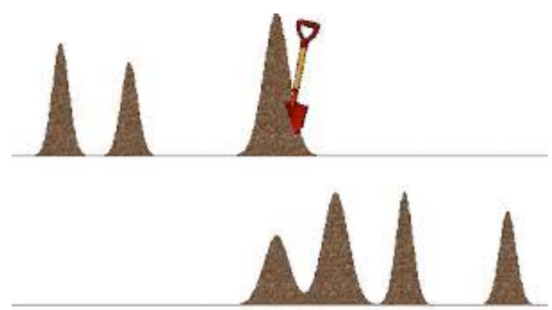
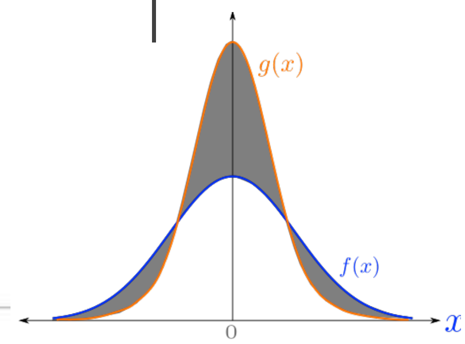
Topics in probability

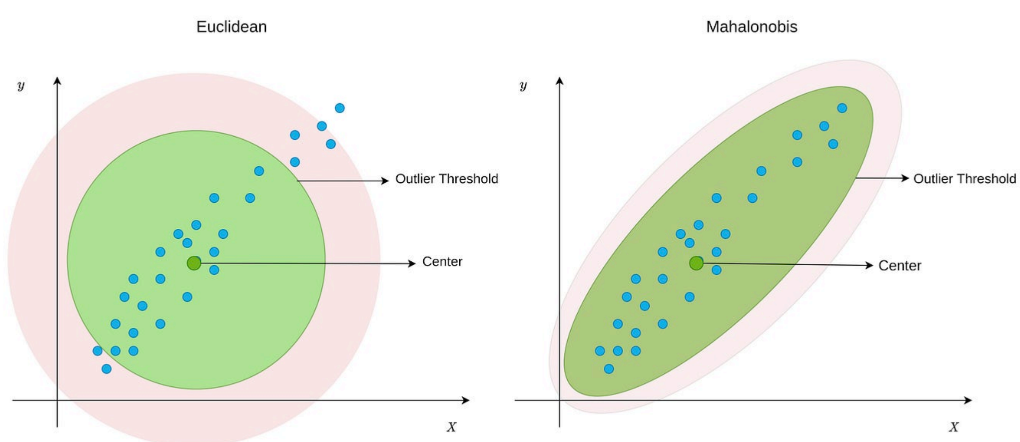
- ~~Axioms~~
- ~~Bayes' theorem~~
- ~~Distributions~~
- Statistical distances
- Information theory
- HEP data

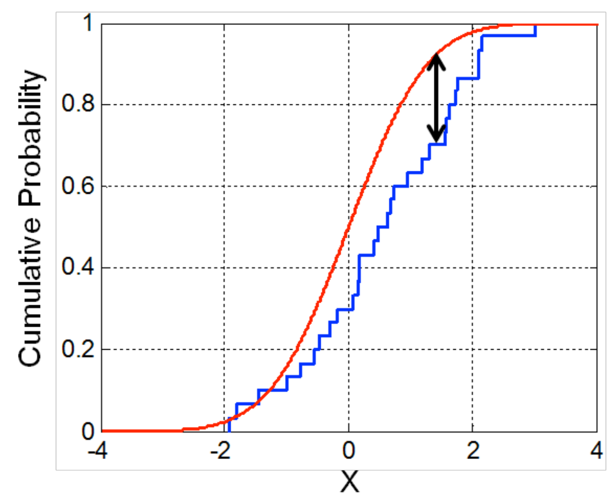
Statistical distances

- Metrics (i.e. triangle)
  - Integral probability metrics: $D_{\mathscr{F}}(P,Q) = \sup_{f\in\mathscr{F}} \left| \int f(x)dP(x) - \int f(x)dQ(x) \right|$
    - Total variation distance
    - Wasserstein (earth-mover's)
    - Kolmogorov metric (KS-test)
  - P(D) through stochastic process analysis
  - Mahalanobis distance









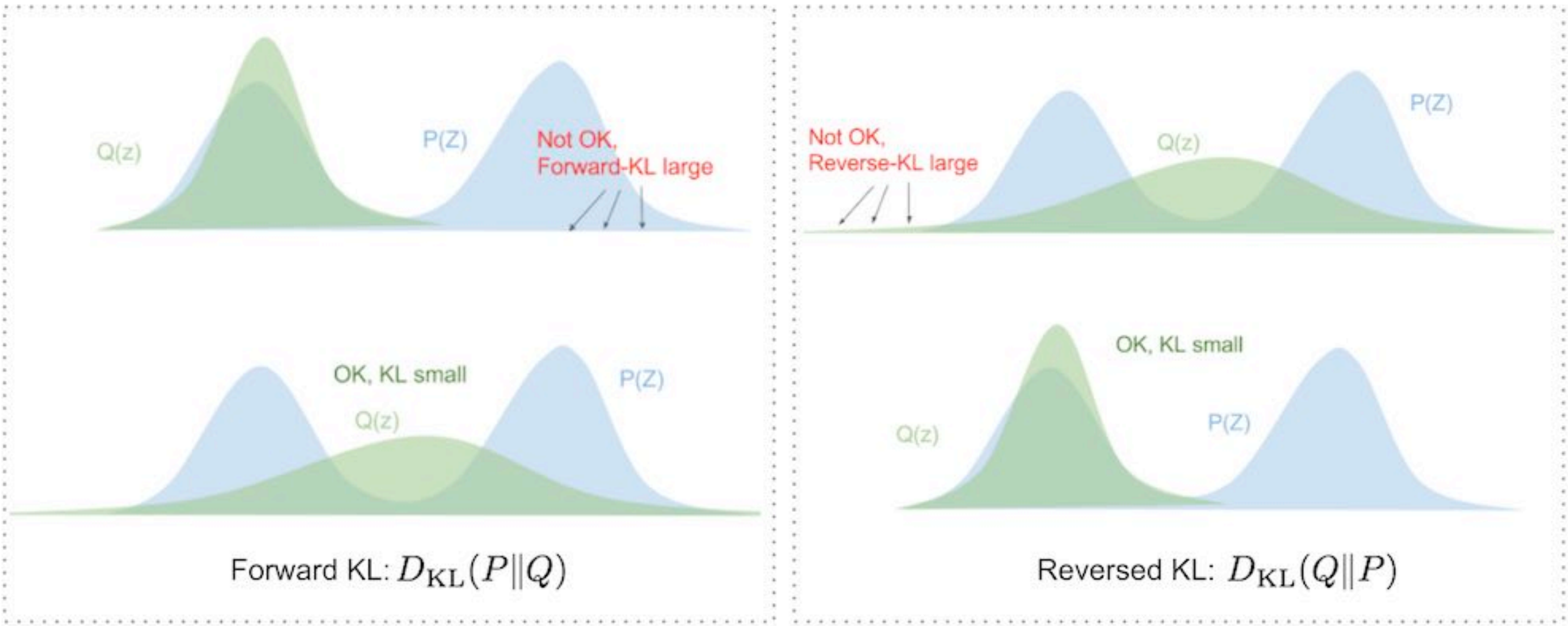
Fermilab

Statistical distances

- Divergences (information geometry)
  - F-divergences: $D_f(P \parallel Q) = \int f\left(\frac{dP}{dQ}\right) dQ$
    - Kullback–Leibler divergence (relative entropy)
      - e.g. in $\mathbb{R}$, $D_{KL}(P \parallel Q) = \int p(x) \ln \frac{p(x)}{q(x)} dx$
    - Jensen–Shannon divergence
      - Symmetrized K-L: $D_{JS}(P \parallel Q) = (D_{KL}(P \parallel M) + D_{KL}(Q \parallel M))/2$ where $M = (P+Q)/2$
      - The square root is then a metric
      - For any distribution $0 \leq D_{JS} \leq \ln 2$
- Mutual information
  - $I(X, Y) = D_{KL}(P(x, y) \parallel P(x) \otimes P(y))$
    - Divergence between joint distribution and direct product of marginals

K-L in pictures

- Kullback–Leibler divergence (relative entropy)
  - e.g. in $\mathbb{R}$, $D_{KL}(P \parallel Q) = \int p(x) \ln \frac{p(x)}{q(x)} dx$



https://hugocisneros.com/notes/kullback_leibler_divergence/

Entropy

- As K-L is relative entropy, Shannon entropy $H(x)$ is K-L w.r.t. the base measure
  - e.g. counting measure for bytes
  - Unit:
    - Bits if $\log_2$ used in K-L
    - Nats if $\ln$ used
- Compression algorithms increase entropy per byte
  - Maximum entropy: 8 bits per byte

[18]:
```python
import gzip
import numpy as np

def entropy(data: bytes):
    data_ints = np.frombuffer(data, dtype="u1")
    _, counts = np.unique(data_ints, return_counts=True)
    probs = counts / counts.sum()
    return -(probs * np.log2(probs)).sum()
```

[19]:
```python
data = b"Hodor hodor hodor hodor hodor hodor. Hodor."
print(entropy(data))
print(entropy(gzip.compress(data)))
```
```
2.548930957111943
4.571374711042188
```

[20]:
```python
data = b"The quick brown fox jumped over the lazy dog"
print(entropy(data))
print(entropy(gzip.compress(data)))
```
```
4.368522527728206
5.2730810667284835
```

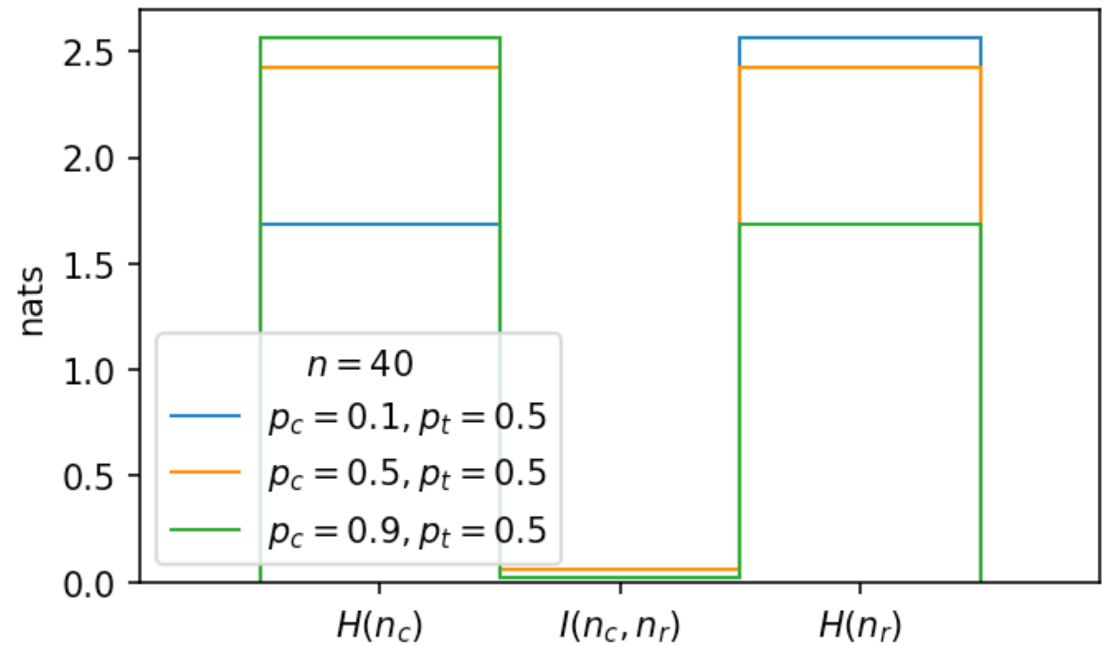
[21]:
```python
data = b"<6?hB:wj9eApZK[F^uw~$4(':"
print(entropy(data))
print(entropy(gzip.compress(data)))
```
```
4.483856189774723
4.922749974675523
```

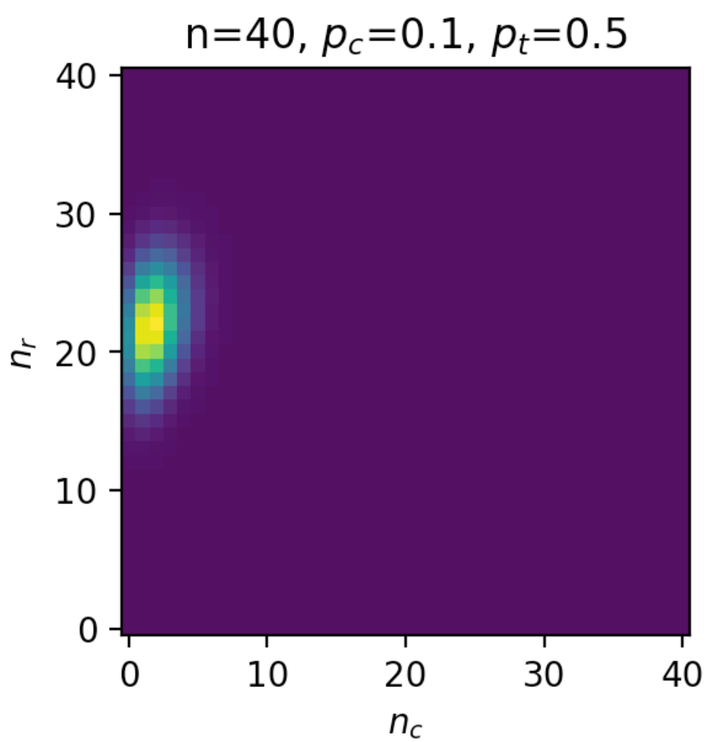
[22]:
```python
data = bytes(range(256))
print(entropy(data))
print(entropy(gzip.compress(data)))
```
```
8.0
7.9269186236261255
```

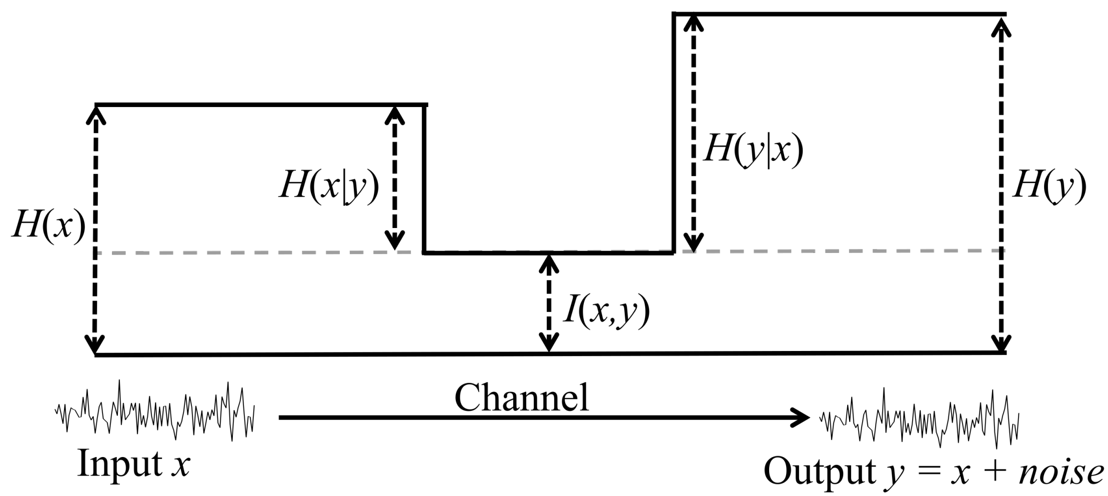

Fermilab

Information theory

- We can measure the mutual information between n_r and n_c in our P(cheat) example



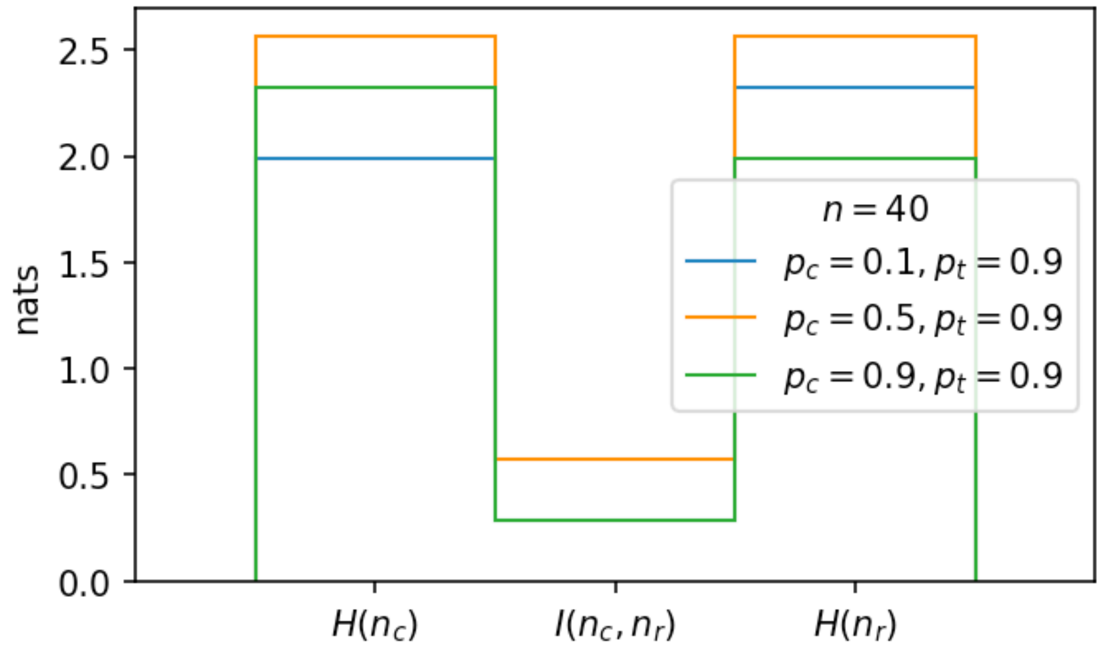


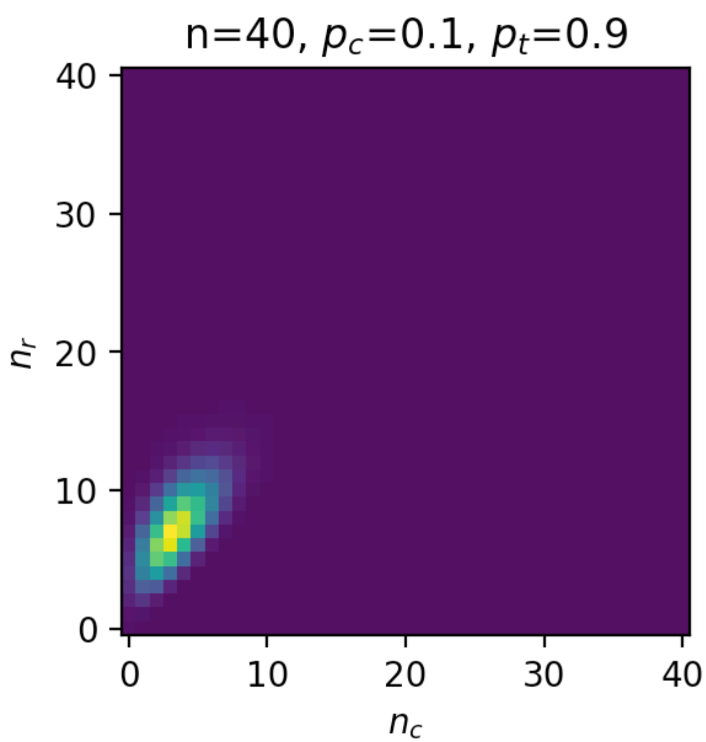


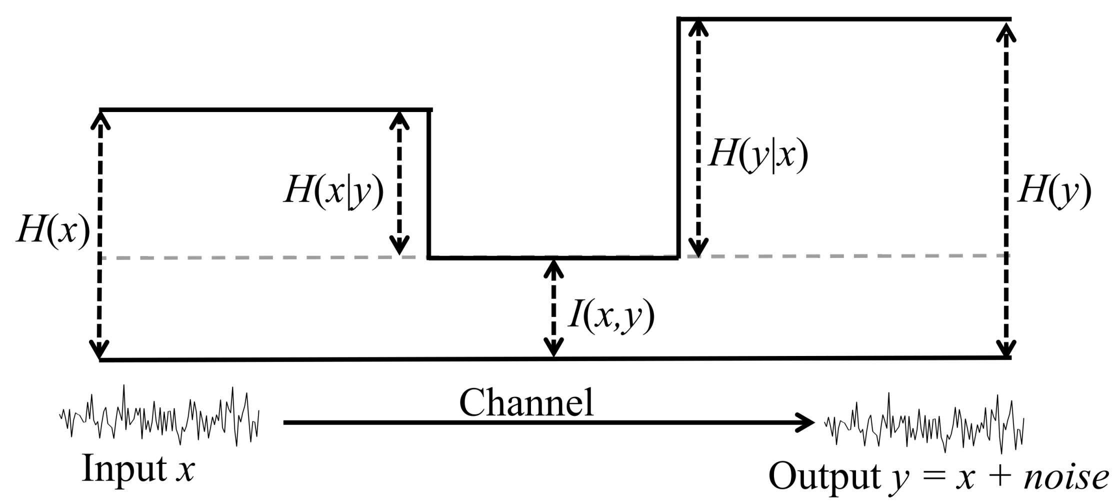
Fermilab

Information theory

- We can measure the mutual information between n_r and n_c in our P(cheat) example
  - If the coin was biased T the channel would have less noise







Fermilab

Probability in HEP

- In HEP we often collect a variable amount of data
  - $n \sim f_P(n; \lambda)$
  - Each event $x_i \sim f(x; \ldots)$ for some distribution f
  - Total probability density of sample $P(\{x_1, \ldots x_n\}) = f_P(n; \lambda) \prod_i^n f(x_i; \ldots) d^n x$
- Often we bin the data: overall PDF is a joint distribution over disjoint regions
  - $P(n_1, \ldots, n_b) = \prod_i^b f_P(n_i; \lambda_i)$
  - This can be shown in the infinitely small bin limit to be equivalent to the above
    - R. Barlow, *Extended maximum likelihood*

Poisson process

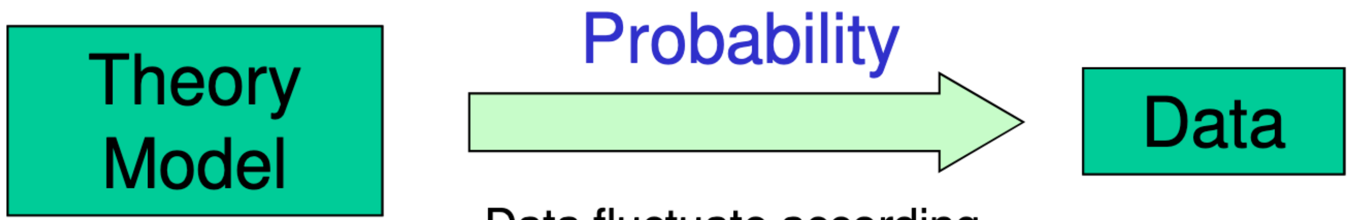
- In CMS, collision events occur at a rate $\lambda(x,t) = L(t)\,\sigma_{pp\to X}(x)\,\epsilon(x,t)$
  - Where (omitting model parameters)
    - $L(t)$ is the instantaneous luminosity
    - $\sigma_{pp\to X}$ is some cross section (differential w.r.t. observables x, e.g. muon 4-momentum)
    - ϵ is our detector acceptance/efficiency (hopefully mild t-dependence!)
- Integrate $\lambda(x,t)$ over some region B ("a bin") to get a Poisson pmf
  - $$\Lambda_i = \int_{B_i} \lambda(x,t)dxdt, \qquad P(N_i\,|\,\Lambda_i) = \frac{\Lambda_i^{N_i} e^{-\Lambda_i}}{N_i!}$$
- This is a Poisson Process
  - Binned model: overall PDF is a joint distribution (product) over disjoint regions
    - $$P(N_1,\ldots,N_b) = \prod_i^b f_P(N_i;\Lambda_i)$$
  - Un-binned model: conditional on N, λ can be interpreted as a PDF (integrating t)
    - $$P(\{x_1,\ldots x_n\}) = f_P(N;\Lambda)\prod_i^N \lambda(x_i)dx_i$$

Inference

Outline: inference

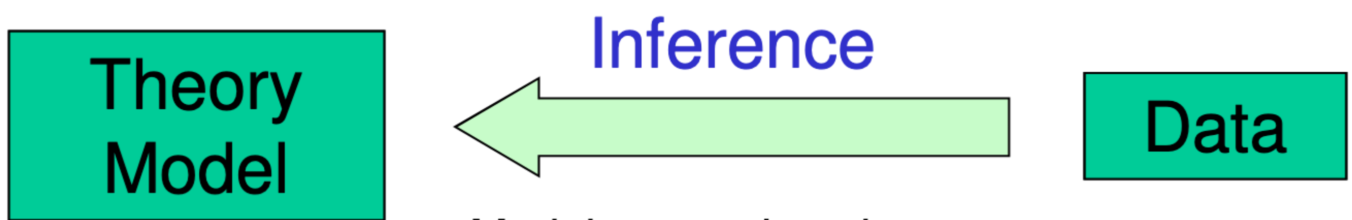
- Bayesian inference
- Maximum likelihood point estimation

Inference



Data fluctuate according to process randomness

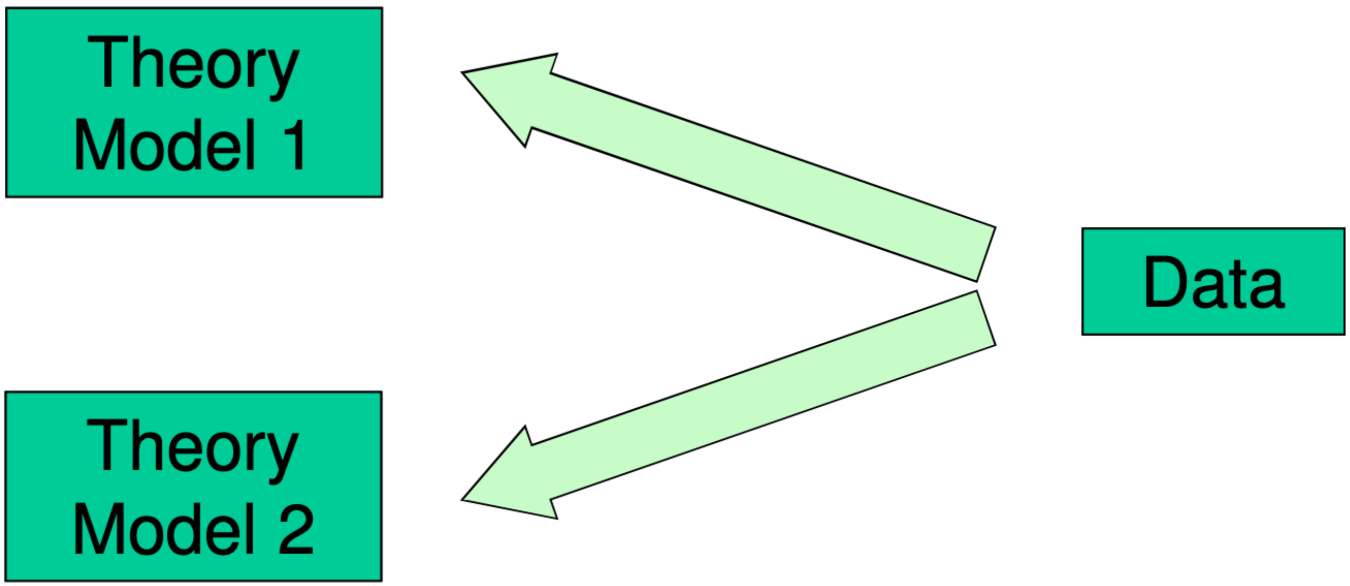
$$x_{obs} \sim P(x \,|\, \theta)$$



Model uncertainty due to fluctuations of the data sample

$$P(\theta \,|\, x_{obs})??$$

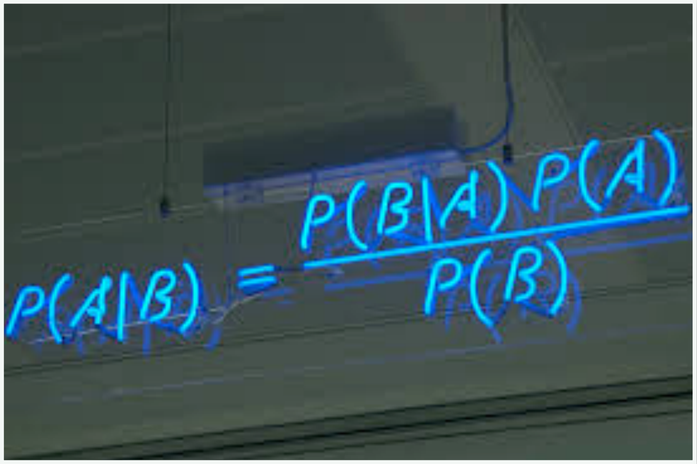
Hypothesis tests



Which hypothesis is the most consistent with the experimental data?

Bayes' theorem revisited

- Consider a joint probability space $P(x, \theta)$ on the space (X, Θ)
  - Can condition on θ, i.e. $P(x, \theta) = P(x \mid \theta)P(\theta)$
- Take observation $x \sim P(x \mid \theta_t)$ drawn for fixed *but unknown* $\theta_t \in \Theta$
- We can define $P(\theta \mid x) = \dfrac{P(x \mid \theta)P(\theta)}{P(x)}$
  - To infer distribution of θ
  - We will never know θ_t with absolute certainty (is it even in Θ?)
- The terms have names:
  - $P(\theta \mid x)$ is the *posterior*
  - $P(x \mid \theta)$ is the *likelihood*
  - $P(\theta)$ is the *prior*
  - $P(x) = \displaystyle\int P(x \mid \theta) dP(\theta)$ is the *evidence*



Priors

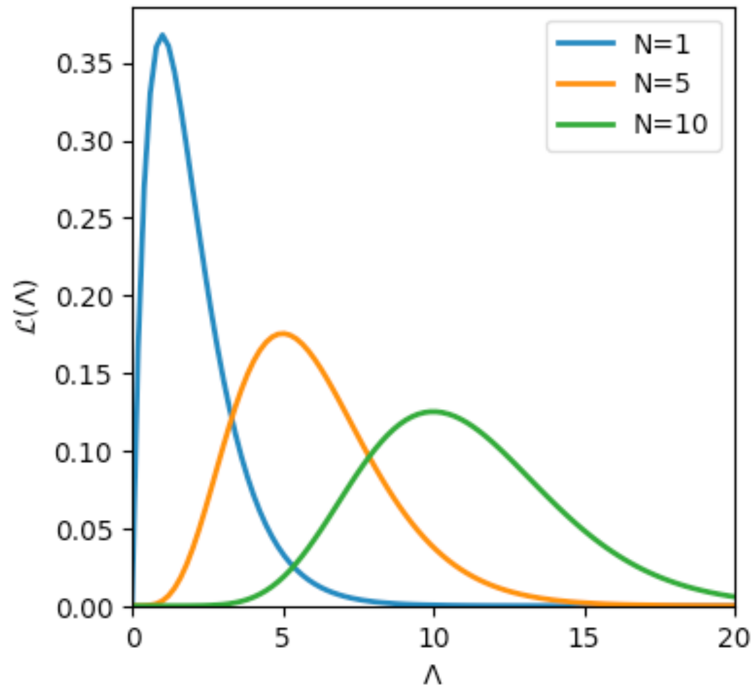
- What is our prior $P(\theta)$? We have a few options
- Subjective Bayesian - whatever *you* feel is a good prior
  - Probably challenging for science
  - *Everyone's a Bayesian in their head*
- Objective Bayesian - a fixed recipe, given the likelihood
  - Jeffrey's prior: $\pi(\theta) \propto \sqrt{|I(\theta)|}$ where $I(\theta)$ is the Fisher information
  - Maximum entropy prior: maximize $H[\theta \sim \pi(\theta)]$
    - Uniform, or exponential family if constrained moments
- Conjugate prior: the posterior is in the same function space
  - For exponential-family likelihoods we are guaranteed a conjugate prior
  - Example: $f_{Bi}(k; n, p)$ with known n, the conjugate is

$$f_{Beta}(p; \alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

    - The posterior parameters are $\alpha' = \alpha + k, \beta' = \beta + n - k$

Maximum likelihood point estimate

- For a fixed *observation*, can define likelihood $\mathscr{L}(\theta) = P(x_{obs}; \theta)$
  - This is a function of θ, but ***not** a probability density*
- $\hat{\theta}$ that maximizes this function is the *maximum likelihood estimate* (MLE)
  - $\hat{\theta} = \text{arg max}_\theta[\mathscr{L}(\theta)] = \text{arg min}_\theta[-\ln \mathscr{L}(\theta)]$
  - This is a random variable
  - We usually minimize the negative log-likelihood (NLL) numerically
    - The core job of MINUIT's MIGRAD routine
- Poisson example: $\hat{\Lambda} = N$



Fermilab

Maximum likelihood trivia

- Absolute value of $\mathscr{L}(\hat{\theta})$ is usually not meaningful
- The MLE has good limiting properties as sample size→∞
  - Consistent: sequence of MLEs converges to true value
  - Efficient: variance of MLE saturates the Cramér–Rao lower bound
    - Asymptotic variance of unbiased estimator at least $I^{-1}(\theta)$
  - Asymptotically unbiased
    - Bias can exist for finite samples, can be corrected (with increase in variance)
- The likelihood (and it's maximum) is invariant under change of variables
  - Again, it is not a PDF!

- Ok, but this is just a *point*. Can frequentists say more without a prior?
  - Likelihood ratios let us make relative statements, **but** the statements are always of the form "assuming a value of θ, would x_{obs} be a likely outcome?" This is not $P(\theta)$
  - Hypothesis tests and frequentist confidence intervals/sets
    - Next time

Fisher information

$$E_{x\sim f(x)}[g(x)] = \int g(x)f(x)dx$$

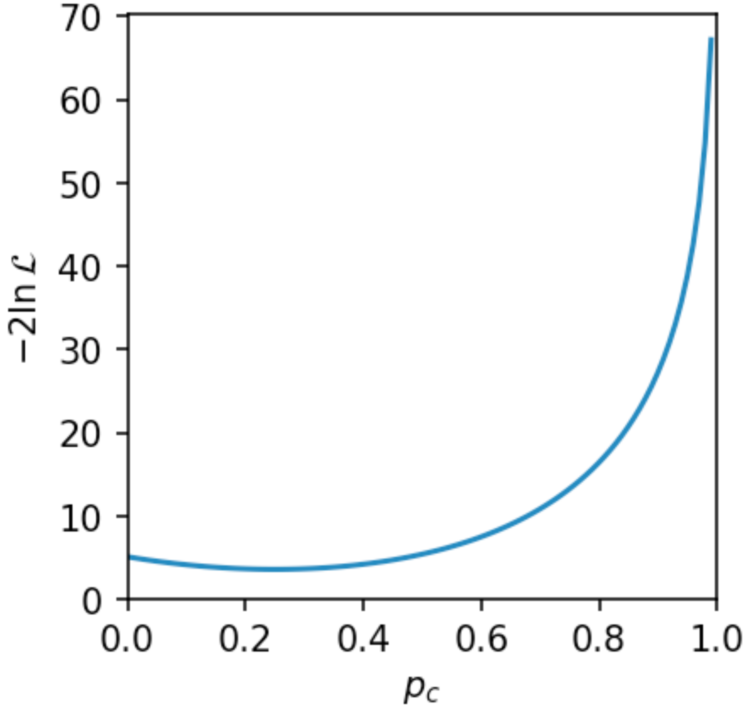
- This is defined as the expectation
  - $I(\theta) = E_{x\sim f(x;\theta)}\left[\left(\frac{\partial}{\partial\theta}\ln f(x;\theta)\right)^2\right]$
  - i.e. the variance of the *score*
  - Under typical conditions, $I(\theta) = -E_{x\sim f(x;\theta)}\left[\frac{\partial^2 \ln \mathscr{L}(\theta)}{\partial^2\theta}\right]$
- Empirical Fisher information is the Hessian of $-\ln\mathscr{L}$
  - Let's expand this function around the MLE $\hat{\theta}$:
  - $-\ln\mathscr{L}(\theta) \approx -\ln\mathscr{L}(\hat{\theta}) + \frac{\partial(-\ln\mathscr{L})}{\partial\theta}\Big|_{\hat{\theta}} \cdot (\theta-\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})^\top I(\hat{\theta})(\theta-\hat{\theta}) + \cdots$
  - $\approx \mathscr{L}_0 + f_N(\theta;\hat{\theta}, I^{-1}(\hat{\theta}))$
  - Looks like a χ^2 up to a factor 2! This is why we often plot $-2\Delta\ln\mathscr{L}$
    - Will revisit with Wilk's theorem later
  - MINUIT HESSE

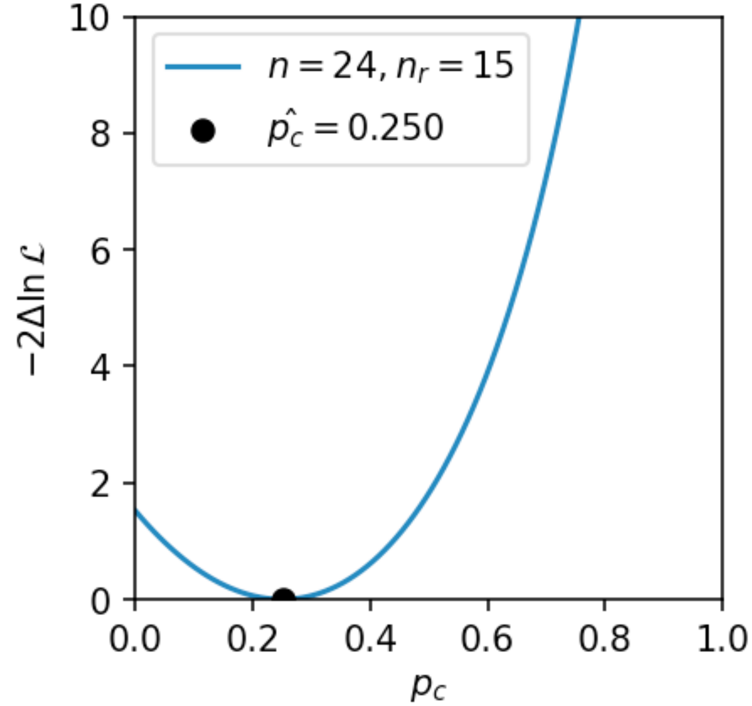
Inferring P(cheat)

- Setup reminder: flip a fair coin without showing anyone the result
  - If heads (H), raise your hand (✋)
  - If tails (T): raise your hand ✋ *only if* you have ever cheated on your homework
- We have a model describing probability of observing n_r given n, p_c
  - $$f(n_r; n, p_c) = \sum_{n_t=0}^{n} f_{Bi}(n_t; n, 1/2) f_{Bi}(n_r + n_t - n; n_t, p_c)$$
- We *observed* $n_r = 15$ and *know* $n = 24$. What can we *infer* about p_c?
  - Frequentists: determine how *likely* n_r is given a *hypothesized* but *fixed* p_c
  - Bayesians: promote p_c to a random variate and use Bayes' theorem
    - Need a measure on the space $p_c \in [0,1]$, call it $\pi(p_c)$
    - The *joint* probability is then $f(n_r, p_c; n) = f(n_r \mid p_c; n)\pi(p_c)$
- Use $f(n_r \mid p_c)$ vs. $f(n_r; p_c)$ to distinguish random variates from parameters
  - I can barely keep this up

Inferring P(cheat) as a frequentist

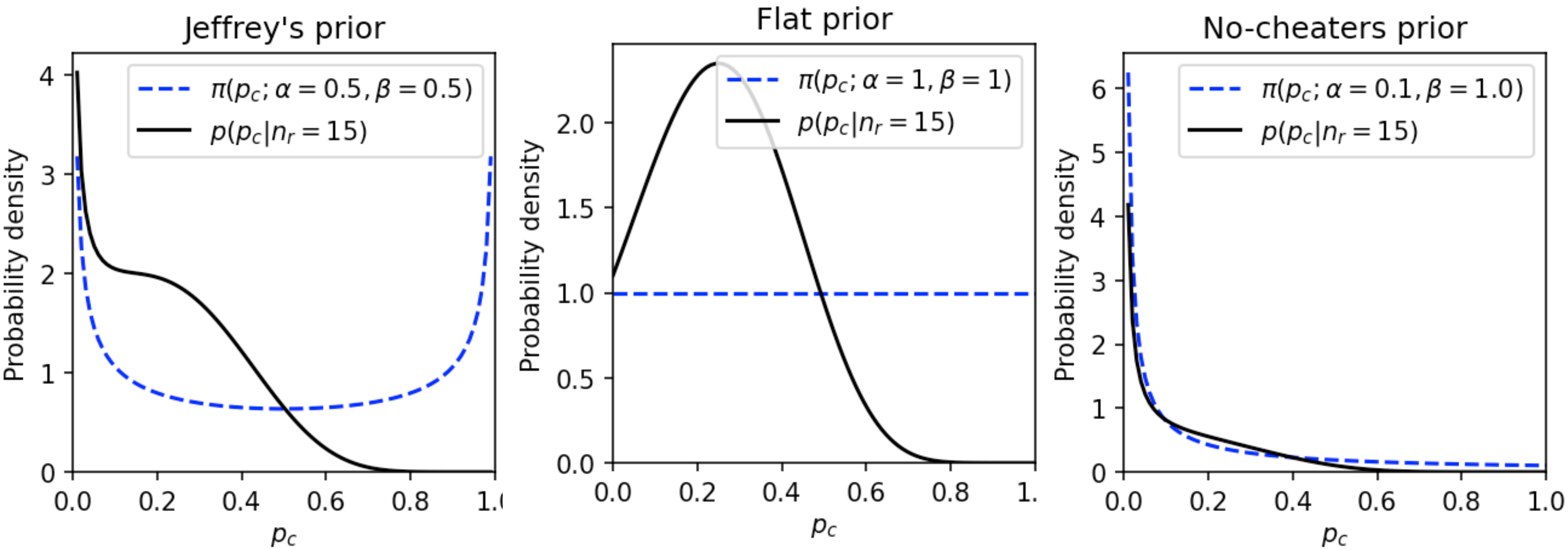
- Scan $-2\ln\mathscr{L}(p_c) = -2\ln f(n_r; n, p_c)$
- Find minimum
- ???
- Profit





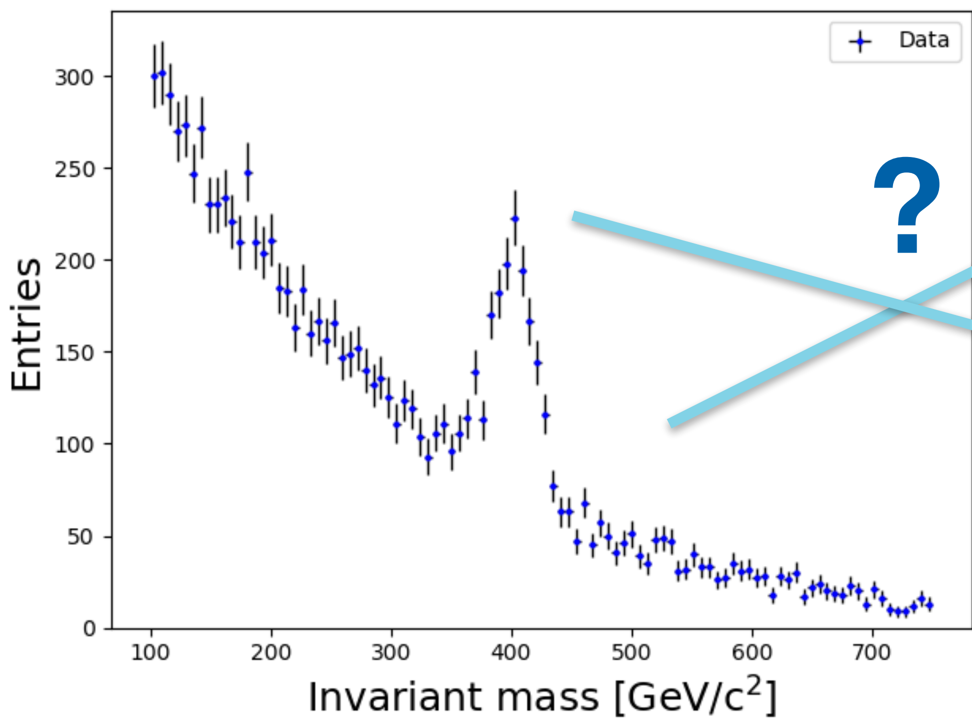
Inferring P(cheat) as a Bayesian

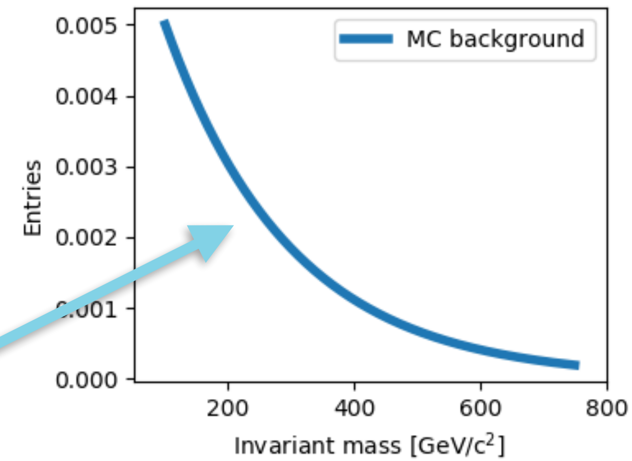
- What is our prior $\pi(p_c)$? We have a few options
  - Likelihood is not exponential-family, hard to find conjugate
  - Try Binomial conjugate $f_{Beta}(p_c; \alpha, \beta)$
    - Jeffrey's* prior: $\alpha = \beta = 1/2$
    - Maximum entropy prior: $\alpha = \beta = 1$
    - "They're good kids" prior: $\alpha \ll 1, \beta = 1$

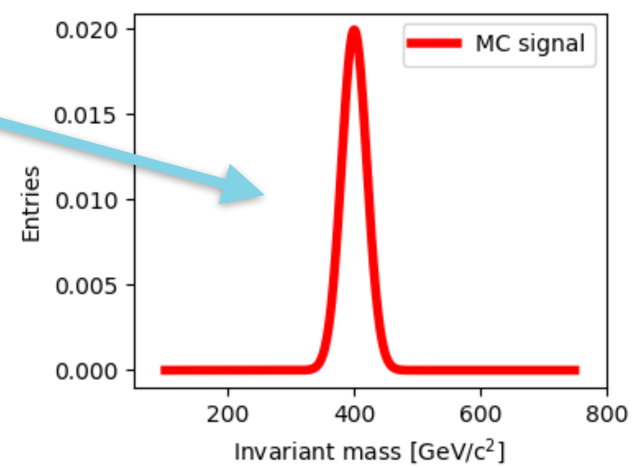


Inference example in HEP

- Given this data and a model for signal and background, I might infer:
  - The amount of signal present (a *parameter of interest*, or POI)
  - The functional form of the background, if a-priori unknown
    - Parameterized by *nuisance parameters* (more later)
  - Which hypothesis (S+B or B-only) is more consistent







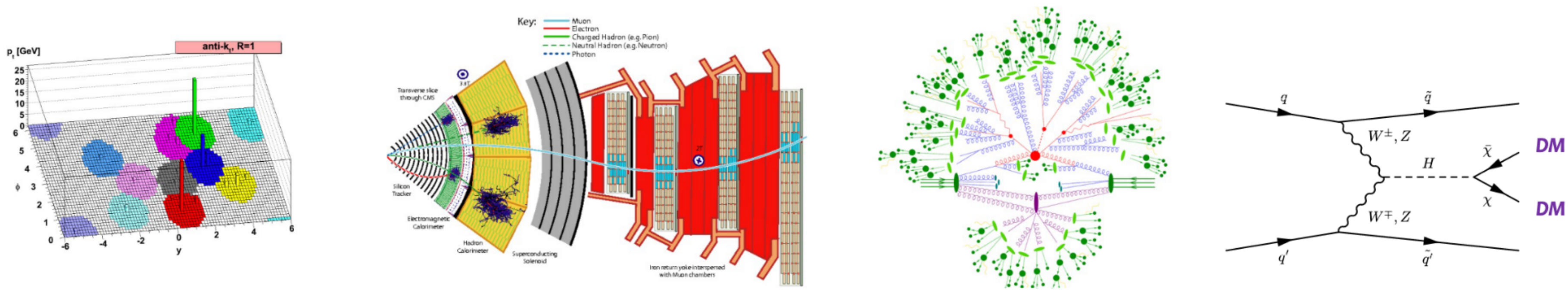
Fermilab

Inference for collider simulation

- The whole picture is more complex
  - We often cannot compute $P(x \mid \theta)$, but we can efficiently sample it
    - ➜ surrogate model using Monte Carlo (MC) estimates of bin yields

Observables	Detector interactions	Shower splittings	Parton-level momenta	Theory parameters

$$x \longleftarrow z_d \longleftarrow z_s \longleftarrow z_p \longleftarrow \theta$$



$$\underbrace{p(x|\theta)}_{\text{Features}} = \int \mathrm{d}z_d \int \mathrm{d}z_s \int \mathrm{d}z_p \; p(x|z_d) \; \underbrace{p(z_d|z_s) \quad p(z_s|z_p) \quad p(z_p|\theta)}_{\text{Latent Variables}}$$

Model Parameters

diagram: K. Cranmer

Templates

- We often build models via template histograms derived from MC
  - Typically to infer signal strength μ = normalization of signal template