PHYS/ASTR 8150
Introduction to probabilities and statistics

Fabien Baron

Georgia State University

baron@chara.gsu.edu

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)

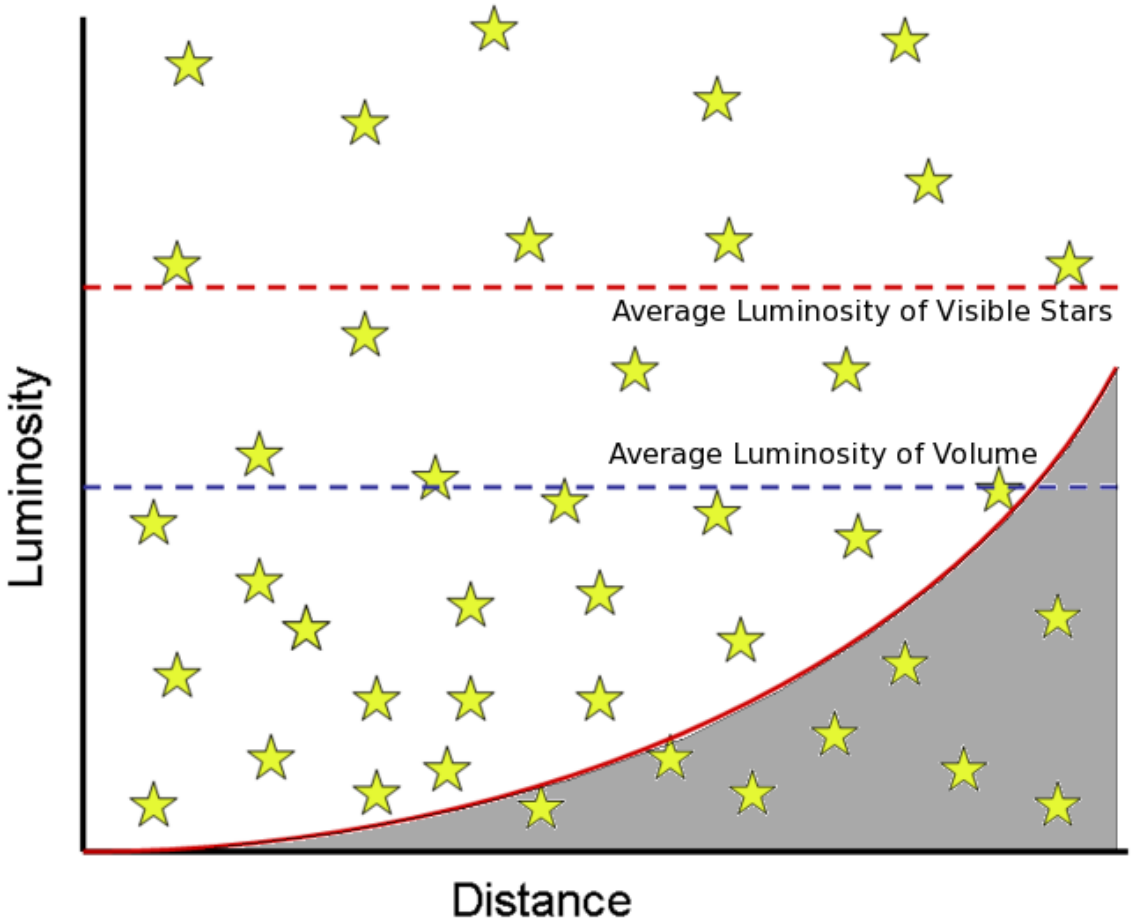
Statistics are a mean to summarize concisely yet rather precisely some of the characteristics of our data, while probabilities are a way to understand how the data is likely to behave.
Both are used to state the uncertainties ("errors") in measurements:

- **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



Biases (2)

- 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

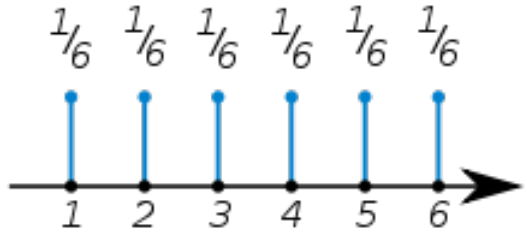
- Data is affected by noise, which prevents from getting to the exact mathematical laws, and behaves as a **random variable**
- Understanding the noise = we may recover the real data
- Probabilities are the best tool to describe the noise **stochastic processes** or the data probability distribution
- 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

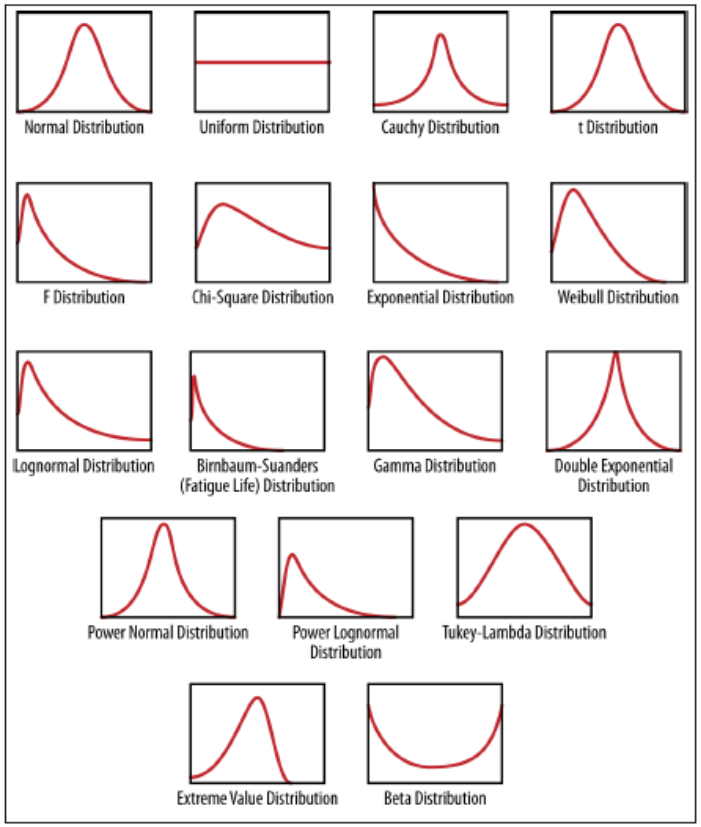


Figure 2-1. A bunch of continuous density functions (aka probability distributions)

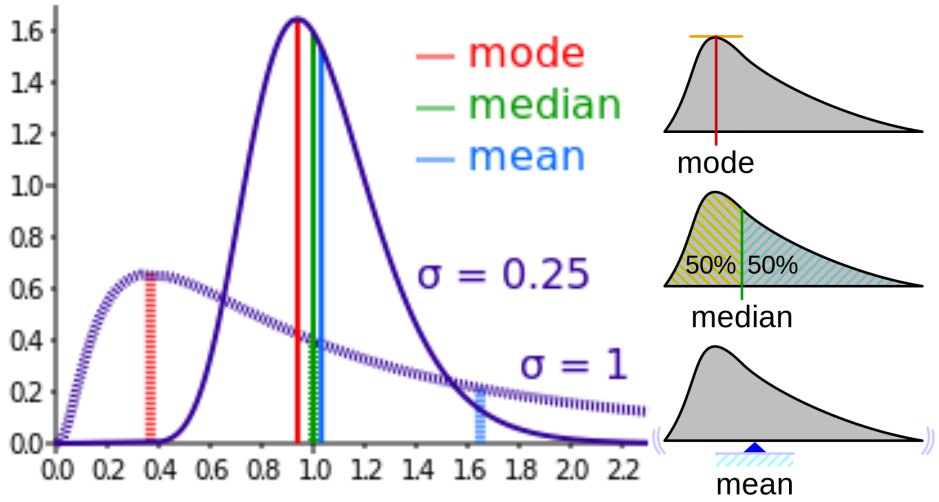
- Continuous case: if X a continuous random variable, the **probability density function** $f_X(X)$ is a continuous probability distribution
- **Support** of a probability function is the set of points where the probability density is not zero-valued
- The probability mass (not density) of getting any 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, $\mathrm{E}[X] = \sum x \Pr(X = x)$
- For a continuous one $\mathrm{E}[X] = \int x f_X(x)\,\mathrm{d}x$, 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 $\mathrm{P}(X \leq x) = \frac{1}{2}$ and $\mathrm{P}(X \geq x) = \frac{1}{2}$.

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

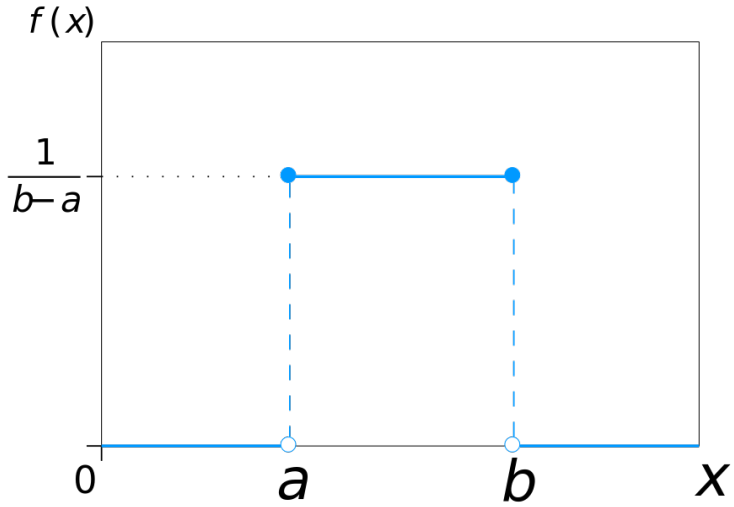
Sample Variance

- 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

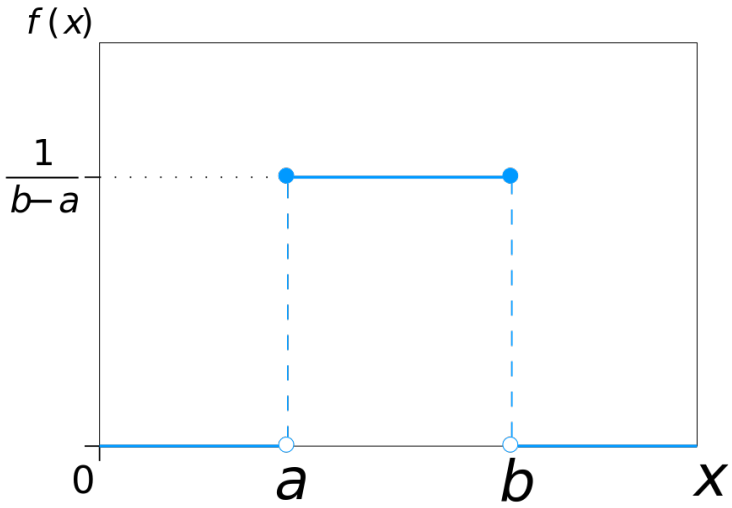
- The **cumulative distribution function** measures the probability that the variable takes a value less than or equal to x.
- CDF: $F_X(x) = \mathrm{P}(X \leq x)$
- Consequently $\mathrm{P}(X > a) = 1 - F_X(a)$
- $\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a)$
- In the continous case: $F_X(x) = \int_{-\infty}^{x}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

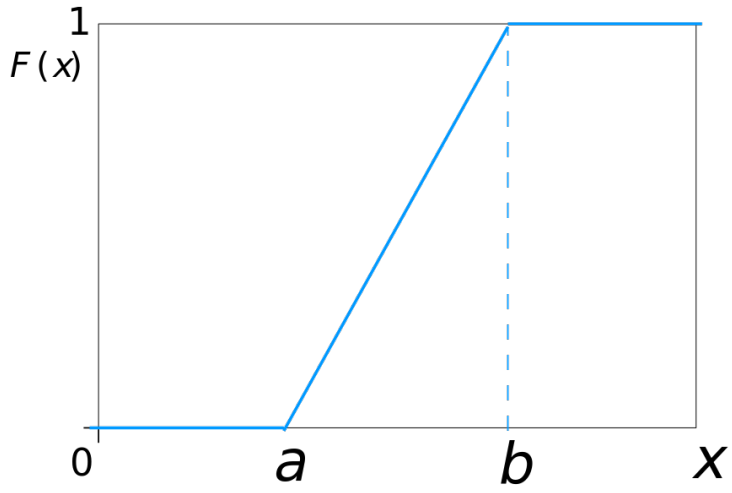
Examples of distributions: continuous uniform distribution



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

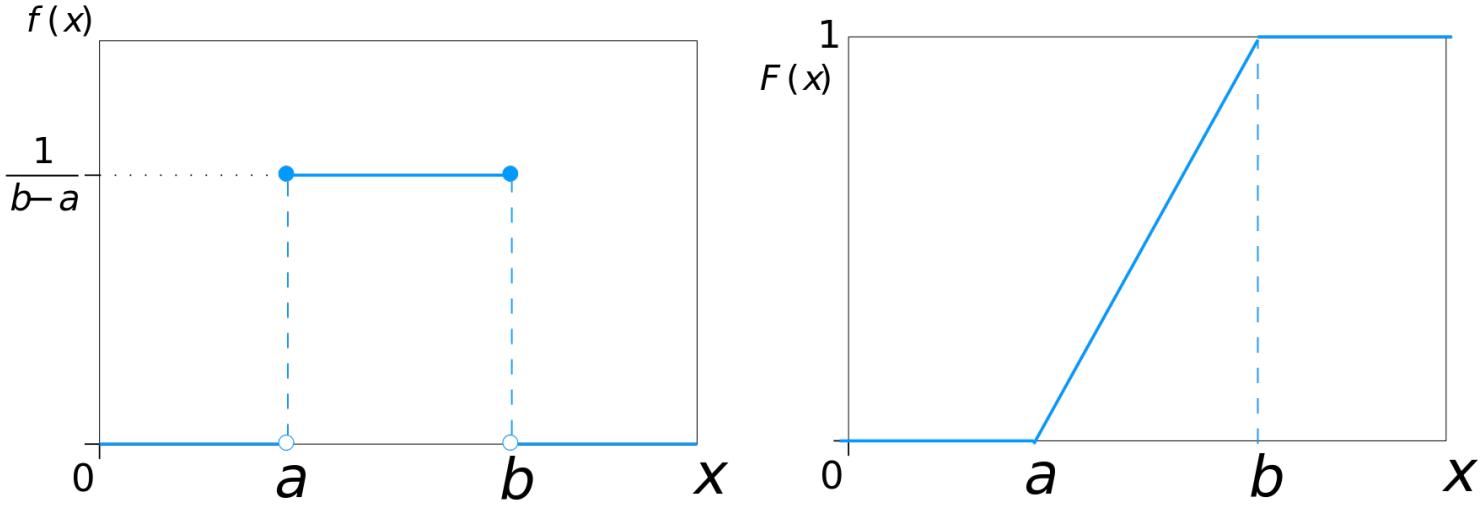
Examples of distributions: continuous uniform distribution





- 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)

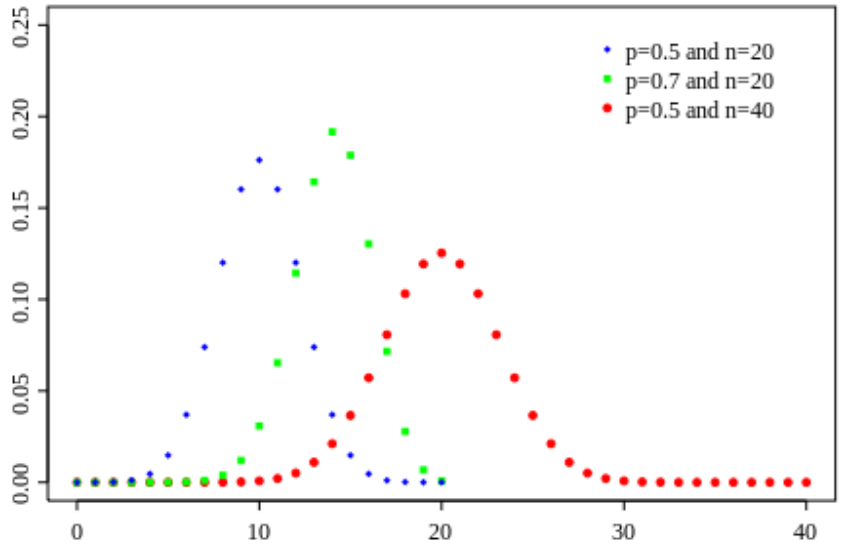


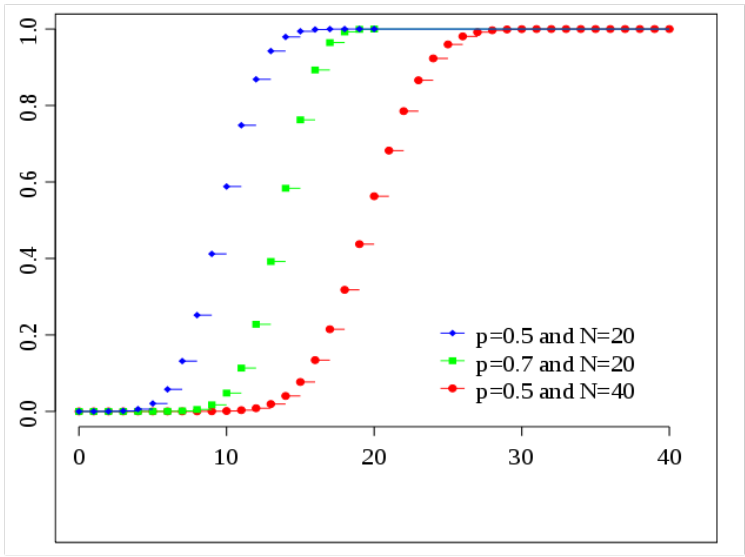
- Mean, Median $\frac{1}{2}(a+b)$, multimodal in (a,b)
- Variance $\frac{1}{12}(b-a)^2$, standard deviation is $\frac{b-a}{2\sqrt{3}}$
- Consequence: if data on an angle (by definition in $[-\pi,+\pi]$) has a standard deviation comparable to $\pi/\sqrt{3}\,\mathrm{rad} \simeq 104^o$, it is possibly completely 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 (occurences 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: $f(k; n, p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- $E[X] = np$, $\mathrm{Var}\{X\} = np(1-p)$
- 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:

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

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

$n \geq 16$ 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 LESS 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

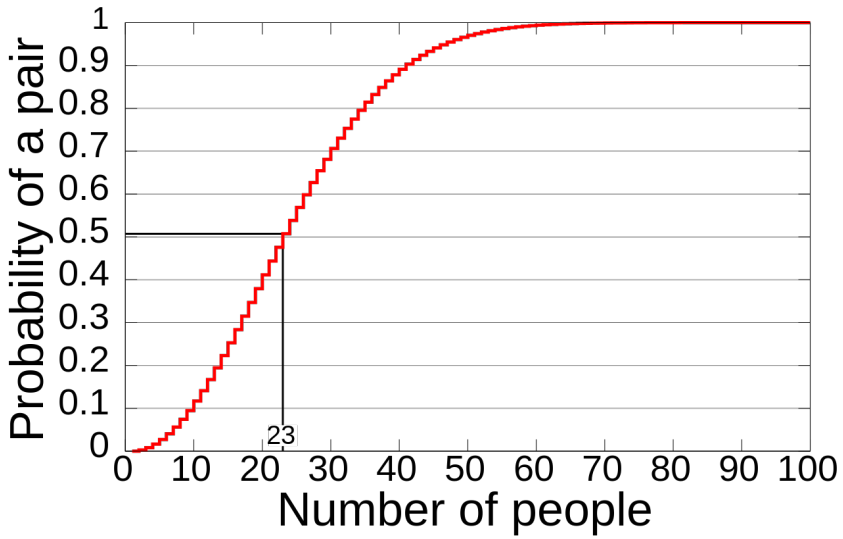
The binomial expected value (and standard deviation) should give you a strong clue something's amiss: $E[X] = np$ is n times the probability of a single event, reminding you it's treating the events as independent. Using the binomial probability distribution is **not** the right approach for this birthday 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: $365 \times 364 \times \ldots \times (365 - n + 1) = 365!/(365 - n)!$.
- 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 overfloat (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 $p(i) = p_A(i) + p_B(i)$:

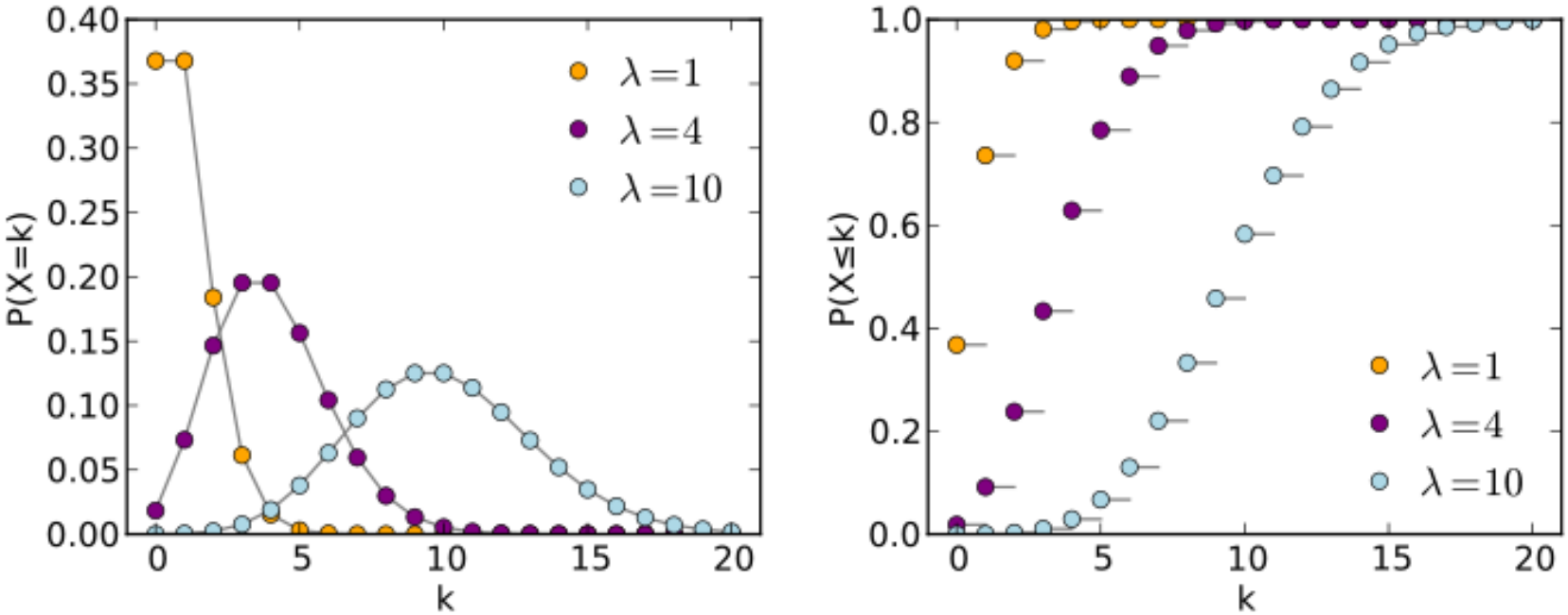
$$p_A(1) = \frac{365}{365.25}, \forall i > 1 : p_A(i) = \frac{365 - i + 1}{365} p_A(i-1)$$

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

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: $F(k;\lambda) = \Pr(X \leq k) = e^{-\lambda} \sum_{i=0}^{k} \frac{\lambda^i}{i!}$
- $\mathrm{E}[X] = \lambda$, $\mathrm{Var}(X) = \lambda$.
- 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 less 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 $= \frac{8}{20000} n$

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

$$n \geq 1733$$

- What is the maximum number of lines I can have my boss read insuring he will see less 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

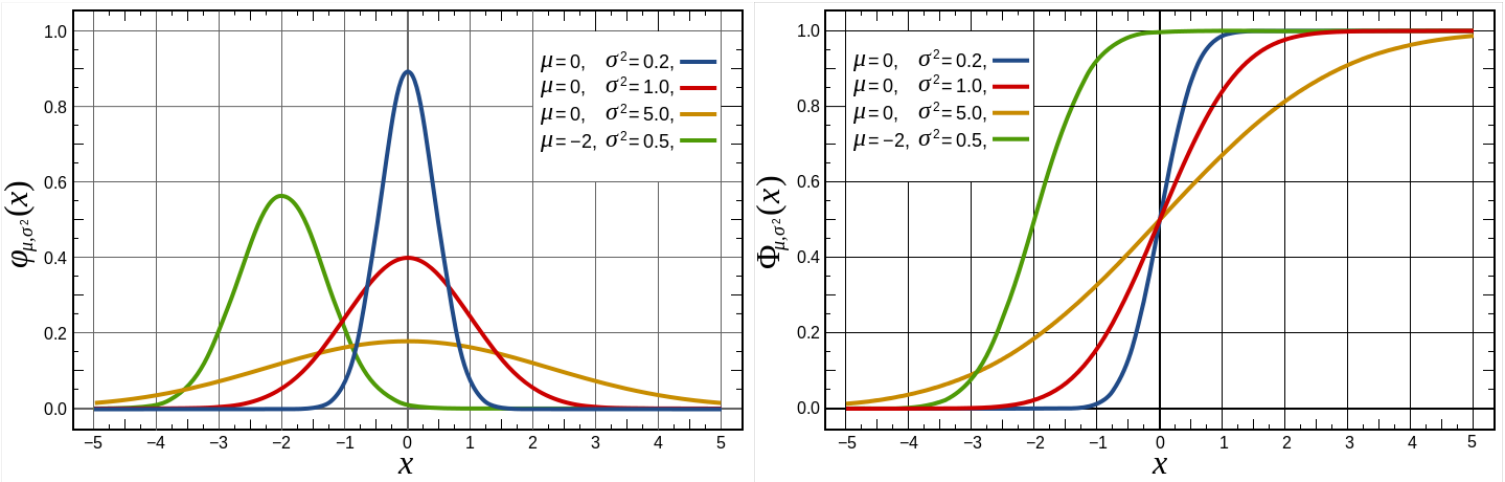
- **Law of rare events**: if $X \sim B(n,p)$, when $n \to \infty$ and $p \to 0$ and $\lim_{n\to\infty} np = \lambda$, then $X \sim P(\lambda)$. $p = \lambda/n$ is the probability of success of each of the k trial.

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

Examples of distributions: Poisson and Binomial (2)

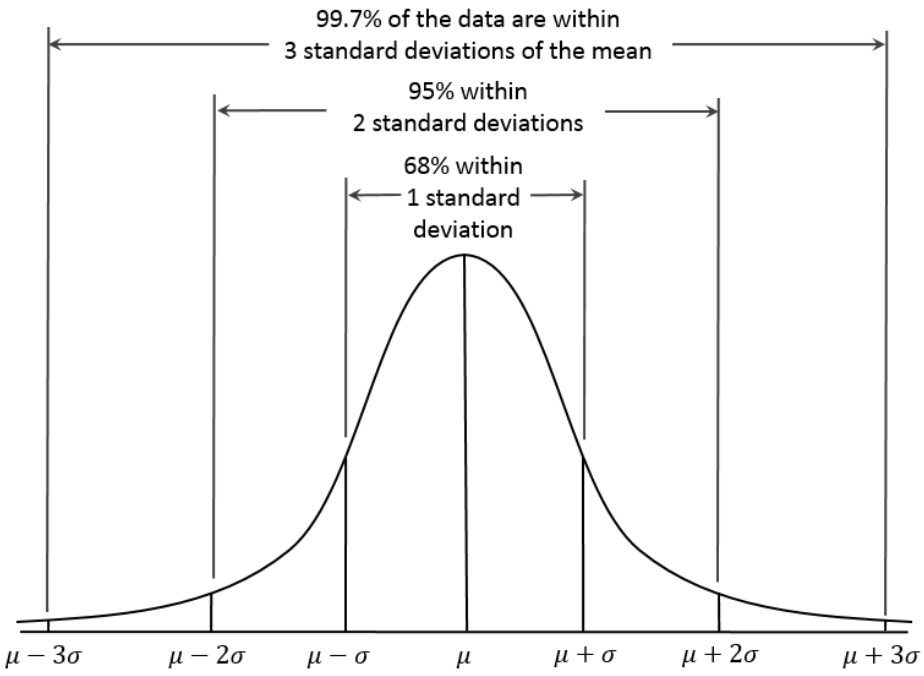
- **Poisson limit theorem**: another name for the law of rare events, if $n \to \infty$ and $p \to 0$ and $np \to \lambda$, then the binomial distribution tends to a Poisson distribution.
- Poisson distribution is the equivalent of a binomial law with a large number of attempts, each with with low probability, 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 $\operatorname{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2}\, dt$
- $E[X] = \text{Mode}(X) = \text{Med}(X) = \mu$, $\text{Var}(x) = \sigma^2$

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:

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

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

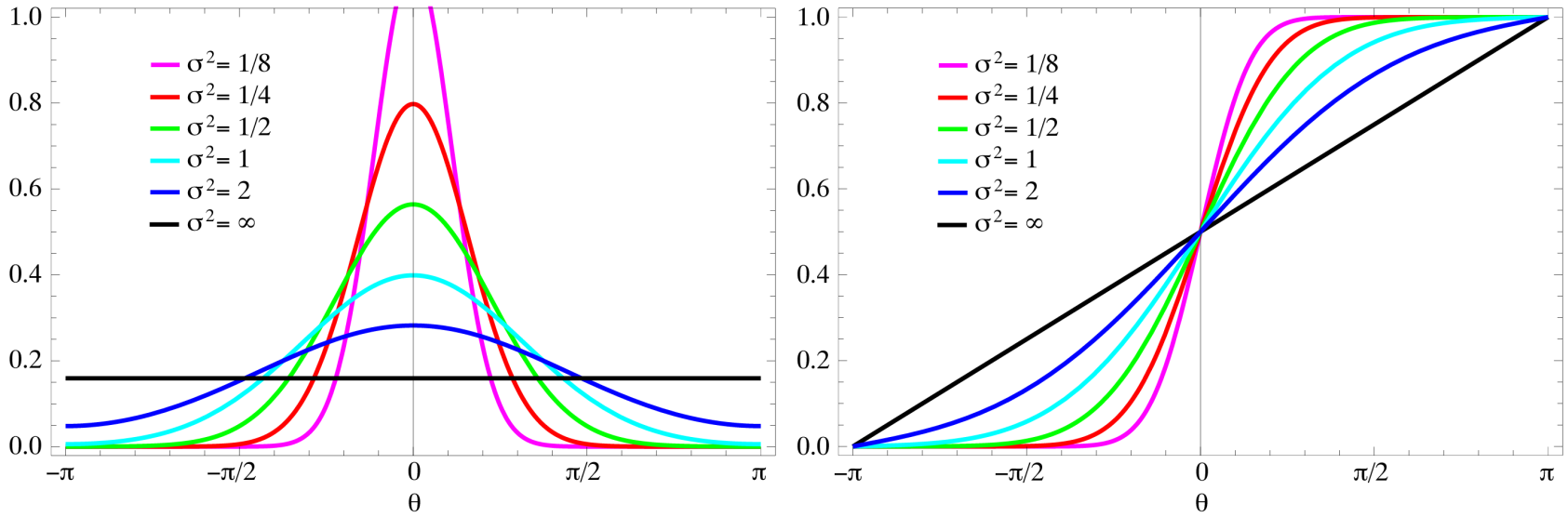
$$\begin{aligned}
\Pr(\mu - \ \sigma \leq x \leq \mu + \ \sigma) &\approx 68.27\% \\
\Pr(\mu - 2\sigma \leq x \leq \mu + 2\sigma) &\approx 95.45\% \\
\Pr(\mu - 3\sigma \leq x \leq \mu + 3\sigma) &\approx 99.73\%
\end{aligned}$$

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 $\mathcal{N}(\mu, \sigma/\sqrt{N})$
- 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 !

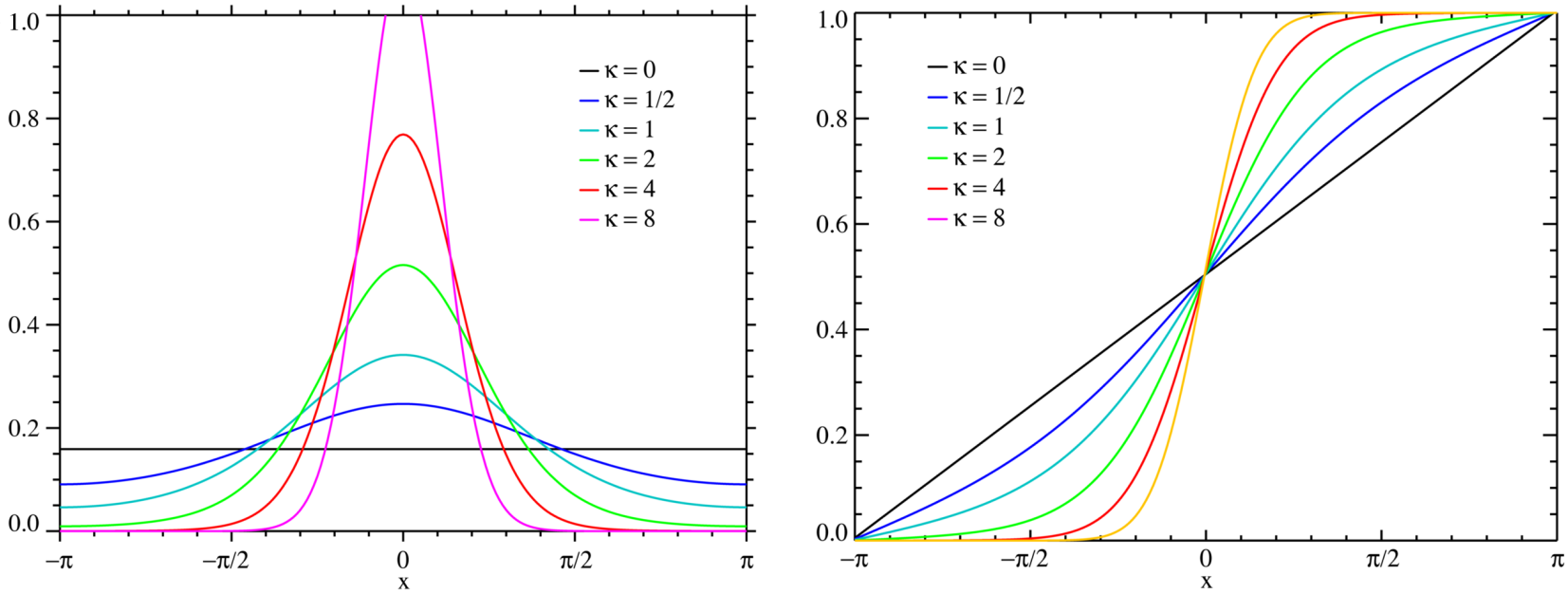
Wrapped normal distribution



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 $\mathrm{Var}\{e^{i\theta}\} = 1 - e^{-\sigma^2}$
- A pain to deal with... No analytic CDF and tricky to manipulate.

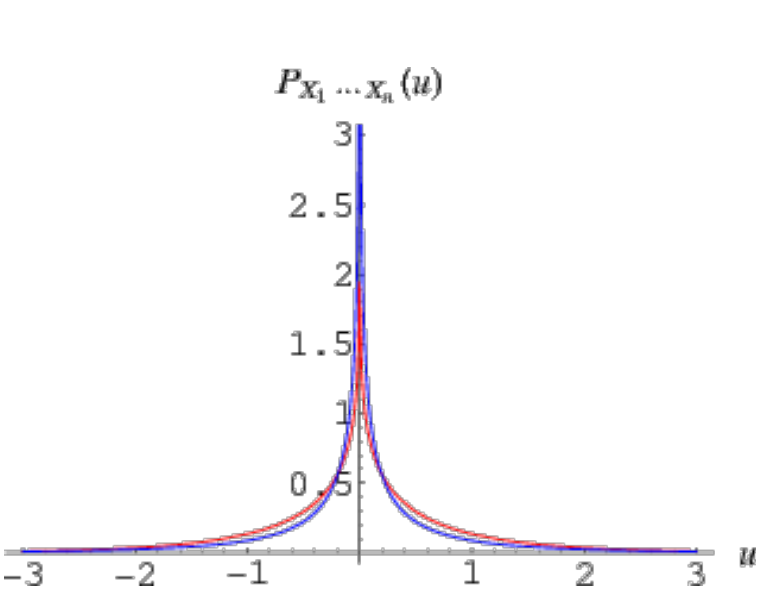
von Mises distribution

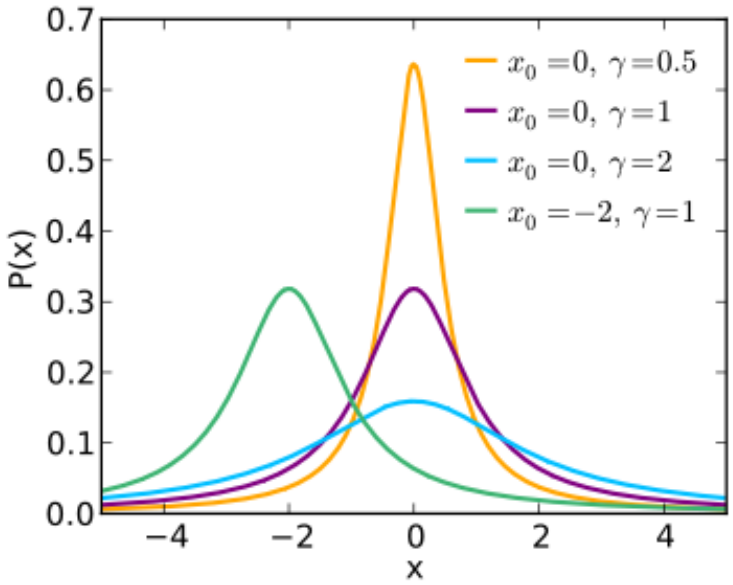


von Mises with $\mu = 0$ and support $[-\pi, +\pi]$

- 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, $\mathrm{Var}\{e^{i\theta}\} = 1 - I_1(\kappa)/I_0(\kappa)$
- Still no analytic CDF.

Combination of normal variables





- If X and Y are independent normallly distributed random variables:
  - $Z = X + Y$ and $Z = X - Y$ are normally distributed.
  - $Z = XY$ follows a **normal product distribution** (above, left).
  - $Z = X/Y$ 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 normallly distributed, X^2 follows a χ^2 distribution.

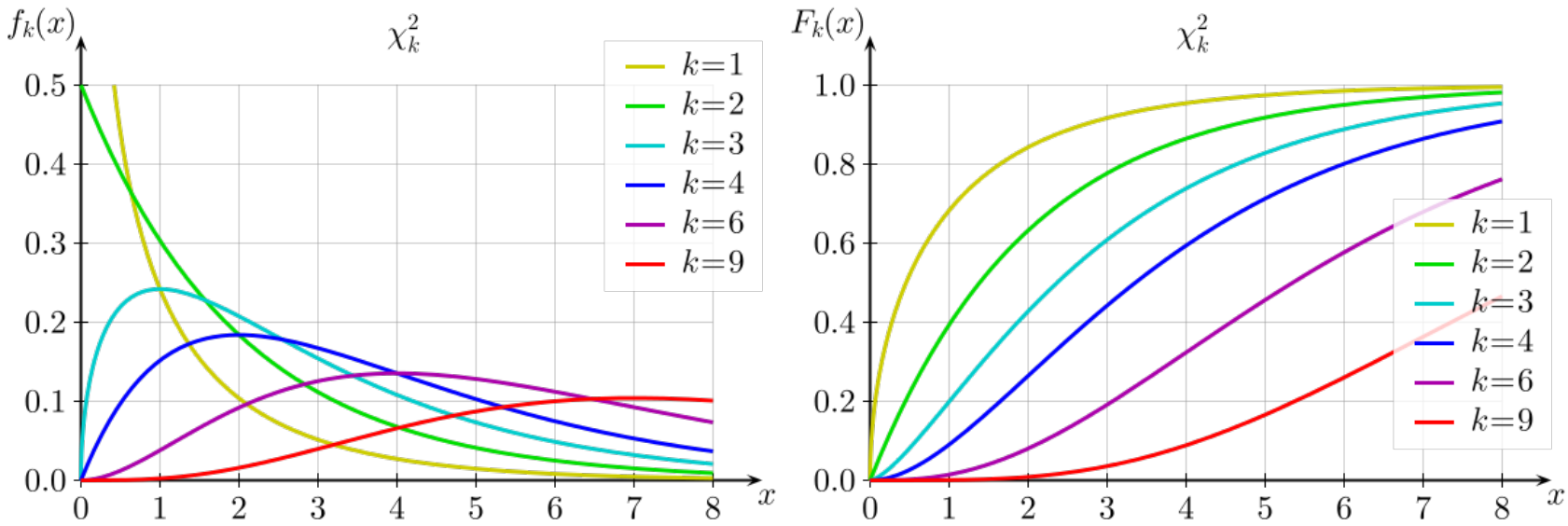
Examples of distributions: χ^2 distribution

- 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: $f_k(x) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$
- CDF: $F_k(x) = \Pr(X \leq x; k) = \frac{1}{\Gamma\left(\frac{k}{2}\right)} \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$
- 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 $\chi^2_r = \chi^2/k \sim 1 \pm \sqrt{2/k} \xrightarrow{k\to\infty} 1$
- For reasonably high k, whenever the correct model is employed to predict μ_i, $\chi^2_r \simeq 1$ (not $\chi^2_r \simeq 0$) 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 $\chi^2_r = \chi^2/N$. 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 $\chi^2_r \leq 1$ and $\chi^2_r > 2$ in each case.
- Put your code with reasonable amount of comments + plots into a single PDF file (LaTeX preferred if you can) and email me with email title "[ASTR/PHYS 8150] Homework 1".

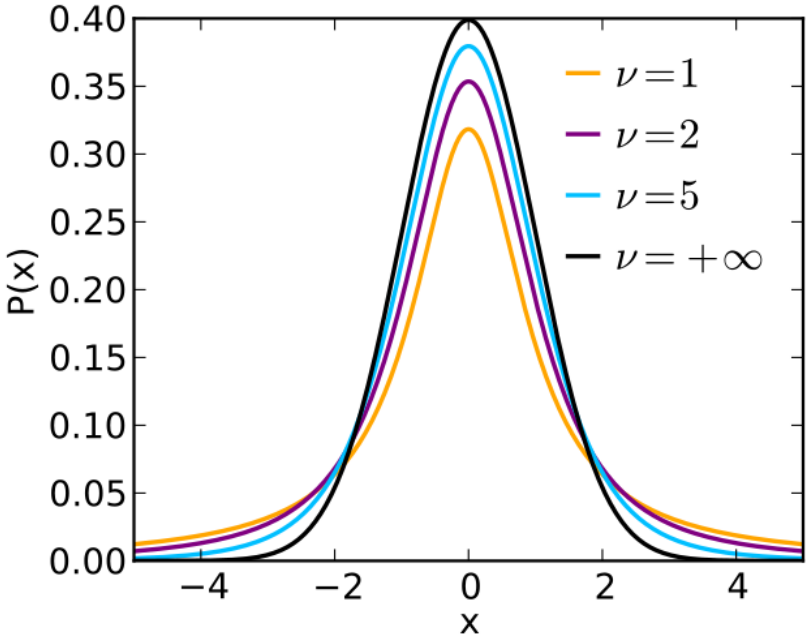
How a sample is approaching a population (1)

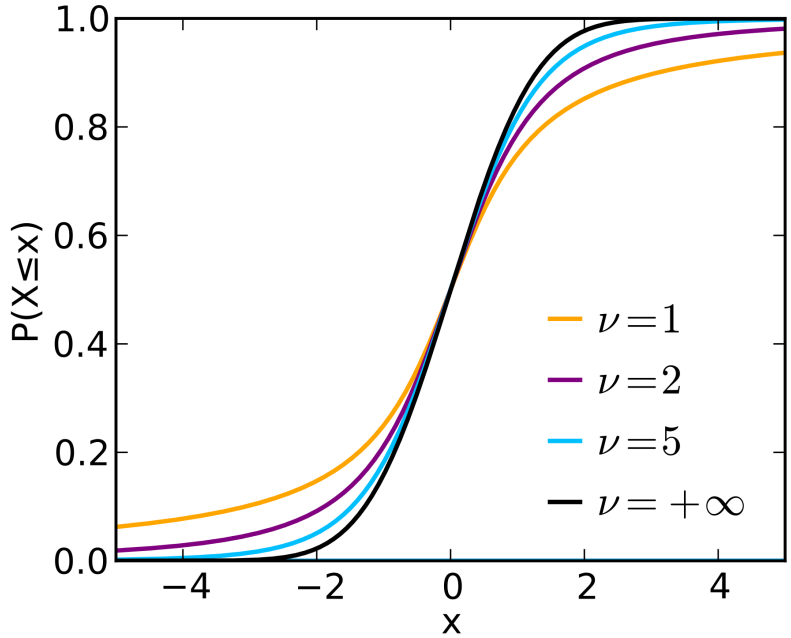
- Let's suppose we made N **independent** observations of a random variable X from a Gaussian-distributed population with unknown parameters $\mathcal{N}(\mu, \sigma^2)$.
- $\bar{X} = \frac{1}{N} \sum_{i=1}^{n} X_i$ is the sample mean.
- $S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})$ is the sample variance.
- A **standard error** of a statistic is the estimated standard deviation of this statistic. The **standard error of the mean** is the standard deviation of the error in the sample mean with respect to the true (population) mean. The standard error of the sample mean is: $\mathrm{SE}(\bar{X}) = \frac{\sigma}{\sqrt{N}} \simeq \frac{S}{\sqrt{N}}$
- Standard error on the sample median $\mathrm{SE}(\mathrm{Med}\{X\}) \simeq \sqrt{\frac{\pi}{2}} \frac{\sigma}{\sqrt{N}}$, the median more subject to sampling fluctuations than the mean.
- Standard error on sample standard deviation $\mathrm{SE}(var(X)) \simeq \frac{\sigma^2}{\sqrt{2(N-1)}}$.

How a sample is approaching a population (2)

- Let's examine the 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 t-distribution.

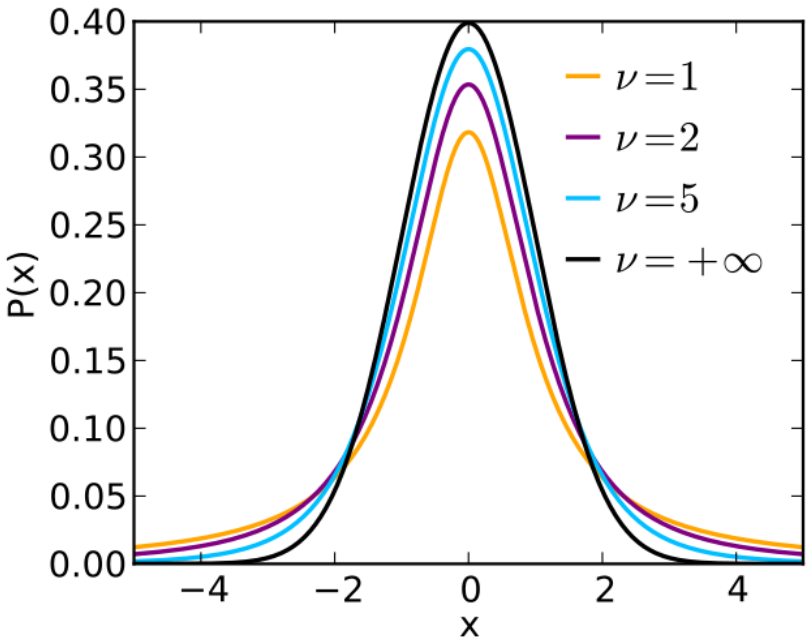
Examples of distributions: Student's t-distribution

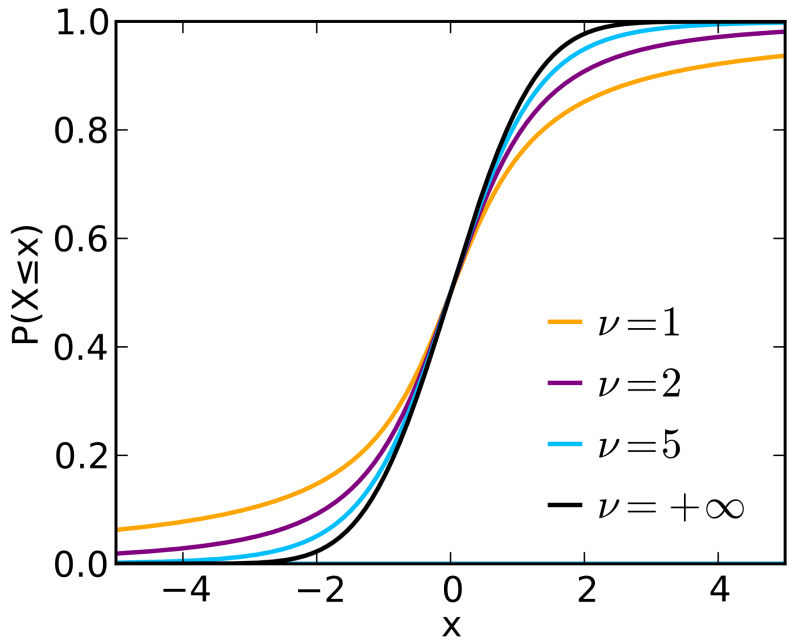




- Notation $X \sim t(\nu)$ where ν is the number of degrees of freedom. Typically $\nu = N - 1$ for N samples/measurements.
- PDF: $f_\nu(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
- CDF: $F_\nu(x)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

Examples of distributions: Student's t-distribution





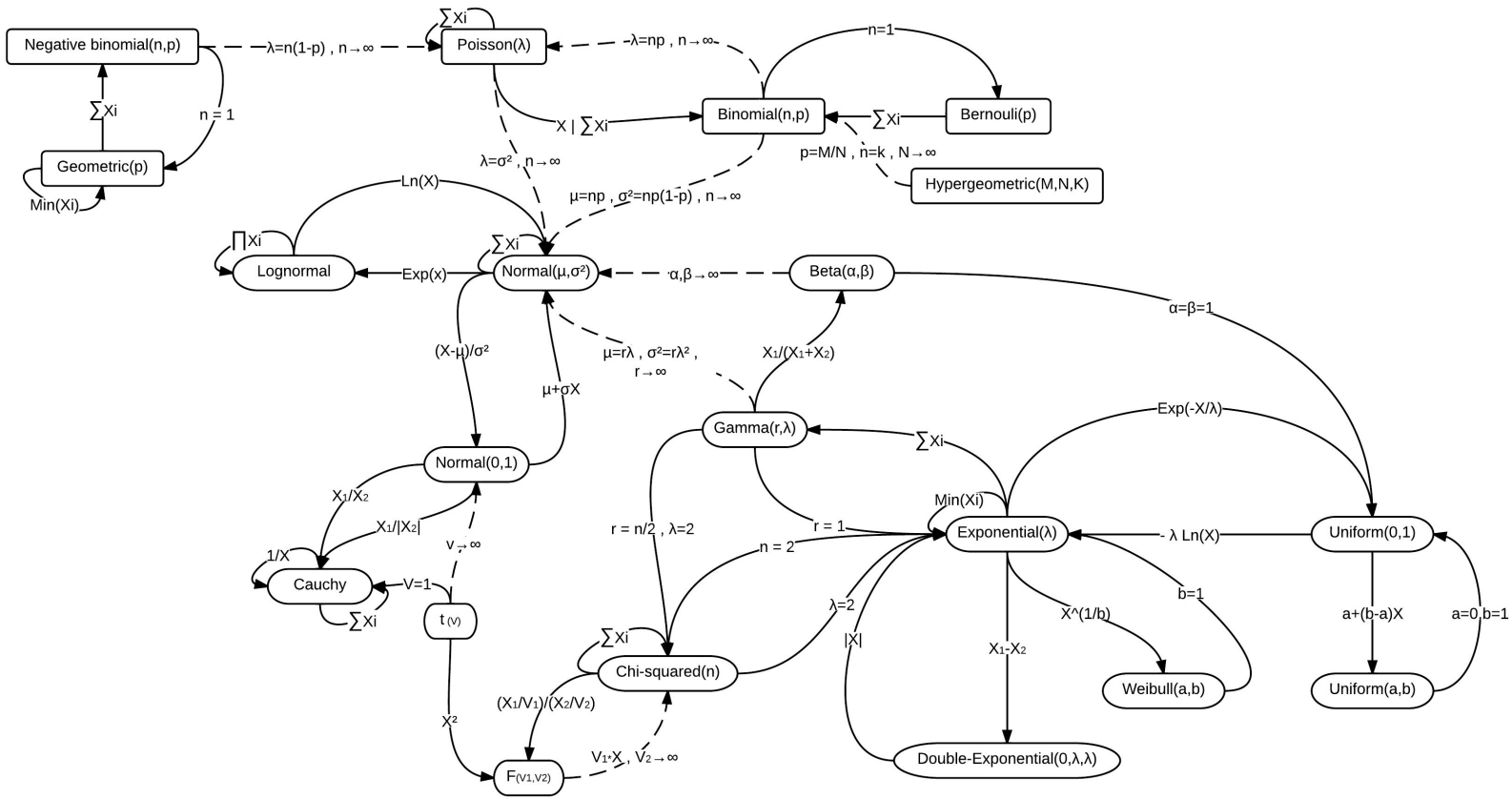
- For small ν, Student's t distribution has heavier tails than the normal distribution. More extreme values are likely to arise. Most often, Student's t distribution is used when the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown.
- For larger ν, this distribution tends to normal (see the $+\infty$ above).

Other distributions

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 computer languages have an implementation of the uniform distribution (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: $U \sim F(X) \rightarrow X \sim F^{-1}(U)$.
- 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 - KS & Kuiper's tests

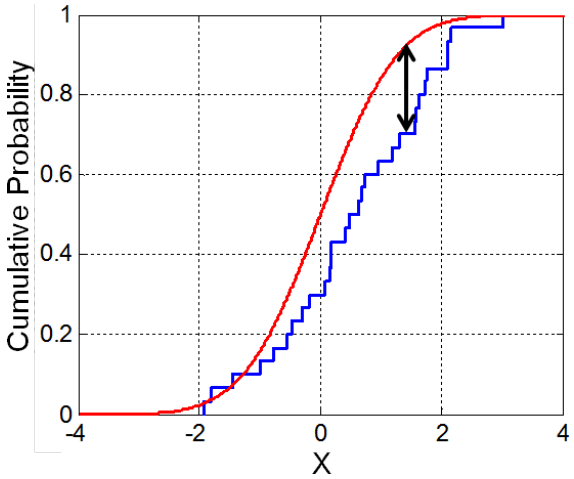
- You want to compare the distributions of two samples, $X_{1,i}$, $i = 1 \dots n_1$ and $X_{2,i}$, $i = 1 \dots n_2$.
- Empirical Distribution Functions: $F_1(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x)$ where I is the indicator function, here equal to 1 if $X_i \leq x$ and equal to 0 otherwise. For samples, $F_1(x) = n_1(x)/n$, $n_1(x)$ 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:

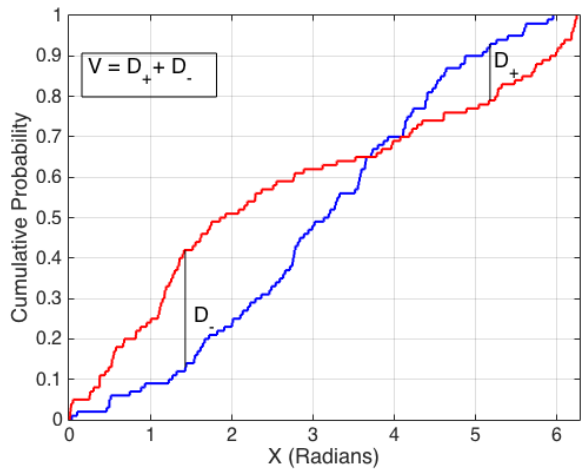
$$D_{\mathrm{KS}} = \max |F_1(X_1) - F_2(X_2)|$$

- Kuiper's test is a variant:

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

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(α) is gotten from tabulated values.
- 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 (derived from KS), Cramer-von Mises/Watson and Anderson-Darling (using quadratic distances), ShapiroWilk's test (specialized to test for Gaussianity).
- 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, f test.
- There is extensive litterature on the benefits/drawback of all these methods.

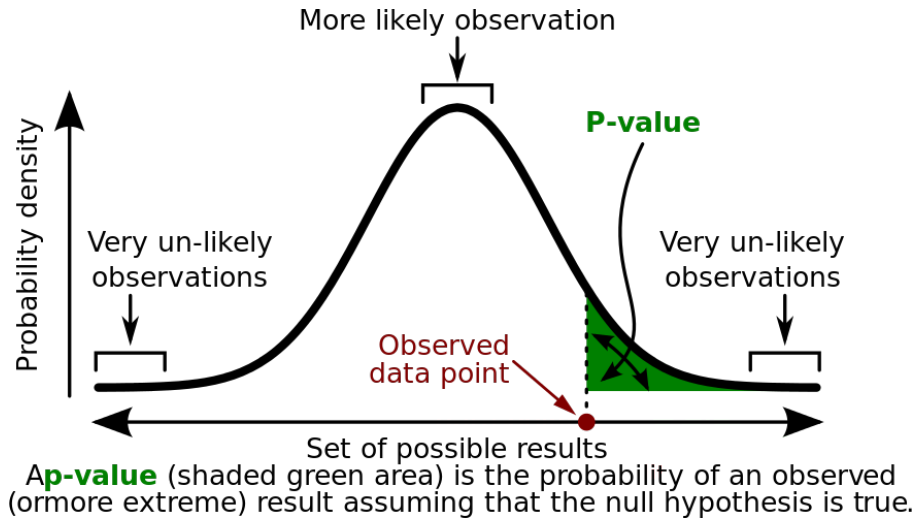
Two fallacies to start thinking about probabilities

- **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 occured", 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**

More on p-value

- 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

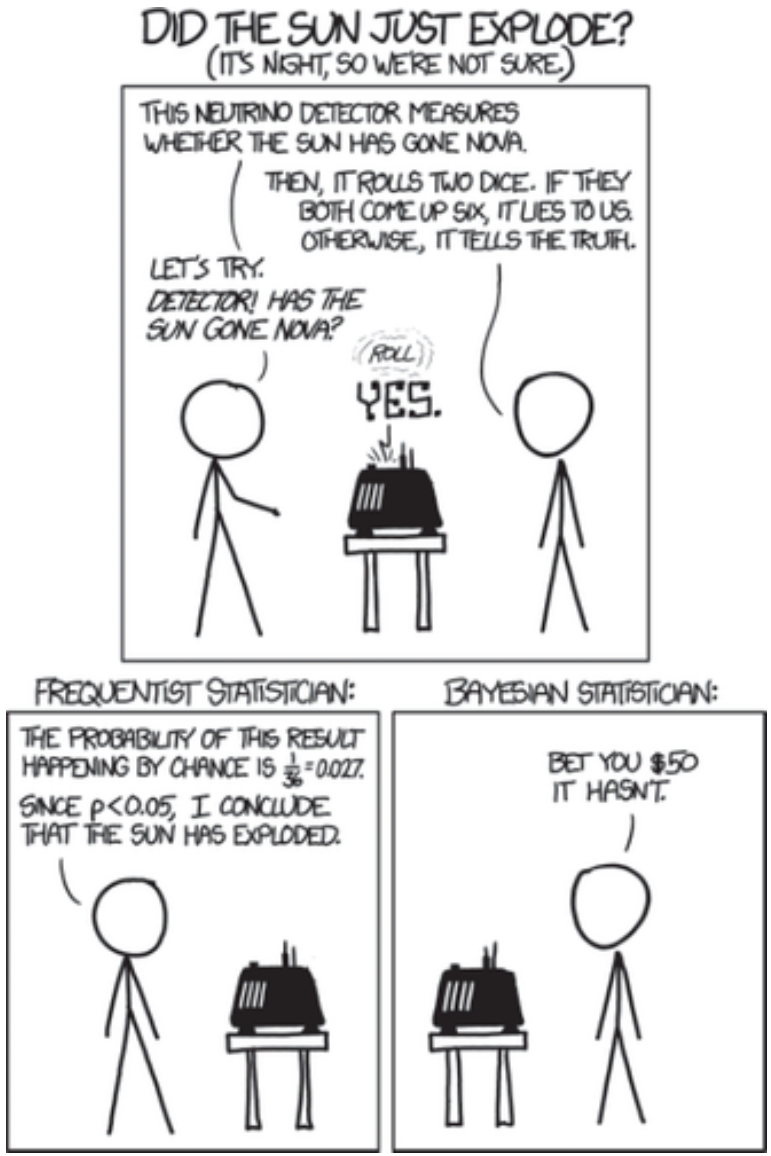
- Example: **Null hypothesis** H_0: a coin is fair, H: coin is not fair/ biased in some way. H and the null hypothesis H_0 are mutually exclusive and have to be the only possibilities. $\Pr(H) \neq \Pr(H_0)$ in general, but 50%-50% here.
- Applied to χ^2: we have a set of data at hand onto which we fitted a model, and found χ^2_0, then
  p-value $= \Pr(\chi^2 \geq \chi^2_0 | H_0) = 1 - \Pr(\chi^2 < \chi^2_0 | H_0)$.
- 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 $X = \{x_1, \dots x_N\}$ with $S = \{\sigma_1, \dots \sigma_N\}$ (heteroskedasticity). We also have a model M based on parameters $\theta = \{\theta_1, \dots \theta_p\}$, predicting values $\mu = \{\mu_1, \dots \mu_N\}$.
- The **likelihood** of θ given the data is equal to the probability of the observed data given those parameter values

$$\mathcal{L}(\theta|X) := \Pr(X|\theta)$$

- The likelihood is a function of θ given X. It is **not** a probability density function (function of X given θ).
- The likelihood is **NOT** the probability that θ are the right ones given X as we will see if Bayes equation. Consequently it is generally unnormalized, i.e. $\int_\theta \mathcal{L}(\theta|X)d\theta \neq 1$, while $\int_X \Pr(X|\theta)dX = 1$.
- Probability and statistical inference are dealing with different problems. Probability theory studies processes modeled by random variables. Statistical inference tries to find models that explain given observations.

Model-fitting: likelihood and χ^2

- For independent data points,

$$\mathcal{L}(\theta|X) = \prod_i^N \Pr(x_i|M)$$

- The **log-likelihood** is

$$\log \mathcal{L}(\theta|X) = \sum_i^N \log \Pr(x_i|M)$$

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

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

Model-fitting and Maximum likelihood

- The most likely values for the model parameters μ are found by maximizing the likelihood.
- Maximizing the likelihood is minimizing the negative log-likelihood:

$$\begin{aligned}\underset{\theta}{\operatorname{argmax}}\,\mathcal{L}(\theta|X) &= \underset{\theta}{\operatorname{argmin}}\,\{-\log\mathcal{L}(\theta|X)\}\\ &= \underset{\theta}{\operatorname{argmin}}\left(\mathrm{cnst}+\frac{1}{2}\sum_{i}^{N}\left(\frac{x_i-\mu_i(\theta)}{\sigma_i}\right)^2\right)\\ &= \underset{\theta}{\operatorname{argmin}}\,\chi^2(\theta)\end{aligned}$$

- 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 x_i, σ_i^2, $i = 1 \ldots N$, x_i know to be normally distributed with variance σ_i^2. E.g. we could want to combine $T = 300 \pm 50K$ and $T = 326 \pm 12K$.

Application of Maximum likelihood: Inverse variance weighting

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

$$\tilde{\mu} = \underset{\mu}{\text{argmin}} \sum_i \left(\frac{x_i - \mu}{\sigma_i} \right)^2 \implies 0 = \sum_i \frac{1}{\sigma_i^2} 2(x_i - \tilde{\mu}) \implies \tilde{\mu} = \frac{\sum_i \frac{x_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}}$$

- The **inverse-variance weighted average** is $\text{Var}\{\tilde{\mu}\} = \frac{1}{\sum_i 1/\sigma_i^2}$ and can be shown to have the least variance among all weighted averages.
- For $T = 300 \pm 50K$ and $T = 326 \pm 12K$,
  $\tilde{\mu} = \frac{300/50^2 + 326/12^2}{1/50^2 + 1/12^2} \simeq 325$ K and $\tilde{\sigma} \simeq \sqrt{\frac{1}{1/50^2 + 1/12^2}} \simeq 12$ 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, $i = 1 \ldots N$, $M_i \cap M_j = \varnothing$, but exhaustive $M_1 \cup \ldots \cup M_N = \mathbb{U}$ then

$$\Pr(D) = \Pr\left((D \cap M_1) \cup \ldots \cup (D \cap M_N)\right) = \sum_i \Pr(D|M_i) \Pr(M_i)$$

- 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

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

- $\Pr(T)$ 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)