
Image credit: Marguerite Tonjes

Statistics

Nick Smith (Fermilab/CMS)
COFI 2023

Goal of these lectures

Evolved from a CMS tutorial given last February

There, the goal was to give plausible familiarity with sentences like:

> An observed (expected) upper limit is placed on the signal strength μ, using the profile likelihood ratio test statistic, following the CL_s criterion, under asymptotic assumptions, and found to be …"

Expanded to talk about:

- Alternatives to the standard CMS methods
- ML-related topics

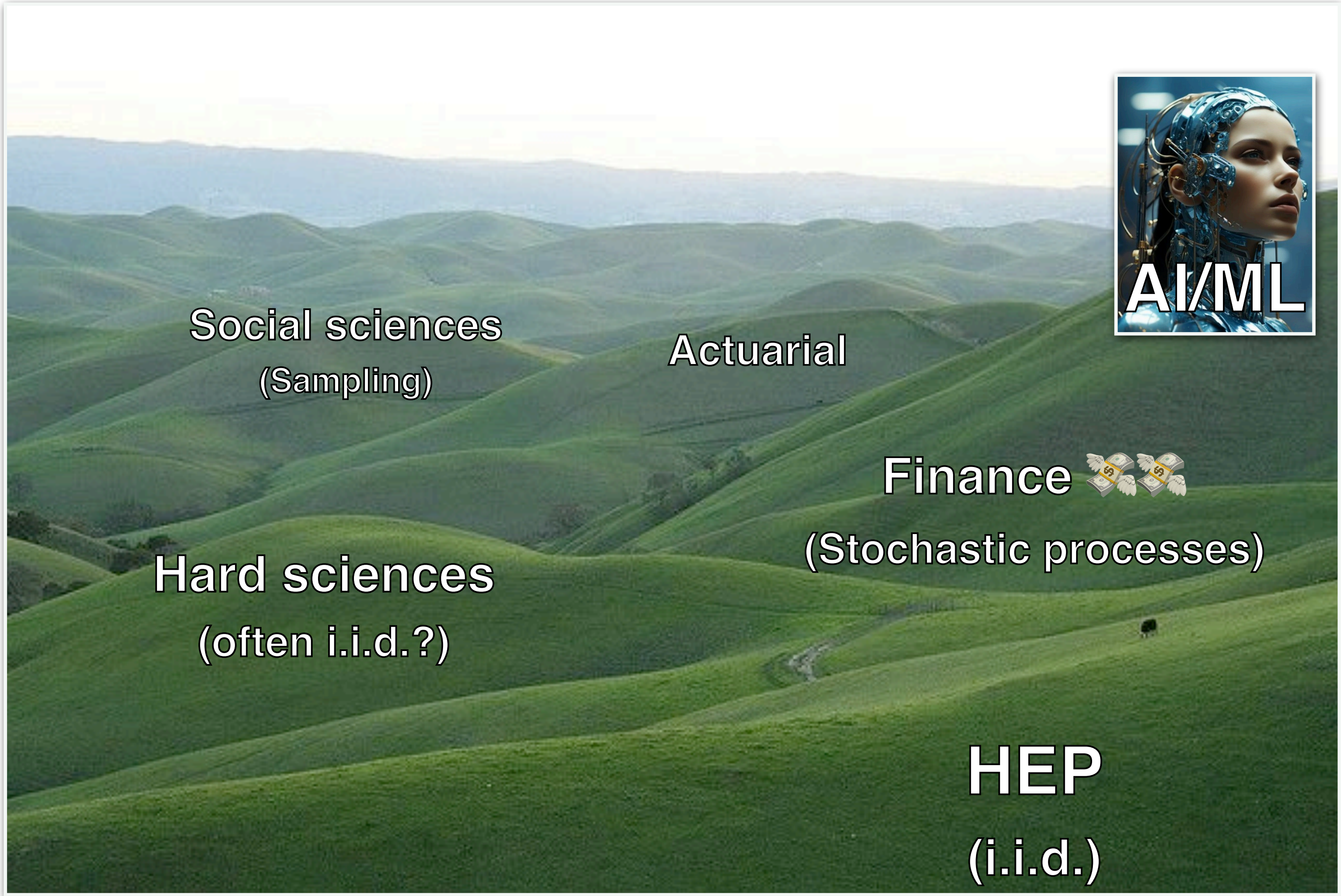
Resources:

- PDG: probability, statistics
- Past lectures to HEP audiences
  - R. Cousins, N. Wardle, K. Cranmer, J. Duarte
- Wikipedia, Youtube, (3b1b, Simons TV, …)

The current plan

- Probability
- Inference
- Intervals
- Uncertainties

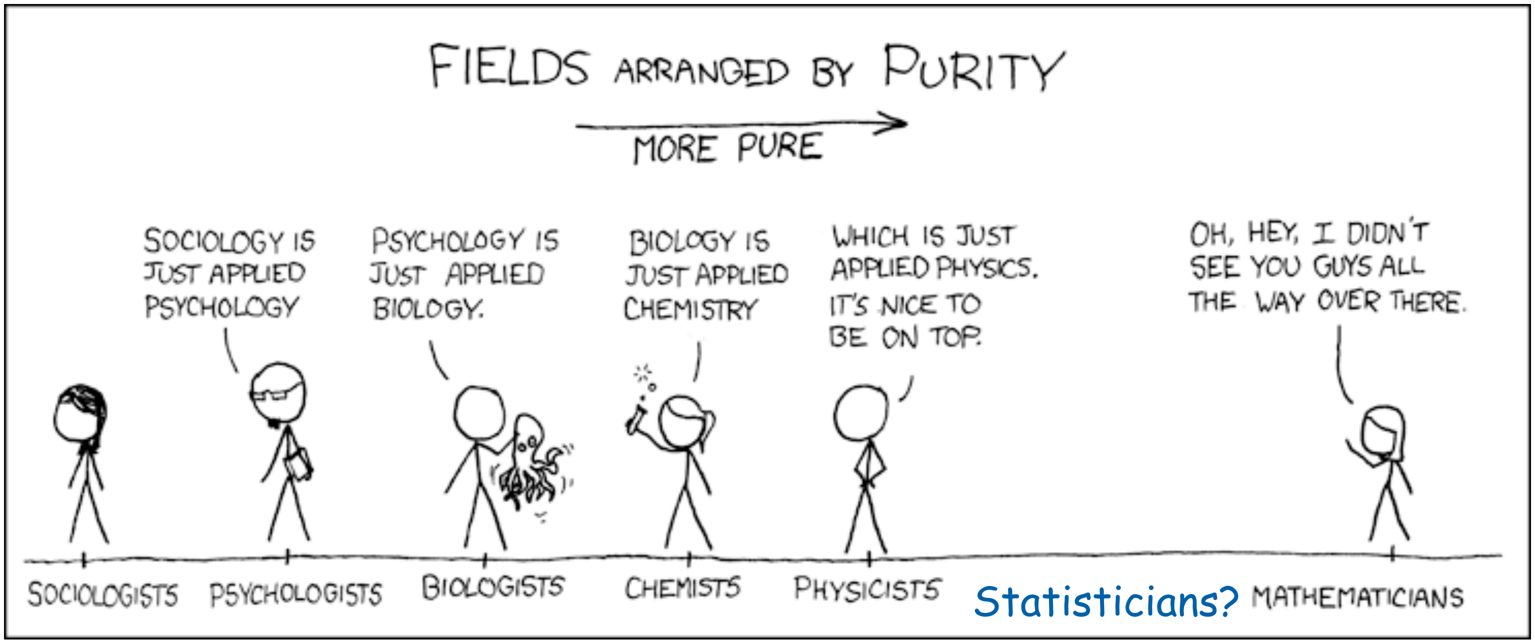
Statistical landscape



The AI hill


terbox
Tracks weight & important litterbox habits for multiple cats using AI
Detects changes that may indicate weight issues or health conditions requiring veterinary diagnosis
Smart Litterbox Monitor
Proactively Manage Your Cat's Health

Statistical purity



More is different: Broken symmetry and the nature of the hierarchical structure of science (P.W. Anderson)

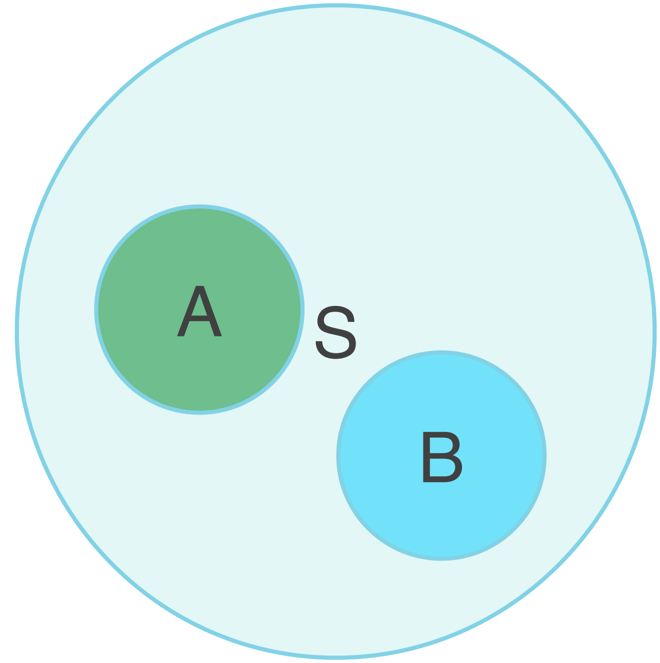
Probability

Topics in probability

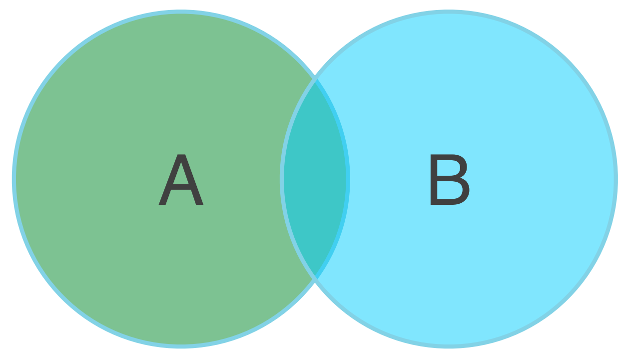
- Axioms
- Bayes' theorem
- Distributions
- Statistical distances
- Information theory
- HEP data

Probability

- Kolmogorov axioms: for a sample space S, we have
  - $\forall A \subset S \quad P(A) \geq 0$
  - $\forall A, B \subset S, A \cap B = \emptyset \quad P(A \cup B) = P(A) + P(B)$
  - $P(S) = 1$



- Conditional probability
  - $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$



- Total probability:
  - For a partition $S = \bigcup_i A_i$ where $i \neq j \implies A_i \cap A_j = \emptyset,\ P(B) = \sum_i P(B \mid A_i) P(A_i)$

Example application: total probability

- 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
- If I count n_r hands raised in n attendees, we have:
  - P(T) = n_t/n
  - P(✋|H) = 1
  - P(✋|T) = n_c/n_t
  - P(✋) = P(✋|H)P(H) + P(✋|T)P(T) = $1 - n_t/n + n_c/n = n_r/n$
  - $\implies n_c/n = n_r/n + n_t/n - 1$
- Can we find P(cheat) for this audience?
  - Which variables are known here? Which are random variates? Which are parameters?
  - How will we interpret the result?

Probability the hard way

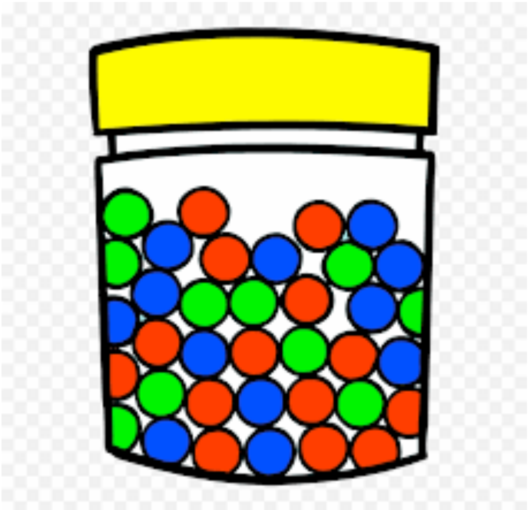
- Measure-theoretic take: [Radon-Nikodym theorem]
- On a measurable space (X, Σ)
  - e.g. $X = \{1,2,3\}, \Sigma = \{\{1\},\{2\},\{3\},\{1,2\},\ldots,\{1,2,3\}\}$ or $X = \mathbb{R}$ with the Borel algebra
- If measure ν is absolutely continuous w.r.t. μ
  - i.e. for $A \in \Sigma$, $\mu(A) = 0 \implies \nu(A) = 0$
  - μ could be a counting measure (for X above) or Lebesgue measure (for $\mathbb{R}^n$)
- Then there is a measurable function $f : X \to [0,\infty)$
  - s.t. $\nu(A) = \int_A f \, d\mu$
  - $f = \dfrac{d\nu}{d\mu}$ is the Radon-Nikodym derivative
- If $\nu(X) = 1$ we have a probability measure
  - i.e. $\nu = P$ can now have continuous support if needed
- Glad I'm not a mathematician

Probability mass

- Probability mass function (pmf)
  - probability of observing a specific outcome
  - defined over a support (space of outcomes/observables/samples)
  - may be parameterized

Examples:

- Marbles: P(draw 2 red, 2 green, 1 blue from jar)
  - pmf Multivariate hypergeometric distribution
  - $$f(\mathbf{k};\mathbf{K}) = \frac{\prod_{i=1}^{c}\binom{K_i}{k_i}}{\binom{\sum K_i}{\sum k_i}}$$
  - Support: $k_i \in \{0,1,\ldots\}$; Parameters: $K_i \in \{1,2,\ldots\}$



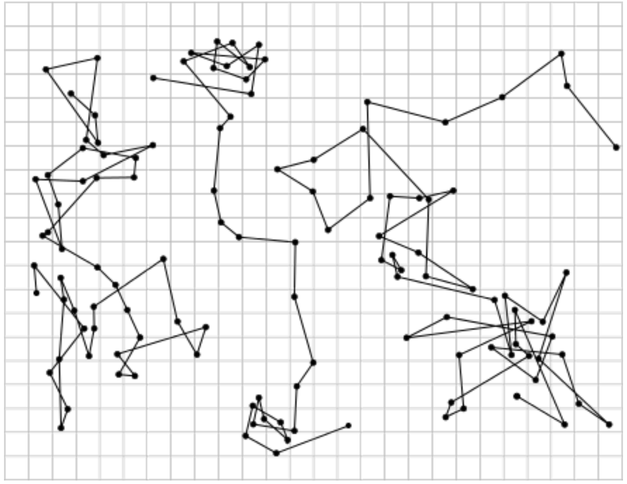
- Counts in a particle detector after some time
  - pmf Poisson distribution
  - $$f(n;\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$
  - Support: $n \in \{0,1,\ldots\}$; Parameters: $\lambda \in [0,\infty)$

Probability density

- Probability density function (pdf), e.g. $f(x)$
  - a *differential* probability of observing an outcome: e.g. for 1D,
    - $P(a < x < b) = \int_a^b f(x)\,dx$, sometimes write $P(x \in A) = \int_A dP$
  - defined over a support (space of outcomes/observables/samples)
  - implies a cumulative (cdf), percentile (inverse cdf), etc. in 1D
  - may be parameterized

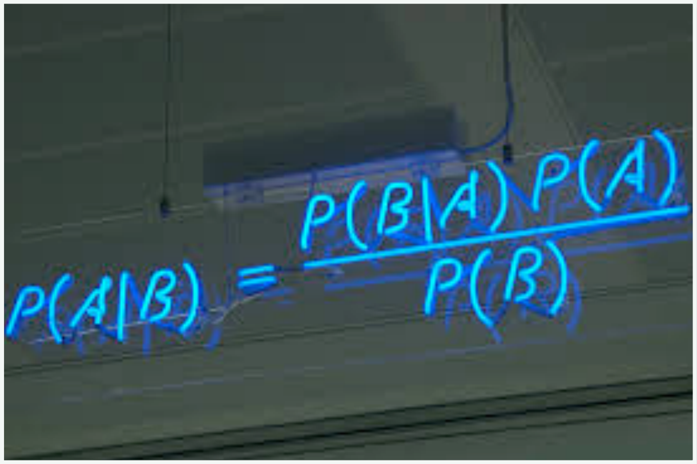
Example:

- Brownian motion: P(displacement after some time)
  - pdf Normal distribution
    - $f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
  - Support: $x \in \mathbb{R}$; Parameters: $\mu, \sigma \in \mathbb{R}, \sigma > 0$



Bayes' theorem

- From conditional probability,
  - $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
  - $\implies P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$
  - $\implies P(A \mid B) = \dfrac{P(B \mid A)P(A)}{P(B)}$

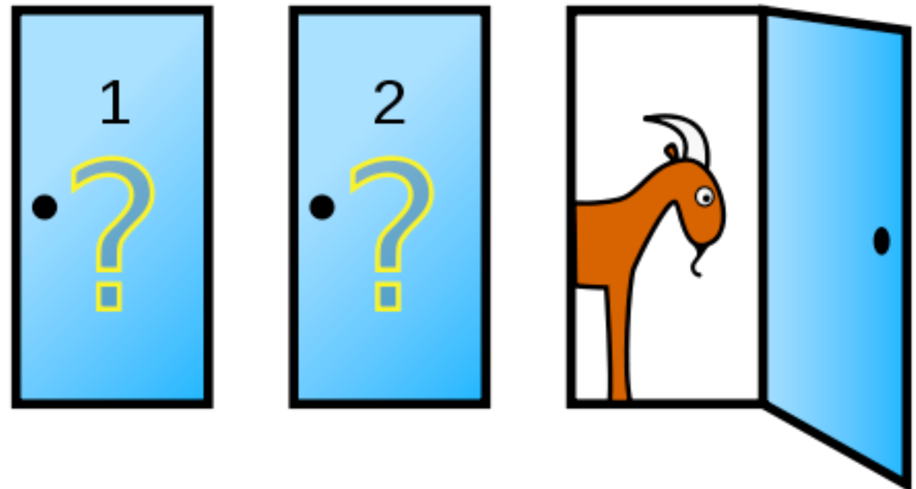


- Note: from total probability,
  - If $\bigcup_i A_i = S$ then we can also write $P(B) = \sum_i P(B \mid A_i)P(A_i)$
  - So $P(B)$ is a normalization
- You don't have to be Bayesian to use Bayes' theorem

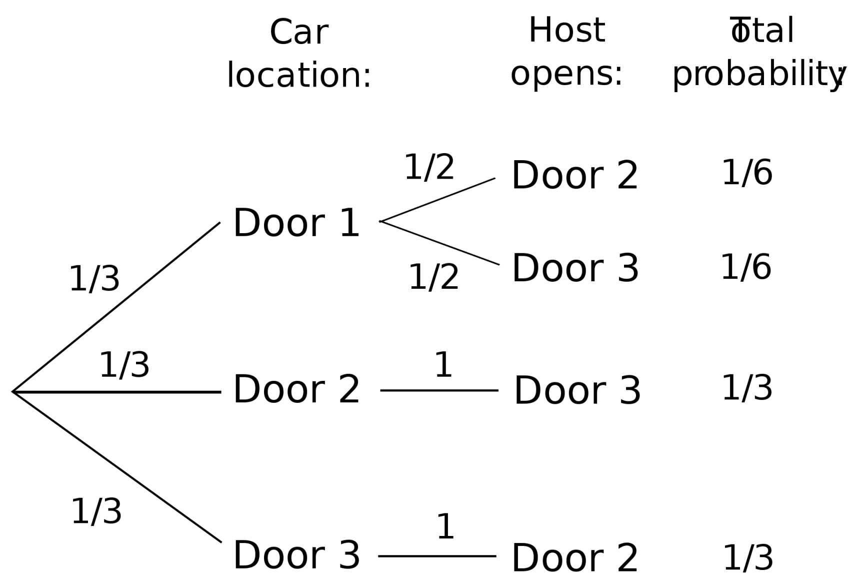
Example application of Bayes' theorem

- Monty hall problem:

> Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?



- You should switch
  - P(door 1 wins) = 1/3
  - P(host opens door 3 I door 1 wins) = 1/6
  - P(host opens door 3) = 1/6 + 1/3
  - ⇒P(door 1 wins I host opens door 3) = 1/3
    - P(door 2 wins I host opens door 3) = 2/3
    - P(door 3 wins I host opens door 3) = 0

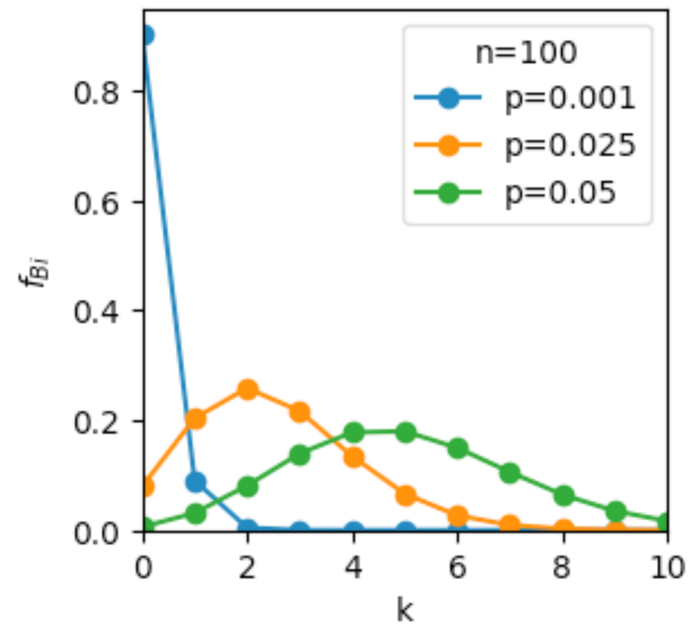


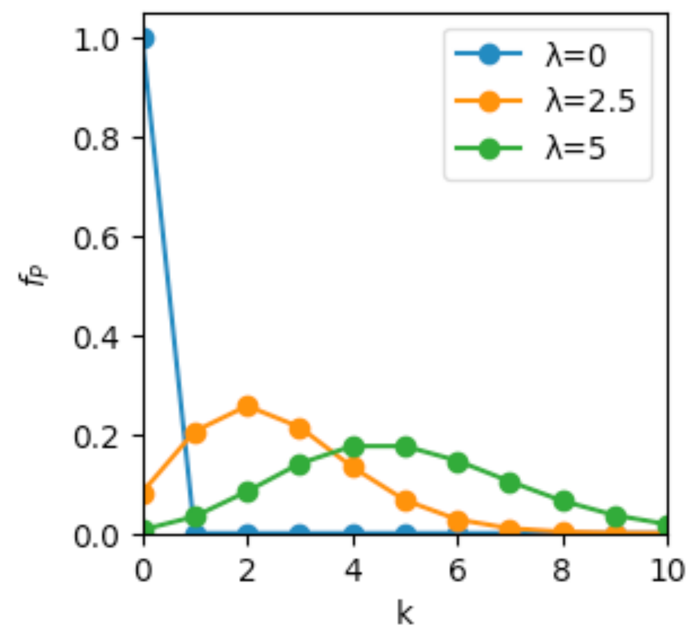
Cool discrete distributions

- Bernoulli: weighted coin flip
  - $f_B(k; p) = p[k = 1] + (1 - p)[k = 0]$

- Binomial: *k* success in *n* trials
  - $f_{Bi}(k; n, p) = \binom{n}{k} p^k (1 - p)^{n-k}$

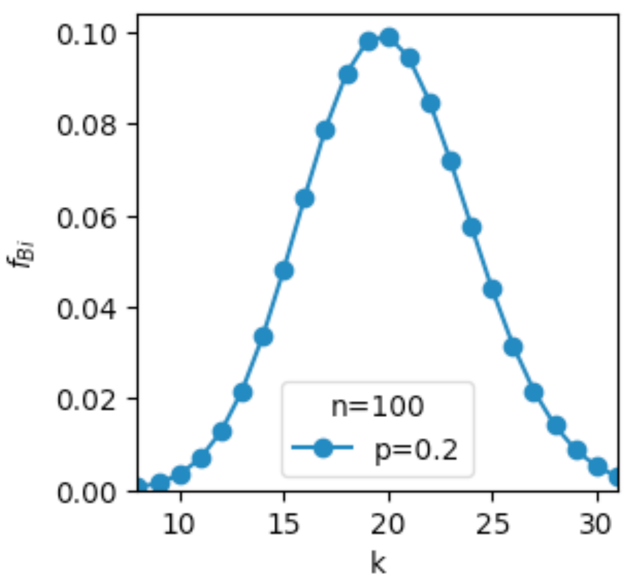
- Poisson: a limiting case of Binomial
  - $f_P(k; \lambda) = \lim_{n \to \infty, np = \lambda} f_{Bi}(k; n, p) = \frac{\lambda^k e^{-\lambda}}{k!}$

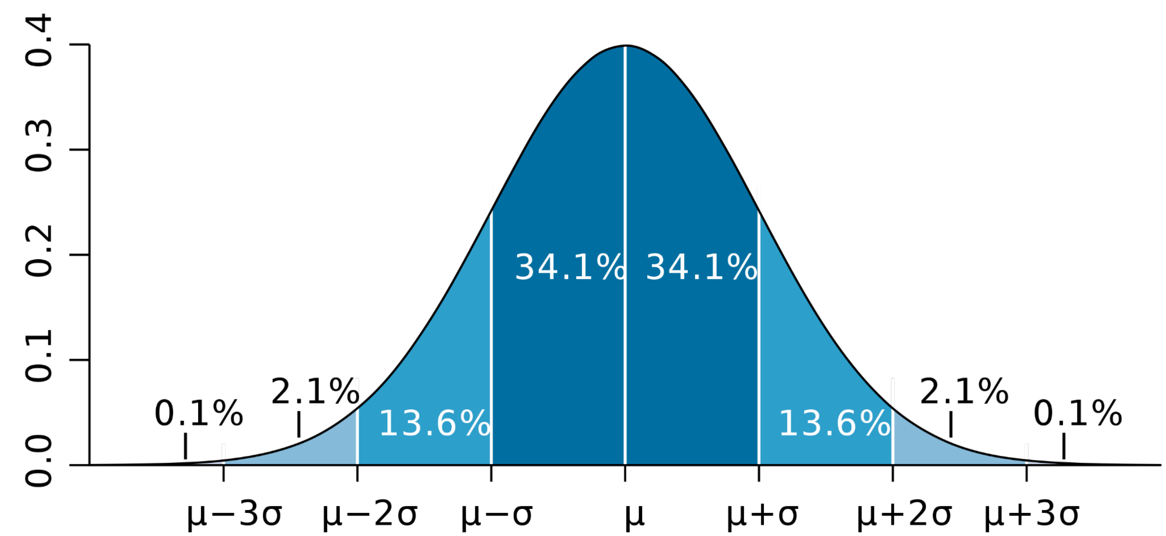




Cool continuous distributions

- Normal: another limiting case of Binomial
  - $f_N(x;\mu = np, \sigma = \sqrt{np(1-p)}) = \lim_{n\to\infty} f_{Bi}(x;n,p)$
  - [Central limit theorem](#):
    - sums of independent random-distributed variables tend towards a Normal-distributed variable
  - [Standard (Z) score](#):
    - Convention for interesting percentiles: “1σ” = 0.6827…, “2σ” = 0.9545…, “5σ” = 5.7e-7
      - These are 2-sided. Can also define 1-sided (common in HEP.) Often quote 95 %-ile





Cool continuous distributions - continued

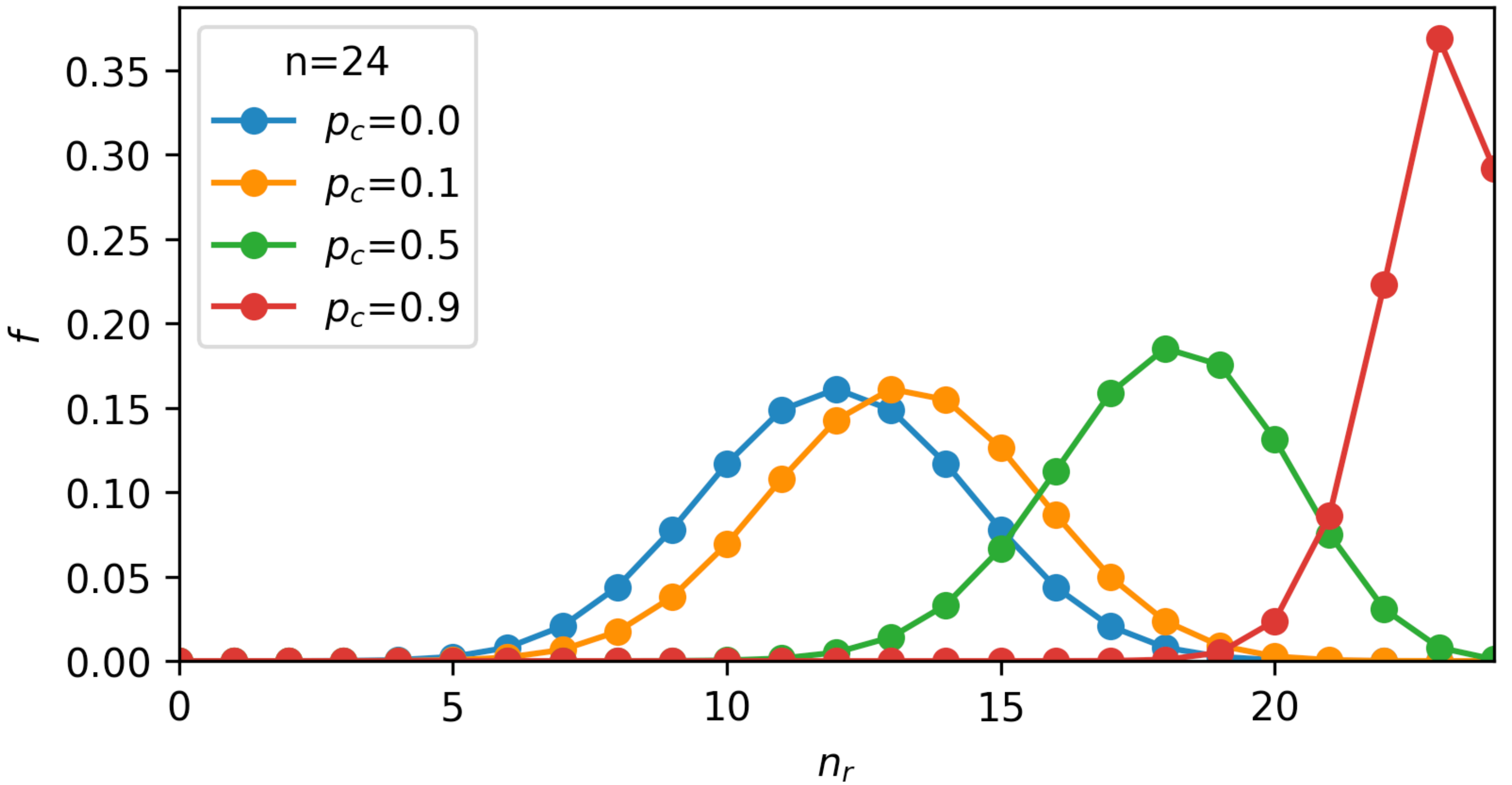
- Chi-square: squared *distance* from mean of unit multivariate normal
  - $\mathbf{x} \sim f_N(\mathbf{x}; \mathbf{0}, \mathbb{1}) \implies \mathbf{x} \cdot \mathbf{x} = d^2 \sim f_{\chi^2}(d^2; n)$
- Log-normal distribution
  - Definition: Normal in log-space (change of variables: $y = \ln(x), dy = x^{-1}dx$)
  - Corollary to central limit theorem:
    - *products* of […] tend towards a Log-normal distributed variable
    - Common model for calibration uncertainties (more later)
- All of the above (both discrete and continuous): *exponential family*
  - *Exponential, gamma, beta, Dirichlet, categorical, Wishart, geometric, Pareto, …*

Coming back to 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
- Sample space: n_t coins are tails, n_c people who flip tails raised their hand
  - $n_t \sim f_{Bi}(n_t; n, p_t)$ and $n_c \sim f_{Bi}(n_c; n_t, p_c)$
- The joint distribution factorizes
  - $$f(n_t, n_c) = f(n_c \,|\, n_t) f(n_t) = \binom{n}{n_t}\binom{n_t}{n_c} p_t^{n_t}(1-p_t)^{n-n_t} p_c^{n_c}(1-p_c)^{n_t-n_c}$$
  - Also an exponential family
- We can *marginalize* over the *latent* variable n_t
  - Also substituting $n_c = n_r + n_t - n$ and $p_t = 1/2$
  - $$f(n_r; n, p_c) = \sum_{n_t=0}^{n} f_{Bi}(n_t; n, 1/2) f_{Bi}(n_r + n_t - n; n_t, p_c)$$
  - We don't have to assume p_t but it seems reasonable (more later)

Coming back to P(cheat)

- Visualizing $f(n_r; n, p_c) = \sum_{n_t=0}^{n} f_{Bi}(n_t; n, 1/2) f_{Bi}(n_r + n_t - n; n_t, p_c)$



Exponential family

- A exponential family is a set of distributions with a pdf/pmf of the form
  - $f(x;\theta) = h(x)e^{\eta(\theta)\cdot T(x) - A(\theta)}$
  - This generalizes to multiple dimensions, implicitly $\dim(\eta) = \dim(T)$
- The terms have an interpretation:
  - $\eta(\theta)$ is the *natural parameter* of the distribution. Values η where f is integrable define the space of the parameter. This space is convex!
  - $T(x)$ is the *sufficient statistic:* it holds *all* data that x provides.
    - For i.i.d. samples $x_i \sim f$, the sufficient statistic of the joint distribution $T(x_1, x_2, \ldots) = \sum T(x_i)$
  - $A(\eta) = \ln \left[\int h e^{\eta T} dx \right]$ is the *log-partition* function.
    - Moments of the sufficient statistic can be found by differentiating $A(\eta)$
- A few interesting properties:
  - Exponential families are the only families with sufficient statistics that can summarize arbitrary amounts of i.i.d. data using a fixed number of values
  - All distributions in this family have *conjugate priors* $\pi(\eta;\chi,\nu) = f(\chi,\nu)e^{\eta\cdot\chi - \nu A(\eta)}$
  - The relative entropy (KL-divergence) can be computed using $A(\eta)$ and its derivative

Exponential family

- Example: Normal distribution with *known* variance
  - $h(x) = (2\pi\sigma^2)^{-1} e^{-x^2/2\sigma^2}$
  - $\eta(\mu) = \mu/\sigma$
  - $T(x) = x/\sigma$
  - $A(\eta) = \eta^2/2$
    - $E[T] = dA/d\eta = \eta \implies E[x] = \mu$
    - $E[T^2] = d^2A/d\eta^2 = 1 \implies E[x^2] = \sigma^2$
  - Conjugate prior $\pi(\eta; \chi, \nu) = f(\chi, \nu) e^{\eta\chi - \nu\eta^2/2}$
    - Complete the square: Normal distribution
- Example: Poisson distribution $f_P(k; \lambda)$
  - $h(k) = 1/k!$
  - $\eta(\lambda) = \ln \lambda$
  - $T(k) = k$
  - $A(\eta) = e^\eta$
  - Conjugate prior $\pi(\eta; \chi, \nu) = f(\chi, \nu) e^{\eta\chi - \nu e^\eta} \implies \pi(\lambda) \propto \lambda^\chi e^{-\nu\lambda} \lambda^{-1}$
    - A Gamma distribution