# Statistics for Astronomers: Lecture 10, 2020.11.04

Prof. Sundar Srinivasan

IRyA/UNAM

# Review

Confidence intervals

CI for Gaussian with known $\sigma$ (Standardisation).

CI for Gaussian with unknown $\sigma$ (Studentisation).

# Example: CI for (normal approx of) Binomial distribution

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

Recall: for a Binomial distribution with $N$ trials and $k$ successes, if $\theta = P(1 \text{ success})$,
$\mathbb{E}[k] = N\theta$, 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): $\hat{\theta}_{\mathrm{MLE}} = \frac{k}{N} = 0.75$.

# Example: CI for (normal approx of) Binomial distribution

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

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

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

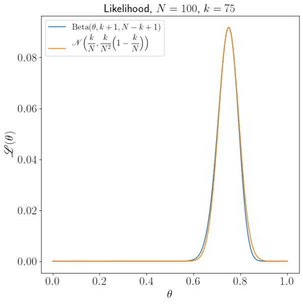
MLE (See Lecture 6, Slide 4): $\hat{\theta}_{\text{MLE}} = \frac{k}{N} = 0.75$.

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

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

Asymmetric function, construct CI with care.
However, this problem satisfies conditions for a Gaussian approximation:

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




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 the MLE for $P(\mathrm{Head})$? 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.

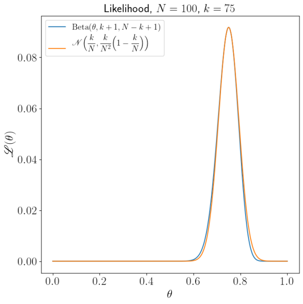
MLE (See Lecture 6, Slide 4): $\hat{\theta}_{\mathrm{MLE}} = \frac{k}{N} = 0.75.$

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

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

Asymmetric function, construct CI 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 $2\sigma$ CI: $[0.75 - 2 \times 0.043,\ 0.75 + 2 \times 0.043] \approx [0.66, 0.84]$.

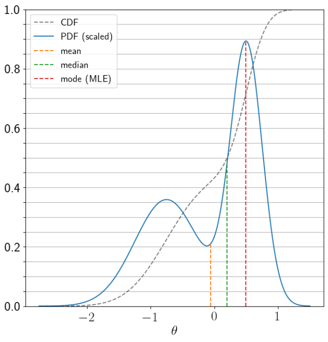
# CIs for asymmetric distributions

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!



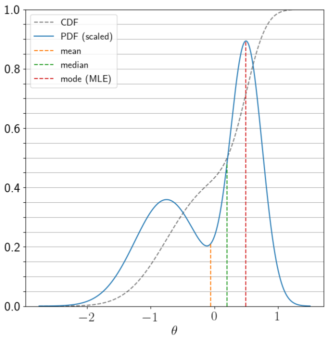

Code for plot available here.

# 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.

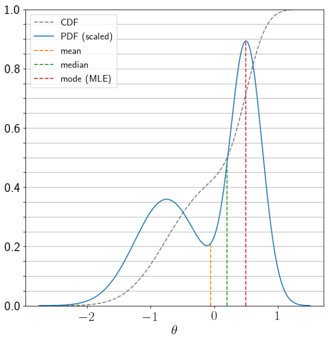
# CIs for asymmetric distributions

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.

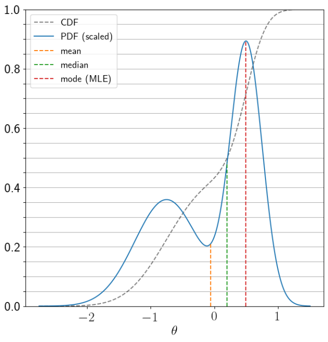
# CIs for asymmetric distributions

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.

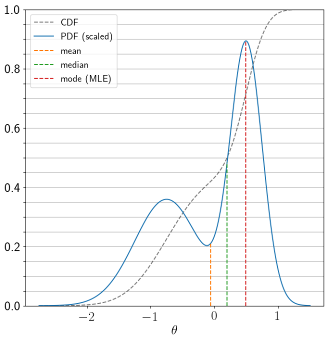
# CIs for asymmetric distributions

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

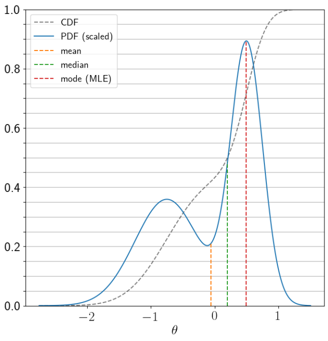
Example: $\mathscr{L}(\theta) = f\mathcal{N}(\mu_1, \sigma_1^2) + (1-f)\mathcal{N}(\mu_2, \sigma_2^2)$ with $0 < f < 1$ (mixture of Gaussians).

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

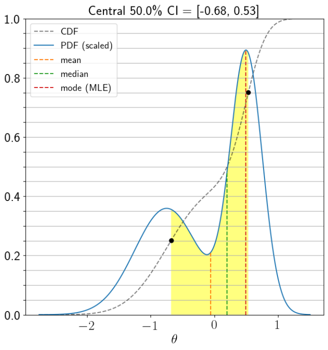
A $100(1-\alpha)\%$ central CI is $[\theta_-,\ \theta_+]$ such that $P(\hat{\theta} \leq \theta_-) = P(\hat{\theta} \geq \theta_+) = \alpha/2$.

# Asymmetric distributions: Central (equal tail) CI

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

Only one equal-tail CI is possible for a given $\alpha$.

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 $P(\hat{\theta} \leq \theta_-) = P(\hat{\theta} \geq \theta_+) = \alpha/2$.

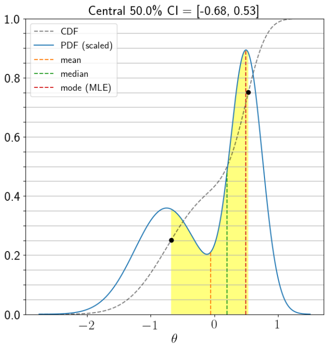
Only one equal-tail CI is possible for a given $\alpha$.

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: $0.53 + 0.68 = 1.21$.




Code for plot available here

# Asymmetric distributions: Central (equal tail) CI

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

Only one equal-tail CI is possible for a given $\alpha$.

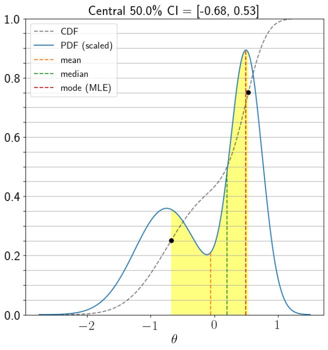
The central CI is the sensible choice in most cases.

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: $0.53 + 0.68 = 1.21$.

What happens to the central CI as its width shrinks?




Code for plot available here

# Asymmetric distributions: Shortest CI

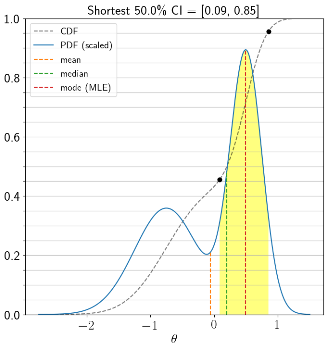
The shortest $100(1-\alpha)\%$ CI is $[\theta_-,\ \theta_+]$ such that, for that $\alpha$, $\theta_+ - \theta_-$ is minimum.

# Asymmetric distributions: Shortest CI

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

Only one shortest CI is possible for a given $\alpha$ [specify rule for multimodal distributions!].

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




Code for plot available here.

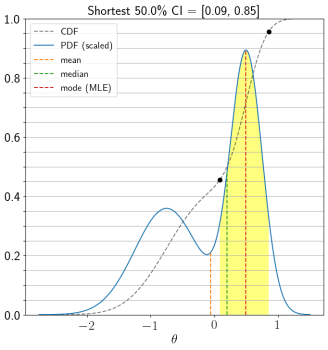
# Asymmetric distributions: Shortest CI

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

Only one shortest CI is possible for a given $\alpha$ [specify rule for multimodal distributions!].

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

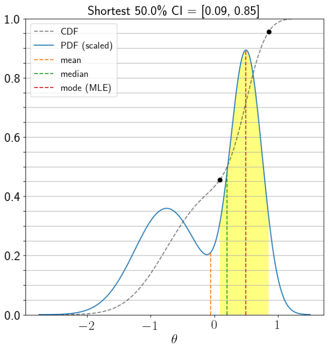
The shortest $100(1-\alpha)\%$ CI is $[\theta_-, \ \theta_+]$ such that, for that $\alpha$, $\theta_+ - \theta_-$ is minimum.

Only one shortest CI is possible for a given $\alpha$ [specify rule for multimodal distributions!].

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 $\alpha$, $\theta_+ - \theta_-$ is minimum.

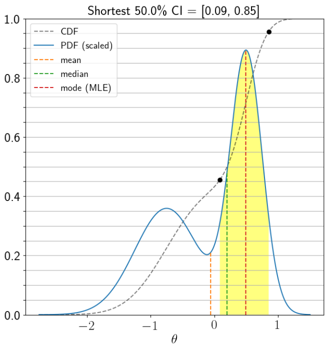
Only one shortest CI is possible for a given $\alpha$ [specify rule for multimodal distributions!].

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: $P(\text{left}) + P(\text{right}) \approx 0.45 + (1 - 0.95) = 50\%$.
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 $\alpha$, $\theta_+ - \theta_-$ is minimum.

Only one shortest CI is possible for a given $\alpha$ [specify rule for multimodal distributions!].

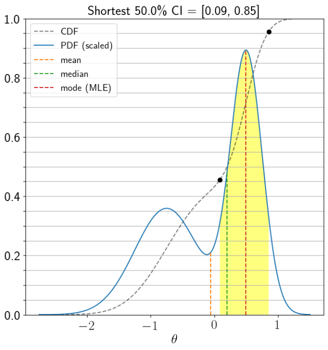
The shortest CI picks out the densest (highest total probability per unit width) part of the distribution.

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

What happens to the shortest 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 $\alpha$, $\theta_+ - \theta_-$ is minimum.

Only one shortest CI is possible for a given $\alpha$ [specify rule for multimodal distributions!].

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.

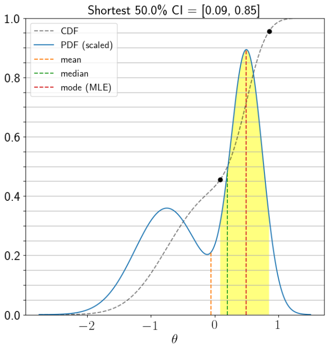
Bayesian estimation: likelihood $\rightarrow$ posterior probability distribution for the parameter. The shortest CI is called the highest posterior density (HPD) interval.

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

What happens to the shortest CI as its width shrinks?

Caution!

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




Code for plot available here.

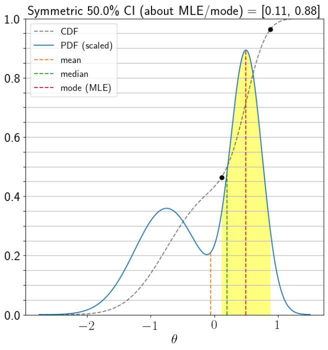
# Asymmetric distributions: Symmetric CI

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.



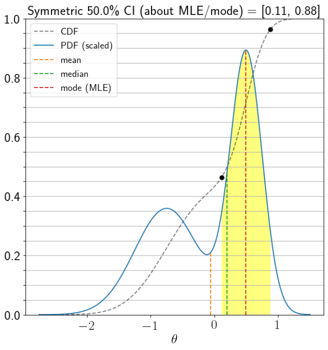

Code for plot available here.

# 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.

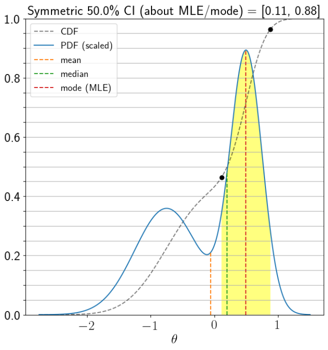
# Asymmetric distributions: Symmetric CI

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: $0.88 - 0.11 = 0.77$.



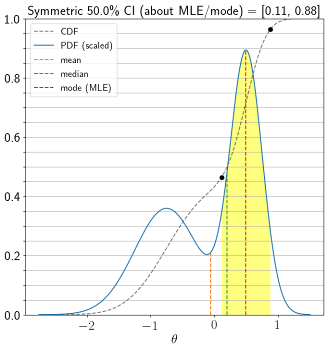

Code for plot available here.

# 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 $P(\theta_- \leq \hat{\theta} \leq \theta_+) = 1 - \alpha$ and $\hat{\theta}_0 - \theta_- = \theta_+ - \hat{\theta}_0$ (equal width on either side of $\hat{\theta}_0$).




Verify: $P(\mathrm{left}) + P(\mathrm{right}) \approx 0.45 + (1 - 0.95) = 50\%$.
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 $\alpha$ (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 MLE of $P(\mathrm{Head})$? 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 MLE of $P(\mathrm{Head})$? 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 MLE of $P(\mathrm{Head})$? 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: $\mathscr{L}(\theta) \propto \theta^k (1-\theta)^{N-k}$.

$\hat{\theta}_{\mathrm{MLE}} = \dfrac{k}{N} = 0.2 \quad \hat{\sigma}(\hat{\theta}_{\mathrm{MLE}}) = \sqrt{\dfrac{\hat{\theta}_{\mathrm{MLE}}(1-\hat{\theta}_{\mathrm{MLE}})}{N}} \approx 0.179$, 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 MLE of $P(\mathrm{Head})$? 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!).

# Example of an asymmetric distribution: Binomial

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

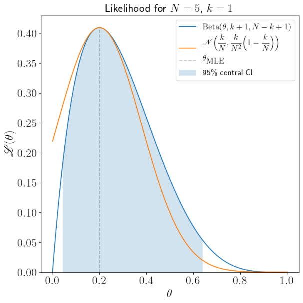
Once again, $\mathbb{E}[k] = N\theta$, and $\mathrm{Var}[k] = N\theta(1-\theta)$. Likelihood: $\mathscr{L}(\theta) \propto \theta^k (1-\theta)^{N-k}$.

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

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: $\mathrm{Beta}(\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$.
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 MLE of $P(\mathrm{Head})$? 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: $\mathscr{L}(\theta) \propto \theta^k(1-\theta)^{N-k}$.

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

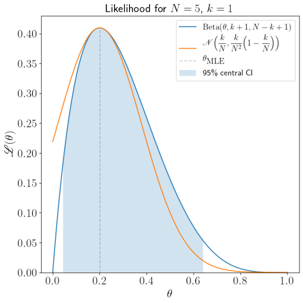
Can't use Gaussian approximation (*e.g.*, mode close to zero, Gaussian will result in non-negligible probability for negative values, unphysical!).

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

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

Python to the rescue:

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




Code for plot available here

# Example of an asymmetric distribution: Binomial

Flip a coin $N = 5$ times. Observe: 1 head, 4 tails.
What is MLE of $P(\text{Head})$? 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: $\mathscr{L}(\theta) \propto \theta^k (1-\theta)^{N-k}$.

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

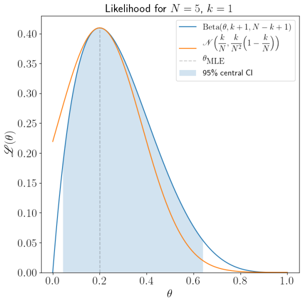
Can't use Gaussian approximation (*e.g.*, mode close to zero, Gaussian will result in non-negligible probability for negative values, unphysical!).

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

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

Python to the rescue:

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




Code for plot available here.

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

# Algorithm for constructing the shortest CI

You must have either have the CDF or must compute it from the data ("empirical distribution function" – you've already encountered concept!).

```
from statsmodels.distributions.empirical_distribution import ECDF
ecdf = ECDF(x)
plt.plot(ecdf.x, ecdf.y)
```

# Algorithm for constructing the shortest CI

You must have either have the CDF or must compute it from the data ("empirical distribution function" – you've already encountered concept!).

```
from statsmodels.distributions.empirical_distribution import ECDF
ecdf = ECDF(x)
plt.plot(ecdf.x, ecdf.y)
```

In principle, we must try to construct a $100(1-\alpha)\%$ CI for each point on the CDF, then find the shortest interval among these.

# Algorithm for constructing the shortest CI

You must have either have the CDF or must compute it from the data ("empirical distribution function" – you've already encountered concept!).

```
from statsmodels.distributions.empirical_distribution import ECDF
ecdf = ECDF(x)
plt.plot(ecdf.x, ecdf.y)
```

In principle, we must try to construct a $100(1-\alpha)\%$ CI for each point on the CDF, then find the shortest interval among these.

In practice, the largest value of $x_{-}$ (lower limit for CI) will be $x_{(j)}$ such that $P(X \leq x_{(j)}) \leq \alpha$ (only these can "extend" over a mass $\geq 1-\alpha$).

# Algorithm for constructing the shortest CI

You must have either have the CDF or must compute it from the data ("empirical distribution function" – you've already encountered concept!).

```
from statsmodels.distributions.empirical_distribution import ECDF
ecdf = ECDF(x)
plt.plot(ecdf.x, ecdf.y)
```

In principle, we must try to construct a $100(1-\alpha)\%$ CI for each point on the CDF, then find the shortest interval among these.

In practice, the largest value of $x_-$ (lower limit for CI) will be $x_{(j)}$ such that $P(X \leq x_{(j)}) \leq \alpha$
(only these can "extend" over a mass $\geq 1-\alpha$).

Similarly, the largest value of $x_+$ (upper limit for CI) will be $x_{(j)}$ such that $P(X \geq x_{(j)}) \leq \alpha$
(only these already contain mass $\geq 1-\alpha$).

# Algorithm for constructing the shortest CI

You must have either have the CDF or must compute it from the data ("empirical distribution function" – you've already encountered concept!).

```
from statsmodels.distributions.empirical_distribution import ECDF
ecdf = ECDF(x)
plt.plot(ecdf.x, ecdf.y)
```

In principle, we must try to construct a $100(1-\alpha)\%$ CI for each point on the CDF, then find the shortest interval among these.

In practice, the largest value of $x_-$ (lower limit for CI) will be $x_{(j)}$ such that $P(X \leq x_{(j)}) \leq \alpha$
(only these can "extend" over a mass $\geq 1-\alpha$).

Similarly, the largest value of $x_+$ (upper limit for CI) will be $x_{(j)}$ such that $P(X \geq x_{(j)}) \leq \alpha$
(only these already contain mass $\geq 1-\alpha$).

1. Sort the points in ascending order if they aren't already sorted.
2. Identify the first $j$ points such that $P(X \leq x_{(j)}) \equiv \mathrm{CDF}(x_{(j)}) \leq \alpha$.
3. For each of these points $x_-$, find $x_+$ such that $P(x_- \leq X < x_+) \geq 1-\alpha$. Compute the interval width $x_+ - x_-$.
4. Find the $x_-, x_+$ pair such that the width is minimum.

# Algorithm for constructing the shortest CI

You must have either have the CDF or must compute it from the data ("empirical distribution function" – you've already encountered concept!).

```
from statsmodels.distributions.empirical_distribution import ECDF
ecdf = ECDF(x)
plt.plot(ecdf.x, ecdf.y)
```

In principle, we must try to construct a $100(1-\alpha)\%$ CI for each point on the CDF, then find the shortest interval among these.

In practice, the largest value of $x_-$ (lower limit for CI) will be $x_{(j)}$ such that $P(X \leq x_{(j)}) \leq \alpha$ (only these can "extend" over a mass $\geq 1-\alpha$).

Similarly, the largest value of $x_+$ (upper limit for CI) will be $x_{(j)}$ such that $P(X \geq x_{(j)}) \leq \alpha$ (only these already contain mass $\geq 1-\alpha$).

1. Sort the points in ascending order if they aren't already sorted.
2. Identify the first $j$ points such that $P(X \leq x_{(j)}) \equiv \mathrm{CDF}(x_{(j)}) \leq \alpha$.
3. For each of these points $x_-$, find $x_+$ such that $P(x_- \leq X < x_+) \geq 1-\alpha$. Compute the interval width $x_+ - x_-$.
4. Find the $x_-, x_+$ pair such that the width is minimum.

You will write code for this as part of the midterm!

# Midterm exam

Monday, 2020.11.09, 09:00 – Tuesday, 2020.11.10, 17:00.

Discussion on Zoom: Tuesday, 2020.11.10, 11:00 – 12:00.
(email anytime!).

"Open Internet" exam
but first consult the lecture notes and homework solutions!