLE}} = \overline{x}$.

$$\mathscr{L}(\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right] = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^2\right].$$

Noting $\displaystyle\sum_{i=1}^{N}(x_i-\mu)^2 = \sum_{i=1}^{N}(x_i-\overline{x})^2 + \sum_{i=1}^{N}(\overline{x}-\mu)^2 = N\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 + (\overline{x}-\mu)^2\right)$,

$$\mathscr{L}(\mu) \propto \exp\left[-\frac{N}{2}\left(\frac{\overline{x}-\mu}{\sigma}\right)^2\right] = \exp\left[-\frac{1}{2}\left(\frac{\mu-\overline{x}}{\sigma/\sqrt{N}}\right)^2\right]$$

Example: CI for mean of normal with known variance

Case 1: $X \sim \mathcal{N}(\mu, \sigma^2)$. Single observation x. MLE for μ: $\hat{\mu}_{\rm MLE} = x$.

$$\mathscr{L}(\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{\mu-x}{\sigma}\right)^2\right]$$

$\Longrightarrow \mathscr{L}(\mu) \propto \mathcal{N}(\hat{\mu}_{\rm MLE}, \sigma^2(\hat{\mu}_{\rm MLE}))$, with $\hat{\mu}_{\rm MLE} = x$ and $\sigma(\hat{\mu}_{\rm MLE}) = \sigma$.

Case 2: $X \sim \mathcal{N}(\mu, \sigma^2)$. N observations $\{x_i\}$ $(i = 1, \cdots, N)$. MLE for μ: $\hat{\mu}_{\rm MLE} = \overline{x}$.

$$\mathscr{L}(\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right] = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^2\right].$$

Noting $\displaystyle\sum_{i=1}^{N}(x_i-\mu)^2 = \sum_{i=1}^{N}(x_i-\overline{x})^2 + \sum_{i=1}^{N}(\overline{x}-\mu)^2 = N\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 + (\overline{x}-\mu)^2\right)$,

$$\mathscr{L}(\mu) \propto \exp\left[-\frac{N}{2}\left(\frac{\overline{x}-\mu}{\sigma}\right)^2\right] = \exp\left[-\frac{1}{2}\left(\frac{\mu-\overline{x}}{\sigma/\sqrt{N}}\right)^2\right]$$

$\Longrightarrow \mathscr{L}(\mu) \propto \mathcal{N}(\hat{\mu}_{\rm MLE}, \sigma^2(\hat{\mu}_{\rm MLE}))$, where $\hat{\mu}_{\rm MLE} = \overline{x}$ and $\sigma(\hat{\mu}_{\rm MLE}) = \sigma/\sqrt{N}$.

Example: CI for mean of normal with known variance

Case 1: $X \sim \mathcal{N}(\mu, \sigma^2)$. Single observation x. MLE for μ: $\hat{\mu}_{\rm MLE} = x$.

$$\mathscr{L}(\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{\mu-x}{\sigma}\right)^2\right]$$

$\implies \mathscr{L}(\mu) \propto \mathcal{N}(\hat{\mu}_{\rm MLE}, \sigma^2(\hat{\mu}_{\rm MLE}))$, with $\hat{\mu}_{\rm MLE} = x$ and $\sigma(\hat{\mu}_{\rm MLE}) = \sigma$.

Case 2: $X \sim \mathcal{N}(\mu, \sigma^2)$. N observations $\{x_i\}$ $(i = 1, \cdots, N)$. MLE for μ: $\hat{\mu}_{\rm MLE} = \overline{x}$.

$$\mathscr{L}(\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right] = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^2\right].$$

Noting $\sum_{i=1}^{N}(x_i-\mu)^2 = \sum_{i=1}^{N}(x_i-\overline{x})^2 + \sum_{i=1}^{N}(\overline{x}-\mu)^2 = N\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 + (\overline{x}-\mu)^2\right)$,

$$\mathscr{L}(\mu) \propto \exp\left[-\frac{N}{2}\left(\frac{\overline{x}-\mu}{\sigma}\right)^2\right] = \exp\left[-\frac{1}{2}\left(\frac{\mu-\overline{x}}{\sigma/\sqrt{N}}\right)^2\right]$$

$\implies \mathscr{L}(\mu) \propto \mathcal{N}(\hat{\mu}_{\rm MLE}, \sigma^2(\hat{\mu}_{\rm MLE}))$, where $\hat{\mu}_{\rm MLE} = \overline{x}$ and $\sigma(\hat{\mu}_{\rm MLE}) = \sigma/\sqrt{N}$.

2σ CI ($=$ 95% CI for Gaussian) centered at $\hat{\mu}_{\rm MLE}$ with variance $\sigma^2(\hat{\mu}_{\rm MLE})$: $\left[\hat{\mu}_{\rm MLE} - 2\sigma(\hat{\mu}_{\rm MLE}),\ \hat{\mu}_{\rm MLE} + 2\sigma(\hat{\mu}_{\rm MLE})\right]$.

Example: CI for mean of normal with known variance

Case 1: $X \sim \mathcal{N}(\mu, \sigma^2)$. Single observation x. MLE for μ: $\hat{\mu}_{\rm MLE} = x$.

$$\mathscr{L}(\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{\mu-x}{\sigma}\right)^2\right]$$

$\Longrightarrow \mathscr{L}(\mu) \propto \mathcal{N}(\hat{\mu}_{\rm MLE}, \sigma^2(\hat{\mu}_{\rm MLE}))$, with $\hat{\mu}_{\rm MLE} = x$ and $\sigma(\hat{\mu}_{\rm MLE}) = \sigma$.

Case 2: $X \sim \mathcal{N}(\mu, \sigma^2)$. N observations $\{x_i\}$ $(i = 1, \cdots, N)$. MLE for μ: $\hat{\mu}_{\rm MLE} = \overline{x}$.

$$\mathscr{L}(\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right] = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^2\right].$$

Noting $\sum_{i=1}^{N}(x_i-\mu)^2 = \sum_{i=1}^{N}(x_i-\overline{x})^2 + \sum_{i=1}^{N}(\overline{x}-\mu)^2 = N\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 + (\overline{x}-\mu)^2\right)$,

$$\mathscr{L}(\mu) \propto \exp\left[-\frac{N}{2}\left(\frac{\overline{x}-\mu}{\sigma}\right)^2\right] = \exp\left[-\frac{1}{2}\left(\frac{\mu-\overline{x}}{\sigma/\sqrt{N}}\right)^2\right]$$

$\Longrightarrow \mathscr{L}(\mu) \propto \mathcal{N}(\hat{\mu}_{\rm MLE}, \sigma^2(\hat{\mu}_{\rm MLE}))$, where $\hat{\mu}_{\rm MLE} = \overline{x}$ and $\sigma(\hat{\mu}_{\rm MLE}) = \sigma/\sqrt{N}$.

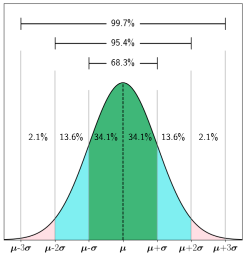
2σ CI (= 95% CI for Gaussian) centered at $\hat{\mu}_{\rm MLE}$ with variance $\sigma^2(\hat{\mu}_{\rm MLE})$: $\left[\hat{\mu}_{\rm MLE} - 2\sigma(\hat{\mu}_{\rm MLE}),\ \hat{\mu}_{\rm MLE} + 2\sigma(\hat{\mu}_{\rm MLE})\right]$.

$\Longrightarrow$ Case 1: $[x - 2\sigma,\ x + 2\sigma]$. Case 2: $\left[\overline{x} - 2\frac{\sigma}{\sqrt{N}},\ \overline{x} + 2\frac{\sigma}{\sqrt{N}}\right]$.

Example: CI for Gaussian uncertainties

A single measurement of the mass of a rock results in a value of 0.2 kg.

The 1σ measurement uncertainty due to the resolution of the mass measuring device is 0.05 kg.



Code for plot available here.

Example: CI for Gaussian uncertainties

A single measurement of the mass of a rock results in a value of 0.2 kg.

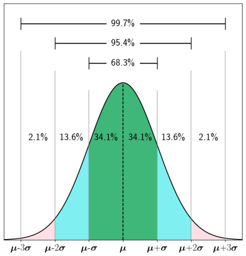
The 1σ measurement uncertainty due to the resolution of the mass measuring device is 0.05 kg.

Construct a 99.7% CI on the true mass of the rock.

Using the Empirical Rule, 99.7% corresponds approx. to 3σ.

90.7% CI $= [\widehat{\mu} - 3\sigma, \widehat{\mu} + 3\sigma] = [0.2 - 0.15, 0.2 + 0.15] = [0.05, 0.35]$ kg.

"The mass of the rock is (0.2 ± 0.15) kg (3σ)".



Code for plot available here.

Example: CI for Gaussian uncertainties

A single measurement of the mass of a rock results in a value of 0.2 kg.

The 1σ measurement uncertainty due to the resolution of the mass measuring device is 0.05 kg.

Construct a 99.7% CI on the true mass of the rock.

Using the Empirical Rule, 99.7% corresponds approx. to 3σ.

90.7% CI $= [\widehat{\mu} - 3\sigma, \widehat{\mu} + 3\sigma] = [0.2 - 0.15, 0.2 + 0.15] = [0.05, 0.35]$ kg.

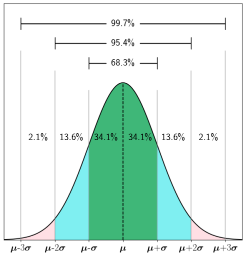
"The mass of the rock is (0.2 ± 0.15) kg (3σ)".

Construct a 82% CI on the true mass of the rock.

$100(1 - \alpha) = 82 \implies \alpha = 0.18$.

```
scipy.stats.norm.ppf(0.18/2) = -1.341
scipy.stats.norm.ppf(1-0.18/2) = 1.341
```
#"1.341 sigma confidence interval"

CI: $[\widehat{\mu} - 1.341\sigma, \widehat{\mu} + 1.341\sigma] = [0.2 - 0.067, 0.2 + 0.067] \approx [0.133, 0.267]$ kg.



Code for plot available here

Example: CI for Gaussian uncertainties

A single measurement of the mass of a rock results in a value of 0.2 kg.

The 1σ measurement uncertainty due to the resolution of the mass measuring device is 0.05 kg.

Construct a 99.7% CI on the true mass of the rock.

Using the Empirical Rule, 99.7% corresponds approx. to 3σ.

90.7% CI $= [\widehat{\mu} - 3\sigma, \widehat{\mu} + 3\sigma] = [0.2 - 0.15, 0.2 + 0.15] = [0.05, 0.35]$ kg.

"The mass of the rock is (0.2 ± 0.15) kg (3σ)".

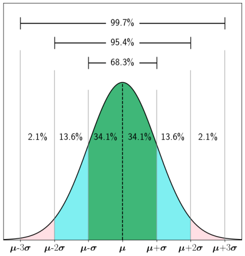
Construct a 82% CI on the true mass of the rock.

$100(1 - \alpha) = 82 \implies \alpha = 0.18$.

```
scipy.stats.norm.ppf(0.18/2) = -1.341
scipy.stats.norm.ppf(1-0.18/2) = 1.341
```
#"1.341 sigma confidence interval"

CI: $[\widehat{\mu} - 1.341\sigma, \widehat{\mu} + 1.341\sigma] = [0.2 - 0.067, 0.2 + 0.067] \approx [0.133, 0.267]$ kg.

What confidence is associated with the interval $[0.00545, 0.3946]$ kg?



Code for plot available here

Example: CI for Gaussian uncertainties

A single measurement of the mass of a rock results in a value of 0.2 kg.

The 1σ measurement uncertainty due to the resolution of the mass measuring device is 0.05 kg.

Construct a 99.7% CI on the true mass of the rock.

Using the Empirical Rule, 99.7% corresponds approx. to 3σ.

90.7% CI $= [\widehat{\mu} - 3\sigma, \widehat{\mu} + 3\sigma] = [0.2 - 0.15, 0.2 + 0.15] = [0.05, 0.35]$ kg.

"The mass of the rock is (0.2 ± 0.15) kg (3σ)".

Construct a 82% CI on the true mass of the rock.

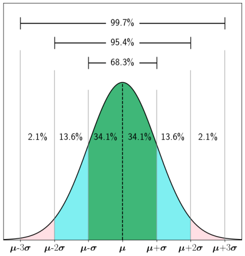
$100(1 - \alpha) = 82 \implies \alpha = 0.18$.

`scipy.stats.norm.ppf(0.18/2) = -1.341`

`scipy.stats.norm.ppf(1-0.18/2) = 1.341` # "1.341 sigma confidence interval"

CI: $[\widehat{\mu} - 1.341\sigma, \widehat{\mu} + 1.341\sigma] = [0.2 - 0.067, 0.2 + 0.067] \approx [0.133, 0.267]$ kg.

What confidence is associated with the interval $[0.00545, 0.3946]$ kg?



Code for plot available here.

We just said something probabilistic about a true parameter with only **one** data point!

Assumptions: uncertainties are Gaussian, we know the standard deviation.

Example: CI for Gaussian uncertainties with unknown σ

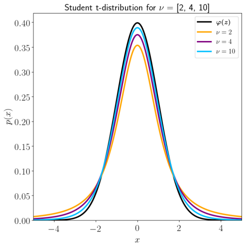
Three measurements of the mass of a rock results in values of $0.2, 0.35$, and 0.25 kg.

As usual, $\hat{\mu} = \overline{x}$.

σ unknown, estd. from data $\Longrightarrow$ functional form of $\mathscr{L}(\mu)$: Student's t-distribution around $\hat{\mu}$.

```
m = np.array([0.2, 0.35, 0.25]); m_mean = m.mean(); m_std = m.std(ddof = 1)
```

$\widehat{\mu} = \overline{x} = 0.267$ kg. #dof $= N - 1 = 2$. $\widehat{\sigma} = 0.076$ kg (Bessel-corrected).



Code for plot available here.

Example: CI for Gaussian uncertainties with unknown σ

Three measurements of the mass of a rock results in values of $0.2, 0.35$, and 0.25 kg.

As usual, $\hat{\mu} = \overline{x}$.

σ unknown, estd. from data $\Longrightarrow$ functional form of $\mathscr{L}(\mu)$: Student's t-distribution around $\hat{\mu}$.

```
m = np.array([0.2, 0.35, 0.25]); m_mean = m.mean(); m_std = m.std(ddof = 1)
```

$\widehat{\mu} = \overline{x} = 0.267$ kg. #dof $= N - 1 = 2$. $\widehat{\sigma} = 0.076$ kg (Bessel-corrected).

Use methods in `scipy.stats.t` for the following:

Construct a 95% CI on the true mass of the rock.

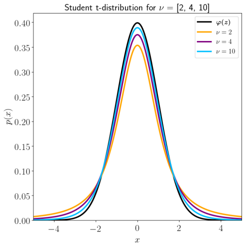
Find value of t (Studentised) for which $P(|T| \leq t) = 0.95$.

```
k95 = t.ppf((1-0.95)/2, df = 2) #number of std dev from mean
```

95% CI $= [\widehat{\mu} - k95 \cdot \widehat{\sigma}, \widehat{\mu} + k95 \cdot \widehat{\sigma}]$

$= [0.267 - 4.303 \times 0.076, 0.267 + 4.303 \times 0.076] = [0.06, 0.59]$ kg.

In general, 95% CI for Student's t wider than 95% CI for Gaussian.



Code for plot available here.

Example: CI for Gaussian uncertainties with unknown σ

Three measurements of the mass of a rock results in values of $0.2, 0.35$, and 0.25 kg.

As usual, $\hat{\mu} = \overline{x}$.

σ unknown, estd. from data $\Longrightarrow$ functional form of $\mathscr{L}(\mu)$: Student's t-distribution around $\hat{\mu}$.

```
m = np.array([0.2, 0.35, 0.25]); m_mean = m.mean(); m_std = m.std(ddof = 1)
```

$\widehat{\mu} = \overline{x} = 0.267$ kg. $\#\mathrm{dof} = N - 1 = 2$. $\widehat{\sigma} = 0.076$ kg (Bessel-corrected).

Use methods in `scipy.stats.t` for the following:

Construct a 95% CI on the true mass of the rock.

Find value of t (Studentised) for which $P(|T| \leq t) = 0.95$.

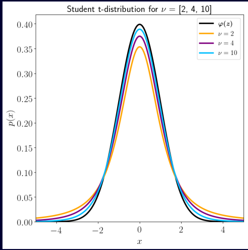
```
k95 = t.ppf((1-0.95)/2, df = 2) #number of std dev from mean
```

95% CI $= [\widehat{\mu} - k95 \cdot \widehat{\sigma}, \widehat{\mu} + k95 \cdot \widehat{\sigma}]$

$= [0.267 - 4.303 \times 0.076, 0.267 + 4.303 \times 0.076] = [0.06, 0.59]$ kg.

In general, 95% CI for Student's t wider than 95% CI for Gaussian.

What confidence is associated with the interval $[-0.0484, 0.5824]$ kg?



Code for plot available here.

Example: CI for Gaussian uncertainties with unknown σ

Three measurements of the mass of a rock results in values of $0.2, 0.35$, and 0.25 kg.

As usual, $\hat{\mu} = \overline{x}$.

σ unknown, estd. from data $\Longrightarrow$ functional form of $\mathscr{L}(\mu)$: Student's t-distribution around $\hat{\mu}$.

```
m = np.array([0.2, 0.35, 0.25]); m_mean = m.mean(); m_std = m.std(ddof = 1)
```

$\widehat{\mu} = \overline{x} = 0.267$ kg. #dof $= N - 1 = 2$. $\widehat{\sigma} = 0.076$ kg (Bessel-corrected).

Use methods in `scipy.stats.t` for the following:

Construct a 95% CI on the true mass of the rock.

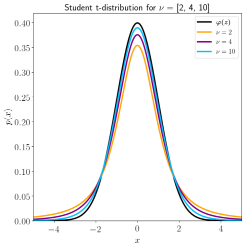
Find value of t (Studentised) for which $P(|T| \leq t) = 0.95$.

```
k95 = t.ppf((1-0.95)/2, df = 2) #number of std dev from mean
```

$$95\%\ \text{CI} = [\widehat{\mu} - k95 \cdot \widehat{\sigma}, \widehat{\mu} + k95 \cdot \widehat{\sigma}]$$
$$= [0.267 - 4.303 \times 0.076, 0.267 + 4.303 \times 0.076] = [0.06, 0.59]\ \text{kg}.$$

In general, 95% CI for Student's t wider than 95% CI for Gaussian.

What confidence is associated with the interval $[-0.0484, 0.5824]$ kg?



Code for plot available here.

See Section 7.2 in Barlow for an interpretation of negative values in the CI in such cases!

Example: CI for (normal approx of) Binomial distribution

Flip a coin $N = 100$ times. Observe: 75 heads, 25 tails.
What is $P(\mathrm{Head})$? What is the 95% CI for this estimate?

Example: CI for (normal approx of) Binomial distribution

Flip a coin $N = 100$ times. Observe: 75 heads, 25 tails.

What is $P(\mathrm{Head})$? What is the 95% CI for this estimate?

Recall: for a Binomial distribution with N trials and k successes, if $\theta = P(1\ \mathrm{success})$, $\mathbb{E}[k] = N\theta$, and $\mathrm{Var}[k] = N\theta(1-\theta)$.

Example: CI for (normal approx of) Binomial distribution

Flip a coin $N = 100$ times. Observe: 75 heads, 25 tails.
What is $P(\mathrm{Head})$? What is the 95% CI for this estimate?

Recall: for a Binomial distribution with N trials and k successes, if $\theta = P(1\ \mathrm{success})$, $\mathbb{E}[k] = N\theta$, and $\mathrm{Var}[k] = N\theta(1-\theta)$.

Likelihood: $\mathscr{L}(\theta) \propto \theta^k (1-\theta)^{N-k}$ – Beta distribution.

Example: CI for (normal approx of) Binomial distribution

Flip a coin $N = 100$ times. Observe: 75 heads, 25 tails.
What is $P(\mathrm{Head})$? What is the 95% CI for this estimate?

Recall: for a Binomial distribution with N trials and k successes, if $\theta = P(1\ \mathrm{success})$,
$\mathbb{E}[k] = N\theta$, and $\mathrm{Var}[k] = N\theta(1-\theta)$.

Likelihood: $\mathscr{L}(\theta) \propto \theta^k (1-\theta)^{N-k}$ – Beta distribution.

MLE (See Lecture 6, Slide 4): $\hat{\theta}_{\mathrm{MLE}} = \dfrac{k}{N} = 0.75$.

Example: CI for (normal approx of) Binomial distribution

Flip a coin $N = 100$ times. Observe: 75 heads, 25 tails.
What is $P(\mathrm{Head})$? What is the 95% CI for this estimate?

Recall: for a Binomial distribution with N trials and k successes, if $\theta = P(1\ \mathrm{success})$,
$\mathbb{E}[k] = N\theta$, and $\mathrm{Var}[k] = N\theta(1-\theta)$.

Likelihood: $\mathscr{L}(\theta) \propto \theta^k (1-\theta)^{N-k}$ – Beta distribution.

MLE (See Lecture 6, Slide 4): $\hat{\theta}_{\mathrm{MLE}} = \dfrac{k}{N} = 0.75.$

$$\mathrm{Var}[\hat{\theta}_{\mathrm{MLE}}] = \mathrm{Var}\left[\frac{k}{N}\right] = \frac{1}{N^2}\mathrm{Var}[k] = \frac{\hat{\theta}_{\mathrm{MLE}}(1-\hat{\theta}_{\mathrm{MLE}})}{N} \approx 0.0019$$

$$\implies \hat{\sigma}(\hat{\theta}_{\mathrm{MLE}}) \approx 0.043.$$

Example: CI for (normal approx of) Binomial distribution

Flip a coin $N = 100$ times. Observe: 75 heads, 25 tails.
What is $P(\mathrm{Head})$? What is the 95% CI for this estimate?

Recall: for a Binomial distribution with N trials and k successes, if $\theta = P(1\ \mathrm{success})$,
$\mathbb{E}[k] = N\theta$, and $\mathrm{Var}[k] = N\theta(1-\theta)$.

Likelihood: $\mathscr{L}(\theta) \propto \theta^k(1-\theta)^{N-k}$ – Beta distribution.

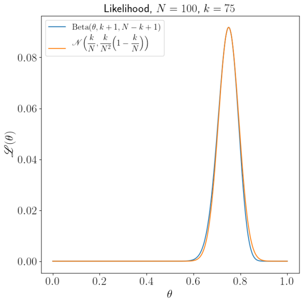
MLE (See Lecture 6, Slide 4): $\hat{\theta}_{\mathrm{MLE}} = \frac{k}{N} = 0.75$.

$$\mathrm{Var}[\hat{\theta}_{\mathrm{MLE}}] = \mathrm{Var}\left[\frac{k}{N}\right] = \frac{1}{N^2}\mathrm{Var}[k] = \frac{\hat{\theta}_{\mathrm{MLE}}(1-\hat{\theta}_{\mathrm{MLE}})}{N} \approx 0.0019$$

$$\implies \hat{\sigma}(\hat{\theta}_{\mathrm{MLE}}) \approx 0.043.$$

asymmetric function, so CI needs to be constructed with care.
However, this problem satisfies conditions for a Gaussian approximation:

$$\mathscr{L}(\theta) \approx \mathscr{N}\left(\hat{\theta}_{\mathrm{MLE}}, \widehat{\sigma^2}(\hat{\theta}_{\mathrm{MLE}})\right) = \mathscr{N}(0.75, (0.043)^2).$$



Code for plot available here.

Example: CI for (normal approx of) Binomial distribution

Flip a coin $N = 100$ times. Observe: 75 heads, 25 tails.
What is $P(\mathrm{Head})$? What is the 95% CI for this estimate?

Recall: for a Binomial distribution with N trials and k successes, if $\theta = P(1\ \mathrm{success})$,
$\mathbb{E}[k] = N\theta$, and $\mathrm{Var}[k] = N\theta(1-\theta)$.

Likelihood: $\mathscr{L}(\theta) \propto \theta^k (1-\theta)^{N-k}$ – Beta distribution.

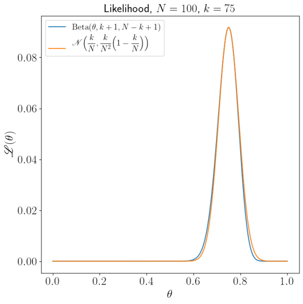
MLE (See ▸ Lecture 6, Slide 4): $\hat{\theta}_{\mathrm{MLE}} = \dfrac{k}{N} = 0.75$.

$$\mathrm{Var}[\hat{\theta}_{\mathrm{MLE}}] = \mathrm{Var}\left[\frac{k}{N}\right] = \frac{1}{N^2}\mathrm{Var}[k] = \frac{\hat{\theta}_{\mathrm{MLE}}(1-\hat{\theta}_{\mathrm{MLE}})}{N} \approx 0.0019$$

$$\implies \hat{\sigma}(\hat{\theta}_{\mathrm{MLE}}) \approx 0.043.$$

asymmetric function, so CI needs to be constructed with care.
However, this problem ▸ satisfies conditions for a Gaussian approximation:

$$\mathscr{L}(\theta) \approx \mathscr{N}\left(\hat{\theta}_{\mathrm{MLE}}, \widehat{\sigma^2}(\hat{\theta}_{\mathrm{MLE}})\right) = \mathscr{N}(0.75, (0.043)^2).$$



Code for plot available ▸ here.

A 95% CI for this problem is also a 2σ CI: $[0.75 - 2\times 0.043,\ 0.75 + 2\times 0.043] \approx [0.66, 0.84]$.

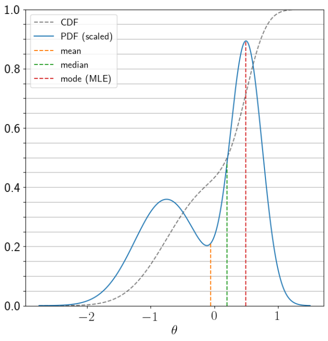
CIs for asymmetric distributions

Example: $\mathscr{L}(\theta) = f\mathcal{N}(\mu_1, \sigma_1^2) + (1-f)\mathcal{N}(\mu_2, \sigma_2^2)$ with $0 < f < 1$ (mixture of Gaussians).

CIs for asymmetric distributions

Example: $\mathscr{L}(\theta) = f\mathcal{N}(\mu_1, \sigma_1^2) + (1-f)\mathcal{N}(\mu_2, \sigma_2^2)$ with $0 < f < 1$ (mixture of Gaussians).

Highly asymmetric: mean $\neq$ median $\neq$ mode!

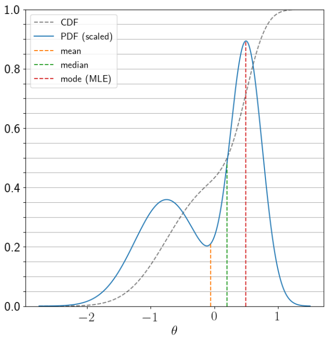


Code for plot available here.

CIs for asymmetric distributions

Example: $\mathscr{L}(\theta) = f\mathcal{N}(\mu_1, \sigma_1^2) + (1-f)\mathcal{N}(\mu_2, \sigma_2^2)$ with $0 < f < 1$ (mixture of Gaussians).
Highly asymmetric: mean $\neq$ median $\neq$ mode!

Three different ways to specify a CI:



Code for plot available here.

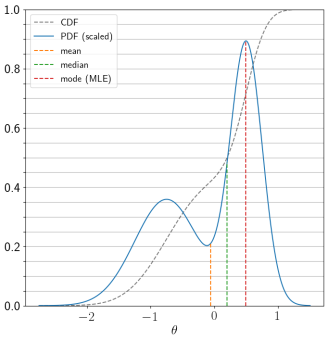
CIs for asymmetric distributions

Example: $\mathscr{L}(\theta) = f\mathscr{N}(\mu_1, \sigma_1^2) + (1-f)\mathscr{N}(\mu_2, \sigma_2^2)$ with $0 < f < 1$ (mixture of Gaussians).

Highly asymmetric: mean $\neq$ median $\neq$ mode!

Three different ways to specify a CI:

1. Central ("equal tail") CI: equal areas rejected on either side (therefore associated with median).



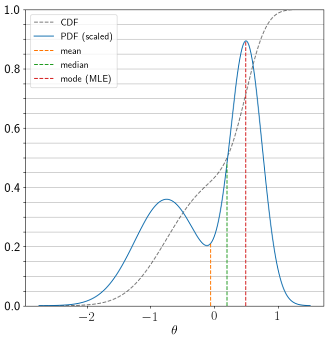
Code for plot available here.

CIs for asymmetric distributions

Example: $\mathscr{L}(\theta) = f\mathcal{N}(\mu_1, \sigma_1^2) + (1-f)\mathcal{N}(\mu_2, \sigma_2^2)$ with $0 < f < 1$ (mixture of Gaussians).
Highly asymmetric: mean $\neq$ median $\neq$ mode!

Three different ways to specify a CI:

1. Central ("equal tail") CI: equal areas rejected on either side (therefore associated with median).
2. Shortest CI: interval chosen closest to region of highest density (therefore usually contains mode).



Code for plot available here.

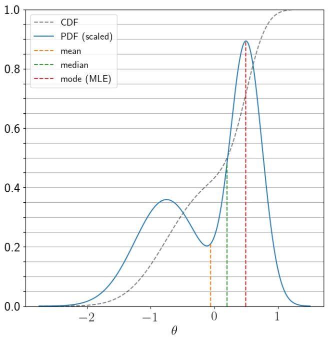
CIs for asymmetric distributions

Example: $\mathscr{L}(\theta) = f\mathcal{N}(\mu_1, \sigma_1^2) + (1-f)\mathcal{N}(\mu_2, \sigma_2^2)$ with $0 < f < 1$ (mixture of Gaussians).

Highly asymmetric: mean $\neq$ median $\neq$ mode!

Three different ways to specify a CI:

1. Central ("equal tail") CI: equal areas rejected on either side (therefore associated with median).
2. Shortest CI: interval chosen closest to region of highest density (therefore usually contains mode).
3. Symmetric CI: upper and lower boundaries equidistant from location parameter (in this case, the MLE, = the mode).



Code for plot available here.

CIs for asymmetric distributions

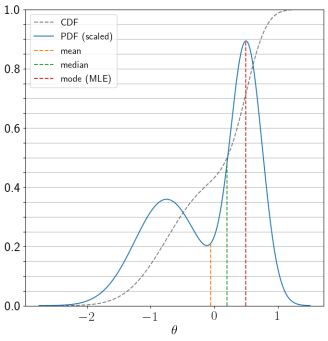
Example: $\mathscr{L}(\theta) = f\mathcal{N}(\mu_1, \sigma_1^2) + (1-f)\mathcal{N}(\mu_2, \sigma_2^2)$ with $0 < f < 1$ (mixture of Gaussians).

Highly asymmetric: mean $\neq$ median $\neq$ mode!

Three different ways to specify a CI:

1. Central ("equal tail") CI: equal areas rejected on either side (therefore associated with median).
2. Shortest CI: interval chosen closest to region of highest density (therefore usually contains mode).
3. Symmetric CI: upper and lower boundaries equidistant from location parameter (in this case, the MLE, $=$ the mode).

Let's construct 50% CIs of each type...



Code for plot available here.

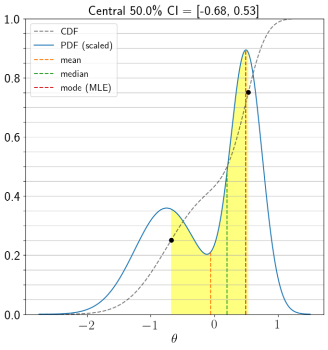
Asymmetric distributions: Central (equal tail) CI

A $100(1-\alpha)\%$ central CI is $[\theta_-,\ \theta_+]$ such that $P(\hat{\theta} \leq \theta_-) = P(\hat{\theta} \geq \theta_+) = \alpha/2$.

Asymmetric distributions: Central (equal tail) CI

A $100(1-\alpha)\%$ central CI is $[\theta_-,\ \theta_+]$ such that $P(\hat{\theta} \leq \theta_-) = P(\hat{\theta} \geq \theta_+) = \alpha/2$.
Only one equal-tail CI is possible for a given α.

The central CI is the sensible choice in most cases.



Code for plot available here.

Asymmetric distributions: Central (equal tail) CI

A $100(1-\alpha)\%$ central CI is $[\theta_-,\ \theta_+]$ such that $P(\hat{\theta} \leq \theta_-) = P(\hat{\theta} \geq \theta_+) = \alpha/2$.

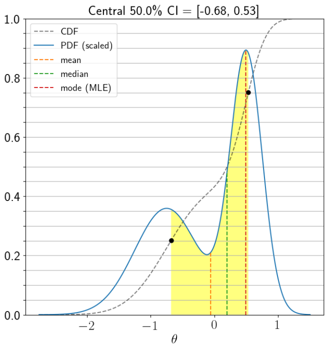
Only one equal-tail CI is possible for a given α.

The central CI is the sensible choice in most cases.

In our specific Gaussian-mixture example, the 50% central CI encloses the mean, median, and mode of the distribution.

Verify: $P(\mathrm{left}) = P(\mathrm{right}) = 50/2 = 25\%$.

Central CI width for this example: $0.53 + 0.68 = 1.21$.



Code for plot available here

Asymmetric distributions: Central (equal tail) CI

A $100(1-\alpha)\%$ central CI is $[\theta_-, \ \theta_+]$ such that $P(\hat{\theta} \leq \theta_-) = P(\hat{\theta} \geq \theta_+) = \alpha/2$.

Only one equal-tail CI is possible for a given α.

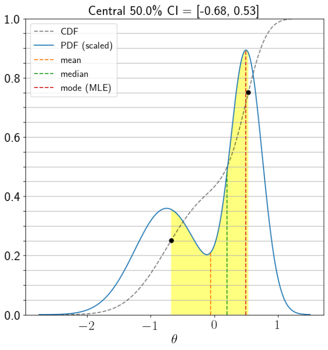
The central CI is the sensible choice in most cases.

In our specific Gaussian-mixture example, the 50% central CI encloses the mean, median, and mode of the distribution.

Verify: $P(\text{left}) = P(\text{right}) = 50/2 = 25\%$.

Central CI width for this example: $0.53 + 0.68 = 1.21$.

What happens to the central CI as its width shrinks?



Code for plot available here.

Asymmetric distributions: Shortest CI

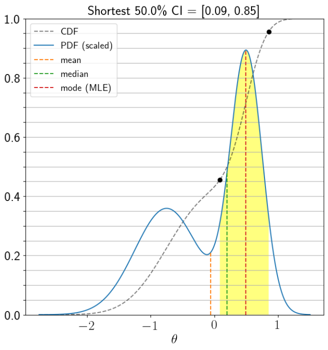
The shortest $100(1-\alpha)\%$ CI is $[\theta_-,\ \theta_+]$ such that, for that α, $\theta_+ - \theta_-$ is minimum.

Asymmetric distributions: Shortest CI

The shortest $100(1-\alpha)\%$ CI is $[\theta_-,\ \theta_+]$ such that, for that α, $\theta_+ - \theta_-$ is minimum.

Only one shortest CI is possible for a given α.

The shortest CI picks out the densest (highest total probability per unit width) part of the distribution.



Code for plot available here.

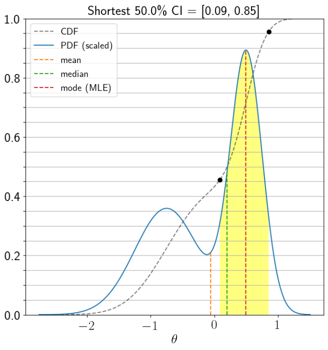
Asymmetric distributions: Shortest CI

The shortest $100(1-\alpha)\%$ CI is $[\theta_-,\ \theta_+]$ such that, for that α, $\theta_+ - \theta_-$ is minimum.

Only one shortest CI is possible for a given α.

The shortest CI picks out the densest (highest total probability per unit width) part of the distribution.

Useful for multimodal distributions such as this example – selects the global maximum of the distribution. Useful in multidimensional space.



Code for plot available here.

Asymmetric distributions: Shortest CI

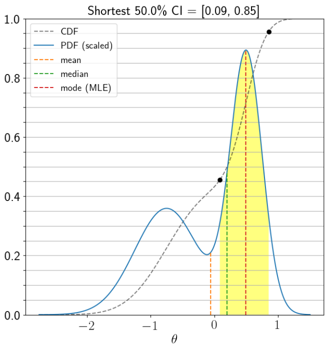
The shortest $100(1-\alpha)\%$ CI is $[\theta_-,\ \theta_+]$ such that, for that α, $\theta_+ - \theta_-$ is minimum.

Only one shortest CI is possible for a given α.

The shortest CI picks out the densest (highest total probability per unit width) part of the distribution.

Useful for multimodal distributions such as this example – selects the global maximum of the distribution. Useful in multidimensional space.

Bayesian estimation: likelihood $\rightarrow$ posterior probability distribution for the parameter. The shortest CI is called the highest posterior density (HPD) interval.



Code for plot available here.

Asymmetric distributions: Shortest CI

The shortest $100(1-\alpha)\%$ CI is $[\theta_-,\ \theta_+]$ such that, for that α, $\theta_+ - \theta_-$ is minimum.

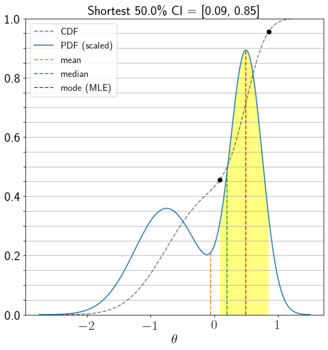
Only one shortest CI is possible for a given α.

The shortest CI picks out the densest (highest total probability per unit width) part of the distribution.

Useful for multimodal distributions such as this example – selects the global maximum of the distribution. Useful in multidimensional space.

Bayesian estimation: likelihood $\rightarrow$ posterior probability distribution for the parameter. The shortest CI is called the highest posterior density (HPD) interval.

Verify: $P(\mathrm{left}) + P(\mathrm{right}) \approx 0.45 + (1 - 0.95) = 50\%$.
Shortest CI width for this example: $0.85 - 0.09 = 0.76$.



Code for plot available here.

Asymmetric distributions: Shortest CI

The shortest $100(1-\alpha)\%$ CI is $[\theta_-,\ \theta_+]$ such that, for that α, $\theta_+ - \theta_-$ is minimum.

Only one shortest CI is possible for a given α.

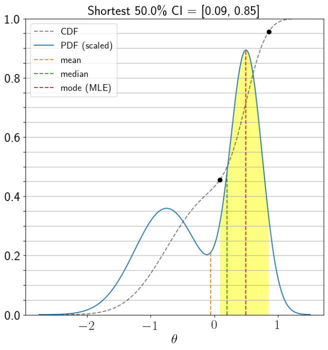
The shortest CI picks out the densest (highest total probability per unit width) part of the distribution.

Useful for multimodal distributions such as this example – selects the global maximum of the distribution. Useful in multidimensional space.

Bayesian estimation: likelihood $\rightarrow$ posterior probability distribution for the parameter. The shortest CI is called the highest posterior density (HPD) interval.

Verify: $P(\mathrm{left}) + P(\mathrm{right}) \approx 0.45 + (1 - 0.95) = 50\%$.
Shortest CI width for this example: $0.85 - 0.09 = 0.76$.

What happens to the central CI as its width shrinks?



Code for plot available here.

Asymmetric distributions: Shortest CI

The shortest $100(1-\alpha)\%$ CI is $[\theta_-, \ \theta_+]$ such that, for that α, $\theta_+ - \theta_-$ is minimum.

Only one shortest CI is possible for a given α.

The shortest CI picks out the densest (highest total probability per unit width) part of the distribution.

Useful for multimodal distributions such as this example – selects the global maximum of the distribution. Useful in multidimensional space.

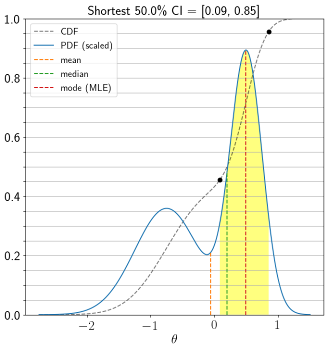
Bayesian estimation: likelihood $\rightarrow$ posterior probability distribution for the parameter. The shortest CI is called the highest posterior density (HPD) interval.

Verify: $P(\mathrm{left}) + P(\mathrm{right}) \approx 0.45 + (1 - 0.95) = 50\%$.
Shortest CI width for this example: $0.85 - 0.09 = 0.76$.

What happens to the central CI as its width shrinks?

Caution!

Sharply peaked, close local maxima – shortest CI may be composed of disconnected regions.



Code for plot available here.

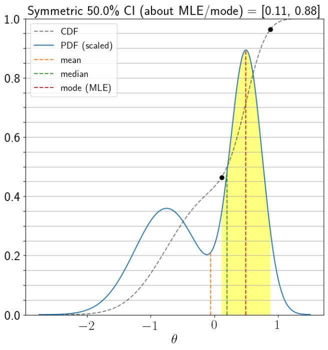
Asymmetric distributions: Symmetric CI

Unlike the other two types of CI, symmetric CIs are not unique. They depend on the choice of centre.

Asymmetric distributions: Symmetric CI

Unlike the other two types of CI, symmetric CIs are not unique. They depend on the choice of centre.

One possible choice is the mean value. In MLE, the more obvious choice is the ML estimate, which is also the mode of the likelihood function.



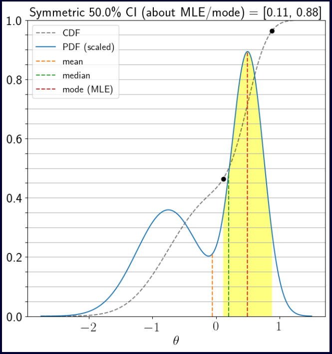
Code for plot available here.

Asymmetric distributions: Symmetric CI

Unlike the other two types of CI, symmetric CIs are not unique. They depend on the choice of centre.

One possible choice is the mean value. In MLE, the more obvious choice is the ML estimate, which is also the mode of the likelihood function.

A symmetric CI around a point estimate $\hat{\theta}_0$ is $[\theta_-,\ \theta_+]$ such that $P(\theta_- \leq \hat{\theta} \leq \theta_+) = 1 - \alpha$ and $\hat{\theta}_0 - \theta_- = \theta_+ - \hat{\theta}_0$ (equal width on either side of $\hat{\theta}_0$).



Code for plot available here.

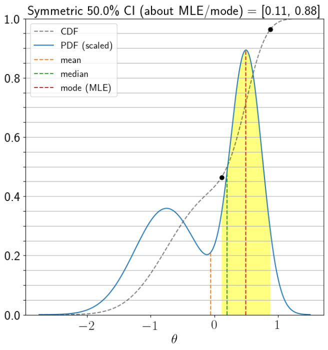
Asymmetric distributions: Symmetric CI

Unlike the other two types of CI, symmetric CIs are not unique. They depend on the choice of centre.

One possible choice is the mean value. In MLE, the more obvious choice is the ML estimate, which is also the mode of the likelihood function.

A symmetric CI around a point estimate $\hat{\theta}_0$ is $[\theta_-, \theta_+]$ such that $P(\theta_- \leq \hat{\theta} \leq \theta_+) = 1 - \alpha$ and $\hat{\theta}_0 - \theta_- = \theta_+ - \hat{\theta}_0$ (equal width on either side of $\hat{\theta}_0$).

Verify: $P(\text{left}) + P(\text{right}) \approx 0.45 + (1 - 0.95) = 50\%$.
Symmetric CI width for this example: $0.88 - 0.11 = 0.77$.



Code for plot available here.

Asymmetric distributions: Symmetric CI

Unlike the other two types of CI, symmetric CIs are not unique. They depend on the choice of centre.

One possible choice is the mean value. In MLE, the more obvious choice is the ML estimate, which is also the mode of the likelihood function.

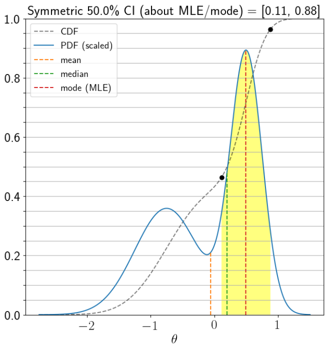
A symmetric CI around a point estimate $\hat{\theta}_0$ is $[\theta_-,\ \theta_+]$ such that $P(\theta_- \le \hat{\theta} \le \theta_+) = 1 - \alpha$ and $\hat{\theta}_0 - \theta_- = \theta_+ - \hat{\theta}_0$ (equal width on either side of $\hat{\theta}_0$).

Verify: $P(\mathrm{left}) + P(\mathrm{right}) \approx 0.45 + (1 - 0.95) = 50\%$.
Symmetric CI width for this example: $0.88 - 0.11 = 0.77$.

Caution!

Multimodal, (almost-)symmetric functions – MLE might pick one peak over the other!

Highly asymmetric functions: if centre of the CI is very far from median, not possible to define a symmetric CI for small α (also in this example!).



Code for plot available here

Example of an asymmetric distribution: Binomial

Flip a coin $N = 5$ times. Observe: 1 head, 4 tails.

What is $P(\mathrm{Head})$? What is the 95% central CI on this estimate?

Example of an asymmetric distribution: Binomial

Flip a coin $N = 5$ times. Observe: 1 head, 4 tails.

What is $P(\mathrm{Head})$? What is the 95% central CI on this estimate?

Once again, $\mathbb{E}[k] = N\theta$, and $\mathrm{Var}[k] = N\theta(1-\theta)$. Likelihood: $\mathscr{L}(\theta) \propto \theta^k(1-\theta)^{N-k}$.

Example of an asymmetric distribution: Binomial

Flip a coin $N = 5$ times. Observe: 1 head, 4 tails.
What is $P(\mathrm{Head})$? What is the 95% central CI on this estimate?

Once again, $\mathbb{E}[k] = N\theta$, and $\mathrm{Var}[k] = N\theta(1-\theta)$. Likelihood: $\mathscr{L}(\theta) \propto \theta^k (1-\theta)^{N-k}$.

$$\hat{\theta}_{\mathrm{MLE}} = \frac{k}{N} = 0.2 \quad \hat{\sigma}(\hat{\theta}_{\mathrm{MLE}}) = \sqrt{\frac{\hat{\theta}_{\mathrm{MLE}}(1-\hat{\theta}_{\mathrm{MLE}})}{N}} \approx 0.179\text{, but not as useful in this case.}$$

Example of an asymmetric distribution: Binomial

Flip a coin $N = 5$ times. Observe: 1 head, 4 tails.
What is $P(\mathrm{Head})$? What is the 95% central CI on this estimate?

Once again, $\mathbb{E}[k] = N\theta$, and $\mathrm{Var}[k] = N\theta(1-\theta)$. Likelihood: $\mathscr{L}(\theta) \propto \theta^k (1-\theta)^{N-k}$.

$$\hat{\theta}_{\mathrm{MLE}} = \frac{k}{N} = 0.2 \quad \hat{\sigma}(\hat{\theta}_{\mathrm{MLE}}) = \sqrt{\frac{\hat{\theta}_{\mathrm{MLE}}(1-\hat{\theta}_{\mathrm{MLE}})}{N}} \approx 0.179,$$ but not as useful in this case.

Can't use Gaussian approximation (*e.g.*, mode close to zero, Gaussian will result in non-negligible probability for negative values, unphysical!).

Example of an asymmetric distribution: Binomial

Flip a coin $N = 5$ times. Observe: 1 head, 4 tails.
What is $P(\mathrm{Head})$? What is the 95% central CI on this estimate?

Once again, $\mathbb{E}[k] = N\theta$, and $\mathrm{Var}[k] = N\theta(1-\theta)$. Likelihood: $\mathscr{L}(\theta) \propto \theta^k(1-\theta)^{N-k}$.

$$\hat{\theta}_{\mathrm{MLE}} = \frac{k}{N} = 0.2 \quad \hat{\sigma}(\hat{\theta}_{\mathrm{MLE}}) = \sqrt{\frac{\hat{\theta}_{\mathrm{MLE}}(1-\hat{\theta}_{\mathrm{MLE}})}{N}} \approx 0.179\text{, but not as useful in this case.}$$

Can't use Gaussian approximation (*e.g.*, mode close to zero, Gaussian will result in non-negligible probability for negative values, unphysical!).

To compute CI, need to know CDF of normalised version of $\mathscr{L}(\theta)$ – does it resemble any standard PDF?

Example of an asymmetric distribution: Binomial

Flip a coin $N = 5$ times. Observe: 1 head, 4 tails.
What is $P(\mathrm{Head})$? What is the 95% central CI on this estimate?

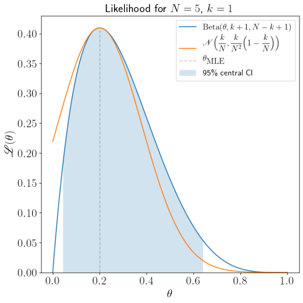
Once again, $\mathbb{E}[k] = N\theta$, and $\mathrm{Var}[k] = N\theta(1-\theta)$. Likelihood: $\mathscr{L}(\theta) \propto \theta^k (1-\theta)^{N-k}$.

$$\hat{\theta}_{\mathrm{MLE}} = \frac{k}{N} = 0.2 \quad \hat{\sigma}(\hat{\theta}_{\mathrm{MLE}}) = \sqrt{\frac{\hat{\theta}_{\mathrm{MLE}}(1-\hat{\theta}_{\mathrm{MLE}})}{N}} \approx 0.179, \text{ but not as useful in this case.}$$

Can't use Gaussian approximation (*e.g.*, mode close to zero, Gaussian will result in non-negligible probability for negative values, unphysical!).

To compute CI, need to know CDF of normalised version of $\mathscr{L}(\theta)$ – does it resemble any standard PDF?

Beta distribution: $\mathrm{Beta}(\alpha, \beta) \propto x^{\alpha-1}(1-x)^{\beta-1}$.
By comparison, $\alpha = k + 1 = 2$, $\beta = N - k + 1 = 5$.



Code for plot available here

Example of an asymmetric distribution: Binomial

Flip a coin $N = 5$ times. Observe: 1 head, 4 tails.
What is $P(\mathrm{Head})$? What is the 95% central CI on this estimate?

Once again, $\mathbb{E}[k] = N\theta$, and $\mathrm{Var}[k] = N\theta(1-\theta)$. Likelihood: $\mathscr{L}(\theta) \propto \theta^k(1-\theta)^{N-k}$.

$$\hat{\theta}_{\mathrm{MLE}} = \frac{k}{N} = 0.2 \quad \hat{\sigma}(\hat{\theta}_{\mathrm{MLE}}) = \sqrt{\frac{\hat{\theta}_{\mathrm{MLE}}(1-\hat{\theta}_{\mathrm{MLE}})}{N}} \approx 0.179, \text{ but not as useful in this case.}$$

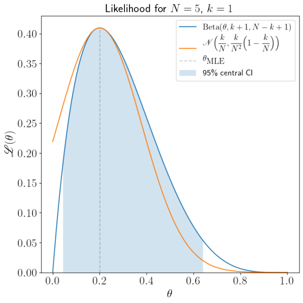
Can't use Gaussian approximation (*e.g.*, mode close to zero, Gaussian will result in non-negligible probability for negative values, unphysical!).

To compute CI, need to know CDF of normalised version of $\mathscr{L}(\theta)$ – does it resemble any standard PDF?

Beta distribution: $\mathrm{Beta}(\alpha, \beta) \propto x^{\alpha-1}(1-x)^{\beta-1}$.
By comparison, $\alpha = k+1 = 2$, $\beta = N - k + 1 = 5$.

Python to the rescue:

```
scipy.stats.beta.ppf(0.025) = 0.043 #lower bound of interval
scipy.stats.beta.ppf(1-0.025) = 0.641 #upper bound of interval
```



Code for plot available here.

Example of an asymmetric distribution: Binomial

Flip a coin $N = 5$ times. Observe: 1 head, 4 tails.
What is $P(\text{Head})$? What is the 95% central CI on this estimate?

Once again, $\mathbb{E}[k] = N\theta$, and $\text{Var}[k] = N\theta(1-\theta)$. Likelihood: $\mathscr{L}(\theta) \propto \theta^k(1-\theta)^{N-k}$.

$$\hat{\theta}_{\text{MLE}} = \frac{k}{N} = 0.2 \qquad \hat{\sigma}(\hat{\theta}_{\text{MLE}}) = \sqrt{\frac{\hat{\theta}_{\text{MLE}}(1-\hat{\theta}_{\text{MLE}})}{N}} \approx 0.179, \text{ but not as useful in this case.}$$

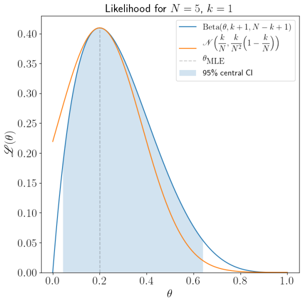
Can't use Gaussian approximation (*e.g.*, mode close to zero, Gaussian will result in non-negligible probability for negative values, unphysical!).

To compute CI, need to know CDF of normalised version of $\mathscr{L}(\theta)$ – does it resemble any standard PDF?

Beta distribution: $\text{Beta}(\alpha, \beta) \propto x^{\alpha-1}(1-x)^{\beta-1}$.
By comparison, $\alpha = k+1 = 2$, $\beta = N - k + 1 = 5$.

Python to the rescue:

```
scipy.stats.beta.ppf(0.025) = 0.043 #lower bound of interval
scipy.stats.beta.ppf(1-0.025) = 0.641 #upper bound of interval
```



Code for plot available here.

95% CI: $[0.043, 0.641]$.