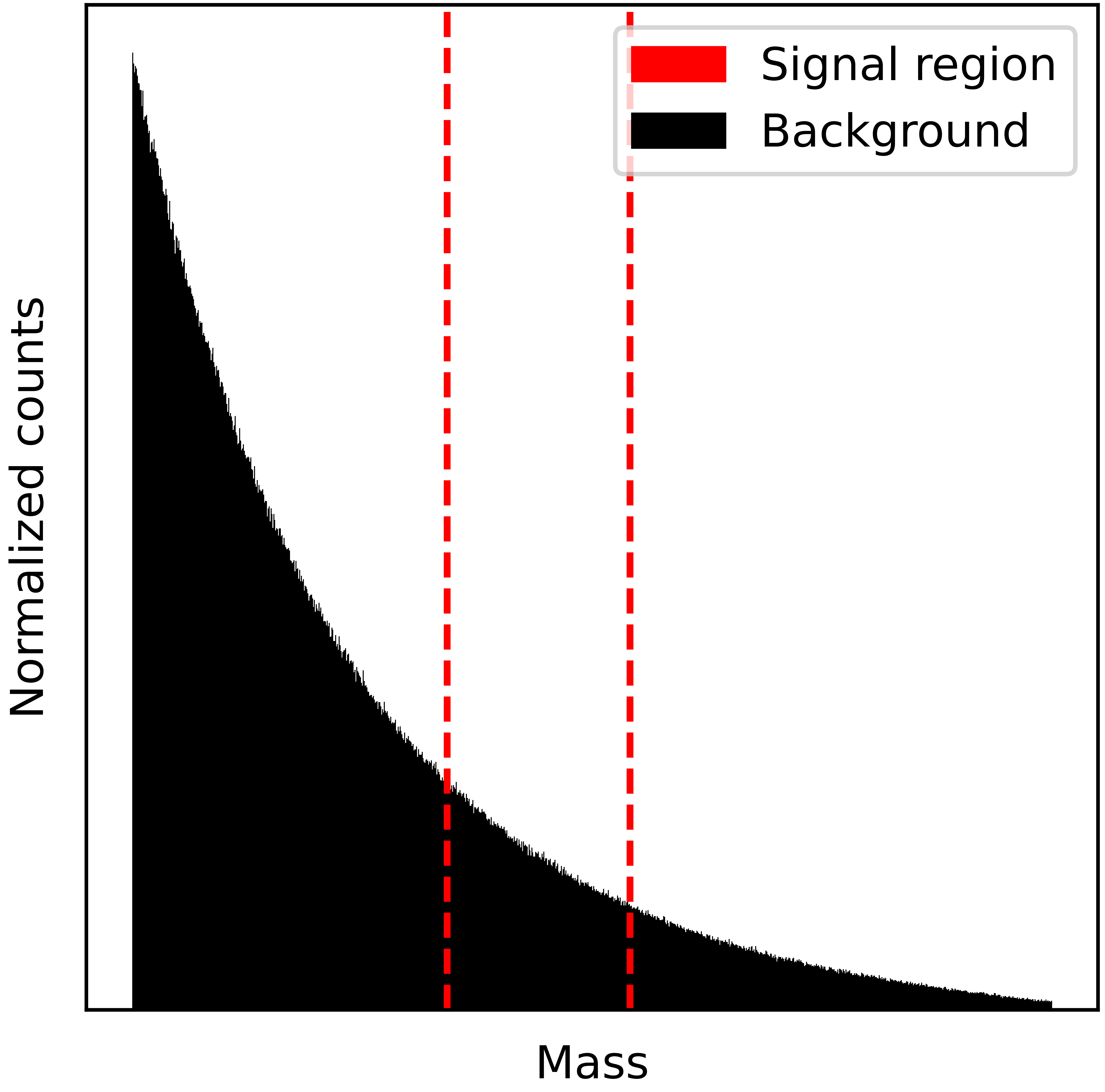
Uncertainty quantification via influence functions

Lucas Kania
(CMU)

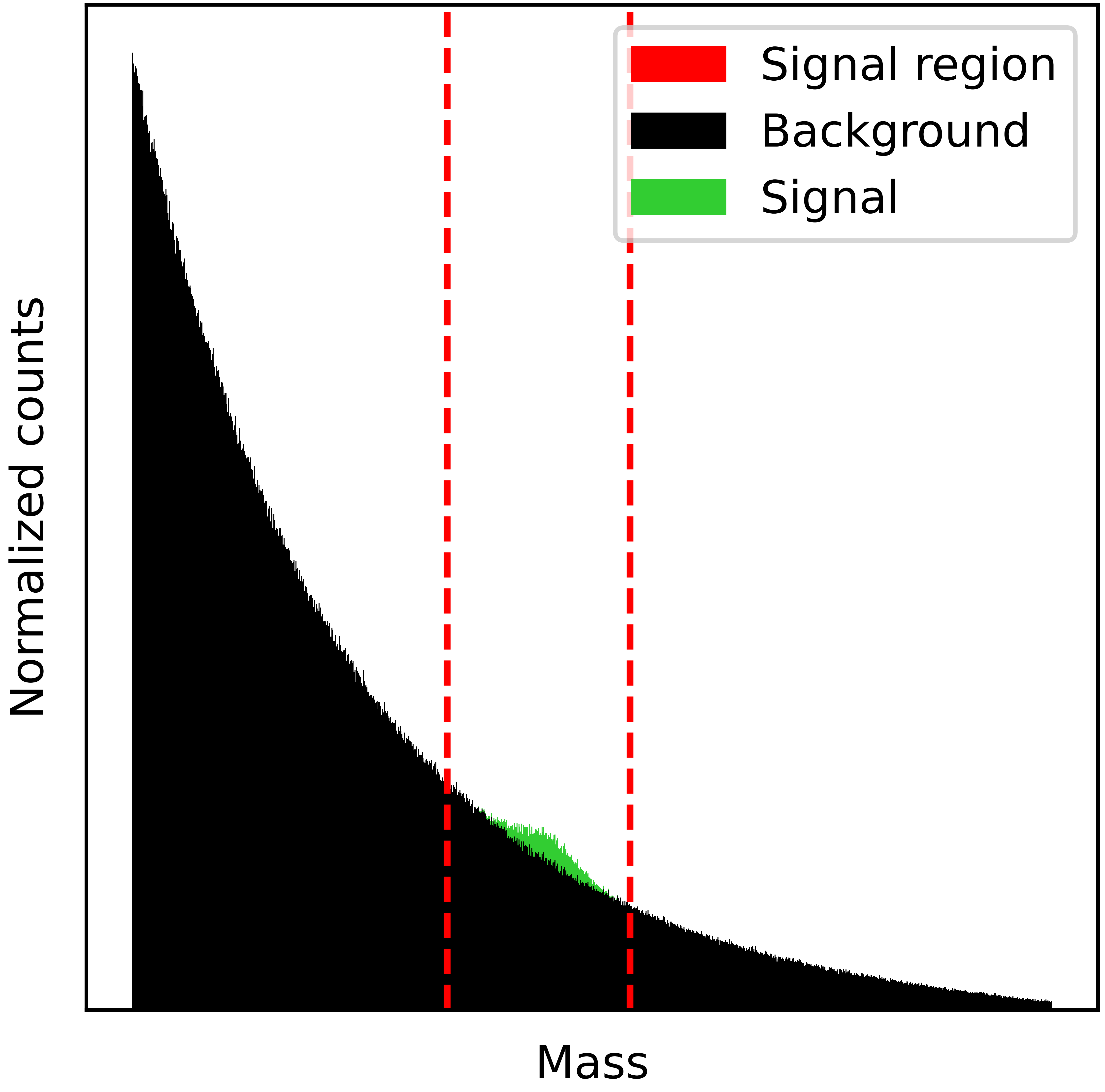
Joint work with: Larry Wasserman (CMU)
Mikael Kuusela (CMU)
Olaf Behnke (DESY)

Is there a new fundamental particle?

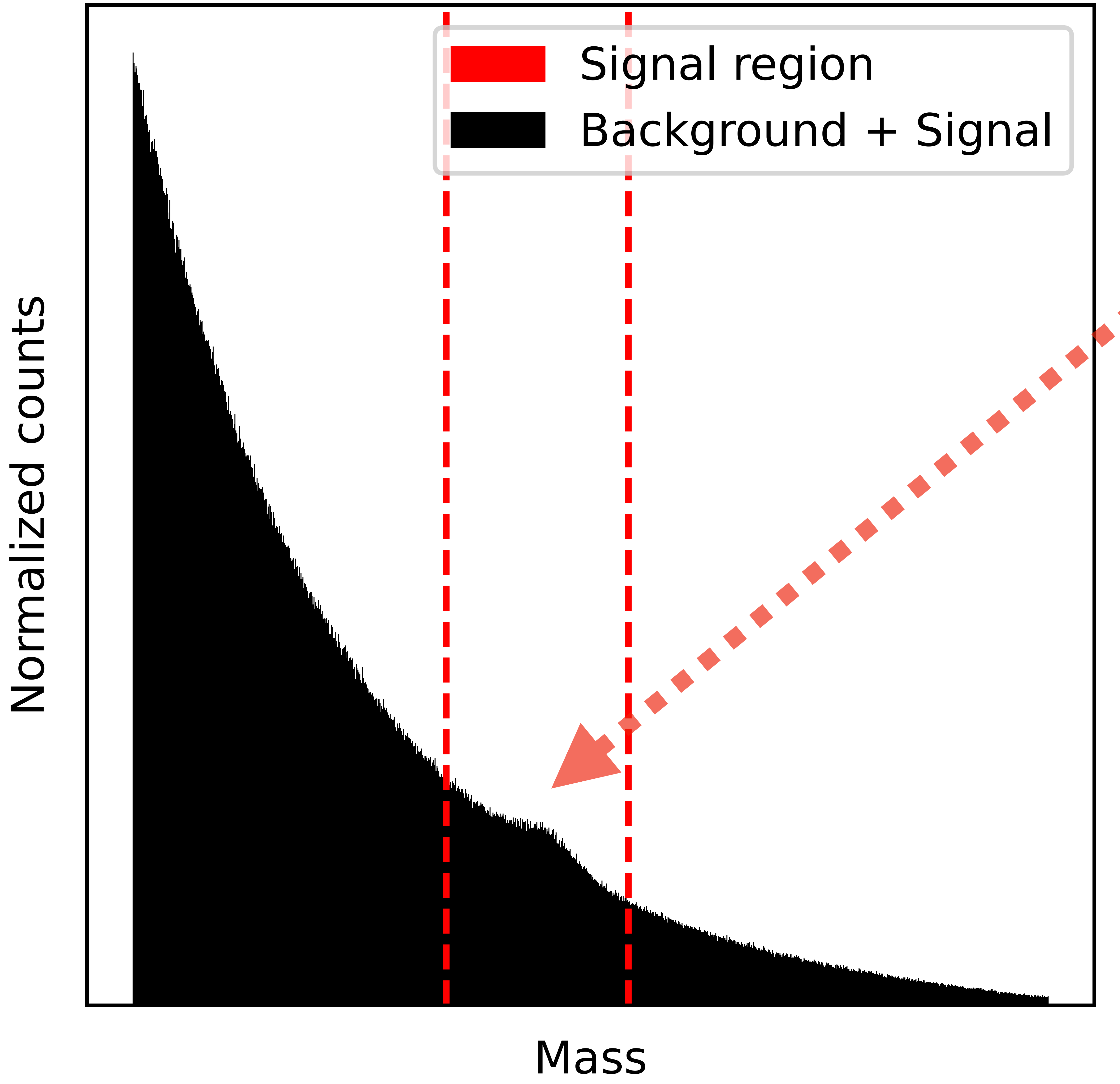
No new fundamental particle = only background



New fundamental particle = back. + signal



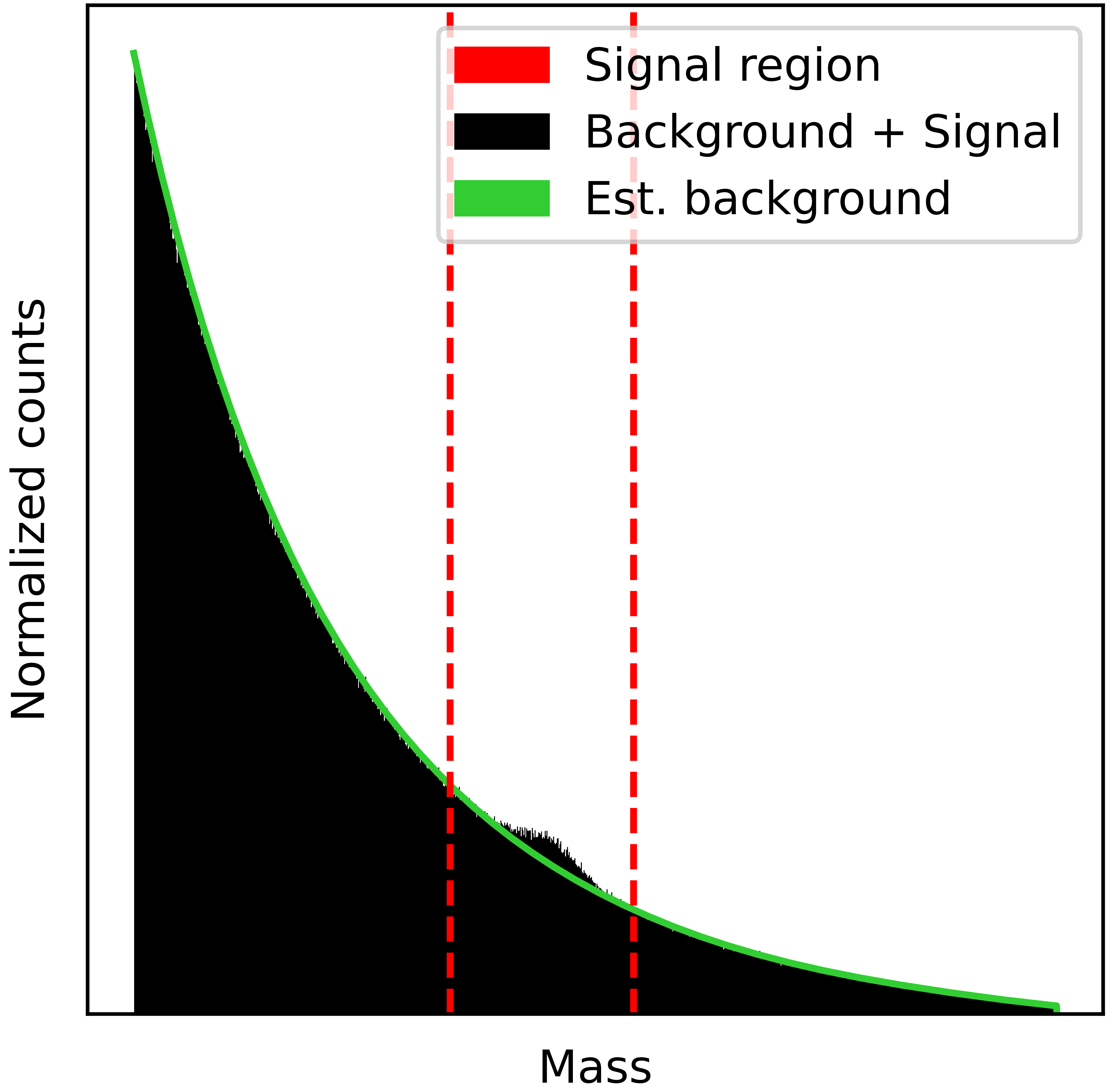
Signal detection requires estimating the background



We only know the theorized signal region

is the bump a significant deviation from the background?

Signal detection requires estimating the background



We only know the theorized signal region

is the bump a significant deviation from the background?

Signal detection

1. Estimate the background

2. Check if the bump is far away from the est. background

Deviance test [Algeri '19], LRT [Cowan'11]

Statistical problem

Model the data as a mixture of background and signal densities

$$X \sim F \;:\; dF(x) = (1-\lambda)\cdot dB(x) \;+\; \lambda \cdot dS_\theta(x)$$

Signal detection construct a confidence interval for the **signal strength** λ

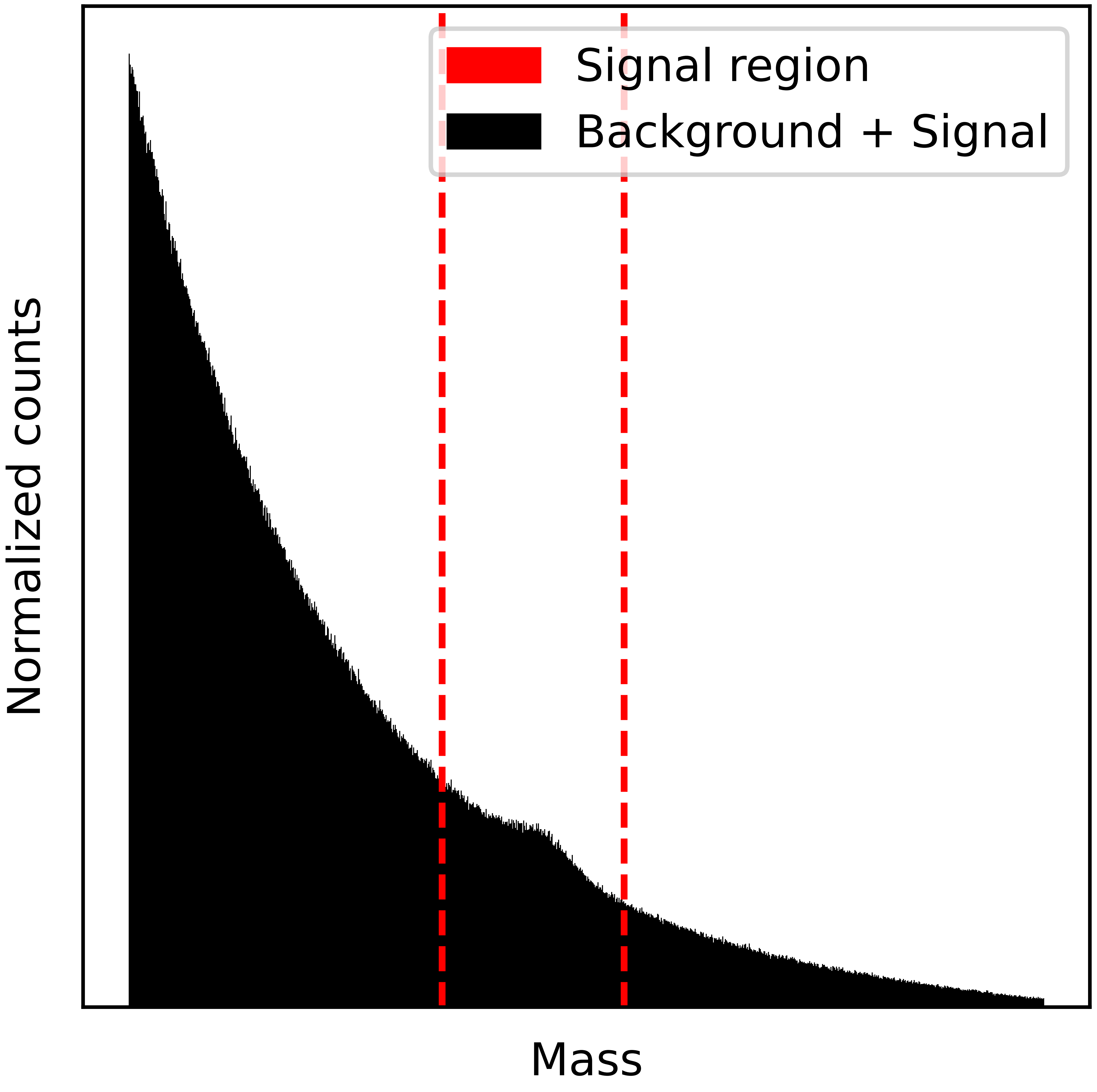
This talk

We focus on constructing confidence intervals for λ

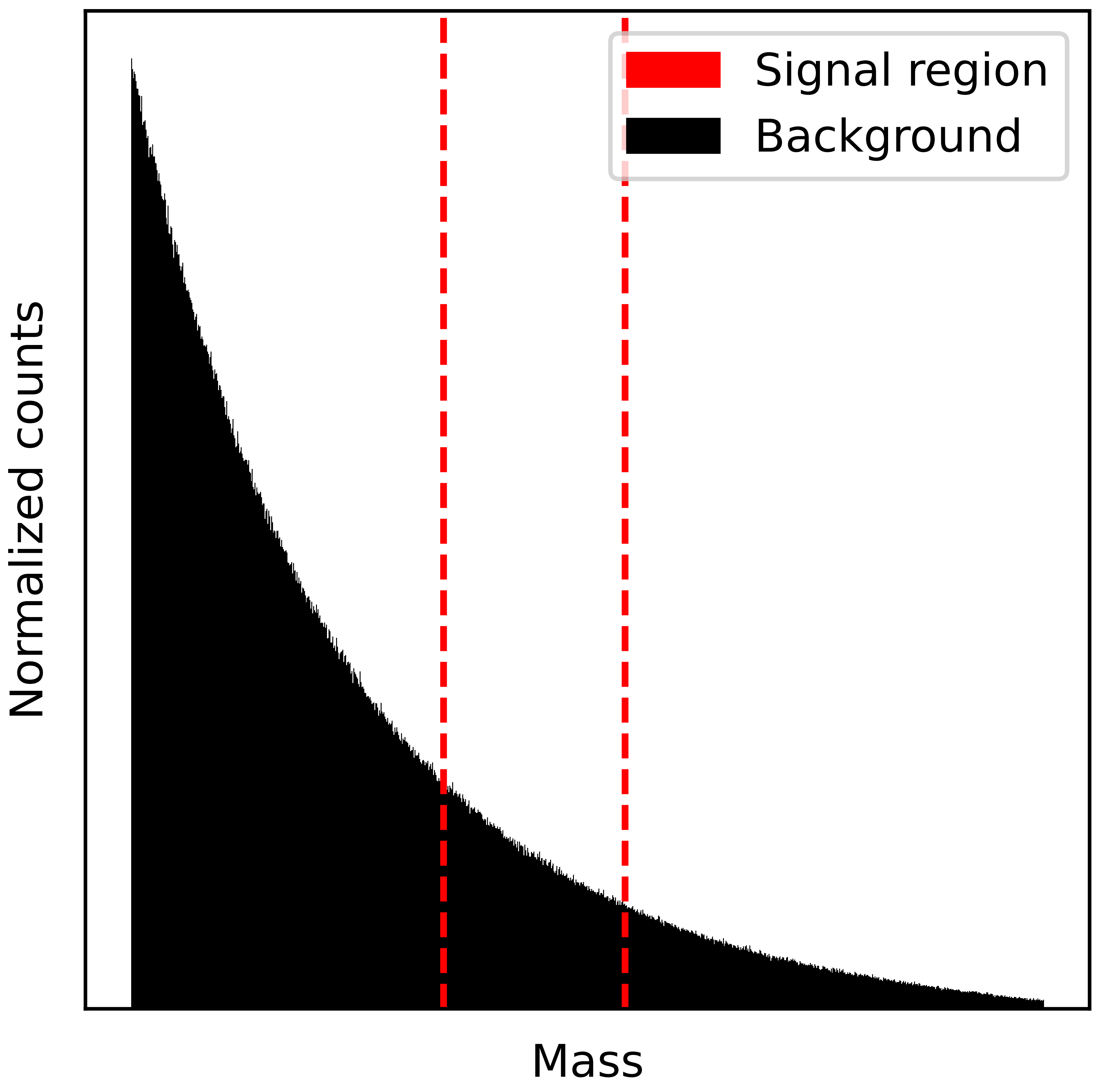
Characterise λ as a fixed function of $F \to \lambda(F)$

Quantify the uncertainty due to using $\hat{F} \to \lambda(\hat{F})$

Approach 1: use two samples

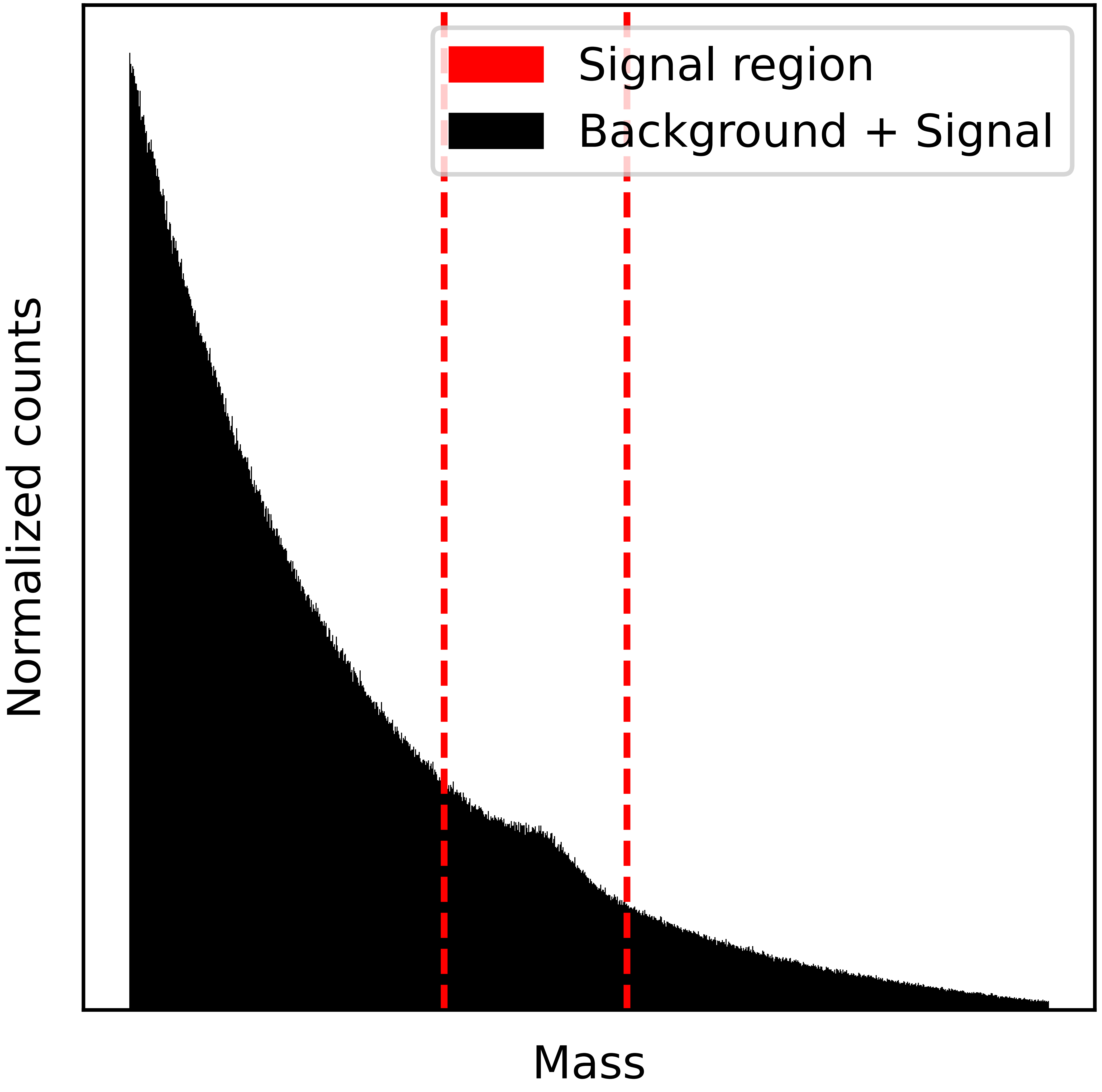


Assume access to a sample only from the background



Problem: it might not be available in all experiments

Approach 2: assume a signal model



Assume a signal model a do joint optimisation

Model the data as a mixture of background and signal densities

$$X \sim F \ : \ dF(x) = (1-\lambda) \cdot dB(x) \ + \ \lambda \cdot dS_\theta(x)$$

polynomials gaussian

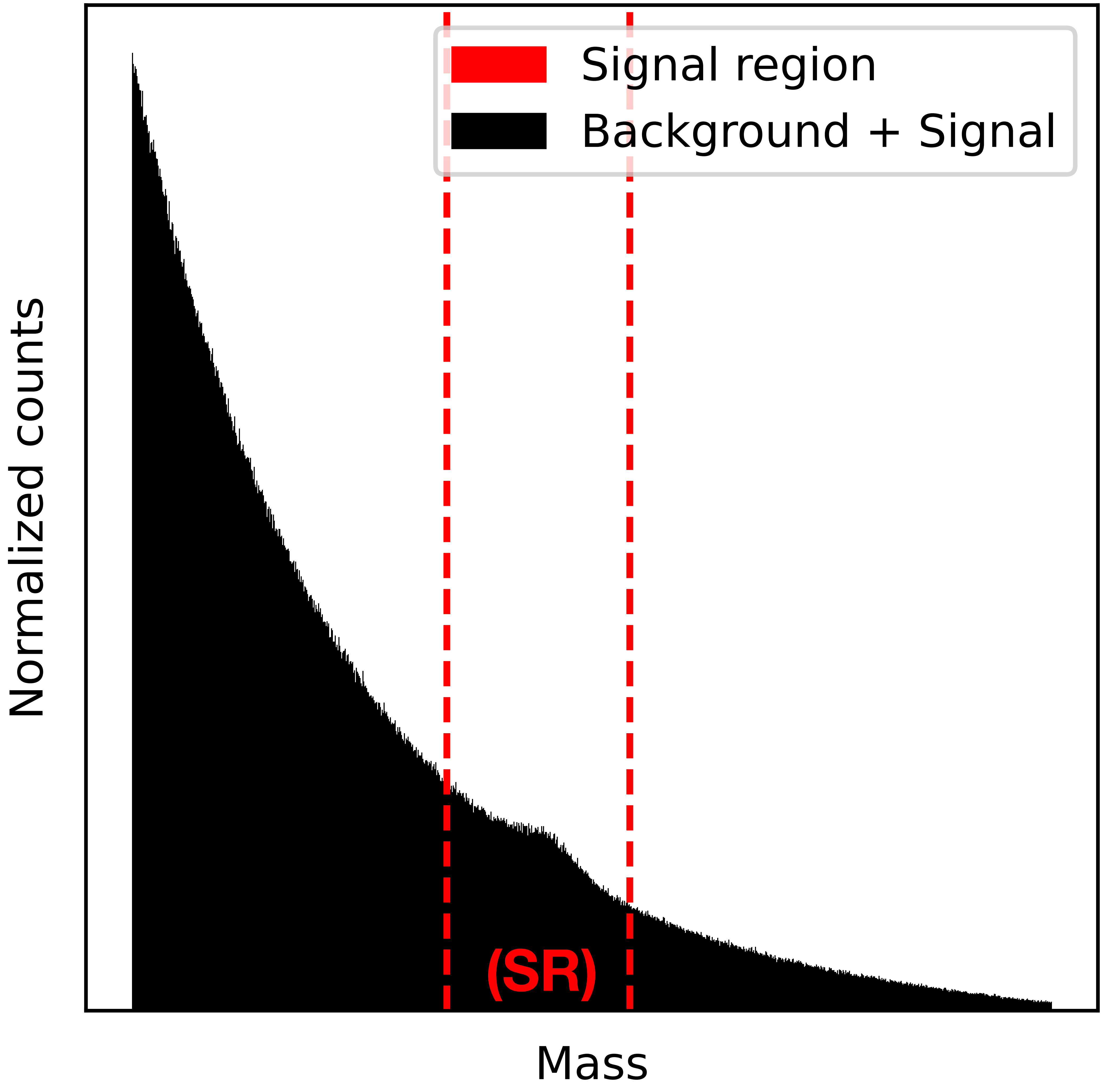
Fit background, signal and signal strength together

$$\lambda_{MLE}(F) = \arg \min_{\lambda, \theta, dB \in \mathscr{B}} \text{KL} \left(dF \ , \ (1-\lambda) \cdot dB + \lambda \cdot dS_\theta \right)$$

Problems the background density might fit some of the signal

if $\lambda = 0$ the model is over-parametrised

Approach 3: assume a control region



We don't assume access to a pure sample from the background

We assume that the signal vanishes outside the signal region **(SR)**

$$X \sim F \ : \ dF(x) = (1 - \lambda) \cdot dB(x) + \lambda \cdot dS_\theta(x)$$

$$dS_\theta(x) = 0 \quad \forall x \notin SR$$

Outside the signal region the data follows a scaled background

$$dF(x) = (1 - \lambda) \cdot dB(x) \quad \forall x \notin SR$$

Represent signal strength as a function of measure F and B

Prob. that $X \sim F$ falls in SR

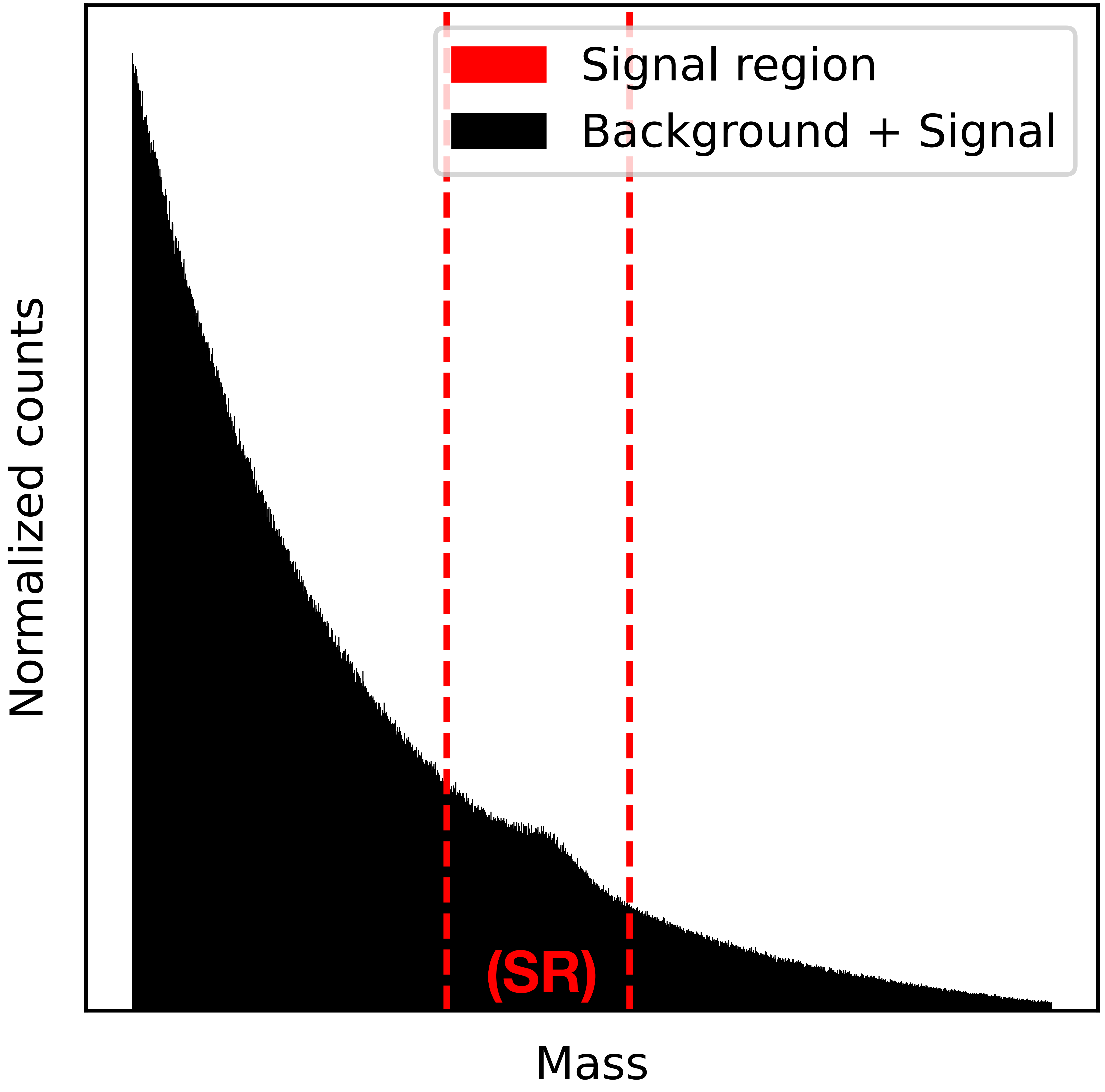
$$\lambda(F, B) = 1 - \frac{1 - F(SR)}{1 - B(SR)}$$

Prob. that $X \sim B$ falls in SR

We have data from observations from F but not from B

counting experiments [Behnke '13]

Approach 3: assume a control region



We assume that the signal vanishes outside the signal region **(SR)**

Represent signal strength as a function of measure F and B

$$\lambda(F, B) = 1 - \frac{1 - F(SR)}{1 - B(SR)}$$

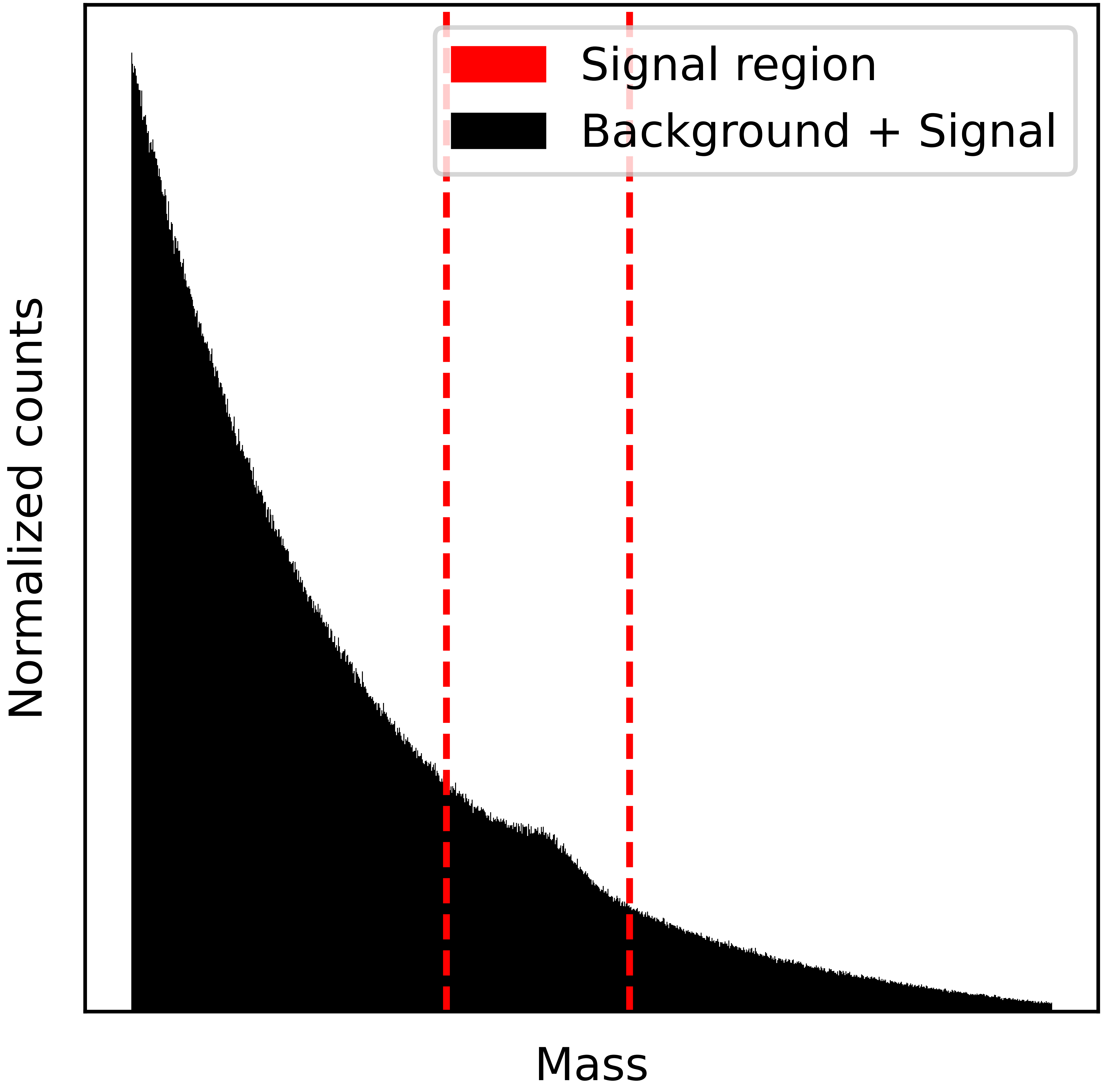
The conditional density of the data on the control region is the conditional background density

$$X \mid X \notin SR \sim \frac{dF(x)}{1 - F(SR)} = \frac{dB(x)}{1 - B(SR)}$$

Assuming that the background can be identified using only the data outside the signal region

$$\lambda(F) = 1 - \frac{1 - F(SR)}{1 - B_F^*(SR)} \quad \text{where} \quad dB_F^* = \arg\min_{d\tilde{B} \in \mathscr{B}} d\left(\frac{dF}{1 - F(SR)}, \frac{d\tilde{B}}{1 - \tilde{B}(SR)} \right)$$

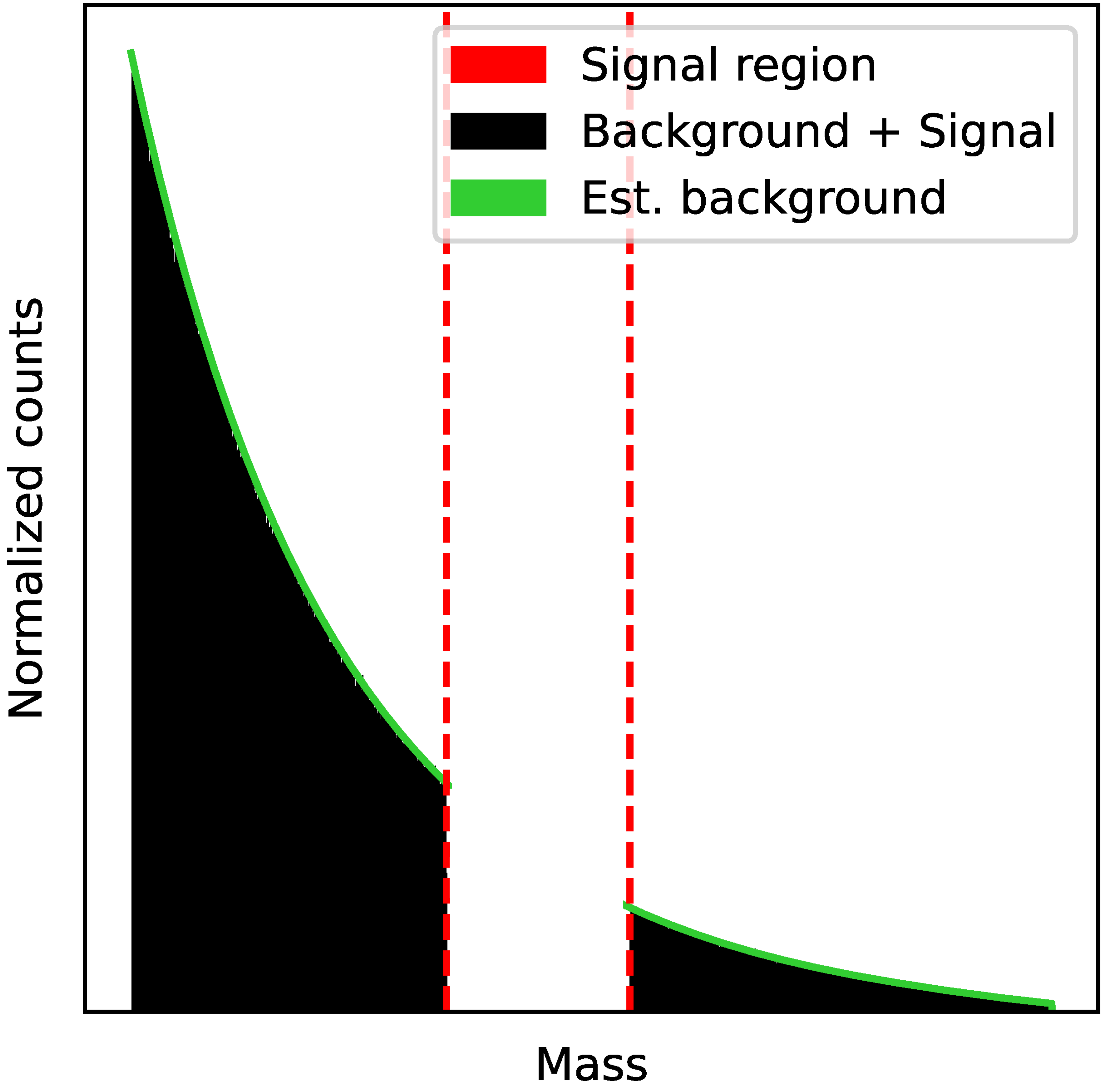
Approach 3: assume a control region



0. Under our assumption, outside the signal region **(SR)**

$$dF(x) = (1 - \lambda) \cdot dB(x) \quad \forall \notin SR$$

Approach 3: assume a control region



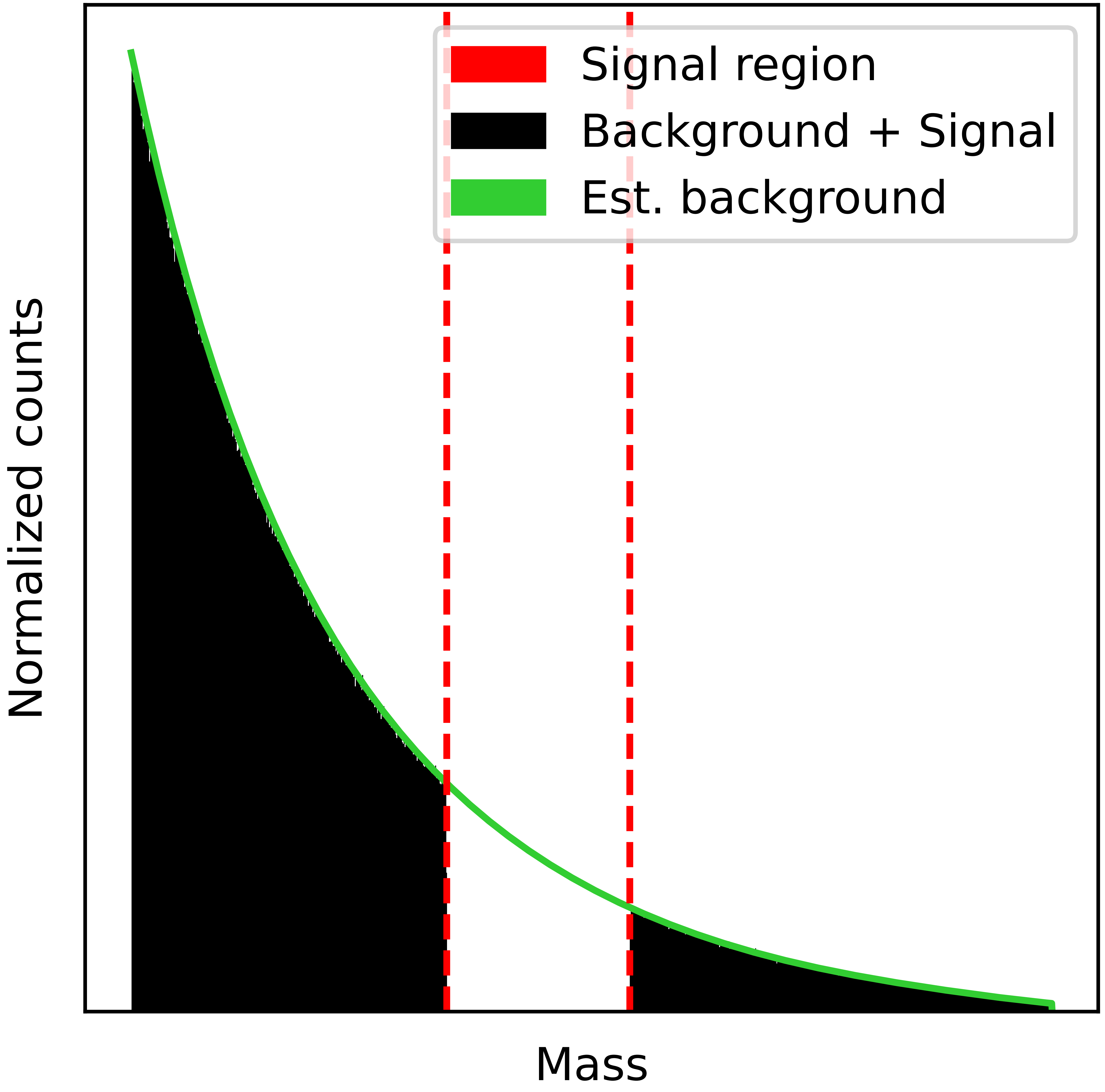
0. Under our assumption, outside the signal region **(SR)**

$$dF(x) = (1-\lambda) \cdot dB(x) \quad \forall \notin SR$$

1. Fit the background without the signal region

$$dB_F^* = \arg\min_{d\tilde{B} \in \mathscr{B}} \text{KL}\left(\frac{dF}{1-F(SR)}, \frac{d\tilde{B}}{1-\tilde{B}(SR)} \right)$$

Approach 3: assume a control region



0. Under our assumption, outside the signal region **(SR)**

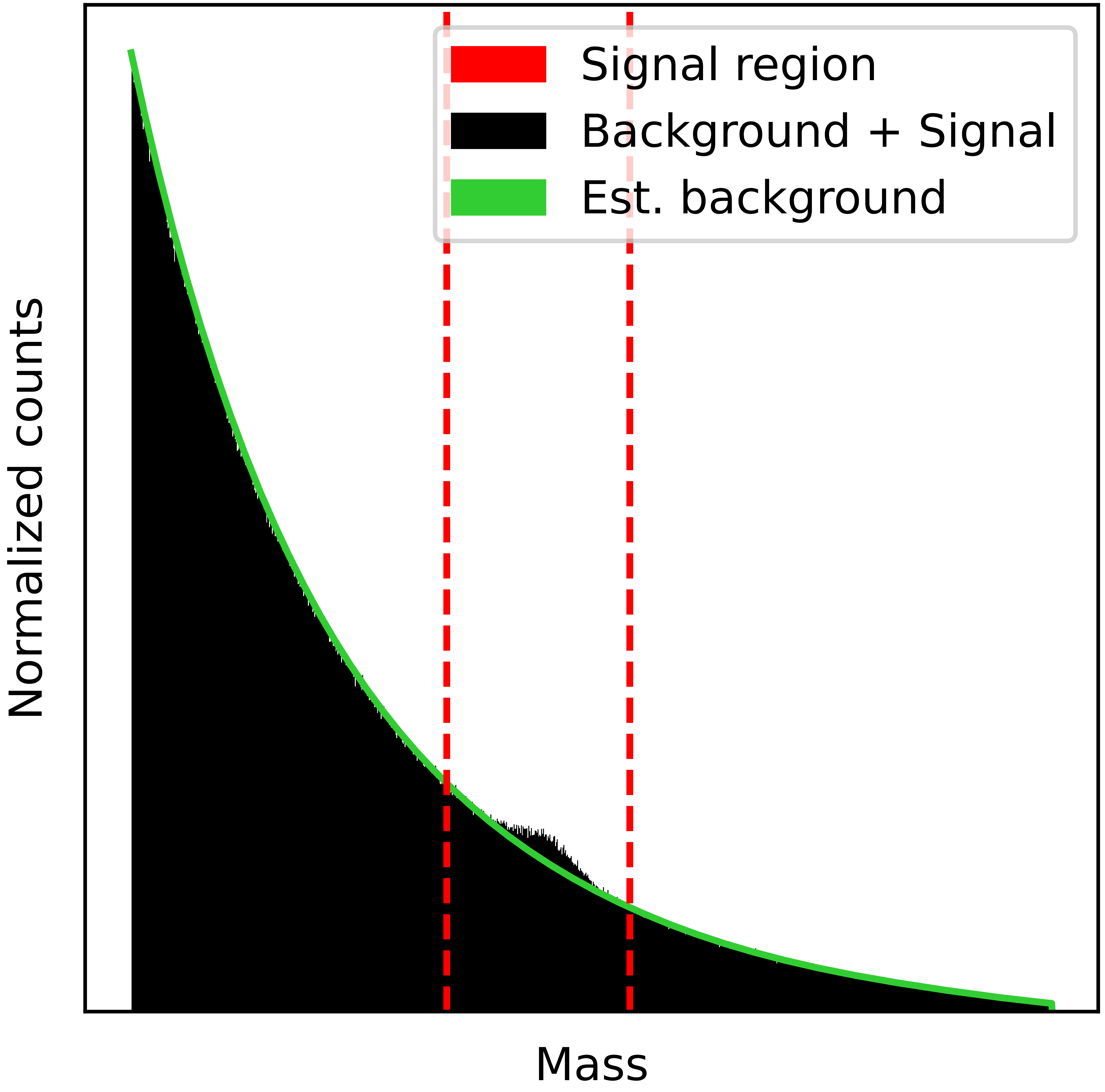
$$dF(x) = (1-\lambda)\cdot dB(x) \quad \forall \notin SR$$

1. Fit the background without the signal region

$$dB_F^* = \arg\min_{d\tilde{B}\in\mathscr{B}} \text{KL}\left(\frac{dF}{1-F(SR)}, \frac{d\tilde{B}}{1-\tilde{B}(SR)}\right)$$

2. Extrapolate the background to the signal region

Approach 3: assume a control region



0. Under our assumption, outside the signal region **(SR)**

$$dF(x) = (1-\lambda)\cdot dB(x) \quad \forall \notin SR$$

1. Fit the background without the signal region

$$dB_F^* = \arg\min_{d\tilde{B}\in\mathscr{B}} \text{KL}\left(\frac{dF}{1-F(SR)}, \frac{d\tilde{B}}{1-\tilde{B}(SR)}\right)$$

2. Extrapolate the background to the signal region

3. Check if the bump is far away from the background

Prob. that $X \sim F$ falls in SR

$$\lambda(F) = 1 - \frac{1-F(SR)}{1-B_F^*(SR)}$$

Prob. that $X \sim B_F^*$ falls in SR

Confidence intervals via functional delta method

Given a distribution, we compute the parameter of interest as a fixed function of it

$$F \;\to\; \lambda(F) = 1 - \frac{1 - F(SR)}{1 - B^*_F(SR)}$$

Given a sample from the distribution, we estimate the distribution and do the same

$$X_1, \ldots, X_n \sim F \;\to\; \hat{F}(A) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in A) \;\to\; \lambda(\hat{F})$$

Functional derivatives tell us how $\lambda(F)$ changes as we move from F to $\hat{F}$

$$\lim_{\epsilon \to \infty} \underbrace{\frac{\lambda(F + \epsilon\ H) - \lambda(F)}{\epsilon}}_{\text{Hadamard derivative}} = \int \underbrace{\psi(x, F)}_{\text{Influence function}} \cdot dH(x) \quad \text{where} \quad \int \psi(x, F)\ dF(x) = 0$$

$$\sqrt{n}\left(\lambda(\hat{F}) - \lambda(F)\right) \asymp \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \psi(X_i, F)\right) \xrightarrow{d} \mathcal{N}(\ 0\ ,\ \sigma^2(F)\ )$$

Confidence intervals via functional delta method

Given a distribution, we compute the parameter of interest as a fixed function of it

$$F \;\to\; \lambda(F) = 1 - \frac{1 - F(SR)}{1 - B^*_F(SR)}$$

Given a sample from the distribution, we estimate the distribution and do the same

$$X_1, \ldots, X_n \sim F \;\to\; \hat{F}(A) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in A) \;\to\; \lambda(\hat{F})$$

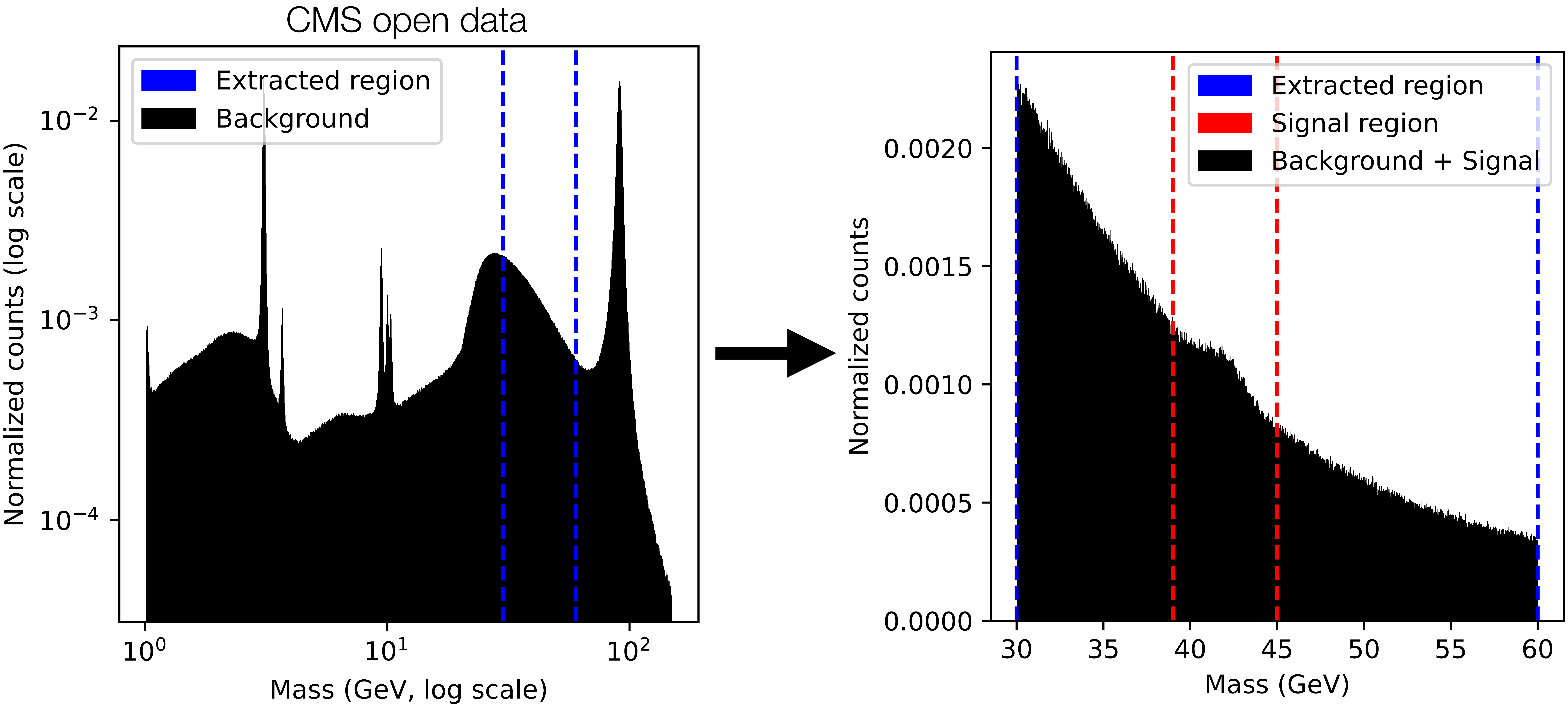
We need to understand how $\lambda(F)$ changes as we move from F to $\hat{F}$

$$\sqrt{n}\left(\lambda(\hat{F}) - \lambda(F)\right) \asymp \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\psi(X_i, F)\right) \xrightarrow{d} \mathcal{N}(\,0\,,\,\sigma^2(F)\,)$$

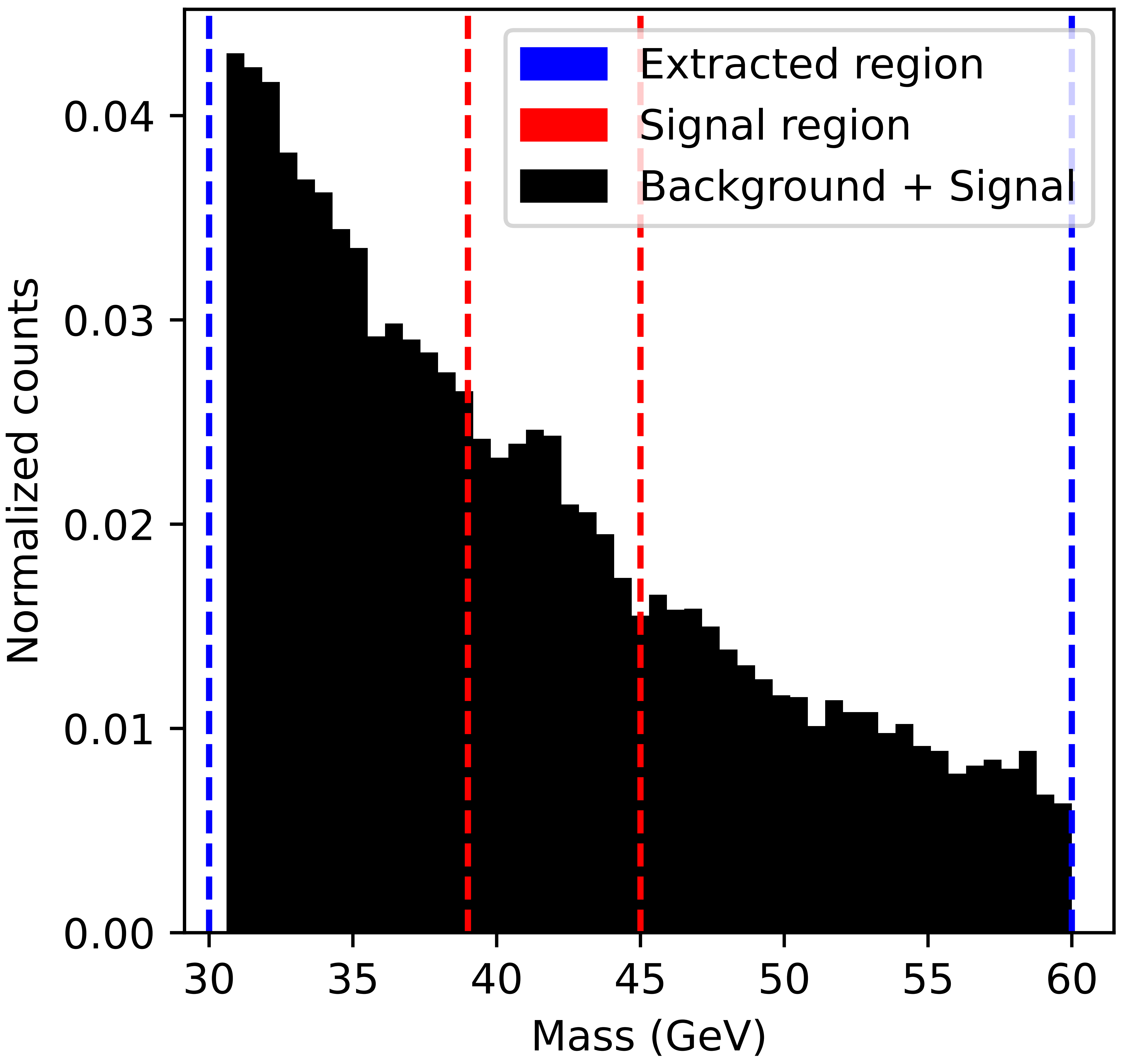
Compute the 95% asymptotically valid confidence interval $\lambda(\hat{F}) \pm 1.96 \cdot \sqrt{\frac{\sigma^2(\hat{F})}{n}}$

Delta method [van der Vaart '98], influence functions [Hines'22], error propagation in unfolding [Adye'11]

Toy problem: we take a background and add a signal

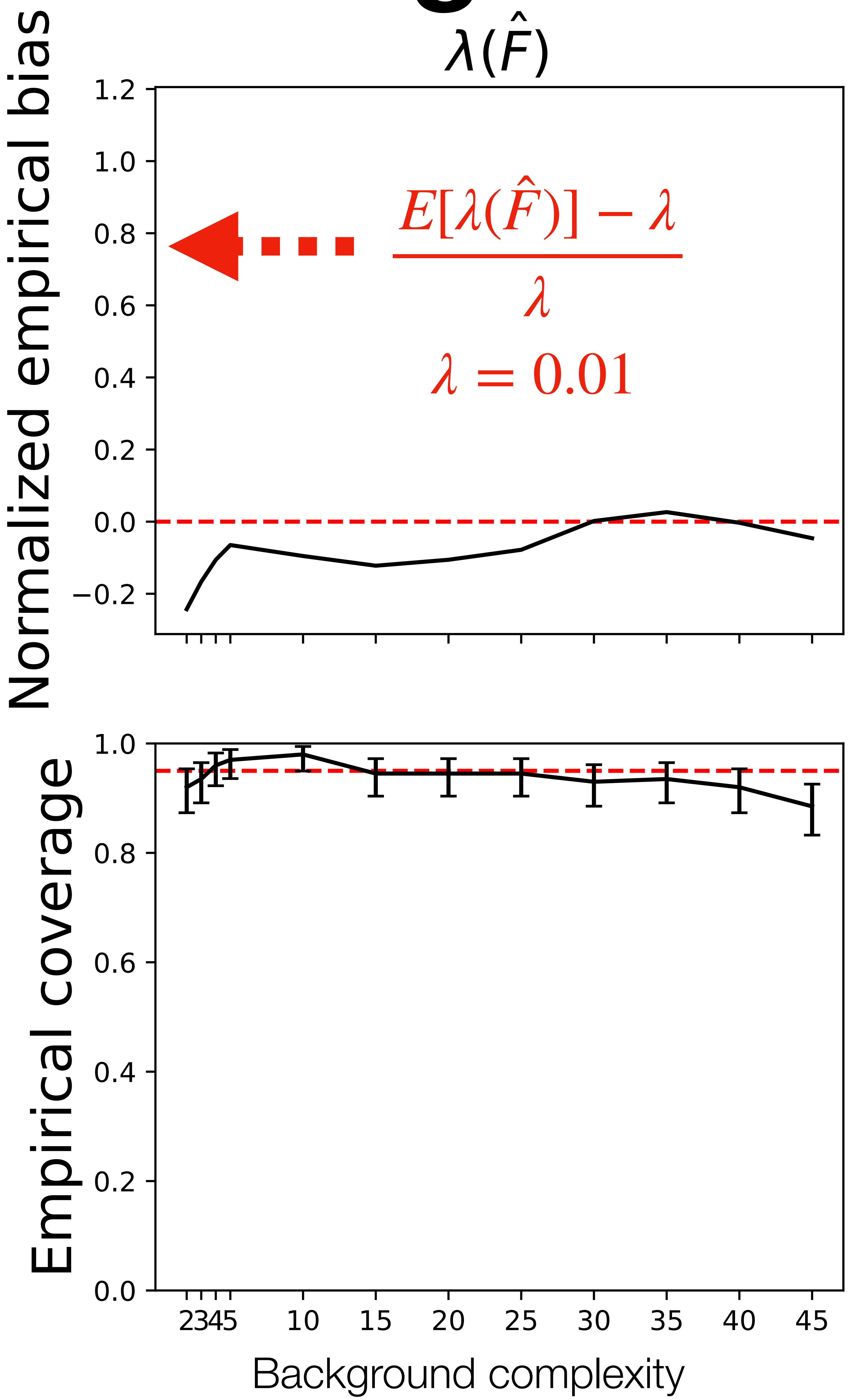


Toy problem: we take a background and add a signal



We produce 200 datasets of 20 000 observations

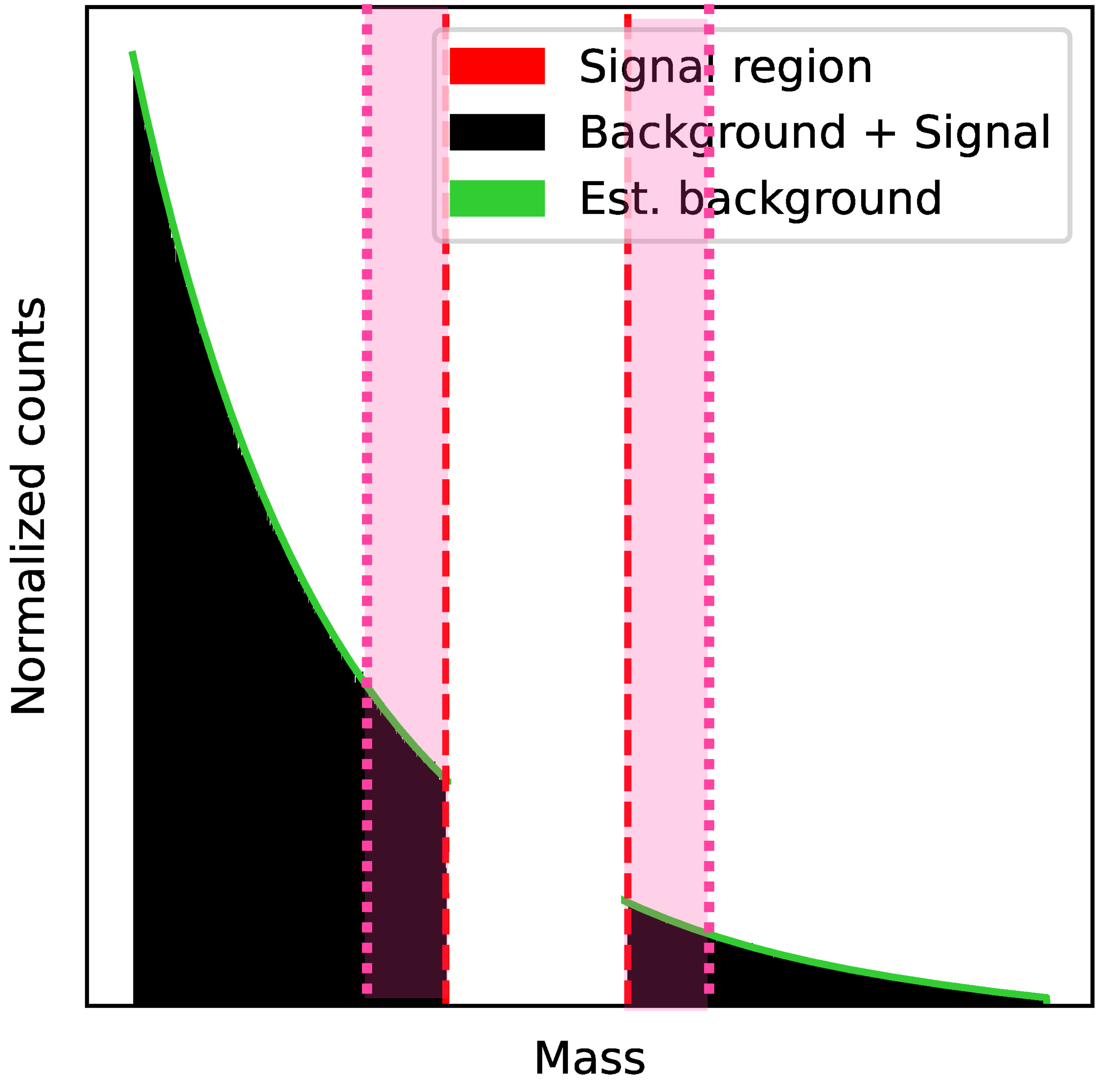
Background model selection



Estimate without signal modelling

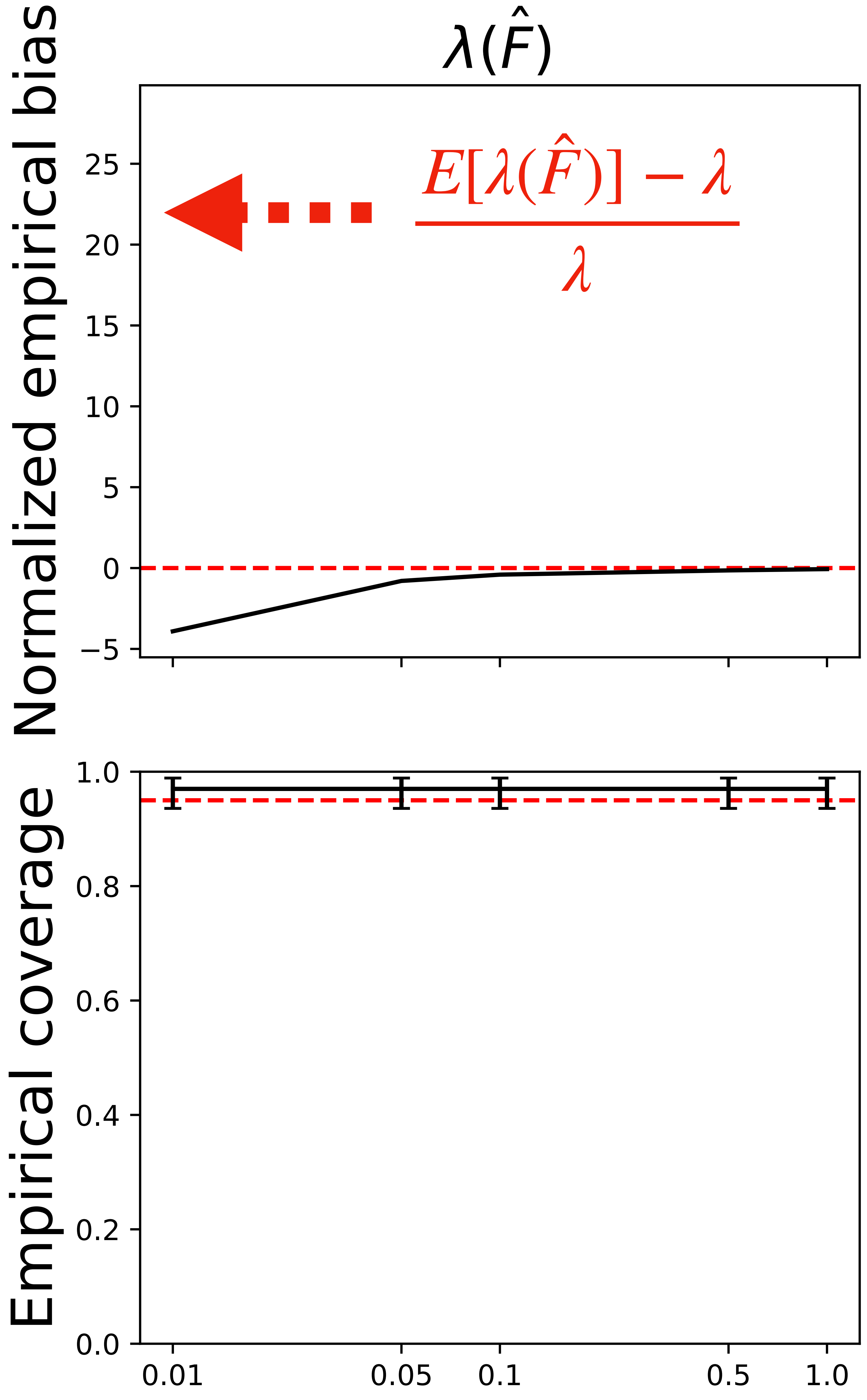
$$\lambda(F) = 1 - \frac{1 - F(SR)}{1 - B_F^*(SR)}$$

Choose the model
with the smallest extrapolation error



donut selection [Barreca'11]

Coverage as the signal strength is varied



Percentage of the data from the signal (log scale)

Estimate without signal modelling

$$\lambda(F) = 1 - \frac{1 - F(SR)}{1 - B_F^*(SR)}$$

Summary + Future work

Assuming: The mixture $X \sim F \;\; : \;\; dF(x) = (1 - \lambda) \cdot dB(x) + \lambda \cdot dS_\theta(x)$

The signal vanishes outside the signal region

The background can be identified from the control region

Re-define target : Rewrite the target parameter $\lambda(F, B)$ as $\lambda(F)$

Construct confidence interval for $\lambda(F)$ using $\lambda(\hat{F})$ and influence functions

Next step: We can reduce the statistical uncertainty of estimating $\lambda(F)$ by using

sample splitting + influence functions

$$X_1, \ldots, X_n \sim F$$

$$X_1, \ldots, X_{n/2} \rightarrow \hat{F}_1$$

$$X_{1+n/2}, \ldots, X_n \rightarrow \hat{F}_2$$

$$\lambda(\hat{F}_1) + \int \psi(x, \hat{F}_1)\; d\hat{F}_2(x)$$

Thanks! Questions?