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.

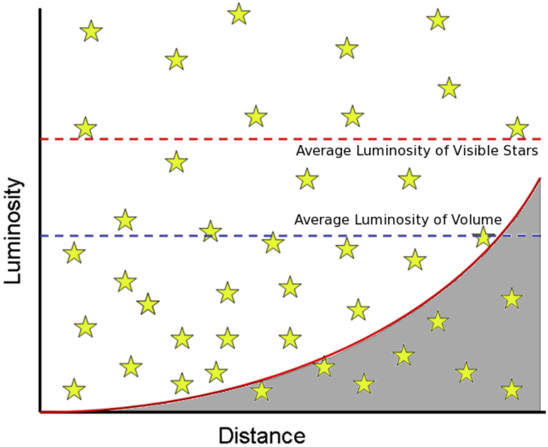
Both are used to state the uncertainties ("errors") in measurements:

- **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

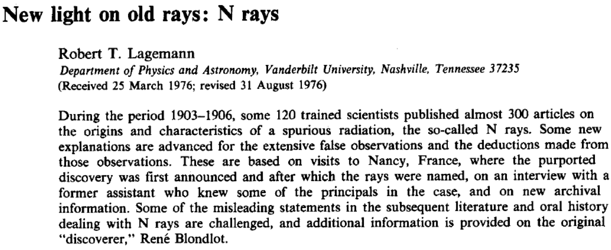
Statistics and probabilities can prevent us from being fooled by physical and psychological biases:

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


Average Luminosity of Visible Stars
Average Luminosity of Volume
Luminosity
Distance

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.

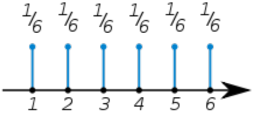
- Small samples: leading to noisy results
- A posteriori: using the same dataset which motivated a hypothesis to

Noise and probabilities

- Data is affected by noise, which prevents from getting to the exact mathematical laws, and behaves as a **random variable**
- Need to beat the noise: understanding the noise means we may get to the exact laws more easily
- 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. 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

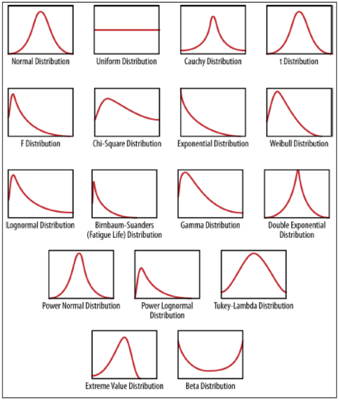


Figure 2-1. A bunch of continuous density functions (aka probability distributions)

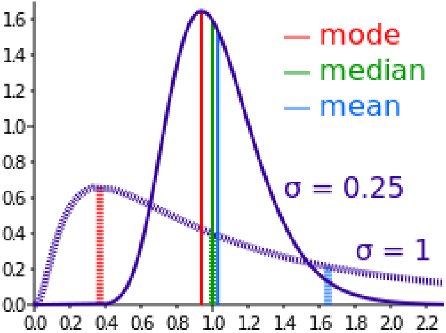
- Continuous case: if X a continuous random variable, the **probability density function** $f_X(X)$ is a continuous probability distribution
- **Support** of a probability function is the set of points where the probability density is not zero-valued
- The probability mass (not density) of getting any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

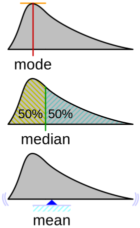
Measures of central tendency - Discrete case

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

- **Mean**: $\bar{X} = \frac{1}{N} \sum_{i=0}^{N} x_i$. Note that $\bar{X}$ is the sample mean, as opposed to the the population mean/**expected value** $E[X]$; this latter often used for long-run/time averages, or Monte Carlo simulations.
- **Mode**: the element that occurs most often in the collection; if not unique, sample is said to be multimodal.
- **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 - Continuous





- Expected value: $\mu = \mathsf{E}[X] = \int x f(x)\,\mathrm{d}x$, 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 $\mathsf{P}(X \leq x) = \frac{1}{2}$ and $\mathsf{P}(X \geq x) = \frac{1}{2}$.

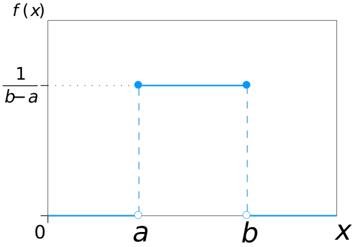
Variance and covariance

- If $\mu = \mathrm{E}[X]$, $\mathrm{Var}(X) = \mathrm{E}\left[(X-\mu)^2\right] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2$
- Continuous: $\mathrm{Var}(X) = \sigma^2 = \int (x-\mu)^2 f(x)\, dx \ = \int x^2 f(x)\, dx \ - \mu^2$
- Biased 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 the true (population) mean is unknown/computed as the sample mean, then the sample variance is a biased estimator: it underestimates the variance by the Bessel's correction factor $(N-1)/N$. **Homework 1 part 1: read on Bessel's correction.**
- Unbiased sample variance: $\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2$

Cumulative distribution functions

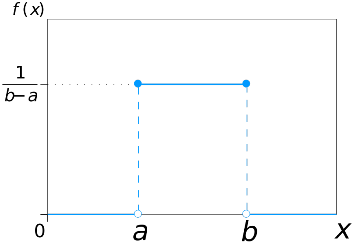
- The **cumulative distribution function** measures the probability that the variable takes a value less than or equal to x
- CDF: $F_X(x) = \mathrm{P}(X \leq x)$
- $\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a)$
- In the continous case: $F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$

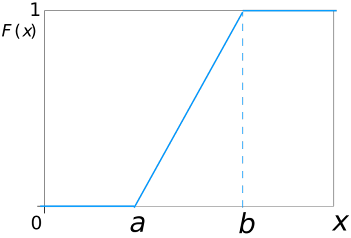
Examples of distributions: continuous uniform distribution



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

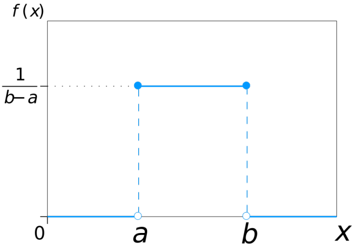
Examples of distributions: continuous uniform distribution

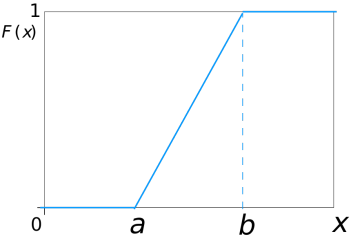




- Support $x \in [a, b]$
- Probability density function $f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$
- CDF $= \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b) \\ 1 & \text{for } x \geq b \end{cases}$

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.

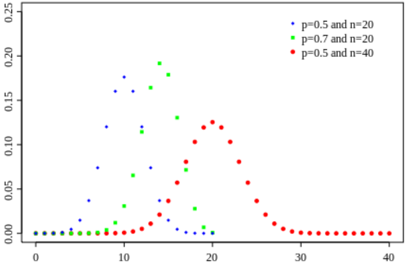
Variance computation and rolling statistics

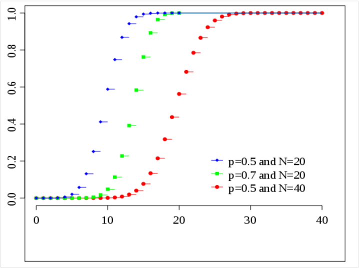
- Moving/rolling/running mean and variance are often used when dealing with streaming data
- Mean subject to issues with summation such as imprecision (research Kahan summation algorithm) or overfloat
- Algorithm 1: first compute the mean, then use it to compute the variance using classic expression. Two pass, and needs access to all the history of the data.
- Algorithm 2: accumulate sums of x_i and x_i^2. One pass only, does not need access to history.
- Welford's method: $(N-1)s_N^2-(N-2)s_{N-1}^2=(x_N-\bar{x}_N)(x_N-\bar{x}_{N-1})$, one pass and needs only updates.
- **Homework 1 part 2**: implement these three methods in Julia using for loops, then compare to native Julia functions for very large arrays (profile with @time or tic() and toc()).

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)
- Typical outcomes: detection or non-detection, belonging or not to a class of objects
- Special cases: tossing biased (Bernouilli) or unbiased coin (Rademacher distribution).

Examples of distributions: Binomial distribution





- Support: number of successes, $k \in 0, 1, 2 \ldots$
- 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

Homework 1 part 3: do the following

- 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 less than 50% right answers (4 or less good 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

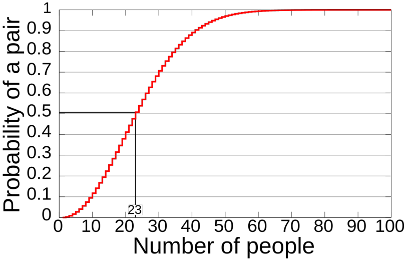
The binomial expected value (and standard deviation) should give you a strong clue something's amiss: $E[X] = np$ is n times the probability of a single event, reminding you it's treating the events as independent. Using the binomial probability distribution is **not** the right approach for this birthday 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 made a single student encounter 25 other students, and compare his birthday 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 $p(i) = p_A(i) + p_B(i)$:

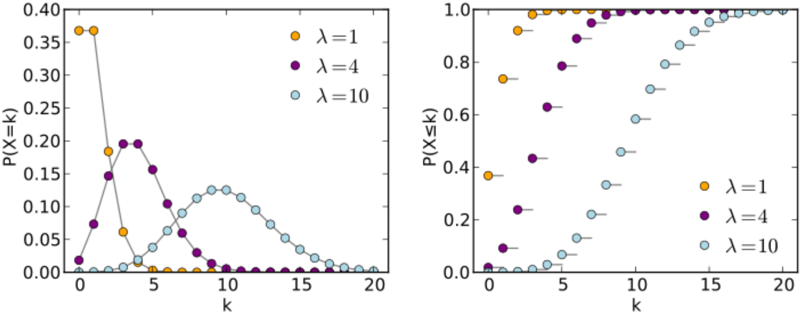
$$p_A(1) = \frac{365}{365.25}, \forall i > 1 : p_A(i) = \frac{365 - i + 1}{365} p_A(i-1)$$

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

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.
- Bad estimate of Poisson error: "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

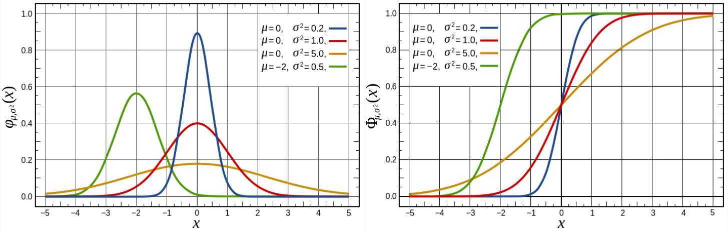
- **Law of rare events**: if $X \sim B(n,p)$, when $n \to \infty$ and $p \to 0$ and $\lim_{n\to\infty} np = \lambda$, then $X \sim P(\lambda)$. $p = \lambda/n$ is the probability of success of each of the k trial.

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

Examples of distributions: Poisson and Binomial (2)

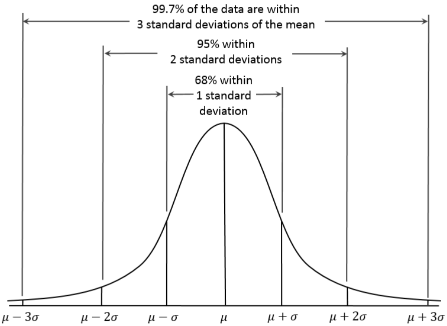
- **Poisson limit theorem**: another name for the law of rare events, if $n \to \infty$ and $p \to 0$ and $np \to \lambda$, then the binomial distribution tends to a Poisson distribution.
- Poisson distribution is the equivalent of a binomial law with a large number of attempts, each with with low probability, 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

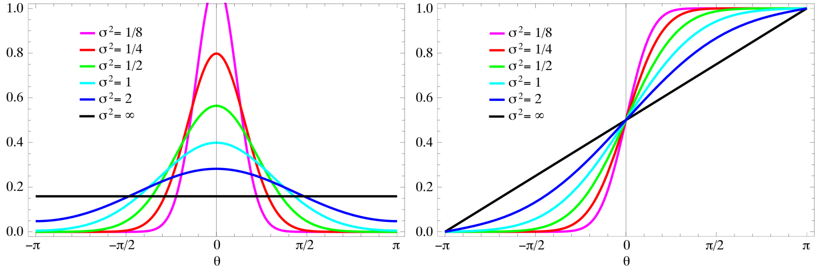
$$\begin{aligned}\Pr(\mu - \ \ \sigma \leq x \leq \mu + \ \ \sigma) &\approx 68.27\%\\ \Pr(\mu - 2\sigma \leq x \leq \mu + 2\sigma) &\approx 95.45\%\\ \Pr(\mu - 3\sigma \leq x \leq \mu + 3\sigma) &\approx 99.73\%\end{aligned}$$

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 !

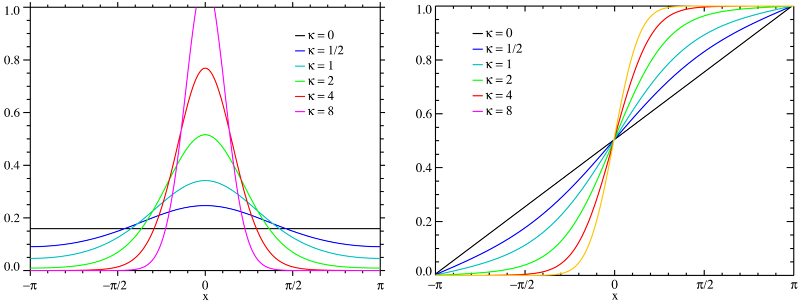
Wrapped normal distribution



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

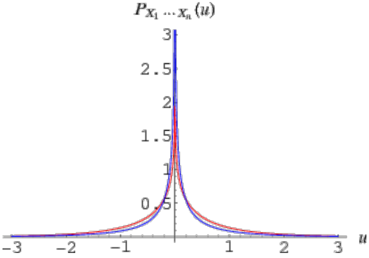
von Mises distribution

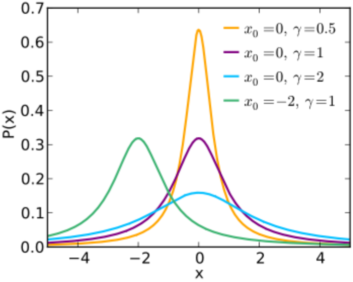


von Mises with $\mu = 0$ and support $[-\pi, +\pi]$

- A close approximation of the wrapped normal distribution PDF: $f_{WN}(\theta; \mu, \kappa) = \frac{e^{\kappa \cos(x-\mu)}}{2\pi I_0(\kappa)}$. $1/\kappa$ and σ^2 are related but not equal !
- Mean $=$ Median $=$ Mode $= \mu$, $\text{Var}\{e^{i\theta}\} = 1 - I_1(\kappa)/I_0(\kappa)$
- Still no analytic CDF.

Combination of normal variables





- If X and Y are independent 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.

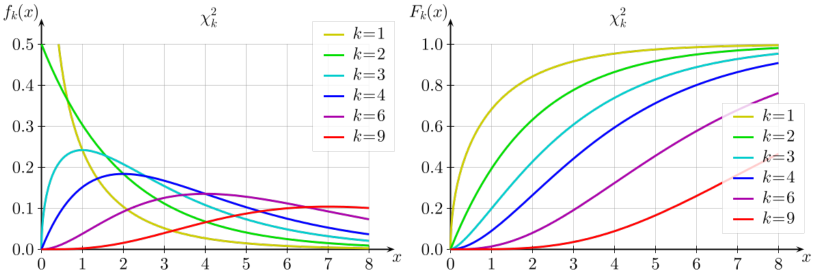
Examples of distributions: χ^2 distribution

- 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.
- This arises in particular when looking at:

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

- Each residual X_i (if the μ_i and σ_i are correct !) should be $X_i \sim \mathcal{N}(0, 1)$.
- For low k you may have bad deviations or outliers bogging down the χ^2. As k increases, you get more data and your estimate of the χ^2 get closer to its population mean.

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

Homework 2

- **Point of this homework**: learn to make publication-ready plots with a new language and learn about reduced χ^2 ...
- **Task**: imagine you're fitting N data samples, each normally distributed, using the χ^2 method. Let's define the reduced-χ^2 as $\chi_r^2 = \chi^2/N$. Plot the probability density distributions $\Pr(\chi_r^2 = x; N)$ on the continuous range $0 \leq x \leq 4$ for $N = 8, 36, 200, 1000$, 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 $\chi_r^2 \leq 1$ for each case.

How a sample is approaching a population

- **Weak Law of Large Numbers/Bernouilli'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.
- Let's suppose we made N **independent** observations of a random variable X from a Gaussian-distributed population with unknown parameters $\mathcal{N}(\mu, \sigma^2)$. The **standard error of the mean** is the standard deviation of the error in the sample mean with respect to the true (population) mean.
- We have $\bar{X} = \frac{1}{N}\sum_i X_i \sim \mathcal{N}(\mu, \sigma^2/N)$, or $\frac{\bar{X}-\mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1)$
- And: $S^2(X) = \frac{\sum_i (X_i - \bar{X})}{N-1} \sim \chi^2(N-1)$.
- $\frac{\bar{X}-\mu}{S/\sqrt{N}} \sim t(N-1)$: 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 $\sqrt{N}$.

Standard Errors on sample statistics (2)

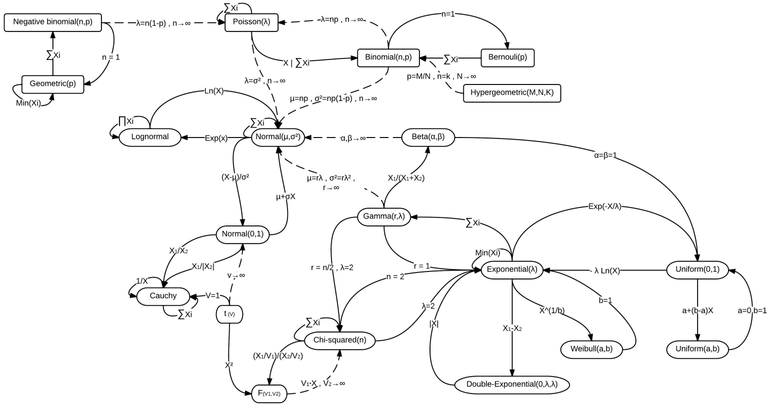
- **Student's t-distribution**: distribution of the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown. Symmetric and bell-shaped, like the normal distribution, but has heavier tails, meaning that it is more prone to producing values that fall far from its mean. Often should use this instead of Normal distribution. In our example, 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.
- A **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 Student t-distribution: $\mathrm{SE}(\bar{X}) = \sqrt{\mathrm{Var}(barX)} = \frac{S}{\sqrt{N}}$.
- Standard error on the sample median $\mathrm{SE}(\mathrm{Med}\{X\}) \simeq \sqrt{\frac{\pi}{2}}\frac{S}{\sqrt{N}}$: median more subject to sampling fluctuations than the mean.
- Standard error on sample standard deviation $\mathrm{SE}(S(X)) \simeq \frac{S}{\sqrt{2(N-1)}}$.

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. Length of time between phone calls, length of time until laptop failure, etc.
- 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

- All useful computer languages have an implementation of the uniform distribution (in Julia: `rand()`).
- To generate random numbers following a given distribution, the analytic expression of the inverse of the CDF then force $F(X)$ to follow a uniform distribution $U = F(X) \rightarrow X = F^{-1}(U)$.
- 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

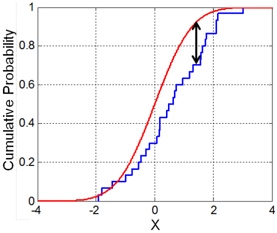
- You want to compare the distributions of two samples, $X_{1,i}$, $i = 1 \ldots n_1$ and $X_{2,i}$, $i = 1 \ldots n_2$.
- Empirical Distribution Functions: $F_1(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x)$ where I is the indicator function, here equal to 1 if $X_i \leq x$ and equal to 0 otherwise.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

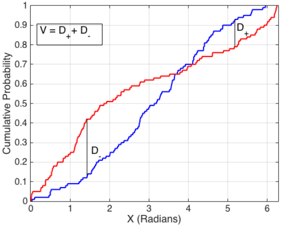
$$D_{\mathrm{KS}} = \max |F_1(X_1) - F_2(X_2)|$$

- 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* 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.

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.

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 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, i.e. **not** a probability density function (function of X given θ), i.e. it is not the probability that the model parameters are the right ones, given the data.
- Consequently the likelihood is generally unnormalized, i.e. $\int\limits_{\theta} \mathcal{L}(\theta|X) d\theta \neq 1$, while $\int\limits_{X} \mathcal{L}(\theta|X) dX = 1$

Model-fitting: likelihood and χ^2

- For independent data points,

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

- The **log-likelihood** is

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

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

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

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:

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

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

χ^2 and reduced χ^2

- χ^2 is used for model-fitting (parameter estimation).

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

- Assuming Normal distribution or Poisson in the high count limit ($\sigma_i^2 \simeq \mu_i$), the residual $R_i = (x_i - \mu_i)/\sigma_i$ is $\mathcal{N}(0,1)$.
- Such a χ^2 follows a χ^2 distribution with degrees of freedom k, i.e. $E\{\chi^2\} = k$ and $\mathrm{Var}\{\chi^2\} = 2k$.
- When fitting a model with p parameters on a sample of N independent data points, litterature often picks $k = N - p$ (we'll see later why this is not an optimal choice).
- Reduced χ^2: $\chi_r^2 = \chi^2/k \sim 1 \pm \sqrt{2/k} \xrightarrow{k \to \infty} 1$

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:

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

- The **inverse-variance weighted average** is $\mathrm{Var}\{\tilde{\mu}\} = \frac{1}{\sum_i 1/\sigma_i^2}$ and can be shown to have the least variance among all weighted averages.