25.02.25 - 28.02.25

Bring the Noise

Exact inference from noisy simulations in collider physics

arXiv:2502.08157

Christopher Chang

Collider Reinterpretation Forum 2025

UNIVERSITY
OF OSLO

Noisy Monte Carlo simulations result in biased likelihood estimates using common estimators.

Motivation / Notation

Interpretation of BSM Collider searches is often done by generating Monte Carlo events for signals from new physics at colliders.

Events passed through series of cuts to approximate experimental analyses.

Predict a number of expected events, λ, for each signal region in an analysis.

Probability of observing o events given λ expected can be modelled by a **Poisson distribution**.

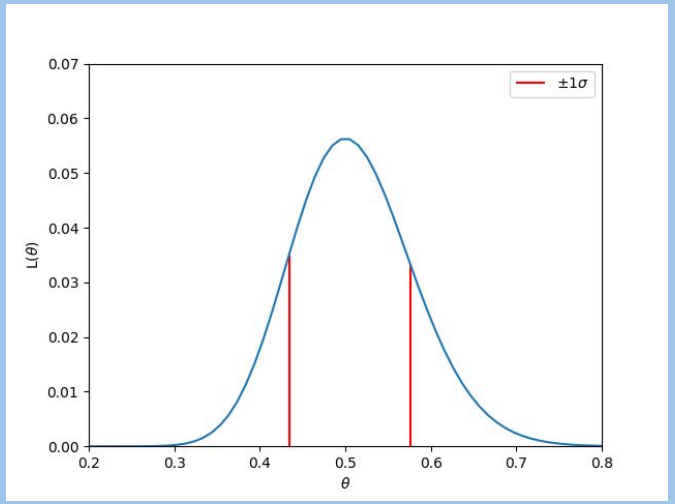
λ has a SM background (b) and a new physics signal (k) component.

Credible regions can be set on new physics parameters by performing Markov-Chain Monte Carlo (MCMC).

Poisson Distribution probability:

$$P(o|\lambda) = \frac{e^{-\lambda}\lambda^{o}}{o!}$$

$$\lambda = b + k$$



Motivation / Notation

MC simulation is noisy estimate: randomness will give a different likelihood each time you calculate it.

The probability of a single signal event ending up in a given bin can be modelled by a **Binomial distribution**.

Given these simulated events, we construct a Maximum Likelihood Estimate of the Poisson likelihood.

The maximum likelihood estimator (MLE) in this case is $\hat{\epsilon}$

The explicit likelihood is shown by $\hat{L}_{\text{MLE}}$

NOTE: hat notation -> estimate from MC simulation

Binomial Distribution probability:

$$P(k|n,p) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

Maximum Likelihood Estimator: $\hat{\epsilon} = \frac{k}{n_{MC}}$

$$\begin{aligned}\hat{L}_{\text{MLE}} &= Po(o|b+\hat{\epsilon}n_{\text{LHC}}) \\ &= \frac{e^{-(b+\hat{\epsilon}n_{\text{LHC}})}(b+\hat{\epsilon}n_{\text{LHC}})^o}{o!}\end{aligned}$$

n_{LHC} Number of events expected at LHC

n_{MC} Number of MC events generated

Why are they biased?

Fixed time experiment: Poisson Distribution

Fixed number of samples: Binomial Distribution

We generate a fixed number of MC samples, and try to model it by a Poisson likelihood.

Example:

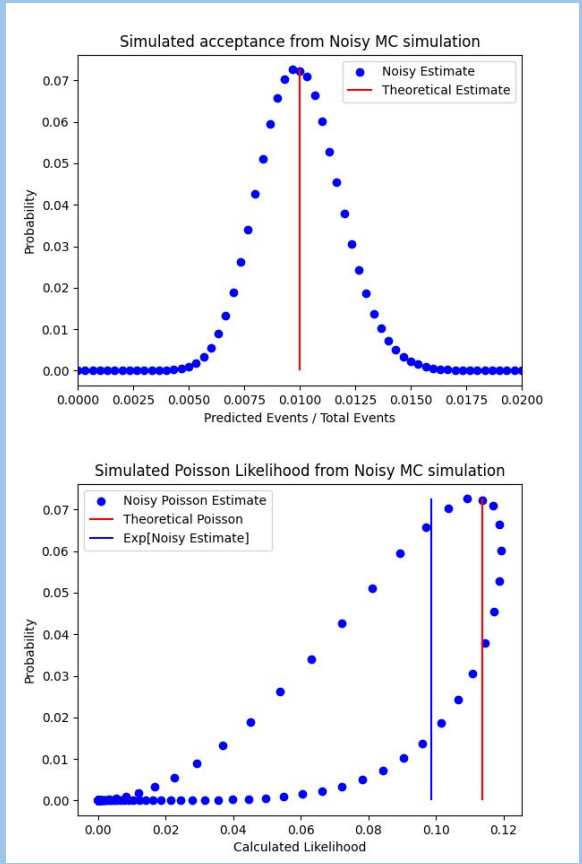
-> true expected acceptance = 0.01

-> **n_{MC}** = 3k, **n_{LHC}** = 1k, **obs** = 11

Expected value of Poisson estimate deviates from our true value from theory -> **biased**

Exp[L] = 0.0988
Exp[L_{theory}] = 0.1137

$$\langle \hat{L}_{MLE} \rangle \neq L$$



Note: Upper/lower curves are from over/under predicting number of events

How can we avoid bias?

Bias reduces to zero as the number of simulated events approaches ∞.

-> but calculations are computationally expensive

In the absence of this, we would like to change our likelihood estimator to one that is not biased

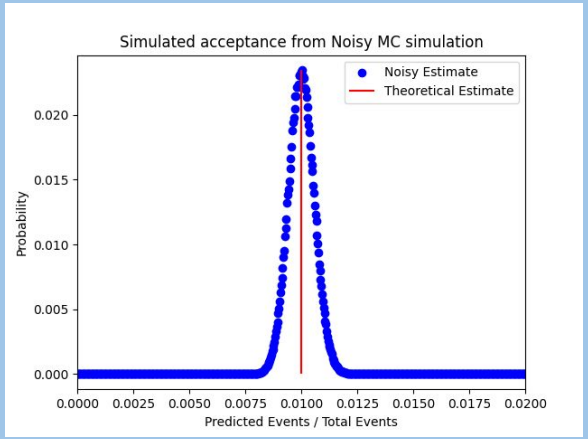
-> **Uniformly Minimum Variance Unbiased Estimator** (UMVUE)

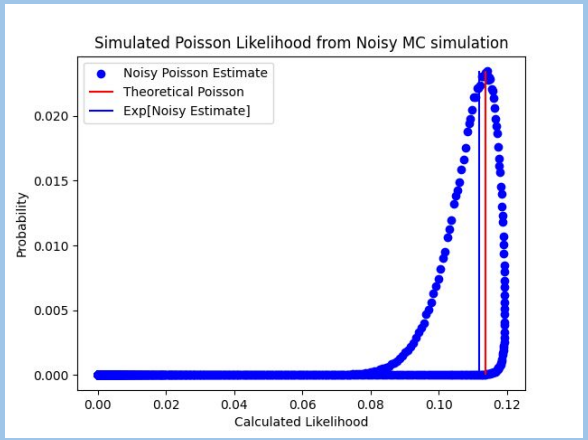
Example (10x as many MC samples):

-> true expected acceptance = 0.01

-> $\mathbf{n_{MC}}$ = 30k $\mathbf{n_{LHC}}$ = 1k, **obs** = 11

Exp[L] = 0.1118
Exp[L_{theory}] = 0.1137





Uniformly Minimum Variance Unbiased Estimator (UMVUE)

Uniformly Minimum Variance: Want an estimator with as small of a variance as possible

Unbiased estimator: expectation value should match the theoretical for a Poisson

Key Point: **the number of MC events simulated is Poisson distributed, as opposed to a fixed number**

Still UMVUE if the likelihood is a product of Poisson likelihoods (even if generated from the same underlying set of MC events).

-> Don't need to generate a new set of MC events for each signal region to be unbiased.

o: observed events, **b**: background events, **k**: signal events

n_{LHC}: number of events expected at LHC

n_{MC}: number of events simulated

$$\hat{L}_{\mathrm{UMVUE}} = \sum_{i=0}^{k} Po(o-i|b)\mathrm{Binom}(i|k,f)$$

$$f = \frac{n_{\mathrm{LHC}}}{n_{\mathrm{MC}}}$$

Uniformly Minimum Variance Unbiased Estimator (UMVUE)

No longer have a fixed grid of possible estimator values.

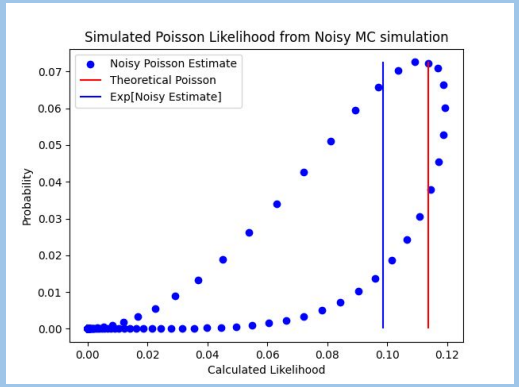
Cost: larger variance

Noise should average out over a MCMC chain

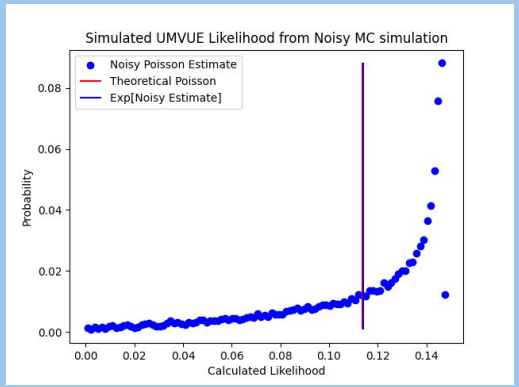
$$\hat{L}_{\mathrm{UMVUE}} = \sum_{i=0}^{k} Po(o-i|b)\mathrm{Binom}(i|k,f)$$

$$f = \frac{n_{\mathrm{LHC}}}{n_{\mathrm{MC}}}$$

MLE



UMVUE (binned)



$\mathrm{Exp}[L]_{\mathrm{UMVUE}} = 0.1137$
$\mathrm{Exp}[L_{\mathrm{theory}}] = 0.1137$

1D Toy model

1D toy model with the selection efficiency (ε) taken as an input parameter.

Perform MCMC.

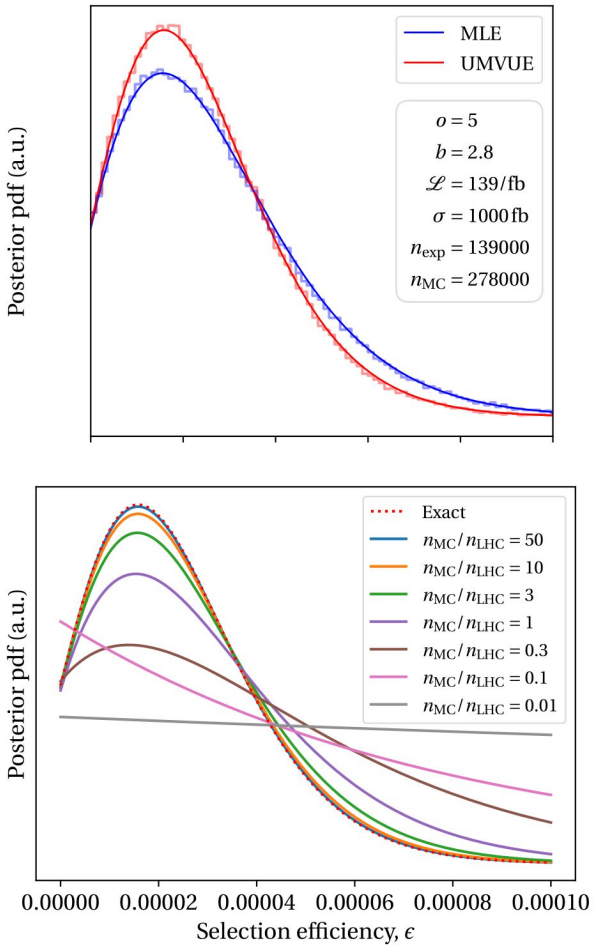
Assume a flat prior on ε, and fix the cross-section.

Top: Fixed number simulated for MLE = 2 x expected from 139 invfb of data

Bottom: Posterior for MLE approaches exact result when ~50x as many MC events simulated.

As opposed to ~2x for UMVUE

Stepped histogram: MCMC results



1D Toy model

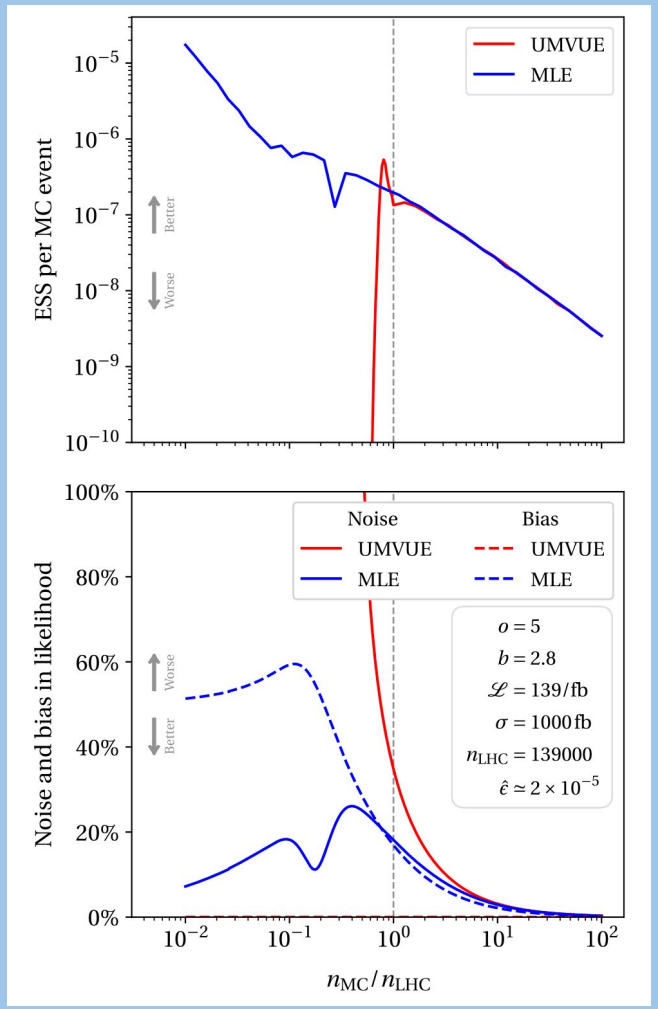
As simulated number of events becomes large, both MLE and UMVUE should become equal.

UMVUE estimator is VERY noisy when n_{MC} is smaller than n_{LHC}.

ESS: Effective sample size (measure of amount of info in event)

ESS for UMVUE (red) behaves very poorly when $n_{MC} < n_{LHC}$ because we downweight negative values.

Negative estimator values are known to occur for noisy estimators, referred to as the 'sign problem' [1].



[1] L. Lin, K.F. Liu and J.H. Sloan, *A Stochastic Monte Carlo algorithm*, Phys. Rev. D **61** (2000) 074505

2D Toy Model

Simplified model based on TChiWZ topology:

Mass-degenerate chargino_1 and Neutralino_2 particles are pair produced and decay exclusively through:

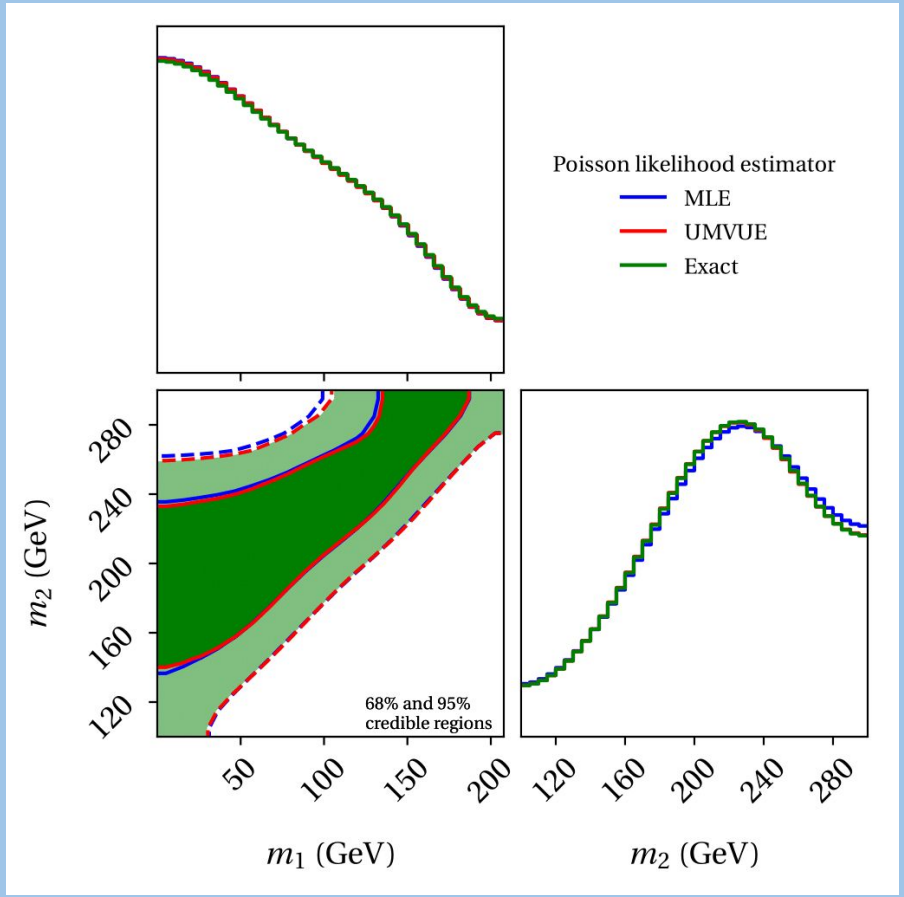
chargino_1 -> W Neutralino_1

Neutralino_2 -> Z Neutralino_1

Cross-sections fixed to 1000 fb

Compute a selection efficiency for the SRWZ_15 signal region provided by ATLAS.

Differences between MLE and UMVUE in the 95% credible regions are slight, but noticeable.



2D Toy model + cross-sections

Simplified model based on TChiWZ topology:

Mass-degenerate chargino_1 and Neutralino_2 particles are pair produced and decay exclusively through:

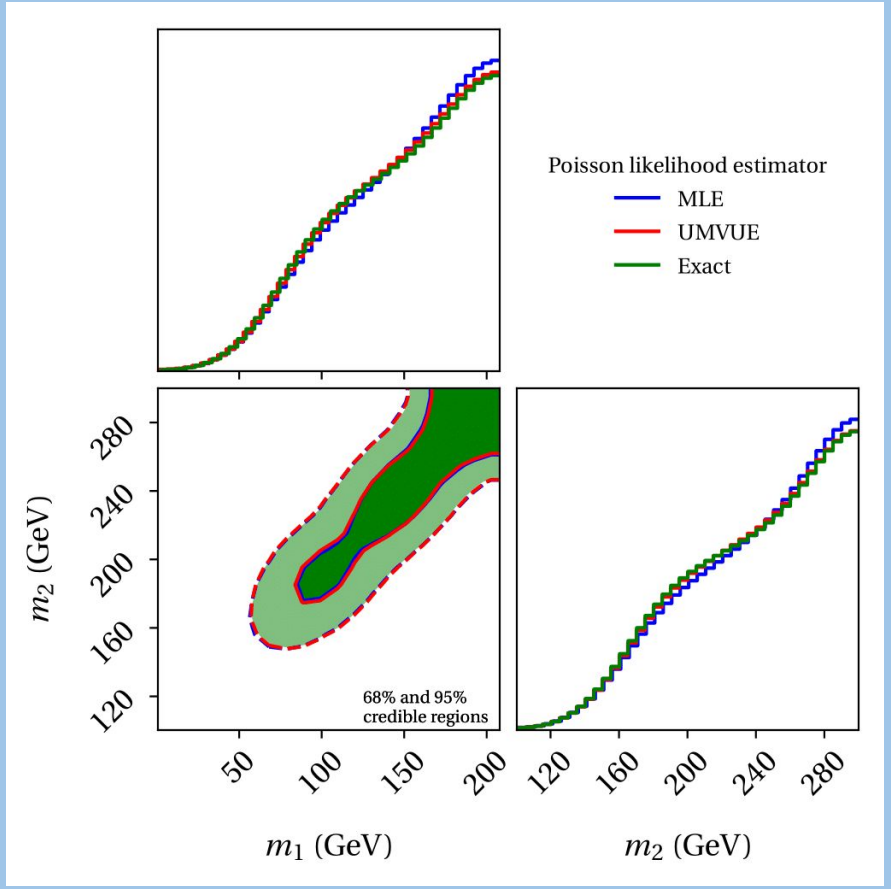
chargino_1 -> W Neutralino_1

Neutralino_2 -> Z Neutralino_1

Cross-sections computed at each point.

Compute a selection efficiency for the SRWZ_15 signal region provided by ATLAS.

Differences between MLE and UMVUE in the 95% credible regions are noticeable at higher masses.



Summary

Key Takeaway: **Using a fixed number of MC events generates biased Poisson likelihoods.**

UMVUE estimators can provide exact inference, at the cost of additional noise

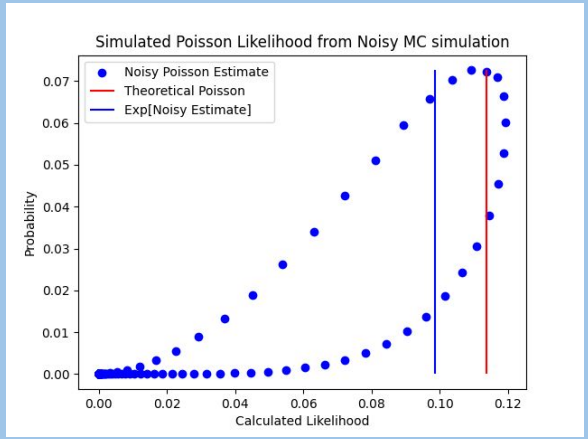
I've demonstrated the use of these in the context of collider inference.

The added noise can be averaged out over a Bayesian model scan.

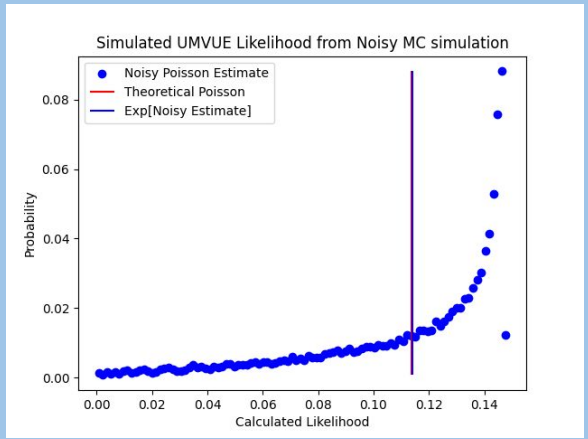
UMVUE estimator inside latest version of ColliderBit (more about ColliderBit in Are's talk this arvo)

$$\hat{L}_{\mathrm{UMVUE}} = \sum_{i=0}^{k} Po(o-i|b)\mathrm{Binom}(i|k,f)$$

MLE



UMVUE (binned)



Backup: many Signal regions + background uncertainty

All my examples have been assuming no background uncertainty, and a single signal region.

Not very useful in real Collider applications.

Procedure is not very different to using regular estimators.

Background uncertainties, and covariance matrices for nuisance parameters, γ_i, are marginalised over.

$$L_{\mathrm{SR}} = \prod_{i=1}^{n_{SR}} L_{\mathrm{UMVUE}}(k_i, b_i, \gamma_i) \frac{1}{\sqrt{det 2\pi\Sigma}} e^{\frac{1}{2}\boldsymbol{\gamma}^T \boldsymbol{\Sigma} \boldsymbol{\gamma}}$$

Backup: Proof the estimator is UMVUE

A linear combination of UMVUE estimators is itself an UMVUE estimator.

1. Write Poisson likelihood as a series expansion in s.
2. Construct the UMVUE by replacing each power of s by the UMVUE of s.
3. Substitute into Poisson likelihood series and rearrange.

This proof follows the approach in [1]

$$\mathrm{Po}(o\,|\,s) = \frac{e^{-s}s^o}{o!} = \sum_{i=0}^{\infty} \frac{(-1)^i}{i!o!}\, s^{o+i}$$

$$\mathrm{UMVUE}\left[s^n\right] = \begin{cases} \left(\frac{n_{\mathrm{LHC}}}{n_{\mathrm{MC}}}\right)^n \frac{k!}{(k-n)!} = f^n \frac{k!}{(k-n)!} & k \geq n \\ 0 & k < n \end{cases}$$

$$\begin{aligned} \mathrm{UMVUE}\left[\mathrm{Po}(o\,|\,s)\right] &= \sum_{i=0}^{k-o} \frac{(-1)^i f^{o+i} k!}{i!o!(k-o-i)!} \\ &= {}^{k}C_o\, f^o \sum_{i=0}^{k-o} {}^{k-o}C_i (-1)^i f^i \\ &= {}^{k}C_o\, f^o (1-f)^{k-o} = \mathrm{Binom}(o\,|\,k,f) \end{aligned}$$

[1] G.J. Glasser, *Minimum Variance Unbiased Estimators for Poisson Probabilities,* Technometrics **4** (1962) 409