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.

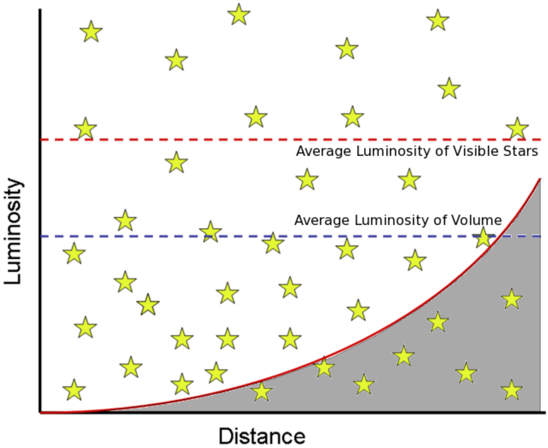
Both are used to state the uncertainties ("errors") in measurements:

- **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

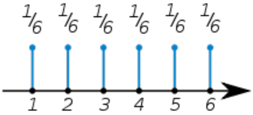
- Data is affected by noise, which prevents from getting to the exact mathematical laws, and behaves as a **random variable**
- Need to beat the noise: understanding the noise means we may get to the exact laws more easily
- Probabilities are the best tool to describe the noise **stochastic processes**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

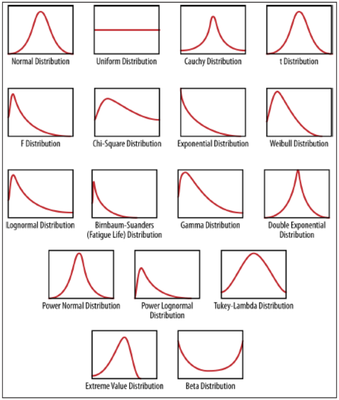


Figure 2-1. A bunch of continuous density functions (aka probability distributions)

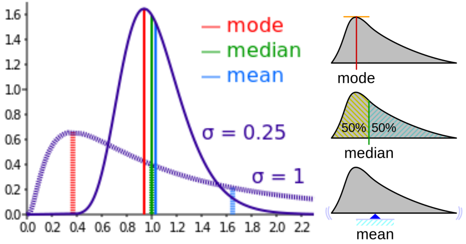
- Continuous case: if X a continuous random variable, the **probability density function** $f_X(X)$ is a continuous probability distribution
- **Support** of a probability function is the set of points where the probability density is not zero-valued
- The probability mass (not density) of getting any 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 - 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.

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: $\mathrm{Var}(X) = \int (x-\mu)^2 f(x)\, dx = \int x^2 f(x)\, dx - \mu^2$

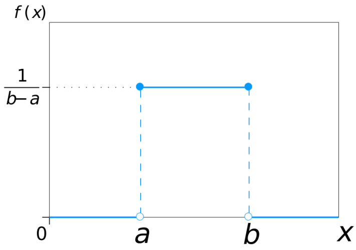
Sample Variance

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

Cumulative distribution functions

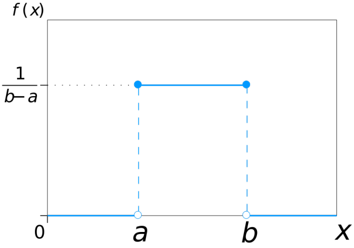
- The **cumulative distribution function** measures the probability that the variable takes a value less than or equal to x
- CDF: $F_X(x) = \mathrm{P}(X \leq x)$
- $\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a)$
- In the continous case: $F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$

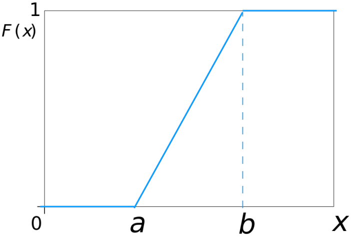
Examples of distributions: continuous uniform distribution



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

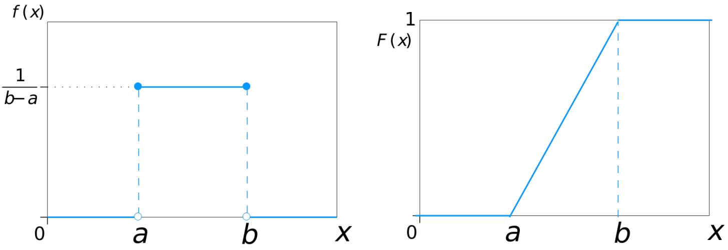
Examples of distributions: continuous uniform distribution





- 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)

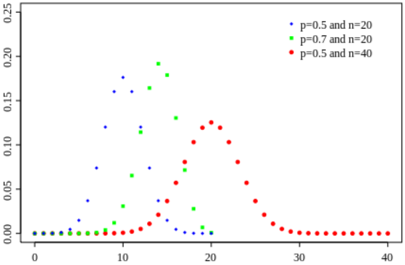


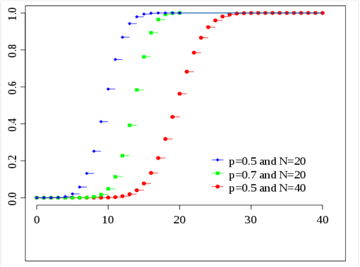
- Mean, Median $\frac{1}{2}(a+b)$, multimodal in (a, b)
- Variance $\frac{1}{12}(b-a)^2$, standard deviation is $\frac{b-a}{2\sqrt{3}}$
- Consequence: if data on an angle (by definition in $[-\pi, +\pi]$) has a standard deviation comparable to $\pi/\sqrt{3}\mathrm{rad} \simeq 104^o$, it is possibly completely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

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 (Bernoulli) or unbiased coin (Rademacher distribution).

Examples of distributions: Binomial distribution





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

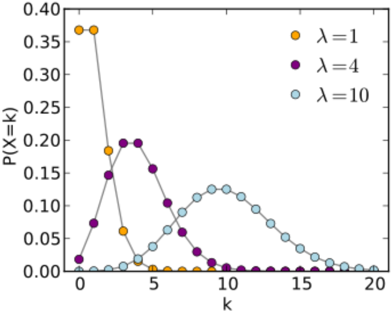
Application of the Binomial distribution

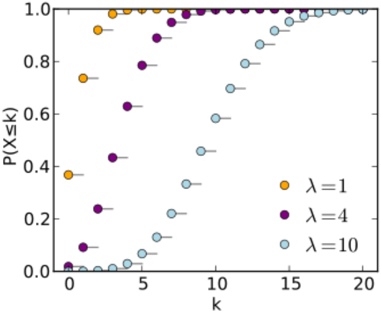
- Professor Henry has a test with 10 multiple choice questions, with 5 possible answers per question. The right answer is unique for each question. Failing the test means getting less than 50% right answers (4 or less good answers in this case). For students that answer questions at random, Professor 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 ?

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 ?