

PHYS/ASTR 8150

Introduction to probabilities and statistics

Fabien Baron

Georgia State University

baron@chara.gsu.edu

Fall 2023

The use of statistics and probabilities

- **Parameter estimation**: given some data, what is our best estimate of a particular parameter? What is the uncertainty in our estimate?
- Identify data **correlations** : are two variables we have measured correlated with each other, implying a possible physical connection?
- **Model/hypothesis testing** : given some data and one or more models, are our data consistent with the models? Which model best describes the data?

The use of statistics and probabilities (2)

Statistics are a mean to summarize concisely yet rather precisely some of the characteristics of our data, while probabilities are a way to understand how the data is likely to behave.

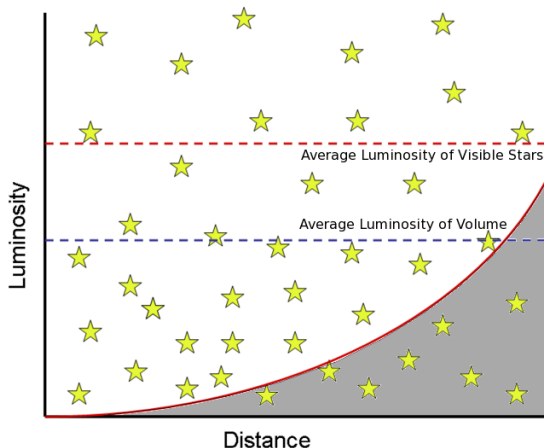
Both are used to state the uncertainties ("errors") in measurements:

- **Random errors:** always present in a measurement, inherently unpredictable fluctuations in the readings of a measurement apparatus or in the experimenter's interpretation of the instrumental reading (**stochastic** noise). Random errors show up as different noise occurrences and can be estimated by comparing multiple measurements, and reduced by averaging multiple measurements.
- **Systematic errors:** typically constant or proportional to the true value, caused by imperfect calibration of measurement instruments or imperfect methods of observation, or interference with the measurement process. Always affect the results of an experiment in a **predictable** direction. Systematic errors cannot be discovered or estimated by comparing/averaging occurrences, as they always push measurements in the same direction.

Biases

Statistics and probabilities can prevent us from being fooled by physical and psychological biases:

- Selection effects leading to spurious correlations, for example Malmquist bias



- Confirmation bias : conclusions distorted by our preconceived idea about what the result should be

New light on old rays: N rays

Robert T. Lagemann

Department of Physics and Astronomy, Vanderbilt University, Nashville, Tennessee 37235

(Received 25 March 1976; revised 31 August 1976)

During the period 1903–1906, some 120 trained scientists published almost 300 articles on the origins and characteristics of a spurious radiation, the so-called N rays. Some new explanations are advanced for the extensive false observations and the deductions made from those observations. These are based on visits to Nancy, France, where the purported discovery was first announced and after which the rays were named, on an interview with a former assistant who knew some of the principals in the case, and on new archival information. Some of the misleading statements in the subsequent literature and oral history dealing with N rays are challenged, and additional information is provided on the original “discoverer,” René Blondlot.

- Small samples: leading to noisy results
- A posteriori: using the same dataset which motivated a hypothesis to

Noise and probabilities

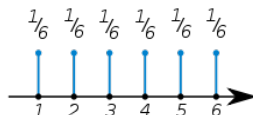
- Data is affected by noise, which prevents from getting to the exact mathematical laws, and behaves as a **random variable**
- Understanding the noise = we may recover the real data
- Probabilities are the best tool to describe the noise **stochastic processes** or the data probability distribution
- Statistics cannot be understood if the noise **distribution** is unknown
- **Population**: the ensemble of all the samples with the characteristic one wishes to understand, i.e. the total set of observations that can be made. A statistical population can be a group of actually existing objects (e.g. the set of all stars within the Milky Way galaxy) or a hypothetical and potentially infinite group of objects conceived as a generalization from experience (e.g. the set of all possible hands in a game of poker).
- **Sample**: a set of data collected and/or selected from a statistical population by a defined procedure, possibly repeated. It is a subset of the population, often chosen to represent the population in a statistical analysis.

Probability mass functions: discrete distributions

- **Discrete distribution:** data X is a discrete random variable, i.e. X takes integer values only.
- **Probability mass function** $f_X(x) = \Pr(X = x)$, gives the probability that X is exactly equal to some value x .
- Coin flip (**Rademacher distribution**): -1 to tails and 1 to heads, random variable X has a 50% chance of each,

$$f_X(x) = \begin{cases} 1/2 & \text{if } x = -1, \\ 1/2 & \text{if } x = +1, \\ 0 & \text{otherwise.} \end{cases}$$

- Uniform distribution for dices, e.g. for a 6-sided dice:



- A **discrete probability distribution** is a probability distribution characterized by a probability mass function with $\sum_x \Pr(X = x) = 1$

Probability density functions

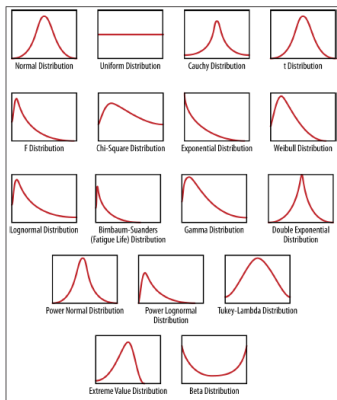


Figure 2-1. A bunch of continuous density functions (aka probability distributions)

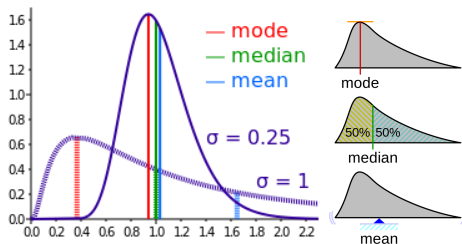
- Continuous case: if X a continuous random variable, the **probability density function** $f_X(X)$ is a continuous probability distribution
- **Support** of a probability function is the set of points where the probability density is not zero-valued
- The probability mass (not density) of getting any exact value is zero:
 $\Pr(X = a) = 0$
- In any interval $[a, b]$,
 $\Pr(a \leq X \leq b) = \int_a^b f_X(x) dx$

Measures of central tendency - on a sample

Given a sample [1, 3, 6, 6, 6, 6, 7, 7, 12, 12, 17] with equal probabilities, find:

- **Sample Mean:** $\bar{X} = \frac{1}{N} \sum_{i=0}^N x_i$. \bar{X} is the sample mean, as opposed to the population mean/**expected value** $E[X]$ which would be based on the underlying probability distribution; this latter would be used for long-run/time averages, or Monte Carlo simulations.
- **Sample Mode:** the element that occurs most often in the collection; if not unique, sample is said to be multimodal.
- **Sample Median:** the element separating the higher half of a data sample from the lower half; arranging all the observations from lowest value to highest value and picking the middle one. If there is an even number of observations, then there is no single middle value; the median is then usually defined to be the mean of the two central values.

Measures of central tendency - on a population/distribution



- **Population mean = "Expected value".**
- For a discrete distribution, $E[X] = \sum x \Pr(X = x)$
- For a continuous one $E[X] = \int x f_X(x) dx$, over all the support of the distribution for the population.
- Mode: the peak value x where the probability density is maximum
- Median: x so that $P(X \leq x) = \frac{1}{2}$ and $P(X \geq x) = \frac{1}{2}$.

Measures of central tendency - sample vs population

- I roll a six-sided die three times and get $[1, 3, 2]$.
- the sample mean is $(1 + 3 + 2)/3 = 2$.
- the population mean is $(1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$, which means I would have expected to get an average of 3.5.
- **Weak Law of Large Numbers/Bernoulli's theorem:** the sample mean converges towards the population mean for large samples (but only for distributions with existing means and variances). Example: coin flipping will average to 0.5 for probabilities of heads/tails.

Population Variance

- If $\mu = E[X]$, $\text{Var}(X) = E[(X - \mu)^2] = E[X^2] - E[X]^2$
- The population variance for a discrete distribution is:
$$\text{Var}(X) = \sum x^2 \Pr(X = x) - (\sum x \Pr(X = x))^2$$
- The population variance for a continuous distribution is:
$$\text{Var}(X) = \int (x - \mu)^2 f(x) dx = \int x^2 f(x) dx - \mu^2$$

Sample Variance

- The sample variance: $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 = \left(\frac{1}{N} \sum_{i=1}^N x_i^2 \right) - \mu^2$
- If μ , the true (population) mean is unknown, it is often computed as the sample mean. Then the sample variance becomes a **biased** estimator: it will underestimate the variance by the Bessel's correction factor $(N - 1)/N$. Then an unbiased/debiased sample variance is:

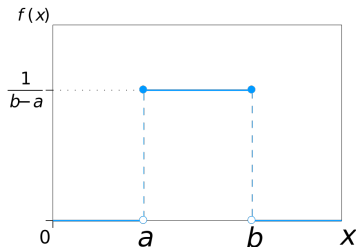
$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2$$

- Bessel's correction corrects the bias in the estimation of the population variance, not all of the bias in the estimation of the sample standard deviation, which will be underestimated. Homework 1 will deal with this.

Cumulative distribution functions

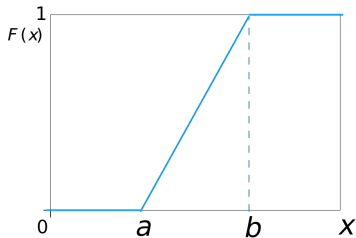
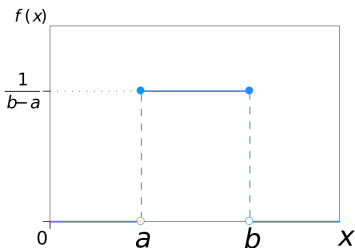
- The **cumulative distribution function** measures the probability that the variable takes a value less than or equal to x .
- CDF: $F_X(x) = P(X \leq x)$
- Consequently $P(X > a) = 1 - F_X(a)$
- $P(a < X \leq b) = F_X(b) - F_X(a)$
- In the continuous case: $F_X(x) = \int_{-\infty}^x f_X(t) dt$

Examples of distributions: continuous uniform distribution



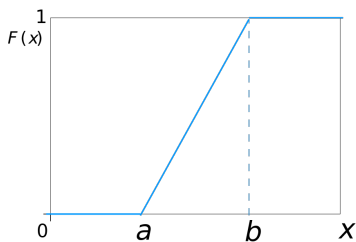
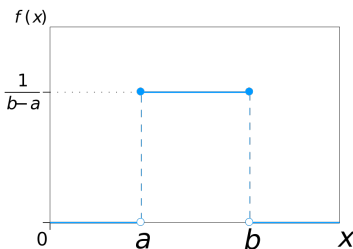
- Support ?
- Probability density function ?
- Cumulative distribution function ?
- Mean ?
- Median ?
- Mode ?
- Variance, and application to a randomly distributed variable in $[-\pi, +\pi]$?

Examples of distributions: continuous uniform distribution



- Support $x \in [a, b]$, Notation $X \sim \mathcal{U}(a, b)$.
- Probability density function $f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$
- CDF = $\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b] \\ 1 & \text{for } x \geq b \end{cases}$

Examples of distributions: continuous uniform distribution (2)

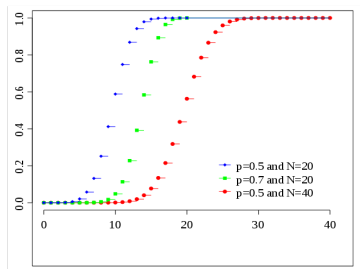
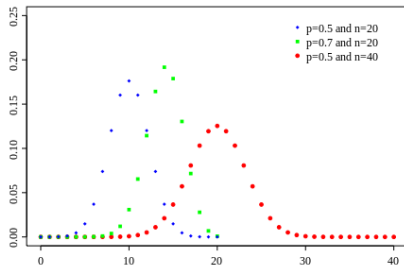


- Mean, Median $\frac{1}{2}(a + b)$, multimodal in (a, b)
- Variance $\frac{1}{12}(b - a)^2$, standard deviation is $\frac{b-a}{2\sqrt{3}}$
- Consequence: if data on an angle (by definition in $[-\pi, +\pi]$) has a standard deviation comparable to $\pi/\sqrt{3} \text{ rad} \simeq 104^\circ$, it is possibly completely random.

Examples of distributions: Binomial distribution

- n **independent** trials of a random process with two mutually exclusive outcomes with probabilities p and $1 - p$, and k successes (occurrences of p probability). Notation: $X \sim B(n, p)$.
- Typical outcomes: detection or non-detection, belonging or not to a class of objects
- Special cases: tossing biased (Bernoulli) or unbiased coin (Rademacher distribution).

Examples of distributions: Binomial distribution



- Support: number of successes, $k \in [0 \dots n]$.
- PMF: $f(k; n, p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- $E[X] = np$, $\text{Var}\{X\} = np(1-p)$
- CDF:

$$\Pr(X \leq k) = \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i} = B_I(1-p, n-k, 1+k) / B(n-k, 1+k)$$

where B is the beta function and B_I the incomplete beta function.

Application of the Binomial distribution

- Professor Henry has a test with 10 multiple choice questions, with 5 possible answers per question. The right answer is unique for each question. Failing the test means getting fewer than 50% right answers (≤ 4 right answers in this case). For students that answer questions at random, Professor Baron wants to get the same failure rate with only 4 choices. How many questions are needed ?
- In the RECONS stellar catalog, 60% of stellar systems are binaries. How large should a stellar system sample be to have 99% or more chance of having at least two binary systems in the sample ?
- Find the probability of at least two students having the same birthday in a class of 25. Is the binomial distribution the right approach to the problem ? What would be the meaning of the binomial expectation here ?

Solution to the first two problems

- Failing the test: the events are independent with constant p , the binomial pdf applies:

$$\Pr(X \leq 4; n = 10; p = .2) \leq \Pr(X \leq \lceil \frac{n}{2} \rceil - 1; n; p = \frac{1}{4})$$

$$.967 \leq \left(\frac{3}{4}\right)^n + n \left(\frac{1}{4}\right) \left(\frac{3}{4}\right)^{n-1} + \dots + \binom{n}{\lceil \frac{n}{2} \rceil - 1} \left(\frac{1}{4}\right)^{\lceil \frac{n}{2} \rceil - 1} \left(\frac{3}{4}\right)^{\lceil \frac{n}{2} \rceil}$$

$n \geq 16$ questions as $\Pr(X \leq 8; 8; p = .25) = .973$

- Binaries: binomial applies with $p = 0.6$, and we want the number of samples n to have the probability of finding FEWER than 2 binaries (i.e. 0 or 1) below $100\% - 99\% = 1\%$.

$$\Pr(X = 0) + \Pr(X = 1) \leq 0.01$$

$$\binom{n}{0} p^0 (1-p)^{n-0} + \binom{n}{1} p (1-p)^{n-1} \leq 0.01$$

$$0.4^n + n \times 0.6 \times 0.4^{n-1} \leq 0.01 \rightarrow n \geq 8$$

Analyzing the birthday problem

The binomial expected value (and standard deviation) should give you a strong clue something's amiss: $E[X] = np$ is n times the probability of a single event, reminding you it's treating the events as independent. Using the binomial probability distribution is **not** the right approach for this birthday problem, as the elementary events (having similar birthdays) are **not independent**. Let's look at the binomial expression terms:

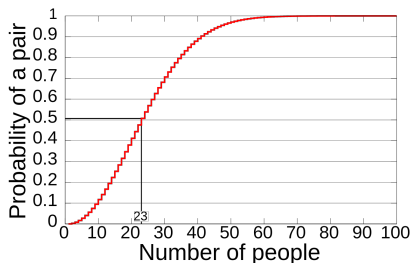
- p^k : probability of having k independent successes.
- $(1 - p)^{n-k}$: probability of having $(n - k)$ independent failures.
- $\binom{n}{k}$: number of possible ways to get k successes in n trials (e.g. success at first trial then third, or 4th and 6th, etc.).

There is no general formula to solve a problem with non-independent conditional probabilities (except Bayes theorem, later...).

The binomial distribution would only apply if you took a given student's birthday and compared it to all other 25, one by one. In this case the probability reaches $\sim 6.8\%$.

How to solve the birthday problem

- We want $1 - P(X = 0)$, where $P(X = 0)$ means $n = 25$ people with 0 common birthdays
- Easy (365 days) or hard problem (365.25 days, Feb 29th is counted)
- Classic way: first find the cardinal of the population ensemble, aka how many possible combination of birthdays can there be: 365^n .
- Then find how many possible ways of assigning birthdays without overlaps: $365 \times 364 \times \dots \times (365 - n + 1) = 365! / (365 - n)!$.
- Therefore answer is $1 - P(X = 0) = 1 - 365! / (365^n (365 - n)!)$.



How to solve the birthday problem - Numerical method

- Using the classic formula has drawbacks: calculation of factorials can overflow (Stirling and similar formula), and is undefined for non-integers (e.g. Mars year is 668.6 sols; would the gamma function solve this ? no). Solution = use **recursion**
- For 365 days, evaluate $p(n)$ so that:

$$p(i) = \frac{365 - i + 1}{365} p(i - 1) \quad \text{with} \quad p(1) = 1$$

For 365.25 days, $p_A(n)$ = prob. n different birthdays other than Feb 29, $p_B(n)$ = prob. n different birthdays, including one on Feb 29th.
so that $p(i) = p_A(i) + p_B(i)$:

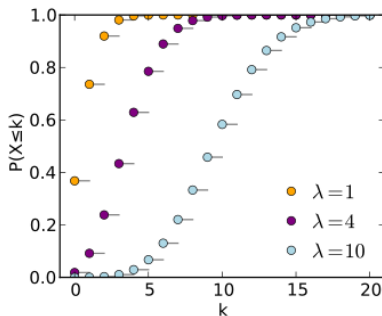
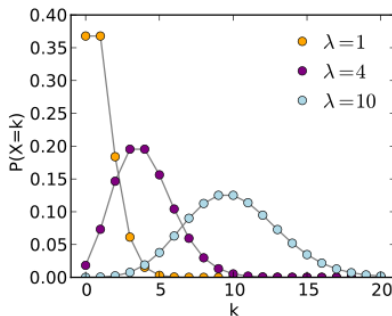
$$p_A(1) = \frac{365}{365.25}, \forall i > 1 : p_A(i) = \frac{365 - i + 1}{365} p_A(i - 1)$$

$$p_B(1) = \frac{0.25}{365.25}, \forall i > 1 : p_B(i) = \frac{365 - i + 2}{365} p_B(i - 1) + \frac{0.25}{365.25} p_A(i - 1)$$

Examples of distributions: Poisson distribution

- **Discrete process**, measures number of events happening within a fixed interval of time if these events occur with a **known average rate** and **independently** of the time since the last event.
- Examples: junk mails per day, number of phone calls received by a call center per hour, decay events per second from a radioactive source, shot noise on cameras. Warning: the lapse of time between Poisson events does not follow a Poisson distribution, but an **exponential distribution** which is a continuous distribution.
- Shot noise caused by statistical quantum fluctuations in the number of photons sensed at a given exposure level. Noises at different pixels are independent of one another.
- Estimate of Poisson noise: "Flux in pixel is k counts, therefore estimate of standard deviation \sqrt{k} counts", assumes the mean count is the observed count, which is not true for low counts.

Examples of distributions: Poisson distribution



- PMF: $f(k; \lambda) = \Pr(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$, where λ the known average rate of occurrence, and k is the number of actual events observed.
- CDF: $F(k; \lambda) = \Pr(X \leq k) = e^{-\lambda} \sum_{i=0}^k \frac{\lambda^i}{i!}$
- $E[X] = \lambda$, $\text{Var}(X) = \lambda$.
- Poisson distribution does approach a normal distribution about its mean for large λ (typically $\lambda \geq 20$).

Application of the Poisson distribution

- The density of bugs in my code is 8 per 20,000 lines. How many lines should I read to ensure more than 50% chance of finding a bug? What is the maximum number of lines I can have my boss read ensuring he will see fewer than 4 bugs?

Solution

- We have a mean rate, so Poisson can be applied. Note we also have a true/false statement, so binomial could apply too.
- Expected number of bugs is $= \frac{8}{20000} n$

$$\Pr(X = 0) = \exp\left(-\frac{8}{20000}n\right) \leq 0.5$$
$$n \geq 1733$$

- What is the maximum number of lines I can have my boss read insuring he will see fewer than 4 bugs? We need some additional information, i.e. acceptable probability. Otherwise there is always a probability, admittedly small, that there could be 5 bugs in the first line. We could set a threshold of 99%.

$$\Pr(X \leq 4) \geq 99\%$$

Examples of distributions: Poisson and Binomial

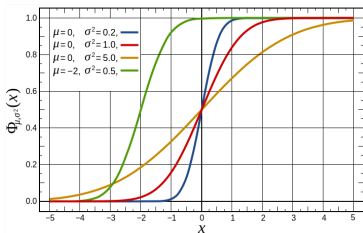
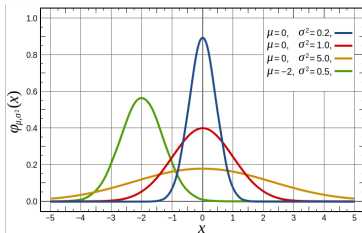
- **Law of rare events:** if $X \sim B(n, p)$, when $n \rightarrow \infty$ and $p \rightarrow 0$ and $\lim_{n \rightarrow \infty} np = \lambda$, then $X \sim P(\lambda)$. $p = \lambda/n$ is the probability of success of each of the k trial.

$$\begin{aligned}\lim_{n \rightarrow \infty} P(X = k) &= \lim_{n \rightarrow \infty} \binom{n}{k} p^k (1-p)^{n-k} \\&= \lim_{n \rightarrow \infty} \frac{n!}{(n-k)!k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\&= \lim_{n \rightarrow \infty} \left(\frac{\lambda^k}{k!}\right) \rightarrow \frac{\lambda^k}{k!} \\&\times \left(\frac{n}{n}\right) \left(\frac{n-1}{n}\right) \dots \left(\frac{n-k+1}{n}\right) \rightarrow 1 \\&\times \left(1 - \frac{\lambda}{n}\right)^{-k} \rightarrow 1 \\&\times \left(1 - \frac{\lambda}{n}\right)^n \rightarrow e^{-\lambda}\end{aligned}$$

Examples of distributions: Poisson and Binomial (2)

- **Poisson limit theorem:** another name for the law of rare events, if $n \rightarrow \infty$ and $p \rightarrow 0$ and $np \rightarrow \lambda$, then the binomial distribution tends to a Poisson distribution.
- Poisson distribution is the equivalent of a binomial law with a large number of attempts, each with with low probability, so that the expected average rate is a "reasonable" (not too low, not too high) number
- in practice, $n > 20$ and $p < 0.05$ works.

Examples of distributions: Gaussian/Normal distribution



- Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$
- PDF: $f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- CDF: $F(x; \mu, \sigma) = \Pr(X \leq x) = \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{x-\mu}{\sigma\sqrt{2}} \right) \right]$, with $\operatorname{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^x e^{-t^2} dt$
- $E[X] = \operatorname{Mode}(X) = \operatorname{Med}(X) = \mu$, $\operatorname{Var}(x) = \sigma^2$

Error bars and Gaussian/Normal distribution

- $X = \mu \pm \sigma$ means we assume X is a random variable normally distributed.
- Using the CDF we get:

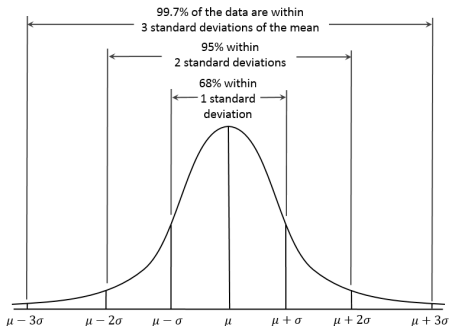
$$\begin{aligned}\Pr(\mu - n\sigma < X \leq \mu + n\sigma) \\ &= F(\mu + n\sigma) - F(\mu - n\sigma) \\ &= \operatorname{erf}\left(\frac{n}{\sqrt{2}}\right)\end{aligned}$$

- 68-95-99.7 or "3-sigmas" rule

$$\Pr(\mu - \sigma \leq x \leq \mu + \sigma) \approx 68.27\%$$

$$\Pr(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 95.45\%$$

$$\Pr(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 99.73\%$$

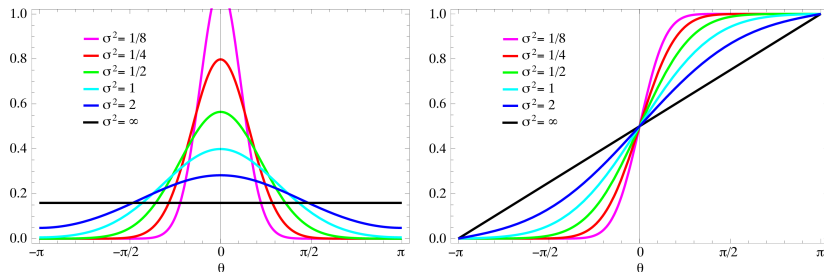


Everything's normal, everything's fine...

Why is "everything" in papers normally distributed ?

- In the high "number of events" regime, Poisson for ($\lambda > 20$) and binomial distributions (very large n , De Moire-Laplace theorem) behave like Normal distributions.
- **Central limit theorem**: the mean of a large number of **independent and identically distributed** random variables, each with a well-defined expected value μ and well-defined variance σ , will be approximately normally distributed, **regardless of the underlying distribution**, as $\mathcal{N}(\mu, \sigma^2/N)$
- Take a large number of independent observations, and **average the results**; repeat this and note the distributions of theses averages: the central limit theorem says they will be normally distributed.
- The underlying distribution does not have to be unimodal !

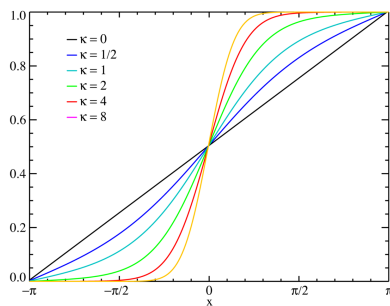
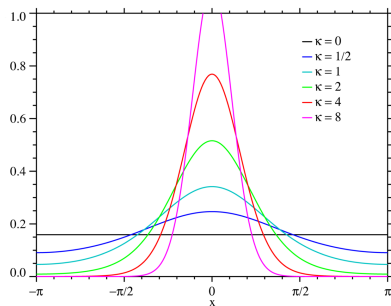
Wrapped normal distribution



Wrapped normal with $\mu = 0$ and support $[-\pi, +\pi]$

- PDF: $f_{WN}(\theta; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} \exp\left[-\frac{(\theta - \mu + 2\pi k)^2}{2\sigma^2}\right]$
- Mean = Median = Mode = μ , circular variance $\text{Var}\{e^{i\theta}\} = 1 - e^{-\sigma^2}$
- A pain to deal with... No analytic CDF and tricky to manipulate.

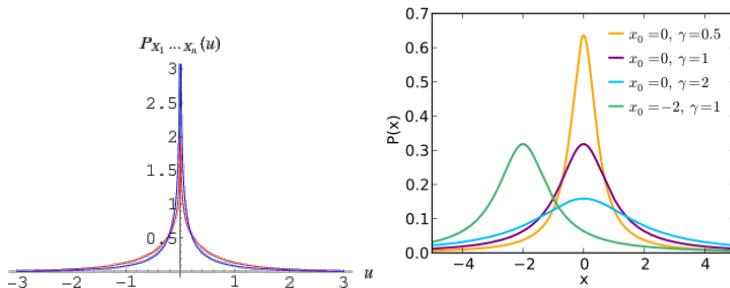
von Mises distribution



von Mises with $\mu = 0$ and support $[-\pi, +\pi]$

- A close (and convex) approximation of the wrapped normal distribution whose derivative is non-discontinuous
- PDF: $f_{WN}(\theta; \mu, \kappa) = \frac{e^{\kappa \cos(x-\mu)}}{2\pi I_0(\kappa)}$. $1/\kappa$ and Gaussian σ^2 play a similar role.
- Mean = Median = Mode = μ , $\text{Var}\{e^{i\theta}\} = 1 - I_1(\kappa)/I_0(\kappa)$
- Still no analytic CDF.

Combination of normal variables



- If X and Y are independent normally distributed random variables:
 - $Z = X + Y$ and $Z = X - Y$ are normally distributed.
 - $Z = XY$ follows a **normal product distribution** (Bessel function, left panel).
 - $Z = X/Y$ follows a **ratio distribution**. In the case where $X \sim \mathcal{N}(0, \sigma_X^2)$ and $Y \sim \mathcal{N}(0, \sigma_Y^2)$, it is a Cauchy distribution (above, right), else a Hinkley distribution. The mean and variance are in general undefined.
- If X is normally distributed, X^2 follows a χ^2 distribution.

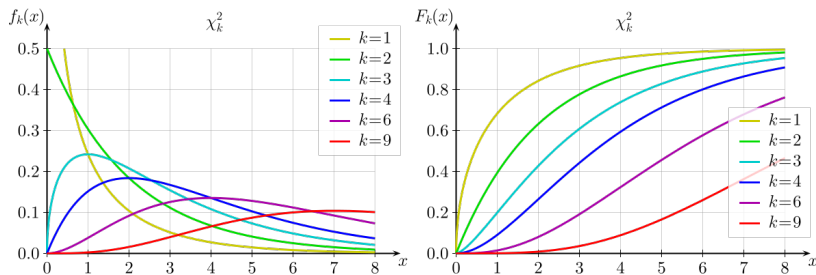
Examples of distributions: χ^2 distribution

- The chi-squared distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables.
- A χ^2 analysis is used for model-fitting (parameter estimation), when collected data points x_i and their (Gaussian distributed) associated uncertainties σ_i are confronted to model predictions μ_i .
- The χ^2 measure the distance between data and model predictions:

$$\chi^2 = \sum_{i=1}^k \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 = \sum_{i=1}^k X_i^2$$

- Each residual X_i (if the μ_i and σ_i are correct !) should be $X_i \sim \mathcal{N}(0, 1)$.
- Hence the χ^2 follows a aptly-named χ^2 distribution with k degrees of freedom.

Examples of distributions: χ^2 distribution



- PDF: $f_k(x) = \frac{1}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$
- CDF: $F_k(x) = \Pr(X \leq x; k) = \frac{1}{\Gamma(\frac{k}{2})} \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$
- Γ : Gamma function, γ : lower incomplete gamma function.
- $E[X] = k$, $\text{Mode}\{X\} = \max(k-2, 0)$, $\text{Med}\{X\} = k(1 - \frac{2}{9k})^2$, $\text{Var}\{X\} = 2k$.
- For $k < 10$, distribution very skewed/non-Gaussian.
- For $k > 50$, central limit theorem implies $\chi^2 \sim \mathcal{N}(k, 2k)$.

Reduced χ^2

- The reduced χ^2 is defined $\chi_r^2 = \chi^2/k \sim 1 \pm \sqrt{2/k} \xrightarrow{k \rightarrow \infty} 1$
- For reasonably high k , whenever the correct model is employed to predict μ_i , $\chi_r^2 \simeq 1$ (not $\chi_r^2 \simeq 0$) is expected due to statistical fluctuations .
- For low k , outliers may dominate the χ^2 . As k increases, you get more data and your estimate of the χ^2 get closer to its population mean. A reduced χ^2 analysis is essentially meaningless when attempted on fewer than 10 points.
- When fitting a model law with p parameters onto a sample of N independent data points, conventional wisdom often says to $k = N - p$, but this is actually arbitrary.
- While a χ^2 analysis enables you to compare how well different models fit the data, it cannot tell you which model is more probable (which requires a Bayesian framework).

Homework 1 - reduced χ^2

- **Task:** imagine you're characterizing or fitting a law through N data samples $\{x_i\}_{i=1}^N$, each x_i being normally distributed $\sim \mathcal{N}(\mu, \sigma)$.
- Take $N = 1000, \mu = 5, \sigma = 3$ and demonstrate numerically (= using code + plots) that the sample mean follows a Gaussian distribution and the sample variance a χ^2 distribution.
- Let's define $\chi_r^2 = \chi^2/N$. Plot the probability density distributions $f_N(x)$ on the continuous range $0 \leq x \leq 4$ for $N = 4, 10, 20, 100, 1000$, each distribution being "renormalized" so that its maximum is 1. Add the corresponding legend and X and Y axis titles, and give the probability of getting $\chi_r^2 \leq 1$ and $\chi_r^2 > 2$ in each case.
- Showcase an example of the central limit theorem with the continuous uniform distribution and Poisson distribution.
- Put your code with reasonable amount of comments + plots into a single PDF file (L^AT_EX preferred if you can) and email me with email title "[ASTR/PHYS 8150] Homework 1".

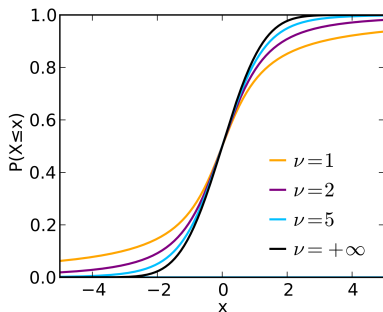
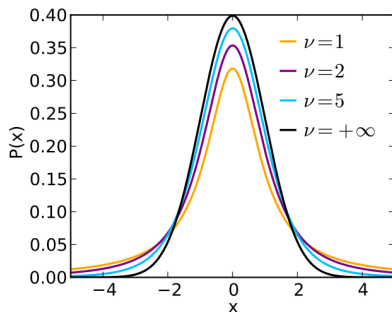
How a sample is approaching a population (1)

- Let's suppose we made N **independent** observations of a random variable X from a Gaussian-distributed population with unknown parameters $\mathcal{N}(\mu, \sigma^2)$.
- $\bar{X} = \frac{1}{N} \sum_{i=1}^n X_i$ is **the sample mean** and $S^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$ is **the sample variance**.
- As $N \rightarrow +\infty$, $\bar{X} \rightarrow \mu$ and $S^2 \rightarrow \sigma^2$. The **standard error** of a statistic is the estimated standard deviation of this statistic. The **standard error of the sample mean** is the standard deviation of the error in the sample mean with respect to the true (population) mean. It can be shown that $\text{SE}(\bar{X}) = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} \simeq \frac{S}{\sqrt{N}}$. We also have $\sigma_{\text{med}(X)} \simeq \sqrt{\frac{\pi}{2}} \frac{\sigma}{\sqrt{N}}$; the median is more subject to sampling fluctuations than the mean.
- Standard error on sample variance: $\sigma_{S^2} = \sigma^2 \sqrt{\frac{2}{N-1}}$. The standard error on sample standard deviation is tricky since S is a biased estimator of σ for low N , but $\sigma_S \simeq \frac{1}{\sqrt{2(N-1)}}$ for $N > 10$.

How a sample is approaching a population (2)

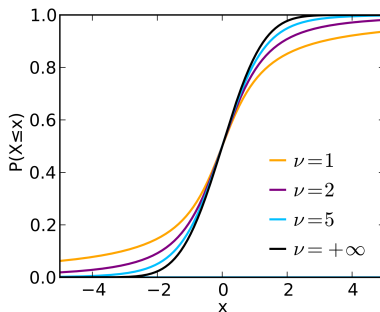
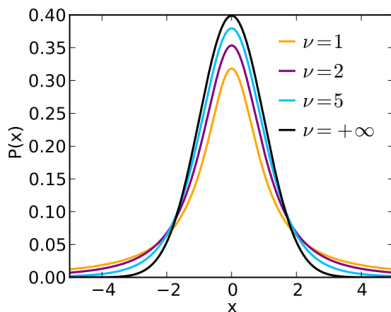
- Let's examine the complete distributions of the sample mean and variances.
- $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i \sim \mathcal{N}(\mu, \sigma^2/N)$, or $\frac{\bar{X}-\mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0, 1)$
- $S^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$; $\frac{N-1}{\sigma^2} S^2 \sim \chi^2(N-1)$ and not, counter-intuitively, $\chi^2(N)$.
- To know how well the sample mean approach μ , we need to rely on sample variance.
- $\frac{\bar{X}-\mu}{S/\sqrt{N}} \sim t(N-1)$: t-distribution with $N-1$ degrees of freedom, the distribution of the location of the sample mean relative to the true mean, divided by the sample standard deviation, after multiplying by the standardizing term \sqrt{N} .
- The t-distribution can be used to estimate whether any given range would contain the true mean. When a true underlying distribution is known to be Gaussian, although with unknown σ , the resulting estimated distribution follows the Student's t distribution.

Examples of distributions: Student's t-distribution



- Notation $X \sim t(\nu)$ where ν is the number of degrees of freedom. Typically $\nu = N - 1$ for N samples/measurements.
- PDF: $f_{\nu}(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi} \Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
- CDF: $F_{\nu}(x) = \frac{1}{2} + x \Gamma\left(\frac{\nu+1}{2}\right) \frac{{}_2F_1\left(\frac{1}{2}, \frac{\nu+1}{2}; \frac{3}{2}; -\frac{x^2}{\nu}\right)}{\sqrt{\pi\nu} \Gamma(\frac{\nu}{2})}$ with ${}_2F_1$ the hypergeometric function.

Examples of distributions: Student's t-distribution

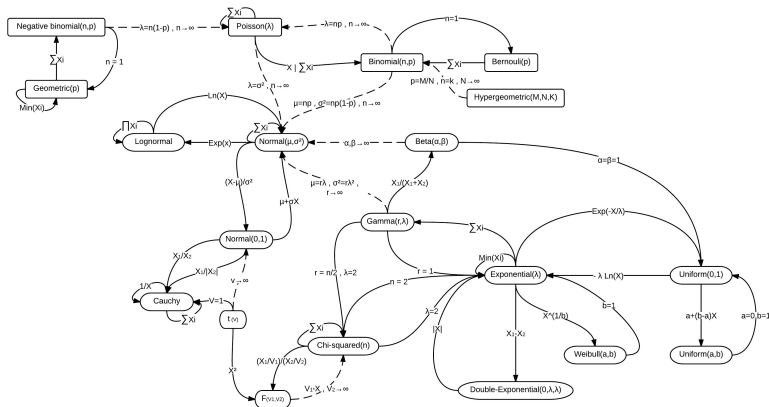


- For small ν , Student's t distribution has heavier tails than the normal distribution. More extreme values are likely to arise. The distribution of the sample mean drawn from a normal population of unknown variance is not normal but follow the Student's t distribution. This most often arises when the sample size is small.
- For larger ν , this distribution tends to normal (see the $+\infty$ above).

Less well-known, but of interest in physics and astronomy:

- Exponential distribution: measures the distribution of time intervals between Poisson events, i.e. a process in which events occur continuously and independently at a constant average rate: e.g. length of time between phone calls, length of time until laptop failure.
- lognormal distribution: distribution of a random variable whose logarithm is normally distributed

Relationships between distributions



- Even at low N , random variables following other distributions can often be transformed into normal variables. Anscombes transform $G(P) \mapsto 2\sqrt{P + \frac{3}{8}}$, P Poisson variable.

Generating random numbers following a given distribution

- Most (all) computer languages have an implementation of the uniform distribution between 0 and 1 (in Julia: `rand()`).
- To generate random numbers following a given distribution, we use the analytic expression of the inverse of the CDF to force $F(X)$ to follow a uniform distribution: $F(X) \sim \mathcal{U}(0, 1) \rightarrow X \sim F^{-1}(\mathcal{U}(0, 1))$.

- Proof:

$$F_X(x) = P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$$

- When there is no analytic solution, the inverse may be found numerically since $P(X)$ is a increasing monotonic function of X .
- Exercice: try to generate random numbers following the exponential distribution which PDF is given here: $f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$

Distances between distributions - K-S & Kuiper's tests

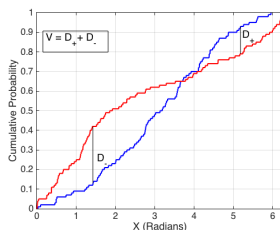
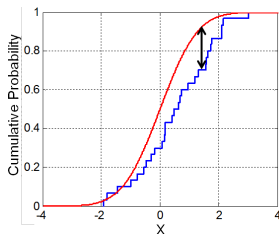
- You may want to compare the distributions of two samples, $X_{1,i}$, $i = 1 \dots n_1$ and $X_{2,i}$, $i = 1 \dots n_2$ ("two-sample test").
- Or you may want to test if it is likely a sample is drawn from a given known distribution ("exact one-sample test").
- Empirical Distribution Functions: $F_1(x) = \frac{1}{n} \sum_{i=1}^n I(x_i \leq x)$ where I is the indicator function, here equal to 1 if $X_i \leq x$ and equal to 0 otherwise. For samples, $F_1(x) = n(x)/n_1$, $n(x)$ the number of X_1 values $\leq x$. Do the same for the second sample, $F_2(X_2)$.
- The Kolmogorov-Smirnov metric measures the distance between two distributions:

$$D_{KS} = \sup |F_1(X_1) - F_2(X_2)|$$

- Kuiper's test is a variant:

$$D_{\text{Kuiper}} = \sup(|F_1(X_1) - F_2(X_2)|) + \sup(|F_2(X_1) - F_1(X_2)|)$$

Distances between distributions - KS & Kuiper's tests



- Both tests are *frequentist* tests, testing the *null hypothesis* (see later in these slides) that the distributions are equal. They are often used to test versus a reference distribution. If $D_n > Q(\alpha)$, the hypothesis "x follows F_{test} distribution" fails at the confidence level $1 - \alpha$ (typically $\alpha = 1\%$ or 5%). $Q(\alpha) \simeq \sqrt{-\frac{1}{2} \log \frac{\alpha}{2}}$.
- For KS and $n > 10$, $Q(\alpha) \simeq 1.63/\sqrt{n}$ for $\alpha = 1\%$ and $Q(\alpha) \simeq 1.36/\sqrt{n}$ for $\alpha = 5\%$, where $n = \frac{n_1 n_2}{n_1 + n_2}$.

Distances between distributions - KS & Kuiper's tests

- Caveats: K-S test only applies to continuous distributions; if testing versus a reference distribution, it must be fully specified (i.e. parameters not fitted from data); the test is more sensitive near the center of the distribution than at the tail.
- Other tests exist: Lilliefors (similar to KS but makes use of sampled means/variances), Cramer-von Mises/Watson and Anderson-Darling (using quadratic distances), Shapiro–Wilk's test (test for Normality).
- Some tests are tailored to compare only specific statistics of sample distributions (e.g. comparing the means): U test, Mann-Whitney-Wilcoxon test.
- Some tests are parametric: t test (check if two sample means are significantly different), f test (check if two Normal distributions have the same variance).
- There is extensive literature on the benefits/drawback of all these methods.

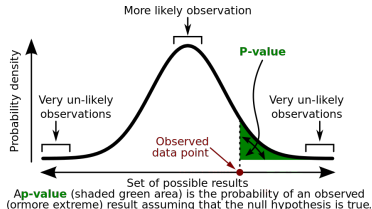
Two fallacies to start thinking about probabilities

- **Prosecutor's fallacy:** the murderer has a tatoo, there is a $1/1,000,000$ chance of anyone having the same tatoo, therefore the accused (who has the tatoo) has $1/1,000,000$ chances of being innocent.
- **Defense attorney's fallacy:** there are 320 millions people in the US, therefore around 320 matches for this tatoo, and my client has only $1/320$ chance of being the murderer.

Frequentist approach to probabilities

- D: data
- M: Model
- Everything to the right of " $|$ " means: "on the condition that these have occurred", or "given these are true"
- Example: probability of $D = 5$ given $M = \mathcal{N}(3, 9)$
- Frequentist statistics assigns: $\Pr(D|M)$
- Frequentist probabilities are understood as though experiments where a population is repeatedly sampled and probabilities are used to express the proportions of outcomes.
- A model is rejected if $\Pr(D|M)$ is below a chosen threshold.

Frequentist approach to hypothesis testing: p-value



- p-value is the number one statistics used in "soft" sciences.
- **Null hypothesis:** default hypothesis (e.g. data originates from random noise, default distribution, etc.).
- **Alternative hypothesis:** any model different from the null hypothesis.
- p-value is the probability of getting our current result or an even less probable one if the null hypothesis were true.
- p-value is the **conditional probability** $\Pr(X \geq x | H_0)$
- X is often not an observation/measurement but a **test statistics** (e.g. KS test, χ^2 , other quantity derived from observations).

- Example: what is the p-value of getting 5 heads in a row: what can we define as the test statistics ? Null hypothesis ? Alternative hypothesis ?

Interpretation of p-value

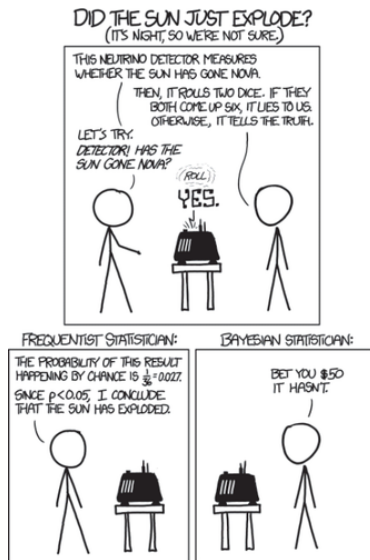
- Example: **Null hypothesis** H_0 : a coin is fair, H : coin is not fair/ biased in some way. H and the null hypothesis H_0 are mutually exclusive and have to be the only possibilities. $\Pr(H) \neq \Pr(H_0)$ in general, but 50%-50% here.
- Applied to χ^2 : we have a set of data at hand onto which we fitted a model, and found χ_0^2 , then
p-value = $\Pr(\chi^2 \geq \chi_0^2 | H_0) = 1 - \Pr(\chi^2 < \chi_0^2 | H_0)$.
- Exercice: compute p-value at 5-sigma.
- The p-value is often compared to a significance level α , the proportion of false alarm or detection we are willing to tolerate. Typically $\alpha = 5\%$ or $\alpha = 1\%$. **We can only conclude one of two things:**
 - if $p > \alpha$, H_0 is rejected
 - if $p < \alpha$, H_0 cannot be rejected at that significance level (beware ! this not imply that H_0 is true).
- α gives the rate of falsely rejecting the null hypothesis
- The p-value is the lowest α for which the null hypothesis can be rejected for a given data set.

Frequent issue: misunderstanding the significance of p-value

- The p-value is a minefield, very easy to misunderstand unless studied properly.
- A p-value is a **frequentist** tool and **cannot be used to figure out the probability of any hypothesis being true** (including H and H_0).
- The p-value $p = \Pr(X \geq x | H_0)$ is not:
 - $\Pr(X)$: probability of getting the data we got
 - $\Pr(H_0)$: probability the null hypothesis is true
 - $\Pr(H)$: probability our hypothesis is true
 - $\Pr(H|X)$: probability our hypothesis is true given the data
 - $\Pr(X|H_0)$: the probability that a finding is "merely a fluke" or "the results are due to chance", i.e. that H_0 is correct.
 - the probability of falsely rejecting H_0
 - the probability that replicating the experiment would yield the same conclusion

Bayesian approach to probabilities

- Bayesian statistics assigns probabilities to models given the data $\Pr(M|D)$ and to models and data themselves $\Pr(M)$ and $\Pr(D)$!
- Bayesian probabilities update model probabilities based on new data.
- Models are not rejected, just assigned low probabilities



Model-fitting: likelihood

- We obtain N data points through experiments $\mathbf{D} = \{d_1, \dots, d_N\}$ with $\sigma = \{\sigma_1, \dots, \sigma_N\}$ (heteroskedasticity). We also have a model \mathbf{M} entirely described by parameters $\theta = \{\theta_1, \dots, \theta_p\}$.
- The **likelihood** of the model \mathbf{M} or θ given the data is a function of θ (given \mathbf{D}) equal to the probability of the observed data given those parameter values,

$$\mathcal{L}(\mathbf{M}|\mathbf{D}) = \mathcal{L}(\theta|\mathbf{D}) := \Pr(\mathbf{D}|\theta)$$

- The likelihood is **not** a probability density function which is a function of \mathbf{D} (given θ). It is in fact not a probability and generally unnormalized, i.e. $\int_{\theta} \mathcal{L}(\theta|\mathbf{D})d\theta \neq 1$, while $\int_{\mathbf{D}} \Pr(\mathbf{D}|\theta)d\mathbf{D} = 1$.
- The likelihood is not $\Pr(\theta|\mathbf{D})$, i.e. the probability that θ is the right one given \mathbf{D} , as Bayes' theorem will demonstrate.
- Probability theory studies processes modeled by random variables; statistical inference tries to find models that explain given observations. Model fitting is statistical inference.

Model-fitting: likelihood and χ^2

- For independent data points,

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{D}) = \Pr(\mathbf{D}|\boldsymbol{\theta}) = \prod_i^N \Pr(d_i|\boldsymbol{\theta}) = \prod_i^N \Pr(d_i|m_i)$$

where $\mathbf{m}(\boldsymbol{\theta}) = \{m_1, \dots, m_N\}$ are the model-predicted values.

- The **log-likelihood** is

$$\log \mathcal{L}(\boldsymbol{\theta}|\mathbf{D}) = \sum_i^N \log \Pr(d_i|m_i)$$

- In particular if we assume the data normally distributed, the log-likelihood is:

$$\sum_i^N \log \Pr(d_i|m_i) = \sum_i^N \log \frac{e^{-\frac{(d_i-m_i)^2}{2\sigma_i^2}}}{\sigma_i\sqrt{2\pi}} = \text{cnst} - \frac{1}{2} \sum_i^N \left(\frac{d_i - m_i}{\sigma_i} \right)^2$$

Model-fitting and Maximum likelihood

- The most likely values for the model parameters θ given the data \mathbf{D} are found by maximizing the likelihood.
- Maximizing the likelihood is minimizing the negative log-likelihood:

$$\begin{aligned}\operatorname{argmax}_{\theta} \mathcal{L}(\theta|\mathbf{D}) &= \operatorname{argmin}_{\theta} \{-\log \mathcal{L}(\theta|\mathbf{D})\} \\ &= \operatorname{argmin}_{\theta} \left(\text{cnst} + \frac{1}{2} \sum_i^N \left(\frac{d_i - m_i(\theta)}{\sigma_i} \right)^2 \right) \\ &= \operatorname{argmin}_{\theta} \chi^2(\theta)\end{aligned}$$

- χ^2 **minimization** results from applying the **maximum likelihood** approach to a model-fitting problem with normally-distributed data.

Application of Maximum likelihood: Inverse variance weighting

- We attempt to **combine independent estimates** of a single quantity, using data $\{d_i, \sigma_i^2\}$, $i = 1 \dots N$, d_i is assumed/measured to be normally distributed with uncertainty variances σ_i^2 . E.g. we could want to combine $T = 200 \pm 50K$ and $T = 326 \pm 12K$.

Application of Maximum likelihood: Inverse variance weighting

- We want to obtain the most probable estimate for our model value m , plus an error bar. As χ^2 is a quadratic function of m , $\tilde{\mu}$ can be found by differentiation:

$$\tilde{m} = \underset{m}{\operatorname{argmin}} \sum_i \left(\frac{d_i - m}{\sigma_i} \right)^2 \implies 0 = \sum_i \frac{2}{\sigma_i^2} (d_i - \tilde{m}) \implies \tilde{m} = \frac{\sum_i \frac{d_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}}$$

- This is the **inverse-variance weighted average** estimator, whose variance $\operatorname{Var}\{\tilde{m}\} = (\sum_i 1/\sigma_i^2)^{-1}$ can be shown to have the least variance among all weighted averages.
- For $T = 200 \pm 50\text{K}$ and $T = 326 \pm 12\text{K}$,
$$\tilde{m} = \frac{200/50^2 + 326/12^2}{1/50^2 + 1/12^2} \pm \sqrt{\frac{1}{1/50^2 + 1/12^2}} \simeq 319 \pm 12 \text{ K}.$$

Bayesian terms and Bayes' theorem

- The following equations are valid for any ensembles, but here I give them in the context of model-fitting
- $\Pr(M, D) = \Pr(M \cap D)$: joint probability (here of model and data)

$$\Pr(M, D) = \Pr(D) \Pr(M|D)$$

but we also have

$$\Pr(M, D) = \Pr(M) \Pr(D|M)$$

- $\Pr(M|D)$: posterior probability of the model
- $\Pr(M)$: prior probability of model, sometimes written $\pi(M)$
- $\Pr(D|M)$ probability of the data given the model, or interpreted as likelihood of the model $\mathcal{L}(M|D)$
- $\Pr(D)$: evidence, marginal likelihood, "probability of the data"
- Bayes' theorem results from eq. (1) and (2)

$$\Pr(M|D) = \frac{\Pr(D|M) \Pr(M)}{\Pr(D)}$$

Evidence expression for mutually exclusive outcomes

- If there are only two possible models, M and \bar{M} (e.g. a statement is true or false), then the marginalization gives:

$$\Pr(D) = p(M)p(D|M) + p(\bar{M})p(D|\bar{M})$$

- Bayes' equation becomes:

$$\Pr(M|D) = \frac{\Pr(D|M) \Pr(M)}{\Pr(M) \Pr(D|M) + \Pr(\bar{M}) \Pr(D|\bar{M})}$$

Marginalization of likelihood/evidence

- More generally, for mutually exclusive models/events M_i , $i = 1 \dots N$, $M_i \cap M_j = \emptyset$, but exhaustive $M_1 \cup \dots \cup M_N = \mathbb{U}$ then

$$\Pr(D) = \Pr((D \cap M_1) \cup \dots \cup (D \cap M_N)) = \sum_i \Pr(D|M_i) \Pr(M_i)$$

- Very often applied to a continuous model parameter θ

$$\Pr(\theta|D) = \frac{\Pr(D|\theta) \Pr(\theta)}{\Pr(D)}$$

- The marginal likelihood $P(D)$ can be understood as a normalization factor, so that the posterior $\Pr(\theta|D)$ is normalized
- **Bayesian marginalization** of the likelihood over the parameter θ :

$$\Pr(D) = \int_{\theta} \Pr(D, \theta) d\theta = \int_{\theta} \Pr(D|\theta) \Pr(\theta) d\theta = \int_{\theta} \mathcal{L}(\theta|D) \pi(\theta) d\theta$$

- The parameter has been marginalized = "**integrated out**"

Thinking about probabilities with Bayes

- **Prosecutor's fallacy:** the murderer has a tatoo, there is a $1/1,000,000$ chance of anyone having the same tatoo, therefore the accused who has the tatoo has $1/1,000,000$ chances of being innocent.
- Bayesian analysis: I = innocent, T =tatoo. We have

$$\Pr(I|T) = \frac{\Pr(T|I) \Pr(I)}{\Pr(T)} \quad (1)$$

- $\Pr(T)$ is the unconditional probability of having an accused with this tatoo. If the accused was picked out of a database of people who have this tatoo, we have a **data fishing issue**.
- $\Pr(I)$ is the unconditional probability of being innocent irrespective of tatoo (depends on other suspects, rest of the evidence, existence other suspects).
- The mistake in the prosecutor's fallacy is to considers $\Pr(I|T) \simeq \Pr(T|I)$ i.e. that $\Pr(I) \simeq \Pr(T)$, which is not true in general.

Thinking about probabilities with Bayes

- **Defense attorney's fallacy:** there are 320 millions people in the US, therefore around 320 matches for this tatoo, and my client has only $1/320$ chance of being the murderer.
- Defense: forgot that the probability of anyone in the general population to be arrested in connection with the murder is likely to be low. Which one is this ? Think about this at home.

Application of Bayes theorem: stellar survey

- Stellar survey in the neighborhood find only three types of stars:

Type	% total	Probability of multiplicity
F	0.45	0.10
G	0.15	0.45
K	0.40	0.20

- I randomly select a multiple star, what is the probability it is F-type ?
- Be extra careful when setting up your Bayesian analysis. What is being observed ? What constitutes a "model" ?

Application of Bayes theorem: stellar survey solution

Type	% total	Probability of multiplicity
F	0.45	0.10
G	0.15	0.45
K	0.40	0.20

- We want $\Pr(F|M)$ and we have $\Pr(M|F) = 0.1$ and $\Pr(F) = 0.45$.
- Bayes' theorem: $\Pr(F|M) = \frac{\Pr(M|F) \times \Pr(F)}{\Pr(M)}$
- We still need $\Pr(M)$, and we can marginalize over the stellar type to find it (as we know we only have 3 possibilities/types):
$$\Pr(M) = \Pr(M|G) \Pr(G) + \Pr(M|F) \Pr(F) + \Pr(M|K) \Pr(K) = .1925$$
- So the answer is $\Pr(F|M) = \frac{.1 \times .45}{.1925} = .23$

Bayesian statistics: the zombie disease example

- Only 1% of the population who participate to a zombie plague screening have the plague. 95% of plague carriers will get tested positive, and 10% of non-carriers also get tested positive. Fabien just got tested positive. What is the probability he is infected ?
- Sensitivity of the test: prob. of true positives
- Specificity of the test: prob. of true negatives, here $100\% - 10\% = 90\%$
- Type I errors: prob. of false positives, i.e. testing someone as positive who is in fact negative, given to be 10% here (other example: computer anti-virus flagging a safe file)
- Type II errors: prob. of false negatives, i.e. testing someone as negative who is in fact positive, here $100\% - 95\% = 5\%$ (example: computer anti-virus failing to detect actual viruses)

Bayesian statistics: the zombie disease example

- Only 1% of the population who participate to a zombie plague screening have the plague. 95% of plague carriers will get tested positive, and 10% of non-carriers also get tested positive. Fabien just got tested positive. What is the probability he is infected ?
- M : Fabien is infected, D : tested positive
- Prior probability $\Pr(M) = .01$, and $\Pr(\bar{M}) = .99$
- Sensitivity: $\Pr(D|M) = .95$, specificity: $\Pr(\bar{D}|\bar{M}) = 1 - \Pr(D|\bar{M}) = .9$
- Prob. type I: $\Pr(D|\bar{M}) = .1$, Prob. type II $\Pr(\bar{D}|M) = 1 - \Pr(D|M)$.

$$\Pr(M|D) = \frac{\Pr(D|M) \Pr(M)}{\Pr(M) \Pr(D|M) + \Pr(\bar{M}) \Pr(D|\bar{M})} \simeq 8.8\%$$

- What is the probability he is infected if a second test also comes positive ?

Changing variables in a Bayesian context

- Changing variables requires using the **jacobian** modulus $\left| \frac{\partial Y}{\partial X} \right|$:

$$X \mapsto Y = f(X), f \text{ monotonic} \implies \Pr(X) = \Pr(Y) \left| \frac{\partial Y}{\partial X} \right|$$

- if $Y = f(X)$ is not a one to one transformation, $\Pr(X)$ is given by the sum over all values of Y which correspond to X .
- Generalization:

$$\Pr(X_i) = \Pr(Y_i) \times \left| \frac{\partial(Y_1, Y_2, \dots, Y_n)}{\partial(X_1, X_2, \dots, X_n)} \right|$$

- Multivariate **jacobian determinant**:

$$\left| \frac{\partial(Y_1, Y_2, \dots, Y_n)}{\partial(X_1, X_2, \dots, X_n)} \right| = \det \begin{bmatrix} \frac{\partial Y_1}{\partial X_1} & \dots & \frac{\partial Y_1}{\partial X_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial Y_m}{\partial X_1} & \dots & \frac{\partial Y_m}{\partial X_n} \end{bmatrix}$$

Changing variables in a Bayesian context: example

- Isotropic bivariate Gaussian given as a function of Cartesian coordinates (x, y)

$$\Pr(x, y) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{x^2 + y^2}{2\sigma^2}\right]$$

- We want it as a function of (r, θ) ... and maybe just r ?

Changing variables in a Bayesian context: example

- Isotropic bivariate Gaussian given as a function of Cartesian coordinates (x, y)

$$\Pr(x, y) = \frac{1}{2\pi\sigma^2} \exp \left[-\frac{x^2 + y^2}{2\sigma^2} \right]$$

- We want it as a function of (r, θ) .

- $x = r \cos \theta, y = r \sin \theta, \left| \frac{\partial(x, y)}{\partial(r, \theta)} \right| = \begin{vmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{vmatrix} = r$

$$\Pr(r, \theta) = \Pr(x, y) \left| \frac{\partial(x, y)}{\partial(r, \theta)} \right| = \frac{r}{2\pi\sigma^2} \exp \left[-\frac{r^2}{2\sigma^2} \right]$$

- Marginalization:

$$\Pr(r) = \int_0^{2\pi} \Pr(r, \theta) d\theta = \frac{r}{\sigma^2} \exp \left[-\frac{r^2}{2\sigma^2} \right]$$

Bayesian error propagation: $Z = f(X, Y)$

- Propagation of uncertainties, for $X = x_0 \pm \sigma_x$ and $Y = y_0 \pm \sigma_y$

$$\begin{aligned}\Pr(Z) &= \iint \Pr(Z|X, Y) \Pr(X, Y) dX dY \\ &= \iint \Pr(Z - f(X, Y)) \Pr(X, Y) dX dY \\ &= \iint \Pr(X, Y) \delta(Z - f(X, Y)) dX dY\end{aligned}$$

- if X and Y are independent: $\Pr(X, Y) = \Pr(X) \Pr(Y)$

$$\Pr(Z) = \int dX \Pr(X) \int \Pr(Y) \delta(Z - f(X, Y)) dY$$

Error propagation: $Z = X + Y$

- For $Z = X + Y$

$$\begin{aligned}\Pr(Z) &= \int \Pr(X) \Pr(Y = Z - X) dX \\ &= \frac{1}{2\pi\sigma_x\sigma_y} \int_{-\infty}^{+\infty} \exp\left[-\frac{(X - x_0)^2}{2\sigma_x^2}\right] \exp\left[-\frac{(Z - X - y_0)^2}{2\sigma_y^2}\right] dX \\ &= \frac{1}{\sqrt{2\pi}\sigma_z} \exp\left[-\frac{(Z - z_0)^2}{2\sigma_z^2}\right]\end{aligned}$$

with $z_0 = x_0 + y_0$ and $\sigma_z^2 = \sigma_x^2 + \sigma_y^2$.

- $Z = X - Y$ gives the same PDF, with $z_0 = x_0 - y_0$ and $\sigma_z^2 = \sigma_x^2 + \sigma_y^2$.
- This method is more general than "classic" error propagation

Classic error propagation

- For $Z = X - Y$, $\delta X = X - x_0$, $\delta Y = Y - y_0$, $(\delta X)^2 = \sigma_x^2$, $(\delta Y)^2 = \sigma_y^2$

$$\delta Z = \delta X - \delta Y$$

$$(\delta Z)^2 = (\delta X)^2 + (\delta Y)^2 - 2\delta X\delta Y$$

$$\langle (\delta Z)^2 \rangle = \langle (\delta X)^2 \rangle + \langle (\delta Y)^2 \rangle - 2 \langle \delta X \delta Y \rangle$$

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2\sigma_{xy}$$

- For $Z = X/Y$,

$$\delta Z = \frac{Y\delta X - X\delta Y}{Y^2}$$

$$\frac{\delta Z}{Z} = \frac{\delta X}{X} - \frac{\delta Y}{Y}$$

$$\frac{\langle (\delta Z)^2 \rangle}{z_0^2} = \frac{\langle (\delta X)^2 \rangle}{x_0^2} + \frac{\langle (\delta Y)^2 \rangle}{y_0^2} - 2 \frac{\langle \delta X \rangle \langle \delta Y \rangle}{x_0 y_0}$$

$$\frac{\sigma_z^2}{z_0^2} = \frac{\sigma_x^2}{x_0^2} + \frac{\sigma_y^2}{y_0^2} - 2 \frac{\sigma_{xy}}{x_0 y_0}$$

The square modulus problem

- Classic error propagation two main drawbacks:
 - Relies on expansion around a center value which may sometimes fail when the function is submitted to constraints
 - Hides the fact that the PDF of the result often has changed nature (e.g. $Z = XY$ or $Z = X/Y$ are strongly non-Gaussian).
- Example of power spectrum measurement: $S = |X|^2$, $S = s_0 \pm \sigma_S$. S is normally distributed, and we know $X \geq 0$
- We would like to find the most probable distribution and give an answer under the form: $X_0 \pm \sigma_X$. How should we proceed ?

Failures of classic error propagation: square modulus

- We would like the most probable X , the classic analysis would set $X_0 = \sqrt{S_0}$ and

$$\begin{aligned}\delta S &= 2X_0\delta X \\ \langle (\delta S)^2 \rangle &= 4X_0^2 \langle (\delta X)^2 \rangle = 4S_0 \langle (\delta X)^2 \rangle \\ \sigma_S^2 &= 4S_0\sigma_X^2\end{aligned}$$

- This leads to $X = \sqrt{S_0} \pm \frac{\sigma_S}{2\sqrt{S_0}}$, which obviously will not work when $S_0 < 0$, e.g. $S = 8 \pm 5$ works but not $S = -1 \pm 4$
- We failed to express the problem in a Bayesian framework
- **Bayesian inference** encompasses all our knowledge of the problem by using as needed three tools: **Bayes' theorem, marginalization, and change of variables**

Success of Bayesian analysis: square modulus

- Given the data $D = \{S_0, \sigma_S\}$, we are looking at the most probable value for X , i.e. we are looking for $X_0 = \operatorname{argmax} \Pr(X|D)$
- What we means by measurement $S = S_0 \pm \sigma_S$ is that as a function of S the probability of the data is Gaussian and expressed by:

$$\Pr(D|S) = \exp \left[-\frac{(S - S_0)^2}{2\sigma_S^2} \right] \quad (2)$$

- We also have a prior on S :

$$\Pr(S) = \begin{cases} 0 & S < 0 \\ \text{const} & S \geq 0 \end{cases} \quad (3)$$

- Due to the prior, central expansion (classic error propagation) fails in this case, but Bayesian change of variable still works
- Finally, Bayes equation allows us to express our whole knowledge of S :

$$\Pr(S|D) = \Pr(D|S) \Pr(S) \quad (4)$$

Success of Bayesian analysis: square modulus

- For $X < 0$, $\Pr(X|D) = 0$, and for $X \geq 0$:

$$\Pr(X|D) = \left| \frac{\partial S}{\partial X} \right| \Pr(S|D) = 2X \exp \left[-\frac{(X^2 - S_0)^2}{2\sigma_S^2} \right]$$

- This is a non-Gaussian function, and truncated for negative X !
- We want a **Gaussian approximation** so that $X_0 = \underset{X}{\operatorname{argmax}} \Pr(X|D)$:

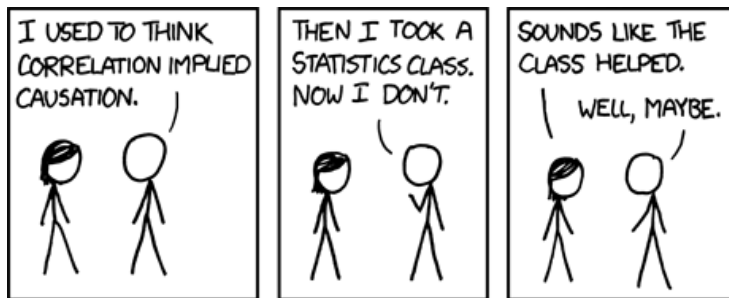
$$\left. \frac{\partial \log \Pr(X|D)}{\partial X} \right|_{X_0} = 0 \implies X_0 = \left[\frac{S_0 + (S_0^2 + 2\sigma_S^2)^{\frac{1}{2}}}{2} \right]^{\frac{1}{2}}$$

$$\left. \frac{\partial^2 \log \Pr(X|D)}{\partial X^2} \right|_{X_0} = -\frac{1}{\sigma_X^2} \implies \sigma_X = \left[\frac{1}{X_0^2} + \frac{2(3X_0^2 - S_0)}{\sigma_S^2} \right]^{-\frac{1}{2}}$$

True/false ?

- *Correlation is not evidence of causation*, Anonymous
- *Absence of evidence is not evidence of absence*, Martin Rees
- *Knowing that apples are green does not tells you anything about the color of ravens*, Bob the physicist

Demonstrate the following



These are **true** as Bayesian statements:

- *Correlation is not causation but it sure is a hint*, Edward Tufte
- *Absence of evidence is evidence of absence*
- *Knowing that apples are green provide evidential support to the proposition "All ravens are black"*, Carl Gustav Hempel
- Note that A is evidence of/for B if $Pr(B|A) > Pr(B)$, or if $Pr(B|A) > Pr(B|\bar{A})$.

Bayesian statistics and model-fitting: Maximum a posteriori

- Based on the same principle as Maximum likelihood
- At the root of many numerical methods we are going to see:
 - Monte Carlo Markov Chains
 - Error propagation
 - Optimization for model-fitting
 - Deconvolution
 - ...
- Maximum a posteriori (MAP) method: we want to find the most probable parameter θ of a model given the data, i.e. maximize $\Pr(\theta|D)$
- For a given data set, $\Pr(D)$ is constant:

$$\operatorname{argmax}_{\theta} \Pr(\theta|D) \propto \operatorname{argmax}_{\theta} \{\Pr(D|\theta) \Pr(\theta)\} = \operatorname{argmax}_{\theta} \{\mathcal{L}(\theta|D) \pi(\theta)\}$$

Examples of priors

- Prior knowledge about the problem: constraints or preferences on parameters, knowledge of their distribution, ...
- Constraints: positivity, specific bounds (e.g. ratio between 0 and 1)
- Preferences: smoothness, sparsity, correlation between parameters.
- Prior chosen as **non-committal** as possible in order not to bias the result. **Uninformative** priors express weak information about the problem
- An **improper prior** is not truly a prior, but a function that does not normalize to unity or only does so by assigning an infinitesimal probability to each value e.g. the scale $\pi(\theta) = \frac{1}{\theta}$. It can be made into a prior by normalizing it within reasonable bounds.
- Example: we measure $\theta = 40 \pm 10$ (normal distribution) and we know that $\Pr(\theta)$ follows $\pi(\theta) \propto \frac{1}{\theta}$, and that $1 \leq \theta \leq 100$. What is the most probable θ ?

Inverse problems and regularization

- **Inverse problem:** recovering a wanted set of parameters from noise-corrupted data (e.g. pixel fluxes in image reconstruction)
- **Ill-posed** inverse problem: when the effective number of parameters to recover is greater than the effective number of data points, the solution may not exist or may not be unique
- **Regularization:** introduction of priors (known as regularizers) which will try to compensate for our lack of data. As usual, the prior/regularizer expresses the *a priori* probability of an object when we have no data.
- Some regularizers can take the convenient form:

$$\Pr(\theta) \propto \exp[-R(\theta)] \implies -\log \Pr(\theta) = R(\theta)$$

- The solution $\tilde{\theta}$ of MAP for Gaussian-distributed data and N regularizers with weights w_1, w_2, \dots is found as:

$$\tilde{\theta} = \underset{\theta}{\operatorname{argmin}} \left[\chi^2(\theta) + \sum_{k=1}^N w_k R_k(\theta) \right]$$

A famous regularizer: the Maximum Entropy

- The Shannon entropy $S(\theta)$ is a **non-committal prior**. When used as regularizer, it expresses the expected absence of correlations between the parameters to recover (i.e. we expect θ_i and θ_j to be uncorrelated):

$$S(\theta) = - \sum_i \theta_i \log \theta_i$$

- Maximum Entropy is most often (but not exclusively) applied to a set of proportions, i.e. when $\sum \theta_i = 1$.
- In our Maximum a Posteriori framework, $R(\theta) = -S(\theta)$ and so we actually want to maximize entropy S or minimize negentropy R .

Maximum Entropy principle

- Maximum entropy is based on the following principle: *If the proportion of some entity which has a given property is known to be p , then the most probable estimate of the proportion in some subclass which has that property is (in the absence of any information connecting the subclass with the property) the same number p*
- We all know that two thirds of physicists are right handed, and that only one in 10 physicists drink Coke (the remaining 9/10th preferring Pepsi). These are true facts. What is the probability that a randomly encountered physicist is a left-handed Coke drinker? Find a unique answer, show that it is equivalent to non-correlation between the parameters.
- What is the regularizer R that correspond to a prior Poisson distribution ? Conclusion ?

Maximum entropy method: solution 1

	Coke	Pepsi	
Left	p	$1/3 - p$	$\Pr(\text{Left}) = 1/3$
Right	$1/10 - p$	$2/3 - 1/10 + p$	$\Pr(\text{Right}) = 2/3$
	$\Pr(\text{Coke}) = 1/10$	$\Pr(\text{Pepsi}) = 9/10$	

- The table above is called a **contingency table**
- Method 1: non-correlation of handedness with soda:
 $\Pr(\text{Coke}|\text{Left}) = \Pr(\text{Coke}|\text{Right})$. We can express the joint probabilities as $\Pr(\text{Coke}, \text{Left}) = \Pr(\text{Coke}|\text{Left}) \Pr(\text{Left})$ and $\Pr(\text{Coke}, \text{Right}) = \Pr(\text{Coke}|\text{Right}) \Pr(\text{Right})$, leading to:

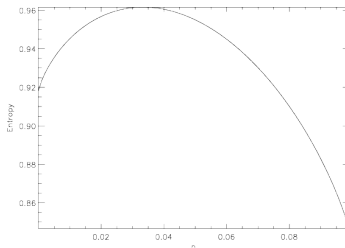
$$\frac{\Pr(\text{Coke}, \text{Left})}{\Pr(\text{Left})} = \frac{\Pr(\text{Coke}, \text{Right})}{\Pr(\text{Right})}$$
$$\implies \frac{p}{1/3} = \frac{1/10 - p}{2/3} \implies p = 1/30$$

Maximum entropy method: solution 1

- Method 2: express the entropy and maximize it

$$S = - \sum_i p_i \log p_i = - \left[p \log p + \left(\frac{1}{3} - p \right) \log \left(\frac{1}{3} - p \right) + \left(\frac{1}{10} - p \right) \log \left(\frac{1}{10} - p \right) + \left(\frac{2}{3} - \frac{1}{10} + p \right) \log \left(\frac{2}{3} - \frac{1}{10} + p \right) \right]$$

- $\frac{\partial S}{\partial p} = 0 \implies p = 1/30$



Bayesian model selection

- We have two models M_1 and M_2 and we want to do **model selection**, i.e. determine the best model given data.
- Applying Bayes theorem in both cases gives the **Bayes factor** K :

$$K_{12} = \frac{\Pr(M_1|D)}{\Pr(M_2|D)} = \frac{\Pr(D|M_1)}{\Pr(D|M_2)} \times \frac{\Pr(M_1)}{\Pr(M_2)} \quad (5)$$

- The **prior odds** $\frac{\Pr(M_1)}{\Pr(M_2)}$ measures if we have an *a priori* preference for one model
- The interpretation of $\frac{\Pr(D|M_1)}{\Pr(D|M_2)}$ depends on the nature of the models being compared: different parameter values for the same law, or different model laws.

Bayesian model selection: interpretation of Bayes factor

- Table from Robert E. Kass and Adrian E. Raftery (1995). "Bayes Factors" (PDF). Journal of the American Statistical Association 90 (430): 791.

$2 \log K$	K	Strength of evidence
0 to 2	1 to 3	weak evidence
2 to 6	3 to 20	positive
6 to 10	20 to 150	strong
> 10	> 150	very strong

Bayesian model selection: example

- Evolution models show that 80% of a certain type of flares are due to G stars, and 20% by F stars. Our new instrument is custom-made to detect this flares, but is more likely to detect G stellar flares than F with probabilities 60% (G) vs 40% (F). We detect a flare, which detection is more likely ?
- Astronomers believe that 70% of the public prefers donuts to cookies; physicists believe that only 60% do. A poll out of 100 people finds 62%. Compute the Bayes factor. Same if astronomers say $70 \pm 5\%$ and physicists $60 \pm 10\%$.

Bayesian model selection: different parameter values

- The two models can be **different parameter values** for a similar underlying model laws. Examples: different μ and σ for a Gaussian, different eccentricity/semi-major axis for heliocentric model
- If the models M_i predict exact values θ_i for the parameter θ , then $\Pr(D|M_i) = \mathcal{L}(\theta_i|D)$ is the likelihood of the model. Then for equiprobable model parameters, K becomes a **likelihood ratio**; and for Gaussian-distributed data, the most probable model parameters are the ones minimizing the χ^2 .
- When the models M_i include a prior distribution on model parameter θ , $\Pr(D|M_i)$ is a **marginal likelihood**, and needs to be found via marginalization, $\Pr(D|M_i) = \int \Pr(D|\theta, M_i) \Pr(\theta|M_i) d\theta$.

Bayesian model selection: different model laws

- The two models can be **different model laws**, i.e. Gaussian vs Poisson, a two-parameter law vs 3-parameter law, heliocentric vs geocentric model, different number of spots on a star, etc.
- $\Pr(D|M_i)$, the marginal likelihood is in this case often referred to as the **evidence** for the model.
- To estimate the evidence, $\Pr(D|M_i)$ needs to be computed by marginalization over the entire set of parameters in M_i .
- Bayes factor is the only valid way of comparing different model laws: the χ^2 or reduced χ^2 approaches are invalid approaches as soon as priors are involved (\simeq always). Bayes factors guards against overfitting as it naturally includes a penalty for including too much model structure, e.g. the evidence for models that can "explain anything" is low. Example: in the 17th century, Bayes factor would have been overwhelmingly in favor of the heliocentric model, even though geocentric had better χ^2 .

Bayesian model selection in practice

- In complex cases, Monte Carlo methods have to be used to compute the evidence. At the moment the best algorithms are the Nested Sampling (for a low number of parameters) and Parallel Tempering.
- They cleverly sample the unnormalized distribution of the posterior probability $\Pr(\theta|D, M_i) \propto \Pr(D|\theta, M_i) \Pr(\theta|M_i)$, then integrate it to estimate $\Pr(D|M_i)$.
- Probabilities/likelihoods are cumbersome on computers due to floating point accuracy, so algorithms rely on log-likelihoods to compute $\log Z_i := \log \Pr(D|M_i)$ for a given model M_i .
- Since $\log Z_i - \log Z_j = \log \frac{\Pr(D|M_i)}{\Pr(D|M_j)} = \log K_{ij}$, models with higher $\log Z$ are more probable.
- There is no way to find the "best ever" model out of nowhere, one can only compare the probabilities of models (given the data). Unless one had a way to devise generic models (e.g. Mixture Models).

Bayesian model selection: other criteria

- Several simplified criteria for model selection have been derived over the years, such as the AIC or BIC.
- The Akaike information criterion (AIC). If k be the number of estimated parameters in the model, and \mathcal{L}_{\max} the maximum value of the likelihood function for the model, then the AIC value of the model is $\text{AIC} = 2k - 2 \log \mathcal{L}_{\max}$. The preferred model is the one with the minimum AIC value. A modified criterion for small sample size n is also available, $\text{AICc} = \text{AIC} + \frac{2k^2 + 2k}{n - k - 1}$.
- The Bayesian Information criterion $\text{BIC} = k \log n - 2 \log \mathcal{L}_{\max}$ where n = the number of data points. Here again the model with the lowest BIC is preferred.
- In a χ^2 model-fitting context, $\text{BIC} = \chi^2 + k \log n$ and $\text{AIC} = \chi^2 + 2k$.
- Assumptions underlying BIC and AIC have to be understood to use them. My take on them is only use them if you cannot estimate the $\log Z$ through adequate algorithms.